Πρόγραμμα Μεταπτυχιακών Σπουδών

στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

<u>Διπλωματική Εργασία</u>

Θέμα: «Ανάλυση καταναλωτικής συμπεριφοράς και εφαρμογή κατάλληλης στρατηγικής μάρκετινγκ με τη χρήση της ανάλυσης RFM και του μοντέλου μηχανικής μάθησης kmeans clustering»

της

Λαζαρίδου Βασιλικής του Γρηγορίου

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

Ιούλιος 2023

Ευχαριστίες

Η παρούσα διπλωματική εργασία πραγματοποιήθηκε στα πλαίσια του Προγράμματος Μεταπτυχιακών Σπουδών στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων του τμήματος Οργάνωσης και Διοίκησης Επιχειρήσεων του Πανεπιστημίου Μακεδονίας. Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κύριο Λεωνίδα Χατζηθωμά για τη συνεργασία και την πολύτιμη βοήθεια και καθοδήγησή του καθώς επίσης και την ευελιξία που μου προσέφερε κατά τη συγγραφή της διπλωματικής εργασίας. Επιπλέον, θα ήθελα να ευχαριστήσω τον καθηγητή μου, κύριο Γεώργιο Παλταγιάν για την βοήθεια, τη στήριξη και τις συμβουλές που μου προσέφερε κατά την διάρκεια της φοίτησής μου αλλά και της συγγραφής της εργασίας.

Πίνακας Περιεχομένων

Πίνακας Εικόνων

Πίνακας Διαγραμμάτων

Περίληψη

Στο σύγχρονο περιβάλλον, οι επιχειρήσεις και οι επαγγελματίες του μάρκετινγκ έχουν πλέον στη διάθεσή τους τεράστιες ποσότητες δεδομένων σχετικά με τους πελάτες τους και οι προσπάθειες που πραγματοποιούν για να αναλύσουν την αγορά επικεντρώνονται ολοένα και περισσότερο γύρω από την προσπάθεια για αποδοτική ερμηνεία και σωστής αξιοποίησής τους. Η επιτυχία των στρατηγικών και των σχεδίων μάρκετινγκ που προσπαθούν να επιτύχουν οι επιχειρήσεις, καθώς ο αριθμός των πόρων τους είναι περιορισμένος, βασίζεται στη σωστή διαχείριση των δεδομένων, ώστε να εντοπιστούν και να στοχευφθούν τα κατάλληλα και τα πιο επικερδή τμήματα της αγοράς. Μια από τις επικρατέστερες μεθόδους τμηματοποίησης με βάση την αγοραστική συμπεριφορά των πελατών είναι το μοντέλο RFM, το οποίο τμηματοποιεί και κατατάσσει τους πελάτες σε συστάδες (clusters) με βάση τρία χαρακτηριστικά της αγοραστικής τους συμπεριφοράς: Recency (διάστημα από την τελευταία αγορά), Frequency (συχνότητα αγορών) και Monetary (νομισματική αξία των αγορών). Η ανάλυση RFM συχνά χρησιμοποιείται συνδυαστικά με τεχνικές ομαδοποίησης, όπως είναι η ανάλυση k-means, ώστε οι πελάτες της επιχείρησης να ταξινομηθούν σε clusters και στη συνέχεια να αναλυθούν από τα στελέχη μάρκετινγκ του οργανισμού. Στο πρακτικό μέρος της εργασίας, θα πραγματοποιηθεί αντίστοιχη ανάλυση στο dataset μιας επιχείρησης ηλεκτρονικού εμπορίου με είδη δώρων. Από τα αποτελέσματα της έρευνας προέκυψε ότι πιο αποδοτική μέθοδος σχετικά με την αποτελεσματικότερη τμηματοποίηση των πελατών είναι η εφαρμογή του μοντέλου rfm, σύμφωνα με το οποίο τα μεγαλύτερα ποσοστά πελατών συγκεντρώνονται σε συστάδα που περιέχει αποδοτικότερους και μεγαλύτερης αξίας πελάτες. Επίσης, προτείνονται εξατομικευμένες στρατηγικές για κάθε μια συστάδα, με βάση τις ανάγκες τους και τέλος, σχολιάζεται η περίοδος με τα μικρότερα έσοδα κατά τη διάρκεια του έτους, προτείνοντας στην επιχείρηση να εστιάσει στα αίτια που δημιουργούν αυτήν τη χαμηλή καταναλωτική ζήτηση.

Λέξεις Κλειδιά: Ανάλυση RFM, kmeans clustering, data-driven marketing, τμηματοποίηση, καταναλωτική συμπεριφορά, μηχανική μάθηση, αλγόριθμος

Εισαγωγή

Η παρούσα εργασία είναι μια έρευνα εξόρυξης δεδομένων που πραγματεύεται τη μέθοδο τμηματοποίησης πελατών, γνωστή ως ανάλυση RFM και του αλγορίθμου μηχανικής μάθησης, kmeans clustering, με σκοπό να παρέχει στην επιχείρηση, την πληροφόρηση που απαιτείται για να αυξήσει τα κέρδη της και να ενισχύσει την αξία της. Η συγγραφή της εργασίας καλείται να συνεισφέρει ερευνητικά στην υπάρχουσα βιβλιογραφία ως προς τη χρήση της ανάλυσης RFM και συνδυαστικά με την ανάλυση μηχανικής μάθησης (ML), με τη χρήση του αλγορίθμου kmeans clustering σε επιχειρήσεις με είδη δώρων. Επίσης, η έρευνα επιχειρεί να καλύψει το ερευνητικό κενό που υπάρχει σχετικά με την σύγκριση των δυο μεθόδων και την επιλογή της πιο συμφέρουσας εξ' αυτών, για επιχειρήσεις που επιθυμούν να πραγματοποιήσουν τμηματοποίηση αγοράς και ανάλυση καταναλωτικής συμπεριφοράς. Τα αποτελέσματα της έρευνας ανέδειξαν την ανάγκη για περαιτέρω διερεύνηση συγκεκριμένης αγοράς και την ανάλυση του καλαθιού για να μπορέσει να αντιμετωπιστεί και το φαινόμενο του customer churn που πιθανόν να υπάρχει.

Επιτυγχάνοντας τη στοχευμένη προσέγγιση του καταναλωτικού κοινού, από την εφαρμογή της ανάλυσης, οι επαγγελματίες του μάρκετινγκ εφοδιάζονται με σημαντικές πληροφορίες που προσφέρουν αξία στην επιχείρηση. Αυτό, βοηθάει την εταιρεία να αναλύσει την καταναλωτική συμπεριφορά και να εστιάσει στα κατάλληλα τμήματα, με σωστά προϊόντα και κατάλληλες προωθητικές ενέργειες, εφαρμόζοντας το κατάλληλο στρατηγικό μάρκετινγκ και αξιοποιώντας με βέλτιστο τρόπο τους διαθέσιμους πόρους. Η καταναλωτική συμπεριφορά, ενώ έχει κάνει την εμφάνισή της από τη δεκαετία του 1960, αρχικά ως μια μονοδιάστατη έννοια, τα σύγχρονα χρόνια, έχει αποκτήσει μια πιο ευρεία και πολυεπίπεδη έννοια, που εκτός από τη συμπεριφορά του καταναλωτή για το προϊόν και τη διαδικασία λήψης απόφασης μελετά και αναλύει και τη συμπεριφορά του πριν και μετά την ολοκλήρωση της αγοράς καθώς και τις ενέργειες και αξιολογήσεις στις οποίες προβαίνει ο καταναλωτής. Από την άλλη πλευρά, η κατάτμηση της αγοράς, αποτελεί βασικό συστατικό για την εφαρμογή του μίγματος μάρκετινγκ καθώς και τις στρατηγικές ενέργειες που επιθυμεί να εφαρμόσει μια επιχείρηση ανάλογα με τους στόχους της.

Στόχος της εργασίας είναι να αναλυθεί η πελατειακή βάση που μελετάται, ώστε να διαχωριστεί το σύνολο των πελατών και να δημιουργηθούν ομάδες με διαφορετικά γνωρίσματα για να έχουν τη δυνατότητα οι επαγγελματίες του μάρκετινγκ να εντοπίζουν τις πιο κερδοφόρες ομάδες που πρέπει να επενδύσουν και να διαμορφώνουν τις κατάλληλες στρατηγικές μάρκετινγκ. Η τμηματοποίηση της αγοράς, η οποία αποτελεί μια από τις βασικότερες διαδικασίες του μάρκετινγκ, είναι εκείνη η διαδικασία που επιτρέπει στις επιχειρήσεις να πραγματοποιήσουν το διαχωρισμό αυτό και να εντοπίσουν τις αγορές-στόχους, στις οποίες πρέπει να εστιάσουν και να διαθέσουν πόρους για να υπάρχει καλύτερη απόδοση της επιχείρησης και εξατομικευμένη προσέγγιση των πελατών. Από την άλλη πλευρά, η ανάλυση της καταναλωτικής συμπεριφοράς, είναι το μέσο με το οποίο μπορούν να αναγνωριστούν οι πραγματικές ανάγκες-επιθυμίες της κάθε υπο-ομάδας που δημιουργείται και να διαμορφωθούν εξατομικευμένες στρατηγικές και ενέργειες μάρκετινγκ για την στοχευμένη προσέγγιση του καταναλωτικού κοινού. Στη συγκεκριμένη εργασία, αυτοί οι στόχοι επιτυγχάνονται με τη βοήθεια της ανάλυσης RFM και του αλγορίθμου ομαδοποίησης kmeans, ώστε να συγκριθούν τα αποτελέσματα των δυο αναλύσεων και να επιλεγεί η πιο συμφέρουσα μέθοδος που δίνει τη μεγαλύτερη αξία στις επιχειρήσεις.

Τα τελευταία χρόνια, καθώς ο επιχειρηματικός κλάδος εξελίσσεται με μεγάλους ρυθμούς, τα επίπεδα του ανταγωνισμού ολοένα και αυξάνονται. Οι επιχειρήσεις θέτουν ως πρωταρχικό μέλημά τους τις ανάγκες των καταναλωτών και την καλύτερη εξυπηρέτησή τους με εξατομικευμένο τρόπο. Για να μπορούν οι επιχειρήσεις να πετυχαίνουν ολοένα και μεγαλύτερα κέρδη, επιδιώκουν να έχουν συνεχώς ικανοποιημένους πελάτες, αποκωδικοποιώντας τις ανάγκες και τις επιθυμίες τους. Είναι φανερό ότι τα τελευταία χρόνια, ολοένα και περισσότερες επιχειρήσεις στρέφουν τις προσπάθειές τους, στην κατάτμηση της αγοράς, ώστε να αναγνωρίζουν την αξία της κάθε ομάδας για να εξυπηρετούν καλύτερα το καταναλωτικό τους κοινό. Παράλληλα, η μελέτη και η ανάλυση της συμπεριφοράς των καταναλωτών, αποτελεί ένα πολύ βασικό στοιχείο για τις επιχειρήσεις και έχει μεγάλη σημασία για το Μάρκετινγκ, καθώς συμβάλλει ουσιαστικά στην επιλογή της κατάλληλης στρατηγικής, πετυχαίνοντας μια καλύτερη θέση στην αγορά. Έτσι, ο κύριος στόχος που προσπαθούν να επιτύχουν οι επαγγελματίες μάρκετινγκ είναι να διασφαλίσουν την ανάπτυξη και την επιβίωση της επιχείρησης, εντοπίζοντας τις καταναλωτικές απαιτήσεις και ικανοποιώντας τους

πελάτες με το βέλτιστο δυνατό τρόπο, διατηρώντας παράλληλα το ανταγωνιστικό πλεονέκτημα.

Στην σύγχρονη εποχή, ο τεράστιος όγκος δεδομένων που είναι διαθέσιμος σε ότι αφορά τους πελάτες και την καταναλωτική συμπεριφορά τους, δίνει την δυνατότητα στις επιχειρήσεις να αξιοποιήσουν αυτά τα δεδομένα με τον καλύτερο τρόπο έτσι ώστε να προσεγγίσουν τα επιθυμητά τμήματα πελάτων εφαρμόζοντας εξατομικευμένες στρατηγικές προώθησης. Η σημαντικότητα της τμηματοποίησης της αγοράς για τις επιχειρήσεις καθώς και για τη λειτουργία του μάρκετινγκ είναι μεγάλη, καθώς παραχωρεί στους οργανισμούς τη δυνατότητα να ανιχνεύουν τμήματα της αγοράς - στόχου με όμοια γνωρίσματα ή τα πιο επικερδή τμήματα της εταιρείας και να διαμορφώνουν τις στρατηγικές που ταιριάζουν στις αντίστοιχες ανάγκες τους.

Το μοντέλο RFM έκανε την εμφάνισή του τη δεκαετία του 1994 από τον Hughes και από τότε θεωρείται μια από τις πιο διαδεδομένες και αποτελεσματικές τεχνικές τμηματοποίησης πελατών με βάση την αξία τους για την επιχείρηση. Τα τελευταία χρόνια, το μοντέλο συνδέεται στενά με την έρευνα μάρκετινγκ, καθώς ενισχύει τα κέρδη μιας εταιρείας, αποτελεί πηγή πληροφοριών, εντοπίζοντας τμήματα με αυξημένη ζήτηση, παρέχει πληροφορίες για το ιστορικό συναλλαγών των πελατών, απαντάει σε σημαντικά ερωτήματα για τη διατήρηση και ανάπτυξη της επιχείρησης και συμβάλλει στην επιλογή εξατομικευμένης στρατηγικής. Ωστόσο, παρά τα πολλά οφέλη που προσφέρει η εφαρμογή του μοντέλου, παρουσιάζει και αρκετά μειονεκτήματα που προβληματίζουν τους επαγγελματίες μάρκετινγκ σχετικά με την εφαρμογή της. Ένα από τα πιο βασικά, είναι η επικέντρωση μόνο στους καλύτερους πελάτες. Πολλές φορές εφαρμόζεται συνδυαστικά και με άλλες αναλύσεις ή αλγορίθμους για να επιτευχθούν αποτελεσματικότερες και πιο στοχευμένες προσεγγίσεις, καλύτερα αποτελέσματα και περισσότερες πληροφορίες.

Η δομή της εργασίας οργανώνεται ως εξής: στην πρώτη ενότητα της παρούσας εργασίας, παρουσιάζεται το εισαγωγικό κομμάτι της, ώστε να μπορέσει να εισάγει τον αναγνώστη στο θέμα και να τον βοηθήσει να κατανοήσει τα βασικότερα σημεία της έρευνας. Παράλληλα παρουσιάζεται το αντικείμενο μελέτης και ο σκοπός της εργασίας. Στο πρώτο κεφάλαιο, γίνεται γνωστός ο ρόλος και η έννοια του μάρκετινγκ και αναπτύσσεται η έννοια και η χρησιμότητα του μάρκετινγκ με γνώμονα τα δεδομένα, τα προβλήματα

που αντιμετωπίζονται και η στάση των επιχειρήσεων στον σημερινή εποχή. Επιπλέον, παρουσιάζονται τομείς που ενδέχεται να εμφανίσουν αυξημένο ενδιαφέρον για το φαινόμενο του data-driven τα επόμενα χρόνια.

Στο δεύτερο κεφάλαιο παρουσιάζεται το θεωρητικό υπόβαθρο για την τμηματοποίηση των πελάτων, τη χρησιμότητα αυτής της τεχνικής και αναφέρονται λεπτομερώς οι τύποι τμηματοποίησης. Παρακάτω, αναλύεται το μοντέλο RFM, η χρησιμότητα και η εφαρμογή του, καθώς επίσης και τα πλεονεκτήματα από τη χρήση του. Επιπλέον, αναλύεται η διαχείριση των αποτελεσμάτων του μοντέλου και παρουσιάζεται ο συνδυασμός του με τεχνικές και αλγορίθμους ομαδοποίησης (clustering). Ακολούθως, παρουσιάζεται η έννοια και η συμβολή της μηχανικής μάθησης καθώς και η μέθοδος της συσταδοποίησης που βοηθάει στην κατάτμηση των καταναλωτών. Τέλος, αναλύονται συνοπτικά οι σημαντικότεροι αλγόριθμοι μηχανικής μάθησης, με έμφαση στον αλγόριθμο k-means clustering, οι οποίοι συχνά συνδυάζονται με την ανάλυση RFM στην εφαρμογή αλλά και στις ερευνητικές εργασίες.

Στο τρίτο κεφάλαιο, το οποίο αποτελεί μαζί με τα επόμενα κεφάλαια το πρακτικό μέρος της εργασίας, πραγματοποιείται μια ανάλυση dataset μιας επιχείρησης ηλεκτρονικού λιανεμπορίου είδη δώρων με ανάλυση RFM σε συνδυασμό με ανάλυση μηχανικής μάθησης, και συγκεκριμένα με τον αλγόριθμο k-means clustering, με σκοπό μέσα από τα αποτελέσματα να προκύψουν προτάσεις για τις στρατηγικές μάρκετινγκ που θα εφαρμοστούν σε κάθε συστάδα (cluster) πελατών, ώστε να επιτευχθεί η μέγιστη δυνατή αποδοτικότητα των διαθέσιμων πόρων της επιχείρησης και η αποτελεσματικότητα των στρατηγικών μάρκετινγκ. Επιπλέον, υπάρχει περιγραφική ανάλυση που αφορά τον καθαρισμό και την περιγραφή των δεδομένων που χρησιμοποιήθηκαν, τη μεθοδολογία και ολόκληρη τη διαδικασία εφαρμογής του μοντέλου RFM αλλά και του αλγορίθμου kmeans clustering. Στο τέταρτο κεφάλαιο εμφανίζονται τα αποτελέσματα από την ανάλυση που πραγματοποιήθηκε με την RFM ανάλυση και τον αλγόριθμο μηχανική μάθησης, kmeans clustering, στο πέμπτο παρατίθενται τα συμπεράσματα που προέκυψαν από τις αναλύσεις που διεξάχθηκαν και προτείνονται κατάλληλες στρατηγικές μάρκετινγκ με στοχευμένες προωθητικές ενέργειες για κάθε συστάδα προσαρμοσμένες σύμφωνα με τις ανάγκες τους. Τέλος, στο έκτο κεφάλαιο παρατίθενται προτάσεις για μελλοντικές διερευνήσεις που μπορούν να πραγματοποιηθούν και να ενισχύσουν ακόμη περισσότερο την ανάπτυξη της επιχείρησης και τα κέρδη της.

Η διαδικασία για την ανάλυση των δεδομένων πραγματοποιήθηκε βάσει 7 βημάτων. Αρχικά, εφαρμόστηκε περιγραφική ανάλυση δεδομένων, η οποία πραγματοποιείται για την κατανόηση του dataset με όλες τις μεταβλητές, την άντληση απαραίτητων πληροφορίων και τον έλεγχο στατιστικών στοιχείων με σκοπό να υλοποιηθούν ο απαραίτητος καθαρισμός και οι κατάλληλες ενέργειες σε κάθε μεταβλητή.

Στη συνέχεια (βήμα 2), πραγματοποιείται καθαρισμός δεδομένων ώστε να αφαιρεθεί ο «θόρυβος» από τα δεδομένα και να αντληθούν πιο αξιόπιστες πληροφορίες για την εξαγωγή των αποτελεσμάτων. Επιπλέον, δημιουργούνται διαγράμματα, για να απεικονιστούν και να κατανοηθούν καλύτερα τα δεδομένα και βοηθητικές μεταβλητές, ώστε να αποθηκευτούν συγκεκριμένες τιμές για τη δημιουργία των διαγραμμάτων.

Ακολούθως, στο βήμα 3 διενεργείται διερευνητική ανάλυση δεδομένων συλλέγοντας πληροφορίες για συγκεκριμένες μεταβλητές όπως και επίσης και ομαδοποίηση μεταβλητών, μέσω συνάρτησης, με σκοπό να δημιουργηθούν τα επιδιωκόμενα αποτελέσματα. Ταυτόχρονα, περνώντας στο βήμα 4 δημιουργούνται διαγράμματα ώστε να μελετηθούν οι απεικονίσεις, να αντληθούν οι απαραίτητες πληροφορίες και να κατανοηθούν πιθανά μοτίβα ή τάσεις για την εξαγωγή των αποτελεσμάτων.

Κατόπιν, ακολουθεί το βήμα 5 όπου εφαρμόζεται η ανάλυση RFM, με σκοπό να αναλυθεί η συμπεριφορά των πελατών και να πραγματοποιηθεί τμηματοποίηση της αγοράς, κατηγοριοποιώντας τους πελάτες σε συστάδες με βάση τα χαρακτηριστικά και τη συμπεριφορά τους. Παράλληλα, πραγματοποιούνται διαγραμματικές απεικονίσεις για τη δημιουργία συσχετίσεων μεταξύ των μετρικών.

Στο βήμα 6, πραγματοποιείται εφαρμογή διαδικασιών για τη διαχείριση της λοξότητας των δεδομένων και της διαφορετικής μέσης τιμής και τυπικής απόκλισης, για να μπορέσει να γίνει εφαρμογή του kmeans αλγορίθμου.

Τέλος, στο βήμα 7, πραγματοποιείται εφαρμογή του kmeans αλγορίθμου, ως δεύτερη μέθοδος, για τη δημιουργία clusters, ώστε να συγκριθούν τα αποτελέσματα των δυο αναλύσεων και αργότερα να επιλεγεί η πιο συμφέρουσα μέθοδος για την επιχείρηση. Επιπλέον, εφαρμόζεται η «elbow method», για να βρεθεί ο βέλτιστος αριθμός για τη δημιουργία συστάδων.

Κεφάλαιο 1 : Βιβλιογραφική Ανασκόπηση

1.1. Εισαγωγή

Το μάρκετινγκ ως επιστήμη ορίζεται ως: «*η δραστηριότητα, το σύνολο των θεσμών και οι διαδικασίες για τη δημιουργία, την επικοινωνία, την παράδοση και την ανταλλαγή προσφορών που έχουν αξία για τους πελάτες, τους πελάτες, τους συνεργάτες και την κοινωνία γενικότερα*» (American Marketing Association, 2017). Το επιτυχημένο μάρκετινγκ επιτυγχάνεται με την εφαρμογή του κατάλληλου μείγματος μάρκετινγκ (4P's), πραγματοποιώντας συνδυασμό με τα σωστά προϊόντα (product), στις σωστές τιμές (price), χρησιμοποιώντας τα κατάλληλα κανάλια διανομής (place) και εφαρμόζοντας εξατομικευμένες ενέργειες προώθησης (promotion) των προϊόντων και υπηρεσιών ενός οργανισμού (Perreault et al., 2012).

Στη σημερινή εποχή, το μάρκετινγκ αποτελεί αναπόσπαστο κομμάτι όλων των επιχειρήσεων και ο ρόλος και η σημαντικότητα του είναι καθοριστικοί για την επιτυχία μιας επιχείρησης. Το γεγονός ότι ο τομέας του μάρκετινγκ αυξάνεται συνεχώς με ραγδαίους ρυθμούς, αναγκάζει τους επαγγελματίες του μάρκετινγκ να ακολουθούν στενά τα βήματα του και να προσπαθούν συνεχώς να προσαρμόζονται ώστε να συμβαδίζουν με τις εξελίξεις αυτές, καθώς η ανάπτυξη, η εξέλιξη και η πορεία των επιχειρήσεων βασίζεται σχεδόν εξ 'ολοκλήρου στο μάρκετινγκ (Kotler et al., 2015). Τα τελευταία χρόνια το μάρκετινγκ επικεντρώνεται περισσότερο στην αγορά και στον πελάτη. Ακολουθώντας αυτή την πελατοκεντρική προσέγγιση, οι επιχειρήσεις μελετούν αρχικά εκτενώς τα δεδομένα που διαθέτουν και εντοπίζουν τις απαιτούμενες ανάγκες και επιθυμίες των πελατών και στη συνέχεια προβαίνουν στην παραγωγή και τη διάθεση των αγαθών στο καταναλωτικό κοινό. Με τον τρόπο αυτό, οι επιχειρήσεις εστιάζουν όσο το δυνατόν καλύτερο τον απαιτούμενο στόχο, χρησιμοποιούν αποδοτικά όλους τους πόρους τους, αυξάνουν τα ποσοστά των ικανοποιημένων πελατών και μειώνουν τα κόστη της επιχείρησης (Pride & Ferrell, 2019).

Σύμφωνα με έρευνα που διενεργήθηκε από τους Reichheld & Kenny (1990) σχετικά με την καταναλωτική συμπεριφορά προέκυψε ότι:
i) η προσέλκυση νέου καταναλωτικού κοινού μπορεί να επιβαρύνει οικονομικά την επιχείρηση πέντε φορές περισσότερο απ'ότι η διατήρηση αυτών που υπάρχουν ήδη, ii)

μια επιχείρηση μέσα σε ένα έτος μπορεί να χάσει κατά μέσο όρο το 10% του πελατολογίου της, iii) μετά από κάποιο χρονικό διάστημα μελλοντικά, το κέρδος του καταναλωτή αυξάνει iv) μια ενδεχόμενη μείωση του πελατολογίου μιας επιχείρησης της τάξεως του 5% μπορεί να συμβάλλει σε μια 25%-85% αύξηση των κερδών αυτής, v) όλοι οι πελάτες μιας επιχείρησης δεν μπορεί να είναι ικανοποιημένοι με αυτήν αλλά αυτό συνεπάγεται και ότι δεν υπάρχει δέσμευση ότι αυτοί που είναι ικανοποιημένοι θα συνεχίσουν να είναι πελάτες και μακροπρόθεσμα και τέλος vi) μια ενδεχόμενη μείωση του προσωπικού κατά 10% και μια ενδεχόμενη αύξηση στη διατήρηση των πελάτων παρουσιάζει όμοιο αποτέλεσμα σε ότι αφορά το οικονομικό κομμάτι για την επιχείρηση.

Ο εντοπισμός και η κατανόηση των αναγκών και επιθυμιών των καταναλωτών καθώς και η αποτελεσματική ικανοποίηση τους με παροχή καλύτερων προϊόντων από αυτά των ανταγωνιστών, συμβάλλουν στη δημιουργία ενός επιτυχημένου μάρκετινγκ (Mulvenna et al., 1998). Οι σύγχρονοι καταναλωτές είναι εξελιγμένοι και διαθέτουν γνώση και δύναμη. Το γεγονός ότι η αγοραστική συμπεριφορά των καταναλωτών μεταβάλλεται συνεχώς, απαιτεί την εύρεση του κατάλληλου καταναλωτικού κοινού και την κατανόηση των προτύπων αγοράς τους (Kotler et al., 2015). Λόγω της εμφάνισης ραγδαίων συνεχόμενων τεχνολογικών προόδων, η ψηφιοποίηση και η ανάλυση μάρκετινγκ – ειδικά η εξόρυξη δεδομένων (Data Mining) – αποτελούν ανεκτίμητα εργαλεία για τις επιχειρήσεις και τους διευθυντές μάρκετινγκ και θεωρούνται σημαντικά στοιχεία της έρευνας μάρκετινγκ (Hauser, 2007).

1.2. Data Driven Marketing

Οι επιχειρηματικές αποφάσεις που καλούνται να λάβουν οι επιχειρήσεις και οι υπεύθυνοι λήψης αποφάσεων επηρεάζουν αρκετούς τομείς της επιχείρησης, όπως είναι η απόδοση μιας επένδυσης μέχρι και τη διαχείριση της επωνυμίας (brand) της. Για να μπορέσουν να ληφθούν αυτές οι αποφάσεις, οι επιχειρήσεις στηρίζονται στην άντληση πληροφοριών που πραγματοποιείται από τους τεράστιους όγκους δεδομένων που έχουν πλέον οι επιχειρήσεις στη διάθεσή τους (Grandhi, Patwa & Saleem, 2020). Η ραγδαία αύξηση που υπάρχει τα τελευταία χρόνια στην εισροή των δεδομένων καθώς και οι τεχνολογικές πρόοδοι καθιστούν την εξόρυξη δεδομένων ένα από τα σημαντικότερα εργαλεία της έρευνας μάρκετινγκ. Οι τεράστιοι όγκοι δεδομένων που διαχειρίζονται οι οργανισμοί αποκαλούνται «Μεγάλα Δεδομένα» (Big Data) και με σωστή επεξεργασία και ανάλυση

αποφέρουν αποδοτικότερες επενδύσεις, καλύτερα διαμορφωμένες στρατηγικές και αυξανόμενα κέρδη (Mazzei & Noble, 2017).

Ωστόσο, ένα από τα μεγαλύτερα προβλήματα που αντιμετωπίζουν οι επιχειρήσεις χρησιμοποιώντας τα δεδομένα αυτά είναι να καταφέρουν να διαχειριστούν και να ερμηνεύσουν σωστά τα των μεγάλο όγκο των δεδομένων για να αντλήσουν αξία από αυτά. Οι υπεύθυνοι λήψης αποφάσεων διαθέτουν υπερβολικό χρόνο για να καταφέρουν να λάβουν τις κατάλληλες επιχειρηματικές αποφάσεις που θα δημιουργήσουν αξία για την επιχείρηση. Είναι σημαντικό μια επιχείρηση να χρησιμοποιεί analytics κατά τη διαδικασία λήψης αποφάσεων ώστε να έχει όλα τα απαραίτητα εφόδια για την σωστή ανάλυση και αξιοποίηση αυτών των δεδομένων (Pearson and Wegener, 2013).

Στη σημερινή εποχή, η εξατομικευμένη προσέγγιση και η δημιουργία σχέσεων αλληλεπίδρασης με τους πελάτες, έτσι ώστε να μπορούν οι επιχειρήσεις να αναγνωρίζουν και να κατανοούν τις επιθυμίες και ανάγκες των καταναλωτών είναι δυο από τα πιο σημαντικά πράγματα που επιδιώκει να επιτύχει μια επιχείρηση. Παράλληλα, με την ανάπτυξη και την εφαρμογή μοντέλων, οι επιχειρήσεις και οι υπεύθυνοι μάρκετινγκ προσπαθούν να αντιληφθούν τις αγοραστικές συνήθειες και τις στάσεις των καταναλωτών απέναντι σε προϊόντα, προωθητικές καμπάνιες και πολιτικές μάρκετινγκ και να εφαρμόσουν τις κατάλληλες στρατηγικές μάρκετινγκ. Είναι αρκετά σημαντικό για τους επαγγελματίες του μάρκετινγκ να χρησιμοποιούν τον μεγάλο όγκο των διαθέσιμων δεδομένων για να στοχεύσουν στην ενίσχυση του επιπέδου εμπειρίας τους και στην προσέλκυση των καταναλωτών (Armstrong & Kotler, 2011).

Η διαρκώς αυξανόμενη ζήτηση-ανάγκη για ψηφιοποίηση όλων των τομέων της κοινωνίας καθώς και η συνεχόμενη στροφή των ανθρώπων στον κόσμο του διαδικτύου, λόγω της ραγδαίας τεχνολογικής εξέλιξης, ήταν η αφορμή για εκτεταμένη συγκέντρωση πληροφοριών. Το μάρκετινγκ είναι ένας από τους σημαντικούς τομείς μιας επιχείρησης που εφαρμόζεται η ψηφιοποίηση και μάλιστα με δεδομένα μεγάλου όγκου (Nazarov, 2019). Το data-driven μάρκετινγκ είναι μια διαδικασία, η οποία αναλύοντας μεγάλα δεδομένα, στηρίζεται στις γνώσεις που αποκτά από τη ανάλυση και προβλέπει τις μελλοντικές συμπεριφορές των καταναλωτών, με σκοπό να βελτιώσει τις δραστηριότητες μάρκετινγκ και την απόδοση της εταιρείας (Cohen, 2018). Επίσης, αναγνωρίζει τις αγοραστικές συνήθειες των πελατών και προτείνει εξατομικευμένες

στρατηγικές, με σκοπό να κατανοήσει και να προσεγγίσει καλύτερα τους πελάτες (Mulvenna et al., 1998).

Σύμφωνα με τον ερευνητή Bass (1993) έγινε μια συνειδητή προσπάθεια μέσω ποσοτικών τεχνικών (όπως ο γραμμικός προγραμματισμός και οι μαρκοβιανές αλυσίδες), να μετατραπεί το μάρκετινγκ από περιγραφικός κλάδος σε επιστήμη πρόβλεψης για τη βελτιστοποίηση των διαδικασιών. Επιπλέον, το μάρκετινγκ με γνώμονα τα δεδομένα, ενώ αποτελούσε μια καινοτόμο προσέγγιση, πλέον έχει μετατραπεί σε θεμελιώδες μέρος της στρατηγικής των επιχειρήσεων και της διαφήμισης (Nazarov, 2019).Τελικά, αυτή η προσέγγιση με επίκεντρο τα δεδομένα μπορεί να οδηγήσει τις επιχειρήσεις και τους διευθυντές μάρκετινγκ σε αύξηση της αποδοτικότητας της εταιρείας, σε αποφυγή ή μείωση κόστους, αποτελεσματικότερη προσέλκυση των πελατών και κατάλληλα διαμορφωμένες τεχνικές για επιλογή βιώσιμης στρατηγικής. Έτσι, η ανάλυση δεδομένων μέσω του μάρκετινγκ με γνώμονα τα δεδομένα μπορεί να εφοδιάσει τις εταιρείες με πολύτιμες γνώσεις που θα είναι αποτελεσματικές και κερδοφόρες για αυτές (Higuera, 2016).

Τα τελευταία χρόνια, οι επικοινωνίες του μάρκετινγκ, οι συναλλαγές ακόμα και σχόλια των καταναλωτών πραγματοποιούνται μέσω του διαδικτύου, με αποτέλεσμα να έχει δημιουργηθεί ένας τεράστιος όγκος από αδόμητα δεδομένα (π.χ. μηνύματα κειμένου). Συνδυαστικά με αυτό, η πολύ μεγάλη ανάπτυξη που υπάρχει στα μέσα κοινωνική δικτύωσης και ο προσανατολισμός στην ψηφιοποίηση ενισχύουν την τάση αυτή. Τα μέσα κοινωνική δικτύωσης είναι το μέσο με το οποίο οι επιχειρήσεις αλληλοεπιδρούν με τους πελάτες τους (He et al., 2013). Μερικά εργαλεία κοινωνικών μέσων είναι το Instagram, Facebook, What's App και YouTube και από την εκτεταμένη χρήση τους παράγονται όλο και περισσότερα αδόμητα δεδομένα καθώς στις εφαρμογές αυτές γίνεται αυξημένη χρήση μηνυμάτων κειμένου και βίντεο. Οι νέες τεχνικές πλέον δημιουργούνται από τα δεδομένα και βασίζονται σε επαγωγικές στατιστικές τεχνικές (Sheth & Kellstadt, 2021).

Οι **τομείς** του data driven μάρκετινγκ στους οποίους προβλέπεται ραγδαία ανάπτυξη τα επόμενα χρόνια είναι οι (Sheth & Kellstadt, 2021):

- **Εξόρυξη κειμένου (Text Mining)** (Netzer et al., 2012): Τα μηνύματα κειμένου σε διαδικτυακές πλατφόρμες δημιουργούν μεγάλους όγκους δεδομένων με

αποτέλεσμα να δημιουργηθεί μια νέα τεχνική, η επεξεργασία φυσικής γλώσσας (NLP). Η εφαρμογή της NLP στο μάρκετινγκ με γνώμονα τα δεδομένα είναι ολοένα και περισσότερο αυξημένη παρά το γεγονός ότι προήλθε από τον κλάδο της γλωσσολογίας και της γνωστικής ψυχολογίας. Είναι ένα εργαλείο τυπολογίας και ταξινόμησης και όχι ένα εργαλείο πρόβλεψης. Οι αριθμητικές μετρήσεις, μετά την διαδικασία της ταξινόμησης των λέξεων, δημιουργεί παραμετρικές και μη παραμετρικέ στατιστικές για επαγωγικούς και για προγνωστικούς σκοπούς.

- **Emoji Analytics**: Η χρήση των emoji στα μηνύματα κειμένου αυξάνεται συνεχώς, με αποτέλεσμα να δημιουργούνται με τον τρόπο αυτό δεδομένα που έχουν προγνωστικές δυνάμεις. Αυτό συνεπάγεται ότι, οι γνώσεις και τα συναισθήματα των καταναλωτών που χρησιμοποιούν στοιχεία emoji θα έχουν αυξημένη και ταχεία αποδοχή και χρήση (Shiha & Ayvaz, 2017).
- **Video Analytics**: Ένας άλλος τομέας αυξανόμενου ενδιαφέροντος για τα επόμενα χρόνια είναι τα Video Analytics, των οποίων ο ρυθμός αυξάνεται εκθετικά με τα χρόνια. Η πληροφόρηση και οι δυνατότητες που παρέχουν τα Video Analytics καθιστούν την παρουσία και την εξέλιξή τους απαραίτητο εργαλείο του μάρκετινγκ. Για να μπορέσουν να προσελκυστούν και να διατηρηθούν πελάτες, είναι σημαντικό να γίνεται μέτρηση της ελαστικότητας της επικοινωνίας με τη χρήση βίντεο, ενώ οι δαπάνες μάρκετινγκ μετατίθενται από την έντυπη και την τηλεόραση στα μέσα κοινωνικής δικτύωσης (He, 2013).
- **Βιομετρικές βάσεις δεδομένων**: Ένας άλλος τομέας αυξημένου ενδιαφέροντος μελλοντικά είναι οι βιομετρικές βάσεις δεδομένων, οι ζήτηση και η εφαρμογή των οποίων αυξάνονται εκθετικά τα τελευταία χρόνια, τόσο σε προηγμένες όσο και σε αναδυόμενες οικονομίες. Οι πρώτες εφαρμογές έχουν γίνει ήδη εδώ και κάποια χρόνια σε ορισμένες χώρες όπως η Ινδία, η οποία διαθέτει το βιομετρικό προφίλ πάνω από ένα δισεκατομμύριο ατόμων της χώρας. Επιπλέον, προσφέρουν πρόσβαση στην σύγχρονη τραπεζική, χωρίς εξάρτηση από κάρτες και υπολογιστές, παρέχει κοινωνικά και οικονομικά οφέλη σε άτομα με χαμηλό εισόδημα και η εφαρμογή τους είναι εγκεκριμένη και για τις κινητές συσκευές (Derawi et al., 2010).
- **Αναγνώριση προτύπων**: Σε έναν κόσμο ο οποίος κατακλύζεται από πληροφορίες, η αναγνώριση των προτύπων είναι απαραίτητη. Για το λόγο αυτό, ο εντοπισμός αδύναμων μοτίβων και η πρόβλεψη του μέλλοντος είναι πολύ

σημαντική για την αποδοτικότητα των στρατηγικών μάρκετινγκ και τις τιμολογιακές αλλαγές, τις εισαγωγές νέων προϊόντων, καθώς και την αποτύπωση των δημογραφικών και ψυχογραφικών τάσεων ως μοτίβα αγοραστικής συμπεριφοράς (Sheth & Kellstadt, 2021).

- **Forensic Research**: Αποτελεί ένα πολύ χρήσιμο εργαλείο για τυχόν αντιδικίες όπως η προστασία της πνευματικής ιδιοκτησίας, των εμπορικών σημάτων και άλλων περιουσιακών στοιχείων μάρκετινγκ. Επίσης, έχει σημαντικό ρόλο στην εσωτερική διακυβέρνηση και σε αντιδικίες δημόσιας πολιτικής, όπως παραπλανητικές πρακτικές μάρκετινγκ και η μη συμμόρφωση με τους κανονισμούς. Τέλος, η λειτουργία του μάρκετινγκ είναι ευάλωτη από κυβερνοεπιθέσεις σε πελατειακές βάσεις δεδομένων (Sheth & Kellstadt, 2021).

1.3. CRM

Η διαχείριση πελατειακών σχέσεων (CRM) είναι μια στρατηγική διαχείρισης των σχέσεων αλληλεπίδρασης μεταξύ των επιχειρήσεων και των πελατών. Ουσιαστικά πρόκειται για διαδικασία ενοποίησης των λειτουργιών του μάρκετινγκ, των πωλήσεων και της εξυπηρέτησης πελάτων με σκοπό να αποκτήσει αξία η επιχείρηση και να ελαχιστοποιηθεί το κόστος. Στόχος του CRM είναι η προσέλκυση και διατήρηση των πελατών και η δημιουργία κατάλληλων στρατηγικών με πελατοκεντρικές προσεγγίσεις σκοπεύοντας να προσθέσει αξία στην επιχείρηση και τους πελάτες και να αυξήσει την κερδοφορία αξιοποιώντας αποτελεσματικά την τεχνολογία (Chalmeta, 2006). Σύμφωνα με τους Kracklauer et al. (2004), υπάρχουν τέσσερις διαστάσεις για το CRM, οι οποίες είναι η αναγνώριση των πελατών, η προσέλκυση, η διατήρηση και η ανάπτυξη των πελατών. Η πρώτη διάσταση, είναι εξαιρετικά σημαντική καθώς για να μπορέσει μια επιχείρηση να εφαρμόσει την κατάλληλη στρατηγική που απαιτείται θα πρέπει να αναγνωρίσει και να κατανοήσει πλήρως τους πελάτες της και τις ανάγκες τους μέσω διαδικασίας της τμηματοποίησης.

Στη σημερινή εποχή, η εισβολή των μέσων κοινωνικής δικτύωσης στις ζωές των καταναλωτών και οι συνεχόμενες τεχνολογικές εξελίξεις οδήγησαν τους επαγγελματίες μάρκετινγκ στην διαρκή προσαρμογή στα νέα δεδομένα που προκύπτουν. Αυτό έχει άμεση επιρροή και διαφοροποιεί τη διαχείριση των πελατών. Οι επιχειρήσεις διαθέτουν πλέον αρκετά εργαλεία και μεθόδους για να καταφέρουν να επιτύχουν υψηλότερα

επίπεδα ικανοποίησης των πελατών αλλά και να ανευρίσκουν τακτικές για τη διατήρηση της αφοσίωσής τους. Ένα από αυτά τα εργαλεία είναι και το μοντέλο RFM, με το οποίο οι επιχειρήσεις και η υπεύθυνοι μάρκετινγκ έχουν την δυνατότητα να αναγνωρίσουν τους καλύτερους και πιο κερδοφόρους πελάτες της επιχείρησης, στηριζόμενοι σε τρεις βασικούς παράγοντες: την πρόσφατη αγορά, τη συχνότητα των αγορών και τη νομισματική αξία των αγορών (Hosseini et al., 2010).

Κεφάλαιο 2 : Τμηματοποίηση πελατών

2.1. Ορισμός και Χρησιμότητα

Η έννοια της τμηματοποίησης πελατών – customer segmentation – αναπτύχθηκε για πρώτη φορά από τον Αμερικανό ειδικό μάρκετινγκ, Wendell R. Smith το 1956. Η κατάτμηση της αγοράς αποτελεί μια από τις θεμελιώδεις αρχές του μάρκετινγκ και ένα ισχυρό στρατηγικό εργαλείο (Littler, 1995). Ένας από τους σημαντικότερους λόγους για την εφαρμογή της τμηματοποίησης είναι στοχευμένη και καλύτερη εξυπηρέτηση των πελατών μιας επιχείρησης, καθώς αυτό δεν είναι εφικτό όταν αυτοί ανήκουν σε μια μόνο ομάδα (McDonald et al., 2003). Με την προσέγγιση αυτή οι επιχειρήσεις έχουν τη δυνατότητα να εφαρμόζουν κατάλληλες ενέργειες μάρκετινγκ σε κάθε τμήμα καταναλωτών αντίστοιχα με στόχο την καλύτερη κατανόηση και διαχείριση των καταναλωτικών αναγκών και την αύξηση των κερδών (Armstrong et al., 2014).

Υπάρχουν αρκετοί ορισμοί για την τμηματοποίηση της αγοράς. Ένας από αυτούς την ορίζει ως μια διαδικασία κατάτμησης μιας αγοράς σε μικρότερα τμήματα-ομάδες, με διαφορετικές ανάγκες και καταναλωτικές συμπεριφορές, που απαιτούν ξεχωριστά προϊόντα ή έχουν διαφορετική ανταπόκριση, χρησιμοποιώντας διαφορετικά είδη μάρκετινγκ και έχοντας σκοπό να κατανοήσουν και να διαχειριστούν καλύτερα τους πελάτες (Armstrong, 2009). Οι πελάτες που ανήκουν στις ίδιες ομάδες παρουσιάζουν μεταξύ τους ομοιότητες, ενώ τα χαρακτηριστικά μεταξύ των διαφορετικών τμημάτων δεν παρουσιάζουν όμοια χαρακτηριστικά (Yelmen et al., 2020).

Ουσιαστικά, πραγματοποιείται μείωση στον μεγάλο όγκο δεδομένων που διαχειρίζεται μια επιχείρηση με σκοπό να δημιουργηθούν πολλές μικρές ομάδες με κοινά

χαρακτηριστικά. Με τον εντοπισμό των ομάδων, μπορούν να πραγματοποιηθούν προβλέψεις σχετικά με την ανταπόκριση και τις αντιδράσεις των ομάδων σε επερχόμενα προϊόντα και καταστάσεις, να διαμορφωθούν κατάλληλες πολιτικές και να προσδιοριστούν στοχευμένες στρατηγικές μάρκετινγκ (Yankelovich & Meer, 2006).

Βασικός στόχος των επαγγελματιών του μάρκετινγκ είναι η απόκτηση ανταγωνιστικού πλεονεκτήματος, η μεγιστοποίηση των κερδών της αλλά και η επιτυχία και η διατήρηση της επιχείρησης. Για να επιτευχθεί αυτό, θα πρέπει οι επιχειρήσεις να αναγνωρίζουν και να κατανοούν τις καταναλωτικές ανάγκες χρησιμοποιώντας εξατομικευμένες στρατηγικές μάρκετινγκ (Kotler, 2001). Ο στρατηγικός σχεδιασμός μάρκετινγκ βασίζεται στο **μοντέλο STP – Τμηματοποίηση (Segmentation), Στόχευση (Targeting), Τοποθέτηση (Positioning)** – για να επιτύχει το καλύτερο μίγμα μάρκετινγκ. Το μοντέλο αυτό βοηθάει τις επιχειρήσεις να εστιάσουν στις αγοραστικές συμπεριφορές και τις καταναλωτικές ανάγκες και να αναγνωρίσουν τα πιο επικερδή τμήματα, προσαρμόζοντας το κατάλληλο μίγμα μάρκετινγκ και τα κατάλληλα προϊόντα στο κάθε τμήμα (Kalam, 2020). Συγκεκριμένα, η εφαρμογή του μοντέλου οδηγεί α) στη δημιουργία και κατανόηση της αγοράς – στόχου, β) στη διαμόρφωση στόχων και στρατηγικής μάρκετινγκ, γ) στον εντοπισμό των αποδοτικότερων αγορών και στην σωστή τοποθέτηση των προϊόντων μέσω των αντιληπτικών χαρτών. Η διαδικασία του STP μοντέλου είναι ένα εργαλείο που επιτρέπει στο μάρκετινγκ και στις επιχειρήσεις να αναπτύσσουν σχέσεις με τους πελάτες, το προϊόν και την επωνυμία - brand της επιχείρησης και συμβάλλει ώστε οι οργανισμοί να αποκτήσουν ανταγωνιστικό πλεονέκτημα (Kotler, 2001).

Οι επιχειρήσεις που ακολουθούν την προσέγγιση της τμηματοποίησης ενισχύουν την αποτελεσματικότητα του μάρκετινγκ και αξιοποιούν τις ευκαιρίες μάρκετινγκ (Beane & Ennis, 1987). Επιπλέον, δημιουργούν περισσότερους πιστούς πελάτες καθώς εστιάζουν στη βέλτιστη κατανόηση των αναγκών τους και επιτυγχάνουν αποτελεσματικότερη κατανομή των διαθέσιμων πόρων, καθώς διανέμουν τους πόρους στις ελκυστικότερες αγορές. Όσο περισσότερο αυξάνεται το ποσοστό των ικανοποιημένων πελατών, οι επιχειρήσεις πετυχαίνουν μεγαλύτερη αύξηση της κερδοφορίας μέσω της καλύτερης διαφήμισης που πραγματοποιείται από αυτούς (McDonald, 2012).

2.2. Βασικοί τύποι τμηματοποίησης

Η αποτελεσματικότητα και η κερδοφορία των στρατηγικών μάρκετινγκ σε σχέση με την τμηματοποίηση επηρεάζεται σημαντικά από έξι κριτήρια, σύμφωνα με έρευνα που πραγματοποιήθηκε από τον Kotler (1988) (Wendel & Kakamura, 2000). Αυτά είναι:

- **Αναγνωρισιμότητα**: αναφέρεται στο *βαθμό* στον οποίο τα στελέχη μπορούν να αναγνωρίσουν τα ξεχωριστά τμήματα στην αγορά.
- **Ουσιαστικότητα:** αναφέρεται στο *μέγεθος του τμήματος* και ικανοποιείται εάν το τμήμα είναι επαρκές ώστε να εξασφαλίζει την κερδοφορία των ενεργειών μάρκετινγκ.
- **Προσβασιμότητα**: αναφέρεται στο *βαθμό της επικοινωνίας μάρκετινγκ* να φτάσει τους στοχούμενους πελάτες, μέσω διαφημιστικών ή προωθητικών ενεργειών.
- **Ανταπόκριση:** αφορά στον τρόπο με τον οποίο τα διαφορετικά τμήματα *ανταποκρίνονται στις στρατηγικές μάρκετινγκ* της εταιρείας, δηλαδή σε διαφημιστικές καμπάνιες και αλλαγές τιμών.
- **Σταθερότητα**: θα πρέπει να υπάρχει *σταθερότητα* στα τμήματα σε σχέση με τη σύνθεση αλλά και τη συμπεριφορά των ατόμων, για να εφαρμοστεί η στρατηγική μάρκετινγκ.
- **Δυνατότητα δράσης**: αναφέρεται στην ευθυγράμμιση των τμημάτων των πελατών και των ενεργειών μάρκετινγκ με τους στόχους και τις δυνατότητες της επιχείρησης.

Σύμφωνα με τους Armstrong et al., 2014, κατά τη διαδικασία της τμηματοποίησης της αγοράς είναι σημαντικό να καθοριστούν οι μεταβλητές για την τμηματοποίηση και να αναλυθεί το προφίλ των τμημάτων. Οι βασικότερες κατηγορίες τμηματοποίησης είναι:

- **Γεωγραφική τμηματοποίηση**: υποδηλώνει μια αγορά χωρισμένη με βάση την τοποθεσία, όπως η πόλη, ο νομός, η χώρα, ο πληθυσμός. Η γεωγραφική κατάτμηση βασίζεται στην πεποίθηση ότι οι καταναλωτές που ζουν στην ίδια περιοχή τείνουν να εμφανίζουν και τις ίδιες ανάγκες και επιθυμίες καθώς και οι άνθρωποι που βρίσκονται σε διαφορετικές περιοχές είναι πιθανό να έχουν διαφορετικές ανάγκες.

- **Δημογραφική Τμηματοποίηση**: πραγματοποιείται με βάση δημογραφικούς παράγοντες όπως η ηλικία, το φύλο, το εισόδημα, το επάγγελμα. Συχνά τα άτομα που ανήκουν σε όμοιες δημογραφικές ομάδες τείνουν να αντιδρούν με τον ίδιο τρόπο στη διαφήμιση. Επομένως ένας οργανισμός μπορεί να επιλέξει πιο εύκολα τους καταναλωτές που θα στοχεύσει.
- **Συμπεριφορική Τμηματοποίηση**: επικεντρώνονται κυρίως στη συμπεριφορά των καταναλωτών και στον τρόπο λήψης αποφάσεων, όπως ποσότητα αγοράς, τρόπος σκέψης, μοτίβα καταναλωτικής συμπεριφοράς. Έτσι η επιχείρηση έχει τη δυνατότητα να αναγνωρίσει τους πιστούς πελάτες και να εστιάσει στα κατάλληλα σημεία για αύξηση της κερδοφορίας.
- **Ψυχογραφική Τμηματοποίηση**: βασίζεται στον τρόπο ζωής των πελατών, στις καταναλωτικές συνήθειες και στις δραστηριότητες τους, όπως τρόπος ζωής, χόμπι, προσωπικότητα, καταναλωτικές συνήθειες. Η ψυχογραφική τμηματοποίηση αναπτύχθηκε από ερευνητές μάρκετινγκ για να συσχετίσει την προσωπικότητα με τις επωνυμίες.

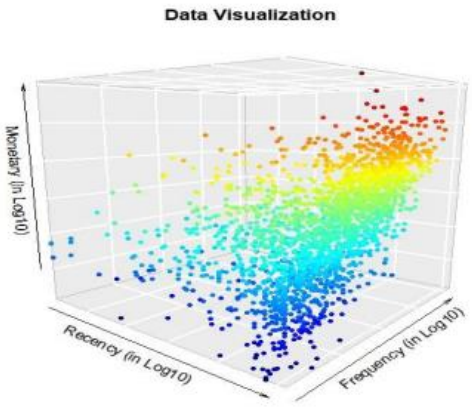
Ένα από τα βασικότερα μοντέλα τμηματοποίησης που βασίζεται στην καταναλωτική συμπεριφορά των πελατών είναι το μοντέλο RFM, το οποίο θα εξετάσουμε και αναλυτικότερα στην παρούσα εργασία.

2.3. RFM Μοντέλο τμηματοποίησης

2.3.1. Σχετικές αναφορές σε RFM model και kmeans clustering

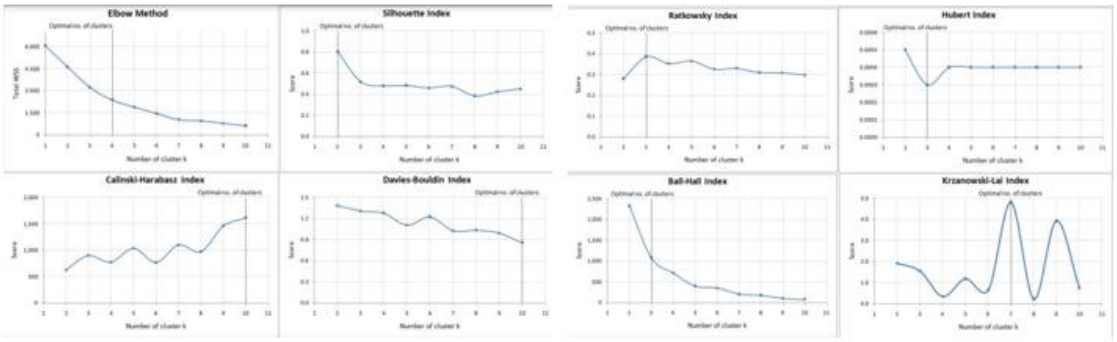
Οι Gustriansyah et al. (2020) στην έρευνα που πραγματοποίησαν, βασίστηκαν στο γεγονός ότι καθώς ο αριθμός των συναλλαγών σε μια επιχείρηση μεγαλώνει, αυτομάτως δημιουργείται και δυσκολία διαχείρισης των βάσεων δεδομένων των προϊόντων για τη διαχείριση των αποθεμάτων. Για το λόγο αυτό, στην έρευνά τους προσπάθησαν να πραγματοποιήσουν τμηματοποίηση προϊόντων σε ένα φαρμακείο στο Palembang, σε δεδομένα που αφορούσαν την διάρκεια του έτους 2015. Για να καταφέρουν να πετύχουν το στόχο τους, προτίμησαν να εφαρμόσουν την ανάλυση RFM. Χρησιμοποίησαν rfm scores με την quantile method, προσδίδοντας το 20% του συνόλου των δεδομένων σε κάθε ομάδα, μετατρέποντας έτσι τις τιμές σε αριθμούς από 1-5. Το πρώτο 20% των τιμών

των μετρικών, με τις παλαιότερες ημερομηνίες πωλήσεων, τις χαμηλότερες πωλήσεις και τις χαμηλότερες χρηματικές αξίες έλαβε την τιμή 1, το επόμενο 20% έλαβε την τιμή 2, μέχρις ότου οι πιο πρόσφατες πωλήσεις και υψηλότερες συχνότητες και χρηματικές αξίες έλαβαν την τιμή 5. Τέλος, οι κωδικοποιημένες αυτές τιμές των τριών μετρικών αποτέλεσαν το rfm score κάθε προϊόντος. Καθώς υπήρχαν ακραίες τιμές και λοξότητα στα δεδομένα, οι ερευνητές χρησιμοποίησαν λογάριθμο για να μετασχηματίσουν τα δεδομένα. Από την εφαρμογή, όπως φαίνεται και στο σχήμα, τα δεδομένα απέκτησαν μια καλύτερη κατανομή. Τα προϊόντα με την καλύτερη τιμή rfm βρίσκονται επάνω δεξιά με κόκκινο χρώμα, ενώ τα προϊόντα με τις χαμηλότερες τιμές εμφανίζονται κάτω αριστερά με σκούρο μπλε χρώμα. Με γαλάζιο, πράσινο και κίτρινο χρώμα είναι προϊόντα με μέσες τιμές rfm.



Εικόνα 1: R,F,M log-transformed metrics 3d visualization

Στη συνέχεια, για την ομαδοποίηση των δεδομένων, επιλέχθηκε ένας από τους δημοφιλέστερους αλγορίθμους, ο kmeans clustering. Για να οριστεί ο βέλτιστος αριθμός των συστάδων επιλέχθηκαν 8 διαφορετικοί δείκτες εγκυρότητας, που συμβάλλουν στη βελτίωση της ακρίβειας στη διαδικασία διαχείρισης των αποθεμάτων και προσδίδουν μια πιο αντικειμενική ομαδοποίηση των προϊόντων. Πιο συγκεκριμένα, η elbow method ανέδειξε τη δημιουργία 4 clusters, η Silhouette Index τη δημιουργία 2 clusters, ο Calinski-Harabasz Index και Davies-Bouldin Index 10 clusters, ο Ratkowski Index, Hubert Index και Ball-Hall Index 3 clusters ενώ ο Krzanowski-Lai Index 7 clusters. Μετά τον έλεγχο, επιλέχθηκε ο αριθμός k = 3.



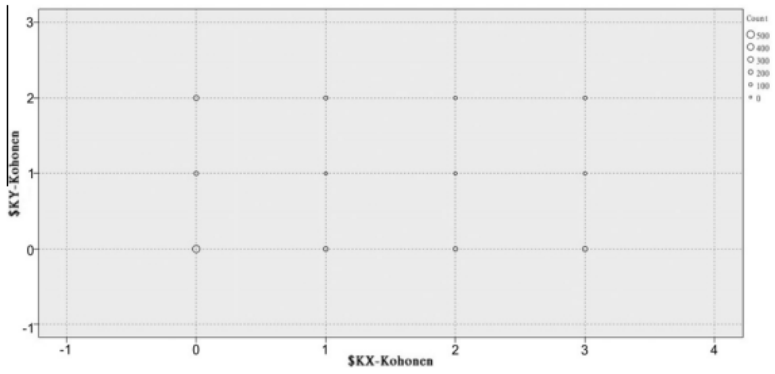
Εικόνα 2: Results from 8 validity indexes

Από το evaluation (αξιολόγηση) που πραγματοποιήθηκε, η μέση διακύμανση των τριών μετρικών ήταν ίση με 0,19113, δηλαδή πολύ κοντά στο μηδέν και επομένως το k = 3 ήταν σωστή επιλογή. Τέλος, ελέγχοντας τα αποτελέσματα της τμηματοποίησης των προϊόντων, παρατηρείται ότι το cluster 3 περιείχε υψηλές μέσες συχνότητες και χρηματικές αξίες αλλά όχι πρόσφατες πωλήσεις, ενώ το cluster 2 περιείχε τις πιο πρόσφατες πωλήσεις αλλά χαμηλές μέσες συχνότητες και χρηματικές αξίες.

Ακόμη μια έρευνα που ασχολήθηκε με την ανάλυση RFM για να κατανοήσει και να αναλύσει την καταναλωτική συμπεριφορά, ήταν και αυτή των Anitha & Patil (2022). Η έρευνά τους βασίστηκε στην ανάλυση ιστορικού πωλήσεων καθώς και της αγοραστικής συμπεριφοράς των καταναλωτών, με στόχο να εντοπιστούν οι κερδοφόροι πελάτες που αποφέρουν και υψηλότερα έσοδα στην επιχείρηση. Σύμφωνα με τη μεθοδολογία που ακολούθησαν, πραγματοποιήθηκε διερευνητική ανάλυση (EDA) και προεπεξεργασία στα δεδομένα, στη συνέχεια εφαρμόστηκε το μοντέλο RFM, ώστε να οριστούν οι τιμές της πρόσφατης, της συχνότητας και της νομισματικής για να αναλύσουν την συμπεριφορά των καταναλωτών και να ανακαλύψουν πιθανά μοτίβα ή τάσεις. Στο επόμενο βήμα, για την ομαδοποίηση των πελατών, χρησιμοποιήθηκε ο αλγόριθμος μηχανικής μάθησης kmeans clustering. Ο βέλτιστος αριθμός των συστάδων επιλέχθηκε με τη βοήθεια της Silhouette Coefficient και ορίστηκε k = 3 (Silhouette score = 0.3622), καθώς απέδιδε καλύτερα αποτελέσματα έναντι του k = 5 (Silhouette score=0,3491).

Οι Wei et al. (2012) στη δική τους έρευνά τους επέλεξαν να χρησιμοποιήσουν μια παραλλαγή της RFM συνδυάζοντας παράλληλα και την τεχνική self-organizing map (SOM) – χάρτης αυτοοργάνωσης - για να πραγματοποιήσουν την τμηματοποίηση. Πιο

συγκεκριμένα, πρόσθεσαν στο μοντέλο την μεταβλητή length (L – ημέρες από την ημερομηνία πρώτης επίσκεψης μέχρι την τελευταία επίσκεψη – LRFM), και αξιοποίησαν τα δεδομένα από μια παιδική οδοντιατρική κλινική στην Taiwan με 2258 πελάτες, με σκοπό να εξετάσουν τις στρατηγικές μάρκετινγκ της κλινικής και να εντοπίσουν τους πολύτιμους πελάτες για τις ενέργειες μάρκετινγκ οδοντιατρικών υπηρεσιών. Για την επιλογή αυτής της επέκτασης της ανάλυσης, οι ερευνητές επικεντρώθηκαν στο γεγονός ότι η διαχείριση των μακροχρόνιων σχέσεων μεταξύ των επιχειρήσεων και των πελατών αποτελεί το βασικό στοιχείο για την πιστότητα των πελατών. Με βάση τα αποτελέσματα της ανάλυσης, παρατηρείται ότι το πλήθος μεταξύ των δυο φύλων κατανέμεται σχεδόν ισόποσα, με αυτό των ανδρών (1158) να υπερτερεί κατά 58 ασθενείς από αυτό του γυναικείου φύλου (1100). Αντίστοιχα, οι ασθενείς διαχωρίστηκαν σε 4 ηλικιακές ομάδες, α) κάτω των 5 ετών, β) 6-10 ετών, γ) 11-15 ετών και δ) 16 ετών και άνω με το μεγαλύτερο πλήθος να εμφανίζεται στην ηλικιακή ομάδα 6-10 (1001 άτομα). Σε ότι αφορά τις μέσες τιμές της συχνότητας (F) παρουσίαζαν ένα καθοδικό μοτίβο, γεγονός που υποδηλώνει ότι παιδιά μικρότερης ηλικίας έκαναν περισσότερες επισκέψεις στο οδοντίατρο από παιδιά μεγαλύτερης ηλικίας.



Εικόνα 3: Generated clusters by SOM technique

Οι ερευνητές με βάση το SOM, κατέταξαν τους πελάτες σε 12 ομάδες, εκ των οποίων το cluster 11 (421) και το cluster 9 (332) περιείχε τα περισσότερα άτομα, ενώ το cluster 5 περιείχε μόλις 8 άτομα. Τα παιδιά με τις μικρότερες ηλικίες εμφανίζονται στα cluster 3,9 και 11. Τέλος, οι ερευνητές για να εφαρμόσουν τις κατάλληλες στρατηγικές σε κάθε ένα cluster, δημιούργησαν έναν Πίνακα Αξίας Πελατών (customer value matrix), στον οποίο οι πελάτες χαρακτηρίστηκαν ως loyal customers, potential customers, new customers και uncertain customers, με τους πρώτους να είναι οι πιο κερδοφόροι.

Ομοίως, οι Wei et al. (2013) στην έρευνά τους συνδύασαν την RFM με τον kmeans και την τεχνική εξόρυξης δεδομένων SOM για να αναγνωρίσουν τους πελάτες υψηλής αξίας, ενός κομμωτηρίου στην Taiwan με βάση τις καταναλωτικές τους συνήθειες και να αναπτύξουν μοναδικές στρατηγικές σε κάθε ομάδα. Από την εφαρμογή του SOM και του αλγορίθμου kmeans τα clusters που δημιουργήθηκαν ήταν 11. Οι ερευνητές, μετά την εφαρμογή της μεθόδου RFM, αφού ανέλυσαν τα προφίλ των πελατών, τις προτιμήσεις και τις συνήθειες τους, ομαδοποίησαν τις συστάδες σε 4 clusters (loyal customers, potential customers, new customers και lost customers) και σχεδίασαν από μια στρατηγική με σκοπό να δημιουργήσουν πιο πιστούς πελάτες και να αποφέρουν μεγαλύτερα κέρδη στην επιχείρηση.

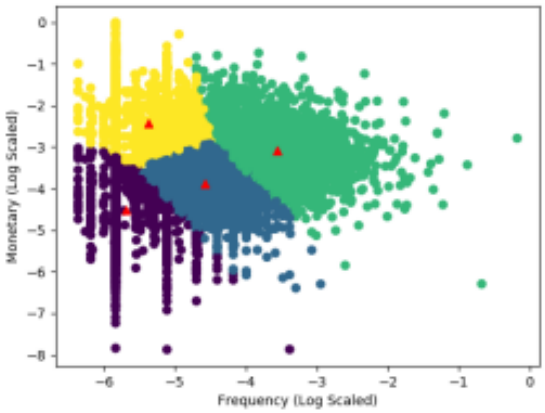
Οι Christy et al. (2021) προσπάθησαν επίσης μια πελατοκεντρική προσέγγιση με τον μοντέλο RFM και τις τεχνικές ομαδοποίησης kmeans, Fuzzy C-Means και RM kmeans (Repetitive Median based kmeans), τμηματοποιώντας το καταναλωτικό τους κοινό και να αναλύοντας τις συμπεριφορές τους σχετικά με τις αγορές τους ώστε να βελτιστοποιήσουν τα κέρδη της εταιρείας. Ως αρχή τους είχαν το γεγονός ότι η προσπάθεια διατήρησης των πελατών είναι πολύ πιο σημαντική από την απόκτηση νέων πελατών που πιθανόν κοστίζουν περισσότερα στην επιχείρηση. Αρχικά, πραγματοποιήθηκε προεπεξεργασία στα δεδομένα για να αφαιρεθούν οι μη επιθυμητές τιμές, υπολογίστηκαν τα z-scores για να διαχειριστούν οι ακραίες τιμές – outliers και στη συνέχεια εφαρμόστηκε το μοντέλο rfm. Όλοι οι πελάτες έλαβαν ένα rfm score (555 – potential customers, 111 – lost customers), και ταξινομήθηκαν βάση αυτού σε ένα από τα 5 cluster που ορίστηκαν (potential, promising, can’t lose them, at risk, lost). Οι ερευνητές καθώς παρατήρησαν λοξότητα στα δεδομένα, έκαναν normalization με τον MinMax scaler και εφάρμοσαν τους αλγορίθμους kmeans και fuzzy c-means. Παράλληλα, πρότειναν και μια νέα προσέγγιση του αλγορίθμου kmeans για καλύτερα και ταχύτερα αποτελέσματα. Πρόκειται για τον αλγόριθμο RM kmeans, ο οποίος διαφοροποιείται στον τρόπο επιλογής των κεντροειδών – βρίσκονται με τη διάμεσο των μετρικών, R,F,M – και μειώνει τον αριθμό των επαναλήψεων του αλγορίθμου. Από τα αποτελέσματα, δημιουργήθηκαν k = 10 clusters και το silhouette score για τους τρεις αλγορίθμους ήταν kmeans = 0.33, fuzzy c-means = 0.43 και RM kmeans = 0.49, όπου ο τελευταίος απέδωσε και το καλύτερο αποτέλεσμα, με το λιγότερο χρόνο και τις λιγότερες επαναλήψεις.

Comparative Analysis of RM K-Means.

	K-Means	Fuzzy C-Means	RM K-Means
Iterations	4	193	2
Time Taken (in seconds)	2.0035	24.7988	1.4917
Average silhouette width	0.33	0.43	0.49

Εικόνα 4: Results from 3 algorithms

Τέλος, οι Tavakoli et al. (2018) επέλεξαν να χρησιμοποιήσουν στην έρευνά τους τη μέθοδο RFM για την τμηματοποίηση των πελατών μιας εταιρείας ηλεκτρονικών αγορών σε συνδυασμό με τον αλγόριθμο kmeans. Κατά την εφαρμογή, αντί για την απλή μορφή της μεθόδου RFM, επέλεξαν μια παραλλαγή αυτής. Συγκεκριμένα, απομόνωσαν το Recency από τις άλλες δυο μετρικές, Frequency και Monetary, και χρησιμοποίησαν τις δυο τελευταίες μαζί, καθώς θεωρούσαν ότι το Recency αποτυπώνει μόνο την ημερομηνία της αγοράς και τίποτα περισσότερο. Αντίθετα, για τη συχνότητα και τη νομισματική, πίστευαν ότι μαζί υποδηλώνουν τον βαθμό πιστότητας των πελατών, οπότε και παρέχουν μεγαλύτερη βοήθεια στους υπεύθυνους μάρκετινγκ να εφαρμόζουν σωστές εξατομικευμένες στρατηγικές. Κατά την εφαρμογή της ανάλυσης RFM, οι ερευνητές επιχείρησαν να πραγματοποιήσουν rfm scoring στους πελάτες, καθώς όμως τα αποτελέσματα δεν ήταν ικανοποιητικά, επιλέχθηκε για τον ορισμό του recency, ο υπολογισμός του αριθμού των ημερών από την τελευταία αγορά. Οι ερευνητές διαχώρισαν τα δεδομένα σε ίσες ομάδες και δημιούργησαν 3 segments για τη μεταβλητή recency: α) Active – ενεργοί πελάτες (τελευταία αγορά < 90 ημέρες), β) Lapsing – δεν αγόρασαν πρόσφατα και μπορεί να χαθούν (90 < τελευταία αγορά < 360) και γ) Lapsed – χαμένοι πελάτες (τελευταία αγορά > 365 ημέρες). Στη συνέχεια, για να ορίσουν τα segments για τα frequency και monetary, χρησιμοποίησαν έναν γραμμικό συνδυασμό με τις βαρύτητές τους, $F * W_F + M * W_M$. Μετά από την αφαίρεση των ακραίων τιμών με την Interquartile method (IQR) και την κανονικοποίηση των δεδομένων με τον MinMax scaler. Παρατηρήθηκε ότι τα δεδομένα δεν ακολουθούν την κανονική κατανομή και υπάρχει λοξότητα - long tail, την οποία οι ερευνητές προσπάθησαν να διαχειριστούν εφαρμόζοντας log transformation στα δεδομένα. Από την εφαρμογή του kmeans, ώστε να εντοπιστεί ο αριθμός των segments για τα frequency-monetary, προέκυψαν 4 segments για τους active πελάτες: 1) high value, 2) medium value with high monetary, 3) medium value with high frequency, 4) low value και αντίστοιχα 3 segments για τους lapsing και lapsed: 1) high value, 2) medium value, 3) low value.



Εικόνα 5: Cluster segmentation

Τέλος, αφού προσδιόρισαν μια εξατομικευμένη στρατηγική για κάθε segment, για καλύτερη προσέγγιση και απόδοση της επιχείρησης, οι ερευνητές επιχείρησαν μια καμπάνια SMS, ώστε να βελτιώσουν την απόδοση των πελάτων και να ενισχύσουν την πιστότητα τους. Από τα αποτελέσματα της καμπάνιας, φαίνεται ότι οι πελάτες αντέδρασαν θετικά στα κουπόνια που προσέφερε η επιχείρηση καθώς υπάρχει αύξηση των εσόδων και του πλήθους των αγορών συγκριτικά με παλαιότερες καμπάνιες της επιχείρησης.

Segment		Average Monetary (USD)			
Recency	Monetary and Frequency	Control Users		Campaign Users	
		Before Campaign	After Campaign	Before Campaign	After Campaign
Active	High Value	74.2	73.8	88.2	89.2
	Medium Value with High Monetary	100.2	97.6	104.6	105.2
	Medium Value with High Frequency	32	35.2	35.4	49.7
	Low Value	50.7	53.2	56.4	65.2

Εικόνα 6: Results from sms campaign for active segment

2.3.2. Ανάλυση του μοντέλου

Στη σημερινή εποχή, η διαχείριση των πελατειακών σχέσεων (CRM) θεωρείται ένα από τα βασικότερα εργαλεία μιας επιχείρησης, έτσι ώστε να μπορέσει να διατηρήσει τους πελάτες της, να ικανοποιήσουν τις ανάγκες τους και να δημιουργήσει μακροχρόνιες σχέσεις μαζί τους (Nguyen et al., 2007). Το CRM θεωρείται μία φιλοσοφία ή μια στρατηγική επιχειρηματικής λειτουργίας που ο κύριος στόχος της, επικεντρώνεται στην προσέλκυση και διατήρηση πελατών, την αύξηση της αξίας και της αφοσίωσής των

πελατών, προσφέροντας προϊόντα προσαρμοσμένα στις ανάγκες τους, καθώς και την εφαρμογή πελατοκεντρικών προσεγγίσεων (Rababah et al., 2011). Προκειμένου να διαχειρίζονται αποτελεσματικά τις σχέσεις τους με τους πελάτες και να γνωρίζουν το ιστορικό και τις αγοραστικές τους συμπεριφορές, οι επιχειρήσεις πρέπει πλέον να υιοθετήσουν πελατοκεντρικές προσεγγίσεις, και όχι προϊοντοκεντρικές όπως ήταν παλαιότερα. Έτσι, οι επιχειρήσεις μπορούν να έχουν στη διάθεσή τους πληροφορίες για το ποιοι είναι οι πελάτες, ποιες οι ανάγκες τους, ποιες οι καταναλωτικές τους συνήθειες, ποια χρηματικά ποσά ξοδεύουν και ποια η αξία του κάθε πελάτη (Xu et al., 2002).

Το μοντέλο ανάλυσης RFM τμηματοποιεί την αγορά, αναλύοντας μεγάλα δεδομένα, και κατηγοριοποιεί τους πελάτες μιας επιχείρησης, με βάση τις καταναλωτικές συνήθειες και συμπεριφορές τους με βάση τρεις μετρικές. Οι μεταβλητές αυτές είναι το διάστημα που μεσολάβησε από την τελευταία αγορά του πελάτη (Recency), η συχνότητα αγορών (Frequency) και την αξία των αγορών (Monetary). Οι τρείς μεταβλητές του μοντέλου RFM περιγράφονται ως εξής (Birant, 2011):

- **Recency (R)**: είναι το διάστημα μεταξύ της τελευταίας αγοράς – συναλλαγής που πραγματοποιήθηκε και του παρόντος. Όσο μικρότερο είναι το διάστημα, τόσο καλύτερο είναι το R.
- **Frequency (F)**: είναι η συχνότητα με την οποία πραγματοποιεί αγορές ο πελάτης, σε ένα συγκεκριμένο χρονικό διάστημα. Όσο μεγαλύτερη είναι η συχνότητα, τόσο μεγαλύτερο είναι το F.
- **Monetary (M)**: είναι η χρηματική αξία των αγορών του πελάτη, δηλαδή το χρηματικό ποσό που καταναλώθηκε σε μια συγκεκριμένη περίοδο.

Σύμφωνα με τους μελετητές Cheng and Chen (2009), μελετώντας τις έρευνες των Hughes (1994) και Stone (1995) για την ανάλυση RFM, κατέληξαν ότι οι απόψεις των δυο μελετητών είναι αντικρουόμενες προς τη θεώρηση τους για τις τρεις μετρικές του μοντέλου. Ο Hughes υποστήριξε ότι οι βαρύτητες των τριών μετρικών είναι ίσες μεταξύ τους και έχουν την ίδια σημασία, ενώ αντίθετα ο Stone υποστήριξε ότι οι τρεις μετρικές διαφέρουν ως προς τη βαρύτητά τους και η σημασία τους καθορίζεται από τα χαρακτηριστικά τις κάθε επιχείρησης που εξετάζεται σε κάθε περίπτωση.

Από την έρευνα που πραγματοποιήθηκε από τους Neslin et al., (2013), έχει παρατηρηθεί ότι το φαινόμενο του customer churn είναι συχνό σε επιχειρήσεις, καθώς πολλές φορές πέφτουν στην παγίδα του recency και θεωρούν ότι πελάτες με αυξημένη πρόσφατη, δεν

θα πραγματοποιήσουν εκ νέου κάποια άλλη αγορά – συναλλαγή με αποτέλεσμα να αγνοούν τους πελάτες αυτούς ή να τους χάνουν. Στην πραγματικότητα, δεν πρόκειται για χαμένους πελάτες αλλά για πελάτες που με κάποια αφυπνιστική ενέργεια από την επιχείρηση θα μπορέσουν να παρακινηθούν ξανά και να πραγματοποιήσουν μια νέα αγορά, αποκτώντας και πάλι αξία για την επιχείρηση.

Με την ανάλυση του μοντέλου, η επιχείρηση είναι ικανή, να πραγματοποιήσει μελλοντικές προβλέψεις και αναλύσεις και διατηρώντας υψηλά τα επίπεδα ικανοποίησης των πελατών, να αυξήσει την κερδοφορία της και να εφαρμόσει τις κατάλληλες στρατηγικές που απαιτούνται (Wei et al., 2010).

2.3.3. Χρησιμότητα

Η τμηματοποίηση πελατών ενισχύει και βελτιώνει τη σχέση των επιχειρήσεων με τους πελάτες και η διαδικασία σχεδιασμού του μάρκετινγκ μπορεί να πραγματοποιηθεί με μεγαλύτερη σαφήνεια (Dibb, 1998). Η ανάλυση RFM χρησιμοποιείται για να μπορέσουν οι επιχειρήσεις να εξηγήσουν την αγοραστική συμπεριφορά των καταναλωτών με βάση τα δεδομένα παλαιότερων συναλλαγών που έχουν στη διάθεσή τους. Οι κατάτμηση των πελατών σε διαφορετικές ομάδες – συστάδες που δημιουργούνται με παρόμοια χαρακτηριστικά, παρέχει στις επιχειρήσεις τη δυνατότητα να εστιάζουν και να κατανοούν αποτελεσματικότερα τις ανάγκες των πελατών, να εφαρμόζουν στοχευμένες καμπάνιες και εξατομικευμένες προωθητικές ενέργειες σε κάθε ομάδα ξεχωριστά και να διαχειρίζονται αποτελεσματικότερα τους διαθέσιμους πόρους που έχει η κάθε επιχείρηση, εφαρμόζοντας διαφορετικές στρατηγικές μάρκετινγκ στην κάθε ομάδα. Η υψηλότερη αξία αντιστοιχεί σε πελάτες που διαθέτουν την υψηλότερη νομισματική αξία και συχνότητα αγορών και τη χαμηλότερη πρόσφατη συναλλαγή (Tsiptsis & Chorianopoulos, 2011; Miglautsch, 2002).

Με την εφαρμογή της μεθόδου, μια επιχείρηση είναι σε θέση να απαντήσει σε σημαντικά ερωτήματα, που συνεισφέρουν στην επιτυχία αυτής και στην αύξηση της κερδοφορίας της. Μερικά από τα σημαντικά αυτά ερωτήματα είναι: α) ποιοι είναι οι καλύτεροι και ποιοι οι χειρότεροι πελάτες της επιχείρησης, β) ποιοι είναι οι πελάτες που μπορεί να αποχωρήσουν από την επιχείρηση (customer churn), γ) ποιοι πελάτες πραγματοποιούν υψηλές αγορές και ποιοι αυτοί που αποφέρουν τα λιγότερα κέρδη στην επιχείρηση, δ)

ποιοι θα ανταποκριθούν σε συγκεκριμένες επικείμενες προωθητικές ενέργειες και ε) πόσο συχνά αγοράζει ένας πελάτης. Η εφαρμογή της μεθόδου μπορεί να πραγματοποιηθεί σχεδόν σε όλους τους κλάδους δραστηριοποίησης, όπως το ηλεκτρονικό εμπόριο, το λιανικό και χονδρικό εμπόριο, οι χρηματοοικονομικές υπηρεσίες, τράπεζες, κλπ (Ernawati, 2021).

2.3.4. Εφαρμογή-Εκτέλεση

Σύμφωνα με τους Wei et al. (2010), κατά την εφαρμογή της ανάλυσης RFM, οι πελάτες σε πρώτα στάδιο ταξινομούνται με βάση τις μετρικές. Αρχικά υπολογίζεται το διάστημα που μεσολαβεί από την πιο πρόσφατη αγορά των πελατών με την ημερομηνία αναφοράς που έχει οριστεί (recency), μετά υπολογίζεται η συχνότητα (frequency) και μετά η χρηματική αξία από τη αγορά – συναλλαγή που πραγματοποιήθηκε (monetary). Στη συνέχεια, οι πελάτες, με βάση την τιμή τις κάθε μετρικής τους, διαχωρίζονται με τη βοήθεια της μεθόδου quintile, σε πέντε ίσες βαθμίδες (20% σε κάθε ομάδα),για κάθε μια μετρική. Με τον τρόπο αυτό κατατάσσεται καθεμία τιμή στην αντίστοιχη βαθμίδα και στη συνέχεια γίνεται η ομαδοποίηση των πελατών.

Σύμφωνα με την ίδια μελέτη, η βαθμολογία ενός πελάτη μπορεί να είναι από το ένα έως το πέντε. Για το recency, η καλύτερη τιμή της πρόσφατης είναι το μικρότερο διάστημα που μεσολαβεί από την τελευταία αγορά με την ημερομηνία διερεύνησης, οπότε και ο πελάτης παίρνει τον αριθμό πέντε (5) και για τον μεγαλύτερο διάστημα, τον αριθμό ένα (1). Αντίθετα, για το frequency και monetary ισχύει το αντίθετο, καθώς οι πελάτες με τη μεγαλύτερη συχνότητα και χρηματική αξία λαμβάνουν την τιμή πέντε (5) και με την χαμηλότερη την τιμή ένα (1). Με τον τρόπο αυτό δημιουργούνται τα λεγόμενα RFM scores. Επομένως, η βαθμολογία ενός πελάτη μπορεί να κυμαίνεται από 555 που είναι το υψηλότερο σκορ, δηλαδή, ο αριθμός αυτός αντικατοπτρίζει τους καλύτερους πελάτες της επιχείρησης, έως το 111 που είναι χαμηλότερο σκορ και αντικατοπτρίζει τους χειρότερους πελάτες της επιχείρησης. Ο μελετητής επισήμανε επίσης ότι μετά τον υπολογισμό των RFM scores μπορεί πλέον να οριστούν οι ομάδες των πελατών (segments) και να πραγματοποιηθεί η τμηματοποίηση τους με βάση τα scores που προέκυψαν.

Σύμφωνα με μελέτη που πραγματοποιήθηκε από τον Birant (2011), υπάρχουν διάφορες εκδοχές για την ανάλυση του μοντέλου RFM, ανάλογα με το επιθυμεί να εξετάσει η επιχείρηση σε κάθε περίπτωση. Στην εκδοχή RFD (recency, frequency, duration) οι αναλυτές λαμβάνουν υπόψη αντί για το χρηματικό ποσό, τη διάρκεια παραμονής του κάθε πελάτη στο φυσικό ή ηλεκτρονικό κατάστημα, ενώ στην εκδοχή WRFM, οι μετρικές πολλαπλασιάζονται και υπολογίζονται με βάση την βαρύτητα που προσδίδεται στην κάθε μια μετρική ξεχωριστά. Σε μια άλλη εκδοχή, η RFM ορίζεται με βάση την εποχικότητα των αγαθών και καλείται ως TRFM και σε μια διαφορετική (RFL) λαμβάνει υπόψη της την πιστότητα των πελατών και συνήθως χρησιμοποιείται σε επιχειρήσεις που ασχολούνται με ετήσιες συναλλαγές.

Οι ερευνητές Wu et al. (2009) πραγματοποίησαν μια έρευνα στη βάση δεδομένων ενός κατασκευαστή (outfitter) στην πόλη Taipei της Ταϊβάν με σκοπό να μελετήσουν και να εντοπίσουν ποιοι είναι οι πελάτες που διαθέτουν την μεγαλύτερη αξία και να προτείνουν τις κατάλληλες στρατηγικές μάρκετινγκ που απαιτούνται στην κάθε μια συστάδα που δημιουργείται, εφαρμόζοντας την ανάλυση RFM και τον αλγόριθμο k-means clustering. Τα αποτελέσματα της έρευνάς τους έδειξαν ότι δημιουργήθηκαν έξι διαφορετικές συστάδες από τι οποίες αυτές με τη μεγαλύτερη αξία ήταν οι συστάδες πέντε και έξι, που μπορούσαν να αποφέρουν μεγαλύτερα κέρδη στην επιχείρηση, ενώ αντίθετα οι συστάδες ένα και δυο δεν είχαν καμία ή είχαν χαμηλή αξία για την επιχείρηση.

Τα τελευταία χρόνια, όλο και περισσότεροι ερευνητές εστιάζουν τις μελέτες τους στο συνδυασμό μεθόδων για την ανάπτυξη μοντέλων ομαδοποίησης. Από μια έρευνα που πραγματοποιήθηκε σε παγκόσμια εταιρεία πιτσαριών της Τουρκίας, βρέθηκε ότι υπάρχει αυξημένη συσχέτιση μεταξύ των μεταβλητών όταν συνδυάζεται το σταθμισμένο μοντέλο της RFM, όπου τα βάρη της κάθε μετρικής παίζουν σημαντικό ρόλο (WRFM) μαζί με τα δημογραφικά χαρακτηριστικά (Sarvari et al., 2016). Επίσης, οι Hosseini et al. (2010) για να καταφέρουν να βελτιώσουν τη Διαχείριση Σχέσεων Πελατών (CRM) για επιχειρήσεις συνδύασαν την σταθμισμένη εκδοχή της RFM σε K-Means αλγόριθμο και οι Gustriansyah et al. (2020) για να τμηματοποιήσουν με τον καλύτερο δυνατό τρόπο τα προϊόντα ενός φαρμακείου στο Palembang εφάρμοσαν το μοντέλο RFM συνδυάζοντας το με τον αλγόριθμο K-Means.

Η ενοποίηση διαφόρων τεχνικών ομαδοποίησης εφαρμόστηκε και από τους Coussement et al. (2010), οι οποίοι επέλεξαν να συνδυάσουν και να συγκρίνουν τρεις τεχνικές

εξόρυξης δεδομένων: την ανάλυση RFM, τη λογιστική παλινδρόμηση και τα δέντρα αποφάσεων, με σκοπό να τμηματοποιηθούν οι πελάτες μιας εταιρείας για εφαρμογή άμεσου μάρκετινγκ. Από την έρευνα προέκυψε ότι τα δέντρα αποφάσεων εμφανίζουν καλυτέρα αποτελέσματα για την αποδοτικότερη ανάλυση των δεδομένων ακόμα και με προβλήματα ακρίβειας δεδομένων.

2.3.5. Διαχείριση Αποτελεσμάτων

Οι επαγγελματίες και οι ερευνητές έχουν αναπτύξει τα τελευταία χρόνια, μεγάλο ενδιαφέρον για το μάρκετινγκ βάσεων δεδομένων (McCarty & Hastak, 2007). Ο στόχος του μάρκετινγκ βάσεων δεδομένων είναι να βελτιώσει τις πωλήσεις, να διατηρήσει πολύτιμους πελάτες στην επιχείρηση, ώστε να αποφέρουν όσο το δυνατόν μεγαλύτερα κέρδη σε αυτήν και να διατηρήσει σε υψηλά επίπεδα την ικανοποίηση των πελατών, μέσω της αποτελεσματικότερης εξυπηρέτησής τους. Επομένως, θεωρείται ότι το μάρκετινγκ βάσεων δεδομένων μπορεί να αποτελέσει ανταγωνιστικό πλεονέκτημα για τις επιχειρήσεις καθώς είναι ένα πολύ ισχυρό εργαλείο που χρησιμοποιείται από αυτές (Punj & Stuart, 1983).

Τα τελευταία χρόνια η ανάλυση RFM χρησιμοποιείται ολοένα και περισσότερο στον κλάδο των επιχειρήσεων και αποτελεί ένα ισχυρό εργαλείο μάρκετινγκ βάσεων δεδομένων, καθώς χρησιμοποιείται ευρέως για τη μέτρηση της αξίας των πελατών με βάση το ιστορικό των αγορών τους από προηγούμενα έτη (Nimbalkar, 2013). Με τη χρήση της μεθόδου αυτής, οι πελάτες χωρίζονται σε συστάδες – ομάδες (clusters) με βάση παρόμοια αγοραστικά χαρακτηριστικά που εμφανίζουν (Akaah et al., 1995; McCarty and Hastak, 2007). Οι υπεύθυνοι αποφάσεων και οι διαχειριστές είναι σε θέση να κατανοήσουν και να εφαρμόσουν εύκολα αυτήν την τεχνική εξόρυξης δεδομένων (Marcus, 1998). Από την τμηματοποίηση που προκύπτει από την εφαρμογή του μοντέλου, μπορούν να αναπτυχθούν προσαρμοσμένες στρατηγικές μάρκετινγκ για συγκεκριμένες ομάδες πελατών (Ernawati et al., 2021).

Κάθε ομάδα πελάτη είναι εφικτό να τμηματοποιηθεί με διαφορετικές στρατηγικές μάρκετινγκ για την καλύτερη και αποτελεσματικότερη προσέγγιση των πελάτων της. Μια ενδεικτική τμηματοποίηση μπορεί να είναι η παρακάτω (Webengage, χ.χ.):

Εικόνα 7: Τμηματοποίηση RFM. Πηγή: https://www.annexcloud.com/blog/revolutionising-segmentation-individualisation-using-rfm-to-step-further/

- **Champions:** Τα χαρακτηριστικά συμπεριφοράς αυτής της ομάδας πελατών συνήθως χαρακτηρίζονται από τις αγορές που έχουν πραγματοποιηθεί πρόσφατα (υψηλό R), την υψηλότερη συχνότητα (υψηλό F) και τα υψηλότερα χρηματικά ποσά (υψηλό M). Αυτοί είναι οι καλύτεροι πελάτες της επιχείρησης και είναι υπεύθυνοι για ένα μεγάλο μερίδιο των εσόδων του οργανισμού. Η διατήρηση της ικανοποίησής τους θα πρέπει να διατηρηθεί σε υψηλά επίπεδα. Οι προσπάθειες μάρκετινγκ θα πρέπει να επικεντρωθούν σε ιδιαίτερες ανταμοιβές, στην παροχή ειδικών προσφορών ή προνομίων που θα τους κάνουν να νιώσουν ξεχωριστοί, στην έκδοση καρτών VIP μελών, εξατομικευμένες καμπάνιες, προτεραιότητα στην διεκπεραίωση ή/και αποστολή παραγγελίας. Επίσης, από αυτό το τμήμα θα πρέπει να λαμβάνεται τακτική ανατροφοδότηση για να ενημερώνεται η επιχείρηση για τις προτιμήσεις και νέες τάσεις του καταναλωτικού κοινού αλλά και πιθανές για αστοχίες ή βελτιώσεις στα προϊόντα της. Επίσης είναι χρήσιμο τμήμα για έρευνα μάρκετινγκ σχετικά με το λανσάρισμα νέων προϊόντων.
- **Loyal Customers:** Οι πελάτες που ανήκουν σε αυτήν την ομάδα συνήθως χαρακτηρίζονται από υψηλές πρόσφατες αγορές (υψηλό R), υψηλή συχνότητα (υψηλό F) και υψηλές δαπάνες (υψηλό M). Πρόκειται για πολύ ενεργούς και πολύτιμους πελάτες. Ενημερώνονται για τα νέα του οργανισμού και έχουν εγγραφεί στα ενημερωτικά δελτία. Στόχος είναι η διατήρηση τους, η εφαρμογή τεχνικών πώλησης Up-Selling και Cross-Selling, με σκοπό να δημιουργηθούν πλασματικές ανάγκες που θα αυξήσουν το αγοραστικό τους ενδιαφέρον και συνεπώς θα προβούν σε περισσότερες ή και ακριβότερες αγορές. Εφαρμογή

εξατομικευμένων ενεργειών και προσωποποιημένου μάρκετινγκ, customer service καθώς και οποιαδήποτε άλλη ενέργεια μπορεί να αυξήσει τη δέσμευσή τους με τον οργανισμό. Δεδομένου ότι στο τμήμα αυτό αγοράζουν ακριβά προϊόντα, αυξάνοντας τη συχνότητα των πελατών, το αποτέλεσμα στην κερδοφορία θα είναι εντυπωσιακό.

- **Potential Loyalists:** Τα χαρακτηριστικά συμπεριφοράς αυτής της ομάδας πελατών συνήθως χαρακτηρίζονται από πρόσφατες αγορές (υψηλό R), υψηλή συχνότητα αγορών (υψηλό F) και μέτριες δαπάνες (μέτριο M). Αυτοί οι πελάτες έχουν ήδη αγοράσει από τον οργανισμό, αλλά το μέγεθος του καλαθιού τους δεν ήταν πολύ μεγάλο. Υπάρχει περιθώριο για παρακίνηση της αύξησης των αγορών μέσω συστάσεων για συμπληρωματικά ή επιπλέον προϊόντα.
- **Promising:** Οι πελάτες της ομάδας αυτής συνήθως χαρακτηρίζονται από πρόσφατες αγορές (υψηλό R), με μέτρια συχνότητα (μέτριο F) και υψηλές δαπάνες (υψηλό M). Οι πελάτες αυτοί προέβησαν σχετικά πρόσφατα σε αγορά υψηλής αξίας, αλλά δεν είναι τακτικοί πελάτες. Καθώς αξίζει να επενδυθούν πόροι στο τμήμα αυτό, προτείνεται η ένταξή τους σε κάποιο πρόγραμμα πιστότητας ή ανταμοιβής. Οι προτάσεις αγορών ή η αποστολή newsletters είναι επίσης μία τακτική που θα μπορούσε να φέρει τα επιθυμητά αποτελέσματα.
- **New Customers:** Οι πελατών συνήθως χαρακτηρίζονται από πρόσφατες αγορές (υψηλό R), χαμηλή συχνότητα (χαμηλό F) και χαμηλές ή μέτριες δαπάνες (μέτριο M). Αυτοί οι πελάτες προέβησαν σχετικά πρόσφατα σε αγορά, με μέση ή κάτω από τη μέση τιμή καλαθιού και δεν είναι συχνοί πελάτες. Πιθανώς ορισμένοι από αυτούς να πραγματοποιούν την πρώτη τους αγορά από την επιχείρηση. Είναι σημαντικό να διαπιστωθούν οι ανάγκες και οι προτιμήσεις τους, και να τους δοθούν κίνητρα για αγορά. Η επιχείρηση μπορεί κάνει ειδικές εκπτώσεις, να προσφέρει κάποιο δώρο μαζί με την πρώτη αγορά ή να παραχωρήσει κάποιο εκπτωτικό κουπόνι για την επόμενη αγορά. Επίσης μπορεί να διενεργηθεί έρευνα ικανοποίησης μέσω phone μάρκετινγκ ή μέσω website για αξιολόγηση (rate) και σχολιασμό των προϊόντων. Η προσπάθεια επικεντρώνεται στη διαφοροποίηση της επιχείρησης στα μάτια των πελατών και συνεπώς να λειτουργήσει αυτόματα το word of mouth για να διαφημιστούν τα προϊόντα και το όνομα της επιχείρησης περισσότερο και γρηγορότερα.

- **Need Attention:** Είναι πελάτες με πρόσφατες ή σχετικά πρόσφατες αγορές (μέτριο R), μέτρια συχνότητα (μέτριο F) και υψηλές ή μέτριες δαπάνες (μέτριο M). Για τους πελάτες αυτούς, είναι σημαντικό να δοθούν κίνητρα για αγορά και τονιστούν τα μοναδικά χαρακτηριστικά της προσφοράς του οργανισμού, ούτως ώστε να μετατραπούν σε συχνούς πελάτες. Το τμήμα αυτό ανταποκρίνεται καλύτερα σε προσφορές περιορισμένης χρονικής διάρκειας ή ενέργειες εξατομίκευσης, επιθετικό μάρκετινγκ, προσωποποιημένη επικοινωνία, χαμηλού κόστους προωθητικές αυτοματοποιημένες καμπάνιες, customer services και διαφημιστικές ενέργειες.
- **About to Sleep:** Οι πελάτες χαρακτηρίζονται από λιγότερο πρόσφατες αγορές (μέτριο R), μέτρια ή χαμηλή συχνότητα (μέτριο F), μέτριες ή χαμηλές δαπάνες (χαμηλό M). Οι πελάτες σε αυτό το τμήμα δεν έχουν αγοράσει για μεγάλο χρονικό διάστημα, αλλά όχι σε βαθμό που να είναι μη προσεγγίσιμοι. Με τις κατάλληλες στρατηγικές μπορούν πάλι να ενεργοποιηθούν. Ως εκ τούτου, το ενδιαφέρον τους μπορεί να αναζωπυρωθεί με την συνεχή ενημέρωση και παρουσία. Εκπτώσεις, ενέργειες 1+1 δώρο, email μάρκετινγκ, reminder emails για εκπτωτικά προϊόντα ή υπενθύμιση ότι έχουν καιρό να πραγματοποιήσουν κάποια αγορά, εκπτωτικές ενέργειες σε προϊόντα σχετικά με παλαιότερες προτιμήσεις τους, θα ήταν ικανοποιητικά κίνητρα για την αφύπνισή τους.
- **Can’t Lose Them:** Η ομάδα αυτή περιέχει πελάτες με όχι πρόσφατες αγορές (χαμηλό R), μέτρια ή υψηλή συχνότητα (μέτριο F) και υψηλής αξίας συναλλαγές (υψηλό M). Η αξία αυτών των πελατών είναι μεγάλη, ανεξάρτητα που δεν αγοράζουν πρόσφατα και συχνά, είναι πολύτιμοι για την επιχείρηση καθώς όταν αγοράσουν ξοδεύουν αρκετά μεγάλα ποσά. Για την αφύπνισή τους, προτείνεται προσωποποιημένη τηλεφωνική επικοινωνία, win back καμπάνιες, προγράμματα αφοσίωσης, απαλλαγή από έξτρα χρεώσεις (έξοδα αποστολής ή δωρεάν επιστροφή προϊόντων). Μια καμπάνια προσφορών (εκπτώσεις ή δώρων) σε προϊόντα που αγόραζαν στο παρελθόν ή έχουν αναζητήσει θα ήταν μια ενδεδειγμένη στρατηγική για την επαναπροσέγγισή τους. Σημαντική επίσης θα ήταν και η λήψη ανατροφοδότησης από αυτούς για να υπάρχει καλύτερη προσέγγιση στις προτιμήσεις τους και τις ανάγκες τους. Καθώς το τμήμα είναι σημαντικό για την κερδοφορία του οργανισμού, θα πρέπει να επενδυθούν πόροι σε αυτό.

- **At Risk:** Οι πελάτες συνήθως έχουν όχι τόσο πρόσφατες αγορές (χαμηλό R), υψηλή ή μέτρια συχνότητα (μέτριο F) και δεν έχουν δαπανήσει αρκετά χρήματα (μέτριο M). Οι πελάτες αυτοί δεν ολοκληρώνουν σημαντικό μέρος των αγορών τους, αλλά είναι πιθανό να ανταποκριθούν σε προσπάθειες προσωποποιημένης επικοινωνίας ή emails για υπενθύμιση ότι έχουν ξεχάσει προϊόντα στο καλάθι. Σε κάθε περίπτωση, θα πρέπει να διαπιστωθεί ο λόγος για τον οποίο δεν ολοκληρώνουν τις αγορές τους.
- **Hibernating:** Τα χαρακτηριστικά συμπεριφοράς αυτής της ομάδας πελατών συνήθως χαρακτηρίζονται από οχι πρόσφατες αγορές (μέτριο R), χαμηλή συχνότητα (χαμηλό F) και χαμηλές δαπάνες (χαμηλό M). Δεν συνιστάται ιδιαίτερη επένδυση πόρων για αυτήν την ομάδα, καθώς η απόδοσή της δεν είναι πιθανό να είναι θετική αν ληφθεί υπόψη ότι οι πελάτες αυτοί έχουν στο παρελθόν δείξει απροθυμία στις προσπάθειες προσέγγισης.
- **Lost:** Πρόκειται για πελάτες με παλιές αγορές (χαμηλό R), πολύ χαμηλή συχνότητα (χαμηλό F) και με τις χαμηλότερες δαπάνες (χαμηλό M). Συνήθως είναι πιο κερδοφόρο για μια επιχείρηση να προσεγγίσει νέους πελάτες παρά να σπαταλήσει χρόνο και χρήμα για τη διατήρηση αυτών των πελατών. Στρατηγικές όπως αποστολή ενημερωτικών emails, διαφημίσεις σε εορταστικές περιόδους ή κατά την διάρκεια των εκπτώσεων, εκπτωτικά κουπόνια με μικρά χρηματικά ποσά, ίσως αποτελούσαν κάποιο κίνητρο για μια νέα αγορά.

2.3.6. Πλεονεκτήματα RFM

Το μοντέλο RFM είναι δημοφιλές εδώ και δεκαετίες, καθώς αποτελεί ένα από τα πιο ισχυρά εργαλεία εξόρυξης δεδομένων και ανάλυσης της αξίας των πελάτων (Kaymak, 2001). Αρχικά, ένα από τα πλεονεκτήματα χρήσης της μεθόδου είναι ότι προσδιορίζει στις επιχειρήσεις τους πολυτιμότερους πελάτες της καθώς και τους πελάτες που έχουν μεγάλη αξία για την επιχείρηση (Wang, 2010). Σε ότι αφορά τις μεταβλητές του μοντέλου, οι πληροφορίες σχετικά με τις συναλλαγές των πελατών, συλλέγονται από μια εσωτερική βάση δεδομένων και όχι από κάποια εξωτερική βάση δεδομένων (Kaymak, 2001). Επιπλέον, είναι ένα μοντέλο με μικρή πολυπλοκότητα, καθώς για την εξαγωγή βάσεων δεδομένων χρησιμοποιείται ένας μικρός αριθμός μεταβλητών, ικανός για να προβλέπονται αποτελέσματα με ακρίβεια (Wei et al., 2010) και επιπλέον τα δεδομένα αποθηκεύονται σε ηλεκτρονική μορφή που είναι εύκολα προσβάσιμη και επεξεργάσιμη

(Lumsden et al., 2008). Είναι μια μέθοδος ανάλυσης που η εφαρμογή της, μπορεί να κατανοηθεί εύκολα από τους υπεύθυνους λήψης αποφάσεων (McCarty και Hastak, 2007; Wang, 2010).

Υπάρχουν διάφοροι λόγοι που η ανάλυση RFM θεωρείται μια από τις σημαντικότερες μεθόδους τμηματοποίησης. Η ανάλυση RFM είναι μια οικονομικά αποδοτική μέθοδος, που συμβάλλει στην ποσοτικοποίηση της συμπεριφοράς των πελατών (Kahan, 1998; Miglautsch, 2000). Επίσης, συμβάλλει στη μέτρηση της ισχύος της σχέσης με τους πελάτες μιας επιχείρησης (Wang, 2010). Ένα άλλο πλεονέκτημα από την εφαρμογή του μοντέλου είναι η δυνατότητα ανάπτυξης διαφορετικών στρατηγικών μάρκετινγκ στις αντίστοιχες ομάδες πελάτων που προβλέπεται να ανταποκριθούν επικερδώς σε συγκεκριμένες ενέργειες (Rani et al., 2013). Τέλος, τα ποσοστά απόκρισης των πελάτων στο μοντέλο είναι αυξημένα καθώς παρέχει στις επιχειρήσεις τη δυνατότητα ενίσχυσης των κερδών τους σε σύντομο χρονικό διάστημα (Rani et al., 2013) και το κόστος εφαρμογής του είναι χαμηλό (Baecke & Van den Poel, 2011).

Η ανάλυση RFM έχει χρησιμοποιηθεί σε διάφορους επιχειρηματικούς τομείς με διάφορες εφαρμογές, όπως η διαχείριση του φαινομένου customer churn στον κλάδο της αεροπορικής βιομηχανίας (Ran & Cheng, 2021), η αποτελεσματική εφαρμογή πελατοκεντρικού μάρκετινγκ στον τομέα της διαδικτυακού εμπορίου (Chen et al., 2012), η κατάταξη των πελατών με βάση την πιστότητα στον τραπεζικό κλάδο (Zaheri et al., 2012), οι συστάσεις προϊόντων σε πελάτες με βάση κανόνες που σχετίζονται με το ιστορικό συναλλαγών (Liu et al., 2009) κ.α.

2.4. Μηχανική Μάθηση (Machine Learning)

Η μηχανική μάθηση (ML) είναι ένα πεδίο του κλάδου των υπολογιστών που εφαρμόζεται τα τελευταία χρόνια από τις επιχειρήσεις καθώς η ζήτηση της έχει αυξηθεί ραγδαία. Η ραγδαία αυτή εξέλιξη οφείλεται κυρίως στην ανάγκη για επίλυση προβλημάτων όπως η εξόρυξη δεδομένων, η ψηφιοποίηση αρχείων, η αναγνώριση προτιμήσεων των καταναλωτών, η αποφυγή απάτης κ.α. Η μηχανική μάθηση βοηθά στην επίλυση πολύπλοκων προβλημάτων μελετώντας αλγορίθμους και στην αποτελεσματικότερη διαχείριση των δεδομένων (Γεωργούλη, 2016). Με την χρήση της μηχανικής μάθησης, οι αλγόριθμοι εκπαιδεύονται να ανακαλύπτουν μοτίβα και να κάνουν προβλέψεις από

μεγάλα σύνολα δεδομένων, χωρίς να είναι απαραίτητη η προσαρμογή συγκεκριμένων κανόνων (Zhou, 2022). Η βελτίωση της απόδοσης επιτυγχάνεται με την «εμπειρία» που αποκτούν οι αλγόριθμοι από προηγούμενους υπολογισμούς με αποτέλεσμα να παράγονται ακόμα πιο αξιόπιστα αποτελέσματα (Rana & Bhushan, 2022). Ο αλγόριθμος που θα χρησιμοποιηθεί για την επίλυση ενός προβλήματος δεν είναι πάντα ο ίδιος αλλά η επιλογή αυτού εξαρτάται από παράγοντες όπως το μοντέλο που θα χρησιμοποιηθεί, τη φύση του προβλήματος και τις μεταβλητές που υπάρχουν. Ο τομέας της μηχανικής μάθησης συνδυάζεται και με άλλες τομείς όπως είναι η βιολογία, η στατιστική, η τεχνική νοημοσύνη κ.α. για την αποτελεσματικότερη προσέγγιση του προβλήματος και των αποτελεσμάτων (Mahesh, 2020).

Οι βασικότερες κατηγορίες της μηχανικής μάθησης είναι η επιβλεπόμενη μάθηση (supervised learning), η μη επιβλεπόμενη μάθηση (unsupervised learning) και η ενισχυτική μάθηση (reinforcement learning) (Nasteski, 2017). Στην επιβλεπόμενη μάθηση, το μοντέλο εκπαιδεύεται με ένα σύνολο δεδομένων και μαθαίνει από παραδείγματα με ετικέτα. Υπάρχει πάντα μια τιμή εισόδου, όπου αφορά τα δεδομένα που καταχωρούνται στον αλγόριθμο και μια τιμή εξόδου, που αφορά την τιμή που θα προβλεφθεί. Τα δεδομένα εισόδου διαχωρίζονται σε δεδομένα εκπαίδευσης (training set), μαζί με την ετικέτα κατηγοριοποίησης (label) και σε δεδομένα δοκιμής (test set), με τα οποία γίνεται και η πρόβλεψη. Στην ουσία ο αλγόριθμος αναλύει από το σύνολο εκπαίδευσης δεδομένα με ετικέτα για να μπορεί μετά να αναγνωρίζει παραδείγματα χωρίς ετικέτα και να τα ταξινομεί σωστά. Η επιβλεπόμενη μάθηση χρησιμοποιείται σε προβλήματα ταξινόμησης (classification) και παλινδρόμησης (regression) (Alzubi et al., 2018).

Αντίθετα, στη μη επιβλεπόμενη μάθηση, οι αλγόριθμοι προσπαθούν να ανακαλύψουν συσχετίσεις και μοτίβα που επαναλαμβάνονται από δεδομένα που δεν διαθέτουν ετικέτα, για να πραγματοποιήσουν τις προβλέψεις. Στη διαδικασία αυτή δεν υπάρχει ανθρώπινη παρέμβαση και καμία βάση σε προηγούμενη εκπαίδευση. Τα δεδομένα τμηματοποιούνται σε ομάδες με βάση χαρακτηριστικά και ομοιότητες για να αναλυθούν τα δεδομένα και προβλεφθεί το αποτέλεσμα. Η μη επιβλεπόμενη μάθηση εφαρμόζεται σε προβλήματα ομαδοποίησης (clustering) και προβλήματα ανάλυσης συσχέτισης (association analysis) (Alzubi et al,. 2018). Στην ενισχυτική μάθηση, ο αλγόριθμος μέσω της αλληλεπίδρασης που έχει με το περιβάλλον εκπαιδεύεται σε μια σειρά από ενέργειες

στρατηγικού χαρακτήρα και είναι κατάλληλη για προβλήματα σχεδιασμού (Γεωργούλη, 2016).

2.5. Συσταδοποίηση στη μηχανική μάθηση (Clustering)

Οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται για να επιλυθούν ποικίλλα προβλήματα μεταξύ των οποίων είναι και προβλήματα που απαιτούν συσταδοποίηση ή ομαδοποίηση των δεδομένων (clustering). Η συσταδοποίηση, όπως αναφέρθηκε και παραπάνω, εμπίπτει στην κατηγορία της μάθησης χωρίς επίβλεψη και είναι μια μέθοδος που βοηθάει στην εξόρυξη δεδομένων, η οποία πραγματοποιεί κατάτμηση ενός μεγάλου συνόλου δεδομένων σε συστάδες (ομάδες - clusters) με βάση κάποια κοινά χαρακτηριστικά-κριτήρια (Βερύκιος κ.α, 2016). Οι συστάδες που δημιουργούνται δεν έχουν όμοια χαρακτηριστικά μεταξύ τους αλλά τα μέλη κάθε συστάδας διαθέτουν όμοια χαρακτηριστικά μεταξύ τους (Gustriansyah et al., 2020).

Με βάση τον τρόπο προσέγγισης, δυο είναι οι βασικές κατηγορίες της συσταδοποίησης:

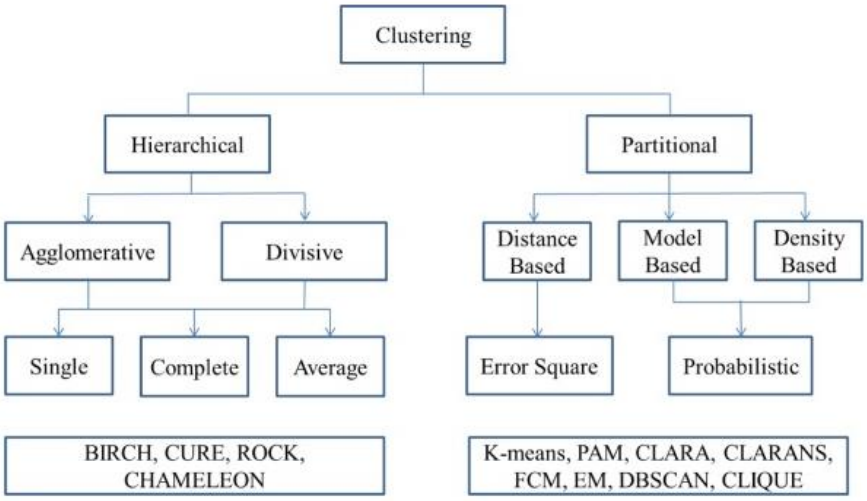
Διαχωριστική συσταδοποίηση (Partinional clustering): είναι ο διαμερισμός κατά τον οποίο κάθε αντικείμενο ενός συνόλου διαχωρίζεται σε ένα συγκεκριμένο αριθμό ομάδων. Ο αριθμός των ομάδων (clusters) πρέπει να είναι γνωστός από την αρχή και συνήθως συμβολίζεται με k. Ένας από τους δημοφιλέστερους αλγορίθμους, που θα αναλυθεί εκτενέστερα και παρακάτω είναι ο αλγόριθμος ομαδοποίησης kmeans.

Ιεραρχική συσταδοποίηση (Hierarchical clustering): είναι η συγχώνευση των μεμονωμένων αντικειμένων σε μεγαλύτερα συμπλέγματα ή η διαίρεση ενός συνόλου αντικειμένων σε πολλαπλά μικρά συμπλέγματα, ανάλογα αν πρόκειται για συσσωρευτική ή διαιρετική μέθοδο. Ο αριθμός των συμπλεγμάτων δεν είναι απαραίτητο να ορισθεί από την αρχή (Kononenko & Kukar, 2007).

2.6. Αλγόριθμοι Συσταδοποίησης

Οι αλγόριθμοι συσταδοποίησης ή ομαδοποίησης έχουν την ικανότητα να διαχειρίζονται αποτελεσματικά μεγάλους όγκους δεδομένων και να δίνουν αποτελέσματα με μεγάλη ακρίβεια. Οι τεχνικές ομαδοποίησης συχνά λειτουργούν συνδυαστικά τους αλγορίθμους

στην επίλυση ενός προβλήματος για την καλύτερη προσέγγιση του προβλήματος και για πιο στοχευμένα αποτελέσματα (Gustriansyah et al., 2020).



Εικόνα 8: Τεχνικές Ομαδοποίησης. Πηγή: Saxena et al., 2022

Σύμφωνα με τη βιβλιογραφία, διατίθεται μεγάλη ποικιλία αλγορίθμους ομαδοποίησης ανάλογα με το εκάστοτε πρόβλημα που πρέπει να αντιμετωπιστεί, παρακάτω όμως θα αναφερθούμε στους δύο πιο βασικούς από αυτούς (Ahuja et al., 2019).

2.6.1. K-means Clustering

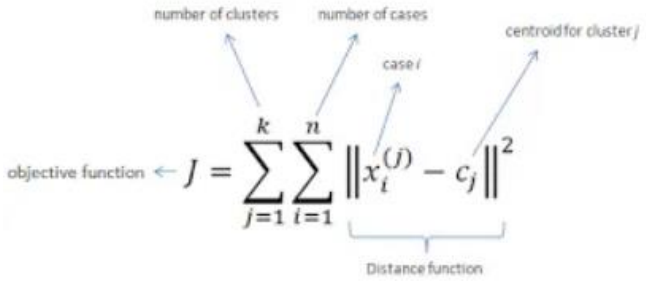
Ο αλγόριθμος k-means είναι ένας από τους πιο δημοφιλείς αλγορίθμους που ανήκει στην κατηγορία της μη επιβλεπόμενης μάθησης και συγκεκριμένα στην διαχωριστική ομαδοποίηση. Χρησιμοποιήθηκε από τον MacQueen το 1967 και μια από τις κύριες εφαρμογές του είναι η τμηματοποίηση των πελάτων (MacQueen, 1967). Είναι ένας επαναληπτικός αλγόριθμος που για την εφαρμογή του, ο αριθμός (k) των ομάδων (clusters) που θα δημιουργηθούν απαιτείται να είναι γνωστός εκ των προτέρων. Για την εύρεση της απόστασης μεταξύ των σημείων από τα κοντινότερα κεντροειδή (centroids) χρησιμοποιείται η Ευκλείδεια απόσταση ή η απόσταση Manhattan. Η επιλογή των αρχικών σημείων k που θα είναι και τα κεντροειδή των ομάδων (centroids) ορίζεται

τυχαία. Η εκχώρηση κάθε σημείου γίνεται στο σύμπλεγμα με το πλησιέστερο κεντροειδές.

Βήματα ομαδοποίησης K-means:

- Βήμα 1: Προσδιορισμός του αριθμού των clusters k.
- Βήμα 2: Τυχαία επιλογή σημείων k που θα αντιμετωπίζονται αρχικά ως κεντροειδή (centroids).
- Βήμα 3: Εκχώρηση όλων των σημείων στα πλησιέστερα κεντροειδή αντίστοιχα που θα έχουν ως αποτέλεσμα τον σχηματισμό k clusters.
- Βήμα 4: Εκ νέου υπολογισμός των κεντροειδών.
- Βήμα 5: Εκ νέου κατανομή όλων των σημείων δεδομένων στα αντίστοιχα νέα κεντροειδή τους. Εάν πραγματοποιηθεί κάποια τέτοια ανακατανομή, πηγαίνουμε ξανά στο βήμα 4 διαφορετικά τερματίζουμε το βρόχο (Gustriansyah, 2020).

Μαθηματικός τύπος αλγορίθμου:



Εικόνα 9: kmeans algorithm mathematical type. Πηγή: https://heartbeat.comet.ml/understanding-the-mathematics-behind-k-means-clustering-40e1d55e2f4c

Μερικά πλεονεκτήματα του αλγορίθμου είναι (i) η εύκολη εφαρμογή του, (ii) απαιτεί μόνο μια παράμετρο για είσοδο στον αλγόριθμο και (iii) ότι είναι ταχύτερος σε σχέση με άλλους αλγορίθμους (Anitha & Patil, 2022). Αντίθετα, μερικά περιορισμοί του είναι (i) ο αρχικός καθορισμός της τιμής k, (ii) ο ορισμός των αρχικών κεντροειδών καθώς επηρεάζουν τα τελικά αποτελέσματα και (iii) δεν προτείνει απαραίτητα την βέλτιστη λύση (Xu & Wunch, 2005).

2.6.2. Αλγόριθμος Ιεραρχικής Ομαδοποίησης

Στον ιεραρχικό αλγόριθμο, τα δεδομένα διαχωρίζονται με ιεραρχικό τρόπο, δημιουργώντας εμφωλευμένες συστάδες για την εύρεση του αριθμού των συστάδων αλλά και της δομής τους. Η αναπαράσταση γίνεται με ένα δενδρόγραμμα και ο τελικός αριθμός των συστάδων εξαρτάται από το κλάδεμα που θα γίνει στο δενδρόγραμμα (Brown & Gunderson, 2002). Οι τεχνικές που μπορεί να ακολουθήσει αυτός ο αλγόριθμος είναι είτε η διαιρετική (divisive) είτε η συσσωρευτική (agglomerative). Στη διαιρετική ομαδοποίηση (από επάνω προς τα κάτω), δημιουργούνται πολλά μικρά clusters με σκοπό να διαιρεθούν μεμονωμένα όλα τα δεδομένα, τα οποία ανήκουν αρχικά σε μια ενιαία ομάδα ενώ στη συσσωρευτική (από κάτω προς τα επάνω), πραγματοποιείται συγχώνευση των μεμονωμένων αντικειμένων σε ένα μεγάλο σύμπλεγμα (Saxena et al., 2017).

Κεφάλαιο 3 : Μεθοδολογία

3.1. Πλαίσιο και Μονάδα Δειγματοληψίας

Το σύνολο των δεδομένων που χρησιμοποιήθηκε ήταν μια βάση δεδομένων ενός ηλεκτρονικού καταστήματος ειδών δώρων το οποίο εδρεύει στο Ηνωμένο Βασίλειο. Το πελατολόγιο της επιχείρησης απαρτίζεται και από B2B πελάτες αλλά και από B2C πελάτες, γεγονός που δικαιολογεί και τα μεγάλα νούμερα που παρουσιάζονται στις ποσότητες των προϊόντων, οπότε και τα μεγάλα ποσά. Μονάδα δειγματοληψίας για την έρευνα που πραγματοποιήθηκε αποτέλεσε ο κάθε πελάτης που ήταν καταχωρημένος στην πελατειακή βάση της επιχείρησης (CustomerID) και το μέγεθος του δείγματος ήταν ίσο με n = 4334 μοναδικοί πελάτες με τουλάχιστον μια παραγγελία ο καθένας.

3.2. Περιγραφή Συνόλου Δεδομένων

Το συγκεκριμένο dataset περιείχε δεδομένα των πελατών και των αγορών τους που πραγματοποιήθηκαν κατά τη διάρκεια μεταξύ 01/12/2010 και 09/12/2011 από διαφορετικές χώρες. Το κύριο αντικείμενο της εταιρείας ήταν η πώληση ειδών δώρων μέσω του ηλεκτρονικού καταστήματος που διαθέτει. Το σύνολο δεδομένων που αναλύθηκε ήταν ίσο με 541.909 εγγραφές και 8 στήλες. Τα δεδομένα που εξετάστηκαν αντλήθηκαν από την πηγή: https://www.kaggle.com/code/mahmoudelfahl/cohort-analysis-customer-segmentation-with-rfm/data?select=Online+Retail.xlsx και αφορούσαν ηλεκτρονικές αγορές ειδών δώρων. Πιο συγκεκριμένα, η πρώτη μεταβλητή (InvoiceNo) υποδήλωνε έναν 6ψήφιο αριθμό που αντιστοιχούσε στον αριθμό του τιμολογίου που είχε εκδοθεί για την κάθε συναλλαγή του πελάτη. Το γράμμα ‘c’ στην αρχή κάθε τέτοιου κωδικού σημαίνει ότι πρόκειται για επιστροφή του προϊόντος. Στη δεύτερη στήλη (StockCode) ήταν αποτυπωμένος ο 5ψηφιος σειριακός αριθμός του εκάστοτε προϊόντος, η οποία περιείχε ένα 5ψήφιο κωδικό αριθμό, ενώ η τρίτη μεταβλητή περιείχε την περιγραφή του κάθε προϊόντος (Description). Η τέταρτη στήλη περιείχε την ποσότητα (Quantity) των προϊόντων ανά παραγγελία που είχαν αγοραστεί, η πέμπτη στήλη περιείχε την ημερομηνία και ώρα της αγοράς (InvoiceDate). Η έκτη στήλη περιείχε την τιμή μονάδος (UnitPrice) για το κάθε προϊόν, στην έβδομη στήλη ήταν καταγεγραμμένος ο μοναδικός αριθμός του κάθε πελάτη (CustomerID). Τέλος, η όγδοη μεταβλητή περιείχε το όνομα της χώρας (Country) από την οποία πραγματοποιήθηκε η αγορά.

Μεταβλητές του dataset:

1. **InvoiceNo**: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter ‘c’, it indicates a cancellation.
2. **StockCode**: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
3. **Description**: Product (item) name. Nominal
4. **Quantity**: The quantities of each product (item) per transaction. Numeric.
5. **InvoiceDate**: Invoice Date and time. Numeric, the day and time when each transaction was generated.
6. **UnitPrice**: UnitPrice. Numeric, Product price per unit in sterling.
7. **CustomerID**: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
8. **Country**: Country name. Nominal, the name of the country where each customer resides.

3.3. Καθαρισμός Δεδομένων και Επεξεργασία Δεδομένων

Ο καθαρισμός των δεδομένων είναι ένα από τα βασικότερα και πιο σημαντικά στάδια στη διαδικασία εξόρυξης δεδομένων για τις επιχειρήσεις. Ο τεράστιος όγκος δεδομένων που έχουν στην διάθεση τους οι επιχειρήσεις, είναι δεδομένα που πρέπει να υποστούν κάποια επεξεργασία για να μπορέσουν να χρησιμοποιηθούν και να αξιοποιηθούν. Πολλές φορές τα δεδομένα περιέχουν χαμένες τιμές (missing values), μηδενικές ή αρνητικές τιμές, κενά πεδία, διπλές καταχωρήσεις, εσφαλμένες τιμές, πράγμα που δημιουργεί προβλήματα κατά την επεξεργασία των δεδομένων και συνεπώς η πληροφορία που αποκτάται δεν είναι αξιόπιστη.

Στην παρούσα εργασία πραγματοποιήθηκε περιγραφική και διερευνητική ανάλυση των δεδομένων (EDA) για να συλλεχθούν απαραίτητες πληροφορίες. Μετά τον καθαρισμό των δεδομένων βρέθηκαν 396.356 παρατηρήσεις. Για τον καθαρισμό των δεδομένων χρησιμοποιήθηκε το Jupyter Notebook για τη συγγραφή του κώδικα. Αρχικά, με την βοήθεια της εντολής import γίνεται εισαγωγή των κατάλληλων βιβλιοθηκών για να μπορέσει να πραγματοποιηθεί αρχικά ο καθαρισμός των δεδομένων και στη συνέχεια να δημιουργηθούν διαγράμματα μέσω των αντίστοιχων βιβλιοθηκών. Οι βιβλιοθήκες που

χρησιμοποιήθηκαν ήταν η ***Pandas*** και η ***NumPy*** για την επεξεργασία των δεδομένων, η ***Matplotlib*** για την οπτικοποίηση των δεδομένων με τη δημιουργία γραφημάτων, η ***Sklearn*** για την πρόβλεψη και η ***Scipy*** για την εφαρμογή ελέγχων μέσω συναρτήσεων. Στη συνέχεια εισάγεται το αρχείο excel της βάσης δεδομένων. Με τη χρήση της εντολής df.shape αποδίδονται οι διαστάσεις του dataset, οι οποίες είναι 541.909 εγγραφές και 8 στήλες. Για να μπορέσει να πραγματοποιηθεί ο καθαρισμός θα πρέπει να μελετηθούν σε πρώτο στάδιο τα δεδομένα, ώστε να γίνουν τα σωστά βήματα στην επεξεργασία. Με την εντολή df.head() και df.tail() εμφανίζονται οι πρώτες δέκα (10) και οι τελευταίες πέντε (5) εγγραφές του dataset αντίστοιχα.

```
df.head(10)
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047.0	United Kingdom

Εικόνα 10: Απεικόνιση πρώτων εγγραφών

Στη συνέχεια, με την εντολή df.info(), παρέχονται πληροφορίες για το ποιες είναι οι μεταβλητές του dataset, ποιος είναι ο συνολικός αριθμός της κάθε στήλης καθώς και ο τύπος της κάθε μεταβλητής σε κάθε μια στήλη. Έπειτα, με την εντολή df.describe() εμφανίζονται τα στατιστικά στοιχεία του dataset ώστε να αποκτηθούν βασικές στατιστικές παράμετροι για το dataset, δηλαδή το πλήθος, η μέση τιμή, η τυπική απόκλιση, η μέγιστη και η ελάχιστη τιμή για κάθε στήλη και τα εκατοστημόρια αυτών. Από τα αποτελέσματα που προκύπτουν γίνεται κατανοητό ότι οι μεταβλητές Quantity και UnitPrice περιέχουν αρνητικές τιμές, που σημαίνει πως έχουν πραγματοποιηθεί επιστροφές προϊόντων. Οι τιμές αυτές αποτελούν «θόρυβο» για την επιχείρηση και θα πρέπει να αφαιρεθούν και αυτές. Επιπλέον, οι μεταβλητές Description και CustomerID περιέχουν ελλείπουσες τιμές – missing values, οι οποίες τιμές θα πρέπει ομοίως να αφαιρεθούν. Στις στήλες Quantity και UnitPrice παρατηρώντας την μέγιστη τιμή της κάθε μεταβλητής (max), είναι κατανοητό πως υπάρχουν ακραίες τιμές.

```
df.describe()
```

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

Εικόνα 11: Descriptive statistics

Στην συνέχεια αφαιρέθηκαν όλες οι ακυρωμένες παραγγελίες που υπήρχαν, παραγγελίες που εμφανίζονταν με αρνητικές τιμές στο Quantity και χαρακτηρίζονταν από το γράμμα «C» στον αριθμό τιμολογίου καθώς και όλες οι μηδενικές τιμές από τη στήλη UnitPrice που πιθανότατα χαρακτήριζαν δώρα ή ελαττωματικά προϊόντα. Δημιουργήθηκαν αντίστοιχα boxplots για να απεικονιστούν οι αρνητικές και μηδενικές τιμές και αφού πραγματοποιήθηκε η αφαίρεση των τιμών δημιουργήθηκαν εκ νέου boxplots για επιβεβαίωση. Τέλος, αποκλείστηκαν όλες οι τιμές από το StockCode που παρουσίαζαν τιμές όπως: D (discount), M (manual), S (Sample), C2 (Carriage), POST, AMAZONFEE (amazon fee) και BANK CHARGES. Επιπλέον εμφανίστηκε ο μοναδικός αριθμός των πελατών (4334), όπου θα είναι και ο αριθμός των πελατών που θα χρησιμοποιηθεί αργότερα στην ανάλυση RFM και ο μοναδικός αριθμός των χωρών (37).

```
df.isnull().sum()
```

```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID     135080
Country            0
dtype: int64
```

Εικόνα 12: Missing values in Descriptive-CustomerID

Παράλληλα με τη διαδικασία καθαρισμού των δεδομένων και μετά από αυτή, δημιουργήθηκαν κάποια διαγράμματα για την καλύτερη απεικόνιση και κατανόηση των

δεδομένων. Επιπλέον δημιουργήθηκαν δυο νέες μεταβλητές, η negative_values στην οποία αποθηκεύεται το πλήθος τον αρνητικών και μη αρνητικών τιμών της στήλης Quantity και η label στην οποία αποθηκεύεται τα ονόματα για τον χαρακτηρισμό των αγορών (non cancelled, cancelled). Από το διάγραμμα πίτας που δημιουργήθηκε, προέκυψε ότι το 2,19% των παραγγελιών είναι ακυρωμένες παραγγελίες. Ομοίως για τη δημιουργία του διαγράμματος πίτας για τα δώρα/ελαττωματικά προϊόντα, δημιουργήθηκε η μεταβλητή label_1 και η zero_values που περιείχαν τα labels (non zero, zero) για τον διαχωρισμό των τιμών και το πλήθος των μηδενικών και μη τιμών στο UnitPrice αντίστοιχα.

```
#Count number of negative values in Quantity and save it in a new variable #Cancelled orders 8905/406829 2.19%
negative_values = (df['Quantity'] < 0).value_counts()
negative_values
```

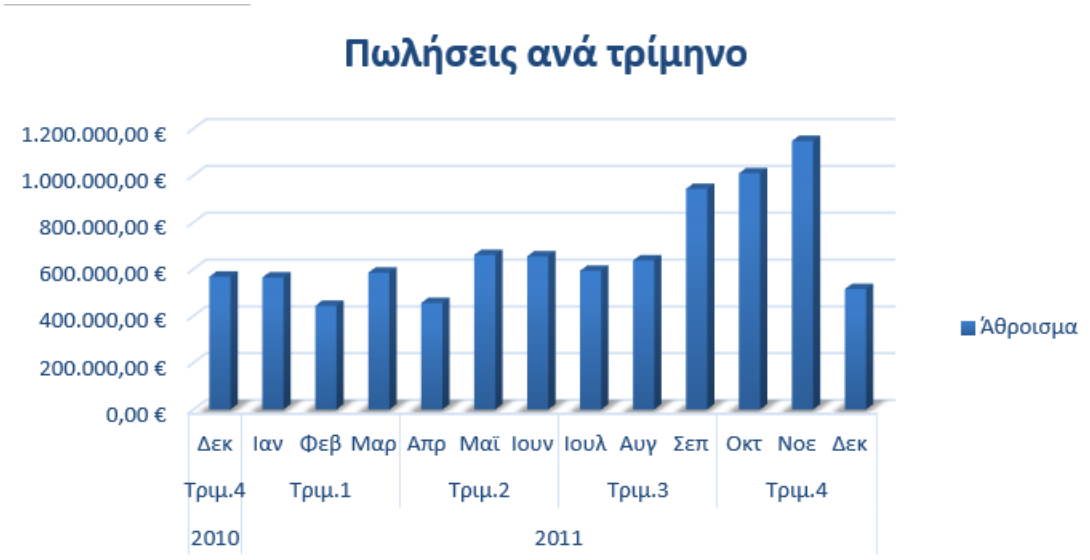
```
False    397924
True       8905
Name: Quantity, dtype: int64
```

```
#Define a new variable and store a list with the labels for the pie chart
label = ['Non cancelled: quantity > 0','Cancelled: quantity < 0']
```

Εικόνα 13: Create values to calculate cancelled orders

Τέλος, δημιουργήθηκαν ομαδοποιήσεις μεταξύ μεταβλητών για να μπορέσει να γίνει διερευνητική ανάλυση δεδομένων και δημιουργήθηκαν αντίστοιχα διαγράμματα για την απεικόνιση των αποτελεσμάτων. Ενδεικτικά αναφέρεται ότι υπολογίστηκαν οι καλύτεροι πελάτες της επιχείρησης, ο αριθμός των παραγγελιών ανά χώρα, τα συνολικά έσοδα ανά χώρα, οι πωλήσεις ανά μήνα, τα προϊόντα με την μεγαλύτερη ζήτηση. Επιπλέον κατά την διάρκεια της διερευνητικής ανάλυσης κρίθηκε αναγκαίο να αφαιρεθεί η ώρα από τη στήλη InvoiceDate, για να μπορέσουν να γίνουν διαγράμματα ανά μήνα ή μέρα. Από τις διαδικασίες, το τελικό dataset προς ανάλυση ήταν 4.334 μοναδικές εγγραφές. Στο διάγραμμα που ακολουθεί απεικονίζονται οι πωλήσεις ανά τρίμηνο μετά την εκκαθάριση των δεδομένων. Οι μήνες με τις υψηλότερες πωλήσεις φαίνεται οτι είναι ο Σεπτέμβριος, ο Οκτώβριος και ο Νοέμβριος, που εμφανίζουν μεγάλη αύξηση, με τον τελευταίο μήνα να έχει ξεκάθαρα της μεγαλύτερη αύξηση και συνεπώς τις περισσότερες πωλήσεις. Οι μήνες Μάρτιος και Απρίλιος αναμενόταν να έχουν μεγαλύτερα ποσά, λαμβάνοντας υπόψη τις επερχόμενες γιορτές του Πάσχα, όπου και πάλι η ζήτηση συνήθως είναι

αυξημένη. Τους υπόλοιπους μήνες υπάρχει μια σχετική ομοιομορφία με μικρές αυξομειώσεις.



Εικόνα 14: Πωλήσεις ανά τρίμηνο μετά την εκκαθάριση των δεδομένων

3.4. Εφαρμογή RFM Μοντέλου

Για να μπορέσει να πραγματοποιηθεί η ανάλυση RFM, ήταν απαραίτητο να δημιουργηθεί μια καινούργια μεταβλητή, TotalPrice, μετά των καθαρισμό των δεδομένων, η οποία προκύπτει από τον πολλαπλασιασμό της ποσότητας με το κόστος ανά τεμάχιο (Quantity*UnitPrice), ώστε να μπορέσει να υπολογισθεί αργότερα η μετρική Monetary της RFM Analysis.

```
df['TotalPrice'] = df['Quantity'] * df['UnitPrice']
df
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalPrice
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34

Εικόνα 15: Δημιουργία μεταβλητής TotalPrice

Η εφαρμογή του μοντέλου πραγματοποιήθηκε με τη βοήθεια του excel. Αρχικά στο υπολογιστικό φύλλο *RFM data* υπολογίσθηκε η διαφορά που προκύπτει από την *ημερομηνία της τελευταίας αγοράς* κάθε πελάτη από την *ημερομηνία αναφοράς* που ορίστηκε (diff_days) ώστε να μπορέσει να υπολογιστεί το recency. Σημειώνεται ότι ως ημερομηνία αναφοράς ορίστηκε η επόμενη ήμερα από την τελευταία αγορά (09/12/2011) που πραγματοποιήθηκε, δηλαδή 10/12/2011. Στο υπολογιστικό φύλλο *Sales per quarter* δημιουργήθηκε ένα PivotTable που υπολόγιζε τις πωλήσεις της επιχείρησης χωρισμένες σε τρίμηνα και δημιουργήθηκε ένα γράφημα για την απεικόνιση των πωλήσεων.

Στο υπολογιστικό φύλλο *PivotTable* δημιουργήθηκε ένας συγκεντρωτικός πίνακας με σκοπό να δημιουργηθούν οι τρεις μετρικές της RFM. Με την βοήθεια του συγκεντρωτικού πίνακα πραγματοποιήθηκε groupby με βάση το CustomerID, υπολογίζοντας τη μετρική recency, με την μεταβλητή diff_days (min), δηλαδή το μικρότερο χρονικό διάστημα από την τελευταία αγορά κάθε πελάτη, το frequency με την μεταβλητή InvoiceNo (count) το πλήθος δηλαδή των αγορών που πραγματοποίησε ο κάθε πελάτης και το monetary με την μεταβλητή TotalPrice (sum), το συνολικό χρηματικό ποσό που έχει ξοδέψει κάθε πελάτης από τις συναλλαγές του.

Για να μπορέσει η επιχείρηση να αντλήσει σημαντική πληροφορία από τις μετρικές αυτές θα πρέπει να ορίσει μια βαθμολογία στους πελάτες για κάθε μια μετρική ξεχωριστά, τα λεγόμενα rfm scores. Με τα scores η επιχείρηση μπορεί να πραγματοποιήσει τμηματοποίηση των πελατών ανάλογα με τις ανάγκες και τις καταναλωτικές τους συνήθειες διαχωρίζοντας τους στις αντίστοιχες ομάδες (segment) με βάση το rfm score που θα έχουν. Η μεθοδολογία που ακολουθήθηκε παρουσιάζεται παρακάτω. Με τη βοήθεια της συνάρτησης percentile.inc() δημιουργήθηκαν στο υπολογιστικό φύλλο του excel *RFM scores*, τρεις πίνακες με τα cutoffs των μετρικών ώστε να γίνει διακριτικοποίηση (discretization) στις τιμές αποκτήσουν οι πελάτες έναν αριθμό κατάταξης μέσα στον πίνακα ανάλογα με τις τιμές που διαθέτουν. Για τον διαχωρισμό των τμημάτων με ποσοστά χρησιμοποιήθηκαν τα 0, 20%, 40%, 60%, 80%, που αντίστοιχα αντιπαρατίθενται με τα 4 τεταρτημόρια της συνάρτησης quantile.inc(). Σημειώνεται ότι ο αριθμός 5 είναι η υψηλότερη τιμή και το 1 η χαμηλότερη τιμή.

Recency cutoffs			Frequency cutoffs			Monetary cutoffs	
1	5		1	1		4 €	1
15	4		14	2		249 €	2
33	3		29	3		485 €	3
72	2		58	4		925 €	4
180	1		121	5		2.043 €	5

Εικόνα 16: Διακριτικοποίηση μετρικών

Για τον εντοπισμό των R score, F score, M score έγινε χρήση της συνάρτησης VLOOKUP() για κάθε μεταβλητή ξεχωριστά, για να αναζητηθεί η τιμή της κάθε μεταβλητής μέσα στον αντίστοιχο πίνακα των cutoffs και να οριστεί σε ποιο τμήμα 1,2,3,4,5 κατατάσσεται ο πελάτης ανάλογα με την τιμή του. Για το RFM score το αποτέλεσμα προήλθε από το υπολογισμό: R*100+F*10+M. Τα segments που ορίστηκαν τελικά ήταν έξι (6) και ήταν τα εξής: *Hybernating, Can't Loose, Need Attention, New Customers, Loyal Customers και Champions*. Εφόσον επιλέχθηκαν οι ομάδες που θα χρησιμοποιηθούν έπρεπε να αντιστοιχιστούν οι τιμές των rfm scores στη σωστή ομάδα.

Για το χαρακτηρισμό του κάθε πελάτη στα αντίστοιχα segments που ορίστηκαν έγινε ξανά χρήση της συνάρτησης VLOOKUP() στη στήλη segment ώστε να μπορέσει να γίνει αναζήτηση της κάθε τιμής rfm score στο υπολογιστικό φύλλο *segments*, όπου βρίσκονται οι χαρακτηρισμοί των ομάδων, και να αντιστοιχιστούν με τα scores. Τα segments διαμορφώθηκαν ως εξής:

Segments	Scores
Champions	544, 545, 554, 555
Loyal Customers	342,343,344,345,353,354,355,431,432,433,434,435,443,444,445,453,454,455,531,523,533,534,551,535,543,553,442,452,532,542,552,541,441,423,424,425,451,524,525,,421,422,521,522
New Customers	414,415,511,512,513,514,515,411,412,413
Need attention	211,212,213,214,215,221,222,223,224,225,231,232,233,234,235,241,242,243,244,245,251,252,253,254,255,311,312,313,314,323,324,332,333,334,335,315,321,322,325,331,341,351,352
Can't lose	123,124,125,133, 134,135, 143, 144,145,153,154, 155,151,152,141,142,131,132
Hybernating	111, 112, 121, 122, 113, 114, 115

Εικόνα 17: Segments

Τέλος, με την βοήθεια κώδικα υπολογίστηκαν κάποια γραφήματα σχετικά με την κατανομή των πελάτων στα segments καθώς και τη σχέση τους με τις τρεις μετρικές.

3.5. K-means clustering

Η ομαδοποίηση των πελατών μπορεί να πραγματοποιηθεί και με διάφορα προβλεπτικά μοντέλα όπως είναι ο αλγόριθμος kmeans. Για να μπορέσει να εξεταστεί και αυτή η εναλλακτική αποφασίστηκε να χρησιμοποιηθεί και μια δεύτερη μεθοδολογία στην εργασία, αυτή της μηχανικής μάθησης (ML) ώστε να αποκτηθούν περισσότερες πληροφορίες και να συγκριθούν τα αποτελέσματα και με τα δυο μοντέλα για να εκτιμηθεί ποιο από τα δυο δίνει καλυτέρα αποτελέσματα για την επιχείρηση. Για να εφαρμοστεί ο αλγόριθμος μηχανικής μάθησης kmeans στην παρούσα εργασία ήταν απαραίτητη η προεπεξεργασία των δεδομένων.

Από τα περιγραφικά στοιχεία που υπολογίσθηκαν με τη βοήθεια του Jupyter, παρατηρήθηκε ότι η μέση τιμή και η τυπική απόκλιση διαφέρουν και οι κατανομές των μεταβλητών δεν ακολουθούν την κανονική κατανομή (ασύμμετρες κατανομές λοξές δεξιά), όπως φαίνεται και στα διαγράμματα που θα ακολουθήσουν. Το προγνωστικό μοντέλο απαιτεί τα δεδομένα να βρίσκονται στην ίδια κλίμακα για να μπορέσουν να εκπαιδευτούν μέσω του αλγορίθμου. Για να αντιμετωπιστεί το πρόβλημα αυτό έπρεπε να μετασχηματιστούν οι τιμές των μεταβλητών σε παρόμοιο εύρος ή να έχουν την ίδια μέση τιμή και τυπική απόκλιση, ώστε να συμπεριφέρεται καλύτερα το μοντέλο. Για το λόγο αυτό, αφού έγινε αρχικά μια διαγραμματική απεικόνιση των κατανομών για τις τρεις μετρικές ξεχωριστά, ελέγχθηκε με τη συνάρτηση skew() η λοξότητα που υπάρχει και έγινε το τεστ κανονικότητας Shapiro-Wilk, με τη βοήθεια του shapiro() τεστ για να ελεγχθεί η κανονικότητα των δεδομένων. Το τεστ Shapiro-Wilk ελέγχει την μηδενική υπόθεση ότι τα δεδομένα είναι κανονικά κατανεμημένα. Αν η p-value είναι μικρότερη του alpha level (5%), τότε η μηδενική υπόθεση απορρίπτεται, καθώς υπάρχουν επαρκή στοιχεία ότι τα δεδομένα δεν ακολουθούν την κανονική κατανομή. Ενώ αν p-value μεγαλύτερο του 0.05, τότε η μηδενική υπόθεση ότι τα δεδομένα είναι κανονικά κατανεμημένα γίνεται αποδεκτή.

Στη συνέχεια θεωρήθηκε χρήσιμο να ελεγχθούν οι συσχετίσεις μεταξύ των τριών μετρικών. Παρακάτω, για να μπορέσει να αντιμετωπιστεί η λοξότητα που βρέθηκε, έγινε λογαριθμοποίηση των δεδομένων με τη χρήση της np.log, ώστε τα δεδομένα να προσεγγίζουν όσο το δυνατόν περισσότερο την κανονική κατανομή και ακολούθησαν διαγραμματικές απεικονίσεις σχετικά με τις συσχετίσεις των μεταβλητών, ώστε να κατανοηθεί πως συμπεριφέρονται τα δεδομένα. Στη συνέχεια έγινε κανονικοποίηση

(standardization) των λογαριθμοποιημένων τιμών με τη χρήση του StandardScaler(), για να μπορέσει να αντιμετωπιστεί το πρόβλημα με τη μέση τιμή και την τυπική απόκλιση.

```
scale = StandardScaler()
scale.fit(logdata)
normalized = scale.transform(logdata)
normalized
```

Εικόνα 18: Standardization

Για την εφαρμογή του kmeans clustering αλγόριθμου, αρχικά αποφασίστηκε η δημιουργία πέντε clusters και εφαρμόστηκε ο αλγόριθμος στα κανονικοποιημένα δεδομένα. Για την δημιουργία των clusters συντάχθηκε ο παρακάτω κώδικας στον οποίο εφαρμόζεται ο αλγόριθμος kmeans με αριθμό συστάδων πέντε, γίνεται εφαρμογή του αλγορίθμου στα δεδομένα και πρόβλεψη του cluster στο οποίο ανήκει ο κάθε πελάτης. Τέλος δημιουργήθηκαν τα κεντροειδή των πέντε clusters. Για να θεωρηθεί ότι η επιλογή στο αριθμό των clusters ήταν σωστή, επιλέχθηκε μια από τις δημοφιλέστερες μεθόδους, η μέθοδος του Αγκώνα, γνωστή και ως Elbow Method, η οποία προσδιορίζει τον βέλτιστο αριθμό των συστάδων. Ορίστηκε ένα εύρος τιμών για τον αριθμό των συστάδων και στη συνέχεια η συνθήκη ώστε να βρεθεί ο αριθμός των βέλτιστων clusters. Τέλος, γίνεται η γραφική αναπαράσταση της μεθόδου. Παρακάτω παρουσιάζεται ο κώδικας που γράφτηκε για την εφαρμογή της.

```
kmeans = KMeans(n_clusters=5)
kmeans.fit(normalized)
predict = kmeans.predict(normalized)
df['k_clusters'] = predict
kmeans.cluster_centers_
```

Εικόνα 19: kmeans 5 clusters

```
sse = []
r = range(1, 10)
for k in r :
    kmeans = KMeans(n_clusters = k)
    kmeans.fit(norm)
    sse.append(kmeans.inertia_)

plt.figure(figsize = (10,6))
plt.plot(r, sse, marker = 'o')
plt.xlabel('Number of Clusters')
plt.ylabel('Sum of Squares Error (SSE)')
plt.title('Elbow for kmeans clustering')
plt.show()
```

Εικόνα 20: Elbow method

Τα αποτελέσματα της μεθόδου υπέδειξαν τη δημιουργία τριών clusters, οπότε έπρεπε να εφαρμοστεί ξανά ο αλγόριθμος για να βγουν τα νέα αποτελέσματα. Παρακάτω φαίνεται ο νέος κώδικας.

```
kmeans = KMeans(n_clusters=3)
kmeans.fit(normalized)
predict = kmeans.predict(normalized)
df['k_clusters'] = predict
centroids = kmeans.cluster_centers_
centroids
```

Εικόνα 21: kmeans 3 clusters

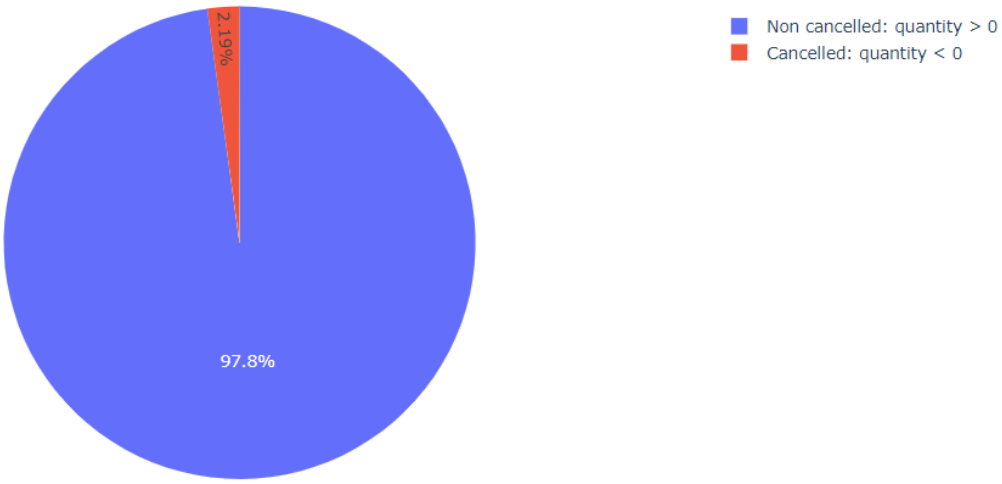
Τέλος, εμφανίστηκε το πλήθος των πελάτων σε κάθε cluster, οι μέσες τιμές των τριών μετρικών ανά cluster και διαγράμματα για τα cluster, τις συσχετίσεις και τις κατανομές μετά την κανονικοποίηση των δεδομένων.

Η σύγκριση των αποτελεσμάτων των δυο μοντέλων πραγματοποιήθηκε στο excel. Τα αποτελέσματα αφορούσαν την σύγκριση των δυο αναλύσεων που πραγματοποιήθηκαν για να μπορέσει η επιχείρηση να κατανοήσει ποια από τις δυο είναι η καλύτερη. Η αριστερή στήλη απεικόνιζε την rfm ανάλυση, όπου οι πελάτες ήταν χωρισμένοι σε segments ανάλογα με το rfm score που είχαν. Οι καλύτεροι πελάτες της επιχείρησης ήταν οι πελάτες με rfm score *555* (*champions*) ενώ οι χειρότεροι, αυτοί με τον αριθμό *111 (hybernating)*. Στη συνέχεια για να μπορέσει να γίνει η σύγκριση με τα k-cluster, διαχωρίστηκαν και κωδικοποιήθηκαν όλες οι τιμές με 0-2 με τη βοήθεια της συνάρτησης VLOOKUP(). Θεωρήθηκε ότι οι hybernating και can’t lose πελάτες θα αποτελούν μαζί μια ομάδα, καθώς έχουν χαμηλότερες τιμές στις μετρικές από τα άλλα segments και θα

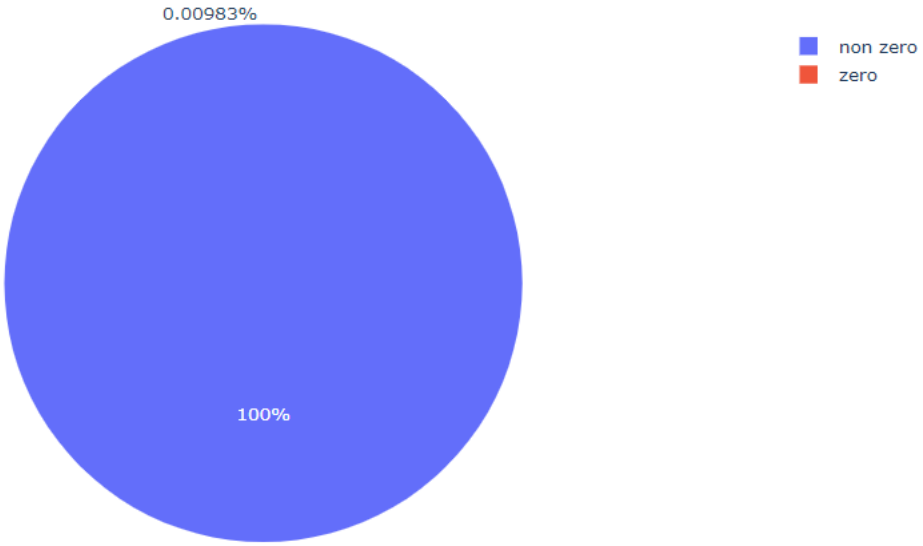
έχουν τον κωδικό 0. Οι need attention και new customers θα έχουν τον κωδικό 1 και οι champions και loyal customers θα αποτελούν μια ομάδα με κωδικό 2, καθώς περιέχουν αρκετά υψηλές τιμές στις τρεις μετρικές. Η δεξιά στήλη απεικόνιζε τα αποτελέσματα από τον αλγόριθμο kmeans, όπου οι πελάτες διαχωρίστηκαν σε cluster. Οι καλύτεροι πελάτες της επιχείρησης ήταν αυτοί που ανήκαν στο cluster *high value*, ενώ οι χειρότεροι πελάτες ανήκαν στο cluster *low value.* Το *mid value* cluster περιείχε πελάτες με μέση αξία συναλλαγής, μέση συχνότητα και οι παραγγελίες τους δεν ήταν ούτε πρόσφατες αλλά ούτε και πολύ παλιές. Η αντιστοιχία του κάθε πελάτη στο αντίστοιχο cluster πραγματοποιήθηκε με τη VLOOKUP(), αναζητώντας κάθε μια τιμή στον πίνακα με τα ονόματα των clusters. Η στήλη compare με τη χρήση της συνάρτησης if() σύγκρινε τα δυο αποτελέσματα και εμφάνιζε τη σύγκριση μεταξύ των δυο μοντέλων, αν δηλαδή ταυτίζονται ή όχι.

Κεφάλαιο 4: Αποτελέσματα

Τα δεδομένα που εξετάστηκαν στο συγκεκριμένο dataset υποβλήθηκαν σε καθαρισμό ώστε να αφαιρεθούν τιμές που αποτελούσαν θόρυβο για τα δεδομένα και συνεπώς θα επηρεάζονταν άμεσα και τα αποτελέσματα. Έτσι, μαζί με τις ελλείπουσες τιμές (missing values) και κάποιες ειδικές μεταβλητές, αφαιρέθηκαν και τιμές που αντιστοιχούσαν σε 2.19% ακυρωμένες παραγγελίες και 0.0098% προωθητικά δώρα.

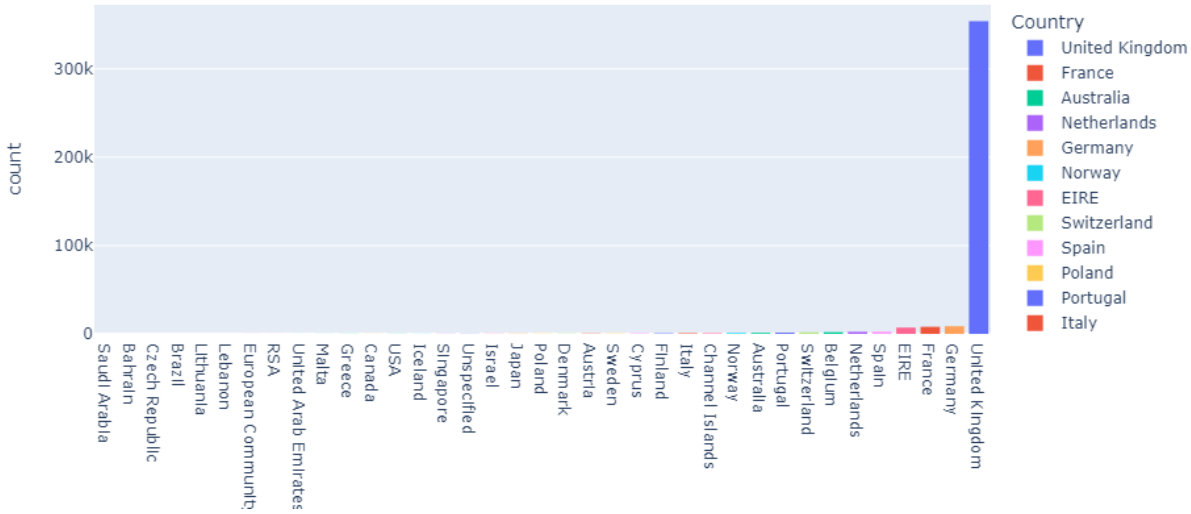


Διάγραμμα 1: Ποσοστό ακυρωμένων παραγγελιών



Διάγραμμα 2: Ποσοστό προωθητικών δώρων ή ελαττωματικών προϊόντων

Από την ανάλυση που πραγματοποιήθηκε στο συγκεκριμένο dataset παρατηρήθηκε ότι η πλειοψηφία του αριθμού των παραγγελιών (354.004) συγκεντρώνεται στη χώρα United Kingdom, η οποία είναι και η χώρα με τα περισσότερα συνολικά έσοδα. Αυτή η μεγάλη συγκέντρωση στη συγκεκριμένη χώρα μπορεί να οφείλεται στο γεγονός ότι η συγκεκριμένη επιχείρηση εδρεύει στην Αγγλία, οπότε και η προτίμηση του καταναλωτικού κοινού είναι περισσότερο αυξημένη. Ακολούθως τα επόμενα μεγαλύτερα αποτελέσματα σε σχέση με τις πωλήσεις ανήκουν στην Γερμανία με 8.658 πωλήσεις, στην Γαλλία με 8.034 και ακολούθως στην Ιρλανδία (EIRE) με 7.136 αγορές.



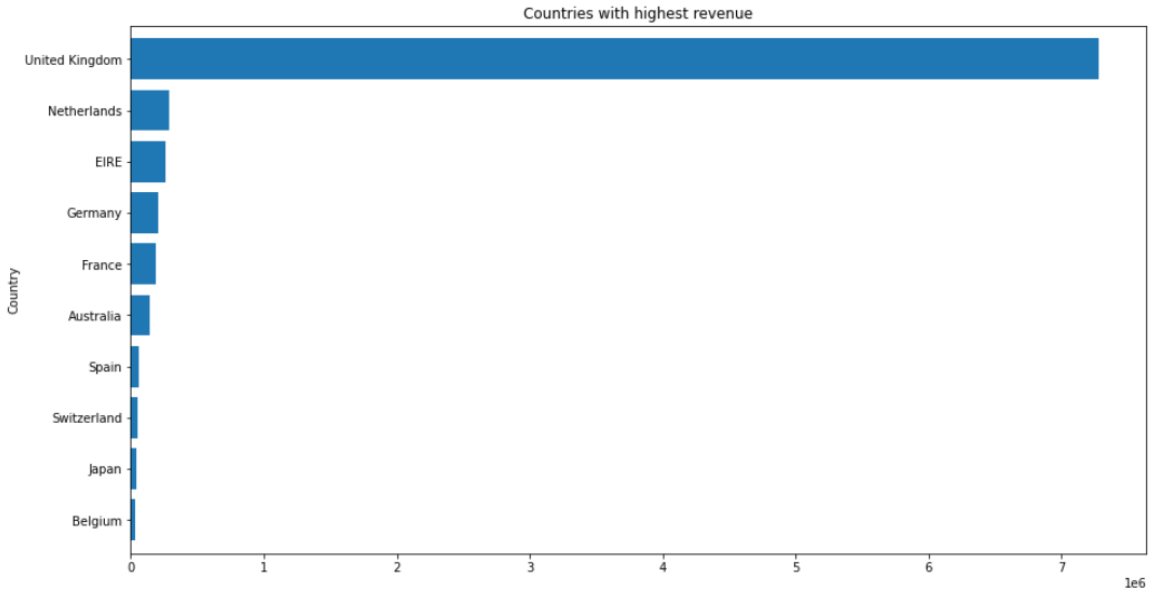
Διάγραμμα 3: Σύνολο παραγγελιών ανά χώρα

Country	
United Kingdom	354004
Germany	8658
France	8034
EIRE	7136
Spain	2422
Netherlands	2322
Belgium	1935
Switzerland	1810

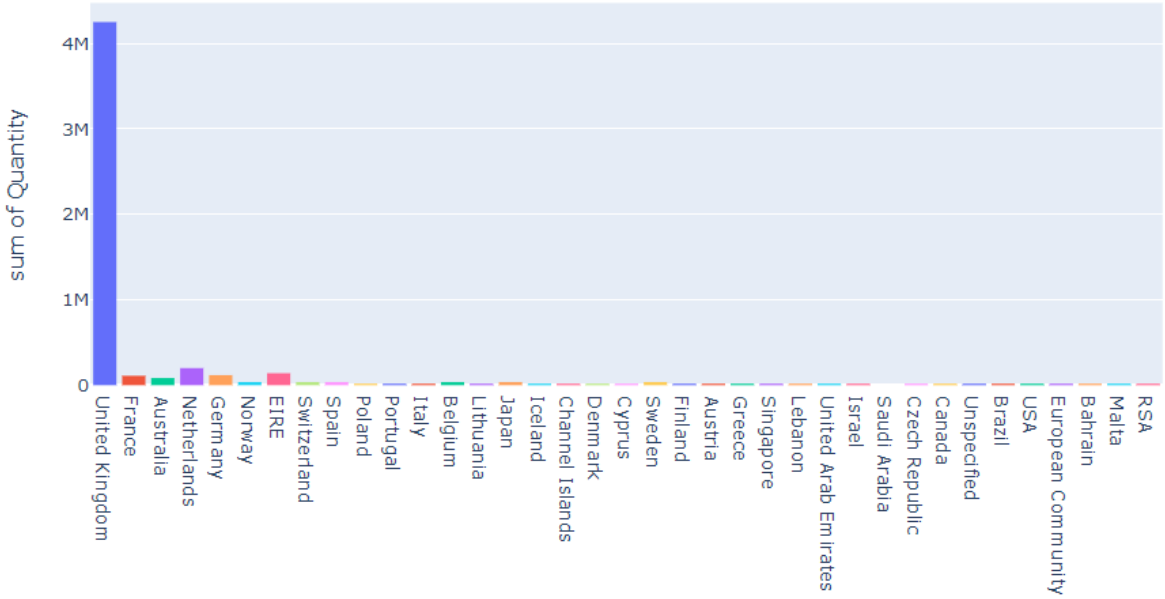
Εικόνα 22: Κατάταξη χωρών με βάση τις πωλήσεις

Στο κάτωθι διάγραμμα απεικονίζεται η κατάταξη των χωρών (37 μοναδικές χώρες) με βάση τα συνολικά έσοδα που είχαν κατά το έτος 2010-2011. Η χώρα που συγκέντρωσε

τη μεγαλύτερη συνολική αξία αγορών (7.277.768.59£) είναι το Ηνωμένο Βασίλειο. Ακολουθεί η Ολλανδία με 283.889£ και η Ιρλανδία με 257.296£. Ομοίως και για τις συνολικές ποσότητες των παραγγελιών ο μεγαλύτερος όγκος πωλήσεων παρατηρείται στο Ηνωμένο Βασίλειο, την Ολλανδία, την Ιρλανδία και την Γερμανία με 4.249.526, 200.258, 140.175 και 118.139 προϊόντα αντίστοιχα, ενώ οι χώρες που εμφάνισαν τα λιγότερα νούμερα όσον αφορά τον όγκο πωλήσεων είναι οι χώρες Bahrain και Saudi Arabia με την τελευταία να έχει μόλις 80 προϊόντα συνολικά κατά τη διάρκεια του έτους.



Διάγραμμα 4: Best countries by TotalPrice



Διάγραμμα 5: Best Countries by Quantity

Επιπλέον, από τα περιγραφικά στοιχεία παρατηρείται ότι, η μέση τιμή της ποσότητας των προϊόντων μετά τον καθαρισμό των δεδομένων είναι 13 τεμάχια με μέγιστη ποσότητα 80.995 τεμάχια, ενώ η μέση τιμή των εσόδων της επιχείρησης είναι 22£, με τυπική απόκλιση 309£. Επιπλέον η στήλη Quantity και TotalPrice παρουσιάζουν ακραίες τιμές.

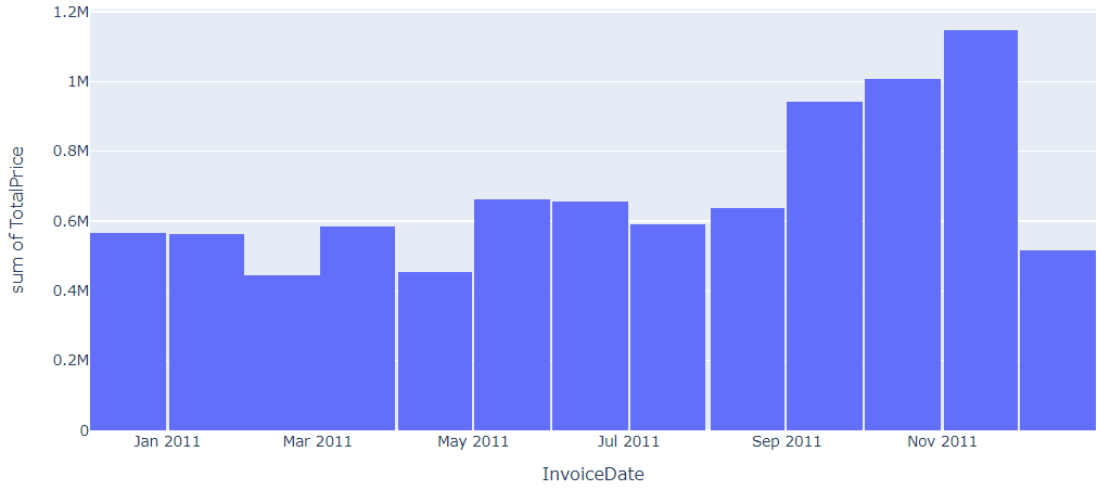
	Quantity	UnitPrice	CustomerID	TotalPrice
count	396356.000000	396356.000000	396356.000000	396356.000000
mean	13.011972	2.898124	15301.640399	22.134074
std	179.636958	7.171965	1709.940131	308.943701
min	1.000000	0.001000	12346.000000	0.001000
25%	2.000000	1.250000	13975.000000	4.680000
50%	6.000000	1.950000	15159.000000	11.800000
75%	12.000000	3.750000	16803.000000	19.800000
max	80995.000000	1599.260000	18287.000000	168469.600000

Εικόνα 23: Περιγραφικά στοιχεία για Quantity, UnitPrice και TotalPrice

Το επόμενο διάγραμμα έγινε για να μπορέσει να κατανοήσει η επιχείρηση καλύτερα το πως κατανέμονται οι πωλήσεις της κατά την διάρκεια των μηνών. Παρατηρείται ξεκάθαρα μια σημαντική αύξηση τους μήνες Σεπτέμβριο με 942.582£ και όγκο πωλήσεων 544.076 τεμάχια, Οκτώβριο με 1.009.051£ και 592.296 τεμάχια και Νοέμβριο με 1.147.351£ και 664.677 τεμάχια προϊόντων. Συγκεκριμένα, το μήνα Νοέμβριο οι πωλήσεις εκτινάσσονται, συγκεντρώνοντας διπλάσιες πωλήσεις και σχεδόν διπλάσιες ποσότητες προϊόντων από τους πρώτους 9 μήνες. Η ξαφνική αυτή αύξηση πιθανότατα οφείλεται στην περίοδο των εκπτώσεων και τις επερχόμενες γιορτές των Χριστουγέννων και της Πρωτοχρονιάς, όπου παρατηρείται πάντα αύξηση της ζήτησης. Τους υπόλοιπους μήνες οι πωλήσεις κυμαίνονται περίπου στα ίδια επίπεδα.

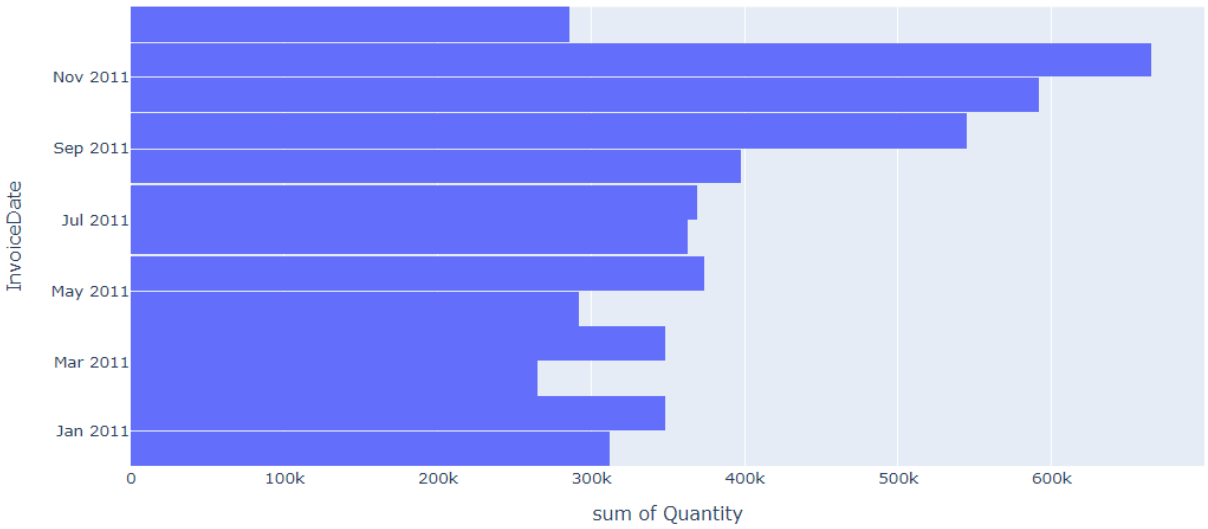
InvoiceDate	Quantity	UnitPrice	CustomerID	TotalPrice
2010-12-31	311840	79955.160	404811804.0	567490.720
2011-01-31	348803	65158.710	320635913.0	564026.640
2011-02-28	265380	61313.280	304208993.0	443346.020
2011-03-31	348164	81394.320	411056140.0	584562.850
2011-04-30	291973	66912.851	344736863.0	455266.911
2011-05-31	373325	87118.370	429644366.0	660481.900
2011-06-30	363410	81668.920	415244123.0	654432.560
2011-07-31	369114	71709.941	409239025.0	592731.901
2011-08-31	397848	74057.510	410641894.0	636818.410
2011-09-30	544076	112878.311	608798185.0	942582.641
2011-10-31	592296	142368.940	753419887.0	1009051.660
2011-11-30	664677	178641.300	988271689.0	1147351.110
2011-12-31	286467	45511.230	264188100.0	514829.690

Εικόνα 24: Συνολικές πωλήσεις και όγκος πωλήσεων ανά μήνα



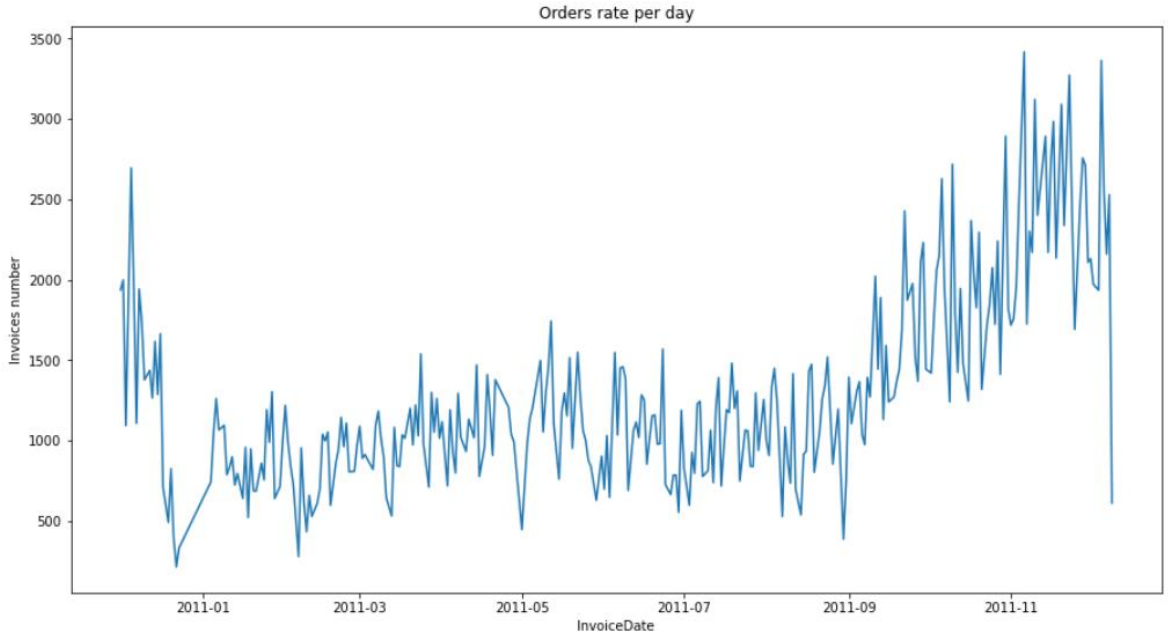
Διάγραμμα 6: Συνολικές πωλήσεις ανά μήνα

Ομοίως και σε ότι αφορά των όγκο πωλήσεων των προϊόντων, αυτός φαίνεται να είναι αυξημένος τους ίδιους μήνες με τους οποίους παρατηρείται και η αύξηση στα συνολικά έσοδα της επιχείρησης, όπως αυτό είναι αναμενόμενο με τον Νοέμβριο να συγκεντρώνει και πάλι το μεγαλύτερο αριθμό τεμαχίων (664.677).



Διάγραμμα 7: Συνολική ποσότητα προϊόντων ανά μήνα

Από τις ημερήσιες πωλήσεις παρατηρείται ότι καθ'όλη την διάρκεια των ημερών σημειώνονται αυξομειωτικές τάσεις, με την μεγαλύτερη αυξητική πορεία να εμφανίζεται το τελευταίο τρίμηνο πριν την περίοδο των Χριστούγεννων.

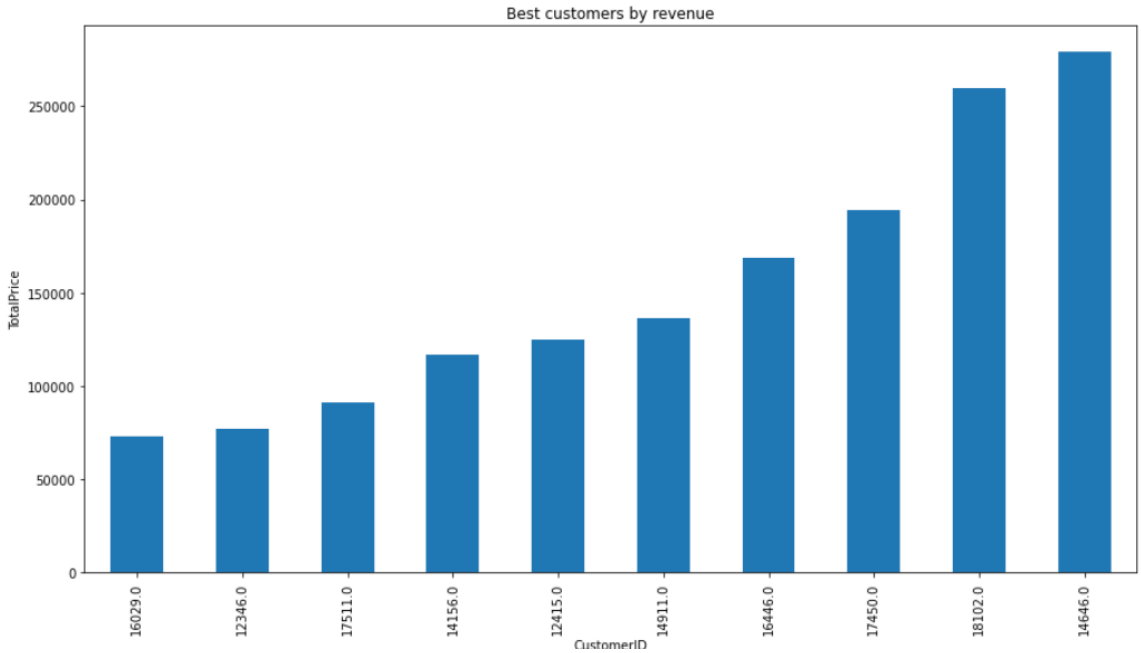


Διάγραμμα 8: Ημερήσιες πωλήσεις

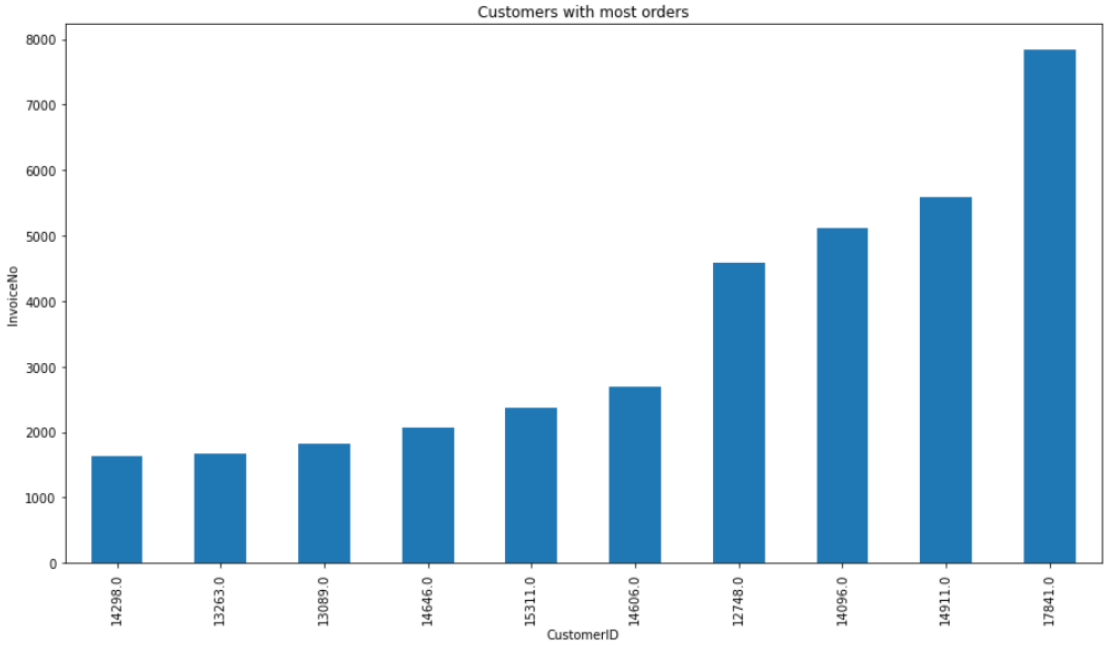
Στη συνέχεια αποτυπώνονται οι δέκα καλύτεροι πελάτες της επιχείρησης με βάση τα συνολικά έσοδα, οι δέκα καλύτεροι που πραγματοποίησαν τις περισσότερες αγορές μέσα στο έτος καθώς και τα προϊόντα με τη μεγαλύτερη ζήτηση. Έτσι, η επιχείρηση μπορεί να πραγματοποιήσει μελλοντικά περαιτέρω αναλύσεις και στρατηγικές για αυτούς τους πελάτες ή ομάδες πελατών και να εφαρμόσει κατάλληλες προωθητικές ενέργειες.

```
CustomerID
16029.0     72882.09
12346.0     77183.60
17511.0     91062.38
14156.0    116729.63
12415.0    124564.53
14911.0    136275.72
16446.0    168472.50
17450.0    194550.79
18102.0    259657.30
14646.0    279138.02
Name: TotalPrice, dtype: f
```

Εικόνα 25: Καλύτεροι πελάτες με βάση την τιμή



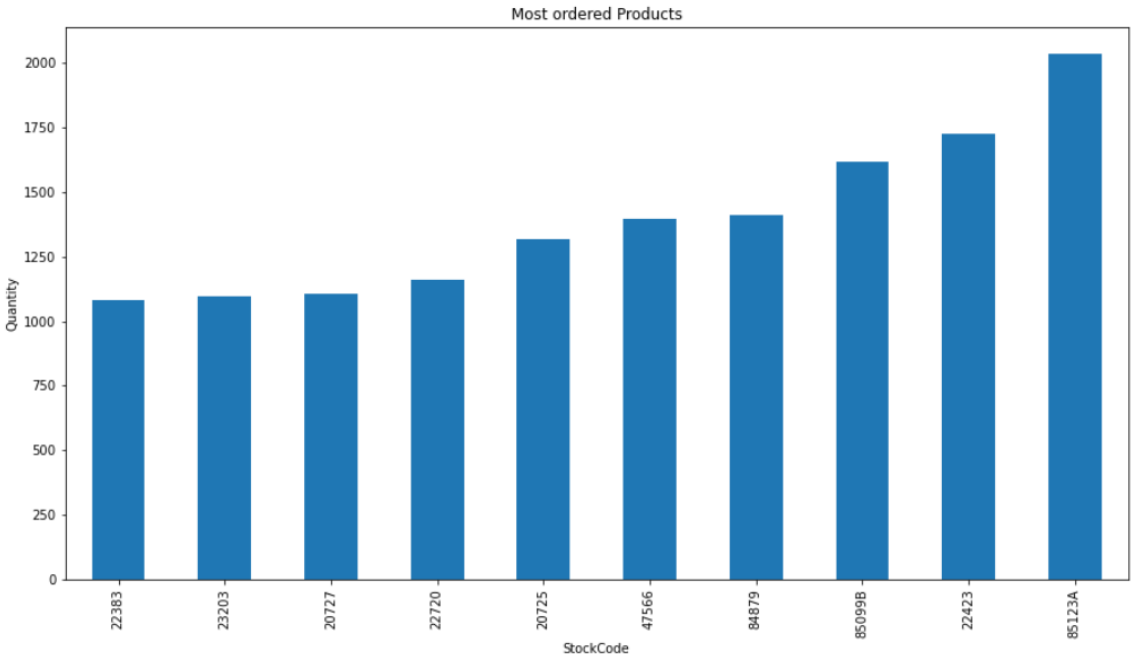
Διάγραμμα 9: Best customers by revenue



Διάγραμμα 10: Most frequent customers

```
StockCode
22383     1083
23203     1098
20727     1105
22720     1159
20725     1317
47566     1396
84879     1408
85099B    1618
22423     1723
85123A    2035
Name: Quantity, dtype: int64
```

Εικόνα 26: Stockcodes quantity



Διάγραμμα 11: Most ordered products

Στη συνέχεια, τα παρακάτω διαγράμματα προέκυψαν μετά από την εφαρμογή του μοντέλου RFM στο excel. Αφού κατατάχθηκαν οι τιμές σε βαθμίδες, ώστε να κατηγοριοποιηθούν οι τιμές με βάση την τιμή τους, στη συνέχεια βρέθηκαν οι μεταβλητές r score, f score, m score και rfm score για να γίνει η τμηματοποίηση των πελατών στο αντίστοιχο segment. Για χάρη της εργασίας θεωρήθηκαν ως καλύτεροι πελάτες οι 555 και χειρότεροι οι 111 αλλά στις αντίστοιχες συστάδες ορίστηκαν περισσότεροι από ένας συνδυασμοί. Τα αποτελέσματα από την ομαδοποίησή τους φαίνεται παρακάτω:

R score	F score	M score	RFM score	Segment
1	1	5	115	Hybernating
5	5	5	555	Champions
2	2	4	224	Need attention
4	4	4	444	Loyal Customers
1	2	2	122	Hybernating
3	4	4	344	Loyal Customers
1	1	1	111	Hybernating
1	4	4	144	Can't lose
1	1	2	112	Hybernating
4	4	5	445	Loyal Customers
3	5	5	355	Loyal Customers
5	2	4	524	Loyal Customers
3	5	5	355	Loyal Customers
3	5	5	355	Loyal Customers
1	1	1	111	Hybernating
5	5	5	555	Champions
2	2	3	223	Need attention
5	4	4	544	Champions
1	2	2	122	Hybernating
5	1	1	511	New Customers
3	5	5	355	Loyal Customers
3	4	4	344	Loyal Customers
2	3	4	234	Need attention
1	1	2	112	Hybernating
4	3	3	433	Loyal Customers

Εικόνα 27: Δημιουργία r, f, m scores και rfm total score – Customer segment

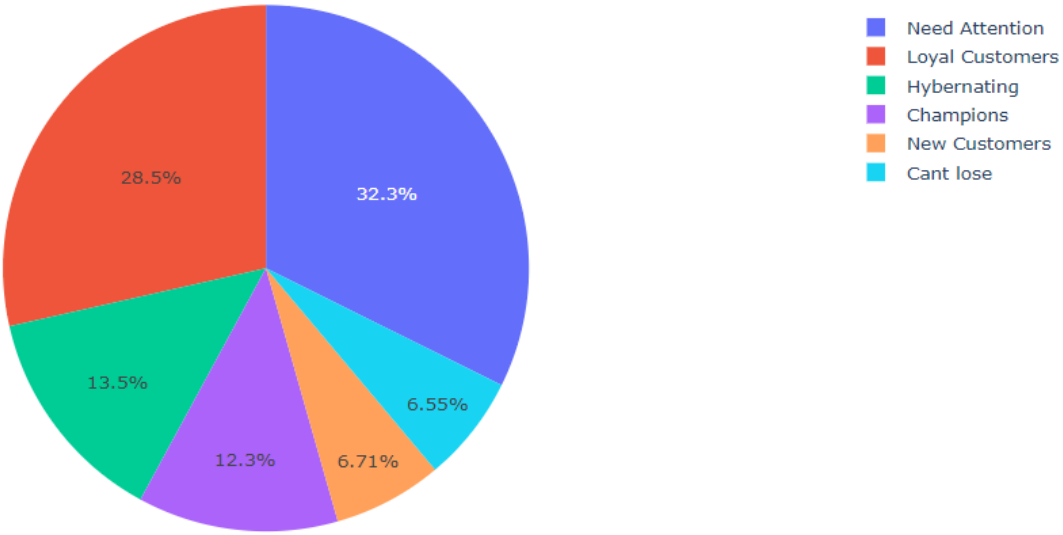
Από τα αποτελέσματα της ανάλυσης παρατηρείται ότι το μεγαλύτερο ποσοστό (32,3%) των πελάτων κατανέμεται στην ομάδα Need Attention με 1401 πελάτες και μέση συχνότητα αγορών 39 αγορές, μέση αξία συναλλαγής 775£ και η τελευταία τους αγορά πραγματοποιήθηκε 91 ημέρες πριν από την ημερομηνία αναφοράς που ορίστηκε 10/12/2011. Οι loyal customers κατέχουν το αμέσως επόμενο μεγαλύτερο ποσοστό 28,5% του συνόλου των πελατών με 1237 πελάτες, μέση αξία συναλλαγής 2.174£, 120 αγορές και η τελευταία αγορά πραγματοποιήθηκε 27 ημέρες πριν από την ημερομηνία αναφοράς. Ακολουθεί το cluster Hybernating που καταλαμβάνει το 13,5% με μέση συχνότητα αγορών 11 παραγγελίες, 547£ αξία συναλλαγής, 587 πελάτες και 273 ημέρες απέχει η τελευταία τους αγορά από την ημερομηνία αναφοράς. Οι πελάτες αυτοί είναι οι χειρότεροι πελάτες της επιχείρησης. Ακολουθεί το cluster με τους champions πελάτες, που είναι η ομάδα με τους καλύτερους πελάτες της επιχείρησης. Το cluster αυτό καταλαμβάνει ποσοστό 12,3%, 534 πελάτες, η τελευταία τους αγορά έγινε 6 ημέρες πριν από την ημερομηνία αναφοράς και η μέση αξία συναλλαγών τους είναι 7.755£ με μέση συχνότητα αγορών 311 συναλλαγές. Στο cluster, new customers, εντάσσεται το 6,71% των πελατών με 291 άτομα, οι οποίοι συγκεντρώνουν αρκετά υψηλή μέση αξία συναλλαγής 1021£, με 13 συναλλαγές και η τελευταία συναλλαγή τους έγινε 17 ημέρες πριν την ημερομηνία αναφοράς που είναι η 10/12/2011. Στο τελευταίο cluster can’t lose

εντάσσονται 284 πελάτες, οι οποίοι κατέχουν και αυτοί ένα πολύ μικρό ποσοστό, 6,55% του συνόλου των πελατών έχοντας αρκετά υψηλή μέση αξία συναλλαγής, 839£ με 54 αγορές και η τελευταία αγορά τους έγινε 260 ημέρες πριν από την ημερομηνία αναφοράς.

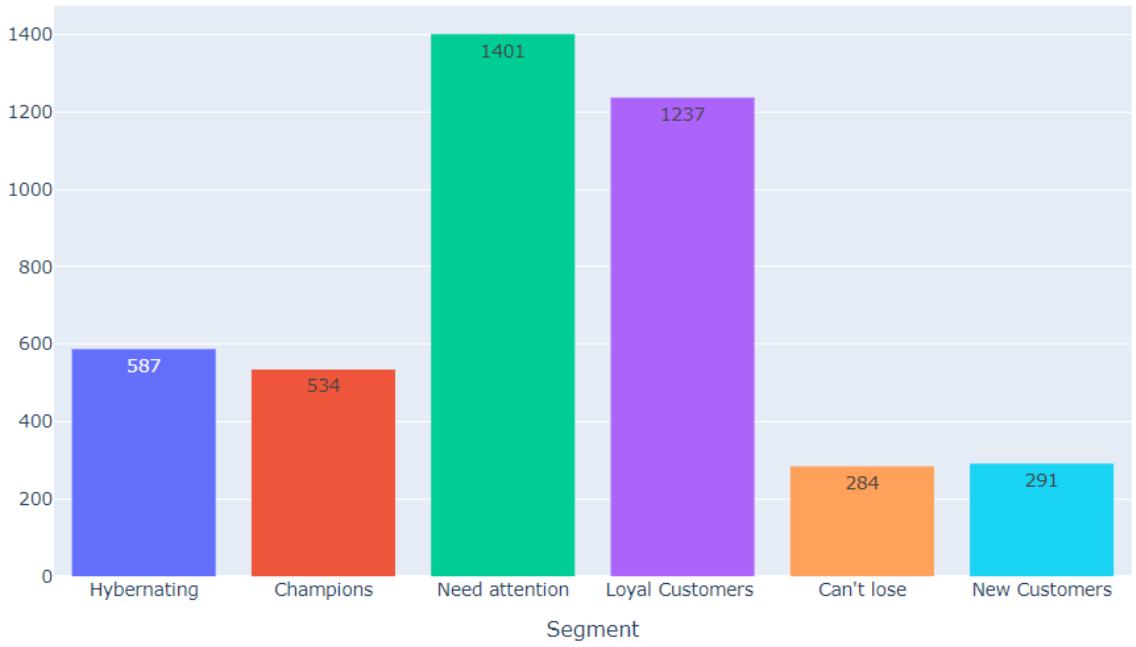
```
Need attention      1401
Loyal Customers     1237
Hybernating          587
Champions            534
New Customers        291
Can't lose           284
Name: Segment, dtype: int64
```

Segment	Recency	Frequency	Monetary
Can't lose	260.049296	54.228873	838.812820
Champions	6.211610	311.406367	7754.730225
Hybernating	273.107325	11.833049	547.276286
Loyal Customers	27.447858	120.639450	2174.284842
Need attention	91.098501	38.992148	774.807896
New Customers	17.103093	13.261168	1021.913540

Εικόνα 28: Πλήθος πελατών ανά segment - Mean R, F, M metrics by segment

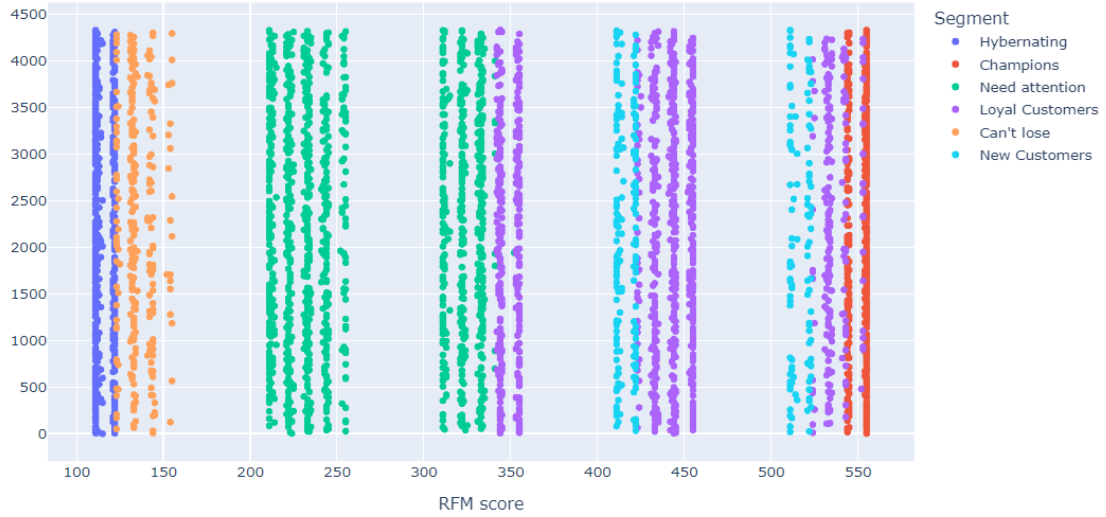


Διάγραμμα 12: Ποσοστά κατανομής ανά segment

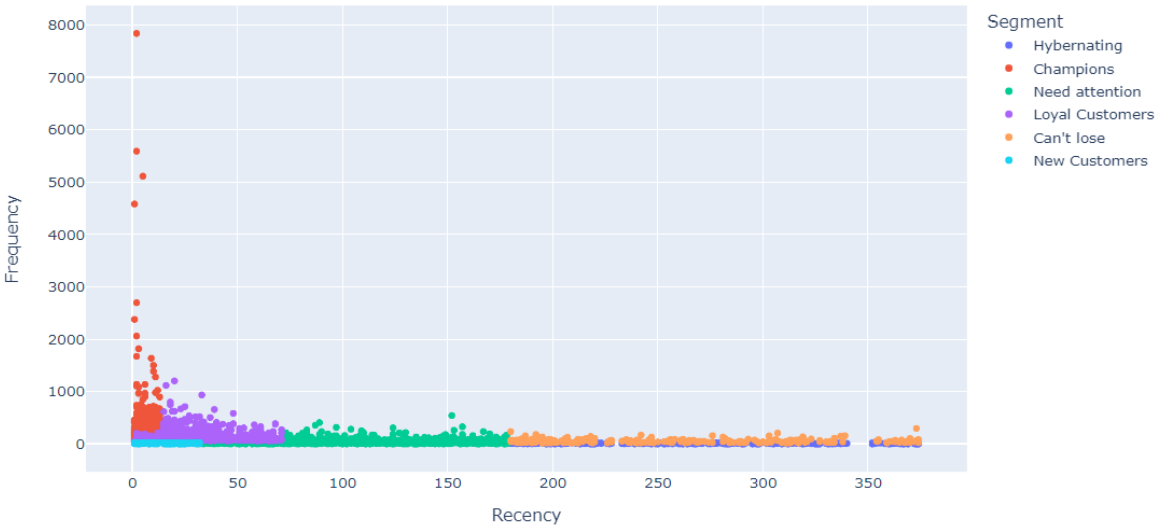


Διάγραμμα 13: Πλήθος πελατών ανά cluster

Στα παρακάτω διαγράμματα απεικονίζεται ο διαχωρισμός των segments σε σχέση με τη μέση συχνότητα αγορών και πρόσφατης και σε σχέση με όλες τις μετρικές μαζί.

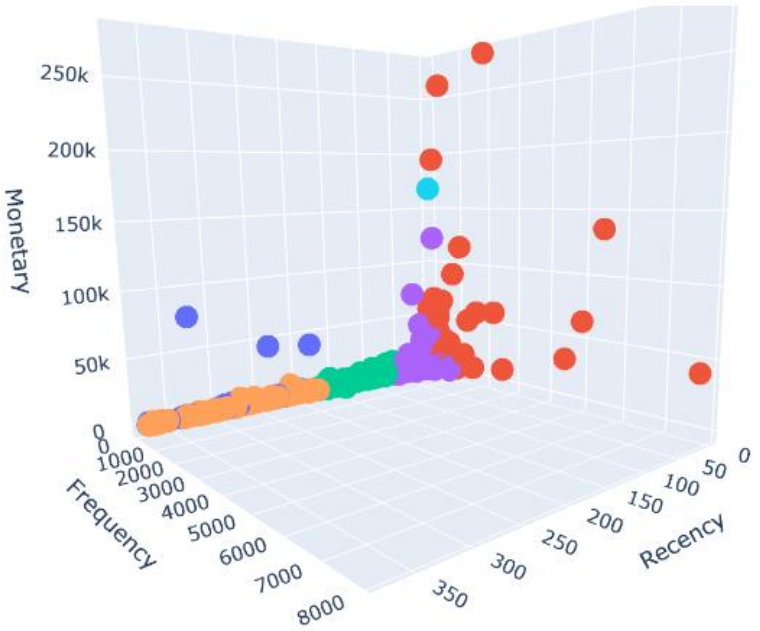


Διάγραμμα 14: Segments



Διάγραμμα 15: Segments by recency – frequency

Όπως παρατηρείται οι champions πελάτες συγκεντρώνονται κοντά στην αρχή των αξόνων και φτάνουν οι τιμές τους αρκετά υψηλά, καθώς οι ημέρες από την τελευταία αγορά τους σε σχέση με την ημερομηνία είναι λίγες και το ποσό συναλλαγής τους είναι υψηλό. Αντίθετα οι hybernating πελάτες μαζί με τους can't lose συγκεντρώνονται πάντα στην εξωτερική πλευρά του άξονα x, καθώς έχουν μεγάλο διάστημα να πραγματοποιήσουν αγορές (recency υψηλό) και κοντά στην αρχή του άξονα y καθώς δεν πραγματοποιούν συχνά αγορές. Ομοίως σχετικά με την αξία συναλλαγής το ποσό για το segment αυτό είναι και πάλι πολύ χαμηλά, κοντά στην αρχή των αξόνων.



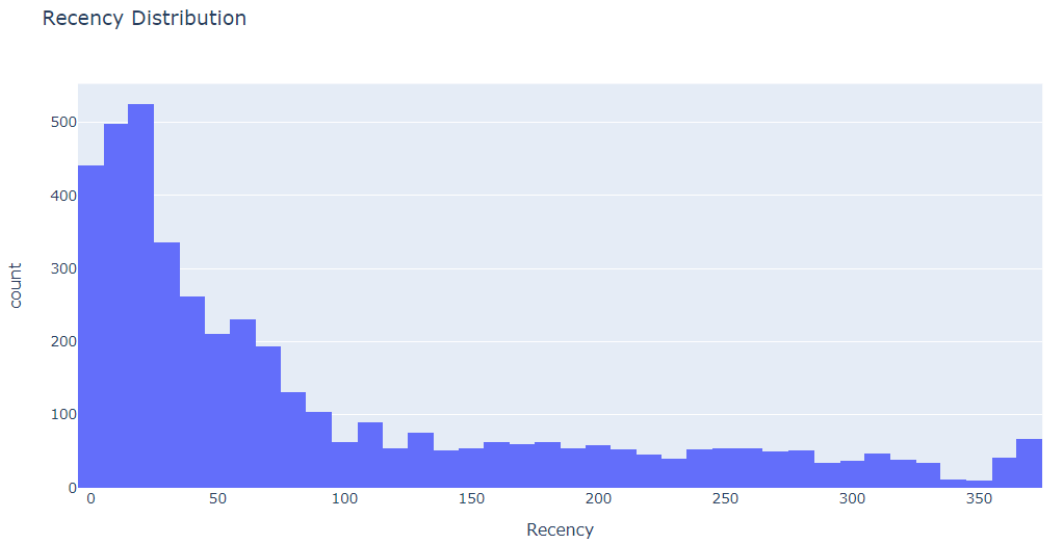
Διάγραμμα 16: Segments - RFM metrics

Για την καλύτερη προσέγγιση των αποτελεσμάτων θεωρήθηκε σκόπιμο και χρήσιμο, για την επιχείρηση να εφαρμοστεί και μια δεύτερη μεθοδολογία με μηχανική μάθηση (ML) για να ελέγχει πως συμπεριφέρονται τα δεδομένα και σε αυτήν την περίπτωση και να συγκριθούν τα αποτελέσματα. Για να μπορέσει να εφαρμοστεί ο αλγόριθμος kmeans ελέγχθηκε η μέση τιμή και η τυπική απόκλιση των τιμών και οι κατανομές των μετρικών recency, frequency και monetary. Παρατηρείται ότι, οι κατανομές δεν ακολουθούν την κανονική κατανομή αλλά είναι ασύμμετρες, λοξές δεξιά γεγονός που δικαιολογείται από την ύπαρξη ακραίων τιμών. Θα πρέπει να γίνει κανονικοποίηση των τιμών.

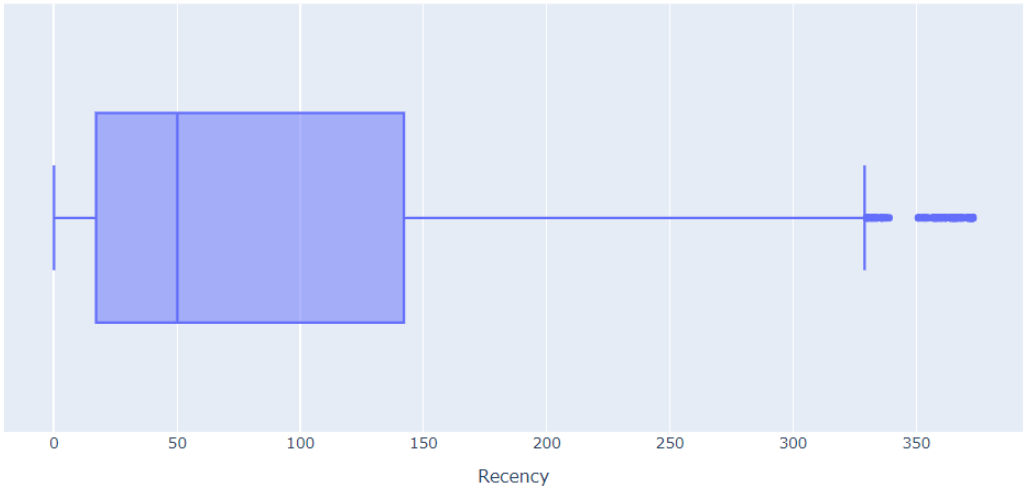
	Count	CustomerID	Recency	Frequency	Monetary
count	4334.000000	4334.000000	4334.000000	4334.000000	4334.000000
mean	2167.500000	15299.251731	92.216325	91.452700	2024.220815
std	1251.262363	1721.994109	100.176290	228.077332	8925.125833
min	1.000000	12346.000000	0.000000	1.000000	3.750000
25%	1084.250000	13812.250000	17.072396	17.000000	305.560000
50%	2167.500000	15297.500000	50.126042	41.000000	668.125000
75%	3250.750000	16778.750000	142.079861	100.000000	1631.622500
max	4334.000000	18287.000000	373.122917	7838.000000	279138.020000

Εικόνα 29: Περιγραφικά στοιχεία R, F, M

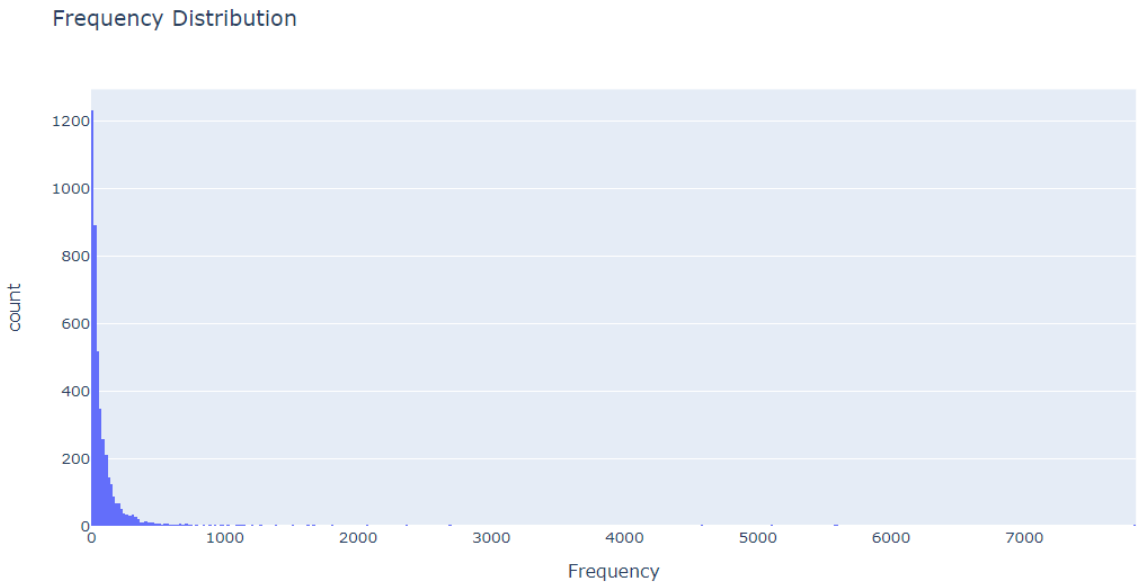
Ακολουθούν οι κατανομές των μετρικών και η αναπαράσταση τους με θηκογράμματα:



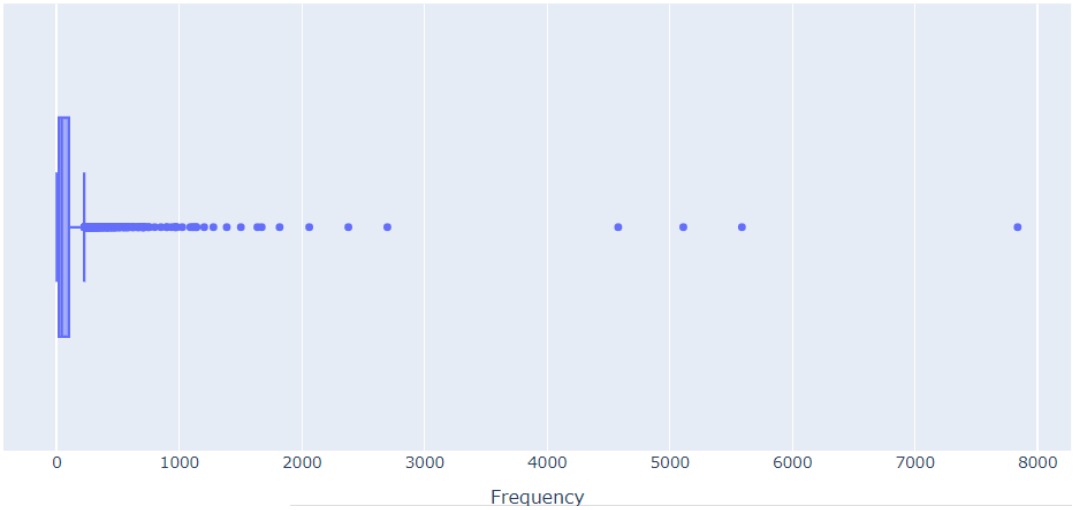
Διάγραμμα 17: Distribution of recency



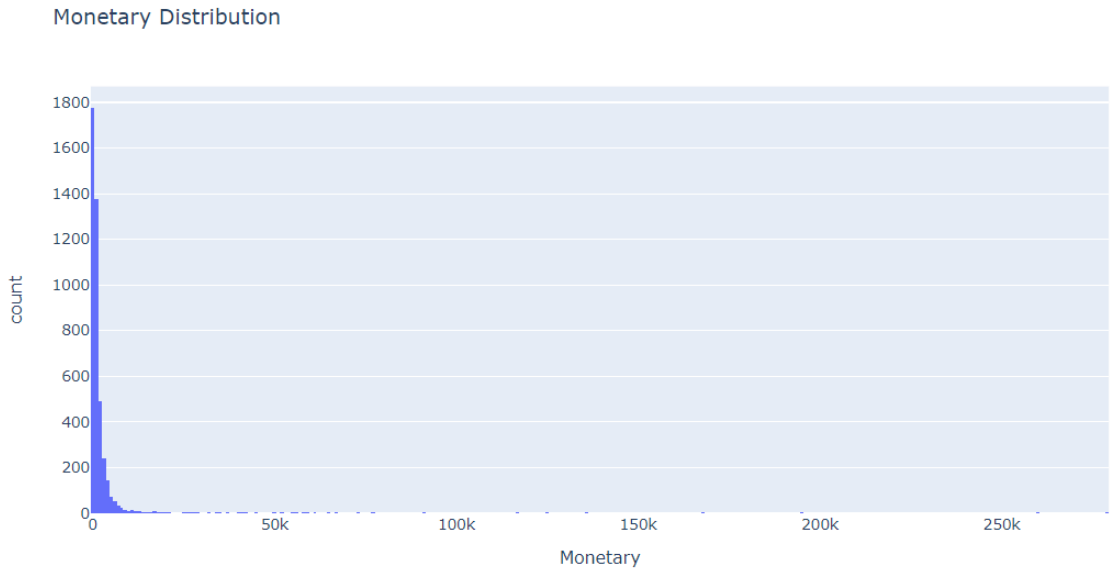
Διάγραμμα 18: Recency boxplot



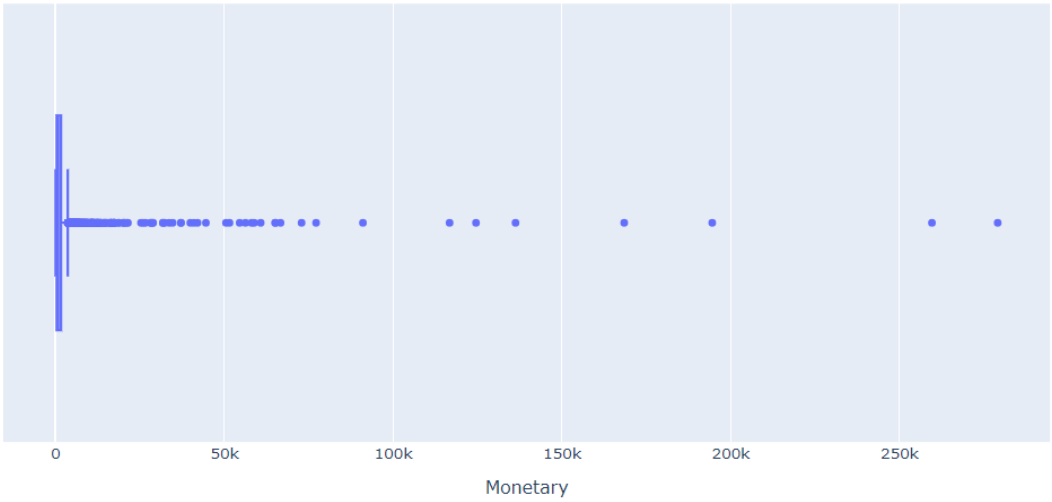
Διάγραμμα 19: Distribution of frequency



Διάγραμμα 20: Frequency boxplot



Διάγραμμα 21: Distribution of monetary



Διάγραμμα 22: Monetary boxplot

Επιπλέον τα αποτελέσματα από το τεστ Shapiro-Wilk (normality test) που πραγματοποιήθηκε έδειξαν ότι η p-value < 0.05, δηλαδή υπάρχουν επαρκή στοιχεία για να θεωρηθεί ότι τα δεδομένα δεν ακολουθούν την κανονική κατανομή, άρα η μηδενική υπόθεση απορρίπτεται.

```
ShapiroResult(statistic=0.08236300945281982, pvalue=0.0)
```

Εικόνα 30: Shapiro results

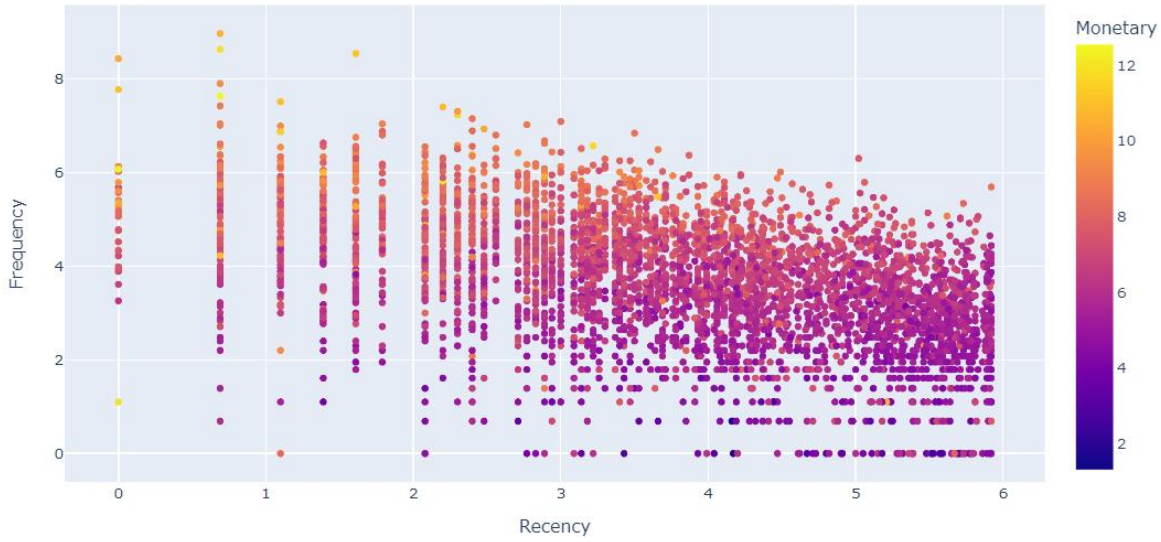
Για να μπορέσει να αντιμετωπιστεί το πρόβλημα της λοξότητας, ελέγχθηκαν οι συσχετίσεις μεταξύ των μεταβλητών.

	Recency	Frequency	Monetary
Recency	1.000000	-0.205984	-0.121355
Frequency	-0.205984	1.000000	0.417190
Monetary	-0.121355	0.417190	1.000000

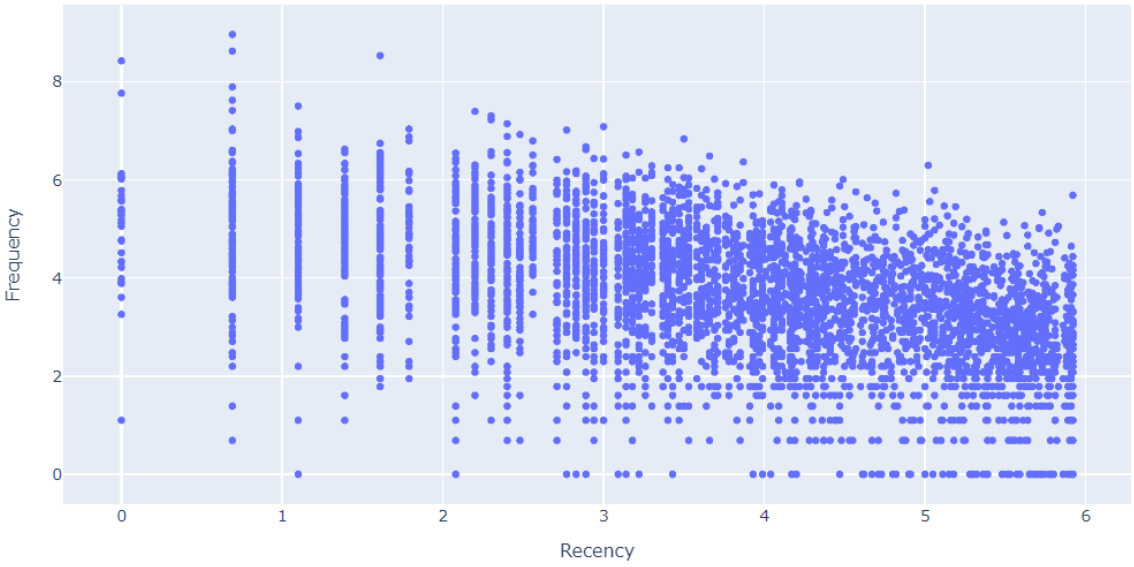
Εικόνα 31: RFM correlations

Ακολούθως παρουσιάζονται οι συσχετίσεις μεταξύ των μεταβλητών recency, frequency και monetary. Φαίνεται ότι η μόνη θετική συσχέτιση παρατηρείται στη σχέση μεταξύ των μεταβλητών frequency και monetary, καθώς όσο αυξάνεται η συχνότητα αγορών,

αυξάνεται και η αξία συναλλαγής. Οι υπόλοιπες σχέσεις μεταξύ των μεταβλητών recency-frequency, recency-monetary καθώς και η σχέση recency-frequency-monetary είναι αρνητικές, με το μεγαλύτερο πλήθος των δεδομένων να συγκεντρώνεται πάντα στα σημεία όπου είναι πιο πυκνή η συγκέντρωση και πιο έντονο το χρώμα των σημείων.

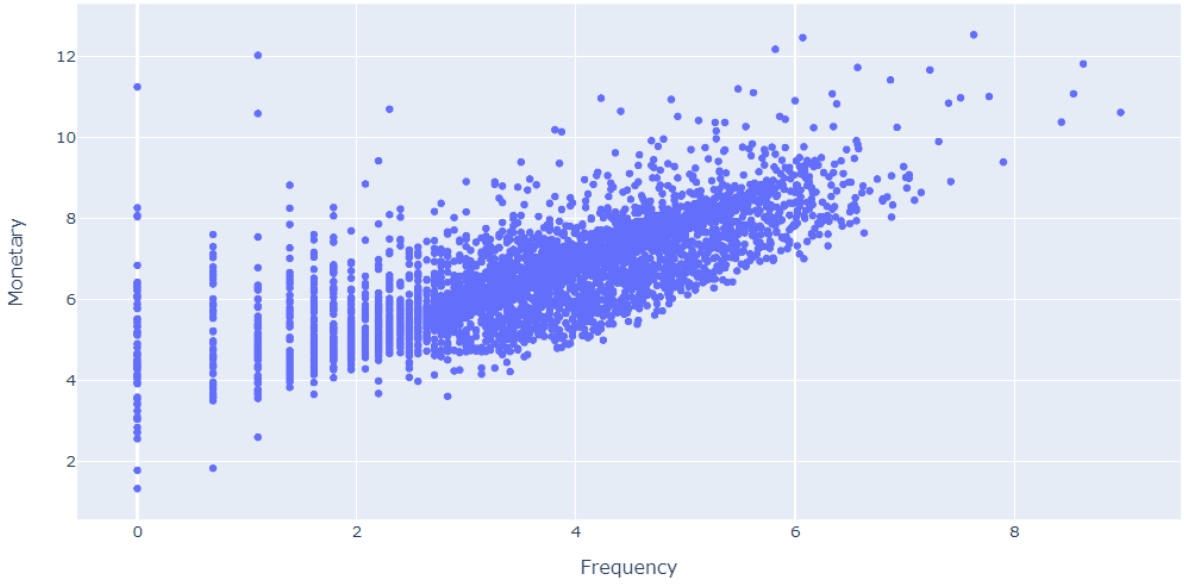


Διάγραμμα 23: Σχέση recency – frequency – monetary



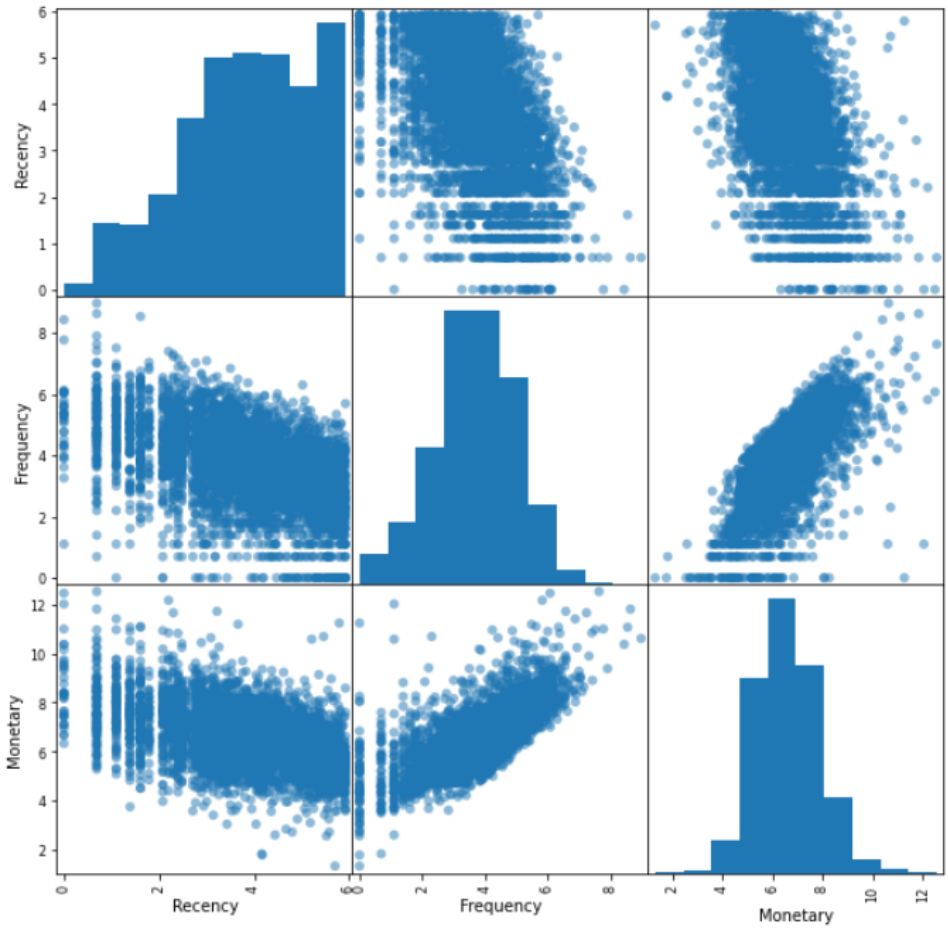
Διάγραμμα 24: Σχέση recency – frequency

Διάγραμμα 25: Σχέση recency – monetary



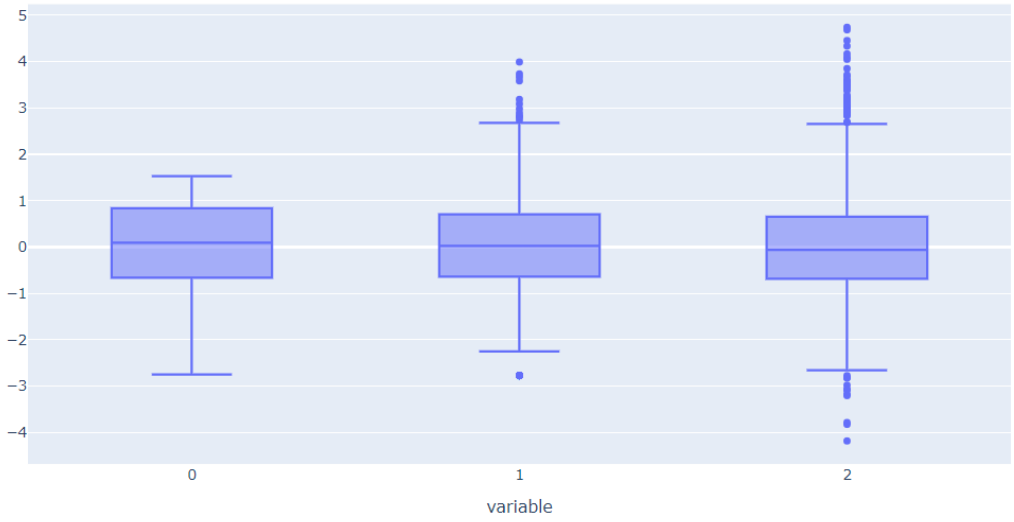
Διάγραμμα 26: Συσχέτιση frequency-monetary

Συνοψίζοντας τις συσχετίσεις και τις κατανομές μετά το log transformation, τα δεδομένα παρουσιάζονται έτσι:



Διάγραμμα 27: Correlations and distributions after log transform

Παρακάτω παρουσιάζονται τα αποτελέσματα από το μοντέλο μηχανικής μάθησης με τη βοήθεια του αλγορίθμου kmeans clustering, αφού έγινε πρώτα η κανονικοποίηση των δεδομένων με τον StandardScaler(). Με την κανονικοποίηση των δεδομένων και αφού προηγουμένως αντιμετωπίστηκε και η λοξότητα που υπήρχε, παρατηρείται ότι τα δεδομένα έχουν πλέον μέση τιμή μηδέν και η κατανομή που ακολουθούν είναι ή τείνει να είναι η κανονική. Παρατηρείται μια μικρή λοξότητα στο recency αλλά συγκριτικά με τα αρχικά αποτελέσματα υπάρχει σημαντική βελτίωση.



Διάγραμμα 28: Data after standardization

Σχετικά με τον αλγόριθμο, αρχικά ο αριθμός των clusters αποφασίστηκε να είναι πέντε (5). Για να βρεθεί ο βέλτιστος αριθμός συστάδων χρησιμοποιήθηκε η μέθοδος του αγκώνα «Elbow method», η οποία, όπως φαίνεται από το γράφημα, υπέδειξε ως βέλτιστο αριθμό συστάδων τον αριθμό τρία (3). Έτσι πραγματοποιήθηκε ξανά η διαδικασία για να οριστούν τα νέα clusters. Από την εφαρμογή του αλγορίθμου στα κανονικοποιημένα δεδομένα, βρέθηκαν τα κεντροειδή των τριών clusters.

Πρώτη δοκιμή με cluster = 5

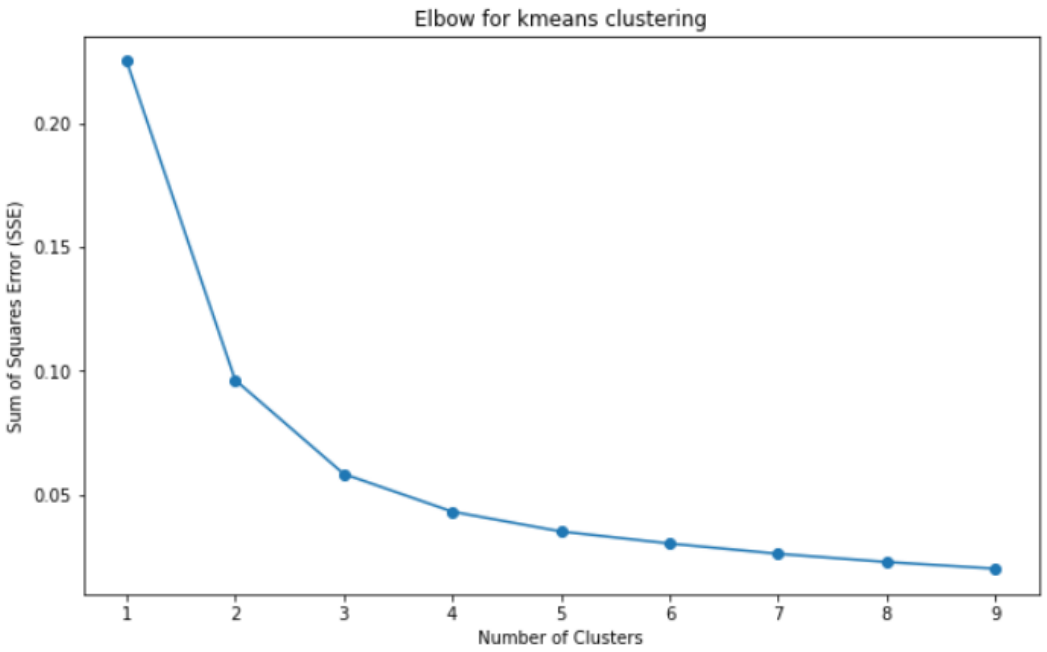
```
kmeans = KMeans(n_clusters=5)
kmeans.fit(normalized)
predict = kmeans.predict(normalized)
df['k_clusters'] = predict
kmeans.cluster_centers_
```

Εικόνα 32: kmeans 5 clusters

```
[[ 0.73573512, -1.49365875, -1.2486778 ],
 [ 0.08788193,  0.65678316,  0.63827442],
 [-0.82464217, -0.14472306, -0.27613612],
 [ 0.82728106, -0.30883439, -0.41114179],
 [-1.37416242,  1.27927802,  1.38881746]])
```

Εικόνα 33: Centroids 5 clusters

Από τον έλεγχο του πραγματοποιήθηκε μέσω του elbow method ο βέλτιστος αριθμός των cluster φαίνεται ότι είναι ο αριθμός τρία.



Εικόνα 34: Elbow method - 3 clusters

Εφαρμογή του αλγορίθμου για clusters = 3

```
kmeans = KMeans(n_clusters=3)
kmeans.fit(normalized)
predict = kmeans.predict(normalized)
df['k_clusters'] = predict
centroids = kmeans.cluster_centers_
centroids
```

Εικόνα 35: Εφαρμογή kmeans 3 clusters

```
[ 0.73812999, -0.97364294, -0.91835741],
[-0.00565806,  0.21892094,  0.13196137],
[-1.18426404,  1.14725552,  1.22833324]])
```

Εικόνα 36: Centroids 3 clusters

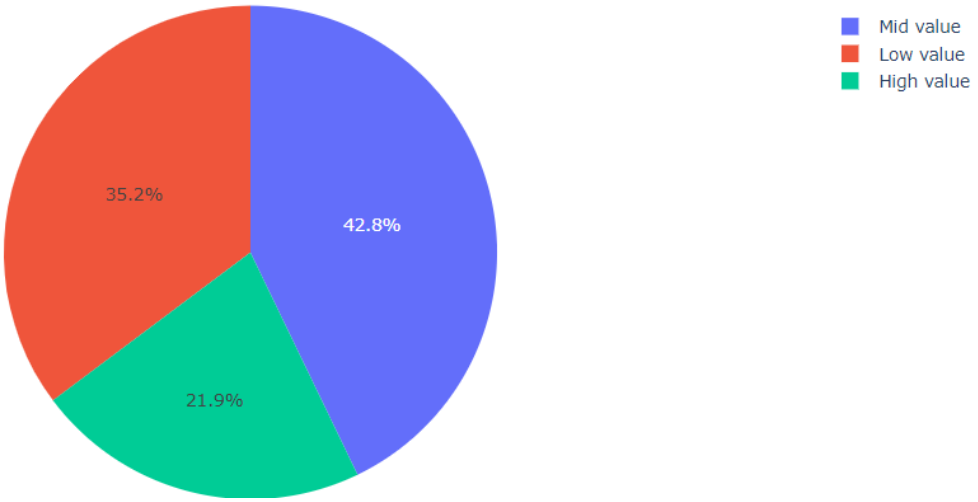
```
1    1857
0    1527
2     950
Name: k_clusters
```

Εικόνα 37: Πλήθος πελατών ανά cluster

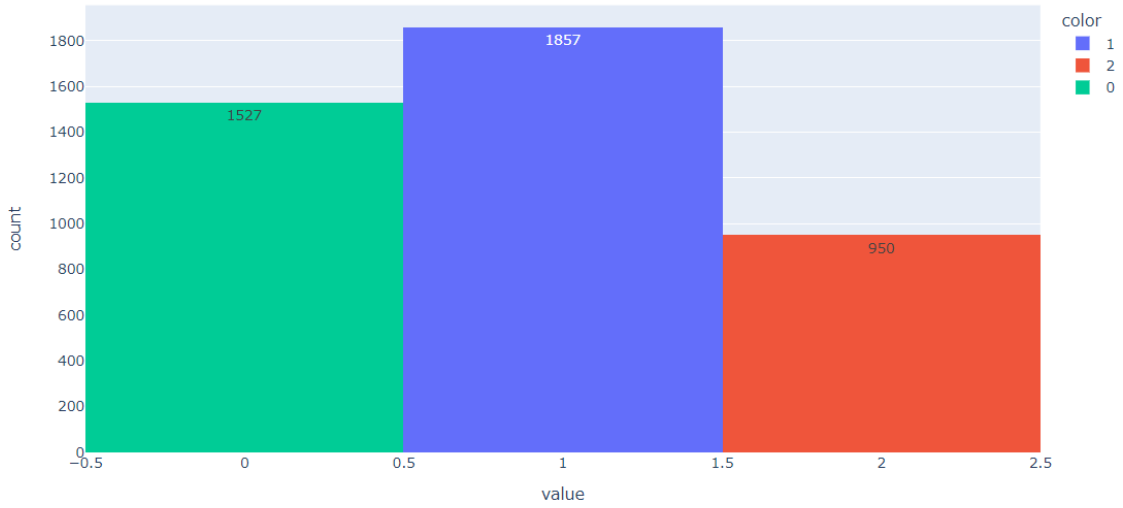
Ακολουθούν οι μέσες τιμές των μετρικών ανά cluster καθώς και οι κατανομές των κανονικοποιημένων μεταβλητών όπου φαίνεται ότι πλέον ακολουθούν την κανονική κατανομή. Παρατηρείται ότι το cluster 0 (low value), που αποτελεί αρκετά μεγάλο ποσοστό, περιέχει τους χειρότερους πελάτες της επιχείρησης (1527 πελάτες) με 35,2%, οι οποίοι έχουν μέση αξία συναλλαγής 290£, με 15 αγορές και recency 172 ημέρες απόσταση από την ημερομηνία αναφοράς. Το cluster 1 (mid value) περιέχει το μεγαλύτερο ποσοστό, 42,8% του συνόλου με 1.857 πελάτες με 1.155£ μέση αξία συναλλαγής, 66 μέσες συνολικές αγορές και 69 ημέρες διάστημα μεταξύ της αγοράς και της ημερομηνίας αναφοράς. Τέλος το cluster 2 (high value), με 21,9%, περιέχει τους καλύτερους πελάτες της επιχείρησης, με αριθμό πελατών 950 άτομα, 265 αγορές, 6.510£ μέση αξία συναλλαγής και 14 ημέρες διάστημα από την ημερομηνία αναφοράς.

k_clusters	Recency	Frequency	Monetary
0	171.827767	14.965291	290.063072
1	69.119548	65.729672	1155.291310
2	14.008421	264.677895	6510.179726

Εικόνα 38: RFM metrics for k clusters

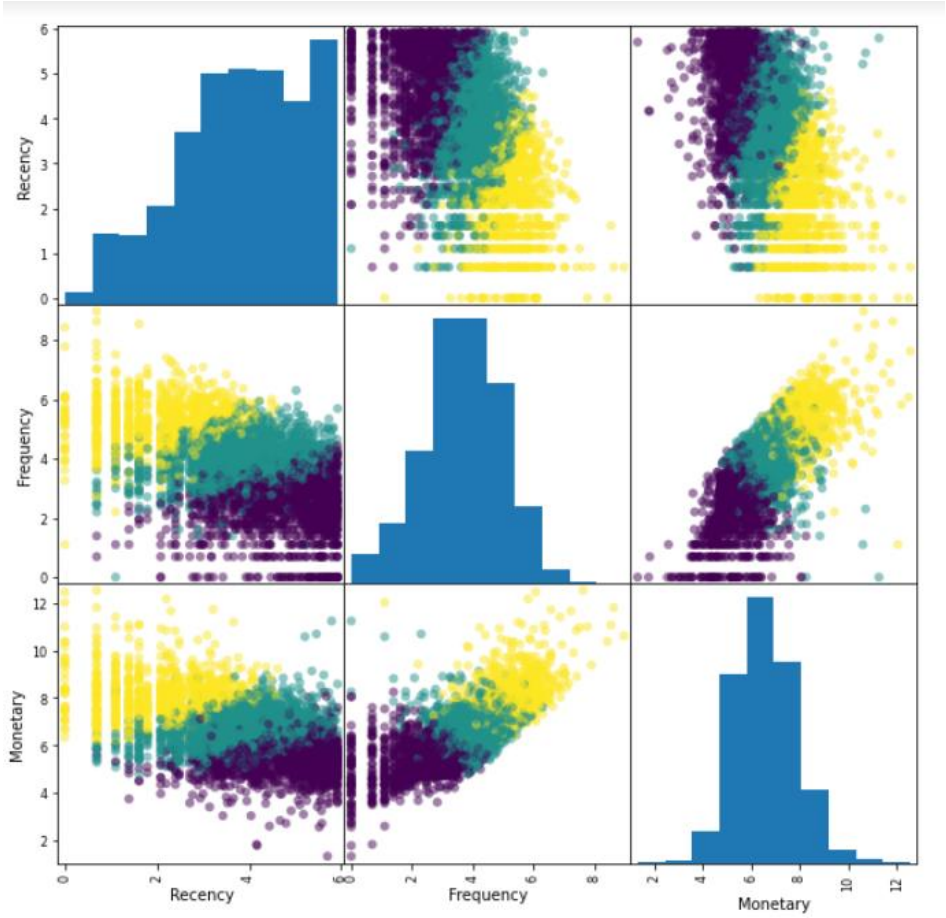


Εικόνα 39: Proportion by clusters

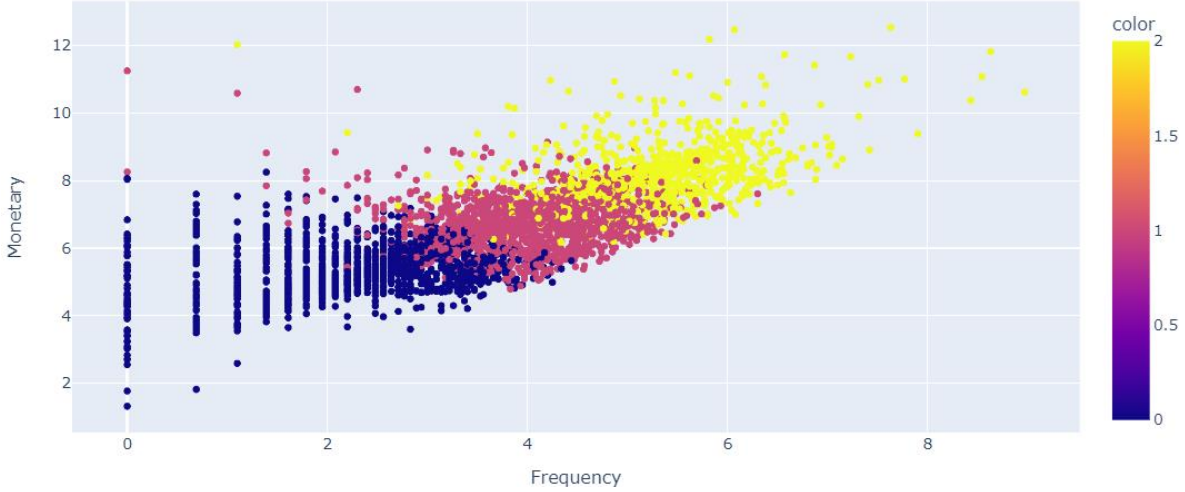


Εικόνα 40: Customer number by cluster

Από το συγκεντρωτικό διάγραμμα παρατηρείται ότι οι τιμές πλέον να ακολουθούν ή ακολουθούν την κανονική κατανομή και η μόνη θετική σχέση είναι και πάλι μεταξύ του frequency και monetary. Παρατηρείται ότι οι τιμές με κίτρινο χώμα είναι το cluster 2 με τους high value πελάτες, όπου η συχνότητα αγορών και συναλλαγής είναι υψηλά ενώ το recency χαμηλό. Αντίθετα με πράσινο χρώμα αναπαρίσταται το cluster 1, mid value πελάτες, όπου οι πελάτες σε αυτό το cluster κάνουν αρκετά συχνά αγορές, ξοδεύοντας όχι πολλά αλλά ούτε και λίγα χρήματα, ενώ με μπλε χρώμα είναι το cluster 0 με τους low value πελάτες, οι οποίοι δεν πρόσφατες και συχνές αγορές και ξοδεύουν πάντα μικρά ποσά.



Διάγραμμα 29: Correlations by clusters and distributions after standardization



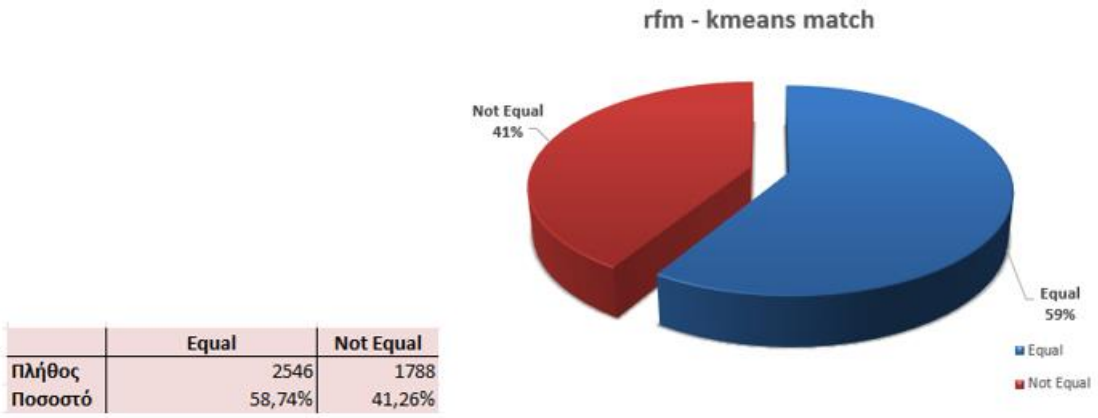
Διάγραμμα 30: Positive correlation R-F and clusters

Στη συνέχεια, τα συγκεντρωτικά αποτελέσματα από το excel, αφορούσαν την σύγκριση των δυο αναλύσεων που πραγματοποιήθηκαν για να μπορέσει η επιχείρηση να κατανοήσει ποια από τις δυο είναι η καλύτερη.

RFM Analysis			Compare	Kmeans Clustering	
RFM score	Segment	rfm segment code	Compare clusters	k_clusters	kmeans clusters
115	Hybernating	0	Not Equal	1	Mid value
555	Champions	2	Equal	2	High value
224	Need attention	1	Equal	1	Mid value
444	Loyal Customers	2	Not Equal	1	Mid value
122	Hybernating	0	Equal	0	Low value
344	Loyal Customers	2	Not Equal	1	Mid value
111	Hybernating	0	Equal	0	Low value
144	Can't lose	0	Not Equal	1	Mid value
112	Hybernating	0	Equal	0	Low value
445	Loyal Customers	2	Not Equal	1	Mid value
355	Loyal Customers	2	Equal	2	High value
524	Loyal Customers	2	Equal	2	High value
355	Loyal Customers	2	Equal	2	High value
355	Loyal Customers	2	Not Equal	1	Mid value
111	Hybernating	0	Equal	0	Low value
555	Champions	2	Equal	2	High value
223	Need attention	1	Not Equal	0	Low value
544	Champions	2	Equal	2	High value
122	Hybernating	0	Equal	0	Low value
511	New Customers	1	Not Equal	0	Low value
355	Loyal Customers	2	Equal	2	High value

Εικόνα 41: Comparison rfm - kmeans

Η ταύτιση μεταξύ των αποτελεσμάτων των δυο μοντέλων συγκέντρωσε μεγάλο ποσοστό, κοντά στο 59%, με 2.546 παρατηρήσεις να ανήκουν στο ίδιο cluster. Το ποσοστό αυτό όμως δεν ήταν επαρκές ώστε να θεωρηθεί ότι δεν επηρεάζονται τα αποτελέσματα αρνητικά. Το 41,26% των clusters που δεν ταυτίστηκε (1.788 clusters) επηρεάζει την κατανομή των πελάτων στα clusters.



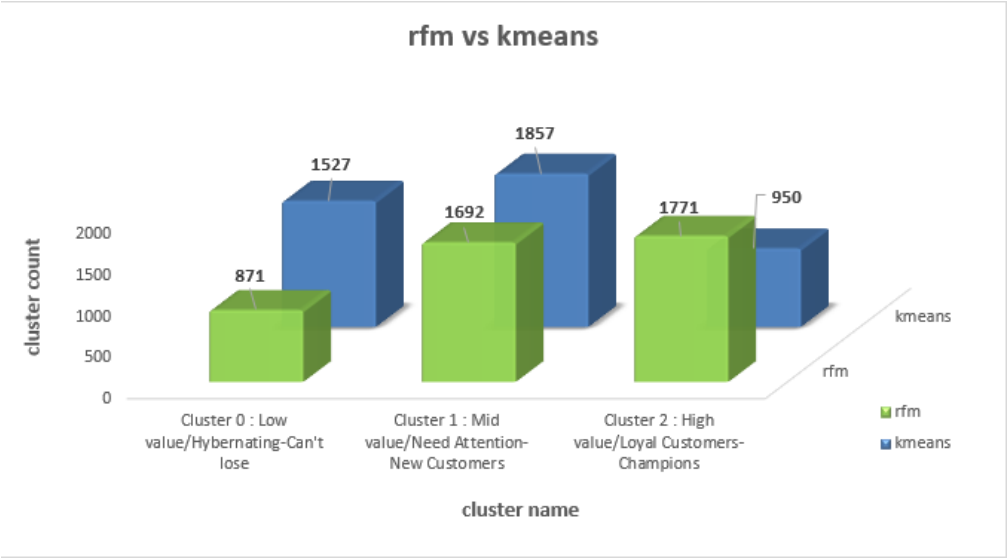
	Equal	Not Equal
Πλήθος	2546	1788
Ποσοστό	58,74%	41,26%

Διάγραμμα 31: Comparison rfm – kmeans

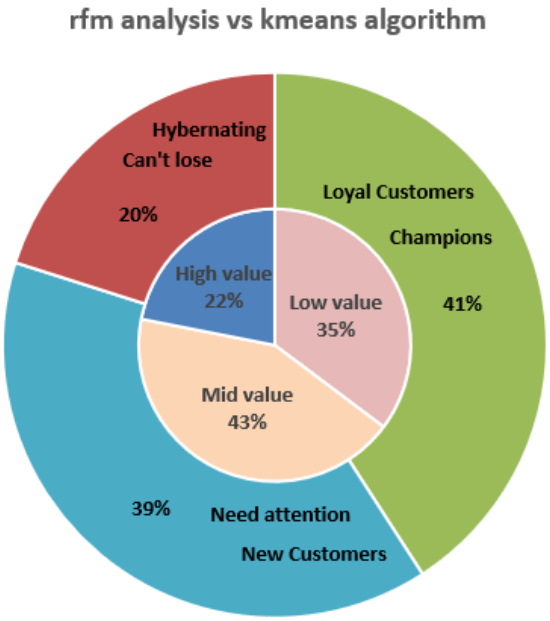
Τα παρακάτω διαγράμματα απεικονίζουν τα αποτελέσματα από την σύγκριση των δυο μοντέλων.

clusters name	rfm	kmeans
Cluster 0 : Low value/Hybernating-Can't lose	871	1527
Cluster 1 : Mid value/Need Attention-New Customer	1692	1857
Cluster 2 : High value/Loyal Customers-Champions	1771	950

Εικόνα 42: Results from rfm - kmeans



Διάγραμμα 32: rfm vs kmeans



Διάγραμμα 33: Proportions from clusters

Κεφάλαιο 5: Συμπεράσματα

Σύμφωνα με τη rfm ανάλυση και την ομαδοποίηση που πραγματοποιήθηκε για χάρη της σύγκρισης με το μοντέλο του kmeans μοντέλου, το μεγαλύτερο ποσοστό (41%) των πελατών συγκεντρώνεται στο cluster που περιέχει τους champions και loyal customers (1.771 άτομα), πελάτες δηλαδή που έχουν μεγάλη αξία για την επιχείρηση, το 39% συγκεντρώνεται στο cluster με τους need attention και new customers πελάτες (1.692 άτομα) και αφορούν πελάτες που χαρακτηρίζονται από ουδετερότητα, πραγματοποιούν δηλαδή αγορές αλλά όχι τόσο συχνά και το καλάθι αγορών τους δεν είναι τόσο μεγάλο. Το μικρότερο ποσοστό (20%) εμφανίζεται στο cluster με τους can't lose και hybernating, πελάτες, δηλαδή πελάτες που δεν έχουν μεγάλη αξία για την επιχείρηση (871 άτομα).

Αντίθετα σύμφωνα με τα αποτελέσματα του αλγόριθμου kmeans, το μεγαλύτερο ποσοστό των πελάτων συγκεντρώνεται στο *cluster 1*, που περιέχει τους *mid value* πελάτες, με 42,8% και 1.857 άτομα. Είναι δηλαδή πελάτες που δεν είναι ούτε οι καλύτεροι αλλά ούτε και οι χειρότεροι της επιχείρησης, τους χαρακτηρίζει μια ουδετερότητα και έχουν μέτρια αξία για την επιχείρηση. Ακολουθεί επίσης με μεγάλο ποσοστό, 35,2%, το *cluster 0*, το οποίο περιέχει τους *low value* πελάτες, που συγκεντρώνει 1.527 άτομα και πρόκειται για πελάτες με χαμηλή ή και καθόλου αξία για την επιχείρηση. Τέλος, το *cluster 2*, περιέχει τους *high value* πελάτες που συγκεντρώνουν το 21,9% του συνόλου των πελατών με 950 άτομα.

Μετά από τις δυο αναλύσεις και τα αποτελέσματα που διεξάχθηκαν, το μοντέλο που προτείνεται στην επιχείρηση να ακολουθήσει για να μπορέσει να πετύχει αύξηση της κερδοφορίας της, είναι το μοντέλο της rfm, καθώς τα αποτελέσματα του και η κατανομή των πελάτων που γίνεται στα segments είναι περισσότερο ωφέλιμα για αυτό τον σκοπό. Η ανάλυση rfm συγκεντρώνει τα μεγαλύτερα ποσοστά σε ομάδες με υψηλή ή μέτρια αξία για την επιχείρηση, που και αυτός είναι ο κύριος στόχος της επιχείρησης, ενισχύοντας φυσικά με αυτόν τον τρόπο την αύξηση των κερδών. Από την αντίθετη πλευρά, τα μικρότερα ποσοστά εμφανίζονται στα segments που περιέχουν τους πελάτες που δεν έχουν μεγάλη αξία για την εταιρεία, επομένως και η επιχείρηση θα παρουσιάσει περισσότερο ικανοποιητικά αποτελέσματα με μια τέτοια κατανομή. Έτσι, η επιχείρηση ανάλογα και με τους πόρους που έχει στη διάθεσή της θα αποφασίσει ποιες ενέργειες και στρατηγικές μπορεί να εφαρμόσει στα αντίστοιχα segments. Από την άλλη πλευρά ο αλγόριθμος kmeans δεν προτείνεται καθώς η μεγαλύτερη συγκέντρωση των πελατών

γίνεται στα cluster 0 και cluster 1 που περιέχουν πελάτες με μέτρια και πολύ χαμηλή αξία και συχνότητα αγορών, κάτι που σίγουρα δεν είναι καθόλου αποδοτικό και βοηθητικό για να μπορέσει η επιχείρηση να ενισχύσει τα κέρδη της, αντίθετα θα την αναγκάσει να δαπανήσει περισσότερους πόρους και προσπάθεια για να καταφέρει να έχει κάποιο κέρδος από τις ομάδες αυτές ή για να διατηρήσει αυτούς τους πελάτες.

Έτσι, κρίνεται απαραίτητο να εφαρμόσει η επιχείρηση κατάλληλες στρατηγικές σε κάθε ένα cluster ή ακόμα και σε κάθε segment που περιέχεται μέσα στα τρία clusters ξεχωριστά, καθώς ο διαχωρισμός των segments της rfm ανάλυσης έγινε για την σύγκριση με τον αλγόριθμο kmeans. Για το cluster 2 που περιέχει τους champions και loyal customers πελάτες, η επιχείρηση θα πρέπει να επικεντρωθεί σε ενέργειες που θα ενισχύσουν το αίσθημα εμπιστοσύνης που υπάρχει ήδη από αυτούς και να τους κάνει να νιώθουν ξεχωριστοί. Και τα δυο segments αφορούν ενεργούς, πολύτιμους και πιστούς πελάτες που ευθύνονται και για μεγάλο μερίδιο των εσόδων της επιχείρησης.

Για τους champions πελάτες δεν χρειάζεται μεγάλη προσπάθεια και προσέγγιση με μεγάλες εκπτώσεις καθώς είναι πελάτες που είναι διατεθειμένοι να προβούν σε νέες αγορές και να δοκιμάσουν νέα προϊόντα ανεξαρτήτου τιμής. Προτείνεται δοθούν ειδικά προνόμια, κάρτες επιβράβευσης και προτεραιότητα στην διεκπεραίωση των παραγγελιών τους. Ακόμα προτείνεται δυνατότητα εγγραφής τους ως VIP μέλη και στη συνέχεια εφαρμογή εξατομικευμένων καμπανιών μόνο για πελάτες VIP. Επίσης, στο τμήμα αυτό, η επιχείρηση μπορεί να δοκιμάζει τα νέα προϊόντα που θέλει να λανσάρει στην αγορά καθώς είναι πελάτες δεκτικοί σε αυτό και μπορεί μέσω αυτής της διαδικασίας να εφοδιαστεί η επιχείρηση με πολύτιμες πληροφορίες σχετικά με νέες τάσεις και προτιμήσεις του καταναλωτικού κοινού με τη χρήση ηλεκτρονικών ερωτηματολογίων. Με τον τρόπο αυτό η επιχείρηση θα είναι σε θέση να παίρνει την ανατροφοδότηση και να αναγνωρίζει πιθανές αστοχίες και βελτιώσεις που χρειάζονται τα προϊόντα της.

Για τους loyal customers, καθώς είναι ήδη πελάτες με υψηλή αξία αγορών και αυξημένη συχνότητα, προτείνεται η εφαρμογή τεχνικών πώλησης cross-selling και up-selling για να δημιουργηθούν πλασματικές ανάγκες που θα οδηγήσουν σε περισσότερες ή ακόμα και ακριβότερες αγορές. Εφαρμογή προσωποποιημένου μάρκετινγκ, εξατομικευμένες ενέργειες και προγράμματα πιστότητας με συλλογή πόντων ή προγράμματα αφοσίωσης μπορούν να λειτουργήσουν αποτελεσματικά. Επίσης, ειδικά προνόμια όπως απαλλαγή από τα έξοδα αποστολής, δώρα ή ειδικές προσφορές κάθε μήνα μπορούν να

λειτουργήσουν θετικά στην παρακίνηση της ομάδας αυτής. Ακόμη, το τμήμα έρευνας μάρκετινγκ μπορεί να πραγματοποιήσει έρευνα σχετικά με τα προϊόντα που αγοράζονται συχνότερα και να εφαρμόσει κατάλληλες και στοχευμένες στρατηγικές.

Το cluster 1 που περιλαμβάνει τους new customers και τους need attention πελάτες. Η στρατηγική που προτείνεται στην επιχείρηση για τους new customers είναι το email μάρκετινγκ και η προσωποποιημένη επικοινωνία για να μπορέσει να δημιουργήσει μια σχέση δέσμευσης μαζί με τους πελάτες. Επιπλέον πραγματοποίηση έρευνας σχετικά με τις ανάγκες και τις προτιμήσεις των πελατών, ώστε να εφαρμοστούν στοχευμένες ενέργειες μάρκετινγκ, εκπτώσεις, επιβράβευση με την πρώτη παραγγελία και δώρα ή εκπτωτικά κουπόνια μετά από κάποιον αριθμό παραγγελιών. Εξαιρετικά χρήσιμη θεωρείται και η εφαρμογή post-sale υπηρεσιών, για τη συλλογή πληροφοριών σχετικά με την ικανοποίηση των πελάτων σε σχέση με το προϊόν μετά την πώληση, όπως για παράδειγμα η τηλεφωνική επικοινωνία μέσω των customer supporters.

Για τους need attention προτείνεται εφαρμογή επιθετικού μάρκετινγκ καθώς πρόκειται για πελάτες που δεν ξοδεύουν πολλά και δεν αγοράζουν συχνά και επιδιώκεται από την επιχείρηση να τους μετατρέψει σε πιστούς πελάτες. Εφαρμογή προσωποποιημένου email μάρκετινγκ και μαζικές αυτοματοποιημένες καμπάνιες με χαμηλό κόστος, ώστε να μην επιβαρύνεται με πολλά επιπλέον κόστη η επιχείρηση. Επιπλέον, παροχή προσφορών ή εκπτωτικών κουπονιών με συγκεκριμένο χρονικό διάστημα λήξης, ώστε να υπάρχει η δέσμευση για να παρακινούνται οι πελάτες και προβαίνουν άμεσα σε μια αγορά. Εποχιακές προσφορές με ειδικές παροχές όπως δωρεάν επιστροφή προϊόντων, διαφημίσεις με στοχευμένο περιεχόμενο και αποστολή ενημερωτικών newsletters και μεγάλες εκπτώσεις σε συγκεκριμένα προϊόντα σε εορταστικές ή εκπτωτικές περιόδους, όπου η ζήτηση παρουσιάζεται πάντα αυξημένη.

Τέλος, οι στρατηγικές που προτείνονται για το cluster 0, είναι αρχικά για τους can’t lose πελάτες, η παροχή εκπτώσεων στις επόμενες παραγγελίες, ειδικές προσφορές ή κουπόνια σε περιόδους εκπτώσεων ώστε να καταφέρει η επιχείρηση να αφυπνίσει τους πελάτες αυτούς. Πρόκειται για καλούς πελάτες, οι οποίοι όμως αυτήν την στιγμή δεν έχουν κάποια αξία για την επιχείρηση, καθώς έχουν αρκετό διάστημα να πραγματοποιήσουν κάποια αγορά. Προτείνεται επίσης τηλεφωνική επικοινωνία με τους πελάτες για να μπορέσει να κατανοηθεί η αιτία αυτής της αδράνειας και να γίνει προσπάθεια επαναπροσέγγισης των πελατών αυτών ή sms μάρκετινγκ προτείνοντάς τους προϊόντα με

υψηλή ζήτηση. Τέλος, θα μπορούσε να πραγματοποιηθεί έρευνα σχετικά με τα προϊόντα που αναζητούσαν αυτοί οι πελάτες και στη συνέχεια να γίνουν προσωποποιημένες διαφημιστικές ενέργειες για να μπορέσει να προσελκυστεί και πάλι το ενδιαφέρον τους.

Για τους hybernating, καθώς είναι ήδη χαμένοι πελάτες, προτείνεται στην επιχείρηση να κάνει μια προσπάθεια για την επαναπροσέγγισή τους αλλά με οικονομικά συμφέρουσες στρατηγικές, καθώς πολύ πιο κοστοβόρο για μια επιχείρηση να προσπαθήσει να διατηρήσει αυτήν την ομάδα πελατών από το να προσελκύσει νέους πελάτες. Έτσι, η παροχή κουπονιών μικρής αξίας σε περιόδους ενδιάμεσων εκπτώσεων ή Black Friday, ημέρα αγίου Βαλεντίνου κ.α ή η εφαρμογή email μάρκετινγκ για την υπενθύμιση της εταιρείας θα ήταν ικανοποιητικές ενέργειες για να τους δοθεί μια δεύτερη ευκαιρία.

Τέλος, από την έρευνα παρατηρήθηκε ότι τα μεγαλύτερα έσοδα της επιχείρησης εμφανίζονται το τρίμηνο πριν από την εορταστική περίοδο των Χριστουγέννων, καθώς υπήρχε θεαματική άνοδος των πωλήσεων γεγονός που δικαιολογείται, γιατί οι καταναλωτές αγοράζουν πολλά δώρα. Από την άλλη πλευρά, οι μήνες πριν την εορταστική περίοδο του Πάσχα, δεν εμφανίζουν μεγάλα έσοδα, παρόλο που αναμενόταν καθώς είναι επίσης μια περίοδος με αυξημένη καταναλωτική ζήτηση. Προτείνεται στην επιχείρηση να εστιάσει στην προσπάθεια βελτίωσης αυτών των μηνών, εξετάζοντας τις αιτίες που μπορεί να οφείλονται, όπως χαμηλή επισκεψιμότητα, μη επαρκές εμπόρευμα, πιθανόν χαμηλό εποχικό εμπόρευμα (Πασχαλιάτικα είδη δώρων) ή πιθανή αυξημένη τιμολόγηση σε σχέση με των ανταγωνιστών της.

Κεφάλαιο 6: Προτάσεις για μελλοντικές έρευνες

Καθώς τα δεδομένα που χρησιμοποιήθηκαν αφορούσαν μόνο τη διάρκεια ενός έτους, χρονική περίοδος από το Δεκέμβριο 2010 έως το Δεκέμβριο 2011, θα ήταν περισσότερο σκόπιμο για να συλλέξει η επιχείρηση πιο ακριβή στοιχεία και καλύτερα αποτελέσματα να διερευνηθεί μια περίοδος μεγαλύτερης διάρκειας και πιο πρόσφατη.

Επίσης, παρατηρήθηκε ότι το μεγαλύτερο ποσοστό των πελάτων ανήκει στη χώρα του Ηνωμένου Βασιλείου, οπότε θα ήταν απαραίτητο να πραγματοποιηθεί επιπλέον μια ανάλυση για περαιτέρω διερεύνηση αυτής της αγοράς. Θα εξεταστούν αναλυτικότερα ερωτήματα όπως: ποια είναι τα προϊόντα που αγοράζονται πιο συχνά, ποια προϊόντα αγοράζονται μαζί, να αναλυθούν γεωγραφικά και δημογραφικά στοιχεία των πελάτων, να βρεθούν αγοραστικές τάσεις ανά περιοχή, τι δουλειά κάνει ο κάθε πελάτης και διάφορα άλλα τέτοια ερωτήματα ώστε να μπορέσει η επιχείρηση να εφαρμόσει καλύτερες και περισσότερο στοχευμένες στρατηγικές.

Τέλος, είναι σημαντικό να πραγματοποιηθεί ανάλυση καλαθιού (basket analysis) σχετικά με τα προϊόντα που αγοράζουν οι καταναλωτές ή με τα προϊόντα που τελικά δεν αγόρασαν και έμειναν στο καλάθι. Αυτό θα εφοδιάσει την επιχείρηση με απαραίτητες πληροφορίες σχετικά με τα συμπληρωματικά προϊόντα που μπορεί να προωθήσει στους πελάτες ή στην περίπτωση που τα προϊόντα παρέμειναν στο καλάθι και η αγορά δεν πραγματοποιήθηκε θα μπορέσει να κατανοήσει γιατί ο πελάτης έφυγε και δεν ολοκλήρωσε την αγορά (customer churn).

Βιβλιογραφική Αναφορά

Ξενόγλωσση Βιβλιογραφία

Ahuja, R., Chug, A., Gupta, S., Ahuja, P., & Kohli, S. (2020). Classification and clustering algorithms of machine learning with their applications. *Nature-Inspired Computation in Data Mining and Machine Learning*, 225-248.

Akaah, I. P., Korgaonkar, P. K., & Lund, D. (1995). Direct marketing attitudes. *Journal of Business Research, 34*(3), 211-219.

Alzubi, J., Nayyar, A., & Kumar, A. (2018, November). Machine learning from theory to algorithms: an overview. Journal of Physics: Conference Series (Vol. 1142, p. 012012). IOP Publishing.

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. Journal of King Saud University-Computer and Information Sciences, 34(5), 1785-1792.

Armstrong, G. (2009). Marketing: an introduction. Pearson education.

Armstrong, G. & Kotler, P. (2011). Principles of marketing (14th edition). Pearson Prentice Hall.

Armstrong, G., Adam, S., Denize, S., & Kotler, P. (2014). Principles of marketing. Pearson Australia.

Baecke, P., & Van den Poel, D. (2011). Data augmentation by predicting spending pleasure using commercially available external data. *Journal of Intelligent Information Systems, 36*(3), 367-383.

Bass, F. M. (1993). The future of research in marketing: marketing science. Journal of Marketing Research, 30(1), 1-6.

Beane, T. P., & Ennis, D. M. (1987). Market segmentation: a review. European journal of marketing. 21, 20-42.

Birant, D. (2011). Data mining using RFM analysis. *Knowledge-Oriented Applications in Data Mining*. IntechOpen.

Brown, D. E., & Gunderson, L. R. (2002). Using data mining to discover the preferences of computer criminals. Advances in Computers, 343-373, (Vol. 56, pp. 343-373). Elsevier.

Chalmeta, R. (2006). Methodology for customer relationship management. Journal of systems and software, 79(7), 1015-1024.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing & Customer Strategy Management, 19(3), 197-208.

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, *36*(3), 4176-4184.

Christy, A. 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), 1251–1257.

Cohen, M. C. (2018). Big data and service operations. Production and Operations Management, 27(9), 1709-1723.

Coussement, K., Van den Bossche, F. A., & De Bock, K. W. (2014). Data Accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. *Journal of Business Research*, *67*(1), 2751-2758.

Derawi, M. O., Nickel, C., Bours, P., & Busch, C. (2010, October). Unobtrusive user-authentication on mobile phones using biometric gait recognition. In 2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (pp. 306-311). IEEE.

Dibb, S. (1998). Market segmentation: Strategies For Success. *Marketing Intelligence & Planning*, *16*(7), 394-406.

Ernawati, E., Baharin, S. S., & Kasmin, F. (2021, April). A review of data mining methods in RFM-based customer segmentation. Journal of Physics: Conference Series, Vol. 1869, No 1, p. 012085, IOP Publishing.

Grandhi, B., Patwa, N., & Saleem, K. (2020). Data-driven marketing for growth and profitability. EuroMed Journal of Business, 16(4), 381-398.

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), 470-477.

Hauser, W. J. (2007). Marketing analytics: the evolution of marketing research in the twenty-first century. Direct marketing: an international journal, 1(1), 38-54.

He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. Computers in Human Behavior, 29(1), 90-102.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. International journal of information management, 33(3), 464-472.

Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, *37*(7), 5259-5264.

Kahan, R. (1998). Using database marketing techniques to enhance your one-to-one marketing initiatives. *Journal of Consumer Marketing*, *15*(5), 491-493.

Kalam, K. K. (2020). Market segmentation, targeting and positioning strategy adaptation for the global business of Vodafone Telecommunication Company. International Journal of Research and Innovation in Social Science, 4(6), 427-430.

Kaymak, U. (2001, July). Fuzzy target selection using RFM variables. *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569*) (Vol. 2, pp. 1038-1043). IEEE.

Kononenko, I., & Kukar, M. (2007). Chapter 12-Cluster Analysis. Machine Learning and Data Mining, 321-358.

Kotler, P. (2001). Marketing management, millenium edition. Prentice-Hall, Inc.

Kotler, P., Burton, S., Deans, K., Brown, L., & Armstrong, G. (2015). Marketing. Pearson Higher Education AU.

Kracklauer, A. H., Mills, D. Q., & Seifert, D. (2004). Customer management as the origin of collaborative customer relationship management. Collaborative customer relationship management: taking CRM to the next level, 3-6.

Littler, D. (1995). Market segmentation. Marketing Theory and Practice, 90-103.

Liu, D. R., Lai, C. H., & Lee, W. J. (2009). A hybrid of sequential rules and collaborative filtering for product recommendation. *Information Sciences, 179*(20), 3505-3519.

Lumsden, S. A., Beldona, S., & Morrison, A. M. (2008). Customer value in an all-inclusive travel vacation club: An application of the RFM framework. *Journal of Hospitality & Leisure Marketing, 16*(3), 270-285.

MacQueen, J. (1967, June). Classification and analysis of multivariate observations. In 5th Berkeley Symp. Math. Statist. Probability (pp. 281-297). Los Angeles LA USA: University of California.

Mahesh, B. (2020). Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR), 9, 381-386.

Marcus, C. (1998). A practical yet meaningful approach to customer segmentation. *Journal of Consumer Marketing, 15*(5), 494-504.

Mazzei, M. J., & Noble, D. (2017). Big data dreams: A framework for corporate strategy. Business Horizons, 60(3), 405-414.

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research, 60*(6), 656-662.

McDonald, M., Christopher, M., Bass, M. (2003). Market segmentation (pp. 41-65). Macmillan Education UK.

McDonald, M. (2012). Market segmentation: How to do it and how to profit from it. John Wiley & Sons.

Miglautsch, J. (2002). Application of RFM principles: What to do with 1–1–1 customers?. Journal of Database Marketing & Customer Strategy Management, 9(4), 319-324.

Miglautsch, J. R. (2000). Thoughts on RFM scoring. *Journal of Database Marketing & Customer Strategy Management, 8*(1), 67-72.

Mulvenna, M., Norwood, M., & Büchner, A. (1998). Data-driven marketing. Electronic Markets, 8(3), 32-35.

Nainggolan, R., Perangin-angin, R., Simarmata, E., & Tarigan, A. F. (2019, November). Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method. In Journal of Physics: Conference Series (Vol. 1361, No. 1, p. 012015). IOP Publishing.

Neslin, S. A., Taylor, G. A., Grantham, K. D., & McNeil, K. R. (2013). Overcoming the "recency trap" in customer relationship management. *Journal of the Academy of Marketing Science, 41*(3), 320-337.

Netzer, O., Feldman, R., Goldenberg, J., & Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. Marketing Science, 31(3), 521-543.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons.* b, 4, 51-62.

Nazarov, A. D. (2019, December). Big Data Driven Marketing. In International Scientific and Practical Conference on Digital Economy (ISCDE 2019) (pp. 677-680). Atlantis Press.

Nguyen, T. H., Sherif, J. S., & Newby, M. (2007). Strategies for successful CRM implementation. Information Management & Computer Security, 15(2), 102-115.

Nimbalkar, D. D., & Shah, P. (2013). Data mining using RFM Analysis. International Journal of Scientific & Engineering Research (IJSRE), 4(12), 940-943.

Pearson, T., & Wegener, R. (2013). Big data: the organizational challenge. Bain Co.

Perreault Jr, W., Cannon, J., & McCarthy, E. J. (2012). Basic marketing. McGraw-Hill Higher Education.

Pride, W. M., & Ferrell, O. C. (2019). Marketing. Cengage Learning.

Punj, G., & Stewart, D. W. (1983). Cluster Analysis in Marketing Research: Review and Suggestions For Application. *Journal of Marketing Research, 20*(2), 134-148.

Rababah, K., Mohd, H., & Ibrahim, H. (2011). Customer relationship management (CRM) processes from theory to practice: The pre-implementation plan of CRM system. *International Journal of e-Education, e-Business, e-Management and e-Learning*, *1*(1), 22-27.

Ran, J., & Cheng, X. (2021, November). Airline Customer Value Analysis and Customer Churn Prediction Based on LRFMC Model and K-means Algorithm. In *2021 2nd International Conference on Computer Science and Management Technology (ICCSMT)* (pp. 185-193). IEEE.

Rana, M., & Bhushan, M. (2022). Machine learning and deep learning approach for medical image analysis: diagnosis to detection. Multimedia Tools and Applications, 1-39.

Rani, P., Mishra, N., & Diwedi, S. (2013). An efficient multi-set HPID3 algorithm based on RFM model. *International Journal of Computer Applications, 69*(1), 44-47.

Reichheld, F. F., & Kenny, D. W. (1990). The hidden advantages of customer retention. Journal of Retail Banking, 12(4), 19-24

Sarvari, P. A., Ustundag, A., & Takci, H. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. *Kybernetes*, *45*(7), 1129-1157.

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J, Ding, W., Lin, C. (2017). A review of clustering techniques and developments. *Neurocomputing*, *267*, 664-681.

Sheth, J., & Kellstadt, C. H. (2021). Next frontiers of research in data driven marketing: Will techniques keep up with data tsunami?. Journal of Business Research, 125, 780-784.

Shiha, M., & Ayvaz, S. (2017). The effects of emoji in sentiment analysis. International Journal of Computer Electrical Engineering (IJCEE.), 9(1), 360-369.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP conference series: materials science and engineering (Vol. 336, p. 012017). IOP Publishing

Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., & Rahmani, R. (2018). Customer segmentation and strategy development based on User Behavior Analysis, RFM model and data mining techniques: A case study. 2018 IEEE 15th International Conference on E-Business Engineering (ICEBE).

Tsiptsis, K. K., & Chorianopoulos, A. (2011). Data mining techniques in CRM: inside customer segmentation. John Wiley & Sons.

Wang, C. H. (2010). Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques. *Expert Systems with Applications*, *37*(12), 8395-8400.

Wendel, M., & Kamakura, W. A. (2000). Market segmentation: Conceptual and methodological foundations. Springer Science & Business Media.

Wei, J.-T., Lin, S.-Y., Weng, C.-C., & Wu, H.-H. (2012). A case study of applying LRFM model in market segmentation of a children's Dental Clinic. Expert Systems with Applications, 39(5), 5529–5533.

Wei, J. T., Lin, S. Y., & Wu, H. H. (2010). A review of the application of RFM model. African Journal of Business Management, 4(19), 4199.

Wu, H. H., Chang, E. C., & Lo, C. F. (2009). Applying RFM model and K-means method in customer value analysis of an outfitter. In Global Perspective for Competitive Enterprise, Economy and Ecology: Proceedings of the 16th ISPE International Conference on Concurrent Engineering (pp. 665-672). Springer London.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.

Xu, Y., Yen, D. C., Lin, B., & Chou, D. C. (2002). Adopting customer relationship management technology. Industrial management & data systems, *102*(8), 442-452.

Yankelovich, D., & Meer, D. (2006). Rediscovering market segmentation. Harvard business review, 84(2), 122.

YELMEN, İ., ÜSTEBAY, S., & Zontul, M. (2020). Customer Segmentation Based on Self-Organizing Maps: A Case Study on Airline Passengers. Journal of Aeronautics and Space Technologies, 13(2), 227-233.

Zaheri, F., Farughi, H., Soltanpanah, H., Alaniazar, S., & Naseri, F. (2012). Using multiple criteria decision making models for ranking customers of Bank network based on loyalty properties in weighted RFM model. *Management Science Letters*, *2*(2), 697-704.

Zhou, Z. (2022). Machine learning. S.l.: SPRINGER VERLAG, SINGAPOR, Chapter 1, 2-23.

Ελληνική Βιβλιογραφία

Βερύκιος, Β., Καγκλής, Β., & Σταυρόπουλος, Η. (2016). Η επιστήμη των δεδομένων μέσα από τη γλώσσα R.

Γεωργούλη, Α. (2016). Μηχανική Μάθηση. *Τεχνητή νοημοσύνη*. Kallipos, Open Academic Editions.

Ιστοσελίδες

American Marketing Association (2017). Definitions of Marketing. Retrieved March 28, 2023, from: https://www.ama.org/the-definition-of-marketing-what-is-marketing/

Higuera, V. (2016, October 26). Definition of data driven market research. Retrieved March 28, 2023, from: https://smallbusiness.chron.com/definition-data-driven-market-research-38715.html

Klingensmith, K. (2023, January 23). K-means clustering: An introductory guide and practical application. Retrieved March 28, 2023, from: https://towardsdatascience.com/k-means-clustering-an-introductory-guide-and-practical-application-dce70bfa4249

Sharma, N. (2021, September 27). Understanding the mathematics behind K-means clustering. Retrieved April 5, 2023, from: https://heartbeat.comet.ml/understanding-the-mathematics-behind-k-means-clustering-40e1d55e2f4c

WebEngage (χ.χ.). RFM analysis. Retrieved March 28, 2023, from: https://knowledgebase.webengage.com/docs/predictive-rfm-segmentation

ΠΑΡΑΡΤΗΜΑ

Cleaning script:

```
#Import Libraries
import pandas as pd
import datetime as dt
import plotly.express as px
import matplotlib.pyplot as plt
```

```
#Read the excel file of dataset
df = pd.read_excel(r'C:\Users\vasol\OneDrive\Υπολογιστής\Ergasia\Online Retail 2010-2011.xlsx')
df
```

```
#Unterstanding data
```

```
#Dimensions of DataFrame, number of rows and columns
df.shape
```

```
#First 10 rows of dataset
df.head(10)
```

```
#Last 5 rows of dataset
df.tail(5)
```

```
#Return information about the dataset (names of columns, number of observations and dtype of data)
#Missing values are obsereved in Description and CustomerID
df.info()
```

```
#Descreptive statistics of data
#There are also negative values in Quantuty and UnitPrice,need to remove them
df.describe()
```

```
#Data cleaning and Exploratory Data Analysis
```

```
#Check missing values in Description and CustomerID and return total number of them
df.isnull().sum()
```

```
#Remove missing values from dataset
df.dropna(inplace=True)
```

```
#Check that missing values removed
df.isnull().sum()
```

```
#New shape of DataFrame
df.shape
```

```
#Cancelled orders: observe negative values in Quantity and the same time letter 'C' at InvoiceNo. Should remove them
df[df['Quantity'] < 0].head(10)
```

```
#Count number of negative values in Quantity and save it in a new variable #Cancelled orders 8905/406829 2.19%
negative_values = (df['Quantity'] < 0).value_counts()
negative_values
```

```
#Define a new variable and store a list with the labels for the pie chart
label = ['Non cancelled: quantity > 0','Cancelled: quantity < 0']
```

```
#Pie chart for Quantity
fig = px.pie(df,
             values = negative_values,
             names = label ,
             title = 'Proportion of cancelled orders: positive/negative quantity values')
fig.show()
```

```
#Zero values in UnitPrice, probably are gifts or defective products. Should remove them
#There are many zero values with no customerID
df[df['UnitPrice'] == 0].head(5)
```

```
#Define a new variable and store a list with the labels to create the pie chart
label_1 = ['non zero', 'zero']
```

```
#Count number of zero and non zero values in UnitPrice and store it in a new variable
zero_values = (df['UnitPrice'] == 0).value_counts()
zero_values
```

```
#Pie chart for zero UnitPcice values
fig = px.pie (df,
              values = zero_values,
              names = label_1,
              title = 'Proportion of gifts/defective products: UnitPrice zero/non zero values')
fig.show()
```

```
#Boxplot - Negative and zero values in Quantity and UnitPrice respectively
df[['Quantity','UnitPrice']].boxplot()
```

```
#Remove negative and zero values from Quantity and UnitPrice
df = df[(df['Quantity'] > 0) & (df['UnitPrice'] > 0)]
df.shape
```

```
#Boxplot to check again the values
df[['Quantity','UnitPrice']].boxplot()
```

```
#Clean data from noise-specific values (ex. discounts, manual, sample, carriage, post, amazon fee, bank charges)
df = df[df['StockCode'] != 'D']
df = df[df['StockCode'] != 'M']
df = df[df['StockCode'] != 'S']
df = df[df['StockCode'] != 'C2']
df = df[df['StockCode'] != 'POST']
df = df[df['StockCode'] != 'AMAZONFEE']
df = df[df['StockCode'] != 'BANK CHARGES']
df.shape
```

```
#Data preperation
```

```
#Create new variable multiplying Quantity with UnitPrice
df['TotalPrice'] = df['Quantity'] * df['UnitPrice']
df
```

```
#Exploratory Data Analysis (EDA)
#Explore Country
```

```
df.describe()
```

```
#Unique number of customers
df['CustomerID'].nunique()
```

```
#Unique number of countries
df['Country'].nunique()
```

```
#Invoices number per country
df.groupby('Country')['InvoiceNo'].count().sort_values(ascending = False)
```

```
#Invoices number by country
fig = px.histogram (df,
                    x = 'Country',
                    title = 'Invoices by Countries',
                    color = 'Country').update_xaxes(categoryorder = 'total ascending')
fig.show()
```

```
#Explore TotalPrice
```

```
#Total amount by country
df.groupby('Country')['TotalPrice'].sum().sort_values(ascending=False).round(2)
```

```
#10 best countries by revenue
df.groupby('Country')['TotalPrice'].sum().sort_values(ascending=True).round(2).tail(10)
```

```
#10 best countries by revenue
df.groupby('Country')['TotalPrice'].sum().sort_values(ascending=True).round(2).tail(10).plot.barh (ylabel = 'Country',
                                                                                                   width = 0.8,
                                                                                                   title = 'Countries with highe
                                                                                                   figsize = (15,8))
```

```
#Explore InvoiceDate and Quantity
```

```
#First and last selling date from dataset
df['InvoiceDate'].min()
```

```
df['InvoiceDate'].max()
```

```
#Total sales per month
df.groupby(pd.Grouper(key = 'InvoiceDate', freq = '1M')).sum()
```

```
#Sales per month
fig = px.histogram(df,
                   x = 'InvoiceDate',
                   y='TotalPrice',
                   title = 'Total sales per month',
                   nbins = 13)
fig.show()
```

```
#Find total quantity of purchases by country
#best country by quantity
df.groupby('Country')['Quantity'].sum().sort_values()
```

```
#Volume sales by country
fig = px.histogram(df,
              x = 'Country',
              y = 'Quantity',
              color = 'Country',
              title = 'Total quantity by country')
fig.show()
```

```
#Sales volume per month
fig = px.histogram(df,
                   x='Quantity',
                   y = 'InvoiceDate',
                   title = 'Volume sales per month',
                   nbins = 13)
fig.show()
```

```
#To remove time from InvoiceDate column, need to convert
#InvoiceDate type from object to datetime
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])
df['InvoiceDate']
```

```
#Select only date without time
df['InvoiceDate'] = df['InvoiceDate'].dt.date
df['InvoiceDate']
```

```
#Explore InvoiceNo and StockeCode
```

```
#Orders per day and sort values from lowest to highest
orders_perday = df.groupby('InvoiceDate')['InvoiceNo'].count().sort_values()
orders_perday
```

```
#Orders rate per day
plt.figure(figsize = (15,8))
df.groupby('InvoiceDate')['InvoiceNo'].count().plot(ylabel = 'Invoices number',
                                                    title = 'Orders rate per day')
```

```
#Total amount per day and sort values from lowest to highest
revenue_perday = df.groupby('InvoiceDate')['TotalPrice'].sum().sort_values()
revenue_perday
```

```
#Total invoices per customer #best customers by order
df.groupby('CustomerID')['InvoiceNo'].count().sort_values(ascending = True).tail(10)
```

```
#Most frequent customers
plt.figure(figsize = (15,8))
df.groupby('CustomerID')['InvoiceNo'].count().sort_values(ascending = True).tail(10).plot(kind = 'bar',
                                                                                          ylabel = 'InvoiceNo',
                                                                                          title = 'Customers with most orders')
```

```
#Total amount of purchases per customerid#best customers by totalprice
df.groupby('CustomerID')['TotalPrice'].sum().sort_values(ascending = True).tail(10)
```

```
#Best customers by revenue
df.groupby('CustomerID')['TotalPrice'].sum().sort_values(ascending = True).tail(10).plot(kind = 'bar',
                                                                                         ylabel = 'TotalPrice',
                                                                                         title = 'Best customers by revenue',
                                                                                         figsize = (15,8))
```

```
#Most ordered products
df.groupby('StockCode')['Quantity'].count().sort_values().tail(10)
```

```
#Bar chart for most ordered products
df.groupby('StockCode')['Quantity'].count().sort_values().tail(10).plot(kind = 'bar',
                                                                        ylabel = 'Quantity',
                                                                        title = 'Most ordered Products',
                                                                        figsize = (15,8))
```

```
#RFM analysis to excel
```

```
excel_file = pd.ExcelWriter('Data_cleaning_process.xlsx')
```

```
df.to_excel(excel_file)
```

```
excel_file.save()
```

RFM – kmeans script:

```
#kmeans clustering
```

```
import numpy as np
import pandas as pd
import plotly.express as px
import matplotlib.pyplot as plt

from scipy.stats import skew
from scipy.stats import shapiro
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import StandardScaler
```

```
#Read the excel file of rfm dataset
df = pd.read_excel(r'C:\Users\vasol\OneDrive\Υπολογιστής\Ergasia\RFM_Analysis.xlsx', sheet_name = 'K-Means Clustering_data')
df
```

```
#Customer number by segment
seg = df['Segment'].value_counts()
seg
```

```
#Mean for  R,F,M metrics
cust_seg = df.groupby('Segment')[['Recency','Frequency','Monetary']].mean()
cust_seg
```

```
segments = ['Need Attention', 'Loyal Customers', 'Hybernating', 'Champions', 'New Customers', 'Cant lose']
```

```
#Segments proportion/Customer distribution by segment
fig = px.pie(df,
             values = seg,
             names = segments,
             title = 'Proportion of Segments')
fig.show()
```

```
fig = px.histogram(df, x = 'Segment', color = 'Segment', text_auto=True)
fig.show()
```

```
#Segments
fig = px.scatter(df, x= 'RFM score', color='Segment')
fig.show()
```

```
#Segments by recency - frequency
fig = px.scatter(df, x = 'Recency', y = 'Frequency', color = 'Segment')
fig.show()
```

```
#Segments by rfm metrics
fig = px.scatter_3d(df,
                    x = 'Recency',
                    y = 'Frequency',
                    z = 'Monetary',
                   color = 'Segment')
fig.show()
```

```
#KMeans Algorithm
#Apply kmeans clustering algorithm to create clusters.
#To apply kmeans clustering should check mean, variance and distributions
#Transform the data with log to handle skewness and then use
#Standardization to handle the problem with mean and standard deviation
#Distributions are not normal but right skewed (positive)
```

```
#Descriptive statistics to check R,F,M metrics values for kmeans
df.describe()
```

```
#Recency distribution and boxplot
fig = px.histogram (df, x = 'Recency',
                    title = 'Recency Distribution')
fig.show()
```

```
fig = px.box(df, x = 'Recency',
             title = 'Recency boxplot')
fig.show()
```

```
#Frequency distribution and boxplot
fig = px.histogram (df, x = 'Frequency',
                    title = 'Frequency Distribution')
fig.show()
```

```
fig = px.box(df, x = 'Frequency',
             title = 'Frequency boxplot')
fig.show()
```

```
#Monetary distribution and boxplot
fig = px.histogram (df, x = 'Monetary',
                    title = 'Monetary Distribution')
fig.show()
```

```
fig = px.box(df, x = 'Monetary',
             title = 'Monetary boxplot')
fig.show()
```

```
#First should check the skewness with skew() function
#Recency has skewness but frequency and monetary have extremly high skewness
#Need to handle the skewness with log transformation

df[['Recency','Frequency','Monetary']].skew(axis = 0)
```

```
#Shapiro-Wilk normality test - null hypothsesis is that data are normally distributed
#If p-value<0.05, then there is sufficient evidence that data are not normally distributed and null hypothesis is rejected
#If p-value>0.05. then there data are normally distributed, null hypothesis is accepted
shapiro(df[['Recency','Frequency','Monetary']])
```

```
#Should check the correlations between the values
values = df[['Recency','Frequency','Monetary']]
```

```
#Correlations between metrics
values.corr()
```

```
#Distrubutions before log transformation (pairplot)
pd.plotting.scatter_matrix(df[['Recency','Frequency','Monetary']], figsize=(10,10), marker = 'o')
```

```
#Log transformation with np.log to manage the skewness
logdata = df[['Recency','Frequency','Monetary']].apply(np.log).round(2)
logdata
```

```
#Check skewness again
#There is skewness but it is very small. Data tend to follow the normal distribution
logdata.skew(axis = 0)
```

```
#Correlation R-F-M
fig = px.scatter(logdata,
                 x = 'Recency',
                 y = 'Frequency',
                 color = 'Monetary')
fig.show()
```

```
#Correlation R-F
fig = px.scatter(logdata,
                 x = 'Recency',
                 y = 'Frequency')
fig.show()
```

```
#Correlation R-M
fig = px.scatter(logdata,
                 x = 'Monetary',
                 y = 'Recency')
fig.show()
```

```
#Correlation F-M
fig = px.scatter(logdata,
                 x = 'Frequency',
                 y = 'Monetary')
fig.show()
```

```
#Distrubutions after log transformation
pd.plotting.scatter_matrix(logdata, figsize=(10,10), marker = 'o')
```

```
#Apply scaler to logdata to fix the scale
#Apply kmeans algorithm
#Define optimal number of clusters with elbow method
```

```
#Standardization logdata to scale data with mean = 0 and std = 1
scale = StandardScaler()
scale.fit(logdata)
normalized = scale.transform(logdata)
normalized
```

```
#Box and whisker plots for normalized data
fig = px.box(normalized)
fig.show()
```

```
#Kmeans clustering algorithm - Build the model
#Create centroids for 5 clusters
kmeans = KMeans(n_clusters = 5)
kmeans.fit(normalized)
predict = kmeans.predict(normalized)
df['k_clusters'] = predict
kmeans.cluster_centers_
```

```
#Cluster index for each point
predict
```

```
#Create a list to hold sse values (sum of squared error) for each k and
#visualize the results to define best number of clusters
sse = []
r = range(1, 10)
for k in r :
    kmeans = KMeans(n_clusters = k)
    kmeans.fit(normalized)
    sse.append(kmeans.inertia_)

plt.figure(figsize = (10,6))
plt.plot(r, sse, marker = 'o')
plt.xlabel('Number of Clusters')
plt.ylabel('Sum of Squares Error (SSE)')
plt.title('Elbow for kmeans clustering')
plt.show()
```

```
#Apply again with optimal number of clusters - 3 centroids created
kmeans = KMeans(n_clusters = 3)
kmeans.fit(normalized)
predict = kmeans.predict(normalized)
df['k_clusters'] = predict
centroids = kmeans.cluster_centers_
centroids
```

```
#Cluster index for each point
predict
```

```
#Check k_clusters column in dataset
df
```

```
#Customer number per cluster
clusters = df['k_clusters'].value_counts()
clusters
```

```
#Means for k_clusters
df.groupby('k_clusters')[['Recency','Frequency','Monetary']].mean()
```

```
clusters_labels = ['Mid value','Low value','High value']
```

```
#Clusters proportion
fig = px.pie(df,
             values = clusters,
             names = clusters_labels,
             title = 'Proportion of k_clusters')
fig.show()
```

```
#Customer number by cluster
fig = px.histogram(df['k_clusters'], color = df['k_clusters'], text_auto=True)
fig.show()
```

```
pd.plotting.scatter_matrix(logdata, figsize = (10,10), marker = 'o', c = df['k_clusters'])
```

```
#Correllations by cluster
fig = px.scatter_matrix(logdata,
                        dimensions = ['Recency','Frequency', 'Monetary'],
                        color = df['k_clusters'],
                        title = 'Correllations')
fig.show()
```

```
#Frequency - Monetary by k_clusters - created clusters
fig = px.scatter(logdata,
                 x = 'Frequency',
                 y = 'Monetary',
                 color = df['k_clusters'],
                 title = 'Frequency - Monetary correllation')
fig.show()
```

```
#Export to excel to compare the results from rfm analysis and kmeans clustering
```

```
excel_file = pd.ExcelWriter('Data2.xlsx')
```

```
df.to_excel(excel_file)
```

```
excel_file.save()
```

Πρόγραμμα Μεταπτυχιακών Σπουδών

στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

Διπλωματική Εργασία

Θέμα: «Ανάλυση καταναλωτικής συμπεριφοράς και εφαρμογή κατάλληλης στρατηγικής μάρκετινγκ με τη χρήση της ανάλυσης RFM και του μοντέλου μηχανικής μάθησης kmeans clustering»

της

Λαζαρίδου Βασιλικής του Γρηγορίου

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

Ιούλιος 2023