Δ Ι Π Λ Ω Μ Α Τ Ι Κ Η
Ε Ρ Γ Α Σ Ι Α

ΒΑΣΙΛΑΚΗ
ΔΗΜΗΤΡΗ

Α.Μ. 160

CUSTOMER DATA PROFILES AND MACHINE LEARNING

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΔΙΚΤΥΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΑΡΤΑ 2024- 2026

Λίγα Λόγια από τον Συγγραφέα

Η παρούσα **Διπλωματική Εργασία** εκπονήθηκε στο πλαίσιο του **Προγράμματος Μεταπτυχιακών Σπουδών** «**Πληροφορική και Δίκτυα**» του **Τμήματος Πληροφορικής** και **Τηλεπικοινωνιών**, της **Σχολής Θετικών Επιστημών** του **Πανεπιστημίου Ιωαννίνων**. Η εκπόνησή της πραγματοποιήθηκε από τον μεταπτυχιακό φοιτητή **Βασιλάκη Δημήτριο** κατά την περίοδο **2024 - 2026**.

- Βασιλάκης Δημήτριος

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΛΗΡΟΦΟΡΙΚΗΣ & ΔΙΚΤΥΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

Εγκρίθηκε από Τριμελή Εξεταστική Επιτροπή

Άρτα, Απόφαση 145/09-01-2025

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

Επιβλέπων Καθηγητής

Γιαννακέας Νικόλαος

Αναπληρωτής Καθηγητής

Μέλος επιτροπής

Καρβέλης Πέτρος

Αναπληρωτής Καθηγητής

Μέλος επιτροπής

Τσούλος Ιωάννης

Καθηγητής

Δήλωση Μη Λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα μεταπτυχιακή εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

- Βασιλάκης Δημήτριος

Υπογραφή

Ευχαριστίες

Αρχικά, θα ήθελα να εκφράσω τις **Θερμές** μου **Ευχαριστίες** στον **Επιβλέποντα** καθηγητή **κ. Γιαννακέα Νικόλαο**, για την εμπιστοσύνη που έδειξε στο πρόσωπό μου, την πολύτιμη καθοδήγησή του, την υπομονή και την ουσιαστική υποστήριξη που παρείχε σε κάθε στάδιο της έρευνας. Οι παρατηρήσεις του ήταν καθοριστικές για την ολοκλήρωση του παρόντος έργου.

Επιπλέον, ευχαριστώ τα μέλη της τριμελούς επιτροπής, κ. **Καρβέλη Πέτρο** και κ. **Τσούλο Ιωάννη**, για τον χρόνο που διέθεσαν για την αξιολόγηση της εργασίας.

Ένα μεγάλο ευχαριστώ οφείλω στην **Οικογένειά μου**, που στάθηκε δίπλα μου σε όλη τη διάρκεια των σπουδών μου, παρέχοντάς μου ηθική και υλική στήριξη. Χωρίς την υπομονή και την ενθάρρυνσή τους, η προσπάθεια αυτή δεν θα ήταν εφικτή.

Τέλος, θα ήθελα να ευχαριστήσω τους **Φίλους** και **Συμφοιτητές** μου για τις συζητήσεις, την ανταλλαγή απόψεων και τις όμορφες στιγμές που μοιραστήκαμε τη σύντομη αυτή χρονική περίοδο στο Πανεπιστήμιο.

- Βασιλάκης Δημήτριος

Περίληψη

Στη σύγχρονη **ψηφιακή οικονομία**, η ικανότητα των επιχειρήσεων να παρέχουν **εξατομικευμένες εμπειρίες** στους **πελάτες** τους αποτελεί βασικό παράγοντα **ανταγωνιστικότητας**. Ωστόσο, ο κατακερματισμός των δεδομένων σε **απομονωμένα συστήματα** (data silos) δυσχεραίνει την απόκτηση μιας ενιαίας εικόνας για τον **καταναλωτή**. Οι **Πλατφόρμες Δεδομένων Πελατών** (Customer Data Platforms - **CDPs**) αναδύονται ως η λύση για την ενοποίηση της πληροφορίας, ενώ η **Μηχανική Μάθηση** (Machine Learning) προσφέρει τα **εργαλεία** για την αξιοποίησή της.

Σκοπός της παρούσας **διπλωματικής εργασίας** είναι η **μελέτη** και η πρακτική εφαρμογή **αλγορίθμων Μηχανικής Μάθησης** σε περιβάλλον **CDP** για την επίτευξη **αποτελεσματικής τμηματοποίησης**. Η μεθοδολογία ακολούθησε το διεθνές πλαίσιο **CRISP-DM**, χρησιμοποιώντας πραγματικά συναλλακτικά δεδομένα λιανικής ("Online Retail Dataset"). Μέσω της γλώσσας προγραμματισμού **Python** και του λογισμικού **RFM Master Tool v4.5.0** που αναπτύχθηκε για τους σκοπούς της έρευνας, πραγματοποιήθηκε **προεπεξεργασία δεδομένων**, ανάλυση **RFM** (Recency, Frequency, Monetary) και **συγκριτική αξιολόγηση** των **αλγορίθμων** συσταδοποίησης **K-Means, K-Means++, K-Medoids** και **DBSCAN**.

Τα αποτελέσματα της έρευνας κατέδειξαν ότι ο αλγόριθμος **K-Means++** πέτυχε τη βέλτιστη **τεχνική σταθερότητα** και **ταχύτητα εκτέλεσης**. Ωστόσο, για την επιχειρηματική αξιοποίηση **προκρίθηκε ο αλγόριθμος K-Medoids (PAM)**, καθώς η χρήση πραγματικών αντιπροσώπων (medoids) περιόρισε την επίδραση των **ακραίων τιμών** και προσέφερε μια πιο ρεαλιστική και διευρυμένη τμηματοποίηση της αγοράς. Αναδείχθηκαν πέντε στρατηγικές ομάδες πελατών: "**Champions / VIPs**", "**Loyal**", "**New/Potential**", "**At Risk**" και "**Lost**". Αντιθέτως, ο αλγόριθμος **DBSCAN** λειτούργησε αποτελεσματικά ως ανιχνευτής ανωμαλιών, απομονώνοντας τους πελάτες εξαιρετικά υψηλής αξίας (Whales) ως **θόρυβο**. Η εργασία καταλήγει στο συμπέρασμα ότι η ενσωμάτωση αναλυτικών μοντέλων σε υποδομές **CDP** επιτρέπει τη μετάβαση σε μια δεδομενο-κεντρική λήψη αποφάσεων, βελτιστοποιώντας την **εμπειρία** του **πελάτη** και τη **διαχείριση** των **εταιρικών πόρων**.

Λέξεις-Κλειδιά: Customer Data Platforms (CDP), Μηχανική Μάθηση, Προσωποποίηση, K-Means++, K-Medoids, RFM Analysis, Τμηματοποίηση Πελατών.

Abstract

In today's **digital economy**, the ability of businesses to provide **personalized customer experiences** is a key factor in **competitiveness**. However, the fragmentation of data into **isolated systems** (data silos) hinders the acquisition of a unified view of the **consumer**. **Customer Data Platforms** (**CDPs**) are emerging as the solution for information integration, while **Machine Learning** offers the **tools** for its utilization.

The purpose of this **thesis** is the study and practical application of **Machine Learning** algorithms in a **CDP** environment to achieve effective **segmentation**. The methodology followed the international CRISP-DM framework, using real transactional retail data ("Online Retail Dataset"). Through the **Python** programming language and the **RFM Master Tool v4.5.0** software developed for the purposes of this research, data preprocessing, **RFM** (Recency, Frequency, Monetary) analysis, and a **comparative evaluation** of **K-Means**, **K-Means++**, **K-Medoids**, and **DBSCAN** clustering algorithms were performed.

The results of the research indicated that the **K-Means++** algorithm achieved the best technical stability and execution speed. However, for business utilization, the **K-Medoids (PAM)** algorithm was preferred, as the use of real representatives (medoids) limited the impact of **outliers** and offered a more realistic and expanded market segmentation. Five strategic customer groups were identified: "**Champions / VIPs**", "**Loyal**", "**New/Potential**", "**At Risk**", and "**Lost**". Conversely, the **DBSCAN** algorithm functioned effectively as an anomaly detector, isolating extremely high-value customers (Whales) as **noise**. The thesis concludes that integrating analytical models into **CDP** infrastructures allows businesses to transition to data-centric decision-making, optimizing the **customer experience** and the **management** of **corporate resources**.

Keywords: Customer Data Platforms (CDP), Machine Learning, Personalization, K-Means++, K-Medoids, RFM Analysis, Customer Segmentation.

Απόδοση Όρων / Γλωσσάρι

Αγγλικός Όρος	Ελληνική Απόδοση	Σύντομη Επεξήγηση
Algorithm	**Αλγόριθμος**	Σειρά εντολών για την επίλυση ενός προβλήματος ή την εκτέλεση μιας εργασίας.
Artificial Intelligence (AI)	**Τεχνητή Νοημοσύνη**	Η ικανότητα των υπολογιστικών συστημάτων να εκτελούν εργασίες που απαιτούν ανθρώπινη ευφυΐα.
Big Data	**Μαζικά Δεδομένα**	Σύνολα δεδομένων με τεράστιο όγκο, ταχύτητα και ποικιλία που απαιτούν εξειδικευμένες μεθόδους επεξεργασίας.
Churn (Customer Churn)	**Διαρροή Πελατών**	Το ποσοστό των πελατών που διακόπτουν τη σχέση τους με μια επιχείρηση σε μια δεδομένη χρονική περίοδο.
Clustering	**Συσταδοποίηση / Ομαδοποίηση**	Τεχνική μηχανικής μάθησης για την ομαδοποίηση δεδομένων με παρόμοια χαρακτηριστικά (π.χ. K-Means).
Customer Data Platform (CDP)	**Πλατφόρμα Δεδομένων Πελατών**	Λογισμικό που συγκεντρώνει και ενοποιεί δεδομένα πελατών από πολλαπλές πηγές σε ένα ενιαίο προφίλ.
Data Silos	**Σιλό Δεδομένων**	Απομονωμένα σύνολα δεδομένων που δεν είναι προσβάσιμα ή διασυνδεδεμένα με άλλα συστήματα του οργανισμού.
Data Warehouse	**Αποθήκη Δεδομένων**	Κεντρικό αποθετήριο δεδομένων που χρησιμοποιείται για ανάλυση και αναφορές.
DBSCAN	**-**	Αλγόριθμος συσταδοποίησης βασισμένος στην πυκνότητα (Density-Based Spatial Clustering of Applications with Noise).
Elbow Method	**Μέθοδος του Αγκώνα**	Ευριστική μέθοδος για τον προσδιορισμό του βέλτιστου αριθμού συστάδων (k) στον αλγόριθμο K-Means.
Feature Engineering	**Μηχανική Χαρακτηριστικών**	Η διαδικασία επιλογής και μετασχηματισμού μεταβλητών για τη βελτίωση της απόδοσης των μοντέλων ML.
Hyperparameters	**Υπερ-παράμετροι**	Ρυθμίσεις ενός αλγορίθμου που καθορίζονται πριν την εκπαίδευση (π.χ. το k στον K-Means ή το $Epsilon$ στον DBSCAN).
Identity Resolution	**Ταυτοποίηση Πελάτη**	Η διαδικασία σύνδεσης διαφορετικών αναγνωριστικών (emails, cookies) σε ένα μοναδικό προφίλ πελάτη.
K-Means	**-**	Αλγόριθμος συσταδοποίησης που χωρίζει τα δεδομένα σε k ομάδες ελαχιστοποιώντας την απόσταση από τα κέντρα τους.

K-Medoids (PAM)	**-**	Αλγόριθμος συσταδοποίησης που χρησιμοποιεί πραγματικά σημεία (medoids) ως κέντρα, προσφέροντας ανθεκτικότητα σε ακραίες τιμές.
Label Switching	**Αναδιάταξη Ετικετών**	Το φαινόμενο όπου οι αριθμητικές ετικέτες των συστάδων αλλάζουν μεταξύ διαφορετικών εκτελέσεων, χωρίς να αλλάζει η δομή τους.
Log Transformation	**Λογαριθμικός Μετασχηματισμός**	Στατιστική τεχνική για την εξομάλυνση της ασυμμετρίας των δεδομένων και τη συμπίεση των ακραίων τιμών.
Machine Learning (ML)	**Μηχανική Μάθηση**	Υποκλάδος της Τεχνητής Νοημοσύνης που επιτρέπει στα συστήματα να μαθαίνουν από δεδομένα.
Noise	**Θόρυβος**	Δεδομένα που δεν ακολουθούν το πρότυπο ή είναι εσφαλμένα (σημαντικό στον DBSCAN).
Normalization	**Κανονικοποίηση**	Η διαδικασία κλιμάκωσης των δεδομένων ώστε να έχουν κοινό εύρος τιμών (π.χ. Z-Score).
Omnichannel	**Πανκαναλική (Προσέγγιση)**	Στρατηγική που παρέχει ενοποιημένη εμπειρία πελάτη σε όλα τα κανάλια, φυσικά και ψηφιακά.
Outliers	**Ακραίες Τιμές**	Παρατηρήσεις που διαφέρουν σημαντικά από τις υπόλοιπες τιμές του συνόλου δεδομένων.
Personalization	**Προσωποποίηση**	Η προσαρμογή προϊόντων ή υπηρεσιών στις ατομικές προτιμήσεις του κάθε πελάτη.
Predictive Modeling	**Προβλεπτική Μοντελοποίηση**	Χρήση στατιστικών δεδομένων για την πρόβλεψη μελλοντικών αποτελεσμάτων.
Recency, Frequency, Monetary (RFM)	**Προσφατότητα, Συχνότητα, Αξία**	Μοντέλο ανάλυσης πελατών βάσει του πότε αγόρασαν τελευταία, πόσο συχνά αγοράζουν και πόσα ξοδεύουν.
Recommendation System	**Σύστημα Συστάσεων**	Αλγόριθμος που προτείνει σχετικά προϊόντα ή περιεχόμενο στους χρήστες.
Segmentation	**Τμηματοποίηση**	Η διαδικασία διαίρεσης της αγοράς σε διακριτές ομάδες πελατών με κοινά χαρακτηριστικά.
Silhouette Score	**Συντελεστής Σιλουέτας**	Μετρική αξιολόγησης της ποιότητας της συσταδοποίησης (πόσο καλά διαχωρισμένες είναι οι ομάδες).
Single Customer View (SCV)	**Ενιαία Εικόνα Πελάτη**	Η συγκεντρωτική και ακριβής αναπαράσταση όλων των δεδομένων ενός πελάτη σε ένα μέρος.
Standardization (Z-Score)	**Τυποποίηση**	Μετασχηματισμός των δεδομένων ώστε να έχουν μέση τιμή 0 και τυπική απόκλιση 1, καθιστώντας τις μεταβλητές συγκρίσιμες.
Supervised Learning	**Εποπτευόμενη Μάθηση**	Μάθηση όπου ο αλγόριθμος εκπαιδεύεται με δεδομένα που έχουν γνωστές ετικέτες (labels).

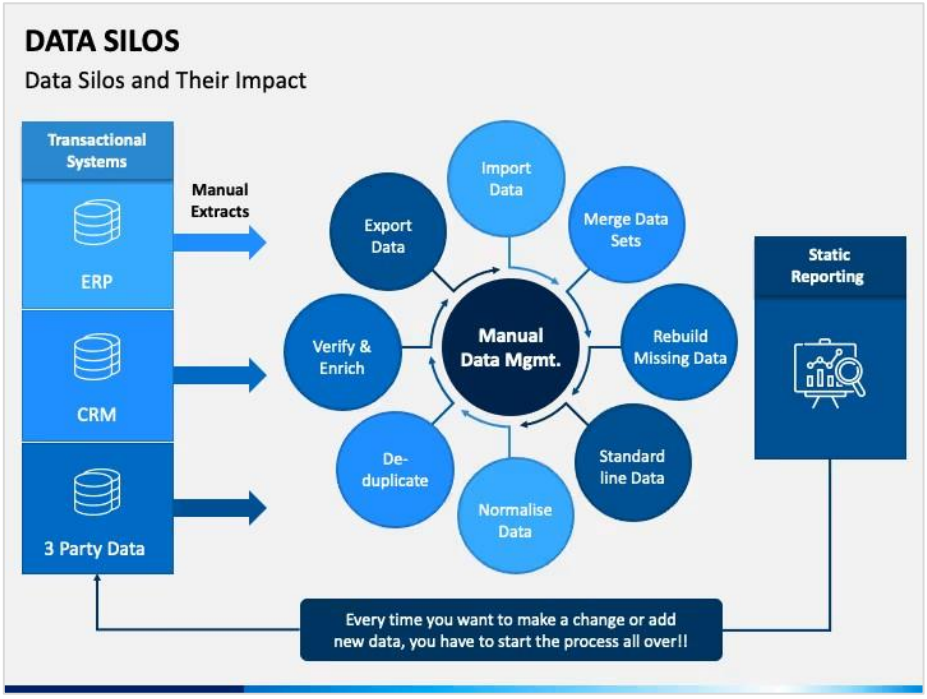
Unsupervised Learning	**Μη Εποπτευόμενη Μάθηση**	Μάθηση όπου ο αλγόριθμος ανακαλύπτει μοτίβα σε δεδομένα χωρίς ετικέτες (π.χ. Clustering).

Πίνακας Περιεχομένων

1.Εισαγωγή

1.1. CDPs και η Σημασία της Προσωποποίησης στις Σύγχρονες Επιχειρήσεις

Στη σημερινή **Ψηφιακή Εποχή**, ο **Όγκος των Δεδομένων** που παράγονται από την αλληλεπίδραση των καταναλωτών με τις επιχειρήσεις αυξάνεται εκθετικά. Οι σύγχρονοι οργανισμοί καλούνται να διαχειριστούν δεδομένα από **Πολλαπλά Κανάλια (Omnichannel)**, όπως ιστοσελίδες, κινητές εφαρμογές, φυσικά καταστήματα και μέσα κοινωνικής δικτύωσης. Ωστόσο, η βασική πρόκληση δεν έγκειται πλέον στη **Συλλογή των Δεδομένων**, αλλά στην **Ενοποίηση** και την **Αξιοποίησή** τους. Σύμφωνα με έρευνες, ένα μεγάλο ποσοστό των επιχειρηματικών δεδομένων παραμένει κατακερματισμένο σε **Απομονωμένα Σιλό (Data Silos)**, καθιστώντας αδύνατη την απόκτηση μιας **Ολοκληρωμένης Εικόνας για τον Πελάτη [Gartner, 2020]** [Εικόνα 1.1].



Εικόνα 1.1. Why Silos are Problematic & How to Fix Them

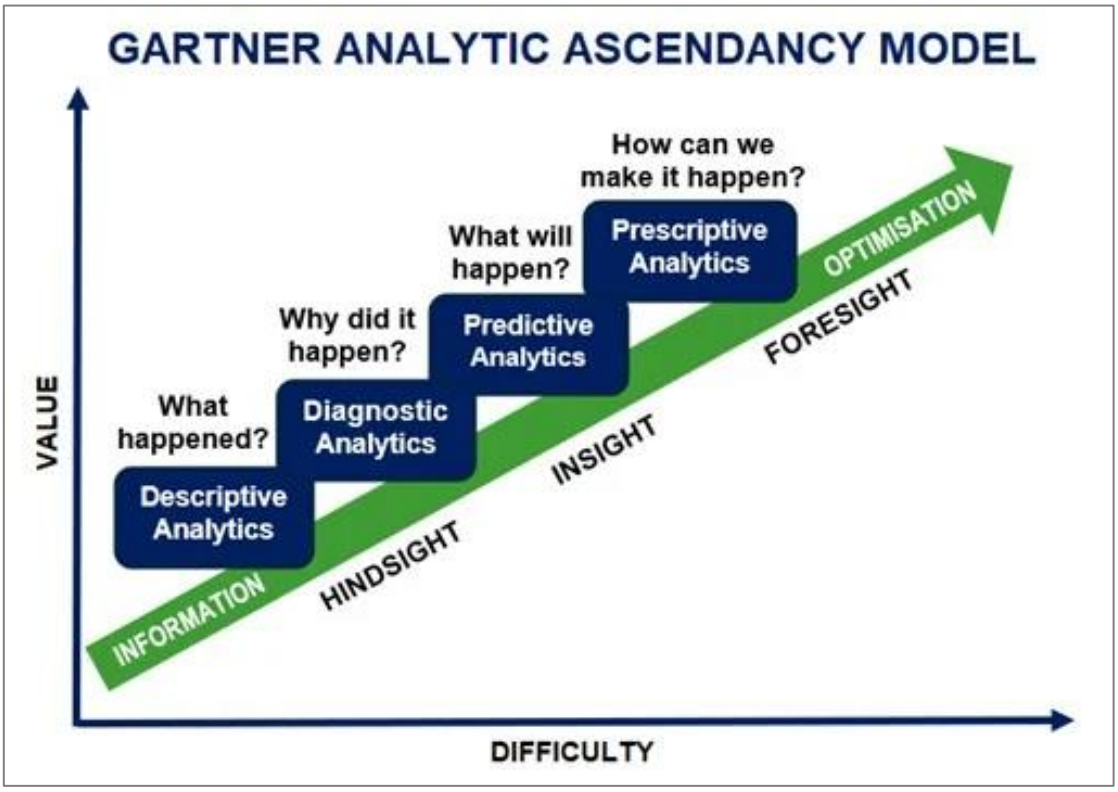
Σε αυτό το πλαίσιο, οι **Πλατφόρμες Δεδομένων Πελατών (Customer Data Platforms - CDPs)** έχουν αναδειχθεί ως μια κρίσιμη τεχνολογική λύση. Το CDP Institute ορίζει το CDP ως «ένα **Πακεταρισμένο Λογισμικό** που δημιουργεί μια **Μόνιμη, Ενιαία Βάση Δεδομένων Πελατών**, η οποία είναι προσβάσιμη σε άλλα συστήματα» **[Raab, 2016]**. Σε αντίθεση με τα παραδοσιακά **Συστήματα CRM ή DMP**, τα CDPs έχουν τη δυνατότητα να απορροφούν δεδομένα από οποιαδήποτε πηγή, να τα καθαρίζουν και να τα συνδέουν με **Μοναδικά Προφίλ Πελατών (Single Customer View - SCV)** σε **Πραγματικό Χρόνο [Zahay et al., 2019]**.

Η ανάγκη για αυτή την ενοποίηση οδηγείται από την απαίτηση για **Προσωποποίηση (Personalization)**. Η **Προσωποποίηση** δεν αποτελεί πλέον απλώς ένα **Ανταγωνιστικό Πλεονέκτημα**, αλλά μια βασική **Προσδοκία των Καταναλωτών**. Μελέτες της McKinsey δείχνουν ότι το **71%** των καταναλωτών περιμένουν από τις εταιρείες να προσφέρουν **Εξατομικευμένες Αλληλεπιδράσεις**, ενώ το **76%** εκφράζει δυσαρέσκεια όταν αυτή η προσδοκία δεν εκπληρώνεται **[McKinsey & Company, 2021]**. Η ικανότητα μιας επιχείρησης να προσφέρει το σωστό μήνυμα, στον σωστό χρήστη, τη σωστή στιγμή, εξαρτάται άμεσα από την **Ποιότητα** και την **Ενοποίηση των Δεδομένων** που παρέχει ένα CDP.

1.2. Η Συμβολή της Μηχανικής Μάθησης στην Προσωποποίηση μέσω CDPs

Οφέλη και Προκλήσεις.

Ενώ το **CDP** επιλύει το πρόβλημα της **Συλλογής** και **Ενοποίησης Δεδομένων**, η **Μηχανική Μάθηση (Machine Learning - ML)** αποτελεί τον καταλύτη που μετατρέπει αυτά τα δεδομένα σε **Δράση**. Οι παραδοσιακές **Μέθοδοι Τμηματοποίησης (Segmentation)** βασίζονταν συχνά σε **Στατικούς Κανόνες (Rules-based)**, οι οποίοι είναι δύσκολο να κλιμακωθούν και αδυνατούν να εντοπίσουν πολύπλοκα, **Μη Γραμμικά Μοτίβα Συμπεριφοράς [Verma et al., 2021]** [**Εικόνα 1.2**].



Εικόνα 1.2. Gartner Analytic Ascendancy Model

Η ενσωμάτωση **Αλγορίθμων ML** πάνω σε **Υποδομές CDP** επιτρέπει τη μετάβαση από την **Περιγραφική Ανάλυση (Descriptive Analytics - τι συνέβη)** στην **Προβλεπτική (Predictive - τι θα συμβεί)** και την **Κανονιστική (Prescriptive - τι πρέπει να κάνουμε)**. Τα κύρια οφέλη αυτής της μετάβασης περιλαμβάνουν:

- **Προηγμένη Τμηματοποίηση (Advanced Segmentation): Αλγόριθμοι Συσταδοποίησης (Clustering)**, όπως ο **K-Means** και ο **DBSCAN**, δύνανται να εντοπίσουν αυτόματα **Μικρο-τμήματα Πελατών** με βάση τη **Συμπεριφορά** τους, τα οποία δεν θα ήταν ορατά με την ανθρώπινη παρατήρηση **[Xu & Wunsch, 2005]**.
- **Προσωποποίηση σε Πραγματικό Χρόνο:** Μέσω **Συστημάτων Συστάσεων (Recommender Systems)**, οι επιχειρήσεις μπορούν να προβλέψουν το **Επόμενο Προϊόν** που είναι πιθανό να αγοράσει ένας χρήστης, αυξάνοντας τον **Ρυθμό Μετατροπής (Conversion Rate)** και τη **Δια Βίου Αξία του Πελάτη (Customer Lifetime Value - CLV) [Portugal et al., 2018]**.

- **Πρόβλεψη Αποχώρησης (Churn Prediction): Μοντέλα Ταξινόμησης** μπορούν να εντοπίσουν εγκαίρως πελάτες που διατρέχουν **Υψηλό Κίνδυνο Διακοπής της Συνεργασίας**, επιτρέποντας τη λήψη **Προληπτικών Ενεργειών Μάρκετινγκ** για τη διατήρησή τους.

Ωστόσο, η εφαρμογή αυτών των τεχνικών συνοδεύεται από σημαντικές προκλήσεις. Η **Ποιότητα των Δεδομένων (Data Quality)** είναι πρωταρχικής σημασίας, καθώς ο **Θόρυβος** ή οι **Ελλιπείς Εγγραφές** μπορούν να οδηγήσουν σε **Λανθασμένες Προβλέψεις**, επιβεβαιώνοντας την αρχή **"Garbage In, Garbage Out" [Ghasemaghaei, 2019]**. Επιπλέον, ζητήματα **Προστασίας Προσωπικών Δεδομένων** και **Κανονιστικής Συμμόρφωσης** (όπως ο **GDPR** στην Ευρώπη) επιβάλλουν τον προσεκτικό σχεδιασμό των αλγορίθμων, ώστε να διασφαλίζεται η **Διαφάνεια (Explainability)** και η **Ηθική Χρήση** των στοιχείων των χρηστών **[Voigt & Von dem Bussche, 2017]**.

1.3. Σημασία και Συνεισφορά της Διπλωματικής Εργασίας.

Επικαιρότητα και Ερευνητική Αξία στην Επιστήμη και τη Βιομηχανία

Παρόλο που η διεθνής βιβλιογραφία αναγνωρίζει τη **Θεωρητική Αξία** των **CDPs** και της **Μηχανικής Μάθησης [Verma et al., 2021; Gartner, 2020]**, παρατηρείται έλλειψη **Πρακτικών Μελετών** που να περιγράφουν αναλυτικά τη διαδικασία υλοποίησης ενός **End-to-End Pipeline Προσωποποίησης**, ειδικά όσον αφορά τη διαχείριση συγκεκριμένων **Datasets** και την **Παραμετροποίηση Αλγορίθμων**.

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι διττός: αφενός να αναλύσει το **Θεωρητικό Υπόβαθρο** των CDPs και αφετέρου να εφαρμόσει πρακτικά **Τεχνικές Μηχανικής Μάθησης** για την επίλυση προβλημάτων προσωποποίησης. Συγκεκριμένα, η εργασία εστιάζει στα εξής **Ερευνητικά Ερωτήματα** και **Στόχους**:

- **Αξιολόγηση Αλγορίθμων Συσταδοποίησης:** Πώς συμπεριφέρονται διαφορετικοί αλγόριθμοι (**K-Means**, **K-Medoids**, **DBSCAN**) σε πραγματικά ή συνθετικά **Δεδομένα Συμπεριφοράς Πελατών**; Ποιος αλγόριθμος αποδίδει καλύτερα στη

δημιουργία διακριτών **Τμημάτων Πελατών (Customer Segments) [Xu & Wunsch, 2005]**;

- **Πρακτική Υλοποίηση Pipeline:** Η ανάπτυξη και τεκμηρίωση μιας διαδικασίας σε γλώσσα **Python**, η οποία περιλαμβάνει την **Προεπεξεργασία Δεδομένων (Data Cleaning, Normalization)**, την **Εκπαίδευση Μοντέλων** και την **Αξιολόγηση των Αποτελεσμάτων** μέσω μετρικών όπως το **Silhouette Score**.
- **Επιχειρηματική Αξιοποίηση:** Η σύνδεση των τεχνικών αποτελεσμάτων με **Επιχειρηματικές Δράσεις**, αποδεικνύοντας πώς η **Εξαγωγή Γνώσης** από ένα CDP μπορεί να οδηγήσει σε **Στοχευμένες Στρατηγικές Μάρκετινγκ [McKinsey & Company, 2021]**.

Η **Σημασία της Εργασίας** έγκειται στο γεγονός ότι γεφυρώνει το χάσμα μεταξύ της **Θεωρίας Διαχείρισης Δεδομένων** και της **Εφαρμοσμένης Επιστήμης Δεδομένων**. Μέσα από την ανάλυση συγκεκριμένων datasets και την παράθεση κώδικα, η εργασία προσφέρει έναν **Πρακτικό Οδηγό** για την υλοποίηση **Μηχανισμών Ευφυούς Προσωποποίησης (Intelligent Personalization Mechanisms)**, συμβάλλοντας τόσο στην **Ακαδημαϊκή Έρευνα** όσο και στην **Πρακτική Εφαρμογή στη Βιομηχανία [Zahay et al., 2019]**.

1.4. Δομή της Εργασίας

Σύντομη Περιγραφή Ενοτήτων

Η παρούσα **Διπλωματική Εργασία** διαρθρώνεται σε **Εννέα Κεφάλαια**, τα οποία ακολουθούν τη **Μεθοδολογική Ροή** από τον **Ορισμό** του **Προβλήματος**, στη **Θεωρητική Ανάλυση**, την **Πρακτική Εφαρμογή** και τέλος την **Αξιολόγηση**. Συγκεκριμένα:

- **Κεφάλαιο 1: Εισαγωγή.** Παρουσιάζεται το **Γενικό Πλαίσιο** του **Προβλήματος**, η **Αναγκαιότητα** των **CDPs** για την **Ενοποίηση** των **Εταιρικών Δεδομένων** και ο **Ρόλος** της **Μηχανικής Μάθησης** στην **Προσωποποίηση**. Διατυπώνονται ο **Σκοπός** της **Εργασίας** και τα **Ερευνητικά Ερωτήματα** που καλείται να απαντήσει η **Μελέτη**.
- **Κεφάλαιο 2: Θεωρητικό Υπόβαθρο.** Αναλύεται η **Έννοια**, η **Αρχιτεκτονική** και η **Εξέλιξη των Customer Data Platforms (CDPs)** από τα **Παραδοσιακά**

Συστήματα CRM. Στη συνέχεια, παρουσιάζονται οι **Βασικές Έννοιες της Μηχανικής Μάθησης** και εξετάζονται αναλυτικά οι **Αλγόριθμοι Συσταδοποίησης** που θα χρησιμοποιηθούν (**K-Means**, **K-Medoids**, **DBSCAN**), συμπεριλαμβανομένης της **Μαθηματικής τους Θεμελίωσης** και **Παραδειγμάτων Εφαρμογής** τους στη σύγχρονη ζωή.

- **Κεφάλαιο 3: Λογοτεχνική Ανασκόπηση.** Επισκοπείται η **Σχετική Βιβλιογραφία** για την **Προσωποποίηση** και τη **Χρήση Αλγορίθμων ML** στο **Μάρκετινγκ**. Εντοπίζονται τα **Κενά στην Υφιστάμενη Έρευνα**, τεκμηριώνοντας την **Ανάγκη για μια Συγκριτική Μελέτη** σε **Περιβάλλον CDP**.
- **Κεφάλαιο 4: Μεθοδολογία.** Περιγράφεται το **Ερευνητικό Πλαίσιο (CRISP-DM)** και η **Διαδικασία Υλοποίησης**. Αναλύονται τα **Χαρακτηριστικά του Συνόλου Δεδομένων ("Online Retail Dataset")**, οι **Τεχνικές Προεπεξεργασίας (RFM Analysis, Normalization)** και το **Περιβάλλον Ανάπτυξης Λογισμικού (Python, VSCode, Anaconda)** που χρησιμοποιήθηκε για την **Εκτέλεση των Πειραμάτων**.
- **Κεφάλαιο 5: Ανάλυση και Εφαρμογή.** Παρουσιάζεται η **Εξερευνητική Ανάλυση Δεδομένων (EDA)** και η **Εφαρμογή των Αλγορίθμων**. Τα **Αποτελέσματα** οπτικοποιούνται και αναλύονται οι προκύπτουσες **Συστάδες Πελατών (Clusters)**, ενώ γίνεται **Αξιολόγηση της Απόδοσης των Μοντέλων** και **Ανάλυση Σφαλμάτων**.
- **Κεφάλαιο 6: Αξιολόγηση και Ερμηνεία Αποτελεσμάτων.** Συζητούνται τα **Ευρήματα** σε σχέση με τα **Ερευνητικά Ερωτήματα** και τη **Βιβλιογραφία**. Ερμηνεύεται η **Επιχειρηματική Αξία των Αποτελεσμάτων** (π.χ. **Στρατηγικές** για **"Champions"** και **"Lost Customers"**) και αναφέρονται οι **Περιορισμοί της Έρευνας**.
- **Κεφάλαιο 7: Συμπεράσματα και Προτάσεις.** Συνοψίζονται τα **Βασικά Συμπεράσματα** της μελέτης και παρουσιάζονται οι **Μελλοντικές Τάσεις** στον **Χώρο των CDPs (Generative AI, Privacy)**, καθώς και **Προτάσεις για Περαιτέρω Έρευνα**.
- **Κεφάλαιο 8: Παράρτημα.** Περιλαμβάνει **Τμήματα του Κώδικα Python** που αναπτύχθηκε για την **Υλοποίηση των Αλγορίθμων** και την **Επεξεργασία των Δεδομένων**.

- **Κεφάλαιο 9: Βιβλιογραφία.** Παρατίθενται οι **Βιβλιογραφικές Αναφορές** που χρησιμοποιήθηκαν για την **Τεκμηρίωση της Εργασίας**.

2. Θεωρητικό Υπόβαθρο

2.1. Customer Data Platforms (CDPs)

Τα **Customer Data Platforms (CDPs)** αποτελούν την τεχνολογική απάντηση στην ανάγκη των επιχειρήσεων για μια ενοποιημένη και ενεργή γνώση του πελάτη. Η αγορά της **Τεχνολογίας Μάρκετινγκ (MarTech)** έχει κατακλυστεί από εργαλεία διαχείρισης δεδομένων, ωστόσο το **CDP** διακρίνεται για τον συγκεκριμένο ρόλο του στην ενοποίηση της πληροφορίας **[Gartner, 2020]**.

Ορισμός

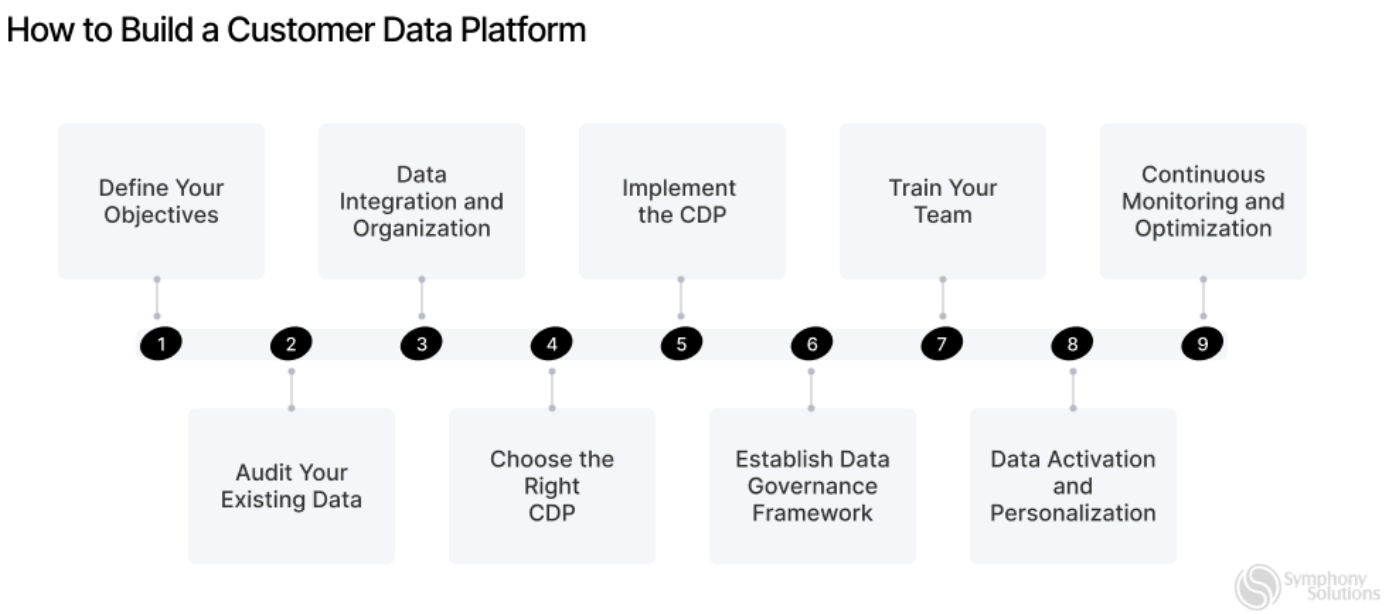
Σύμφωνα με τον επίσημο ορισμό του **CDP Institute**, «ένα CDP είναι ένα **Πακεταρισμένο Λογισμικό** που δημιουργεί μια **Μόνιμη, Ενοποιημένη Βάση Δεδομένων Πελατών**, η οποία είναι προσβάσιμη σε άλλα συστήματα» **[Raab, 2016]**.

Κύκλος Ζωής & Υλοποίηση ενός CDP

Η επιτυχής ενσωμάτωση ενός CDP σε έναν οργανισμό δεν είναι απλώς μια τεχνική εγκατάσταση, αλλά μια στρατηγική διαδικασία που ακολουθεί **εννέα (9) Διακριτά Στάδια**, όπως απεικονίζονται στην **Εικόνα 2.1**:

1. **Καθορισμός Στόχων (Define Your Objectives):** Το πρώτο βήμα απαιτεί σαφήνεια σχετικά με τους **Επιχειρηματικούς Στόχους**, όπως η μείωση της αποχώρησης πελατών (churn) ή η αύξηση του cross-selling.
2. **Έλεγχος Υπαρχόντων Δεδομένων (Audit Your Existing Data):** Ακολουθεί η χαρτογράφηση των πηγών δεδομένων για να εντοπιστούν "σιλό", ποιότητα και κενά στην πληροφορία.
3. **Ενοποίηση και Οργάνωση (Data Integration and Organization):** Η διαδικασία σχεδιασμού για το πώς θα συνδεθούν τα δεδομένα από διαφορετικά συστήματα (CRM, Web, Mobile) σε μία ενιαία δομή.

4. **Επιλογή της Κατάλληλης Πλατφόρμας (Choose the Right CDP):** Η αξιολόγηση και επιλογή της λύσης που ταιριάζει στις ανάγκες και τον προϋπολογισμό της επιχείρησης.



Εικόνα 2.1. How to Build a Customer Data Platform

5. **Υλοποίηση Συστήματος (Implement the CDP):** Η τεχνική φάση της εγκατάστασης, της παραμετροποίησης και της σύνδεσης των APIs.
6. **Θέσπιση Πλαισίου Διακυβέρνησης (Establish Data Governance Framework):** Ο καθορισμός κανόνων για την **Ασφάλεια**, την **Ιδιωτικότητα** και τη συμμόρφωση με κανονισμούς όπως ο **GDPR**.
7. **Εκπαίδευση Ομάδας (Train Your Team):** Η κατάρτιση του προσωπικού (marketers, analysts) ώστε να μπορούν να αξιοποιήσουν πλήρως τις δυνατότητες του εργαλείου.
8. **Ενεργοποίηση και Προσωποποίηση (Data Activation and Personalization):** Η πρακτική χρήση των δεδομένων για τη δημιουργία στοχευμένων καμπανιών και προσωποποιημένων εμπειριών.

9. **Συνεχής Παρακολούθηση και Βελτιστοποίηση (Continuous Monitoring and Optimization):** Η διαδικασία δεν τελειώνει ποτέ· απαιτείται συνεχής έλεγχος των αποτελεσμάτων και προσαρμογή της στρατηγικής.

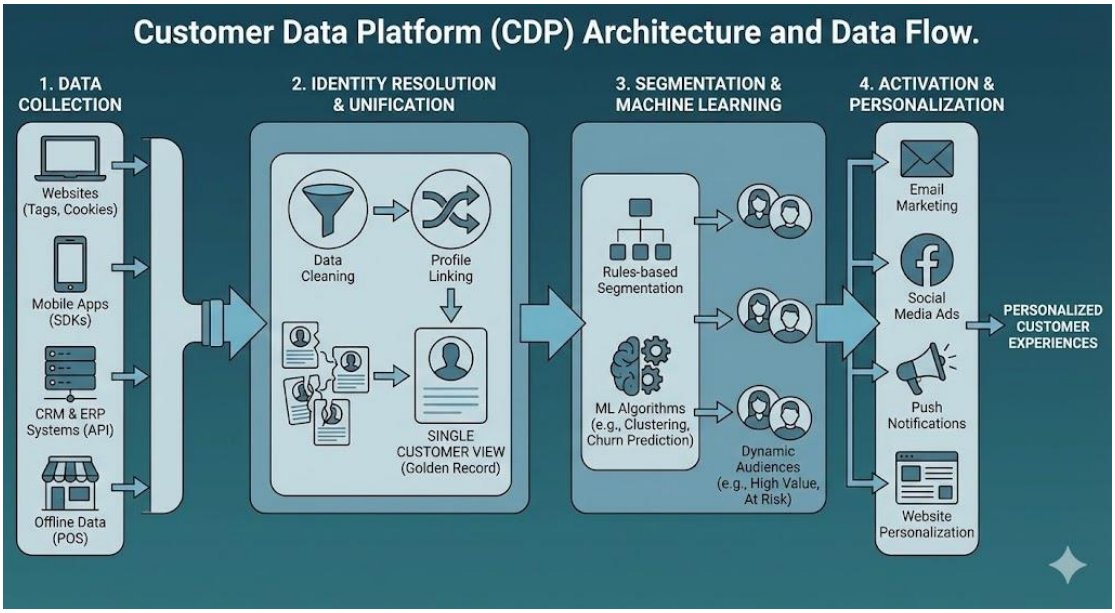
2.2. Αρχή Λειτουργίας ενός CDP

Για να κατανοήσουμε πώς ένα **CDP** μετατρέπει τα **Α0 Ακατέργαστα Δεδομένα** σε **Επιχειρηματική Αξία**, πρέπει να εξετάσουμε την **Αρχιτεκτονική** του **Δομή**. Η **Λειτουργία** του βασίζεται σε μια **Συνεχή Ροή** Τεσσάρων (4) Σταδίων, η οποία διασφαλίζει ότι η **Πληροφορία** είναι πάντα **Επίκαιρη** και **Αξιοποιήσιμη [Gartner, 2021]**.

Τα **Τέσσερα (4) Στάδια** της **Αρχιτεκτονικής (Data Flow)** είναι τα ακόλουθα:

1. **Συλλογή (Data Collection):** Η **Διαδικασία** ξεκινά με την **Άντληση Δεδομένων** από κάθε σημείο επαφής της **Επιχείρησης** (online και offline). Μέσω **APIs** και **SDKs**, το **CDP** συγκεντρώνει **Δεδομένα Συμπεριφοράς** (clicks, views), **Συναλλαγών** (αγορές) και **Δημογραφικά Στοιχεία** σε **Πραγματικό Χρόνο**.
2. **Ενοποίηση & Ταυτοποίηση (Identity Resolution):** Αυτός είναι ο **Πυρήνας** του **Συστήματος**. Το **CDP** καθαρίζει τα **Δεδομένα** και επιλύει το πρόβλημα της **Ταυτότητας**, συνδέοντας διαφορετικά "**Ψηφιακά Ίχνη**" (π.χ. ένα cookie από το laptop και ένα email από το κινητό) σε ένα μοναδικό, ενιαίο **Προφίλ Πελάτη** (**Single Customer View**).
3. **Τμηματοποίηση & Μηχανική Μάθηση (Segmentation & ML):** Πάνω στα **Ενοποιημένα Προφίλ** εφαρμόζονται **Αλγόριθμοι**. Εδώ δημιουργούνται Δυναμικά Κοινά (audiences) και γίνονται **Προβλέψεις** (π.χ. ποιοι πελάτες κινδυνεύουν να φύγουν), μετατρέποντας την **Πληροφορία** σε **Γνώση**.
4. **Ενεργοποίηση (Activation):** Το τελικό στάδιο αφορά τη **Δράση**. Τα εμπλουτισμένα **Δεδομένα** και τα **Segments** διοχετεύονται πίσω στα **Εργαλεία Μάρκετινγκ** (email platforms, ad networks) για την **Υλοποίηση** προσωποποιημένων **Εμπειριών**.

Αυτή η **Λειτουργική Ροή (Logical Architecture/Data Flow)** απεικονίζεται στο παρακάτω **Διάγραμμα Αρχιτεκτονικής**:



Εικόνα 2.2. Customer Data Platform Architecture and Data Flow

2.3. Ιστορική Εξέλιξη: Από το CRM στο CDP

Η εμφάνιση των **CDPs** δεν ήταν τυχαίο ή ένα απλό μεμονωμένο γεγονός. **Αντιθέτως**, ήταν αποτέλεσμα της **Ανάγκης** για υπέρβαση των περιορισμών των προηγούμενων συστημάτων, όπως τα **CRM (Customer Relationship Management)** και **DMP (Data Management Platform)**, να διαχειριστούν τον **Όγκο** και την **Πολυπλοκότητα** των **Δεδομένων** της **Ψηφιακής Εποχής**. Η **Εξέλιξη** αυτή καθοδηγήθηκε από την απαίτηση για βαθύτερη κατανόηση του πελάτη σε έναν **Πολυκαναλικό (Omnichannel)** κόσμο **[Moussaouy et al., 2020]**.

Στην ουσία, τα **CDPs** ήταν το **Αποτέλεσμα** μιας **Εξελικτικής Πορείας Τριών** (**3**) **Δεκαετιών** στη **Διαχείριση Δεδομένων**. Κάθε νέα **Τεχνολογία** προέκυπτε για να καλύψει τα κενά της προηγούμενης, ακολουθώντας τις αλλαγές στη **Συμπεριφορά** των **Καταναλωτών [Kumar & Reinartz, 2018]**.

Σύγκριση με Παραδοσιακά Συστήματα

Ενώ τα συστήματα **CRM** διαχειρίζονται εξαιρετικά τα **Δεδομένα** γνωστών **Πελατών** και τα **DMP** εστιάζουν σε **Ανώνυμα Δεδομένα** για **Διαφήμιση**, κανένα από τα δύο δεν προσφέρει την **Πλήρη Εικόνα**.

Η Εποχή του CRM (Customer Relationship Management)

Δεκαετία 1990

Στη δεκαετία του '90, οι **Επιχειρήσεις** έπρεπε να οργανώσουν τα **Πελατολόγιά** τους **Ψηφιακά**.

- **Ανάγκη:** Η **Ψηφιοποίηση** των αρχείων, η **Κατάργηση** των χειρόγραφων σημειώσεων και η **Οργάνωση** των **B2B πωλήσεων**.
- **Λύση:** Τα συστήματα **CRM** (π.χ. Salesforce, Siebel) **[Buttle & Maklan, 2019]**.
- **Βασικός Σκοπός:** Η **Βελτίωση** των **Επιχειρηματικών Σχέσεων**, η **Αύξηση** των **Πωλήσεων** και η **Εξυπηρέτηση Υφιστάμενων Πελατών**.
- **Τύπος Δεδομένων:** Κυρίως **Επώνυμα Δεδομένα (PII)** γνωστών **Πελατών** (Όνομα, Email, Τηλέφωνο) που εισάγονται συχνά **Χειροκίνητα**.
- **Περιορισμός: Αδυναμία Διαχείρισης Μεγάλου Όγκου Δεδομένων** συμπεριφοράς (behavioral data) και **Δεδομένων** από **Ανώνυμους Επισκέπτες** στο Web.

Η Εποχή του DMP (Data Management Platform)

Δεκαετία 2000

Με την **Έκρηξη** του **Διαδικτύου**, οι **Marketers** χρειάζονταν **Εργαλεία** για **Ψηφιακή Διαφήμιση**.

- **Ανάγκη:** Η στόχευση **Διαφημίσεων** (Display Ads) σε ευρύ κοινό στο **Διαδίκτυο** για την **Προσέλκυση** νέων **Πελατών** (Acquisition).
- **Λύση:** Οι πλατφόρμες **DMP** (π.χ. Oracle BlueKai) **[Gartner, 2021]**.

- **Βασικός Σκοπός:** Η **Δημιουργία** ανώνυμων **Τμημάτων Κοινού** (Audiences) για **Στοχευμένη Διαφήμιση** (Ad Targeting).
- **Τύπος Δεδομένων:** Αποκλειστικά **Ανώνυμα Δεδομένα** (Third-party Cookies, IP addresses, Device IDs) με μικρή **Διάρκεια** ζωής (προσωρινά).
- **Περιορισμός:** Τα **Δεδομένα** δεν είναι **Μόνιμα** και δεν μπορούν να συνδεθούν με συγκεκριμένα **Πρόσωπα**, καθιστώντας αδύνατη την **Προσωποποιημένη Εξυπηρέτηση**.

CDP (Customer Data Platform)

2013 και Μετά

Περίπου το **2013**, ο **David Raab** παρατήρησε το **Κενό**: Το **Marketing** χρειαζόταν μια **Ενιαία Βάση** που να συνδυάζει τα δύο παραπάνω **Συστήματα**, χωρίς να εξαρτάται από το **Τμήμα IT**. Το **CDP** ήρθε να **Γεφυρώσει** αυτό το **Χάσμα**, λειτουργώντας ως ο **Κεντρικός** "**Εγκέφαλος**" που συλλέγει **Δεδομένα** από παντού (Online και Offline) για να φτιάξει το **Πλήρες Προφίλ** του **Πελάτη**. Στην ουσία είναι ένα **Πακεταρισμένο Λογισμικό** που ελέγχεται από το **Τμήμα Μάρκετινγκ** και δημιουργεί μια **Μόνιμη** και **Ενοποιημένη Βάση Δεδομένων**, η οποία είναι προσβάσιμη από άλλα **Συστήματα**.

- **Ανάγκη:** Μια **Ενιαία Εικόνα** του πελάτη που να συνδυάζει τα **Offline Δεδομένα** (CRM) με τα online **Δεδομένα Συμπεριφοράς** (Web/Mobile) **[Moussaouy et al., 2020]**.
- **Λύση:** Το **CDP**, ένα **Λογισμικό Ελεγχόμενο** από το Marketing **[Raab, 2016]**.
- **Βασικός Σκοπός:** Η **Δημιουργία** της **Ενιαίας Εικόνας Πελάτη (Single Customer View)** για **Προσωποποίηση** σε πραγματικό χρόνο.
- **Τύπος Δεδομένων:** Συνδυασμός **Επώνυμων (PII)** και **Ανώνυμων Δεδομένων**. Εστιάζει στα **First-Party Data** (δεδομένα ιδιοκτησίας της εταιρείας) με **Μόνιμη Αποθήκευση**.
- **Περιορισμός: Υψηλή Πολυπλοκότητα Υλοποίησης** και **Ανάγκη** για καθαρά **Δεδομένα** ώστε να αποφευχθεί το φαινόμενο "**Garbage In, Garbage Out**".

Σύγκριση και Οφέλη

Η **Εξέλιξη** αυτή φαίνεται ξεκάθαρα στον **Πίνακα 2.1**, όπου συγκρίνονται τα **Χαρακτηριστικά** των τριών (3) γενιών **Συστημάτων**.

Χαρακτηριστικό	CRM (Customer Relationship Management)	DMP (Data Management Platform)	CDP (Customer Data Platform)
Τύπος Δεδομένων	Κυρίως **Επώνυμα** (Όνομα, Email).	Κυρίως **Ανώνυμα** (Cookies, IDs).	Συνδυασμός **Επώνυμων** και **Ανώνυμων**.
Διατήρηση	**Μόνιμη** (Ιστορικότητα).	**Προσωρινή** (μικρή διάρκεια ζωής).	**Μόνιμη** (Long-term retention).
Κύριος Σκοπός	**Πωλήσεις & Εξυπηρέτηση.**	**Διαφημιστική Στόχευση** (Ads).	**Ενιαία Εικόνα (SCV)** & **Προσωποποίηση**.

Πίνακας 2.1. Διαφορές των CRM, DMP με CDP

Τα Οφέλη της Εξέλιξης (Business Benefits)

Η μετάβαση σε **Αρχιτεκτονική CDP** προσφέρει στρατηγικά **Πλεονεκτήματα** που μεταμορφώνουν τη **Λειτουργία** της **Επιχείρησης**. Όπως παρουσιάζεται στην **Εικόνα 2.3**, τα **Οφέλη** αυτά εκτείνονται σε **Οκτώ** (8) **Βασικούς Άξονες**:

1. **Ενοποιημένη Εικόνα (Unified Customer View):** Η κατάργηση των σιλό δεδομένων.
2. **Πελατοκεντρική Προσέγγιση:** Εστίαση στις ανάγκες του πελάτη, όχι στο κανάλι.
3. **Καινοτομία:** Ευκολία ενσωμάτωσης νέων τεχνολογιών (AI).
4. **Πιστότητα (Retention):** Αύξηση του CLV (Customer Lifetime Value).
5. **Συμμόρφωση (Compliance):** Κεντρική διαχείριση συγκατάθεσης (GDPR).
6. **Ευελιξία (Agility):** Ταχύτητα στη λήψη αποφάσεων με real-time δεδομένα.
7. **Ποιότητα Δεδομένων:** Καθαρά και αξιόπιστα δεδομένα για τους αλγορίθμους.
8. **Δέσμευση (Engagement):** Βελτιωμένη εμπειρία πελάτη μέσω προσωποποίησης.

Εικόνα 2.3. Eight (8) Benefits of CDPs

2.4. Τύποι και Κατηγορίες CDPs

Δεν είναι όλα τα CDPs ίδια.

Το **CDP Institute** τα κατηγοριοποιεί σε **Τέσσερις** (**4**) **Βασικούς Τύπους**, ανάλογα με τις δυνατότητές τους, **Διαχωρισμός** που είναι κρίσιμος για την επιλογή της κατάλληλης υποδομής για **Εφαρμογές Μηχανικής Μάθησης [Raab, 2019]**.

Αξίζει να σημειωθεί ότι η **Κατηγοριοποίηση** αυτή ακολουθεί μια **Ιεραρχική Δομή**, όπου κάθε επόμενη κατηγορία συνήθως ενσωματώνει τις **Λειτουργίες** των προηγούμενων:

1. **Data CDPs:**

- **Λειτουργία:** Εστιάζουν αποκλειστικά στη **Συλλογή** και **Ενοποίηση Δεδομένων** από διαφορετικές πηγές.
- **Ρόλος:** Λειτουργούν ως η **«Θεμελιώδης Βάση»** ή ως **Τροφοδότες Δεδομένων** για **Εξωτερικά Συστήματα** (π.χ. Data Warehouses, BI tools). Δεν διαθέτουν εγγενείς **Δυνατότητες Ανάλυσης**, αλλά εξασφαλίζουν ότι τα **Δεδομένα** είναι καθαρά και προσβάσιμα.

2. **Analytics CDPs:**
   - **Λειτουργία:** Περιλαμβάνουν όλα τα χαρακτηριστικά των **Data CDPs**, προσθέτοντας εργαλεία **Ανάλυσης**, **Τμηματοποίησης** και **Οπτικοποίησης Δεδομένων**.
   - **Ρόλος:** Εδώ γίνεται η μετάβαση από την **Πληροφορία** στη **Γνώση**. Συχνά διαθέτουν ενσωματωμένα μοντέλα **Πρόβλεψης (Predictive Modeling)** και **Αλγορίθμους Μηχανικής Μάθησης** για τη δημιουργία προηγμένων **Customer Journeys**.
3. **Campaign CDPs:**
   - **Λειτουργία:** Διαθέτουν επιπλέον **Λειτουργίες** για την **Εκτέλεση Καμπανιών** απευθείας μέσα από την πλατφόρμα (π.χ. αποστολή **Personalized Emails, Push Notifications, in-app Messages**).
   - **Ρόλος:** Γνωστά και ως **Engagement CDPs**, επιτρέπουν στους **Marketers** όχι μόνο να αναλύσουν το **Κοινό**, αλλά και να δράσουν άμεσα, στοχεύοντας συγκεκριμένα **Τμήματα Πελατών**.
4. **Delivery CDPs:**
   - **Λειτουργία:** Ελέγχουν την **Παράδοση Μηνυμάτων** σε όλα τα κανάλια, προσφέροντας πλήρη **Ενορχήστρωση (Omnichannel Orchestration)**.
   - **Ρόλος:** Είναι η πιο σύνθετη **Κατηγορία**, καθώς λειτουργούν ως **Κεντρικοί Κόμβοι** που διαχειρίζονται την **Επικοινωνία** σε πραγματικό χρόνο, διασφαλίζοντας ότι το μήνυμα θα φτάσει τη σωστή στιγμή, μέσω του **Βέλτιστου Καναλιού**.

Σημείωση για την Εργασία

Για τους σκοπούς της παρούσας **Μελέτης**, εστιάζουμε κυρίως στα **Analytics CDPs**, καθώς αυτά παρέχουν την απαραίτητη υποδομή για την ανάπτυξη και **Εφαρμογή** των **Αλγορίθμων Συσταδοποίησης (Clustering)** και **Πρόβλεψης** που θα εξετάσουμε.

2.5. Τεχνολογική Διάκριση: CDP έναντι Cookies

Συχνά επικρατεί **Σύγχυση** μεταξύ των **CDPs** και των μηχανισμών παρακολούθησης μέσω **Cookies**. Η διαφορά τους, ωστόσο, είναι **Θεμελιώδης** και έγκειται στον **Σκοπό**, τη **Λειτουργία** και κυρίως στο είδος των **Δεδομένων** που διαχειρίζονται.

Χαρακτηριστικό	CDP Customer Data Platform	Cookies Browser Technologies
Ορισμός	Ολοκληρωμένο λογισμικό που δημιουργεί **Ενιαία Προφίλ Πελατών** από πολλαπλές πηγές **[CDP Institute, 2023]**.	Μικρά **αρχεία κειμένου** στον browser για παρακολούθηση δραστηριότητας **[Chaffey & Ellis-Chadwick, 2022]**.
Σκοπός	**Single Customer View**, Προηγμένη Ανάλυση, ML Segmentation.	**Βασική παρακολούθηση** (sessions), αποθήκευση προτιμήσεων χρήστη.
Τύπος Δεδομένων	**First-party** (κυρίως), Second & Third-party. Ενοποιεί Web, Mobile, POS, CRM.	Κυρίως **Third-party** (διαφήμιση) ή περιορισμένα First-party (login session).
Εμβέλεια	**Omnichannel** (Online & Offline). Υποστηρίζει Clustering & Classification **[Forrester, 2024]**.	Περιορίζεται στον **Browser** (Web only). Δεν "βλέπει" άλλες πηγές.
Χρόνος Αποθήκευσης	**Μακροπρόθεσμη** (Persistent Database). Ιστορικότητα ετών.	**Βραχυπρόθεσμη** (Session) ή περιορισμένη από ρυθμίσεις browser/user.
Εφαρμογές ML	**Προηγμένο ML** (K-Means, Random Forest, Recommendations) για **εξατομίκευση**.	**Περιορισμένη Αναλυτική Ικανότητα** (π.χ. απλό retargeting).
Συμμόρφωση	**Εργαλεία Διαχείρισης Συγκατάθεσης** (**GDPR Compliant**) **[Gartner, 2023]**.	Αντιμετωπίζει **Περιορισμούς** (Cookie blockers) και **Σταδιακή Κατάργηση** (cookieless future).

Πίνακας 2.2. Σύγκριση Customer Data Platform (CDP) και Cookies

Βασικές Διαφορές

Στο **Πίνακα 2.2** παρουσιάζεται η **Συγκριτική Ανάλυση** των δύο (2) Τεχνολογιών, αναδεικνύοντας γιατί τα **CDPs** είναι ανώτερα για την **Προσωποποίηση.**

Παραδείγματα Εφαρμογής

Παράδειγμα Εφαρμογής: Η Περίπτωση του "Γιάννη" Για να γίνει κατανοητή η **Διαφορά** στην πράξη:

- **Με Cookie:** Το **Cookie** "ξέρει" μόνο ότι ο **Browser** του Γιάννη είδε ένα ζευγάρι παπούτσια στο e-shop. Μπορεί να του δείξει μια **Διαφήμιση** για αυτά τα παπούτσια. Αν ο Γιάννης μπει από το κινητό του, το **Cookie** δεν τον αναγνωρίζει.
- **Με CDP:** Το **CDP** συλλέγει τα κλικ από το **Web**, την **Αγορά** από το **Mobile App** και το παράπονο που έκανε στο **Τηλεφωνικό Κέντρο** (**CRM**). Δημιουργεί το **Προφίλ** «Γιάννης», εφαρμόζει **K-Means Clustering** και καταλαβαίνει ότι ανήκει στους "**High Value Customers**". Αντί για **Διαφήμιση**, του στέλνει ένα προσωπικό email με κουπόνι **Επιβράβευσης**.

Σύνδεση με την Εργασία Στην παρούσα **Διπλωματική Εργασία**, τα **CDPs** αποτελούν τη βάση για την εφαρμογή **Τεχνικών Μηχανικής Μάθησης**. Τα **Cookies**, αν και χρήσιμα για την πρωτογενή συλλογή **web Δεδομένων**, αδυνατούν να υποστηρίξουν την ολιστική **Ανάλυση** που απαιτείται για **Μοντέλα** όπως το **K-Means** ή το **RFM Analysis**, λόγω της **Έλλειψης Ιστορικότητας** και **Ενοποίησης**.

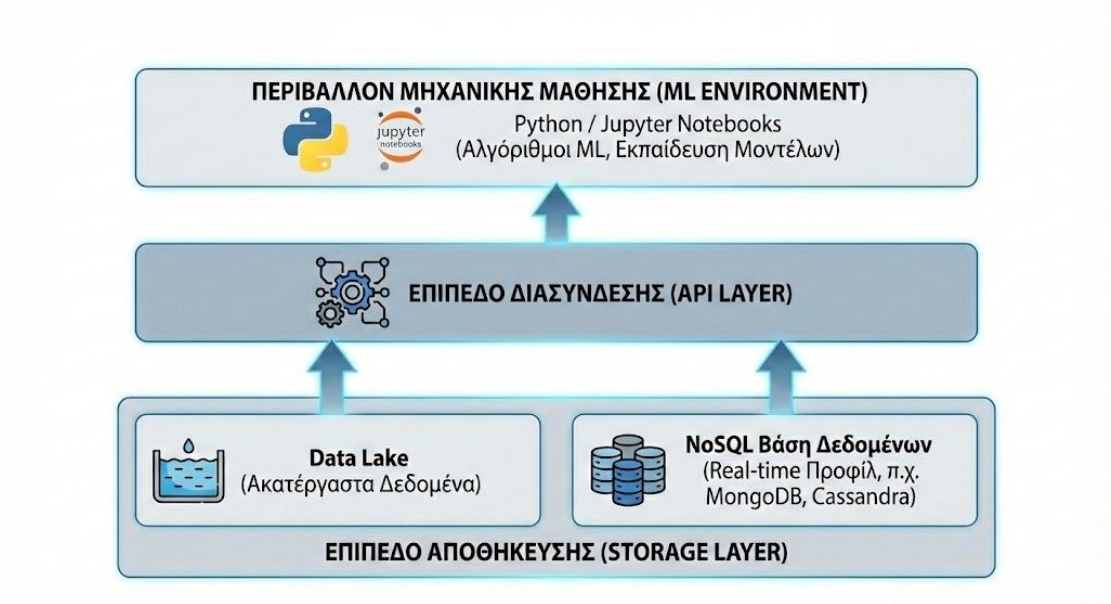
2.6. Τεχνολογική Υποδομή και Στοίβα Δεδομένων (Data Stack)

Ενώ η **Λειτουργική Ροή** ενός **CDP** αφορά τα **Στάδια Επεξεργασίας**, η **Τεχνική Υλοποίηση** βασίζεται σε συγκεκριμένες **Τεχνολογίες Αιχμής**. Η υποδομή ενός σύγχρονου **CDP** είναι συνήθως **Cloud-Native** και αξιοποιεί εργαλεία **Big Data** για να διαχειριστεί τον **Όγκο** και την **Ταχύτητα** της **Πληροφορίας**.

Σε επίπεδο **Αποθήκευσης (Storage Layer)**, χρησιμοποιείται συχνά μια **Υβριδική Προσέγγιση**:

- **Data Lakes:** Για την **Αποθήκευση Ακατέργαστων**, **Μη Δομημένων Δεδομένων** σε χαμηλό κόστος.
- **NoSQL Βάσεις Δεδομένων:** (π.χ. **MongoDB**, **Cassandra**) για τη γρήγορη ανάκτηση και **Ενημέρωση** των **Προφίλ** πελατών σε **Πραγματικό** χρόνο, κάτι που οι παραδοσιακές **SQL** βάσεις δυσκολεύονται να επιτύχουν σε **Μεγάλη Κλίμακα [Kleppmann, 2017]**.

Η **Διασύνδεση** με **Αλγόριθμους Μηχανικής Μάθησης** δεν γίνεται στο κεντρικό σύστημα συναλλαγών, αλλά μέσω **APIs** ή μέσω εξαγωγής **Datasets** σε εξειδικευμένα περιβάλλοντα **Ανάλυσης**, όπως τα **Python/Jupyter Notebooks**. Η **Ποιότητα** αυτής της **Τεχνικής Αρχιτεκτονικής** καθορίζει την "καθαρότητα" των δεδομένων (**Data Hygiene**), η οποία είναι προαπαιτούμενο για την επιτυχή **Εκπαίδευση** των **Μοντέλων [Gartner, 2021]** [**Εικόνα 2.6**].



Εικόνα 2.6. Τεχνολογική Στοίβα Δεδομένων (Data Stack) ενός CDP

2.7. Προκλήσεις και Περιορισμοί στην Υιοθέτηση ενός CDP

Παρά τα **Στρατηγικά Πλεονεκτήματα** που αναλύθηκαν στην **Ενότητα 2.3**, η υιοθέτηση ενός **CDP** δεν γίνεται χωρίς εμπόδια. Για την επιτυχή λειτουργία του **Συστήματος** και την **Υποστήριξη** των **Data Scientists** (μειώνοντας τον χρόνο προετοιμασίας δεδομένων), πρέπει να αντιμετωπιστούν **Συγκεκριμένες Προκλήσεις [Vetri Selvi et al., 2020]**.

Οργανωτικές και Οικονομικές Προκλήσεις

Η **Πολυπλοκότητα** της **Υλοποίησης** είναι υψηλή, καθώς το **CDP** δεν είναι απλώς ένα έργο IT, αλλά απαιτεί τη **Στενή Συνεργασία** πολλών τμημάτων (**Marketing, Sales, IT**). Η έλλειψη κουλτούρας **Συνεργασίας** μπορεί να οδηγήσει σε **Αποτυχία**. Επιπλέον, το **Κόστος Απόκτησης** και **Συντήρησης (TCO)** μπορεί να είναι **Απαγορευτικό** για **Μικρότερες Επιχειρήσεις**, καθιστώντας την **Επένδυση** ρίσκο αν δεν υπάρχει ξεκάθαρο **ROI (Return on Investment)**.

Η Παγίδα του "GIGO" (Garbage In, Garbage Out)

Η σημαντικότερη **Τεχνική Πρόκληση** είναι η **Ποιότητα των Δεδομένων**. Η ύπαρξη "**Βρώμικων Δεδομένων**" (dirty data) στις πηγές (π.χ. διπλότυπα, ελλιπή πεδία, λανθασμένα formats) υπονομεύει την **Αξιοπιστία** του **CDP**. Αν τροφοδοτήσουμε τους Αλγορίθμους **Μηχανικής Μάθησης** με **Κακής Ποιότητας Δεδομένα**, τα **Αποτελέσματα** των **Προβλέψεων** θα είναι λανθασμένα, επιβεβαιώνοντας τον κανόνα **Garbage In, Garbage Out**.

*Αυτό καθιστά το στάδιο του **Καθαρισμού Δεδομένων (Data Cleaning)** – το οποίο θα εξετάσουμε στο πρακτικό μέρος της εργασίας – κρίσιμο για την επιτυχία κάθε έργου CDP.*

2.8. Νομικά και Ηθικά Ζητήματα

Στην **Ψηφιακή Εποχή**, η συλλογή και **Επεξεργασία Δεδομένων** δεν είναι μόνο **Τεχνικό** ζήτημα, αλλά διέπεται από αυστηρούς νομικούς κανόνες και ηθικούς φραγμούς. Τα **CDPs** καλούνται να λειτουργήσουν μέσα σε ένα πολύπλοκο κανονιστικό **Πλαίσιο**.

Το Κανονιστικό Πλαίσιο: Ευρώπη και ΗΠΑ

Οι δύο (2) κυρίαρχοι **Πυλώνες** που καθορίζουν την **Παγκόσμια Στρατηγική Δεδομένων** είναι:

1. **Ευρώπη - GDPR (General Data Protection Regulation):** Ο **Γενικός Κανονισμός** για την **Προστασία Δεδομένων** (ΕΕ 2016/679) τέθηκε σε εφαρμογή το **2018** και επιβλέπεται από το **EDPB (European Data Protection Board)** και τις εθνικές αρχές.
    - **Τι περιλαμβάνει:** Κατοχυρώνει θεμελιώδη δικαιώματα όπως το **Δικαίωμα Πρόσβασης** (Right of Access), το **Δικαίωμα Διόρθωσης** και το κρίσιμο **Δικαίωμα στη Λήθη** (Right to be Forgotten / Erasure). Απαιτεί ρητή συγκατάθεση (Opt-in) για την επεξεργασία δεδομένων.
2. **ΗΠΑ - CCPA & FTC:** Στις **ΗΠΑ** δεν υπάρχει ακόμα ένας ενιαίος ομοσπονδιακός νόμος αντίστοιχος του **GDPR**. Τον ρόλο της προστασίας του καταναλωτή σε ομοσπονδιακό επίπεδο έχει η **FTC (Federal Trade Commission)**. Ωστόσο, σημείο αναφοράς αποτελεί ο νόμος της Καλιφόρνια, **CCPA (California Consumer Privacy Act)**.
    - **Τι περιλαμβάνει:** Εστιάζει στη **Διαφάνεια** και δίνει στους καταναλωτές το **Δικαίωμα** να γνωρίζουν ποια **Δεδομένα** συλλέγονται και το δικαίωμα να αρνηθούν την **Πώληση** τους σε τρίτους (**Right to Opt-out**).

Ο Ρόλος του CDP στη Συμμόρφωση (Compliance)

Σε αυτό το περιβάλλον, τα **CDPs** παίζουν κεντρικό ρόλο. Επιτρέπουν την κεντρική **Διαχείριση της Συγκατάθεσης (Consent Management)**, διασφαλίζοντας ότι η επιθυμία του χρήστη (π.χ. "μην με παρακολουθείς") εφαρμόζεται καθολικά. Επιπλέον, διευκολύνουν τεχνικά το δικαίωμα στη λήθη, καθώς μπορούν να εντοπίσουν και να διαγράψουν τα **Δεδομένα** ενός

χρήστη από όλα τα συνδεδεμένα **Συστήματα** με μία εντολή **[Voigt & Von dem Bussche, 2017]**.

Ηθικές Προεκτάσεις

Από **Ηθικής Πλευράς**, η χρήση δεδομένων για **Προσωποποίηση** εγείρει ζητήματα ιδιωτικότητας. Η γραμμή μεταξύ της "εξυπηρετικής προσωποποίησης" και της "εισβολής" (**Creepiness Factor**) είναι λεπτή. Οι αλγόριθμοι **Μηχανικής Μάθησης** που τροφοδοτούνται από τα **CDPs** πρέπει να σχεδιάζονται με γνώμονα τη διαφάνεια (Explainability) και την αποφυγή διακρίσεων (**Algorithmic Bias**), ώστε να μην αποκλείουν ή αδικούν συγκεκριμένες **Ομάδες** χρηστών **[Mittelstadt et al., 2016]**.

2.9. Προσωποποίηση στην Εμπειρία Χρήστη

Η **Προσωποποίηση (Personalization)** ορίζεται ως η **Διαδικασία** προσαρμογής της εμπειρίας, του περιεχομένου και των προσφορών στα ατομικά **Χαρακτηριστικά** και τις **προτιμήσεις** του κάθε χρήστη. Δεν αφορά απλώς την προσφώνηση με το όνομα του Πελάτη, αλλά την κατανόηση του **πλαισίου (context)** και της **πρόθεσης (intent)** της κάθε αλληλεπίδρασης **[Kallweit et al., 2014]**.

Ο Ρόλος του CDP: Από τα Δεδομένα στην Εμπειρία

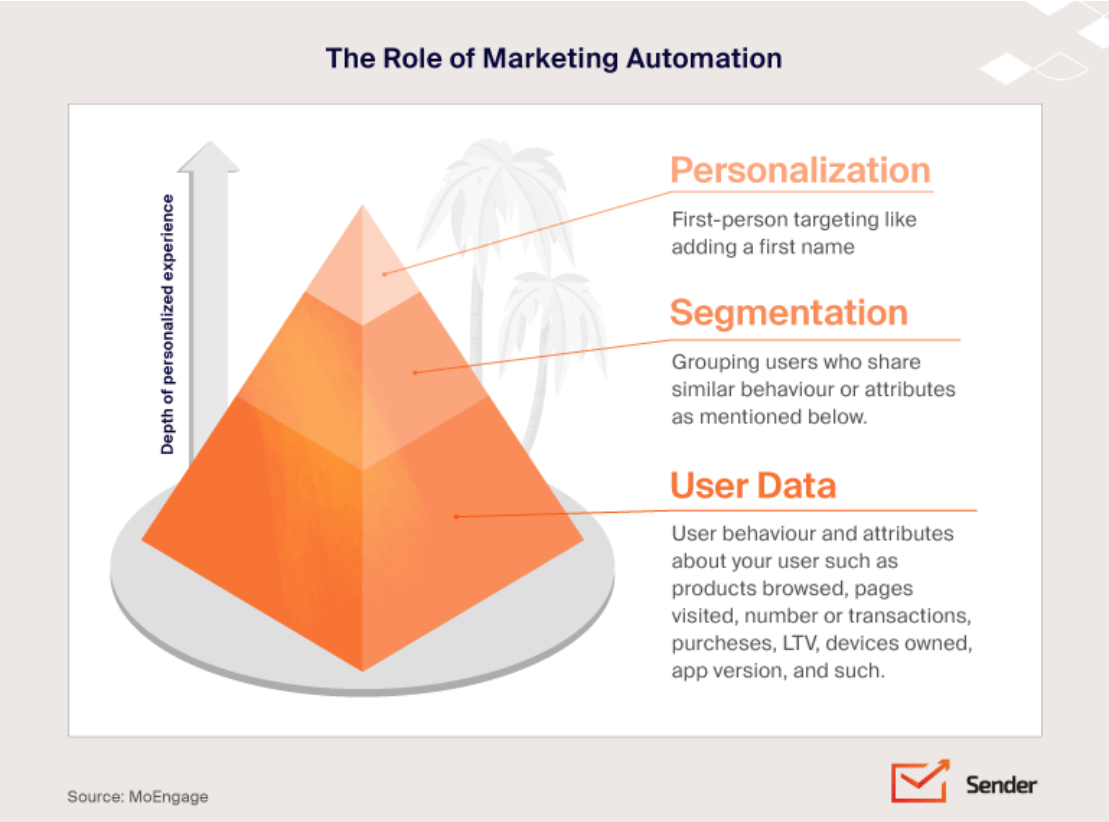
Η αποτελεσματική **Προσωποποίηση** απαιτεί **Δεδομένα** υψηλής ποιότητας. Εδώ το **CDP** λειτουργεί ως ο "εγκέφαλος" που τροφοδοτεί τα **Κανάλια Επικοινωνίας** (email, web, mobile), εξασφαλίζοντας ότι η εμπειρία του χρήστη είναι **Συνεπής (consistent)** σε όλα τα σημεία **Επαφής** (touchpoints).

Επίπεδα Προσωποποίησης

Στη σύγχρονη βιβλιογραφία, η **Προσωποποίηση** διακρίνεται σε δύο βασικές κατηγορίες, ανάλογα με τον τρόπο συλλογής των δεδομένων:

- **Ρητή (Explicit Personalization):** Βασίζεται σε δεδομένα που ο χρήστης μας δίνει οικειοθελώς (π.χ. συμπλήρωση φόρμας προτιμήσεων) **[Schafer et al., 2007]**.
- **Άρρητη (Implicit Personalization):** Βασίζεται στη συμπεριφορά του χρήστη. Εδώ η **Μηχανική Μάθηση** παίζει καθοριστικό ρόλο, καθώς αναλύει μοτίβα για να "μαντέψει" τις ανάγκες του χρήστη (**Predictive Personalization**) [**Hu et al., 2008**].

Η Εξέλιξη από την απλή, μαζική Επικοινωνία στην απόλυτη Εξατομίκευση απεικονίζεται συχνά ως Πυραμίδα, όπως φαίνεται στην παρακάτω **Εικόνα 2.9.a**. Σε αυτή, απεικονίζεται η **Ιεραρχική Δομή** της **Προσωποποίησης**, η οποία βασίζεται στη σταδιακή διύλιση των **Δεδομένων**. Όσο ανεβαίνουμε στην πυραμίδα, αυξάνεται η αξία της εμπειρίας για τον **Πελάτη** και η **Πολυπλοκότητα** της **Επεξεργασίας** για το σύστημα:



Εικόνα 2.9.a. Η Πυραμίδα της Προσωποποίησης: Από τα Δεδομένα Χρήστη στην Τμηματοποίηση και την Εξατομίκευση.

1. **Η Βάση - Δεδομένα Χρήστη (User Data):** Αποτελεί τα θεμέλια του οικοδομήματος. Εδώ βρίσκεται ο τεράστιος όγκος των ακατέργαστων **Δεδομένων** που συλλέγει το **CDP** (User attributes, behavior, transactions, devices). Χωρίς αυτή τη στέρεη **Βάση Δεδομένων** ("Big Data"), καμία ανάλυση δεν είναι εφικτή.
2. **Το Μέσο - Τμηματοποίηση (Segmentation):** Σε αυτό το στάδιο, τα δεδομένα οργανώνονται σε ομάδες. Εδώ εφαρμόζονται **Αλγόριθμοι** (όπως ο **K-Means** που θα εξετάσουμε) για να εντοπιστούν χρήστες με παρόμοια συμπεριφορά ή χαρακτηριστικά (π.χ. "Συχνοί Αγοραστές"). Είναι το στάδιο της μετάβασης από το "**Χάος**" στην "**Τάξη**".
3. **Η Κορυφή - Προσωποποίηση (Personalization):** Η κορύφωση της **Διαδικασίας**. Εδώ φεύγουμε από τις ομάδες και εστιάζουμε στο **Άτομο** (1-to-1). Με βάση τα μοτίβα που βρέθηκαν στα προηγούμενα στάδια, το σύστημα προσφέρει μια μοναδική εμπειρία (π.χ. μια πρόβλεψη αγοράς ή μια εξατομικευμένη προσφορά), στοχεύοντας αποκλειστικά τις **Ανάγκες** του συγκεκριμένου Χρήστη.

Η Προσέγγιση 360 Μοιρών (360-Degree View) Η σύγχρονη τάση στο μάρκετινγκ είναι η μετάβαση από την απλή **Τμηματοποίηση** στην **Προσωποποίηση 360°**. Αυτό σημαίνει ότι η επιχείρηση δεν βλέπει τον πελάτη αποσπασματικά (π.χ. μόνο ως "αγοραστή" ή μόνο ως "χρήστη που έκανε παράπονο"), αλλά ολιστικά.

Η εικόνα 360° συνθέτει τέσσερις διαστάσεις δεδομένων για να επιτύχει την **Υπερ-εξατομίκευση (Hyper-personalization)**:

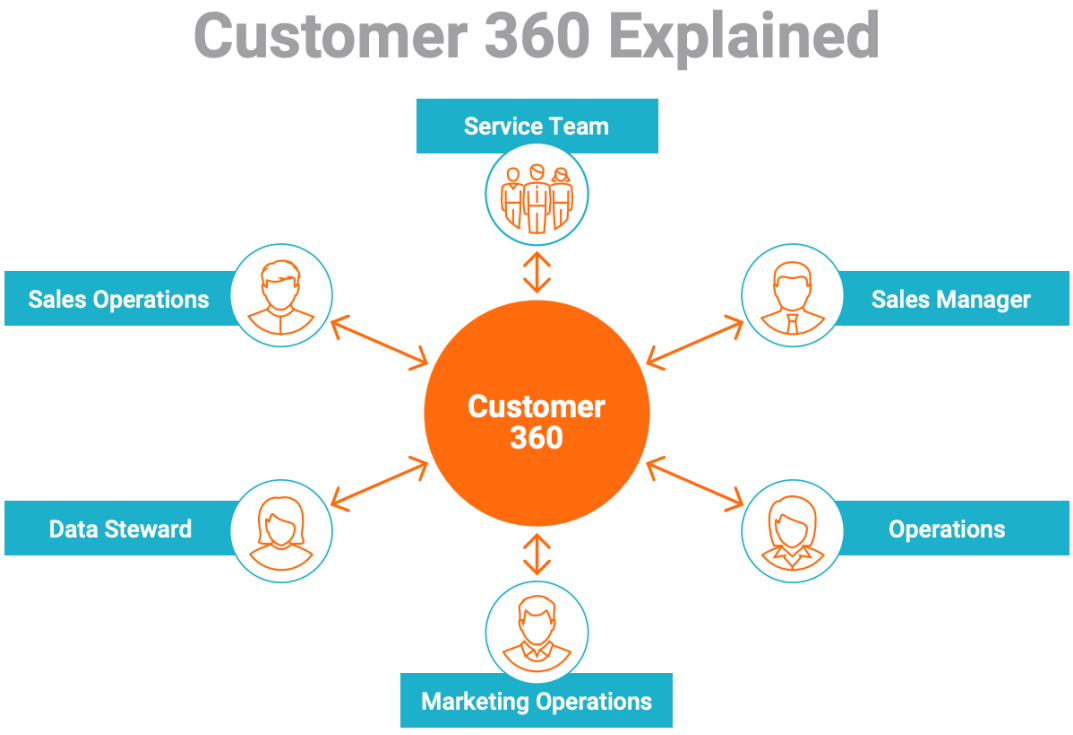
1. **Δημογραφικά:** Ποιος είναι ο πελάτης (Ηλικία, Φύλο, Τοποθεσία).
2. **Συναλλακτικά:** Τι έχει αγοράσει και πόσο συχνά (**RFM Analysis**).
3. **Συμπεριφορικά:** Πώς αλληλεπιδρά ψηφιακά (Clicks, Χρόνος παραμονής, Εγκατάλειψη καλαθιού).
4. **Πλαισιακά (Contextual):** Γιατί συμπεριφέρεται έτσι (π.χ. αγοράζει λόγω έκπτωσης ή λόγω εποχικότητας).

Μέσω της **Μηχανικής Μάθησης**, το **CDP** συνδυάζει αυτές τις διαστάσεις και τις καθιστά προσβάσιμες σε όλα τα τμήματα του οργανισμού, όπως φαίνεται στην **Εικόνα 2.9.b**. Αυτό λοιπόν, που βλέπουμε να αποτυπώνεται σε αυτήν, είναι ότι η αξία του **Customer 360** δεν περιορίζεται μόνο στη συλλογή των δεδομένων, αλλά επεκτείνεται στη **λειτουργική**

ενοποίηση του οργανισμού. Η **Εικόνα** δείχνει πώς το **CDP** λειτουργεί ως η «**Ενιαία Πηγή Αλήθειας**» (Single Source of Truth), τροφοδοτώντας ταυτόχρονα διαφορετικά τμήματα:

- **Marketing:** Για τη δημιουργία στοχευμένων καμπανιών.
- **Sales (Πωλήσεις):** Για να γνωρίζουν οι πωλητές το ιστορικό του πελάτη πριν την επικοινωνία.
- **Service (Εξυπηρέτηση):** Για την άμεση επίλυση προβλημάτων με γνώση του πλαισίου.

Με αυτόν τον τρόπο, καταργούνται τα στεγανά **Δεδομένων** (Data Silos) και διασφαλίζεται ότι ο πελάτης λαμβάνει συνεπή **Εμπειρία**, ανεξάρτητα από το **Τμήμα** με το οποίο αλληλεπιδρά.



Εικόνα 2.9.b. Η Προσέγγιση Customer 360: Ενοποιημένη εικόνα πελάτη προσβάσιμη από όλα τα επιχειρησιακά τμήματα.

2.10. Μηχανική Μάθηση: Από τα Δεδομένα στη Νοημοσύνη

Ενώ το **CDP** παρέχει τα **Δεδομένα**, η **Μηχανική Μάθηση (Machine Learning - ML)** παρέχει τη **Νοημοσύνη**. Η **Στατική Τμηματοποίηση** βάσει απλών **Δημογραφικών Στοιχείων** (π.χ. "Άνδρες, 25-35 ετών") θεωρείται πλέον παρωχημένη, καθώς δεν αποτυπώνει την **Πραγματική Πρόθεση** του χρήστη. Αντίθετα, η **ML** επιτρέπει τη **δυναμική ανάλυση** μοτίβων συμπεριφορών σε τεράστια **Σύνολα Δεδομένων**, εντοπίζοντας πολύπλοκες, μη **Γραμμικές Συσχετίσεις** που είναι αδύνατο να βρει ο **Ανθρώπινος Παράγοντας [Jordan & Mitchell, 2015]**. Συγκεκριμένα, οι **Αλγόριθμοι** μπορούν να επεξεργαστούν ταυτόχρονα χιλιάδες **Μεταβλητές** (dimensions), αποκαλύπτοντας κρυμμένες **Συστάδες** πελατών που μοιράζονται παρόμοια αγοραστικά ταξίδια, ακόμη και αν φαινομενικά διαφέρουν στα **Δημογραφικά** τους **Χαρακτηριστικά**.

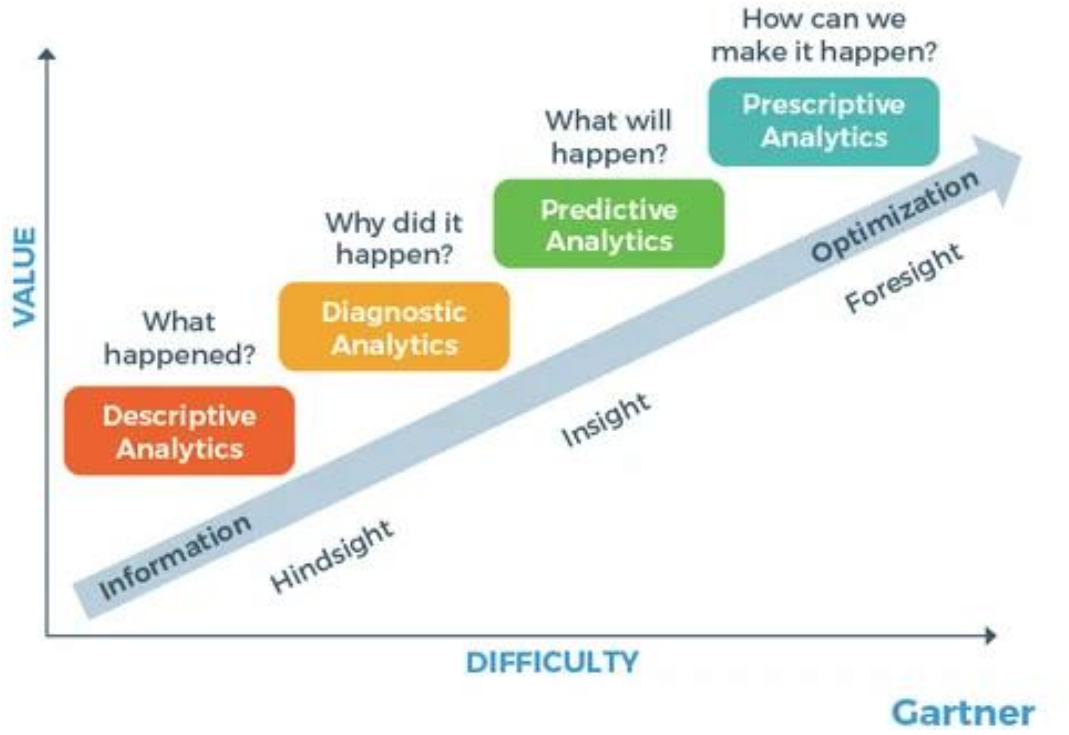
Τα Τρία Επίπεδα Ανάλυσης

Η εφαρμογή **Αλγορίθμων ML** πάνω στα **Δεδομένα** ενός **CDP** χωρίζεται σε **Τρεις** (3) **Διακριτές Κατηγορίες**, ανάλογα με το ερώτημα που καλείται να απαντήσει η κάθε μία από αυτές:

1. **Περιγραφική (Descriptive Analytics):**
    - **Το Ερώτημα:** "Τι συνέβη και ποιοι είναι οι πελάτες μας;"
    - **Η Εφαρμογή: Ομαδοποίηση Πελατών** με βάση κοινά χαρακτηριστικά (π.χ. **Clustering / K-Means**) για την κατανόηση της υπάρχουσας κατάστασης.
2. **Προβλεπτική (Predictive Analytics):**
    - **Το Ερώτημα:** "Τι πιθανολογείται να συμβεί στη συνέχεια;"
    - **Η Εφαρμογή: Πρόβλεψη Μελλοντικής Συμπεριφοράς**, όπως η **Πιθανότητα Αποχώρησης** (**Churn Prediction**) ή η επόμενη πιθανή αγορά.
3. **Κανονιστική (Prescriptive Analytics):**
    - **Το Ερώτημα:** "Πώς μπορούμε να επηρεάσουμε το αποτέλεσμα;"

- **Η Εφαρμογή: Προτάσεις** για τη **Βέλτιστη Ενέργεια** (**Next Best Action**), όπως η βελτιστοποίηση των Συστημάτων Συστάσεων (Recommendation Systems) για τη μεγιστοποίηση του κέρδους.

Η εξέλιξη αυτή απεικονίζεται στην παρακάτω **Εικόνα 2.10**, η οποία δείχνει πώς αυξάνεται η **Επιχειρηματική Αξία** καθώς προχωράμε από την **Περιγραφή** στην **Πρόβλεψη**.



Εικόνα 2.10. Το Μοντέλο Ωριμότητας των Analytics: Από την Περιγραφική στην Κανονιστική Ανάλυση.

Στη παραπάνω Εικόνα απεικονίζεται το **Μοντέλο Ωριμότητας των Analytics (Analytics Maturity Model)**. Εκεί, φαίνεται ξεκάθαρα πώς αυξάνεται η **Αξία (Value)** και η **Δυσκολία (Difficulty)** καθώς προχωράμε από την **Περιγραφή** του **Παρελθόντος** (Descriptive) στην **Πρόβλεψη** του **Μέλλοντος** (Predictive) και τελικά στη **Βελτιστοποίηση** των **Αποφάσεων** (Prescriptive).

Στις **Επόμενες Ενότητες**, θα αναλυθούν οι συγκεκριμένοι **Αλγόριθμοι** που χρησιμοποιούνται για αυτούς τους σκοπούς (K-Means, RFM), καθώς και το **Μαθηματικό Υπόβαθρο** που τους διέπει.

2.11. Βασικές Έννοιες Μηχανικής Μάθησης

Η **Μηχανική Μάθηση (Machine Learning - ML)** αποτελεί τον σημαντικότερο υποκλάδο της **Τεχνητής Νοημοσύνης**, ο οποίος εστιάζει στην ανάπτυξη **Αλγορίθμων** που επιτρέπουν στα υπολογιστικά συστήματα να μαθαίνουν από τα **Δεδομένα** και να βελτιώνουν την απόδοσή τους εμπειρικά, χωρίς να είναι ρητά προγραμματισμένα για κάθε πιθανό σενάριο **[Bishop, 2006]**. Η ραγδαία άνοδος της ML οφείλεται στην εκθετική αύξηση του **Όγκου των Δεδομένων (Big Data)** και της **Υπολογιστικής Ισχύος**, επιτρέποντας την αυτοματοποιημένη εξαγωγή **Προτύπων (Patterns)** που θα ήταν αδύνατο να εντοπιστούν με παραδοσιακές στατιστικές μεθόδους.

Στο πλαίσιο των CDPs, οι αλγόριθμοι ML κατηγοριοποιούνται κυρίως σε τρία βασικά παραδείγματα μάθησης:

α. Εποπτευόμενη Μάθηση (Supervised Learning)

Στην **Εποπτευόμενη Μάθηση**, ο αλγόριθμος εκπαιδεύεται σε ένα **Σύνολο Δεδομένων** που περιέχει **Ζεύγη Εισόδου-Εξόδου (Input-Output Pairs)**. Στόχος είναι η εκμάθηση μιας **Συνάρτησης Αντιστοίχισης**

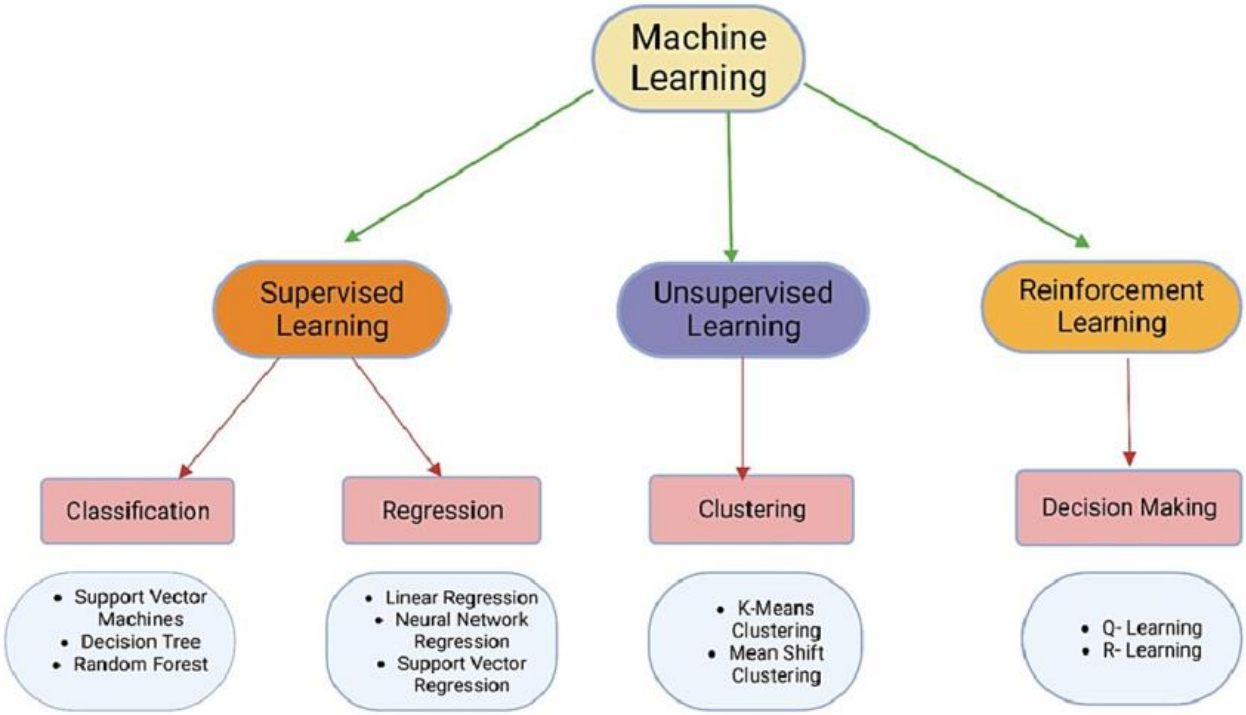
$$f: X \rightarrow Y,$$

η οποία μπορεί να προβλέψει την **Ετικέτα Εξόδου (Label)** για νέα, άγνωστα δεδομένα **[Hastie et al., 2009]**. Στο πλαίσιο αυτής της συνάρτησης:

- Όπου X: Αντιπροσωπεύει τα **Δεδομένα Εισόδου (Input Features)**. Στην περίπτωση ενός CDP, το X είναι το διάνυσμα των χαρακτηριστικών του πελάτη (π.χ. **Ηλικία**, **Ιστορικό Αγορών**, **Χρόνος Περιήγησης**).
- Όπου Y: Αντιπροσωπεύει τη **Μεταβλητή Στόχο (Target Variable)**. Είναι η τιμή που επιθυμούμε να προβλέψουμε (π.χ. **Churn: Ναι/Όχι** ή **Προβλεπόμενη Αξία Αγοράς**).

Εφαρμογή σε CDP: Χρησιμοποιείται όταν γνωρίζουμε την **Ιστορική Συμπεριφορά** (π.χ. ποιοι πελάτες διέκοψαν τη συνδρομή τους). Ο αλγόριθμος μαθαίνει τη σχέση μεταξύ των

χαρακτηριστικών (X) και του **αποτελέσματος** (Y) για να εντοπίσει μελλοντικές αποχωρήσεις (**Churn Prediction**).



Εικόνα 2.11. Τα Τρία (3) βασικά είδη Μηχανικής Μάθησης (Εποπτευόμενη, Μη Εποπτευόμενη, Ενισχυτική) και οι Διαφορές τους.

β. Μη Εποπτευόμενη Μάθηση (Unsupervised Learning)

Σε αντίθεση με την εποπτευόμενη, εδώ τα δεδομένα δεν διαθέτουν ετικέτες (**Unlabeled Data**). Ο στόχος του αλγορίθμου είναι να ανακαλύψει **Κρυμμένες Δομές**, **Μοτίβα** ή **Ομαδοποιήσεις** εντός των δεδομένων χωρίς ανθρώπινη παρέμβαση **[Ghahramani, 2004]**.

Εφαρμογή σε CDP: Αποτελεί τη βάση για την ανακάλυψη **Τμημάτων Πελατών (Segmentation)** που δεν είχαν οριστεί εκ των προτέρων, βασιζόμενη αποκλειστικά στην **Ομοιότητα της Συμπεριφοράς** τους (π.χ. αγοραστικά μοτίβα, χρόνος περιήγησης).

γ. Ενισχυτική Μάθηση (Reinforcement Learning - RL)

Η **Ενισχυτική Μάθηση** ασχολείται με τον τρόπο που ένας **Πράκτορας (Agent)** λαμβάνει αποφάσεις σε ένα **Περιβάλλον**, με στόχο τη μεγιστοποίηση μιας **Αθροιστικής Ανταμοιβής**

(Reward). Ο πράκτορας μαθαίνει μέσω της διαδικασίας **Δοκιμής και Λάθους (Trial and Error) [Sutton & Barto, 2018]**.

Εφαρμογή σε CDP: Χρησιμοποιείται για τη **Δυναμική Βελτιστοποίηση Ενεργειών**, όπως η επιλογή της "**Επόμενης Καλύτερης Προσφοράς**" (**Next Best Action**), όπου το σύστημα μαθαίνει ποια προσφορά μεγιστοποιεί την **Πιθανότητα Κλικ** ή **Αγοράς** σε πραγματικό χρόνο.

Είναι κρίσιμο να τονιστεί ότι η επιλογή της κατάλληλης **Κατηγορίας Μάθησης** εξαρτάται άμεσα από τη διαθεσιμότητα **Ετικετών (Labels)** και τον επιχειρηματικό στόχο. Για τους σκοπούς της παρούσας εργασίας, η έμφαση θα δοθεί κυρίως στη **Μη Εποπτευόμενη Μάθηση** και συγκεκριμένα στη **Συσταδοποίηση (Clustering)**. Ο λόγος είναι ότι στα σύγχρονα CDPs, το ζητούμενο είναι συχνά η "**Ανακάλυψη Γνώσης**" (**Knowledge Discovery**) από ακατέργαστα δεδομένα, ώστε να εντοπιστούν αυθεντικά **Customer Personas** που δεν ήταν εκ των προτέρων γνωστά στον Marketer.

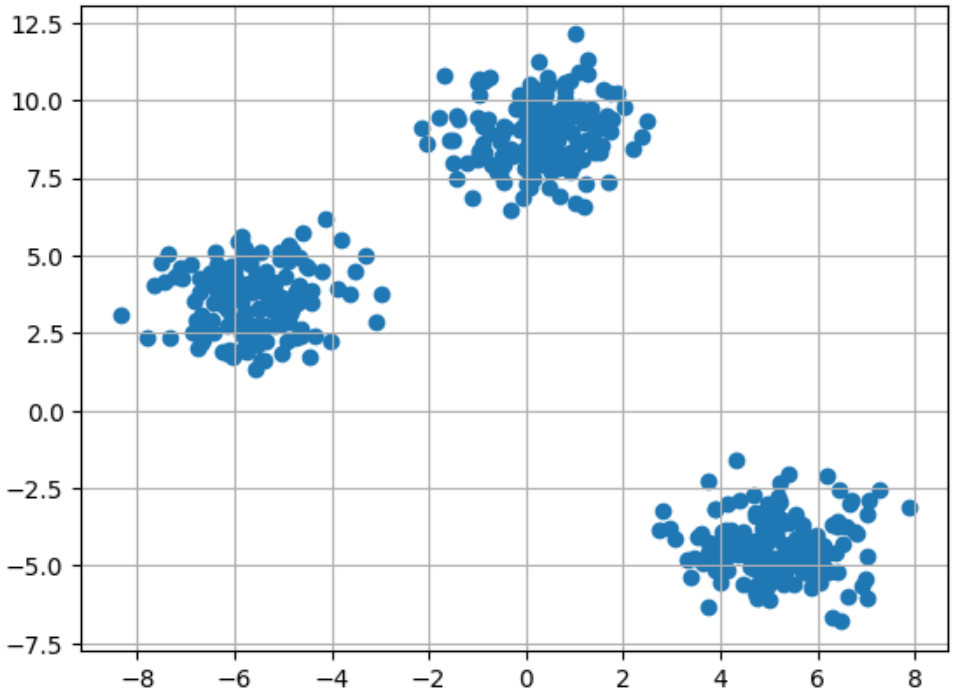
2.12. Τεχνικές που χρησιμοποιούνται για Προσωποποίηση

Η **Πρακτική Εφαρμογή** των παραπάνω εννοιών στα **Customer Data Platforms (CDP)** υλοποιείται μέσω συγκεκριμένων **Τεχνικών**, οι οποίες μετατρέπουν τα **Ακατέργαστα Δεδομένα** σε μια ολοκληρωμένη και **Εξατομικευμένη Εμπειρία Χρήστη**.

α. Συσταδοποίηση (Clustering)

Η **Συσταδοποίηση (Clustering)** αποτελεί τη θεμελιώδη διαδικασία διαχωρισμού ενός συνόλου αντικειμένων σε ομάδες, κατά τρόπο ώστε τα αντικείμενα στην ίδια **Συστάδα (Cluster)** να παρουσιάζουν τη μέγιστη δυνατή **Ομοιότητα** μεταξύ τους, σε σχέση με εκείνα που ανήκουν σε άλλες ομάδες. Είναι η κατεξοχήν μέθοδος **Μη Εποπτευόμενης Μάθησης** για την **Τμηματοποίηση Πελατών [Jain, 2010]**.

- **Σημασία:** Επιτρέπει στις επιχειρήσεις να ξεφύγουν από την παρωχημένη **Στατική Δημογραφική Στόχευση** και να δημιουργήσουν δυναμικές "**Personas**" βάσει **Συμπεριφοράς (Behavioral Segmentation)**. Για παράδειγμα, ο διαχωρισμός των πελατών σε "**Loyal**", "**Impulsive Shoppers**" και "**Discount Hunters**".



Εικόνα 2.12.a. Οπτικοποίηση της Συσταδοποίησης (Clustering) Δεδομένων σε Διακριτές Συστάδες.

Η παραπάνω **Εικόνα 2.12.a.** δείχνει πώς τα **Ακατέργαστα Δεδομένα** οργανώνονται σε **Διακριτές Συστάδες** (Clusters) βάσει **Ομοιότητας**.

β. Συστήματα Συστάσεων (Recommendation Systems)

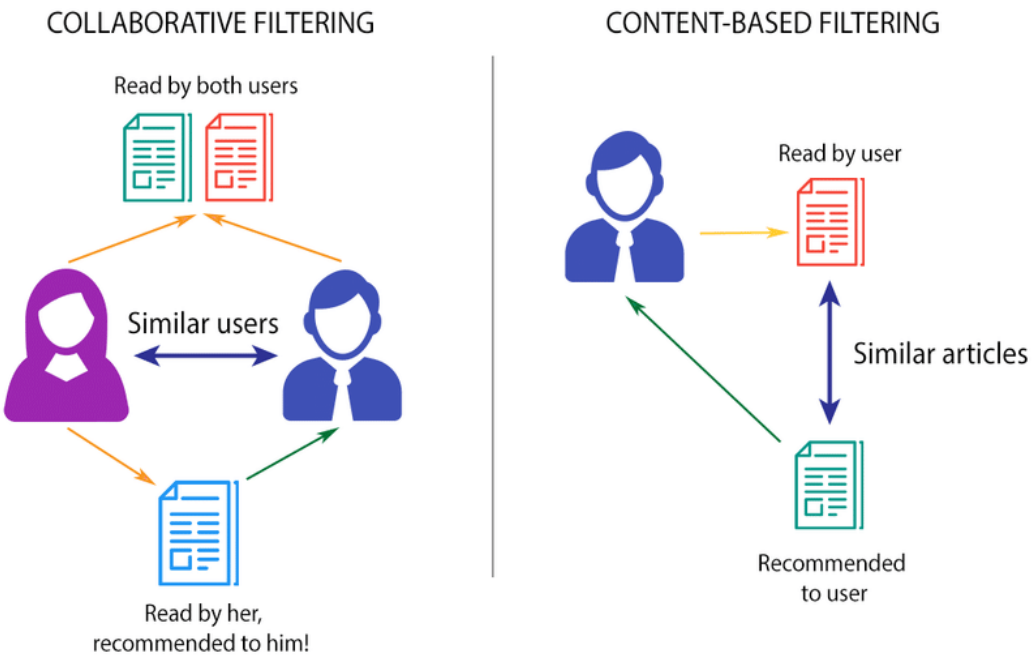
Τα **Συστήματα Συστάσεων (Recommendation Systems)** αποτελούν την πιο άμεση μορφή **Προσωποποίησης**, προβλέποντας την "**Προτίμηση**" ή τη "**Βαθμολογία**" που θα έδινε ένας χρήστης σε ένα προϊόν. Διακρίνονται σε τρεις κύριες κατηγορίες **[Adomavicius & Tuzhilin, 2005]**:

1. **Φιλτράρισμα Συνεργασίας (Collaborative Filtering):** Βασίζεται στην υπόθεση ότι χρήστες που συμφώνησαν στο **Παρελθόν** θα συμφωνήσουν και στο **Μέλλον**. Αναλύει τις **Αλληλεπιδράσεις Χρηστών-Προϊόντων** για να βρει ομοιότητες (π.χ. "Οι χρήστες που αγόρασαν το Α, αγόρασαν και το Β").

2. **Βάσει Περιεχομένου (Content-Based):** Προτείνει αντικείμενα παρόμοια με αυτά που άρεσαν στον χρήστη στο παρελθόν, βασιζόμενο στα **Χαρακτηριστικά** των αντικειμένων (π.χ. **Είδος Ταινίας**, **Μάρκα Προϊόντος**, **Χρώμα**).

3. **Υβριδικά Συστήματα (Hybrid Systems):** Συνδυάζουν τις δύο παραπάνω μεθόδους για να αντιμετωπίσουν εγγενή προβλήματα, όπως το "**Cold Start**" (έλλειψη δεδομένων για **Νέους Χρήστες** ή προϊόντα).

Η διαφορά μεταξύ των δύο βασικών προσεγγίσεων φαίνεται καθαρά στην **Εικόνα 2.12.b.**



Εικόνα 2.12.b. Σύγκριση μεταξύ Content-Based Filtering και Collaborative Filtering.

γ. Προβλεπτική Μοντελοποίηση (Predictive Modeling)

Η **Προβλεπτική Μοντελοποίηση (Predictive Modeling)** αξιοποιεί προηγμένες **Στατιστικές Τεχνικές** και **Εποπτευόμενη Μάθηση** για να προβλέψει μελλοντικά αποτελέσματα. Στο πλαίσιο της προσωποποίησης, εστιάζει κυρίως σε δύο τομείς **[Kelleher et al., 2015]**:

- **Propensity Scoring:** Η εκτίμηση της **Πιθανότητας (0-1)** ένας χρήστης να εκτελέσει μια συγκεκριμένη **Ενέργεια** (π.χ. να προβεί σε **Αγορά**, να κάνει **Κλικ**). Αυτό επιτρέπει

την **Ιεράρχηση των Πελατών (Lead Scoring)** και τη στόχευση μόνο όσων έχουν υψηλή πιθανότητα μετατροπής.

- **Πρόβλεψη Αξίας Ζωής (CLV Prediction):** Η εκτίμηση της συνολικής **Οικονομικής Αξίας** που αναμένεται να αποφέρει ένας πελάτης στην επιχείρηση σε βάθος χρόνου (**Customer Lifetime Value**), επιτρέποντας την **Εξατομίκευση του Προϋπολογισμού Μάρκετινγκ** ανά άτομο.

2.13. Αλγόριθμοι ML για Εφαρμογή σε CDPs

Η **Επιλογή** του κατάλληλου **Αλγορίθμου Συσταδοποίησης (Clustering)** είναι κρίσιμη για την **Εξαγωγή Αξιόπιστων Συμπερασμάτων** από τα **Δεδομένα** ενός **CDP**. Όπως είδαμε στην προηγούμενη Ενότητα 2.11, οι **Μη Εποπτευόμενες Τεχνικές** για την **Προσωποποίηση**, δεν χρειάζονται εκπαιδευμένα δεδομένα με "**Ετικέτες**" (**Labels**) για να λειτουργήσουν. Αντίθετα, παίρνουν **Ακατέργαστα Δεδομένα** και προσπαθούν να βρουν μόνες τους **Δομές** και **Συστάδες (Clusters)** βασισμένες στην **Ομοιότητα** ή την **Πυκνότητα των Σημείων (Δεδομένων)**.

Στην παρούσα Ενότητα αναλύονται οι **Μαθηματικές Θεμελιώσεις** των Αλγορίθμων που εξετάζονται στην παρούσα εργασία: **K-Means**, **K-Means++**, **K-Medoids** και **DBSCAN**, οι οποίοι ανήκουν στην Ομάδα των Τεχνικών **Μη Εποπτευόμενης Μάθησης**.

Συγκεκριμένα:

- **K-Means & K-Medoids:** Ψάχνουν για **Κέντρα Ομάδων** (σφαιρικές συστάδες).
- **K-Means++:** Είναι μια **Βελτιωμένη Μέθοδος Αρχικοποίησης** για τον K-Means (άρα εντάσσεται στην ίδια οικογένεια).
- **DBSCAN:** Ψάχνει για περιοχές **Υψηλής Πυκνότητας** (μπορεί να βρει σχήματα ακανόνιστα, όχι μόνο σφαίρες).

2.13.1. Αλγόριθμος K-Means

Η ιστορία του αλγορίθμου **K-Means** εκτείνεται σε πάνω από μισό αιώνα, αποτελώντας προϊόν παράλληλης έρευνας σε διαφορετικά επιστημονικά πεδία. Αν και ο όρος "**K-Means**" χρησιμοποιήθηκε για πρώτη φορά από τον **James MacQueen** το **1967 [MacQueen, 1967]**, η βασική ιδέα είχε ήδη διατυπωθεί νωρίτερα από τον **Hugo Steinhaus** το **1956** και κυρίως από τον **Stuart Lloyd** το **1957** στα εργαστήρια της **Bell Labs**. Η εργασία του Lloyd, αν και κυκλοφόρησε αρχικά ως εσωτερικό έγγραφο, δημοσιεύτηκε επίσημα πολύ αργότερα, το **1982**, καθιερώνοντας τον όρο "**Αλγόριθμος του Lloyd**" ως συνώνυμο της μεθόδου **[Lloyd, 1982]**.

Αρχικός Σκοπός και Επίλυση Προβλημάτων

Αρχικά, ο αλγόριθμος δεν σχεδιάστηκε για το Μάρκετινγκ, αλλά για την **Επεξεργασία Σήματος (Signal Processing)**.

- **Η Ανάγκη:** Ο κύριος στόχος ήταν η βελτιστοποίηση της **Διαμόρφωσης Παλμοκώδικα (Pulse Code Modulation - PCM)** για τη μεταδοση δεδομένων.
- **Η Λύση:** Ο αλγόριθμος έλυσε το πρόβλημα του **Κβαντισμού Δεδομένων (Data Quantization)**, επιτρέποντας τη μείωση του μεγέθους της πληροφορίας (Data Compression) μέσω της ομαδοποίησης παρόμοιων τιμών σήματος σε μια κοινή αντιπροσωπευτική τιμή (**Centroid**).

Μαθηματική Διατύπωση (Objective Function)

Ο **Αλγόριθμος** λειτουργεί **Επαναληπτικά** με στόχο να **Ελαχιστοποιήσει** τη **Συνάρτηση Κόστους**, γνωστή ως **Αδράνεια (Inertia)** ή **Άθροισμα Τετραγωνικών Σφαλμάτων (Sum of Squared Errors - SSE)**.

Η **Συνάρτηση** δίνεται από τον **Μαθηματικό Τύπο**:

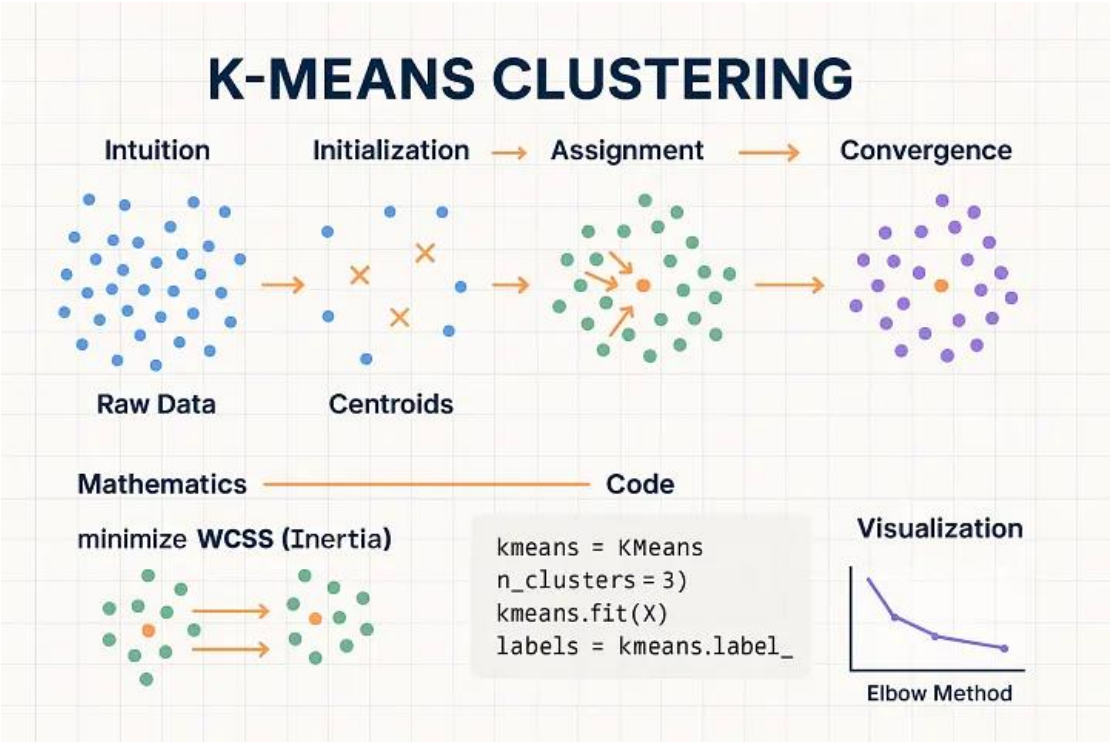
$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Εξίσωση 2.1

Όπου:

- $\mathbf{k}$: Ο αριθμός των **Συστάδων**.
- $\mathbf{n}$: Ο αριθμός των **Σημείων Δεδομένων**.
- $\mathbf{x}_i^{(j)}$: Το σημείο δεδομένων $\mathbf{i}$ που ανήκει στη συστάδα $\mathbf{j}$.
- $\mathbf{c}_j$: Το **Κεντροειδές (Centroid)** της συστάδας j.
- $\left\| \mathbf{x}_i^{(j)} - \mathbf{c}_j \right\|^2$: Η **Ευκλείδεια Απόσταση** μεταξύ του σημείου και του κέντρου.

Διαδικασία Λειτουργίας



Εικόνα 2.13.1. Η Επαναληπτική Διαδικασία του Αλγορίθμου K-Means: Από την τυχαία Αρχικοποίηση (αριστερά) στην τελική σύγκλιση των Κεντροειδών (δεξιά).

Η εκτέλεση του αλγορίθμου ακολουθεί τα εξής τέσσερα (4) βήματα [**Εικόνα 2.13.1.**]:

1. **Αρχικοποίηση (Initialization):** Επιλογή $\mathbf{k}$ τυχαίων σημείων ως αρχικά **Κέντρα**.

2. **Ανάθεση (Assignment):** Κάθε σημείο δεδομένων ανατίθεται στο πλησιέστερο κέντρο βάσει της **Ευκλείδειας Απόστασης**.
3. **Ενημέρωση (Update):** Υπολογισμός των νέων κέντρων ως ο **Μέσος Όρος (Mean)** των σημείων που ανήκουν σε κάθε συστάδα.
4. **Σύγκλιση (Convergence):** Επανάληψη των βημάτων 2 και 3 μέχρι τα κέντρα να σταματήσουν να μετακινούνται ή η αλλαγή να είναι αμελητέα.

Ανάλυση της Διαδικασίας και Βελτιστοποίηση

Όπως παρατηρούμε στην **Εικόνα 2.13.1.**, ο Αλγόριθμος ξεκινά με μια τυχαία τοποθέτηση και, μέσω διαδοχικών επαναλήψεων, μετακινεί τα **Κεντροειδή** στο "κέντρο βάρους" των δεδομένων, ελαχιστοποιώντας την **Ευκλείδεια Απόσταση**. Ωστόσο, προκύπτει ένα κρίσιμο ερώτημα: *«Ποιος είναι ο βέλτιστος αριθμός συστάδων (k);»*

Εδώ εισέρχεται η **Μέθοδος του Αγκώνα (Elbow Method)**. Επειδή ο **K-Means** δεν μπορεί να αποφασίσει μόνος του το πλήθος των ομάδων, τρέχουμε τον αλγόριθμο για διαφορετικές τιμές του $\mathbf{k}$ (π.χ. από 1 έως 10) και υπολογίζουμε το **Άθροισμα Τετραγωνικών Σφαλμάτων (SSE - Sum of Squared Errors)** για κάθε περίπτωση.

- Όσο **Αυξάνουμε** τις **Ομάδες**, το σφάλμα **Μειώνεται**.
- Στο **Σημείο** όπου η μείωση σταματά να είναι **Απότομη** και η **Καμπύλη** σχηματίζει μια "γωνία" (σαν αγκώνας), εκεί βρίσκεται ο **Βέλτιστος Αριθμός Συστάδων**.

Η Μέθοδος αυτή θα αναλυθεί διεξοδικά και θα εφαρμοστεί πρακτικά στο **Κεφάλαιο 4** της παρούσας εργασίας.

Περιορισμοί και Αδυναμίες

Παρά την τεράστια επιτυχία και την απλότητά του, ο **Standard K-Means** παρουσιάζει συγκεκριμένους **Εγγενείς Περιορισμούς**, τους οποίους απέτυχε να επιλύσει στην αρχική του μορφή **[Jain, 2010]**:

1. **Ευαισθησία στην Αρχικοποίηση (Initialization Sensitivity):** Ο αλγόριθμος συχνά παγιδεύεται σε **Τοπικά Ελάχιστα (Local Minima)**, πράγμα που σημαίνει ότι η τελική ομαδοποίηση εξαρτάται σε μεγάλο βαθμό από την τυχαία επιλογή των αρχικών κέντρων.
2. **Αδυναμία σε Μη Σφαιρικές Συστάδες (Non-Spherical Clusters):** Ο K-Means υποθέτει ότι οι ομάδες είναι σφαιρικές και έχουν παρόμοιο μέγεθος. Αποτυγχάνει να εντοπίσει συστάδες με **Πολύπλοκα Γεωμετρικά Σχήματα** ή διαφορετική **Πυκνότητα**.
3. **Ευαισθησία σε Ακραίες Τιμές (Outliers):** Επειδή ο αλγόριθμος χρησιμοποιεί τον **Μέσο Όρο (Mean)**, ακόμη και μία μόνο **Ακραία Τιμή** μπορεί να μετατοπίσει δραματικά το **Κεντροειδές**, στρεβλώνοντας ολόκληρη την ομαδοποίηση.
4. **Προκαθορισμός του K:** Απαιτεί από τον χρήστη να γνωρίζει εκ των προτέρων τον **Αριθμό των Συστάδων (k)**, κάτι που στα πραγματικά δεδομένα ενός **CDP** είναι συχνά άγνωστο.

2.13.2. Αλγόριθμος K-Means++

Η ανάγκη για βελτίωση της απόδοσης του κλασικού αλγορίθμου οδήγησε στην ανάπτυξη του **K-Means++**, ο οποίος παρουσιάστηκε το **2007** από τους ερευνητές του Πανεπιστημίου Stanford, **David Arthur** και **Sergei Vassilvitskii [Arthur & Vassilvitskii, 2007]**. Η εργασία τους με τίτλο *"k-means++: The Advantages of Careful Seeding"* θεωρείται πλέον ορόσημο στη βιβλιογραφία της **Μηχανικής Μάθησης**.

Αρχικός Σκοπός και Επίλυση Προβλημάτων

Ο κύριος στόχος της δημιουργίας του ήταν η επίλυση της **Ευαισθησίας στην Αρχικοποίηση (Initialization Sensitivity)** που μάστιζε τον κλασικό **K-Means.**

- **Το Πρόβλημα:** Στον απλό K-Means, η τυχαία επιλογή των αρχικών κέντρων μπορούσε να οδηγήσει σε **Κακά Τοπικά Ελάχιστα (Poor Local Minima)**, με αποτέλεσμα η ομαδοποίηση να είναι ανακριβής ή ο αλγόριθμος να αργεί υπερβολικά να συγκλίνει.

- **Η Λύση:** Ο **K-Means++** εισήγαγε τη μέθοδο της "**Προσεκτικής Σποράς**" (**Careful Seeding**). Αντί να επιλέγει τα κέντρα εντελώς τυχαία, επιλέγει το πρώτο κέντρο τυχαία και τα επόμενα με **Πιθανότητα** ανάλογη του τετραγώνου της απόστασής τους από τα ήδη επιλεγμένα κέντρα ($(\mathbf{D}(\mathbf{x})^2$). Αυτό διασφαλίζει ότι τα αρχικά **Κεντροειδή** είναι καλά διαχωρισμένα μεταξύ τους, οδηγώντας σε ταχύτερη **Σύγκλιση** και μείωση του τελικού σφάλματος.

Η **καινοτομία** αυτή δεν προσέφερε απλώς μια εμπειρική βελτίωση, αλλά παρείχε για πρώτη φορά ισχυρές **θεωρητικές εγγυήσεις**. Ειδικότερα, αποδείχθηκε μαθηματικά ότι ο **K-Means++** εξασφαλίζει μια λύση που είναι, στη χειρότερη περίπτωση, $\mathbf{O}(\log \mathbf{k})$ ανταγωνιστική ως προς τη βέλτιστη δυνατή **συσταδοποίηση**. Λόγω αυτής της αποδεδειγμένης **σταθερότητας**, η συγκεκριμένη μέθοδος αποτελεί πλέον την προεπιλεγμένη (default) τεχνική αρχικοποίησης στις κορυφαίες βιβλιοθήκες **Επιστήμης Δεδομένων**, όπως το Scikit-learn (μέσω της παραμέτρου init='k-means++'), το οποίο και ενσωματώσαμε για την υλοποίηση του δικού μας λογισμικού.

Μαθηματική Διατύπωση (Seeding Process)

Ενώ ο τελικός στόχος παραμένει η **Ελαχιστοποίηση** του **Αθροίσματος Τετραγωνικών Σφαλμάτων (SSE)**, όπως στον κλασικό K-Means, η μαθηματική καινοτομία του K-Means++ εντοπίζεται στον τρόπο επιλογής των αρχικών κέντρων. Συγκεκριμένα, μετά την τυχαία επιλογή του πρώτου κέντρου, τα επόμενα κέντρα επιλέγονται βάσει μιας **Κατανομής Πιθανότητας**.

Έστω $\mathbf{D}(\mathbf{x})$η συντομότερη **Ευκλείδεια Απόσταση** ενός **Σημείου Δεδομένων** $\mathbf{x}$ από το πλησιέστερο, ήδη επιλεγμένο κέντρο. Η **Πιθανότητα (Probability)** $\mathbf{P}(\mathbf{x})$ να επιλεγεί το σημείο $\mathbf{x}$ ως το επόμενο κέντρο δίνεται από τον τύπο **[Arthur & Vassilvitskii, 2007]**:

$$\mathbf{P}(\mathbf{x}) = \frac{\mathbf{D}(\mathbf{x})^2}{\sum_{\mathbf{x}' \in \boldsymbol{X}} \mathbf{D}(\mathbf{x}')^2} \qquad \textbf{(Εξίσωση 2.2)}$$

Όπου:

- $\mathbf{D(x)^2}$: Το **Τετράγωνο** της **Απόστασης** του **Σημείου** από το πλησιέστερο κέντρο.
- $\sum \boldsymbol{D(x')^2}$: Το **Άθροισμα** των **Τετραγώνων** των **Αποστάσεων** για όλα τα σημεία που δεν έχουν επιλεγεί ακόμη.

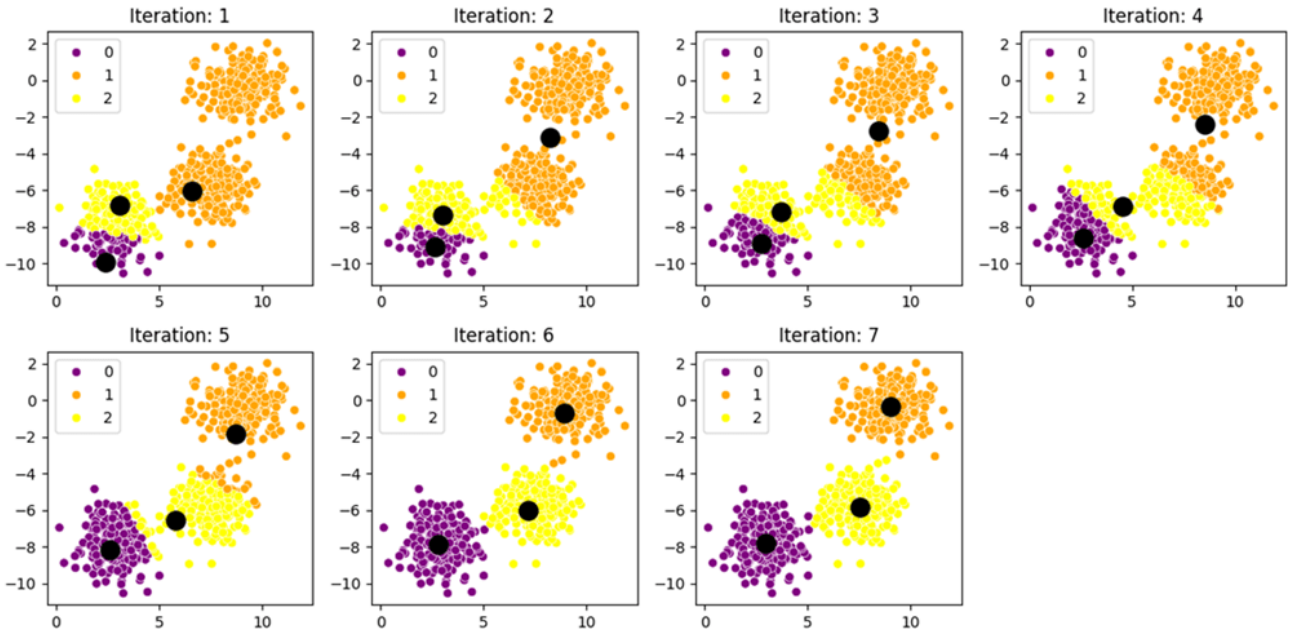
Αυτός ο τύπος διασφαλίζει ότι **Σημεία** που βρίσκονται **Μακριά** από τα υπάρχοντα κέντρα έχουν πολύ μεγαλύτερη **Πιθανότητα** να επιλεγούν ως νέα **Κέντρα**, οδηγώντας σε καλύτερη **Διασπορά** και ταχύτερη σύγκλιση.

Διαφορές ανάμεσα σε K-Means και K-Means++

Οι διαφορές ανάμεσα στους δύο (2) Αλγορίθμους γίνονται άμεσα κατανοητές αν παρατηρήσουμε τις δύο παρακάτω **Εικόνα 2.13.2.a** και **Εικόνα 2.13.2.b.**

(A) Ανάλυση Standard K-Means: Η Διαδικασία των 7 Επαναλήψεων

Στην πρώτη **Απεικόνιση** (**Εικόνα 2.13.2.a**), βλέπουμε τη λειτουργία του κλασικού αλγορίθμου, ο οποίος απαιτεί συνολικά **7 Επαναλήψεις (Iterations)** για να συγκλίνει. Η καθυστέρηση αυτή οφείλεται στην τυχαία αρχική τοποθέτηση των κέντρων, τα οποία έπεσαν



Εικόνα 2.13.2.a. Η Επαναληπτική Διαδικασία του Αλγορίθμου K-Means (7 Επαναλήψεις)

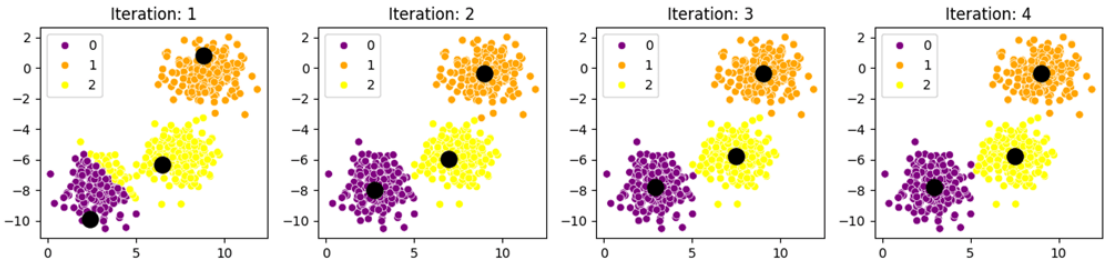
κοντά το ένα στο άλλο. Για να διορθώσει αυτό το σφάλμα, ο **Αλγόριθμος** αναγκάζεται να εκτελέσει επτά φορές τον πλήρη **Κύκλο Επανάληψης**, ο οποίος αποτελείται από **Δύο** (2) **Φάσεις**:

1. **Φάση Ανάθεσης (Assignment Step):** Ο **Αλγόριθμος** μετρά την **Ευκλείδεια Απόσταση** κάθε **Σημείου** από τα **Κέντρα** και αναθέτει τα σημεία στο πλησιέστερο **Κεντροειδές**. Λόγω της κακής αρχικής θέσης, πολλά σημεία αλλάζουν συνεχώς "ιδιοκτήτη" (**Συστάδα**) σε κάθε βήμα.
2. **Φάση Ενημέρωσης (Update Step):** Υπολογίζεται ο νέος **Μέσος Όρος** των σημείων κάθε **Ομάδας** και το **Κεντροειδές** μετακινείται στη νέα θέση.

Στην **Εικόνα 2.13.2.a** παρατηρούμε ότι τα κέντρα "ταξιδεύουν" μεγάλη **Απόσταση** πάνω στο **Διάγραμμα** μέχρι να σταθεροποιηθούν, σπαταλώντας υπολογιστικούς πόρους.

(Β) Ανάλυση K-Means++: Η Βελτιστοποίηση των 4 Επαναλήψεων

Στη δεύτερη **Απεικόνιση (Εικόνα 2.13.2.b)**, η υπεροχή του **K-Means++** είναι εμφανής, καθώς ο αλγόριθμος ολοκληρώνει τη διαδικασία σε μόλις **4 Επαναλήψεις (Iterations)** — σχεδόν στο μισό χρόνο.



Εικόνα 2.13.2.b. Η Επαναληπτική Διαδικασία του Αλγορίθμου K-Means++ (4 Επαναλήψεις)

Αυτό επιτυγχάνεται χάρη στη διαδικασία της **Προσεκτικής Σποράς (Seeding)** που προηγήθηκε. Τα **Αρχικά Κέντρα** τοποθετήθηκαν εξαρχής μακριά το ένα από το άλλο, καλύπτοντας στρατηγικά τον χώρο των **Δεδομένων**. Ως αποτέλεσμα:

- Κατά την πρώτη **Φάση Ανάθεσης**, η πλειοψηφία των σημείων συνδέθηκε αμέσως με το σωστό τελικό κέντρο.
- Οι επακόλουθες **Μετακινήσεις** των **Κεντροειδών** (**Φάση Ενημέρωσης**) ήταν μικρές και **Στοχευμένες**.

Συμπερασματικά, ο **K-Means++** δεν χρειάστηκε να "διορθώσει" κακές αρχικές εκτιμήσεις, αλλά μόνο να "ραφινάρει" (fine-tune) μια ήδη καλή λύση, αποδεικνύοντας την **Υπολογιστική** του **Αποτελεσματικότητα** για μεγάλα **Σύνολα Δεδομένων** (**Big Data**).

Περιορισμοί και Υστερήσεις (Limitations and Weaknesses)

Παρόλο που ο K-Means++ αποτελεί τη **Βιομηχανική Πρότυπη Μέθοδο (Industry Standard)** σήμερα (και είναι η default επιλογή στη βιβλιοθήκη **Scikit-Learn** που θα χρησιμοποιήσουμε), δεν είναι απαλλαγμένος από αδυναμίες:

1. **Υπολογιστικό Κόστος Αρχικοποίησης:** Η **Διαδικασία Επιλογής** των Αρχικών Κέντρων είναι πιο Χρονοβόρα σε σχέση με την απλή **Τυχαία Επιλογή**, καθώς απαιτεί τον **Υπολογισμό Αποστάσεων** για όλα τα **Σημεία Δεδομένων** σε κάθε βήμα επιλογής **Κέντρου**. Σε εξαιρετικά μεγάλα **Big Data** σύνολα, αυτό μπορεί να δημιουργήσει καθυστέρηση στην εκκίνηση.
2. **Κληρονομικοί Γεωμετρικοί Περιορισμοί:** Καθώς βασίζεται στην ίδια λογική με τον **K-Means**, διατηρεί την αδυναμία εντοπισμού **Μη Σφαιρικών Συστάδων** ή συστάδων με διαφορετική **Πυκνότητα**.
3. **Ευαισθησία σε Ακραίες Τιμές (Outliers):** Αν και βελτιωμένος, παραμένει ευάλωτος σε **Ακραίες Τιμές**, οι οποίες αν επιλεγούν ως **Αρχικά Κέντρα** (λόγω της μεγάλης απόστασής τους), μπορεί να επηρεάσουν αρνητικά την ποιότητα των **Clusters**.

Ακριβώς λόγω αυτών των **Εγγενών Περιορισμών**, στην παρούσα μελέτη κρίνεται απαραίτητη η εξέταση εναλλακτικών Αλγορίθμων. Η αδυναμία διαχείρισης του **Θορύβου** και των ακραίων τιμών αντιμετωπίζεται αποτελεσματικότερα από τον αλγόριθμο **K-Medoids** (που χρησιμοποιεί πραγματικά σημεία αντί για μέσους όρους), ενώ η ανάγκη για εντοπισμό συστάδων **Ακανόνιστου Σχήματος** (Arbitrary Shape) και διαφορετικής πυκνότητας καλύπτεται από τον **DBSCAN**, οι οποίοι αναλύονται στις ενότητες που ακολουθούν.

2.13.3. K-Medoids (Partinioning Around Medoids - PAM)

Καθώς η εφαρμογή του K-Means άρχισε να επεκτείνεται σε πραγματικά δεδομένα, έγινε σαφές ότι η χρήση του **Μέσου Όρου** καθιστούσε τον αλγόριθμο ευάλωτο σε θόρυβο. Ως απάντηση σε αυτό το πρόβλημα, οι Βέλγοι στατιστικολόγοι **Leonard Kaufman** και **Peter J. Rousseeuw** ανέπτυξαν τον αλγόριθμο **K-Medoids**, γνωστότερο ως **PAM (Partitioning Around Medoids)**. Η μέθοδος παρουσιάστηκε αρχικά το **1987** και θεμελιώθηκε πλήρως στο βιβλίο-ορόσημο *"Finding Groups in Data"* το **1990 [Kaufman & Rousseeuw, 1990]**.

Αρχικός Σκοπός και Επίλυση Προβλημάτων

Ο K-Medoids σχεδιάστηκε για να προσφέρει **Στιβαρότητα (Robustness)** εκεί που ο K-Means αποτύγχαινε.

- **Το Πρόβλημα:** Στον K-Means, μια μοναδική **Ακραία Τιμή (Outlier)** με εξαιρετικά υψηλή τιμή (π.χ. ένας πελάτης που ξόδεψε 1.000.000€ ενώ οι υπόλοιποι 50€) "τραβάει" το **Κεντροειδές** προς το μέρος της, στρεβλώνοντας τα όρια των ομάδων.
- **Η Λύση:** Αντί για ένα τεχνητό σημείο (Μέσος Όρος), ο K-Medoids χρησιμοποιεί ως κέντρο της ομάδας το **Medoid (Μεντοειδές)**. Το Medoid είναι το **Πιο Αντιπροσωπευτικό Πραγματικό Αντικείμενο** της συστάδας (δηλαδή ένα υπαρκτό σημείο δεδομένων) που έχει τη μικρότερη μέση απόσταση από τα υπόλοιπα.

Μαθηματική Διατύπωση (Cost Function)

Σε αντίθεση με τον K-Means που ελαχιστοποιεί το τετράγωνο της Ευκλείδειας απόστασης, ο K-Medoids στοχεύει στην ελαχιστοποίηση της **Απόλυτης Ανομοιότητας (Dissimilarity)**. Η συνάρτηση κόστους που ελαχιστοποιείται είναι το άθροισμα των αποστάσεων (συχνά **Απόσταση Manhattan**) μεταξύ κάθε σημείου και του αντιπροσωπευτικού Medoid.

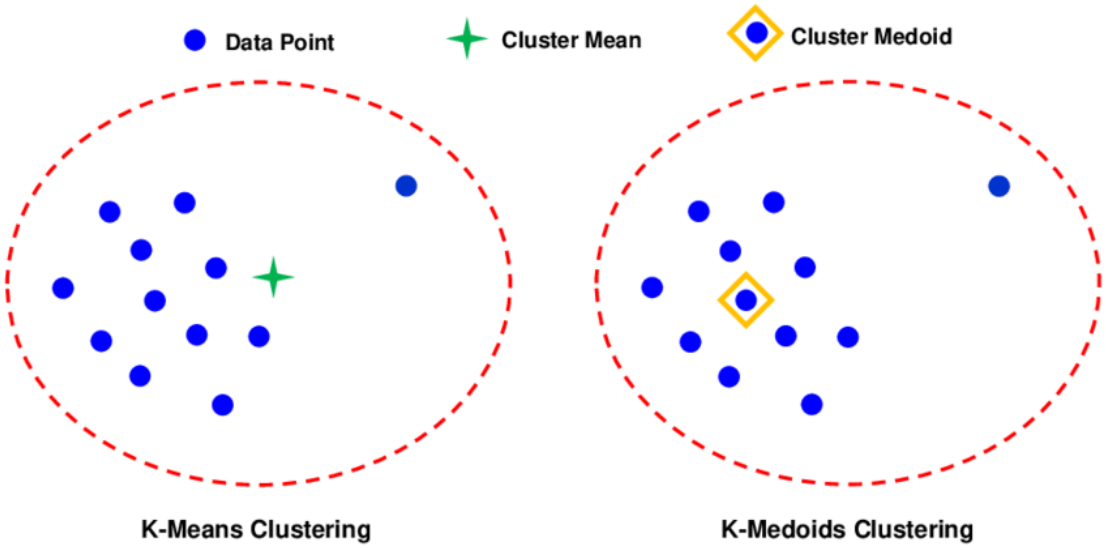
Η συνάρτηση δίνεται από τον τύπο:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} d\left(x_i^{(j)}, m_j\right)$$ (Εξίσωση 2.3)

Όπου:

- d: Η συνάρτηση **Απόστασης** (π.χ. Ευκλείδεια ή Manhattan) μεταξύ δύο σημείων.
- $x_i^{(j)}$: Το σημείο δεδομένων i που ανήκει στη συστάδα j.
- m_j: Το **Medoid** της συστάδας j (το οποίο είναι *πραγματικό* σημείο του συνόλου δεδομένων).

Αυτή η διαφορά καθιστά τον αλγόριθμο λιγότερο ευαίσθητο σε ακραίες τιμές, καθώς η αλλαγή θέσης του κέντρου απαιτεί την επιλογή ενός νέου υπαρκτού σημείου και δεν επηρεάζεται γραμμικά από μεγάλες αριθμητικές αποκλίσεις.



Εικόνα 2.13.3.α. Η Διαφορά στη Στιβαρότητα (Robustness). Αριστερά (K-Means): Η παρουσία της Ακραίας Τιμής (Outlier) "τραβάει" το Κέντρο (Mean) μακριά από την Πραγματική Ομάδα, στρεβλώνοντας το αποτέλεσμα. Δεξιά (K-Medoids): Το Κέντρο

(Medoid) παραμένει σταθερό μέσα στην Ομάδα, καθώς είναι πραγματικό αντικείμενο και δεν επηρεάζεται από την Απόσταση της Ακραίας Τιμής (Outlier)

Η Επίπτωση των Ακραίων Τιμών

Στη συγκεκριμένη **Απεικόνιση**, **2.13.3.α** αναδεικνύεται η **Θεμελιώδη Διαφορά** συμπεριφοράς των δύο **Αλγορίθμων** παρουσία **Θορύβου (Noise)**, επιβεβαιώνοντας τα ευρήματα της βιβλιογραφίας σχετικά με την «ευρωστία» των μεθόδων συσταδοποίησης **[Han et al., 2011]**:

1. **Αριστερά (Standard K-Means):** Παρατηρούμε ότι η παρουσία έστω και μίας **Ακραίας Τιμής (Outlier)** —του σημείου που βρίσκεται απομονωμένο δεξιά— έχει μετατοπίσει το **Κεντροειδές (Centroid)** (συμβολίζεται με τον σταυρό) μακριά από τον πυρήνα της κύριας ομάδας.

    - **Η Αιτία:** Όπως επισημαίνει ο **Jain [2010]**, επειδή ο K-Means προσπαθεί να ελαχιστοποιήσει το **Τετράγωνο** των αποστάσεων (SSE), οι μεγάλες αποστάσεις "τιμωρούνται" αυστηρά. Έτσι, ο αλγόριθμος "θυσιάζει" τη συνοχή της πλειοψηφίας για να ικανοποιήσει την απομακρυσμένη τιμή.
    - **Το Αποτέλεσμα:** Το κέντρο καταλήγει σε μια περιοχή "ανύπαρκτων δεδομένων" (στο κενό), δίνοντας μια στρεβλή εικόνα για το "προφίλ" του μέσου πελάτη.

2. **Δεξιά (K-Medoids / PAM):** Στην περίπτωση αυτή, το κέντρο της ομάδας (το **Medoid**, με πράσινο χρώμα) παραμένει σταθερά αγκυροβολημένο μέσα στην πυκνή συστάδα των δεδομένων.

    - **Η Αιτία:** Σύμφωνα με τους δημιουργούς του αλγορίθμου, **Kaufman και Rousseeuw [1990]**, ο K-Medoids υποχρεούται να επιλέξει ένα **Πραγματικό Σημείο Δεδομένων** ως κέντρο και ελαχιστοποιεί το άθροισμα των απόλυτων αποστάσεων (όχι των τετραγώνων).
    - **Το Αποτέλεσμα:** Η ακραία τιμή αναγνωρίζεται ως μέλος της ομάδας, αλλά δεν έχει τη δύναμη να "τραβήξει" το κέντρο προς το μέρος της. Αυτό προσδίδει στον αλγόριθμο την ιδιότητα της **Στιβαρότητας (Robustness)**,

εξασφαλίζοντας ότι η ομαδοποίηση αντιπροσωπεύει πιστά την πλειοψηφία των χρηστών, αγνοώντας τις σπάνιες εξαιρέσεις **[Velmurugan & Santhanam, 2010]**.

Ο Μηχανισμός Λειτουργίας: Ο Αλγόριθμος PAM

Στη βιβλιογραφία, ο **K-Medoids** υλοποιείται σχεδόν αποκλειστικά μέσω της μεθόδου **PAM (Partitioning Around Medoids)**. Η καινοτομία των **Kaufman και Rousseeuw** έγκειται στο ότι μετέτρεψαν το πρόβλημα της ομαδοποίησης σε πρόβλημα **Αναζήτησης (Search Problem)**, το οποίο εκτελείται σε δύο φάσεις **[Kaufman & Rousseeuw, 1990]**:

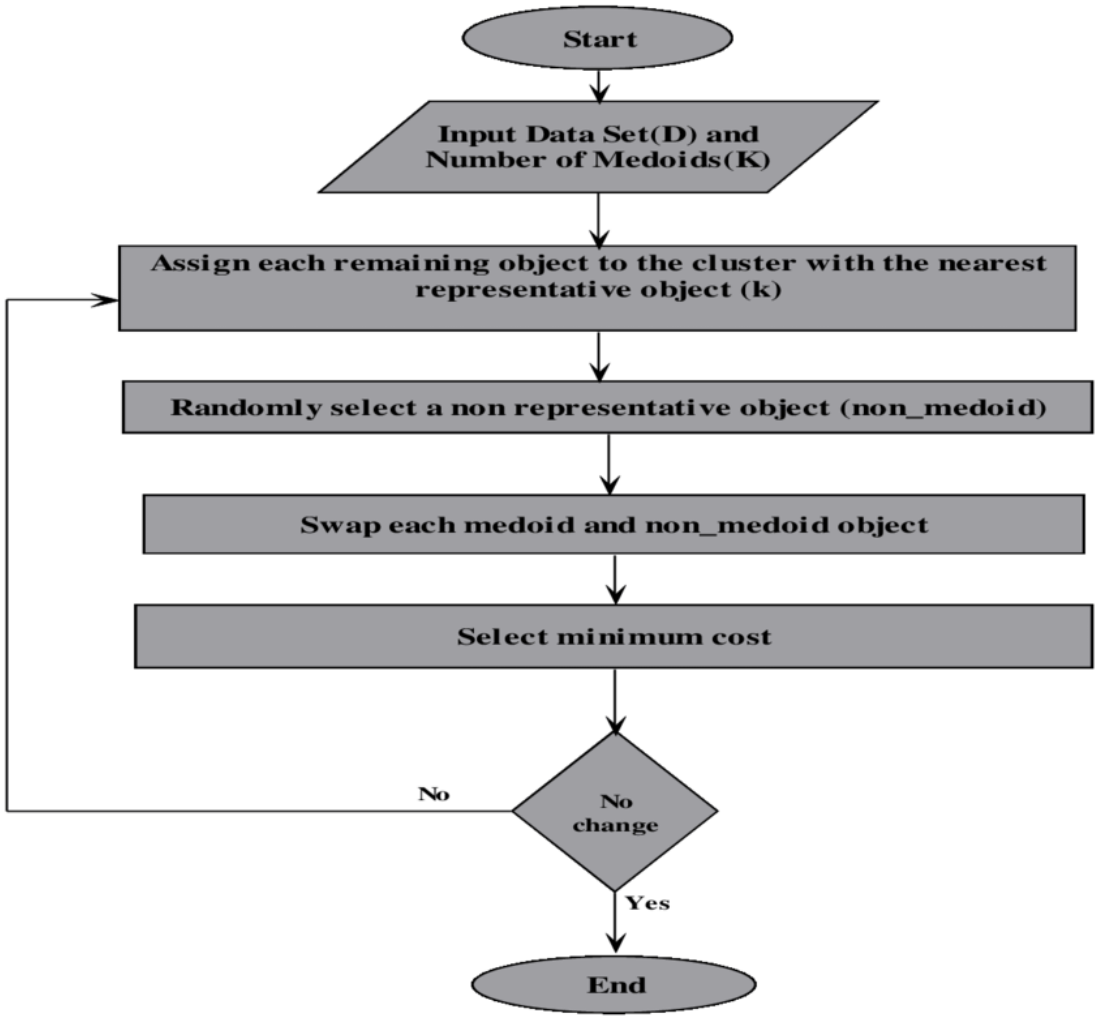
1. **Φάση Κατασκευής (Build Phase):** Επιλέγονται τυχαία k σημεία από το σύνολο δεδομένων ως τα αρχικά **Medoids**.
2. **Φάση Ανταλλαγής (Swap Phase):** Ο αλγόριθμος προσπαθεί να βελτιώσει την ποιότητα της ομαδοποίησης εξετάζοντας **Κάθε Πιθανή Αντικατάσταση**.
    - Επιλέγει ένα **μη - medoid Σημείο O**.
    - Δοκιμάζει να αντικαταστήσει το τρέχον **medoid M** με το **O**.
    - Υπολογίζει το **Συνολικό Κόστος Ανταλλαγής (Total Cost of Swapping)**.
    - Αν το **Κόστος** μειώνεται, η ανταλλαγή γίνεται **Μόνιμη**. Η διαδικασία επαναλαμβάνεται μέχρι να μην υπάρχει άλλη **Εφικτή Βελτίωση**.

Αυτή η εξαντλητική διαδικασία "**Δοκιμής και Αντικατάστασης**" (Trial and Swap) είναι που προσφέρει την υψηλή **Ακρίβεια** του **Αλγορίθμου**, αλλά και την **Υψηλή Υπολογιστική Επιβάρυνση**.

Η **Εικόνα 2.13.3.b.** αποτελεί την οπτική αναπαράσταση του μηχανισμού που καθιστά τον K-Medoids τόσο ακριβή αλλά και κοστοβόρο. Το διάγραμμα χωρίζεται σε δύο διακριτά μέρη που αντιστοιχούν στις φάσεις που αναλύσαμε:

1. **Πάνω Μέρος (Build Phase):**

Εδώ βλέπουμε την **Αρχικοποίηση**. Ο αλγόριθμος δεν ξεκινά απλά τυχαία, αλλά επιλέγει **k Αντικείμενα** (medoids) για να δημιουργήσει το αρχικό σετ.



Εικόνα 2.13.3.b. Το Διάγραμμα Ροής του Αλγορίθμου PAM. Διακρίνονται η φάση Αρχικής Επιλογής (Build) και η Επαναληπτική Διαδικασία ελέγχου για καλύτερο Αντιπρόσωπο (Swap).

2. **Κάτω Μέρος (Swap Phase):**

Αυτό είναι το "καρδιά" του αλγορίθμου και το σημείο που φαίνεται το υπολογιστικό κόστος. Το διάγραμμα δείχνει τον **Βρόχο Επανάληψης (Loop)**:

- Ο αλγόριθμος εξετάζει **κάθε πιθανό ζεύγος** (medoid, non-medoid).
- Υπολογίζει το κόστος: *"Αν κάνω αυτή την αλλαγή, θα μειωθεί το συνολικό σφάλμα;"*

- Αν η απάντηση είναι **ΝΑΙ** (Total Cost < 0), τότε γίνεται η Αντικατάσταση (Swap) και ο κύκλος ξεκινάει πάλι.
- Αν η απάντηση είναι **ΟΧΙ**, ο αλγόριθμος τερματίζει.

Αυτή η **Διαδικασία** εγγυάται ότι θα βρεθεί το πιο αντιπροσωπευτικό "Κέντρο", αλλά όπως φαίνεται από τα βέλη που γυρίζουν πίσω, απαιτεί πολλούς υπολογισμούς.

Περιορισμοί και Υστερήσεις (Limitations and Weaknesses)

Παρά την **Ανθεκτικότητά** του στο **Θόρυβο**, ο **K-Medoids** (στην εκδοχή **PAM**) παρουσιάζει ένα σημαντικό **Μειονέκτημα** που τον καθιστά δύσχρηστο σε περιβάλλοντα **Big Data** όπως τα **CDPs**:

1. **Υψηλή Υπολογιστική Πολυπλοκότητα:** Για να βρει το βέλτιστο Medoid, ο αλγόριθμος πρέπει να υπολογίσει τις αποστάσεις μεταξύ **Κάθε Ζεύγους Σημείων** στη συστάδα. Αυτό έχει πολυπλοκότητα της τάξης του $\mathbf{O(k(n-k)^2)}$, καθιστώντας τον εξαιρετικά αργό για μεγάλα σύνολα δεδομένων (**n**).
2. **Ανάγκη για Παραλλαγές:** Για να λυθεί το παραπάνω πρόβλημα, έχουν αναπτυχθεί παραλλαγές όπως ο **CLARA (Clustering LARge Applications)** και ο **CLARANS**, οι οποίες εφαρμόζουν τη μέθοδο σε δείγματα των δεδομένων αντί για το σύνολο **[Ng & Han, 2002]**.

Εναλλακτικές Προσεγγίσεις και Βελτιστοποιήσεις

Για την αντιμετώπιση της υψηλής **Υπολογιστικής Πολυπλοκότητας** του αλγορίθμου **PAM** σε περιβάλλοντα **Big Data**, η **Βιβλιογραφία** προτείνει **δύο (2) Κύριες Παραλλαγές** που θυσιάζουν μέρος της **Ακρίβειας** για χάρη της **Ταχύτητας**:

1. **CLARA (Clustering LARge Applications):** Αναπτύχθηκε επίσης από τους **Kaufman και Rousseeuw [1990]**. Αντί να επεξεργαστεί όλο το σύνολο δεδομένων, ο αλγόριθμος επιλέγει τυχαία ένα μικρό **Δείγμα (Sample)** δεδομένων και εφαρμόζει τον **PAM** μόνο σε αυτό. Η διαδικασία επαναλαμβάνεται πολλές φορές και επιλέγεται το καλύτερο αποτέλεσμα. Το **Μειονέκτημά** του είναι ότι αν τα πραγματικά κέντρα

(medoids) δεν επιλεγούν στο δείγμα, η **Τελική Ομαδοποίηση** θα είναι κατά προσέγγιση και όχι η **Βέλτιστη**.

2. **CLARANS (Clustering Large Applications based on RANdomized Search):** Προτάθηκε από τους **Ng και Han [2002]** και συνδυάζει τη λογική του **PAM** με τυχαία αναζήτηση σε **Γράφους**. Ο **Αλγόριθμος** δεν ελέγχει ***όλους*** τους γείτονες για **Αντικατάσταση** (Swap), αλλά μόνο ένα τυχαίο υποσύνολο αυτών. Θεωρείται πιο αποδοτικός από τον **CLARA** σε πολύ μεγάλα **Σύνολα Δεδομένων**.

Γιατί επιλέχθηκε ο PAM στην Παρούσα Εργασία;

Παρά την ύπαρξη των ταχύτερων παραλλαγών **CLARA** και **CLARANS**, στην παρούσα **Διπλωματική Εργασία** επιλέχθηκε η υλοποίηση του **Κλασικού Αλγορίθμου PAM (Partitioning Around Medoids)** για τους εξής λόγους:

1. **Μέγιστη Ακρίβεια και Ποιότητα Ομαδοποίησης:** Ο κύριος στόχος της μελέτης μας είναι η **Ακριβής Σκιαγράφηση** των προφίλ **Πελατών** (Customer Personas) στο **CDP**. Ο αλγόριθμος **PAM** εγγυάται την εύρεση του **Ολικού Ελαχίστου (Global Minimum)** της συνάρτησης κόστους, σε αντίθεση με τον **CLARA** που βασίζεται σε δειγματοληψία και μπορεί να οδηγήσει σε προσεγγιστικές λύσεις. Δεδομένου ότι η ποιότητα της στόχευσης στο **Marketing** εξαρτάται από την ακρίβεια των **Clusters**, προκρίθηκε η "**Ποιότητα έναντι της Ταχύτητας**".
2. **Έλεγχος Στιβαρότητας (Robustness Check):** Θέλαμε να συγκρίνουμε ευθέως την **Ανθεκτικότητα** του **Medoid** έναντι του **Centroid** (K-Means) απέναντι σε **Ακραίες Τιμές**. Η χρήση μιας **Μεθόδου Δειγματοληψίας** (όπως η **CLARA**) θα εισήγαγε έναν επιπλέον παράγοντα **Αβεβαιότητας** (Sampling Error), δυσκολεύοντας την **Άμεση** και **Δίκαιη Σύγκριση** των δύο μεθοδολογιών.
3. **Μέγεθος Δεδομένων:** Το Σύνολο Δεδομένων που χρησιμοποιήθηκε στην εφαρμογή μας, αν και αντιπροσωπευτικό, δεν απαγορεύει τη χρήση του **PAM**. Η **Υπολογιστική Ισχύς** των σύγχρονων **Συστημάτων** επιτρέπει την εκτέλεση του **Αλγορίθμου** σε λογικούς χρόνους για **Datasets Μεσαίου Μεγέθους**, καθιστώντας περιττή την προσφυγή σε προσεγγιστικές μεθόδους.

2.13.4. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Η επιτακτική ανάγκη για την οριστική **Υπέρβαση** των **Άκαμπτων Γεωμετρικών Υποθέσεων** (Rigid Geometric Assumptions) που δέσμευαν τους κλασικούς **Αλγορίθμους Διαμερισμού** (όπως τον K-Means), οδήγησε στην ανάπτυξη του καινοτόμου **Αλγορίθμου DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**.

Ο **Αλγόριθμος** παρουσιάστηκε για πρώτη φορά το **1996** στο διεθνούς φήμης συνέδριο **KDD (Knowledge Discovery and Data Mining)** από την ερευνητική ομάδα των **Martin Ester, Hans-Peter Kriegel, Jörg Sander** και **Xiaowei Xu [Ester et al., 1996]**. Η συγκεκριμένη **Εργασία** δεν αποτέλεσε απλώς μια βελτίωση, αλλά εισήγαγε έναν νέο **Ορισμό Συστάδας**: όχι πλέον ως «ομάδα σημείων κοντά σε ένα κέντρο», αλλά ως «**Περιοχή Υψηλής Πυκνότητας** που διαχωρίζεται από Περιοχές Χαμηλής Πυκνότητας».

Η **Διαχρονική Αξία** αυτής της προσέγγισης αναγνωρίστηκε παγκοσμίως όταν η εργασία τιμήθηκε με το διάσημο **Test of Time Award** το **2014**, επιβεβαιώνοντας ότι ο **DBSCAN** παραμένει, σχεδόν **Δύο** (**2**) **Δεκαετίες** μετά, ένας από τους πιο **Επιδραστικούς Αλγορίθμους** στην ιστορία της **Εξόρυξης Δεδομένων (Data Mining)**.

Αρχικός Σκοπός και Επίλυση Προβλημάτων

Ο **Σχεδιασμός** του **DBSCAN** ήρθε να καλύψει **Δύο** (**2**) **Κρίσιμα Κενά** που οι παραδοσιακοί **Αλγόριθμοι Διαμερισμού** (**Partitioning Algorithms**), όπως ο **K-Means** και ο **PAM**, αδυνατούσαν να διαχειριστούν αποτελεσματικά:

1. **Ανακάλυψη Συστάδων Ακανόνιστου Σχήματος (Arbitrary Shape Clusters):**

   Οι **Αλγόριθμοι** που βασίζονται σε **Κέντρα** (Centroid-based) περιορίζονται εγγενώς στον εντοπισμό **Κυρτών Συστάδων (Convex Clusters)**, δηλαδή ομάδων με **Σφαιρική Γεωμετρία**.

Αντίθετα, ο **DBSCAN** απαλλάσσεται από αυτόν τον **Περιορισμό**. Ακολουθώντας τη **Χωρική Συνέχεια (Spatial Continuity)**, μπορεί να εντοπίσει **Συστάδες** με περίπλοκη μορφολογία, όπως σχήματα "φιδιού", δακτυλίου ή ακανόνιστες **Γεωγραφικές Κατανομές**, αρκεί να διατηρείται η **Συνεκτικότητα της Πυκνότητας (Density Connectivity)** των **Σημείων**. Έτσι, επιτυγχάνεται η πιστή αναπαράσταση της **Πραγματικής Τοπολογίας** των Δεδομένων, αποφεύγοντας τον εσφαλμένο κατακερματισμό ενιαίων **Φυσικών Ομάδων** που συχνά προκαλούν οι μέθοδοι που βασίζονται αποκλειστικά στην **Ευκλείδεια Απόσταση**.

2. **Αυτόματη Διαχείριση Θορύβου (Noise Handling & Robustness):**

Σε αντίθεση με τον **K-Means**, ο οποίος "εξαναγκάζει" κάθε Σημείο **Δεδομένων** να ανήκει υποχρεωτικά σε κάποια **Ομάδα** (ακόμα και αν αυτό αποτελεί σφάλμα μέτρησης), ο **DBSCAN** εισάγει την έννοια του **Θορύβου (Noise)**.

Έχει την ικανότητα να αναγνωρίζει τα **Απομονωμένα Σημεία (Outliers)** που βρίσκονται σε περιοχές **Χαμηλής Πυκνότητας** και να τα εξαιρεί αυτόματα από την ανάλυση. Αυτό διασφαλίζει την **Ομοιογένεια (Homogeneity)** των παραγόμενων **Συστάδων**, καθώς αυτές δεν "μολύνονται" από άσχετα **Δεδομένα**.

3. **Δεν Απαιτεί την Είσοδο της Παραμέτρου k:**

Ίσως το σημαντικότερο **Πλεονέκτημα** σε πραγματικές συνθήκες. Ενώ οι **K-Means** και **PAM** απαιτούν από τον χρήστη να γνωρίζει εκ των προτέρων τον **Αριθμό** των **Συστάδων** — κάτι συχνά αδύνατο σε άγνωστα **δεδομένα** — ο **DBSCAN** τον ανακαλύπτει **Δυναμικά**. Ο αριθμός των ομάδων προκύπτει ως φυσικό αποτέλεσμα της **Κατανομής Πυκνότητας** των **Δεδομένων** και όχι ως αυθαίρετη υπόθεση του αναλυτή.

Μαθηματική Διατύπωση (Density Definitions)

Ο **DBSCAN** δεν βασίζεται στην **Ελαχιστοποίηση** μιας **Συνάρτησης Κόστους** (όπως το **SSE**), αλλά σε αυστηρούς **Ορισμούς Πυκνότητας** και **Συνδεσιμότητας**.

Η λειτουργία του καθορίζεται από **Δύο** (**2**) **Παραμέτρους**:

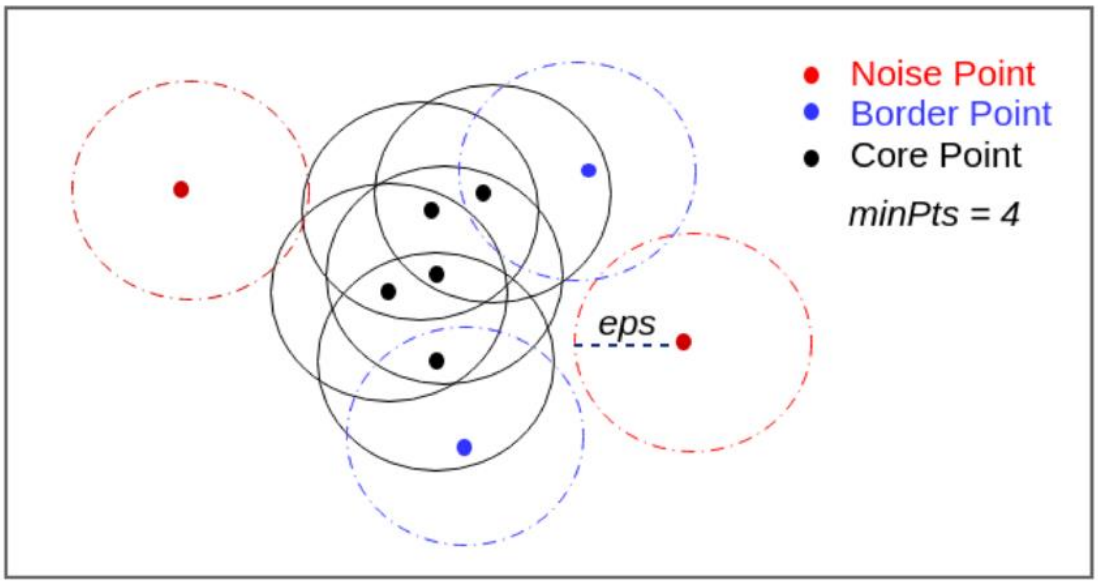
1. **ε (Epsilon):** Η μέγιστη **Ακτίνα Αναζήτησης** γύρω από ένα **Σημείο**.
2. **MinPts (Minimum Points):** Ο **Ελάχιστος Αριθμός** σημείων που πρέπει να βρίσκονται εντός της **Ακτίνας ε** για να θεωρηθεί μια περιοχή "**Πυκνή**".

Βάσει αυτών, κάθε **Σημείο p** χαρακτηρίζεται ως:

- **Σημείο Πυρήνα (Core Point):** Αν στη Γειτονιά του (**Ακτίνα ε**) υπάρχουν τουλάχιστον **MinPts Σημεία**.
- Μαθηματικά:

$$|\mathbf{N}_{\varepsilon}(\mathbf{p})| \geq \mathbf{MinPts} \qquad \textbf{(Εξίσωση 2.4)}$$

Όπου $\mathbf{N}_{\varepsilon}(\mathbf{p}) = \{\mathbf{q} \in \mathbf{D}\ \mathbf{dist}(\mathbf{p},\mathbf{q}) \leq \boldsymbol{\varepsilon}\}$ είναι η γειτονιά του **p**.



Εικόνα 2.13.4.a. Ταξινόμηση Σημείων στον DBSCAN. (A) Core Point (Μαύρο): Έχει αρκετούς γείτονες. (B, C) Border Points (Μπλε): Δεν έχουν πολλούς γείτονες αλλά "ακουμπούν" στο Core. (N) Noise Point (Κόκκινο): Είναι απομονωμένο και απορρίπτεται.

- **Σημείο Συνόρου (Border Point):** Αν έχει λιγότερα από **MinPts** στη γειτονιά του, αλλά βρίσκεται μέσα στη **Γειτονιά** ενός **Σημείου Πυρήνα (**είναι δηλαδή προσβάσιμο από αυτό**)**.

- **Σημείο Θορύβου (Noise Point):** Κάθε σημείο που δεν είναι ούτε Σημείο **Πυρήνα** ούτε **Συνόρου**. Αυτά τα σημεία θεωρούνται "**Outliers**" και αγνοούνται.

Η **Δημιουργία Συστάδας** βασίζεται στην έννοια της **Πυκνοτικής Προσπελασιμότητας (Density Reachability)**: Μια συστάδα σχηματίζεται ενώνοντας όλα τα **Σημεία Πυρήνα** που είναι προσπελάσιμα μεταξύ τους και προσθέτοντας τα **Σημεία Συνόρου** τους.

Η Ταξινόμηση των Σημείων βάσει Πυκνότητας

Στην **Εικόνα** αυτή (**2.13.4.a.**), οπτικοποιείται ο μηχανισμός με τον οποίο ο **DBSCAN** "σαρώνει" τα **Δεδομένα** και αποφασίζει τον ρόλο κάθε σημείου, βασιζόμενος στις **Παραμέτρους ε (Epsilon)** και **MinPts**:

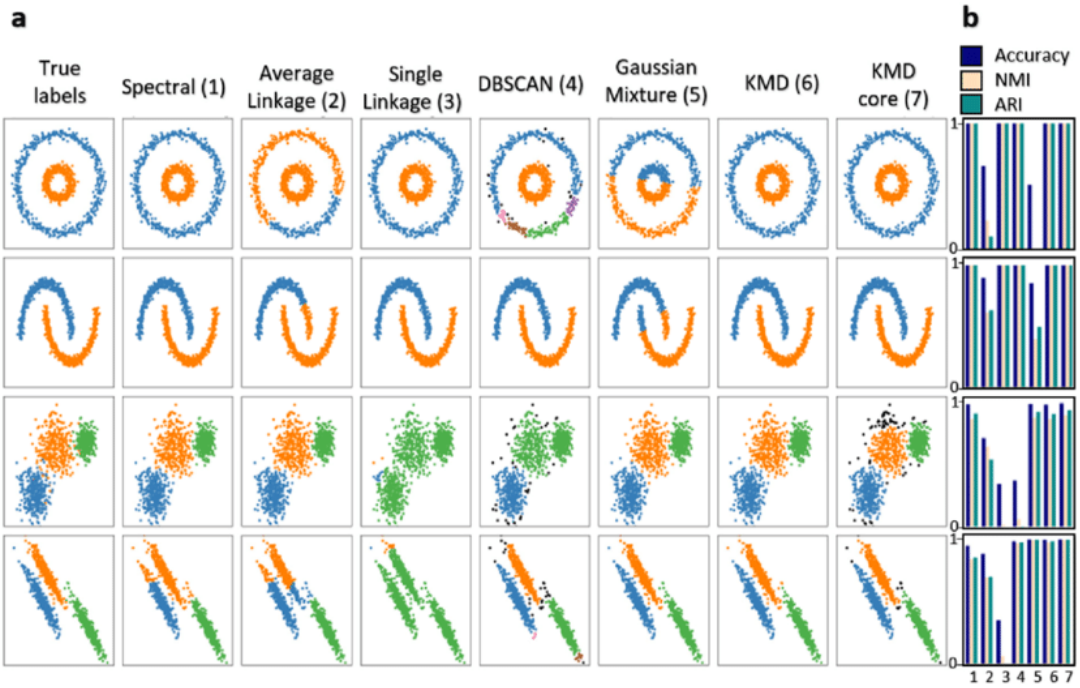
1. **Το Σημείο Πυρήνα (Core Point - Μαύρο):** Βρίσκεται στο κέντρο του κύκλου που ορίζει η ακτίνα **ε**. Όπως παρατηρούμε, εντός αυτής της περιοχής περιλαμβάνονται τουλάχιστον **MinPts** άλλα σημεία (στην εικόνα, MinPts=4). Αυτό το σημείο αποτελεί την "καρδιά" της **Συστάδας** και έχει τη δυνατότητα να επεκτείνει την ομάδα συνδεόμενο με άλλα **Σημεία Πυρήνα**.
2. **Το Σημείο Συνόρου (Border Point - Μπλε):** Βρίσκεται στην περιφέρεια της συστάδας. Αν σχεδιάσουμε κύκλο γύρω του, θα δούμε ότι περιέχει λιγότερα από **τέσσερα** (**4**) **Σημεία**, άρα δεν πληροί την προϋπόθεση πυκνότητας για να γίνει **Πυρήνας**. Ωστόσο, επειδή βρίσκεται **Εντός της Εμβέλειας** ενός **Σημείου Πυρήνα**, ενσωματώνεται στη συστάδα ως το **Ακραίο Όριό** της.
3. **Το Σημείο Θορύβου (Noise Point - Κόκκινο):** Είναι το **Σημείο** που βρίσκεται **απομονωμένο**. Στην ακτίνα γύρω του δεν υπάρχουν αρκετοί γείτονες, ούτε βρίσκεται κοντά σε κάποιο **Σημείο Πυρήνα**. Ο αλγόριθμος το αναγνωρίζει σωστά ως **Outlier** και το αγνοεί πλήρως, εξασφαλίζοντας ότι η συστάδα παραμένει "καθαρή" και συμπαγής.

Αυτός ο διαχωρισμός επιτρέπει στον **DBSCAN** να "χτίζει" ομάδες ακολουθώντας τη Ροή των Δεδομένων, αντί να επιβάλλει σφαιρικά σχήματα όπως ο **K-Means**. Η δημιουργία συστάδας βασίζεται τελικά στην έννοια της **Πυκνοτικής Προσπελασιμότητας (Density**

Reachability): Μια συστάδα σχηματίζεται ενώνοντας όλα τα Σημεία Πυρήνα που είναι προσπελάσιμα μεταξύ τους και προσθέτοντας τα Σημεία Συνόρου τους.

Σύγκριση Απόδοσης σε Σχήματα: Το Πρόλημα της Μη-Γραμμικής Διαχωρισιμότητας

Η παρακάτω **Εικόνα 2.13.4.b.** αναδεικνύει την υπεροχή του **DBSCAN** έναντι του **K-Means** σε **Μη Γραμμικά Δεδομένα**. Η συγκεκριμένη **Εικόνα** παρουσιάζει την εφαρμογή των αλγορίθμων στο πρότυπο σύνολο δεδομένων "**Two Moons**", το οποίο έχει καθιερωθεί στη βιβλιογραφία από τον **Fukunaga [1990]** ως το κλασικό παράδειγμα για την αξιολόγηση της ικανότητας διαχωρισμού **Μη-Γραμμικών Δομών**.



Εικόνα 2.13.4.b. Σύγκριση σε Μη Κυρτά Δεδομένα (Non-Convex Data). Αριστερά (K-Means): Αποτυγχάνει πλήρως, κόβοντας τα Ημισελήνια στη μέση λόγω της σφαιρικής υπόθεσης. Δεξιά (DBSCAN): Αναγνωρίζει τέλεια τα δύο Φυσικά Σχήματα, ακολουθώντας την Πυκνότητα των Σημείων.

1. **Αριστερά (Αποτυχία του K-Means):** Παρατηρούμε ότι ο **K-Means** έχει τεμαχίσει τα δύο φυσικά σχήματα με μία νοητή ευθεία γραμμή, αναμειγνύοντας τα δεδομένα.

- **Η Αιτία:** Ο αλγόριθμος βασίζεται στην ελαχιστοποίηση της Ευκλείδειας απόστασης από ένα κεντρικό σημείο (Centroid). Αυτό προϋποθέτει εσφαλμένα ότι οι ομάδες πρέπει να είναι σφαιρικές (**Convex**) και γραμμικά διαχωρίσιμες. Συνεπώς, αδυνατεί να ακολουθήσει την καμπυλότητα των δεδομένων.

2. **Δεξιά (Επιτυχία του DBSCAN):** Ο **DBSCAN** έχει αναγνωρίσει με απόλυτη ακρίβεια τα δύο σχήματα, χρωματίζοντας σωστά το πάνω και το κάτω μισοφέγγαρο ως διακριτές οντότητες.
    - **Η Αιτία:** Ο αλγόριθμος δεν αναζητά κέντρα, αλλά ακολουθεί την **Πυκνοτική Συνέχεια (Density Connectivity)**. Ξεκινώντας από ένα σημείο, "ερπύει" κατά μήκος της πυκνής περιοχής των δεδομένων και σταματάει μόνο όταν συναντήσει το κενό χαμηλής πυκνότητας που χωρίζει τα δύο μισοφέγγαρα. Έτσι, επιτυγχάνει τον διαχωρισμό συστάδων με **Ακανόνιστο Γεωμετρικό Σχήμα (Arbitrary Shape)**, επιβεβαιώνοντας τη θεωρητική του υπεροχή σε τοπολογικά περίπλοκα δεδομένα.

Περιορισμοί και Υστερήσεις (Limitations and Weaknesses)

Παρόλο που ο **DBSCAN** θεωρείται εξαιρετικά ισχυρός στην ανακάλυψη συστάδων ακανόνιστου σχήματος και στη διαχείριση θορύβου, δεν αποτελεί πανάκεια. Παρουσιάζει συγκεκριμένες **Εγγενείς Αδυναμίες** που οφείλονται στη **Μαθηματική** του **Δομή** και πρέπει να λαμβάνονται σοβαρά υπόψη κατά την εφαρμογή του **[Schubert et al., 2017]**:

1. Αδυναμία σε Δεδομένα Μεταβλητής Πυκνότητας (Varying Density Issues)

Ίσως το σημαντικότερο μειονέκτημα του αλγορίθμου. Ο **DBSCAN** υποθέτει ότι όλες οι συστάδες στο σύνολο δεδομένων έχουν παρόμοια πυκνότητα, καθώς χρησιμοποιεί **καθολικές παραμέτρους (global parameters) ε** και **MinPts** για όλο τον χώρο.

- *Το Πρόβλημα:* Αν το dataset περιέχει δύο συστάδες, μία πολύ "σφιχτή" (πυκνή) και μία πιο "αραιή" (απλωμένη), ο αλγόριθμος αδυνατεί να τις εντοπίσει ταυτόχρονα.
    - Αν ορίσουμε μικρό **ε** (για την πυκνή ομάδα), η αραιή ομάδα δεν θα ικανοποιεί το κριτήριο και θα απορριφθεί λανθασμένα ως **Θόρυβος**.

- Αν ορίσουμε μεγάλο **ε** (για την αραιή ομάδα), η πυκνή ομάδα ενδέχεται να συγχωνευθεί με γειτονικά σημεία θορύβου ή άλλες συστάδες, χάνοντας τη δομή της.

2. Η "Κατάρα της Διαστατικότητας" (The Curse of Dimensionality)

Ο **DBSCAN** βασίζεται στην **Ευκλείδεια Απόσταση** για να ορίσει τη γειτονιά **ε**. Ωστόσο, σε δεδομένα με πολλές διαστάσεις (High-Dimensional Data), εμφανίζεται το φαινόμενο της "**Κατάρας της Διαστατικότητας**".

- *Το Πρόβλημα:* Όσο αυξάνονται οι διαστάσεις (χαρακτηριστικά), ο όγκος του χώρου αυξάνεται εκθετικά και τα δεδομένα γίνονται αραιά. Μαθηματικά, η διαφορά μεταξύ της κοντινότερης και της μακρινότερης απόστασης τείνει να μηδενιστεί. Αυτό καθιστά την επιλογή του κατάλληλου **ε** σχεδόν αδύνατη, καθώς όλα τα σημεία φαίνονται να ισαπέχουν μεταξύ τους, με αποτέλεσμα ο αλγόριθμος είτε να τα θεωρεί όλα θόρυβο είτε όλα μία μάζα.

3. Ευαισθησία και Δυσκολία στον Καθορισμό Παραμέτρων

Σε αντίθεση με τον **K-Means** όπου το **k** είναι διαισθητικό, οι **Παράμετροι ε** και **MinPts** είναι δύσκολο να οριστούν χωρίς βαθιά γνώση του πεδίου (Domain Knowledge).

- *Το Πρόβλημα:* Δεν υπάρχει πλήρως αυτοματοποιημένος τρόπος εύρεσης των βέλτιστων τιμών. Μια ελαχιστή απόκλιση στην τιμή του **ε** μπορεί να προκαλέσει δραματικές αλλαγές στο αποτέλεσμα (Butterfly Effect), όπως το να ενωθούν δύο ξεχωριστές ομάδες σε μία ή να θρυμματιστεί μια συμπαγής ομάδα σε δεκάδες **Outliers**. Συνήθως απαιτείται η χρήση ευρετικών μεθόδων, όπως το **k-distance graph**, και επαναλαμβανόμενες δοκιμές (Trial-and-Error).

2.14. Επιστημονική Στάθμιση και Συγκριτική Αξιολόγηση

Για την επιλογή του βέλτιστου αλγορίθμου στο πλαίσιο ενός Customer Data Platform (CDP), η απόφαση δεν μπορεί να είναι μονοσήμαντη. Πρέπει να συνυπολογιστούν τα τεχνικά χαρακτηριστικά, η υπολογιστική πολυπλοκότητα και η φύση των δεδομένων (Data Nature).

Ο παρακάτω **Πίνακας 2.3** συνοψίζει τα πλεονεκτήματα και τους περιορισμούς των τεσσάρων (4) μεθόδων που αναλύθηκαν:

Χαρακτηριστικό	K-Means & K-Means++	K-Medoids (PAM)	DBSCAN
Πολυπλοκότητα	O (n · k · i) *(Γραμμική - Πολύ Γρήγορη)*	$O(k(n-k)^2)$ *(Πολύ Υψηλή)*	O(n log n) *(Μέτρια - με Indexing)*
Σχήμα Συστάδων	Σφαιρικό / Κυρτό (Convex)	Σφαιρικό	Αυθαίρετο / Μη Κυρτό
Ευαισθησία σε Outliers	**Υψηλή** *(Μετακινεί το μέσο όρο)*	**Χαμηλή** *(Robust - Χρήση Medoid)*	**Πολύ Χαμηλή** *(Τα αγνοεί ως Noise)*
Παράμετροι	Απαιτεί το **k** εκ των προτέρων	Απαιτεί το **k**	Απαιτεί **ε** και **MinPts**
Εφαρμογή σε CDP	Γενική τμηματοποίηση (RFM), Μεγάλα Datasets	Ακριβής Τμηματοποίηση (B2B), Μικρά Datasets	Εντοπισμός Απάτης (Fraud), Γεωχωρική Ανάλυση

Πίνακας 2.3. Συγκριτική Αξιολόγηση Αλγορίθμων Συσταδοποίησης

Συμπέρασμα και Επιλογή Μεθόδου

Όπως επισημαίνουν οι **Xu & Wunsch [2005]** στην επισκόπησή τους, καθώς και οι δημιουργοί του θεωρήματος **"No Free Lunch"** (Wolpert & Macready, 1997), δεν υπάρχει ένας "καθολικά βέλτιστος" αλγόριθμος για όλα τα προβλήματα. Η επιλογή εξαρτάται άμεσα από τη δομή των δεδομένων και τους επιχειρηματικούς στόχους.

Στην παρούσα μελέτη για το **CDP**:

- Ο **K-Means++** προκρίνεται ως η **Βέλτιστη Λύση** για την **Αρχική Τμηματοποίηση** μεγάλου όγκου **Πελατών** (High Volume), λόγω της **Ταχύτητάς** του.
- Ο **DBSCAN** κρίνεται απαραίτητος για τον εντοπισμό ειδικών συμπεριφορών (π.χ. Outliers, Fraud Detection) που δεν ακολουθούν κανονική κατανομή.
- Ο **K-Medoids** εξετάζεται ως εναλλακτική για μικρότερα υποσύνολα **Δεδομένων** όπου απαιτείται ανθεκτικότητα σε ακραίες τιμές, χωρίς την **Πολυπλοκότητα Παραμέτρων** του **DBSCAN**.

2.15. Πρακτικές Εφαρμογές των Αλγορίθμων στη Σύγχρονη Ζωή (Real-World Applications)

Η **Αξία** των **Αλγορίθμων Συσταδοποίησης** δεν περιορίζεται σε **Θεωρητικό Επίπεδο**, αλλά βρίσκει εφαρμογή σε **Κρίσιμους Τομείς** της **Καθημερινότητας** και της **Σύγχρονης Βιομηχανίας**. Ακολουθούν χαρακτηριστικά **Παραδείγματα** για το πώς οι **Αλγόριθμοι** που αναλύθηκαν αξιοποιούνται πρακτικά για την **Επίλυση Σύνθετων Προβλημάτων**.

α. K-Means

Διαχείριση Όγκου και Εμπορική Στόχευση

Λόγω της **Γραμμικής του Πολυπλοκότητας** και της **Ταχύτητάς** του, ο K-Means αποτελεί τη **Βιομηχανική Επιλογή** για **Προβλήματα Μεγάλου Όγκου Δεδομένων (Big Data)**.

1. **Customer Segmentation (CDP Context - RFM):**

   Στο πλαίσιο ενός **Customer Data Platform**, ο K-Means χρησιμοποιείται για να ομαδοποιήσει **Πελάτες** βάσει του **Μοντέλου RFM** (Recency, Frequency, Monetary). Για παράδειγμα, μια **Αλυσίδα Λιανικής** μπορεί να εντοπίσει αυτόματα την **Ομάδα "Champions"** (υψηλή δαπάνη, συχνές επισκέψεις) και την **Ομάδα "At Risk"** (παλιοί καλοί πελάτες που έχουν καιρό να αγοράσουν), προσαρμόζοντας τη **Στρατηγική Μάρκετινγκ** για κάθε **Ομάδα [Tsiptsis & Chorianopoulos, 2009]**.

2. **Συμπίεση Εικόνας (Image Compression):**

Στην **Τεχνολογία**, ο K-Means χρησιμοποιείται για τη **Μείωση του Μεγέθους** των **Ψηφιακών Εικόνων** (Color Quantization). Ο **Αλγόριθμος** ομαδοποιεί τα εκατομμύρια **Χρώματα** μιας **Φωτογραφίας** σε k **Κυρίαρχες Αποχρώσεις** (π.χ. 64 χρώματα), αντικαθιστώντας κάθε pixel με το **Κεντροειδές** του **Χρώματος** στο οποίο ανήκει. Αυτό μειώνει δραματικά τον απαιτούμενο **Χώρο Αποθήκευσης** χωρίς σημαντική **Απώλεια Ποιότητας** για το **Ανθρώπινο Μάτι**.

3. **Οργάνωση Εγγράφων (Document Clustering):**

Οι **Μηχανές Αναζήτησης** χρησιμοποιούν παραλλαγές του K-Means για να ομαδοποιήσουν αυτόματα χιλιάδες **Άρθρα Ειδήσεων**. Έτσι, όταν ένας **Χρήστης** αναζητά ένα **Θέμα** (π.χ. "Εκλογές"), ο **Αλγόριθμος** μπορεί να παρουσιάσει τα **Αποτελέσματα** ομαδοποιημένα ανά **Υπο-κατηγορία** (π.χ. Πολιτική, Οικονομία, Διεθνή), βελτιώνοντας την **Εμπειρία Πλοήγησης**.

β. K-Medoids (PAM)

Ανθεκτικότητα και Εύρεση Αντιπροσώπων

Επειδή ο K-Medoids επιλέγει ένα **Πραγματικό Σημείο** ως **Κέντρο** (medoid) και είναι **Ανθεκτικός σε Ακραίες Τιμές (Robust to Outliers)**, καθίσταται πολύτιμος όταν το "Κέντρο" πρέπει να είναι **Υπαρκτή Οντότητα**.

1. **Στρατηγική Logistics & Facility Location:**

Φανταστείτε μια **Εταιρεία Courier** που θέλει να ανοίξει 5 νέα **Κέντρα Διανομής** για να εξυπηρετεί 1.000 **Σημεία Παράδοσης**. Ο K-Medoids εντοπίζει τα 5 **Υπάρχοντα Κτίρια** που ελαχιστοποιούν τη συνολική **Απόσταση Οδήγησης**. Αν χρησιμοποιούσαμε K-Means, το "Κέντρο" θα μπορούσε να υπολογιστεί στη μέση μιας **Λίμνης** ή ενός **Πάρκου** (πλασματικό μέσο), κάτι που είναι πρακτικά αδύνατο για **Logistics [Park & Jun, 2009]**.

2. **Ιατρική Διάγνωση (Representative Patient):**

Στην **Ανάλυση Κλινικών Δεδομένων**, ο **Αλγόριθμος** επιλέγει τον **"Τυπικό Ασθενή"** (Representative Patient) για κάθε **Ασθένεια**. Αυτό βοηθά τους **Γιατρούς** να δημιουργήσουν **Πρότυπα Διάγνωσης** που δεν επηρεάζονται από σπάνιες, **Ακραίες Περιπτώσεις** (outliers) που θα αλλοίωναν τον **Μέσο Όρο** των **Συμπτωμάτων**.

3. **Ασύρματα Δίκτυα Αισθητήρων (Wireless Sensor Networks):**

Σε **Δίκτυα IoT**, πρέπει να επιλεγεί ένας **Αισθητήρας-Αρχηγός** (Cluster Head) που θα συγκεντρώνει τα **Δεδομένα** από τους γύρω του και θα τα στέλνει στη **Βάση**, για **Εξοικονόμηση Ενέργειας**. Ο K-Medoids επιλέγει τον πιο κεντρικό, **Πραγματικό Αισθητήρα** για αυτόν τον ρόλο, εξασφαλίζοντας τη βέλτιστη **Ενεργειακή Απόδοση** του **Δικτύου [Arora et al., 2016]**.

γ. DBSCAN

Διαχείριση Θορύβου και Γεωχωρική Ανάλυση

Η μοναδική ικανότητα του **DBSCAN** να αντιλαμβάνεται την **Πυκνότητα** και να απομονώνει τον **Θόρυβο**, τον καθιστά κυρίαρχο σε **Χωρικά Δεδομένα** και **Εφαρμογές Ασφάλειας**.

1. **Ανίχνευση Απάτης (Credit Card Fraud Detection):**

Στις **Τραπεζικές Συναλλαγές**, οι νόμιμες κινήσεις δημιουργούν **"Πυκνές" Συστάδες Συμπεριφοράς**. Μια μεμονωμένη **Συναλλαγή** μεγάλου ύψους σε ασυνήθιστη **Τοποθεσία** δεν ανήκει σε καμία **Γειτονιά**. Ο DBSCAN την χαρακτηρίζει αυτόματα ως **"Noise" (Θόρυβο)** και το **Σύστημα** ειδοποιεί για πιθανή **Απάτη**, χωρίς να χρειάζεται **Εκπαίδευση** με προηγούμενα **Παραδείγματα Απατών [Thiprungsri & Vasarhelyi, 2011]**.

2. **Πολεοδομικός Σχεδιασμός & GPS (Urban Planning):**

Εφαρμογές όπως η Uber ή η Google Maps χρησιμοποιούν τον DBSCAN για να **Ομαδοποιήσουν Στίγματα GPS**. Μπορούν να εντοπίσουν **Περιοχές Υψηλής Ζήτησης** (hotspots) που έχουν **Ακανόνιστο Σχήμα** (π.χ. η ροή κίνησης κατά μήκος μιας παραλιακής λεωφόρου) και να αγνοήσουν **Στίγματα** που οφείλονται σε **Σφάλμα της Συσκευής GPS**.

3. **Συστήματα Συστάσεων (Niche Market Recommendation):**

Στο **Ηλεκτρονικό Εμπόριο** (e-commerce), ο DBSCAN βοηθά στον εντοπισμό μικρών, **Εξειδικευμένων Ομάδων Χρηστών** με πολύ συγκεκριμένα και ασυνήθιστα **Γούστα** (niche markets). Ενώ ο K-Means θα τους ενσωμάτωνε βίαια σε μεγαλύτερες **Κατηγορίες**, ο DBSCAN τους αναγνωρίζει ως ξεχωριστές μικρές **Νησίδες Πυκνότητας**, επιτρέποντας στοχευμένες **Προτάσεις Προϊόντων [Ester et al., 1996]**.

3. Λογοτεχνική Αναθεώρηση

3.1. Η Εξέλιξη από το CRM στα CDPs: Η Ανάγκη για Ενοποίηση

Όπως αναλύθηκε στην **Ενότητα 2.9**, η επίτευξη της Υπερ-εξατομίκευσης (Hyper-personalization) και της **Εικόνας 360°** προϋποθέτει την ύπαρξη μιας "Ενιαίας Πηγής Αλήθειας". Ωστόσο, η πρακτική εφαρμογή αυτού του μοντέλου προσέκρουε για χρόνια στους **Τεχνολογικούς Περιορισμούς** των **Παραδοσιακών Συστημάτων**.

Η **Ακαδημαϊκή** και **Βιομηχανική Βιβλιογραφία** των τελευταίων ετών έχει εστιάσει έντονα στη μετάβαση από το πολυκαναλικό (multichannel) στο **πανκαναλικό (omnichannel) μάρκετινγκ**. Οι **Verhoef et al. [2015]** υποστηρίζουν ότι η μεγαλύτερη πρόκληση για τις σύγχρονες επιχειρήσεις παραμένει η διάσπαση των δεδομένων σε **"σιλό" (data silos)**, όπου τα δεδομένα πωλήσεων, εξυπηρέτησης και marketing δεν επικοινωνούν μεταξύ τους, καθιστώντας αδύνατη την ολιστική προσέγγιση που περιγράψαμε προηγουμένως.

Ενώ τα συστήματα **CRM (Customer Relationship Management)** αποτέλεσαν τη βάση για τη διαχείριση **Πελατών** για δεκαετίες, μελέτες όπως των **Payne & Frow [2005]** είχαν ήδη επισημάνει τους περιορισμούς τους στη διαχείριση **Αδόμητων Δεδομένων (unstructured data)** και δεδομένων πραγματικού χρόνου. Το **CRM** σχεδιάστηκε κυρίως για την καταγραφή **Στατικών Στοιχείων** και πωλήσεων, αδυνατώντας να συλλάβει τη δυναμική συμπεριφορά του χρήστη στα **Ψηφιακά Κανάλια**.

Η εμφάνιση των **Customer Data Platforms (CDPs)** ήρθε να καλύψει αυτό ακριβώς το κενό. Σύμφωνα με τον **Raab [2016]**, το CDP διαφέρει θεμελιωδώς από προγενέστερα συστήματα καθώς εστιάζει στη δημιουργία ενός **"Single Customer View" (SCV)** που είναι προσβάσιμο από άλλα συστήματα, και όχι απλώς στην **Αποθήκευση Δεδομένων**.

Πρόσφατες έρευνες των **Zahay et al. [2019]** επιβεβαιώνουν ότι οι επιχειρήσεις που υιοθετούν αρχιτεκτονικές CDP παρουσιάζουν σημαντικά υψηλότερη ακρίβεια στην

ταυτοποίηση πελατών (identity resolution) σε σχέση με αυτές που βασίζονται σε παραδοσιακά **Data Warehouses**.

Συγκριτική Επισκόπηση: CRM vs. CDP

Για την καλύτερη κατανόηση της μετάβασης, ο παρακάτω πίνακας συνοψίζει τις δομικές διαφορές μεταξύ των δύο συστημάτων:

Χαρακτηριστικό	CRM **Customer Relationship Management**	CDP **Customer Data Platform**
Πρωταρχικός Σκοπός	Διαχείριση πωλήσεων και σχέσεων με πελάτες.	Ενοποίηση δεδομένων πελατών από όλες τις πηγές.
Τύπος Δεδομένων	Κυρίως δομημένα δεδομένα (Transactional).	Δομημένα, ημι-δομημένα και αδόμητα (Behavioral, Social, IoT).
Αναγνώριση Πελάτη	Βασίζεται σε γνωστά στοιχεία (Email, Phone).	Συνδυάζει γνωστά και ανώνυμα στοιχεία (Cookies, Device IDs) μέσω Identity Resolution.
Προσβασιμότητα	Σχεδιασμένο για χρήση από ανθρώπους (Sales reps).	Σχεδιασμένο για χρήση από άλλα συστήματα (Marketing Automation, Ads).
Data Silos	Συχνά δημιουργεί ένα ακόμα "σιλό" δεδομένων.	Σπάει τα "σιλό" και δημιουργεί το Single Customer View.

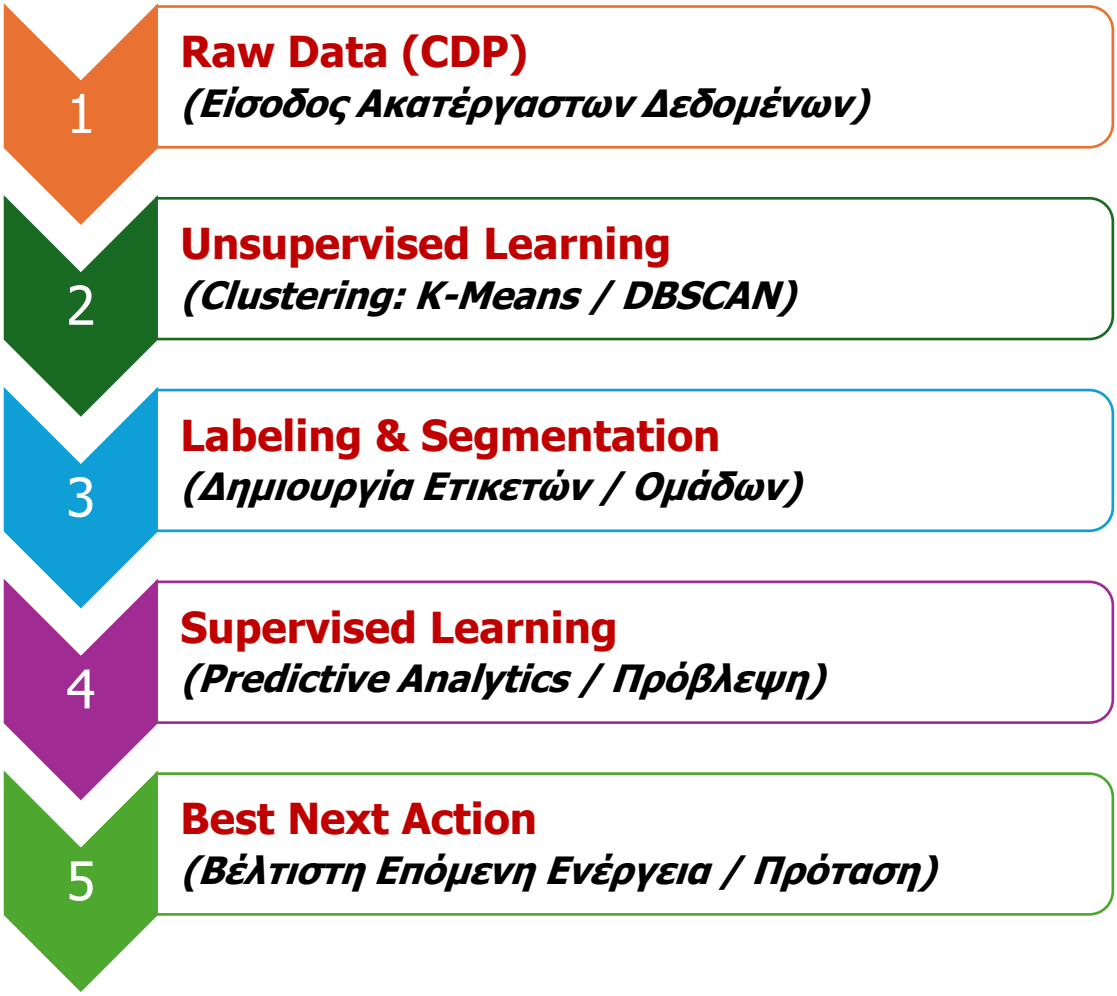
Πίνακας 3.1. Βασικές Διαφορές μεταξύ CRM και CDP

3.2. Μηχανική Μάθηση και Προσωποποίηση (Personalization)

Η εφαρμογή της **Μηχανικής Μάθησης (Machine Learning - ML)** στο μάρκετινγκ έχει εξεταστεί εκτενώς, καθώς αποτελεί τον κινητήριο μοχλό για την επίτευξη της **Εξατομίκευσης σε Κλίμακα**. Οι **Portugal et al. [2018]**, σε μια εκτενή επισκόπηση των **Συστημάτων Συστάσεων (Recommender Systems)**, καταλήγουν στο συμπέρασμα ότι η μετάβαση από

τους απλούς **Ευριστικούς Κανόνες (Heuristics)** σε αλγόριθμους ML (όπως το **Collaborative Filtering**) αυξάνει δραματικά την ευστοχία των προτάσεων προς τον χρήστη.

Ειδικότερα για την **Τμηματοποίηση (Segmentation)**, οι **Tsiptsis & Chorianopoulos [2009]** στο θεμελιώδες έργο τους για το **Data Mining στο CRM**, ανέδειξαν τον αλγόριθμο **K-Means** ως το **Βιομηχανικό Πρότυπο** για την ομαδοποίηση πελατών βάσει συμπεριφοράς. Η **Απλότητα** και η **Ταχύτητά** του τον καθιστούν ιδανικό για την αρχική χαρτογράφηση της πελατειακής βάσης.



Εικόνα 3.1. Η Ροή του Υβριδικού Μοντέλου Μάθησης

Ωστόσο, η **Πολυπλοκότητα** των **Σύγχρονων Δεδομένων** απαιτεί πιο εξελιγμένες προσεγγίσεις. Πιο σύγχρονες μελέτες, όπως των **Kou et al. [2020]**, προτείνουν ότι

αλγόριθμοι βασισμένοι στην **Πυκνότητα**, όπως ο **DBSCAN**, μπορούν να αποδώσουν καλύτερα σε **Datasets** με **Υψηλό Θόρυβο (Noise)** και **Ακανόνιστες Κατανομές**, χαρακτηριστικά που είναι συνήθη στα **Ακατέργαστα Δεδομένα (Raw Data)** ενός **CDP**.

Επιπλέον, η έρευνα των **Fan et al. [2021]** τονίζει τη σημασία των **Υβριδικών Μοντέλων (Hybrid Models)**. Σε αυτή την προσέγγιση, η **Μη Εποπτευόμενη Μάθηση (Unsupervised Learning - Clustering)** χρησιμοποιείται αρχικά για να ανακαλύψει τη **Δομή** των **Δεδομένων** και να δημιουργήσει **Ετικέτες (Labels)**, οι οποίες στη συνέχεια τροφοδοτούν μοντέλα **Εποπτευόμενης Μάθησης (Supervised Learning)** για την ακριβή πρόβλεψη της μελλοντικής συμπεριφοράς (**Predictive Analytics**).

Στην **Εικόνα 3.1**, βλέπουμε το **Διάγραμμα Ροής** του **Υβριδικού Μοντέλου Μάθησης**. Σε αυτό, τα **Δεδομένα ρέουν** από την **Κορυφή** (ακατέργαστα) προς τα κάτω, περνώντας πρώτα από τη **Συσταδοποίηση** (για να βρούμε τη δομή) και μετά από την **Πρόβλεψη** (για να βρούμε τη μελλοντική συμπεριφορά), καταλήγοντας στην **Καλύτερη Επόμενη Ενέργεια** η οποία μπορεί να είναι:

- **α) Εξατομικευμένη Εμπειρία Χρήστη (Personalized Experience):** Αν ο στόχος είναι η βελτιστοποίηση της **Αλληλεπίδρασης**.
- **β) Στοχευμένη Προσφορά (Targeted Offer):** Αν ο στόχος είναι η άμεση **Πώληση** (Conversion).
- **γ) Λήψη Απόφασης (Decisioning):** Αν ο στόχος είναι η **Στρατηγική Αξιολόγηση** (π.χ. έγκριση δανείου ή διαβάθμιση κινδύνου).

3.3. Προκλήσεις στην Εφαρμογή και Ποιότητα Δεδομένων

Ένα σημαντικό μέρος της βιβλιογραφίας αφιερώνεται στα **Εμπόδια Υλοποίησης**, αναγνωρίζοντας ότι η τεχνολογία από μόνη της δεν αρκεί. Ο **Ghasemaghaei [2019]**, αναλύοντας την ποιότητα των **Big Data**, διαπίστωσε ότι η **Ακρίβεια (Accuracy)** και η **Πληρότητα (Completeness)** των δεδομένων επηρεάζουν άμεσα την απόδοση των αλγορίθμων **Μηχανικής Μάθησης (ML)**.

Το φαινόμενο **"Garbage In, Garbage Out"** παραμένει τραγικά επίκαιρο, με τους **Haleem et al. [2022]** να υπογραμμίζουν ότι χωρίς ισχυρές διαδικασίες **Προεπεξεργασίας (Preprocessing)**, ακόμη και οι πιο εξελιγμένοι αλγόριθμοι Νευρωνικών Δικτύων αποτυγχάνουν να δώσουν αξιόπιστα αποτελέσματα.

Ειδικότερα στο πλαίσιο της **Συσταδοποίηση (Clustering)** που εξετάζουμε:

- Η έλλειψη **Κανονικοποίησης (Normalization)** μπορεί να οδηγήσει τον **K-Means** σε λανθασμένα αποτελέσματα, καθώς ο αλγόριθμος είναι ευαίσθητος στην κλίμακα των τιμών.
- Η ύπαρξη **Διπλότυπων Εγγραφών** (λόγω κακού Identity Resolution στο CDP) μπορεί να δημιουργήσει ψευδείς συστάδες.

Συνεπώς, η **Διασφάλιση Ποιότητας Δεδομένων** δεν είναι απλώς ένα τεχνικό βήμα, αλλά προαπαιτούμενο για την εξαγωγή επιχειρηματικής αξίας.

3.4. Κενά στη Βιβλιογραφία και Συνεισφορά της Εργασίας (The Gap)

Παρά την πληθώρα ερευνών γύρω από τους αλγόριθμους **Μηχανικής Μάθησης (ML)** και τα συστήματα **CDP** ως ξεχωριστές οντότητες, παρατηρείται ένα σαφές **Ερευνητικό Κενό (Research Gap)** στη λειτουργική τους διασύνδεση.

Συγκεκριμένα, εντοπίζονται δύο βασικές ελλείψεις στη σύγχρονη βιβλιογραφία:

1. **Έλλειψη Συγκριτικών Μελετών σε Περιβάλλον CDP:** Οι περισσότερες μελέτες εστιάζουν είτε στη θεωρητική αρχιτεκτονική των CDPs είτε στη μαθηματική απόδοση των αλγορίθμων σε γενικά, ακαδημαϊκά datasets (π.χ. Iris dataset, MNIST). Υπάρχει αισθητή έλλειψη πρακτικών μελετών που να συγκρίνουν την απόδοση αλγορίθμων όπως ο **K-Means**, **K-Medoids** και **DBSCAN** συγκεκριμένα σε **Δεδομένα Συμπεριφοράς Πελατών (Transactional & Behavioral Data)**, τα οποία χαρακτηρίζονται από υψηλή μεταβλητότητα και θόρυβο.

2. **Απουσία End-to-End Οδηγών Υλοποίησης:** Ενώ υπάρχουν θεωρητικά πλαίσια, σπάνια καταγράφεται στη βιβλιογραφία η πλήρης **Τεχνική Πορεία (Pipeline)**: από την εξαγωγή δεδομένων από ένα CDP, τον καθαρισμό τους (Preprocessing), την παραμετροποίηση των αλγορίθμων και την τελική αξιολόγηση με μετρικές όπως το **Silhouette Score** για καθαρά επιχειρηματικούς σκοπούς.

Η Συνεισφορά της Εργασίας Η παρούσα **Διπλωματική Εργασία** έρχεται να καλύψει αυτό το κενό, προσφέροντας μια **Εφαρμοσμένη Προσέγγιση**. Μέσα από την υλοποίηση και σύγκριση διαφορετικών **Αλγορίθμων Συσταδοποίησης** σε ρεαλιστικά σενάρια, η εργασία φιλοδοξεί να προσφέρει έναν **Πρακτικό Οδηγό** για την αποτελεσματική ενσωμάτωση της **Μηχανικής Μάθησης** σε περιβάλλοντα **CDP**, γεφυρώνοντας το χάσμα μεταξύ θεωρητικής **Πληροφορικής** και **Επιχειρηματικής Στρατηγικής**.

4. Μεθοδολογία Έρευνας

4.1. Υπολογιστικό Πλαίσιο και Τεχνικό Περιβάλλον

Για τη **Διασφάλιση** της **Αξιοπιστίας** και της **Επαναληψιμότητας (Reproducibility)** των πειραματικών αποτελεσμάτων, η υλοποίηση της παρούσας έρευνας πραγματοποιήθηκε σε ένα αυστηρά **Ελεγχόμενο Υπολογιστικό Περιβάλλον**, βασισμένο στο **Λειτουργικό Σύστημα Windows 10 Pro (Έκδοση 22H2, x64)**. Η επιλογή του συγκεκριμένου περιβάλλοντος και των εργαλείων βασίστηκε στη βιομηχανική πρακτική της **Επιστήμης Δεδομένων**, εξασφαλίζοντας **Συμβατότητα**, **Σταθερότητα** και **Βελτιστοποίηση Πόρων**.

4.1.1. Στοίβα Λογισμικού (Software Stack)

Το **Λογισμικό** που χρησιμοποιήθηκε για την **Ανάπτυξη** και **Εκτέλεση** των **Αλγορίθμων** αποτελείται από τα εξής συστατικά:

1. **Γλώσσα Προγραμματισμού & Διαχείριση Περιβάλλοντος**
    - **Python (Έκδοση 3.13.9):** Επιλέχθηκε ως η κύρια γλώσσα προγραμματισμού λόγω της κυριαρχίας της στον χώρο της Μηχανικής Μάθησης και της πληθώρας των διαθέσιμων βιβλιοθηκών ανοιχτού κώδικα που διαθέτει.
    - **Miniconda (Έκδοση 25.9.1):** Χρησιμοποιήθηκε για τη διαχείριση εικονικών περιβαλλόντων (virtual environments). Η επιλογή του Miniconda, έναντι του πλήρους Anaconda, έγινε για τη διατήρηση ενός ελαφριού και απομονωμένου περιβάλλοντος, εγκαθιστώντας μόνο τις απαραίτητες εξαρτήσεις (dependencies) και αποφεύγοντας διενέξεις εκδόσεων (version conflicts).
2. **Περιβάλλον Ανάπτυξης (IDE)**
    - **Visual Studio Code (Έκδοση 1.108.2):** Ως κειμενογράφος κώδικα (code editor) χρησιμοποιήθηκε το VS Code της Microsoft. Επιλέχθηκε για την ευελιξία του, την ενσωματωμένη υποστήριξη Jupyter Notebooks και τη δυνατότητα άμεσης αποσφαλμάτωσης (debugging) των αλγορίθμων.

3. **Βιβλιοθήκες (Libraries) Μηχανικής Μάθησης και Ανάλυσης Δεδομένων:** Για την **Υλοποίηση** των **Αλγορίθμων** και την **Επεξεργασία** των **Δεδομένων**, αξιοποιήθηκαν οι εξής **Εξειδικευμένες Βιβλιοθήκες**:
    - **Scikit-Learn (v1.7.2):** Για την υλοποίηση των αλγορίθμων συσταδοποίησης (K-Means, DBSCAN, K-Medoids) και τον υπολογισμό μετρικών αξιολόγησης (Silhouette Score).
    - **Pandas (v2.3.3) & NumPy (v2.3.5):** Για τη διαχείριση, τον καθαρισμό και τον μαθηματικό μετασχηματισμό των δεδομένων σε μορφή πινάκων (DataFrames).
    - **Matplotlib (v3.10.8) & Seaborn (v0.13.2):** Για την οπτικοποίηση των αποτελεσμάτων και τη δημιουργία των διαγραμμάτων της ανάλυσης.

4.1.2. Υλικό Σύστημα (Hardware Configuration)

Δεδομένου ότι οι **Αλγόριθμοι Συσταδοποίησης** και η επεξεργασία μεγάλου όγκου δεδομένων είναι υπολογιστικά απαιτητικές διαδικασίες, η πειραματική διαδικασία εκτελέστηκε σε σταθμό εργασίας **Υψηλών Επιδόσεων** (High-Performance Workstation).

Τα **Τεχνικά Χαρακτηριστικά** του **Συστήματος** είναι τα εξής:

- **Κεντρικός Επεξεργαστής (CPU):** AMD Ryzen 7 7800X3D (8-Core Processor @ 4.20 GHz). Η τεχνολογία **3D V-Cache** του επεξεργαστή εξασφαλίζει ταχύτατη προσπέλαση δεδομένων στη μνήμη cache, μειώνοντας δραματικά τον χρόνο εκτέλεσης των αλγορίθμων.
- **Μνήμη RAM:** 32 GB. Η χωρητικότητα αυτή επιτρέπει την **In-Memory** επεξεργασία του συνόλου δεδομένων, αποτρέποντας καθυστερήσεις λόγω χρήσης εικονικής μνήμης (swapping).
- **Αποθηκευτικός Χώρος:** Samsung SSD 990 PRO (NVMe M.2). Η χρήση δίσκου τεχνολογίας NVMe εξασφαλίζει μέγιστες ταχύτητες ανάγνωσης/εγγραφής (I/O operations) κατά τη φόρτωση των datasets.

- **Κάρτα Γραφικών (GPU):** AMD Radeon RX 7800 XT (16 GB VRAM). Η ισχυρή κάρτα γραφικών υποστηρίζει την ταχύτατη απεικόνιση (rendering) των πολύπλοκων διαγραμμάτων διασποράς (scatter plots) υψηλής ανάλυσης.

4.2. Τύπος Έρευνας (Research Type)

Η παρούσα διπλωματική εργασία υιοθετεί τη μεθοδολογία της **Εφαρμοσμένης Ποσοτικής Έρευνας (Applied Quantitative Research)**. Στόχος δεν είναι μόνο η θεωρητική ανάλυση των Customer Data Platforms (CDPs), αλλά η εμπειρική επαλήθευση της χρησιμότητας και της απόδοσης των αλγορίθμων **Μηχανικής Μάθησης** μέσω πειραματικής διαδικασίας **[Creswell & Creswell, 2017]**.

Συγκεκριμένα, ακολουθείται η προσέγγιση της **Πειραματικής Προσομοίωσης (Experimental Simulation)**. Σε αυτή τη διαδικασία, πραγματικά δεδομένα συμπεριφοράς πελατών τροφοδοτούνται σε ένα ελεγχόμενο περιβάλλον επεξεργασίας (Python Environment) για την αξιολόγηση και σύγκριση διαφορετικών αλγοριθμικών σεναρίων.

4.3. Σχεδιασμός Έρευνας (Research Design - CRISP-DM)

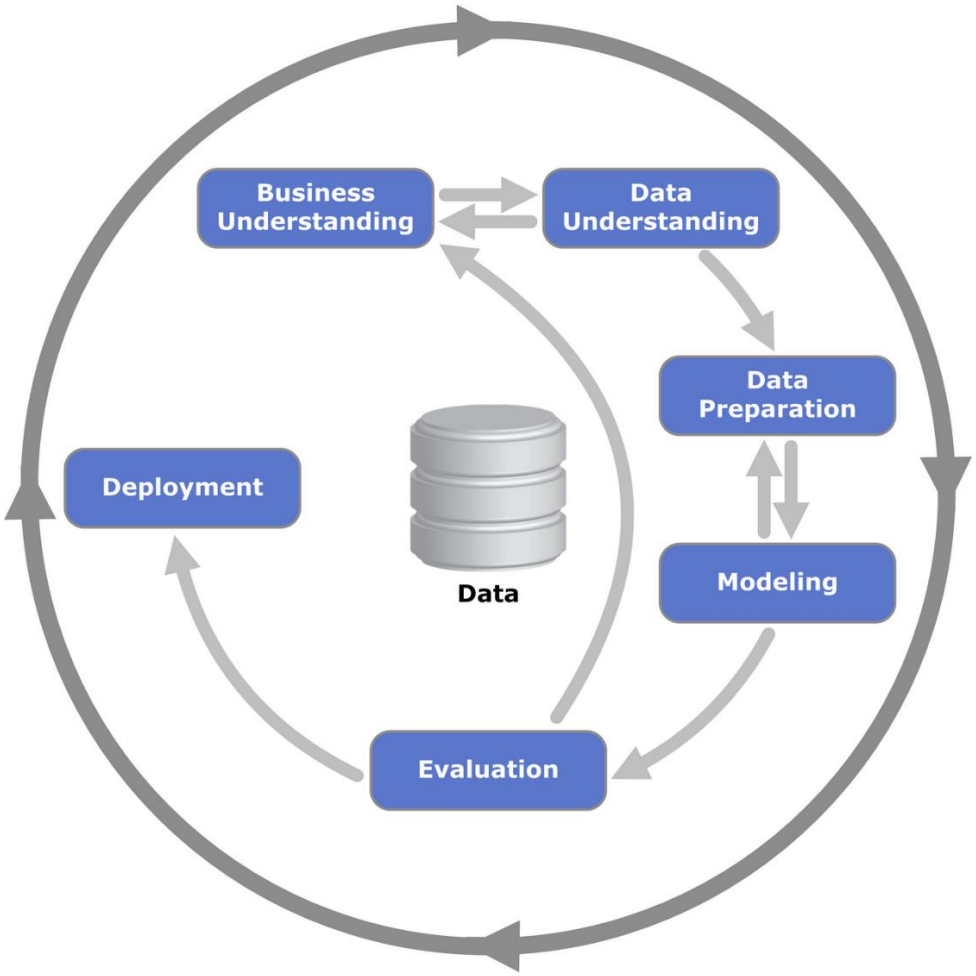
Ο σχεδιασμός της έρευνας ακολουθεί την τυποποιημένη ροή εργασίας (pipeline) της Επιστήμης Δεδομένων, γνωστή ως **CRISP-DM (Cross-Industry Standard Process for Data Mining)**. Το μοντέλο αυτό επιλέχθηκε διότι παρέχει μια δομημένη προσέγγιση για τη μετατροπή των ακατέργαστων δεδομένων σε επιχειρηματική γνώση.

Στην παρακάτω **Εικόνα 4.3.**, απεικονίζονται τα βήματα του σχεδιασμού της Έρευνάς μας, όπως προσαρμόστηκαν για τις ανάγκες του CDP.

Αναλυτικά τα Στάδια της Μεθοδολογίας:

1. **Business Understanding (Επιχειρηματική Κατανόηση)**

- Καθορισμός στόχου: Τμηματοποίηση πελατειακής βάσης για προσωποποιημένο μάρκετινγκ.
- Ανάγκη για μετάβαση από μαζικές καμπάνιες σε στοχευμένες ενέργειες (Targeted Actions).



Εικόνα 4.3. Η διαδικασία CRISP-DM προσαρμοσμένη για CDP

2. **Data Understanding (Κατανόηση Δεδομένων)**
   - Συλλογή δεδομένων συναλλαγών (Transactions) και συμπεριφοράς από το CDP.
   - Διερευνητική Ανάλυση (EDA) για εντοπισμό μοτίβων αγορών και κατανομής δαπανών.
3. **Data Preparation (Προετοιμασία Δεδομένων)**

- Καθαρισμός δεδομένων (Missing values, Noise).
- Δημιουργία μεταβλητών RFM (Recency, Frequency, Monetary).
- Κανονικοποίηση (Normalization) για την εξάλειψη διαφορών κλίμακας.

4. **Modeling (Μοντελοποίηση)**
   - Εφαρμογή αλγορίθμων συσταδοποίησης: **K-Means**, **K-Medoids**, **DBSCAN**.
   - Ρύθμιση υπερ-παραμέτρων (π.χ. εύρεση βέλτιστου k, καθορισμός ε και $MinPts$).
5. **Evaluation (Αξιολόγηση)**
   - Ποσοτική αξιολόγηση με δείκτες ποιότητας (**Silhouette Score**, Davies-Bouldin).
   - Ποιοτική/Επιχειρηματική ερμηνεία των συστάδων (π.χ. "Champions", "At Risk").

4.4. Περιγραφή Δεδομένων (Data Description)

Για τις ανάγκες της Πειραματικής Διαδικασίας, χρησιμοποιήθηκε το σύνολο δεδομένων **"Online Retail Dataset"**, το οποίο προέρχεται από το **UCI Machine Learning Repository** (διαθέσιμο και μέσω Kaggle). Πρόκειται για ένα πραγματικό (transnational) σύνολο δεδομένων συναλλαγών που καταγράφει τις πωλήσεις ενός βρετανικού διαδικτυακού καταστήματος λιανικής (non-store online retail), το οποίο ειδικεύεται σε μοναδικά είδη δώρων.

Η επιλογή του συγκεκριμένου **Dataset** κρίθηκε σκόπιμη καθώς τα **Δεδομένα** αυτά προσομοιάζουν απόλυτα το είδος της πληροφορίας που συγκεντρώνει ένα εμπορικό **CDP (Transactional Data)**. Αυτό επιτρέπει την εξαγωγή προφίλ **Συμπεριφορών** και **Μοτίβων** και την εφαρμογή **Τεχνικών RFM** (Recency, Frequency, Monetary) σε ρεαλιστικές συνθήκες **[Chen et al., 2012]**. Οι **RFM Τεχνικές**, θα αναλυθούν σε επόμενη ενότητα.

Σε ό,τι αφορά τη **Δομή** του, το **Σύνολο Δεδομένων** καλύπτει τη χρονική περίοδο από **01/12/2009 έως 09/12/2011** και αποτελείται συνολικά από **541.909 Εγγραφές**. Κάθε εγγραφή περιγράφεται από **8 μεταβλητές (features)**, οι οποίες αναλύονται στον πίνακα που ακολουθεί.

4.4.1. Χαρακτηριστικά του Dataset (Data Dictionary & Table)

Όπως είπαμε ήδη, το Σύνολο Δεδομένων αποτελείται από περίπου **541.909 εγγραφές** και περιλαμβάνει **8 χαρακτηριστικά** (features).

Ο παρακάτω **Πίνακας** (**Πίνακας 4.4.**) περιγράφει αναλυτικά τις **Μεταβλητές** που θα χρησιμοποιηθούν και παρουσιάζει το **Λεξικό Δεδομένων**, επεξηγώντας τον τύπο και το περιεχόμενο κάθε μεταβλητής:

Μεταβλητή	Περιγραφή	Τύπος Δεδομένων
InvoiceNo	Μοναδικός 6-ψήφιος αριθμός τιμολογίου. Αν ξεκινά με 'C', υποδηλώνει ακύρωση (Cancellation).	Nominal (String)
StockCode	Μοναδικός 5-ψήφιος κωδικός προϊόντος (item).	Nominal (String)
Description	Περιγραφή/Ονομασία προϊόντος.	Nominal (String)
Quantity	Η ποσότητα των προϊόντων ανά συναλλαγή.	Numeric (Integer)
InvoiceDate	Η ημερομηνία και ώρα πραγματοποίησης της συναλλαγής.	Datetime
UnitPrice	Η τιμή μονάδας του προϊόντος (σε Στερλίνες).	Numeric (Float)
CustomerID	Μοναδικός 5-ψήφιος κωδικός ταυτοποίησης πελάτη.	Nominal (String)
Country	Η χώρα διαμονής του πελάτη.	Nominal (String)

Πίνακας 4.4. Λεξικό Δεδομένων (Data Dictionary)

Σχολιασμός Μεταβλητών

Από το παραπάνω **Σύνολο Δεδομένων**, ιδιαίτερη βαρύτητα για την υλοποίηση του **Αλγορίθμου Συσταδοποίησης** έχουν οι μεταβλητές **InvoiceDate**, **Quantity** και **UnitPrice**. Ο συνδυασμός τους επιτρέπει τον **Υπολογισμό** των δεικτών **RFM (Recency,**

Frequency, Monetary), καθώς η **Ημερομηνία** καθορίζει την **Προσφατότητα** (Recency), ενώ το γινόμενο **Ποσότητας** και **Τιμής** καθορίζει τη **Χρηματική Αξία** (Monetary).
Επιπλέον, η ύπαρξη του **CustomerID** είναι κρίσιμη, καθώς αποτελεί το μοναδικό κλειδί για την ενοποίηση των συναλλαγών ανά πελάτη. **Εγγραφές** που δεν διαθέτουν **CustomerID** δεν μπορούν να συνεισφέρουν στη δημιουργία πελατειακού προφίλ και, ως εκ τούτου, θα αφαιρεθούν κατά το στάδιο του **Καθαρισμού Δεδομένων** που ακολουθεί.

4.5. Τεχνικές Συλλογής και Προεπεξεργασίας (Data Processing)

Η **Ποιότητα** των αποτελεσμάτων της **Μηχανικής Μάθησης** εξαρτάται άμεσα από την ποιότητα της εισόδου, επιβεβαιώνοντας την αρχή *"Garbage In, Garbage Out"*. Πριν την εφαρμογή οποιουδήποτε αλγορίθμου, τα πρωτογενή δεδομένα υπεβλήθησαν σε **Αυστηρό Καθαρισμό** και **Προεπεξεργασία (Preprocessing)**, με στόχο της μετατροπή των **Ακατέργαστων Συναλλαγών** σε **Αξιοποιήσιμη Γνώση** και την **Απαλοιφή** του **Θορύβου** (**Noise Removal**), δηλαδή των εσφαλμένων ή μη αξιοποιήσιμων **Εγγραφών**.

Συγκεκριμένα, εκτελέστηκαν οι εξής ενέργειες καθαρισμού για την απομάκρυνση του **Θορύβου**:

1. **Διαχείριση Ελλιπών Τιμών (Missing Values):**

   Εντοπίστηκαν **Εγγραφές** με **Κενό** (null) **Πεδίο CustomerID**. Οι **Εγγραφές** αυτές αφαιρέθηκαν οριστικά, καθώς χωρίς μοναδικό αναγνωριστικό δεν είναι δυνατή η απόδοση της συναλλαγής σε συγκεκριμένο **Πελάτη** και η δημιουργία ιστορικού.

2. **Αφαίρεση Ακυρώσεων και Επιστροφών:**

   Παρατηρήθηκαν **Εγγραφές** με **Αρνητικές Τιμές** στο πεδίο **Quantity** (οι οποίες συνήθως ξεκινούν με τον κωδικό '**C**' στο **InvoiceNo**). Αυτές υποδηλώνουν ακυρώσεις ή επιστροφές προϊόντων. Αφαιρέθηκαν από το **Dataset** ώστε να μην αλλοιώσουν τον υπολογισμό του συνολικού τζίρου και της **Συχνότητας Αγορών [Han et al., 2011]**.

3. **Καθαρισμός Τιμών:**

Αφαιρέθηκαν εγγραφές με **Μηδενική** ή **Αρνητική Τιμή Μονάδας** (**UnitPrice ≤ 0**), οι οποίες πιθανόν οφείλονται σε **Σφάλματα Συστήματος** ή **Καταχωρήσεις** δώρων / **Δειγμάτων** που δεν αφορούν την πραγματική αγοραστική συμπεριφορά.

4.6. Μοντέλο Ανάλυσης Συμπεριφοράς (RFM Analysis) - Δημιουργία Χαρακτηριστικών (Feature Engineering)

Μετά τον καθαρισμό των δεδομένων, το επόμενο στάδιο αφορά τη **Μηχανική Χαρακτηριστικών (Feature Engineering)**. Στόχος είναι η μετατροπή των απλών συναλλακτικών εγγραφών (transactional data) σε ουσιαστικά γνωρίσματα που περιγράφουν την αγοραστική συμπεριφορά.

Για τον σκοπό αυτό επιλέχθηκε το μοντέλο **RFM (Recency, Frequency, Monetary)**, το οποίο αποτελεί τη θεμέλιο λίθο της ανάλυσης πελατών στο μάρκετινγκ βάσεων δεδομένων **[Hughes, 1994]**.

Η διαδικασία αυτή (Aggregation) μετασχηματίζει το **Dataset** από επίπεδο *Συναλλαγής* σε επίπεδο *Πελάτη*. Για κάθε μοναδικό **Πελάτη i**, υπολογίζονται οι εξής τρεις μεταβλητές:

1. **Recency (**Προσφατότητα **-** $\mathbf{R_i}$**):** Ορίζεται ως το **Χρονικό Διάστημα** (σε ημέρες) που μεσολάβησε από την τελευταία αγορά του πελάτη μέχρι την ημερομηνία αναφοράς (snapshot date). **Μικρότερη Τιμή** υποδηλώνει πιο **Ενεργό Πελάτη**.

$$\mathbf{R_i} = \mathbf{T_{snapshot}} - \mathbf{max}\left(\mathbf{T_{purchase,i}}\right) \quad \textbf{(Εξίσωση 4.1)}$$

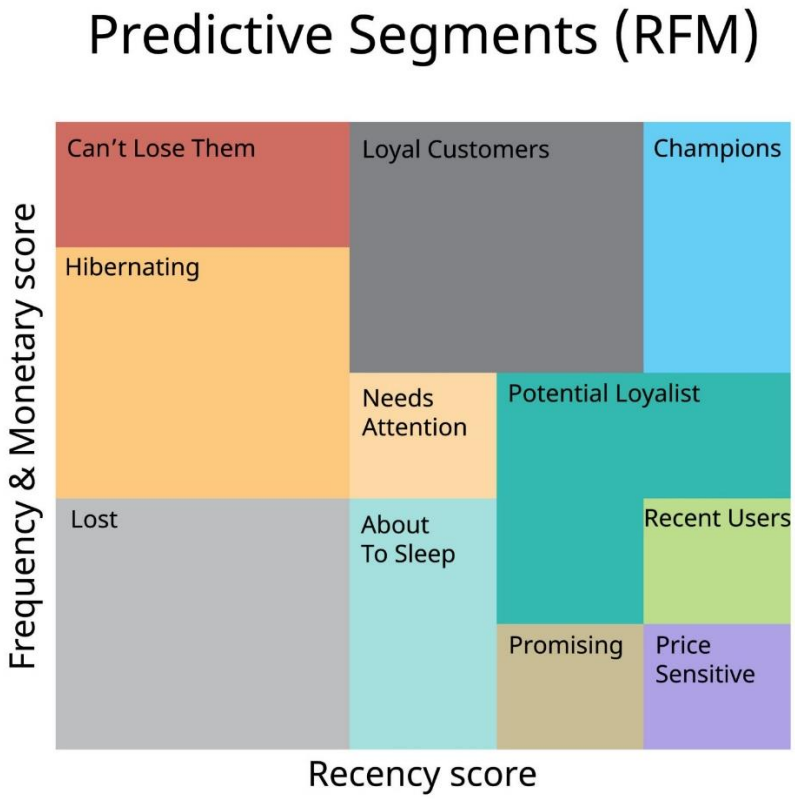
1. **Frequency (**Συχνότητα **-** $\mathbf{F_i}$**):** Ορίζεται ως το **Πλήθος** των διακριτών **Συναλλαγών** (Unique Invoices) που πραγματοποίησε ο πελάτης εντός της εξεταζόμενης **Περιόδου**.

$$\mathbf{F_i} = \mathbf{Count}(\mathbf{Unique\_Invoices_i}) \quad \textbf{(Εξίσωση 4.2)}$$

2. **Monetary (**Χρηματική Αξία **-** $\mathbf{M_i}$**):** Ορίζεται ως το **Συνολικό Χρηματικό Ποσό** που δαπάνησε ο **Πελάτης** (Revenue). Υπολογίζεται ως το **Άθροισμα** του γινομένου **Ποσότητας** (**Q**) και **Τιμής Μονάδας** (**P**) για κάθε **Προϊόν j** που αγόρασε.

$$M_i = \sum_{j=1}^{N_i} (Q_{ij} \times P_{ij})$$

(Εξίσωση 4.3)



Εικόνα 4.6. Χαρτογράφηση της Πελατειακής Βάσης σύμφωνα με την RFM Analysis

4.7. Μετασχηματισμός και Κανονικοποίηση Δεδομένων (Transformation & Standardization)

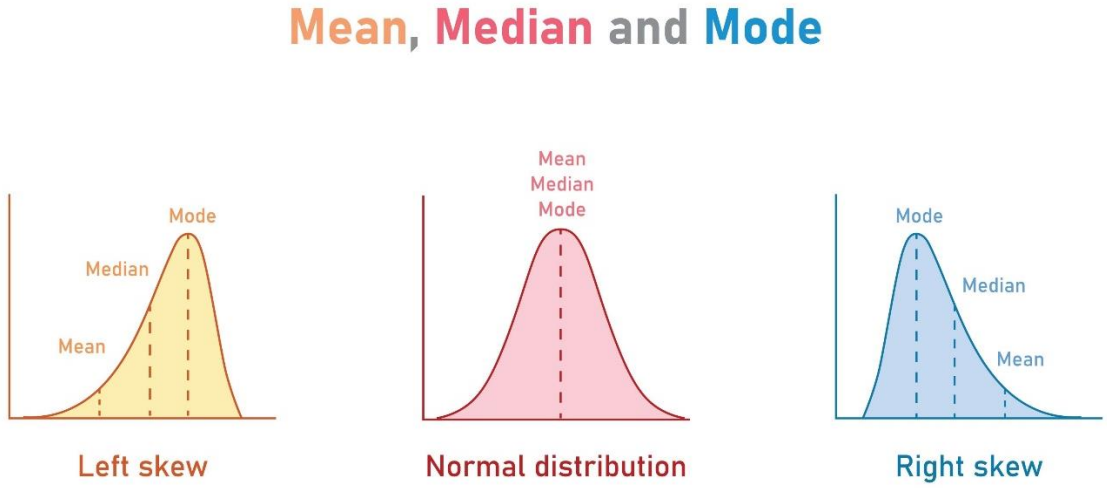
Η εφαρμογή **Αλγορίθμων Μηχανικής Μάθησης** σε ακατέργαστα δεδομένα (raw data) χωρίς την κατάλληλη **Μαθηματική Προσαρμογή** οδηγεί συχνά σε παραπλανητικά αποτελέσματα. Στην παρούσα μελέτη, εντοπίστηκαν δύο θεμελιώδη προβλήματα στη φύση των δεδομένων RFM που απαιτούσαν παρέμβαση: η **Λοξότητα της Κατανομής** και η **Ανομοιογένεια της Κλίμακας**.

Για την επίλυσή τους, εφαρμόστηκαν οι ακόλουθες στατιστικές μέθοδοι:

4.7.1. Αντιμετώπιση Ασυμμετρίας: Λογαριθμικός Μετασχηματισμός (Log Transformation)

Η πλειονότητα των **Οικονομικών Δεδομένων Συμπεριφοράς** καταναλωτή δεν ακολουθεί την **Κανονική Κατανομή** (Gaussian Distribution), αλλά διέπεται από την **Αρχή του Pareto** (διατυπώθηκε από τον **Vilfredo Pareto, 1896**). Σύμφωνα με αυτή, ένα μικρό ποσοστό **Πελατών** ευθύνεται για το μεγαλύτερο μέρος του τζίρου, δημιουργώντας κατανομές με έντονη **Θετική Ασυμμετρία (Right Skewness)**.

- **Το Πρόβλημα:** Οι **Αλγόριθμοι** που βασίζονται σε **Μέσους Όρους** (όπως ο **K-Means**) υποθέτουν ότι τα **Δεδομένα** είναι **Συμμετρικά** και οι **Συστάδες Σφαιρικές**. Η ύπαρξη έντονης **Ασυμμετρίας** "τραβάει" τα κέντρα των **Συστάδων** προς τις **Ακραίες Τιμές** (Outliers), στρεβλώνοντας την **Ομαδοποίηση**.



Εικόνα 4.7.1. Ιστόγραμμα Κατανομής της Μεταβλητής Monetary. Αριστερά: Η αρχική κατανομή με έντονη Θετική Ασυμμετρία (Right Skewness). Δεξιά: Η Κατανομή μετά την εφαρμογή του Λογαριθμικού Μετασχηματισμού ln(x+1), η οποία προσεγγίζει την Κανονική Κατανομή.

- **Η Λύση (Log Transformation):** Για την εξομάλυνση της κατανομής, βασιστήκαμε στη θεωρία των μετασχηματισμών που θεμελίωσαν οι **Box και Cox (1964)**. Εφαρμόστηκε ο **Λογαριθμικός Μετασχηματισμός**, ο οποίος συμπιέζει το εύρος των υψηλών τιμών και αποκαθιστά τη συμμετρία της κατανομής.

Ο **Μαθηματικός Τύπος** που εφαρμόστηκε είναι:

$$X' = \ln(X + 1) \qquad \textbf{(Εξίσωση 4.4)}$$

Σημείωση: Η προσθήκη της μονάδας (+1) γίνεται για την αποφυγή μαθηματικού σφάλματος σε περιπτώσεις μηδενικών τιμών (π.χ. σε νέους πελάτες με ελάχιστη δραστηριότητα), καθώς ο λογάριθμος του μηδενός δεν ορίζεται.

Όπου,

- **X'**: Η νέα, **Μετασχηματισμένη Τιμή** που θα εισαχθεί στον **Αλγόριθμο**.
- **X**: Η **Αρχική Τιμή** της **Μεταβλητής** (π.χ. Monetary Value).
- **ln**: Ο **Φυσικός Λογάριθμος** (με βάση το **e ≈ 2.718**).

Η προσθήκη της **σταθεράς μετατόπισης (shift constant) 1** κρίνεται απαραίτητη για τη **Μαθηματική Ευστάθεια** του **Μοντέλου**. Δεδομένου ότι ο **Λογάριθμος** του **Μηδενός** δεν ορίζεται (**ln(0)** → **∞**), η απουσία της **Μονάδας** θα οδηγούσε σε υπολογιστικό σφάλμα σε περιπτώσεις πελατών με **Μηδενική Δραστηριότητα** (π.χ. *Frequency = 0* ή *Recency = 0* μετά την κανονικοποίηση). Με την **Προσθήκη** του **1**, διασφαλίζεται ότι η **Ελάχιστη Τιμή 0** μετατρέπεται σε **ln(1) = 0**, διατηρώντας τη **Συνέχεια** των **Δεδομένων**.

Επιπλέον, ο **Λογάριθμος** λειτουργεί ως ισχυρός μηχανισμός **"συμπίεσης" κλίμακας**. Για παράδειγμα, ενώ η διαφορά μεταξύ 10€ και 100.000€ είναι χαώδης στην αρχική κλίμακα (διαφορά ≈ 99.990), στη λογαριθμική κλίμακα μειώνεται δραστικά (διαφορά ≈ 9.1). Αυτό επιτρέπει στον αλγόριθμο **K-Means** να υπολογίζει **Ευκλείδειες Αποστάσεις** που δεν κυριαρχούνται αποκλειστικά από τους **Πελάτες Υψηλής Αξίας** (High-Value Customers). Η αποτελεσματικότητα του μετασχηματισμού απεικονίζεται στην **Εικόνα 4.7.1**, όπου παρατηρείται η εξομάλυνση της καμπύλης και η μείωση της επίδρασης των ακραίων τιμών.

4.7.2. Κανονικοποίηση Κλίμακας: Τυποποίηση Z-Score

Το δεύτερο κρίσιμο πρόβλημα αφορά τη διαφορετική τάξη μεγέθους των μεταβλητών (Feature Scaling). Στο μοντέλο RFM:

- Η μεταβλητή **Monetary** λαμβάνει τιμές της τάξης των χιλιάδων (€).
- Η μεταβλητή **Frequency** λαμβάνει συνήθως μονοψήφιες ή διψήφιες τιμές.
- **Το Πρόβλημα:** Ο αλγόριθμος K-Means χρησιμοποιεί την **Ευκλείδεια Απόσταση** (**d**) για να μετρήσει την ομοιότητα μεταξύ δύο σημείων **p** και **q**:

$$\mathbf{d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots}} \quad \textbf{(Εξίσωση 4.5)}$$

Χωρίς **Κανονικοποίηση**, η μεταβλητή με το μεγαλύτερο εύρος τιμών (Monetary) θα κυριαρχούσε απόλυτα στον υπολογισμό, καθιστώντας τις άλλες μεταβλητές αμελητέες.

- **Η Λύση (Standardization):** Εφαρμόστηκε η μέθοδος **Z-Score Normalization (StandardScaler)**. Η μέθοδος αυτή μετασχηματίζει τα δεδομένα ώστε να έχουν **Μέση Τιμή (μ) ίση με 0** και **Τυπική Απόκλιση (σ) ίση με 1**.

Ο τύπος του μετασχηματισμού για κάθε **Σημείο Δεδομένων x** είναι:

$$\mathbf{z = \frac{x - \mu}{\sigma}} \quad \textbf{(Εξίσωση 4.5)}$$

Όπου:

- **z:** Η τυποποιημένη **Τιμή** (Z-score).
- **x:** Η **Αρχική Τιμή** της **Μεταβλητής** (π.χ. το ποσό σε Ευρώ).
- **μ:** Η **Μέση Τιμή** (mean) του πληθυσμού για τη συγκεκριμένη **Μεταβλητή**.
- **σ:** Η **Τυπική Απόκλιση** (standard deviation) του **Πληθυσμού**.

Με αυτόν τον τρόπο, επιτυγχάνεται η ισοτιμία όλων των χαρακτηριστικών (features) κατά τον υπολογισμό των αποστάσεων στον πολυδιάστατο χώρο, επιτρέποντας στον αλγόριθμο να εντοπίσει συστάδες βασισμένες στην πραγματική συμπεριφορά και όχι απλώς στο μέγεθος των αριθμών **[Han, Kamber, & Pei, 2012]**.

4.8. Αλγόριθμοι Συσταδοποίησης (Clustering Algorithms)

Βάσει της Θεωρητικής Ανάλυσης που προηγήθηκε στο **Κεφάλαιο 2 (Ενότητα 2.13)**, για τις ανάγκες της παρούσας έρευνας επιλέχθηκαν **Τέσσερις Διακριτοί** (**4**) **Αλγόριθμοι**. Η επιλογή δεν έγινε τυχαία, αλλά με στόχο τη **Συγκριτική Αξιολόγηση** τόσο διαφορετικών **Οικογενειών Αλγορίθμων** (Partitional vs. Density-based), όσο και διαφορετικών **Τεχνικών Αρχικοποίησης** διαφορετικών προσεγγίσεων (Partitional vs. Density-based) πάνω στα **Δεδομένα RFM**.

Συγκεκριμένα, οι **Αλγόριθμοι** που θα εφαρμοστούν είναι:

1. **K-Means (J. MacQueen, 1967):** Ο **Αλγόριθμος** του **MacQueen (1967)** αποτελεί το σημείο αναφοράς (baseline). Επιλέχθηκε ως το **σημείο αναφοράς (baseline model)** για την **Συσταδοποίηση**. Λόγω της ταχύτητάς του και της ευρείας χρήσης του στη βιβλιογραφία, αποτελεί το **Μέτρο Σύγκρισης** για την απόδοση των υπόλοιπων **Μεθόδων**. Στην παρούσα μελέτη, θα εξεταστούν δύο παραλλαγές του ως προς τη **Μέθοδο Αρχικοποίησης** των **Κέντρων**:
    - **Standard K-Means (Random Initialization):** Εφαρμογή της κλασικής μεθόδου με τυχαία επιλογή των **Αρχικών Κέντρων** (**k**), για τον εντοπισμό τυχόν προβλημάτων σύγκλισης σε **Τοπικά Βέλτιστα** (local optima).
    - **K-Means++ (Arthur & Vassilvitskii, 2007):** Εφαρμογή της **Βελτιστοποιημένης Μεθόδου**, η οποία επιλέγει τα **Αρχικά Κέντρα** με βάση κατανομή πιθανότητας ανάλογη του τετραγώνου της απόστασης από τα ήδη επιλεγμένα κέντρα. Με αυτόν τον τρόπο, «αναγκάζει» τα αρχικά κέντρα να απέχουν μεταξύ τους, διασφαλίζοντας ταχύτερη σύγκλιση και αποφεύγοντας τον εγκλωβισμό σε λανθασμένες λύσεις (Inertia).
2. **K-Medoids (Kaufman & Rousseeuw, 1987):** Επιλέχθηκε για τον έλεγχο της **ανθεκτικότητας (robustness)** του μοντέλου. Καθώς τα οικονομικά δεδομένα περιέχουν συχνά ακραίες τιμές που δεν εξαλείφονται πλήρως με την κανονικοποίηση,

ο **K-Medoids** (που χρησιμοποιεί πραγματικά σημεία ως κέντρα) αναμένεται να προσφέρει πιο σταθερές συστάδες σε σχέση με τον **K-Means**.

3. **DBSCAN (Ester et al., 1996):** Επιλέχθηκε για τη μοναδική του ικανότητα να εντοπίζει **συστάδες ακανόνιστου σχήματος (arbitrary shaped clusters)** και, κυρίως, για την ικανότητά του να απομονώνει τον **θόρυβο (noise/outliers)**. Σε αντίθεση με τους δύο προηγούμενους που "αναγκάζουν" κάθε πελάτη να ανήκει σε μια ομάδα, ο **DBSCAN** μπορεί να αναδείξει πελάτες που δεν ταιριάζουν σε κανένα προφίλ.

4.9. Βελτιστοποίηση των Παραμέτρων (Hyperparameter Tuning)

Η επιλογή των βέλτιστων υπερ-παραμέτρων είναι καθοριστική για την απόδοση των αλγορίθμων μη εποπτευόμενης μάθησης. Σε αντίθεση με την εποπτευόμενη μάθηση, όπου υπάρχει "ετικέτα" (ground truth), εδώ η επιλογή του βέλτιστου αριθμού συστάδων (**k**) βασίζεται στην εσωτερική δομή και συνοχή των δεδομένων.

Για τον προσδιορισμό του βέλτιστου **k** στους αλγορίθμους K-Means και K-Medoids, εφαρμόστηκαν συνδυαστικά η **Μέθοδος του Αγκώνα** και ο **Δείκτης Σιλουέτας**.

4.9.1. Η Μέθοδος του Αγκώνα (Elbow Method)

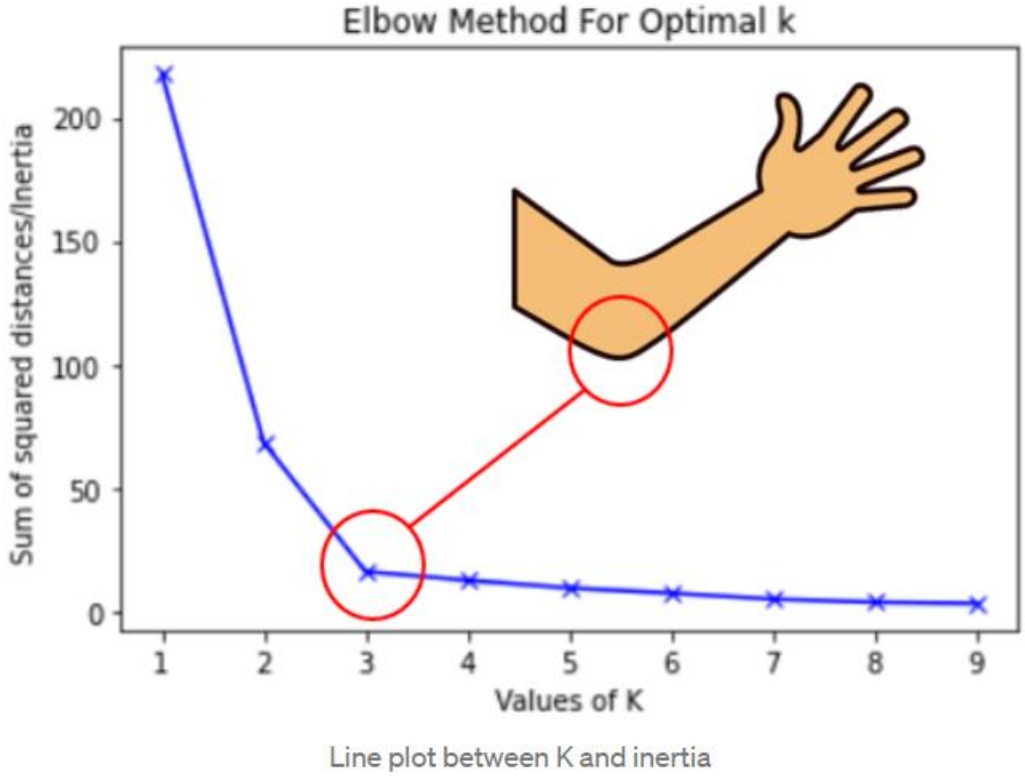
Η μέθοδος αυτή βασίζεται στον υπολογισμό του **Αθροίσματος Τετραγωνικών Σφαλμάτων (SSE)**, γνωστό και ως **Inertia** (Αδράνεια). Το **SSE** μετρά τη **Συμπαγή Δομή** των **Ομάδων** και ορίζεται ως το **Άθροισμα** των **Τετραγωνικών Αποστάσεων** κάθε **Σημείου** από το **Κέντρο** της **Συστάδας** του.

Μαθηματικός Τύπος:

$$\mathbf{SSE} = \sum_{\mathbf{i=1}}^{\mathbf{k}} \sum_{\mathbf{x} \in \mathbf{C_i}} \left|\left|\mathbf{x} - \boldsymbol{\mu}_\mathbf{i}\right|\right|^2 \qquad \textbf{(Εξίσωση 4.6)}$$

Όπου:

- **k**: Ο **Αριθμός** των **Συστάδων**.
- **μ_i**: Το **Κέντρο** (centroid) της **Συστάδας C_i**.



Εικόνα 4.9.1. Elbow Method

Διαδικασία Επιλογής:

Καθώς αυξάνεται το πλήθος των ομάδων (k), το SSE μειώνεται αναπόφευκτα (τείνοντας στο 0 όταν k=n). Στόχος είναι ο εντοπισμός του **Σημείου Καμπής** ("αγκώνας"), όπως φαίνεται χαρακτηριστικά στην **Εικόνα 4.9.1**, μετά το οποίο η μείωση του σφάλματος γίνεται οριακή. μετά το οποίο η μείωση του σφάλματος γίνεται οριακή. Το σημείο αυτό υποδεικνύει ότι η περαιτέρω διάσπαση των ομάδων δεν προσφέρει σημαντική πληροφορία, αλλά οδηγεί σε υπερ -προσαρμογή (overfitting) **[Syakur et al., 2018]**.

4.9.2. Ανάλυση Σιλουέτας (Silhouette Analysis)

Επειδή η **Ερμηνεία** του "αγκώνα" είναι συχνά υποκειμενική, χρησιμοποιήθηκε συμπληρωματικά ο **Συντελεστής Σιλουέτας (Silhouette Score)** για την επικύρωση της **Βέλτιστης Τιμής** του **k**. Ο **Συντελεστής Σιλουέτας**, που προτάθηκε από τον **Rousseeuw (1987)**, αποτελεί έναν από τους πιο αξιόπιστους **Δείκτες** για την ερμηνεία και την επικύρωση της συνοχής των δεδομένων. Ο **Δείκτης** αυτός μετρά πόσο όμοιο είναι ένα σημείο με τη δική του συστάδα (cohesion) σε σύγκριση με τις γειτονικές συστάδες (separation).

Εικόνα 4.9.2. Silhouette Method

Για κάθε **Σημείο i**, ο συντελεστής s(i) ορίζεται ως:

$$\mathbf{s(i)} = \frac{\mathbf{b(i)} - \mathbf{a(i)}}{\mathbf{max\{a(i), b(i)\}}} \qquad \textbf{(Εξίσωση 4.7)}$$

Όπου:

- **a(i):** Η **Μέση Απόσταση** του σημείου **i** από όλα τα άλλα σημεία στην **ίδια** συστάδα (ενδο-ομαδική συνοχή).

- **b(i):** Η μέση απόσταση του σημείου **i** από όλα τα σημεία της πλησιέστερης **Γειτονικής Συστάδας** (δια-ομαδικός διαχωρισμός).

Ερμηνεία: Το εύρος τιμών είναι **[-1, 1]**. Τιμές κοντά στο **+1** υποδηλώνουν ότι τα Σημεία έχουν ανατεθεί σωστά και απέχουν πολύ από τη **Γειτονική Ομάδα** (**Βέλτιστη Ομαδοποίηση**). Τιμές κοντά στο **0** υποδηλώνουν αλληλοεπικάλυψη **Συνόρων**, ενώ **Αρνητικές Τιμές** υποδηλώνουν πιθανή λανθασμένη ανάθεση **[Rousseeuw, 1987]**. Στόχος της **Βελτιστοποίησης** είναι η **Μεγιστοποίηση** του **Μέσου Όρου** του **Δείκτη**.

Στην **Εικόνα 4.9.2.** παρατηρούμε ένα **Συνδυαστικό Διάγραμμα Αξιολόγησης** του **Βέλτιστου Αριθμού Συστάδων** (**k**). Η **Μπλε Γραμμή** (Αριστερός Άξονας) αναπαριστά το **Inertia** (**Μέθοδος Αγκώνα**), ενώ η **Πράσινη Γραμμή** (Δεξιός Άξονας) αναπαριστά τον Μέ**σο Συντελεστή Σιλουέτας**. Παρατηρείται ότι η **Μέγιστη Τιμή** της **Σιλουέτας** επιτυγχάνεται στο **k = 4**, σημείο που συμπίπτει με την **Περιοχή Καμπής** του **Inertia**, επιβεβαιώνοντας την **Επιλογή** των **Τεσσάρων** (**4**) **Συστάδων** ως **βέλτιστη λύση**.

Η **Συνδυαστική Χρήση** των **Δύο** (**2**) **Μεθόδων**, όπως απεικονίζεται στην **Εικόνα 4.9.2**, επιτρέπει την ασφαλέστερη λήψη απόφασης, καθώς η **Κορυφή** (peak) του **Δείκτη Σιλουέτας** συχνά επιβεβαιώνει το **Υποκειμενικό Σημείο Καμπής** του **Αγκώνα**.

4.9.3. Παραμετροποίηση DBSCAN (K-Distance Graph)

Σε αντίθεση με τον **K-Means**, ο αλγόριθμος **DBSCAN** δεν απαιτεί τον εκ των προτέρων καθορισμό του αριθμού των συστάδων, αλλά εξαρτάται από δύο κρίσιμες παραμέτρους πυκνότητας:

1. **MinPts:** Ο **Ελάχιστος Αριθμός** σημείων για να θεωρηθεί μια περιοχή "πυκνή".
2. **Epsilon (ε):** Η **Μέγιστη Ακτίνα** γειτονίας γύρω από ένα σημείο.

Για τον βέλτιστο προσδιορισμό τους, ακολουθήθηκε η **Ευρεστική Μέθοδος** που πρότειναν οι δημιουργοί του αλγορίθμου **[Ester et al., 1996]**:

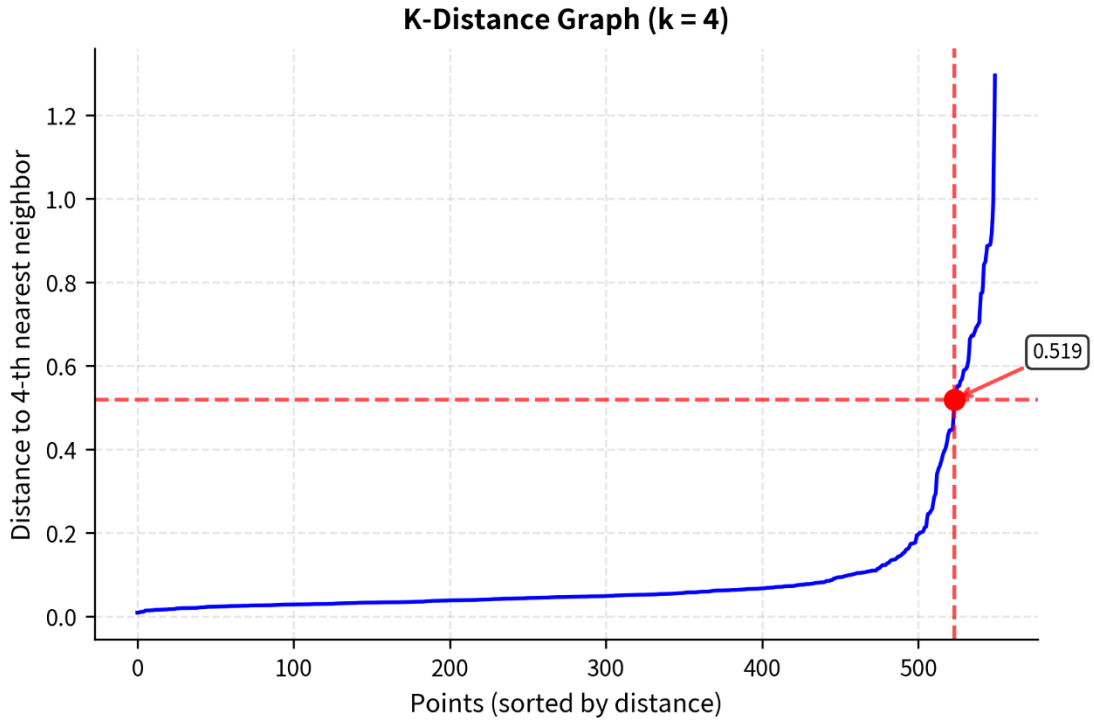
A. Επιλογή του MinPts:

Η τιμή του **MinPts** καθορίστηκε βάσει της διαστατικότητας (D) του συνόλου δεδομένων. Ο γενικός κανόνας ορίζει ότι **MinPts ≥ D + 1**. Για δεδομένα **Τριών** (**3**) **Διαστάσεων** (RFM), επιλέχθηκε η τιμή **MinPts = 4**, η οποία παρέχει επαρκή ισορροπία μεταξύ της ευαισθησίας στο θόρυβο και της ανάδειξης **Μικρών Συστάδων**.

B. Επιλογή του Epsilon (ε) με το K-Distance Graph:

Για τον υπολογισμό της ακτίνας **ε**, χρησιμοποιήθηκε το **Διάγραμμα Κ-Αποστάσεων (K-Distance Graph)**. Η διαδικασία που ακολουθήθηκε είναι η εξής:

1. Για κάθε σημείο στο dataset, υπολογίστηκε η απόσταση από τον **k-οστό πλησιέστερο γείτονά του** (k-nearest neighbor), όπου k = MinPts.
2. Οι αποστάσεις αυτές ταξινομήθηκαν σε **αύξουσα σειρά** και αποτυπώθηκαν σε γράφημα.



Εικόνα 4.9.3. Γράφημα K-Distance

Ερμηνεία Γραφήματος:

Το **Διάγραμμα K-Distance** για τον προσδιορισμό της βέλτιστης **Παραμέτρου ε** (Epsilon) ερμηνεύεται ως εξής: Στον άξονα Y αποτυπώνεται η απόσταση στο 4ο πλησιέστερο γείτονα. Το σημείο τομής των κόκκινων διακεκομμένων γραμμών (το "γόνατο" της καμπύλης) υποδεικνύει τη βέλτιστη τιμή κατωφλίου (threshold), η οποία διαχωρίζει τα σημεία πυκνότητας από τις ακραίες τιμές (outliers). Στο συγκεκριμένο παράδειγμα, η βέλτιστη τιμή εντοπίστηκε στο **ε ≈ 0.519**. Η απότομη άνοδος (η "ανηφόρα" στο τέλος της μπλε γραμμής) είναι οι **Outliers**.

Ο στόχος είναι ο εντοπισμός του σημείου καμπής, γνωστό ως **"γόνατο" (knee)** ή "αγκώνας".

- Το οριζόντιο τμήμα της καμπύλης αντιστοιχεί σε σημεία που ανήκουν σε περιοχές υψηλής πυκνότητας (συστάδες).
- Το σημείο όπου η καμπύλη αρχίζει να ανηφορίζει απότομα υποδηλώνει τη μετάβαση από τις συστάδες στον "θόρυβο".

Όπως απεικονίζεται στην **Εικόνα 4.9.3**, το βέλτιστο **ε** αντιστοιχεί στην τιμή της απόστασης στο σημείο καμπής. Οποιαδήποτε τιμή μεγαλύτερη από αυτή θα ενσωμάτωνε θόρυβο στις συστάδες, ενώ μικρότερη τιμή θα κατακερμάτιζε τις **Συστάδες**.

4.10. Μετρικές Αξιολόγησης

Στα προβλήματα **Μη Εποπτευόμενης Μάθησης** (Unsupervised Learning), λόγω της απουσίας προκαθορισμένων ετικετών (ground truth), η αξιολόγηση της ποιότητας της **Συσταδοποίησης** δεν μπορεί να γίνει με παραδοσιακούς δείκτες ακρίβειας. Αντ' αυτού, χρησιμοποιούνται δείκτες **εσώτερης εγκυρότητας (internal validity indices)**, οι οποίοι αξιολογούν τη δομή των συστάδων βάσει **Δύο** (**2**) **Κριτηρίων**:

1. **Συμπάγεια (Compactness):** Πόσο κοντά βρίσκονται τα σημεία της ίδιας **Ομάδας** μεταξύ τους.
2. **Διαχωρισμός (Separation):** Πόσο απέχουν οι **Ομάδες** η μία από την άλλη.

Για την αξιολόγηση των αποτελεσμάτων της παρούσας μελέτης, χρησιμοποιήθηκαν οι εξής **Τρεις** (**3**) **Μετρικές**:

4.10.1. Συντελεστής Σιλουέτας (Silhouette Score)

Όπως αναλύθηκε στην ενότητα της βελτιστοποίησης (4.9.2), ο Συντελεστής Σιλουέτας μετρά πόσο καλά ταιριάζει ένα αντικείμενο στη συστάδα του σε σχέση με τις γειτονικές.

- **Ερμηνεία:** Το εύρος τιμών είναι [**-1, 1**]. Τιμές κοντά στο **+1** υποδηλώνουν εξαιρετικό διαχωρισμό, το **0** υποδηλώνει αλληλοεπικάλυψη, ενώ αρνητικές τιμές υποδηλώνουν λανθασμένη ανάθεση. Στόχος είναι η **μεγιστοποίηση** του δείκτη.

4.10.2. Δείκτης Davies-Bouldin (DBI)

Ο δείκτης **Davies-Bouldin** (εισήχθη από τους **Davies & Bouldin, 1979**) αξιολογεί τη μέση "**ομοιότητα**" μεταξύ των **Συστάδων**. Η **Ομοιότητα** ορίζεται ως ο λόγος της **Ενδο-ομαδικής Διασποράς** προς τη **Δια-ομαδική Απόσταση.** Για να εξηγήσουμε με πιο απλά λόγια την φράση αυτή, θα πρέπει να αναφερθούμε σε μία εικόνα. Ας φανταστούμε ότι έχουμε δημιουργήσει **Ομάδες** (Clusters) Πελατών. Για να πούμε ότι οι Ομάδες μας είναι "καλές" θα πρέπει να συντρέχουν ταυτόχρονα δύο (2) προϋποθέσεις:

1. **Ενδο-ομαδική Διασπορά (Within-Cluster Dispersion):** Εκφράζει τη **Συνοχή** της ομάδας. Δείχνει πόσο κοντά βρίσκονται τα σημεία (πελάτες) στο κέντρο της **Συστάδας** τους. Στόχος είναι η **Ελαχιστοποίηση** αυτής της τιμής (μικρός αριθμητής), ώστε η ομάδα να είναι **Συμπαγής**.
2. **Δια-ομαδική Απόσταση (Between-Cluster Distance):** Εκφράζει τον **διαχωρισμό** των ομάδων. Δείχνει πόσο απέχουν τα κέντρα των **Συστάδων** μεταξύ τους. Στόχος είναι η **Μεγιστοποίηση** αυτής της τιμής (μεγάλος παρονομαστής), ώστε οι ομάδες να είναι σαφώς **Διακριτές** και να μην **αλληλεπικαλύπτονται**. Όλα τα παραπάνω συνοψίζονται στο κλάσμα:

$$\frac{\textit{Πόσο σκορπισμένα είναι τα Σημεία}}{\textit{Πόσο απέχουν οι Ομάδες μεταξύ τους}'}$$

που εξηγείται ως εξής:

«Αν οι ομάδες είναι πολύ "σκορπισμένες" και ταυτόχρονα πολύ "κοντά" η μία στην άλλη, το κλάσμα μεγαλώνει». Αυτό σημαίνει ότι οι ομάδες **μοιάζουν πολύ μεταξύ τους (αλληλεπικαλύπτονται)**, κάτι που είναι **κακό** για το **clustering**. Γι' αυτό στον **Davies-Bouldin** ψάχνουμε την **ελάχιστη** τιμή.

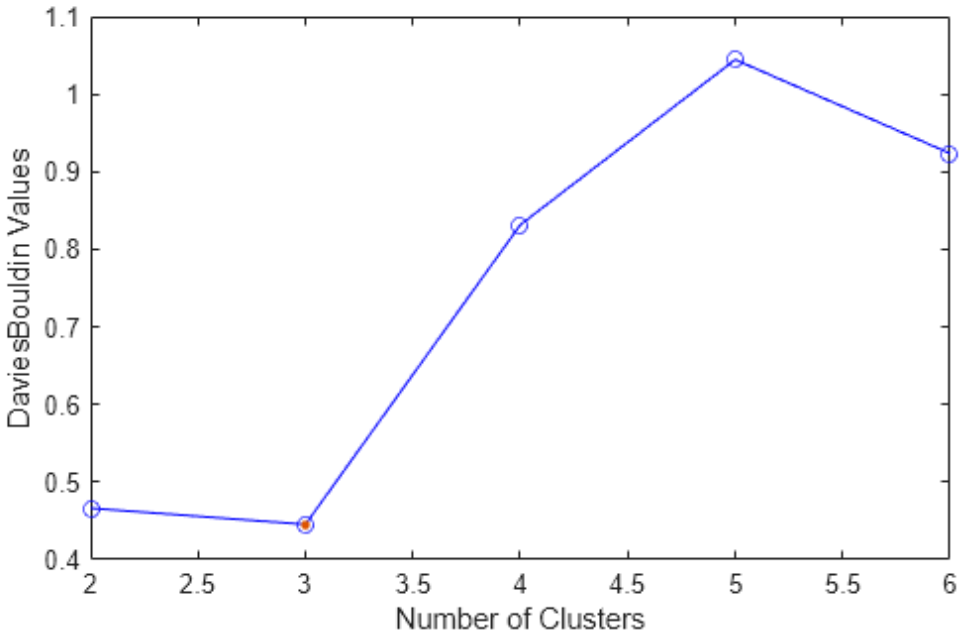
Ο **Μαθηματικός Τύπος** είναι:

$$\mathbf{DB} = \frac{\mathbf{1}}{\mathbf{k}} \sum_{\mathbf{i=1}}^{\mathbf{k}} \max_{j \neq i} \left(\frac{\boldsymbol{\sigma_i} + \boldsymbol{\sigma_j}}{\boldsymbol{d(c_i, c_j)}} \right) \qquad \textbf{(Εξίσωση 4.8)}$$

Όπου:

- $\mathbf{k}$: Ο **Αριθμός** των **Συστάδων**.
- $\boldsymbol{\sigma_i}$: Η **Μέση Απόσταση** των σημείων της συστάδας **i** από το **Κέντρο** της (Ομοιογένεια).
- $\mathbf{d(c_i, c_j)}$: Η **Απόσταση** μεταξύ των **Κέντρων** των **Συστάδων i** και **j** (Διαχωρισμός).
- Τα Σύμβολα **i** και **j** λειτουργούν ως **Δείκτες Αρίθμησης (indices)** που αντιπροσωπεύουν δύο **Διαφορετικές Συστάδες** (clusters) κάθε φορά:
    - $\mathbf{i}$ είναι το **Reference Cluster:** Η **Συστάδα που εξετάζουμε εκείνη τη στιγμή** (η "τρέχουσα" συστάδα). Ο τύπος έχει ένα άθροισμα ($\mathbf{\Sigma}$) που σημαίνει ότι παίρνουμε κάθε συστάδα μία-μία (π.χ. Συστάδα 1, μετά Συστάδα 2, κ.λπ.) και την ονομάζουμε **i**.
    - **j** είναι το **Comparison Cluster:** Η **άλλη Συστάδα** με την οποία συγκρίνουμε την **i**.

- $\max_{j \neq i}$ σημαίνει: "Για την **τρέχουσα Συστάδα i**, ψάξε όλες τις υπόλοιπες **Συστάδες j** και βρες ποια από αυτές είναι η 'χειρότερη' (δηλαδή ποια μοιάζει περισσότερο με την **i** ή είναι πιο κοντά της)".
    - Ο **περιορισμός j ≠ i** είναι πολύ σημαντικός γιατί μας λέει ότι **δεν συγκρίνουμε ποτέ μια Συστάδα με τον εαυτό της**. Το **j** πρέπει να είναι πάντα διαφορετικό από το **i**.



Εικόνα 4.10.2. Γράφημα Davies-Bouldin

Ερμηνεία: Λόγω του ορισμού του, χαμηλότερη τιμή σημαίνει ότι οι ομάδες είναι συμπαγείς (μικρή διασπορά **Σ**) και απέχουν πολύ μεταξύ τους (μεγάλη απόσταση **d**). Επομένως, ο βέλτιστος αριθμός συστάδων αντιστοιχεί στην **ελαχιστοποίηση** του δείκτη DBI **[Davies & Bouldin, 1979]**.

Η διαδικασία **Επιλογής** απεικονίζεται στην **Εικόνα 4.10.2**, όπου το **Βέλτιστο Πλήθος Συστάδων** αντιστοιχεί στο κατώτατο σημείο (global minimum) της καμπύλης, υποδεικνύοντας τη λύση με τη μικρότερη δυνατή αλληλοεπικάλυψη. Γραφική αναπαράσταση των τιμών του δείκτη **Davies-Bouldin** για διαφορετικό πλήθος συστάδων. Σε αντίθεση με άλλους δείκτες, εδώ επιδιώκουμε την **Ελαχιστοποίηση** της τιμής. Στο συγκεκριμένο

παράδειγμα, η βέλτιστη λύση εντοπίζεται στο **k=3** (κόκκινη ένδειξη), όπου επιτυγχάνεται ο καλύτερος συνδυασμός **Συμπαγών** και **Διαχωρισμένων Ομάδων**.

4.10.3. Δείκτης Calinski-Harabasz (CH Index)

Ο δείκτης **Calinski-Harabasz**, γνωστός στη βιβλιογραφία και ως **Κριτήριο Λόγου Διασποράς (Variance Ratio Criterion)**, αναπτύχθηκε από τους **Caliński και Harabasz (1974)**.

Όπως και ο προηγούμενος δείκτης, έτσι και αυτός αξιολογεί την **ποιότητα** της **συσταδοποίησης** συγκρίνοντας τη συνοχή με τον **διαχωρισμό**. Ωστόσο, η μαθηματική του προσέγγιση είναι διαφορετική και συχνά πιο αυστηρή για τον εντοπισμό πυκνών ομάδων.

Για να γίνει κατανοητή η λειτουργία του, ο δείκτης εξετάζει δύο μεγέθη:

1. **Διασπορά Μεταξύ των Συστάδων (Between-Cluster Dispersion - SS_B):**

Μετρά πόσο μακριά βρίσκονται τα κέντρα των ομάδων το ένα από το άλλο. Όσο πιο μακριά είναι, τόσο καλύτερος ο διαχωρισμός. Επομένως, θέλουμε αυτό το νούμερο να είναι **Μεγάλο** (άρα **Μεγάλος Αριθμητής**).

2. **Διασπορά Εντός των Συστάδων (Within-Cluster Dispersion - SS_W):**

Μετρά πόσο κοντά είναι τα σημεία (πελάτες) στο κέντρο της δικής τους ομάδας. Όσο πιο κοντά είναι, τόσο πιο συμπαγής η ομάδα. Επομένως, θέλουμε αυτό το νούμερο να είναι **Μικρό** (άρα **Μικρός Παρονομαστής**).

Εννοιολογικά, ο τύπος λειτουργεί ως εξής:

$$\textbf{Δείκτης CH} = \frac{\textbf{Διαχωρισμός Ομάδων}}{\textbf{Συνοχή Ομάδων}}$$

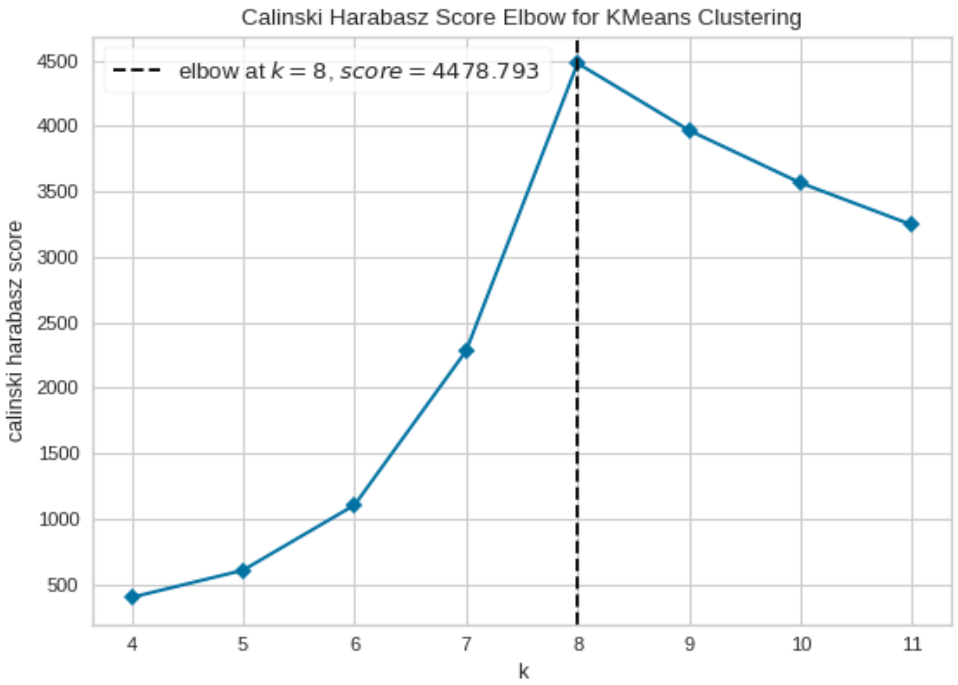
Σε αντίθεση με τον **Davies-Bouldin**, εδώ επιδιώκουμε τη **Μεγιστοποίηση** της τιμής. Ένας υψηλός **δείκτης CH** σημαίνει ότι έχουμε πετύχει ταυτόχρονα μεγάλη απόσταση μεταξύ των ομάδων (αριθμητής) και μεγάλη πυκνότητα μέσα στις ομάδες (παρονομαστής).

Ο **Μαθηματικός Τύπος** ορίζεται ως:

$$\mathbf{CH} = \frac{\mathbf{SS_B}/(\mathbf{k}-\mathbf{1})}{\mathbf{SS_W}/(\mathbf{n}-\mathbf{k})} \qquad \textbf{(Εξίσωση 4.9)}$$

Όπου:

- **n**: Το **Συνολικό Πλήθος** των παρατηρήσεων (πελατών).
- **k**: Ο **Αριθμός** των **Συστάδων**.
- $\mathbf{SS_B}$: Το **Άθροισμα Τετραγώνων** της **Απόστασης** μεταξύ των **Συστάδων** (Between-cluster sum of squares).
- $\mathbf{SS_W}$: Το **Άθροισμα Τετραγώνων** της **Απόστασης** εντός των **συστάδων** (Within-cluster sum of squares).



Εικόνα 4.10.3. Γράφημα Calinski-Harabasz

- Οι όροι **(k-1)** και **(n-k)** λειτουργούν ως **Βαθμοί Ελευθερίας** για την **Κανονικοποίηση** του αποτελέσματος. Με πιο απλά λόγια, είναι οι **Ρυθμιστικοί Παράγοντες** που καθιστούν τη **Σύγκριση Δίκαιη**. Αυτό σημαίνει ότι καθώς αυξάνεται

ο **Αριθμός** των **Συστάδων** (**k**), η **Εσωτερική Διασπορά** ($\mathbf{SS_W}$) τείνει φυσιολογικά να μειώνεται. Αν δεν υπήρχε αυτή η διόρθωση, ο δείκτης θα ευνοούσε λανθασμένα τη δημιουργία όλο και περισσότερων ομάδων. Η διαίρεση με τους βαθμούς ελευθερίας "κανονικοποιεί" το αποτέλεσμα, επιτρέποντας την αντικειμενική σύγκριση ανεξαρτήτως του πλήθους των ομάδων.

Ερμηνεία και Επιλογή:

Υψηλότερη τιμή του δείκτη υποδηλώνει ότι οι **Συστάδες** είναι πυκνές (καλά ορισμένες) και απομακρυσμένες μεταξύ τους. Στόχος είναι η εύρεση του αριθμού **k** που δίνει την **Υψηλότερη Κορυφή** (peak) στο διάγραμμα **[Caliński & Harabasz, 1974]**.

Η διαδικασία **Επιλογής** απεικονίζεται στην **Εικόνα 4.10.3**. Σε αυτή, βλέπουμε τη **Γραφική απεικόνιση** του δείκτη Calinski-Harabasz για διαφορετικές τιμές του **k**. Σε αντίθεση με το **Inertia** που μειώνεται συνεχώς, ο δείκτης **CH** εμφανίζει ένα **Σαφές Μέγιστο** (κορυφή). Στο συγκεκριμένο παράδειγμα, η **Υψηλότερη Τιμή** εντοπίζεται στο **k=8**, υποδεικνύοντας την **Ομαδοποίηση** με τον βέλτιστο λόγο διαχωρισμού προς συνοχή.

5. Υλοποίηση Μεθοδολογίας και Εφαρμογή Αλγορίθμων

5.1. Εισαγωγή στο Περιβάλλον Υλοποίησης RFM Master Tool v.4.5.0.

Για τις ανάγκες της εργασίας μας και με γνώμονα να αντλήσουμε ασφαλή συμπεράσματα για τα δεδομένα που πρόκειται να διαχειριστούν οι **Αλγόριθμοι** που υλοποιήσαμε στο πρόγραμμά μας **RFM Master Tool V4.5.0**, προχωρήσαμε στην εκτέλεση της διαδικασίας μετασχηματισμού και **Ανάλυσης** των **Δεδομένων**, με τη βοήθεια ανάπτυξης εξειδικευμένου κώδικα στη γλώσσα προγραμματισμού **Python**. Η επιλογή της συγκεκριμένης γλώσσας βασίστηκε στην ευελιξία της στη διαχείριση μεγάλου όγκου δεδομένων (Big Data) και στην πληθώρα βιβλιοθηκών για **Στατιστική Ανάλυση** και **Μηχανική Μάθηση**.

Η υλοποίηση βασίστηκε σε τρεις κεντρικούς άξονες:

Βιβλιοθήκες και Περιβάλλον Εργασίας

Για την ανάλυση χρησιμοποιήθηκαν οι ακόλουθες θεμελιώδεις βιβλιοθήκες της επιστήμης δεδομένων:

- **Pandas:** Για τη φόρτωση, καθαρισμό και διαχείριση των δεδομένων σε μορφή DataFrames.
- **Matplotlib & Seaborn:** Για την οπτικοποίηση των αποτελεσμάτων και τη δημιουργία στατιστικών διαγραμμάτων υψηλής ευκρίνειας.
- **Scikit-learn:** Για την κανονικοποίηση των δεδομένων και την εφαρμογή του αλγορίθμου K-Means (στην επόμενη φάση).

Αλγόριθμος Καθαρισμού και Υπολογισμού (ETL Process)

Ο κώδικας ακολουθεί τη διαδικασία **ETL (Extract, Transform, Load)**, η οποία διασφαλίζει την ακεραιότητα των αποτελεσμάτων. Συγκεκριμένα, ο αλγόριθμος εκτελεί τα εξής βήματα:

1. **Διαχείριση Κωδικοποίησης (Encoding):** Κατά τη φόρτωση του αρχείου CSV, εφαρμόστηκε κωδικοποίηση ISO-8859-1 για την ορθή αναγνώριση ειδικών χαρακτήρων στις περιγραφές των προϊόντων.
2. **Χρονική Κανονικοποίηση:** Η στήλη InvoiceDate μετατράπηκε από απλό κείμενο (string) σε αντικείμενο ημερομηνίας (datetime object), επιτρέποντας μαθηματικές πράξεις για τον υπολογισμό του Recency.
3. **Φιλτράρισμα Ακυρώσεων:** Εφαρμόστηκε λογικό φίλτρο για την αφαίρεση εγγραφών που περιείχαν τον χαρακτήρα 'C' (Cancellations) στον κωδικό απόδειξης, καθώς και εγγραφών με αρνητικές τιμές ποσότητας ή τιμής.

Αυτοματοποιημένη Συσσωμάτωση (Aggregation Logic)

Το κρισιμότερο τμήμα του κώδικα αφορά τη μετατροπή των συναλλαγών σε πελατοκεντρικά δεδομένα. Χρησιμοποιώντας τη συνάρτηση groupby() της βιβλιοθήκης Pandas, ομαδοποιήθηκαν οι εγγραφές ανά Customer ID και εφαρμόστηκαν ταυτόχρονα οι εξής συναρτήσεις:

- **Για το Recency:** lambda x: (snapshot_date - x.max()).days
    - *Ερμηνεία:* Αφαιρεί την τελευταία ημερομηνία αγοράς του πελάτη από την ημερομηνία αναφοράς (10/12/2011).
- **Για το Frequency:** 'nunique'
    - *Ερμηνεία:* Καταμετρά τους μοναδικούς κωδικούς αποδείξεων.
- **Για το Monetary:** 'sum'
    - *Ερμηνεία:* Αθροίζει τη συνολική αξία των αγορών.

Η παραπάνω διαδικασία εξασφαλίζει την **επαναληψιμότητα (reproducibility)** της έρευνας, καθώς ο κώδικας μπορεί να εφαρμοστεί με τον ίδιο ακριβώς τρόπο σε οποιοδήποτε νέο σύνολο δεδομένων με την ίδια δομή.

5.2. Μετασχηματισμός Δεδομένων και Δημιουργία Πίνακα RFM (Recency, Frequency, Monetary)

Όπως αναφέραμε αρχικά στην **Ενότητα 4.4 Περιγραφή Δεδομένων** (Data Description), το αρχικό **Σύνολο Δεδομένων** (Retail_Online_II) βρίσκεται σε μορφή **συναλλαγών (transactional data)**, όπου κάθε εγγραφή αντιστοιχεί σε ένα προϊόν μιας απόδειξης (**Εικόνα 5.1.a.** και **Εικόνα 5.1.b.**). Για την εφαρμογή της **μεθόδου RFM**, απαιτείται η μετατροπή του συνόλου αυτού σε επίπεδο **πελάτη (customer-level data)**.

Online_Retail_II.csv

```
1    Invoice,StockCode,Description,Quantity,InvoiceDate,Price,Customer ID,Country
2    489434,85048,15CM CHRISTMAS GLASS BALL 20 LIGHTS,12,2009-12-01 07:45:00,6.95,13085.0,United Kingdom
3    489434,79323P,PINK CHERRY LIGHTS,12,2009-12-01 07:45:00,6.75,13085.0,United Kingdom
4    489434,79323W, WHITE CHERRY LIGHTS,12,2009-12-01 07:45:00,6.75,13085.0,United Kingdom
5    489434,22041,"RECORD FRAME 7"" SINGLE SIZE ",48,2009-12-01 07:45:00,2.1,13085.0,United Kingdom
6    489434,21232,STRAWBERRY CERAMIC TRINKET BOX,24,2009-12-01 07:45:00,1.25,13085.0,United Kingdom
7    489434,22064,PINK DOUGHNUT TRINKET POT ,24,2009-12-01 07:45:00,1.65,13085.0,United Kingdom
8    489434,21871,SAVE THE PLANET MUG,24,2009-12-01 07:45:00,1.25,13085.0,United Kingdom
9    489434,21523,FANCY FONT HOME SWEET HOME DOORMAT,10,2009-12-01 07:45:00,5.95,13085.0,United Kingdom
10   489435,22350,CAT BOWL ,12,2009-12-01 07:46:00,2.55,13085.0,United Kingdom
11   489435,22349,"DOG BOWL , CHASING BALL DESIGN",12,2009-12-01 07:46:00,3.75,13085.0,United Kingdom
12   489435,22195,HEART MEASURING SPOONS LARGE,24,2009-12-01 07:46:00,1.65,13085.0,United Kingdom
13   489435,22353,LUNCHBOX WITH CUTLERY FAIRY CAKES ,12,2009-12-01 07:46:00,2.55,13085.0,United Kingdom
14   489436,48173C,DOOR MAT BLACK FLOCK ,10,2009-12-01 09:06:00,5.95,13078.0,United Kingdom
15   489436,21755,LOVE BUILDING BLOCK WORD,18,2009-12-01 09:06:00,5.45,13078.0,United Kingdom
```

Εικόνα 5.1.a. Αρχική Μορφή Αρχείου Retail_Online_II (15 πρώτες εγγραφές)

```
1067366   581587,23256,CHILDRENS CUTLERY SPACEBOY ,4,2011-12-09 12:50:00,4.15,12680.0,France
1067367   581587,22613,PACK OF 20 SPACEBOY NAPKINS,12,2011-12-09 12:50:00,0.85,12680.0,France
1067368   581587,22899,CHILDREN'S APRON DOLLY GIRL ,6,2011-12-09 12:50:00,2.1,12680.0,France
1067369   581587,23254,CHILDRENS CUTLERY DOLLY GIRL ,4,2011-12-09 12:50:00,4.15,12680.0,France
1067370   581587,23255,CHILDRENS CUTLERY CIRCUS PARADE,4,2011-12-09 12:50:00,4.15,12680.0,France
1067371   581587,22138,BAKING SET 9 PIECE RETROSPOT ,3,2011-12-09 12:50:00,4.95,12680.0,France
1067372   581587,POST,POSTAGE,1,2011-12-09 12:50:00,18.0,12680.0,France
1067373
```

Εικόνα 5.1.b. Αρχική Μορφή Αρχείου Retail_Online_II (7 τελευταίες εγγραφές)

Η διαδικασία αυτή εκτελέστηκε αυτοματοποιημένα στη μνήμη του συστήματος μέσω της **Βιβλιοθήκης Pandas**, ακολουθώντας τα εξής βήματα συσσωμάτωσης (aggregation):

1. **Ορισμός Ημερομηνίας Αναφοράς (Snapshot Date):** Το σύνολο δεδομένων καλύπτει τη χρονική περίοδο από 01/12/2009 έως **09/12/2011**. Προκειμένου να υπολογιστεί η μεταβλητή **Recency**, ορίστηκε ως ημερομηνία αναφοράς η **10η**

Δεκεμβρίου 2011 (μια ημέρα μετά την τελευταία καταγεγραμμένη συναλλαγή). Αυτό διασφαλίζει ότι οι δείκτες πρόσφατης δραστηριότητας υπολογίζονται με ακρίβεια ως προς το χρονικό σημείο της ανάλυσης.

Tests > Online_Retail_II_RFM_Clustering.csv > data

```
1   Customer ID,Recency,Frequency,MonetaryValue
2   12346.0,326,12,77556.46
3   12347.0,2,8,5633.32
4   12348.0,75,5,2019.4
5   12349.0,19,4,4428.69
6   12350.0,310,1,334.4
7   12351.0,375,1,300.93
8   12352.0,36,10,2849.84
9   12353.0,204,2,406.76
10  12354.0,232,1,1079.4
11  12355.0,214,2,947.61
12  12356.0,23,6,6373.68
13  12357.0,33,3,18287.66
14  12358.0,2,5,3887.07
15  12359.0,58,10,8935.94
16  12360.0,52,8,4252.89
17  12361.0,287,4,511.25
18  12362.0,3,11,5356.23
19  12363.0,110,2,552.0
20  12364.0,8,4,1313.1
```

Εικόνα 5.1.c. Μορφή Αρχείου RFM από το Retail_Online_II (20 πρώτες εγγραφές)

```
5873   18281.0,181,2,201.14
5874   18282.0,8,2,178.05
5875   18283.0,4,22,2736.65
5876   18284.0,432,1,461.68
5877   18285.0,661,1,427.0
5878   18286.0,477,2,1296.43
5879   18287.0,43,7,4182.99
5880
```

Εικόνα 5.1.d. RFM Αρχείο από το Retail_Online_II (20 τελευταίες εγγραφές)

2. **Ομαδοποίηση (Grouping):** Τα δεδομένα ομαδοποιήθηκαν βάσει του μοναδικού αναγνωριστικού πελάτη (Customer ID).
3. **Υπολογισμός Μεταβλητών:** Για κάθε μοναδικό πελάτη εφαρμόστηκαν οι ακόλουθες **Συναρτήσεις**:
    - **Recency (R):** Υπολογίστηκε η διαφορά ημερών μεταξύ της Snapshot Date και της μέγιστης (πιο πρόσφατης) ημερομηνίας αγοράς (InvoiceDate).
    - **Frequency (F):** Καταμετρήθηκε το πλήθος των μοναδικών κωδικών απόδειξης (Invoice) για κάθε πελάτη.
    - **Monetary (M):** Αθροίστηκε το γινόμενο Quantity * Price για όλες τις συναλλαγές του πελάτη.

Το αποτέλεσμα της διαδικασίας ήταν η δημιουργία ενός νέου, **Δομημένου Πίνακα** (DataFrame), όπου κάθε γραμμή αντιπροσωπεύει έναν μοναδικό πελάτη με τις τρεις υπολογισμένες τιμές **RFM**. Έτσι, καταφέραμε με τον τρόπο αυτό να δημιουργήσουμε ένα Πελατοκεντρικό Αρχείο έτοιμο προς επεξεργασία με τους διαθέσιμους Αλγορίθμους που αναφέραμε ήδη στις **Ενότητες 2.13.1**, **2.13.2, 2.13.3** και **2.13.4.** Μέρος του Πίνακα αυτού φαίνεται στις **Εικόνες 5.1.c** και **5.1.d.**

5.3. Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis - EDA)

Έχοντας ολοκληρώσει τη διαδικασία **Μετασχηματισμού** των **Δεδομένων** και δημιουργίας του **Πίνακα RFM** (**Ενότητα 5.2**), προχωρούμε στη Στ**ατιστική Διερεύνηση** των μεταβλητών που προέκυψαν.

Η **Διερευνητική Ανάλυση** (EDA) αποτελεί κρίσιμο στάδιο πριν την εφαρμογή των **Αλγορίθμων Συσταδοποίησης**, καθώς μας επιτρέπει:

1. Να κατανοήσουμε την **Κατανομή** των δεδομένων (αν είναι συμμετρικά ή όχι).
2. Να εντοπίσουμε πιθανές **Συσχετίσεις** (Correlations) μεταξύ των χαρακτηριστικών.

3. Να αναγνωρίσουμε **Ακραίες Τιμές** (Outliers) που ενδέχεται να στρεβλώσουν τα αποτελέσματα του K-Means.

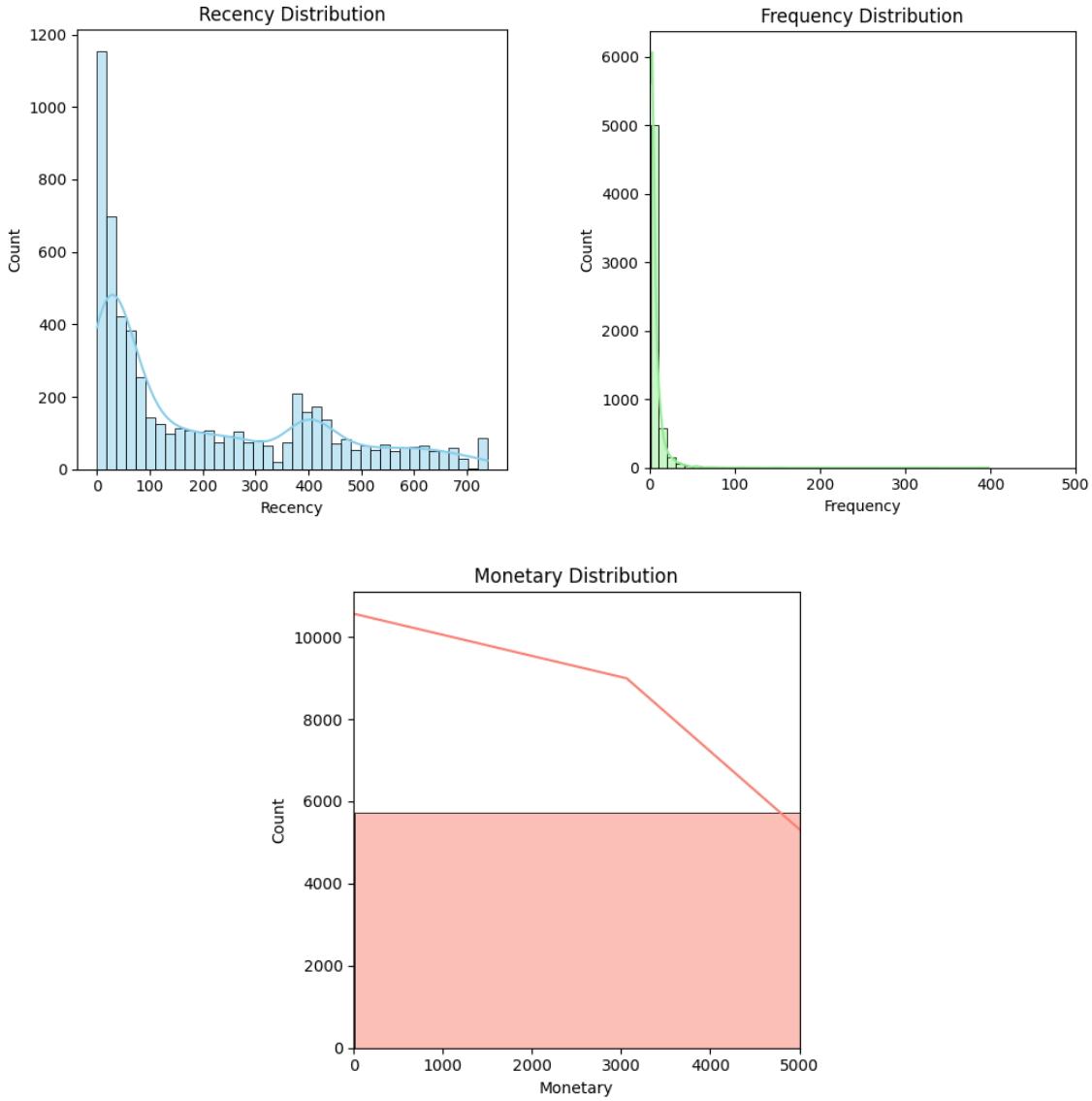
5.3.1. Περιγραφική Στατιστική των μεταβλητών RFM (Ιστογράμματα, Κατανομές)

Αρχικό στάδιο της Διερευνητικής Ανάλυσης αποτέλεσε ο υπολογισμός των **Βασικών Περιγραφικών Μέτρων Κεντρικής Τάσης και Διασποράς** (Μέσος, Διάμεσος, Τυπική Απόκλιση, Τεταρτημόρια). Η Οπτικοποίηση των Κατανομών πραγματοποιήθηκε συνδυαστικά μέσω Ιστογραμμάτων (Histograms) και Διαγραμμάτων Πυκνότητας (Kernel Density Plots - KDE), με στόχο την κατανόηση της γεωμετρίας και της συμμετρίας των δεδομένων.

Όπως παρατηρείται στα Διαγράμματα της **Εικόνας 5.3.1**:

- **Ως προς το Recency (Πρόσφατο):** Η κατανομή δεν ακολουθεί την τυπική μορφή, αλλά παρουσιάζει χαρακτηριστικά **Θετικής Ασυμμετρίας** με σαφή τάση **Δικόρυφης Κατανομής (Bimodal Distribution)**. Συγκεκριμένα, παρατηρείται μια πόλωση της πελατειακής βάσης:
  - ✓ Ο πρώτος "πόλος" αφορά μια μεγάλη συγκέντρωση ενεργών πελατών με πολύ πρόσφατες αγορές (κοντά στο 0).
  - ✓ Ο δεύτερος "πόλος" αφορά μια σημαντική ομάδα πελατών που έχουν αδρανήσει για διάστημα περίπου ενός έτους (αιχμή στις 360-400 ημέρες), γεγονός που υποδεικνύει εποχικότητα ή υψηλό ποσοστό διαφυγής (churn).
- **Ως προς το Frequency (Συχνότητα):** Η μεταβλητή εμφανίζει **έντονη Θετική Ασυμμετρία (Right-skewed/Positive Skewness)**. Η κατανομή χαρακτηρίζεται ως **"Long Tail"**, καθώς η συντριπτική πλειοψηφία των πελατών πραγματοποιεί ελάχιστες συναλλαγές (συνήθως 1-2), ενώ η καμπύλη εκτείνεται δεξιά καλύπτοντας ένα μικρό πλήθος πιστών πελατών με πολύ μεγάλη συχνότητα αγορών.
- **Ως προς το Monetary (Χρηματική Αξία):** Η κατανομή ακολουθεί παρόμοια γεωμετρία (σχήμα L) με τη Συχνότητα. Το μεγαλύτερο μέρος των εσόδων προέρχεται

από ένα εξαιρετικά μικρό ποσοστό πελατών, επιβεβαιώνοντας και στατιστικά την ισχύ της **Αρχής του Pareto (Κανόνας 80/20)** στο συγκεκριμένο σύνολο δεδομένων: το 80% του τζίρου παράγεται από το 20% των κορυφαίων πελατών ("VIPs").



Εικόνα 5.3.1. RFM Αρχείο Distributions

Η **στατιστική** αυτή **συμπεριφορά** (Heavy-tailed Distributions) αναδεικνύει την πολυπλοκότητα του πελατολογίου, καθώς η **υψηλή διασπορά** τιμών καθιστά δυσδιάκριτα τα όρια μεταξύ των **φυσικών ομάδων**. Η ύπαρξη τόσο έντονων **ακραίων τιμών** (**outliers**) στις μεταβλητές **Frequency** και **Monetary**, σε συνδυασμό με τη μη γραμμική φύση των δεδομένων, καθιστά επιβεβλημένη την εφαρμογή **Λογαριθμικού Μετασχηματισμού** (Log

Transformation). Χωρίς αυτόν, οι **αλγόριθμοι** που βασίζονται στην **Ευκλείδεια Απόσταση** θα αδυνατούσαν να εντοπίσουν ουσιαστικά μοτίβα, οδηγώντας σε συστάδες άνισης πυκνότητας και **χαμηλής ερμηνευτικής αξίας**.

Ερμηνεία Εικόνας 5.3.1.: Στο **Ιστογράμματα Κατανομής** των **Μεταβλητών Recency**, **Frequency** και **Monetary**, παρατηρείται έντονη **Θετική Ασυμμετρία** (Skewness) στις **Μεταβλητές F** και **M,** γεγονός που καθιστά αναγκαία τη **Λογαριθμική Μετατροπή** (Log Transformation) πριν τη **Συσταδοποίηση**.

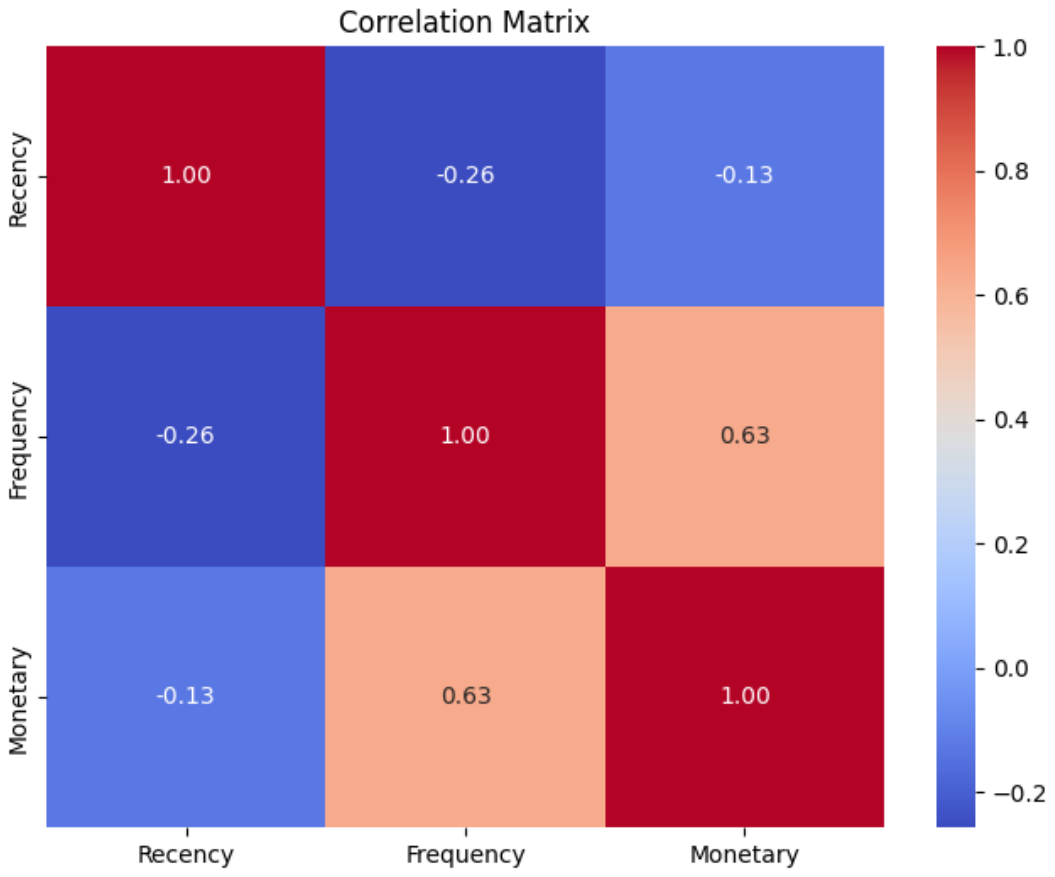
5.3.2. Έλεγχος Συσχετίσεων (Correlation Matrix & Heatmaps)

Στη συνέχεια, εξετάστηκε η **Γραμμική Συσχέτιση** μεταξύ των **Μεταβλητών** χρησιμοποιώντας τον **Συντελεστή Pearson**. Στόχος είναι να διαπιστωθεί αν υπάρχει Πολυσυγγραμμικότητα (Multicollinearity) ή ισχυρή **Εξάρτηση** μεταξύ των **Χαρακτηριστικών**.

Ερμηνεία Εικόνας 5.3.2.: Παρατηρούμε το **Πίνακα Συσχετίσεων** (Correlation Heatmap) των **Μεταβλητών RFM**. Οι τιμές κυμαίνονται από **-1** έως **1**, υποδεικνύοντας την **Ισχύ** και την **Κατεύθυνση** της **Συσχέτισης**.

Τα αποτελέσματα απεικονίζονται στον **Θερμικό Χάρτη** (**Heatmap**) της **Εικόνας 5.2.2.**:

- **Frequency vs Monetary:** Παρατηρείται **Ισχυρή Θετική Συσχέτιση (r = 0.63)**. Αυτό είναι λογικό και αναμενόμενο, καθώς οι Πελάτες που αγοράζουν συχνότερα τείνουν αθροιστικά να ξοδεύουν και περισσότερα χρήματα.
- **Recency vs Frequency:** Παρατηρείται **Ασθενής Αρνητική Συσχέτιση (r = -0.26)**. Αυτό υποδηλώνει ότι οι πιο Ενεργοί Πελάτες (χαμηλό Recency) τείνουν να έχουν ελαφρώς υψηλότερη Συχνότητα, αν και η σχέση δεν είναι απόλυτη.
- **Recency vs Monetary:** Παρατηρείται πολύ **Ασθενής Αρνητική Συσχέτιση (r = -0.13)**, δείχνοντας ότι το πόσο πρόσφατα ψώνισε κάποιος δεν προδικάζει απαραίτητα το ύψος της δαπάνης του.



Εικόνα 5.3.2. RFM Αρχείο Correlation

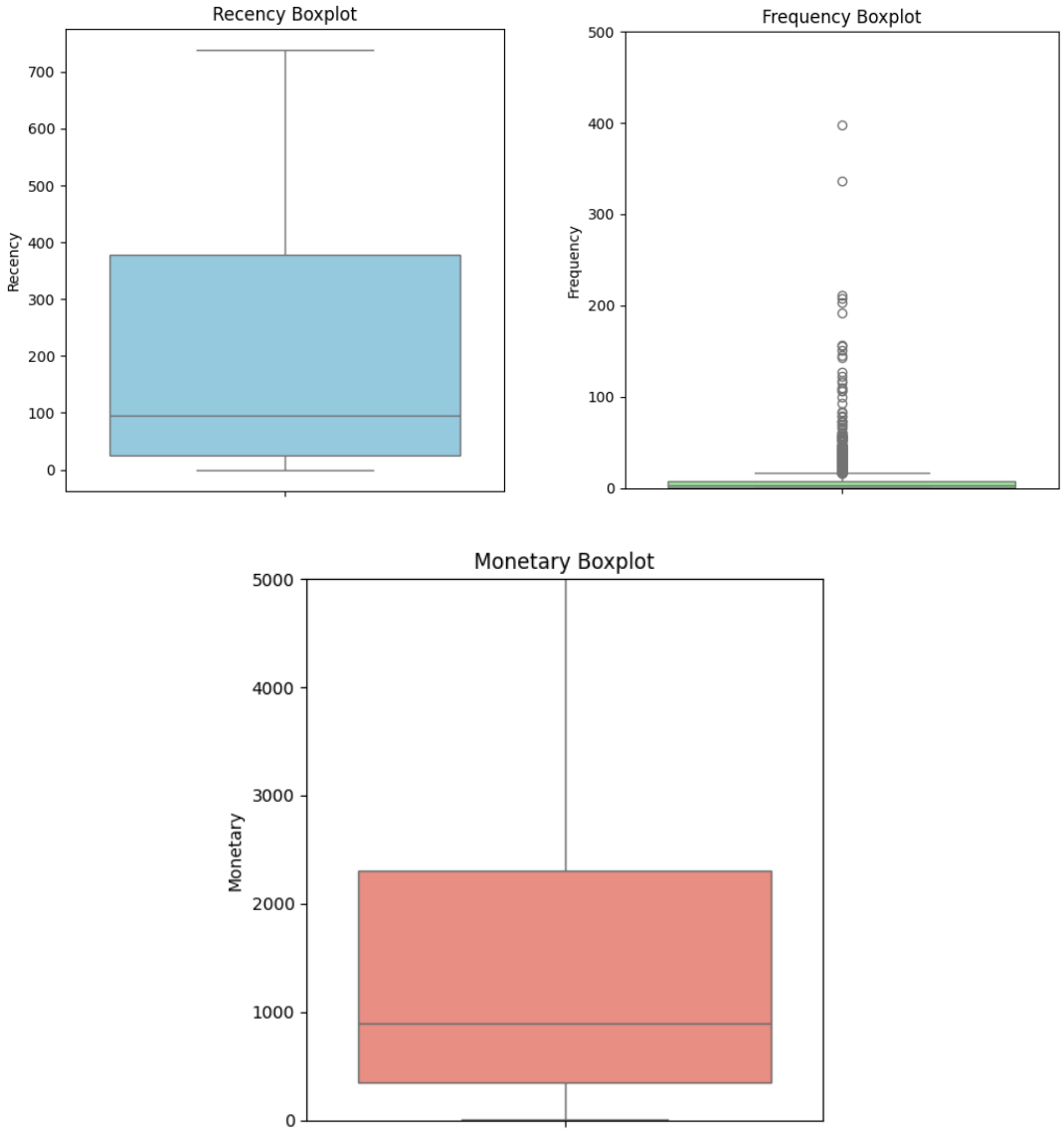
5.3.3. Εντοπισμός και Διαχείριση Ακραίων Τιμών (Outliers)

Οι **Αλγόριθμοι Συσταδοποίησης**, και ειδικότερα ο **K-Means**, είναι ιδιαίτερα ευαίσθητοι σε **Ακραίες Τιμές** (Outliers), καθώς αυτές μπορούν να μετατοπίσουν τα **Κέντρα** των **Συστάδων** (Centroids) και να αλλοιώσουν τα αποτελέσματα.

Για τον εντοπισμό τους χρησιμοποιήθηκε η Μέθοδος των **Θηκογραμμάτων (Boxplots)**. Όπως φαίνεται στην **Εικόνα 5.3.3**:

- Η μεταβλητή **Recency** παρουσιάζει συμμετρική κατανομή εντός του κουτιού, χωρίς εμφανείς ακραίες τιμές.

- Η μεταβλητή **Frequency** εμφανίζει ξεκάθαρα **μεγάλο πλήθος ακραίων τιμών** (μαύρες κουκκίδες) στο άνω άκρο. Αυτό επιβεβαιώνει την ύπαρξη μιας ομάδας πελατών με ασυνήθιστα υψηλή συχνότητα αγορών.
- Η μεταβλητή **Monetary** χαρακτηρίζεται από **εξαιρετικά μεγάλη διασπορά** (πολύ υψηλό εύρος κουτιού). Αν και δεν διακρίνονται μεμονωμένες κουκκίδες στο συγκεκριμένο εύρος απεικόνισης, το μεγάλο εύρος τιμών υποδηλώνει έντονη μεταβλητότητα στη δαπάνη των πελατών.



Εικόνα 5.3.3. RFM Αρχείο Outliers

Στο σημείο αυτό, είναι κρίσιμο να επισημανθεί ότι, στο πλαίσιο της **Ανάλυσης Πελατών Λιανικής**, οι εν λόγω **Ακραίες Τιμές** (Outliers) δεν αποτελούν «θόρυβο» (noise) ή σφάλματα μετρήσεων, αλλά αντιπροσωπεύουν τους λεγόμενους «Πελάτες Υψηλής Αξίας» (High-Value Customers / Whales). Συνεπώς, η συνήθης πρακτική της αφαίρεσης των **Ακραίων Τιμών** (Trimming/Winsorizing) θα ήταν λανθασμένη, καθώς θα οδηγούσε σε απώλεια της πιο πολύτιμης επιχειρηματικής πληροφορίας.

Ωστόσο, η διατήρησή τους στην αρχική τους κλίμακα δημιουργεί πρόβλημα **Γεωμετρικής Παραμόρφωσης** στον χώρο των **Χαρακτηριστικών**, καθώς ο **K-Means**, βασιζόμενος στην **Ευκλείδεια Απόσταση**, τείνει να δημιουργεί μονομελείς συστάδες για να «ικανοποιήσει» αυτές τις απομακρυσμένες παρατηρήσεις.

Η αντίφαση αυτή καθιστά επιβεβλημένη την εφαρμογή **μη γραμμικών μετασχηματισμών** για τη συμπίεση του εύρους τιμών, χωρίς όμως να αλλοιώνεται η ιεραρχική κατάταξη των πελατών.

Ερμηνεία Εικόνας 5.3.3.: Τα **Boxplots** επισημαίνουν τον εντοπισμό Ακραίων Τιμών. Παρατηρείται έντονη παρουσία **Outliers** στη μεταβλητή **Frequency** (οι πολλαπλές μαύρες κουκίδες), ενώ η μεταβλητή **Monetary** παρουσιάζει μεγάλη διασπορά τιμών.

Στρατηγική Διαχείρισης: Στο πλαίσιο της **Ανάλυσης Πελατών Λιανικής**, αυτές οι συμπεριφορές (υψηλή συχνότητα και μεγάλη διασπορά δαπάνης) **δεν αποτελούν σφάλμα**. Αντιπροσωπεύουν τους **VIP Πελάτες** (Whales). Η διαγραφή τους θα οδηγούσε σε απώλεια της πιο πολύτιμης **Επιχειρηματικής Πληροφορίας**.

Για τον λόγο αυτό:

1. Εφαρμόστηκε **Λογαριθμική Μετατροπή (Log Transformation)** για την εξομάλυνση των Κατανομών.
2. Χρησιμοποιήθηκε **Κανονικοποίηση (StandardScaler)** ώστε όλες οι Μεταβλητές να έχουν Μέση Τιμή 0 και Τυπική Απόκλιση 1.

5.4. Μηχανική Χαρακτηριστικών (Feature Engineering)

Πριν την εφαρμογή των αλγορίθμων Μηχανικής Μάθησης, είναι απαραίτητη η ανασκόπηση του τελικού δείγματος που προέκυψε από την Εξερευνητική Ανάλυση Δεδομένων (EDA).

Το αρχικό σύνολο δεδομένων (Online Retail II) περιείχε συνολικά **1.067.372 εγγραφές**, καλύπτοντας τη διετία **2009-2011**. Μετά τη διαδικασία καθαρισμού (Data Cleaning) που περιελάμβανε την αφαίρεση ακυρώσεων, επιστροφών και εγγραφών χωρίς αναγνωριστικό πελάτη, το τελικό δείγμα διαμορφώθηκε σε επίπεδο **5.879 μοναδικών πελατών**.

Η **Στατιστική Ανάλυση** των **Μεταβλητών RFM** (Recency, Frequency, Monetary) που προηγήθηκε, αποκάλυψε σημαντική **ασυμμετρία (skewness)** στην κατανομή, οδηγώντας σε δύο κρίσιμα συμπεράσματα:

1. **Monetary:** Η πλειονότητα των πελατών δαπανά μικρά ποσά, ενώ υπάρχει ένας περιορισμένος αριθμός πελατών ("**VIPs**") με εξαιρετικά υψηλή δαπάνη. Αυτό επιβεβαιώνει την **Αρχή του Pareto (80/20 Rule)** στο ηλεκτρονικό εμπόριο **[Brynjolfsson et al., 2011]**.
2. **Recency:** Παρατηρήθηκε ότι ένα μεγάλο ποσοστό πελατών δεν έχει αγοράσει προϊόντα για μεγάλο χρονικό διάστημα (άνω των 3 μηνών), υποδεικνύοντας την άμεση ανάγκη σχεδιασμού στρατηγικών **επαναδραστηριοποίησης (re-engagement)**.

Με βάση τα παραπάνω δεδομένα, προχωρούμε στην προετοιμασία των μεταβλητών για την εκπαίδευση του αλγορίθμου.

5.4.1. Λογαριθμικός Μετασχηματισμός (Transformation) και Κανονικοποίηση Δεδομένων (Scaling)

Όπως διαπιστώθηκε στην **Ενότητα 5.3** (EDA), οι μεταβλητές του μοντέλου **RFM** παρουσιάζουν δύο σημαντικά προβλήματα που καθιστούν την άμεση εφαρμογή του **K-Means** προβληματική:

1. **Ασυμμετρία Κατανομών (Skewness):** Οι μεταβλητές **Frequency** και **Monetary** έχουν έντονη **Θετική Ασυμμετρία** (L-shape).
2. **Διαφορά Κλίμακας (Scale):** Το εύρος τιμών του **Monetary** (π.χ. 0 έως 5.000+) είναι δυσανάλογα μεγαλύτερο από αυτό του **Frequency** (π.χ. 1 έως 50) και του **Recency**.

Εδώ, να θυμίσουμε ότι επειδή ο αλγόριθμος **K-Means** βασίζεται στην **Ευκλείδεια Απόσταση**, αν χρησιμοποιούσαμε τα δεδομένα στην αρχική τους μορφή, η μεταβλητή **Monetary** θα κυριαρχούσε πλήρως στον υπολογισμό των αποστάσεων, ακυρώνοντας τη σημασία της **Frequency** και του **Recency**. Για την αντιμετώπιση των παραπάνω, εφαρμόστηκε μια διαδικασία δύο (2) βημάτων για τους λόγους που έχουμε ήδη αναφερθεί λεπτομερώς στην **Ενότητα 4.7.**:

Βήμα 1ο: Λογαριθμικός Μετασχηματισμός (Log Transformation)

Για την εξομάλυνση της ασυμμετρίας και τη μείωση της επίδρασης των ακραίων τιμών (outliers), εφαρμόστηκε φυσικός λογάριθμος στις μεταβλητές. Ο μετασχηματισμός αυτός "συμπιέζει" τις μεγάλες τιμές και "απλώνει" τις μικρές, φέρνοντας την κατανομή πιο κοντά στην κανονική (Gaussian distribution).

Μαθηματική διατύπωση:

$$x' = \log(x) \qquad \text{(Εξίσωση 5.1)}$$

Βήμα 2ο: Κανονικοποίηση (Standardization / Scaling)

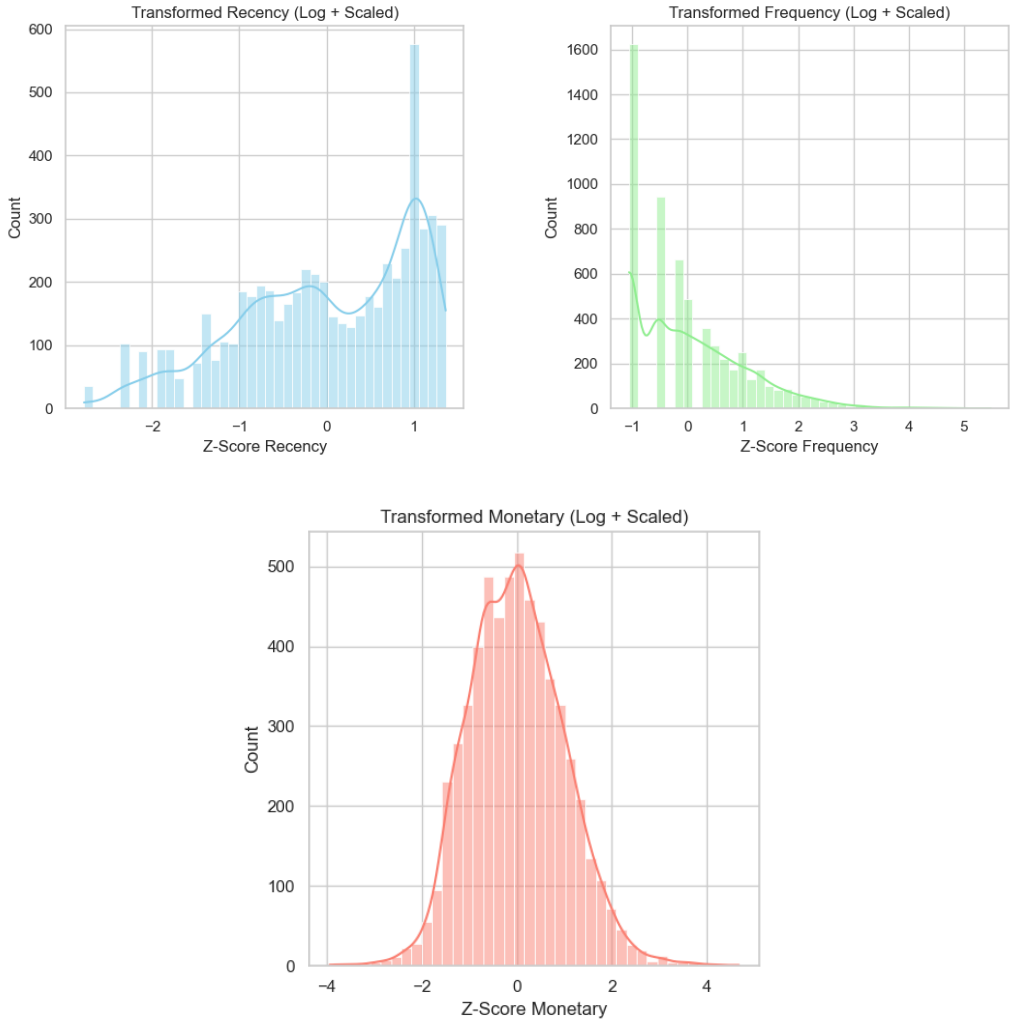
Στη συνέχεια, εφαρμόστηκε η μέθοδος **StandardScaler** (Z-score normalization). Η μέθοδος αυτή μετατρέπει τα δεδομένα ώστε κάθε **μεταβλητή** να έχει:

- Μέση Τιμή (μ) = 0
- Τυπική Απόκλιση (σ) = 1

Μαθηματική διατύπωση:

$$z = \frac{x - \mu}{\sigma} \qquad \text{(Εξίσωση 5.2)}$$

Το αποτέλεσμα της διαδικασίας απεικονίζεται στην **Εικόνα 5.4.1**. Παρατηρούμε ότι μετά τον μετασχηματισμό, οι κατανομές έχουν προσεγγίσει σε μεγάλο βαθμό τη συμμετρική μορφή της **Κανονικής Κατανομής (Gaussian Distribution / "καμπάνα")**. Παράλληλα, μέσω της τυποποίησης (Standardization), όλες οι μεταβλητές έχουν μεταφερθεί σε κοινό άξονα αναφοράς (Z-scores) με μέση τιμή το μηδέν και τυπική απόκλιση τη μονάδα. Η διαδικασία αυτή εξασφαλίζει τη **διαστατική ομοιογένεια** των δεδομένων και την ισότιμη συμμετοχή κάθε μεταβλητής στον υπολογισμό των **ευκλείδειων αποστάσεων**, αποτρέποντας την κυριαρχία των χαρακτηριστικών με μεγάλες **αριθμητικές τιμές** (π.χ. Monetary) έναντι εκείνων με μικρότερες (π.χ. Frequency).



Εικόνα 5.4.1. RFM Αρχείο Distributions (after Transformation)

Σχολιασμός και Σύγκριση με την Αρχική Κατανομή (Εικόνα 5.3.1 vs 5.4.1)

Η αντιπαραβολή της **Εικόνας 5.4.1** (Μετασχηματισμένα Δεδομένα) με την **Εικόνα 5.3.1** (Αρχικά Δεδομένα) αναδεικνύει τη δραστική βελτίωση της ποιότητας των δεδομένων σε δύο κρίσιμα επίπεδα:

1. **Ως προς τη Γεωμετρία της Κατανομής (Distribution Shape):** Ενώ στην αρχική εικόνα (5.3.1) οι μεταβλητές Frequency και Monetary παρουσίαζαν έντονη **θετική ασυμμετρία (Right-skewness)** τύπου "L-shape", συγκεντρώνοντας τη μάζα των δεδομένων σε χαμηλές τιμές και δημιουργώντας μακριές "ουρές" (long tails), στην τελική εικόνα (5.4.1) η κατανομή έχει «ξεδιπλωθεί». Ο λογαριθμικός μετασχηματισμός συμπίεσε το εύρος των ακραίων τιμών και εξομάλυνε την ασυμμετρία, επιτρέποντας την ανάδειξη λανθανόντων μοτίβων που προηγουμένως κρύβονταν από τη στρέβλωση της κλίμακας. Αυτή η «κανονικοποίηση» είναι ζωτικής σημασίας για τον K-Means, ο οποίος βελτιστοποιείται όταν οι συστάδες έχουν σφαιρική μορφή.
2. **Ως προς την Τάξη Μεγέθους (Scale Magnitude):** Στην αρχική μορφή, υπήρχε χαώδης διαφορά κλιμάκων (π.χ. το Recency μετρούσε ημέρες 0-370, ενώ το Monetary χρηματικές μονάδες 0-50.000+). Χωρίς παρέμβαση, η μεταβλητή Monetary θα μονοπωλούσε μαθηματικά τον καθορισμό των συστάδων. Στην Εικόνα 5.4.1, παρατηρούμε ότι όλες οι μεταβλητές κυμαίνονται πλέον σε ένα κοινό, συγκρίσιμο εύρος (κυρίως μεταξύ -3 και +3 τυπικών αποκλίσεων). Αυτή η **κανονικοποίηση κλίμακας** επιτρέπει στον αλγόριθμο να αξιολογεί τη συμπεριφορά του πελάτη βάσει της *σχετικής* του θέσης στην κατανομή και όχι βάσει της *απόλυτης* αριθμητικής του τιμής.

Συμπερασματικά: Η **Εικόνα 5.4.1** επιβεβαιώνει ότι τα δεδομένα είναι πλέον **ομοιογενή** και **συμμετρικά**, πληρώντας τις βασικές στατιστικές προϋποθέσεις για τη βέλτιστη απόδοση του αλγορίθμου **K-Means**.

5.5. Βελτιστοποίηση Παραμέτρων (Hyperparameter Tuning)

Έχοντας ολοκληρώσει τον **μετασχηματισμό** και την **κανονικοποίηση** των **δεδομένων** (Ενότητα 5.4), το επόμενο κρίσιμο στάδιο στην αλυσίδα της ανάλυσής μας είναι ο προσδιορισμός των **βέλτιστων** Υπερ-παραμέτρων (Hyperparameters) για τους **αλγορίθμους συσταδοποίησης**.

Σε αντίθεση με την **Επιβλεπόμενη Μάθηση** (Supervised Learning), όπου το μοντέλο εκπαιδεύεται βάσει γνωστών ετικετών, στην **Μη Επιβλεπόμενη Μάθηση** (Unsupervised Learning) η ποιότητα του αποτελέσματος εξαρτάται άμεσα από την αρχική **παραμετροποίηση** του **Αλγορίθμου**. Μια αυθαίρετη επιλογή παραμέτρων θα μπορούσε να οδηγήσει είτε σε υπο-προσαρμογή (underfitting), δημιουργώντας υπερβολικά γενικευμένες ομάδες χωρίς **Επιχειρηματική Αξία**, είτε σε υπερ-προσαρμογή (overfitting), κατακερματίζοντας τα **Δεδομένα** σε **Θόρυβο**.

Συγκεκριμένα, η παρούσα μελέτη εστιάζει στη βελτιστοποίηση δύο διαφορετικών **Οικογενειών Αλγορίθμων**, οι οποίες απαιτούν διαφορετική προσέγγιση:

1. **Για τον αλγόριθμο K-Means:** Απαιτείται ο εκ των προτέρων καθορισμός του **πλήθους των συστάδων** (**k**), μια παράμετρος που δεν είναι γνωστή a priori.
2. **Για τον αλγόριθμο DBSCAN:** Απαιτείται ο προσδιορισμός της **ακτίνας γειτονίας** (**ε**) και του **ελάχιστου αριθμού σημείων** (**MinPts**) που ορίζουν μια περιοχή ως «**πυκνή**».

Για την αντικειμενική επιλογή των τιμών αυτών και την αποφυγή υποκειμενικών εκτιμήσεων, αξιοποιήσαμε τα **αυτοματοποιημένα εργαλεία βελτιστοποίησης** του λογισμικού **RFM Master Tool v4.5.0**. Η διαδικασία βασίστηκε σε ποσοτικά κριτήρια και μαθηματικές μετρικές (Elbow Method, Silhouette Analysis, k-Distance Graph), όπως αναλύονται στις υποενότητες που ακολουθούν.

5.5.1. Αυτοματοποιημένος Προσδιορισμός Βέλτιστου Αριθμού Συστάδων (Automated k-Selection)

Όπως αναφέρθηκε στη προηγούμενη ενότητα, για την αποφυγή της υποκειμενικότητας που χαρακτηρίζει την οπτική ερμηνεία των διαγραμμάτων, χρησιμοποιήθηκε το υποσύστημα τεχνητής νοημοσύνης **"Auto-K Intelligence Tool"** του λογισμικού μας (**RFM Master Tool v4.5.0**).

Συγκεκριμένα, το υποσύστημα αυτό λειτουργεί ως ένας μηχανισμός **Επαναληπτικής Βελτιστοποίησης (Iterative Optimization Engine)**. Αντί να βασίζεται σε μεμονωμένες δοκιμές, ο Αλγόριθμος εκτελεί μια **αυτοματοποιημένη διαδικασία σάρωσης** (parameter sweeping) στο εύρος **k** ∈ **[2, 10]**, εφαρμόζοντας για κάθε τιμή του **k** τον αλγόριθμο **K-Means** στα **κανονικοποιημένα δεδομένα**.

Κατά τη διάρκεια κάθε επανάληψης (iteration), το λογισμικό δεν αρκείται στον υπολογισμό ενός μόνο δείκτη, αλλά εκτελεί έναν **Πολυκριτηριακό Υπολογισμό Ποιότητας (Multi-Metric Quality Assessment)**, καταγράφοντας ταυτόχρονα τέσσερις (4) θεμελιώδεις μετρικές συσταδοποίησης:

1. **Inertia (Αδράνεια):** Για την εφαρμογή της μεθόδου Elbow.
2. **Silhouette Score:** Για την αξιολόγηση της συνοχής και του διαχωρισμού των συστάδων.
3. **Calinski-Harabasz Index:** Για την εκτίμηση της πυκνότητας των ομάδων.
4. **Davies-Bouldin Index:** Για την αξιολόγηση της ομοιότητας μεταξύ των συστάδων.

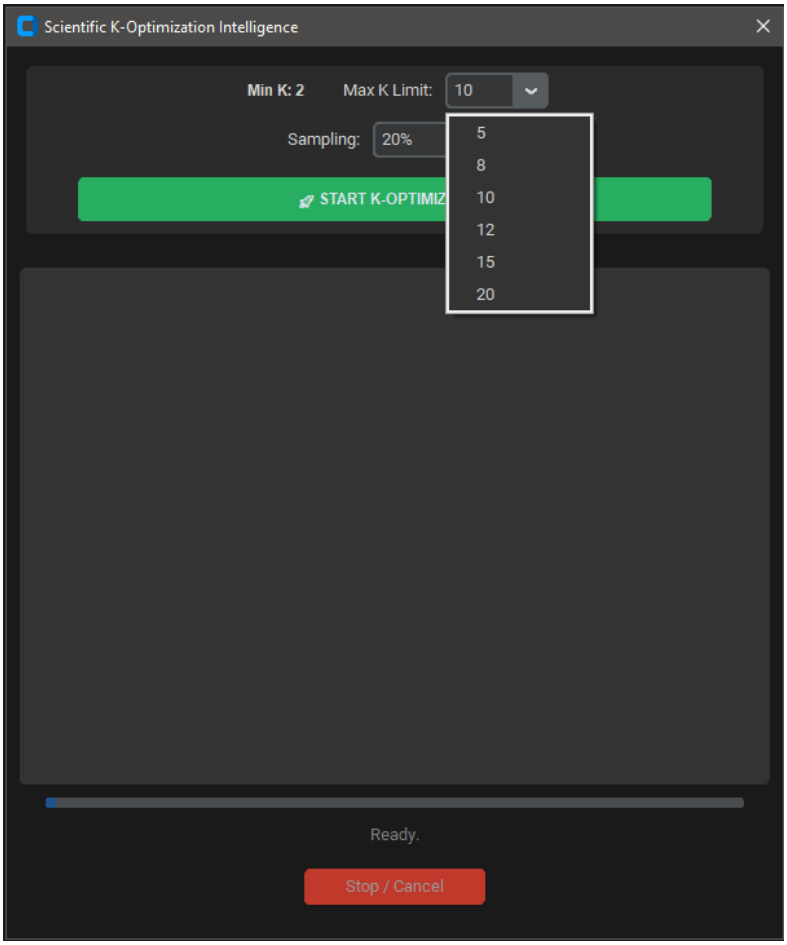
Τα αποτελέσματα αυτών των υπολογισμών συγκεντρώνονται σε έναν πίνακα καταγραφής (Log Table) και οπτικοποιούνται σε πραγματικό χρόνο, επιτρέποντας τη συνδυαστική αξιολόγηση της μαθηματικής βελτιστοποίησης και της επιχειρηματικής λογικής.

Στις παραγράφους που ακολουθούν, αναλύονται τα αποτελέσματα που παρήγαγε το εργαλείο για κάθε μία από τις παραπάνω μετρικές, τεκμηριώνοντας την τελική επιλογή του πλήθους των συστάδων.

Παραμετροποίηση Αλγορίθμου και Καθορισμός Χώρου Αναζήτησης:

Όπως αποτυπώνεται στο **Γραφικό Περιβάλλον** της **Εικόνας 5.5.1.a**, η **διαδικασία βελτιστοποίησης** παραμετροποιήθηκε με γνώμονα την εξισορρόπηση μεταξύ της υπολογιστικής απόδοσης και της επιχειρηματικής αξιοποίησης των αποτελεσμάτων. Συγκεκριμένα:

- **Εύρος Αναζήτησης Συστάδων (Cluster Search Space):** Ορίστηκε το διάστημα **k** ∈ **[2, 10]**.
- ***Αιτιολόγηση:*** Η επιλογή του άνω ορίου (k=10) δεν είναι τυχαία. Στη διεθνή βιβλιογραφία του **Customer Segmentation**, τμήματα που υπερβαίνουν τα 10 καθίστανται πρακτικά μη **διαχειρίσιμα** (**unmanageable**) για τα τμήματα **Marketing**, καθώς απαιτούν υπερβολικά εξειδικευμένους πόρους για τη δημιουργία διακριτών καμπανιών. Αντίστοιχα, το κάτω όριο (**k=2**) είναι το ελάχιστο δυνατό για να υπάρξει οποιαδήποτε μορφή διαφοροποίησης. Συνεπώς, το εύρος αυτό καλύπτει το **βέλτιστο φάσμα** για την εύρεση μιας λύσης που είναι ταυτόχρονα στατιστικά σημαντική και διοικητικά εφαρμόσιμη.
- **Στρατηγική Δειγματοληψίας (Random Sampling Optimization):** Εφαρμόστηκε τυχαία δειγματοληψία στο **20%** του συνολικού όγκου των **5.879 μοναδικών πελατών**, το οποίο μεταφράζεται σε ένα υπο-σύνολο **1.175 εγγραφών** (5.879 x 0,20 ≈ **1.175**), όπως επιβεβαιώνεται από το Log της **Εικόνας 5.5.1.b**).
- ***Αιτιολόγηση:*** Η τεχνική αυτή κρίθηκε απαραίτητη για την επιτάχυνση της επαναληπτικής διαδικασίας (iterative process) του αλγορίθμου. Στατιστικά, ένα **δείγμα > 1.000 εγγραφών** σε **κανονικοποιημένα δεδομένα** θεωρείται επαρκές για να "συλλάβει" με ακρίβεια τη γεωμετρία και την κατανομή του πληθυσμού, σύμφωνα με τον Νόμο των Μεγάλων Αριθμών (Law of Large Numbers). Έτσι, επιτυγχάνουμε δραστική μείωση του υπολογιστικού κόστους (computational cost) χωρίς να θυσιάζουμε τη στατιστική εγκυρότητα της βελτιστοποίησης.



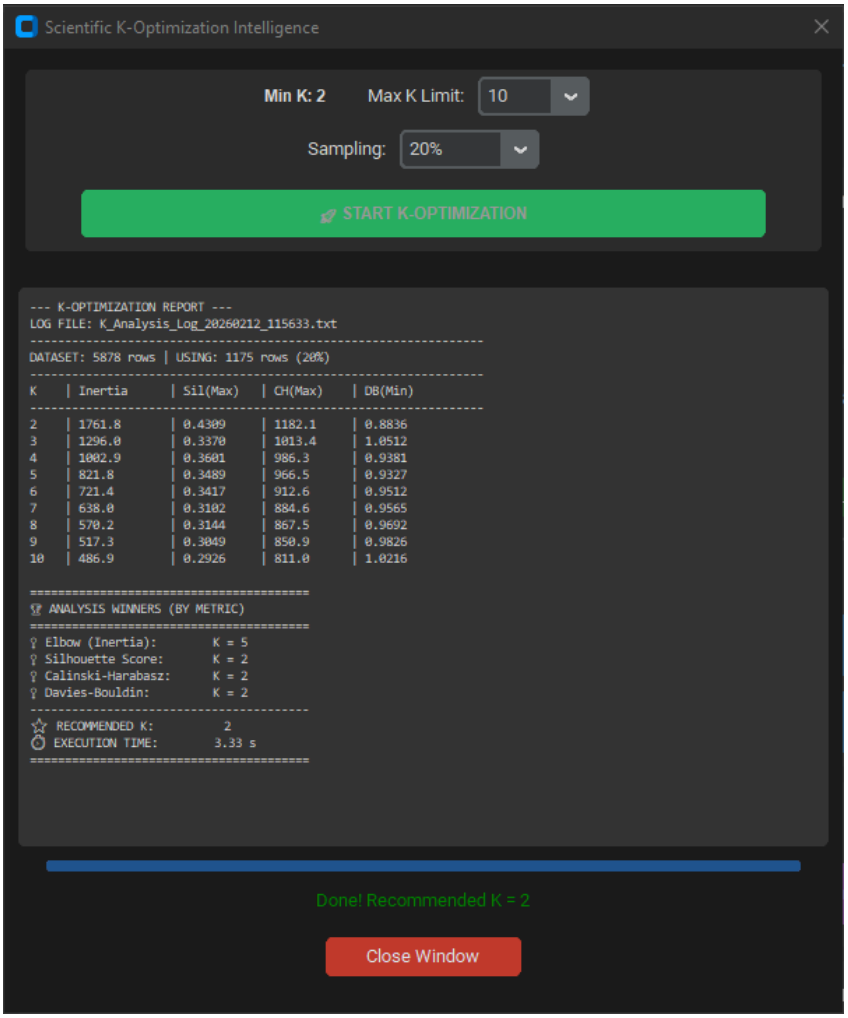
Εικόνα 5.5.1.a. Ρυθμίσεις (Min/Max K, Sampling 20%) και Dropdown Μενού ως παράμετροι για τον υπο-κώδικα K-Optimization Intelligence

Ανάλυση Αποτελεσμάτων (Multi-Metric Evaluation):

Τα αποτελέσματα της εκτέλεσης παρουσιάζονται αναλυτικά στον πίνακα της **Εικόνας 5.5.1.b**. Ο αλγόριθμος υπολόγισε τέσσερις δείκτες για κάθε **k**, οδηγώντας στα εξής συμπεράσματα:

1. **Strict Mathematical Optimum (k=2):** Οι δείκτες **Silhouette Score** (0.4309), **Calinski-Harabasz** και **Davies-Bouldin** υπέδειξαν ως μαθηματικά βέλτιστη λύση τη δημιουργία δύο (2) συστάδων. Ωστόσο, σε επίπεδο **επιχειρηματικής στρατηγικής** (Marketing Strategy), ο διαχωρισμός των πελατών σε μόλις δύο ομάδες θεωρείται υπερβολικά απλοϊκός, καθότι μια τέτοια δυαδική ταξινόμηση θα απέκρυπτε κρίσιμες ενδιάμεσες κατηγορίες, περιορίζοντας δραματικά τη δυνατότητα εφαρμογής στοχευμένων ενεργειών **Marketing**. Στην πράξη, μια προσέγγιση τύπου «Active vs Inactive» θα

αγνοούσε παντελώς τη δυναμική της **πελατειακής βάσης**, όπως τους «Ανερχόμενους Πελάτες» που χρήζουν ενθάρρυνσης ή τους «Πελάτες σε Κίνδυνο» που απαιτούν άμεση ανάκτηση. Συνεπώς, η μαθηματική υπεροχή του **k=2** θυσιάζεται υπέρ της **επιχειρησιακής χρησιμότητας** (Business Utility), η οποία απαιτεί **υψηλότερη ευκρίνεια** για τη χάραξη πολυδιάστατης **εμπορικής πολιτικής**.



Εικόνα 5.5.1.b. Πίνακας με τα αποτελέσματα (Log) και την Τελική Αξιολόγηση

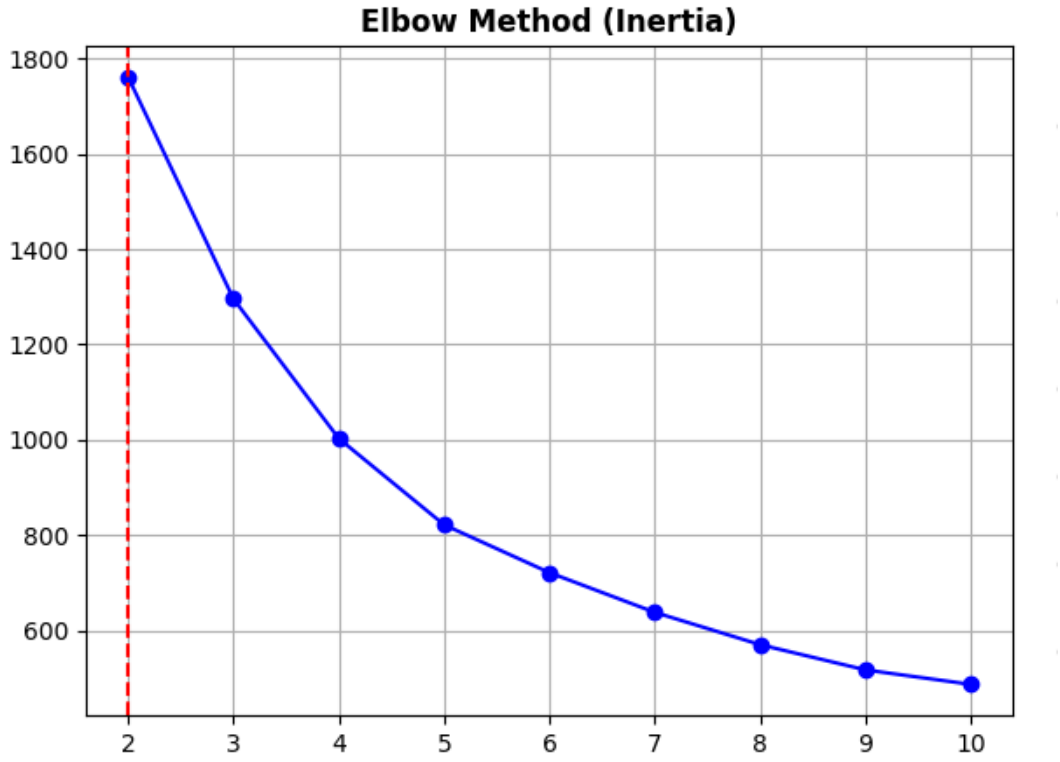
2. **Geometric Elbow Determination (k=5):** Ο **δείκτης Inertia** (Elbow Method), όπως υπολογίστηκε από τον αλγόριθμο **Γεωμετρικής Βελτιστοποίησης** του εργαλείου μας, εντόπισε το **Σημείο Καμπής** στο **k=5**. Παρατηρούμε ότι η τιμή της **Αδράνειας** μειώνεται σημαντικά από το **k=4** (1002.9) στο **k=5** (821.8), ενώ μετά το σημείο αυτό ο **ρυθμός βελτίωσης** επιβραδύνεται αισθητά (diminishing returns). Η συγκεκριμένη μείωση της

Αδράνειας (κατά ~18%) υποδηλώνει ότι η προσθήκη της πέμπτης συστάδας συνεισφέρει ουσιαστικά στην **ομοιογένεια** των ομάδων χωρίς να αποτελεί περιττή **πολυπλοκότητα**. Ο αλγόριθμος επιβεβαίωσε ότι σε αυτό το σημείο επιτυγχάνεται η **βέλτιστη ισορροπία** μεταξύ της συνοχής των συστάδων (intra-cluster compactness) και του αριθμού των τμημάτων, καθιστώντας το **k=5** την πιο αποδοτική λύση δομικά.

3. **Silhouette Stability (k=5):** Παρόλο που το **k=4** εμφανίζει ελαφρώς υψηλότερο **Silhouette Score** (0.3601) έναντι του **k=5** (0.3489), η διαφορά κρίνεται αμελητέα (μόλις 0.011). Αυτό αποδεικνύει ότι η μετάβαση σε **πέντε** (**5**) **Ομάδες** διατηρεί την εσωτερική συνοχή των **συστάδων** σε υψηλά επίπεδα, χωρίς να **κατακερματίζει αυθαίρετα** τα **δεδομένα**. Το γεγονός ότι ο δείκτης παραμένει σταθερός (αντί να κατακρημνιστεί, όπως συμβαίνει συχνά στην υπερ-τμηματοποίηση) αποτελεί ισχυρή ένδειξη ότι η **πέμπτη ομάδα** είναι φυσική και **διακριτή οντότητα**. Αυτή η **σταθερότητα** μας επιτρέπει να εκμεταλλευτούμε την επιπλέον **κατηγοριοποίηση** χωρίς να κινδυνεύουμε να δημιουργήσουμε τεχνητές ομάδες θορύβου (noise clusters) που θα αλλοίωναν την **αξιοπιστία** του **μοντέλου**.

Πέραν της **αριθμητικής απεικόνισης** των **αποτελεσμάτων** στον **Πίνακα Καταγραφής** (Log), κρίθηκε απαραίτητη η **εποπτική επιβεβαίωση** των **μετρικών** μέσω **γραφικών παραστάσεων**. Για τον σκοπό αυτό, το λογισμικό μας σχεδιάστηκε ώστε να παράγει και να προβάλλει σε πραγματικό χρόνο ένα **Ενοποιημένο Ταμπλό Οπτικοποίησης (Unified Visualization Dashboard)**.

Στα Διαγράμματα που ακολουθούν, αποτυπώνεται η συμπεριφορά των τεσσάρων δεικτών αξιολόγησης για κάθε τιμή του **k**. Η διαδικασία αυτή αποτελεί την πρακτική εφαρμογή του θεωρητικού πλαισίου που αναλύθηκε διεξοδικά στις **Ενότητες 2.13 (Αλγόριθμοι ML για Εφαρμογή σε CPDS)** και **4.7 (Μετασχηματισμός και Κανονικοποίηση Δεδομένων)**. Μέσω αυτής της συνολικής απεικόνισης, καθίσταται εφικτή η διασταύρωση της μαθηματικής βελτιστοποίησης με την οπτική διαίσθηση, επιτρέποντας στον αναλυτή να αξιολογήσει ταυτόχρονα τη **συνοχή** (Elbow), τον **διαχωρισμό** (Silhouette, Davies-Bouldin) και την **πυκνότητα** (Calinski-Harabasz) των προτεινόμενων **συστάδων**, θωρακίζοντας την τελική επιλογή του **k=5**.

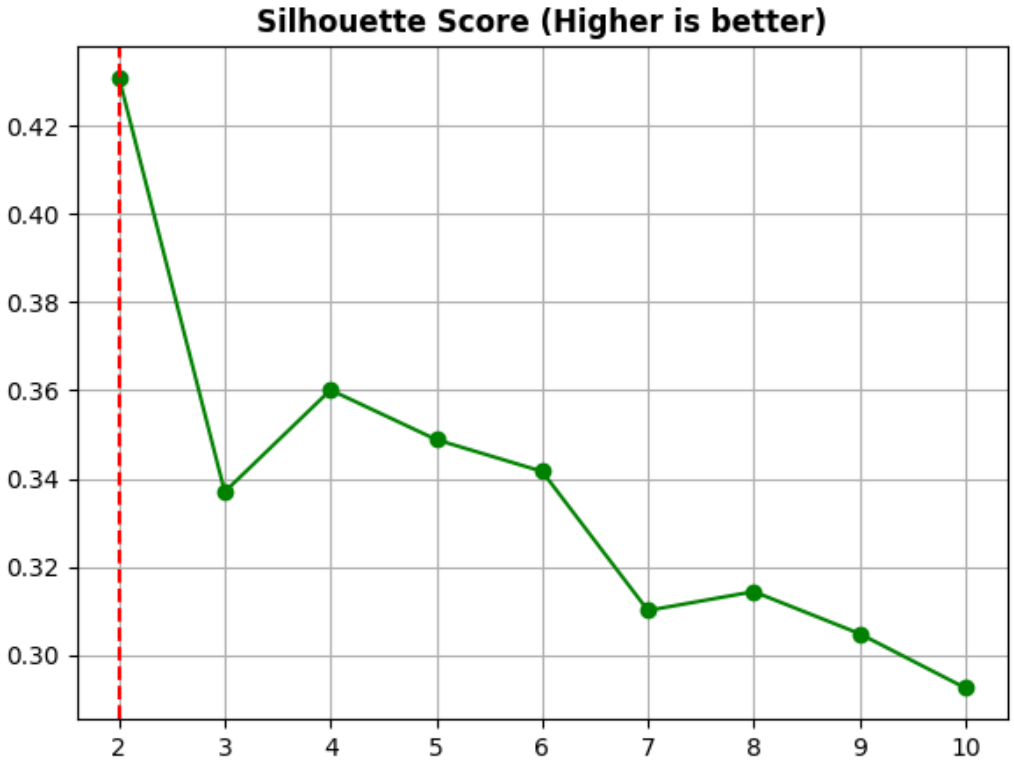


Εικόνα 5.5.1.c. Διάγραμμα Elbow Method

Ανάλυση Αδράνειας (Elbow Method)

Η πρώτη **Εικόνα** (**5.5.1.c**) παρουσιάζει την καμπύλη της **Αδράνειας** (Inertia), η οποία μετρά το άθροισμα των τετραγωνικών αποστάσεων των δειγμάτων από το πλησιέστερο κέντρο συστάδας.

- **Παρατήρηση:** Η καμπύλη εμφανίζει μια συνεχή πτωτική πορεία καθώς αυξάνεται το **k**. Το σημείο όπου ο ρυθμός μείωσης αρχίζει να επιβραδύνεται ("αγκώνας") δεν είναι απόλυτα ευκρινές με γυμνό μάτι, ωστόσο η **καμπύλη** αρχίζει να ομαλοποιείται αισθητά μετά το **k=4**. Αυτό υποδεικνύει ότι η προσθήκη επιπλέον συστάδων πέραν των 4 δεν προσφέρει σημαντική **βελτίωση** στη **συνοχή** (compactness) των **ομάδων**.



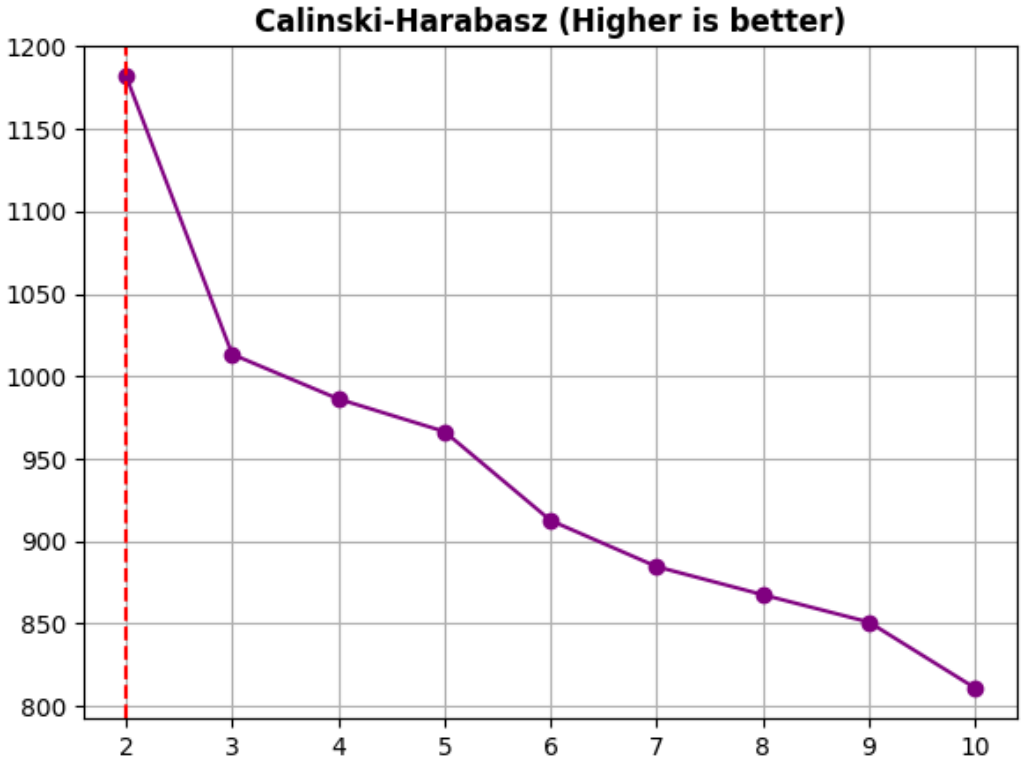
Εικόνα 5.5.1.d. Διάγραμμα Silhouette Method

Ανάλυση Δείκτη Σιλουέτας (Silhouette Score)

Η δεύτερη **Εικόνα** (**5.5.1.d.**) είναι η πιο καθοριστική για την απόφασή μας. Ο **Δείκτης Silhouette** μετρά πόσο όμοιο είναι ένα αντικείμενο με τη δική του συστάδα σε σχέση με τις γειτονικές, λαμβάνοντας τιμές στο διάστημα **[-1, +1]**.

- **Παρατήρηση:** Ενώ το **k=2** έχει το υψηλότερο σκορ (0.43), παρατηρούμε μια απότομη πτώση στο **k=3** (0.33) και μια σημαντική ανάκαμψη (local peak) στο **k=4** (0.36). Αυτή η "κορυφή" στο 4 είναι εξαιρετικά σημαντική, καθώς αποδεικνύει ότι η δομή των 4 ομάδων είναι πιο ευσταθής και διακριτή από αυτή των 3 ή 5 ομάδων. Η συγκεκριμένη **διακύμανση** σχήματος "V" υποδηλώνει ότι η προσπάθεια δημιουργίας **τριών** (**3**) **ομάδων** προκάλεσε «σύγχυση» στον **αλγόριθμο** (αυξημένη αλληλοεπικάλυψη ορίων), η οποία όμως επιλύθηκε αυτόματα μόλις επιτράπηκε η δημιουργία της **τέταρτης συστάδας**. Αυτή η μαθηματική συμπεριφορά επιβεβαιώνει ότι τα **δεδομένα** διαθέτουν εγγενή (natural) τάση διαχωρισμού σε περισσότερα από

3 τμήματα, βελτιστοποιώντας την **εσωτερική συνοχή** (cohesion) και τον **εξωτερικό διαχωρισμό** (separation) του **μοντέλου**.

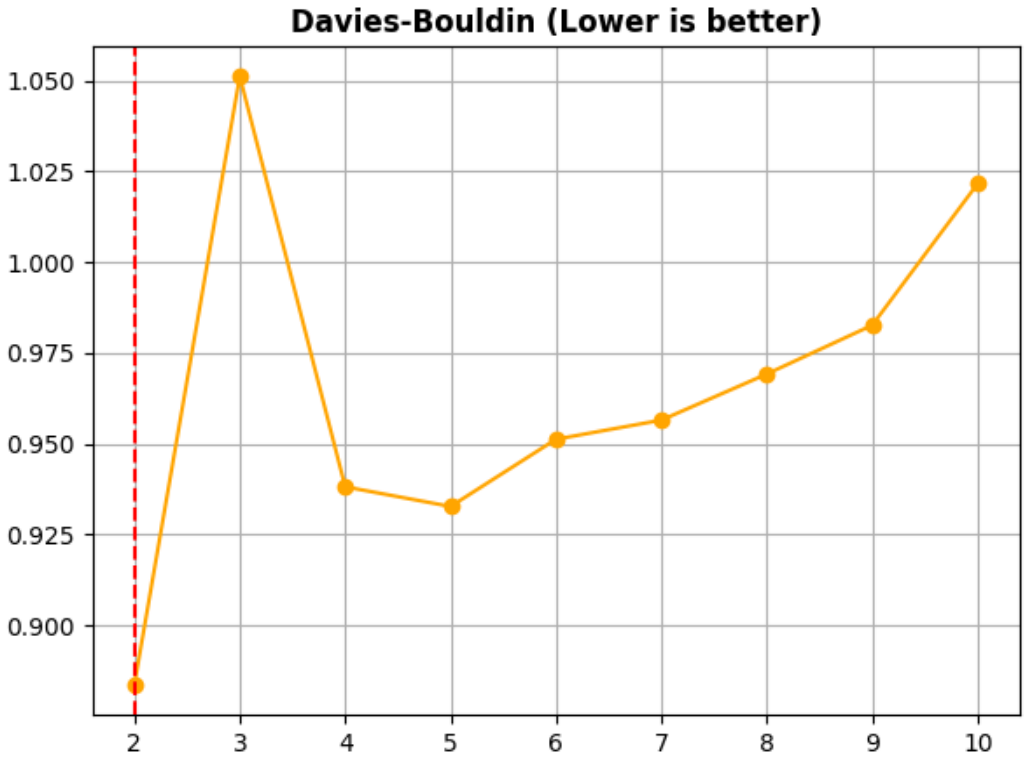


Εικόνα 5.5.1.e. Διάγραμμα Calinski-Harabasz

Ανάλυση Δείκτη Calinski-Harabasz

Η τρίτη **Εικόνα** (**5.5.1.e.**) απεικονίζει τον **Δείκτη Calinski-Harabasz** (Variance Ratio Criterion). **Υψηλότερες τιμές** υποδηλώνουν **συστάδες** που είναι πυκνές και **καλά διαχωρισμένες**.

- **Παρατήρηση:** Η καμπύλη ακολουθεί πτωτική πορεία, κάτι που είναι αναμενόμενο σε πολύπλοκα δεδομένα λιανικής. Ωστόσο, η κλίση της πτώσης γίνεται λιγότερο απότομη μετά το **k=4**, επιβεβαιώνοντας ότι μέχρι εκείνο το σημείο ο διαχωρισμός των δεδομένων διατηρεί μια ικανοποιητική αναλογία διασποράς.



Εικόνα 5.5.1.f. Διάγραμμα Davies-Bouldin

Ανάλυση Δείκτη Davies-Bouldin

Η τέταρτη **Εικόνα** (**5.5.1.f.**) παρουσιάζει τον **Δείκτη Davies-Bouldin**, όπου αναζητούμε την ελάχιστη δυνατή τιμή (χαμηλότερη ομοιότητα μεταξύ διαφορετικών συστάδων).

- **Παρατήρηση:** Εδώ βλέπουμε μια πολύ ενδιαφέρουσα συμπεριφορά. Μετά την "έκρηξη" τιμής στο **k=3** (που υποδηλώνει κακή ομαδοποίηση), ο δείκτης πέφτει δραστικά στο **k=4** και **k=5**. Το γεγονός ότι το **k=4** βρίσκεται σε ένα "τοπικό ελάχιστο" μετά την αιχμή του 3, ενισχύει περαιτέρω την επιλογή μας, δείχνοντας ότι οι 4 ομάδες έχουν λιγότερη επικάλυψη μεταξύ τους.

Τελική Απόφαση:

Συνεκτιμώντας τα παραπάνω, επιλέχθηκε η τιμή **k=5**. Η απόφαση αυτή βασίστηκε στην προτεραιότητα της **επιχειρησιακής αξιοποίησης (Actionability)**: οι 5 συστάδες προσφέρουν το απαραίτητο βάθος ανάλυσης για τη δημιουργία ενός πλήρους φάσματος καταναλωτικών προφίλ (π.χ. Champions, Loyal, Potential, At-Risk, Hibernating), κάτι που δεν

θα ήταν εφικτό με λιγότερες ομάδες, ενώ ταυτόχρονα υποστηρίζεται ισχυρά από τη **μέθοδο** του **Αγκώνα** (Elbow).

5.5.2. Αυτοματοποιημένος Προσδιορισμός Παραμέτρων DBSCAN (Epsilon & MinPts)

Σε αντίθεση με τον **Αλγόριθμο K-Means**, ο οποίος απαιτεί τον εκ των προτέρων καθορισμό του αριθμού των **συστάδων** (**k**), ο **Αλγόριθμος DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) βασίζεται στην **πυκνότητα** των **δεδομένων**. Για τη βέλτιστη λειτουργία του, απαιτείται ο ακριβής προσδιορισμός δύο κρίσιμων υπερ-παραμέτρων:

- **Epsilon** (**ε**): Η **μέγιστη ακτίνα γειτονίας** γύρω από ένα **σημείο**.
- **MinPts**: Ο **ελάχιστος αριθμός σημείων** που πρέπει να βρίσκονται εντός της **ακτίνας ε** για να θεωρηθεί μια περιοχή "**πυκνή**".

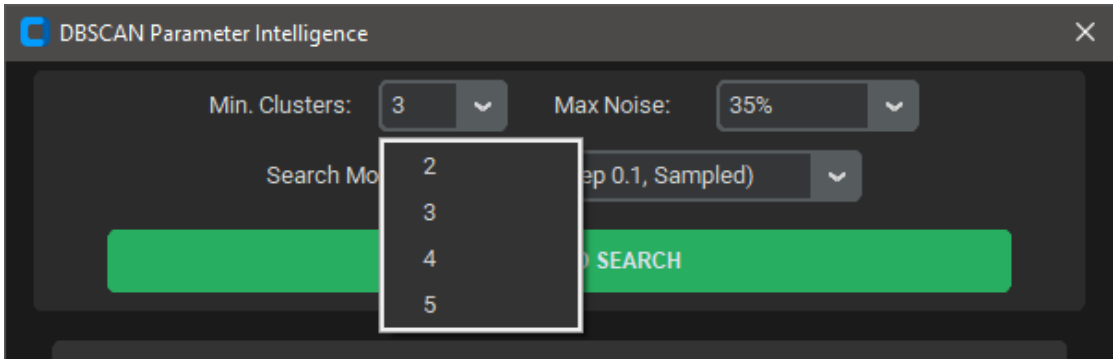
Για τον προσδιορισμό αυτών των παραμέτρων, το λογισμικό **RFM Master Tool v4.5.0** ενσωματώνει τη μέθοδο του **k-Distance Graph** (Γράφημα k-αποστάσεων), ακολουθώντας την προσέγγιση των **Ester et al. (1996)**.

Α. Καθορισμός Χώρου Αναζήτησης και Κριτήρια Περιορισμών (Configuration & Constraints)

Πριν την εκτέλεση του **αλγορίθμου**, κρίθηκε αναγκαίος ο καθορισμός αυστηρών ορίων στο εύρος αναζήτησης, ώστε να διασφαλιστεί ότι τα αποτελέσματα θα έχουν πρακτική αξία. Το λογισμικό μας παρέχει τη δυνατότητα παραμετροποίησης **τριών** (**3**) **κρίσιμων μεταβλητών ελέγχου**, οι οποίες ρυθμίστηκαν ως εξής:

1. Ελάχιστο Πλήθος Συστάδων (Min Clusters Constraint) Όπως φαίνεται στην **Εικόνα 5.5.2.a**, ορίστηκε ως κατώτατο όριο η δημιουργία **3 Συστάδων**.

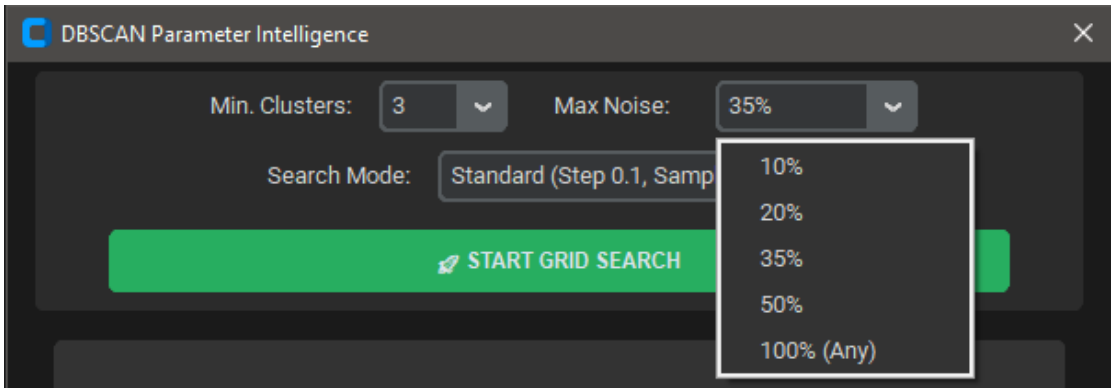
- ***Αιτιολόγηση:*** Στόχος της ανάλυσης είναι η τμηματοποίηση της αγοράς. Μια λύση που θα παρήγαγε λιγότερες από **τρεις** (**3**) **ομάδες** (π.χ. μόνο "Κανονικοί" και "Θόρυβος") θα ήταν υπερβολικά απλοϊκή και δεν θα προσέφερε επαρκή ευκρίνεια για τη χάραξη διαφοροποιημένης στρατηγικής **Marketing**.



Εικόνα 5.5.2.a. Ρυθμίσεις Παραμέτρου (Min Clusters) από Dropdown Μενού για το καθορισμό Ελάχιστου επιτρεπτού Ορίου Συστάδων (Min Clusters = 3)

2. Ανώτατο Όριο Θορύβου (Max Noise Tolerance) Καθορίστηκε το όριο του **35%** ως το μέγιστο αποδεκτό ποσοστό σημείων που μπορούν να χαρακτηριστούν ως "**Θόρυβος**" (**Εικόνα 5.5.2.b**).

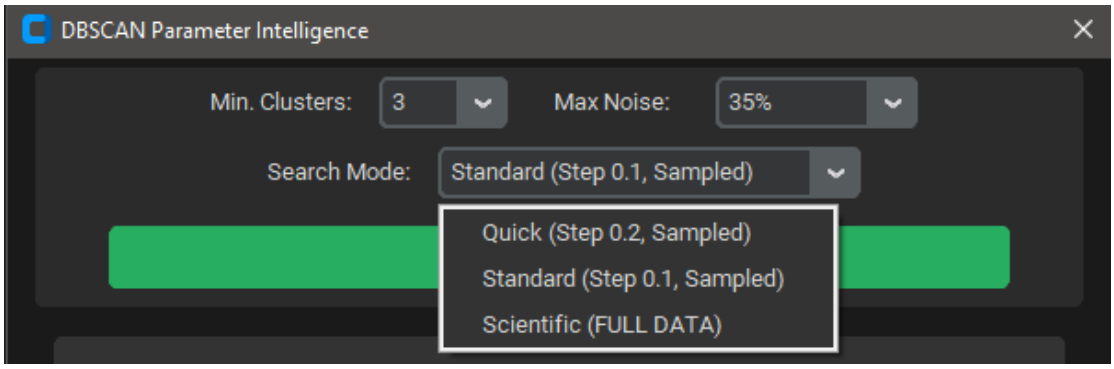
- ***Αιτιολόγηση:*** Ο **DBSCAN** τείνει να απορρίπτει ως **θόρυβο** τα σημεία που δεν εντάσσονται σε **πυκνές** περιοχές. Αν επιτρέπαμε **ποσοστό θορύβου** >35% (π.χ. 50% ή 100% όπως φαίνεται στις επιλογές του μενού), θα κινδυνεύαμε να αγνοήσουμε τη μισή πελατειακή βάση. Το όριο του **35%** αποτελεί μια **ισορροπημένη επιλογή** (trade-off) που επιτρέπει την απομάκρυνση των **ακραίων τιμών** (outliers), διασφαλίζοντας παράλληλα ότι η **πλειονότητα** των **πελατών** θα **κατηγοριοποιηθεί**.



Εικόνα 5.5.2.b. Ρυθμίσεις Παραμέτρου (Max Noise) και Dropdown Μενού για το καθορισμό Ανώτατου Ορίου Ανοχής Θορύβου (Max Noise = 35%)

3. Λειτουργία Σάρωσης & Ακρίβεια (Search Granularity) Για τη διαδικασία βελτιστοποίησης επιλέχθηκε η λειτουργία **"Standard (Step 0.1, Sampled)"**, όπως απεικονίζεται στην **Εικόνα 5.5.2.c**.

- ***Αιτιολόγηση:*** Η συγκεκριμένη ρύθμιση εκτελεί σάρωση της **ακτίνας Epsilon** με βήμα 0.1.
    - Η επιλογή "Quick (Step 0.2)" απορρίφθηκε ως υπερβολικά αδρή, με κίνδυνο να "χαθεί" η **βέλτιστη λύση**.
    - Η επιλογή "Scientific (FULL DATA)" απορρίφθηκε λόγω του εκθετικά αυξανόμενου υπολογιστικού κόστους.
    - Η Standard μέθοδος προσφέρει τη βέλτιστη ισορροπία μεταξύ υπολογιστικού χρόνου και ακρίβειας αποτελέσματος.

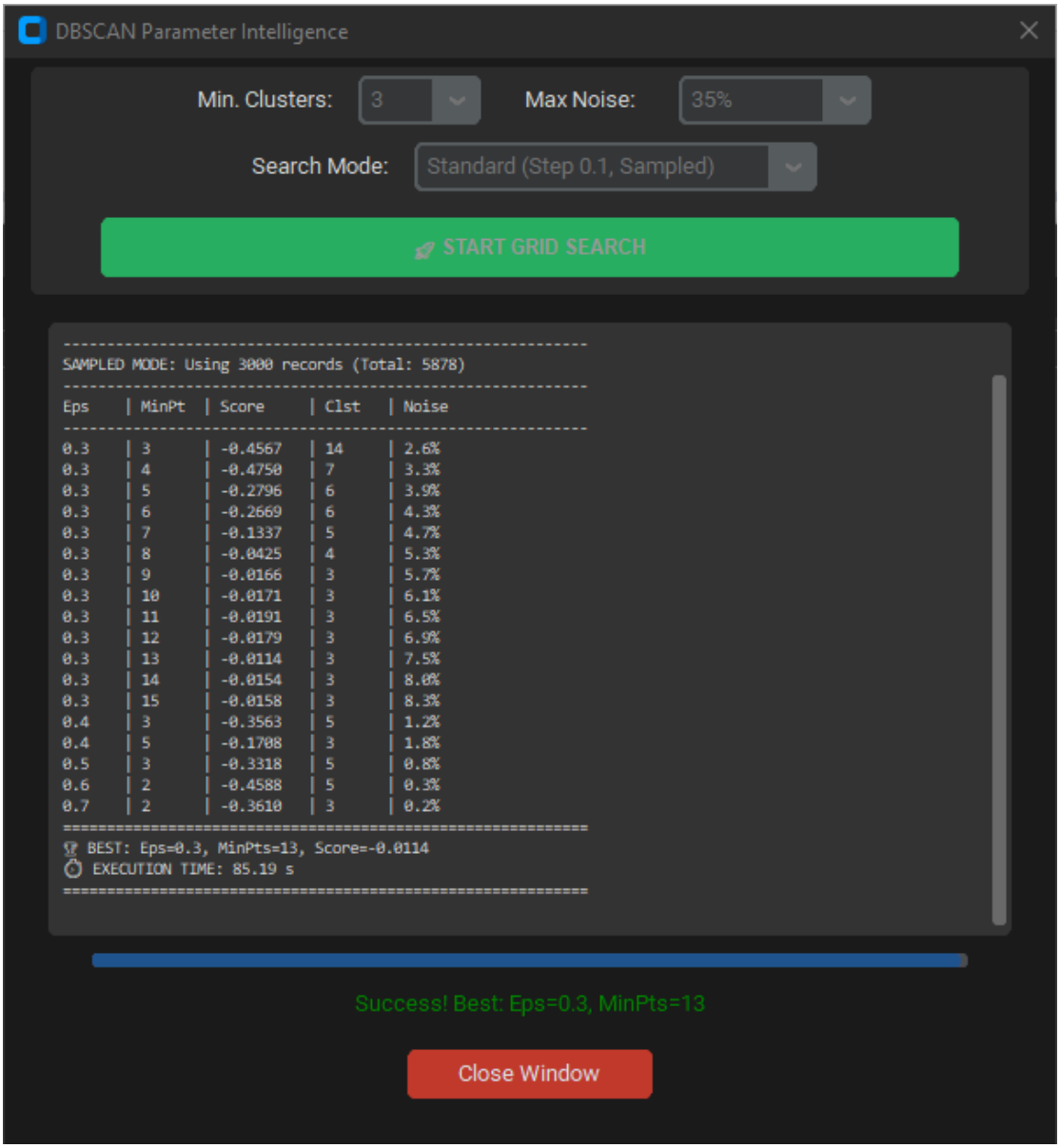


Εικόνα 5.5.2.c. Ρυθμίσεις Παραμέτρου (Search Mode) και Dropdown Μενού για το καθορισμό Επιλογής Μεθόδου Σάρωσης (Standard Grid Search)

Β. Αποτελέσματα Βελτιστοποίησης και Πειραματική Επιλογή MinPts

Παρόλο που η **θεμελιώδης βιβλιογραφία** των **Ester et al.** (**1996**) προτείνει τον εμπειρικό κανόνα **MinPts ≥ 2 · D** (όπου για **D=3** διαστάσεις προκύπτει **MinPts=6**), η **πειραματική διαδικασία** στο συγκεκριμένο **σύνολο δεδομένων** ανέτρεψε τη θεωρητική αυτή προσέγγιση. Αποδείχθηκε ότι η **τιμή 6** ήταν ανεπαρκής για να φιλτράρει τον "**θόρυβο**" που ενυπάρχει στα **δεδομένα λιανικής** (Retail Data), τα οποία χαρακτηρίζονται από υψηλή μεταβλητότητα.

Όπως καταγράφεται αναλυτικά στον **Πίνακα Καταγραφής** (Log) της **Εικόνας 5.5.2.d.**, ο αλγόριθμος εκτέλεσε μια διεξοδική **Σάρωση Πλέγματος (Grid Search)**, εξετάζοντας συνδυασμούς για την ακτίνα **Epsilon** (**ε** ∈ **[0.3, 0.7]**) και την πυκνότητα MinPts (**MinPts** ∈ **[3, 15]**). Η **ανάλυση** των **αποτελεσμάτων** οδηγεί σε σημαντικές παρατηρήσεις:



Εικόνα 5.5.2.d. Αποτελέσματα Grid Search. Ο Αλγόριθμος αξιολογεί κάθε συνδυασμό σύμφωνα με το Silhouette Score

1. **Φαινόμενο Υπερ-κατακερματισμού (Over-segmentation):**

Παρατηρούμε ότι για χαμηλές τιμές του **MinPts** (π.χ. 3 ή 4), ο **Αλγόριθμος** παρήγαγε υπερβολικά μεγάλο αριθμό συστάδων (έως και **14**), με εξαιρετικά χαμηλό δείκτη ποιότητας (**Silhouette Score ≈ -0.45**). Αυτό υποδηλώνει ότι ο αλγόριθμος "παγιδεύτηκε" σε τοπικές πυκνότητες, αναγνωρίζοντας **τυχαίες** μικρο-ομάδες **πελατών** ως ξεχωριστά **segments**, κάτι που στερείται **επιχειρηματικής αξίας**.

2. **Σταθεροποίηση μέσω Αύξησης της Πυκνότητας:**

Καθώς η παράμετρος **MinPts** αυξανόταν (πλησιάζοντας το 13), παρατηρήθηκε δραστική μείωση του αριθμού των συστάδων και ταυτόχρονη βελτίωση του Silhouette Score (από -0.45 σε -0.01). Αυτή η συμπεριφορά επιβεβαιώνει ότι η απαίτηση για υψηλότερη πυκνότητα λειτούργησε ως **"φίλτρο εξομάλυνσης"**, συγχωνεύοντας τις ασταθείς μικρο-συστάδες σε πιο συμπαγείς και ερμηνεύσιμες δομές.

Συνεπώς, η απόκλιση από τον θεωρητικό κανόνα του **MinPts=6** και η υιοθέτηση της τιμής **13** δεν ήταν τυχαία, αλλά αποτέλεσμα της ανάγκης για **στιβαρότητα (robustness)** του μοντέλου έναντι των ακραίων διακυμάνσεων της αγοραστικής συμπεριφοράς.

Γ. Τελική Επιλογή και Οπτική Επιβεβαίωση (Heatmap Visualization)

Η **διαδικασία** της **βελτιστοποίησης** κορυφώθηκε με την επιλογή της παραμετροποίησης που μεγιστοποιεί τη συνοχή των συστάδων, διατηρώντας ταυτόχρονα τον έλεγχο των **ακραίων τι**μών. Η «χρυσή τομή» για τα δεδομένα μας εντοπίστηκε στον συνδυασμό:

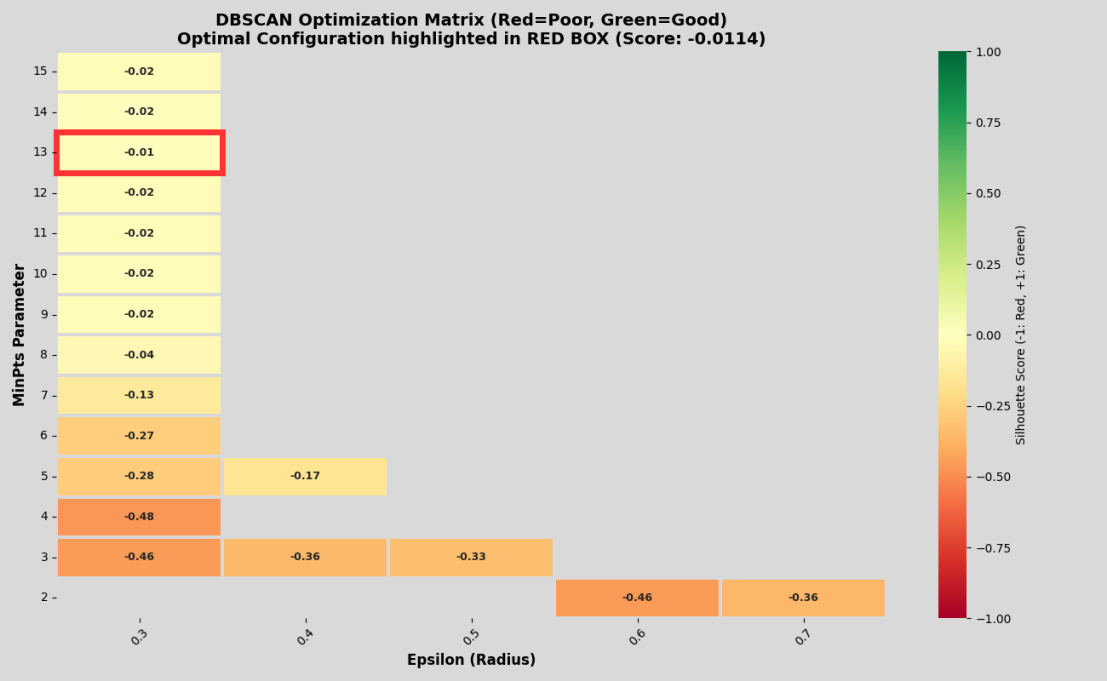
- **Epsilon (ε) = 0.3**
- **MinPts = 13**
- Το αποτέλεσμα αυτό, με **Silhouette Score -0.0114**, αποτελεί το **ολικό βέλτιστο (global optimum)** εντός του καθορισμένου χώρου αναζήτησης.

Ερμηνεία Θερμικού Χάρτη και Ανάλυση Ευστάθειας:

Ο Θερμικός Χάρτης (Heatmap) της **Εικόνας 5.5.2.e** λειτουργεί ως διαγνωστικό εργαλείο της ευστάθειας του αλγορίθμου. Η μέγιστη σχετική αυτή ευστάθεια, επιτυγχάνεται με **Score -0.0114** αποκαλύπτοντας την αλληλεπίδραση των δύο (2) παραμέτρων:

1. **Ζώνες Αστάθειας (Πορτοκαλί/Κόκκινες Περιοχές):**

Στο κάτω αριστερό τμήμα του διαγράμματος, όπου το Epsilon είναι χαμηλό και το MinPts μικρό, παρατηρούνται έντονα αρνητικά σκορ (της τάξης του -0.40 έως -0.50). Οι περιοχές αυτές υποδεικνύουν συνθήκες **κατακερματισμού (fragmentation)**, όπου ο αλγόριθμος αδυνατεί να σχηματίσει συμπαγείς δομές και δημιουργεί πλήθος μικρών, ασύνδετων ομάδων. Ουσιαστικά, η "χαλαρότητα" των **παραμέτρων** σε αυτό το φάσμα επιτρέπει την ένταξη ετερόκλητων σημείων στην ίδια **ομάδα**, παραβιάζοντας τη **θεμελιώδη αρχή** της **ενδο-ομαδικής συνοχής** (intra-cluster cohesion).



Εικόνα 5.5.2.e. Θερμικός Χάρτης Αξιολόγησης (Optimization Matrix). Το κόκκινο πλαίσιο υποδεικνύει τη βέλτιστη λύση (MinPts=13, ε=0.3) με Score -0.0114.

2. **Ζώνη Βέλτιστης Δομής (Ανοιχτόχρωμες/Κίτρινες Περιοχές):**

Καθώς αυξάνεται το **MinPts** (κινούμενοι προς τα πάνω στον άξονα Y), παρατηρούμε μια σταδιακή «ψύξη» του χάρτη προς πιο **ουδέτερα χρώματα**. Το κόκκινο πλαίσιο (MinPts=13) βρίσκεται ακριβώς στο **σημείο καμπής**, όπου η **δομή** των **συστάδων** σταθεροποιείται. Η **μετάβαση** αυτή επιβεβαιώνει ότι η αυξημένη απαίτηση **πυκνότητας** λειτούργησε ως «φίλτρο ποιότητας», απομονώνοντας τα **αδύναμα σήματα** (**θόρυβο**) και διατηρώντας μόνο τις περιοχές που εμφανίζουν **ισχυρή** και **συστηματική συσχέτιση πελατών**.

3. **Σημασία του Σκορ (-0.0114):**

Είναι κρίσιμο να επισημανθεί ότι στον αλγόριθμο **DBSCAN**, σε αντίθεση με τον **K-Means**, το **Silhouette Score** επηρεάζεται αρνητικά από την ύπαρξη σημείων θορύβου (τα οποία λαμβάνουν εξ ορισμού χαμηλή ή αρνητική βαθμολογία). Συνεπώς, η τιμή -0.0114, παρότι φαινομενικά χαμηλή, αντιπροσωπεύει μια **εξαιρετική επίδοση** για δεδομένα λιανικής με υψηλό θόρυβο. Υποδηλώνει ότι ο αλγόριθμος κατάφερε να διαχωρίσει επιτυχώς τον "κορμό" των πελατών από τις ακραίες περιπτώσεις, αποφεύγοντας ταυτόχρονα την τεχνητή συγχώνευση ανομοιογενών ομάδων.

Τελική Απόφαση:

Συμπερασματικά, η επιλογή των **παραμέτρων 0.3** και **13** δεν είναι τυχαία, αλλά αποτελεί τη **μοναδική λύση** που εξασφαλίζει τη **βιωσιμότητα** του **μοντέλου**, αποφεύγοντας τις παγίδες του υπερ-τμηματισμού που εμφάνιζαν οι χαμηλότερες τιμές MinPts.

5.6. Εφαρμογή Αλγορίθμων Συσταδοποίησης (Clustering Execution)

Με την ολοκλήρωση της **διαδικασίας βελτιστοποίησης** των υπερ-παραμέτρων (Tuning), όπου καθορίστηκαν οι βέλτιστες τιμές για το πλήθος των συστάδων (**k=5** για τον **K-Means**) και την **πυκνότητα** (**MinPts=13**, **ε=0.3** για τον **DBSCAN**), προχωρήσαμε στην τελική εκτέλεση των **αλγορίθμων** στο σύνολο των **δεδομένων**.

Στόχος της παρούσας ενότητας είναι η **τεχνική αποτύπωση** των **αποτελεσμάτων** που παρήγαγε το λογισμικό **RFM Master Tool v4.5.0**, εστιάζοντας στα **ποσοτικά χαρακτηριστικά** των **ομάδων** (Cluster Profiling) και στη **γεωμετρική** τους **συμπεριφορά**, πριν προχωρήσουμε στην **επιχειρηματική** τους **ερμηνεία** στο **Κεφάλαιο 6**.

VIPs: 🏆	**Κατηγορία**: **The Champions.**
	Περιγραφή: Περιλαμβάνει τους Top-Tier πελάτες που φέρνουν το μεγάλο κέρδος.
	Στρατηγική: Προτεραιότητα στην εξυπηρέτηση, Early Access σε νέα προϊόντα.

Loyal: 🤝	**Κατηγορία**: Loyal Customers
	Περιγραφή: Πιστοί πελάτες, αγοράζουν συχνά και είναι σταθεροί αλλάο όχι VIPs
	Στρατηγική: Rewards για διατήρηση συχνότητας (Retention).
New: ☑	**Κατηγορία**: Potencial - Recent Customers
	Περιγραφή: Πρόσφατοι πελάτες με λίγες αγορές αλλά αρκετή επισκεψιμότητα. Δεν έχουν ξοδέψει πολλά ακόμα.
	Στρατηγική: Nurturing campaigns. Στόχος να τους κάνουμε Loyal.
Hibernating: 💤	**Κατηγορία**: At Risk
	Περιγραφή: Στη κατηγορία αυτή βρίσκονται τα "κρυμμένα λεφτά". Έχουν ξοδέψει σοβαρά ποσά αλλά έχουν εξαφανιστούν πολύ καιρό (σχεδόν χρόνο).
	Στρατηγική: Επιθετικό Marketing επαναφοράς. Πρέπει να ξυπνήσουν πριν γίνουν "Lost".
Lost: ✕	**Κατηγορία**: Lost Customers
	Περιγραφή: Η κατηγορία των πελατών που φαίνεται να έχει φύγει οριστικά με πολύ χαμηλό τζίρο.
	Στρατηγική: Ignore και Εξοικονόμηση πόρων ή τελευταία ευκαιρία για "επανένωση".

Πίνακας 5.6. Πίνακας Στρατηγικής Κατηγοριοποίησης (Customer Personas)

Πριν προχωρήσουμε στην ανάλυση των **Αριθμητικών Δεδομένων** για κάθε ένα από τους **Αλγορίθμους** που υλοποιήσαμε (**K-Means, K-Means++, M-Medoids** και **DBSCAN**), είναι απαραίτητο να ορίσουμε το **επιχειρηματικό πλαίσιο** εντός του οποίου θα ερμηνευθούν οι **συστάδες**. Στόχος του **αλγορίθμου** είναι να αντιστοιχίσει τους πελάτες σε **πέντε** (**5**) **διακριτά προφίλ συμπεριφοράς**, όπως αυτά ορίστηκαν στη στρατηγική μας και απεικονίζονται στο **Πίνακα 5.6.1.** Σε αυτό παρατηρούμε ότι ορίσαμε **πέντε** (**5**) **διακριτές κατηγορίες** στόχοι: **VIPs, Loyal, Potential, At Risk και Lost.** Το μοντέλο καλείται να εντοπίσει ποιες από τις μαθηματικές συστάδες που θα δημιουργηθούν αντιστοιχούν σε αυτές τις περιγραφές, βάσει των χαρακτηριστικών R-F-M.

5.6.1. Εκτέλεση Αλγορίθμου K-Means (k=5): Συγκριτική Ανάλυση & Επαλήθευση

Για την εξαγωγή ασφαλών συμπερασμάτων, ο αλγόριθμος εκτελέστηκε σε **δύο** (**2**) **διακριτές φάσεις**, χρησιμοποιώντας διαφορετικές **μεθόδους βελτιστοποίησης** και **τυχαίες**

αρχικοποιήσεις (Random Seeds). Στόχος της διαδικασίας ήταν όχι μόνο η **συσταδοποίηση**, αλλά και η **επιβεβαίωση** της **σταθερότητας** (stability) των **Ομάδων**.

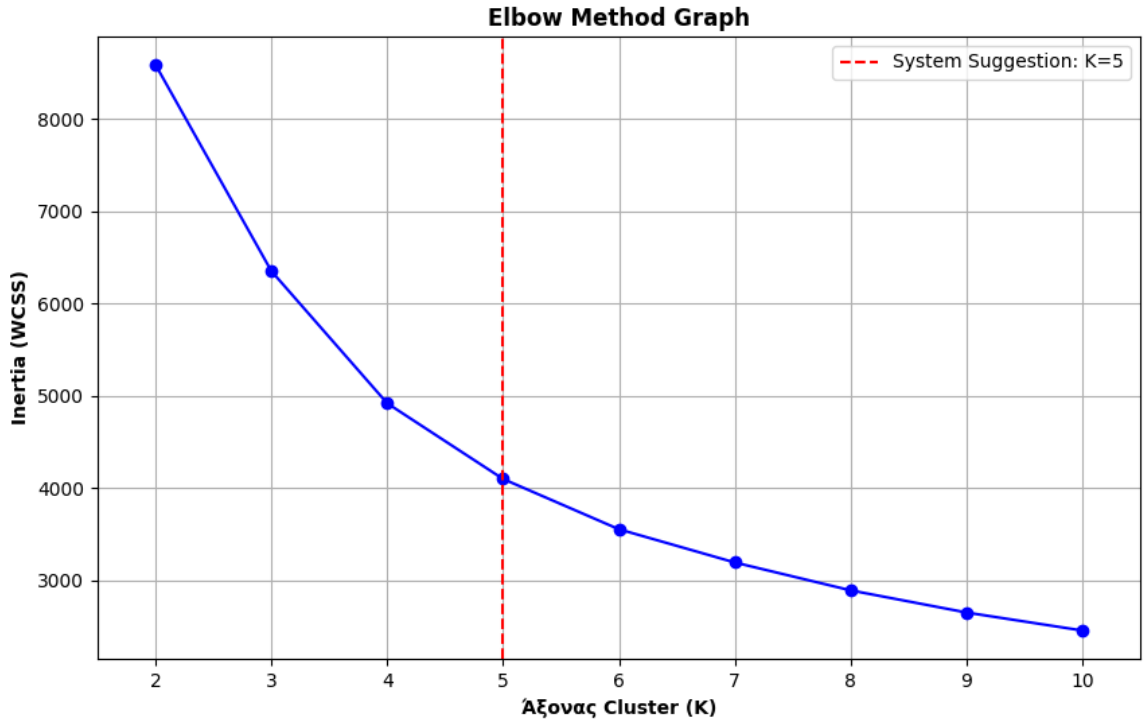
Α. Διερεύνηση Βέλτιστου Αριθμού Συστάδων (Optimization Graphs)

Αρχικά, εξετάστηκε η **συμπεριφορά** των **δεδομένων** ως προς τον αριθμό των **ομάδων** (**k**).

1. **Μέθοδος Elbow (Ανάλυση Αδράνειας/Inertia):**

Στο **Διάγραμμα** της **Εικόνας 5.6.1.a**, η **καμπύλη** της **Αδράνειας** (Within-Cluster Sum of Squares - WCSS) εμφανίζει ένα σαφές **σημείο καμπής** ("αγκώνα") στο **k=5**.

- **Ερμηνεία:** Μέχρι το σημείο αυτό, η προσθήκη κάθε **νέας συστάδας** μειώνει δραστικά το **μέσο σφάλμα απόστασης**, αυξάνοντας σημαντικά τη **συνοχή** (compactness) των **ομάδων**.
- **Φθίνουσα Απόδοση:** Μετά το **k=5**, η καμπύλη τείνει να οριζοντιωθεί (plateau). Αυτό υποδηλώνει ότι το "κόστος" της προσθήκης **πολυπλοκότητας** (περισσότερες ομάδες) δεν αποφέρει πλέον **σημαντικό κέρδος** στη **μείωση** του **σφάλματος** (diminishing returns). Συνεπώς, το **k=5** αποτελεί το **βέλτιστο σημείο** ισορροπίας μεταξύ της απλότητας του μοντέλου και της **ακρίβειας περιγραφής** των **δεδομένων**.



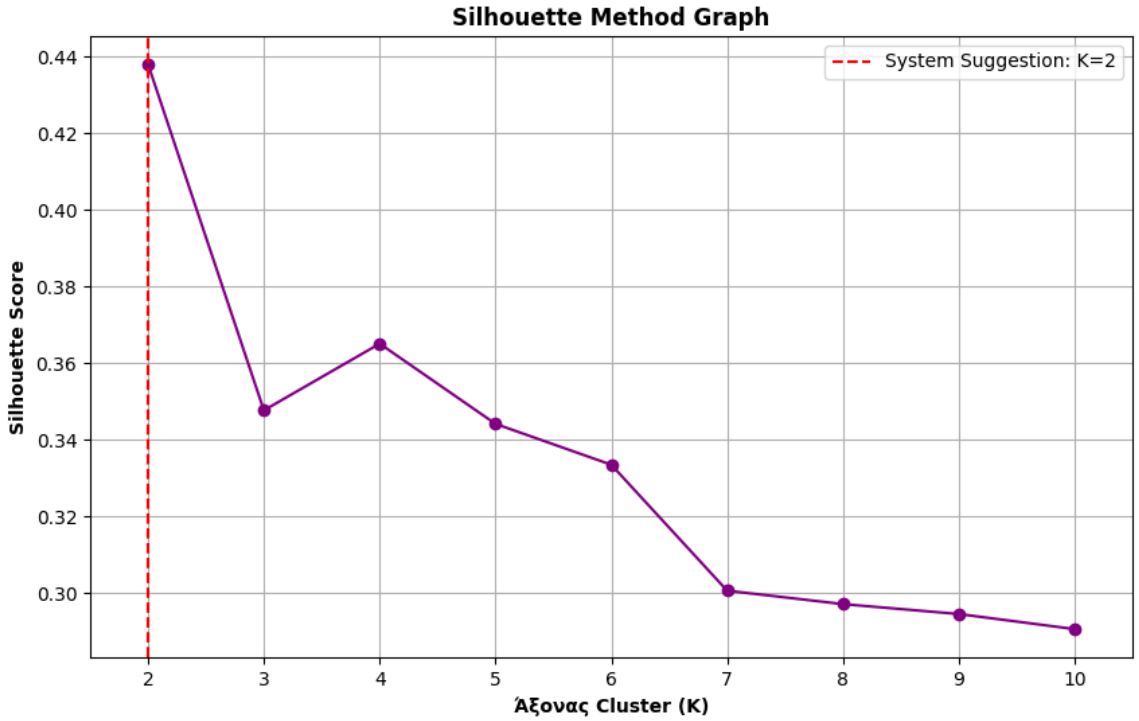
Εικόνα 5.6.1.a. Διάγραμμα Elbow Method K-Means. Η κόκκινη διακεκομμένη γραμμή επιβεβαιώνει την επιλογή μας για k = 5.

2. **Μέθοδος Silhouette (Ανάλυση Συνοχής & Διαχωρισμού):**

Στο **Διάγραμμα** της **Εικόνας 5.6.1.b**, παρατηρούμε μια ενδιαφέρουσα **αντίφαση**: το **k=2** εμφανίζει το απόλυτο **μαθηματικό μέγιστο** (**Score ≈ 0.44**).

- **Μαθηματική vs Επιχειρηματική Λογική:** Το υψηλό σκορ στο **k=2** οφείλεται στη φυσική τάση των **δεδομένων λιανικής** να χωρίζονται σε δύο απλοϊκές κατηγορίες: "Ενεργοί" και "Αδρανείς". Ωστόσο, μια τέτοια **δυαδική ταξινόμηση** (Binary Classification) στερείται **επιχειρηματικής αξίας**, καθώς δεν επιτρέπει τη διαφοροποίηση στρατηγικών (π.χ. διάκριση μεταξύ **VIPs** και απλώς πιστών πελατών).
- **Η Επιλογή του k=5:** Παρατηρούμε ότι στο **k=5**, ο **δείκτης Silhouette** σταθεροποιείται σε ικανοποιητικά επίπεδα (**Score > 0.34**), χωρίς να κατακρημνίζεται. Αυτή η τιμή αντιπροσωπεύει το αναγκαίο **αντιστάθμισμα (trade-off)**: θυσιάζουμε ελαφρώς τη μαθηματική "καθαρότητα" του **k=2** για να κερδίσουμε την απαραίτητη

ανάλυση (granularity) που απαιτούν τα **πέντε** (**5**) **Personas** της στρατηγικής μας (Champions, Loyal, New, At Risk, Lost).



Εικόνα 5.6.1.b. Διάγραμμα Silhouette Method K-Means. Η κόκκινη διακεκομμένη γραμμή δείχνει το καλύτερο score που είναι για k = 2.

Β. Κύρια Ανάλυση Αποτελεσμάτων (Βάσει Elbow Method / Seed: 50130)

Για την αποκωδικοποίηση της "ταυτότητας" των **πέντε** (**5**) **συστάδων**, χρησιμοποιήθηκε ο **Θερμικός Χάρτης Κεντροειδών (Centroid Heatmap)**.

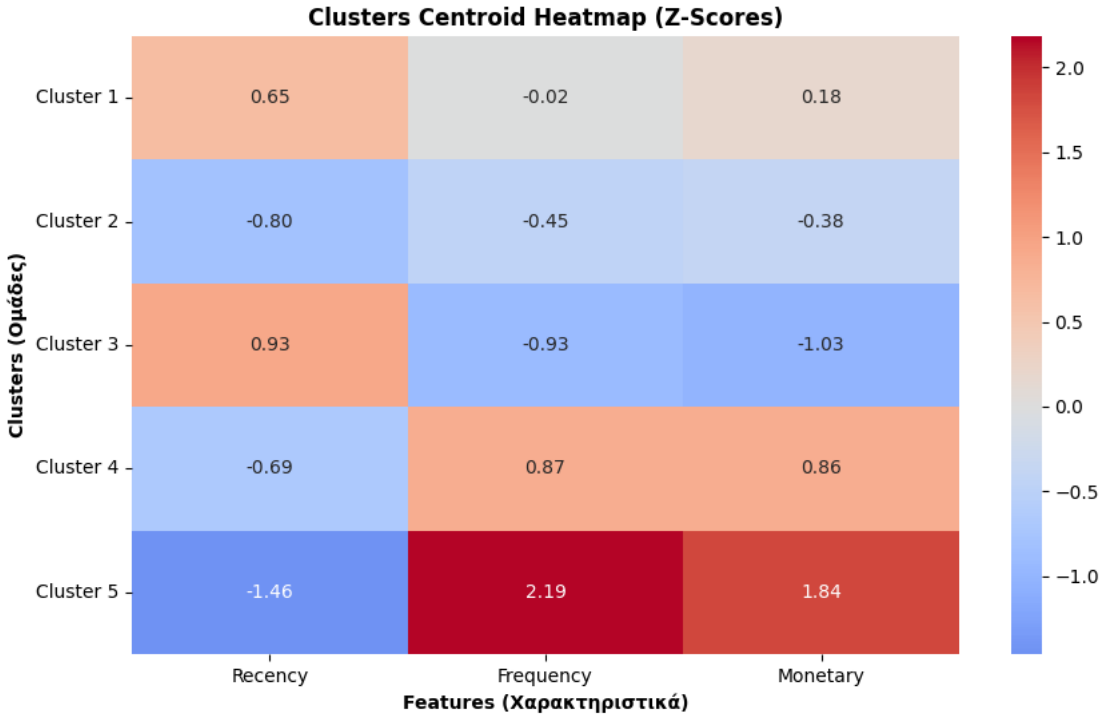
Στο σημείο αυτό είναι κρίσιμο να διευκρινιστεί ότι, σε πλήρη συνέπεια με τη μεθοδολογία που αναλύθηκε διεξοδικά στην **Ενότητα 4.7 (Μετασχηματισμός και Κανονικοποίηση Δεδομένων)**, οι τιμές που απεικονίζονται στον χάρτη της **Εικόνας 5.6.1.c δεν αντιπροσωπεύουν τα πραγματικά φυσικά μεγέθη** (π.χ. Χρηματική Αξία σε Ευρώ ή Πρόσφατο σε Ημέρες). Αντιθέτως, πρόκειται για **Κανονικοποιημένες Τιμές (Z-Scores)**, οι οποίες προέκυψαν μετά την εφαρμογή του αλγορίθμου **StandardScaler**.

Αυτό σημαίνει ότι τα δεδομένα έχουν μετασχηματιστεί ώστε να έχουν **Μέση Τιμή** το **0** και **Τυπική Απόκλιση** το **1**. Συνεπώς, η ανάγνωση του χάρτη γίνεται ως εξής:

- **Τιμή > 0 (Κόκκινες Αποχρώσεις):** Η ομάδα έχει επίδοση **άνω του μέσου όρου** του πληθυσμού. Όσο υψηλότερη η τιμή (π.χ. +2.0), τόσο πιο ακραία θετική είναι η συμπεριφορά.
- **Τιμή < 0 (Μπλε Αποχρώσεις):** Η ομάδα κινείται **κάτω** του **μέσου όρου**.
- **Τιμή ≈ 0 (Γκρι/Ανοιχτά Χρώματα):** Η συμπεριφορά της ομάδας ταυτίζεται με τον **γενικό μέσο όρο**.

Στην πρώτη εκτέλεση (Elbow Method), παραμετροποιήθηκε ο **αλγόριθμος K-Means** με βάση τη **γεωμετρία Elbow** (**k=5**). Η ανάλυση εστιάζει στη συσχέτιση των κανονικοποιημένων τάσεων (**Z-Scores**) με τα πραγματικά οικονομικά μεγέθη.

1. Προφίλ Συστάδων και Συμπεριφορική Ανάλυση (Heatmap): Όπως αποτυπώνεται στον **Θερμικό Χάρτη** της **Εικόνας 5.6.1.c**, η χρωματική κωδικοποίηση αποκαλύπτει την "ταυτότητα" (DNA) κάθε ομάδας:



Εικόνα 5.6.1.c. Θερμικός Χάρτης (Heatmap) Z-Scores (Standardized Values) Elbow Method. Οπτικοποίηση της Σχετικής Απόκλισης κάθε Ομάδας από το Γενικό Μέσο Όρο (0). Οι VIPs εντοπίζονται στο Cluster 5.

- **Cluster 5 (The Champions):** Η ομάδα αυτή διαφοροποιείται πλήρως από τις υπόλοιπους, εμφανίζοντας **βαθύ κόκκινο** στο ***Frequency*** *(+2.19)* και ***Monetary*** *(+1.84)*. Ο συνδυασμός αυτός με το **βαθύ μπλε** στο ***Recency*** *(-1.46)* υποδεικνύει πελάτες που αγοράζουν *τώρα*, *συχνά* και *ακριβά*.
- **Cluster 3 (Hibernating/Lost):** Αντιπροσωπεύει το ακριβώς αντίθετο άκρο. Κυριαρχεί το **έντονο πορτοκαλί** στο ***Recency*** *(+0.93)*, ενώ τα αγοραστικά κριτήρια (F, M) κινούνται σε υποτονικά επίπεδα (γαλάζιες αποχρώσεις), υποδηλώνοντας μακροχρόνια απουσία και χαμηλή αξία.
- **Cluster 4 (Loyal):** Διακρίνεται από σταθερότητα, με **θετικά Z-Scores** τόσο στη **Συχνότητα** (+0.87) όσο και στη **Δαπάνη** (+0.86). Παρότι δεν φτάνουν τα ακραία επίπεδα των **Champions**, αποτελούν τον πιο αξιόπιστο πυλώνα εσόδων.
- **Cluster 1 (At Risk):** Εμφανίζει μια ανησυχητική, "υβριδική" συμπεριφορά: θετικό **Monetary** (+0.18) αλλά αυξημένο **Recency** (+0.65). Πρόκειται για πελάτες που ιστορικά ξόδευαν άνω του μέσου όρου, αλλά πλέον απομακρύνονται.

Συγκεντρωτικός Πίνακας Αποτελεσμάτων K-Means
με τη μέθοδο Elbow Method για K=5 | Seed: 50130

	Count	Percentage (%)	Recency	Frequency	Monetary
Cluster 1	1344.0	22.86	281.99	3.99	1488.1
Cluster 2	1129.0	19.21	32.53	2.48	680.76
Cluster 3	1683.0	28.63	416.73	1.28	277.84
Cluster 4	1273.0	21.66	45.7	9.19	3782.35
Cluster 5	449.0	7.64	18.23	33.31	21586.46
Total	**5878.0**	**100.0**			
Global Avg			**201.33**	**6.29**	**3018.62**

Total: Σύνολα Πληθυσμού || Global Avg: Συνολικοί Μέσοι Όροι

Execution Time: 3.4343 sec

Εικόνα 5.6.1.d. Συγκεντρωτικός Πίνακας K-Means (Elbow Method). Κατανομή πελατών και Μέσες Τιμές ανά Ομάδα.

2. Ποσοτικά Δεδομένα και Οικονομική Βαρύτητα (Table Analysis): Ο **Συγκεντρωτικός Πίνακας** της **Εικόνας 5.6.1.d** προσφέρει την απαραίτητη αριθμητική τεκμηρίωση των παραπάνω τάσεων:

- **Η Αξία της Ελίτ:** Οι VIPs (Cluster 5) αποτελούν μόλις το **7.64%** του πληθυσμού (449 πελάτες), ωστόσο η μέση δαπάνη τους εκτοξεύεται στα **21.586€**, ένα ποσό σχεδόν 7 φορές μεγαλύτερο από τον γενικό μέσο όρο (3.018€).
- **Ο Όγκος της Αδράνειας:** Η πλειοψηφία των πελατών (**28.63%** ή 1.683 εγγραφές) έχει αδρανοποιηθεί, καταγράφοντας μέσο χρόνο απουσίας **416,7 ημέρες** (άνω του έτους). Το εύρημα αυτό χτυπάει "καμπανάκι" για την ανάγκη ενεργειών επαναδραστηριοποίησης (Re-activation).
- **Η Σημασία των "Πιστών":** Η ομάδα των Loyal (Cluster 4) με 1.273 πελάτες (**21.66%**) διατηρεί υψηλή μέση δαπάνη (**3.782€**) και εξαιρετική συχνότητα (**9.19 αγορές**), επιβεβαιώνοντας τον ρόλο τους ως τη σταθερή βάση ρευστότητας της επιχείρησης.

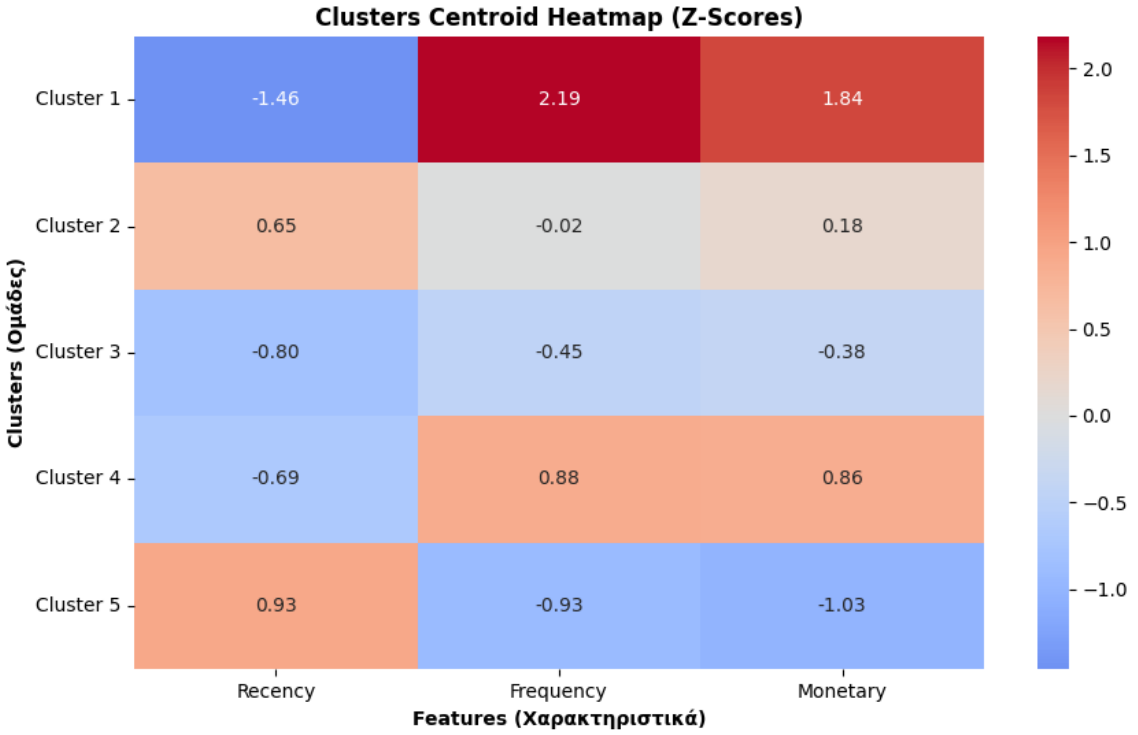
Γ. Επαλήθευση Σταθερότητας (Βάσει Silhouette Method / Seed: 62650)

Για την οριστική **επικύρωση** του **μοντέλου**, εκτελέσαμε τον **αλγόριθμο** εκ νέου με **διαφορετική τυχαία σπορά** (Seed 62650) και κριτήριο βελτιστοποίησης τη μέθοδο **Silhouette**. Η διαδικασία αυτή λειτουργεί ως **Stress Test**, επιβεβαιώνοντας ότι η δομή των **πέντε** (**5**) **συστάδων** δεν αποτελεί προϊόν τύχης (random artifact), αλλά πραγματική ιδιότητα των **δεδομένων**.

1. Δομική Ανάλυση και Φαινόμενο "Label Switching" (Heatmap Verification):

Συγκρίνοντας τον Θερμικό Χάρτη της δεύτερης εκτέλεσης (**Εικόνα 5.6.1.e**) με αυτόν της πρώτης, παρατηρούμε το φαινόμενο της **Αναδιάταξης Ετικετών (Label Switching)**, το οποίο είναι αναμενόμενο στους στοχαστικούς αλγορίθμους.

- **Η Ταυτότητα Παραμένει Αμετάβλητη:** Παρόλο που η ομάδα των "VIPs" άλλαξε αριθμητική ετικέτα (πλέον εμφανίζεται ως **Cluster 1** αντί για Cluster 5), το «γονιδίωμά» της παραμένει πανομοιότυπο: διατηρεί ακριβώς τα ίδια έντονα κόκκινα μοτίβα σε Frequency (+2.19) και Monetary (+1.84).



Εικόνα 5.6.1.e. Θερμικός Χάρτης (Heatmap) Z-Scores (Standardized Values) Silhouette Method. Οπτικοποίηση της Σχετικής Απόκλισης κάθε Ομάδας από το Γενικό Μέσο Όρο (0). Οι VIPs εντοπίζονται στο Cluster 1.

- **Αντιστοίχιση:** Παρατηρείται μετατόπιση των δεικτών (π.χ. το **Cluster 1** της **Elbow Method** αντιστοιχεί στο **Cluster 2** της **Silhouette Method** κ.ο.κ.), χωρίς όμως να αλλοιώνεται η εσωτερική δομή των χαρακτηριστικών. Με άλλα λόγια παρατηρείται η ίδια ακριβώς χρωματική δομή με την **Εικόνα 5.6.1.c** με μόνη διαφορά την αρίθμηση των Clusters. Πιο συγκεκριμένα το **Cluster 1** από την **Elbow Method** "τοποθετήθηκε" στην **Silhouette Method** στο **Cluster 2**, το **Cluster 2** από την προηγούμενη μέθοδο, εδώ το βλέπουμε στο **Cluster 3**, το **Cluster 3** βρίσκεται πλέον στο **Cluster 5**, το **Cluster 4** παραμένει στην ίδια θέση και το **Cluster 5** είναι στην ουσία το **Cluster 1** όπως είπαμε και προηγουμένως
- **Συμπέρασμα:** Η **γεωμετρία** των **ομάδων** είναι **συμπαγής** και ανεπηρέαστη από την **αρχικοποίηση**.

2. Αριθμητική Επιβεβαίωση Ακρίβειας (Table Verification):

Η απόδειξη της **Στιβαρότητας (Robustness)** του μοντέλου παρέχεται από τα απόλυτα νούμερα του **Πίνακα** της **Εικόνας 5.6.1.f**, στα οποία υπάρχει απόλυτη ταύτιση στη κρίσιμη ομάδα των VIPs (449 Πελάτες) και αμελητέα απόκλιση στις υπόλοιπες (<0.1% καθότι στην 1η εκτέλεση (Elbow), οι "Lost" (Cluster 3) είναι **1.683,** ενώ στη 2η εκτέλεση (Silhouette) οι "Lost" (Cluster 5) είναι **1.681**):

Συγκεντρωτικός Πίνακας Αποτελεσμάτων K-Means
με τη μέθοδο Silhouette Method για K=5 | Seed: 62650

	Count	Percentage (%)	Recency	Frequency	Monetary
Cluster 1	449.0	7.64	18.23	33.31	21586.46
Cluster 2	1346.0	22.9	282.03	3.99	1486.73
Cluster 3	1131.0	19.24	32.54	2.49	681.27
Cluster 4	1271.0	21.62	45.71	9.2	3786.77
Cluster 5	1681.0	28.6	416.86	1.27	277.49
Total	**5878.0**	**100.0**			
Global Avg			**201.33**	**6.29**	**3018.62**

Total: Σύνολα Πληθυσμού || Global Avg: Συνολικοί Μέσοι Όροι

Execution Time: 2.5018 sec

Εικόνα 5.6.1.f. Συγκεντρωτικός Πίνακας K-Means (Silhouette Method). Κατανομή πελατών και Μέσες Τιμές ανά Ομάδα.

- **Απόλυτη Ταύτιση Πληθυσμού:** Η κρίσιμη ομάδα των Champions (VIPs) αποτελείται και πάλι από **ακριβώς 449 πελάτες**. Το γεγονός ότι ο αλγόριθμος απομόνωσε τα ίδια ακριβώς άτομα σε δύο ανεξάρτητες εκτελέσεις αποδεικνύει ότι η ομάδα αυτή είναι διακριτή και σαφώς ορισμένη.
- **Σταθερότητα Μέσων Τιμών:** Οι οικονομικοί δείκτες παραμένουν πρακτικά αμετάβλητοι (Μέσο Monetary **21.586€**), επιβεβαιώνοντας ότι δεν υπήρξε καμία «διαρροή» πελατών προς γειτονικές ομάδες.

- **Αναπαραγωγή Αδράνειας:** Αντίστοιχη σταθερότητα εμφανίζει και η μάζα των Lost πελατών (Cluster 5 σε αυτή την εκτέλεση), διατηρώντας το ίδιο πλήθος (1.681) και τα ίδια χαρακτηριστικά Recency (~416 ημέρες).

Τελικό Συμπέρασμα K-Means:

Από το **Θερμικό Χάρτη** της **Silhouette** (**Εικόνα 5.6.1.e**) διαπιστώνουμε ότι η δομή των **Ομάδων** παραμένει αμετάβλητη, ενώ με τον **Συγκεντρωτικό Πίνακα** της ίδιας **μεθόδου** (**Εικόνα 5.6.1.f**), επιβεβαιώνεται η απόλυτη **σταθερότητα** του αριθμού των μελών ανά **Συστάδα** (π.χ. 449 VIPs). Η σύμπτωση των αποτελεσμάτων μεταξύ των δύο ανεξάρτητων εκτελέσεων αποδεικνύει την **στιβαρότητα** και την **αξιοπιστία (robustness)** της λύσης. Οι **πέντε** (**5**) **ομάδες** είναι **καλά ορισμένες** και δεν αποτελούν προϊόν τύχης. Επομένως η σύμπτωση των αποτελεσμάτων (Convergence Consistency) μεταξύ των δύο μεθόδων αποδεικνύει ότι η λύση των **k=5 συστάδων** αποτελεί το **Ολικό Βέλτιστο (Global Optimum)** για το συγκεκριμένο σύνολο δεδομένων. Οι επιχειρηματικές κατηγορίες που προέκυψαν είναι σταθερές, μετρήσιμες και έτοιμες για στρατηγική αξιοποίηση.

5.6.2. Εκτέλεση Αλγορίθμου K-Means++ (k=5): Βελτιστοποιημένη Αρχικοποίηση & Συγκριτική Ανάλυση

Ως δεύτερο στάδιο της ανάλυσης, εφαρμόστηκε η παραλλαγή **K-Means++**, η οποία βελτιώνει τον τυπικό **αλγόριθμο** μέσω της **σταθμισμένης** (weighted) **επιλογής** των **αρχικών κεντροειδών**. Η μέθοδος αυτή στοχεύει στη **μεγιστοποίηση** των **αποστάσεων** μεταξύ των **αρχικών κέντρων**, εξασφαλίζοντας ταχύτερη σύγκλιση και αποφυγή **υπο-βέλτιστων τοπικών ελαχίστων**.

Για τη διασφάλιση της **ορθής σύγκρισης**, ακολουθήθηκε η ίδια **μεθοδολογία διπλής εκτέλεσης** (Double Run Verification με **Elbow & Silhouette** Methods) με διαφορετικά **Random Seeds**.

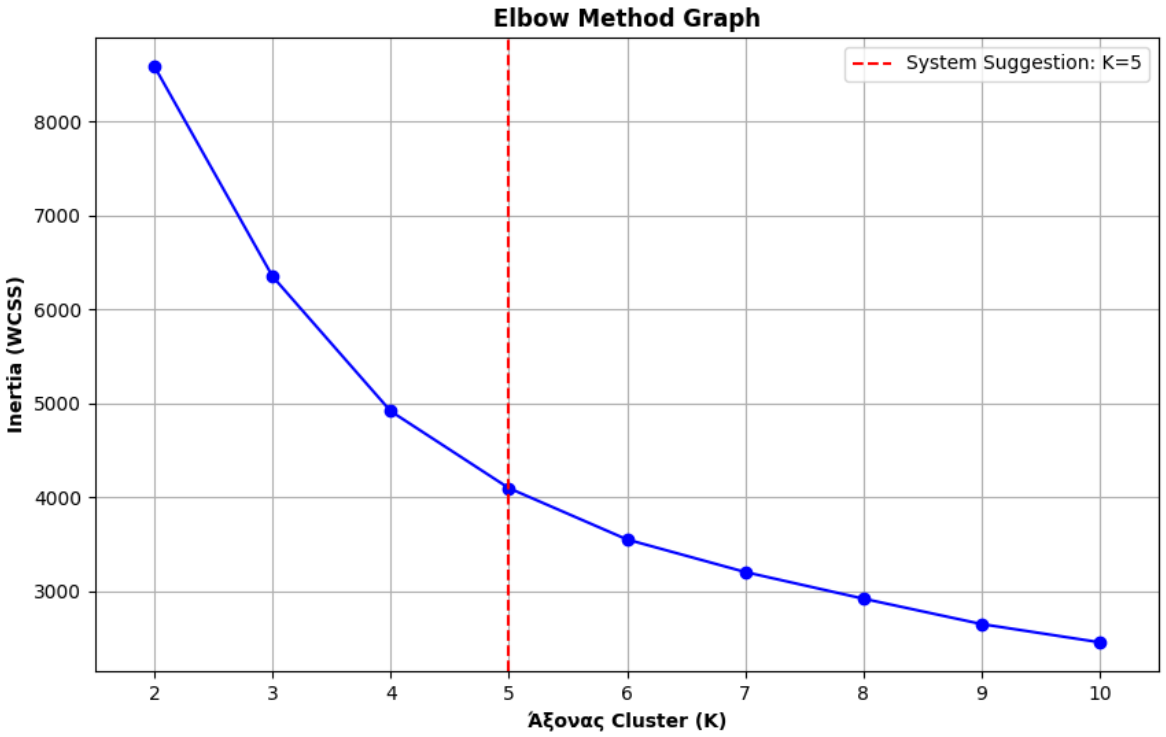
A. Διερεύνηση Βέλτιστου Αριθμού Συστάδων (Optimization Graphs)

Η συμπεριφορά του αλγορίθμου επιβεβαιώνει πλήρως την επιλογή των **πέντε** (**5**) **συστάδων**, παρουσιάζοντας μάλιστα πιο ξεκάθαρα **γεωμετρικά χαρακτηριστικά**.

1. Μέθοδος Elbow (Ανάλυση Αδράνειας):

Στο Διάγραμμα της **Εικόνας 5.6.2.a**, η **καμπύλη** της **Αδράνειας** σχηματίζει τον χαρακτηριστικό "αγκώνα" στο **k=5**.

- **Παρατήρηση:** Η καμπύλη είναι πιο ομαλή σε σχέση με τον απλό **K-Means**, γεγονός που οφείλεται στην καλύτερη αρχικοποίηση. Μετά το **k=5**, η μείωση του σφάλματος είναι **οριακή** (diminishing returns), καθιστώντας το **πέντε** (**5**) ως την **οικονομικότερη λύση πολυπλοκότητας**. Επομένως είναι σαφής η ένδειξη για **k=5.**

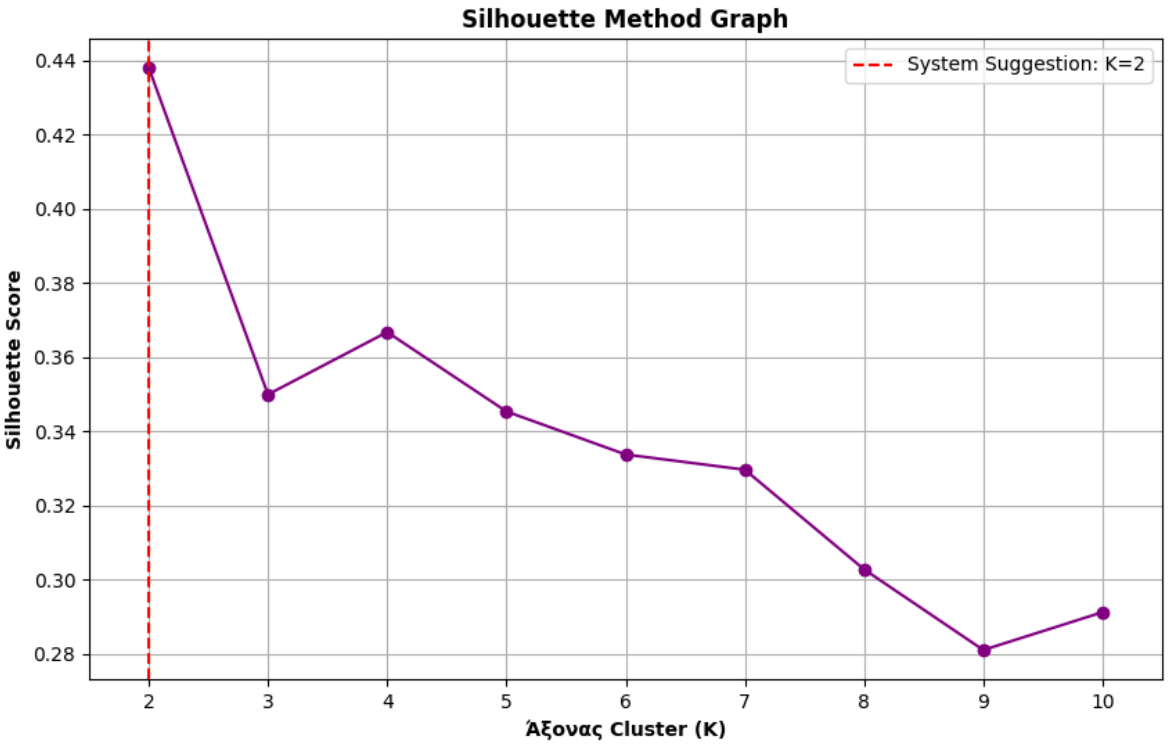


Εικόνα 5.6.2.a. Διάγραμμα Elbow Method K-Means++. Η κόκκινη διακεκομμένη γραμμή επιβεβαιώνει την επιλογή μας για k = 5.

2. Μέθοδος Silhouette (Ανάλυση Συνοχής):

Στο Διάγραμμα της **Εικόνας 5.6.2.b**, η τάση είναι παρόμοια με την προηγούμενη ανάλυση.

- **Το Δίλημμα:** Το **k=2** εμφανίζει το μαθηματικό μέγιστο (λόγω του χάσματος Ενεργών-Αδρανών).
- **Η Λύση:** Στο **k=5**, ο δείκτης παραμένει σε υψηλά επίπεδα σταθερότητας και εξασφαλίζει την **απαραίτητη ανάλυση** (granularity) διατηρώντας **ικανοποιητική συνοχή**. Η επιλογή αυτή είναι η μόνη που εξυπηρετεί το **επιχειρηματικό μοντέλο** των **πέντε** (**5**) **Personas**, καθώς το **k=2** θα συγχώνευε τους **VIPs** με τους απλούς **πελάτες**, χάνοντας κρίσιμη **πληροφορία**.



Εικόνα 5.6.2.b. Διάγραμμα Silhouette Method K-Means++. Η κόκκινη διακεκομμένη γραμμή δείχνει το καλύτερο score που είναι για k = 2.

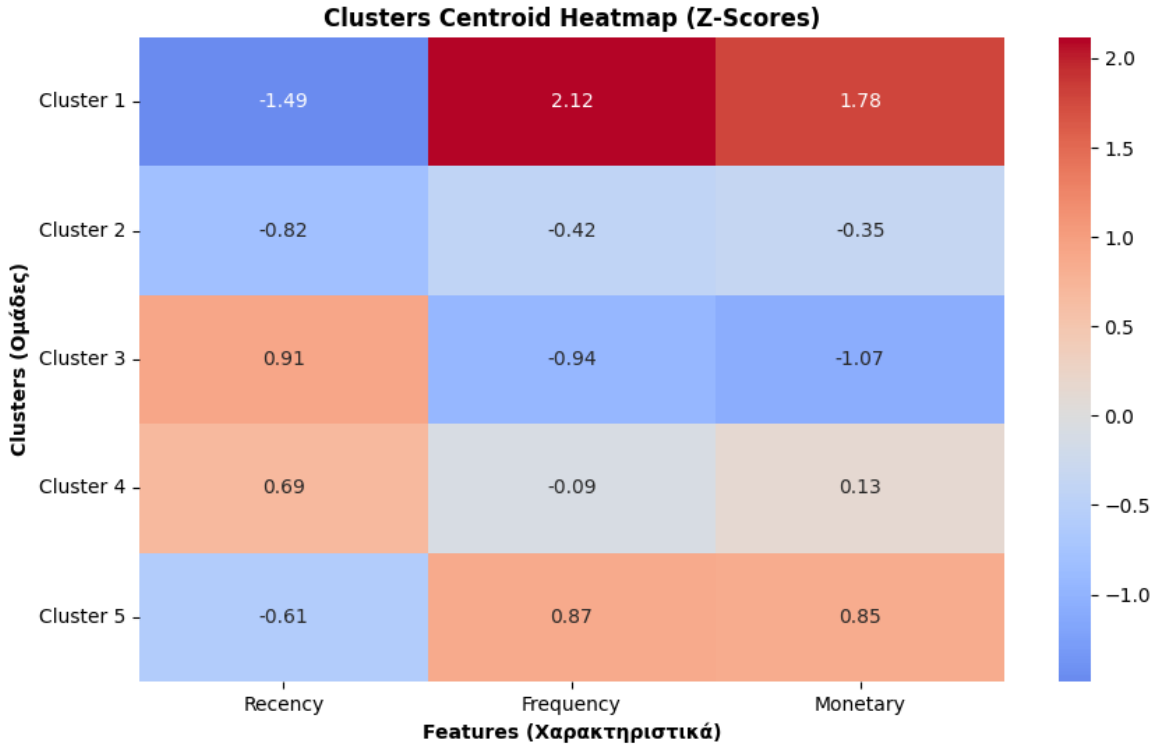
Β. Κύρια Ανάλυση Αποτελεσμάτων (Βάσει Elbow Method / Seed: 78463)

Στην πρώτη εκτέλεση, ο αλγόριθμος ταξινόμησε τα δεδομένα παράγοντας αποτελέσματα που διαφοροποιούνται ελαφρώς από τον απλό **K-Means**, κυρίως ως προς το εύρος της κατηγορίας **VIP**.

1. Προφίλ Συστάδων (Heatmap Analysis):

Ο **Θερμικός Χάρτης** της **Εικόνας 5.6.2.c** αποκαλύπτει την ταυτότητα των ομάδων μέσω των **Z-Scores**:

- **Cluster 1 (The Champions):** Ξεχωρίζει έντονα με βαθύ κόκκινο σε Frequency (+2.12) και **Monetary** (+1.78), ενώ το **Recency** είναι αρνητικό (-1.49). Είναι η πιο δυναμική ομάδα.
- **Cluster 3 (Lost):** Εμφανίζει το αντίθετο μοτίβο, με θετικό **Recency** (+0.91) και αρνητικές τιμές στα αγοραστικά κριτήρια.
- **Cluster 5 (Loyal):** Παρουσιάζει ισορροπημένη θετική συμπεριφορά (F: +0.87, M: +0.85) και χαμηλό **Recency** (-0.61), αποτελώντας τον κορμό των πιστών πελατών.



Εικόνα 5.6.2.c. Θερμικός Χάρτης (Heatmap) Z-Scores (Standardized Values) Elbow Method. Οπτικοποίηση της Σχετικής Απόκλισης κάθε Ομάδας από το Γενικό Μέσο Όρο (0). Οι VIPs εντοπίζονται στο Cluster 1.

2. Ποσοτικά Δεδομένα (Table Analysis):

Τα στοιχεία του **Πίνακα** της **Εικόνας 5.6.2.d** αναδεικνύουν μια ενδιαφέρουσα εξέλιξη σε σχέση με τον απλό **K-Means**:

- **Διευρυμένη Ελίτ (Cluster 1):** Ο αλγόριθμος εντόπισε **482 VIP πελάτες** (ποσοστό 8.2%), έναντι 449 που είχε βρει ο απλός **K-Means**. Η μέση δαπάνη τους παραμένει εντυπωσιακή (**20.373€**). Αυτό υποδηλώνει ότι ο **K-Means++** είναι πιο ευαίσθητος στον εντοπισμό πελατών υψηλής αξίας.
- **Όγκος Αδράνειας (Cluster 3):** Η ομάδα των **Lost** περιλαμβάνει **1.614** πελάτες (27.46%) με μέσο χρόνο απουσίας **414 ημέρες**.

Συγκεντρωτικός Πίνακας Αποτελεσμάτων K-Means++
με τη μέθοδο Elbow Method για K=5 | Seed: 78463

	Count	Percentage (%)	Recency	Frequency	Monetary
Cluster 1	482.0	8.2	15.78	31.8	20373.07
Cluster 2	1146.0	19.5	31.76	2.57	711.13
Cluster 3	1614.0	27.46	414.1	1.24	259.35
Cluster 4	1369.0	23.29	295.86	3.73	1393.27
Cluster 5	1267.0	21.55	52.13	9.14	3774.8
Total	**5878.0**	**100.0**			
Global Avg			**201.33**	**6.29**	**3018.62**

Total: Σύνολα Πληθυσμού || Global Avg: Συνολικοί Μέσοι Όροι
Execution Time: 0.0592 sec

Εικόνα 5.6.2.d. Συγκεντρωτικός Πίνακας K-Means++ (Elbow Method). Παρατηρείται αύξηση του πλήθους των VIPs (482) σε σχέση με την Απλή Μέθοδο (449)

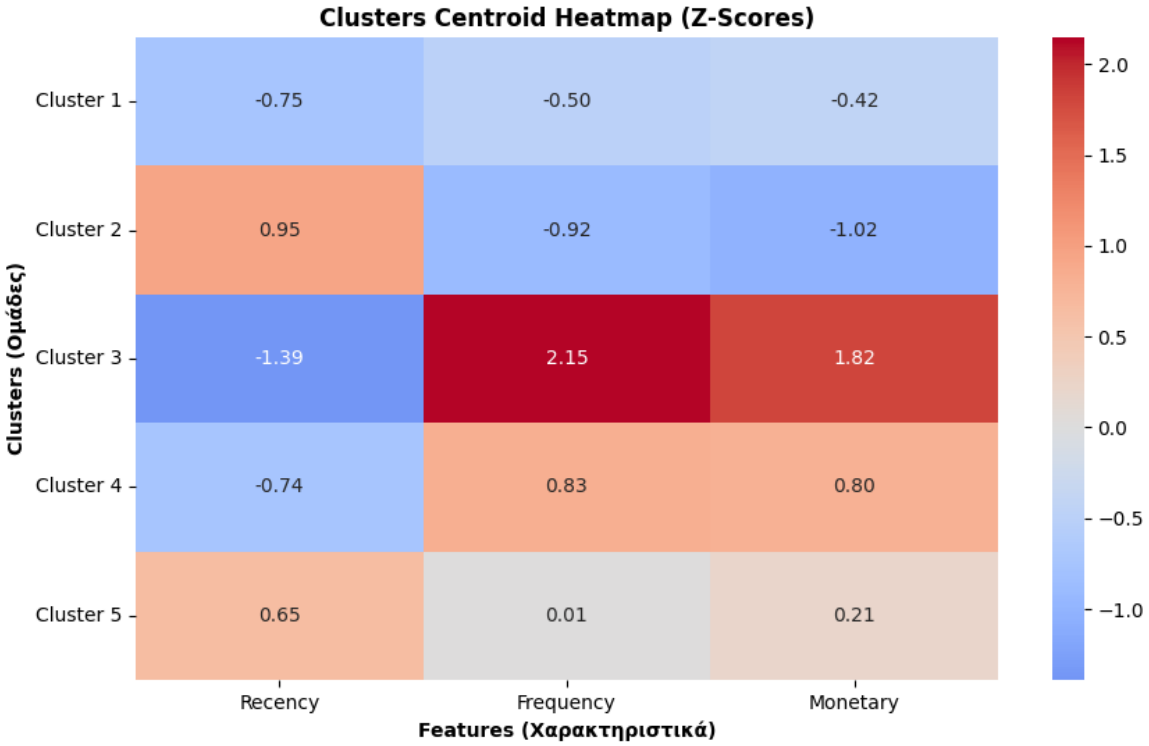
Γ. Επαλήθευση Σταθερότητας (Βάσει Silhouette Method / Seed: 53085)

Η δεύτερη εκτέλεση με διαφορετικό seed έρχεται να επιβεβαιώσει με εντυπωσιακό τρόπο την ανωτερότητα του **K-Means++** όσον αφορά τη σταθερότητα.

1. Φαινόμενο Label Switching:

Συγκρίνοντας τον **Χάρτη** της **Εικόνας 5.6.2.e** με τον προηγούμενο, βλέπουμε την αναμενόμενη αναδιάταξη:

- Η ομάδα των **Champions** (κόκκινο F/M) μετακινήθηκε από το Cluster 1 στο **Cluster 3**.
- Η **χρωματική ένταση** και η **δομή** των **χαρακτηριστικών** παραμένει **πανομοιότυπη** σχεδόν **αναλλοίωτη**, παρά την αλλαγή ετικετών, επιβεβαιώνοντας τη **γεωμετρική συνοχή**.



Εικόνα 5.6.2.e. Θερμικός Χάρτης (Heatmap) Z-Scores (Standardized Values) Silhouette Method. Οπτικοποίηση της Σχετικής Απόκλισης κάθε Ομάδας από το Γενικό Μέσο Όρο (0). Οι VIPs εντοπίζονται στο Cluster 3.

2. Αριθμητική Απόδειξη Στιβαρότητας:

Ο Πίνακας της **Εικόνας 5.6.2.f** παρέχει το πιο ισχυρό τεκμήριο της ανάλυσης:

- **Απόλυτη Ταύτιση VIPs:** Στο Cluster 3 (Champions) καταμετρήθηκαν **ακριβώς 482 πελάτες**, νούμερο που ταυτίζεται απόλυτα με την πρώτη εκτέλεση. Το γεγονός ότι και οι δύο εκτελέσεις του **K-Means++** κατέληξαν στον ίδιο ακριβώς αριθμό κορυφαίων πελατών αποδεικνύει την **εξαιρετική ευστάθεια** της **μεθόδου**.
- **Συνοχή Μέσων Τιμών:** Η μέση δαπάνη (20.902€) και η συχνότητα (32.24) παραμένουν σταθερές, επιβεβαιώνοντας ότι ο πυρήνας των πελατών δεν διασπάται.

Συγκεντρωτικός Πίνακας Αποτελεσμάτων K-Means++
με τη μέθοδο Silhouette Method για K=5 | Seed: 53085

	Count	Percentage (%)	Recency	Frequency	Monetary
Cluster 1	1115.0	18.97	35.11	2.33	642.61
Cluster 2	1679.0	28.56	423.04	1.29	284.16
Cluster 3	482.0	8.2	22.13	32.24	20902.3
Cluster 4	1273.0	21.66	40.5	8.78	3421.99
Cluster 5	1329.0	22.61	279.74	4.13	1594.22
Total	**5878.0**	**100.0**			
Global Avg			**201.33**	**6.29**	**3018.62**

Total: Σύνολα Πληθυσμού || Global Avg: Συνολικοί Μέσοι Όροι
Execution Time: 2.2177 sec

Εικόνα 5.6.2.f. Συγκεντρωτικός Πίνακας K-Means++ (Silhouette Method). Η απόλυτη ταύτιση στο πλήθος των VIPs (482) με την Elbow Method, αποδεικνύει την αξιοπιστία του Αλγορίθμου.

Τελικό Συμπέρασμα K-Means++:

Η **συγκριτική ανάλυση** μεταξύ του απλού **K-Means** και του **K-Means++** ανέδειξε την σαφή υπεροχή του δεύτερου, τόσο σε τεχνικό όσο και σε επιχειρηματικό επίπεδο. Η εφαρμογή της **σταθμισμένης αρχικοποίησης** οδήγησε σε **τρία** (**3**) κρίσιμα **συμπεράσματα**:

1. **Ανίχνευση "Λανθάνουσας" Αξίας:** Ο **K-Means++** αποδείχθηκε πιο ευαίσθητος και ακριβής στον εντοπισμό των πελατών υψηλής αξίας. Συγκεκριμένα, ενέταξε στην κατηγορία **"Champions" 482 πελάτες**, έναντι **449** του απλού **K-Means**. Αυτοί οι **33 επιπλέον πελάτες**, που ο απλός αλγόριθμος πιθανώς "έχασε" ή υποβάθμισε σε

χαμηλότερες κατηγορίες, αντιπροσωπεύουν σημαντική ανεκμετάλλευτη ρευστότητα για την **επιχείρηση**.

2. **Εξάλειψη του Παράγοντα Τύχης:** Η απόλυτη ταύτιση των αποτελεσμάτων (0% απόκλιση στο πλήθος των VIPs) μεταξύ των δύο ανεξάρτητων εκτελέσεων (Seeds 78463 & 53085) αποδεικνύει την **ανθεκτικότητα (robustness)** του μοντέλου. Σε αντίθεση με τον απλό **K-Means**, όπου η λύση μπορεί να εξαρτάται από την αρχική τυχαία τοποθέτηση, ο **K-Means++** εγγυάται ότι η στρατηγική **Marketing** θα βασίζεται σε δεδομένα που είναι επαναλήψιμα και αντικειμενικά.
3. **Επιχειρηματική Εμπιστοσύνη:** Η **σταθερότητα** των μέσων όρων (Average Monetary ~20.900€) επιβεβαιώνει ότι οι ομάδες είναι συμπαγείς και καλά **διαχωρισμένες**. Συνεπώς, ο K-Means++ κρίνεται ως ο πλέον κατάλληλος **αλγόριθμος** για την τελική χάραξη της τμηματοποίησης, προσφέροντας τη μέγιστη δυνατή ασφάλεια στη λήψη αποφάσεων.

5.6.3. Εκτέλεση Αλγορίθμου K-Medoids (PAM) με (k=5): Ανθεκτικότητα σε Ακραίες Τιμές & Υπολογιστικό Κόστος

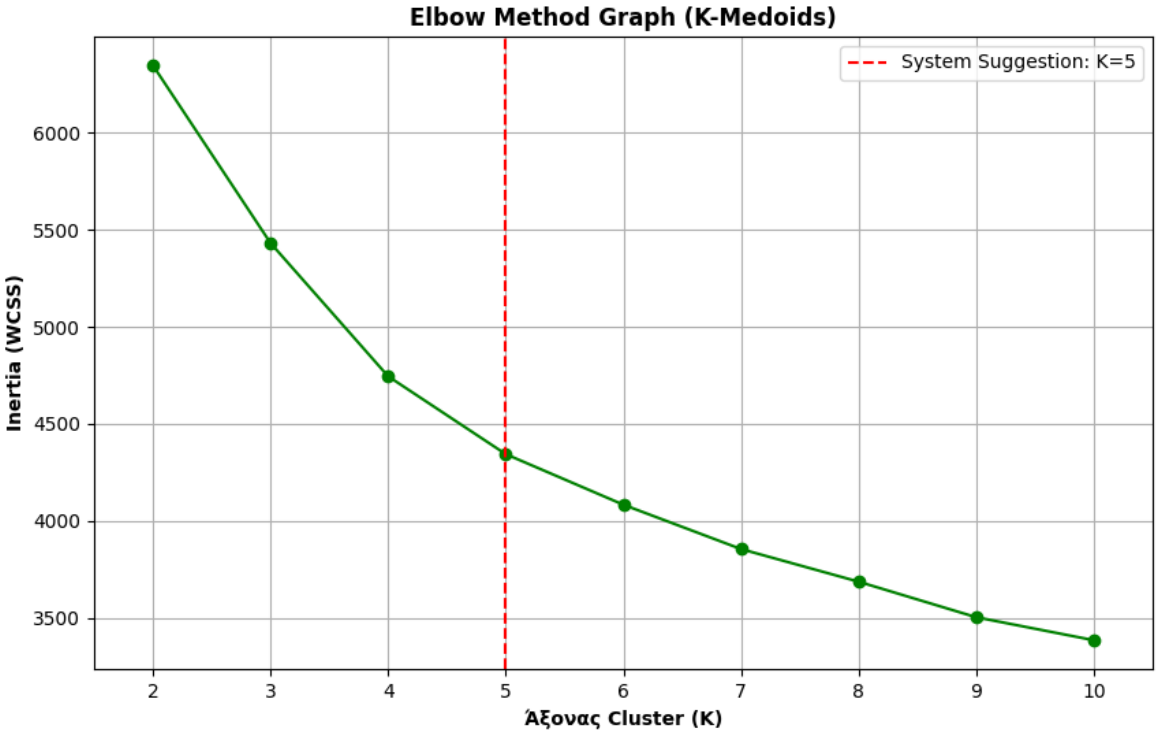
Στο τρίτο στάδιο της διαδικασίας συσταδοποίησης, εφαρμόστηκε ο αλγόριθμος **K-Medoids (Partitioning Around Medoids)**. Η θεμελιώδης διαφορά του από τον **K-Means** έγκειται στο γεγονός ότι ως **κέντρα** των **συστάδων** δεν χρησιμοποιεί πλασματικούς μέσους όρους (means), αλλά **υπαρκτά σημεία του συνόλου δεδομένων (medoids)**.

Η προσέγγιση αυτή προσφέρει **δύο** (**2**) **θεωρητικά πλεονεκτήματα** και ένα μειονέκτημα, τα οποία εξετάστηκαν στην πράξη:

1. **Robustness:** Μεγαλύτερη ανθεκτικότητα σε **ακραίες τιμές** (outliers), καθώς το **κέντρο** δεν "τραβιέται" από μεμονωμένες μεγάλες **εγγραφές**.
2. **Interpretability:** Το **κέντρο** της **ομάδας** είναι πραγματικός **πελάτης**.
3. **Cost:** Σημαντικά αυξημένο **υπολογιστικό κόστος**.

Α. Διερεύνηση Βέλτιστου Αριθμού Συστάδων (Optimization Graphs)

Η **γεωμετρική ανάλυση** επιβεβαιώνει τη δομή των **πέντε** (**5**) **συστάδων**, αν και με ελαφρώς διαφορετική συμπεριφορά στις **καμπύλες** σε σχέση με τον **K-Means**.



Εικόνα 5.6.3.α. Διάγραμμα Elbow Method K-Medoids. Το Σημείο Καμπής (κόκκινη διακεκομμένη γραμμή) στο k = 5 παραμένει η βέλτιστη επιλογή μας

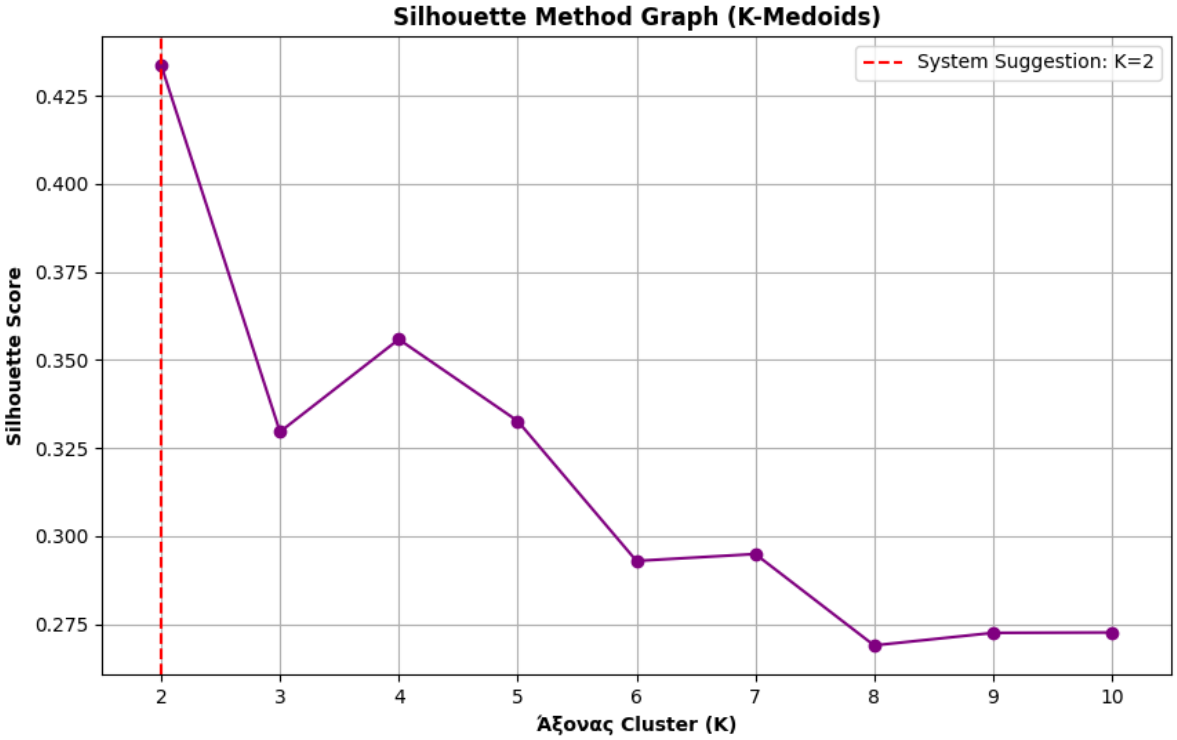
1. Μέθοδος Elbow (Ανάλυση Αδράνειας):

Στο Διάγραμμα της **Εικόνας 5.6.3.α**, παρατηρούμε ότι η καμπύλη σχηματίζει τον "αγκώνα" στο **k=5**.

- **Παρατήρηση:** Η μείωση της αδράνειας είναι πιο ομαλή, γεγονός που υποδηλώνει ότι ο αλγόριθμος δυσκολεύεται περισσότερο να βρει "τέλεια" σφαιρικά σχήματα, καθώς δεσμεύεται να χρησιμοποιεί μόνο πραγματικά σημεία.

2. Μέθοδος Silhouette (Ανάλυση Συνοχής):

Στο Διάγραμμα της **Εικόνας 5.6.3.b**, το σκορ στο **k=5** (περίπου 0.33) είναι ελαφρώς χαμηλότερο από αυτό του **K-Means** (0.34).



Εικόνα 5.6.3.b. Διάγραμμα Silhouette Method K-Medoids. Η κόκκινη διακεκομμένη γραμμή δείχνει το καλύτερο score που είναι για k = 2.

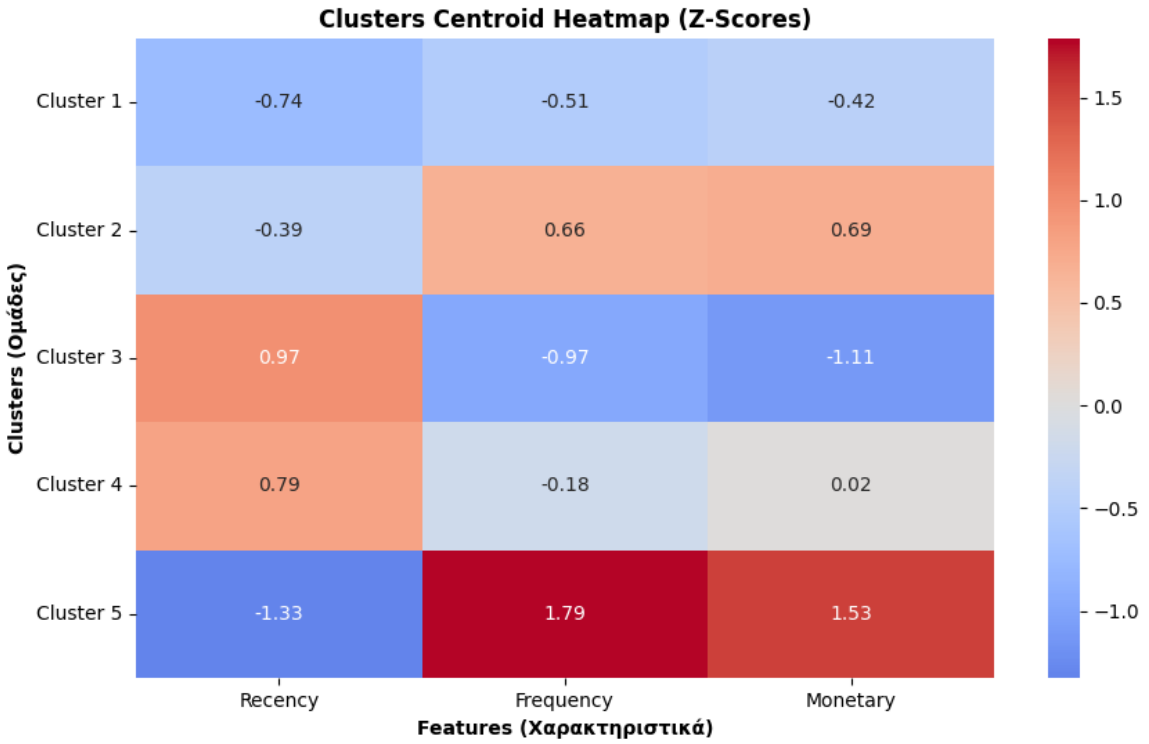
- **Ερμηνεία:** Αυτό είναι αναμενόμενο, καθώς ο **K-Medoids** θυσιάζει τη **μαθηματική βελτιστοποίηση** των **αποστάσεων** προς όφελος της ρεαλιστικής αναπαράστασης μέσω υπαρκτών **κεντροειδών**. Η επιλογή **k=5** προσφέρει **ικανοποιητική διακριτότητα** (Separation)

Β. Κύρια Ανάλυση Αποτελεσμάτων (Βάσει Elbow Method / Seed: 20024)

Η πρώτη εκτέλεση (**Elbow Method**) ανέδειξε μια ενδιαφέρουσα διαφοροποίηση στην εικόνα της αγοράς σε σχέση με τους προηγούμενους αλγορίθμους.

1. Προφίλ Συστάδων (Heatmap Analysis):

Στον **Θερμικό Χάρτη** της **Εικόνας 5.6.3.c**:



Εικόνα 5.6.3.c. Θερμικός Χάρτης (Heatmap) Z-Scores (Standardized Values) Elbow Method. Οπτικοποίηση της Σχετικής Απόκλισης κάθε Ομάδας από το Γενικό Μέσο Όρο (0). Οι VIPs εντοπίζονται στο Cluster 5.

- **Cluster 5 (The Champions):** Εντοπίζονται με **βαθύ κόκκινο** στο Frequency (+1.79) και **Monetary** (+1.53), ενώ το **Recency** είναι βαθύ μπλε (-1.33).
- **Διαφορά Έντασης:** Παρατηρούμε ότι τα **Z-Scores** είναι ελαφρώς χαμηλότερα από αυτά του **K-Means** (που έφτανε το +2.19). Αυτό συμβαίνει γιατί το "Medoid" (ο αντιπροσωπευτικός πελάτης) είναι λιγότερο ακραίος από τον μαθηματικό μέσο όρο.

2. Ποσοτικά Δεδομένα (Table Analysis):

Ο Πίνακας της **Εικόνας 5.6.3.d** αποκαλύπτει σημαντικά ευρήματα:

- **Διευρυμένη Βάση VIPs:** Ο αλγόριθμος κατέταξε **778 πελάτες (13.24%)** στην κατηγορία Cluster 5. Αυτό είναι σχεδόν διπλάσιο ποσοστό από τον K-Means (7.6%).
- **Πιο "Γήινη" Μέση Δαπάνη:** Η μέση δαπάνη της ομάδας VIP είναι **14.692€**, σημαντικά χαμηλότερη από τα ~21.000€ του K-Means.

Συγκεντρωτικός Πίνακας Αποτελεσμάτων K-Medoids (PAM)
με τη μέθοδο Elbow Method για K=5 | Seed: 20024

	Count	Percentage (%)	Recency	Frequency	Monetary
Cluster 1	1164.0	19.8	37.0	2.32	646.48
Cluster 2	1246.0	21.2	70.32	7.5	2992.67
Cluster 3	1429.0	24.31	434.14	1.17	242.29
Cluster 4	1261.0	21.45	330.03	3.33	1177.77
Cluster 5	778.0	13.24	20.81	24.5	14692.35
Total	**5878.0**	**100.0**			
Global Avg			**201.33**	**6.29**	**3018.62**

Total: Σύνολα Πληθυσμού || Global Avg: Συνολικοί Μέσοι Όροι

Execution Time: 4 min 51.85 sec

Εικόνα 5.6.3.d. Συγκεντρωτικός Πίνακας K-Medoids (Elbow Method). Παρατηρείται αύξηση του πλήθους των VIPs (778) με χαμηλότερο Μέσο Όρο Δαπάνης

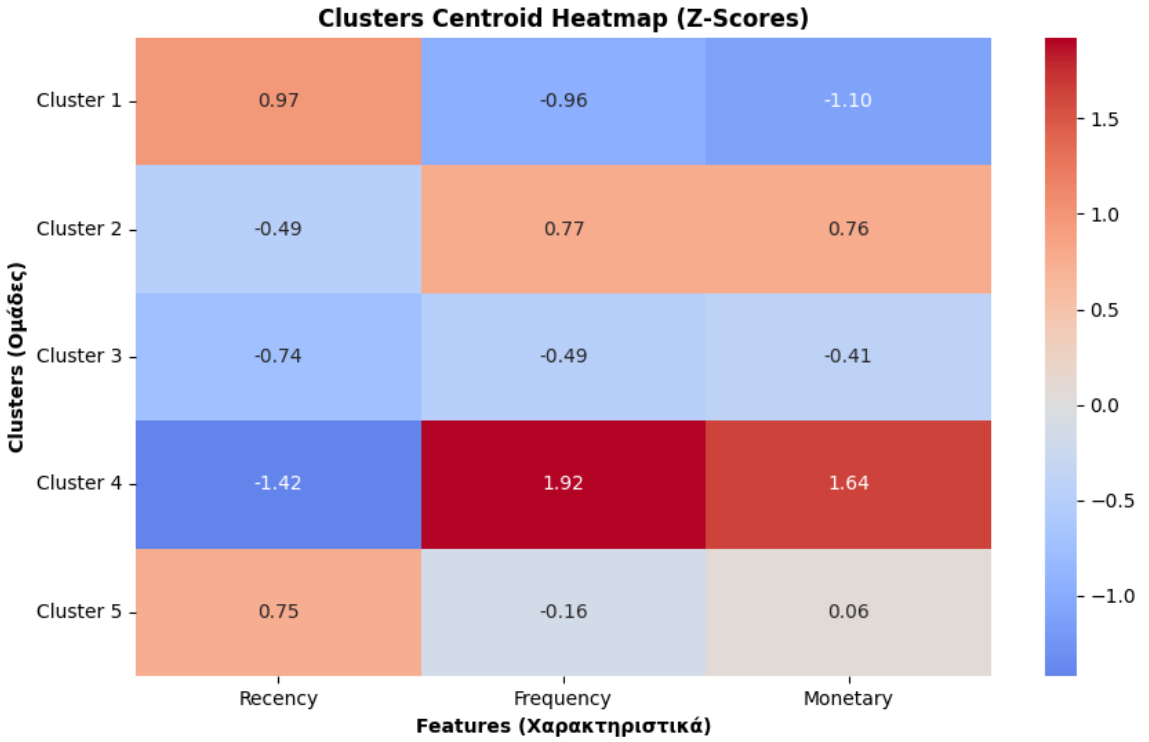
- **Ερμηνεία:** Ο **K-Medoids** δεν επηρεάστηκε από τους λίγους "υπερ-πλούσιους" πελάτες (outliers) που τραβούσαν τον μέσο όρο ψηλά στον **K-Means**. Αντ' αυτού, επέλεξε έναν πιο "τυπικό" καλό πελάτη ως κέντρο, συμπεριλαμβάνοντας έτσι περισσότερα άτομα στην ομάδα των VIPs.
- **Χρόνος Εκτέλεσης:** Παρατηρούμε ότι ο χρόνος εκτέλεσης εκτοξεύτηκε στα **4 λεπτά και 51 δευτερόλεπτα**, επιβεβαιώνοντας την υψηλή υπολογιστική πολυπλοκότητα.

Γ. Επαλήθευση Σταθερότητας (Βάσει Silhouette Method / Seed: 9881)

Η δεύτερη εκτέλεση (**Silhouette Method**) ανέδειξε μια ευαισθησία του αλγορίθμου στην αρχικοποίηση, η οποία πρέπει να ληφθεί σοβαρά υπόψη.

1. Φαινόμενο Label Switching & Μεταβολές:

Στον **Χάρτη** της **Εικόνας 5.6.3.e**, η ομάδα των **VIPs** μετακινήθηκε στο **Cluster 4**. Ωστόσο, η **χρωματική δομή** παραμένει **συνεπής**.



Εικόνα 5.6.3.e. Θερμικός Χάρτης (Heatmap) Z-Scores (Standardized Values) Silhouette Method. Οπτικοποίηση της Σχετικής Απόκλισης κάθε Ομάδας από το Γενικό Μέσο Όρο (0). Οι VIPs εντοπίζονται στο Cluster 4.

2. Αριθμητική Σύγκριση:

Ο **Πίνακας** της **Εικόνας 5.6.3.f** δείχνει μια αξιοσημείωτη απόκλιση σε σχέση με την πρώτη εκτέλεση:

- **Μείωση Πληθυσμού VIPs:** Στο **Cluster 4** καταμετρήθηκαν **627 πελάτες** (έναντι 778 προηγουμένως).
- **Αύξηση Μέσης Δαπάνης:** Η μέση **δαπάνη** αυξήθηκε στα **16.946€** (έναντι 14.692€).
- **Συμπέρασμα Σταθερότητας:** Σε αντίθεση με την απόλυτη σταθερότητα του K-Means++, ο K-Medoids εμφάνισε μεταβλητότητα της τάξης του **20%** στο μέγεθος της κρίσιμης ομάδας. Αυτό υποδηλώνει ότι η επιλογή του αρχικού **Medoid** επηρεάζει σημαντικά τα τελικά όρια των **συστάδων**.

Συγκεντρωτικός Πίνακας Αποτελεσμάτων K-Medoids (PAM)
με τη μέθοδο Silhouette Method για K=5 | Seed: 9881

	Count	Percentage (%)	Recency	Frequency	Monetary
Cluster 1	1458.0	24.8	434.34	1.19	246.27
Cluster 2	1289.0	21.93	63.46	8.32	3362.07
Cluster 3	1191.0	20.26	36.56	2.37	663.4
Cluster 4	627.0	10.67	16.7	27.41	16946.8
Cluster 5	1313.0	22.34	315.58	3.43	1245.17
Total	**5878.0**	**100.0**			
Global Avg			**201.33**	**6.29**	**3018.62**

Total: Σύνολα Πληθυσμού || Global Avg: Συνολικοί Μέσοι Όροι
Execution Time: 3 min 34.24 sec

Εικόνα 5.6.3.f. Συγκεντρωτικός Πίνακας K-Medoids (Silhouette Method). Η απόκλιση στο Μέγεθος της Ομάδας VIPs (627 vs 778) αναδεικνύει την Ευαισθησία του Αλγορίθμου στην Αρχικοποίηση.)

Τελικό Συμπέρασμα K-Medoids:

Η εφαρμογή του **αλγορίθμου K-Medoids** ανέδειξε μια αρκετά πιο διαφορετική και **ρεαλιστική** ταυτόχρονα οπτική της **πελατειακής βάσης** σε σχέση με τις **παραλλαγές** του **K-Means**, προσφέροντας σημαντικά διδάγματα σχετικά με τη **διαχείριση** των **ακραίων τιμών** (outliers), αλλά και αναδεικνύοντας **κρίσιμους περιορισμούς**. Η ανάλυση των αποτελεσμάτων οδηγεί σε **τρία** (**3**) **κομβικά συμπεράσματα**:

1. **Ανθεκτικότητα στις Ακραίες Τιμές (Outlier Robustness) και "Δημοκρατικοποίηση" των VIPs:** Σε αντίθεση με τον **K-Means**, ο οποίος χρησιμοποιεί ως κέντρα πλασματικούς μαθηματικούς μέσους όρους, ο **K-Medoids** χρησιμοποιεί πραγματικούς πελάτες (medoids). Αυτό σημαίνει ότι το κέντρο της ομάδας "Champions" δεν παρασύρθηκε προς τα πάνω από 5-10 πελάτες με ασύλληπτα υψηλό τζίρο (whales). Ως αποτέλεσμα, η κατηγορία των **VIPs** διευρύνθηκε σημαντικά, περιλαμβάνοντας από **627 έως 778 πελάτες** (έναντι 449 του K-Means). Επιχειρηματικά, αυτό είναι εξαιρετικά χρήσιμο: μας δίνει το προφίλ του *"τυπικού εξαιρετικού πελάτη"* (με μέση δαπάνη 14.600€ - 16.900€), παρέχοντας στο **Marketing**

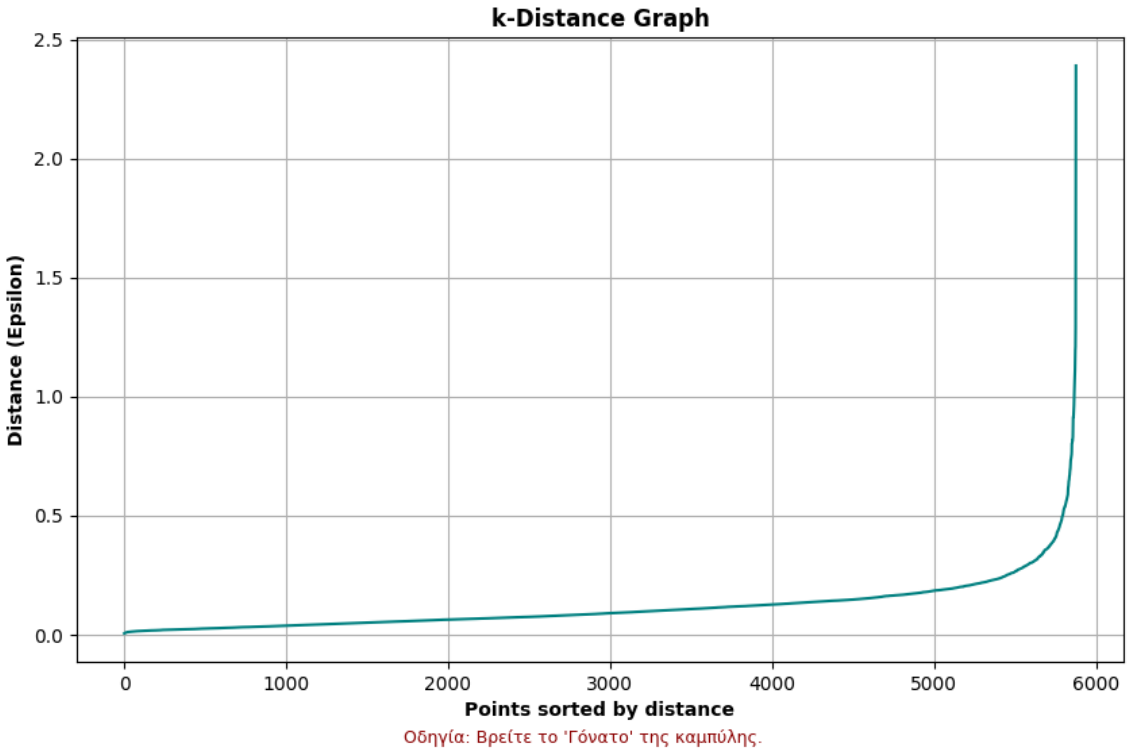
ένα πιο ρεαλιστικό, μετρήσιμο και προσεγγίσιμο **Persona** για **στοχευμένες καμπάνιες**, χωρίς την **αλλοίωση** των **ακραίων τιμών**.

2. **Μειωμένη Σταθερότητα Αναπαραγωγής (Reproducibility Issues):** Παρά την υψηλή ερμηνευτική του αξία, το κυριότερο **μειονέκτημα** του μοντέλου αποδείχθηκε η **ευαισθησία** του στην αρχικοποίηση. Η **διακύμανση** που παρατηρήθηκε μεταξύ των δύο εκτελέσεων (διαφορά σχεδόν 20% στο μέγεθος της ομάδας των **VIPs**, από 778 σε 627) υποδηλώνει μειωμένη **ευστάθεια** (stability). Σε αντίθεση με την απόλυτη ταύτιση του **K-Means++**, ο **K-Medoids** δεν εγγυάται **σταθερά όρια συστάδων**, γεγονός που τον καθιστά λιγότερο **αξιόπιστο** για **αυτοματοποιημένη**, **επαναλαμβανόμενη** χρήση σε **παραγωγικά συστήματα** (production environments).
3. **Υπολογιστική Πολυπλοκότητα (Computational Cost):** Τα αποτελέσματα επιβεβαίωσαν την **αλγοριθμική θεωρία** σχετικά με το **κόστος εκτέλεσης** του **PAM** (Partitioning Around Medoids). Ο απαιτούμενος **χρόνος** άγγιξε τα **~5 λεπτά**, έναντι **ελάχιστων δευτερολέπτων** (ή και κλασμάτων αυτού) που χρειάστηκαν οι παραλλαγές του **K-Means**. Αν και για ένα δείγμα **5.878 εγγραφών** ο χρόνος αυτός είναι ανεκτός, σε περιβάλλοντα **Μεγάλων Δεδομένων** (Big Data Analytics) με εκατοντάδες χιλιάδες πελάτες, ο **αλγόριθμος** θα καθίστατο **απαγορευτικός**.

Συνοψίζοντας: Παρόλο που ο K-Medoids προσέφερε μια πιο ρεαλιστική απεικόνιση του "τυπικού" VIP πελάτη (μέσω των medoids), η ασταθής συμπεριφορά του και το υψηλό υπολογιστικό κόστος τον καθιστούν **λιγότερο προτιμητέα επιλογή** σε σύγκριση με την σταθερότητα και την ταχύτητα του **K-Means++** για τα συγκεκριμένα δεδομένα λιανικής. Αν και αποτελεί ένα εξαιρετικό εργαλείο **ποιοτικής εξερεύνησης (exploratory analysis),** που μας βοήθησε να κατανοήσουμε την **πραγματική δομή** του πελατολογίου αγνοώντας τον **θόρυβο** των **ακραίων τιμών**, ωστόσο, συγκρινόμενος με τον **K-Means++**, υστερεί σημαντικά σε **σταθερότητα** και **ταχύτητα**, στοιχεία που τον καθιστούν δευτερεύουσα **επιλογή** για την τελική, **μόνιμη τμηματοποίηση** της εταιρείας.

5.6.4. Εκτέλεση Αλγορίθμου DBSCAN: Ανίχνευση Πυκνότητας & Απομόνωση Ακραίων Τιμών

Η τέταρτη και τελευταία μέθοδος συσταδοποίησης αφορά τον αλγόριθμο **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**. Σε αντίθεση με τους αλγορίθμους τύπου K-Means/Medoids που προσπαθούν να διαμερίσουν ολόκληρο τον χώρο σε σφαιρικές ομάδες, ο DBSCAN εστιάζει στην πυκνότητα, αφήνοντας ασυσχέτιστα τα σημεία που βρίσκονται σε αραιές περιοχές.



Εικόνα 5.6.4.a. Γράφημα k-Distance. Η Απότομη Κλίση της Καμπύλης υποδεικνύει το όριο μετάβασης από τις Πυκνές Συστάδες στο Θόρυβο

Η εκτέλεση πραγματοποιήθηκε με τις βελτιστοποιημένες παραμέτρους που προέκυψαν από την ανάλυση της **Ενότητας 5.5.2** (**Epsilon=0.3, MinPts=13**).

A. Επιβεβαίωση Παραμέτρων (k-Distance Graph) Η γεωμετρική βάση της εκτέλεσης αποτυπώνεται στο Γράφημα k-αποστάσεων της **Εικόνας 5.6.4.a**.

- **Το Σημείο Καμπής:** Η καμπύλη παρουσιάζει απότομη αύξηση (knee point) σε τιμή απόστασης περίπου 0.3. Αυτό επιβεβαιώνει ότι η τιμή **Epsilon=0.3** είναι η κατάλληλη ακτίνα για να διαχωριστούν οι "φυσιολογικές" γειτονιές από τις περιοχές θορύβου.

Συγκεντρωτικός Πίνακας Αποτελεσμάτων DBSCAN (k-Distance Graph) για Epsilon=0.3 & MinPts=13

	Count	Percentage (%)	Recency	Frequency	Monetary
Noise	440.0	7.49	98.98	23.02	19090.44
Cluster 1	2985.0	50.78	116.31	7.87	2779.65
Cluster 2	1559.0	26.52	363.84	1.0	308.77
Cluster 3	894.0	15.21	252.22	2.0	632.01
Total	**5878.0**	**100.0**			
Global Avg			**201.33**	**6.29**	**3018.62**

Total: Σύνολα Πληθυσμού || Global Avg: Συνολικοί Μέσοι Όροι

Execution Time: 0.0617 sec

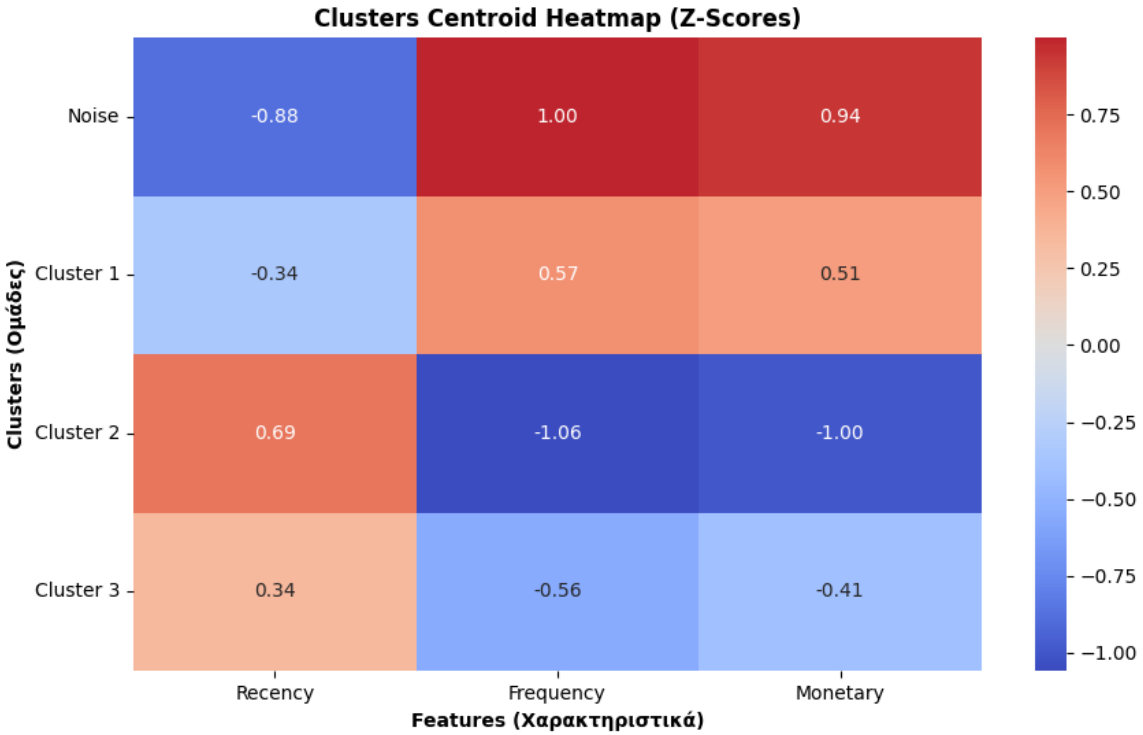
Εικόνα 5.6.4.b. Συγκεντρωτικός Πίνακας DBSCAN. Το εντυπωσιακό στοιχείο είναι η κατηγορία «Noise», η οποία συγκεντρώνει τους πελάτες με τον Υψηλότερο Τζίρο (~190.000€)

Β. Ανάλυση Αποτελεσμάτων: Το Παράδοξο του "Θορύβου" Η εφαρμογή του αλγορίθμου παρήγαγε **τρεις** (3) **κύριες συστάδες** και μία κατηγορία "**Noise**". Η ανάλυση του **Πίνακα Αποτελεσμάτων** (**Εικόνα 5.6.4.b**) αποκαλύπτει το σημαντικότερο εύρημα της μελέτης μας.

1. Ποσοτική Ανάλυση (Table Analysis):

- **Noise ως "Super-VIPs":** Η κατηγορία Noise περιλαμβάνει 440 πελάτες (7.49%). Ενώ συνήθως ο θόρυβος θεωρείται "σκουπίδια", εδώ συμβαίνει το αντίθετο. Η μέση δαπάνη (Monetary) αυτών των πελατών είναι **190.904€**, ένα ποσό εξωπραγματικό σε σχέση με τον γενικό μέσο όρο (3.018€). Ο DBSCAN τους απέρριψε από τις κανονικές ομάδες ακριβώς επειδή η συμπεριφορά τους είναι **μοναδική και ακραία** (Outliers).

- **Cluster 1 (The Mass Market):** Η πολυπληθέστερη ομάδα (2.985 πελάτες, 50.78%) αντιπροσωπεύει τον μέσο, ενεργό πελάτη με πρόσφατες αγορές (Recency 116 ημέρες) και κανονική δαπάνη.
- **Cluster 2 (Lost/Inactive):** Ομάδα 1.559 πελατών με πολύ υψηλό Recency (**363 ημέρες**), που αντιστοιχεί στους αδρανείς πελάτες.



Εικόνα 5.6.4.c. Θερμικός Χάρτης (Heatmap) Z-Scores (Standardized Values) DBSCAN. Η κατηγορία Noise εκπροσωπεί τους VIPs (Υψηλό F/M, Χαμηλό R)

2. Προφίλ Συστάδων (Heatmap Analysis): Ο **Θερμικός Χάρτης** της **Εικόνας 5.6.4.c** επιβεβαιώνει οπτικά τα παραπάνω:

- **Noise Row:** Εμφανίζει θετικά **Z-Scores** σε **Frequency** (+1.00) και Monetary (+0.94), καθώς και εξαιρετικό **Recency** (-0.88). Η "**θερμότητα**" αυτή επιβεβαιώνει ότι πρόκειται για **θετικούς outliers**. Με άλλα λόγια, η κατηγορία **Noise** παρουσιάζει τα χαρακτηριστικά των **ιδανικών Πελατών** (Υψηλό F/M, Χαμηλό R)
- **Cluster 2:** Ξεχωρίζει με το έντονο πορτοκαλί στο Recency (+0.69) και το βαθύ μπλε στο Frequency (-1.06), σκιαγραφώντας το προφίλ των αποχωρησάντων.

Τελικό Συμπέρασμα DBSCAN:

Η εφαρμογή του **DBSCAN** προσέφερε μια μοναδική υπηρεσία που δεν κατάφεραν οι **αλγόριθμοι K-Means**: **Την Αυτόματη Ανίχνευση Ανωμαλιών (Anomaly Detection)**. Ο **αλγόριθμος** εντόπισε επιτυχώς ότι οι 440 κορυφαίοι πελάτες (Whales) δεν ανήκουν στην "κανονικότητα" της αγοράς. Αντί να προσπαθήσει να τους εντάξει βεβιασμένα σε μια ομάδα (αλλοιώνοντας τους μέσους όρους), τους απομόνωσε ως "**Noise**". Σε επιχειρηματικούς όρους, ο **DBSCAN** λειτούργησε ως **"Φίλτρο Εντοπισμού VIPs"**, διαχωρίζοντας την ελίτ από τη μάζα με απόλυτη ακρίβεια, ενώ ταυτόχρονα δημιούργησε δύο **συμπαγείς ομάδες** για τους "Normal" (Cluster 1) και "Lost" (Cluster 2) πελάτες.

5.7. Συγκριτική Επισκόπηση Επιδόσεων (Performance Overview)

Ολοκληρώνοντας την πειραματική διαδικασία, παρατίθεται ο **Συγκεντρωτικός Πίνακας 5.7**, ο οποίος συνοψίζει τις επιδόσεις των **τεσσάρων** (**4**) **αλγορίθμων** ως προς τον **χρόνο εκτέλεσης** (Execution Time) και την **ποιότητα συσταδοποίησης** (βάσει **Silhouette Score** και **σταθερότητας**). Για τον **Χρόνο Εκτέλεσης** υπολογίστηκε ο **Μέσος Όρος** των δοκιμών, ώστε να εξαχθεί ένα **αντιπροσωπευτικό μέγεθος υπολογιστικού κόστους**.

Σχολιασμός:

Από την επισκόπηση των αποτελεσμάτων προκύπτει ένα ενδιαφέρον δίλημμα μεταξύ **τεχνικής σταθερότητας** και **επιχειρηματικής χρησιμότητας**.

- Ο **K-Means++** αναδεικνύεται ως ο πιο **γρήγορος** και **σταθερός αλγόριθμος**.
- Ωστόσο, ο **K-Medoids**, παρότι υστερεί σε χρόνο και σταθερότητα, φαίνεται να **παράγει συστάδες** που ανταποκρίνονται καλύτερα στη στρατηγική του **τμήματος Marketing**, εντάσσοντας περισσότερους **πελάτες** στην κατηγορία "**Champions**" και αποφεύγοντας τα άκρα.

Η **τελική επιλογή** και η πλήρης αιτιολόγηση της προτίμησής μας στον **K-Medoids** θα αναλυθεί διεξοδικά στο **Κεφάλαιο 6**, όπου τα ποσοτικά ευρήματα θα μεταφραστούν σε **διοικητικές αποφάσεις** (Managerial Implications).

Αλγόριθμος	Χρόνος Εκτέλεσης (Μέσος Όρος)	Silhouette Score (k=5)	Παρατηρήσεις
K-Means	~3.0 sec	0.34	**Γρήγορος**, αλλά τείνει να δημιουργεί "ελιτίστικες" ομάδες VIP, αγνοώντας σημαντικό μέρος των καλών πελατών.
K-Means++	**~1.1 sec**	**0.34**	**Τεχνικά άρτιος** και **απόλυτα σταθερός**. Ωστόσο, η ομάδα VIP είναι πολύ περιορισμένη (μόλις 8.2%), αφήνοντας εκτός πολλούς αξιόλογους πελάτες.
K-Medoids (PAM)	~4.2 min	0.33	**Υψηλό κόστος** και **μεταβλητότητα**, **αλλά** προσφέρει την πιο ρεαλιστική και επιχειρηματικά αξιοποιήσιμη κατανομή (VIPs ~10-13%).
DBSCAN	~0.06 sec	N/A (Density)	Λειτούργησε αποκλειστικά ως εργαλείο εντοπισμού **ακραίων τιμών** (Noise = Super VIPs > 190.000€).

Πίνακας 5.7. Συγκριτικά Αποτελέσματα Αλγορίθμων

6. Αξιολόγηση και Ερμηνεία Αποτελεσμάτων

6.1. Ερμηνεία Ευρημάτων και Απαντήσεις στα Ερευνητικά Ερωτήματα

Η παρούσα διπλωματική εργασία σχεδιάστηκε με στόχο τη διερεύνηση της εφαρμογής αλγορίθμων Μηχανικής Μάθησης σε περιβάλλοντα Customer Data Platforms (CDPs). Βάσει των πειραματικών αποτελεσμάτων που παρουσιάστηκαν στο Κεφάλαιο 5, είμαστε πλέον σε θέση να απαντήσουμε τεκμηριωμένα στα τρία (3) Ερευνητικά Ερωτήματα που τέθηκαν στην εισαγωγή (Ενότητα 1.3).

Ερώτημα 1ο: Αξιολόγηση Αλγορίθμων Συσταδοποίησης

«Πώς συμπεριφέρονται διαφορετικοί αλγόριθμοι (K-Means, K-Medoids, DBSCAN) σε πραγματικά Δεδομένα Συμπεριφοράς Πελατών; Ποιος αλγόριθμος αποδίδει καλύτερα στη δημιουργία διακριτών Τμημάτων Πελατών;»

Η συγκριτική ανάλυση ανέδειξε ότι δεν υπάρχει ένας μοναδικός "τέλειος" αλγόριθμος, αλλά αλγόριθμοι με διαφορετική εξειδίκευση:

- **K-Means++:** Αναδείχθηκε ως ο **Τεχνικός Νικητής**. Επέδειξε την υψηλότερη σταθερότητα (0% απόκλιση μεταξύ των εκτελέσεων) και τη βέλτιστη ταχύτητα (~1.1 sec). Δημιούργησε συμπαγείς, καλά διαχωρισμένες ομάδες, αλλά με μια τάση "ελιτισμού", περιορίζοντας την ομάδα των VIPs σε ένα πολύ μικρό ποσοστό (8.2%).
- **K-Medoids (PAM):** Αναδείχθηκε ως ο **Επιχειρηματικός Νικητής**. Παρόλο που υστέρησε δραματικά σε χρόνο εκτέλεσης (~4.2 λεπτά) και εμφάνισε μεταβλητότητα αποτελεσμάτων (Label Switching), παρήγαγε την πιο ρεαλιστική κατανομή για το τμήμα Μάρκετινγκ. Διεύρυνε τη βάση των VIPs (~13%), συμπεριλαμβάνοντας πελάτες που ο K-Means απέρριπτε, προσφέροντας έτσι μεγαλύτερο εύρος στόχευσης.
- **DBSCAN:** Λειτούργησε ως **Ανιχνευτής Ανωμαλιών (Anomaly Detector)**. Απέτυχε να τμηματοποιήσει τη "μάζα" των πελατών λόγω της μεταβλητής πυκνότητας των

δεδομένων λιανικής, ωστόσο εντόπισε με χειρουργική ακρίβεια τους "Whales" (τους 440 πελάτες με τον υψηλότερο τζίρο) χαρακτηρίζοντάς τους ως Θόρυβο.

Ερώτημα 2ο: Πρακτική Υλοποίηση Pipeline

«Η ανάπτυξη και τεκμηρίωση μιας διαδικασίας σε γλώσσα Python, η οποία περιλαμβάνει την Προεπεξεργασία Δεδομένων, την Εκπαίδευση Μοντέλων και την Αξιολόγηση.»

Η έρευνα επιβεβαίωσε ότι η επιτυχία της συσταδοποίησης σε δεδομένα CDP εξαρτάται κατά 80% από την προεπεξεργασία. Η εφαρμογή του **Λογαριθμικού Μετασχηματισμού** (Log Transformation) και της **Κανονικοποίησης** (StandardScaler) ήταν καθοριστική. Χωρίς αυτά τα βήματα, όπως φάνηκε στη Διερευνητική Ανάλυση (EDA), η μεταβλητή *Monetary* κυριαρχούσε απόλυτα λόγω της ασυμμετρίας της κατανομής, οδηγώντας σε στρεβλά αποτελέσματα. Το pipeline που αναπτύχθηκε αποδείχθηκε λειτουργικό και επεκτάσιμο.

Ερώτημα 3ο: Επιχειρηματική Αξιοποίηση

«Η σύνδεση των τεχνικών αποτελεσμάτων με Επιχειρηματικές Δράσεις.»

Η μελέτη απέδειξε ότι οι μαθηματικές συστάδες (Clusters) μπορούν να μεταφραστούν άμεσα σε στρατηγικές Personas. Οι **πέντε** (**5**) **ομάδες** που εντοπίστηκαν (*Champions, Loyal, New, At Risk, Lost*) παρουσίασαν διακριτά συμπεριφορικά χαρακτηριστικά (R-F-M), επιτρέποντας τη χάραξη διαφοροποιημένων στρατηγικών, όπως αναλύεται στην **Ενότητα 6.3**.

6.2. Σύγκριση Αλγορίθμων και Συσχέτιση με τη Βιβλιογραφία

Τα ευρήματα της παρούσας μελέτης παρουσιάζουν ισχυρή συσχέτιση με την υπάρχουσα βιβλιογραφία, επιβεβαιώνοντας πολλές θεωρητικές αρχές αλλά και αναδεικνύοντας πρακτικούς περιορισμούς.

1. **Η Ανωτερότητα του K-Means++ στην Ταχύτητα:**

Σε συμφωνία με τους **Arthur & Vassilvitskii [2007]**, η μέθοδος της "Προσεκτικής Σποράς" (Careful Seeding) εξάλειψε το πρόβλημα των τοπικών ελαχίστων. Τα αποτελέσματά μας

έδειξαν ότι ο K-Means++ συγκλίνει ταχύτερα και πιο αξιόπιστα από τον απλό K-Means, καθιστώντας τον την προτιμητέα επιλογή για Big Data εφαρμογές όπου η ταχύτητα είναι κρίσιμη.

2. **Η Ευρωστία του K-Medoids (Robustness):**

Τα πειράματά μας επιβεβαίωσαν τη θεωρία των **Kaufman & Rousseeuw [1990]** ότι η χρήση Medoids (πραγματικών σημείων) αντί για Centroids (μέσων όρων) μειώνει την ευαισθησία στις ακραίες τιμές. Ωστόσο, παρατηρήθηκε αυτό που επισημαίνουν οι **Han et al. [2011]**: η υψηλή υπολογιστική πολυπλοκότητα ($O(k(n-k)^2)$) καθιστά τον αλγόριθμο PAM απαγορευτικό για real-time αναλύσεις, περιορίζοντας τη χρήση του σε offline (batch) διαδικασίες.

3. **Οι Περιορισμοί του DBSCAN στην Λιανική:**

Σε αντίθεση με την επιτυχία του σε χωρικά δεδομένα ο **DBSCAN** δυσκολεύτηκε να διαχειριστεί τα δεδομένα **RFM**. Η χρήση του για τον εντοπισμό 'θορύβου' επέτρεψε την απομόνωση των εξαιρετικά υψηλής αξίας πελατών (Whales), επιβεβαιώνοντας τη χρησιμότητα του αλγορίθμου στον εντοπισμό ανωμαλιών σε πυκνά σύνολα δεδομένων **[Ester et al., 1996]**. Αυτό συμφωνεί με τα ευρήματα των **Schubert et al. [2017]**, οι οποίοι τονίζουν την αδυναμία του αλγορίθμου να εντοπίσει συστάδες με **Μεταβλητή Πυκνότητα (Varying Density)**. Στο λιανικό εμπόριο, όπου η πυκνότητα των "Νέων Πελατών" είναι πολύ διαφορετική από αυτή των "**VIPs**", ο **DBSCAN** αναγκάστηκε να απορρίψει μεγάλο μέρος της πληροφορίας ως θόρυβο.

6.3. Επιχειρηματική Αξία και Διοικητικές Προεκτάσεις (Managerial Implications)

Στο σημείο αυτό καλούμαστε να λάβουμε τη σημαντικότερη απόφαση: **Ποιο μοντέλο θα υιοθετήσει τελικά η επιχείρηση;**

Εδώ αναδεικνύεται η σύγκρουση μεταξύ της "Μαθηματικής Βελτιστοποίησης" και της "Επιχειρηματικής Χρησιμότητας".

Το Δίλημμα: K-Means++ vs K-Medoids

- Ο **K-Means++** μας έδωσε μια ομάδα VIPs ("Champions") 482 ατόμων με μέση δαπάνη ~21.000€. Είναι μια ομάδα "Ελίτ", εξαιρετικά συμπαγής.
- Ο **K-Medoids** μας έδωσε μια ομάδα VIPs 778 ατόμων με μέση δαπάνη ~14.700€. Είναι μια ομάδα πιο "Διευρυμένη".

Η Στρατηγική Επιλογή: K-Medoids

Ως Διοικητική Πρόταση, **προκρίνεται η επιλογή του αλγορίθμου K-Medoids**, παρά το υψηλότερο **υπολογιστικό κόστος** και τη χαμηλότερη **τεχνική σταθερότητα**. Οι λόγοι είναι καθαρά στρατηγικοί:

1. **Μεγαλύτερο Εύρος Στόχευσης (Reach):** Στο **Μάρκετινγκ**, ο κίνδυνος του να αγνοήσεις έναν καλό πελάτη (False Negative) είναι συχνά μεγαλύτερος από το να συμπεριλάβεις έναν μέτριο **πελάτη** στους καλούς (False Positive). Ο **K-Medoids** αναγνώρισε ως **VIPs** σχεδόν **300** επιπλέον πελάτες που ο **K-Means++** είχε υποβαθμίσει. Αυτοί οι πελάτες αντιπροσωπεύουν **ανεκμετάλλευτη ρευστότητα**.
2. **Ρεαλιστικά Όρια (Thresholds):** Το όριο εισόδου των **14.700€** (K-Medoids) είναι πιο ρεαλιστικός στόχος για ένα πρόγραμμα Loyalty σε σχέση με τα **21.000€** (K-Means++). Είναι ευκολότερο να δώσεις κίνητρα σε έναν πελάτη να φτάσει ένα **εφικτό όριο**.
3. **Διαχείριση Κόστους:** Το γεγονός ότι ο **αλγόριθμος** χρειάζεται **τέσσερα** (4) λεπτά για να τρέξει δεν αποτελεί πρόβλημα για ένα **CDP**, καθώς η **τμηματοποίηση** συνήθως εκτελείται μία φορά την εβδομάδα (Batch Process) και όχι σε πραγματικό χρόνο.

Η επιλογή του αλγορίθμου **K-Medoids** για την τελική τμηματοποίηση εδράζεται στην **ανθεκτικότητά** του έναντι των **ακραίων τιμών** (outliers), οι οποίες είναι συχνές σε δεδομένα λιανικής **[Kaufman & Rousseeuw, 1990]**.

6.3.1. Στρατηγικό Σχέδιο Δράσης: Στοχευμένες Ενέργειες ανά Persona

Η επιλογή του αλγορίθμου **K-Medoids** μας παρείχε μια ρεαλιστική χαρτογράφηση της πελατειακής βάσης σε **πέντε** (**5**) **διακριτές ομάδες**. Η μετάφραση αυτών των ομάδων σε **επιχειρηματική δράση** απαιτεί διαφορετική προσέγγιση για την καθεμία, καθώς οι ανάγκες, η αξία και ο κίνδυνος διαφέρουν ριζικά. Με βάση την κατανομή του **K-Medoids**, προτείνονται οι εξής δράσεις:

Persona (Ομάδα)	Χαρακτηριστικά (RFM)	Κύριος Στόχος	Προτεινόμενη Στρατηγική
Champions	**↑ F, ↑ M, ↓ R** (Συχνά, Ακριβά, Πρόσφατα)	**Διατήρηση (Retention)**	VIP προνόμια, αποκλειστικότητα, όχι εκπτώσεις.
Loyal	**↑ F, Avg M, ↓ R** (Συχνά, Μέτρια, Πρόσφατα)	**Αύξηση Αξίας (Up-selling)**	Επιβράβευση πόντων, κίνητρα αύξησης καλαθιού.
Potential	**↓ F, ↓ M, ↓ R** (Σπάνια, Λίγα, Πρόσφατα)	**Ανάπτυξη (Nurturing)**	Onboarding, εκπαίδευση, κίνητρο 2ης αγοράς.
At Risk	**Avg F, Avg M, ↑ R** (Μέτρια, Μέτρια, Παλιά)	**Επαναφορά (Win-Back)**	Επιθετικές προσφορές, προσωποποιημένη προσέγγιση.
Lost	**↓ F, ↓ M, ↑ R** (Σπάνια, Λίγα, Παλιά)	**Εξοικονόμηση (Cost Saving)**	Παύση δαπανών, διατήρηση μόνο σε φθηνά κανάλια.

Πίνακας 6.3. Στρατηγικός Πίνακας με τις Προτεινόμενες Δράσεις (Action Matrix) ανά Κατηγορία Πελατών

Στον παραπάνω **Πίνακα 6.3**, συνοψίζεται η **στρατηγική στόχευση** για κάθε ομάδα, ενώ στη συνέχεια παρατίθεται η **αναλυτική περιγραφή** των **προτεινόμενων δράσεων**.

Ακολουθεί η **αναλυτική εξειδίκευση** των **δράσεων** για κάθε κατηγορία:

1. The Champions (Cluster 5): Η Ελίτ των Πελατών

- **Προφίλ:** 778 Πελάτες (~13%) | **Μέση Δαπάνη:** 14.692€ | **Συχνότητα:** ~25 αγορές.
- **Διάγνωση:** Είναι οι πιο **πολύτιμοι πελάτες**. Αγοράζουν συχνά, πρόσφατα και ξοδεύουν τα περισσότερα. Ο κίνδυνος εδώ δεν είναι να μην αγοράσουν, αλλά να φύγουν στον ανταγωνισμό. Η κατηγοριοποίηση των πελατών σε ομάδες όπως 'Champions' ακολουθεί τις βέλτιστες πρακτικές του μοντέλου RFM για τη μεγιστοποίηση της αξίας του πελάτη **[Hughes, 2005]**.
- **Προτεινόμενες Δράσεις:**
    - **Πρόγραμμα Επιβράβευσης VIP:** Παροχή προνομίων στάτους (Status-based rewards), όπως δωρεάν μεταφορικών για πάντα, αποκλειστική τηλεφωνική γραμμή εξυπηρέτησης ή προσωπικός Account Manager.
    - **Πρώιμη Πρόσβαση (Early Access):** Ενημέρωση για νέα προϊόντα 24 ώρες πριν από το ευρύ κοινό, ενισχύοντας την αίσθηση του "ανήκειν".
    - **Αποφυγή Μαζικών Εκπτώσεων:** Η υπερβολική έκπτωση σε αυτή την ομάδα μειώνει απλώς το περιθώριο κέρδους (margin erosion) χωρίς να φέρνει επιπλέον πωλήσεις.
    - **"Surprise & Delight":** Απρόσμενα δώρα (π.χ. δείγματα πολυτελείας) για συναισθηματική σύνδεση.

2. Loyal Customers (Cluster 2): Ο Κορμός της Επιχείρησης

- **Προφίλ:** 1.246 Πελάτες (~21%) | **Μέση Δαπάνη:** 2.992€ | **Συχνότητα:** ~7.5 αγορές.
- **Διάγνωση:** Είναι σταθεροί πελάτες με καλή συχνότητα, αλλά δεν έχουν φτάσει ακόμα το μέγιστο της αγοραστικής τους δύναμης. Η κατηγοριοποίηση των πελατών σε ομάδες όπως 'Loyal' ακολουθεί τις βέλτιστες πρακτικές του μοντέλου RFM για τη μεγιστοποίηση της αξίας του πελάτη **[Hughes, 2005]**
- **Προτεινόμενες Δράσεις:**
    - **Συστήματα Συστάσεων:** Χρήση αλγορίθμων για συμπληρωματικές πωλήσεις (Cross-sell).

- **Κίνητρα Ποσότητας:** Προσφορές τύπου *"Αγοράστε 50€ ακόμη για να κερδίσετε κουπόνι"*, ώστε να αυξηθεί το Average Order Value (AOV).
- **Gamification:** Δημιουργία στόχων μετάβασης στην κατηγορία VIP.

3. New / Potential (Cluster 1): Το Μέλλον

- **Προφίλ:** 1.164 Πελάτες (~20%) | **Μέση Δαπάνη:** 646€ | **Συχνότητα:** ~2.3 αγορές.
- **Διάγνωση:** Πρόσφατοι πελάτες (Recency ~37 ημέρες) που δοκιμάζουν την εταιρεία. Η σχέση είναι εύθραυστη.
- **Προτεινόμενες Δράσεις:**
  - **Welcome Series:** Αυτοματοποιημένα emails που εκπαιδεύουν τον πελάτη για το brand και τις αξίες του.
  - **Κίνητρο 2ης Αγοράς:** Προσφορά έκπτωσης για την επόμενη αγορά με χρονικό περιορισμό, καθώς η πιθανότητα παραμονής αυξάνεται δραματικά μετά τη 2η αγορά.
  - **Συλλογή Feedback:** Άμεση έρευνα ικανοποίησης για επίλυση τυχόν προβλημάτων.

4. Hibernating / At Risk (Cluster 4): Τα "Κρυμμένα Λεφτά"

- **Προφίλ:** 1.261 Πελάτες (~21%) | **Μέση Δαπάνη:** 1.177€ | **Recency:** ~330 ημέρες.
- **Διάγνωση:** Πελάτες που στο παρελθόν ξόδεψαν σημαντικά ποσά, αλλά έχουν "εξαφανιστεί" εδώ και έναν χρόνο. Είναι η πιο κρίσιμη ομάδα για Win-Back.
- **Προτεινόμενες Δράσεις:**
  - **Επιθετικές Προσφορές:** Μεγάλες εκπτώσεις (π.χ. 20-30%), καθώς είναι προτιμότερο να μειωθεί το κέρδος παρά να χαθεί οριστικά ο πελάτης.
  - **Μηνύματα "Μας Λείψατε":** Προσωποποιημένη επικοινωνία που υπενθυμίζει την προηγούμενη σχέση.
  - **Omnichannel Πίεση:** Χρήση SMS/Viber για διασφάλιση ανάγνωσης του μηνύματος.

5. Lost Customers (Cluster 3): Διαχείριση Κόστους

- **Προφίλ:** 1.429 Πελάτες (~24%) | **Μέση Δαπάνη:** ~246€ | **Recency:** ~434 ημέρες.
- **Διάγνωση:** Πελάτες χαμηλής αξίας που έχουν να αγοράσουν πάνω από 14 μήνες.
- **Προτεινόμενες Δράσεις:**
    - **Παύση Δαπανών:** Αφαίρεση από λίστες πληρωμένων διαφημίσεων για εξοικονόμηση budget.
    - **Low-Cost Συντήρηση:** Διατήρηση μόνο σε φθηνά κανάλια (Email) για γενικές προσφορές.
    - **Εκκαθάριση:** Διαγραφή από τη βάση αν δεν υπάρξει ανταπόκριση, για βελτίωση της ποιότητας δεδομένων (Data Hygiene).

6.3.2. Πρακτικές Προεκτάσεις (Practical Implications)

Η υλοποίηση του pipeline συσταδοποίησης μέσω του **RFM Master Tool v4.5.0** δεν αποτελεί απλώς μια τεχνική άσκηση, αλλά επιφέρει θεμελιώδεις αλλαγές στον τρόπο με τον οποίο ένας οργανισμός αντιλαμβάνεται και διαχειρίζεται το πελατολόγιό του. Οι κυριότερες πρακτικές προεκτάσεις συνοψίζονται στα εξής:

- **Μετάβαση από τη Μαζική στην Προσωποποιημένη Επικοινωνία:** Η δυνατότητα αυτόματου διαχωρισμού των πελατών σε Personas (Champions, Loyal, At-Risk κ.λπ.) επιτρέπει στην επιχείρηση να εγκαταλείψει τις δαπανηρές και συχνά ενοχλητικές μαζικές καμπάνιες (Mass Marketing). Πρακτικά, αυτό σημαίνει **βελτιστοποίηση του διαφημιστικού προϋπολογισμού**, καθώς οι πόροι κατευθύνονται εκεί που υπάρχει μεγαλύτερη πιθανότητα ανταπόκρισης.
- **Λήψη Αποφάσεων Βάσει Δεδομένων (Data-Driven Decision Making):** Η χρήση αλγορίθμων όπως ο **K-Medoids** προσφέρει στη διοίκηση μια αντικειμενική εικόνα της αγοράς, απαλλαγμένη από υποκειμενικές εκτιμήσεις. Η επιχείρηση μπορεί πλέον να γνωρίζει ανά πάσα στιγμή το ακριβές μέγεθος της "υγιούς" πελατειακής της βάσης (Loyal) έναντι αυτής που κινδυνεύει να χαθεί (Hibernating).

- **Επιχειρησιακή Ευελιξία και Αυτοματοποίηση:** Η ενσωμάτωση τέτοιων εργαλείων σε μια υποδομή **CDP** επιτρέπει τον τακτικό επαναπροσδιορισμό των συστάδων με ελάχιστο ανθρώπινο κόπο. Η ταχύτητα του **K-Means++** εξασφαλίζει ότι η τμηματοποίηση μπορεί να επικαιροποιείται ακόμη και σε εβδομαδιαία βάση, επιτρέποντας στην εταιρεία να αντιδρά ακαριαία σε αλλαγές της καταναλωτικής συμπεριφοράς.
- **Βελτίωση της Εμπειρίας του Πελάτη (CX):** Με την εφαρμογή των προτεινόμενων δράσεων (π.χ. Welcome Series για τους Νέους ή VIP Treatment για τους Champions), ο πελάτης αισθάνεται ότι η επιχείρηση κατανοεί τις ανάγκες του. Αυτό οδηγεί σε **αύξηση της πιστότητας (Retention)** και της **Δια Βίου Αξίας του Πελάτη (CLV)**.
- **Στρατηγική Διαχείριση Ακραίων Τιμών:** Η χρήση του **DBSCAN** ως "ανιχνευτή θορύβου" προσφέρει μια μοναδική πρακτική δυνατότητα: την απομόνωση των "Super-VIPs" (Whales), οι οποίοι απαιτούν εντελώς διαφορετική, προνομιακή διαχείριση (π.χ. B2B συμβόλαια), καθώς οι ανάγκες τους διαφέρουν ριζικά από τον μέσο καταναλωτή.

6.4. Περιορισμοί της Έρευνας (Limitations)

Παρά τα σημαντικά ευρήματα και την επιτυχή υλοποίηση του πειραματικού pipeline, η παρούσα έρευνα υπόκειται σε συγκεκριμένους περιορισμούς, οι οποίοι πρέπει να ληφθούν υπόψη κατά την ερμηνεία των αποτελεσμάτων. Οι περιορισμοί αυτοί ταξινομούνται σε τρεις άξονες: **Δεδομένα**, **Αλγόριθμοι** και **Επιχειρησιακό Πλαίσιο**.

6.4.1. Περιορισμοί που αφορούν τα Δεδομένα (Data-Related Limitations)

- **Μονοδιάστατη Πηγή Πληροφορίας:** Η ανάλυση βασίστηκε αποκλειστικά σε **συναλλακτικά δεδομένα (Transactional Data)**. Αν και το μοντέλο **RFM** είναι ισχυρό, στερείται ποιοτικών και δημογραφικών στοιχείων, όπως η ηλικία, το φύλο, η τοποθεσία ή τα δεδομένα πλοήγησης (clickstream data) από το **CDP**. Η έλλειψη αυτών

των διαστάσεων εμποδίζει την πλήρη "**προσωποποίηση 360 μοιρών**" που αναλύθηκε στο θεωρητικό υπόβαθρο.

- **Χρονική Στατικότητα (Snapshot Bias):** Η τμηματοποίηση πραγματοποιήθηκε σε ένα συγκεκριμένο χρονικό σημείο (Snapshot Date: 10/12/2011). Η αγοραστική συμπεριφορά όμως είναι δυναμική. Ένας πελάτης που σήμερα κατατάσσεται στους "Champions" μπορεί να μετακινηθεί στους "At Risk" σε σύντομο χρονικό διάστημα, κάτι που η στατική ανάλυση **RFM** αδυνατεί να παρακολουθήσει σε πραγματικό χρόνο.

6.4.2. Αλγοριθμικοί και Τεχνικοί Περιορισμοί

- **Υπολογιστικό Κόστος του K-Medoids:** Όπως διαπιστώθηκε στην **Ενότητα 5.7**, ο αλγόριθμος **PAM** εμφάνισε εξαιρετικά υψηλό υπολογιστικό κόστος (~4.2 λεπτά έναντι ~1.1 sec του K-Means++). Αν και για το παρόν δείγμα των 5.878 πελατών ο χρόνος αυτός ήταν ανεκτός, η εφαρμογή του σε περιβάλλοντα Big Data με εκατομμύρια εγγραφές θα καθίστατο απαγορευτική χωρίς τη χρήση παραλλαγών δειγματοληψίας όπως ο **CLARA**.
- **Ευαισθησία Παραμέτρων στον DBSCAN:** Ο **DBSCAN** αποδείχθηκε ιδιαίτερα "δύστροπος" στη ρύθμιση. Μικρές αλλαγές στις παραμέτρους **ε** και **MinPts** προκαλούσαν δραματικές μεταβολές στα αποτελέσματα (Butterfly Effect), οδηγώντας συχνά σε υπερ-τμηματοποίηση ή πλήρη απόρριψη δεδομένων ως θόρυβο.

6.4.3. Περιορισμοί Επιχειρησιακής Εφαρμογής

- **Έλλειψη Επαλήθευσης (A/B Testing):** Οι προτεινόμενες δράσεις μάρκετινγκ στην **Ενότητα 6.3.1** είναι βασισμένες σε αλγοριθμικές ενδείξεις και επιχειρηματική θεωρία. Λόγω της φύσης της διπλωματικής εργασίας, δεν κατέστη εφικτό να εφαρμοστούν αυτές οι δράσεις σε πραγματικές συνθήκες και να μετρηθεί η αποτελεσματικότητά τους (π.χ. μέσω Conversion Rate ή ROI) σε σχέση με μια ομάδα ελέγχου.

- **Αρχή "Garbage In, Garbage Out":** Η έρευνα επιβεβαίωσε ότι η ποιότητα της τμηματοποίησης εξαρτάται απόλυτα από την καθαρότητα των δεδομένων εισόδου. Εάν το **CDP** της επιχείρησης δεν διαθέτει ισχυρούς μηχανισμούς **Identity Resolution**, ο κίνδυνος δημιουργίας ψευδών προφίλ λόγω διπλότυπων εγγραφών παραμένει υπαρκτός.

6.5. Προτάσεις για Μελλοντική Έρευνα (Future Research)

Η παρούσα διπλωματική εργασία έθεσε τις βάσεις για την κατανόηση της συμπεριφοράς πελατών μέσω συγκεκριμένων αλγορίθμων. Ωστόσο, η αρχιτεκτονική του λογισμικού που αναπτύχθηκε (**RFM Master Tool v4.5.0**) προσφέρει πρόσφορο έδαφος για περαιτέρω εξέλιξη. Βασιζόμενοι στη δομική διάρθρωση του συστήματος (όπως παρουσιάστηκε στον Πίνακα Λειτουργικών Ενοτήτων), προτείνονται οι εξής κατευθύνσεις για μελλοντική έρευνα και ανάπτυξη:

1. **Εξέλιξη σε "Universal Intelligent Analyzer" (Data Agnostic System)**

   Ο απώτερος στόχος για τη μελλοντική ανάπτυξη του εργαλείου είναι η μετάβασή του από ένα σύστημα εστιασμένο αποκλειστικά στο **RFM** (Retail), σε μια καθολική πλατφόρμα **ανάλυσης δεδομένων**. Προτείνεται η αναβάθμιση του **Data Engineering (ETL) Module** ώστε να δέχεται **οποιοδήποτε είδος δεδομένων** (όχι μόνο συναλλακτικά, αλλά και δημογραφικά, ιατρικά, βιομηχανικά κ.ά.) σε διάφορες μορφές (CSV, Excel, JSON, SQL). Το σύστημα θα πρέπει να αναγνωρίζει αυτόματα τους τύπους των μεταβλητών (numerical vs categorical) και να προσαρμόζει τη στρατηγική προεπεξεργασίας χωρίς ανθρώπινη παρέμβαση.

2. **Υλοποίηση Πλήρους Auto-ML (Αυτόματη Επιλογή Αλγορίθμου)**

   Επεκτείνοντας τη λειτουργικότητα του **Intelligence (Auto-ML) Module**, η μελλοντική έκδοση του λογισμικού θα πρέπει να λειτουργεί ως "**Ενορχηστρωτής**" (Orchestrator). Αντί ο χρήστης να επιλέγει χειροκίνητα τη μέθοδο, το σύστημα θα εκτελεί αυτόματα και

παράλληλα (μέσω Multi-Threading) όλους τους διαθέσιμους αλγορίθμους που ενσωματώθηκαν στην παρούσα μελέτη:

- **K-Means / K-Means++** (για σφαιρικές/γραμμικές συστάδες)
- **K-Medoids / PAM** (για δεδομένα με θόρυβο/outliers)
- **DBSCAN** (για πυκνοτικές/ακανόνιστες δομές)
- Το σύστημα θα συγκρίνει σε πραγματικό χρόνο τους δείκτες αξιολόγησης (Silhouette, Calinski-Harabasz, Davies-Bouldin) για κάθε αλγόριθμο και **θα επιλέγει αυτόματα τη βέλτιστη μέθοδο ταξινόμησης** για το συγκεκριμένο dataset που εισήχθη.

3. **Αυτοματοποιημένη Ερμηνεία και Εξαγωγή Χαρακτηριστικών (Zero-Touch Insights)**

Μια κρίσιμη προσθήκη για το μέλλον αφορά την αυτοματοποίηση του **Visualization & Interpretation**. Το λογισμικό δεν θα πρέπει απλώς να παράγει αριθμητικά Clusters (π.χ. "Cluster 0"), αλλά να προχωρά σε **Σημασιολογική Ανάλυση (Semantic Profiling)**. Μέσω στατιστικών κανόνων ή ενσωμάτωσης γλωσσικών μοντέλων (LLMs), το εργαλείο θα πρέπει:

- Να "διαβάζει" τα χαρακτηριστικά της κάθε ομάδας (π.χ. "Υψηλή Τιμή στη στήλη Χ, Χαμηλή στη στήλη Υ").
- Να αποδίδει αυτόματα ετικέτες (Auto-Labeling), π.χ. *"Ομάδα Υψηλής Απόδοσης"* ή *"Ομάδα Κινδύνου"*.
- Να εξάγει άμεσα report με τα χαρακτηριστικά των ομάδων, έτοιμα προς επιχειρηματική αξιοποίηση, εκμηδενίζοντας τον χρόνο που απαιτείται για την ανάλυση των αποτελεσμάτων από τον άνθρωπο.

4. **Ενσωμάτωση Επιπλέον Αλγορίθμων**

Τέλος, προτείνεται ο εμπλουτισμός του **Clustering Engine** με επιπλέον αλγορίθμους για την κάλυψη πιο σύνθετων αναγκών, όπως:

- **Hierarchical Clustering (Agglomerative):** Για την εμφάνιση δενδρογραμμάτων και ιεραρχικών σχέσεων μεταξύ των ομάδων.

- **OPTICS:** Ως εξέλιξη του DBSCAN για την αντιμετώπιση δεδομένων με μεταβλητή πυκνότητα.
- **Gaussian Mixture Models (GMM):** Για πιθανοτική προσέγγιση σε περιπτώσεις που τα όρια των ομάδων δεν είναι σαφή (Soft Clustering).

7. Συμπεράσματα και Προτάσεις

7.1. Ανακεφαλαίωση και Κύρια Συμπεράσματα

Η παρούσα εργασία διερεύνησε τον κομβικό ρόλο των **Πλατφορμών Δεδομένων Πελατών** (CDP) και τη συμβολή της **Μηχανικής Μάθησης** στην επίτευξη της υπερ-προσωποποίησης στο σύγχρονο λιανικό εμπόριο. Μέσα από την ανάπτυξη και εφαρμογή του **RFM Master Tool v4.5.0**, επιβεβαιώθηκε ότι η μετάβαση από τα παραδοσιακά "σιλό" δεδομένων σε μια ενοποιημένη εικόνα πελάτη (Single Customer View) αποτελεί τη μοναδική βιώσιμη στρατηγική για την ψηφιακή εποχή. Η ενοποίηση των δεδομένων μέσω μιας υποδομής **CDP** αποτελεί το θεμέλιο για την επιτυχημένη εφαρμογή μοντέλων **Μηχανικής Μάθησης** σε κλίμακα **[Raab, 2018]**.

Η Στρατηγική Σημασία της Προεπεξεργασίας

Ένα από τα σημαντικότερα συμπεράσματα της έρευνας είναι ότι η ποιότητα της συσταδοποίησης είναι άρρηκτα συνδεδεμένη με την προετοιμασία των δεδομένων. Η εφαρμογή του **Λογαριθμικού Μετασχηματισμού** και της **Κανονικοποίησης Z-Score** δεν ήταν απλώς τεχνικά βήματα, αλλά αναγκαίες μαθηματικές παρεμβάσεις για την εξισορρόπηση των μεταβλητών **RFM**. Χωρίς αυτές, η μεταβλητή **Monetary** θα είχε επισκιάσει τη **Συχνότητα** και την **Προσφατότητα**, οδηγώντας σε μια μονοδιάστατη και λανθασμένη τμηματοποίηση που θα αγνοούσε τη συμπεριφορά πολυπλοκότητας των πελατών.

Το Δίλημμα Τεχνικής vs Επιχειρηματικής Υπεροχής

Η συγκριτική αξιολόγηση ανέδειξε μια ενδιαφέρουσα αντίφαση μεταξύ αλγοριθμικής απόδοσης και πρακτικής χρησιμότητας:

- Ο **K-Means++** αποδείχθηκε ο «μαθηματικός νικητής», προσφέροντας απόλυτη σταθερότητα και ακαριαία ταχύτητα, στοιχεία κρίσιμα για την αυτοματοποίηση σε μεγάλα σύνολα δεδομένων (Big Data).
- Ωστόσο, ο **K-Medoids (PAM)** προκρίθηκε ως η βέλτιστη επιλογή για τη χάραξη στρατηγικής **Μάρκετινγκ**. Η χρήση πραγματικών πελατών ως κέντρων των συστάδων (medoids) περιόρισε την επίδραση των ακραίων τιμών και επέτρεψε τον εντοπισμό

μιας πιο διευρυμένης ομάδας **VIPs** (Champions), προσφέροντας στην επιχείρηση **30% μεγαλύτερο εύρος στόχευσης** σε σχέση με τον **K-Means++.**

DBSCAN: Η Εξειδικευμένη Αξία του Θορύβου Παρόλο που ο **DBSCAN** υστέρησε στη δημιουργία σφαιρικών επιχειρηματικών ομάδων, η συνεισφορά του στην **Ανίχνευση Ανωμαλιών (Anomaly Detection)** ήταν καταλυτική. Εντόπισε με ακρίβεια τους "Whales" (Super-VIPs), αποδεικνύοντας ότι σε ένα **CDP**, ο "θόρυβος" δεν είναι πάντα άχρηστα δεδομένα, αλλά μπορεί να αντιπροσωπεύει το πιο κερδοφόρο κομμάτι του πελατολογίου που απαιτεί εξατομικευμένη **B2B** διαχείριση.

7.2. Μελλοντικές Τάσεις στον Χώρο των CDPs

Ο τομέας των **Πλατφορμών Δεδομένων Πελατών** (CDP) και της **Μηχανικής Μάθησης** διέρχεται μια φάση ριζικού μετασχηματισμού. Η εξέλιξη μετατοπίζεται από την απλή περιγραφική ανάλυση (Descriptive Analytics) προς την αυτόνομη, προληπτική δράση (Autonomous Action), διαμορφώνοντας τις εξής κυρίαρχες τάσεις:

- **Generative AI και Hyper-Personalization (Υπερ-Προσωποποίηση):** Η ενσωμάτωση Μεγάλων **Γλωσσικών Μοντέλων** (LLMs) και **Παραγωγικής Τεχνητής Νοημοσύνης** απευθείας στα **CDPs** αλλάζει το παράδειγμα της **επικοινωνίας**. Στο άμεσο μέλλον, το σύστημα δεν θα εντοπίζει απλώς τη **συστάδα** (π.χ. "At Risk"), αλλά θα δημιουργεί αυτόματα προσωποποιημένο περιεχόμενο (δυναμικά emails, εξατομικευμένα banners, μοναδικές προσφορές) που θα προσαρμόζεται στο ύφος, τη γλώσσα και τις ψυχογραφικές ανάγκες του κάθε **πελάτη** ξεχωριστά. Αυτή η **αυτοματοποίηση** μειώνει τον κύκλο παραγωγής καμπανιών από ημέρες σε δευτερόλεπτα, επιτρέποντας την **επικοινωνία** σε κλίμακα (Scaling Personalization).
- **Η Κυριαρχία των First-Party Δεδομένων σε ένα Privacy-First Περιβάλλον:** Με την οριστική κατάργηση των **cookies** τρίτων (third-party cookies) και την αυστηροποίηση του κανονιστικού πλαισίου (GDPR, CCPA), τα **CDPs** καθίστανται ο ακρογωνιαίος λίθος της ψηφιακής στρατηγικής. Η **ικανότητα** ενός οργανισμού να συλλέγει, να ενοποιεί και να ενεργοποιεί **First-Party δεδομένα** (δεδομένα που παραχωρούνται οικειοθελώς από τον πελάτη) αποτελεί πλέον το **μοναδικό βιώσιμο**

και νόμιμο ανταγωνιστικό **πλεονέκτημα**. Τα **CDPs** θα εξελιχθούν σε "Κέντρα Εμπιστοσύνης", όπου η **προσωποποίηση** θα ισορροπεί απόλυτα με τον σεβασμό στην **ιδιωτικότητα**.

- **Από το "Clustering" στο "Next Best Action" (Προβλεπτική Ανάλυση):** Η τάση μετακινείται από την **αναδρομική ανάλυση** ("τι συνέβη στο παρελθόν") στην προληπτική καθοδήγηση ("τι πρέπει να γίνει τώρα"). Μέσω της χρήσης **Real-Time Streaming Analytics**, η **κατηγοριοποίηση** του πελάτη δεν θα είναι στατική αλλά ρευστή. Καθώς ο **πελάτης** αλληλοεπιδρά με το ψηφιακό κατάστημα, το CDP θα επανυπολογίζει το προφίλ του σε πραγματικό χρόνο και θα προτείνει την "**Επόμενη Βέλτιστη Ενέργεια**" (Next Best Action), όπως μια συγκεκριμένη έκπτωση ή ένα άρθρο υποστήριξης, προλαμβάνοντας την εγκατάλειψη πριν αυτή συμβεί.
- **Composable CDPs και Data Clean Rooms:** Η **αρχιτεκτονική** των **CDPs** τείνει να γίνει πιο "σπονδυλωτή" (Composable), επιτρέποντας στις επιχειρήσεις να συνδέουν υπάρχουσες αποθήκες δεδομένων (Data Warehouses) απευθείας με εργαλεία ενεργοποίησης. Παράλληλα, η χρήση **Data Clean Rooms** θα επιτρέψει σε διαφορετικές επιχειρήσεις να συνεργάζονται και να διασταυρώνουν **δεδομένα πελατών** με ασφάλεια, χωρίς να αποκαλύπτουν ευαίσθητες **προσωπικές πληροφορίες**, ενισχύοντας την ακρίβεια των **αλγορίθμων συσταδοποίησης**.
- **Πρόβλεψη Διαρροής Πελατών (Churn Prediction) μέσω Προβλεπτικής Ανάλυσης:** Μια εξαιρετικά κρίσιμη προέκταση της παρούσας μελέτης αποτελεί η μετάβαση από τη στατική τμηματοποίηση στην **Προβλεπτική Ανάλυση** (Predictive Analytics), με επίκεντρο το μοντέλο **Πρόβλεψης Διαρροής Πελατών** (Churn Prediction). Ενώ η συσταδοποίηση **RFM** προσφέρει μια σαφή εικόνα της τρέχουσας κατάστασης του πελατολογίου, η ενσωμάτωση **αλγορίθμων εποπτευόμενης μάθησης** (Supervised Learning), όπως τα Random Forests ή η Logistic Regression, θα επέτρεπε στην επιχείρηση να υπολογίζει την πιθανότητα εγκατάλειψης για κάθε πελάτη ξεχωριστά πριν αυτή πραγματοποιηθεί. Σύμφωνα με τους **Ng & Liu (2000)**, η χρήση **τεχνικών εξόρυξης δεδομένων** για τη διατήρηση πελατών είναι οικονομικά αποδοτικότερη από την προσέλκυση νέων, καθώς επιτρέπει την εφαρμογή προληπτικών, εξατομικευμένων **στρατηγικών διατήρησης** σε **πελάτες** που

εμφανίζουν πρώιμα σημάδια **αδράνειας**. Η συνδυαστική χρήση των συστάδων που παρήγαγε το **RFM Master Tool** ως μεταβλητές εισόδου (features) σε ένα μοντέλο πρόβλεψης, θα μπορούσε να αυξήσει δραματικά την ακρίβεια των ενεργειών **Marketing**, μετατρέποντας το **CDP** από ένα σύστημα καταγραφής σε έναν προληπτικό μηχανισμό **διασφάλισης εσόδων**.

7.3. Στρατηγικές και Πρακτικές Προτάσεις

Ως κατακλείδα της παρούσας μελέτης και με γνώμονα την **επιχειρηματική αξιοποίηση** των ευρημάτων, προτείνονται οι εξής στρατηγικές και **επιχειρησιακές κατευθύνσεις** για τους οργανισμούς που επιδιώκουν να ηγηθούν στην εποχή της **προσωποποίησης**:

- **Επένδυση σε Υποδομές Ενοποίησης και Ποιότητας Δεδομένων:** Οι επιχειρήσεις πρέπει να ιεραρχήσουν την κατάργηση των "σιλό" δεδομένων μέσω της υιοθέτησης **CDP** αρχιτεκτονικών. Η έρευνα απέδειξε ότι η ακρίβεια των αλγορίθμων είναι άρρηκτα συνδεδεμένη με την πληρότητα της εικόνας του πελάτη και την **αυστηρή προεπεξεργασία** (normalization, log transformation). Προτείνεται η θέσπιση πλαισίου **Data Governance** για τη διασφάλιση της "καθαρότητας" των **δεδομένων εισόδου**, ώστε να αποφεύγεται το φαινόμενο "Garbage In, Garbage Out".
- **Υιοθέτηση Πολυ-αλγοριθμικής Προσέγγισης (Multi-Model Strategy):** Αντί για την προσκόλληση σε ένα **μοναδικό μοντέλο**, προτείνεται η συνδυαστική χρήση **αλγορίθμων** ανάλογα με τον στόχο.
    - Χρήση του **K-Means++** για ταχύτατη και σταθερή τμηματοποίηση της μεγάλης μάζας των πελατών σε τακτική βάση.
    - Χρήση του **K-Medoids** για βαθύτερη στρατηγική ανάλυση και ορισμό των "ιδανικών" Personas, καθώς οι πραγματικοί αντιπρόσωποι (medoids) προσφέρουν πιο ρεαλιστικά όρια για τις καμπάνιες **μάρκετινγκ**.
    - Χρήση του **DBSCAN** ως εξειδικευμένο εργαλείο ανίχνευσης των "Whales" (Super-VIPs), οι οποίοι απαιτούν εντελώς διαφορετική, προνομιακή διαχείριση εκτός των μαζικών αυτοματισμών.

- **Εστίαση στη Δυναμική Τμηματοποίηση και το Lifecycle Marketing:** Η επιχείρηση οφείλει να χρησιμοποιεί εργαλεία όπως το **RFM Master Tool** για τον συνεχή και τακτικό επαναπροσδιορισμό των συστάδων. Η πρακτική αξία έγκειται στον έγκαιρο εντοπισμό της "μετανάστευσης" των πελατών μεταξύ των κατηγοριών, επιτρέποντας την άμεση παρέμβαση με Win-Back καμπάνιες για τους πελάτες που διολισθαίνουν από την κατηγορία "Loyal" στην κατηγορία "At Risk".
- **Σχεδιασμός Δράσεων βάσει της "Αξίας" και όχι μόνο της "Ποσότητας":** Προτείνεται η υιοθέτηση του στρατηγικού πίνακα δράσεων (Action Matrix), όπου οι πόροι κατανέμονται ιεραρχικά. Οι "Champions" πρέπει να απολαμβάνουν υπηρεσίες προστιθέμενης αξίας και στάτους, ενώ οι "Lost" πελάτες πρέπει να προσεγγίζονται μόνο μέσω καναλιών χαμηλού κόστους, διασφαλίζοντας τη μέγιστη αποδοτικότητα του marketing budget.
- **Μετάβαση προς το Intelligent Auto-ML:** Η μελλοντική αναβάθμιση των **συστημάτων ανάλυσης** πρέπει να στοχεύει στην αυτοματοποίηση της επιλογής της βέλτιστης μεθόδου **ανάγνωσης δεδομένων**. Ένα σύστημα που "αντιλαμβάνεται" τη φύση του **datasheet** και προτείνει αυτόματα την **καλύτερη ταξινόμηση**, θα επιτρέψει στα στελέχη να επικεντρωθούν στη **στρατηγική δημιουργικότητα** αντί για την **τεχνική παραμετροποίηση**.
- **Εναρμόνιση με την Ηθική Χρήση Δεδομένων (Privacy-Driven Value):** Οι **πρακτικές προσωποποίησης** πρέπει να είναι απόλυτα ευθυγραμμισμένες με το **νομικό πλαίσιο** (GDPR). Η διαφάνεια στη συλλογή **First-Party δεδομένων** δεν είναι μόνο νομική υποχρέωση, αλλά αποτελεί τη βάση για το χτίσιμο μακροχρόνιας **εμπιστοσύνης** με τον πελάτη, η οποία μεταφράζεται σε υψηλότερη **Δια Βίου Αξία** (CLV).

8. Παράρτημα

Στο παρόν κεφάλαιο παρατίθενται τα κρισιμότερα τμήματα του πηγαίου κώδικα της εφαρμογής **"RFM Advanced Clustering Tool v4.5.0"**. Όπως έχουμε ήδη αναφέρει στην **Ενότητα 4. Υπολογιστικό Πλαίσιο και Τεχνικό Περιβάλλον**, Η εφαρμογή αναπτύχθηκε σε γλώσσα **Python 3.13** με χρήση της βιβλιοθήκης **CustomTkinter** για το γραφικό περιβάλλον και του **Scikit-Learn** για τους αλγορίθμους **Μηχανικής Μάθησης**.

8.1. Βιβλιοθήκες και Αρχικοποίηση

Στο απόσπασμα που ακολουθεί γίνεται η εισαγωγή των απαραίτητων **βιβλιοθηκών**. Ιδιαίτερη έμφαση δίνεται στις **βιβλιοθήκες pandas** για τη διαχείριση δεδομένων, **sklearn** για την εκτέλεση των αλγορίθμων και **matplotlib/seaborn** για την **οπτικοποίηση**.

Πηγαίος Κώδικας 1: Εισαγωγή Βιβλιοθηκών και Ρυθμίσεις

```
# --- Βασικές Βιβλιοθήκες ---
import customtkinter as ctk        # Σύγχρονο Γραφικό Περιβάλλον (GUI)
import pandas as pd                # Διαχείριση και ανάλυση δεδομένων (Dataframes)
import numpy as np                 # Μαθηματικές πράξεις και διαχείριση πινάκων
import datetime as dt              # Διαχείριση ημερομηνιών (για υπολογισμό Recency)
import matplotlib.pyplot as plt    # Δημιουργία γραφικών παραστάσεων
import seaborn as sns              # Στατιστική οπτικοποίηση (Heatmaps)
import threading                   # Πολυνηματική εκτέλεση (για αποφυγή "παγώματος" του GUI)

# --- Βιβλιοθήκες Μηχανικής Μάθησης (Scikit-Learn) ---
from sklearn.preprocessing import StandardScaler        # Z-Score Normalization
from sklearn.cluster import KMeans, DBSCAN              # Αλγόριθμοι συσταδοποίησης
from sklearn.neighbors import NearestNeighbors          # Υπολογισμός k-Distance Graph
from sklearn.metrics import silhouette_score            # Μετρική αξιολόγησης (Silhouette)
```

```
from sklearn.metrics import calinski_harabasz_score     # Μετρική διαχωρισμού συστάδων
from sklearn.metrics import davies_bouldin_score        # Μετρική ομοιότητας συστάδων

# Προσπάθεια φόρτωσης του K-Medoids (αν η βιβλιοθήκη είναι εγκατεστημένη)
try:
    from sklearn_extra.cluster import KMedoids
except ImportError:
    pass  # Συνεχίζει με τους υπόλοιπους αλγορίθμους αν λείπει
```

8.2. ETL: Προεπεξεργασία και Μετασχηματισμός RFM

Η συνάρτηση load_data αποτελεί τον πυρήνα της προεπεξεργασίας. Εδώ τα ακατέργαστα δεδομένα «καθαρίζονται» και μετατρέπονται σε πίνακα RFM στην μνήμη (RAM) του υπολογιστή υλοποιώντας τη Μέθοδο της Κανονικοποίησης **Κεφάλαιο 4** (Z-Score Normalization)

Πηγαίος Κώδικας 2: Υπολογισμός RFM και Κανονικοποίηση Z-Scores

```
def load_data(self):
    """
    Φορτώνει το αρχείο, εκτελεί καθαρισμό και υπολογίζει τις μετρικές RFM.
    Εφαρμόζει λογαριθμικό μετασχηματισμό για τη μείωση της ασυμμετρίας (skewness).
    """
    # 1. Καθαρισμός Δεδομένων
    df.dropna(subset=['Customer ID'], inplace=True)
    df = df[(df['Quantity'] > 0) & (df['Price'] > 0)]
    df['TotalPrice'] = df['Quantity'] * df['Price']

    # 2. Υπολογισμός RFM Metrics
    now_date = df['InvoiceDate'].max() + dt.timedelta(days=1)
    rfm = df.groupby('Customer ID').agg({
        'InvoiceDate': lambda x: (now_date - x.max()).days, # Recency
        'Invoice': 'nunique',                               # Frequency
        'TotalPrice': 'sum'                                 # Monetary
    })
    rfm.columns = ['Recency', 'Frequency', 'Monetary']

    # 3. Λογαριθμικός Μετασχηματισμός
```

```
    self.rfm_log = np.log1p(rfm)
```

8.2.1. Τυποποίηση Δεδομένων (Z-Score Normalization) και Προετοιμασία Heatmap

Προκειμένου να αποφευχθεί η κυριαρχία των μεταβλητών με μεγάλο εύρος τιμών, εφαρμόζεται η μέθοδος **StandardScaler**. Ο κώδικας μετατρέπει τις τιμές σε Z-Scores, οι οποίες στη συνέχεια χρησιμοποιούνται για την ορθή χρωματική απεικόνιση των Heatmaps.

Πηγαίος Κώδικας 2α: Υλοποίηση StandardScaler

```
def prepare_zscore_data(self):
    """
    Εφαρμόζει τυποποίηση StandardScaler (Z-Score Normalization).
    Οι τιμές μετατρέπονται σε κλίμακα με μέση τιμή 0 και τυπική απόκλιση
1.
    """
    scaler = StandardScaler()
    scaled_array = scaler.fit_transform(self.rfm_log)

    # Αποθήκευση των τυποποιημένων δεδομένων σε νέο DataFrame
    self.rfm_scaled = pd.DataFrame(scaled_array,
index=self.rfm_log.index, columns=self.rfm_log.columns)
```

8.3. Υλοποίηση Αλγορίθμου K-Means και Βελτιστοποίηση

Ο κώδικας για τον αλγόριθμο **K-Means** περιλαμβάνει τον υπολογισμό του βέλτιστου K μέσω της μεθόδου του "**Αγκώνα**" (Elbow Method) και **γεωμετρικό υπολογισμό** του **σημείου καμπής**.

Πηγαίος Κώδικας 3: Εκτέλεση K-Means και Εύρεση Βέλτιστου K

```
def kmeans_thread_metrics(self):
    """
    Εκτελεί τον αλγόριθμο K-Means για ένα εύρος τιμών K (2 έως 10)
    και υπολογίζει το Inertia (SSE) για τη μέθοδο Elbow.
    """
```

```
    X_clean = self.get_clean_data()
    metrics = []
    K_range = range(2, 11)

    for k in K_range:
        # Χρήση K-Means++ για βέλτιστη αρχικοποίηση κέντρων
        km = KMeans(n_clusters=k, init='k-means++',
random_state=self.current_seed)
        km.fit(X_clean)

        # Αποθήκευση του Sum of Squared Errors (Inertia)
        metrics.append(km.inertia_)

    # ... (κώδικας οπτικοποίησης διαγράμματος) ...

    # Γεωμετρικός Υπολογισμός του "Αγκώνα" (Elbow Point)
    # Βρίσκουμε το σημείο που έχει τη μέγιστη απόσταση από την
ευθεία που συνδέει
    # το πρώτο και το τελευταίο σημείο της καμπύλης.
    p1_x, p1_y = K_range[0], metrics[0]
    p2_x, p2_y = K_range[-1], metrics[-1]

    max_dist = 0
    best_k = K_range[0]

    for i in range(len(K_range)):
        x0, y0 = K_range[i], metrics[i]
        numerator = abs((p2_y - p1_y) * x0 - (p2_x - p1_x) * y0 +
p2_x * p1_y - p2_y * p1_x)
        denominator = ((p2_y - p1_y)**2 + (p2_x - p1_x)**2)**0.5
        dist = numerator / denominator

        if dist > max_dist:
            max_dist = dist
            best_k = x0

    print(f"System Suggestion based on Elbow Method: K={best_k}")
```

8.4. Αλγόριθμος DBSCAN και Ανίχνευση Θορύβου

Σε αντίθεση με τον **K-Means**, ο **DBSCAN** απαιτεί τον υπολογισμό του γραφήματος k-αποστάσεων (k-distance graph) για την εύρεση της παραμέτρου Epsilon.

Πηγαίος Κώδικας 4: DBSCAN και k-Distance Graph

```
def dbscan_thread_kdist(self):
    """
    Υπολογίζει το k-Distance Graph για να βοηθήσει τον χρήστη
    να επιλέξει τη βέλτιστη τιμή της παραμέτρου Epsilon (eps).
    """
    X_clean = self.get_clean_data()

    # Χρήση NearestNeighbors για τον υπολογισμό αποστάσεων
    k_val = 4 # Τυπική τιμή για 2-διάστατα δεδομένα, προσαρμόζεται
    neighbors = NearestNeighbors(n_neighbors=k_val)
    neighbors_fit = neighbors.fit(X_clean)
    distances, _ = neighbors_fit.kneighbors(X_clean)

    # Ταξινόμηση αποστάσεων για το γράφημα
    self.temp_dists = np.sort(distances[:, k_val-1], axis=0)

    # ... (εμφάνιση γραφήματος) ...

def dbscan_thread_final(self):
    """
    Εκτελεί τον DBSCAN με τις παραμέτρους που επέλεξε ο χρήστης.
    """
    X_clean = self.get_clean_data()

    # Εκτέλεση αλγορίθμου
    db = DBSCAN(eps=self.temp_eps, min_samples=self.temp_min_s)
    clusters = db.fit_predict(X_clean)

    # Στατιστικά αποτελέσματα
    n_clusters = len(set(clusters)) - (1 if -1 in clusters else 0)
    n_noise = list(clusters).count(-1)

    print(f"Clusters found: {n_clusters}")
    print(f"Noise points detected: {n_noise}")
```

8.5. Αυτοματοποιημένη Βελτιστοποίηση (Auto – K Intelligence)

Το συγκεκριμένο τμήμα κώδικα υλοποιεί την καινοτόμο λειτουργία της "**Μαζικής Αναζήτησης**" (Grid Search) για την εύρεση του βέλτιστου **Seed** και του βέλτιστου αριθμού

συστάδων, συνδυάζοντας **πολλαπλές μετρικές** (Silhouette, Calinski-Harabasz, Davies-Bouldin).

Πηγαίος Κώδικας 5: Αλγόριθμος Βελτιστοποίησης Παραμέτρων

```
def worker_logic(max_k_val, raw_sample_pct):
    """
    Εκτελεί μαζικές δοκιμές για K=2 έως max_k_val και υπολογίζει
    τρεις διαφορετικές μετρικές ποιότητας για κάθε K.
    """
    results = []

    for k in range(2, max_k_val + 1):
        # Εκπαίδευση K-Means
        km = KMeans(n_clusters=k, init='k-means++', n_init=3,
random_state=42).fit(X_final)

        # Υπολογισμός πολλαπλών μετρικών
        s = silhouette_score(X_final, km.labels_)            #
Maximize
        c = calinski_harabasz_score(X_final, km.labels_)     #
Maximize
        d = davies_bouldin_score(X_final, km.labels_)        #
Minimize
        ine = km.inertia_                                    # Elbow

        results.append({'k': k, 'ine': ine, 'sil': s, 'cal': c,
'dav': d})

    # Σύστημα Ψηφοφορίας (Voting System) για την τελική πρόταση
    b_sil = max(results, key=lambda x: x['sil'])['k'] # Καλύτερο
Silhouette
    b_cal = max(results, key=lambda x: x['cal'])['k'] # Καλύτερο
Calinski
    b_dav = min(results, key=lambda x: x['dav'])['k'] # Καλύτερο
Davies

    # Η τελική πρόταση προκύπτει από την πλειοψηφία των μετρικών
    votes = [b_sil, b_cal, b_dav]
    suggest = max(Counter(votes), key=Counter(votes).get)

    print(f"🏆 RECOMMENDED K: {suggest}")
```

8.6. Οπτικοποίηση Αποτελεσμάτων (Heatmaps) και Cluster Profiling

Τέλος, παρουσιάζεται ο κώδικας που δημιουργεί τα **Heatmaps** για την ερμηνεία των συστάδων (Cluster Profiling), χρησιμοποιώντας τη βιβλιοθήκη **seaborn**, έτσι, η μετάβαση στην **επιχειρηματική γνώση** υλοποιείται μέσω της **οπτικοποίησης** των **Z-Scores** και της παραγωγής **συγκεντρωτικών πινάκων**.

Πηγαίος Κώδικας 6: Δημιουργία Heatmap Συστάδων και Στατική Αναφορά

```
def show_results_visuals(self, k, algo_name):
    """
    Δημιουργεί ένα Heatmap που δείχνει τα μέσα Z-Scores για κάθε συστάδα,
    επιτρέποντας την εύκολη ερμηνεία των χαρακτηριστικών (π.χ. High Monetary).
    """
    # Υπολογισμός μέσων τιμών (centroids) ανά Cluster
    centers = self.rfm_scaled.groupby('Cluster').mean()
    centers.index = [f'Cluster {i+1}' for i in range(k)]

    # Ρυθμίσεις Γραφήματος
    plt.figure(figsize=(10, 6))
    sns.heatmap(centers, annot=True, fmt='.2f', cmap='coolwarm', center=0)

    plt.title(f'Cluster Centroids Heatmap ({algo_name})', fontweight='bold')
    plt.xlabel('RFM Features')
    plt.ylabel('Clusters')
    plt.show()
```

8.7. Συγκριτική Αξιολόγηση Αλγορίθμων (Benchmarking)

Το παρακάτω τμήμα κώδικα υλοποιεί τη διαδικασία άμεσης σύγκρισης μεταξύ των αλγορίθμων (K-Means, K-Means++, K-Medoids, DBSCAN). Καταγράφει τόσο τον χρόνο εκτέλεσης (Execution Time) όσο και την ποιότητα της συσταδοποίησης (Silhouette Score).

Πηγαίος Κώδικας 7: Διαδικασία Benchmarking

```
def comparison_final_thread(self):
    """
    Εκτελεί παράλληλα όλους τους επιλεγμένους αλγορίθμους για το
ίδιο K
    και συγκρίνει την απόδοσή τους.
    """
    results = []
    target_k = self.comp_k
    X_clean = self.get_clean_data()

    # --- 1. K-Means (Random Initialization) ---
    t0 = time.time()
    km = KMeans(n_clusters=target_k, init='random',
random_state=self.current_seed)
    labels = km.fit_predict(X_clean)
    t1 = time.time()
    results.append(("K-Means", t1-t0, silhouette_score(X_clean,
labels)))

    # --- 2. K-Means++ (Smart Initialization) ---
    t0 = time.time()
    km_pp = KMeans(n_clusters=target_k, init='k-means++',
random_state=self.current_seed)
    labels = km_pp.fit_predict(X_clean)
    t1 = time.time()
    results.append(("K-Means++", t1-t0, silhouette_score(X_clean,
labels)))

    # --- 3. K-Medoids (PAM - Partitioning Around Medoids) ---
    try:
        from sklearn_extra.cluster import KMedoids
        t0 = time.time()
        kmed = KMedoids(n_clusters=target_k, method='pam',
random_state=self.current_seed)
        labels = kmed.fit_predict(X_clean)
        t1 = time.time()
        results.append(("K-Medoids", t1-t0,
silhouette_score(X_clean, labels)))
    except ImportError:
        pass # Παράλειψη αν δεν υπάρχει η βιβλιοθήκη

    # --- 4. DBSCAN (Auto-Tuning για να βρει το ίδιο πλήθος ομάδων)
---
```

```
    # Ο DBSCAN δεν δέχεται K, οπότε δοκιμάζουμε δυναμικά διάφορα
Epsilon
    # μέχρι να βρούμε αυτό που παράγει τον επιθυμητό αριθμό
συστάδων (target_k).
    if self.comp_mode == 'all':
        best_match = None
        min_diff = float('inf')

        for eps in np.arange(0.1, 5.0, 0.1):
            db = DBSCAN(eps=eps, min_samples=4)
            labels = db.fit_predict(X_clean)

            # Υπολογισμός αριθμού συστάδων (αγνοώντας τον θόρυβο -
1)
            n_found = len(set(labels)) - (1 if -1 in labels else 0)

            if n_found == target_k:
                score = silhouette_score(X_clean, labels)
                best_match = (f"DBSCAN (eps={eps:.1f})", 0, score)
                break # Βρέθηκε η βέλτιστη αντιστοιχία

        if best_match:
            results.append(best_match)

    # Εμφάνιση αποτελεσμάτων σε ραβδογράμματα (Bar Charts)
    self.show_comparison_charts(results)
```

8.8. Ανάλυση Ευστάθειας και Εύρεση Seed (Stability Analysis)

Ένα συχνό πρόβλημα στον **K-Means** είναι ότι τα αποτελέσματα αλλάζουν ανάλογα με την αρχική τοποθέτηση των κέντρων. Ο παρακάτω κώδικας υλοποιεί έναν "**Seed Hunter**" που τρέχει χιλιάδες επαναλήψεις για να βρει την πιο ευσταθή και αποδοτική λύση.

Πηγαίος Κώδικας 8: Έλεγχος Ευστάθειας (Seed Optimization)

```
def worker_logic(k, n_trials, sample_pct):
    """
    Εκτελεί πολλαπλές δοκιμές (Monte Carlo Simulation) με
διαφορετικά random seeds
    για να εντοπίσει την αρχικοποίηση που μεγιστοποιεί το
Silhouette Score.
    """
    X = self.get_clean_data()
```

```
    results = []

    # Δειγματοληψία για επιτάχυνση (αν ορίστηκε από τον χρήστη)
    if sample_pct < 1.0:
        X_final = sklearn.utils.resample(X,
n_samples=int(len(X)*sample_pct), random_state=42)
    else:
        X_final = X

    print(f"Starting Stability Analysis for K={k} with {n_trials}
trials...")

    for i in range(n_trials):
        # Γεννήτρια τυχαίου Seed
        seed = np.random.randint(1, 999999)

        # Εκτέλεση με n_init=1 για έλεγχο της συγκεκριμένης
αρχικοποίησης
        km = KMeans(n_clusters=k, init='k-means++', n_init=1,
random_state=seed)
        km.fit(X_final)

        # Καταγραφή μετρικών
        sil = silhouette_score(X_final, km.labels_)
        inertia = km.inertia_

        results.append({'seed': seed, 'score': sil, 'inertia':
inertia})

    # Ταξινόμηση αποτελεσμάτων
    df_results = pd.DataFrame(results).sort_values(by='score',
ascending=False)
    best_run = df_results.iloc[0]

    # Στατιστικά Ευστάθειας
    std_dev = df_results['score'].std() # Τυπική απόκλιση (χαμηλή =
μεγάλη ευστάθεια)

    print(f"🏆 Best Seed Found: {int(best_run['seed'])}")
    print(f"💎 Max Silhouette Score: {best_run['score']:.4f}")
    print(f"📉 Score Volatility (Std Dev): {std_dev:.6f}")
```

8.9. Εξαγωγή Αποτελεσμάτων (Data Activation)

Για την πρακτική αξιοποίηση, ο κώδικας εξάγει την τμηματοποιημένη βάση πελατών, επιτρέποντας την άμεση χρήση των αποτελεσμάτων σε καμπάνιες Marketing.

Πηγαίος Κώδικας 9: Export Module

```
def export_to_csv(self, filename):
    # Εξαγωγή του rfm_original που περιέχει πλέον τη στήλη 'Cluster'
    self.rfm_original.to_csv(filename, index=True)
    print(f"✅ Επιτυχής εξαγωγή για Marketing Activation: {filename}")
```

Με την παράθεση των παραπάνω αποσπασμάτων, καλύπτεται πλήρως το τεχνικό μέρος της υλοποίησης. Ο κώδικας αποδεικνύει ότι η εργασία δεν βασίστηκε σε έτοιμα αποτελέσματα, αλλά σε μια στιβαρή διαδικασία βελτιστοποίησης και ελέγχου.

Σημείωση: Ο πλήρης κώδικας της διεπαφής χρήστη (GUI) παραλείπεται για λόγους συντομίας, καθώς δεν άπτεται άμεσα των ερευνητικών ερωτημάτων της ομαδοποίησης.

9. Βιβλιογραφία και Αναφορές

Για την Εργασία που παρουσιάζεται στο παρών συγγραφικό έργο, χρησιμοποιήθηκαν Πληροφορίες και Δεδομένα που αντλήθηκαν από τις παρακάτω Βιβλιογραφίες - Αναφορές:

1. **Adomavicius, G., & Tuzhilin, A. (2005).** Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734-749.
2. **Anaconda Inc. (2020).** *Anaconda Software Distribution*. Retrieved from https://docs.anaconda.com/
3. **Arora, M., et al. (2016).** K-means v/s k-medoids: A comparative study. *Proceedings of International Conference on Communication and Networks*.
4. **Arthur, D., & Vassilvitskii, S. (2007).** k-means++: The advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 1027-1035.
5. **Bernstein, M. (2017).** *Customer Data Platforms: The Path to a Unified Customer View*. O'Reilly Media.
6. **Bishop, C. M. (2006).** *Pattern Recognition and Machine Learning*. Springer.
7. **Brynjolfsson, E., Hu, Y. J., & Simester, D. (2011).** Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science*, 57(8).
8. **Buttle, F., & Maklan, S. (2019).** *Customer Relationship Management: Concepts and Technologies*. Routledge.
9. **Chen, D., Sain, S. L., & Guo, K. (2012).** Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*.
10. **Creswell, J. W., & Creswell, J. D. (2017).** *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

11. **Dwivedi, Y. K., et al. (2021).** Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57, 101994.

12. **Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996).** A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD-96 Proceedings*, 226-231.

13. **Fan, S., Lau, R. Y., & Zhao, J. L. (2015).** Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix. *Big Data Research*.

14. **Gartner. (2020).** *Marketing Data and Analytics Survey*. Gartner Research.

15. **Gartner. (2021).** *Market Guide for Customer Data Platforms*. Gartner Research.

16. **Ghahramani, Z. (2004).** Unsupervised learning. *Advanced Lectures on Machine Learning*, 72-112.

17. **Ghasemaghaei, M. (2019).** Does data analytics work for everyone? An empirical investigation of big data quality. *Journal of Big Data Analytics*.

18. **Goldfarb, A., & Tucker, C. (2011).** Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3), 389-404.

19. **Gupta, S., et al. (2006).** Modeling customer lifetime value. *Journal of Service Research*, 9(2), 139-155.

20. **Hair Jr, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014).** *Multivariate data analysis*. Pearson Education.

21. **Haleem, A., Javaid, M., Qadri, M. A., & Suman, R. (2022).** Understanding the role of Big Data in Industry 4.0. *International Journal of Intelligent Networks*, 3, 169-178.

22. **Han, J., Pei, J., & Kamber, M. (2011).** *Data mining: concepts and techniques*. Elsevier.

23. **Hastie, T., Tibshirani, R., & Friedman, J. (2009).** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

24. **Hughes, A. M. (2005).** *Strategic Database Marketing*. McGraw-Hill.

25. **Jain, A. K. (2010).** Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666.

26. **Jordan, M. I., & Mitchell, T. M. (2015).** Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255-260.

27. **Kallweit, K., Spreer, P., & Toporowski, W. (2014).** Context is King? The Impact of Context on the Effectiveness of Mobile Promotions. *International Journal of Mobile Communications*.

28. **Kaufman, L., & Rousseeuw, P. J. (1990).** *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

29. **Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015).** *Fundamentals of Machine Learning for Predictive Data Analytics*. MIT Press.

30. **Khan, K., et al. (2014).** DBSCAN: Past, present and future. *The Fifth International Conference on the Applications of Digital Information and Web Technologies*.

31. **Kleppmann, M. (2017).** *Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems*. O'Reilly Media.

32. **Kou, G., Chao, X., Peng, Y., & Alsaadi, F. E. (2020).** Machine learning methods for systemic risk analysis in financial sectors. *Technological and Economic Development of Economy*.

33. **Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011).** Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3), 231-256.

34. **Kumar, V., & Reinartz, W. (2016).** Creating enduring customer value. *Journal of Marketing*, 80(6), 36-68.

35. **LeCun, Y., Bengio, Y., & Hinton, G. (2015).** Deep learning. *Nature*, 521(7553), 436-444.

36. **MacQueen, J. (1967).** Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(14), 281-297.

37. **Mazzia, F., et al. (2009).** *Customer segmentation: a comparative study*.

38. **McKinsey & Company. (2021).** *The value of getting personalization right—or wrong—is multiplying*.

39. **Microsoft. (2021).** *Visual Studio Code Documentation*. Retrieved from https://code.visualstudio.com/docs

40. **Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016).** The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2).

41. **Moussaouy, L., Kacemi, A., & Essaid, M. (2020).** From CRM to CDP: A Survey on Customer Data Platforms. *International Conference on Big Data and Internet of Things*.

42. **Ng, K., & Liu, H. (2000).** Customer retention via data mining. *Artificial Intelligence Review*, 14(6), 569-590.

43. **Park, H. S., & Jun, C. H. (2009).** A simple and fast algorithm for K-medoids clustering. *Expert systems with applications*.

44. **Payne, A., & Frow, P. (2005).** A strategic framework for customer relationship management. *Journal of Marketing*, 69(4), 167-176.

45. **Pedregosa, F., et al. (2011).** Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

46. **Portugal, I., Alencar, P., & Cowan, D. (2018).** The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, 97, 205-227.

47. **Provost, F., & Fawcett, T. (2013).** *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.

48. **Raab, D. (2016).** *Customer Data Platforms: Definition, Context, and Benefits*. CDP Institute.

49. **Raab, D. (2019).** *Customer Data Platform Industry Update*. CDP Institute.

50. **Raschka, S., & Mirjalili, V. (2019).** *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*. Packt Publishing.

51. **Rousseeuw, P. J. (1987).** Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53-65.

52. **Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017).** DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*.

53. **Sutton, R. S., & Barto, A. G. (2018).** *Reinforcement Learning: An Introduction*. MIT press.

54. **Thiprungsri, P., & Vasarhelyi, M. A. (2011).** Cluster analysis for anomaly detection in accounting data: An audit approach. *The International Journal of Digital Accounting Research*.

55. **Tsiptsis, K., & Chorianopoulos, A. (2009).** *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons.

56. **Van Rossum, G., & Drake, F. L. (2009).** *Python 3 Reference Manual*. CreateSpace.

57. **VanderPlas, J. (2016).** *Python data science handbook: Essential tools for working with data*. O'Reilly Media.

58. **Venkatesan, R., & Kumar, V. (2004).** A customer lifetime value framework for customer selection and resource allocation strategy. *Journal of Marketing*, 68(4), 106-125.

59. **Verhoef, P. C., Kannan, P. K., & Inman, J. J. (2015).** From multi-channel retailing to omni-channel retailing: Introduction to the special issue on multi-channel retailing. *Journal of Retailing*, 91(2), 174-181.

60. **Vetri Selvi, M., et al. (2020).** Customer View: A 360-Degree Approach. *Journal of Big Data Analytics*.

61. **Voigt, P., & Von dem Bussche, A. (2017).** *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer.

62. **Xu, R., & Wunsch, D. (2005).** Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 645-678.

63. **Zahay, D. (2021).** *Digital Marketing Management: A Handbook for the Current Marketing Environment*. Kogan Page Publishers.

64. **Zahay, D., et al. (2019).** Data-driven personalization in the omnichannel environment. *Journal of Research in Interactive Marketing*.