
Πρόγραμμα Μεταπτυχιακών Σπουδών

στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

Διπλωματική Εργασία

# Περίπτωση ανάλυσης RFM σε βιομηχανική επιχείρηση

του

Σγούρα Γεωργίου του Στεφάνου

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Ιούλιος 2023

# Περίληψη

Η παρούσα έρευνα αφορά μία μεγάλη βιομηχανική εταιρεία κατασκευής ανελκυστήρων η οποία έχει τόσο τυποποιημένα προϊόντα τα οποία είναι make to order, αλλά και ειδικά προϊόντα που δεν είναι τυποποιημένα και είναι make to engineer. Το δίκτυο πωλήσεων της εταιρείας εξυπηρετεί πάνω από 100 χώρες και είναι μία από τις μεγαλύτερες εταιρείες στον κλάδο της.

Η έρευνα στοχεύει στην ομαδοποίηση των πελατών της εταιρείας, με βάση κάποια κοινά χαρακτηριστικά τους, ώστε να προταθούν στρατηγικές μάρκετινγκ για τη σωστή διαχείριση τους.

Η ανάλυση των δεδομένων που αφορούν 54 μήνες, πραγματοποιήθηκε σε δύο στάδια. Στο πρώτο στάδιο επιλέξαμε τη μεθοδολογία RFM Analysis ώστε να γίνει κατηγοριοποίηση των πελατών βάσει των χαρακτηριστικών Recency, Frequency και Monetary. Με το Recency βλέπουμε πότε ήταν η τελευταία συναλλαγή για κάθε πελάτη στο χρονικό πλαίσιο που εξετάζουμε. Με το Frequency, βλέπουμε τη συχνότητα των συναλλαγών που πραγματοποίησε ο κάθε πελάτης. Τέλος με το Monetary βλέπουμε το ύψος της αξίας των συναλλαγών που πραγματοποίησε ο κάθε πελάτης. Χρησιμοποιώντας αυτά τα χαρακτηριστικά μπορούμε να έχουμε μία πρώτη εικόνα για τις κατηγορίες που προκύπτουν.

Στο δεύτερο στάδιο κάνουμε cluster analysis με τον αλγόριθμο k-means. Λαμβάνοντας υπόψιν τα παραπάνω χαρακτηριστικά από την RFM Analysis, βρήκαμε κοινά μοτίβα των πελατών της εταιρείας και βάσει αυτών δημιουργήσαμε ομάδες (clusters). Στη cluster analysis θέλαμε να υπάρχει ομοιογένεια των δεδομένων εντός των cluster που δημιουργούνται αλλά ταυτόχρονα να υπάρχει ετερογένεια μεταξύ των διαφορετικών ομάδων ώστε κάθε cluster να συγκεντρώνει δεδομένα πελατών με διαφορετικά προφίλ. Οι ομάδες που προέκυψαν ελέγχθηκαν με την μετρική του Silhouette score για να επικυρώσουμε την ποιότητα τους.

Από το συνδυασμό αυτών των δύο σταδίων καταλήξαμε σε τέσσερις κατηγορίες πελατών και προτείναμε διαφορετικές στρατηγικές μάρκετινγκ για κάθε ομάδα, ώστε να είναι αποδοτικότερος ο τρόπος προσέγγισης του κάθε πελάτη.

## Abstract

The present study is about a large industrial company which manufactures elevators. The company offers both standardized products (make to order), but also special products (make to engineer). The company's sales network serves over 100 countries globally and is one of the largest companies in its industry.

The research conducted focuses on grouping the customers based on the common characteristics of the customers to provide efficient marketing strategies for them.

We examined transactional data through 54 months and the analysis was carried out in two stages. In the first stage we chose RFM analysis methodology to categorize customers based on Recency, Frequency and Monetary values. With recency we take into consideration the last transaction for each customer, with frequency we examine the frequency of the transactions and with monetary we examine the overall value of the transactions for each customer within the time frame we have gathered data. Considering all the characteristics mentioned above, we come down to the categories of the customers in our data set.

In the second stage we chose cluster analysis with the K-Means algorithm. By considering the above features we already had from RFM Analysis, we found common patterns that created customers' clusters. With the cluster analysis we wanted to have homogeneity of data points within each of the cluster created, but at the same time we wanted to have heterogeneity among different clusters so that each cluster would have customers with different profiles. The validity of the clusters was validated with the metric of Silhouette score to make sure we have high quality clusters.

By combining those two stages we created four different clusters of customers, and we proposed different marketing strategies for each group so we can achieve more efficient ways to approach each customer.

# Περιεχόμενα


1 Εισαγωγή ..... 1
2 Σκοπός ..... 3
3 Ανασκόπηση βιβλιογραφίας ..... 4
3.1 Κατηγοριοποίηση πελατών ..... 4
3.2 Ανάλυση RFM ..... 6
3.3 Clustering ..... 8
3.3.1 Hierarchical Clustering ..... 10
3.3.1.1 Agglomerative hierarchical clustering ..... 11
3.3.1.2 Divisive hierarchical clustering ..... 12
3.3.2 Partitional Clustering ..... 13
3.3.2.1 Hard/Crisp Clustering ..... 15
3.3.2.2 Miscellaneous Clustering ..... 19
3.3.2.3 Mixture Resolving clustering ..... 21
3.3.2.4 Fuzzy Clustering ..... 22
3.3.2.5 Λοιπές κατηγορίες Clustering ..... 23
3.3.3 Ιδιότητες απόδοσης αλγορίθμων Clustering ..... 24
3.3.4 Θέματα προς διερεύνηση στους αλγόριθμους Clustering ..... 26
3.3.5 Μετρικές αξιολόγησης για Clusters ..... 29
3.4 Ανάλυση RFM & K-Means Clustering (Case Studies) ..... 32
3.5 Στρατηγικός σχεδιασμός Μάρκετινγκ ..... 34
3.6 Στάδια ζωής πελάτη ..... 36
3.7 Τμηματοποίηση αγοράς στον κλάδο του ανελκυστήρα - ανυψωτικών ..... 39
4 Μεθοδολογία Έρευνας ..... 46
4.1 Επεξήγηση ερευνητικής διαδικασίας ..... 48
4.2 Ανάλυση δείγματος της έρευνας ..... 48
4.3 Εισαγωγή δεδομένων για επεξεργασία σε Python ..... 56
5 Αποτελέσματα ..... 58
5.1 Υπολογισμός Recency, Frequency, Monetary ..... 58
5.2 Προ - επεξεργασία δεδομένων για Cluster Analysis ..... 69
5.3 Clustering με αλγόριθμο K-Means ..... 72
5.4 Συνδυασμός αποτελεσμάτων RFM & K - Means ..... 75
6 Συμπεράσματα ..... 84
7 Περιορισμοί – Προτάσεις για μελλοντική έρευνα ..... 95
8 Βιβλιογραφία ..... 96

# 1 Εισαγωγή

Η παρούσα έρευνα αφορά βιομηχανική εταιρεία που κατασκευάζει ανελκυστήρες, ανυψωτικά και κυλιόμενους διαδρόμους και αποτελεί μία από τις σημαντικότερες επιχειρήσεις του κλάδου στην παγκόσμια αγορά, με παραγωγή κάθε τύπου ανελκυστήρα για οικιακή, εμπορική, βιομηχανική χρήση, για πρόσωπα ή φορτία. Επίσης στην γκάμα των προϊόντων της, η εταιρεία έχει κυλιόμενες σκάλες, κυλιόμενους διαδρόμους, συστήματα πρόσβασης, συστήματα στάθμευσης, πλατφόρμες, ανελκυστήρες πλοίων, μεμονωμένα υποσυστήματα ανελκυστήρα και ανταλλακτικά και διανέμει τα προϊόντα της σε περισσότερες από 100 χώρες παγκοσμίως.

Το πελατολόγιο της εταιρείας αποτελείται από άλλες εταιρείες εγκατάστασης και συντήρησης (B2B πελατολόγιο) οι οποίες διανέμουν και εγκαθιστούν τα προϊόντα της σε όλο τον κόσμο. Η δομημένη συνεργασία της εταιρείας και των συνεργατών της, εξασφαλίζει την ανάπτυξη ισχυρών μακροχρόνιων δεσμών με τους πελάτες και ταυτόχρονα την παροχή αξιόπιστων υπηρεσιών σε παγκόσμιο επίπεδο.

Με στρατηγικές επενδύσεις, η επιχείρηση έχει προχωρήσει στην εξειδίκευση της παραγωγής και την εξαγωγή προϊόντων σε πολλαπλές αγορές, αναπτύσσοντας το δίκτυο, την διεθνή παρουσία και την γκάμα προϊόντων που προσφέρει.

Οι δύο κύριες κατηγορίες που εστιάζουν οι πωλήσεις και εξυπηρετείται από τη δομή της Εμπορικής Διεύθυνσης, είναι οι πωλήσεις προϊόντων στην ελληνική αγορά που έχει και ηγετική θέση και οι πωλήσεις προϊόντων εκτός Ελλάδος. Από το έτος ίδρυσης της εταιρείας το 1983 μέχρι και το 2004 πάνω από το 90% των πωλήσεων αφορούσε την ελληνική αγορά. Η εταιρεία παρουσίαζε σταδιακή ανάπτυξη κάθε χρόνο, αλλά με την οικονομική κρίση που ξεκίνησε το 2008 οι πωλήσεις που αφορούσαν την Ελλάδα άρχισαν να παρουσιάζουν σημαντική επιβράδυνση. Για να μπορέσει να ανταπεξέλθει και να συνεχίσει να αναπτύσσεται ξεκίνησε την έντονη δραστηριοποίηση εκτός Ελλάδος με τη δημιουργία παραγωγικών μονάδων σε πέντε διαφορετικές χώρες, ενώ ταυτόχρονα άρχισε να δημιουργεί θυγατρικές εταιρίες εμπορικού χαρακτήρα και σε άλλες χώρες με σκοπό την ανάπτυξη της, ως όμιλος εταιρειών. Από το 2008 η εταιρεία έχει αλλάξει την προσέγγιση της όσο αφορά τη στόχευση των πωλήσεων υιοθετώντας ένα έντονο εξαγωγικό χαρακτήρα. Αυτό είναι κάτι που έχει πετύχει δεδομένου ότι πλέον πάνω από το 90% των πωλήσεων της επιχείρησης προέρχονται από πωλήσεις εκτός Ελλάδος.

Η ιδιαιτερότητα που έχει η εξεταζόμενη επιχείρηση, είναι ότι αφορά μία βιομηχανία η οποία προσφέρει μία γκάμα προϊόντων στους πελάτες της, αλλά λόγω της φύσης του προϊόντος, δεν μπορεί να θεωρηθεί πλήρως τυποποιημένο. Τα περισσότερα προϊόντα που πωλούνται μπορούμε να πούμε ότι είναι "make to order" ακολουθώντας δηλαδή κάποιες βασικές προδιαγραφές, αλλά υπάρχουν περιπτώσεις που αυτές οι προδιαγραφές δεν μπορούν να καλύψουν τις ανάγκες κάποιων περιπτώσεων αναγκών του πελάτη. Σε αυτή την περίπτωση ακολουθείται η προσέγγιση "make to engineer" που σημαίνει ότι οι πωλήσεις λαμβάνουν κάποιες ειδικές προδιαγραφές που απαιτούνται από τον πελάτη και το τμήμα μελέτης καλείται να βρει μία νέα λύση η οποία θα μπορούσε να υιοθετηθεί στην εν λόγω περίπτωση. Αυτό σημαίνει ότι υπάρχει μεγάλο περιθώριο "customization" για τον πελάτη ανά περίπτωση, κάτι το οποίο αφενός αυξάνει το κόστος αλλά αφετέρου αποτελεί ένα ανταγωνιστικό πλεονέκτημα για την εταιρεία έναντι των ανταγωνιστών της η οποία μπορεί και πουλάει αυτά τα προϊόντα ως "premium" προϊόντα, καλύπτοντας ένα κενό της αγοράς. Ένα ακόμη ανταγωνιστικό πλεονέκτημα της εταιρείας είναι ο γρήγορος χρόνος παράδοσης κάτι το οποίο εκμεταλλεύεται για την κάλυψη των αναγκών των πελατών σε σχετικά σύντομο χρονικό διάστημα, ιδίως σε σχέση με τον ανταγωνισμό.

Πρόσφατα, η εταιρεία είδε την ανάγκη για μεγαλύτερο περιθώριο τυποποίησης κάποιων προϊόντων της, για να καλύψει τις ανάγκες μίας μερίδας της αγοράς η οποία αποζητάει φθηνότερα προϊόντα. Για το λόγο αυτό η εταιρεία προσφέρει μέσω ενός web portal κάποια τυποποιημένα προϊόντα που δεν επιδέχονται αλλαγών. Μέσω του portal κάποιος πελάτης μπορεί αφού ανοίξει έναν λογαριασμό στην πλατφόρμα, να εισάγει μόνος του κάποια βασικά στοιχεία και να επιλέξει από τα τυποποιημένα προϊόντα ποιο καλύπτει τις ανάγκες του κάνοντας κάποιο μικρό "configuration" μέσα στην πλατφόρμα. Με αυτό τον τρόπο η εταιρεία καταφέρνει να έχει πρόσβαση σε ένα επιπλέον μερίδιο της αγοράς που δεν ενδιαφέρεται για "high end" λύσεις χωρίς να χρειάζεται να απασχολεί σε μεγάλο βαθμό μηχανικούς και χωρίς να πρέπει να κάνει ειδικές παραγγελίες από προμηθευτές για υλικά του τελικού προϊόντος αφού η φιλοσοφία είναι ότι όλα τα υλικά, θα είναι διαθέσιμα για αυτές τις περιπτώσεις. Με αυτό τον τρόπο μπορεί να κατεβάσει αισθητά το κόστος παραγωγής και να αποκομίσει κέρδη από τον μεγαλύτερο όγκο συναλλαγών που μπορούν να έρθουν μέσω αυτής της οδού.

Στην περίπτωση της εταιρείας που εξετάζουμε, είδαμε ότι δεν υπάρχει πλήθος μελετών που να έχει ασχοληθεί με RFM ανάλυση & ομαδοποίηση με k-means σε κλάδο που απευθύνεται σε B2B αγοραστικό κοινό πουλώντας ένα βιομηχανικό προϊόν το οποίο να

έχει περιπτώσεις που να είναι “make to order” αλλά και περιπτώσεις “make to engineer”. Για το λόγο αυτό θεωρήσαμε ότι θα ήταν μία καλή περίπτωση για να γίνει η ομαδοποίηση των πελατών και να εξετάσουμε την πρακτική εφαρμογή βάσει των αποτελεσμάτων που θα προκύψουν.

## 2 Σκοπός

Ένας από τους εντονότερους προβληματισμούς σε κάθε εταιρεία όταν σχεδιάζει το ετήσιο action plan πωλήσεων και το budget της, είναι ποιους νέους πελάτες - αγορές θα πρέπει να προσεγγίσει αλλά και ποιες στρατηγικές μάρκετινγκ θα πρέπει να χρησιμοποιήσει για την διατήρηση του υφιστάμενου πελατολογίου της. Γνωρίζουμε ότι η απόκτηση ενός νέου πελάτη είναι πιο δαπανηρή από την διακράτηση ενός υφιστάμενου πελάτη οπότε δίνεται πολύ μεγάλη σημασία στο υφιστάμενο πελατολόγιο. Για να μπορέσουμε να προτείνουμε στρατηγικές μάρκετινγκ θα πρέπει να γίνει μία ομαδοποίηση με κάποια κοινά χαρακτηριστικά πελατών ώστε να αναγνωρίσουμε μοτίβα καταναλωτικής συμπεριφοράς και να μπορέσουμε να προτείνουμε με περισσότερη επιτυχία συγκεκριμένες στρατηγικές μάρκετινγκ για συγκεκριμένες ομάδες πελατών με παρόμοια χαρακτηριστικά.

Η ανάλυση που θα κάνουμε έχει ως σκοπό την ομαδοποίηση των πελατών της συνολικά. Για αυτό το λόγο επιλέξαμε να χρησιμοποιήσουμε ένα μεγάλο εύρος δεδομένων με στοιχεία συναλλαγών από το 2017 έως το 2022. Αυτή η προσέγγιση θα αποτελέσει και τη βάση για το σχεδιασμό του action plan πωλήσεων της Εμπορικής Διεύθυνσης, αφού θα υπάρχει η δυνατότητα να εφαρμόσουν τις πλέον ενδεδειγμένες στρατηγικές μάρκετινγκ ανάλογα με την κάθε περίπτωση της ομάδας πελατών, η οποία θα προκύπτει με βάση τα κοινά χαρακτηριστικά που προκύπτουν μέσω κάποιων μοτίβων που καλούμαστε να αναγνωρίσουμε. Με αυτό τον τρόπο θα βοηθήσουμε την εταιρεία να κατανείμει πιο σωστά τους πόρους της, ώστε να αποκομίσει το μεγαλύτερο δυνατό όφελος με το μικρότερο δυνατό κόστος.

# 3 Ανασκόπηση βιβλιογραφίας

Στη βιβλιογραφία στα κεφάλαια που ακολουθούν, θα δούμε την κατηγοριοποίηση των πελατών, την ανάλυση RFM, την ανάλυση με Clusters και τις κατηγορίες – ιδιότητες της, τις στρατηγικές μάρκετινγκ με βάση τα στάδια ζωής του πελάτη, κάποιες πρακτικές εφαρμογές της RFM και Cluster ανάλυσης, και τέλος θα δούμε την τμηματοποίηση της αγοράς στον κλάδο του ανελκυστήρα με βάση μία πρόσφατη μελέτη.

## 3.1 Κατηγοριοποίηση πελατών

Οι (Jiang & Tuzhilin, 2009) αναφέρουν ότι η ομαδοποίηση των πελατών και η σωστή στόχευση στο σωστό "target group" αποτελούν μία λογική ακολουθία αλλά το πρόβλημα συνήθως εντοπίζεται στον σωστό συνδυασμό τους. Η προσέγγιση που έχει προταθεί είναι η ομαδοποίηση με K-Classifiers αλγόριθμους με απώτερο σκοπό να διανείμουμε τους πόρους της επιχείρησης με τον πλέον αποδοτικό τρόπο στους πελάτες αξίας που συνεισφέρουν περισσότερο στην εταιρεία

Οι (He & Li, 2016) εστίασαν στη βελτίωση της αξίας του πελάτη κατά τη διάρκεια που η επιχείρηση έχει συναλλαγές μαζί του (CLV), για την ικανοποίηση του, μελετώντας την συμπεριφορά του. Κατέληξαν ότι οι πελάτες είναι διαφορετικοί μεταξύ τους οπότε κατανοούμε και ότι οι ανάγκες τους μπορεί να διαφέρουν. Για να μπορέσει μία επιχείρηση να προσφέρει την καλύτερη δυνατή εμπειρία πρέπει να τους κατηγοριοποιήσει ώστε να μπορέσει να αποκωδικοποιήσει τις προσδοκίες τους.

Ο (Zahrotun, 2017) χρησιμοποίησε δεδομένα από το σύστημα "Customer Relationship Management" (CRM) μίας επιχείρησης μελετώντας τις on-line συναλλαγές της για να αναγνωρίσει το προφίλ των πελατών και να τους ομαδοποιήσει χρησιμοποιώντας "Fuzzy C-Means Clustering Method". Με αυτό τον τρόπο αναγνώρισε τις ανάγκες τους σε κατηγορίες που ανέδειξε με σκοπό να χρησιμοποιήσει αυτή την προσέγγιση για την σωστή προσέγγιση τους και τη μεγιστοποίηση της κερδοφορίας της επιχείρησης.

Οι (Shah & Singh, 2012) πρότειναν έναν νέο αλγόριθμο για "clustering" ο οποίος είναι παρόμοιος με τους K-Means και K-medoids. Ο αλγόριθμός τους δεν παρέχει τη βέλτιστη λύση σε όλες τις περιπτώσεις αλλά παρατήρησαν ότι όσο ο αριθμός των "clusters"

αυξάνεται η νέα μέθοδος χρειάζεται λιγότερο χρόνο σε σχέση με τις παραδοσιακές μεθόδους που αναφέρθηκαν.

Οι (Sheshasaayee & Logeshwari, 2017) σχεδίασαν μία βελτιωμένη ενοποιημένη προσέγγιση με RFM ανάλυση σε συνδυασμό με την LTV (Life Time Value) μέθοδο. Η προσέγγιση τους έχει δύο φάσεις, με την πρώτη να είναι η στατιστική προσέγγιση και η δεύτερη η ομαδοποίηση. Στόχος τους ήταν να χρησιμοποιήσουν K-Means για "clustering" και μετά να χρησιμοποιήσουν "Neural Networks" για να βελτιώσουν την ομαδοποίηση που προκύπτει.

Οι πελάτες μιας επιχείρησης δεν είναι μεταξύ τους ίδιοι ως προς τις ανάγκες που έχουν και ως προς τη δυναμική τους. Αυτό σημαίνει ότι η επιχείρηση πρέπει να είναι σε θέση να κατανοήσει τις ανάγκες τους και να εστιάσει στους πελάτες υψηλής αξίας. Για να μπορέσει να το κάνει αυτό θα πρέπει να εφαρμόσει διαφορετικές στρατηγικές μάρκετινγκ στους πελάτες της κατανέμοντας τους πόρους της κατάλληλα ώστε να γίνει με τον πλέον αποδοτικό και αποτελεσματικό τρόπο (Huang, et al., 2009).

Στο βιβλίο των (Kotler & Armstrong , 2006) τονίζεται ότι η προσέγγιση νέων πελατών σε μία επιχείρηση είναι σημαντική, αλλά ακόμα πιο σημαντική είναι η διατήρηση του υπάρχοντος πελατολογίου της, αφού το να χάσουμε έναν πελάτη που έχουμε ήδη, σημαίνει ότι χάνουμε και όλη την αξία που θα έφερνε ο τελευταίος σε όλη τη διάρκεια του χρόνου που θα πραγματοποιούσε συναλλαγές. Καταλαβαίνουμε οπότε τη σημαντικότητα του να κατανοήσουμε τους πελάτες που έχουμε ήδη, ώστε να μπορέσουμε να τους κρατήσουμε για όσο μεγαλύτερο διάστημα μπορούμε με σκοπό να αποκομίσουμε το μεγαλύτερο δυνατό όφελος σε βάθος χρόνου (Customer Lifetime Value - CLV).

Οι συναλλαγές που πραγματοποιούνται σε μία εταιρεία, με την πάροδο του χρόνου δημιουργούν ένα μεγάλο όγκο δεδομένων οπότε καταλαβαίνουμε την αναγκαιότητα να χωρίσουμε τους πελάτες σε ομάδες (clusters) με κύριο γνώρισμα κάποια κοινά χαρακτηριστικά ανά ομάδα. Βασική προϋπόθεση είναι δηλαδή να υπάρχει ομοιογένεια εντός του ίδιου cluster αλλά ταυτόχρονα να υπάρχει σαφής διαφοροποίηση μεταξύ των clusters που έχουμε δημιουργήσει (Hung & Tsai, 2008).

## 3.2 Ανάλυση RFM

Μία προσέγγιση που μπορούμε να ακολουθήσουμε για την κατηγοριοποίηση των πελατών για τους οποίους κρατάμε αρχείο συναλλαγών βάσει της μοναδικής ταυτότητας τους (ID) είναι η RFM Analysis.

Στο βιβλίο του (Birant, 2011) βλέπουμε ότι η ανάλυση RFM αφορά την ανάλυση στοιχείων συναλλαγών πελατών με βάση τρία κύρια χαρακτηριστικά μέσα σε ένα προκαθορισμένο χρονικό πλαίσιο που έχουμε επιλέξει να εξετάσουμε το συνολικό δείγμα συναλλαγών μιας εταιρείας.

Το πρώτο χαρακτηριστικό είναι το "Recency" λαμβάνοντας υπόψη το χρονικό σημείο στο οποίο έγινε η τελευταία αγορά από τον πελάτη, δηλαδή βλέπουμε πόσος χρόνος έχει περάσει από τότε που ο πελάτης πραγματοποίησε την τελευταία του συναλλαγή. Συνήθως αυτό που παρακολουθούμε είναι η συναλλαγή αλλά σε άλλες παραλλαγές μπορούμε να παρακολουθήσουμε πότε επισκέφτηκε ο πελάτης τελευταία φορά τη σελίδα στο ίντερνετ ή την εφαρμογή της εταιρείας στη συσκευή τηλεφώνου, ανάλογα με το αντικείμενο του προϊόντος ή της υπηρεσίας που πουλάει η εξεταζόμενη επιχείρηση.

Το δεύτερο χαρακτηριστικό είναι το "Frequency" και υποδηλώνει την συχνότητα που ο πελάτης πραγματοποιεί συναλλαγές μέσα στο χρονικό πλαίσιο που εξετάζουμε. Οι πελάτες οι οποίοι κάνουν τακτικότερα συναλλαγές θεωρούνται και πιο πιστοί σε σχέση με άλλους πελάτες που δεν κάνουν συχνά συναλλαγές. Υπάρχει και περίπτωση να έχουμε πελάτες που έχουν κάνει μία μοναδική συναλλαγή και αποτελούν μία ξεχωριστή κατηγορία.

Το τρίτο χαρακτηριστικό που λαμβάνουμε υπόψη είναι το "Monetary" και υποδηλώνει το χρηματικό ποσό που ο πελάτης έχει ξοδέψει μέσα στο χρονικό πλαίσιο το οποίο έχουμε αποφασίσει να εξετάσουμε. Οι πελάτες οι οποίοι ξοδεύουν υψηλότερα ποσά σε σύγκριση με τους υπόλοιπους, αποτελούν και τους πελάτες που είναι και υψηλότερης αξίας για την εταιρεία.

Μέσα από την ανάλυση με αυτά τα τρία χαρακτηριστικά μπορούμε να κατηγοριοποιήσουμε τους πελάτες με το να τους χωρίσουμε σε επιμέρους ομάδες ώστε να δούμε πως θα μπορούσαμε να παρέχουμε προσωποποιημένες υπηρεσίες ανάλογα με το προφίλ τους και με τελικό σκοπό να έχουμε την θετική ανταπόκριση τους σε

επικείμενες προωθητικές ενέργειες της εταιρείας. Με αυτό τον τρόπο η εταιρεία αξιοποιεί αποδοτικότερα το χρηματικό ποσό που έχει διαθέσιμο για το σχεδιασμό μάρκετινγκ.

Αυτά τα τρία χαρακτηριστικά ("Recency", "Frequency", "Monetary") μπορούν να βαθμολογηθούν είτε με κλίμακα από 1 έως 3, είτε από 1 έως 4, είτε από 1 έως 5 χωρίζοντας τα στοιχεία μας συνήθως σε ίσα τμήματα αντίστοιχα. Όσο μεγαλύτερη η τιμή (score) των γνωρισμάτων τόσο καλύτερη θεωρούμε ότι είναι η αντίστοιχη περίπτωση. Με βάση τα score μπορούμε να αναγνωρίσουμε τους πολυτιμότερους για την επιχείρηση πελάτες (Miglautsch, 2002).

Από τα αποτελέσματα (score) που προκύπτουν ομαδοποιούμε τα προφίλ πελατών ανά κατηγορία με σκοπό να υιοθετήσουμε συγκεκριμένες στρατηγικές μάρκετινγκ. Για παράδειγμα σε αυτούς που χαρακτηρίζονται ως "Best Customers" δηλαδή αυτοί που έχουν υψηλό score και στις τρεις κατηγορίες είναι απαραίτητο να τους διατηρήσουμε στο πελατολόγιο μας. Σε αυτή τη κατηγορία οι στρατηγικές που μπορούμε να ακολουθήσουμε είναι τακτική και ποιοτική επικοινωνία ώστε να νιώθουν ξεχωριστοί, έγκαιρη ενημέρωση για νέα προϊόντα και επιλεκτικά κάποιες εκπτώσεις. Για τους "Spenders" αυτούς δηλαδή που έχουν υψηλό Monetary Score η στρατηγική που πρέπει να εφαρμόσουμε είναι να τους ωθήσουμε να κάνουν συχνότερα συναλλαγές. Αυτό μπορεί να γίνει δίνοντας έμφαση στην επικοινωνία μαζί τους δίνοντας τονίζοντας τα ανταγωνιστικά πλεονεκτήματα που διαθέτει η εταιρεία, ώστε να τους ωθήσουμε να συνεχίσουν να συνεργάζονται με την επιχείρηση. Για τους "Loyal Customers" αυτούς δηλαδή που πραγματοποιούν συναλλαγές συχνά, πρέπει να τους ωθήσουμε να αυξήσουν το ποσό που δαπανούν μέσω "bundling", "cross-selling" και "up-selling". Για τους αβέβαιους πελάτες η στρατηγική που θα μπορούσε να εφαρμοστεί είναι να είμαστε επιλεκτικοί σχετικά με το ποιους να προσεγγίσουμε πιο έντονα (πράγμα που σημαίνει να δαπανήσουμε περισσότερους πόρους με δομημένο τρόπο) εστιάζοντας σε αυτούς οι οποίοι είναι καινούριοι (Marcus, 1998).

Τα πλεονεκτήματα της RFM Analysis είναι ότι δεν έχει υψηλό κόστος και είναι εύκολο να ποσοτικοποιήσουμε τη συμπεριφορά των πελατών μέσα από αυτό το μοντέλο με τρόπο που είναι εύκολο να κατανοηθεί (Miglautsch, 2000), δεν χρειάζονται πάρα πολλές μεταβλητές για την υλοποίηση της, λειτουργεί καλύτερα όταν θέλουμε να στοχεύσουμε σε συγκεκριμένους πελάτες με συγκεκριμένο προφίλ και μπορούμε εύκολα να εντοπίσουμε τους πιο πολύτιμους πελάτες της επιχείρησης (Miglautsch, 2002).

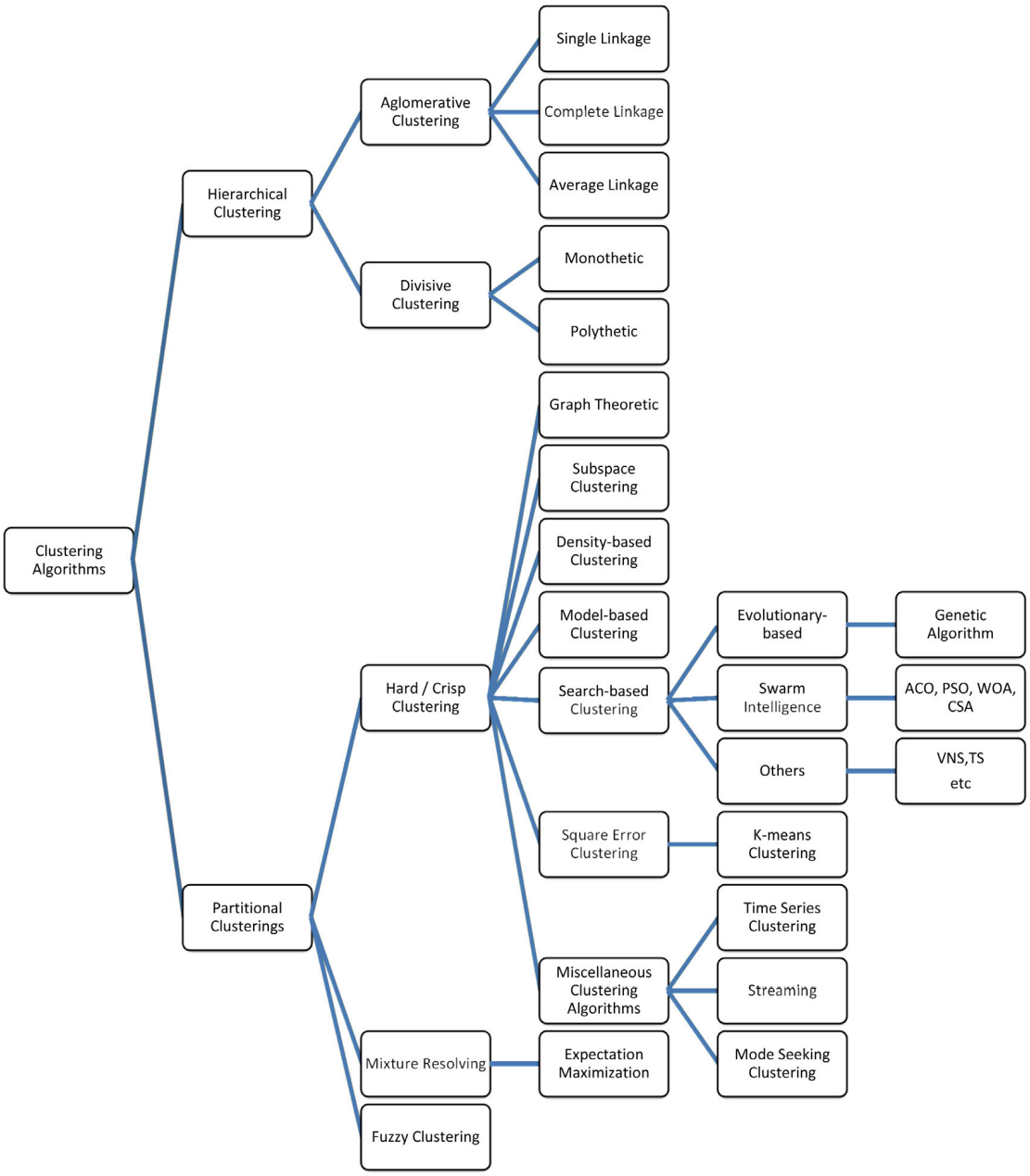
Τα μειονεκτήματα της RFM Analysis είναι ότι εστιάζει κυρίως στους καλύτερους πελάτες της επιχείρησης. Αυτό σημαίνει ότι δεν παρέχει πολλές πληροφορίες για τους πελάτες που έχουν καιρό να προβούν σε κάποια αγορά, με μικρή συχνότητα αγορών και χαμηλής αξίας συναλλαγές. Σε μία επιχείρηση συνήθως το 80% των πωλήσεων προκύπτουν από το 20% των πελατών (Wang, 2010). Αυτό σημαίνει ότι δεν δίνεται μεγάλη βαρύτητα σε μία μεγάλη μερίδα πελατών κάποιοι από τους οποίους είναι πιθανόν να έχουν δυνατότητες (Miglautsch, 2002). Η RFM Analysis εστιάζει σε υφιστάμενους πελάτες δηλαδή δεν μπορεί να εφαρμοστεί σε δυνητικούς νέους πελάτες αφού δεν έχουν ιστορικό συναλλαγών με την επιχείρηση και δεν μπορούμε να τους κατηγοριοποιήσουμε (McCarty & Hastak, 2007). Επίσης στην RFM ανάλυση, υποθέτουμε ότι υπάρχει μία σχετική ομοιογένεια στη βάση δεδομένων για τους πελάτες μας, πράγμα που στην πράξη δεν ισχύει τις περισσότερες φορές αφού παρατηρείται σημαντική ετερογένεια μεταξύ των πελατών (Suh, et al., 1999). Τέλος η RFM δεν αποτελεί ένα ακριβές μοντέλο ποσοτικοποίησης και η σημασία των RFM score δεν είναι ίδια σε όλους τους κλάδους (Yeh, et al., 2008).

## 3.3 Clustering

Το "Clustering" είναι μία τεχνική μη επιβλεπόμενης μηχανικής μάθησης που χρησιμοποιείται για να ομαδοποιήσει μη επισημασμένα ("unlabeled") δεδομένα που έχουν παρόμοια χαρακτηριστικά (Khan & Ahmad, 2004). Πολλοί αλγόριθμοι αυτού του τύπου μπορούν να χρησιμοποιήσουν μόνο κατηγορικά ή μόνο ποσοτικά δεδομένα (Witten & Frank, 2005). Στις περισσότερες περιπτώσεις πραγματικών πρακτικών εφαρμογών σε τομείς όπως το μάρκετινγκ, υγείας, χρηματοπιστωτικός, έχουμε μικτά σύνολα δεδομένων που περιλαμβάνουν κατηγορικά και ποσοτικά δεδομένα που θέλουμε να ομαδοποιήσουμε.

Αν θέλουμε να έχουμε μία συνολική ταξινόμηση των αλγορίθμων για ομαδοποίηση (clustering), μπορούμε να πούμε ότι έχουμε δύο κύριες κατηγορίες, την Ιεραρχική ομαδοποίηση (Hierarchical Clustering) και την Τμηματική ομαδοποίηση (Partitional Clustering). Για την "Hierarchical Clustering" οι κυριότερες υπό-κατηγορίες είναι η "Agglomerative Clustering" και η "Divisive Clustering". Για την κύρια κατηγορία "Partitional Clustering" έχουμε τις υποκατηγορίες "Hard or Crisp Clustering", "Fuzzy Clustering και "Mixture Resolving Clustering". Εκτενέστερα μπορούμε να δούμε κάποια

χαρακτηριστικά τους στη συνέχεια. Μία συνολική εικόνα για τις κατηγορίες “Clustering” μπορούμε να δούμε στην Εικόνα 1 (Anitha & Malini, 2019). Στη συνέχεια, θα επιχειρηθεί μία συνοπτική απεικόνιση των αλγορίθμων ομαδοποίησης μέσω μίας ταξινόμησης τους σε υποκατηγορίες για να έχουμε μία συνολική εικόνα για αυτούς, εξετάζοντας τα πεδία που χρησιμοποιούνται, τις δυνάμεις και αδυναμίες τους και μία περιγραφή των πιο σημαντικών από αυτούς.




*Εικόνα 1*

### 3.3.1 Hierarchical Clustering

Σε αυτή την κατηγορία της Ιεραρχικής ομαδοποίησης τα σημεία των δεδομένων χωρίζονται σε επίπεδα με ιεραρχική δομή. Οι ομάδες που σχηματίζονται είτε έχουν ιεραρχία από πάνω προς τα κάτω (Agglomerative method) είτε από κάτω προς τα πάνω (Divisive method), σχηματίζοντας ένα δέντρο-διάγραμμα. Αυτό επιτρέπει να εξερευνήσουμε τα δεδομένα σε διαφορετικό βαθμό λεπτομέρειας (Saxena, et al., 2017).

Στην "Agglomerative method" οι συστάδες που δημιουργούνται από μεμονωμένα σημεία, συγχωνεύονται μέσω μίας επαναληπτικής διαδικασίας σε μεγαλύτερες συστάδες σχηματίζοντας διαφορετικά επίπεδα ιεραρχίας μέχρι το σύνολο των δεδομένων να έχει πλήρως ταξινομηθεί ή να ικανοποιείται το κριτήριο της διακοπής. Η ακριβώς αντίθετη διαδικασία αφορά την " Divisive method ". Σε αυτή την περίπτωση η αρχική συστάδα περιέχει όλα τα αντικείμενα – σημεία του συνόλου δεδομένων μας και ξεκινάει να διασπάται επαναληπτικά μέχρι κάθε αντικείμενο να είναι μέρος μίας ενιαίας συστάδας ή μέχρι να ικανοποιηθεί το κριτήριο της διακοπής. Και στις δύο προαναφερόμενες περιπτώσεις η συγχώνευση ή η διάσπαση γίνεται με βάση την ομοιότητα ή την διαφορετικότητα των στοιχείων με τελική απεικόνιση μία ιεραρχική ομαδοποίηση η οποία προσομοιάζει σε δέντρο-διάγραμμα (Saxena, et al., 2017).

Στην ιεραρχική ομαδοποίηση η συγχώνευση ή διάσπαση υποσυνόλων των σημείων γίνεται λαμβάνοντας υπόψιν την απόσταση μεταξύ των σημείων που αποτελούν το σύνολο δεδομένων μας. Για να προσδιορίσουμε μέσω της έννοιας της απόστασης την εγγύτητα μεταξύ των σημείων χρησιμοποιούμε τρεις μετρικές σύνδεσης, την απλή σύνδεση, την μέση σύνδεση και την πλήρη σύνδεση (Saxena, et al., 2017).

Η ιεραρχική ομαδοποίηση χρησιμοποιεί έναν πίνακα σύνδεσης N x N διαστάσεων με τις μετρικές σύνδεσης που έχουν χρησιμοποιηθεί για να δημιουργηθούν οι συστάδες. Ο πίνακας που δείχνει την ομοιότητα μεταξύ των σημείων δημιουργείται υπολογίζοντας την ομοιότητα κάθε ζεύγους σημείων στο σύνολο δεδομένων μας, λαμβάνοντας υπόψιν την απόσταση τους. Η μετρική ομοιότητας χρησιμοποιείται για να καθορίσει το σχήμα κάθε συστάδας που δημιουργείται (Ezugwu, et al., 2022).

Η απλή σύνδεση, μετράει την πλησιέστερη απόσταση από οποιοδήποτε μέλος – σημείο μίας συστάδας σε οποιοδήποτε άλλο μέλος σημείο κάθε άλλης συστάδας. Η ομοιότητα

μεταξύ δύο διαφορετικών συστάδων υπολογίζεται μετρώντας την κοντινότερη απόσταση μεταξύ ενός ζεύγους των διαφορετικών συστάδων που εξετάζονται (Rathore, 2018).

Η μέση σύνδεση, είναι αυτή που δίνει την μικρότερη διακύμανση. Αυτό γίνεται βρίσκοντας είτε το μέσο όρο είτε τη διάμεσο των αποστάσεων μεταξύ των σημείων δεδομένων μεταξύ των διαφορετικών συστάδων (Rathore, 2018).

Η πλήρης σύνδεση καθορίζει την απόσταση μεταξύ δύο συστάδων μετρώντας τη μέγιστη απόσταση μεταξύ κάθε αντικειμένου – σημείου μίας συστάδας με κάθε άλλο αντικείμενο – σημείο κάθε άλλης συστάδας. Οι αλγόριθμοι που χρησιμοποιούν αυτή τη μετρική απόστασης είναι πιο συμπαγείς σε σχέση με τους αλγόριθμους απλής σύνδεσης (Rathore, 2018).

Στα πλεονεκτήματα της ιεραρχικής ομαδοποίησης συγκαταλέγεται ότι μπορεί να ανταπεξέλθει σε κάθε μετρική ομοιότητας και έχει ευελιξία όσο αφορά τον βαθμό λεπτομέρειας της ανάλυσης και μπορεί να εφαρμοστεί σε δεδομένα με διαφορετικά χαρακτηριστικά (Rathore, 2018).

Στα μειονεκτήματα, είναι ότι όταν εφαρμόζεται σε μεγάλου εύρους δεδομένων έχει υψηλή πολυπλοκότητα υπολογισμού. Επίσης δεν είναι τόσο ξεκάθαρα κάποιες φορές τα κριτήρια τερματισμού και είναι ευαίσθητη σε ακραίες τιμές. Τέλος έχει παρατηρηθεί ότι ένα μειονέκτημα που υπάρχει στις περιπτώσεις των “Agglomerative method” που χρησιμοποιούν αλγόριθμους που υπολογίζουν την απόσταση μεταξύ των σημείων με την Ευκλείδια απόσταση, ενώ πλεονεκτούν στο χρόνο που χρειάζεται, μειονεκτούν σημαντικά στην μνήμη που χρειάζεται για την εκτέλεση τους (Saxena, et al., 2017).

#### *3.3.1.1 Agglomerative hierarchical clustering*

Σε αυτή την υπό-κατηγορία ιεραρχικής ομαδοποίησης οι συστάδες που δημιουργούνται έχουν ιεραρχία από πάνω προς τα κάτω και μέσω μιας επαναληπτικής διαδικασίας τα σημεία ομαδοποιούνται σε όλο και μεγαλύτερες συστάδες δημιουργώντας τα διαφορετικά επίπεδα ιεραρχίας, μέχρι να ομαδοποιηθούν όλα τα σημεία ή μέχρι να ικανοποιηθεί το κριτήριο τερματισμού. Η αρχική συστάδα κατά την διάρκεια της διαδικασίας ομαδοποίησης αποτελεί την βάση της ιεραρχίας. Τα δύο πλησιέστερα σημεία συνδυάζονται για να δημιουργήσουν μία συστάδα κατά την διάρκεια της ομαδοποίησης (Ezugwu, et al., 2022).

Για τη δημιουργία των "clusters" οι αλγόριθμοι αυτού του τύπου χρειάζονται τα παρακάτω (Ahmad & Khan, 2019):

i) Πίνακας παρόμοιων στοιχείων (Similarity Matrix) – Κατασκευάζεται βρίσκοντας την ομοιότητα μεταξύ κάθε ζεύγους μικτών δεδομένων. Η επιλογή της μετρικής (metric) ομοιότητας επηρεάζει το σχήμα των "clusters".
ii) Μετρική σύνδεσης (linkage metric) – Επηρεάζει την απόσταση μεταξύ των ομάδων των παρατηρήσεων ως συνάρτηση της απόστασης μεταξύ του ζεύγους των παρατηρήσεων.

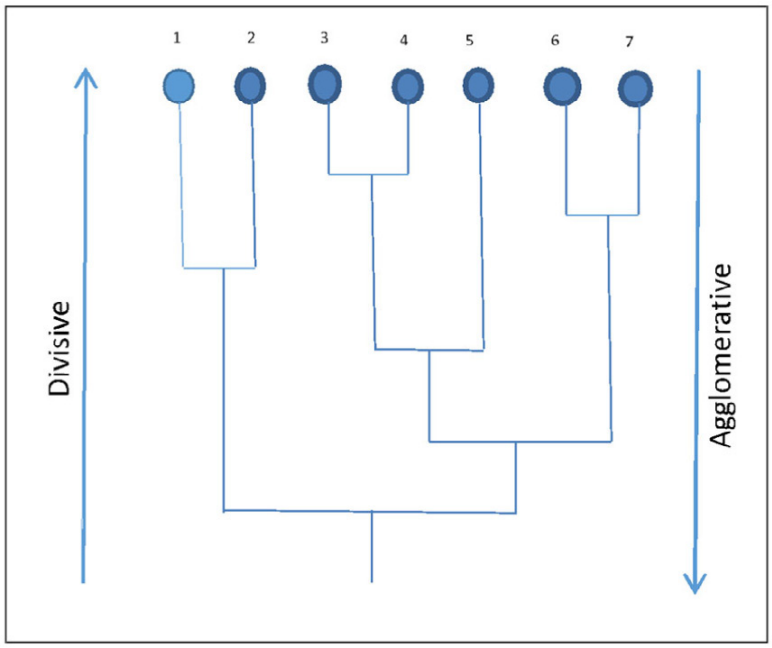
Οι περισσότεροι αλγόριθμοι αυτού του είδους παρουσιάζουν πολυπλοκότητα όσο αφορά το χρόνο και απαιτούν μεγάλη μνήμη. Ενδεικτικός αλγόριθμος αυτού του είδους είναι ο BIRCH (Balanced Iterative Reducing and clustering using hierarchies). Παρόλα αυτά, αξίζει να αναφερθεί ότι ο πίνακας παρόμοιων στοιχείων (Similarity Matrix) εξαρτάται σε μεγάλο βαθμό στον ορισμό που θέτουμε όσο αφορά την ομοιότητα και την απόσταση. Η απόσταση μεταξύ δύο στοιχείων διαφορετικού τύπου δεδομένων δεν είναι εύκολο να ερμηνευθεί και αυτό αποτελεί ένα μειονέκτημα (Ahmad & Khan, 2019).

#### *3.3.1.2 Divisive hierarchical clustering*

Η μέθοδος "Divisive hierarchical clustering" είναι ακριβώς αντίστροφη από την "Agglomerative" που παρουσιάσαμε στην προηγούμενη ενότητα. Διαιρεί κάθε ομάδα σε μικρότερες ξεκινώντας αρχικά από μία ενιαία συστάδα που περιέχει όλα τα σημεία, μέχρι να δημιουργηθούν μέσω μιας επαναληπτικής διαδικασίας όλες οι συστάδες. Η διαφορά της με την μέθοδο "Agglomerative" είναι ότι η διαδικασία ομαδοποίησης είναι από πάνω προς τα κάτω. Σε κάθε φάση της διαδικασίας που γίνεται μία διαίρεση, πρέπει να λαμβάνει υπόψιν και όλες τις προηγούμενες διαιρέσεις που έχουν γίνει. Αυτό σημαίνει ότι η διαδικασία αυτή καθίσταται αρκετά δαπανηρή όσο αφορά το υπολογιστικό κόστος που απαιτείται για την υλοποίηση της (Ezugwu, et al., 2022).

Η διαιρετική μέθοδος, έχει δύο υποκατηγορίες, την "Monothetic" και την "Polythetic". Η "Monothetic", συνδυάζει ένα λογικό χαρακτηριστικό μετασχηματίζοντας το σε μία μεταβλητή που είναι απαραίτητη για την ομαδοποίηση σε μία συστάδα. Η διαδικασία της διαίρεσης και της ανάθεσης κάθε στοιχείου σε μία συστάδα λαμβάνει υπόψιν αν το συγκεκριμένο στοιχείο έχει το χαρακτηριστικό που έχει επιλεχθεί ή όχι. Μπορεί να

θεωρηθεί ως μία παραλλαγή μεθόδου συσχέτισης και έχει εφαρμογή κυρίως για δεδομένα που έχουν δυαδικές μεταβλητές. Η "Polythetic", χρησιμοποιεί το σύνολο των μεταβλητών συνολικά εξετάζοντας την διαφοροποίησης τους ή την απόσταση τους. Στην ουσία δεν βασίζεται δηλαδή κατά τη διαδικασία της διαίρεσης σε μία μόνο μεταβλητή αλλά βασίζεται συνολικά στην απόσταση των σημείων που υποδεικνύουν την διαφορετικότητα μεταξύ τους. Την διαφοροποίηση μεταξύ "Agglomerative method" και "Divisive Method" μπορούμε να την δούμε συνοπτικά στην παρακάτω Εικόνα 2 (Ezugwu, et al., 2022).




*Εικόνα 2*

### 3.3.2 Partitional Clustering

Στην κατηγορία "Partitional Clustering" τα δεδομένα δεν έχουν κάποια ιεραρχική δομή. Χρησιμοποιούνται κυρίως σε περιπτώσεις που έχουμε μεγάλα σύνολα δεδομένων, πράγμα που σημαίνει ότι η δημιουργία δέντρο-διαγραμμάτων δεν αποτελεί λύση λόγω του εξαιρετικά υψηλού υπολογιστικού κόστους και της εξαιρετικά υψηλής πολυπλοκότητας που θα πρόκυπτε από μία τέτοια προσέγγιση. Η προσέγγιση που

υιοθετείται εδώ είναι η εύρεση μοτίβων που μπορεί να ομαδοποιήσει τα δεδομένα σε συστάδες - ομάδες.

Οι (Ahmad & Khan, 2019) αναφέρουν ότι το "Partitional Clustering" είναι το πιο διαδεδομένο και αφορά ομαδοποίηση των σημείων δεδομένων σε "k-clusters" χρησιμοποιώντας μία επαναληπτική διαδικασία με την παρακάτω λογική:

i) Θέτουμε ως αρχή το κέντρο ενός "cluster" το οποίο μπορεί να υποδηλώνει κατηγορικά ή ποσοτικά χαρακτηριστικά ανάλογα τα δεδομένα.
ii) Ορίζουμε μία μετρική απόστασης των σημείων δεδομένων (data points) από το κέντρο.
iii) Μία συνάρτηση κόστους που ελαχιστοποιείται κατόπιν επανάληψης και μπορεί να ανταπεξέλθει και σε μικτά δεδομένα.

Λαμβάνοντας υπόψη τα παραπάνω οι περισσότεροι αλγόριθμοι αυτού του είδους βελτιστοποιούν την παρακάτω συνάρτηση κόστους:

$$\sum_{i=1}^{n} \xi(di, Ci)$$

Όπου n είναι ο αριθμός των σημείων δεδομένων στο σύνολο δεδομένων, $Ci$ το κέντρο του "Cluster" που είναι πλησιέστερα του σημείου δεδομένων $di$, και ξ είναι η απόσταση μεταξύ του $di$ και του $Ci$.

Ένας σημαντικός λόγος που οι "partitional algorithms" είναι από τους πιο διαδεδομένους είναι ότι έχουν γραμμική σχέση με τον αριθμό των σημείων δεδομένων, λειτουργούν καλά σε ευρεία κλίμακα με μεγάλα σύνολα δεδομένων και μπορούν να προσαρμοστούν σε παράλληλα πλαίσια.

Η πιο διαδεδομένη προσέγγιση μέσω αυτής της κατηγορίας, όσο αφορά την κατάτμηση μικτών συνόλων δεδομένων που έχουμε τόσο κατηγορικές όσο και ποσοτικές μεταβλητές είναι ο μετασχηματισμός των κατηγορικών μεταβλητών σε ποσοτικές μέσω κωδικοποίησης και μετά να εφαρμόσουμε κάποιον αλγόριθμο όπως ο K-means για την ομαδοποίηση των δεδομένων.

Στα μειονεκτήματα των αλγορίθμων που βρίσκονται σε αυτή την κατηγορία είναι ότι η έννοια του κέντρου κάθε cluster, δεν είναι πάντα τόσο ξεκάθαρο. Ως εκ τούτου ο συνδυασμός κατηγορικών με ποσοτικές μεταβλητές δεν είναι τόσο ξεκάθαρος και

χρειάζεται αρκετή έρευνα για τον καθορισμό του κέντρου του cluster. Ένα άλλο ζήτημα στους αλγόριθμους αυτής της κατηγορίας είναι η δυσκολία που προκύπτει όσο αφορά την ομοιότητα ανάμεσα στα σημεία και στα κέντρα μεταξύ κάθε cluster. Υπάρχουν μετρικές απόστασης που χρησιμοποιούνται όπως η ευκλείδεια απόσταση μεταξύ δύο σημείων αλλά δεν είναι ξεκάθαρη η κλίμακα όταν έχουμε μικτούς τύπους δεδομένων.

Η εύρεση σωστών clusters είναι το κλειδί για την επιτυχία ομαδοποίησης μέσω αυτών των αλγορίθμων και βοηθάει στην σωστή προσέγγιση και κατανόηση του προβλήματος που καλούμαστε να λύσουμε μέσω των αλγορίθμων που χρησιμοποιούνται.

Μία περαιτέρω κατηγοριοποίηση του "Partitional Clustering" σε ένα επίπεδο πιο κάτω είναι "Hard/Crisp Clustering", "Fuzzy clustering", "Mixture Resolving Clustering". Γενικά μπορούμε να πούμε ότι οι πιο αντιπροσωπευτικοί αλγόριθμοι αυτού του είδους είναι ο K-Means για ποσοτικά δεδομένα, ο K-Modes για κατηγορικά και ο K-Prototypes για μικτά σύνολα δεδομένων.

#### *3.3.2.1 Hard/Crisp Clustering*

Σε αυτή την υπό-κατηγορία του "Partitional Clustering" έχουμε την περίπτωση που κάθε σημείο του συνόλου δεδομένων ανήκει μόνο σε μία συστάδα όπου χρησιμοποιείται ένας αλγόριθμος "Hard/Crisp Clustering". Για την καλύτερη κατανόηση της θα εξετάσουμε τις επιμέρους περιπτώσεις που έχουμε.

- **Graph - theoretic Clustering**

Αυτή η κατηγορία αφορά ένα σύνολο δεδομένων που αποτυπώνεται μέσω κόμβων που συνδέονται για την μοντελοποίηση των σχέσεων μεταξύ των όμοιων χαρακτηριστικών των σημείων που εξετάζουμε και αποτυπώνονται σε γράφους. Οι γραμμές που συνδέουν τους κόμβους / σημεία δεδομένων αποτυπώνουν την εγγύτητα μεταξύ τους. Οι κόμβοι χωρίζονται σε συστάδες βάσει της λογικής ότι οι γραμμές σύνδεσης είναι πιο λιγότερο πυκνές μεταξύ διαφορετικών διακριτών συστάδων, σε σχέση με τις γραμμές σύνδεσης που είναι περισσότερο πυκνές εντός της ίδιας συστάδας. Η αποτύπωση των συστάδων σε γράφους είναι βολική, αλλά δεν είναι ένας συμπαγής τρόπος που μπορεί να αντιμετωπίσει τις περιπτώσεις που έχουμε ακραίες τιμές στα δεδομένα μας. Σε αυτή την

κατηγορία μπορούμε να έχουμε τόσο συστάδες που εντάσσονται εντός μιας προκαθορισμένης ιεραρχίας αλλά μπορούμε να έχουμε επίσης περιπτώσεις που δεν εντάσσονται σε κάποια ιεραρχία. Ένας από τους πιο ευρέως διαδεδομένους αλγόριθμους αυτής της κατηγορίας είναι ο "k-nearest neighbor" (Saxena, et al., 2017).

- **Subspace Clustering**

Μία διαφορετική προσέγγιση είναι αυτή του "Subspace Clustering" η οποία βρίσκει ομάδες (clusters) μέσα σε διαφορετικά υπό-τμήματα του εξεταζόμενου συνόλου δεδομένων. Εδώ χρησιμοποιείται ένας τύπος αλγορίθμου τύπου K-means για την ομαδοποίηση, μέσα στα διαφορετικά υπό-τμήματα του συνόλου δεδομένων (Ahmad & Dey, 2011).

- **Density based Clustering**

Σε αυτή τη μέθοδο στηριζόμαστε στην πυκνότητα των ομάδων (Density based clustering), αντιπαραβάλλοντας ομάδες που παρουσιάζουν υψηλή πυκνότητα σε σχέση με άλλες ομάδες που προκύπτουν και εμφανίζουν χαμηλή πυκνότητα. Ένα χαρακτηριστικό παράδειγμα είναι όταν χρησιμοποιείται μέτρηση μεταξύ των σημείων που βασίζεται σε συντελεστές βαρύτητας που ορίζονται για τις κατηγορικές μεταβλητές του συνόλου δεδομένων. Μετά αυτή η μετρική απόστασης συνδυάζεται με έναν αλγόριθμο που δείχνει την υψηλή πυκνότητα των ομάδων για να προκύψει μία τελική ομαδοποίηση (Rodriguez & Laio, 2014).

- **Model Based Clustering**

Σε αυτή την κατηγορία οι μέθοδοι που συναντάμε υποθέτουν ότι ένα σημείο δεδομένων ταιριάζει με το μοντέλο που στις περισσότερες περιπτώσεις είναι μία στατιστική κατανομή. Τα μοντέλα καθορίζονται από τον χρήστη και αυτό σημαίνει ότι αν επιλεχθεί κάποιο μοντέλο που δεν ταιριάζει από τον χρήστη ή αντίστοιχα αν επιλεχθούν κάποιοι λάθος παράμετροι, αυτό θα έχει αρνητικό αντίκτυπο στα "clusters" που θα δημιουργηθούν. Στα μειονεκτήματα τους συμπεριλαμβάνεται ότι είναι πιο αργοί σε σύγκριση με τους υπόλοιπους "Partitional algorithms". Λόγω του υψηλού υπολογιστικού

τους κόστους επιλέγονται συνήθως για σύνολα δεδομένων με λίγα χαρακτηριστικά (features) (Melnykov & Maitra, 2010).

- **Search-based Clustering**

Σε αυτή την κατηγορία έχουμε μεθευρετικούς αλγόριθμους οι οποίοι θέτουν αυτόματα τη δομή και των αριθμό των συστάδων χωρίς προηγούμενη πληροφόρηση σχετική με τα χαρακτηριστικά των σημείων, σε αντίθεση δηλαδή με τους παραδοσιακούς αλγόριθμους ομαδοποίησης που στηρίζονται σε μετρικές απόστασης μεταξύ των σημείων ή σε ομοιότητα χαρακτηριστικών μεταξύ τους. Η αυτόματη ομαδοποίηση των σημείων σε συστάδες γίνεται ακολουθώντας την προσέγγιση ενός προβλήματος βελτιστοποίησης δύο συναρτήσεων, με το ζητούμενο στη μία να είναι η ελαχιστοποίηση των διαφορών μεταξύ των σημείων εντός της ίδιας συστάδας, ενώ στην άλλη η ταυτόχρονη μεγιστοποίηση των διαφορών που παρουσιάζουν οι διαφορετικές συστάδες μεταξύ τους. Οι δύο κύριες υπό-κατηγορίες των "Search-based Clustering" αλγορίθμων είναι η "Evolutionary" και η "Swarm Intelligence". Τα κοινά βήματα που ακολουθούνται και στις δύο κατηγορίες είναι αρχικά ο τυχαίος διαχωρισμός του πληθυσμού των δεδομένων και μετά η ομαδοποίηση του κάθε εξεταζόμενου σημείου αρχικά με τυχαίο τρόπο. Μετά γίνεται η εκτίμηση των αποτελεσμάτων της ομαδοποίησης ελέγχοντας ποια ομαδοποίηση βελτιστοποιεί τη συνάρτηση που έχουμε θέσει, ανακατανέμοντας επαναληπτικά τα σημεία σε συστάδες.

- **Square error Clustering & K-means**

Σε αυτή την κατηγορία των "Partitional Clustering" αλγορίθμων συναντούμε την μέθοδο όπου τα σημεία του συνόλου δεδομένων μας, ομαδοποιούνται σε έναν προκαθορισμένο αριθμό συστάδων, βασιζόμενοι στο κριτήριο που θέτει μία βασική συνάρτηση που υπολογίζει το άθροισμα της διαφοράς τετραγώνων για να γίνει η επιθυμητή ομαδοποίηση. Οι διαφορές τετραγώνων μεταξύ κάθε σημείου και της εκτιμώμενης τιμής του κέντρου κάθε συστάδας, είναι το βασικό κριτήριο για να ομαδοποιηθεί κάθε σημείο του συνόλου δεδομένων. Στην περίπτωση που το άθροισμα της διαφοράς τετραγώνων για μία συστάδα είναι μηδέν, τα σημεία της συστάδας αυτής είναι πανομοιότυπα, δηλαδή θεωρούμε ότι είναι εξαιρετικά κοντά. Η συνάρτηση αθροίσματος της διαφοράς τετραγώνων που χρησιμοποιείται είναι:

$$Sum\ of\ square\ errors = \sum_{i=1}^{n} (Xi - \underline{X})^2$$

Όπου n είναι ο αριθμός των σημείων δεδομένων, το Xi δείχνει το i σημείο του συνόλου των δεδομένων μας, και $\underline{X}$ το κέντρο κάθε συστάδας (Xu & Wunsch, 2005).

Ο πιο δημοφιλής αλγόριθμος αυτής της κατηγορίας είναι ο K-Means. Η μεθοδολογία που ακολουθεί βασίζεται στα κεντρικά σημεία (centroids), που είναι η βάση για να γίνει η ομαδοποίηση των σημείων σε ένα αριθμό προκαθορισμένων "K" συστάδων. Η κατανομή των παρατηρήσεων σε κάθε συστάδα γίνεται βάσει μίας αντικειμενικής συνάρτησης η οποία διασφαλίζει ότι οι ομοιότητες μεταξύ των αντικειμένων είναι μεγαλύτερες εντός της ίδιας συστάδας σε σύγκριση με τα αντικείμενα που ανήκουν σε μία διαφορετική συστάδα. Τα κεντρικά σημεία (centroids), εκφράζουν μία μετρική για το κέντρο κάθε συστάδας (cluster). Ένας καθορισμένος αριθμός παρατηρήσεων από το σύνολο δεδομένων επιλέγονται τυχαία ως το αντιπροσωπευτικό κέντρο για k συστάδες (clusters). Η ευκλείδεια απόσταση ανάμεσα στα υπόλοιπα σημεία που δεν είχαν αρχικά επιλεχθεί και κάθε υποτιθέμενο κέντρο κάθε συστάδας θα μετρηθεί σε μία επαναληπτική διαδικασία. Η μέτρηση αυτή θα αναδείξει σε ποια συστάδα θα πρέπει να πάει η παρατήρηση με βάση την μικρότερη απόσταση από το κέντρο της συστάδας. Η ομοιογένεια εντός μίας συστάδας βελτιώνεται σε κάθε επανάληψη, αφού μετριέται εκ νέου ένας νέος μέσος που βασίζεται στις παρατηρήσεις που έχουν ήδη καταχωρηθεί σε κάποια συστάδα. Ο νέος μέσος που έχει υπολογιστεί, χρησιμοποιείται για να χωρίσει τις παρατηρήσεις ξανά σε συστάδες. Αυτή η διαδικασία θα επαναλαμβάνεται έως ότου δημιουργηθεί ισορροπία. Η διαδικασία αυτή που βασίζεται στην συνάρτηση που χρησιμοποιεί την ευκλείδεια απόσταση μεταξύ των σημείων, έχει ως αποτέλεσμα την δημιουργία συμπαγών και διακριτών συστάδων, αφού ο αλγόριθμος προσπαθεί να ελαχιστοποιήσει το κριτήριο που βασίζεται στο άθροισμα των διαφορών τετραγώνων των σημείων. Στα πλεονεκτήματα του αλγόριθμου K-Means συγκαταλέγεται ότι είναι σχετικά απλός πράγμα που σημαίνει ότι δεν απαιτεί υψηλή πολυπλοκότητα υπολογισμών. Αυτό τον καθιστά σε ένα από τους πιο διαδεδομένους αλγόριθμους μέχρι και σήμερα σε προβλήματα που απαιτούν "Partitional clustering" (Ezugwu, et al., 2022).

Το πρόβλημα που έχει παρατηρηθεί και συγκαταλέγεται στα μειονεκτήματα του αλγόριθμου, είναι η δυσκολία που παρουσιάζεται όσο αφορά τον αρχικό καθορισμό του

αριθμού των συστάδων που πρέπει να τεθούν. Επίσης ο αλγόριθμος k-means είναι ευαίσθητος στις ακραίες τιμές πράγμα που μπορεί να οδηγήσει στην παραμόρφωση των συστάδων που δημιουργούνται (Saxena, et al., 2017).

#### *3.3.2.2 Miscellaneous Clustering*

Σε αυτή την κατηγορία αλγορίθμων θα δούμε τρεις επιμέρους περιπτώσεις. Την περίπτωση των "Time series clustering", "Streaming clustering", και "Mode seeking clustering". Παρακάτω μπορούμε να εξετάσουμε σε μεγαλύτερο βάθος την κάθε υπό-κατηγορία.

- **Time series clustering**

Η χρονοσειρά (time series), είναι μία ακολουθία πραγματικών αριθμών μέσα σε μία χρονική διάρκεια που έχουμε επιλέξει, όπου κάθε αριθμός αντιπροσωπεύει την τιμή μίας μεταβλητής που θέλουμε να εξετάσουμε. Είναι ένας δυναμικός τύπος δεδομένων αφού αλλάζει με την πάροδο του χρόνου. Κάθε χρονοσειρά αποτελείται από πολλά σημεία δεδομένων – τιμές αλλά μπορεί να θεωρηθεί ταυτόχρονα ως ένα αντικείμενο προς ανάλυση. Η ομαδοποίηση χρονοσειρών μπορεί να παρέχει πληροφορίες σε διαφορετικά πεδία και μπορεί να μας βοηθήσει να ανακαλύψουμε μοτίβα σε μεγάλα σύνολα δεδομένων που θέλουμε να εξετάσουμε (Aghabozorgi, et al., 2015).

Η ομαδοποίηση σε χρονοσειρές μπορεί να θεωρηθεί ως μία μέθοδος διερεύνησης των δεδομένων μας σε πολλές περιπτώσεις και αποτελεί κομμάτι σε περίπλοκους αλγόριθμους για την εξόρυξη δεδομένων. Η οπτική απεικόνιση των ομάδων χρονοσειρών βοηθάει να καταλάβουμε καλύτερα τα δεδομένα προς διερεύνηση, εντοπίζοντας ανωμαλίες που μπορεί να προκύπτουν, τη συνολική δομή των δεδομένων, και την καλύτερη επεξήγηση των δεδομένων μας. Η ανάγκη ανάλυσης μεγάλου όγκου δεδομένων στις περιπτώσεις αυτές, συνήθως έχει ως αντίκτυπο το γεγονός ότι χρειάζεται σημαντικός χρόνος για να γίνει η διαδικασία ομαδοποίησης (Aghabozorgi, et al., 2015).

Η ομαδοποίηση χρονοσειρών μπορεί να κατηγοριοποιηθεί επιμέρους στις εξής κατηγορίες. Στην κατηγορία "whole time series clustering", όπου η ομαδοποίηση γίνεται σε πολλές μεμονωμένες χρονοσειρές με σκοπό την δημιουργία μίας συστάδας με παρόμοιες παρατηρήσεις συνολικά. Την "subsequence clustering" όπου γίνεται ομαδοποίηση σε ένα κυλιόμενο χρονικό πλαίσιο μίας χρονοσειράς για να βρεθούν οι

ομοιότητες και οι διαφορές μεταξύ των παρατηρήσεων που εξετάζονται και να προκύψουν έτσι οι συστάδες κάτω από ένα μοτίβο. Τέλος η "time point clustering", η οποία έχει να κάνει με την προσωρινή εγγύτητα σημείων σε μία χρονική στιγμή που έχει επιλεχθεί να εξεταστεί (Chiş, et al., 2009).

- **Streaming Clustering**

Η ροή δεδομένων (data streaming) αναφέρεται σε μία διηνεκής ροή πολυδιάστατων δεδομένων που αλλάζει συνεχώς κατά τη διάρκεια του χρόνου. Δεν μπορεί να αποθηκευτεί σε μία σταθερή δομή λόγω της φύσης της που αλλάζει συνεχώς και λόγω του ότι δεν υπάρχει έλεγχος στη σειρά με την οποία εισρέουν τα δεδομένα, δεν μπορούν να χρησιμοποιηθούν για την ομαδοποίηση των δεδομένων οι κλασσικοί αλγόριθμοι που χρησιμοποιούνται συνήθως για αυτό τον σκοπό. Τα κυριότερα προβλήματα που πρέπει να αντιμετωπιστούν σε αυτή την περίπτωση είναι η συνεχώς μεταβαλλόμενη συνεχής ροή δεδομένων, τα νούμερα των συστάδων που δημιουργούνται, και η αντιμετώπιση των ακραίων τιμών που συναντούμε. Αυτό σημαίνει ότι οι συστάδες θα πρέπει να αναπροσαρμόζονται συνεχώς όσο αφορά τη δομή τους, ανάλογα με την ροή των δεδομένων, αφού μία στατική προσέγγιση με σταθερές συστάδες δεν θα μπορούσε να αποτυπώσει τις συνεχείς μεταβολές που παρουσιάζονται. Επίσης το ίδιο ισχύει τόσο για τον αριθμό των συστάδων που έχουν δημιουργηθεί, όσο και για την αντιμετώπιση των ακραίων τιμών που μπορεί να προκύψουν κατά τη διάρκεια του χρόνου. Απόρροια των προαναφερθέντων προβλημάτων είναι ο περιορισμένος χρόνος για την επεξεργασία δεδομένων που είναι διαθέσιμος αλλά και ο περιορισμός που προκύπτει στην μνήμη που απαιτείται για να αντιμετωπιστεί η συνεχής μεταβολή των δεδομένων προς εξέταση. Το πλαίσιο που λειτουργεί το "Streaming clustering" βασίζεται στο διαχωρισμό της ροής δεδομένων σε επιμέρους παρτίδες συγκεκριμένου μεγέθους, και εφαρμόζοντας τον αλγόριθμο "K-Median" (Mansalis, et al., 2018).

- **Mode seeking clustering**

Σε αυτή την κατηγορία συναντούμε την προσέγγιση που βασίζεται στην τάση να επιστρέφονται οι επικρατούσες τιμές που παρουσιάζονται πιο συχνά στο σύνολο δεδομένων που εξετάζουμε. Η επικρατούσα τιμή μπορεί να αναφέρεται είτε σε μεταβλητές με ποιοτικά χαρακτηριστικά είτε σε μεταβλητές με ποσοτικά χαρακτηριστικά. Στους αλγόριθμους που βασίζονται στην αναζήτηση επικρατουσών

τιμών μπορούμε να έχουμε περισσότερες από μία τιμές που πληρούν τα κριτήρια, και βάσει αυτών δημιουργούνται οι συστάδες που προκύπτουν λαμβάνοντας υπόψιν την πυκνότητα των υπόλοιπων τιμών του συνόλου δεδομένων γύρω από τις επικρατούσες τιμές. Ο αριθμός των επικρατουσών τιμών καθορίζουν τον αριθμό των συστάδων που θα δημιουργηθούν σε αυτούς τους τύπους αλγορίθμων (Sasaki, et al., 2018).

- **Neural network clustering**

Οι περισσότερες περιπτώσεις σε αυτή την κατηγορία αφορούν κυρίως μικτούς τύπους δεδομένων και επικεντρώνονται κυρίως είτε σε νευρωνικά δίκτυα που χρησιμοποιούνται SOM – "Self-Organizing Maps" ή ART – "Adaptive Resonance Theory" (Lam, et al., 2015). Η περίπτωση SOM είναι ένα νευρωνικό δίκτυο που χρησιμοποιείται σε μη γραμμική προβολή δεδομένων σε έναν χώρο μικρότερων διαστάσεων. Η περίπτωση ART βασίζεται στην θεωρία που βασίζεται και ο ανθρώπινος εγκέφαλος όταν μαθαίνει να κατηγοριοποιεί αυτόνομα και να προβλέπει σε ποια κατηγορία θα ενταχθεί ένα καινούργιο ερέθισμα. Το βασικό χαρακτηριστικό του ART που του προσδίδει πλεονέκτημα στις προβλέψεις είναι η ικανότητα γρήγορης και σταθερής κατηγοριοποίησης τόσο σε επιβλεπόμενα όσο και σε μη επιβλεπόμενα μοντέλα. Τόσο η κατηγορία SOM όσο και η ART μπορούν να χρησιμοποιηθούν σε αριθμητικές μεταβλητές δεδομένων αλλά δεν μπορούν να ανταπεξέλθουν απευθείας σε κατηγορικές μεταβλητές δεδομένων. Για να γίνει αυτό θα πρέπει να προηγηθεί μετασχηματισμός τους κωδικοποιώντας τα πρώτα (Lam, et al., 2015).

Όσο αφορά τα μειονεκτήματα αυτής της κατηγορίας είναι ότι με τη μέθοδο SOM μπορεί να έχουμε κακή χαρτογράφηση πράγμα που θα οδηγήσει σε απόκλιση όταν συγκρίνουμε με την κατανομή των δεδομένων μας στον χώρο. Για την ART έχουμε υψηλή πολυπλοκότητα πράγμα που οδηγεί σε αυξημένο «υπολογιστικό κόστος» (Du, 2010).

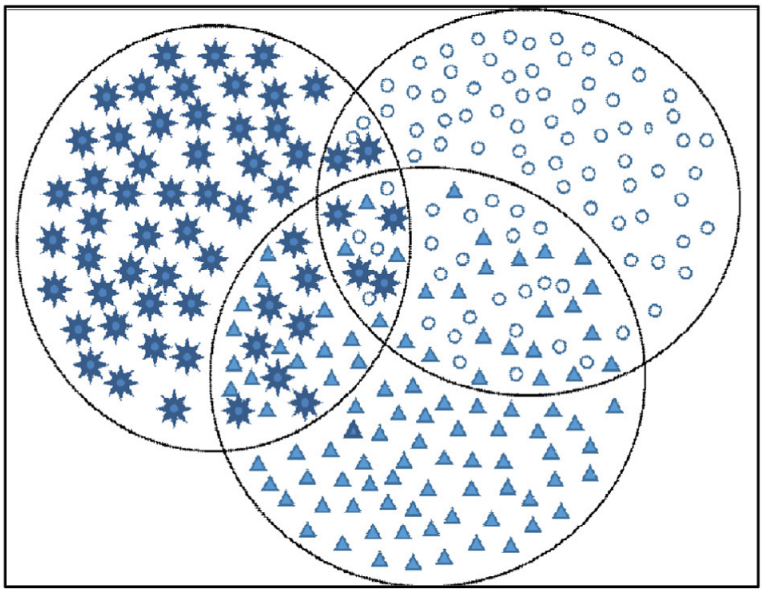
#### *3.3.2.3 Mixture Resolving clustering*

Στην κατηγορία των "Mixture Resolving" αλγορίθμων, υποθέτουμε ότι οι το σύνολο των σημείων που εξετάζουμε από το σύνολο δεδομένων μας, προέρχεται από μία συλλογή στιγμιότυπων που δημιουργούν πιθανολογικές συστάδες. Ως εκ τούτου η ανάθεση κάθε σημείου από το σύνολο δεδομένων μας ορίζεται από το ποια συστάδα ταιριάζει περισσότερο, ανάλογα με την πιθανότητα που έχει κάθε συστάδα να φιλοξενήσει το εν

λόγω σημείο. Η πιθανότητα ορίζεται από μία συνάρτηση σχετική με την πυκνότητα της συστάδας, και υπολογίζει το εξεταζόμενο δείγμα σημείων, για να βρει ποια από αυτά θα μπορούσαν να καταχωρηθούν σε κάθε συστάδα βάσει υπολογισμού πιθανότητας (Ezugwu, et al., 2022).

#### *3.3.2.4 Fuzzy Clustering*

Η ασαφής ομαδοποίηση (Fuzzy Clustering), είναι μία μέθοδος όπου οι ομάδες ορίζονται σε ασαφή σετ, με κάθε μοτίβο να ανήκει ταυτόχρονα σε παραπάνω από μία ομάδες όπως φαίνεται στην εικόνα 3 (Ezugwu, et al., 2022).



*Εικόνα 3*

Τα σημεία του συνόλου δεδομένων, ανατίθενται σε δύο ή και παραπάνω ομάδες έχοντας έναν βαθμό συμμετοχής σε κάθε ομάδα που ανήκουν, χτίζοντας μία μη-δυαδική σχέση. Με αυτό τον τρόπο οι ομάδες έχουν την δυνατότητα να εμφανίζουν αλληλοκαλύψεις, πράγμα που δείχνει την ασάφεια των ορίων των ομάδων που έχουν δημιουργηθεί, απαριθμώντας τον αριθμό των σημείων με σημαντική συμμετοχή στα σημεία που σημειώνεται αλληλοκάλυψη μεταξύ των ομάδων. Αυτή η μέθοδος είναι ιδιαίτερα χρήσιμη στις περιπτώσεις που έχουμε ομάδες με αμφίβολα όρια, που δεν είναι εύκολο να ξεχωρίσουν. Ο βαθμός συμμετοχής στις ομάδες που ανατίθεται τα σημεία που

εξετάζονται, μπορούν να βοηθήσουν στην ανακάλυψη εγγενούς σχέσης μεταξύ του σημείου και των ομάδων στις οποίες είναι μέλος. Οι αλγόριθμοι που αναπτύχθηκαν αρχικά με βάση αυτή την λογική παρουσίαζαν προβλήματα σχετικά με τον αρχικό διαχωρισμό που επηρεάζονταν κυρίως από τις ακραίες τιμές. Οι αλγόριθμοι "Subsequent" που βασίζονται σε αυτή τη μέθοδο, βοήθησαν στην αντιμετώπιση αυτών των προβλημάτων (Saxena, et al., 2017).

Η ασαφής ομαδοποίηση δίνει τη δυνατότητα να αντιμετωπίσουμε θέματα που σχετίζονται με την ανακάλυψη μοτίβων, την ενσωμάτωση δεδομένων που δεν είναι ολοκληρωμένα ή έχουν υψηλό ποσοστό «θορύβου» και δίνουν αποτελέσματα σχετικά γρήγορα. Στα πεδία που μπορεί να χρησιμοποιηθεί όπως η ανάκτηση εικόνων, χρησιμοποιούν κανόνες συσχέτισης και λειτουργικές εξαρτήσεις ανάμεσα στα σημεία δεδομένων. Τα τρία κύρια θέματα που έχουν παρατηρηθεί, είναι ότι δεν καθίσταται δυνατή η εκ των προτέρων γνώση όσο αφορά τον αριθμό των ομάδων που θα προκύψουν και υπάρχει πάντα η ανάγκη να δηλωθεί ένα κριτήριο εγκυρότητας σχηματισμού ομάδας πριν να μπορεί να καθοριστεί ο βέλτιστος αριθμός των ομάδων. Επίσης ο χαρακτήρας και η τοπολογία των ομάδων δεν μπορούν να καθοριστούν εκ των προτέρων, πράγμα που σημαίνει ότι πρέπει αρχικά να επιλεχθούν με τυχαίο τρόπο πράγμα που μπορεί να οδηγήσει σε διαφοροποίηση του σχήματος, της πυκνότητας και τον αριθμό των παρατηρήσεων που θα έχουμε σε κάθε ομάδα. Επιπρόσθετα ο χρόνος που απαιτείται για τους αλγόριθμους αυτού του τύπου είναι σημαντικός (Saxena, et al., 2017).

Τέλος, πρέπει να αναφέρουμε ότι η ασαφής φύση αυτής της μεθόδου με το σχηματισμό ασαφών ομάδων, μπορεί να αφαιρεθεί, με το να ανατίθενται παρόμοια σημεία του συνόλου δεδομένων σε ομάδες που παρουσιάζεται υψηλότερος βαθμός συμμετοχής αυτών των σημείων. Ο πιο γνωστός αλγόριθμος που χρησιμοποιείται για αυτό είναι ο "Fuzzy c-means" (Bezdek, 2013).

#### *3.3.2.5 Λοιπές κατηγορίες Clustering*

Σε αυτή την κατηγορία έχουμε διαφορετικές περιπτώσεις με αλγόριθμους ομαδοποίησης (clustering), που είτε δεν μπορούν να ενταχθούν σε κάποια από τις προηγούμενες διακριτές κατηγορίες είτε δεν έχουν μελετηθεί ακόμη εκτενώς.

Τεχνικές “Spectral clustering” είναι αυτές που έχουν ως προσέγγιση τη μείωση των διαστάσεων των δεδομένων χρησιμοποιώντας ιδιοτιμές (eigenvalues) από το πίνακα ομοιότητας που προκύπτει από τα δεδομένα μας. Αφού υπολογιστεί ο πίνακας ομοιότητας εφαρμόζεται ένας αλγόριθμος “Spectral clustering” για να δημιουργήσει τις ομάδες (clusters). Χρησιμοποιώντας μία μέθοδο “Clustering ensemble” μπορούμε να υπολογίσουμε την ομοιότητα μεταξύ δύο σημείων ξεχωριστά για τις κατηγορικές και τις ποσοτικές μεταβλητές. Μετά ο βαθμός ομοιότητας μεταξύ των δύο τύπων μεταβλητών προστίθεται μεταξύ των σημείων. Τέλος χρησιμοποιείται “Spectral clustering” για να δημιουργηθούν οι ομάδες (clusters) (Luo, et al., 2006).

Μία παραλλαγή της τεχνικής “Spectral clustering” είναι αυτή που προτείνεται από τους David και Averbuch στο άρθρο τους που χρησιμοποιούν τον αλγόριθμο SpectralCAT, μέσω του οποίου χρησιμοποιείται κατηγορικό “Spectral clustering” για την ομαδοποίηση μικτών δεδομένων. Ο αλγόριθμος μετασχηματίζει τις ποσοτικές μεταβλητές σε κατηγορικές και μετά στα μετασχηματισμένα δεδομένα εφαρμόζεται μία μέθοδος “Spectral clustering” (David & Averbuch, 2012).

### 3.3.3 Ιδιότητες απόδοσης αλγορίθμων Clustering

Οι (Al-Jabery, et al., 2019) αναφέρθηκαν στο θέμα της απόδοση των αλγορίθμων ομαδοποίησης κατηγοριοποιώντας τους με βάση κάποιες ιδιότητες όπως θα δούμε παρακάτω.

*Πολλές διαστάσεις:* Μετράει την ικανότητα του αλγορίθμου να ανταπεξέλθει σε δεδομένα που έχουν πολλά χαρακτηριστικά (features), που μερικές φορές μπορεί να είναι και περισσότερα από τα δεδομένα. Σημαντική είναι η ικανότητα να αναγνωρίζουμε ποια είναι τα σχετικά χαρακτηριστικά με το πρόβλημα που καλούμαστε να αντιμετωπίσουμε για να περιγράψουμε καλύτερα τη δομή των δεδομένων μας.

*Στιβαρότητα:* Πολύ συχνά το σύνολο δεδομένων προς εξέταση δεν είναι «καθαρό», για διάφορους λόγους όπως η επεξεργασία που μπορεί να επηρεάσει την ποιότητα των δεδομένων, οι λανθασμένες καταχωρήσεις, οι ακραίες τιμές που μπορεί να υπάρχουν, οι ανακρίβειες που μπορεί να προκύψουν από λάθος κατά τη μεταφορά των δεδομένων. Όλα τα παραπάνω μπορούν να επηρεάσουν τα δεδομένα που έχουμε να επεξεργαστούμε οπότε με τον όρο στιβαρότητα εννοούμε την ικανότητα του αλγορίθμου να εντοπίζει και

να απομακρύνει τις ακραίες τιμές και τον θόρυβο που μπορεί να εντοπιστεί στα δεδομένα μας.

*Εξάρτηση αριθμού k ομάδων:* Η δυνατότητα να γνωρίζουμε εκ των προτέρων τον αριθμό των ομάδων που πρέπει να δημιουργηθούν κατά την διαδικασία της ομαδοποίησης είναι ένα από τα συνηθέστερα ζητήματα που μπορεί να προκύψουν. Πολλοί αλγόριθμοι έχουν ως προαπαιτούμενο τη θέσπιση συγκεκριμένου αριθμού ομάδων κατά την διαδικασία της αρχικής παραμετροποίησης πράγμα που μπορεί να επηρεάσει τα αποτελέσματα που θα πάρουμε. Η αυτόματη ομαδοποίηση είναι ένας από τους τομείς που ερευνητικά υπάρχει μεγάλο ενδιαφέρον αφού στους περισσότερους αλγορίθμους είναι δύσκολο να αποφασίσουμε εκ των προτέρων τον αριθμό των ομάδων που θα πρέπει να δημιουργήσουμε εκ των προτέρων χωρίς καμία επιπλέον πληροφορία. Η ικανότητα να ορίσουμε τον σωστό αριθμό ομάδων μπορεί να βοηθήσει σημαντικά στο να βρούμε τη βέλτιστη λύση στο πρόβλημα της ομαδοποίησης. Ο αλγόριθμος θα πρέπει να μπορεί να καθορίσει τον αριθμό των ομάδων με βάση τις ιδιότητες των δεδομένων προς εξέταση. Ως εκ τούτου οι αλγόριθμοι ομαδοποίησης μπορεί να θεωρηθεί ότι χρησιμοποιούνται για να λύσουν ένα πρόβλημα βελτιστοποίησης. Αυτοί οι μεθευρετικοί αλγόριθμοι μπορούν να χρησιμοποιηθούν για να ανακαλύψουμε τον αριθμό των ομάδων αυτόματα.

*Εξάρτηση παραμέτρων:* Εκτός από την απαίτηση να μπορούμε να ορίσουμε τον αριθμό των ομάδων εκ των προτέρων, υπάρχουν και άλλες παράμετροι στους εν λόγω αλγόριθμους οι οποίες είναι ευαίσθητες και πρέπει να καθοριστούν από τον χρήστη για να λειτουργήσουν σωστά. Αυτό σημαίνει ότι εξαρτώνται σε πολύ μεγάλο βαθμό από την σωστή παραμετροποίηση που θα πραγματοποιήσει ο χρήστης βάσει των υποθέσεων που θα κάνει. Συμπερασματικά, η πρακτική καθοδήγηση που έχει ο χρήστης ώστε να θέσει σωστά αυτές τις παραμέτρους για να αποφευχθούν λάθη, είναι μία καλό μέτρο για την καλή απόδοση του αλγορίθμου.

*Ασυνήθιστο σχήμα της ομάδας:* Η ικανότητα να μπορούμε να καταλάβουμε ότι τα σχήματα των ομάδων που προκύπτουν είναι ασυνήθιστα είναι μία ακόμη πρόκληση όσο αφορά τους αλγόριθμους ομαδοποίησης. Υπάρχουν πολλές εφαρμογές περιπτώσεων οι οποίες μπορεί να μην δώσουν συνηθισμένο σχήμα σε μία ομάδα που θα προκύψει. Για τέτοιου είδους εφαρμογών η βέλτιστη λύση είναι να παρουσιάσουμε τις ομάδες στο φυσικό σχήμα που προκύπτει (κανονικό ή μη κανονικό). Παρόλα αυτά, ένας καλός

αλγόριθμος ομαδοποίησης θα πρέπει να είναι σε θέση να εντοπίσει ασυνήθιστα σχήματα ομάδων που προκύπτουν.

*Εξάρτηση σειράς:* Η δυνατότητα να κρατήσουμε τη σειρά που προκύπτουν τα εισερχόμενα μοτίβα σε ένα σύνολο δεδομένων είναι πολύ σημαντικό για να δημιουργήσουμε σωστές ομάδες. Αυτού του είδους χαρακτηριστικών είναι πολύ συνηθισμένα στις περιπτώσεις που έχουμε συνεχώς μεταβαλλόμενα δεδομένα που εισρέουν (streaming data). Οι αλγόριθμοι ομαδοποίησης για ένα τέτοιου είδους συνόλου δεδομένων, μπορεί να απαιτούν οι λύσεις ομαδοποίησης που προτείνουν να είναι διαφορετικές όσο αφορά τη σειρά που παρουσιάζονται σε σχέση με τα μοτίβα που αναγνωρίζουν ως εισροή. Η μεγαλύτερη πρόκληση σε αυτές τις περιπτώσεις είναι η μειωμένη ευαισθησία στα αποτελέσματα των μοτίβων που αναγνωρίζονται στα δεδομένα.

*Οπτικοποίηση:* Η καλή απεικόνιση της ομαδοποίησης βοηθάει στην καλύτερη ερμηνεία των αποτελεσμάτων, ενώ ταυτόχρονα βοηθάει στην εξαγωγή χρήσιμων συμπερασμάτων μέσω αυτής της πληροφορίας. Με αυτό τον τρόπο είναι πιο εύκολο για τους χρήστες να αναγνωρίσουν τα μοτίβα που προκύπτουν οπτικά και να χρησιμοποιήσουν αυτή την πληροφορία για περαιτέρω επεξεργασία των δεδομένων.

*Μικτοί τύποι δεδομένων:* Οι αλγόριθμοι ομαδοποίησης είναι αρκετά ευέλικτοι ώστε να μπορούν να χειριστούν κάθε τύπο δεδομένων που θα μπορούσε να παρουσιαστεί σε ένα σύνολο δεδομένων. Αυτό είναι πολύ σημαντικό γιατί δεδομένα που έχουν αποκτηθεί από διαφορετικές πηγές ενδέχεται να έχουν διαφορετικά χαρακτηριστικά, δηλαδή μπορεί να είναι διαφορετικού τύπου, κατηγορικά ή ποσοτικά. Συνήθως ο συνδυασμός διαφορετικών τύπων δεδομένων βοηθάει τους αλγόριθμους να δώσουν καλύτερα πιο συμπαγή αποτελέσματα, εφόσον φυσικά είναι σε θέση να χρησιμοποιήσουν δεδομένα μικτού τύπου στο σύνολο δεδομένων που του παρέχουμε.

### 3.3.4 Θέματα προς διερεύνηση στους αλγόριθμους Clustering

Κάποια ανοιχτά θέματα για τους αλγόριθμους ομαδοποίησης μπορούν να κατηγοριοποιηθούν παρακάτω.

*Υπολογιστική πολυπλοκότητα:* Μερικοί αλγόριθμοι ομαδοποίησης αντιμετωπίζουν θέματα υπολογιστικής πολυπλοκότητας, ιδίως σε περιπτώσεις που εφαρμόζονται σε σύνολα δεδομένων με πολλές διαστάσεις. Αυτό το πρόβλημα μπορεί να αντιμετωπιστεί αυξάνοντας τους υπολογιστικούς πόρους έχοντας υψηλή χωρητικότητα GPU. Μία ακόμη προσέγγιση είναι να εκμεταλλευτούμε την ισχύ από παράλληλους υπολογιστικούς πόρους οι οποίοι μπορούν να βοηθήσουν στο να σχεδιαστούν μοτίβα που θα οδηγήσουν σε καλύτερη ομαδοποίηση. Στον αντίποδα η προσέγγιση των παράλληλων υπολογιστικών πόρων αυξάνει την πολυπλοκότητα της υλοποίησης (Shirkhorshidi, et al., 2014).

*Βελτιστοποίηση ομάδων:* Οι ομάδες που προκύπτουν μετά τη διαδικασία της ομαδοποίησης, συχνά χρειάζονται βελτιστοποίηση είτε χρησιμοποιώντας τον ίδιο αλγόριθμο είτε κάποιον διαφορετικό. Αυτό γίνεται για να σιγουρευτούμε ότι τα σημεία που μπορεί να έχουν ανατεθεί σε μία ομάδα λανθασμένα λόγω ανεπαρκών μέτρων ομοιότητας που χρησιμοποιήθηκαν, μπορούν να μετατεθούν σε άλλη ομάδα όπου θα ταιριάζουν καλύτερα. Μερικές μέθοδοι ομαδοποίησης όπως η διαιρετική, εφαρμόζουν δύο προσεγγίσεις για την βελτιστοποίηση της διαδικασίας ομαδοποίησης, ως επί των πλείστων την “monothetic” και την “polythetic”. Ενώ ο προηγούμενος διαχωρισμός της ομαδοποίησης χρησιμοποιεί ένα μόνο χαρακτηριστικό, ο τελευταίος διαχωρισμός γίνεται χρησιμοποιώντας όλα τα χαρακτηριστικά του συνόλου δεδομένων. Αυτή είναι μία εξελικτική μέθοδος που χρησιμοποιείται για να βελτιώσει την ποιότητα των ομάδων που προκύπτουν. Αυτό δείχνει ότι μπορεί να γίνει συνδυασμός δύο μεθόδων σε αυτούς τους μεθευρετικούς αλγόριθμους μετασχηματίζοντας την προσέγγιση μας σε υβριδική ώστε να προκύψει το καλύτερο δυνατό αποτέλεσμα (Ezugwu, et al., 2022).

*Ταχύτητα σύγκλισης:* Ένα μεγάλο εύρος μεθευρετικών αλγορίθμων έχουν βοηθήσει σε προβλήματα βελτιστοποίησης, και πιο συγκεκριμένα με προβλήματα ομαδοποίησης. Ένας καλός βαθμός γρήγορης σύγκλισης των στοιχείων με τις ομάδες, είναι ένας από τους δείκτες που δείχνουν την καλή απόδοση των αλγορίθμων, αφού δείχνουν πόσο καλή ποιότητα έχουν οι ομάδες που προκύπτουν κατά τη διαδικασία της ομαδοποίησης. Ένα ακόμη πρόβλημα που εντοπίζεται σε αυτού του είδους εφαρμογών με ομαδοποίηση είναι η ευαισθησία του αρχικού σταδίου, με τις συναρτήσεις που μετράνε την εσωτερική σχέση των στοιχείων μέσα στην ίδια ομάδα όσο και τη σχέση μεταξύ των ομάδων να επηρεάζει το χρόνο που δημιουργείται η απαιτούμενη σύγκλιση (Xie, et al., 2019).

*Διαστάσεις δεδομένων:* Αλγόριθμοι ομαδοποίησης όπως ο K-Means, Gaussian mixture model (GMM), maximum-margin clustering, δεν μπορούν να εφαρμοστούν εύκολα σε δεδομένα που έχουν πολλές διαστάσεις. Αυτό το πρόβλημα μπορεί να αντιμετωπιστεί αν προβάλλουμε τα δεδομένα σε ένα χώρο με λιγότερες διαστάσεις και μετά να ξεκινήσουμε την διαδικασία της ομαδοποίησης (Wang, et al., 2016).

*Αποτελεσματικότητα και επεκτασιμότητα:* Η έννοια της αποτελεσματικότητας και της επεκτασιμότητας είναι συνυφασμένες με το χώρο των μεγάλων δεδομένων (big data). Μία πιθανή λύση για το συγκεκριμένο θέμα είναι η τεχνική της βαθιάς μάθησης. Επίσης η μείωση της εξάρτησης των αλγορίθμων από την εκτενή παραμετροποίηση που γίνεται στην αρχή μπορεί να βελτιώσει την αποτελεσματικότητα των αλγορίθμων ομαδοποίησης αισθητά (Ezugwu, et al., 2022).

*Αντιπροσωπευτικότητα σημείων δεδομένων:* Η σωστή αντιπροσώπευση των σημείων δεδομένων είναι ένας ακόμη τομέας που χρειάζεται προσοχή στις μεθόδους ομαδοποίησης. Πολλές φορές δεν παρουσιάζονται σε κατάλληλη μορφή ή μπορεί να παρουσιάζονται διαφορετικά σε διαφορετικά πεδία εφαρμογής. Μερικά σημεία δεδομένων μπορεί να παρουσιάζονται ως διανύσματα κάποιων χαρακτηριστικών ενώ κάποια άλλα μπορεί να παρουσιάζονται ως γραφήματα για να δείχνουν την ομοιότητα που υπάρχει μεταξύ τους. Αποτελεσματικός τρόπος για την σωστή αντιπροσώπευση των δεδομένων μπορεί να θεωρηθεί αυτός που μειώνει το επίπεδο υπολογιστικής πολυπλοκότητας. Αυτό επιτρέπει στον αλγόριθμο να επεκταθεί αναγνωρίζοντας νέους τομείς ή περιπτώσεις που μπορεί να έχουμε κάποια «ειδική» κατανομή στα δεδομένα μας (Ezugwu, et al., 2022).

*Μετρικές αξιολόγησης:* Οι μετρικές αξιολόγησης που χρησιμοποιούνται για να αξιολογήσουμε και να συγκρίνουμε την απόδοση των διαφορετικών αλγορίθμων είναι η ακρίβεια, η σταθερότητα του αλγορίθμου, και η κανονικοποίηση των δεδομένων που γίνεται (Benabdellah, et al., 2019).

*Ροή δεδομένων:* Όταν έχουμε περιπτώσεις με συνεχής ροή δεδομένων καθιστά την ομαδοποίηση πιο δύσκολη σε σχέση με τις περιπτώσεις που έχουμε στατικά δεδομένα. Οι μέθοδοι ομαδοποίησης που χρησιμοποιούνται θα πρέπει να είναι αρκετά συμπαγείς ώστε να μπορούν να ανταπεξέλθουν στην ύπαρξη ακραίων τιμών ή «θορύβου» στα δεδομένα μας. Επίσης οι αλγόριθμοι ομαδοποίησης θα πρέπει να είναι σε θέση να ανιχνεύουν τις αλλαγές στα δεδομένα, να εντοπίζουν αν υπάρχει κάποια τάση και να

μπορούν να ομαδοποιούν τα αντικείμενα / σημεία δεδομένων όταν έχουμε συνεχή ροή δεδομένων που δημιουργεί διαφοροποιήσεις με την πάροδο του χρόνου. Αυτές οι περιπτώσεις που αντιμετωπίζουμε αυτές τις συνθήκες είναι συνηθισμένες σε δεδομένα που προέρχονται από μέσα κοινωνικής δικτύωσης, στις οποίες αντιμετωπίζουμε ολοένα και αυξανόμενη εισροή δεδομένων, οπότε η πρόκληση που πρέπει να αντιμετωπίσουμε αφορά την ικανότητα της βελτιστοποίησης της υπολογιστικής πολυπλοκότητας και της μνήμης που πρέπει να χρησιμοποιηθεί από αυτού του είδους των αλγορίθμων. Η έρευνα που χρειάζεται να γίνει σε αυτό το πεδίο αφορά τη δημιουργία μοντέλων που να προσαρμόζονται εύκολα σε δεδομένα που αλλάζουν και εξελίσσονται, βελτιώνοντας τις υπάρχουσες μεθόδους ομαδοποίησης που χρησιμοποιούνται (Kokate, et al., 2018).

*Εξαγωγή συμπερασμάτων:* Ένα άλλο θέμα που εντοπίζουμε στις περιπτώσεις προβλημάτων ομαδοποίησης είναι η εξαγωγή συμπερασμάτων από μεγάλα σύνολα δεδομένων. Αυτό οφείλεται σε μεγάλο βαθμό στην ολοένα αυξανόμενη τάση για όλο και περισσότερες πηγές δεδομένων. Αυτό το θέμα δημιουργεί μια μεγάλη πρόκληση στον αναλυτή δεδομένων (data analyst) αφού είναι δύσκολο να βγάλουν συμπεράσματα από δεδομένα που έχουν όγκο terabytes ή και petabytes. Οι περαιτέρω έρευνες θα μπορούσαν να εστιάσουν στο πως θα μπορούσαμε να ξεπεράσουμε τον περιορισμό που συναντούμε στην εξαγωγή συμπερασμάτων όταν ο όγκος των δεδομένων είναι εξαιρετικά μεγάλος, πιθανότατα εφαρμόζοντας προσεγγίσεις που συμπεριλαμβάνουν αλγόριθμους που αναπτύσσονται παράλληλα ή δημιουργώντας κάποια κατανομή εντός της ομαδοποίησης που δημιουργούν. Επίσης οι μελλοντικές έρευνες θα μπορούσαν να αναπτύξουν καινούργιες μεθόδους ομαδοποίησης στις οποίες ο χρήστης θα μπορούσε να επιλέξει ανάμεσα από έναν μεμονωμένο στόχο ή πολλαπλούς στόχους προς βελτιστοποίηση (Ezugwu, et al., 2022).

### 3.3.5 Μετρικές αξιολόγησης για Clusters

Ένα από τα κυριότερα θέματα που αντιμετωπίζουν οι αλγόριθμοι ομαδοποίησης όπως είδαμε παραπάνω είναι ο προσδιορισμός του αριθμού των ομάδων που προκύπτουν και η αξιολόγηση τους όσο αφορά το βέλτιστο αριθμό που θα μπορούσε να επιλεχθεί αλλά και η ποιότητα των ομάδων που θα προκύψει μετά την ομαδοποίηση.

Για το σκοπό αυτό υπάρχουν κάποιοι δείκτες εγκυρότητας που μπορούμε να χρησιμοποιήσουμε. Για την παρούσα έρευνα εφόσον θα χρησιμοποιήσουμε τον αλγόριθμο k – means για την ομαδοποίηση θα εξετάσουμε τους επικρατέστερους δείκτες που χρησιμοποιούνται για να αποφασίσουμε τον βέλτιστο αριθμό των ομάδων αλλά και για να επικυρώσουμε ότι οι ομάδες που έχουμε δημιουργήσει, πληρούν κάποια χαρακτηριστικά που δείχνουν ότι έχει γίνει μία αποδεκτή κατάτμηση του συνόλου δεδομένων μας σε επιμέρους ομάδες. Οι δείκτες που θα εξετάσουμε είναι οι ακόλουθοι (Gustriansyah, et al., 2020).

- **Elbow method**

Είναι μία μέθοδος που χρησιμοποιείται για να καθορίσουμε το βέλτιστο αριθμό των ομάδων που θα πρέπει να διαλέξουμε κατά την εφαρμογή του αλγορίθμου, εξετάζοντας το ποσοστό της σύγκρισης ανάμεσα στον αριθμό των ομάδων που θα σχηματίσουν μία γωνία στο γράφημα που θα προκύψει. Αν η τιμή της πρώτης ομάδας σχηματίσει μία γωνία που παραπέμπει σε αγκώνα με την τιμή της δεύτερης ομάδας στην καμπύλη που σχηματίζεται ή αν εμφανίζεται μεγάλη μείωση στην καμπύλη, αυτός ο αριθμός ομάδων είναι ο βέλτιστος για να προχωρήσουμε στην ομαδοποίηση. Ο βέλτιστος αριθμός των ομάδων δηλαδή θα επιλεχθεί βάσει του γραφήματος εντοπίζοντας το σημείο καμπής. Η μέθοδος αυτή είναι οπτική, και εξετάζει τη διακύμανση του αθροίσματος τετραγώνων μεταξύ των ομάδων βάσει του δείκτη WSS (Within Clusters Sum of Squares) που είναι μία συνάρτηση των αριθμών των ομάδων που μπορούμε να επιλέξουμε. Όσο μεγαλύτερος είναι ο αριθμός των ομάδων (k) που σχηματίζονται η τιμή του δείκτη WSS θα γίνεται είτε μικρότερη ή μεγαλύτερη. Η συνάρτηση που χρησιμοποιείται είναι:

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

Όπου

k = ο αριθμός των ομάδων

n = ο αριθμός των σημείων δεδομένων

$x_i$ = το στοιχείο i της κάθε ομάδας

$c_j$ = το κεντρικό σημείο j κάθε ομάδας

- **Δείκτης Silhouette**

Ο δείκτης Silhouette χρησιμοποιείται για να μετρήσουμε πόσο καλή είναι μία ομάδα από άποψη ποιοτικών χαρακτηριστικών. Οι συναρτήσεις για τον υπολογισμό του δείκτη Silhouette (SI: Silhouette Index) είναι οι ακόλουθες:

$$SI = \bar{s} = \frac{1}{n}\sum_{i=1}^{n} s(i)$$

Όπου οι επιμέρους υπολογισμοί για κάθε μέρος της γίνονται από τις ακόλουθες συναρτήσεις:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

$$a(i) = \frac{1}{|C_i|-1}\sum_{j\in C_i, i\neq j} d(i,j)$$

$$b(i) = \min_{i\neq j} \frac{1}{|C_j|}\sum_{j\in C_j} d(i,j)$$

Όπου:

j = είναι ένα αντικείμενο – σημείο εντός ενός cluster (C)

d(i,j) = είναι η ευκλείδεια απόσταση μεταξύ των αντικειμένων i με j εντός ενός cluster $C_i$

b(i) = είναι η απόσταση ενός μέσου αντικειμένου i με όλα τα υπόλοιπα αντικείμενα εντός ενός άλλου cluster

SI = Ο συνολικός δείκτης SI, είναι ο μέσος όρος των επιμέρους s(i) αναφορικά με όλα τα αντικείμενα του εξεταζόμενου συνόλου δεδομένων

Οι τιμές που μπορεί να πάρει ο δείκτης SI είναι από -1 έως 1. Όταν παίρνει την τιμή 1 υποδηλώνει ότι τα στοιχεία εντός της ομάδας που αναφέρεται είναι απολύτως συμπαγή με την κάθε ομάδα να είναι μακριά από τις υπόλοιπες ομάδες και τα στοιχεία τους. Η τιμή -1 υποδηλώνει ότι οι ομάδες δεν είναι καθόλου διακριτές και τα σημεία που βρίσκονται εντός των ομάδων που έχουν δημιουργηθεί βρίσκονται σε λάθος ομάδες. Οι τιμές που είναι πάνω από 0 θεωρούνται μέτριες, ενώ τιμές που είναι πάνω από 0,5 υποδεικνύουν ότι οι ομάδες που έχουν σχηματιστεί είναι πολύ καλές.

## 3.4 Ανάλυση RFM & K-Means Clustering (Case Studies)

- ***RFM Case study RFM model for customer purchase behavior using K-Means algorithm.***

Στην έρευνα των (Anitha & Malini, 2019) εξετάζεται μία επιχείρηση του λιανικού κλάδου εστιάζοντας στον εντοπισμό πιθανών πελατών εξετάζοντας το ιστορικό των πωλήσεων και την αγοραστική συμπεριφορά των καταναλωτών με σκοπό να τους ομαδοποιήσουν βάσει κοινών χαρακτηριστικών. Το πρώτο βήμα ήταν να γίνει "Exploratory data analysis" ώστε να ελεγχθούν τα δεδομένα, να αφαιρεθούν ακραίες και αρνητικές τιμές που ενδεχομένως να προέκυψαν από λάθος καταχώρηση.

Μετά συνεχίζουν στην κατηγοριοποίηση των πελατών με βάση την RFM ανάλυση χωρίζοντας τους με βάση τα χαρακτηριστικά της τελευταία συναλλαγής, της συχνότητας συναλλαγών, και τέλος του συνολικού ύψους συναλλαγών για να γίνει μία πρώτη τμηματοποίηση της πελατειακής βάσης.

Επόμενο βήμα είναι να γίνει ομαδοποίηση των πελατών σε συστάδες με βάση μοτίβα που ανιχνεύει ο αλγόριθμος K-means. Ο αλγόριθμος K-means ανήκει στους αλγορίθμους μη επιβλεπόμενης μηχανικής μάθησης και δημιουργεί ομάδες (clusters) βασιζόμενος στις αποστάσεις μεταξύ των σημείων του συνόλου δεδομένων που εξετάζουμε. Ένα σημαντικό βήμα για τον αλγόριθμο είναι να γίνει κανονικοποίηση των δεδομένων και scaling στην προ-επεξεργασία που απαιτείται. Αυτό είναι απαραίτητο γιατί παρατηρείται λοξότητα στην κατανομή των αρχικών δεδομένων οπότε μετασχηματίζονται σε κανονική κατανομή κάνοντας λογαριθμοποίηση. Ο αλγόριθμος K-means σε συνδυασμό με την RFM ανάλυση για να κάνουμε segmentation στους πελάτες που εξετάζουμε είναι από τις πιο διαδεδομένες προσεγγίσεις στον κλάδο του μάρκετινγκ.

Ο αλγόριθμος λειτουργεί παίρνοντας ως input τις εγγραφές του dataset και τον αριθμό των cluster που θέτουμε. Ως output παίρνουμε την ομαδοποίηση των παρατηρήσεων μετά την εφαρμογή του αλγορίθμου στα δεδομένα. Ο τρόπος που τρέχει ο αλγόριθμός είναι:

1) Ανάλογα με τον αριθμό k των cluster που θέτουμε, k τυχαία σημεία των δεδομένων επιλέγονται ως κεντρικά σημεία (centroids).

2) Οι αποστάσεις του κάθε σημείου των δεδομένων από τα centroids που έχουν επιλεγεί αξιολογούνται βάσει των ευκλείδειων αποστάσεων τους.
3) Οι τιμές των αποστάσεων τους συγκρίνονται και το κάθε σημείο των δεδομένων ανατίθεται στην ομάδα του centroid που έχει την μικρότερη απόσταση από αυτό.
4) Τα προηγούμενα βήματα επαναλαμβάνονται. Η διαδικασία σταματάει όταν τα cluster που θα προκύψουν είναι ίδια με τα cluster του προηγούμενου βήματος.

Οι ομάδες που προέκυψαν και εξετάστηκαν από την μελέτη είναι τόσο για τρεις ομάδες όσο και για πέντε. Στην επικύρωση των αποτελεσμάτων της cluster analysis με τον δείκτη Silhouette score, και παρατηρούμε ότι είναι μικρότερος από 0,5 για τις ομάδες που έχουν δημιουργηθεί, πράγμα που δείχνει ότι δεν έχουμε τόσο καλή ποιότητα όσο αφορά το πόσο συμπαγείς είναι. Αυτό είναι πιθανόν να συμβαίνει λόγω του μικρού χρονικού εύρους που είχε η έρευνα με στοιχεία μόλις εννέα μηνών. Βάσει του δείκτη επιλέχθηκε να έχουμε K=3 ομάδες αφού έδινε καλύτερο αποτέλεσμα συγκριτικά.

Βάσει των αποτελεσμάτων ερμηνεύτηκε ότι στο cluster 1 έχουμε πελάτες με κύριο χαρακτηριστικό το υψηλό “Recency” οπότε πρόκειται για νέους πελάτες κυρίως αφού δεν υπήρχε υψηλό “Monetary”. Για το cluster 2 πελάτες με κύριο χαρακτηριστικό το υψηλό “Frequency” και “Monetary” οπότε έχουμε τους καλύτερους πελάτες της εταιρείας. Τέλος στο cluster 3 έχουμε πελάτες με χαμηλά RFM score και στα τρία χαρακτηριστικά οπότε πρόκειται για “Lost / Almost lost customers”. Η έρευνα δεν προχωράει σε πρόταση συγκεκριμένων στρατηγικών μάρκετινγκ, κάτι που το αφήνει στην κρίση της εταιρείας.

- ***RFM ranking – An effective approach to customer segmentation.***

Στην έρευνα των (Christy, et al., 2021) αναλύονται τα δεδομένα συναλλαγών επίσης μίας επιχείρησης on-line retailer αρχικά μέσω ανάλυσης RFM. Η δομή της ανάλυσης είναι ίδια με πριν με τη διαφορά ότι γίνεται ομαδοποίηση των πελατών χρησιμοποιώντας τους αλγόριθμους μη – μηχανικής μάθησης K-means αλλά και Fuzzy C-Means συγκρίνοντας τα αποτελέσματα που έχουμε και για τους δύο.

Στα συμπεράσματα αυτής της έρευνας, συγκρίνεται η ομαδοποίηση με βάση τους δύο αλγόριθμους και στα αποτελέσματα βλέπουμε ότι υπερτερεί ο K-means αφού χρειάστηκε λιγότερο χρόνο για να τρέξει και λιγότερες επαναλήψεις για να τελειώσει. Επίσης ο K-

means φαίνεται να είναι καλύτερος όσο αφορά την ποιότητα των ομάδων που έχουν δημιουργηθεί με τα κεντρικά σημεία που έχουν επιλεχθεί (centroids) να βγάζουν περισσότερο νόημα. Το Silhouette score είναι σχεδόν 0,5 δηλαδή είναι κάπως καλύτερο σε σχέση με την προηγούμενη περίπτωση.

Δεν προτείνονται συγκεκριμένες στρατηγικές μάρκετινγκ, κάτι που το αφήνει στην κρίση της επιχείρησης, αλλά εστιάζει στην υπεροχή του αλγορίθμου K-means για εφαρμογές στον κλάδο του μάρκετινγκ.

## 3.5 Στρατηγικός σχεδιασμός Μάρκετινγκ

Ο στρατηγικός σχεδιασμός μίας εταιρεία όσο αφορά τις στρατηγικές μάρκετινγκ που επιλέγει για να εφαρμόσει μπορεί να αποτυπωθεί σε έναν κύκλο που αποτυπώνει μία κυκλική διαδικασία, ξεκινώντας από την έρευνα μάρκετινγκ, μετά επιλέγοντας τις στρατηγικές μάρκετινγκ βάσει των στοιχείων που έχει συλλέξει και τέλος εφαρμόζοντας το πλάνο βάσει συγκεκριμένων ενεργειών.

***Marketing Research***

Η έρευνα μάρκετινγκ αναφέρεται στην κατανόηση της αγοράς που λειτουργεί μία εταιρεία, των ανταγωνιστών της, και των πελατών της. Αυτό γίνεται μέσω της συλλογής δεδομένων, της ανάλυσης τους μέσω περιγραφικής στατιστικής και τέλος μέσω διάφορων εφαρμογών όπως οι αλγόριθμοι μηχανικής μάθησης που αναλύσαμε στα προηγούμενα κεφάλαια μπορούμε να κατανοήσουμε τους πελάτες μέσω των αγοραστικών συνηθειών τους (Huang & Rust, 2021).

***Marketing Strategy***

Επόμενο βήμα είναι η υιοθέτηση στρατηγικών μάρκετινγκ που βασίζεται σε τρία βασικά στοιχεία. Την τμηματοποίηση (Segmentation), την στοχοθέτηση (Targeting) και την θέση που λαμβάνει η εταιρεία (Positioning).

Η τμηματοποίηση, αναφέρεται στην κατάτμηση της αγοράς σε μικρότερα κομμάτια με διαφορετικούς πελάτες σε κάθε διακριτό τμήμα με βάση τις διαφορετικές ανάγκες που έχουν (Wang, et al., 2017). Η στοχοθέτηση αναφέρεται στην επιλογή των σωστών τμημάτων που πρέπει να εστιάζει η εταιρεία για να εφαρμόσει τις δράσεις μάρκετινγκ που θα επιλέξει (Simester, et al., 2020). Η θέση που λαμβάνει η εταιρεία γεφυρώνει τα χαρακτηριστικά των παρεχόμενων προϊόντων με τα οφέλη που απολαμβάνουν οι πελάτες, προβάλλοντας μία ανταγωνιστική θέση στη συνείδηση των πελατών για το προϊόν. Η τοποθέτηση του προϊόντος μέσω της μάρκας (brand) ή της διαφημιστικής καμπάνιας που επιλέγεται βοηθάει στη τοποθέτηση της εταιρείας στη συνείδηση του πελάτη ως τη λύση για τις ανάγκες που έχει (Daabes & Kharbat, 2017).

### *Marketing Actions*

Όσο αφορά τις δράσεις μάρκετινγκ, έχουμε την τυποποίηση, που βάσει των στοιχείων που έχουμε συλλέξει και των στρατηγικών που έχουν τεθεί επαναλαμβάνονται σε κάθε κύκλο. Η προσωποποίηση (Personalization) που αφορά την πρόταση συγκεκριμένων προϊόντων και προσφορών στους πελάτες βελτιώνοντας την εμπειρία τους με την εταιρεία. Αυτό μπορεί να επιτευχθεί μέσω προσωποποιημένη επικοινωνία μέσω mail και διαφημίσεων για κάθε προφίλ πελάτη. Το "relational marketing" ως δράση αναφέρεται στην δημιουργία και διατήρηση μακροχρόνιων σχέσεων με τους πελάτες. Ξεπερνάει την τυπική συναλλαγή που μπορεί να κάνει ένας πελάτης με την εταιρεία και στοχεύει στην οικοδόμηση ισχυρών σχέσεων που βασίζονται στην πιστότητα και την αμοιβαία αξία που θέλουν να έχουν τόσο η εταιρεία όσο και ο πελάτης. Αυτό έχει ως αποτέλεσμα ο πελάτης να προωθεί κι εκείνος έμμεσα ή άμεσα την εταιρεία στον κύκλο του κάτι που έχει πολλαπλασιαστικά οφέλη για την εταιρεία. Συγκεντρωτικά οι δράσεις που μπορούν να υιοθετηθούν είναι οι παρακάτω (Huang & Rust, 2021):

- Πελατοκεντρική προσέγγιση του πελάτη από την εταιρεία κατανοώντας τις μοναδικές ανάγκες και προτιμήσεις του ώστε να απολαμβάνει την καλύτερη δυνατή εμπειρία.
- Αμφίδρομη επικοινωνία μέσω ανατροφοδότησης που θα επιζητεί η εταιρεία μετά από κάθε συναλλαγή.
- Οικοδόμηση εμπιστοσύνης μέσω της τήρησης των υποσχέσεων της εταιρείας για την παροχή του προϊόντος / υπηρεσίας.

- Προγράμματα επιβράβευσης πιστότητας πελατών επιβραβεύοντας τους πελάτες με ανταμοιβές για να ενθαρρύνει τη συνέχιση της συνεργασίας με την εταιρεία.
- Στρατηγικές διατήρησης πελατών μέσω κινήτρων για να μπορούν να απολαμβάνουν την εξυπηρέτηση που προβάλλει η εταιρεία
- Βελτίωση της εμπειρία του πελάτη κατά την πώληση, ώστε η κάθε συναλλαγή με την εταιρεία να αποτελεί μία θετική εμπειρία για αυτόν.

Συνολικά μπορούμε να δούμε συνοπτικά ολόκληρο τον κύκλο που αφορά όλη τη διαδικασία για την εφαρμογή στρατηγικών μάρκετινγκ από την εταιρεία, παρακάτω στην εικόνα 4.


*Εικόνα 4*

## 3.6 Στάδια ζωής πελάτη

Παρακάτω μπορούμε να δούμε πως συνδέεται ο στρατηγικός σχεδιασμός μάρκετινγκ μίας εταιρείας με τα στάδια ζωής του πελάτη και πως μπορεί να γίνει σύνδεση μεταξύ τους με γνώμονα τα προφίλ που μπορεί να έχει κάθε πελάτης ανάλογα με το στάδιο που βρίσκεται.

Όπως μπορούμε να δούμε στο βιβλίο των (Bleier, et al., 2018) μία από τις κυριότερες στρατηγικές για κάθε εταιρεία είναι η προσήλωση των πελατών της σε αυτή. Ο πελάτης δέχεται πληροφορίες συνεχώς από το εξωτερικό περιβάλλον και για το λόγο αυτό πλέον

οι εταιρείες υιοθετούν εστιασμένες στρατηγικές μάρκετινγκ στους πελάτες τους σε αντίθεση με μαζικές απρόσωπες καμπάνιες.

Η σχέση ενός πελάτη μπορεί να κατηγοριοποιηθεί σε τρία στάδια. Το πρώτο στάδιο είναι η απόκτηση του πελάτη (Acquisition) με νέους πελάτες να πραγματοποιούν αγορές. Μετά ακολουθεί το στάδιο της διακράτησης και εξέλιξης του πελάτη (Retention / Development) που εμφανίζει και τη μεγαλύτερη διάρκεια. Τέλος έχουμε τη φθορά (Attrition) που προκύπτει με την πάροδο του χρόνου. Σε κάθε στάδιο που αναφέραμε αντιστοιχεί συνήθως κάποια προφίλ πελατών.

- ***Acquisition (New Customers & Need attention Customers)***

Το στάδιο της απόκτησης του πελάτη ξεκινάει με την πρώτη συναλλαγή που πραγματοποιείται από έναν νέο πελάτη. Ο κύριος στόχος σε αυτό το στάδιο είναι η εταιρεία να μπορέσει να κεντρίσει το ενδιαφέρον του για να πραγματοποιήσει και μελλοντικά αγορές. Αυτοί αρχικά είναι νέοι πελάτες και μετά από ένα χρονικό διάστημα είναι πελάτες που δεν έχουν προλάβει να κάνουν ακόμη πολλές συναλλαγές. Με την πάροδο του χρόνου, θα μπορούσαν να εξελιχθούν σε πελάτες αξίας ανάλογα σε ποιο σημείο της καμπύλης της εικόνας 5 εστιάζουμε.

.

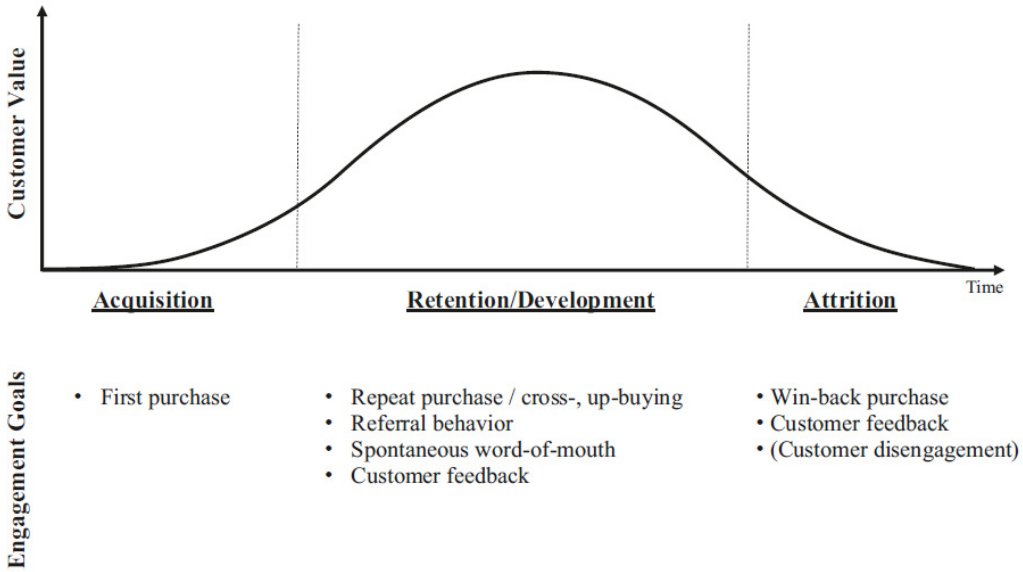
- ***Retention / Development (Best Customers / Need attention)***

Στο στάδιο της διακράτησης – εξέλιξης του πελάτη ο στόχος είναι να πραγματοποιούνται επαναλαμβανόμενες αγορές, να αγοράζονται διαφορετικά προϊόντα συνδυαστικά (cross-selling), να αγοράζονται ακριβότερα προϊόντα (up-selling), ώστε να μεγιστοποιηθούν τα έσοδα της εταιρείας. Επίσης θα πρέπει να δημιουργηθεί μία σύνδεση του πελάτη με την εταιρεία ώστε να την προτείνει στον κύκλο του (word of mouth). Μέσω της έλευσης νέων πελατών από πρόταση υφιστάμενου πελάτη (referral), η εταιρεία έχει τη δυνατότητα να αποκτήσει νέους πελάτες τους οποίους δεν είχε καταφέρει να τους προσεγγίσει με τα παραδοσιακά κανάλια μάρκετινγκ. Τέλος η ανατροφοδότηση που προέρχεται από τους πελάτες σε αυτό το στάδιο, μπορεί να βοηθήσει την εταιρεία να καταλάβει τις επιθυμίες και ανάγκες των πελατών, πράγμα που μπορεί να χρησιμοποιήσει για να βελτιώσει τις προσφορές της αλλά και όλα τα “touchpoints” του πελάτη με την εταιρεία κατά την αγορά. Σε αυτό το στάδιο θα βρούμε τους πελάτες που χαρακτηρίζονται ως “Best

customers" και είναι πελάτες αξίας, ενώ στην καμπύλη που δείχνει με την πάροδο του χρόνου να φθίνει στην εικόνα 5 θα δούμε τους πελάτες που χρήζουν προσοχής (Need attention). Ο στόχος της εταιρείας είναι να κρατήσει τους πελάτες στο κέντρο του διαγράμματος που απεικονίζεται στην εικόνα 5 για όσο μεγαλύτερο χρονικό διάστημα μπορούν.

- ***Attrition (Lost / Almost lost Customers)***

Τέλος έχουμε το στάδιο της φθοράς (Attrition), όπου οι σχέσεις μεταξύ του πελάτη και της εταιρείας χειροτερεύουν, και ενδέχεται να σταματήσουν με την πάροδο του χρόνου. Οι πελάτες που βλέπουμε σε αυτό το στάδιο μπορούνε να χαρακτηριστούν ως "Lost / Almost lost Customers". Αυτό μπορεί να προέρχεται είτε από την πλευρά του πελάτη σε περίπτωση που έχει ξεκινήσει να δυσαρεστείται για κάποιο λόγο, είτε από την πλευρά της εταιρείας σε περίπτωση που ο πελάτης έχει αρχίσει να μην είναι τυπικός στις υποχρεώσεις του με τις πληρωμές. Στην περίπτωση που προέρχεται από την πλευρά του πελάτη, η εταιρεία μπορεί να κάνει κάποιες ενέργειες για να κεντρίσει το ενδιαφέρον του πελάτη εκ νέου, μέσω προσωποποιημένων προσφορών (win-back actions). Επίσης μπορεί να ζητήσει ανατροφοδότηση από τον πελάτη ώστε να καταλάβει τους λόγους που τον ώθησαν να φύγει και να βελτιωθεί σε αυτούς τους τομείς. Σε περίπτωση που η φθορά προέρχεται από μέρους της εταιρείας, μπορεί να έχει επιλέξει την απεμπλοκή από κάποιον συγκεκριμένο πελάτη που δεν προσθέτει κάποια αξία για αυτή.

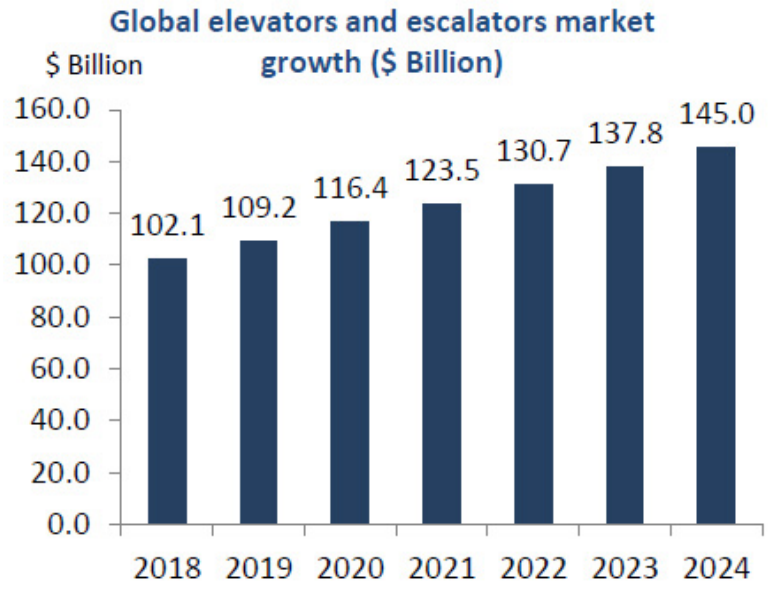



*Εικόνα 5*

## 3.7 Τμηματοποίηση αγοράς στον κλάδο του ανελκυστήρα - ανυψωτικών

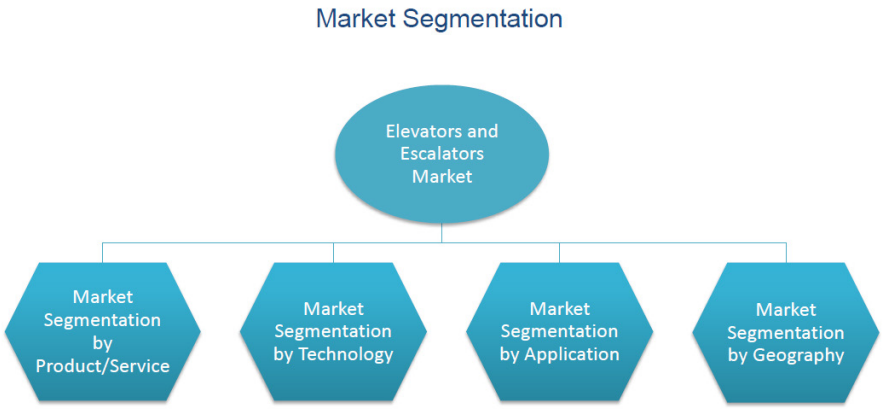
Ο συγκεκριμένος κλάδος αγοράς που θα αναλύσουμε και αφορά τον τομέα του ανελκυστήρα είναι αρκετά ειδικός με αποτέλεσμα να μην υπάρχει πλήθος επιστημονικών δημοσιεύσεων που να μας δίνουν εκτενή στοιχεία μέχρι στιγμής. Για το λόγο αυτό διευρύναμε το πλαίσιο της αναζήτησης μας και συγκεντρώσαμε στοιχεία από έρευνα αγοράς που έγινε από την "Ameliorate - Market Insights Reports" για λογαριασμό της εξεταζόμενης βιομηχανίας ανελκυστήρα (Ameliorate, 2019).

Σε αυτή την έρευνα έχουμε στοιχεία από το 2018 ως βάση, με προβλέψεις έως και το 2024. Βλέπουμε ότι η παγκόσμια αγορά ανελκυστήρα αναμένεται να ξεπεράσει τα 145 δισεκατομμύρια ευρώ το 2024 κάτι το οποίο βασίζεται στην αυξημένη ζήτηση για ανελκυστήρες που στηρίζονται στην ολοένα και μεγαλύτερη ανοικοδόμηση με όλο και ψηλότερα κτήρια. Παρατηρούμε ότι με βάση τα στοιχεία που έχουμε, εξετάζοντας το χρονικό πλαίσιο από το 2018 έως και τη πρόβλεψη του 2024 έχουμε μία σταθερή ετήσια αύξηση περίπου 6%.




*Εικόνα 6*

Η τμηματοποίηση της αγοράς στον κλάδο των ανελκυστήρων μπορεί να γίνει με γνώμονα τέσσερις διαφορετικούς παράγοντες. Με βάση τον τύπο του προϊόντος / υπηρεσίας, με βάση την τεχνολογία που χρησιμοποιείται, με βάση το πεδίο εφαρμογής και τέλος με βάση τη γεωγραφική ζώνη της κάθε αγοράς που απευθύνεται το προϊόν.




*Εικόνα 7*

**Τμηματοποίηση βάσει προϊόντος / υπηρεσίας**

Στον κλάδο του ανελκυστήρα και συστημάτων ανύψωσης έχουμε τρεις διακριτές πηγές εσόδων. Έσοδα που προέρχονται από την πώληση του προϊόντος, έσοδα που προέρχονται

από την εγκατάσταση του προϊόντος και έσοδα που προέρχονται από τη συντήρηση του. Οι βιομηχανική εταιρεία που εξετάζουμε έχει ως κύρια πηγή εσόδων την πώληση του προϊόντος με κύριους πελάτες της άλλες εταιρείες εγκατάστασης και συντήρησης, έχουμε δηλαδή B2B μοντέλο.

Όσο αφορά τους τύπου προϊόντων ανελκυστήρα, έχουμε τους υδραυλικούς ανελκυστήρες με τις υποκατηγορίες του, τους μηχανικούς ανελκυστήρες με τις υποκατηγορίες τους, τα αναβατόρια, τους πνευματικούς ανελκυστήρες και τέλος μία ευρύτερη κατηγορία λοιποί που συγκαταλέγονται ειδικοί τύποι ανελκυστήρων όπως οι ανελκυστήρες για πλοία, για βιομηχανικές εξέδρες, ανελκυστήρες για πάρκινγκ αυτοκινήτων και άλλοι. Παρακάτω (εικόνα 8), μπορούμε να δούμε τα στοιχεία εσόδων από την πώληση νέων ανελκυστήρων παγκοσμίως, από το έτος 2018 και τις προβλέψεις που αναμένεται να φτάσουν μέχρι και το 2024 αναλυτικά για κάθε κατηγορία.

**Market analysis by type ($ Billion)**

| Market Segment | Sub-segment | | | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | CAGR (2018 - 2024)% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Elevators | | | | | | | | | | | |
| | Market analysis by hoist mechanism | | | | | | | | | | |
| | | Hydraulic elevators | | | | | | | | | |
| | | | Holed (conventional) | 2.113 | 2.297 | 2.486 | 2.680 | 2.880 | 3.085 | 3.296 | 7.70 |
| | | | Hole-less | 1.801 | 1.961 | 2.126 | 2.296 | 2.472 | 2.652 | 2.838 | 7.88 |
| | | | Roped | 1.501 | 1.594 | 1.685 | 1.774 | 1.861 | 1.945 | 2.027 | 5.14 |
| | | | Others | 1.117 | 1.206 | 1.295 | 1.386 | 1.479 | 1.573 | 1.668 | 6.90 |
| | | | | | | | | | | | |
| | | | Total | 6.531 | 7.057 | 7.592 | 8.137 | 8.691 | 9.255 | 9.829 | 7.05 |
| | | Traction elevators | | | | | | | | | |
| | | | Geared | 1.833 | 1.989 | 2.148 | 2.311 | 2.478 | 2.650 | 2.825 | 7.47 |
| | | | Gear-less | 1.617 | 1.737 | 1.858 | 1.981 | 2.104 | 2.228 | 2.352 | 6.45 |
| | | | Machine-room-less | 1.330 | 1.443 | 1.559 | 1.678 | 1.800 | 1.925 | 2.053 | 7.50 |
| | | | Others | 1.090 | 1.164 | 1.238 | 1.312 | 1.385 | 1.457 | 1.529 | 5.80 |
| | | | | | | | | | | | |
| | | | Total | 5.870 | 6.333 | 6.804 | 7.281 | 7.767 | 8.259 | 8.759 | 6.90 |
| | | Climbing elevators | | 5.406 | 5.762 | 6.115 | 6.464 | 6.809 | 7.151 | 7.488 | 5.58 |
| | | Pneumatic elevators | | 4.691 | 5.087 | 5.493 | 5.908 | 6.333 | 6.767 | 7.212 | 7.43 |
| | | Others | | 3.497 | 3.725 | 3.951 | 4.175 | 4.395 | 4.613 | 4.829 | 5.53 |
| | | | | | | | | | | | |
| | | Total | | 25.99 | 27.96 | 29.95 | 31.96 | 34.00 | 36.05 | 38.12 | 6.59 |

*Εικόνα 8*

Αντιστοίχως μπορούμε να δούμε από την κατηγορία με τις κυλιόμενες σκάλες και τα συστήματα ανύψωσης (εικόνα 9), μία παρόμοια απεικόνιση με το σύνολο των πληροφοριών παγκοσμίως από το έτος 2018 έως και την πρόβλεψη για το έτος 2024.

**Market analysis by type ($ Billion)**

| Market Segment | Sub-segment | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | CAGR (2018 - 2024)% |
|---|---|---|---|---|---|---|---|---|---|
| Escalators | | | | | | | | | |
| | Wheelchair escalator | 5.138 | 5.513 | 5.890 | 6.268 | 6.648 | 7.029 | 7.411 | 6.30 |
| | Levytator | 4.894 | 5.239 | 5.585 | 5.930 | 6.275 | 6.621 | 6.966 | 6.06 |
| | Moving walkway | 4.703 | 4.931 | 5.147 | 5.350 | 5.540 | 5.719 | 5.885 | 3.81 |
| | Spiral escalator | 4.279 | 4.537 | 4.789 | 5.036 | 5.278 | 5.514 | 5.745 | 5.03 |
| | Parallel | 3.843 | 4.120 | 4.397 | 4.675 | 4.953 | 5.233 | 5.512 | 6.20 |
| | Multiple parallel | 3.358 | 3.615 | 3.874 | 4.136 | 4.400 | 4.667 | 4.936 | 6.63 |
| | Crisscross | 3.067 | 3.298 | 3.532 | 3.767 | 4.004 | 4.243 | 4.484 | 6.54 |
| | Curved escalator | 2.627 | 2.836 | 3.048 | 3.263 | 3.481 | 3.702 | 3.926 | 6.93 |
| | Others | 4.457 | 4.674 | 4.879 | 5.072 | 5.254 | 5.425 | 5.584 | 3.83 |
| | | | | | | | | | |
| | Total | 36.367 | 38.763 | 41.140 | 43.497 | 45.834 | 48.152 | 50.449 | 5.61 |

*Εικόνα 9*

Πέρα από αυτές τις δύο διακριτές κατηγορίες που αφορούν τους ανελκυστήρες και τις κυλιόμενες σκάλες μαζί με τα συστήματα ανύψωσης, έχουμε και έσοδα που δημιουργούνται από την πώληση μεμονωμένων υποσυστημάτων ανελκυστήρα κάτι που συναντούμε σε περιπτώσεις ανακαίνισης, ανταλλακτικών που είναι αναλώσιμα εξαρτήματα, όπως και από την συντήρηση. Υπό αυτό το πρίσμα, μπορούμε να έχουμε μία συνολική, συγκεντρωτική εικόνα από την εικόνα 10 παρακάτω.

**Market analysis by service type ($ Billion)**

| Market Segment | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | CAGR (2018 - 2024)% |
|---|---|---|---|---|---|---|---|---|
| New equipment | 39.474 | 42.289 | 45.110 | 47.937 | 50.771 | 53.612 | 56.459 | 6.15 |
| Maintenance and repair | 34.270 | 36.703 | 39.141 | 41.582 | 44.028 | 46.478 | 48.933 | 6.12 |
| Refurbishing | 28.348 | 30.251 | 32.144 | 34.027 | 35.898 | 37.759 | 39.609 | 5.73 |

*Εικόνα 10*

### Τμηματοποίηση βάσει τεχνολογίας

Μία άλλη κατηγοριοποίηση των προϊόντων / υπηρεσιών όσο αφορά τον εξεταζόμενο κλάδο μπορεί να προκύψει βάσει της τεχνολογίας που χρησιμοποιείται για κάθε προϊόν (Εικόνα 11).

**Market analysis by technology ($ Billion)**

| Market Segment | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 | 2024 | CAGR (2018 - 2024)% |
|---|---|---|---|---|---|---|---|---|
| Port technology | 13.493 | 14.508 | 15.533 | 16.566 | 17.609 | 18.661 | 19.722 | 6.53 |
| Multi-directional | 12.717 | 13.655 | 14.598 | 15.547 | 16.502 | 17.464 | 18.431 | 6.38 |
| Internet of vertical transportation | 11.826 | 12.716 | 13.615 | 14.523 | 15.438 | 16.361 | 17.293 | 6.54 |
| Linear motor technology | 10.335 | 11.107 | 11.886 | 12.671 | 13.462 | 14.260 | 15.063 | 6.48 |
| Magnetic levitation | 9.840 | 10.581 | 11.330 | 12.085 | 12.847 | 13.616 | 14.392 | 6.54 |
| Microprocessor-based controls | 8.714 | 9.363 | 10.018 | 10.677 | 11.342 | 12.011 | 12.686 | 6.46 |
| In-cab sensors | 7.726 | 8.201 | 8.668 | 9.126 | 9.575 | 10.016 | 10.448 | 5.16 |
| Destination dispatch control software | 6.587 | 6.967 | 7.337 | 7.696 | 8.045 | 8.382 | 8.709 | 4.76 |
| Personalized elevator calls | 5.739 | 6.190 | 6.647 | 7.110 | 7.580 | 8.055 | 8.538 | 6.84 |
| Others | 15.115 | 15.954 | 16.764 | 17.545 | 18.298 | 19.022 | 19.717 | 4.53 |

*Εικόνα 11*

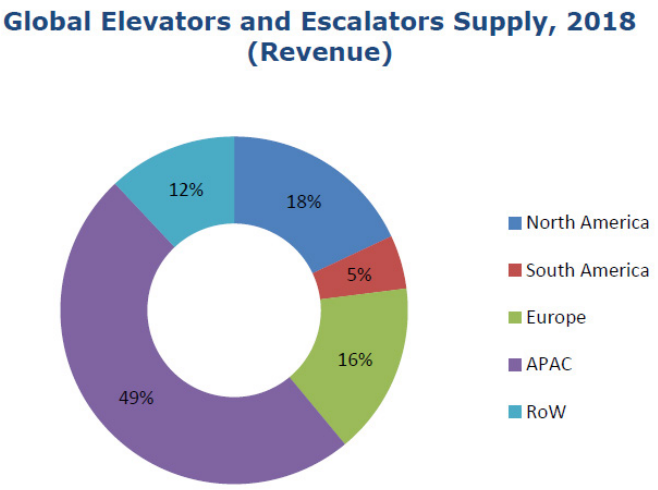
### Τμηματοποίηση βάσει γεωγραφικής περιοχής

Στην έρευνα από την οποία πήραμε τα στοιχεία, η παγκόσμια αγορά έχει χωριστεί σε πέντε διαφορετικές γεωγραφικές ζώνες:

- North America
- South America
- Europe
- APAC (Asia Pacific)
- RoW (Rest of the world)

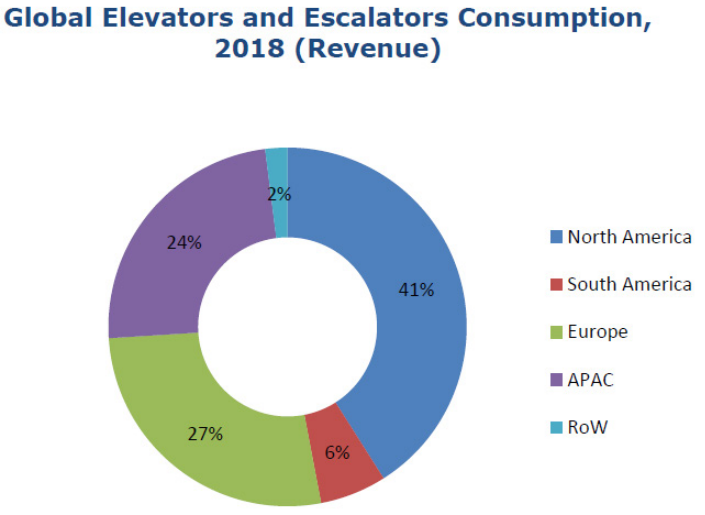
Με βάση αυτό το διαχωρισμό, μπορούμε να δούμε αρχικά την τμηματοποίηση της αγοράς όσο αφορά την προσφορά σε ανελκυστήρες, δηλαδή που παράγονται. Με αυτό τον τρόπο βλέπουμε τις μονάδες παραγωγής του ανταγωνισμού. Παρατηρούμε ότι σχεδόν οι μισοί ανελκυστήρες παράγονται στην γεωγραφική ζώνη APAC σε ποσοστό

49%, ακολουθούν με διαφορά η North America με ποσοστό 18%, η Europe με 16%, ενώ RoW έχει 12% και South America 5% (Εικόνα 12).




*Εικόνα 12*

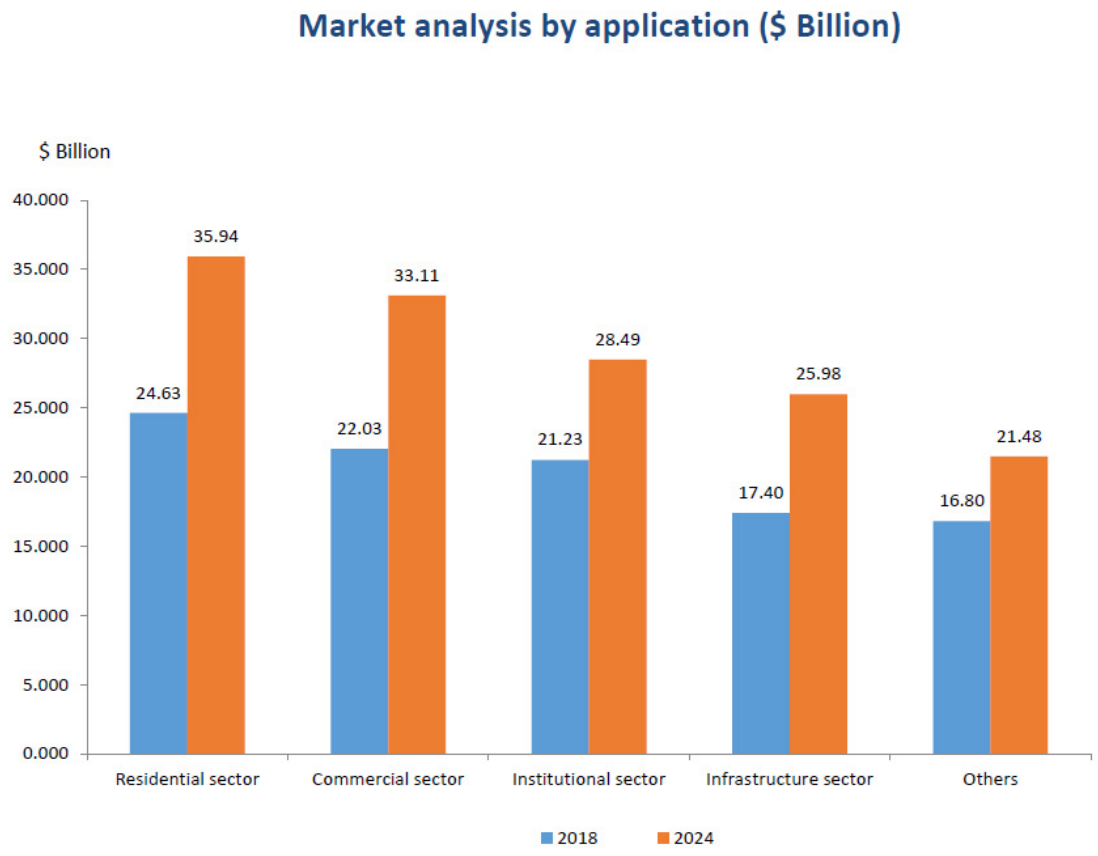
Αντίστοιχα παρακάτω μπορούμε να δούμε τη ζήτηση που σημειώνεται για ανελκυστήρες στις αντίστοιχες γεωγραφικές ζώνες που αντικατοπτρίζει και τη διασπορά των εν δυνάμει πελατών. Παρατηρούμε ότι η υψηλότερη ζήτηση σημειώνεται στη North America με ποσοστό 41%, ακολουθεί η Europe με 27% και πολύ κοντά η APAC με 24%, ενώ η South America και η γεωγραφική ζώνη που έχει προσδιοριστεί ως RoW έχουν μονοψήφια ποσοστά (Εικόνα 13).




*Εικόνα 13*

**Τμηματοποίηση βάσει πεδίου εφαρμογής**

Όσο αφορά το πεδίο εφαρμογής των ανελκυστήρων αυτοί που έχουν αναγνωριστεί από τη έρευνα είναι οι ανελκυστήρες μπορούν να ομαδοποιηθούν ανάλογα με το πεδίο εφαρμογής σε ανελκυστήρες για κατοικία, για εμπορική χρήση όπως εμπορικά κέντρα, για κρατικά ιδρύματα / οργανισμούς / γραφεία, για μέρη όπου υπάρχουν υποδομές είτε βιομηχανικές είτε μεταφορικές, και τέλος μία κατηγορία με όλες τις υπολειπόμενες εφαρμογές που θα μπορούσαν να χρησιμοποιηθούν. Στην εικόνα 14, παρατηρούμε ότι τα προϊόντα που αφορούν κατοικίες είναι η κυρίαρχη κατηγορία, μετά ακολουθεί η κατηγορία με προϊόντα στον εμπορικό τομέα, μετά τα προϊόντα που έχουν ως πεδίο εφαρμογής ιδρύματα, μετά είναι τα προϊόντα που για έργα υποδομών και τέλος έχουμε μία ευρύτερη κατηγορία που συγκαταλέγονται όλα τα υπόλοιπα πεδία εφαρμογής που δεν συμπεριλαμβάνονται στις προαναφερθείσες κατηγορίες.




*Εικόνα 14*

# 4 Μεθοδολογία Έρευνας

Η επιλογή της ερευνητικής διαδικασίας έγινε με γνώμονα την κατηγοριοποίηση των πελατών σε ομογενείς ομάδες που θα βασίζονται σε κοινά χαρακτηριστικά. Όπως είδαμε και στην ανασκόπηση της βιβλιογραφίας, για τις εφαρμογές που θέλουμε να ομαδοποιήσουμε πελάτες στον κλάδο του μάρκετινγκ επιλέγουμε να κάνουμε RFM ανάλυση σε συνδυασμό με ομαδοποίηση με τον αλγόριθμο K – means.

Για την ανάλυση μας, έγινε χρήση δεδομένων πωλήσεων ενός εκτεταμένου εύρους τεσσεράμισι ετών της εξεταζόμενης βιομηχανικής επιχείρησης, τα οποία πήραμε από το CRM (Customer Relationship Management) σύστημα της. Τα δεδομένα αφορούν το χρονικό πλαίσιο από 01/01/2017 έως 30/06/2022.

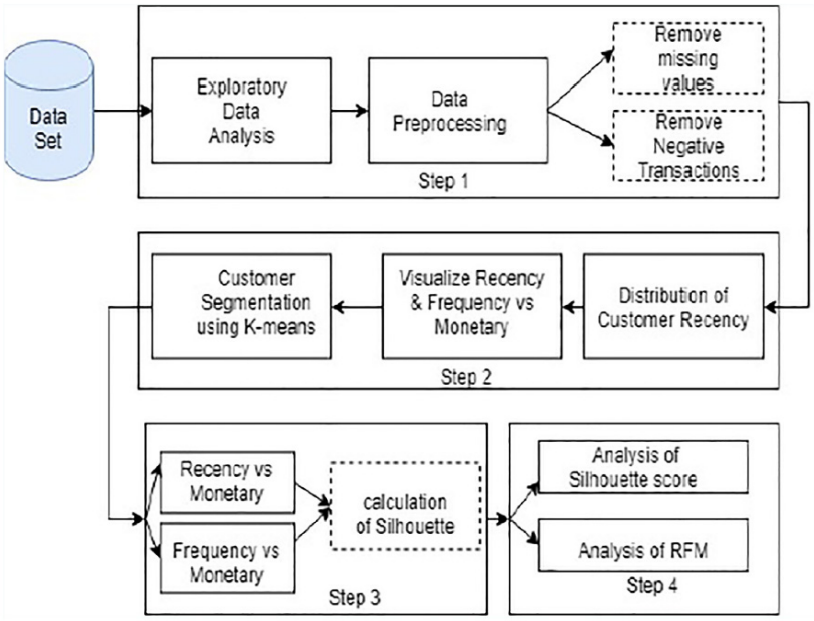
Τα βήματα που έχουν επιλεχθεί για την έρευνα αφού συγκεντρώσαμε το σύνολο δεδομένων που θέλουμε να προχωρήσουμε με την ανάλυση μας είναι να κάνουμε Exploratory data analysis στο χρονικό πλαίσιο που έχουμε επιλέξει, ώστε να μπορέσουμε να καταλάβουμε τα δεδομένα μας αλλά και τη θέση της εταιρείας μέσω των στοιχείων που θα παρουσιαστούν. Για την ανάλυση αυτή έχει χρησιμοποιηθεί το Business Intelligence λογισμικό το οποίο χρησιμοποιεί η εταιρεία όσο και επιπλέον ανάλυση που έχει γίνει σε περιβάλλον της Python.

Επόμενο βήμα που θα εξετάσουμε εκτενώς στο κεφάλαιο των αποτελεσμάτων, είναι να αναλύσουμε τα στοιχεία πωλήσεων ακολουθώντας την RFM Analysis σε Python, χωρίζοντας δηλαδή την βάση των πελατών σε σχέση με κάποια βασικά χαρακτηριστικά που αφορούν την πιο πρόσφατη συναλλαγή κάθε πελάτη (Recency), την συχνότητα των συναλλαγών (Frequency) που έχει πραγματοποιήσει κάθε πελάτης μέσα στο εξεταζόμενο χρονικό πλαίσιο που έχει επιλεχθεί να μελετήσουμε, και τέλος το ύψος των συνολικών συναλλαγών (Monetary) που είχε ο κάθε πελάτης στο ίδιο χρονικό πλαίσιο.

Μετέπειτα για την δημιουργία ομάδων – συστάδων πελατών με παρόμοια χαρακτηριστικά που στηρίζονται στα χαρακτηριστικά των πρόσφατων συναλλαγών, της συχνότητας τους και του συνολικού ύψους συναλλαγών έχει επιλεχθεί να χρησιμοποιήσουμε έναν αλγόριθμο μη επιβλεπόμενης μηχανικής μάθησης, τον K-means με την Python ώστε να δούμε συστάδες πελατών που σχηματίζονται από μοτίβα που ανιχνεύει ο αλγόριθμος στο σύνολο δεδομένων που εξετάζουμε.

Με βάση τα αποτελέσματα της προσέγγισης που έχουμε κάνει μέσω της RFM Analysis και του K-Means Clustering, θα οπτικοποιήσουμε τα αποτελέσματα χρησιμοποιώντας διάφορα γραφήματα και θα επιχειρήσουμε να τα εξηγήσουμε. Επίσης θα χρησιμοποιήσουμε τον δείκτη Silhouette score για να επικυρώσουμε ότι τα αποτελέσματα της ομαδοποίησης που έχουμε πραγματοποιήσει μέσω του clustering algorithm είναι αποδεκτά.

Μία σύνοψη των βημάτων που θα ακολουθήσουμε μπορούμε να δούμε από το παρακάτω διάγραμμα ροής στην Εικόνα 15.




*Εικόνα 15*

Τέλος, θα γίνει σχολιασμός των αποτελεσμάτων, θα βγάλουμε τα συμπεράσματα που προκύπτουν μέσω της ανάλυσης που έγινε και θα κάνουμε τις προτάσεις μας προς την επιχείρηση για τις επόμενες κινήσεις που θα πρέπει να κάνει βάσει της ανάλυσης που έχει πραγματοποιηθεί.

Ένα σημαντικό σημείο στο οποίο θα πρέπει να γίνει αναφορά, είναι ότι στα πλαίσια της συμφωνίας εμπιστευτικότητας που έχει γίνει με την εταιρεία, έχει επιλεχθεί να μην δημοσιευτούν ευαίσθητα οικονομικά στοιχεία ή στοιχεία πελατών και αγορών.

## 4.1 Επεξήγηση ερευνητικής διαδικασίας

Αρχικά εισάγουμε τα δεδομένα σε ένα jupyter notebook και βλέπουμε την δομή του πίνακα και τον τύπο των δεδομένων. Κάνουμε έλεγχο στα δεδομένα μας και κάνουμε καθαρισμό αφαιρώντας εγγραφές (παραγγελίες) οι οποίες είναι εκθεσιακές και δεν αφορούν κάποιους συγκεκριμένους πελάτες οπότε έχουν μηδενική τιμή. Επίσης με την ίδια λογική, αφαιρούμε παραγγελίες αναπλήρωσης θυγατρικών οι οποίες δεν απευθύνονται σε τελικούς πελάτες.

Στη συνέχεια μετασχηματίζουμε τα δεδομένα μετατρέποντας το όνομα του πελάτη σε κατηγορική μεταβλητή, την στήλη προώθηση παραγγελίας σε στήλη τύπου datetime καθώς και μετονομάζουμε τις στήλες σε λατινικά ονόματα ώστε να μην αντιμετωπίσουμε τυχόν προβλήματα με βιβλιοθήκες της python.

## 4.2 Ανάλυση δείγματος της έρευνας

Σε αυτό το σημείο θα δείξουμε κάποια στοιχεία που έχουν να κάνουν με την ανάλυση των πωλήσεων, υπολογίζοντας κάποια στοιχεία που έχουν νόημα για να κατανοήσουμε καλύτερα τα δεδομένα μας.

Αρχικά θέλουμε να δούμε τον αριθμό των διακριτών πελατών οι οποίοι είχαν πραγματοποιήσει συναλλαγή/συναλλαγές μέσα στο διάστημα που εξετάζουμε και βλέπουμε ότι έχουμε συνολικά 2.243 διαφορετικούς πελάτες βάσει του αριθμού Customer ID που αποτελεί στην ουσία την ταυτότητα του πελάτη στο σύστημα της εταιρείας. Ο αριθμός των συναλλαγών που έχουν πραγματοποιηθεί είναι 33.629 και κάθε συναλλαγή αποτυπώνει την πώληση ενός προϊόντος.

Τώρα όσο αφορά τις αγορές που πραγματοποιήθηκαν πωλήσεις, βλέπουμε ότι έγιναν πωλήσεις σε 109 διαφορετικές χώρες στο διάστημα που εξετάσαμε, πράγμα που δείχνει τον έντονα εξαγωγικό χαρακτήρα της επιχείρησης. Οι κυριότερες 10 αγορές που δραστηριοποιείται, φαίνονται παρακάτω στον Πίνακα 1. Εδώ βλέπουμε ότι αυτές οι 10 χώρες εκπροσωπούν περίπου το 69% του συνολικού όγκου εσόδων και αντίστοιχα περίπου το 75% του συνόλου των προϊόντων που έχουν παραχθεί. Η ελληνική αγορά βλέπουμε ότι είναι στην πρώτη θέση και στις δύο κατηγορίες (ποσότητα προϊόντων και έσοδα), πράγμα που δείχνει ότι η εταιρεία είναι ηγέτιδα στην ελληνική αγορά.

| Market | Sales (euros) | Count of Order |
|---|---|---|
| GREECE | 12.22% | 27.30% |
| UNITED KINGDOM | 8.84% | 6.41% |
| AUSTRALIA | 8.35% | 4.48% |
| GERMANY | 7.54% | 7.13% |
| RUSSIAN FEDERATION | 6.01% | 2.64% |
| ROMANIA | 4.85% | 4.77% |
| BELGIUM | 4.13% | 3.87% |
| FRANCE | 3.79% | 2.07% |
| SERBIA | 3.76% | 5.93% |
| SWEDEN | 3.38% | 2.10% |
| TURKEY | 3.37% | 6.09% |
| HUNGARY | 2.68% | 2.10% |

Πίνακας 1

Επόμενο βήμα είναι να υπολογίσουμε την γκάμα των προϊόντων που έχουν πουληθεί μέσα στο χρονικό ορίζοντα που εξετάζουμε και βλέπουμε ότι έχουμε συνολικά 49 διακριτά προϊόντα που έχουν πουληθεί.

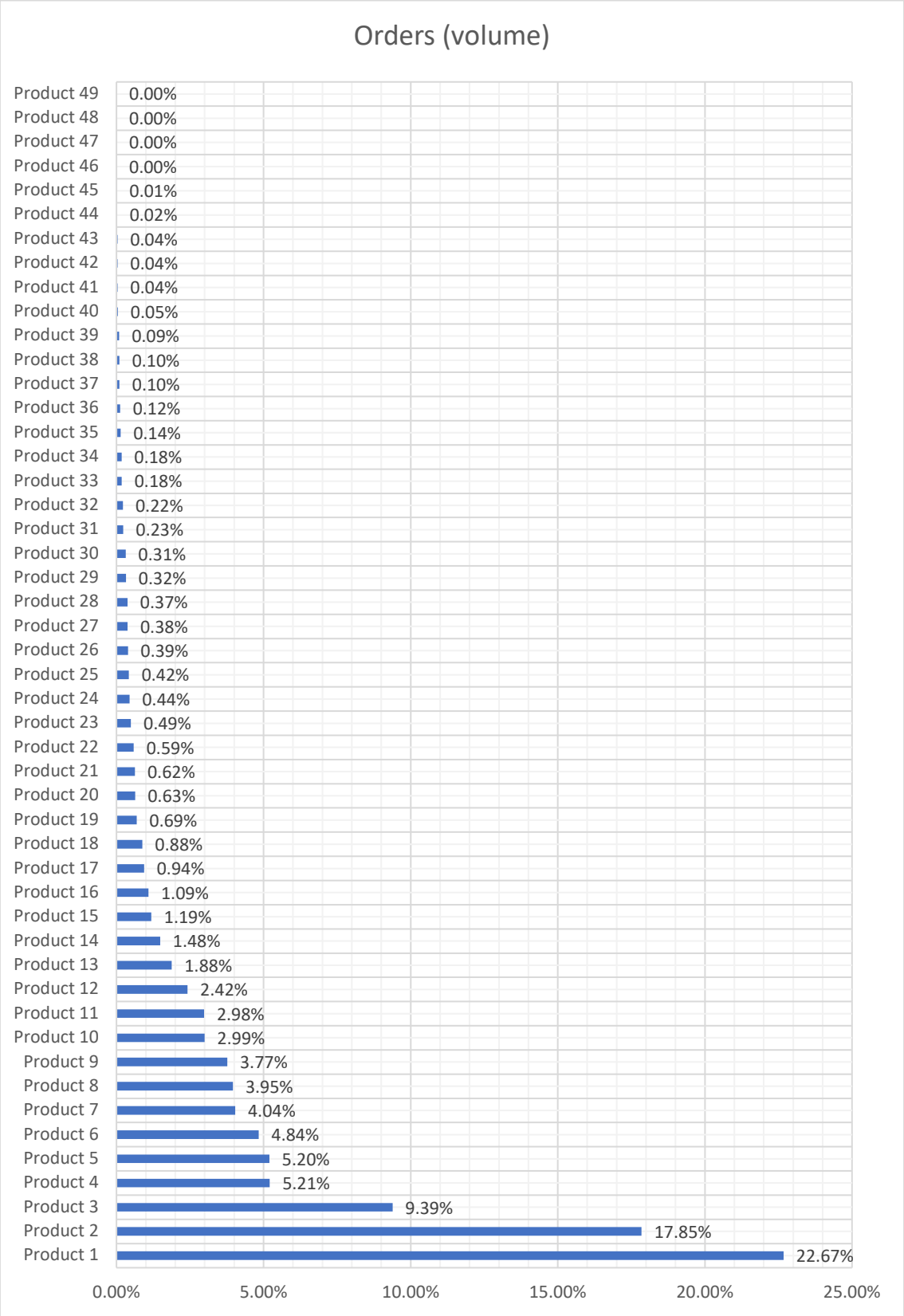
Παρατηρούμε ότι περίπου το 80% του συνολικού όγκου πωλήσεων σε μονάδες προϊόντων αποτελείται από 10 διαφορετικούς τύπους προϊόντων, δηλαδή σε φθίνουσα κατανομή τα Product 1, Product 2, Product 3, Product 4, Product 5, Product 6, Product 7, Product 8, Product 9, Product 10 όπως μπορούμε να δούμε στο Γράφημα 1 - Orders (volume) παρακάτω.

Τώρα αν θέλουμε να δούμε την αντίστοιχη απεικόνιση σε έσοδα που έχει αποφέρει συνολικά ο κάθε τύπος προϊόντος μπορούμε να δούμε το Γράφημα 2 - Orders amount (euros). Εδώ βλέπουμε ότι περίπου τα συνολικά έσοδα που αντιστοιχούν στο 80% προέρχονται από 10 τύπους προϊόντων οι οποίοι είναι σε αύξουσα κατανομή τα Product 1, Product 4, Product 2, Product 3, Product 8, Product10, Product 5, Product 7, Product 11, Product 13.

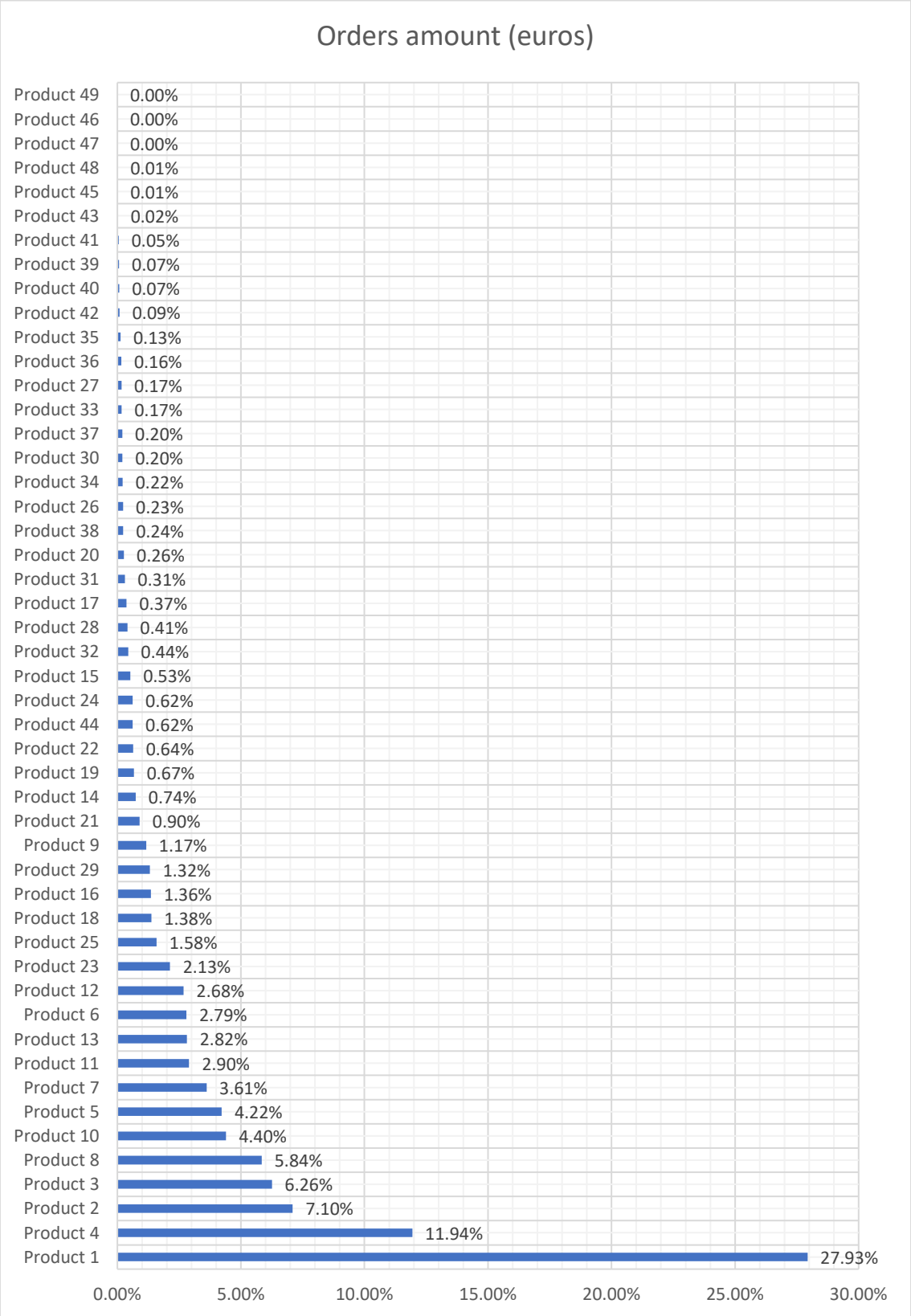
Ένα πρώτο συμπέρασμα που μπορούμε να βγάλουμε είναι ότι το Product 4 που ενώ σε όγκο προϊόντων είναι τέταρτο, σε ύψος εσόδων είναι δεύτερο. Το Product 8 ενώ σε όγκο προϊόντων είναι όγδοο, σε ύψος εσόδων είναι πέμπτο. Το Product 10 ενώ σε όγκο προϊόντων είναι δέκατο, σε ύψος εσόδων είναι έκτο. Το Product 13 ενώ σε όγκο

προϊόντων είναι δέκατο τρίτο σε ύψος εσόδων είναι δέκατο. Αυτά τα προϊόντα που αναφέρθηκαν είναι προϊόντα premium που η εταιρεία έχει ανταγωνιστικό πλεονέκτημα έναντι της αγοράς και μπορεί να τα πουλάει σε ακριβότερη τιμή. Επίσης θα μπορούσαν να αποτελούν στόχο της εμπορικής διεύθυνσης για να κατευθύνουν τους πελάτες που θέλουν να εφαρμόσουν upsell τακτική ώστε να προωθήσουν τα συγκεκριμένα προϊόντα.

Εδώ να σημειώσουμε ότι έχουμε μετασχηματίσει το ύψος των συναλλαγών από € σε ποσοστό επί του συνόλου (κατόπιν συνεννόησης με την επιχείρηση για να μην δημοσιεύσουμε ακριβή οικονομικά στοιχεία) όπως και ότι έχουμε μετονομάσει τα προϊόντα της εταιρείας σε Product X για τον ακριβώς ίδιο λόγο. Στα γραφήματα έχει μετασχηματιστεί και το γράφημα που εκφράζει ποσότητες σε ποσοστό επί του συνόλου για λόγους ομοιογένειας.


Orders (volume)
Product 49 0.00%
Product 48 0.00%
Product 47 0.00%
Product 46 0.00%
Product 45 0.01%
Product 44 0.02%
Product 43 0.04%
Product 42 0.04%
Product 41 0.04%
Product 40 0.05%
Product 39 0.09%
Product 38 0.10%
Product 37 0.10%
Product 36 0.12%
Product 35 0.14%
Product 34 0.18%
Product 33 0.18%
Product 32 0.22%
Product 31 0.23%
Product 30 0.31%
Product 29 0.32%
Product 28 0.37%
Product 27 0.38%
Product 26 0.39%
Product 25 0.42%
Product 24 0.44%
Product 23 0.49%
Product 22 0.59%
Product 21 0.62%
Product 20 0.63%
Product 19 0.69%
Product 18 0.88%
Product 17 0.94%
Product 16 1.09%
Product 15 1.19%
Product 14 1.48%
Product 13 1.88%
Product 12 2.42%
Product 11 2.98%
Product 10 2.99%
Product 9 3.77%
Product 8 3.95%
Product 7 4.04%
Product 6 4.84%
Product 5 5.20%
Product 4 5.21%
Product 3 9.39%
Product 2 17.85%
Product 1 22.67%
0.00% 5.00% 10.00% 15.00% 20.00% 25.00%

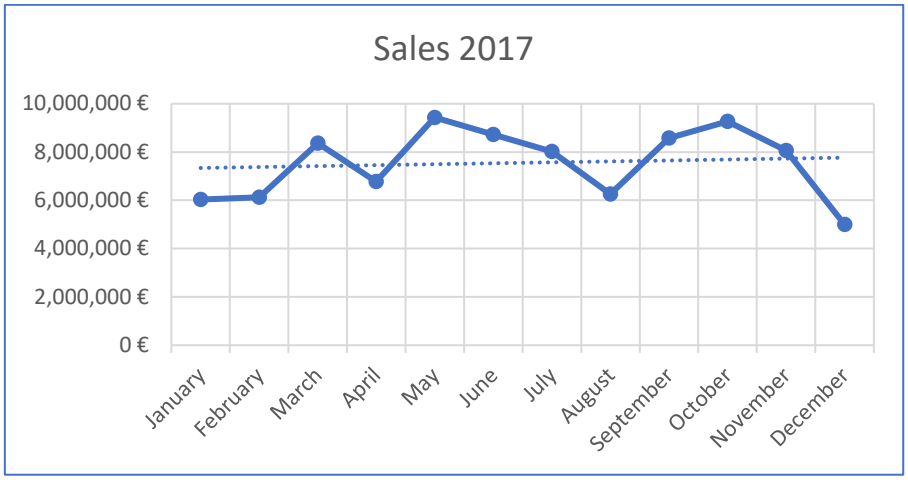
*Γράφημα 1*


Orders amount (euros)
Product 49 0.00%
Product 46 0.00%
Product 47 0.00%
Product 48 0.01%
Product 45 0.01%
Product 43 0.02%
Product 41 0.05%
Product 39 0.07%
Product 40 0.07%
Product 42 0.09%
Product 35 0.13%
Product 36 0.16%
Product 27 0.17%
Product 33 0.17%
Product 37 0.20%
Product 30 0.20%
Product 34 0.22%
Product 26 0.23%
Product 38 0.24%
Product 20 0.26%
Product 31 0.31%
Product 17 0.37%
Product 28 0.41%
Product 32 0.44%
Product 15 0.53%
Product 24 0.62%
Product 44 0.62%
Product 22 0.64%
Product 19 0.67%
Product 14 0.74%
Product 21 0.90%
Product 9 1.17%
Product 29 1.32%
Product 16 1.36%
Product 18 1.38%
Product 25 1.58%
Product 23 2.13%
Product 12 2.68%
Product 6 2.79%
Product 13 2.82%
Product 11 2.90%
Product 7 3.61%
Product 5 4.22%
Product 10 4.40%
Product 8 5.84%
Product 3 6.26%
Product 2 7.10%
Product 4 11.94%
Product 1 27.93%
0.00% 5.00% 10.00% 15.00% 20.00% 25.00% 30.00%

*Γράφημα 2*

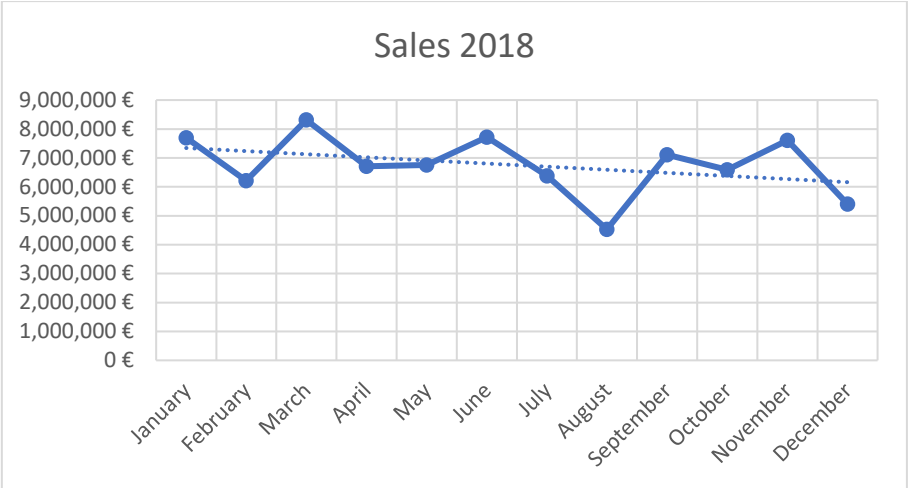
Παρακάτω μπορούμε να δούμε κάποια στοιχεία που αφορούν τις πωλήσεις για κάθε έτος που έχουμε εξετάσει για να δούμε την τάση και την εποχικότητα για κάθε χρονιά.

Ξεκινώντας από το 2017 βλέπουμε στο γράφημα 3 ότι η γραμμή της τάσης είναι ελαφρώς ανοδική και είναι έκδηλο ότι υπάρχει εποχικότητα με υψηλά τον Μάιο και τον Οκτώβριο και χαμηλά Αύγουστο και Δεκέμβριο.



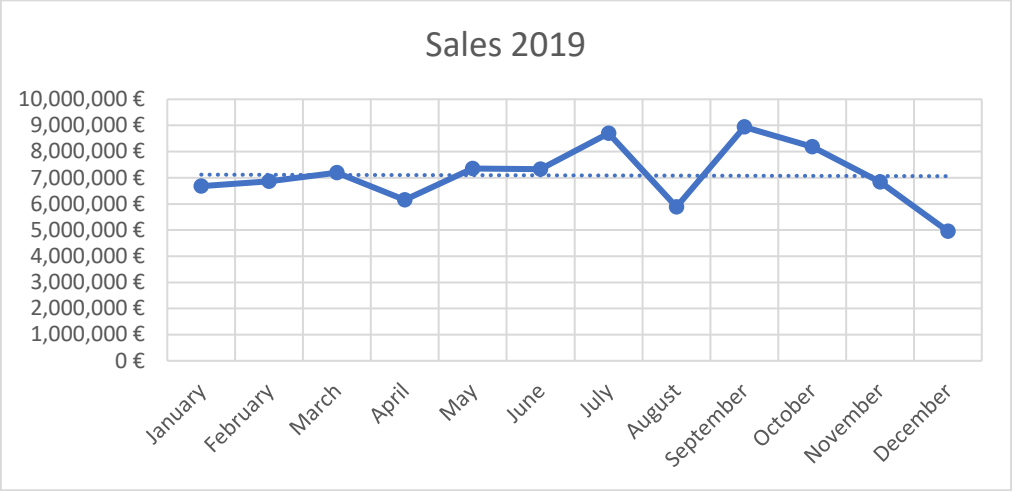

*Γράφημα 3*

Στο γράφημα 4 για το 2018 βλέπουμε ότι έχουμε υψηλά Μάρτιο, Ιούνιο και Νοέμβριο, ενώ χαμηλά Αύγουστο και Δεκέμβριο. Η τάση φαίνεται να αρχίζει να είναι καθοδική.



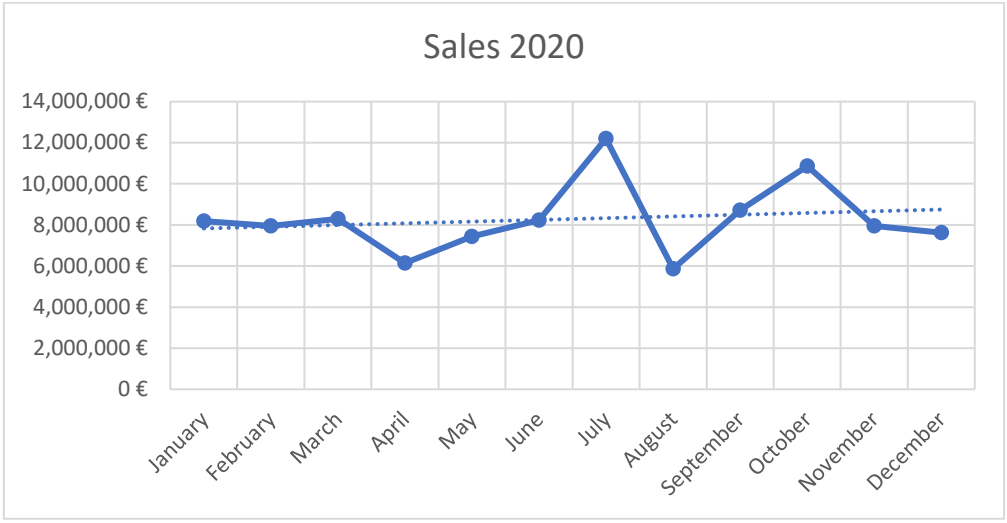

*Γράφημα 4*

Στο γράφημα 5 για το 2019 βλέπουμε ότι έχουμε υψηλά Ιούλιο και Σεπτέμβριο, ενώ χαμηλά Αύγουστο και Δεκέμβριο. Δεν φαίνεται να υπάρχει τάση στο διάγραμμα.



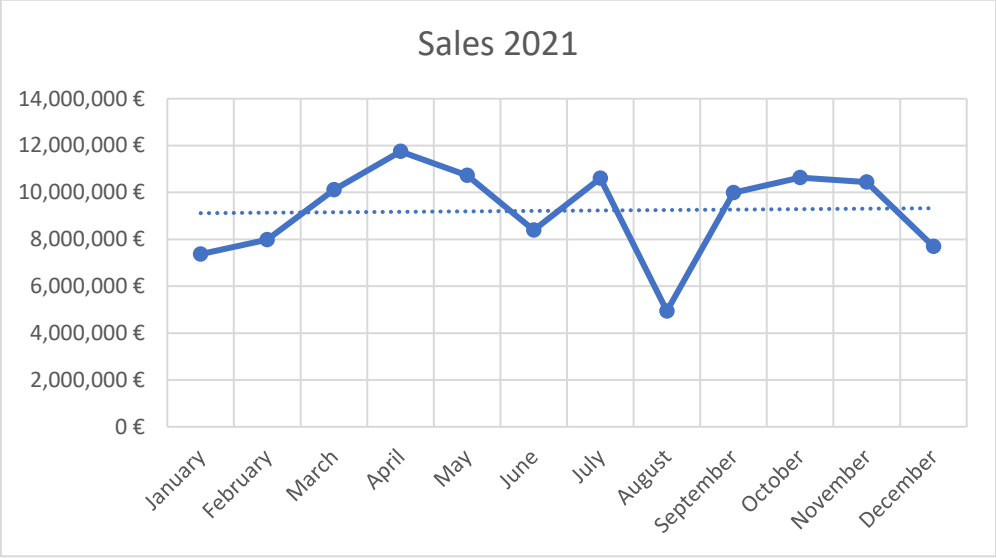

*Γράφημα 5*

Στο γράφημα 6 για το 2020 βλέπουμε ότι έχουμε υψηλά Ιούλιο και Οκτώβριο, ενώ χαμηλά Αύγουστο και Δεκέμβριο. Φαίνεται να υπάρχει ελαφρώς ανοδική τάση.



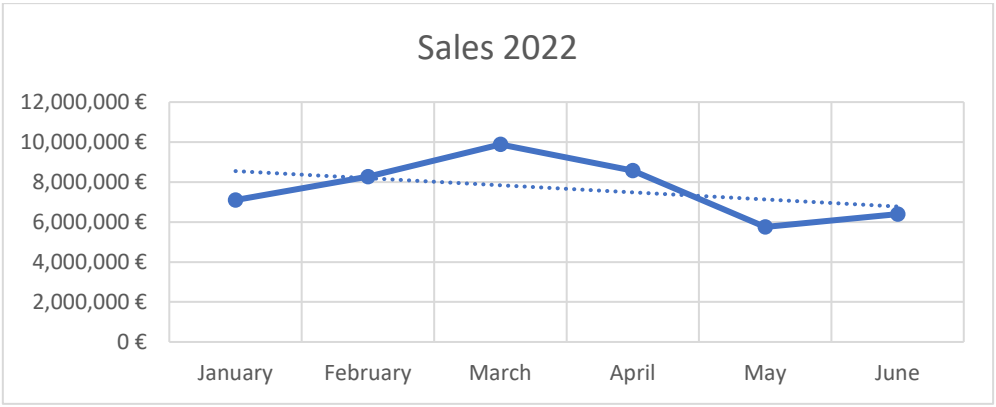

*Γράφημα 6*

Στο γράφημα 7 για το 2021 βλέπουμε ότι έχουμε υψηλά Απρίλιο, Ιούλιο και Οκτώβριο, ενώ χαμηλά Αύγουστο και Δεκέμβριο. Δεν φαίνεται να υπάρχει τάση στο διάγραμμα.




*Γράφημα 7*

Τέλος στο γράφημα 8 για το διάστημα του 2022 που έχουμε στοιχεία, βλέπουμε ότι υπάρχει μία κάμψη τον Μάιο ενώ η τάση φαίνεται να είναι ελαφρώς καθοδική.




*Γράφημα 8*

Συμπερασματικά μπορούμε να πούμε ότι σε ετήσια βάση κυριαρχεί εποχικότητα που είναι σταθερή όσο αφορά τον Αύγουστο και τον Δεκέμβριο, ενώ τα υψηλά είναι συνήθως ακριβώς πριν ή μετά τις περιόδους αυτές.

## 4.3 Εισαγωγή δεδομένων για επεξεργασία σε Python

Αρχικά εισάγουμε στο jupyter notebook που χρησιμοποιούμε, όλες τις βιβλιοθήκες που θα χρειαστούμε για να ακολουθήσουμε τη μεθοδολογία της ανάλυσης που έχουμε περιγράψει σε προηγούμενο κεφάλαιο ότι θα ακολουθήσουμε.

Οι μεταβλητές που έχουμε κρατήσει στα δεδομένα μας, είναι οι ακόλουθες:

**Order ID:** Απεικονίζεται ο μοναδικός κωδικός που δημιουργείται όταν καταχωρείται μία νέα παραγγελία στο σύστημα καταχώρησης παραγγελιών της εταιρείας. Σε κάθε αριθμό παραγγελίας υπάρχει πάντα ένα μόνο προϊόν.

**Customer ID:** Είναι ο αριθμός που αντιστοιχεί στον κωδικό του πελάτη στο σύστημα της εταιρείας. Κάθε πελάτης έχει έναν μοναδικό αριθμό που στην ουσία αποτελεί την ταυτότητα του πελάτη για την εταιρεία.

**Customer name:** Είναι η επωνυμία που έχει ο πελάτης. Υπάρχει περίπτωση κάποιος πελάτης να έχει περισσότερες από μία εταιρείες (θυγατρικές) αλλά σε αυτή την περίπτωση συνήθως υπάρχει κάποια (έστω μικρή διαφοροποίηση) στην επωνυμία. Σε αυτές τις περιπτώσεις κάθε θυγατρική έχει διαφορετικό Customer ID.

**Country:** Μας δείχνει την χώρα στην οποία είναι καταχωρημένος ο κάθε πελάτης.

**Product:** Είναι η ονομασία του τύπου προϊόντος της κάθε παραγγελίας.

**Transaction date:** Είναι η ημερομηνία κατά την οποία έχει γίνει η συναλλαγή. Πιο συγκεκριμένα είναι η ημερομηνία που έχει ενεργοποιηθεί η παραγγελία στο σύστημα της εταιρείας. Δεν είναι η ημερομηνία αποστολής.

**Price:** Είναι η τιμή της παραγγελίας που έχει συμφωνηθεί με τον πελάτη. Σε περίπτωση που ο πελάτης έχει κάποια ειδική έκπτωση, αυτή έχει ήδη υπολογιστεί σε αυτό το πεδίο πράγμα που σημαίνει ότι αντιστοιχεί στο χρηματικό αντίτιμο που έχει πληρώσει ο κάθε πελάτης για την κάθε παραγγελία που έχει καταχωρηθεί στο σύστημα.

**Area code:** Κάθε γεωγραφική περιοχή έχει μία κωδικοποίηση για εσωτερικούς λόγους της εταιρείας. Αυτό γίνεται για λόγους εσωτερικής οργάνωσης ώστε να ομαδοποιούνται πελάτες ανά γεωγραφική ζώνη και να είναι πιο εύκολες οι εσωτερικές διεργασίες της εταιρείας που αφορούν ως επί των πλείστων τα logistics & το marketing.

**Year:** Είναι η χρονιά που αντιστοιχεί στην ημερομηνία που έγινε η συναλλαγή. Για την ακρίβεια είναι μία μεταβλητή που έχουμε δημιουργήσει εμείς για να είναι πιο εύκολη η ανάλυση που κάναμε παραπάνω εξετάζοντας τις πωλήσεις ανά έτος.

**Month:** Είναι ο μήνας που αντιστοιχεί στην ημερομηνία που έγινε η συναλλαγή. Είναι επίσης μία μεταβλητή που έχουμε δημιουργήσει εμείς για να είναι πιο εύκολη η ανάλυση που κάναμε παραπάνω, ομοίως δηλαδή με την μεταβλητή του έτους που αναφέραμε.

Αρχικά μέσω της αναφοράς που τραβήξαμε, στα δεδομένα υπήρχαν ακόμη 8 μεταβλητές οι οποίες όμως δεν χρησίμευαν για την επιμέρους ανάλυση. Συνολικά κρατήσαμε τις 10 μεταβλητές που είδαμε παραπάνω στα δεδομένα μας, που αποτελούν και τα χαρακτηριστικά που έχουμε επιλέξει ότι θα μας χρειαστούν για την ανάλυση που θα κάνουμε. Με αυτό τον τρόπο, όπως είδαμε και στην βιβλιογραφική επισκόπηση παραπάνω, καταφέρνουμε να κάνουμε τον αλγόριθμό μας πιο αποδοτικό αφού μειώνουμε αισθητά τα δεδομένα που έχει να επεξεργαστεί άρα και την υπολογιστική ισχύ, τη μνήμη και τέλος το χρόνο που χρειάζεται για να τρέξει.

Το πρώτο βήμα όσο αφορά τα δεδομένα μας είναι να γίνει ένας αρχικός έλεγχος ώστε να δούμε αν υπάρχουν αρνητικές τιμές, διπλές εγγραφές, ακραίες τιμές που είναι λάθη και γενικά αναντιστοιχίες που αποτελούν «θόρυβο» και που αν κρατήσουμε θα επιφέρουν αλλοίωση στα αποτελέσματα της ανάλυσης που θα κάνουμε. Σε αυτό το στάδιο παρατηρήσαμε ότι είχαμε εγγραφές που αφορούν εκθεσιακά προϊόντα, με μηδενική τιμή οπότε και τα αφαιρέσαμε από το σύνολο δεδομένων μας. Επίσης βλέπουμε ότι έχουμε κάποιες συμπληρωματικές παραγγελίες για πελάτες οι οποίες έγιναν για να καλύψουν κάποια ποιοτικά προβλήματα που είχαν εντοπιστεί στις αρχικές παραγγελίες τους και στάλθηκαν με μηδενική χρέωση προς αντικατάσταση στο πλαίσιο της εγγύησης που παρέχει η εταιρεία, οπότε αφαιρέθηκαν από το σύνολο δεδομένων που κρατήσαμε για να εξετάσουμε.

Επόμενο βήμα είναι να ελέγξουμε τα δεδομένα μας στο jupyter notebook που δημιουργήσαμε για την ανάλυση. Βλέπουμε ότι συνολικά έχουμε κρατήσει 32.629

εγγραφές που αντιστοιχούν σε ισάριθμες παραγγελίες που έχουν πραγματοποιηθεί στο εξεταζόμενο χρονικό πλαίσιο.

Η μέση τιμή για κάθε παραγγελία είναι 15.033 €. Η μικρότερη τιμή παραγγελίας που έχουμε είναι 1.000 € και η μεγαλύτερη τιμή παραγγελίας είναι 524.670 €. Χωρίζοντας τα δεδομένα σε τέσσερα τεταρτημόρια βλέπουμε ότι στο πρώτο τεταρτημόριο έχουμε τιμές έως 4.810 €. Στο δεύτερο τεταρτημόριο έχουμε τιμές μέχρι 13.000 €. Στο τρίτο τεταρτημόριο έχουμε τιμές έως 19.200 €. Τέλος στο τελευταίο τεταρτημόριο έχουμε τιμές έως 524.670 €. Με αυτό τον τρόπο μπορούμε να έχουμε μία πρώτη εικόνα για την κατάτμηση του συνόλου δεδομένων μας με βάση την τιμή πώλησης.

# 5 Αποτελέσματα

Σε αυτό το κεφάλαιο θα δούμε την κατηγοριοποίηση των πελατών που προκύπτει αρχικά μέσω της εφαρμογής της RFM ανάλυσης στο σύνολο δεδομένων μας με βάση τα τρία χαρακτηριστικά Recency, Frequency, Monetary. Έπειτα θα δημιουργήσουμε ομάδες (clusters) χρησιμοποιώντας τον αλγόριθμο k – means, λαμβάνοντας ταυτόχρονα υπόψη μας τα αποτελέσματα που έχουμε από την RFM ανάλυση που έχει προηγηθεί.

## 5.1 Υπολογισμός Recency, Frequency, Monetary

Το επόμενο βήμα στην ανάλυση μας είναι να κάνουμε RFM ανάλυση. Για να γίνει αυτό θα πρέπει να δημιουργήσουμε τρεις νέες μεταβλητές. Η πρώτη μεταβλητή είναι η Recency που υποδηλώνει το χρόνο που έχει μεσολαβήσει από το χρονικό σημείο που έχει πραγματοποιηθεί η τελευταία συναλλαγή του πελάτη με την εταιρεία και τη μετράμε σε πλήθος ημερών. Η δεύτερη μεταβλητή είναι η Frequency που υποδηλώνει τη συχνότητα των συναλλαγών που έχουν πραγματοποιηθεί στη χρονική διάρκεια που εξετάζουμε. Τρίτη μεταβλητή είναι η Monetary που υποδηλώνει το συνολικό ύψος των συναλλαγών που έχει πραγματοποιήσει ο κάθε πελάτης μέσα στο χρονικό πλαίσιο που έχει τεθεί προς εξέταση.

Το πρώτο βήμα στη διαδικασία υπολογισμού της μεταβλητής Recency είναι να βρούμε την ημερομηνία της τελευταίας συναλλαγής στο σύνολο δεδομένων μας. Μετά

δημιουργούμε ένα pinned date αντικείμενο - μεταβλητή το οποίο θα πάρει ως τιμή την επόμενη ημέρα από την τελευταία συναλλαγή. Αυτή η pinned date θα είναι και η βάση για να υπολογίσουμε το Recency της κάθε εγγραφής αφαιρώντας την ημερομηνία κάθε συναλλαγής από αυτή. Βάσει αυτής της λογικής δημιουργούμε μία νέα στήλη που ονομάζεται Recency και υπολογίζει για κάθε εγγραφή πόσες ημέρες έχουν παρέλθει από την ημερομηνία που ορίσαμε ως pinned date. Το αποτέλεσμα είναι να πάρουμε έναν πίνακα ο οποίος θα έχει την παρακάτω μορφή (Εικόνα 16).

| | Customer_ID | Last_Purchase_Date | Recency |
|---|---|---|---|
| **0** | 100040 | 2022-05-17 | 44 |
| **1** | 100061 | 2022-01-17 | 164 |
| **2** | 100085 | 2021-06-25 | 370 |
| **3** | 100086 | 2022-04-19 | 72 |
| **4** | 100088 | 2020-11-20 | 587 |

*Εικόνα 16*

Συνεχίζουμε υπολογίζοντας τη μεταβλητή Frequency που είναι το πλήθος των αγορών που έχει κάνει κάθε πελάτης (δηλαδή η συχνότητα αγορών), και τη Μεταβλητή Monetary που είναι το άθροισμα των συναλλαγών ανά πελάτη στον κώδικά μας. Το αποτέλεσμα είναι να πάρουμε έναν πίνακα που έχει την παρακάτω μορφή (Εικόνα 17).

| | Frequency | Monetary |
|---|---|---|
| **Customer_ID** | | |
| **100040** | 216 | 757495.00 |
| **100061** | 3 | 6739.50 |
| **100085** | 11 | 75691.43 |
| **100086** | 12 | 113221.75 |
| **100088** | 2 | 11670.34 |

*Εικόνα 17*

Τέλος κάνουμε συγχώνευση των υπολογιζόμενων πινάκων, κρατώντας μόνο τις μεταβλητές Recency, Frequency, Monetary σε έναν ενιαίο πίνακα RFM ο οποίος έχει την παρακάτω μορφή (Εικόνα 18).

| | Customer_ID | Recency | Frequency | Monetary |
|---|---|---|---|---|
| **0** | 100040 | 44 | 216 | 757495.00 |
| **1** | 100061 | 164 | 3 | 6739.50 |
| **2** | 100085 | 370 | 11 | 75691.43 |
| **3** | 100086 | 72 | 12 | 113221.75 |
| **4** | 100088 | 587 | 2 | 11670.34 |

*Εικόνα 18*

Σε αυτό το σημείο θα εισάγουμε ακόμη τρεις νέες μεταβλητές ορίζοντας μία κοινή βαθμονόμηση με κωδικοποίηση από 1 έως 4 με τον αριθμό 1 να υποδηλώνει για κάθε μία τη μικρότερη βαθμολογία και με τον αριθμό 4 να υποδηλώνει την υψηλότερη βαθμολογία. Πιο συγκεκριμένα για τη Recency θα εισάγουμε την R, για την Frequency θα εισάγουμε την F και για την Monetary θα εισάγουμε την M. Η βαθμονόμηση για τα R, F, M από 1 έως 4 επιλέχθηκε ώστε να ακολουθήσουμε την λογική ότι ομαδοποιούμε τα σημεία δεδομένων που εξετάζουμε σε τέσσερα τεταρτημόρια. Επίσης εισάγουμε τη μεταβλητή RFM Segment που είναι στην ουσία μία συγχώνευση των τιμών R, F, M, συγχωνεύοντας τα για να δώσουμε μία συνολική εικόνα για κάθε εγγραφή που θα απεικονίζει ταυτόχρονα ποιο είναι το score κάθε ενός από τα R, F, M ταυτόχρονα. Με μία παρόμοια λογική εισάγουμε μία ακόμα μεταβλητή, το RFM Score το οποίο είναι το άθροισμα των τιμών R, F, M. Με βάση όσα έχουμε αναφέρει μέχρι στιγμής ο πίνακας RFM που έχουμε δημιουργήσει θα έχει την παρακάτω μορφή (Εικόνα 19).

| | Customer_ID | Recency | Frequency | Monetary | R | F | M | RFM_Segment | RFM_Score |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 100040 | 44 | 216 | 757495.00 | 4 | 4 | 4 | 444 | 12.0 |
| **1** | 100061 | 164 | 3 | 6739.50 | 4 | 3 | 1 | 431 | 8.0 |
| **2** | 100085 | 370 | 11 | 75691.43 | 3 | 4 | 3 | 343 | 10.0 |
| **3** | 100086 | 72 | 12 | 113221.75 | 4 | 4 | 3 | 443 | 11.0 |
| **4** | 100088 | 587 | 2 | 11670.34 | 2 | 2 | 1 | 221 | 5.0 |

*Εικόνα 19*

**Υπολογισμός κατηγοριών πελατών βάσει ειδικών χαρακτηριστικών.**

Αφού πλέον έχουμε εισάγει τις βασικές μεταβλητές για την RFM ανάλυση, πλέον μπορούμε να ξεκινήσουμε βρίσκοντας κάποιες χρήσιμες κατηγορίες πελατών που έχουν κάποια ειδικά χαρακτηριστικά και ενδιαφέρουν την Εμπορική διεύθυνση της εταιρείας. Αυτές είναι:

Οι πελάτες που έχουν τη μέγιστη βαθμολογία και στα τρία χαρακτηριστικά και είναι αυτοί που έχουν RFM Score = 444 και χαρακτηρίζονται ως οι "Diamond" πελάτες. Βλέπουμε ότι είναι συνολικά 276 και είναι αυτοί οι οποίοι έχουν κάνει τελευταία συναλλαγή πρόσφατα, έχουν μεγάλη συχνότητα συναλλαγών κατά τη διάρκεια του χρονικού ορίζοντα που εξετάζουμε και έχουν δαπανήσει μεγάλα ποσά στις αγορές τους. Είναι πελάτες που η εταιρεία πρέπει να κρατήσει οπωσδήποτε στην πελατειακή της βάση.

Επόμενη ειδική κατηγορία είναι αυτοί οι οποίοι παρουσιάζουν την υψηλότερη συχνότητα αγορών με F = 4 και βλέπουμε ότι συνολικά είναι 561 και χαρακτηρίζονται ως "Loyal". Προφανώς η "Diamond" κατηγορία είναι υποσύνολο της κατηγορίας αυτής.

Τέλος έχουμε την κατηγορία που έχουν RFM Score = 111 που είναι οι "Lost customers" δηλαδή πελάτες που έχουν πάρα πολύ καιρό να κάνουν συναλλαγή, έχουν μικρή συχνότητα συναλλαγών και με μικρή αξία, οπότε η εταιρεία δεν αξίζει να ασχοληθεί για επαναπροσέγγιση αυτών των πελατών ενεργά, αφού δεν είναι πελάτες αξίας. Συνολικά έχουμε 149 πελάτες σε αυτή την κατηγορία.

Στην συνέχεια κάνουμε exploratory analysis στα δεδομένα μας όπως τα έχουμε μετασχηματίσει και υπολογίζουμε τους μέσους και το πλήθος ανά RFM_Score.

Από τα παρακάτω στοιχεία (Εικόνα 20) βλέπουμε για κάθε τιμή του RFM Score ποιος είναι ο μέσος όρος των ημερών από την τελευταία συναλλαγή (Recency Mean), ποια

είναι η μέση συχνότητα συναλλαγών (Frequency Mean) και ποια είναι η μέση αξία συναλλαγών (Monetary Mean) για κάθε κατηγορία. Επίσης βλέπουμε πόσες εγγραφές έχουμε για κάθε κατηγορία RFM Score. Αυτό μας δίνει μία καλή εικόνα για την κατανομή των πελατών για κάθε ένα RFM Score που έχει προκύψει από την προσέγγιση μας.

| | Recency | Frequency | Monetary | |
|---|---|---|---|---|
| | mean | mean | mean | count |
| RFM_Score | | | | |
| **3.0** | 1588.3 | 1.0 | 5419.8 | 149 |
| **4.0** | 1169.8 | 1.1 | 11674.6 | 256 |
| **5.0** | 882.3 | 1.3 | 16805.7 | 223 |
| **6.0** | 695.8 | 1.6 | 23034.4 | 271 |
| **7.0** | 629.4 | 2.2 | 34930.5 | 282 |
| **8.0** | 547.8 | 3.2 | 57482.9 | 226 |
| **9.0** | 555.0 | 7.2 | 160403.4 | 233 |
| **10.0** | 349.9 | 10.6 | 194180.4 | 178 |
| **11.0** | 225.5 | 32.1 | 410164.5 | 149 |
| **12.0** | 80.2 | 82.4 | 1216940.5 | 276 |

*Εικόνα 20*

Συνεχίζοντας την ανάλυση μας θα επιχειρήσουμε την κατηγοριοποίηση των πελατών συνολικά, λαμβάνοντας υπόψιν τα μεγέθη που έχουν προκύψει μέσω των R, F, M μεγεθών, και όχι απλά με κάποια ειδικά χαρακτηριστικά όπως κάναμε παραπάνω. Στην λογική της συνολικής κατηγοριοποίησης των πελατών θα τους ομαδοποιήσουμε ξανά σε τέσσερις επιμέρους κατηγορίες βάσει RFM score. Θέτουμε ως:

***Best Customers:*** Πελάτες οι οποίοι έχουν RFM Score από 10 και πάνω. Είναι οι πελάτες αξίας που έχουν υψηλές τιμές στα R, F, M. Επίσης μέσω της μεταβλητής RFM level που εισάγουμε κωδικοποιούμε αυτή την κατηγορία με RFM level = 3 που είναι το μέγιστο.

***Promising Customers:*** Πελάτες οι οποίοι έχουν RFM Score από 7 έως 9. Είναι πελάτες οι οποίοι έχουν κάποια δυναμική και θα μπορούσε η εταιρεία να εστιάσει σε αυτούς υιοθετώντας είτε στρατηγικές upsell ώστε να αγοράσουν πιο ακριβά προϊόντα και να ανέβουν κατηγορία είτε ελέγχοντας αν οι συγκεκριμένοι μπορεί να αγοράζουν και από

ανταγωνιστές. Στην περίπτωση που δεν είναι αποκλειστικοί πελάτες θα μπορούσε να υιοθετήσει στρατηγική που να εστιάζει σε πρόσθετα κίνητρα που θα ενίσχυε την πιστότητα τους δίνοντας τους κίνητρα για έξτρα εκπτώσεις στην περίπτωση που πιάσουν κάποιους στόχους αγορών, οι οποίοι θα είναι υψηλότεροι από αυτούς που έχουν μέχρι στιγμής. Επίσης μέσω της μεταβλητής RFM level που εισάγουμε, κωδικοποιούμε αυτή την κατηγορία με RFM level = 2.

***Need attention:*** Πελάτες οι οποίοι έχουν RFM Score από 4 έως 6. Είναι πελάτες οι οποίοι δεν έχουν ιδιαίτερα υψηλή βαθμολογία και θα πρέπει η εταιρεία να εξετάσει αν πρόκειται για πελάτες χαμηλής δυναμικής, ή αν πρόκειται για πελάτες που έχουν αρχίσει να απομακρύνονται από την εταιρεία. Στην περίπτωση που ήταν παλιότερα πελάτες υψηλής αξίας και έχουν αρχίσει να απομακρύνονται θα πρέπει να υιοθετήσουμε κάποια καμπάνια επαναπροσέγγισης ώστε να δούμε αν έχει φταίξει κάποιος συγκεκριμένος λόγος που η συνεργασία μας μαζί τους έχει αρχίσει να αποκλιμακώνεται και να εξετάσουμε αν θα μπορούσαμε να το αναστρέψουμε. Επίσης μέσω της μεταβλητής RFM level που εισάγουμε κωδικοποιούμε αυτή την κατηγορία με RFM level = 1.

***Lost Customers:*** Πελάτες που έχουν RFM Score λιγότερο από 4. Αυτό σημαίνει ότι στην ουσία αυτή η κατηγορία έχει πελάτες που έχουν χαθεί και ανήκουν αποκλειστικά στην κατηγορία RFM Segment 111. Είναι πελάτες οι οποίοι δεν έχουν κάνει κάποια συναλλαγή εδώ και πάρα πολύ καιρό, η συχνότητα τους είναι πάρα πολύ χαμηλή και ο τζίρος τους είναι επίσης πολύ χαμηλός. Για αυτή την κατηγορία δεν αξίζει να δαπανήσουμε πολλά χρήματα για την επαναπροσέγγιση τους. Επίσης μέσω της μεταβλητής RFM level που εισάγουμε κωδικοποιούμε αυτή την κατηγορία με RFM level = 0.

Συνοψίζοντας την κατηγοριοποίηση που μόλις αναφέραμε μπορούμε να δούμε μία πρώτη απεικόνιση του προφίλ για κάθε κατηγορία στην εικόνα 21. Παρατηρούμε ότι οι Best Customers είναι συνολικά 603, έχουν μέσο όρο τελευταία συναλλαγή 196 ημέρες, κατά μέσο όρο πραγματοποίησαν 49 συναλλαγές στο χρονικό ορίζοντα που εξετάσαμε, και ο μέσος όρος συναλλαγών τους ήταν 716.000 €.

Οι πελάτες οι οποίοι είναι στην κατηγορία Promising, είναι συνολικά 741, έχουν μέσο όρο τελευταία συναλλαγή 581 ημέρες, με μέσο όρο 4 συναλλαγές και κατά μέσο όρο έχουν δαπανήσει περίπου 81.263 €.

Οι πελάτες που είναι στην κατηγορία Need attention, είναι συνολικά 750, έχουν μέσο όρο τελευταία συναλλαγή 913 ημέρες, με μέσο όρο 1,4 συναλλαγές και κατά μέσο όρο έχουν δαπανήσει περίπου 17.305 €.

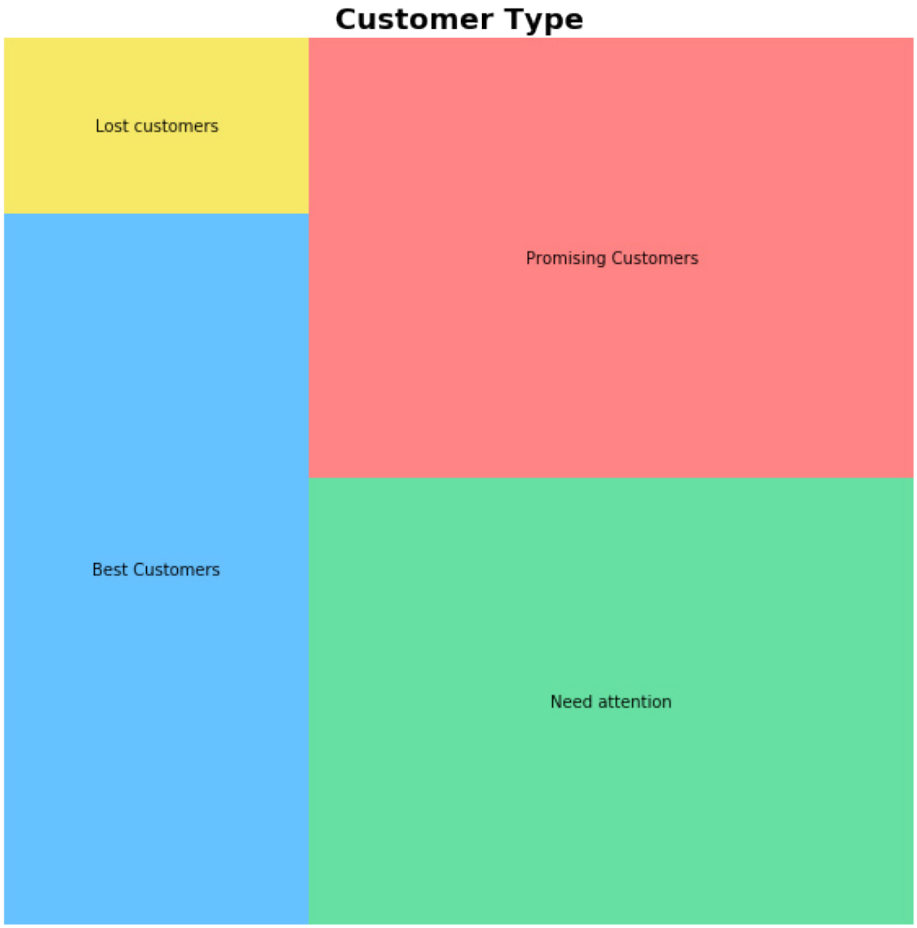
Τέλος, οι πελάτες που είναι στην κατηγορία Lost customers, είναι συνολικά 149, έχουν μέσο όρο τελευταία συναλλαγή 1.588 ημέρες, με μέσο όρο 1 συναλλαγή και κατά μέσο όρο έχουν δαπανήσει περίπου 5.420 €.

| | Recency | Frequency | Monetary | |
|---|---|---|---|---|
| | mean | mean | mean | count |
| Customer_type | | | | |
| **Best Customers** | 195.7 | 48.8 | 715678.6 | 603 |
| **Lost Customers** | 1588.3 | 1.0 | 5419.8 | 149 |
| **Need attention** | 913.1 | 1.4 | 17304.9 | 750 |
| **Promising Customers** | 581.1 | 4.1 | 81262.5 | 741 |

*Εικόνα 21*

Παρακάτω στο γράφημα 9 μπορούμε να δούμε ένα διάγραμμα bar plot που απεικονίζουμε την κατανομή των πελατών για κάθε κατηγορία από τις προαναφερόμενες. Επίσης μία παρόμοια απεικόνιση έχουμε στο γράφημα 10 όπου βλέπουμε τα μεγέθη των κατηγοριών των πελατών που προκύπτουν με βάση τη λογική που έχουμε εισάγει.


*Γράφημα 9*



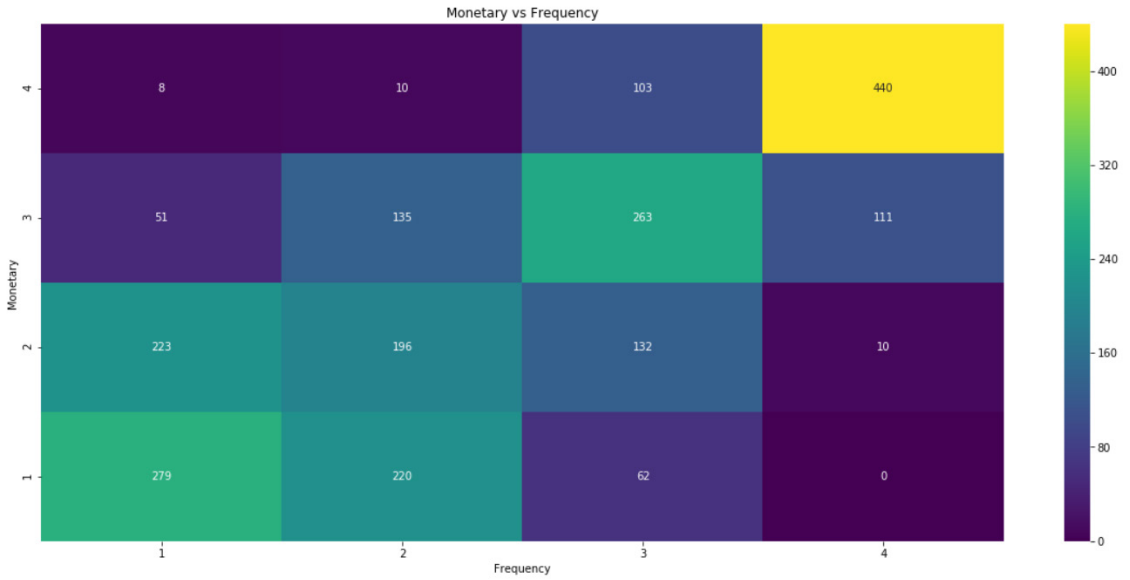

*Γράφημα 10*

Επόμενο βήμα στην ανάλυση μας είναι να δημιουργήσουμε κάποιες οπτικοποιήσεις που θα μας βοηθήσουν να συγκρίνουμε ανά δύο τα μεγέθη R, F, M για να δούμε τον αριθμό των παρατηρήσεων με heatmaps σε κάθε σύγκριση που θα κάνουμε.

**Monetary vs Frequency**

Στο γράφημα 11 συγκρίνοντας το Monetary με το Frequency, βλέπουμε ότι η πλειοψηφία των παρατηρήσεων που αντικατοπτρίζουν πελάτες συγκρίνοντας τους δύο αυτούς παράγοντες συγκεντρώνονται σε παρατηρήσεις που έχουν M = 4 και F = 4 πράγμα το οποίο είναι λογικό αφού πελάτες με υψηλή συχνότητα αγορών είναι λογικό να είναι και πελάτες που ξοδεύουν αρκετά χρήματα συνολικά στις αγορές τους. Επίσης είναι λογικό είναι να μην υπάρχουν πελάτες με υψηλή συχνότητα αγορών F = 4 που ταυτόχρονα παρουσιάζουν M = 0. Επίσης με βάση την ίδια λογική παρατηρούμε ότι η πλειοψηφία

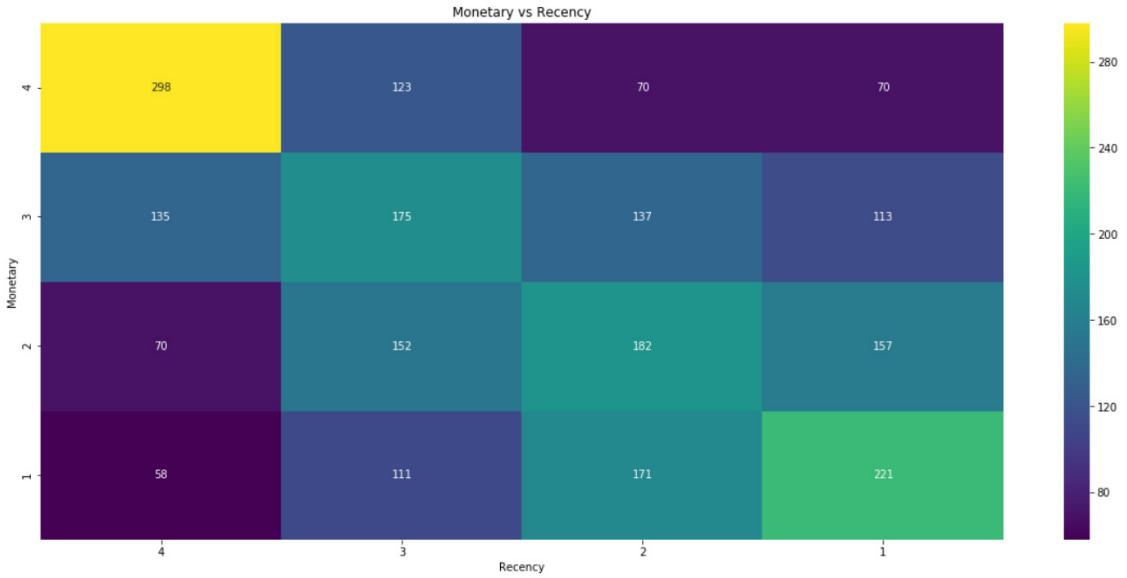
των πελατών συγκρίνοντας αυτές τις δύο μεταβλητές συγκεντρώνονται στη διαγώνια γραμμή από πάνω δεξιά προς τα κάτω αριστερά αφού είναι λογικό ότι αυτές οι δύο μεταβλητές είναι αλληλοεξαρτώμενες και η τιμή της μίας επηρεάζει τις περισσότερες φορές την τιμή της άλλης.




*Γράφημα 11*

## Monetary vs Recency

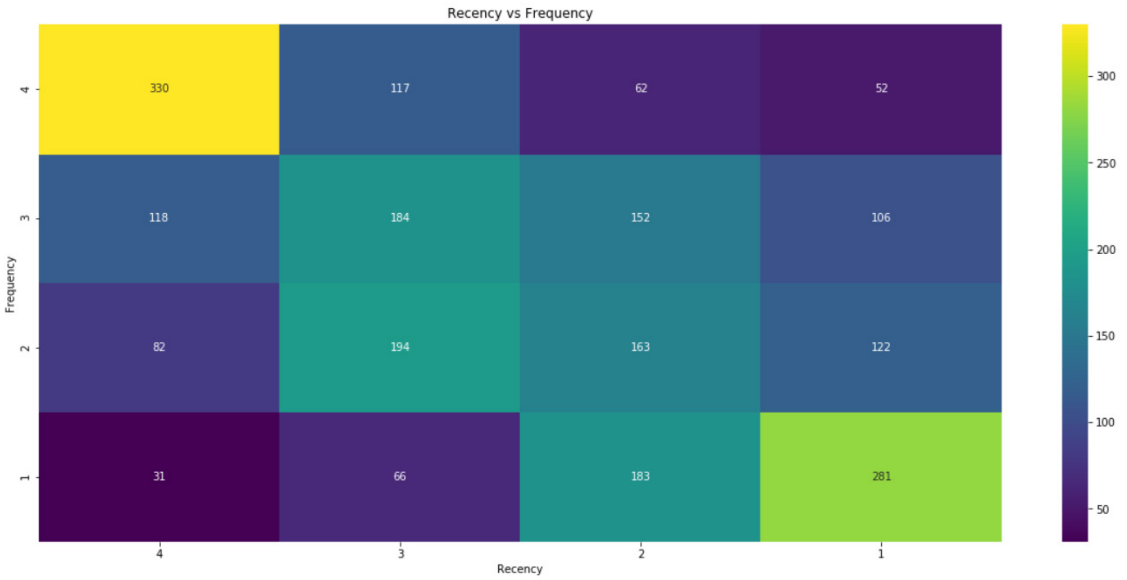
Στο γράφημα 12 βλέπουμε τη σύγκριση του Monetary με Recency όσο αφορά την κατανομή των πελατών σε ένα heatmap. Η πλειοψηφία βλέπουμε ότι ανήκει στην κατηγορία που έχουν M = 4 και R = 4 πράγμα το οποίο είναι λογικό αφού πελάτες με υψηλό τζίρο είναι και πελάτες που έχουν κάνει πρόσφατα συναλλαγές, ανήκουν δηλαδή στους πελάτες αξίας. Επίσης βλέπουμε ότι δεν υπάρχουν πολλοί πελάτες με υψηλή αξία αγορών M = 4 που να έχουν πολύ καιρό να πραγματοποιήσουν συναλλαγή με την εταιρεία, δηλαδή να έχουν χαμηλό δείκτη R.




*Γράφημα 12*

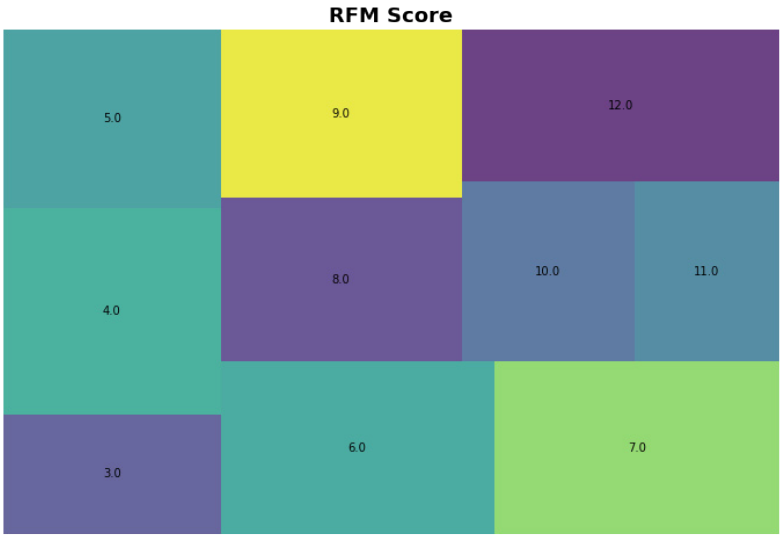
## Recency vs Frequency

Στο γράφημα 13 συγκρίνοντας τους πελάτες με βάση αυτούς τους δύο παράγοντες βλέπουμε ότι στην πλειοψηφία τους είναι πελάτες με R = 4 και F = 4. Εδώ δεν βλέπουμε κάποια μηδενική κατηγορία όπως στις παραπάνω περιπτώσεις.

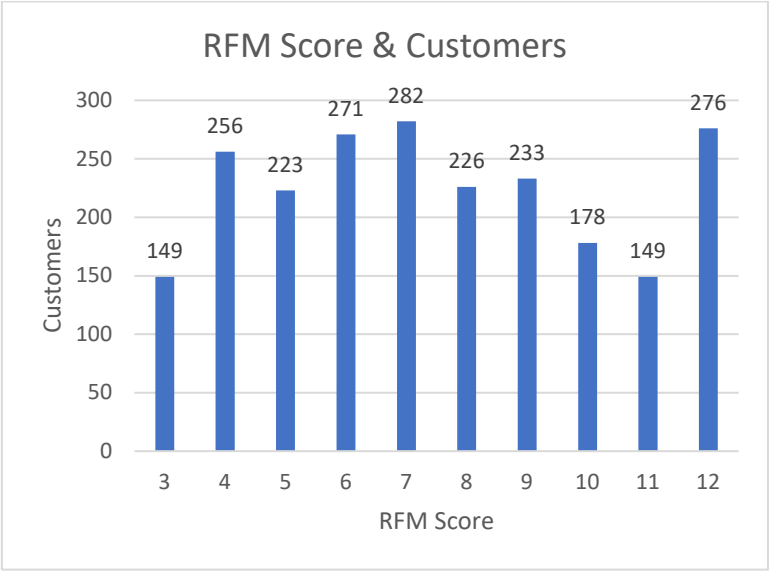



*Γράφημα 13*

Τέλος, στο γράφημα 14 (heatmap) και γράφημα 15 (bar-plot) μπορούμε να δούμε μία συνολική εικόνα της κατανομής των πελατών όσο αφορά το RFM Score. Αρχικά βλέπουμε ότι υπάρχει μία ομοιόμορφη κατανομή με την πλειοψηφία των πελατών να ανήκουν στο RFM Score = 7. Ακολουθεί το RFM Score = 12 και μετά RFM Score = 6. Η μεγάλη συσσώρευση πλήθους πελατών στην κατηγορία 12 δείχνει ότι η εταιρεία έχει μία καλή πελατειακή βάση με αρκετούς πελάτες υψηλής αξίας.




*Γράφημα 14*



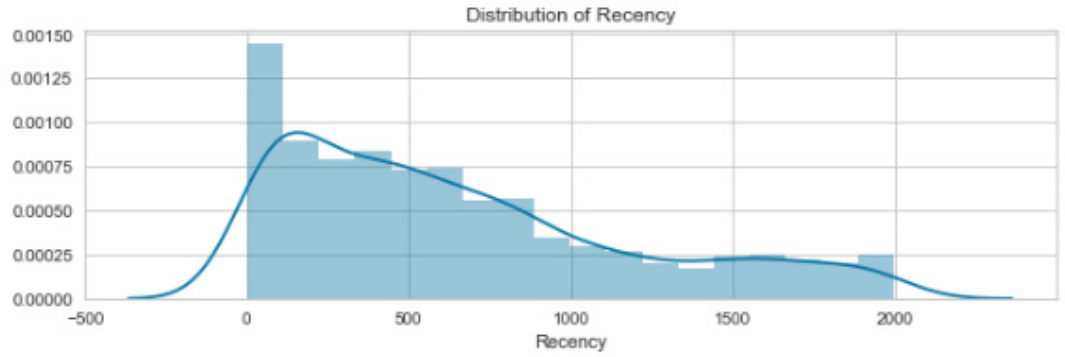

*Γράφημα 15*

## 5.2 Προ - επεξεργασία δεδομένων για Cluster Analysis

Το επόμενο βήμα είναι να εξετάσουμε την κανονικότητα των δεδομένων μας για τα Recency (Γράφημα 16), Frequency (Γράφημα 17), Monetary (Γράφημα 18) ώστε να δούμε αν μπορούμε να προχωρήσουμε σε clustering. Παρατηρούμε ότι υπάρχει λοξότητα αριστερά που είναι αρκετά μεγαλύτερη για το Frequency και το Monetary.

- **Distribution of Recency**

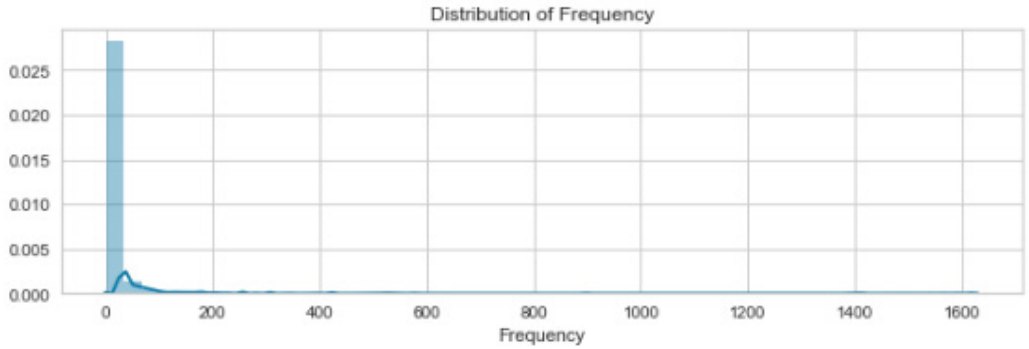
Recency's: Skew: 0.854
Skew test Result (statistic=14.491, pvalue=1.369007951721331e-47)




*Γράφημα 16*

- **Distribution of Frequency**

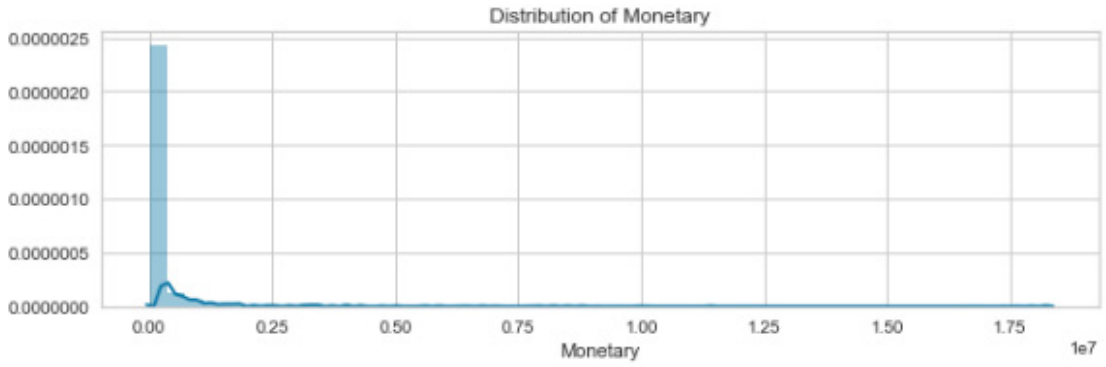
Frequency's: Skew: 14.131
Skew test Result (statistic=56.353, pvalue=0.0)




*Γράφημα 17*

- **Distribution of Monetary**

Monetary's: Skew: 10.018
Skew test Result (statistic=50.891, pvalue=0.0)



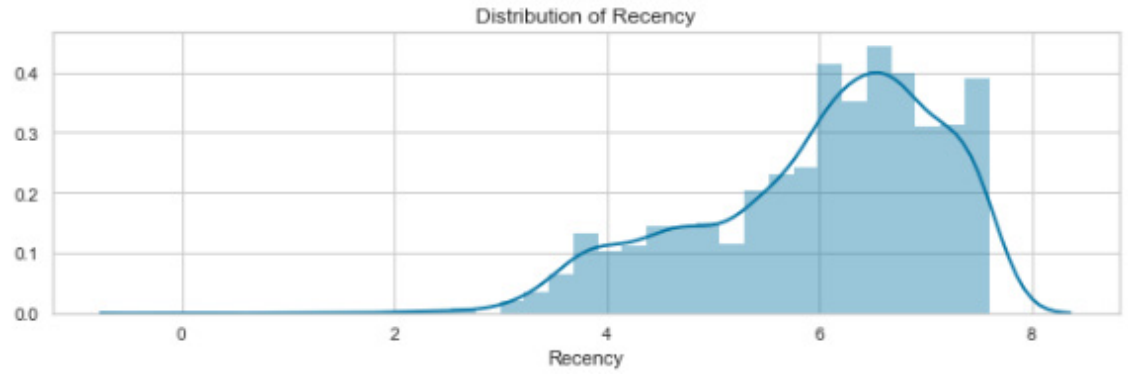

*Γράφημα 18*

**Κανονικοποίηση των δεδομένων**

Παρατηρούμε στα γραφήματα 16, 17, 18 ότι έχουμε λοξότητα για κάθε κατηγορία, οπότε επιλέγουμε να κανονικοποιήσουμε τα δεδομένα μας, κάνοντας λογαριθμοποίηση και μετά επαναλαμβάνοντας τον ίδιο έλεγχο για να δούμε πως έχουν μετασχηματιστεί τα δεδομένα μας. Συνοπτικά μπορούμε πάλι να δούμε παρακάτω τα στατιστικά αποτελέσματα.

- **Distribution of Recency – Normalized**

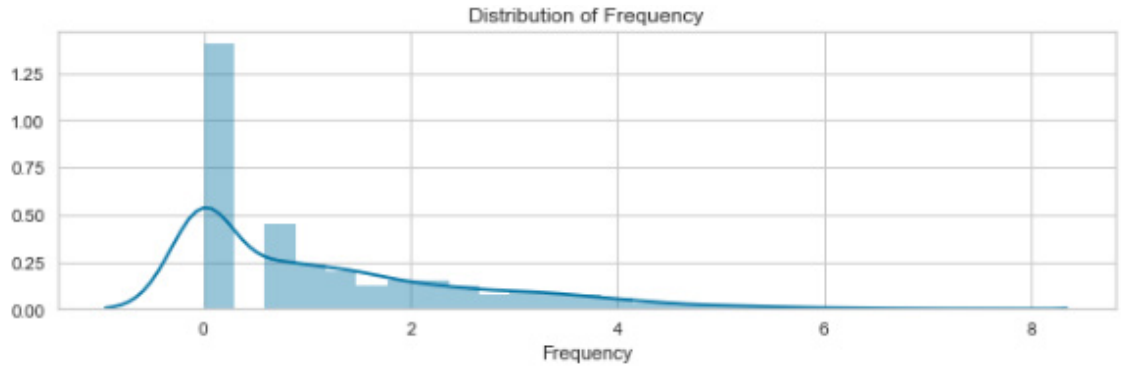
Recency's: Skew: -0.789
Skew test Result (statistic=-13.596, pvalue=4.199420344573936e-42)




*Γράφημα 19*

- **Distribution of Frequency – Normalized**

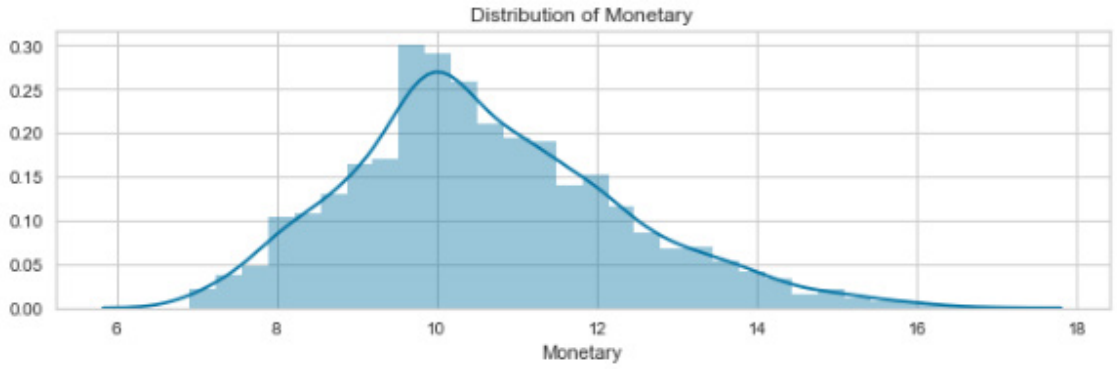
Frequency's: Skew: 1.234
Skew test Result (statistic=19.062, pvalue=5.135283259406427e-81)




*Γράφημα 20*

- **Distribution of Monetary – Normalized**

Monetary's: Skew: 0.504
Skew test Result (statistic=9.254, pvalue=2.157615922572438e-20)

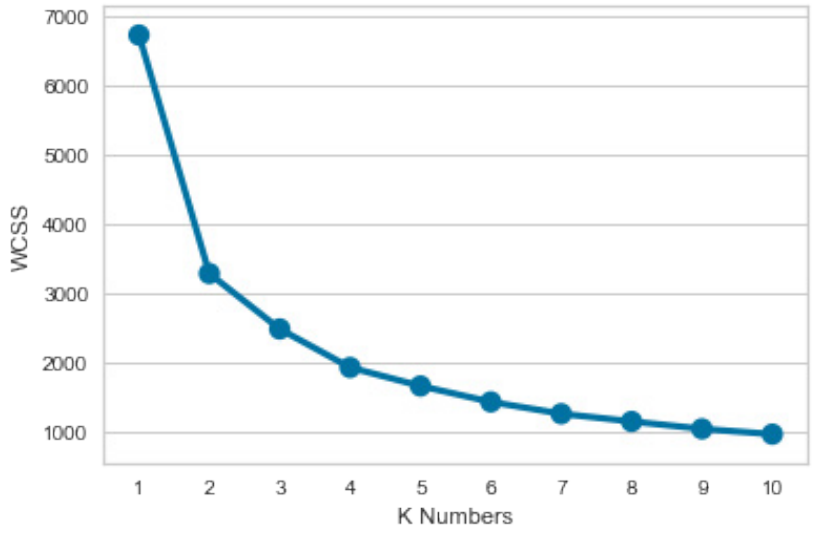



*Γράφημα 21*

Παρατηρούμε ότι το αποτέλεσμα είναι εμφανώς καλύτερο (Γράφημα 19, Γράφημα 20, Γράφημα 21) από την αρχική δομή των δεδομένων οπότε μπορούμε να προχωρήσουμε στο δεύτερο στάδιο της ανάλυσης μας εφαρμόζοντας k-means για να δημιουργήσουμε Clusters.

## 5.3 Clustering με αλγόριθμο K-Means

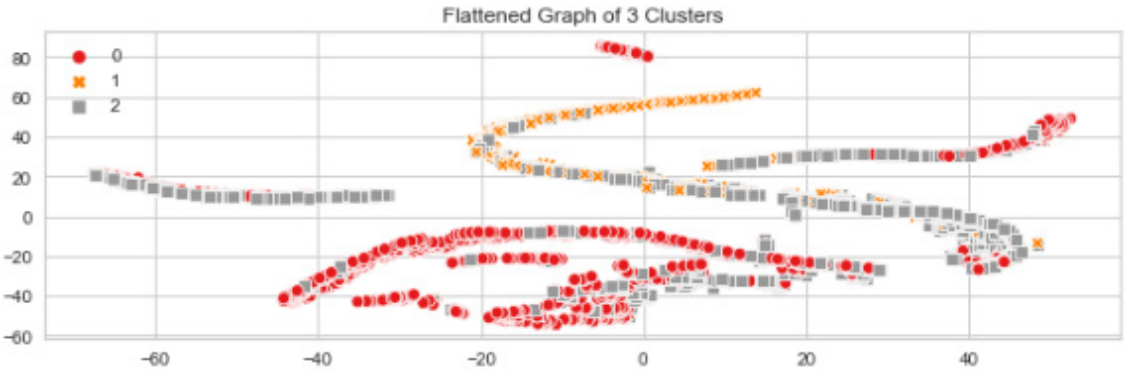
Το πρώτο βήμα είναι να δημιουργήσουμε ένα elbow graph ώστε να δούμε τον βέλτιστο αριθμό των clusters που θα διαλέξουμε.



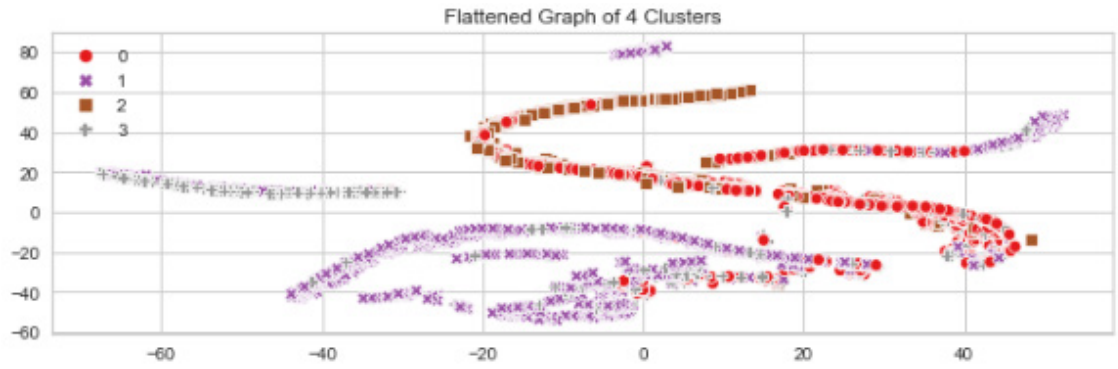

*Γράφημα 22*

Διαλέγουμε τον αριθμό 4 ως το βέλτιστο αριθμό για τα clusters που θα δημιουργήσουμε, καθώς εκεί φαίνεται να δημιουργείται η καμπή στο γράφημα elbow (Γράφημα 22).

Για να είμαστε σίγουροι ότι έχουμε επιλέξει το σωστό αριθμό ομάδων στην ανάλυση μας θα προχωρήσουμε αρχικά σε έναν έλεγχο που θα οπτικοποιήσουμε την ομαδοποίηση μας για K = 3 (Γράφημα 23), K = 4 (Γράφημα 24) και K = 5 (Γράφημα 25) για αντίστοιχες ομάδες, χρησιμοποιώντας τον αλγόριθμο k – means ώστε να δούμε ποια προσέγγιση είναι καλύτερη.




*Γράφημα 23*




*Γράφημα 24*


*Γράφημα 25*

Από τα τρία flattened graphs που βλέπουμε στο Γράφημα 23 για K = 3, στο Γράφημα 24 για K = 4, στο Γράφημα 25 για K = 5, παρατηρούμε ότι καλύτερος διαχωρισμός προκύπτει όντως για K = 4 με τις ομάδες να είναι περισσότερο διακριτές σε σχέση με τα άλλα.

Χρησιμοποιούμε τον αλγόριθμό k-means και κατηγοριοποιούμε τα δεδομένα μας σε 4 clusters όπως φαίνεται παρακάτω στην Εικόνα 22.

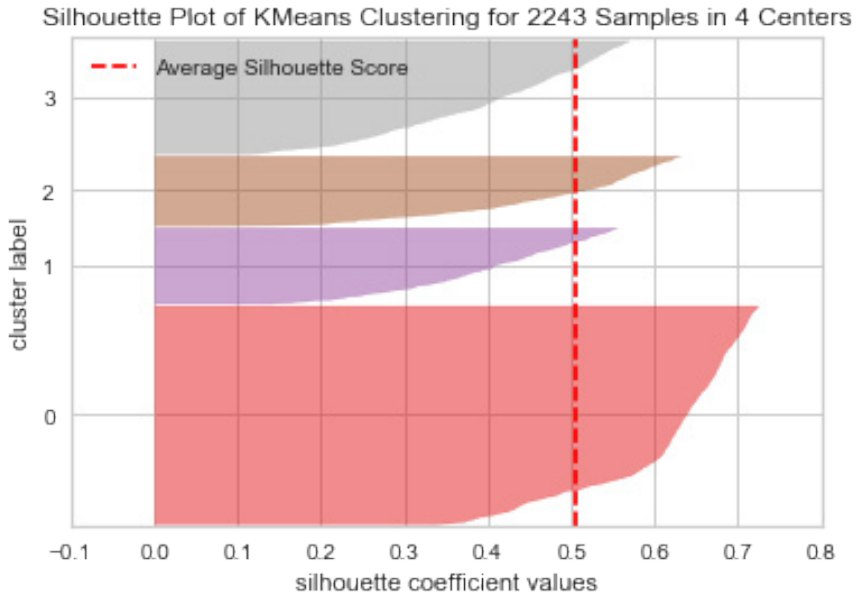
| | Customer_ID | Recency | Frequency | Monetary | R | F | M | RFM_Segment | RFM_Score | RFM_level | K_Cluster |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 100040 | 44 | 216 | 757495.00 | 4 | 4 | 4 | 444 | 12.0 | 3.0 | 2 |
| 1 | 100061 | 164 | 3 | 6739.50 | 4 | 3 | 1 | 431 | 8.0 | 2.0 | 0 |
| 2 | 100085 | 370 | 11 | 75691.43 | 3 | 4 | 3 | 343 | 10.0 | 3.0 | 1 |
| 3 | 100086 | 72 | 12 | 113221.75 | 4 | 4 | 3 | 443 | 11.0 | 3.0 | 2 |
| 4 | 100088 | 587 | 2 | 11670.34 | 2 | 2 | 1 | 221 | 5.0 | 1.0 | 3 |

*Εικόνα 22*

- **Silhouette Score**

Επόμενο βήμα είναι να κάνουμε επικύρωση των clusters που χρησιμοποιήσαμε υπολογίζοντας την μετρική Silhouette Score η οποία χρησιμοποιείται για να δούμε την ποιότητα των ομάδων που προκύπτουν. Ο δείκτης αυτός παίρνει τιμές από -1 έως 1. Κατά γενική ομολογία τιμές που είναι πάνω από 0,5 σηματοδοτούν ότι έχουμε υψηλής

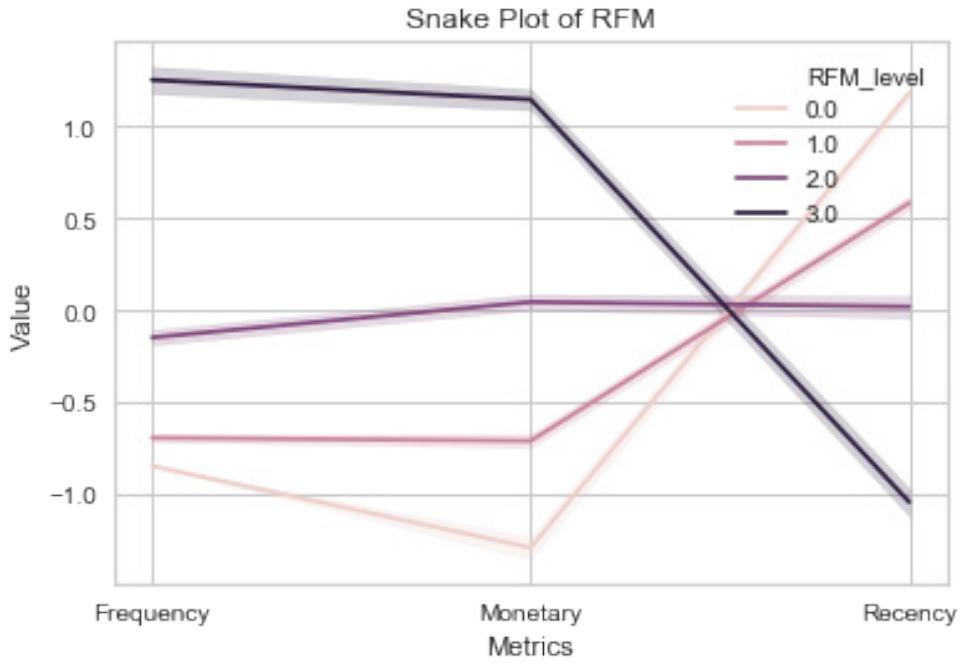
ποιότητας clusters. Στην προκειμένη περίπτωση της ανάλυσης μας ο δείκτης έχει τιμή Silhouette Score = 0,507 που είναι πάνω από 0,5. Κάποια ανυπέρβλητα ποιοτικά χαρακτηριστικά που θα αναλυθούν σε επόμενο κεφάλαιο σχολιασμού των αποτελεσμάτων, μας εμπόδισαν να έχουμε υψηλότερο score. Τη γραφική απεικόνιση των ομάδων μπορούμε να δούμε αναλυτικά και στο Γράφημα 26.



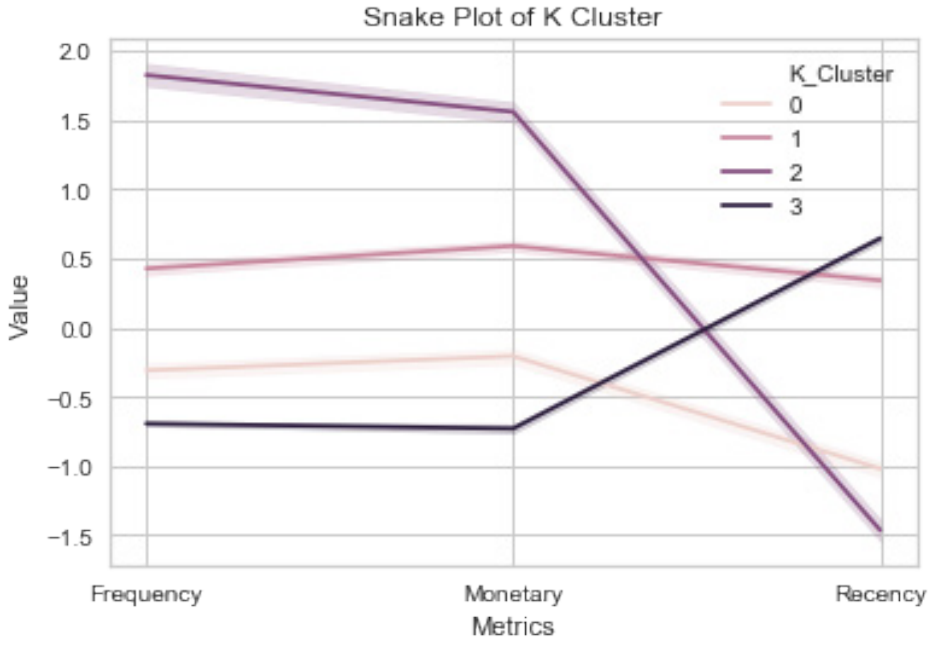

*Γράφημα 26*

## 5.4 Συνδυασμός αποτελεσμάτων RFM & K - Means

Ακολούθως θα κάνουμε μία σύγκριση των ομάδων που προέκυψαν από τον αλγόριθμο K – Means σε σχέση με την RFM ανάλυση που είχε προηγηθεί. Για να το κάνουμε αυτό ως πρώτο βήμα θέτουμε τη δημιουργία snake plots για να συγκρίνουμε το RFM level που είχαμε υπολογίσει (Γράφημα 27) και αντίστοιχα τα K – clusters που δημιουργήσαμε (Γράφημα 28).




*Γράφημα 27*




*Γράφημα 28*

Παρατηρούμε ότι στο snake plot για το RFM η ομάδα που έχει RFM level = 3 και αντιστοιχεί στους Best Customers με χαρακτηριστικό γνώρισμα RFM Score μεγαλύτερο ή ίσο από 10 δηλαδή που αντιστοιχεί σε πελάτες με υψηλή βαθμολογία σε R, F, M έχει παρόμοια απεικόνιση με την ομάδα Cluster = 2 στο snake plot για K clusters.

Προχωρώντας στις συγκρίσεις μας, βλέπουμε ότι στο snake plot για το RFM η ομάδα που έχει RFM level = 2 και αντιστοιχεί στους Promising customers με RFM Score μεγαλύτερο ή ίσο από 7 και μικρότερο από 10, έχει μεγάλη ομοιότητα με την ομάδα Cluster = 1 στο snake plot για K clusters.

Αντίστοιχα βλέπουμε ότι στο snake plot για το RFM η ομάδα που έχει RFM level = 1 και αντιστοιχεί στους πελάτες που έχουμε χαρακτηρίσει ως Need attention με RFM Score μεγαλύτερο ή ίσο από 4 και μικρότερο από 7, έχει κάποια ομοιότητα με την ομάδα Cluster = 0 στο snake plot για K clusters.

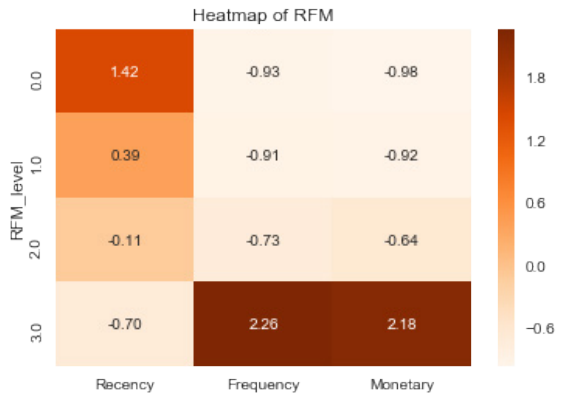
Τέλος, στο snake plot για το RFM η ομάδα που έχει RFM level = 0 και αντιστοιχεί στους Lost customers, με χαρακτηριστικό γνώρισμα RFM Score μικρότερο από 4 δηλαδή αντιστοιχεί σε πελάτες με τη χαμηλότερη βαθμολογία σε R, F, M έχει ομοιότητα με τους πελάτες της ομάδας που απομένει στο snake plot για K clusters με Cluster = 3.

Οι μικρότερες ομοιότητες που παρατηρούμε μεταξύ RFM level = 0 και Cluster = 3 και μεταξύ RFM level = 1 και Cluster = 0 πιθανόν να τις έχουμε γιατί το χρονικό πλαίσιο που έχουμε χρησιμοποιήσει για να τραβήξουμε στοιχεία, περιέχει το διάστημα της καραντίνας που λόγω της πανδημίας covid-19 πολλοί πελάτες είχαν επηρεαστεί από τα lock down που επιβάλλονταν σε διαφορετικά χρονικά σημεία ανάλογα με την εξέλιξη της πανδημίας και τα αντίστοιχα μέτρα που επέβαλλαν οι κυβερνήσεις διαφορετικών χωρών ανά περιόδους από το 2020 έως και τα μέσα περίπου του 2021 δεν ήταν ταυτόχρονα όμοια παντού. Αυτό σημαίνει ότι πολλοί πελάτες που μπορεί να σημείωσαν επιβράδυνση από εξωγενείς παράγοντες και να έπεσαν κατηγορία, από Need attention δηλαδή να μετατοπίστηκαν σε Lost Customers ή αντίστροφα. Αυτό είναι κάτι που θα το δούμε σε μεγαλύτερο βάθος στο κεφάλαιο με τα συμπεράσματα, στο τέλος της ανάλυσης μας.

Μπορούμε να κάνουμε την ίδια σύγκριση με παραπάνω δημιουργώντας γραφήματα heatmaps για να δούμε πως επηρεάζει την κάθε ομάδα τόσο από την RFM ανάλυση όσο και για K clusters ανάλυση που έγινε με K – means οι παράγοντες που έχουμε επικεντρωθεί, δηλαδή το Recency, το Frequency και το Monetary. Ο βαθμός

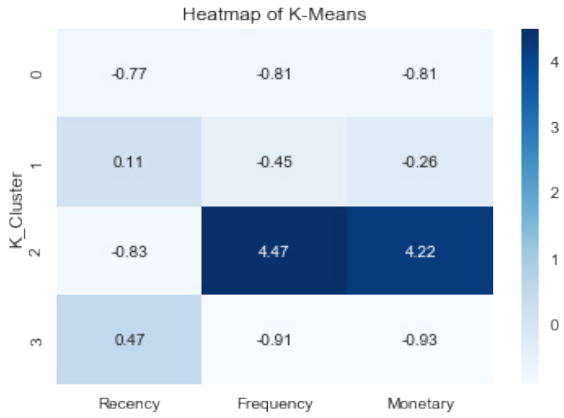
σημαντικότητας κάθε παράγοντα είναι μεγαλύτερος όσο πιο μακριά από το 0 είναι ο δείκτης.

Ξεκινώντας από το RFM βλέπουμε στο heatmap (Γράφημα 29) ότι για RFM level = 0 υψηλότερη σημασία έχει το Recency το οποίο είναι λογικό αφού αυτός είναι ένας από τους πιο κρίσιμους παράγοντες για να θεωρηθεί ένας πελάτης ως Lost Customer. Για RFM level = 3 βλέπουμε ότι περισσότερο σημαντικά είναι το Frequency και το Monetary αφού βάσει αυτών ένας πελάτης μπορεί να θεωρηθεί ως Best customer.



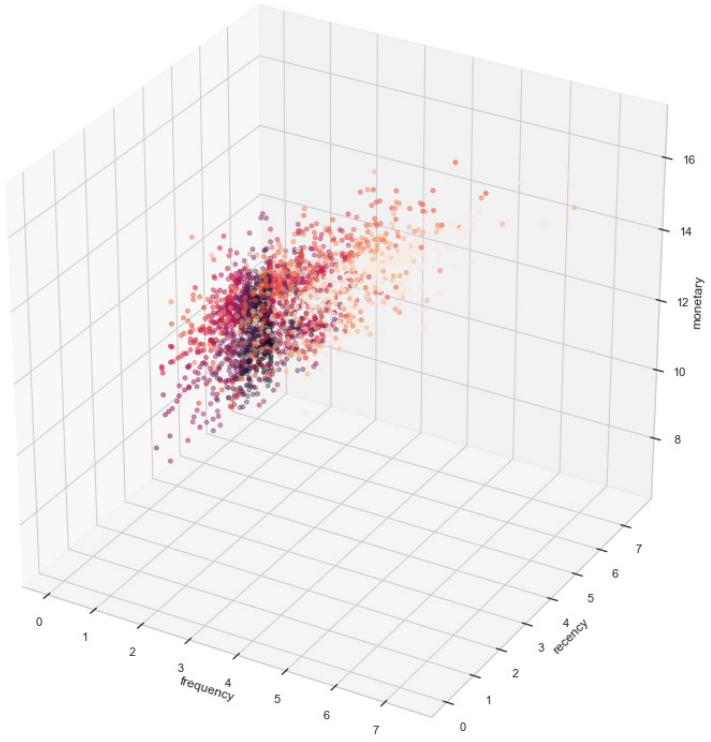

*Γράφημα 29*

Αντίστοιχα στο heatmap (Γράφημα 30) για τα K clusters, βλέπουμε ότι σημαντικότερα είναι για την ομάδα Cluster = 2 τα Recency και Frequency, ενώ για το Cluster = 3 σημαντικότερο είναι το Monetary.




*Γράφημα 30*

Η 3D απεικόνιση (Γράφημα 31) συνδράμει στο να δούμε τη διασπορά των δεδομένων των clusters για τα Recency, Frequency και Monetary σε τρεις διαστάσεις, ώστε να έχουμε μία συνολική εικόνα σχετικά με το πως κατανέμονται στο χώρο.

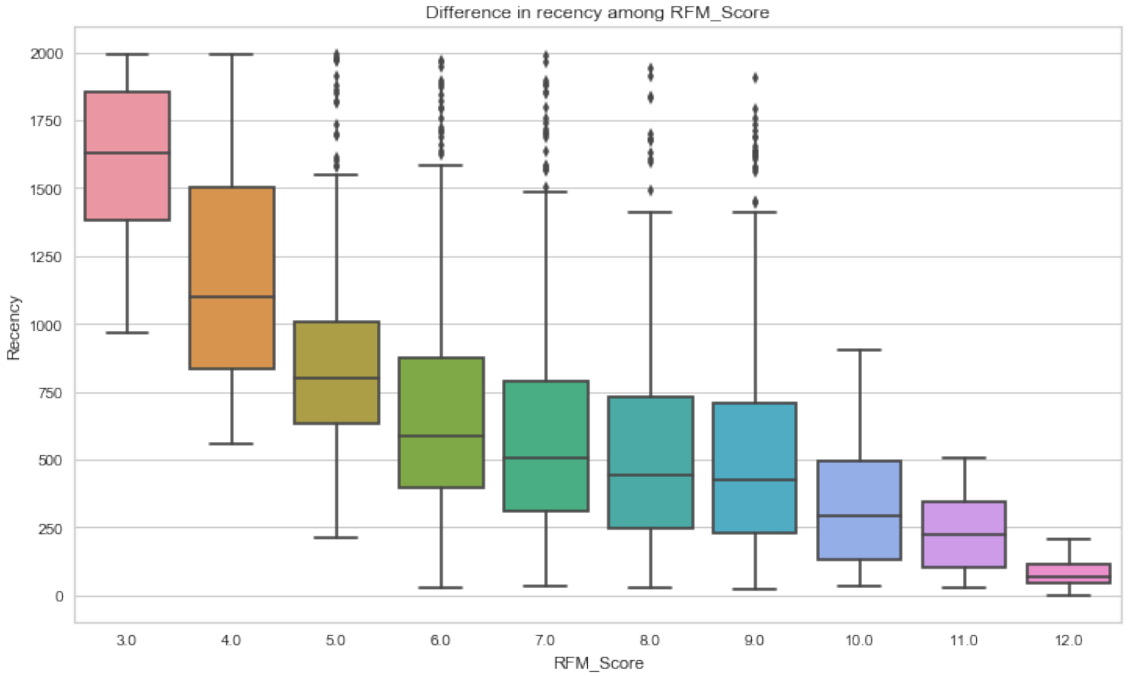



*Γράφημα 31*

- **Σύγκριση Recency, Frequency, Monetary με RFM Score**

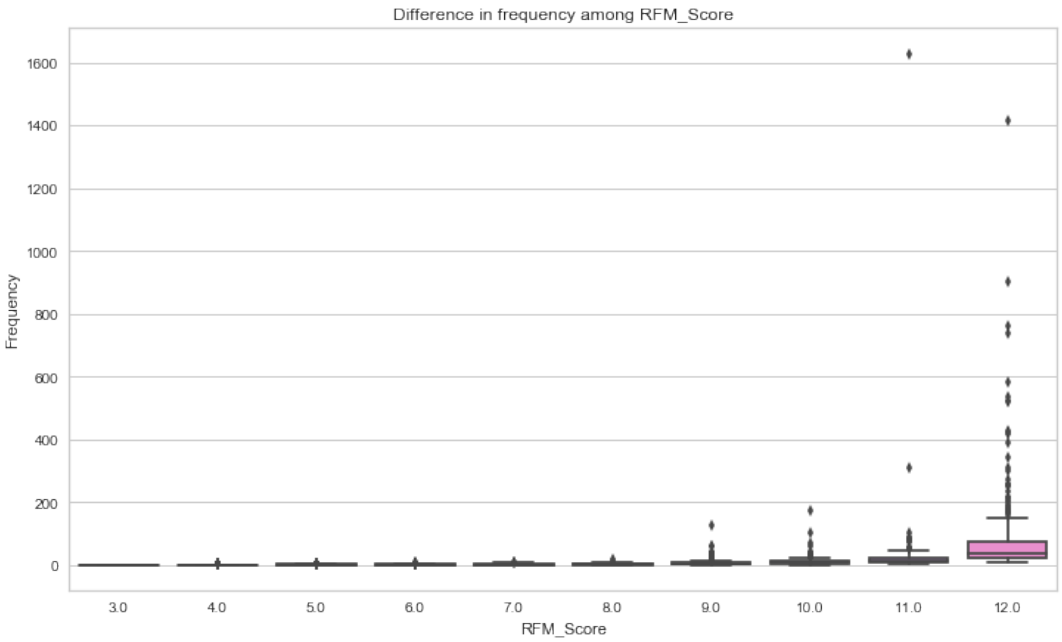
Επόμενο βήμα είναι να δούμε σε boxplots την κατανομή των Recency, Frequency, Monetary αρχικά για κάθε RFM Score. Ξεκινώντας να βλέπουμε στο boxplot (Γράφημα 32) για το RFM Score σε σχέση με το Recency, βλέπουμε ότι όπως είναι αναμενόμενο όσο μεγαλύτερο είναι το RFM Score τόσο χαμηλότερο είναι το Recency. Ενδιαφέρον παρουσιάζει το γεγονός ότι υπάρχουν κάποιες ακραίες τιμές στα RFM Score που είναι ίσα με 5, 6, 7, 8 και μετά λιγότερο στο 9. Αυτά είναι RFM Score στα οποία ανήκουν πελάτες που είναι στο μέσο περίπου της κλίμακας βάσει περιγραφής και το γεγονός ότι

έχουμε κάποιες ακραίες τιμές ενισχύει αυτό που αναφέραμε και πιο πάνω, ότι δηλαδή υπήρξαν περιπτώσεις που κάποιοι πελάτες εμφάνισαν κάποια χαρακτηριστικά που αρχικά δεν συνδέονται τόσο πολύ με την θέση στην οποία βρίσκονται στην κατάταξη αυτή, κάτι το οποίο έχει άμεση σχέση αφενός με την πανδημία που επηρέασε τις αγορές πελατών για μεγάλο διάστημα αλλά επίσης και από την αναστάτωση που υπήρξε στην εφοδιαστική αλυσίδα παγκοσμίως. Τα πιο συμπαγή RFM Score όσο αφορά την κατανομή είναι τα υψηλότερα δηλαδή το 12 και το 11.



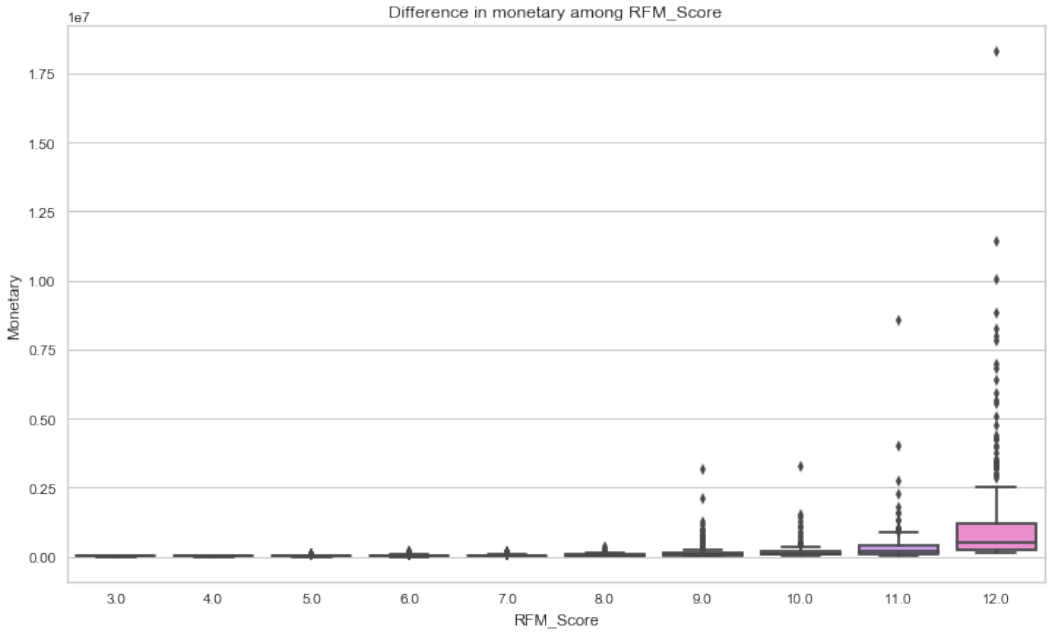

*Γράφημα 32*

Προχωρώντας βλέπουμε το boxplot (Γράφημα 33) για το RFM Score σε σχέση με το Frequency, βλέπουμε πάλι ότι όσο αυξάνεται το RFM Score αυξάνεται ταυτόχρονα και το Frequency, ενώ περισσότερες ακραίες τιμές για RFM Score ίσο με 12, 11, 10 και 9 ενώ παρατηρούμε ότι τα χαμηλότερα score 3 έως και 8 είναι πιο συμπαγή σε σχέση με τα υψηλότερα RFM score που ακολουθούν.




*Γράφημα 33*

Τέλος όσο αφορά το boxplot (Γράφημα 34) για το RFM Score σε σχέση με το Monetary, βλέπουμε ότι τις περισσότερες ακραίες τιμές τις βρίσκουμε για RFM Score 12 και 10 ενώ οι πιο συμπαγής κατηγορίες όσο αφορά την κατανομή τις βρίσκουμε στα RFM score, από 3 έως και 8.

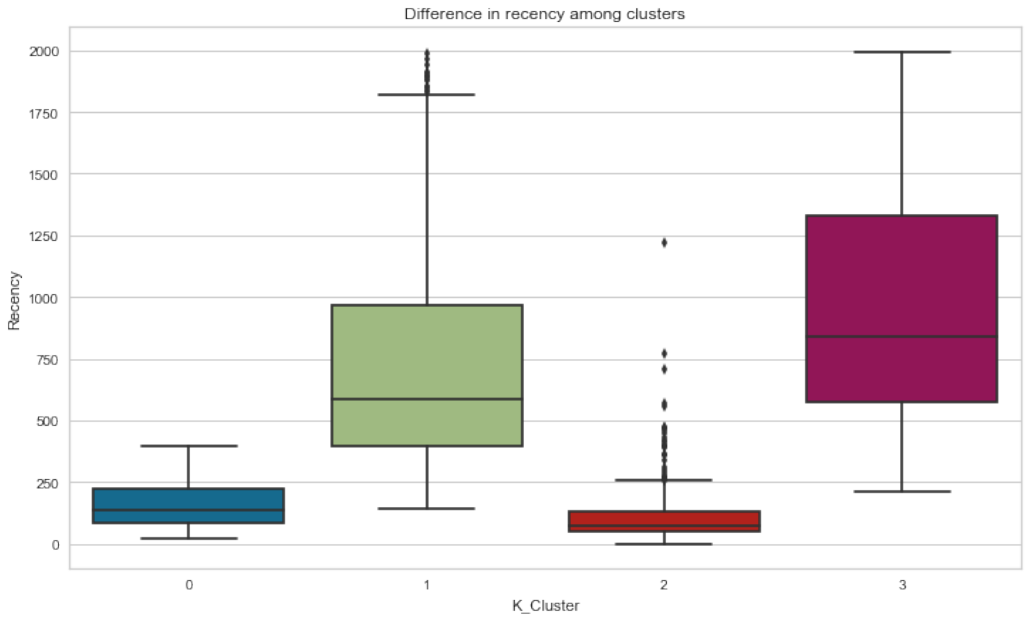



*Γράφημα 34*

- **Σύγκριση Recency, Frequency, Monetary με K Clusters**

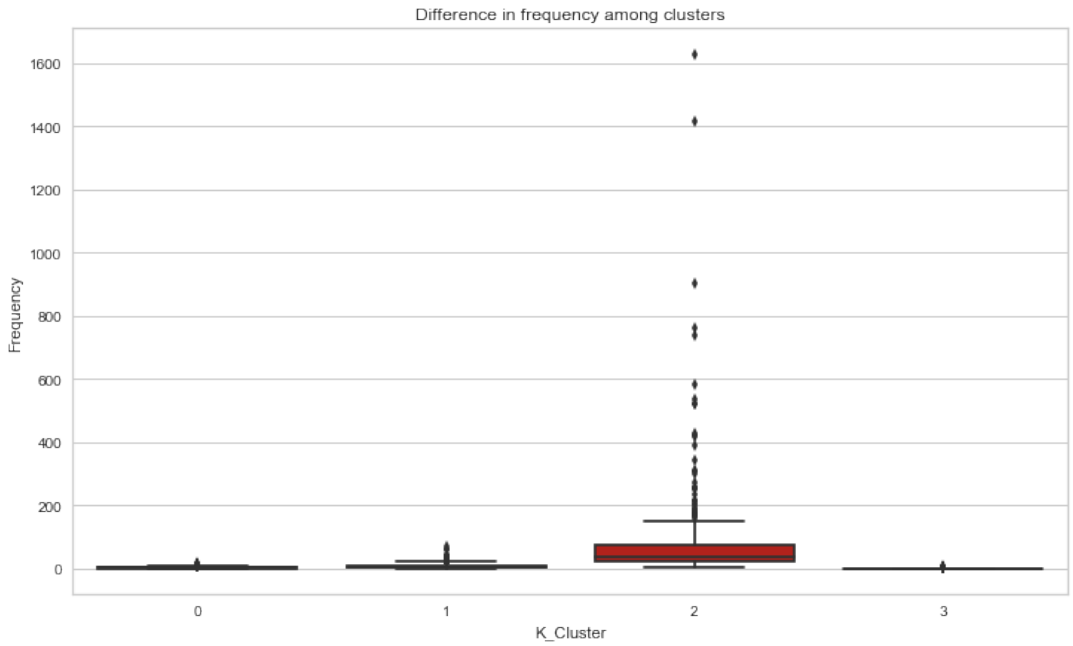
Επόμενο στάδιο είναι να δούμε τα boxplots που μας δείχνουν την κατανομή των δεδομένων για τα K clusters σε σχέση με τα Recency, Frequency, Monetary αντίστοιχα.

Αρχίζοντας από το boxplot (Γράφημα 35) για τα clusters σε σχέση με το Recency, βλέπουμε ότι ακραίες τιμές έχουμε μόνο στο cluster 2 και η πιο συμπαγής ομάδα είναι επίσης το cluster 2.



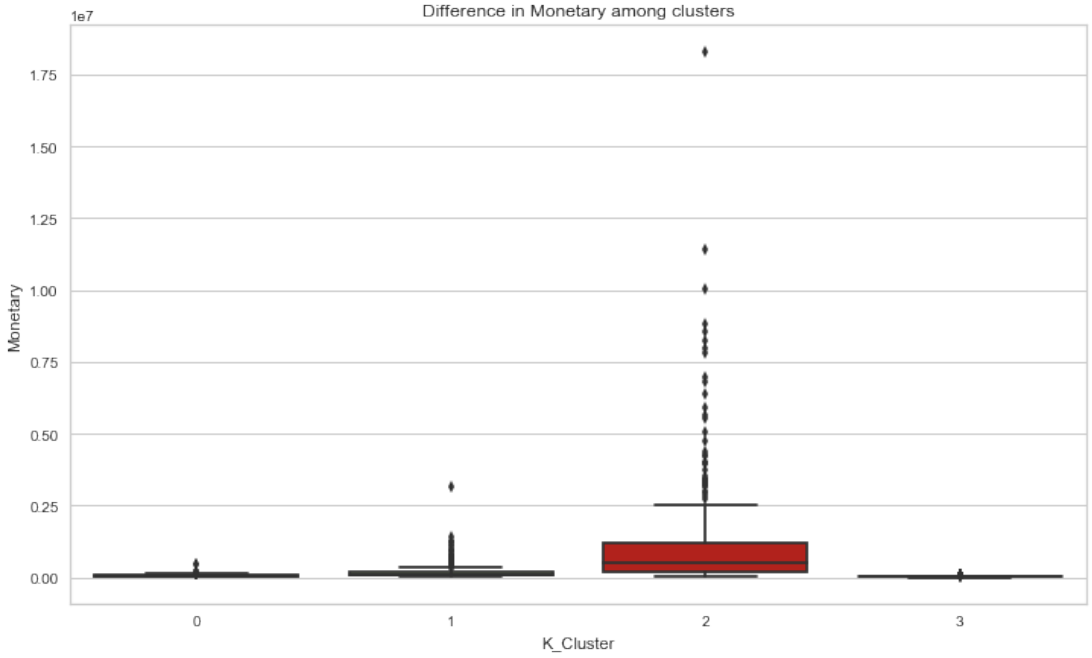

*Γράφημα 35*

Συνεχίζοντας με το boxplot (Γράφημα 36) για τα clusters σε σχέση με το Frequency, βλέπουμε ότι παρατηρούνται κάποιες ακραίες τιμές στο cluster 2 ενώ οι ομάδες 0,1,3 είναι εμφανώς πιο συμπαγείς σε σχέση με την ομάδα 2. Αυτό έχει νόημα γιατί υπάρχει μεγάλη διαφορά στη συχνότητα των συναλλαγών ανάμεσα στους πελάτες που ανήκουν στην υψηλότερη κατηγορία σε σχέση με τους υπόλοιπους.




*Γράφημα 36*

Τέλος βλέπουμε το boxplot (Γράφημα 37) για τα clusters σε σχέση με το Monetary όπου βλέπουμε αρκετές ακραίες τιμές στο cluster 2 και μετά στο cluster 1. Οι πιο συμπαγείς ομάδες είναι τα cluster 0 και 3.




*Γράφημα 37*

# 6 Συμπεράσματα

Στο τελευταίο στάδιο της εργασίας, κατηγοριοποιούμε με βάση τους μέσους όρους των Recency, Frequency και Monetary κάθε ένα από τα 4 clusters που έχουν σχηματιστεί (Εικόνα 23) για να οδηγηθούμε σε συμπεράσματα για την εικόνα κάθε ομάδας.

Με βάση αυτή τη λογική, βλέπουμε ότι στο cluster 0 έχουμε συνολικά 357 πελάτες, με μέσο όρο τελευταίας συναλλαγής πριν 154 ημέρες, συχνότητα συναλλαγών 3 και συνολική αξία συναλλαγών 42.477 €.

Για το cluster 1 βλέπουμε ότι έχουμε συνολικά 532 πελάτες, με μέσο όρο τελευταίας συναλλαγής πριν 726 ημέρες, συχνότητα συναλλαγών 8 και συνολική αξία συναλλαγών 165.862 €.

Για το cluster 2 βλέπουμε ότι έχουμε συνολικά 327 πελάτες, με μέσο όρο τελευταίας συναλλαγή πριν 114 ημέρες, συχνότητα συναλλαγών 82, και συνολική αξία συναλλαγών 1.177.313 €.

Τέλος για το cluster 3 βλέπουμε ότι έχουμε συνολικά 1027 πελάτες, με μέσο όρο τελευταίας συναλλαγής πριν 965 ημέρες, συχνότητα συναλλαγών 1 και συνολική αξία συναλλαγών 16.720 €.

| | Recency | Frequency | Monetary | |
|---|---|---|---|---|
| | mean | mean | mean | count |
| K_Cluster | | | | |
| 0 | 154.0 | 3.0 | 42477.0 | 357 |
| 1 | 726.0 | 8.0 | 165862.0 | 532 |
| 2 | 114.0 | 82.0 | 1177313.0 | 327 |
| 3 | 965.0 | 1.0 | 16720.0 | 1027 |

*Εικόνα 23*

Ακολούθως κάνουμε ακριβώς το ίδιο πράγμα παίρνοντας τους μέσους όρους των Score των Recency, Frequency και Monetary για κάθε ένα από τα 4 clusters που έχουν σχηματιστεί (Εικόνα 24).

Παρατηρούμε ότι:

Το cluster 0 αντιστοιχεί κατά μέσο όρο σε R = 3,71 , F = 2,48 και M = 2,33.

Το cluster 1 αντιστοιχεί κατά μέσο όρο σε R = 2,22 , F = 3,31 και M = 3,38.

Το cluster 2 αντιστοιχεί κατά μέσο όρο σε R = 3,87 , F = 3,99 και M = 3,95.

Τέλος, το cluster 3 αντιστοιχεί κατά μέσο όρο σε R = 1,79 , F = 1,62 και M = 1,64.

| | Rn | Fn | Mn | |
|---|---|---|---|---|
| | mean | mean | mean | count |
| K_Cluster | | | | |
| 0 | 3.71 | 2.48 | 2.33 | 357 |
| 1 | 2.22 | 3.31 | 3.38 | 532 |
| 2 | 3.87 | 3.99 | 3.95 | 327 |
| 3 | 1.79 | 1.62 | 1.64 | 1027 |

*Εικόνα 24*

- **RFM και K – means**

Κάτι το οποίο μπορούμε να παρατηρήσουμε είναι ότι δεν υπάρχει πλήρης ομοιότητα στην κατηγοριοποίηση που αρχικά είχαμε θέσει μέσω μόνο της RFM ανάλυσης και της ανάλυσης με clusters που καταλήξαμε μετά την εφαρμογή του αλγορίθμου K – means για όλες τις ομάδες.

Αρχικά παρατηρούμε ότι μέσω της cluster analysis το **cluster 2** μέσω των μοτίβων που έχει ανιχνεύσει ο αλγόριθμος αντιστοιχούν σε πολύ μεγάλο βαθμό στους πελάτες που αρχικά είχαμε αναγνωρίσει ως Best customers. Είναι βέβαιο ότι οι πελάτες αυτοί είναι πελάτες αξίας για την επιχείρηση αλλά είναι λιγότεροι από τους 603 που είχαμε αναγνωρίσει μόνο μέσω της RFM ανάλυσης. Ο αλγόριθμος ομαδοποίησης τους έχει περιορίσει στον αριθμό των 327 πελατών δίνοντας μας την ευκαιρία να εστιάσουμε πιο συγκεκριμένα σε αυτούς οι οποίοι είναι οι καλύτεροι ανάμεσα στο σύνολο και αποτελούν περίπου το 15 % επί του συνόλου.

Όσο αφορά το **cluster 1** έχουμε ομαδοποιήσει 532 πελάτες αποτελώντας περίπου το 24% των πελατών. Οι πελάτες αυτοί έχουν παρόμοια χαρακτηριστικά με τους Promising

customers που είχαμε αναγνωρίσει μέσω μόνο της RFM analysis οι οποίοι με βάση την αρχική προσέγγιση ήταν συνολικά 741. Αυτό σημαίνει ότι και πάλι έχουμε μία μικρότερη πελατειακή βάση μέσω της διαδικασίας με την cluster ανάλυση η οποία μας βοηθάει να εστιάσουμε σε πελάτες που έχουν μία δυναμική και μπορούμε να επικεντρωθούμε καλύτερα σε αυτούς για να τους προσεγγίσουμε με καλύτερο τρόπο.

Στο **cluster 0** βλέπουμε ότι έχουμε συνολικά 357 πελάτες που αντιστοιχούν περίπου στο 16 % του συνόλου. Μπορούμε να πούμε ότι έχουν παρόμοια χαρακτηριστικά ως ένα βαθμό, με τη λογική που έχουμε θέσει για τους πελάτες και αρχικά είχαμε χαρακτηρίσει ως Need attention. Είναι πελάτες που η εταιρεία μπορεί να εστιάσει ώστε να μην τους χάσει και δώσει κίνητρα για να ανέβουν κατηγορία.

Τέλος στο **cluster 3** βλέπουμε ότι έχουμε την πλειοψηφία των πελατών αφού ο αλγόριθμος έχει ομαδοποιήσει 1.027 συνολικά πελάτες που αντιστοιχούν περίπου στο 45% των πελατών. Η προσέγγιση που υπάρχει μέσω αυτής της κατηγορίας είναι ότι έχουν ομαδοποιηθεί όλοι οι πελάτες που ανήκουν στην πιο αδύναμη κατηγορία.

- **Επεξήγηση ανωμαλιών στα clusters λόγω πανδημίας**

Μία γενική παρατήρηση που μπορούμε να κάνουμε βλέποντας το γράφημα 24 που δείχνει οπτικά πως χωρίζονται οι ομάδες που έχουμε επιλέξει να εισάγουμε, είναι ότι υπάρχει ένα σημείο στο χώρο που το cluster 1 συναντιέται με στοιχεία του cluster 0. Αυτή είναι μια συμπεριφορά που μπορεί εκ πρώτης όψεως να δημιουργήσει κάποιες απορίες αφού βλέπουμε ότι πελάτες που είναι άνω του μετρίου και έχουν χαρακτηριστεί ως Promising customers, φαίνεται να μπλέκονται σε αυτή την ομάδα στο γράφημα με πελάτες που είναι μία κατηγορία κάτω από άποψη δυναμικής και έχουν χαρακτηριστεί ως Need attention.

Προσπαθώντας να αναγνωρίσουμε κάποια επιμέρους κοινά χαρακτηριστικά μεταξύ των περιπτώσεων που εμπίπτουν σε αυτές τις κατηγορίες και συμπίπτουν (cluster 1 και cluster 0), ανακτούμε τους πελάτες αυτών των κατηγοριών για να τους εξετάσουμε περαιτέρω. Κάτι που παρατηρούμε είναι ότι οι κυρίαρχες αγορές σε αυτές τις περιπτώσεις είναι η Ελλάδα, η Σερβία, η Σουηδία, και η Ουγγαρία.

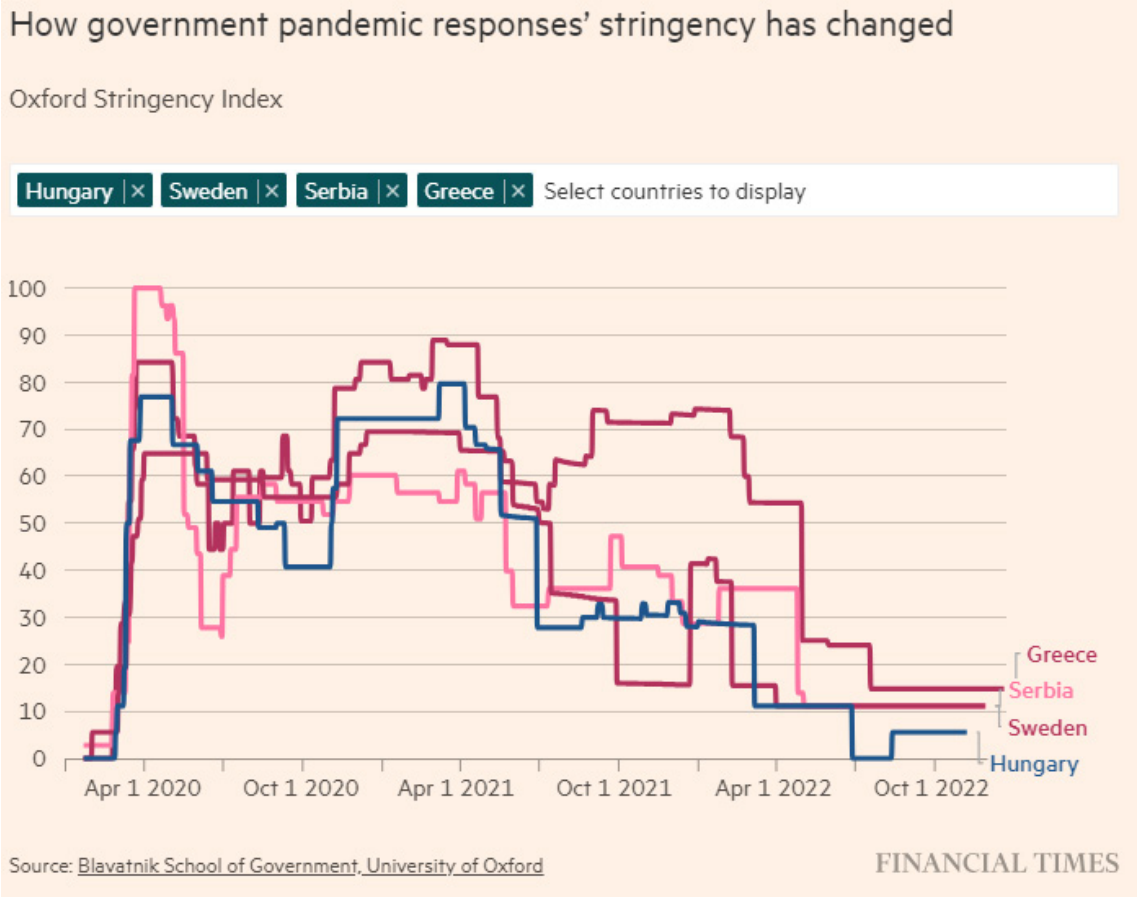
Επιλέξαμε να κάνουμε μία σύνδεση αυτών των αγορών με τα μέτρα που έλαβαν οι χώρες που προαναφέραμε κατά την πανδημία. Για να το καταφέρουμε αυτό, ανατρέξαμε να

βρούμε στοιχεία αρχικά γενικά για την πανδημία και τα μέτρα που λάμβαναν οι χώρες κατά τη διάρκεια της, κάτι που σημαίνει ότι κάνουμε το συσχετισμό ότι όσο πιο αυστηρά μέτρα έπαιρναν οι χώρες τόσο μεγαλύτερη επιβράδυνση σημείωνε η ζήτηση για τα προϊόντα της εταιρείας που εξετάζουμε.

Το πανεπιστήμιο της Οξφόρδης, δημιούργησε έναν δείκτη αυστηρότητας των μέτρων κατά της πανδημίας στον οποίο συνδυάζει τα μέτρα που έλαβε η κάθε χώρα μέσω των πολιτικών που εφάρμοζε κατά τη διάρκεια της πανδημίας (Hale, et al., 2023). Για τη δημιουργία του δείκτη έλαβαν υπόψιν ως παραμέτρους κυρίως το κλείσιμο των σχολείων, το κλείσιμο των επιχειρήσεων, την ακύρωση δημόσιων εκδηλώσεων, τον περιορισμό που αφορά το σύνολο των ατόμων που μπορούν να είναι μαζί στον ίδιο χώρο, το κλείσιμο των μέσων μαζικής μεταφοράς, την απαγόρευση κυκλοφορίας, τους περιορισμούς διεθνών μετακινήσεων, τους περιορισμούς τοπικών μετακινήσεων. Βάσει αυτών των χαρακτηριστικών προέκυψε ο δείκτης αυστηρότητας των μέτρων κατά της πανδημίας και δημιουργήθηκε ένα "online data analytics tool" που ο χρήστης μπορεί να εισάγει διαφορετικές χώρες, και να συγκρίνει την αυστηρότητα των μέτρων που είχε η κάθε μία κατά τη διάρκεια της πανδημίας (Financial Times Journalism, 2023).

Χρησιμοποιώντας αυτό το online data analytics tool μέσω των Financial Times, που επιτρέπει στους χρήστες να εισάγουν τις χώρες που θέλουν για να συγκρίνουν την αυστηρότητα των μέτρων κατά τη διάρκεια της πανδημίας, δημιουργήσαμε ένα γράφημα που συγκρίνει τις χώρες που αναγνωρίσαμε ότι αποτελούν την πλειοψηφία των clusters 1 και 0. Από το γράφημα που προέκυψε και μπορούμε να δούμε παρακάτω στην εικόνα 25, βλέπουμε ότι έχουν σημειωθεί σημαντικές διαφορές όσο αφορά την αυστηρότητα των μέτρων κατά τη διάρκεια της πανδημίας μεταξύ αυτών των χωρών, κάτι που εξηγεί το λόγο που επηρεάστηκαν οι καταναλωτικές συνήθειες των πελατών από αυτές τις χώρες κατά τη διάρκεια της πανδημίας.

Πιο συγκεκριμένα βλέπουμε ότι τον Απρίλιο του 2020 η Σερβία και η Ελλάδα είχαν υιοθετήσει από τα αυστηρότερα μέτρα, σε αντίθεση με χώρες όπως η Σουηδία. Επίσης, στο διάστημα Οκτώβριος του 2020 με Απρίλιο του 2021 βλέπουμε ότι υπάρχει μεγάλη διαφορά μεταξύ Ελλάδος και Σερβίας, όπως και στο διάστημα Οκτώβριος του 2021 με Απρίλιο 2022 βλέπουμε ότι υπάρχει μεγάλη διαφορά μεταξύ Ελλάδος και Σουηδίας.




*Εικόνα 25*

Συνοψίζοντας αυτοί είναι παράγοντες που είναι ικανοί να ωθήσουν έναν πελάτη που αρχικά ήταν στο cluster 0 να πάει στο cluster 1, όπως και αντίστροφα έναν πελάτη που αρχικά ήταν αρχικά στο cluster 1 να μετακινηθεί στο cluster 0.

- **Προτεινόμενες στρατηγικές Μάρκετινγκ**

Μετά από την παραπάνω ανάλυση που έχει προηγηθεί, έχουμε καταλήξει σε συνεννόηση και με την επιχείρηση ότι θα ακολουθήσουμε την ομαδοποίηση που προκύπτει μέσω των clusters για να προβούμε σε συγκεκριμένες προτάσεις που αφορούν στρατηγικές μάρκετινγκ για κάθε μία ομάδα.

Παρακάτω, για κάθε ένα από τα τέσσερα cluster που προέκυψαν, σε συνδυασμό με τα RFM χαρακτηριστικά θα προτείνουμε μία στρατηγική με βάση και τη θεωρεία πάνω στις στρατηγικές μάρκετινγκ που είδαμε στα κεφάλαια 3.5 και 3.6 ώστε να μπορέσουν ο Εμπορικός Διευθυντής και ο Διευθυντής Marketing & Customer Experience να

δημιουργήσουν κατάλληλο μείγμα μάρκετινγκ για την προσέγγιση κάθε στόχου - ομάδας πιο αποτελεσματικά.

- ***Cluster (3) - Πελάτες με χαμηλά R, F M (Lost & Almost Lost Customers).***

Προτεινόμενη στρατηγική: Σε αυτή την ομάδα έχουμε συνολικά 1.027 πελάτες. Είναι πελάτες με χαμηλή δυναμική που θα χαρακτηρίσουμε ως "Lost & Almost Lost Customers". Κρίθηκε σκόπιμο να μην έχουμε ακριβώς την ίδια προσέγγιση απέναντι σε όλους, αφού είναι μεγάλη ομάδα. Με βάση και τη θεωρεία που έχουμε στα κεφάλαια 3.5 και 3.6 συνοπτικά μπορούμε να πούμε ότι πρόκειται κυρίως για πελάτες στο στάδιο ζωής φθοράς και η εταιρεία θα πρέπει να επιλέξει μεταξύ στρατηγικών "win-back" εφόσον κρίνει ότι είναι πελάτες "Almost Lost" και μπορούν να δώσουν αξία στην εταιρεία αν παραμείνουν, ενώ αν πρόκειται για πελάτες που δεν προσθέτουν αξία και είναι "Lost customers" μπορεί να κρίνει ότι το καλύτερο σενάριο είναι η απεμπλοκή. Αναλυτικότερα μπορούμε να δούμε αυτές τις δύο υποκατηγορίες παρακάτω.

Για τους ουραγούς αυτής της κατηγορίας που ανήκουν ταυτόχρονα στην κατηγορία όπου RFM = 111 και αποτελούν υποσύνολο του cluster 3, βλέπουμε ότι έχουμε συνολικά 149 πελάτες (δηλαδή περίπου 7% επί του συνόλου). Για αυτούς προτείνουμε μείωση του κόστους προσέγγισης τους, επειδή δεν αξίζει να δαπανήσουμε πολλούς πόρους της επιχείρησης και θεωρούμε ότι πρόκειται για πελάτες που οριστικά έχουμε χάσει (Lost Customers). Κάτι που θα μπορούσαμε να κάνουμε ως επιπλέον ενέργεια είναι να κάνουμε μία προωθητική ενέργεια για ένα περιορισμένο χρονικό διάστημα, στέλνοντας τους κάποιο mail, που θα λέμε ότι αν κάνουν κάποια αγορά τους επόμενους έξι μήνες θα πάρουν ένα επιπλέον ποσοστό έκπτωσης. Εξαίρεση θα αποτελέσουν πελάτες αυτής της υποκατηγορίας που η εταιρεία θέλει να επιλέξει την απεμπλοκή της με αυτούς – περιπτώσεις με πελάτες που έχουμε οικονομικές εκκρεμότητες.

Για τους υπόλοιπους που είναι εμφανώς πελάτες μειωμένης δυναμικής και τους θεωρούμε Almost Lost Customers, βλέπουμε ότι είναι συνολικά 878 (δηλαδή περίπου 38% επί του συνόλου). Επειδή αντιλαμβανόμαστε ότι αποτελούν μία μεγάλη μερίδα της πελατειακής βάσης καταλαβαίνουμε ότι πρέπει να τους διατηρήσουμε ως πελάτες και να αναγνωρίσουμε τα χαρακτηριστικά που έχουν για να δούμε αν μπορούμε να τους προσεγγίσουμε με εναλλακτικούς τρόπους. Είχαμε αναφέρει ότι η ταυτότητα της εταιρείας έδινε μεγάλη έμφαση στο customization των προϊόντων της για τους πελάτες,

κάτι όμως που ανεβάζει σημαντικά το κόστος παραγωγής και κάνει την εταιρεία μη ανταγωνιστική σε σχέση με άλλους κατασκευαστές που προσφέρουν τυποποιημένα και φθηνά προϊόντα. Για το λόγο αυτό προτείνουμε να ενισχυθεί περαιτέρω η τυποποίηση κάποιων φθηνών (low-end) προϊόντων τα οποία θα προωθήσουμε στους πελάτες αυτής της ομάδας, χρησιμοποιώντας την αυτοματοποιημένη παραγγελιοληψία μέσω του portal στην οποία συμπεριλαμβάνονται τυποποιημένα προϊόντα χαμηλού κόστους για την επιχείρηση. Επίσης μπορούμε να τρέξουμε μία καμπάνια "win-back" για περιορισμένο χρονικό διάστημα με μία προσφορά που απευθύνεται σε αυτούς με σκοπό να ενεργοποιηθούν εκ νέου.

- ***Cluster (0) - Πελάτες με υψηλό R, μέτριο F και μέτριο M (Need attention - Newcomers).***

Προτεινόμενη στρατηγική: Είναι πελάτες που αρχικά χαρακτηρίσαμε ως "Need attention" αλλά λόγω του υψηλού Recency που παρατηρούμε θα προσθέσουμε μία νέα διάσταση αυτή των "Newcomers". Με βάση και τη θεωρεία είναι πελάτες που βρίσκονται στο στάδιο ζωής της απόκτησης.

Η πρώτη περίπτωση αυτής της ομάδας, είναι να έχουμε πελάτες οι οποίοι είναι καινούργιοι. Αυτό μπορούμε να το δούμε αν εξετάσουμε μέσω του CRM συστήματος της εταιρείας πότε άνοιξε ο κωδικός του πελάτη. Για τους καινούργιους πελάτες θα πρέπει να εστιάσουμε στο να τους κρατήσουμε και να τους δώσουμε κίνητρα να συνεχίσουν τη συνεργασία τους μαζί μας. Αυτό μπορεί να γίνει μέσω ενός προγράμματος membership που ο πελάτης θα μπορεί να συμπληρώνει πόντους με τις συναλλαγές του και θα μπορεί να ξεκλειδώσει αποκλειστικά προνόμια σε περίπτωση που θα πετύχει κάποιους στόχους όπως έξτρα εκπτώσεις σε ανταλλακτικά. Με τον τρόπο αυτό θα εισάγουμε και έμμεσα στρατηγική cross-selling αφού θα του δίνουμε κίνητρα για να κάνει αγορές πέρα από τελικά προϊόντα και σε ένα άλλο κομμάτι, αυτό των ανταλλακτικών το οποίο λειτουργεί παράλληλα.

Τέλος, για τους νέους πελάτες, θα πρέπει να υπάρχει εκτενής υποστήριξη τους από το τμήμα After Sales στέλνοντας τεχνικούς της εταιρείας όταν εγκαθιστούν το προϊόν την πρώτη φορά, ώστε να εκπαιδεύσουν τους πελάτες και να μπορούν να κάνουν όσο πιο εύκολα γίνεται την εγκατάσταση των ανελκυστήρων υιοθετώντας τα προϊόντα μας και για μελλοντικά έργα που θα πάρουν.

Η δεύτερη περίπτωση είναι να έχουμε κάποιους πελάτες που δεν είναι καινούργιοι αλλά απλά πραγματοποιούν λίγες συναλλαγές. Πολύ πιθανόν να είναι πελάτες που αγοράζουν προϊόντα και από άλλες ανταγωνιστικές εταιρείες παράλληλα. Αρχικά πρέπει να διερευνήσουμε τους λόγους που δεν πραγματοποιούν συχνά συναλλαγές. Για το λόγο αυτό θα πρέπει να κάνουμε έρευνες που θα εστιάζουν σε αυτούς τους πελάτες παίρνοντας απευθείας feedback μέσω ερωτηματολογίων, για να δούμε αδυναμίες που μπορεί να εμφανίζουμε απέναντι τους και να μην τις έχουμε αναγνωρίσει μέχρι στιγμής. Επειδή ενδεχομένως να είναι πελάτες χαμηλής δυναμικότητας, είναι η κατάλληλη ομάδα πελατών που θα πρέπει να πάρουμε ανατροφοδότηση για την εμπειρία αγορών μέσω του web portal ώστε να δούμε ότι όλα τα touchpoints της σελίδας μας τους ικανοποιούν, είναι εύκολο στη χρήση, εξετάζοντας αν υπάρχουν αλλαγές που θα πρέπει να γίνουν για να είναι καλύτερη η εμπειρία τους μέσω της πλατφόρμας μας.

- ***Cluster (1) - Πελάτες με μέτριο R, μέτριο F και υψηλό M (Promising Customers).***

Προτεινόμενη στρατηγική: Είναι πελάτες που τους χαρακτηρίζουμε "Promising", δηλαδή πελάτες που δεν έχουν προλάβει ακόμα να πραγματοποιήσουν πολλές συναλλαγές αλλά εμφανίζουν δυναμική μέσω συναλλαγών υψηλής αξίας. Σύμφωνα με τη θεωρεία είναι πελάτες που είναι στο στάδιο ζωής διακράτησης και πιο συγκεκριμένα της εξέλιξης, έχουν δηλαδή χαρακτηριστικά που δείχνουν ότι μπορούμε να ενισχύσουμε περαιτέρω τη σχέση της εταιρείας μαζί τους, εφαρμόζοντας "Relational marketing" στρατηγικές.

Αρχικά θα πρέπει να εξετάσουμε για ποιους λόγους δεν έχουν κάνει πρόσφατα κάποια αγορά. Μία στρατηγική θα μπορούσε να είναι να τους προσφέρουμε είτε κάποιες εκπτώσεις/ offers είτε να τους δελεάσουμε με επέκταση εγγύησης των προϊόντων που θα αγοράσουν για τους επόμενους έξι μήνες. Επίσης θα μπορούσαμε να προσφέρουμε κάποια επιπλέον έκπτωση σε παραγγελίες ανταλλακτικών που έχουν και υψηλότερο περιθώριο κέρδους για την επιχείρηση. Οι ενέργειες που αναφέραμε αφορούν στρατηγικές ενεργοποίησης τους εκ νέου.

Μία ακόμη στρατηγική που προτείνουμε να υιοθετήσουμε, είναι η στρατηγική upselling δίνοντας έμφαση στην παροχή πληροφοριών για την υψηλή εξειδίκευσή μας στα ειδικά προϊόντα (make to engineer) προωθώντας λύσεις με υψηλότερο margin για την

επιχείρηση. Αυτός είναι ένας τομέας της εταιρείας που εστιάζει σε προϊόντα υψηλότερου κόστους, ιδιαίτερης αισθητικής όσο αφορά την σχεδίαση, και απευθύνεται σε τελικούς πελάτες που είναι διατεθειμένοι να πληρώσουν παραπάνω για ένα premium προϊόν.

Επίσης, θα μπορούσαν να υιοθετηθούν στρατηγικές cross-selling δηλαδή να τους κάνουν ενημέρωση για την παροχή ανταλλακτικών που χρησιμοποιούν στις συντηρήσεις και δεν συνδέονται άμεσα με την αγορά ενός προϊόντος, με αυτό τον τρόπο τους δίνουμε κίνητρα να προχωρήσουν ταυτόχρονα σε παραγγελία ανταλλακτικών μαζί με την αγορά του προϊόντος.

Τέλος, θα μπορούσε να τρέξει μία καμπάνια που θα ενισχύσει την πιστότητα αυτών των πελατών. Μέσω του προγράμματος membership που αναφέραμε και πιο πάνω, θα μπορούσαμε να τους ωθήσουμε να παραμείνουν πιστοί στην συνεργασία μας, δίνοντας τους πρόσβαση σε έξτρα προνόμια όταν πραγματοποιούν συχνά αγορές, ξεκλειδώνοντας επιμέρους εκπτώσεις σε αυτή την περίπτωση, αλλά και να τους δώσουμε την δυνατότητα έκθεσης τους σε ευρύτερο αγοραστικό κοινό μέσω του δικτύου προβολής της εταιρείας μας, κάτι το οποίο θα τους βοηθήσει να αναπτύξουν περαιτέρω τον κύκλο εργασιών τους.

- ***Cluster (2) - Πελάτες με υψηλά R, F, M (Best Customers).***

Προτεινόμενη στρατηγική: Αυτοί οι πελάτες χαρακτηρίζονται ως “Best Customers”. Θέλουμε να διατηρήσουμε αυτούς τους πελάτες και να εξελίξουμε περαιτέρω τη συνεργασία μας μαζί τους. Όπως είδαμε και στη θεωρεία, είναι πελάτες που βρίσκονται στο στάδιο ζωής της διακράτησης. Επειδή είναι πελάτες υψηλής αξίας, θα πρέπει να υιοθετηθούν στρατηγικές “Personalization” σε συνδυασμό με “Relational marketing”.

Πιο συγκεκριμένα, προτείνουμε οι account managers των πελατών αυτών να έχουν μία προσωποποιημένη επικοινωνία μαζί τους, με τακτικές επισκέψεις στις εγκαταστάσεις τους. Επίσης θα μπορούσαμε να επικοινωνούμε μαζί τους σε περιπτώσεις που δεν σχετίζονται άμεσα με την πώληση προϊόντων όπως γενέθλια, στέλνοντας κάποια δωρεάν δείγματα και δώρα, κάνοντας τους να αισθάνονται ότι η συνεισφορά τους στην σχέση μας μαζί τους αναγνωρίζεται και αποκτά και μία πιο προσωπική διάσταση.

Είναι η κατάλληλη ομάδα πελατών για να εστιάσει το τμήμα customer experience ζητώντας πολύ τακτικά ανατροφοδότηση που σχετίζεται με την εμπειρία πώλησης που απολαμβάνουν και αποζητώντας τη συνεχή βελτίωση μας σε όλα τα touchpoints που

έχουν κατά τη διάρκεια της παραγγελίας. Η πώληση δεν πρέπει να θεωρείται για αυτούς τους πελάτες ένα κομμάτι μίας τυπικής διαδικασίας αλλά μία εμπειρία.

Μέσω του membership προγράμματος που προτείνουμε, πέρα από τις εκπτώσεις και τις προσωποποιημένες προσφορές που θα τους κάνουμε, προτείνουμε να τους δίνουμε πρόσβαση σε κάποια workshops που θα οργανώνουμε ανά τακτικά διαστήματα ώστε να είναι πλήρως ενημερωμένοι για τα προϊόντα μας αλλά και να τους παρέχουμε προσωποποιημένη συνεχή εκπαίδευση που αφορά την εγκατάσταση του προϊόντος μας.

Όσο αφορά την τεχνική υποστήριξη τους, θα πρέπει να δίνεται πρόσβαση στην 24/7 τηλεφωνική γραμμή βοήθειας που έχουμε χωρίς επιπλέον χρέωση, ώστε να μπορούμε να τους υποστηρίξουμε σε περίπτωση προβλήματος κατά την εγκατάσταση, αρχικά τηλεφωνικά, χωρίς να υπάρχει περιορισμός σε ποιο σημείο του κόσμου βρίσκονται με την διαφορά ώρας. Σε περίπτωση που προκύψει κάποιο πρόβλημα που δεν μπορεί να λυθεί απομακρυσμένα κατά την εγκατάσταση προτείνουμε να δίνουμε ως έξτρα παροχή τη δωρεάν αποστολή δικού μας συνεργείου χωρίς χρέωση, για έναν συγκεκριμένο αριθμό περιπτώσεων ανά έτος.

Επίσης θα πρέπει να λαμβάνουμε υπόψιν τα σχόλια τους όχι μόνο για την εμπειρία που έχουν κατά την πώληση αλλά και για την ποιότητα των προϊόντων μας, βρίσκοντας τρόπους είτε να αντιμετωπίσουμε κάποιες αδυναμίες μας έγκαιρα, είτε αναγνωρίζοντας μελλοντικές ανάγκες – τάσεις της αγοράς που θα μπορούσαμε να δώσουμε στο τμήμα έρευνας και ανάπτυξης νέων προϊόντων της εταιρείας, ώστε να εστιάσουν με τη σειρά τους στα σωστά σημεία για νέα προϊόντα.

Προτείνουμε να τους ενημερώνουμε έγκαιρα για τα νέα προϊόντα που θέλουμε να εισάγουμε εφόσον αυτή η ομάδα πελατών είναι πιθανότερο να περιέχει early adopters. Με αυτό τον τρόπο θα μπορούσαμε να τους δώσουμε την αίσθηση ότι έχουν πρόσβαση σε κάποια προϊόντα πριν από τους υπόλοιπους πελάτες. Επίσης για τα προϊόντα αυτά προτείνουμε να τους στέλνουμε χωρίς χρέωση κάποιο εξειδικευμένο συνεργείο μας που θα τους βοηθάει κατά την εγκατάσταση του νέου προϊόντος.

Τέλος, είναι η ιδανική ομάδα πελατών που θα μπορούσαμε να φιλοξενήσουμε δηλώσεις τους (testimonials) για την μεταξύ μας συνεργασία στην ηλεκτρονική σελίδα της εταιρείας αλλά και στα μέσα κοινωνικής δικτύωσης.

- **Συζήτηση**

Εν κατακλείδι, είδαμε ότι δεν υπάρχουν στην βιβλιογραφία έρευνες που να εστιάζουν στην ομαδοποίηση πελατών με RFM και Cluster analysis για βιομηχανικές εταιρείες πάνω στο κομμάτι του ανελκυστήρα, κάτι το οποίο ήρθε να καλύψει η συγκεκριμένη έρευνα.

Συγκρίνοντας τα αποτελέσματα με παρόμοιες έρευνες για άλλους κλάδους που είδαμε ενδεικτικά και στο κεφάλαιο 3.4 παρατηρούμε ότι έχουμε παρόμοια αποτελέσματα με τη διαφορά ότι έχουμε πιο συμπαγείς και ποιοτικές ομάδες με βάση τα αποτελέσματα του Silhouette score, κάτι που οφείλεται στην ποιότητα των δεδομένων που χρησιμοποιήσαμε αλλά και στο μεγαλύτερο χρονικό εύρος που έχουμε επιλέξει να εξετάσουμε.

Επίσης, αναγνωρίσαμε έναν αστάθμητο παράγοντα που επηρέασε την δραστηριότητα των πελατών, την πανδημία covid-19, που οδήγησε πολλές επιχειρήσεις – πελάτες να σταματήσουν τη δραστηριότητα τους ανάλογα με την εξέλιξη των κρουσμάτων και τα μέτρα που επέλεξε να πάρει κάθε χώρα για την αντιμετώπιση της. Αυτό είχε ως αντίκτυπο να αλλάξουν ξαφνικά οι αγοραστικές συνήθειες κάποιων πελατών, ιδίως από συγκεκριμένες χώρες που αναγνωρίσαμε, με αποτέλεσμα στην ομαδοποίηση που προέκυψε να εμφανίζονται σε μικρό βαθμό κάποιες αλληλοκαλύψεις μεταξύ των cluster 0 και cluster 1. Σε αυτό το στάδιο είδαμε ότι υπάρχει συσχετισμός μεταξύ των ομάδων που παρατηρήσαμε ανωμαλίες και των χωρών που με το δείκτη μέτρων αυστηρότητας που χρησιμοποιήσαμε (Hale, et al., 2023).

# 7 Περιορισμοί – Προτάσεις για μελλοντική έρευνα

Ένας από τους σημαντικότερους περιορισμούς της παρούσας έρευνας ήταν η συγκυρία που υπήρξε με την πανδημία, κάτι που προφανώς επηρέασε τις καταναλωτικές συνήθειες των πελατών. Αυτό είχε ως αντίκτυπο την παύση ή επιβράδυνση των αγορών σε χώρες που είχαν επιβληθεί αυστηρά lock down ανάλογα με την εξέλιξη της κάθε φάσης της πανδημίας σε κάθε γεωγραφική ζώνη. Το φαινόμενο αυτό, δεν παρουσίαζε ομοιογένεια παγκοσμίως κατά τις ίδιες χρονικές περιόδους, αφού σε κάθε χώρα ο τρόπο προσέγγισης για τα μέτρα αντιμετώπισης της πανδημίας είχε πολύ μεγάλη εξάρτηση από την κυβέρνηση της κάθε χώρας.

Μία πρόταση για μελλοντική έρευνα είναι να επαναληφθεί η ίδια διαδικασία μετά την πανδημία, αφού συμπληρωθεί ένα εύλογο χρονικό διάστημα ανάλογο με το διάστημα που επιλέξαμε για την έρευνα που έχει παρουσιαστεί, και να συγκριθούν τα αποτελέσματα ώστε να δούμε αν οι ομάδες που προκύπτουν θα είναι περισσότερο συμπαγείς, εφόσον δεν θα υπάρχουν τέτοιου είδους εξωγενείς χαρακτηριστικά όπως είχαμε με την συγκυρία της πανδημίας που ενδέχεται να δημιούργησε κάποιες στρεβλώσεις.

Μία άλλη πρόταση για μελλοντική έρευνα είναι να εστιάσουμε στους πελάτες χαμηλής δυναμικότητας και να γίνει συνδυασμός με μία νέα μελέτη που εστιάζει στους λόγους που οδηγούν έναν πελάτη να φύγει από την εταιρεία, δηλαδή στο κομμάτι του customer churn. Με αυτό τον τρόπο αν αναπτυχθεί ένα μοντέλο που μπορεί να προβλέπει το customer churn η εταιρεία θα μπορούσε για το προφίλ των πελατών αυτών, να αναπροσαρμόζει το πλάνο μάρκετινγκ και να μπορεί να διατηρεί την πελατειακή της βάση χωρίς να χάνει πελάτες, ιδίως αν πρόκειται για πελάτες αξίας.

# 8 Βιβλιογραφία


Aghabozorgi, S., Shirkhorshidi, A. S. & Wah, T. Y., 2015. Time-series clustering – A decade review. *Information Systems,* Τόμος 53, pp. 16-38.

Ahmad, A. & Khan, S., 2019. Survey of State-of-the-Art Mixed Data. *Institute of Electrical and Electronics Engineers (IEEE),* Τόμος 7, pp. 31883-31902.

Ahmad, A. & Dey, L., 2011. Ak-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets. *Pattern Recognit.,* 32(7), pp. 1062-1069.

Al-Jabery, K., Obafemi-Ajayi, T., Olbricht, G. & Wunsch, D., 2019. *Computational learning approaches to data analytics in biomedical applications.* 1 επιμ. s.l.:Academic Press.

Ameliorate, 2019. *Global Elevators Market Report 2019,* Waxhaw NC: Market Insights Reports.

Anitha, P. & Malini, M., 2019. RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences,* Τόμος 34, p. 1785–1792.

Benabdellah, A. C., Benghabrit, A. & Bouhaddou, I., 2019. A survey of clustering algorithms for an industrial context. *Procedia Computer Science,* Τόμος 148, pp. 291-302.

Bezdek, J. C., 2013. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Berlin: Springer.

Birant, D., 2011. *Data Mining Using RFM Analysis, Knowledge - Oriented Applications in Data Mining.* Rijeka: InTech.

Bleier, A., De Keyser, A. & Verleye, K., 2018. *Customer Engagement Through Personalization and Customization.* 1 επιμ. s.l.:Palgrave Macmillan.

Chiş, M., Banerjee, S. & Hassanien, A., 2009. Clustering Time Series Data: An Evolutionary Approach. *Foundations of Computational Intelligence,* Τόμος 206, p. 193–207.

Christy, J., Umamakeswari, A., Priyatharsini, L. & Neyaa, A., 2021. RFM Ranking - An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences,* 33(10), pp. 1251-1257.

Daabes, A. & Kharbat, F., 2017. Customer-based perceptual map as a marketing intelligence source. *International Journal of Economics and Business Research,* 13(4), p. 360–379.

David, G. & Averbuch, A., 2012. SpectralCAT: Categorical spectral clustering of numerical and nominal data. *Pattern Recognit.,* 45(1), pp. 416-433.

Du, K.-L., 2010. Clustering: A neural network approach. *Neural Networks,* 23(1), pp. 89-107.

Ezugwu, A. και συν., 2022. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence,* 110(104743), pp. 1-43.

Financial Times Journalism, F. V. &. D., 2023. *Financial Times.* [Ηλεκτρονικό]
Available at: https://ig.ft.com/coronavirus-lockdowns/
[Πρόσβαση June 2023].

Gustriansyah, R., Suhandi, N. & Antony, F., 2020. Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science,* 18(1), pp. 470-477.

Hale, T. και συν., 2023. *Variation in government responses to COVID-19.* [Ηλεκτρονικό] Available at: https://www.bsg.ox.ac.uk/sites/default/files/2023-06/BSG-WP-2020-032-v15.pdf [Πρόσβαση June 2023].

He, X. & Li, C., 2016. *The research and application of customer segmentation on e-commerce websites.* Guangzhou, 6th International Conference on Digital Home (ICDH), pp. 203-208.

Huang, M. & Rust, R. T., 2021. A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science,* Τόμος 49, pp. 30-50.

Huang, S., Chang, E. & Wu, H., 2009. A case study of applying data mining techniques in an outfitter's customer value analysis. *Expert Systems with Applications*, Issue 36, pp. 5909-5915.

Hung, C. & Tsai, C., 2008. Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand. *Expert Systems with Applications*, Issue 34, pp. 780-787.

Jiang, T. & Tuzhilin, A., 2009. Improving Personalization solutions through optimal segmentation of customer bases. *IEEE Trans. Knowledge Data Eng.,* 21(3), pp. 305-320.

Khan, S. & Ahmad, A., 2004. Cluster center initialization algorithm. *Pattern Recognition Letters,* 25(11), p. 1293–1302.

Kokate, U., Deshpande, A., Mahalle, P. & Patil, P., 2018. Data Stream Clustering Techniques, Applications, and Models: Comparative Analysis and Discussion. *Big data and cognitive computing,* 2(32), pp. 1-30.

Kotler, P. & Armstrong , G., 2006. *Principles of Marketing.* Eleventh Edition επιμ. New Jersey: Pearson Prentice Hall.

Lam, D., Wei, M. & Wunsch, D., 2015. Clustering Data of Mixed Categorical and Numerical Type with Unsupervised Feature Learning. *IEEE Access,* 3(2477216), pp. 1605-1613.

Luo, H., Kong, F. & Li, Y., 2006. *Advanced Data Mining and Applications, Second International Conference.* Xi'an, Springer.

Mansalis, S., Ntoutsi, E., Pelekis, N. & Theodoridis, Y., 2018. An evaluation of data stream clustering algorithms. *Statistical analysis and data mining,* 11(4), pp. 167-187.

Marcus, C., 1998. A practical yet meaningful approach to customer segmentation. *JOURNAL OF CONSUMER MARKETING*, Issue 15, pp. 494-504.

McCarty & Hastak, 2007. Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, pp. 656-662.

Melnykov, V. & Maitra, R., 2010. Finite mixture models and model-based. *Statistics Surveys,* Τόμος 4, p. 80–116.

Miglautsch, 2000. Thoughts on RFM scoring. *Journal of Database Marketing & Customer Strategy Management*, Issue 8, pp. 67-72.

Miglautsch, 2002. Application of RFM principles: What to do with 1-1-1 customers?. *The Journal of Database Marketing*, Issue 9, pp. 319-324.

Rathore, P., 2018. *Big Data Cluster Analysis and ots Applications (Doctor of Philosophy).* Melbourne: University of Melbourne.

Rodriguez, A. & Laio, A., 2014. *Science,* 344(6191), pp. 1492-1496.

Sasaki, H. και συν., 2018. Mode-Seeking Clustering and Density Ridge Estimation via Direct Estimation of Density-Derivative-Ratios. *Journal of Machine Learning Research,* Τόμος 18, pp. 1-47.

Saxena, A. και συν., 2017. A review of clustering techniques and developments. *Neurocomputing,* Τόμος 267, pp. 664-681.

Shah, S. & Singh, M., 2012. *Comparison of a Time efficient Modified K-Mean algorithm with K-Mean and K-Medoid algorithm.* Rajkot, Conference on Communication Systems and Network Technologies , pp. 435-437.

Sheshasaayee, A. & Logeshwari, L., 2017. *An efficiency analysis on the TPA clustering methods for intelligent customer segmentation.* Bangalore, International Conference of innovative mechanisms for industry applications (ICIMIA) , pp. 784-788.

Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y. & Herawan, T., 2014. *Big Data Clustering: A Review.* s.l., Springer International Publishing Switzerland, p. 707–720.

Simester, D., Timoshenko, A. & Zoumpoulis, S., 2020. Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Science,* 66(6), p. 2495–2522.

Suh, E., Noh, S. & Suh, C., 1999. Customer list segmentation using the combined response model. *Expert Systems with Applications*, pp. 89-97.

Wang, 2010. Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques. *Expert Systems with Applications*, Issue 37, pp. 8395-8400.

Wang, Y., Ramanan, D. & Hebert, M., 2017. *Learning to model the tail. 31st conference on neural information processing systems.* Long Beach, CA, USA, Advances in Neural Information Processing Systems.

Wang, Z. και συν., 2016. *Learning a task-specific deep architecture for clustering.* s.l., Society for Industrial and Applied Mathematics, p. 369–377.

Witten, I. H. & Frank, E., 2005. *Data Mining: Practical Machine Learning.* 2nd επιμ. San Mateo, CA, USA: Morgan Kaufmann.

Xie, H. και συν., 2019. Improving K-means clustering with enhanced Firefly Algorithms. *Applied Soft Computing,* 84(105763), pp. 1-22.

Xu, R. & Wunsch, D., 2005. Survey of clustering algorithms. *IEEE Transactions on Neural Networks,* 16(3), pp. 645-678.

Yeh, I., Yang, K. & Ting, T., 2008. Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, Issue 36, pp. 5866-5871.

Zahrotun, L., 2017. *Implementation of data mining technique for customer customer relationship management (CRM) on online shops with fuzzy c-means clustering.* Yogyakarta, 2nd International Conferences on Information Technology, Information Systems & Electrical Engineering (ICITISEE), pp. 299-303.