Πρόγραμμα Μεταπτυχιακών Σπουδών

στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

Διπλωματική Εργασία

ΤΜΗΜΑΤΟΠΟΙΗΣΗ ΚΑΤΑΝΑΛΩΤΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΤΗΣ ΜΕΘΟΔΟΥ RFM ΚΑΙ ΤΟΥ ΕΡΓΑΛΕΙΟΥ ΟΠΤΙΚΟΠΟΙΗΣΗΣ ΔΕΔΟΜΕΝΩΝ MICROSOFT POWER BI

της

Αργυρής Μπαζούκα του Γεωργίου

Επιβλέπων καθηγητής: Λεωνίδας Χατζηθωμάς

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

Δεκέμβριος 2024

Περίληψη

Οι επιχειρήσεις της σύγχρονης εποχής καλούνται πλέον να διαχειριστούν τα υπέρογκα δεδομένα που σχετίζονται με τους πελάτες τους. Η αποτελεσματική ερμηνεία και η αξιοποίησή τους αποτελούν προκλήσεις για το τμήμα μάρκετινγκ, το οποίο φέρει τη διττή ευθύνη για την διαμόρφωση στρατηγικών στόχευσης, οι οποίες ανταποκρίνονται αφενός στις ανάγκες των πελατών, αφετέρου στη βελτίωση της κερδοφορίας της ίδιας της επιχείρησης. Η ανάλυση RFM αποτελεί μία από τις πιο δημοφιλείς μεθόδους τμηματοποίησης πελατών, χάρη στην απλότητά της και την ευκολία ερμηνείας των αποτελεσμάτων της. Η μέθοδος δημιουργεί ομάδες πελατών με βάση χαρακτηριστικά της συμπεριφοράς τους, εστιάζοντας στο χρόνο από την τελευταία αγορά, την συχνότητα αγορών, και τη συνολική χρηματική δαπάνη. Η ανάλυση RFM συνδυάζεται συχνά με αλγόριθμους συσταδοποίησης για την επίτευξη πιο στοχευμένων αποτελεσμάτων τμηματοποίησης. Παράλληλα, λόγω του μεγάλου όγκου των δεδομένων, δημιουργείται η ανάγκη κατανόησης των συμφραζομένων μέσα στα οποία συμβαίνουν οι αγορές, και από όπου τα δεδομένα προκύπτουν. Τα συστήματα επιχειρηματικής ευφυΐας επιτρέπουν στα στελέχη των επιχειρήσεων να κατανοούν και να διαχειρίζονται τα δεδομένα που έχουν στη διάθεσή τους, παρέχοντας δυνατότητες όπως η προεπεξεργασία και η οπτικοποίησή τους. Ο συνδυασμός τους με την ανάλυση RFM επιτρέπει την εμβάθυνση στην συμπεριφορά των καταναλωτών και συνεπώς στην καλύτερη οργάνωσή τους σε ομάδες όμοιων χαρακτηριστικών, η οποία οδηγεί στην επίτευξη της βέλτιστης στόχευσής τους.

Στην παρούσα εργασία παρουσιάζονται τρεις μέθοδοι τμηματοποίησης των πελατών ενός ηλεκτρονικού καταστήματος που εμπορεύεται είδη δώρων. Το σύνολο δεδομένων που χρησιμοποιείται περιλαμβάνει τις συναλλαγές του καταστήματος κατά τη διάρκεια δύο οικονομικών ετών. Βάση για τις τρεις μεθόδους αποτελούν η ανάλυση RFM και ο αλγόριθμος συσταδοποίησης K-means, ο συνδυασμός των οποίων χρησιμοποιείται κατά κόρον στη βιβλιογραφία. Σε αντίθεση όμως με αυτήν, και έχοντας ως είσοδο το προκείμενο σύνολο δεδομένων, η σύγκριση των τριών μεθόδων δείχνει πως η ωφελιμότερη για την επιχείρηση τμηματοποίηση προκύπτει μέσω RFM χωρίς τη χρήση της μεθόδου K-means. Πριν τη σύγκριση αυτή, διενεργείται περιγραφική και διερευνητική ανάλυση του συνόλου δεδομένων. Για την οπτικοποίηση, κατανόηση, και εμβάθυνση στις λεπτομέρειες του συνόλου δεδομένων και των αναλύσεων που διεξάγονται, γίνεται χρήση του εργαλείου επιχειρηματικής ευφυΐας Microsoft Power BI.

Πίνακας Περιεχομένων

Κεφάλαιο 1: Εισαγωγή

1.1 Περιγραφή του προβλήματος

Ένα από τα βασικότερα ζητήματα για την λειτουργία μιας επιχείρησης αποτελεί η επικοινωνία και η σύνδεση των προϊόντων της με το καταναλωτικό κοινό. Τον σκοπό αυτό αναλαμβάνει να υλοποιήσει το τμήμα μάρκετινγκ, το οποίο, μέσω της κατάρτισης του σχεδίου μάρκετινγκ, επιχειρεί να ορίσει τις στρατηγικές θέσης, τιμολόγησης, διανομής, και προώθησης των προσφερόμενων προϊόντων/υπηρεσιών της επιχείρησης στους καταναλωτές.

Βασικό βήμα της ανάπτυξης του σχεδίου μάρκετινγκ αποτελεί η ανάλυση της αγοράς όπου δραστηριοποιείται η επιχείρηση. Η ανάλυση αυτή περιλαμβάνει την μελέτη των πελατών, την κατανόηση των αναγκών τους, και την εξέταση του ανταγωνισμού. Στόχος αυτής της ανάλυσης είναι ο σχεδιασμός των κατάλληλων στρατηγικών που θα επιτρέψουν στην επιχείρηση να πετύχει την αποτελεσματικότερη προσέγγιση των πελατών της, μεγιστοποιώντας παράλληλα τα κέρδη της, και αξιοποιώντας κατά τον βέλτιστο δυνατό τρόπο τους πόρους που διαθέτει.

Σε αυτό το πλαίσιο αναπτύχθηκε από νωρίς η έννοια του διαχωρισμού των πελατών μίας επιχείρησης σε ομάδες που παρουσιάζουν ομοιομορφία και συνοχή στο εσωτερικό τους, αλλά που παραμένουν διακριτές μεταξύ τους, κάτι που σήμερα ονομάζεται *τμηματοποίηση των πελατών της επιχείρησης*. Η τμηματοποίηση αυτή μπορεί να βασίζεται σε δημογραφικά, γεωγραφικά, ψυχογραφικά, και άλλα χαρακτηριστικά. Μία από τις πιο δημοφιλείς τεχνικές τμηματοποίησης είναι η ανάλυση RFM, η οποία είναι μέθοδος που ομαδοποιεί τους πελάτες με βάση χαρακτηριστικά της συμπεριφοράς τους. Συγκεκριμένα βασίζεται σε τρεις μεταβλητές: στον χρόνο που μεσολάβησε από την πιο πρόσφατη αγορά ενός καταναλωτή (***R****ecency*), στην συχνότητα των αγορών του (***F****requency*), και στο συνολικό χρηματικό ποσό που δαπάνησε για τις αγορές του (***M****onetary*). Η ανάλυση RFM αποτέλεσε τη βάση για μεταγενέστερες μεθόδους τμηματοποίησης, κάποιες από τις οποίες την επέκτειναν προσθέτοντας επιπλέον μεταβλητές, ενώ άλλες την χρησιμοποίησαν σε συνδυασμό με τρίτες μεθόδους, συμβάλλοντας στην ανάπτυξη πιο εύρωστων μοντέλων. Σε αυτή την κατηγορία ανήκουν οι μέθοδοι που ερευνούν συνδυασμούς της RFM με αλγόριθμους μη επιβλεπόμενης μηχανικής μάθησης, και ειδικότερα οι - δημοφιλείς - συνδυασμοί της μεθόδου RFM με μεθόδους συσταδοποίησης.

Σύμφωνα με τους Sinaga και Yang (2020), οι αλγόριθμοι συσταδοποίησης χωρίζουν μια συλλογή από σημεία σε ομάδες ανάλογα με τις ομοιότητες και τις διαφορές τους. Ο αλγόριθμος K-means αποτελεί έναν από τους πιο συχνά εφαρμοζόμενους αλγόριθμους συσταδοποίησης, ο οποίος χωρίζει *n* σημεία στον Δ-διάστατο χώρο σε *k* συστάδες βάσει του μέσου όρου της απόστασης των σημείων από το κέντρο κάθε συστάδας. Η απλότητα της υλοποίησης τόσο του αλγόριθμου K-means όσο και της RFM ανάλυσης έχει συντελέσει σημαντικά στην συνδυαστική χρήση των δύο μεθόδων για την τμηματοποίηση πελατών επιχειρήσεων που προέρχονται από διάφορους βιομηχανικούς κλάδους.

Παράλληλα, η σύγχρονη εποχή κυριαρχείται από πληθώρα διαθέσιμων δεδομένων, που προέρχονται από πολλαπλές πηγές, και που κινούνται ταχύτατα. Το φαινόμενο αυτό έχει άμεση επίδραση στις ίδιες τις επιχειρήσεις, οι οποίες καλούνται να χρησιμοποιήσουν τα διαθέσιμα δεδομένα για να προσεγγίσουν με αποτελεσματικό τρόπο το καταναλωτικό κοινό και να κερδίσουν έδαφος έναντι του ανταγωνισμού.

Για να μπορέσουν να διαχειριστούν αυτόν τον ολοένα αυξανόμενο όγκο πληροφοριών, οι επιχειρήσεις έχουν πλέον στη διάθεσή τους πολλαπλά εργαλεία όπως είναι τα συστήματα διαχείρισης πελατειακών σχέσεων (Customer Relationship Management - CRM), τα συστήματα διαχείρισης αποθεμάτων και λογιστικής. Σε αυτά έρχονται να προστεθούν και τα εργαλεία επιχειρηματικής ευφυΐας (Business Intelligence), όπως είναι το Microsoft Power BI και το Tableau, τα οποία επιτρέπουν μία μεγάλη γκάμα ενεργειών γύρω από την σύνδεση, την επεξεργασία και την ανάλυση των επιχειρησιακών δεδομένων. Από την προεπεξεργασία και τον καθαρισμό πρωτογενών (raw) δεδομένων μέχρι την οπτικοποίηση σημαντικών ευρημάτων μέσω της δημιουργίας αναφορών (reports) και διαδραστικών ταμπλό (dashboards) για τον τελικό χρήστη, τα συστήματα επιχειρηματικής ευφυΐας μπορούν να συνδράμουν σημαντικά στη λήψη αποφάσεων σε τομείς όπως το μάρκετινγκ, η διαχείριση ανθρώπινου δυναμικού κ.α.

Συνδυάζοντας τα παραπάνω, μπορούμε συνοπτικά να προσδιορίσουμε το πρόβλημα που αποτελεί αντικείμενο της παρούσας διπλωματικής εργασίας. Αυτό δεν είναι άλλο από την ανεύρεση ενός τρόπου αξιοποίησης του διαθέσιμου όγκου πληροφοριών για αποτελεσματική τμηματοποίηση των πελατών μιας επιχείρησης μέσω RFM και K-means, καθώς και την οπτικοποίηση μέσω Power BI αυτού του όγκου των εν δυνάμει χρήσιμων δεδομένων, με τελικό σκοπό τη συνδρομή των επιχειρησιακών στελεχών στον σχεδιασμό της κατάλληλης στρατηγικής μάρκετινγκ για την προσέγγιση της κάθε ομάδας, βοηθώντας την έτσι να αξιοποιήσει τους πόρους που διαθέτει ώστε να μεγιστοποιήσει τα κέρδη της.

1.2 Στόχοι έρευνας και δομή εργασίας

Στόχος της παρούσας εργασίας είναι η επίλυση του προβλήματος που παρουσιάστηκε στην προηγούμενη ενότητα έχοντας ως βάση τα δεδομένα ενός ηλεκτρονικού καταστήματος που εμπορεύεται είδη δώρων. Ειδικότερα, το ιστορικό των συναλλαγών που πραγματοποιήθηκαν σε χρονικό διάστημα δύο ετών από πελάτες του ηλεκτρονικού καταστήματος αποτελεί το σύνολο δεδομένων που λήφθηκε ως βάση για την υλοποίηση της έρευνας.

Συγκεκριμένα, η παρούσα εργασία συμβάλλει στην έρευνα γύρω από την εφαρμογή της ανάλυσης RFM και του αλγόριθμου μηχανικής μάθησης K-means, υλοποιώντας τρεις διαφορετικές μεθόδους τμηματοποίησης των πελατών που βασίζονται σε αυτές τις τεχνικές. Αυτές παρουσιάζονται στο Κεφάλαιο 3: και συγκεκριμένα στις ενότητες 3.3.2.1, 3.3.2.2 και 3.3.2.3, και ακολουθούν την παρουσίαση του συνόλου δεδομένων που χρησιμοποιήθηκε για την ανάλυση (ενότητα 3.2) και την περιγραφή της ερευνητικής διαδικασίας (ενότητα 3.1) που ακολουθήθηκε συνολικά. Όσον αφορά το κομμάτι της οπτικοποίησης των δεδομένων, η εργασία επιχειρεί να αναδείξει τη συνδρομή του εργαλείου επιχειρηματικής ευφυΐας Power BI τόσο στην προεπεξεργασία των διαθέσιμων δεδομένων όσο και στην παρουσίαση των αποτελεσμάτων της τμηματοποίησης για την λήψη στρατηγικών αποφάσεων από τα στελέχη της ανώτερης διοίκησης. Το Power BI χρησιμοποιήθηκε για την υλοποίηση της διερευνητικής και περιγραφικής ανάλυσης του συνόλου δεδομένων (ενότητα 3.2.2), για την υλοποίηση των μεθόδων τμηματοποίησης (ενότητα 3.3), όσο και κατά την παρουσίαση των αποτελεσμάτων τμηματοποίησης που προέκυψαν στο τέλος της μεθοδολογίας (Κεφάλαιο 4:).

Η υπόλοιπη εργασία διαρθρώνεται ως εξής: το Κεφάλαιο 2: παρουσιάζει την σχετική βιβλιογραφία, εισάγοντας τον αναγνώστη στην έννοια, την προέλευση και τους τύπους της τμηματοποίησης των πελατών (ενότητα 2.1). Ακολουθεί μια εισαγωγή στις βασικές τεχνικές τμηματοποίησης που πραγματεύεται η παρούσα εργασία, και συγκεκριμένα, στην ανάλυση RFM και τις δημοφιλείς παραλλαγές της (ενότητα 2.2) καθώς και στους αλγόριθμους συσταδοποίησης και ειδικότερα στον αλγόριθμο K-means (ενότητα 2.3). Μετά την παρουσίαση των δύο μεθόδων η εργασία προχωρά στην παρουσίαση βασικών άρθρων της βιβλιογραφίας που συντέλεσαν στον συνδυασμό των μεθόδων για την παραγωγή βελτιωμένων αποτελεσμάτων τμηματοποίησης (ενότητα 2.4). Το κεφάλαιο της βιβλιογραφικής επισκόπησης ολοκληρώνεται με την ενότητα 2.5, όπου παρουσιάζονται οι δυνατότητες οπτικοποίησης και διαδραστικότητας που προσφέρουν τα συστήματα επιχειρηματικής ευφυΐας σε αναλύσεις μάρκετινγκ.

Η μεθοδολογία που ακολουθήθηκε στην παρούσα εργασία παρουσιάζεται στο Κεφάλαιο 3:, όπου όπως προαναφέρθηκε υλοποιούνται τρεις διαφορετικές μέθοδοι τμηματοποίησης: η ανάλυση RFM μέσω πεμπτημορίων (*percentiles*), η ανάλυση RFM μέσω συσταδοποίησης ανά στήλη (*column k-means*) και η συσταδοποίηση μέσω K-means. Μετά την παρουσίαση της μεθοδολογίας ακολουθεί το Κεφάλαιο 4:, όπου αρχικά γίνεται μία σύγκριση των αποτελεσμάτων τμηματοποίησης των μεθόδων percentiles και column k-means (ενότητα 4.1), στην συνέχεια παρουσιάζει τα αποτελέσματα της συσταδοποίησης μέσω K-means (ενότητα 4.2) και συνοψίζει συγκρίνοντας τα αποτελέσματα τμηματοποίησης που παρήχθησαν από την εφαρμογή των τριών μεθόδων.

Στο Κεφάλαιο 5: παρατίθενται τα βασικά συμπεράσματα της παρούσας εργασίας (ενότητα 5.1) και γίνεται μία συζήτηση γύρω από αυτά σε σχέση με παρόμοιες έρευνες της βιβλιογραφίας (ενότητα 5.2). Στο τελικό μέρος του κεφαλαίου προτείνονται στρατηγικές μάρκετινγκ προς τους υπεύθυνους λήψης αποφάσεων της επιχείρησης για την προσέγγιση των τμημάτων πελατών που δημιουργήθηκαν (ενότητα 5.3). Η εργασία ολοκληρώνεται με το Κεφάλαιο 6:, όπου παρουσιάζονται οι περιορισμοί της παρούσας έρευνας (ενότητα 6.1) και παρατίθενται προτάσεις για μελλοντική έρευνα (ενότητα 6.2).

Ο κώδικας που αναπτύχθηκε κατά την υλοποίηση της παρούσας εργασίας είναι διαθέσιμος στο δημόσιο αποθετήριο κώδικα GitHub: https://github.com/ArgyroMp/msc-thesis.

Κεφάλαιο 2: Βιβλιογραφική Επισκόπηση

2.1 Τμηματοποίηση πελατών

Η τμηματοποίηση των πελατών παραμένει ακρογωνιαίος λίθος της στρατηγικής μάρκετινγκ διότι επιτρέπει στις επιχειρήσεις να στοχεύουν αποτελεσματικότερα σε συγκεκριμένες ομάδες πελατών, αυξάνοντας τη συνάφεια και τον αντίκτυπο των εκστρατειών μάρκετινγκ. Εστιάζοντας στους σωστούς πελάτες, οι επιχειρήσεις μπορούν να βελτιώσουν την ικανοποίηση των πελατών, την αφοσίωση, και τη συνολική κερδοφορία (Wind & Bell, 2008).

Οι Rust και Verhoef (2005) υπογραμμίζουν τη σημασία των δεδομένων των πελατών στην καθοδήγηση των αποφάσεων μάρκετινγκ, βοηθώντας τις επιχειρήσεις να κατανέμουν καλύτερα τους πόρους και να βελτιστοποιούν τις παρεμβάσεις μάρκετινγκ, θέτοντας τις βάσεις για το μέλλον των στρατηγικών μάρκετινγκ με βάση τα δεδομένα και την εξατομίκευση.

Η βέλτιστη τμηματοποίηση μπορεί να βελτιώσει σημαντικά τις λύσεις εξατομίκευσης, με τις μεθόδους άμεσης ομαδοποίησης να υπερτερούν συχνά των παραδοσιακών προσεγγίσεων που βασίζονται στη στατιστική (Jiang & Tuzhilin, 2006). Συνολικά, η τμηματοποίηση πελατών διαδραματίζει ζωτικό ρόλο στην κατανόηση των αναγκών των πελατών, στη βελτίωση της διατήρησης και στην αύξηση των πωλήσεων στο σημερινό ανταγωνιστικό επιχειρηματικό περιβάλλον (Shinde et al., 2023).

2.1.1 Προέλευση και πρώτες έρευνες για την τμηματοποίηση πελατών

Η έννοια της τμηματοποίησης πελατών προήλθε από τις πρώτες εργασίες για την τμηματοποίηση της αγοράς στην έρευνα μάρκετινγκ. Η πρώτη σημαντική ακαδημαϊκή συζήτηση εισήχθη το 1956 από τον Wendell R. Smith στην επιδραστική εργασία του με τίτλο «Διαφοροποίηση προϊόντος και τμηματοποίηση αγοράς ως εναλλακτικές στρατηγικές μάρκετινγκ» (“Product Differentiation and Market Segmentation as Alternative Marketing Strategies). Ο Smith υποστήριξε ότι η μαζική αγορά θα μπορούσε να χωριστεί σε διακριτές ομάδες με βάση τα κοινά σημεία μεταξύ των καταναλωτών, οδηγώντας σε πιο στοχευμένες στρατηγικές μάρκετινγκ. Τόνισε ότι η τμηματοποίηση επιτρέπει στις εταιρείες να ικανοποιούν τις ανάγκες διαφορετικών ομάδων καταναλωτών πιο αποτελεσματικά, προσαρμόζοντας τα προϊόντα και τις προσεγγίσεις μάρκετινγκ (Smith, 1956).

Καθώς η ιδέα της τμηματοποίησης της αγοράς κέρδισε δημοτικότητα, αρκετοί μελετητές συνέβαλαν στην εξέλιξή της. Για παράδειγμα, ο Frank Bass (1969) ανέπτυξε το μοντέλο διάχυσης Bass, το οποίο χρησιμοποιείται για την ανάλυση του τρόπου με τον οποίο τα νέα προϊόντα υιοθετούνται με την πάροδο του χρόνου, με τους πρώιμους και τους όψιμους υιοθετούντες να σχηματίζουν διακριτά τμήματα της αγοράς. Αυτό το μοντέλο έθεσε τα θεμέλια για την τμηματοποίηση των πελατών με βάση τα πρότυπα υιοθέτησης καινοτομιών, το οποίο έκτοτε επηρέασε πολυάριθμους κλάδους, ιδίως στις αγορές υψηλής τεχνολογίας (Bass, 1969).

Περαιτέρω εξελίξεις στην τμηματοποίηση προήλθαν από την εργασία των Valentine και Powers (1972) για τη δημογραφική τμηματοποίηση, όπου οι ερευνητές διερεύνησαν πώς μεταβλητές όπως το εισόδημα, η εκπαίδευση και η ηλικία επηρεάζουν τη συμπεριφορά και τις προτιμήσεις των καταναλωτών. Ο Philip Kotler (1980) έπαιξε επίσης σημαντικό ρόλο στην επισημοποίηση της πρακτικής της τμηματοποίησης στη βιβλιογραφία του μάρκετινγκ, προσδιορίζοντας διάφορες προσεγγίσεις για την τμηματοποίηση των αγορών, συμπεριλαμβανομένης της δημογραφικής, γεωγραφικής, ψυχογραφικής και συμπεριφορικής τμηματοποίησης (Kotler, 1980).

2.1.2 Τύποι τμηματοποίησης πελατών

- **Δημογραφική τμηματοποίηση**

Η δημογραφική τμηματοποίηση είναι η πιο συχνά χρησιμοποιούμενη μέθοδος διαχωρισμού των αγορών με βάση χαρακτηριστικά όπως η ηλικία, το φύλο, το εισόδημα, η εκπαίδευση, το επάγγελμα και η οικογενειακή κατάσταση. Οι δημογραφικές μεταβλητές είναι εύκολο να συλλεχθούν και συχνά επηρεάζουν άμεσα τη συμπεριφορά των καταναλωτών, καθιστώντας αυτόν τον τύπο τμηματοποίησης μια πρακτική επιλογή για όσους ασχολούνται με το εμπόριο (Dolnicar, 2004).

Για παράδειγμα, οι Chandon et al. (2005) διερεύνησαν πώς δημογραφικά στοιχεία όπως το εισόδημα και το επάγγελμα επηρεάζουν την αγοραστική συμπεριφορά, αποδεικνύοντας ότι τα άτομα με υψηλότερο εισόδημα είναι πιο πιθανό να αγοράσουν αγαθά πολυτελείας. Ομοίως, οι Wedel και Kamakura (2000) τόνισαν ότι η δημογραφική τμηματοποίηση είναι θεμελιώδης σε κλάδους όπως τα καταναλωτικά συσκευασμένα προϊόντα (consumer packaged goods - CPG) και το λιανικό εμπόριο, όπου η ηλικία, το μέγεθος της οικογένειας και το εισόδημα έχουν άμεση επίδραση στην επιλογή προϊόντων και στις καταναλωτικές συνήθειες.

- **Γεωγραφική τμηματοποίηση**

Η γεωγραφική τμηματοποίηση διαχωρίζει τους πελάτες με βάση την τοποθεσία τους, όπως χώρα, περιοχή ή πόλη. Η προσέγγιση αυτή αναγνωρίζει ότι γεωγραφικοί παράγοντες όπως το κλίμα, ο πολιτισμός και οι τοπικές προτιμήσεις μπορούν να επηρεάσουν σε μεγάλο βαθμό την αγοραστική συμπεριφορά. Για παράδειγμα, οι Cavusgil et al. (2004) τόνισαν τη σημασία της γεωγραφίας στο διεθνές μάρκετινγκ, σημειώνοντας ότι οι πολυεθνικές εταιρείες συχνά προσαρμόζουν τις προσφορές τους ώστε να ανταποκρίνονται στις συγκεκριμένες απαιτήσεις των διαφόρων περιοχών ή χωρών.

Επιπλέον, οι Hofstede et al. (2010) επέκτειναν αυτή την κατανόηση συζητώντας πώς οι γεωγραφικοί παράγοντες και οι κουλτούρες των εθνών διαμορφώνουν τις προτιμήσεις των καταναλωτών και τη συμπεριφορά της αγοράς, καθιστώντας έτσι τη γεωγραφική τμηματοποίηση ζωτικής σημασίας για τις παγκόσμιες μάρκες (brands). Το σύστημα PRIZM, που αναπτύχθηκε από την Claritas τη δεκαετία του 1970, αποτελεί ένα παράδειγμα χρήσης τόσο γεωγραφικών όσο και δημογραφικών δεδομένων για τη δημιουργία προφίλ καταναλωτών (Weiss, 1988).

- **Ψυχογραφική τμηματοποίηση**

Η ψυχογραφική τμηματοποίηση περιλαμβάνει την κατηγοριοποίηση των πελατών με βάση τον τρόπο ζωής, τις αξίες, τις προσωπικότητες και τις στάσεις τους. Αυτός ο τύπος τμηματοποίησης υπερβαίνει τις πιο απτές δημογραφικές ή γεωγραφικές μεθόδους προσπαθώντας να κατανοήσει τα υποκείμενα κίνητρα και τις στάσεις που καθοδηγούν τη συμπεριφορά των καταναλωτών. Ο Demby (1974) εισήγαγε τα ψυχογραφικά στοιχεία στο μάρκετινγκ και ο Wells (1975) επέκτεινε αργότερα τη σημασία του τρόπου ζωής στην τμηματοποίηση των καταναλωτών, υποστηρίζοντας ότι τα χαρακτηριστικά της προσωπικότητας και οι αξίες του ατόμου προσφέρουν μια βαθύτερη κατανόηση των αναγκών των καταναλωτών. Για παράδειγμα, ο Plummer (1974) διερεύνησε τα ψυχογραφικά προφίλ, κατηγοριοποιώντας τους πελάτες με βάση τα χαρακτηριστικά της προσωπικότητας και τις συμπάθειές τους για συγκεκριμένα προϊόντα. Οι Kahle και Kennedy (1988) συζήτησαν επίσης τη σημασία της κατανόησης των αξιών των πελατών για την τμηματοποίηση, τονίζοντας ότι τα ψυχογραφικά στοιχεία μπορούν να αποκαλύψουν γιατί οι καταναλωτές λαμβάνουν ορισμένες αγοραστικές αποφάσεις.

Η ψυχογραφική τμηματοποίηση παραμένει ιδιαίτερα χρήσιμη σε κλάδους όπως τα είδη πολυτελείας, η υγεία και η ευεξία και η μόδα, όπου τα κίνητρα και οι αξίες των καταναλωτών

συχνά καθορίζουν την αγοραστική συμπεριφορά περισσότερο από τα δημογραφικά χαρακτηριστικά (Schiffman & Kanuk, 2007).

- **Τμηματοποίηση με βάση την συμπεριφορά των καταναλωτών**

Η συμπεριφορική τμηματοποίηση επικεντρώνεται στις αλληλεπιδράσεις των πελατών με μια εταιρεία ή ένα προϊόν, ομαδοποιώντας τους με βάση παράγοντες όπως το ιστορικό αγορών, η χρήση του προϊόντος, η αφοσίωση και οι απαντήσεις σε εκστρατείες μάρκετινγκ. Σύμφωνα με τους Wedel και Kamakura (2000), αυτός ο τύπος τμηματοποίησης παρέχει πιο εφαρμόσιμες πληροφορίες για τις επιχειρήσεις, καθώς τους επιτρέπει να προσαρμόζουν τις προσπάθειες μάρκετινγκ σε διαφορετικές ομάδες πελατών με βάση την πραγματική συμπεριφορά τους και όχι τα συμπερασματικά χαρακτηριστικά τους.

Οι Blattberg et al. (2008) τόνισαν την αξία της τμηματοποίησης συμπεριφοράς στη διαχείριση πελατειακών σχέσεων (CRM), σημειώνοντας ότι οι επιχειρήσεις μπορούν να στοχεύουν σε πελάτες υψηλής αξίας που παρουσιάζουν συχνή αγοραστική συμπεριφορά ή υψηλή χρηματική αξία. Η τμηματοποίηση συμπεριφοράς χρησιμοποιείται συνήθως στο ηλεκτρονικό εμπόριο και το ψηφιακό μάρκετινγκ, όπου τα δεδομένα των πελατών είναι άμεσα διαθέσιμα, επιτρέποντας τη στόχευση και την εξατομίκευση σε πραγματικό χρόνο (Kumar et al., 2006).

Για παράδειγμα, οι Bauer και Hammerschmidt (2005) χρησιμοποίησαν δεδομένα συμπεριφοράς για να τμηματοποιήσουν τους πελάτες ενός διαδικτυακού λιανοπωλητή, εντοπίζοντας συγκεκριμένα μοτίβα περιήγησης και αγοραστικής συμπεριφοράς που επέτρεψαν στην εταιρεία να βελτιστοποιήσει τις στρατηγικές μάρκετινγκ και προώθησης.

- **Τμηματοποίηση της επιχείρησης**

Στις αγορές B2B (business-to-business), χρησιμοποιείται συνήθως η εταιριογραφική τμηματοποίηση. Η μέθοδος αυτή διαιρεί τις εταιρείες με βάση χαρακτηριστικά όπως ο κλάδος, το μέγεθος της εταιρείας, τα έσοδα και η τοποθεσία. Σύμφωνα με τους Gordon και Chatterjee (2003), η εταιριογραφική τμηματοποίηση είναι απαραίτητη για τις επιχειρήσεις που πωλούν προϊόντα ή υπηρεσίες σε άλλες επιχειρήσεις, διότι τους επιτρέπει να στοχεύουν σε επιχειρήσεις που είναι πιθανότερο να έχουν ανάγκη για τις προσφορές τους με βάση αυτά τα χαρακτηριστικά. Για παράδειγμα, οι Shaw και Adamson (2005) μελέτησαν την εφαρμογή των εταιριογραφικών δεδομένων στον κλάδο των υπηρεσιών πληροφορικής, δείχνοντας ότι οι επιχειρήσεις με μεγαλύτερα έσοδα και βάσεις εργαζομένων είχαν σημαντικά διαφορετική αγοραστική συμπεριφορά σε σύγκριση με τις μικρότερες επιχειρήσεις.

2.2 Ανάλυση RFM

2.2.1 Προέλευση και υλοποίηση

Η ανάλυση RFM είναι μια τεχνική τμηματοποίησης που χρησιμοποιείται στο μάρκετινγκ και τη διαχείριση πελατειακών σχέσεων (Customer Relationship Management) για την κατηγοριοποίηση των πελατών με βάση την αγοραστική τους συμπεριφορά. Η μέθοδος αναπτύχθηκε κατά τις δεκαετίες του 1960 και 1970 αλλά εδραιώθηκε στον τομέα του μάρκετινγκ βάσεων δεδομένων με το έργο του Arthur Hughes (1994), ο οποίος την παρουσίασε ως ένα πρακτικό εργαλείο άμεσου μάρκετινγκ (direct marketing) για τη βελτίωση των ποσοστών ανταπόκρισης στις εκστρατείες μάρκετινγκ. Η ονομασία της προέρχεται από τα αρχικά των παραμέτρων Recency, Frequency και Monetary, οι οποίες αναφέρονται στα εξής δεδομένα:

- Recency (R): Η πιο πρόσφατη αγορά ενός πελάτη, και συγκεκριμένα το διάστημα που μεσολάβησε από την τελευταία αγορά που πραγματοποίησε μέχρι την χρονική στιγμή που ορίζεται στην ανάλυση.
- Συχνότητα (F): Η συχνότητα των αγορών που πραγματοποιεί ο καταναλωτής, και ειδικότερα ο αριθμός των συναλλαγών που πραγματοποίησε σε μια ορισμένη χρονική περίοδο.
- Νομισματική αξία (M): Το ποσό των χρημάτων που ο πελάτης διέθεσε στην επιχείρηση κατά την ορισμένη αυτή χρονική περίοδο.

Σύμφωνα με τον Hughes (1994), το πρώτο βήμα της μεθόδου RFM είναι η ταξινόμηση των εγγραφών των πελατών ανάλογα με το πόσο πρόσφατα έχουν αγοράσει από την επιχείρηση (Recency). Στη συνέχεια, η βάση δεδομένων χωρίζεται σε ίσα πεμπτημόρια (quintiles) και σε αυτά τα πεμπτημόρια αποδίδονται οι βαθμολογίες 1 έως 5: το 20% των πελατών που αγόρασαν πιο πρόσφατα από την εταιρεία λαμβάνει τη βαθμολογία 5, το επόμενο 20% λαμβάνει τον αριθμό 4 και ούτω καθεξής. Το επόμενο βήμα περιλαμβάνει την ταξινόμηση των πελατών σε κάθε πεμπτημόριο με βάση την συχνότητα των αγορών τους (Frequency), όπου τους αποδίδεται μία βαθμολογία από το 1 έως το 5, με το βαθμό 5 να προσδιορίζει τους πελάτες που αγοράζουν πιο συχνά από όλους τους υπόλοιπους. Κάθε μία από αυτές τις ομάδες (25 ομάδες) ταξινομείται ανάλογα με το πόσα χρήματα έχουν ξοδέψει οι πελάτες της εταιρίας (Monetary). Οι κατηγορίες αυτές χωρίζονται σε πεμπτημόρια και τους αποδίδονται βαθμολογίες από 1 έως 5, με το βαθμό 5 να προσδιορίζει τους πελάτες που έχουν ξοδέψει τα περισσότερα χρήματα από όλους τους υπόλοιπους. Επομένως η βάση δεδομένων χωρίζεται σε

5 x 5 x 5 = 125 ισοπληθείς ομάδες σύμφωνα με τις τρεις μετρικές. Ο Miglautsch (2000) σημειώνει το πλεονέκτημα της μεθόδου των πεμπτημορίων για την προβολή της συμπεριφοράς των πελατών εφόσον τα σχήματα τμηματοποίησης δημιουργούνται περιοδικά. Ωστόσο, υποστηρίζει ότι το κύριο μειονέκτημα της μεθόδου των πεμπτημορίων είναι η τάση της να «*ομαδοποιεί πελάτες που έχουν πολύ διαφορετική αγοραστική συμπεριφορά (στην κορυφή) και να διαχωρίζει αυθαίρετα τους πελάτες που έχουν πανομοιότυπη συμπεριφορά (στη βάση)*».

Ο Hughes (1994) τόνισε πώς η ανάλυση των δεδομένων αγοράς των πελατών μέσω των παραμέτρων RFM θα μπορούσε να βοηθήσει τις επιχειρήσεις να εντοπίσουν τους πιο πολύτιμους πελάτες τους και να σχεδιάσουν εξατομικευμένες προσφορές, οδηγώντας σε πιο ισχυρή δέσμευση και αύξηση των πωλήσεων. Η δημοτικότητα της ανάλυσης RFM πηγάζει από την απλότητα και την αποτελεσματικότητα της εφαρμογής της, καθώς είναι μία μέθοδος εύκολα κατανοητή από τις διοικήσεις και τα επιχειρησιακά στελέχη που λαμβάνουν τις αποφάσεις (Marcus, 1998). Επιπλέον, χρησιμοποιείται ευρέως επειδή συνδέει άμεσα τη συμπεριφορά των πελατών με την κερδοφορία, επιτρέποντας στις επιχειρήσεις να βελτιστοποιήσουν τις στρατηγικές μάρκετινγκ τους εστιάζοντας στους πελάτες που είναι πιο πιθανό να ανταποκριθούν σε μελλοντικές προωθητικές ενέργειες (Fader, 2009).

Οι Dursun και Caber (2016) αναφέρουν τα εξής ως κύρια πλεονεκτήματα της RFM ανάλυσης:

- Είναι ένα ισχυρό εργαλείο για την αξιολόγηση της αξίας διάρκειας ζωής των πελατών (customer life value – CLV), το οποίο είναι επίσης ικανό να συνδυαστεί με τεχνικές εξόρυξης συχνών προτύπων/μοτίβων (Hu & Yeh, 2014).
- Θεωρείται ως «*βάση για μια συνεχή ροή τεχνικών για τη βελτίωση της τμηματοποίησης των πελατών*» (Elsner, Krafft, & Huchzemeier, 2003).
- Είναι ένα αποτελεσματικό εργαλείο για την πρόβλεψη της ανταπόκρισης των πελατών σε εκστρατείες μάρκετινγκ και την ενίσχυση των κερδών της εταιρείας σε σύντομο χρονικό διάστημα (Baecke & Van den Poel, 2011).

Από την άλλη μεριά, οι συγγραφείς συγκέντρωσαν τα παρακάτω μειονεκτήματα της ανάλυσης RFM:

- Η μέθοδος, χρησιμοποιώντας μόνο τις τρεις μετρικές R, F και M, αγνοεί ορισμένα άλλα σημαντικά χαρακτηριστικά των πελατών όπως η ηλικία, το εισόδημα, ο τρόπος ζωής και οι παραλλαγές των προϊόντων, γεγονός που την καθιστά ανεπαρκή για τη δημιουργία επιτυχημένων προγραμμάτων μάρκετινγκ (Fitzpatrick, 2001).

- Οι τρεις μετρικές τείνουν να συσχετίζονται σε μεγάλο βαθμό, με τις υψηλότερες συσχετίσεις να παρατηρούνται μεταξύ των μετρικών Frequency και Monetary (Olson et al., 2009).
- Η ανάλυση δεν λαμβάνει υπόψιν της τους δυνητικούς και τους μη κερδοφόρους πελάτες μιας επιχείρησης όπως και την διαφορετική σημασία που έχουν οι μετρικές R, F και M από κλάδο σε κλάδο (Băcilă, Rădulescu, & Marar, 2012).

2.2.2 Παραλλαγές της ανάλυσης RFM

Η ανάλυση RFM έχει εξελιχθεί σημαντικά από την αρχική της διάδοση, με διάφορους ερευνητές και επαγγελματίες να εισάγουν τροποποιήσεις και επεκτάσεις ώστε να ενισχύσουν την αποτελεσματικότητά της σε διαφορετικά επιχειρηματικά πλαίσια. Κάποιες από τις εν λόγω παραλλαγές ενσωματώνουν πρόσθετες διαστάσεις/μεταβλητές συμπεριφοράς, ενώ κάποιες αποσκοπούν στη βελτίωση της ακρίβειας τμηματοποίησης ή στην προσαρμογή της ανάλυσης RFM σε συγκεκριμένους κλάδους. Ακολουθούν ορισμένες σημαντικές παραλλαγές της ανάλυσης RFM που έχουν δημοσιευτεί με την πάροδο των ετών.

Ένα από τα πιο δημοφιλή μοντέλα τμηματοποίησης είναι το LRFM μοντέλο. Οι Reinartz και Kumar (2000) ισχυρίστηκαν ότι το μοντέλο RFM δεν μπορεί να τμηματοποιήσει τους πελάτες που έχουν είτε μακροχρόνια είτε βραχυχρόνια σχέση με την εταιρεία. Υπό αυτές τις συνθήκες, οι Chang και Tsay (2004) εισήγαγαν το μοντέλο LRFM προσθέτοντας την μεταβλητή *Length* (L) στο μοντέλο RFM η οποία αναφέρεται στη διάρκεια της σχέσης του πελάτη με την επιχείρηση, αφού υπολογίζεται ως το χρονικό διάστημα, σε ημέρες, μεταξύ της πρώτης και της τελευταίας επίσκεψης του πελάτη. Όσο μεγαλύτερη είναι η τιμή της μεταβλητής L ενός πελάτη, τόσο μεγαλύτερη αφοσίωση δείχνει ο πελάτης στην επιχείρηση ή στο προϊόν όταν οι άλλες τρεις μετρικές παραμένουν σταθερές. Οι Kao et al. (2011) χρησιμοποίησαν το LRFM μοντέλο σε συνδυασμό με τον αλγόριθμο K-means και τμηματοποίησαν τους πελάτες μιας επιχείρησης ανδρικών ρούχων σε 12 ομάδες.

Οι Peker, Kocyigit και Eren (2017) εισήγαγαν μία νέα προσέγγιση τμηματοποίησης πελατών, το μοντέλο LRFMP, μια επέκταση του παραδοσιακού μοντέλου LRFM. Το μοντέλο LRFMP προσθέτει τη διάσταση *Periodicity* (P), ενισχύοντας τη δυνατότητα αποτελεσματικότερης ταξινόμησης των πελατών με βάση την αγοραστική τους συμπεριφορά. Η μετρική P αντικατοπτρίζει το αν οι πελάτες επισκέπτονται τακτικά τα καταστήματα και ορίζεται ως η τυπική απόκλιση των χρόνων μεταξύ των επισκέψεων των πελατών. Εάν ένας πελάτης έχει

χαμηλή τιμή περιοδικότητας, σημαίνει ότι ο πελάτης αυτός επισκέπτεται ή πραγματοποιεί αγορές σε σχετικά σταθερά διαστήματα και μπορεί να χαρακτηριστεί ως τακτικός. Επιπλέον, η μετρική Recency τροποποιήθηκε για την συγκεκριμένη ανάλυση, καθώς για τον υπολογισμό της λήφθηκαν υπόψη οι τελευταίες N συναλλαγές του πελάτη αντί να λαμβάνεται υπόψη μόνο η πιο πρόσφατη συναλλαγή του. Συγκεκριμένα, η μεταβλητή R ορίστηκε ως η μέση τιμή του αριθμού των ημερών μεταξύ των ημερομηνιών των N πρόσφατων επισκέψεων του πελάτη και της τελευταίας ημερομηνίας της περιόδου παρατήρησης. Στη μεθοδολογία οι συγγραφείς χρησιμοποίησαν το μοντέλο LRFMP σε συνδυασμό με τον αλγόριθμο μηχανικής μάθησης K-means (για περισσότερα βλ. υποενότητα 2.3.3) για την τμηματοποίηση των πελατών μιας αλυσίδας καταστημάτων που δραστηριοποιείται στον κλάδο λιανικής πώλησης ειδών παντοπωλείου στην Τουρκία. Οι ερευνητές τόνισαν την αναγκαιότητα για αποτελεσματική τμηματοποίηση των πελατών των επιχειρήσεων του συγκεκριμένου κλάδου, καθώς ο ανταγωνισμός είναι ιδιαίτερα έντονος, κάτι που παράγει επιπλέον πίεση στους επιχειρηματίες οι οποίοι πασχίζουν να διαχειριστούν αποδοτικά την πελατειακή τους βάση και να αποκτήσουν ανταγωνιστικό πλεονέκτημα μέσα σε αυτό το κορεσμένο περιβάλλον. Σε αυτή την κατεύθυνση το μοντέλο LRFMP μπορεί να τους παράσχει χρήσιμες πληροφορίες για τα διαφορετικά προφίλ πελατών, να βοηθήσει τους υπεύθυνους λήψης αποφάσεων να αναπτύξουν αποτελεσματικές σχέσεις με τους πελάτες και να κατανείμουν με αποδοτικό τρόπο τους διαθέσιμους πόρους των διαφόρων στρατηγικών μάρκετινγκ.

Στην έρευνά τους, οι Liu και Shih (2005) ανέπτυξαν μία νέα μεθοδολογία σύστασης προϊόντων (product recommendation system) δίνοντας ιδιαίτερη έμφαση στην αξία διάρκειας ζωής του πελάτη (CLV) η οποία προσδιορίζεται από τη σταθμισμένη RFM (Weighted RFM). Η προτεινόμενη μεθοδολογία χρησιμοποιεί κυρίως τις τεχνικές της Μεθόδου Αναλυτικής Ιεράρχησης (Analytic Hierarchy Process), συσταδοποίησης (clustering) και εξόρυξης κανόνων συσχέτισης (association rule mining techniques). Αρχικά οι τιμές Recency, Frequency και Monetary χρησιμοποιήθηκαν για την τμηματοποίηση των πελατών σε ομάδες με παρόμοιες τιμές. Η μέθοδος AHP χρησιμοποιήθηκε για την απόδοση βάρους (weight) σε κάθε μεταβλητή R, F και M. Στη συνέχεια, με βάση τη σταθμισμένη RFM τιμή, χρησιμοποιήθηκε η ομαδοποίηση K-means για την ταξινόμηση των καταναλωτών με συγκρίσιμες αξίες διάρκειας ζωής ή επίπεδα αφοσίωσης. Έπειτα εφαρμόστηκε μια προσέγγιση εξόρυξης κανόνων συσχέτισης για την εξαγωγή κανόνων σύστασης, δηλαδή συχνών μοτίβων αγοράς από κάθε ομάδα πελατών. Τα εξαγόμενα συχνά μοτίβα αγοράς αντιπροσωπεύουν την κοινή αγοραστική συμπεριφορά των πελατών με παρόμοιες αγορές προϊόντων. Συνεπώς η συγκεκριμένη

μεθοδολογία συνιστά προϊόντα στους πελάτες με βάση τα συχνά μοτίβα αγοράς πελατών με παρόμοιες αγορές προϊόντων.

Οι Lang et al. (2022) παρατηρούν ότι στην εποχή των «μεγάλων» δεδομένων, η εφαρμογή στατικών μοντέλων βαρών αποδεικνύεται λιγότερο αποτελεσματική, οδηγώντας σε μη βέλτιστη τμηματοποίηση και στόχευση πελατών. Οι συγγραφείς προτείνουν μια δυναμική προσέγγιση στάθμισης (dynamic weighted RFM approach) για κάθε μία από τις μετρικές R, F και M. Αρχικά υπολογίζεται το *υποκειμενικό* βάρος (subjective weight) με τη χρήση της Μεθόδου Αναλυτικής Ιεράρχησης (Analytic Hierarchy Process - AHP), ενώ στη συνέχεια εφαρμόζεται η μέθοδος Εντροπίας (Entropy) για τον υπολογισμό του *αντικειμενικού* βάρους (objective weight). Το τελικό βάρος της κάθε μετρικής προκύπτει από την Integrated Weighting Method. Οι συγγραφείς αρχικά εφάρμοσαν την εν λόγω μέθοδο χρησιμοποιώντας δεδομένα πωλήσεων CD ενός ηλεκτρονικού καταστήματος. Στη συνέχεια χρησιμοποίησαν ένα πολύ ευρύτερο σύνολο δεδομένων που αφορά το ηλεκτρονικό εμπόριο (e-commerce) της Βραζιλίας, θέλοντας να επικυρώσουν την επεκτασιμότητα και την αποδοτικότητα του προτεινόμενου μοντέλου τους σε ένα μοντέρνο και ποικιλόμορφο περιβάλλον ηλεκτρονικού εμπορίου. Οι δύο αναλύσεις έδειξαν ότι η δυναμική απόδοση βαρών στο μοντέλο RFM επηρεάζει θετικά την απόδοση τμηματοποίησης πελατών.

Η έρευνα των Yeh et al. (2009) επέκτεινε το μοντέλο RFM σε μοντέλο RFMTC με τη συμπερίληψη των εξής δύο παραμέτρων: της *Time since first purchase* (T) και της *Churn probability* (C). Η μετρική *T* αναφέρεται στον χρόνο που μεσολαβεί από την πρώτη αγορά που πραγματοποίησε ο πελάτης από την επιχείρηση, ενώ η μετρική *C* αναφέρεται στην πιθανότητα αποχώρησής του από το πελατολόγιο της επιχείρησης. Βασιζόμενοι στην ακολουθία Bernoulli της θεωρίας πιθανοτήτων, οι συγγραφείς εξήγαγαν μαθηματικούς τύπους που μπορούν να προβλέψουν την πιθανότητα ότι ένας πελάτης θα αγοράσει την επόμενη φορά από την επιχείρηση, καθώς και τον αναμενόμενο συνολικό αριθμό των φορών που ο πελάτης θα αγοράσει στο μέλλον από την επιχείρηση. Υπογραμμίζεται ότι το μοντέλο RFMTC αποτελεί ένα ακριβές μοντέλο ποσοτικής πρόβλεψης της συμπεριφοράς των πελατών το οποίο, σε αντίθεση με το παραδοσιακό μοντέλο RFM, δεν χρησιμοποιεί πρόχειρες μεθόδους κατηγοριοποίησής τους, όπως ο χωρισμός τους σε πέντε ισόποσα τμήματα. Ένα ακόμη πλεονέκτημα της συγκεκριμένης μεθόδου είναι ότι δεν προσαρμόζει κάθε φορά τα βάρη των παραμέτρων R, F, M ανάλογα με τις διάφορες βιομηχανίες, αλλά μπορεί να οικοδομήσει αυτόματα το βέλτιστο προγνωστικό μοντέλο με βάση τα δεδομένα των βάσεων δεδομένων μάρκετινγκ διαφορετικών κλάδων. Αναφέρεται επίσης ότι το μοντέλο RFMTC δεν χρειάζεται

να τμηματοποιήσει τους πελάτες σε διαφορετικές ομάδες ώστε να επιβεβαιώσει το ποσοστό ανταπόκρισης κάθε ομάδας στις ενέργειες μάρκετινγκ της επιχείρησης αλλά χρησιμοποιεί μία μόνο ομάδα πελατών, μειώνοντας σημαντικά την αναγκαία ποσότητα δοκιμών σε πελάτες.

Η έρευνα των Heldt et al. (2021) αποσκοπεί στη βελτίωση της πρόβλεψης της αξίας των πελατών εισάγοντας μία νέα επέκταση του παραδοσιακού μοντέλου RFM, το μοντέλο *RFM per Product* (RFM/P). Η μεταβλητή P (product) ενσωματώνει στην ανάλυση δεδομένα για τα συγκεκριμένα προϊόντα που αγοράζουν οι πελάτες, επιτρέποντας έτσι στις επιχειρήσεις να διαφοροποιούν την αξία των πελατών ανά προϊόν, αντί να συγκεντρώνουν όλες τις αγορές μαζί. Το προτεινόμενο μοντέλο παρέχει τον προσδιορισμό των προϊόντων που αφορούν τους πιο πολύτιμους πελάτες και των πελατών που αγοράζουν τα πιο κερδοφόρα προϊόντα. Η εστίαση αποκλειστικά στην κερδοφορία των προϊόντων μπορεί να οδηγήσει την εταιρεία στη διαδικασία που είναι γνωστή ως «σπείρα θανάτου». Από την άλλη πλευρά, η εστίαση μόνο στην κερδοφορία των πελατών μπορεί να οδηγήσει σε αυξημένο συνολικό κίνδυνο της επιχείρησης, ενθαρρύνοντας ενδεχομένως την υπερβολική συγκέντρωση των προσπαθειών μάρκετινγκ σε μια μικρή ομάδα πελατών. Το μοντέλο RFM/P συνδυάζει τις δύο προοπτικές – προϊόντων και πελατών – δίνοντας τη δυνατότητα στους διαχειριστές να εντοπίσουν ευκαιρίες για βελτιώσεις προϊόντων και υπηρεσιών ώστε να ταιριάζουν καλύτερα οι προσφορές της εταιρείας στους πελάτες-κλειδιά, να δρομολογήσουν επεκτάσεις εμπορικών σημάτων για πολύτιμες υφιστάμενες κατηγορίες προϊόντων για την απόκτηση νέων πελατών και να επιτρέψουν στρατηγικές μάρκετινγκ που έχουν θετικό αναμενόμενο αντίκτυπο στην αξία των πελατών (CLV).

2.2.3 Τμήματα RFM: Χαρακτηριστικά & προτεινόμενες στρατηγικές μάρκετινγκ

Η γνώση που αντλούν τα στελέχη των επιχειρήσεων από την τμηματοποίηση των πελατών με βάση την συμπεριφορά τους βοηθά στην ανάπτυξη προσαρμοσμένων στρατηγικών μάρκετινγκ που στοχεύουν σε συγκεκριμένες ομάδες πελατών (Ernawati et al., 2021). Ο πίνακας της Εικόνας 1 απεικονίζει έναν ευρέως αποδεκτό τρόπο δημιουργίας τμημάτων πελατών με βάση το RFM score τους (Cuce & Tiryaki, 2022), ενώ στη συνέχεια παρουσιάζονται οι κατάλληλες στρατηγικές μάρκετινγκ για την προσέγγιση κάθε τμήματος.

Segment	Description	Recency Score	Frequency Score	Monetary Score
Champions	Bought recently, buy often and spend the most.	**4 - 5**	**4 - 5**	**4 - 5**
Loyal Customers	Spend good money. Responsive to promotions.	**2 - 4**	**3 - 4**	**4 - 5**
Potential Loyalists	Recent customers, spent good amount, bought more than once	**3 - 5**	**1 - 3**	**1 - 3**
New Customers	Bought more recently but not often	**4 - 5**	**< 2**	**< 2**
Promising	Recent shoppers but haven't spent much	**3 - 4**	**< 2**	**< 2**
Need Attention	Above average recency, frequency & monetary values	**3 - 4**	**3 - 4**	**3 - 4**
About to Sleep	Below average recency, frequency & monetary values	**2 - 3**	**< 3**	**< 3**
At Risk	Spent big money, purchased often but long time ago	**< 3**	**2 - 5**	**2 - 5**
Can't Lose Them	Made big purchases and often but long time ago	**< 2**	**4 - 5**	**4 - 5**
Hibernating	Low spenders, low frequency and purchased long time ago	**2 - 3**	**2 - 3**	**2 - 3**
Lost	Lowest recency, frequency & monetary values	**< 2**	**< 2**	**< 2**

Εικόνα 1. Τμήματα πελατών με βάση το RFM score (Πηγή: Cuce & Tiryaki, 2022)

- **Champions**

Πρόκειται για τους πελάτες που έχουν πολύ πρόσφατα πραγματοποιήσει αγορές από την επιχείρηση (υψηλό R), έχουν αγοράσει πολλές φορές (υψηλό F) καταναλώνοντας μεγάλα ποσά χρηματικών μονάδων (υψηλό M). Στόχος της επιχείρησης είναι να διατηρήσει την ικανοποίησή τους σε υψηλά επίπεδα καθώς οι πελάτες αυτοί είναι υπεύθυνοι για ένα μεγάλο μερίδιο των εσόδων της αλλά και για την διαφήμιση των προϊόντων της μέσω θετικών σχολίων (positive word of mouth). Είναι επίσης οι πελάτες που είναι πιο πιθανό να αγοράσουν πρώτοι τα νέα προϊόντα που θα λανσάρει η επιχείρηση θέτοντας τη βάση για την καθιέρωσή τους στην αγορά. Η επιχείρηση θα πρέπει να εστιάσει σε ενέργειες μάρκετινγκ με στόχο την επιβράβευση αυτής της ομάδας πελατών με ενέργειες όπως η παροχή αποκλειστικών προσφορών, η προτεραιότητα στη διάθεση των προϊόντων και στην εξυπηρέτηση, και οι εξατομικευμένες προσπάθειες προσέγγισής τους. Τέλος, σε περιπτώσεις που υλοποιείται έρευνα για τον εντοπισμό πιθανών μειονεκτημάτων/αστοχιών των παρεχόμενων προϊόντων/υπηρεσιών αλλά

και νέων τάσεων στις προτιμήσεις των καταναλωτών, η ομάδα ‘champions’ θα παρέχουν την πιο αξιόπιστη ανατροφοδότηση.

- **Loyal Customers**

Οι πελάτες αυτής της κατηγορίας αγόρασαν πολύ πρόσφατα από την επιχείρηση (υψηλό R), παρουσιάζουν συνήθως υψηλή συχνότητα αγορών (υψηλό F) για τις οποίες διαθέτουν σημαντικά χρηματικά ποσά (υψηλό M). Είναι πολύτιμοι και ιδιαίτερα αφοσιωμένοι πελάτες καθώς ανταποκρίνονται θετικά στις καμπάνιες μάρκετινγκ. Λαμβάνουν ενημερωτικά δελτία και ενημερώνονται για τα νέα της επιχείρησης. Η επιχείρηση θα πρέπει να τους κρατήσει χρησιμοποιώντας στρατηγικές upselling και cross-selling για να «κατασκευάσει» νέες ανάγκες που θα τους κεντρίσουν το ενδιαφέρον καταλήγοντας σε περισσότερες ή πιο δαπανηρές αγορές. Θα πρέπει να εφαρμόζονται στην πράξη εξατομικευμένες ενέργειες μάρκετινγκ, προσωποποιημένη εξυπηρέτηση καθώς και κάθε άλλη δραστηριότητα που μπορεί να ενισχύσει την αφοσίωσή τους στην επιχείρηση.

- **Potential Loyalists**

Οι πρόσφατες αγορές (υψηλό R), οι συχνές αγορές (υψηλό F) και οι δαπάνες μέτριου ύψους (μέτριο M) αποτελούν τυπικά χαρακτηριστικά συμπεριφοράς αυτής της κατηγορίας καταναλωτών. Αν και οι πελάτες αυτοί έχουν πραγματοποιήσει στο παρελθόν αρκετές αγορές από την επιχείρηση, το μέγεθος του καλαθιού τους δεν ήταν πολύ μεγάλο. Οι ενέργειες μάρκετινγκ θα πρέπει να επικεντρωθούν στην ενθάρρυνση των συγκεκριμένων πελατών να αυξήσουν τις αγορές τους μέσω προτάσεων για συμπληρωματικά προϊόντα ή επιπλέον προϊόντα αλλά και μέσω ενεργειών που θα τους κάνουν να νιώσουν πολύτιμοι και να επομένως να αυξήσουν την αφοσίωσή τους στην επιχείρηση.

- **Promising**

Οι πελάτες αυτής της κατηγορίας πραγματοποίησαν πρόσφατα κάποια αγορά (υψηλό R), αγοράζουν με μέτρια συχνότητα από την επιχείρηση (μέτριο F) και έχουν διαθέσει ένα αρκετά μεγάλο χρηματικό ποσό (υψηλό M). Παρόλο που αγόρασαν πολύ πρόσφατα από την επιχείρηση, δεν είναι οι πελάτες που πραγματοποιούν αγορές σε τακτά χρονικά διαστήματα. Για να τους προσελκύσει η επιχείρηση θα πρέπει να εστιάσει σε στρατηγικές μάρκετινγκ που θα τους δελεάσουν να αυξήσουν την συχνότητα των αγορών τους, όπως είναι η προσφορά προγραμμάτων συνδρομής, η παροχή προτάσεων για προϊόντα, η αποστολή μικρών δώρων με

κάθε αγορά καθώς και η ενθάρρυνσή τους να κάνουν αξιολόγηση των προϊόντων/υπηρεσιών της επιχείρησης.

- **New Customers**

Οι πελάτες αυτής της ομάδας έχουν πρόσφατα αγοράσει από την επιχείρηση (υψηλό R), δεν αγοράζουν όμως συχνά (χαμηλό F) και έχουν δαπανήσει ποσά μετρίου ή χαμηλού ύψους (μέτρια M). Για κάποιους από αυτούς ίσως είναι η πρώτη φορά που αγοράζουν από την επιχείρηση. Η στρατηγική μάρκετινγκ που συνήθως ακολουθείται στοχεύει κυρίως στη διερεύνηση των απαιτήσεων και των προτιμήσεών των συγκεκριμένων πελατών όπως και στην προσφορά κινήτρων για αγορά. Η επιχείρηση μπορεί να τους παρέχει ένα εκπτωτικό κουπόνι για μελλοντικές αγορές ή να τους προσφέρει ένα δώρο με την πρώτη αγορά. Επιπλέον μια έρευνα ικανοποίησης μέσω τηλεφωνικού μάρκετινγκ ή μιας ιστοσελίδας όπου θα αξιολογήσουν τα προϊόντα, μπορεί να τους κινητοποιήσει σημαντικά. Ο στόχος των προσπαθειών μάρκετινγκ είναι να ξεχωρίσει η επιχείρηση από τον ανταγωνισμό στα μάτια του καταναλωτή και συνεπώς να στραφεί ολοκληρωτικά προς αυτή.

- **Need Attention**

Οι πελάτες αυτής της κατηγορίας αγόρασαν πρόσφατα ή σχετικά πρόσφατα από την επιχείρηση (μέτριο R), αγοράζουν με μέτρια συχνότητα (μέτριο F) και διαθέτουν υψηλά ή μέτρια χρηματικά ποσά (μέτριο M). Για να μετατρέψει αυτούς τους πελάτες σε τακτικούς πελάτες, η επιχείρηση είναι κρίσιμο να αναδείξει τις ιδιαίτερες ιδιότητες των αγαθών/υπηρεσιών που προσφέρει και να παρέχει κίνητρα για περισσότερες αγορές. Οι προσφορές περιορισμένης χρονικής διάρκειας, οι ενέργειες εξατομικευμένου μάρκετινγκ, το επιθετικό μάρκετινγκ, οι εξατομικευμένες επικοινωνίες και τα χαμηλού κόστους προγράμματα αυτοματοποιημένης προώθησης αποδεικνύονται συνήθως αποτελεσματικές ενέργειες για την προσέλκυση αυτής της ομάδας πελατών.

- **About to Sleep**

Οι πελάτες αυτοί έχουν λιγότερες πρόσφατες αγορές (μέτριο R), μέτρια ή χαμηλή συχνότητα αγορών (μέτριο F) και χαμηλές ή μέτριες δαπάνες (χαμηλή M). Οι πελάτες της συγκεκριμένης κατηγορίας δεν έχουν αγοράσει από την επιχείρηση για μεγάλο χρονικό διάστημα, αλλά όχι σε βαθμό που να είναι απρόσιτοι. Μπορούν να γίνουν ξανά ενεργοί με την χρήση των κατάλληλων τακτικών όπως η συνεχής πληροφόρηση. Κάποια κίνητρα για να τους προσελκύσει ξανά θα μπορούσαν να περιλαμβάνουν εκπτώσεις, προσφορές δώρων 1+1, μάρκετινγκ μέσω

ηλεκτρονικού ταχυδρομείου, μηνύματα υπενθύμισης για προϊόντα με έκπτωση ή υπενθυμίσεις ότι έχουν καιρό να αγοράσουν, προϊόντα με έκπτωση παρόμοια με τις προηγούμενες προτιμήσεις τους.

- **Can't Lose Them**

Η ομάδα αυτή περιλαμβάνει πελάτες που έχουν καιρό να πραγματοποιήσουν κάποια αγορά (χαμηλό R), έχουν μέτρια ή υψηλή συχνότητα αγορών (μέτριο F) αλλά έχουν επενδύσει σημαντικά χρηματικά ποσά στην επιχείρηση (υψηλό M). Παρά το γεγονός ότι δεν πραγματοποιούν συχνές ή πρόσφατες αγορές, αυτοί οι καταναλωτές είναι πολύτιμοι για την επιχείρηση, επειδή τείνουν να πραγματοποιούν μεγάλες αγορές. Οι εξατομικευμένες τηλεφωνικές κλήσεις, οι εκστρατείες επιστροφής κέρδους, τα προγράμματα επιβράβευσης και η απαλλαγή από πρόσθετες χρεώσεις (όπως τα έξοδα αποστολής ή οι δωρεάν επιστροφές προϊόντων) αποτελούν στρατηγικές μάρκετινγκ που συνιστώνται για να τους κινητοποιήσουν. Μια εκστρατεία με προσφορές (δώρα ή εκπτώσεις) σε αγαθά που έχουν ήδη αγοράσει ή αναζητήσει θα ήταν μια κατάλληλη τακτική εκ νέου προσέλκυσής τους. Θα ήταν επίσης χρήσιμο για την επιχείρηση να λάβει τη συμβολή τους προκειμένου να κατανοήσει καλύτερα τις απαιτήσεις και τις προτιμήσεις τους. Θα πρέπει να διατεθούν πόροι στο τμήμα αυτό, δεδομένου ότι είναι ζωτικής σημασίας για την κερδοφορία του οργανισμού.

- **At risk**

Οι καταναλωτές αυτοί συνήθως δεν ξοδεύουν αρκετά χρήματα (μέτριο M), δεν έχουν κάνει αρκετές αγορές πρόσφατα (χαμηλό R) και έχουν μέτρια συχνότητα (μέτριο F). Παρόλο που ένα σημαντικό ποσοστό των συναλλαγών που πραγματοποιούνται από αυτούς τους καταναλωτές δεν ολοκληρώνεται, είναι πιθανό να απαντήσουν σε προσπάθειες προσαρμογής της αγοραστικής τους εμπειρίας ή σε μηνύματα ηλεκτρονικού ταχυδρομείου που τούς υπενθυμίζουν τα είδη που δεν κατάφεραν να προσθέσουν στο καλάθι τους. Ανεξάρτητα από αυτό, πρέπει να διερευνηθεί το σκεπτικό που κρύβεται πίσω από τις ανολοκλήρωτες αγορές τους.

- **Hibernating**

Η συμπεριφορά αυτής της ομάδας πελατών χαρακτηρίζεται συνήθως από χαμηλή συχνότητα (χαμηλό F), χαμηλές δαπάνες (χαμηλό M) και μη πρόσφατες αγορές (μέτριο R). Δεδομένου ότι αυτή η ομάδα πελατών έχει ήδη επιδείξει απροθυμία να

συμμετάσχει σε πρωτοβουλίες προσέγγισης, δεν συνιστάται η διάθεση ειδικών πόρων για αυτούς.

- **Lost**

Πρόκειται για τους καταναλωτές με τις χαμηλότερες δαπάνες (χαμηλό M), τις παλαιότερες αγορές (χαμηλό R) και τη χαμηλότερη συχνότητα αγορών (χαμηλό F). Η προσέγγιση νέων πελατών είναι συνήθως πιο επικερδής για μια επιχείρηση από το να δαπανά χρόνο και πόρους προσπαθώντας να διατηρήσει τους υπάρχοντες πελάτες. Τεχνικές όπως η αποστολή εκπαιδευτικών μηνυμάτων ηλεκτρονικού ταχυδρομείου, η προώθηση κατά τη διάρκεια των διακοπών ή των εκπτώσεων και η προσφορά εκπτωτικών κουπονιών με μέτρια ποσά αγοράς θα μπορούσαν να χρησιμεύσουν ως κάποιο είδος κινήτρου για μια νέα αγορά.

2.3 Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένας κλάδος της τεχνητής νοημοσύνης που επικεντρώνεται στη δημιουργία στατιστικών μοντέλων και αλγορίθμων που επιτρέπουν στους υπολογιστές να μαθαίνουν από δεδομένα και να κάνουν κρίσεις ή προβλέψεις χωρίς ρητό προγραμματισμό (Ανδρουτσόπουλος, 2019). Το θέμα αυτό έχει προσελκύσει το ενδιαφέρον πολλών ερευνητών τα τελευταία χρόνια λόγω της ικανότητάς του να εξάγει σημαντικές γνώσεις από μεγάλα σύνολα δεδομένων, καθιστώντας το ιδιαίτερα αποτελεσματικό στην επιχειρησιακή ανάλυση και τη διαχείριση πελατειακών σχέσεων (Brynjolfsson & McAfee, 2017). Οι Alzubi et al. (2018) αναφέρουν ότι η μηχανική μάθηση εφαρμόζεται σε πληθώρα προβλημάτων του πραγματικού κόσμου τα οποία παρουσιάζουν υψηλή πολυπλοκότητα. Για παράδειγμα, μέσω αυτής σχεδιάζονται και προγραμματίζονται αλγόριθμοι υψηλής απόδοσης για το φιλτράρισμα ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου (spam), την ανίχνευση απάτης σε κοινωνικά δίκτυα (fraud detection), τις διαδικτυακές χρηματιστηριακές συναλλαγές, την ανίχνευση προσώπων και σχημάτων, την ιατρική διάγνωση, την πρόβλεψη της κυκλοφορίας στους δρόμους της πόλης, την αναγνώριση χαρακτήρων και τα συστήματα σύστασης προϊόντων (product recommendation). Τα αυτοκινούμενα αυτοκίνητα της Google, το Netflix που παρουσιάζει τις ταινίες και τις σειρές που μπορεί να αρέσουν σε ένα άτομο, οι μηχανές ηλεκτρονικών συστάσεων -όπως οι προτάσεις φίλων στο Facebook, τα «περισσότερα αντικείμενα προς εξέταση» και «πάρε κάτι για τον εαυτό σου» στο Amazon και η ανίχνευση

απάτης με πιστωτικές κάρτες- είναι όλα παραδείγματα εφαρμογής της μηχανικής μάθησης στον πραγματικό κόσμο.

Η θεμελιώδης ιδέα της μηχανικής μάθησης είναι ότι, σε αντίθεση με τον ρητό προγραμματισμό γενικών εφαρμογών, τα συστήματα είναι ικανά να μαθαίνουν από την εμπειρία και να λαμβάνουν αποφάσεις. Αυτό επιτρέπει στους αλγορίθμους μηχανικής μάθησης να βελτιώνουν την απόδοσή τους με την πάροδο του χρόνου και να προσαρμόζονται σε νέα δεδομένα (Mitchell, 1997).

2.3.1 Κατηγορίες Μηχανικής Μάθησης

Οι Mohammed et al. (2016) αναφέρουν ότι οι τέσσερις βασικές κατηγορίες αλγορίθμων μηχανικής μάθησης είναι η επιβλεπόμενη μάθηση (supervised learning), η μη επιβλεπόμενη μάθηση (unsupervised learning), η ημιεπιβλεπόμενη μάθηση (semi-supervised learning) και η ενισχυτική μάθηση (reinforcement learning). Παρακάτω παρουσιάζονται συνοπτικά τα κύρια χαρακτηριστικά της κάθε κατηγορίας:

- **Επιβλεπόμενη Μάθηση (Supervised Learning)**

Η έρευνα των Sandhya και Charanjeet (2016) αναφέρει ότι στη μάθηση με επίβλεψη παρέχεται ένα σύνολο δεδομένων που αποτελείται τόσο από χαρακτηριστικά όσο και από ετικέτες. Το έργο της επιβλεπόμενης μάθησης είναι να κατασκευάσει έναν εκτιμητή που είναι σε θέση να προβλέψει την ετικέτα ενός αντικειμένου δεδομένου του συνόλου χαρακτηριστικών του. Ο αλγόριθμος μάθησης λαμβάνει ένα σύνολο χαρακτηριστικών ως είσοδο (inputs) μαζί με τις αντίστοιχες σωστές εξόδους (outputs) και «μαθαίνει» συγκρίνοντας την πραγματική του έξοδο με τις σωστές εξόδους για να βρει σφάλματα. Στη συνέχεια τροποποιεί το μοντέλο αναλόγως. Η επιβλεπόμενη μάθηση χρησιμοποιείται συνήθως σε εφαρμογές όπου τα ιστορικά δεδομένα προβλέπουν πιθανά μελλοντικά γεγονότα. Κατά τη διαδικασία της εκπαίδευσης, ο αλγόριθμος επιβλεπόμενης μάθησης κατασκευάζει το μοντέλο πρόβλεψης. Μετά την εκπαίδευση, το προσαρμοσμένο μοντέλο θα προσπαθήσει να προβλέψει τις πιο πιθανές ετικέτες για ένα νέο σύνολο δειγμάτων X στα δεδομένα δοκιμής. Οι αλγόριθμοι επιβλεπόμενης μάθησης χρησιμοποιούνται κυρίως για την επίλυση προβλημάτων ταξινόμησης (classification) και προβλημάτων παλινδρόμησης (regression) (Sarker, 2021). Παραδείγματα αλγορίθμων αυτής της κατηγορίας αποτελούν τα δέντρα αποφάσεων (decision trees), η λογιστική παλινδρόμηση (logistic regression), και η γραμμική παλινδρόμηση (linear regression).

- **Μη επιβλεπόμενη Μάθηση (Unsupervised Learning)**

Η μη επιβλεπόμενη μάθηση λειτουργεί με δεδομένα χωρίς ετικέτες (unlabelled data) και αναζητά κρυφές ομαδοποιήσεις ή μοτίβα (Phill, 2024). Σε αυτή την κατηγορία ανήκουν οι αλγόριθμοι συσταδοποίησης (clustering algorithms), όπως είναι ο K-means, ένας από τους πιο απλούς στην εφαρμογή τους αλγορίθμους που όμως χρησιμοποιείται ευρέως από τους ερευνητές για την επίλυση ποικίλων προβλημάτων του πραγματικού κόσμου. Δουλεύοντας με δεδομένα που δεν έχουν ετικέτες, οι αλγόριθμοι μη επιβλεπόμενης μάθησης επιδιώκουν να εντοπίσουν μια φυσική ομαδοποίηση μέσα στα δεδομένα. Εξαιτίας αυτού είναι για παράδειγμα ιδανικοί για τον εντοπισμό υποκείμενων τάσεων στη συμπεριφορά των καταναλωτών χωρίς να χρειάζεται εκ των προτέρων γνώση των επιδιωκόμενων αποτελεσμάτων.

- **Ημιεπιβλεπόμενη Μάθηση (Semi-supervised Learning)**

Η μηχανική μάθηση με ημιεπίβλεψη είναι ένας συνδυασμός επιβλεπόμενων και μη επιβλεπόμενων μεθόδων μηχανικής μάθησης. Ενδέχεται να υπάρχουν περιπτώσεις όπου ορισμένες παρατηρήσεις είναι εφοδιασμένες με ετικέτες, αλλά η πλειονότητα των παρατηρήσεων δεν είναι επισημασμένες λόγω του υψηλού κόστους της επισήμανσης και της έλλειψης εξειδικευμένης ανθρώπινης γνώσης. Σε τέτοιες περιπτώσεις, οι αλγόριθμοι με ημιεπίβλεψη είναι οι καταλληλότεροι για τη δημιουργία μοντέλων (Alzubi et al., 2018). Η μάθηση με ημιεπίβλεψη μπορεί να χρησιμοποιηθεί σε προβλήματα όπως η ταξινόμηση, η παλινδρόμηση και η πρόβλεψη (Sandhya & Charanjeet, 2016).

- **Ενισχυτική Μάθηση (Reinforcement Learning)**

Η εκπαίδευση πρακτόρων λογισμικού και μηχανών ώστε να συμπεριφέρονται στο περιβάλλον τους με τρόπο που μεγιστοποιεί μία ανταμοιβή είναι γνωστή ως ενισχυτική μάθηση (Sutton & Barto, 2018). Ένας αλγόριθμος μαθαίνει πώς να συμπεριφέρεται σε ένα δεδομένο περιβάλλον μέσω της ενισχυτικής μάθησης, σύμφωνα με την οποία οι ενέργειες ανταμείβονται ή τιμωρούνται. Οι Portugal et al. (2018) αναφέρουν σαν παράδειγμα έναν αλγόριθμο μηχανικής μάθησης που παίζει παιχνίδια στον υπολογιστή εναντίον ενός αντιπάλου. Οι κινήσεις που οδηγούν σε νίκες (θετική ανατροφοδότηση) στο παιχνίδι πρέπει να μαθαίνονται και να επαναλαμβάνονται, ενώ οι κινήσεις που οδηγούν σε ήττες (αρνητική ανατροφοδότηση) πρέπει να αποφεύγονται. Σύμφωνα με τον Sarker (2021) η ενισχυτική μάθηση είναι ένα ισχυρό εργαλείο για την εκπαίδευση μοντέλων τεχνητής νοημοσύνης που μπορούν να βοηθήσουν στην αύξηση της αυτοματοποίησης ή στη βελτιστοποίηση της λειτουργικής αποδοτικότητας εξελιγμένων συστημάτων, όπως η ρομποτική, οι εργασίες αυτόνομης οδήγησης, η μεταποίηση

και η εφοδιαστική αλυσίδα, ωστόσο, δεν προτιμάται για την επίλυση βασικών ή απλών προβλημάτων.

2.3.2 Αλγόριθμοι Συσταδοποίησης

Μια βασική μέθοδος της μη επιβλεπόμενης μηχανικής μάθησης είναι η συσταδοποίηση (clustering), η οποία χωρίζει μια συλλογή σημείων σε ομάδες ανάλογα με τις ομοιότητες και τις διαφορές τους (Sinaga & Yang, 2020). Αυτή η προσέγγιση χρησιμοποιείται συχνά στην αναγνώριση προτύπων και στην ανάλυση δεδομένων, ιδίως όταν γίνεται ανάλυση RFM για την τμηματοποίηση των καταναλωτών. Με βάση τις τιμές των μετρικών Recency, Frequency και Monetary, η συσταδοποίηση μπορεί να βοηθήσει στην ανάλυση RFM στη διαίρεση των πελατών σε διακριτές ομάδες (Devarapalli et al., 2022). Επειδή οι αλγόριθμοι συσταδοποίησης περιλαμβάνουν πολλές παραμέτρους, λειτουργούν συχνά σε χώρους πολλών διαστάσεων και πρέπει να διαχειριστούν θορυβώδη, ελλιπή και δειγματοληπτικά δεδομένα, η απόδοσή τους μπορεί να διαφέρει σημαντικά για διαφορετικές εφαρμογές και τύπους δεδομένων (Rodriguez, 2019). Για τους λόγους αυτούς, στη βιβλιογραφία έχουν προταθεί αρκετές διαφορετικές προσεγγίσεις συσταδοποίησης, μερικές από τις οποίες παρουσιάζονται στην παρούσα ενότητα.

- **Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)**

Η ιεραρχική συσταδοποίηση δημιουργεί μια ιεραρχία συστάδων διαχωρίζοντας ή συνδυάζοντας προϋπάρχουσες συστάδες (Ding & He, 2002). Αυτή η μέθοδος είναι ιδιαίτερα χρήσιμη όταν ο αριθμός των συστάδων είναι ακανόνιστος ή αβέβαιος, επειδή μπορεί να αποκαλύψει εμφωλευμένες δομές συστάδων. Η δομή ενός δέντρου που δημιουργείται από την ιεραρχική συσταδοποίηση απεικονίζει τις σχέσεις μεταξύ των συστάδων. Εξαιτίας αυτού, μπορεί να είναι χρήσιμη για την παρουσίαση των συνδέσεων μεταξύ των διαφόρων τμημάτων καταναλωτών που βρέθηκαν από την ανάλυση RFM (John et al., 2023).

- **Αλγόριθμος K-Means**

Ένας από τους πιο δημοφιλείς και συχνά εφαρμοζόμενους αλγορίθμους συσταδοποίησης είναι ο αλγόριθμος K-means. Στόχος του είναι να χωρίσει n παρατηρήσεις σε k συστάδες, στις οποίες κάθε παρατήρηση ανήκει στη συστάδα που έχει τον πλησιέστερο μέσο όρο, χρησιμεύοντας ως πρωτότυπο της συστάδας. Ο μέσος όρος των παρατηρήσεων σε μια συγκεκριμένη συστάδα ορίζει το κέντρο της συστάδας (Alzubi et al., 2018). Ο εν λόγω αλγόριθμος φημίζεται για την απλότητα και την αποτελεσματικότητά του και λειτουργεί ιδιαίτερα καλά με σφαιρικές

συστάδες. Ο αλγόριθμος K-means αναθέτει τα σημεία δεδομένων στο πλησιέστερο κεντροειδές με επαναληπτική ενημέρωση των κεντροειδών (centroids). Για πολλές εφαρμογές, εξακολουθεί να είναι μια από τις πιο επιτυχημένες και αποδοτικές τεχνικές ομαδοποίησης, παρόλο που προϋποθέτει σφαιρικές συστάδες.[1] Στην επόμενη ενότητα του παρόντος κεφαλαίου παρουσιάζονται αναλυτικότερα ο τρόπος εφαρμογής, τα πλεονεκτήματα καθώς και οι περιορισμοί του αλγορίθμου K-means.

- **Συσταδοποίηση βασισμένη στην πυκνότητα**

Η εύρεση συστάδων διαφορετικών μεγεθών και μορφών μπορεί να επιτευχθεί με τη βοήθεια αλγορίθμων συσταδοποίησης με βάση την πυκνότητα, ο πιο διάσημος από τους οποίους είναι ο αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Οι αλγόριθμοι αυτής της κατηγορίας είναι ιδιαίτερα χρήσιμοι σε περιπτώσεις όπου ο αριθμός των συστάδων είναι αβέβαιος ή όταν τα δεδομένα περιλαμβάνουν μεγάλα επίπεδα θορύβου (Ester et al., 1996). Ο αλγόριθμος DBSCAN ομαδοποιεί τα σημεία ανάλογα με το πόσο κοντά και πυκνά είναι το ένα στο άλλο. Αυτό αυξάνει την ανθεκτικότητά του στο θόρυβο και τις ακραίες τιμές, γεγονός που θα μπορούσε να οδηγήσει σε ακριβέστερη κατάτμηση στην ανάλυση RFM (Monalisa & Kurnia, 2019).

2.3.3 Συσταδοποίηση K-means

Οι Jin και Han (2011) αναφέρουν ότι τα βασικά βήματα της δημιουργίας συστάδων με τη χρήση του αλγορίθμου K-means είναι τα εξής:

1. Αρχικοποίηση μέσω τυχαίας επιλογής K κεντροειδών (centroids)
2. Σε κάθε σημείο δεδομένων ανάθεση του πλησιέστερου κεντροειδούς με βάση την Ευκλείδεια απόσταση (Maheswari, 2019).
3. Καθορισμός των νέων κεντροειδών που προκύπτουν από τον υπολογισμό του μέσου όρου των σημείων που έχουν ανατεθεί στην κάθε συστάδα
4. Επανάληψη των βημάτων 2 και 3 έως ότου ικανοποιηθεί κάποιο κριτήριο σύγκλισης (διαφορά στην τιμή της συνάρτησης-παραμόρφωσης) ή διακοπής (περάτωση προκαθορισμένου αριθμού επαναλήψεων).

[1] What is k-means clustering? https://www.ibm.com/topics/k-means-clustering, τελευταία πρόσβαση 26/11/2024

Δεδομένων των αρχικών συνθηκών διασποράς των κεντροειδών, ο K-means μπορεί να προσδιορίσει τις βέλτιστες αναθέσεις συστάδων μέσω της παραπάνω επαναληπτικής διαδικασίας. Ωστόσο το τελικό αποτέλεσμα της συσταδοποίησης μπορεί να επηρεαστεί σημαντικά από την επιλογή των αρχικών κεντροειδών.

Η συσταδοποίηση με τον αλγόριθμο K-means προσφέρει τα ακόλουθα πλεονεκτήματα όταν συνδυάζεται με την ανάλυση RFM:

- Κλιμακωσιμότητα (Scalability): Οι σύγχρονες επιχειρήσεις έχουν να αντιμετωπίσουν μεγάλα σύνολα δεδομένων τα οποία ο αλγόριθμος K-means μπορεί να διαχειριστεί αποτελεσματικά. Η ανάλυση της συμπεριφοράς των πελατών σε πραγματικό χρόνο με βάση τις μετρικές RFM καθίσταται δυνατή χάρη στην υπολογιστική αποδοτικότητα του αλγορίθμου, η οποία του επιτρέπει να χειρίζεται γρήγορα τεράστιους όγκους δεδομένων (Shindler et al., 2011).
- Ερμηνευσιμότητα (Interpretability): Οι προκύπτουσες συστάδες μπορούν να επισημανθούν σύμφωνα με τις ιδιότητές τους και να ερμηνευθούν εύκολα. Μια συστάδα με υψηλή βαθμολογία Recency, για παράδειγμα, θα μπορούσε να ονομαστεί «Πρόσφατοι αγοραστές». Προκειμένου οι επιχειρήσεις να κατανοήσουν τα υποκείμενα μοτίβα στη συμπεριφορά των καταναλωτών και να δημιουργήσουν προσαρμοσμένες στρατηγικές ως απάντηση, η ερμηνευσιμότητα είναι απαραίτητη στην ανάλυση RFM.
- Ταχύτητα: Η μέθοδος K-means είναι κατάλληλη για εφαρμογές πραγματικού χρόνου και για συχνή επανασυγκέντρωση σε ομάδες όταν διατίθενται νέα δεδομένα, καθώς είναι συνήθως ταχύτερη από άλλες τεχνικές ομαδοποίησης. Λόγω της ταχύτητας του k-means, οι επιχειρήσεις μπορούν να παρακολουθούν συνεχώς τις αλλαγές στη συμπεριφορά των πελατών και να προσαρμόζουν γρήγορα τη στρατηγική τους ως απάντηση στις τάσεις της αγοράς ή των προτιμήσεων των πελατών (Fadaei & Khasteh, 2019).

Η συσταδοποίηση K-means χρησιμοποιείται ευρέως, ωστόσο έχει ορισμένα μειονεκτήματα και περιορισμούς:

- Καθορισμός του βέλτιστου αριθμού συστάδων: Η εύρεση του βέλτιστου αριθμού συστάδων (k) στις εφαρμογές K-means είναι ένα από τα πιο δύσκολα ζητούμενα. Για το σκοπό αυτό απαιτείται συχνά δοκιμή-και-σφάλμα (trial-and-error) ή η εφαρμογή άλλων μέτρων, όπως η μέθοδος του «αγκώνα» (Elbow Method) (Syakur et al., 2018) ή η βαθμολογία «σιλουέτας» (Silhouette Method) (Shutaywi & Kachouie, 2021). Παρόλο που

οι τεχνικές αυτές μπορούν να είναι χρήσιμες για τον προσδιορισμό του ιδανικού αριθμού συστάδων, πρέπει και οι ίδιες να ερμηνεύονται προσεκτικά.

- Ευαισθησία στις αρχικές συνθήκες: Το τελικό αποτέλεσμα της ομαδοποίησης επηρεάζεται έντονα από την αρχική τοποθέτηση των κεντροειδών. Ακόμη και για το ίδιο σύνολο δεδομένων, διαφορετικές αρχικοποιήσεις μπορεί να οδηγήσουν σε διαφορετικές αναθέσεις συστάδων. Εξαιτίας αυτής της ευαισθησίας στις αρχικές συνθήκες, συνιστάται η εκτέλεση του k-means πολλές φορές χρησιμοποιώντας εναλλακτικές αρχικοποιήσεις, προκειμένου να διασφαλιστεί η συνέπεια των αποτελεσμάτων σε όλες τις αναλύσεις.[2]
- Υπόθεση σφαιρικών συστάδων: Η μέθοδος K-means προϋποθέτει ότι κάθε συστάδα έχει την ίδια διακύμανση, κάτι που μπορεί να μην ισχύει σε πραγματικές καταστάσεις όπου οι συστάδες έχουν διαφορετικά μεγέθη και σχήματα. Τα τμήματα πελατών μπορεί να εμφανίζουν διαφορετικά επίπεδα ομοιογένειας στην ανάλυση RFM, γεγονός που αντιβαίνει στην υπόθεση των σφαιρικών συστάδων. Λόγω αυτού του περιορισμού, απαιτούνται εναλλακτικές στρατηγικές συσταδοποίησης ή προσαρμογές στον συμβατικό αλγόριθμο k-means.[3]

2.4 Συνδυασμός RFM και K-means

Στην παρούσα ενότητα παρουσιάζονται έρευνες που συνδυάζουν τη μέθοδο RFM με τον αλγόριθμο μηχανικής μάθησης K-means για την τμηματοποίηση των πελατών σε επιχειρήσεις που ανήκουν σε διαφορετικούς κλάδους. Ο αυξανόμενος αριθμός των ερευνών που συνδυάζουν πολλαπλές μεθόδους για την ανάπτυξη νέων μοντέλων τμηματοποίησης υποδηλώνει την κρισιμότητα της αποτελεσματικής διαχείρισης των πελατών τόσο για την βελτίωση της κερδοφορίας των επιχειρήσεων όσο και για την διατήρηση της ικανοποίησης των πελατών σε υψηλά επίπεδα.

[2] The Drawbacks of K-Means Algorithm https://www.baeldung.com/cs/k-means-flaws-improvements, τελευταία πρόσβαση 26/11/2024

[3] Demonstration of K-Means Assumptions https://www.geeksforgeeks.org/demonstration-of-k-means-assumptions/, τελευταία πρόσβαση 26/11/2024

Οι Maryani και Riana (2017) προσπάθησαν να δώσουν λύση στην πρόκληση μιας επιχείρησης να εντοπίσει δυνητικούς πελάτες ώστε να αξιοποιήσει το σύστημα διαχείρισης πελατειακών σχέσεων (CRM) για να υλοποιήσει το βέλτιστο δυνατό σχέδιο μάρκετινγκ για την προσέγγισή τους και τελικά για την αποκόμιση κέρδους για την ίδια. Οι συγγραφείς χρησιμοποίησαν το μοντέλο RFM σε δεδομένα συναλλαγών μιας εταιρίας της βιομηχανίας εξάτμισης μοτοσικλετών και αυτοκινήτων για την κατηγοριοποίηση των πελατών της. Τα στάδια της έρευνας που διεξήχθη αποτελούνται από τέσσερις δραστηριότητες. Αρχικά πραγματοποιήθηκε η ομαδοποίηση των δεδομένων συναλλαγών με τη χρήση του αλγορίθμου K-means. Το δεύτερο στάδιο ήταν ο προσδιορισμός των χαρακτηριστικών κάθε συστάδας με τη μέθοδο του δέντρου αποφάσεων (Decision Tree Method). Κατά το τρίτο στάδιο πραγματοποιήθηκε αξιολόγηση του προφίλ του πελάτη με τη χρήση της οικονομικής θεωρίας του Grid Hill, η οποία διαμορφώνει τη σύσταση 4 χαρακτηριστικών πελατών που ταιριάζουν με τη συναλλαγή εξόρυξης δεδομένων. Στο τελικό στάδιο δημιουργήθηκαν εφαρμογές χαρτογράφησης πελατών με την παραγωγή συστάσεων σε κάθε προφίλ πελάτη ώστε να μεγιστοποιηθεί η ικανοποίησή τους και έτσι η επιχείρηση να μπορεί να τους διατηρεί στο πελατολόγιό της. Τα αποτελέσματα της παρούσας μελέτης μπορούν να χρησιμοποιηθούν ως σύστημα υποστήριξης αποφάσεων στον κλάδο των μέσων ενημέρωσης για τη χαρτογράφηση πελατών και τη γνώση δυνητικών πελατών.

Η έρευνα των Gustriansyah et al. (2020) βασίστηκε στην ιδέα ότι η διαδικασία διατήρησης βάσεων δεδομένων προϊόντων για τη διαχείριση αποθεμάτων γίνεται όλο και πιο δύσκολη καθώς αυξάνεται ο όγκος των συναλλαγών. Οι συγγραφείς ισχυρίζονται ότι μια πιο αποτελεσματική στρατηγική για την επίλυση αυτού του ζητήματος θα ήταν η κατάτμηση όλων των προϊόντων στον κατάλληλο αριθμό συστάδων με βάση κάποιες από τις ομοιότητές τους με την χρήση κάποιας μεθόδου εξόρυξης δεδομένων. Οι αξίες των διαφόρων ομάδων μπορούν στη συνέχεια να υπολογιστούν και να αξιολογηθούν ώστε η διοίκηση να λάβει τεκμηριωμένες αποφάσεις και να κατανείμει τους πόρους με τον πλέον ορθολογικό τρόπο. Για την τεκμηρίωση των προαναφερθέντων οι ερευνητές χρησιμοποίησαν τον αλγόριθμο K-Means για να ομαδοποιήσουν τα δεδομένα των προϊόντων ενός φαρμακείου στην Palembang της Ινδονησίας με βάση τις τιμές RFM. Ο προσδιορισμός του βέλτιστου αριθμού k συστάδων στη μέθοδο K-Means αξιολογήθηκε με τη χρήση των εξής οκτώ δεικτών εγκυρότητας: της μεθόδου Elbow, του δείκτη Silhouette, του δείκτη Calinski-Harabasz, του δείκτη Davies-Bouldin, του δείκτη Ratkowski, του δείκτη Hubert, του δείκτη Ball-Hall και του δείκτη Krzanowski -Lai. Ο λόγος που χρησιμοποιήθηκαν όλοι οι παραπάνω δείκτες ήταν για τη βελτίωση της ακρίβειας στη

διαδικασία διαχείρισης των αποθεμάτων και την υλοποίηση μιας πιο αντικειμενικής ομαδοποίησης των προϊόντων. Για τον αριθμό k που προέκυψε από κάθε δείκτη εγκυρότητας πραγματοποιήθηκε έλεγχος για την ποιότητα των αποτελεσμάτων τμηματοποίησης. Ο εν λόγω έλεγχος είχε σαν μετρική την διακύμανση *R* η οποία αναφέρεται στην τιμή του λόγου μεταξύ της μέσης απόστασης των δεδομένων στην ίδια συστάδα (απόσταση εντός συστάδας) και της μέσης απόστασης των δεδομένων στις άλλες συστάδες (απόσταση μεταξύ συστάδων). Μια τιμή R κοντά στο 0 υποδηλώνει ότι τα δεδομένα στις ίδιες συστάδες παρουσιάζουν μεγάλη ομοιότητα μεταξύ τους. Τα αποτελέσματα της αξιολόγησης έδειξαν ότι ο βέλτιστος αριθμός συστάδων k ήταν οι τρεις συστάδες με τιμή διακύμανσης R=0,19113.

Οι Sarvari, Ustundag και Takci (2016) αξιολόγησαν την απόδοση διαφορετικών προσεγγίσεων τμηματοποίησης πελατών που ενσωματώνουν τόσο μετρήσεις RFM όσο και δημογραφικές πληροφορίες, όπως η ηλικία, το φύλο και η τοποθεσία, στην πρόβλεψη της συμπεριφοράς των πελατών. Οι συγγραφείς χρησιμοποίησαν διάφορες τεχνικές τμηματοποίησης, συμπεριλαμβανομένων της ομαδοποίησης K-means, των αυτό-οργανωτικών χαρτών (Self-Organizing Maps - SOM) και των δέντρων απόφασης (decision trees), οι οποίες έπειτα αξιολογήθηκαν με βάση την ικανότητά τους να ομαδοποιούν τους πελάτες σε τμήματα με ακρίβεια αλλά και με νόημα. Σύμφωνα με τα αποτελέσματα της ανάλυσης, διαπιστώθηκε ότι οι προσεγγίσεις τμηματοποίησης που ενσωματώνουν δημογραφικά δεδομένα με την ανάλυση RFM αποδίδουν σημαντικά καλύτερα από την χρήση του παραδοσιακού μοντέλου RFM από μόνο του. Η ενισχυμένη προσέγγιση προσφέρει στις επιχειρήσεις βαθύτερη κατανόηση των προφίλ των πελατών, οδηγώντας στη βελτίωση της διαχείρισης των πελατειακών σχέσεων, σε καλύτερη στόχευση και αποτελεσματικότερες στρατηγικές μάρκετινγκ.

Στην έρευνά τους οι Christy et al. (2018) τόνισαν τη σημασία της τμηματοποίησης των πελατών για τη βελτίωση της κερδοφορίας της επιχείρησης καθώς και για την διατήρηση του πελατολογίου της καθώς υπογραμμίζουν ότι αποτελεί πιο συμφέρουσα στρατηγική σε σχέση με την προσέγγιση νέων πελατών. Για την ανάλυσή τους σε δεδομένα συναλλαγών μιας επιχείρησης λιανεμπορίου εφάρμοσαν αρχικά το παραδοσιακό μοντέλο RFM ενώ στη συνέχεια επέκτειναν την έρευνά τους χρησιμοποιώντας τρεις αλγόριθμους συσταδοποίησης: τον αλγόριθμο K-means, τον αλγόριθμο Fuzzy C-means και τέλος, τον αλγόριθμο Repetitive Median K-means (RM K-means). Πριν προχωρήσουν στην υλοποίηση των εν λόγω αλγορίθμων και έπειτα στη σύγκριση των αποτελεσμάτων τους, χρειάστηκε να κανονικοποιήσουν τις τιμές R, F και M καθώς παρατηρήθηκε ιδιαίτερα μεγάλη λοξότητα στις κατανομές τους. Ο αλγόριθμος Fuzzy C-Means επιτρέπει σε ένα συγκεκριμένο data point να

ανήκει σε περισσότερες από μία συστάδες. Οι συγγραφείς αναφέρουν ότι το πλεονέκτημα αυτής της μεθόδου έναντι του K-Means είναι ότι εφόσον ένας πελάτης μπορεί να ανήκει σε περισσότερες από μία ομάδες, η επιχείρηση αυξάνει την πιθανότητα διατήρησης των πελατών της καθώς τους προσεγγίζει με ποικίλους τρόπους. Τέλος υλοποιείται ο αλγόριθμος RM K-means, μία προτεινόμενη από τους συγγραφείς παραλλαγή του K-means, η οποία χρησιμοποιεί τις διαμέσους (medians) των μετρικών R, F και M για την αρχικοποίηση των κεντροειδών (centroids) των συστάδων. Ο αλγόριθμος αυτός αποδεικνύεται και ο πιο αποδοτικός συγκριτικά με τους άλλους δύο, καθώς ο χρόνος εκτέλεσής του είναι μικρότερος και οι επαναλήψεις που χρειάζεται να γίνουν λιγότερες.

Οι Wu et al. (2020) χρησιμοποίησαν επίσης τον συνδυασμό της ανάλυσης RFM και του αλγορίθμου K-means για την τμηματοποίηση των πελατών μιας online επιχείρησης στο Πεκίνο της Κίνας. Οι συγγραφείς αφού πραγματοποίησαν αρχικά την προεπεξεργασία και τον καθαρισμό των δεδομένων και υπολόγισαν τις τιμές R, F και M, προχώρησαν σε κανονικοποίηση των μετρικών με τη χρήση της μεθόδου min-max. Στη συνέχεια χρησιμοποιείται η Ανάλυση Κύριων Συνιστωσών PCA (Principal Component Analysis) για την ανάθεση βαρών στο μοντέλο RFM δεδομένου ότι το σύνολο δεδομένων που χρησιμοποιήθηκε για την συγκεκριμένη ανάλυση χαρακτηρίζεται από μεγάλο όγκο εγγραφών. Η ανάλυση PCA είναι μια μέθοδος στατιστικής ανάλυσης που καταφέρνει να μετατρέψει το αρχικό σύνολο δεδομένων που περιέχει μεγάλο αριθμό μεταβλητών σε ένα σύνολο με λιγότερες μεταβλητές μέσω της τεχνικής μείωσης των διαστάσεων του συνόλου δεδομένων. Οι συγγραφείς όρισαν ότι το βάρος κάθε μετρικής ισούται με το ποσοστό συνεισφοράς της διακύμανσης της κύριας συνιστώσας. Έπειτα πραγματοποιήθηκε ομαδοποίηση των πελατών με τη χρήση του αλγορίθμου K-means σε τέσσερις ομάδες και προτάθηκαν στρατηγικές διαχείρισης της κάθε ομάδας για την απόκτηση υψηλού επιπέδου ικανοποίησης των πελατών. Οι συγγραφείς ισχυρίζονται ότι η υιοθέτηση της προτεινόμενης μεθόδου έχει ως αποτέλεσμα την αύξηση του συνολικού όγκου αγορών και του συνολικού ποσού κατανάλωσης.

2.5 Συστήματα επιχειρηματικής ευφυΐας και οπτικοποίηση δεδομένων

Τα συστήματα επιχειρηματικής ευφυΐας (Business Intelligence tools) είναι τύποι λογισμικού που χρησιμοποιούνται για τη συλλογή, την οργάνωση, την οπτικοποίηση και την ανάλυση δεδομένων που συγκεντρώνονται μέσω των επιχειρηματικών λειτουργιών για την ανάδειξη

τάσεων και μοτίβων που επιτρέπουν τη λήψη αποφάσεων βάσει δεδομένων (Tripathi et al., 2020). Οι Saabith et al. (2022) κάνοντας μια καταγραφή των πλεονεκτημάτων της χρήσης των εργαλείων BI αναφέρουν μεταξύ άλλων ότι τα εργαλεία BI:

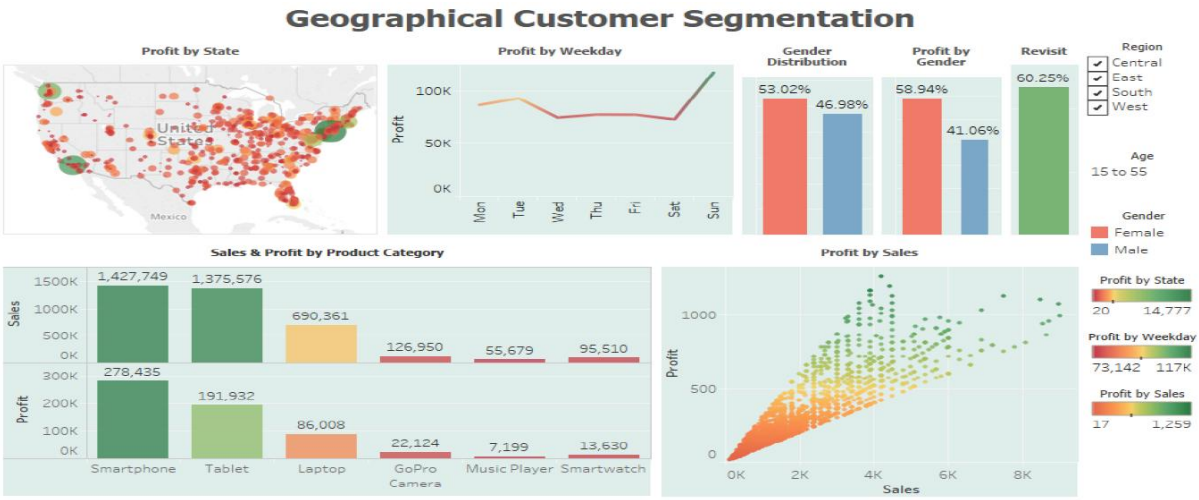
- Βελτιώνουν την ανάλυση μάρκετινγκ καθώς επιτρέπουν την αξιολόγηση των στρατηγικών ή των εκστρατειών μάρκετινγκ, συνεισφέρουν στην καλύτερη ανάλυση της συμπεριφοράς των πελατών και στον εντοπισμό νέων ευκαιριών
- Βοηθούν στην κατανόηση των αγοραστικών προτύπων των πελατών και στη βελτίωση της εμπειρίας τους, επιτρέποντας την υλοποίηση στοχευμένων ενεργειών μάρκετινγκ, όπως η παροχή εκπτώσεων ή προσφορών
- Επιτρέπουν στους μη τεχνικούς χρήστες της επιχείρησης να αποκομίσουν τις πληροφορίες που χρειάζονται λαμβάνοντας εξατομικευμένες αναφορές (tailor made reports)
- Βοηθούν στην ταχύτερη ανίχνευση προβλημάτων ή σφαλμάτων, καθώς μπορούν να παρέχουν ιστορικές, σε πραγματικό χρόνο και προγνωστικές αναφορές

Οι συγγραφείς συνέκριναν δέκα διαφορετικά εργαλεία BI με βάση τις διαφορετικές λειτουργίες που μπορούν να υποστηρίξουν. Μία από τις θεματικές που εξετάζουν αφορά την οπτικοποίηση των δεδομένων και συγκεκριμένα το αν υποστηρίζουν τις παρακάτω λύσεις-εφαρμογές:

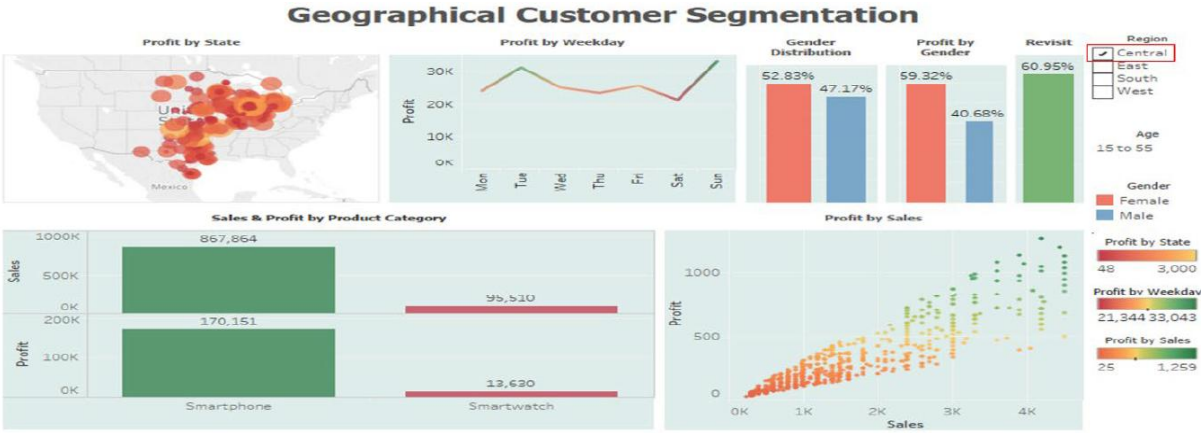
- Advanced Visualizations using Python and R: δημιουργία προηγμένων και εξελιγμένων οπτικοποιήσεων με τη χρήση βιβλιοθηκών και πακέτων προγραμματισμού Python και R
- Animations: παρουσίαση δεδομένων ως κινούμενη εικόνα, κυρίως για την παρουσίαση αλλαγών σε πολλαπλές ομάδες ή χρονικές περιόδους
- Auto-charting: καθοδήγηση των χρηστών προτείνοντας την καταλληλότερη οπτικοποίηση για τη γραφική αναπαράσταση των επιλεγμένων δεδομένων
- Auto-refresh: αυτόματη ανανέωση των διαγραμμάτων και των οπτικοποιήσεων σε ένα ταμπλό σε τακτά χρονικά διαστήματα
- Dashboard Rebranding: δυνατότητα αλλαγής των ρυθμίσεων μορφοποίησης (γραμματοσειρά, λογότυπο, χρώμα κλπ) ενός ταμπλό για την ευθυγράμμισή του με το εμπορικό σήμα του οργανισμού
- Dashboards: δημιουργία οπτικοποιήσεων που παρέχουν μια στιγμιαία προβολή σε μία οθόνη διαφόρων KPIs, μετρήσεων επιχειρηματικής ανάλυσης και κρίσιμων σημείων δεδομένων

- Embed Dashboards and Visualizations in Webpages: ενσωμάτωση ταμπλό και οπτικοποιήσεων σε άλλες ιστοσελίδες
- Interactive Data Visualizations: διαγράμματα, γραφήματα και οπτικοποιήσεις με αλληλεπιδράσεις όπως κλιμάκωση (scaling), συνδεσιμότητα (linking) και υπομνήσεις (tooltips)
- Visualizations with Drill-down and Drill-up: δυνατότητες drill-down και drill-up για την εξερεύνηση πολυδιάστατων και ιεραρχικών δεδομένων απευθείας από οπτικοποιήσεις

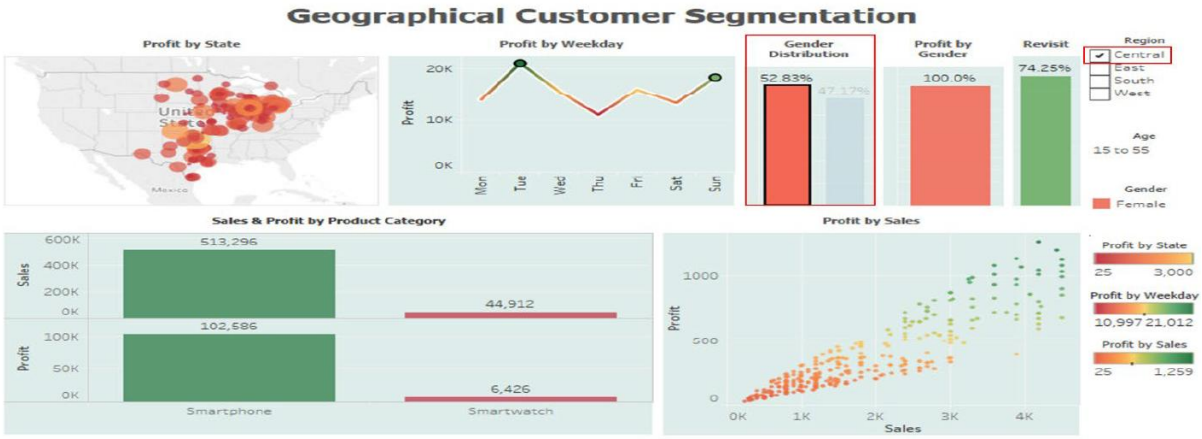
Οι δυνατότητες οπτικοποίησης και διαδραστικότητας που προσφέρουν σε όλο και ευρύτερο φάσμα τα εργαλεία BI (με κυρίαρχα το Microsoft Power BI και το Tableau) μπορούν να βοηθήσουν σημαντικά σε αναλύσεις μάρκετινγκ. Οι Bačić και Fadlalla (2016) υποστήριξαν ότι η τμηματοποίηση πελατών μέσω τεχνικών οπτικοποίησης αυξάνει την ερμηνευσιμότητα των τμημάτων, γεγονός που διευκολύνει την παρουσίαση των αποτελεσμάτων της τμηματοποίησης. Παρόλα αυτά, οι Sheng και Subramanian (2019) σημειώνουν ότι οι διαδραστικές τεχνικές οπτικοποίησης χρησιμοποιούνται ελάχιστα στην υπάρχουσα βιβλιογραφία για την εκτέλεση στοιχειώδους τμηματοποίησης πελατών. Οι συγγραφείς χρησιμοποιούν το Tableau ως ένα ισχυρό εργαλείο για την εφαρμογή της προτεινόμενης μεθόδου διαδραστικής οπτικοποίησης με βάση τη βαθμίδα κατάταξης (rank-based stepwise interactive visualization method). Η στοιχειώδης τμηματοποίηση των πελατών πραγματοποιείται με τη χρήση δημογραφικών στοιχείων για την ανακάλυψη των χαρακτηριστικών των πελατών που έλκονται από ένα συγκεκριμένο προϊόν. Ως εκ τούτου, κατασκευάζεται ένας διαδραστικός πίνακας (dashboard) με τη χρήση του Tableau που επιτρέπει την επισήμανση και το φιλτράρισμα με βάση τα χαρακτηριστικά που σχετίζονται με τα δημογραφικά στοιχεία των πελατών και παρουσιάζεται στις εικόνες 2, 3 και 4. Τα διαδραστικά ταμπλό των εργαλείων BI, τα προσαρμόσιμα φίλτρα και οι δυνατότητες κατάταξης (ranking) επιτρέπουν στον χρήστη να εξερευνεί δυναμικά τα πολυδιάστατα δεδομένα μέσω διαφόρων τύπων γραφημάτων για την εις βάθος κατανόηση των αγοραστικών μοτίβων των πελατών και της σχέσης τους με την επιχείρηση.



Εικόνα 3. Διαδραστικός πίνακας γεωγραφικής τμηματοποίησης πελατών (Πηγή: Sheng και Subramanian, 2019)



Εικόνα 2. Εστίαση στο τμήμα πελατών της κεντρικής περιοχής (Central region) (Πηγή: Sheng και Subramanian, 2019)



Εικόνα 4. Εστίαση στο τμήμα πελατών της κεντρικής περιοχής (Central region) θηλυκού γένους (Female) (Πηγή: Sheng και Subramanian, 2019)

Κεφάλαιο 3: Μεθοδολογία

3.1 Ερευνητική διαδικασία

Το παρόν κεφάλαιο περιλαμβάνει τα βήματα που υλοποιηθήκαν για την τμηματοποίηση των πελατών μίας επιχείρησης με τη χρήση της μεθόδου RFM και του αλγορίθμου συσταδοποίησης K-means. Το σύνολο δεδομένων που χρησιμοποιείται στην παρούσα εργασία αντλήθηκε από την ηλεκτρονική πλατφόρμα Kaggle και περιλαμβάνει τις συναλλαγές των πελατών ενός ηλεκτρονικού καταστήματος με είδη δώρων σε χρονικό ορίζοντα δύο ημερολογιακών ετών. Κατά την υλοποίηση των εν λόγω βημάτων χρησιμοποιήθηκαν εναλλακτικά, και ενίοτε συμπληρωματικά, το Jupyter Notebook και το εργαλείο οπτικοποίησης δεδομένων Microsoft Power BI.

Το πρώτο βήμα της ερευνητικής διαδικασίας παρουσιάζει την περιγραφική και διερευνητική ανάλυση του συνόλου δεδομένων για την εξοικείωση με τα βασικά γνωρίσματα και τις μεταβλητές του (ενότητα 3.2.1). Αρχικά υλοποιήθηκε προεπεξεργασία και καθαρισμός του συνόλου δεδομένων (ενότητα 3.2.2). Καθώς το σύνολο δεδομένων ήταν διαθέσιμο σε δύο αρχεία excel, κρίθηκε απαραίτητος ο έλεγχος για διπλές εγγραφές κατά την ενοποίηση των δύο αρχείων (ενότητα 3.2.2.1). Άλλες ενέργειες που υλοποιήθηκαν κατά το στάδιο της προεπεξεργασίας και του καθαρισμού αποτελούν ο χειρισμός ελλειπουσών τιμών, ο μετασχηματισμός τύπου δεδομένων (*πχ από numerical σε string*) και η διαγραφή εγγραφών του συνόλου δεδομένων οι οποίες δεν ικανοποιούσαν κάποια βασικά κριτήρια που τέθηκαν (ενότητες 3.2.2.2 και 3.2.2.3). Στην ενότητα 3.2.3, για λόγους πληρότητας, παρέχεται μία πλήρης λίστα των ενεργειών μετασχηματισμού του αρχικού συνόλου δεδομένων σε αυτό που χρησιμοποιήθηκε στα πλαίσια της ανάλυσης.

Με την ολοκλήρωση των προαναφερθέντων βημάτων προετοιμασίας των δεδομένων, το dataset μετασχηματίζεται με τρόπο που να επιτρέπει την εφαρμογή της ανάλυσης RFM (ενότητα 3.3). Αυτό μεταφράζεται στην παραγωγή ενός πίνακα, ο οποίος για εγγραφές έχει τους πελάτες και όχι τις συναλλαγές της επιχείρησης. Για κάθε εγγραφή λοιπόν, δηλαδή για κάθε πελάτη, αρχικά υπολογίζονται οι τιμές Recency, Frequency και Monetary από τις στήλες του μετασχηματισμένου συνόλου δεδομένων (ενότητα 3.3.1). Στην επόμενη ενότητα (3.3.2) παρουσιάζονται τρεις διακριτές μέθοδοι τμηματοποίησης πελατών με βάση τις παραπάνω τιμές. Οι δύο πρώτες μέθοδοι κατατάσσουν κάθε πελάτη σε ένα από έντεκα τμήματα

(segments) με βάση το RFM score τους και έναν πίνακα αντιστοίχισης RFM score σε τμήμα. Συγκεκριμένα, ο πρώτος τρόπος παραγωγής των RFM scores προκύπτει από την παραδοσιακή μέθοδο των «πεμπτημορίων» (ενότητα 3.3.2.1). Ως δεύτερος τρόπος παραγωγής των RFM scores προτείνεται η υλοποίηση του αλγορίθμου K-means σε κάθε μία από τις στήλες Recency, Frequency, Monetary (ενότητα 3.3.2.2). Η τρίτη μέθοδος κατάταξης πελατών σε segments (ενότητα 3.3.2.3) εφαρμόζει τη μέθοδο K-means απευθείας στα σημεία του τρισδιάστατου χώρου που έχει ως διαστάσεις του τις μετρικές Recency, Frequency και Monetary, και συνεπώς, εφόσον δεν υπολογίζει ενδιάμεσα RFM scores, δεν απαιτεί πίνακα αντιστοίχισης RFM scores σε τμήματα.

3.2 Το σύνολο δεδομένων

3.2.1 Περιγραφή συνόλου δεδομένων

Για την παρούσα ανάλυση χρησιμοποιήθηκαν δεδομένα από τις συναλλαγές πελατών ενός ηλεκτρονικού καταστήματος με είδη δώρων. Το κατάστημα εδρεύει στο Ηνωμένο Βασίλειο και εξυπηρετεί πελάτες λιανικής αλλά και χονδρικής. Το εν λόγω σύνολο δεδομένων αντλήθηκε από την ιστοσελίδα Kaggle[4], ωστόσο είναι επίσης διαθέσιμο και στην ιστοσελίδα UCI Machine Learning Repository[5], από όπου αντλούμε πληροφορίες για τις μεταβλητές του. Τα δεδομένα βρίσκονται σε μορφή excel και είναι χωρισμένα σε δύο καρτέλες (*Year 2009-2010*, *Year 2010-2011*) σύμφωνα με το έτος πραγματοποίησης των συναλλαγών. Οι συναλλαγές χαρακτηρίζονται από οκτώ μεταβλητές, οι οποίες παρουσιάζονται παρακάτω:

1. **Invoice**: Αριθμός τιμολογίου. Είναι εξαψήφιος για την κάθε συναλλαγή. Αν ξεκινάει με το γράμμα "c" υποδηλώνει ακύρωση τιμολογίου (cancellation).
2. **StockCode**: Κωδικός προϊόντος. Είναι ένας πενταψήφιος αριθμός, μοναδικός για κάθε προϊόν.
3. **Description**: Η ονομασία του προϊόντος.
4. **Quantity**: Η ποσότητα προϊόντων που αντιστοιχεί σε κάθε συναλλαγή.
5. **InvoiceDate**: Η ημερομηνία και η ώρα έκδοσης του τιμολογίου.
6. **Price**: Η τιμή μιας μονάδας προϊόντος εκφρασμένη σε λίρες (£).
7. **Customer ID**: Ο μοναδικός αριθμός καταχώρισης πελάτη. Αποτελείται από 5 ψηφία.

[4] https://www.kaggle.com/datasets/kabilan45/online-retail-ii-dataset
[5] https://archive.ics.uci.edu/dataset/502/online+retail+ii

8. **Country**: Το όνομα της χώρας στην οποία κατοικεί ο κάθε πελάτης.

Για τον καθαρισμό του συνόλου δεδομένων, καθώς και για την πραγματοποίηση της περιγραφικής και διερευνητικής ανάλυσης των δεδομένων (EDA), χρησιμοποιήθηκε το Jupyter Notebook και το εργαλείο οπτικοποίησης δεδομένων Microsoft Power BI.

Αρχικά ενοποιήθηκαν τα δύο excel sheets με την χρήση των εντολών `pd.read_excel()` και `pd.concat()` της Python σε ένα σύνολο δεδομένων, το οποίο ονομάστηκε `df_raw`, όπου με `pd` συμβολίζεται η βιβλιοθήκη `pandas` της Python. Η Εικόνα 5 παρουσιάζει τον κώδικα που χρησιμοποιήθηκε.

```
# Read excel's two sheets and merge them into a single Dataframe
df1_raw = pd.read_excel("online_retail_II.xlsx", sheet_name = "Year 2009-2010")
df2_raw = pd.read_excel("online_retail_II.xlsx", sheet_name = "Year 2010-2011")
df_raw = pd.concat([df1_raw, df2_raw])
```

Εικόνα 5. Ενοποίηση των δύο υποσυνόλων δεδομένων στο "df_raw"

Στην Εικόνα 6 παρατίθενται πληροφορίες για το σύνολο των εγγραφών (1 067 371) και των μεταβλητών (8) του ενοποιημένου dataset, όπως προκύπτει από την εντολή `df_raw.info()` της Python.

```
df_raw.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1067371 entries, 0 to 541909
Data columns (total 8 columns):
 #   Column       Non-Null Count    Dtype
---  ------       --------------    -----
 0   Invoice      1067371 non-null  object
 1   StockCode    1067371 non-null  object
 2   Description  1062989 non-null  object
 3   Quantity     1067371 non-null  int64
 4   InvoiceDate  1067371 non-null  datetime64[ns]
 5   Price        1067371 non-null  float64
 6   Customer ID  824364 non-null   float64
 7   Country      1067371 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 73.3+ MB
```

Εικόνα 6. Πληροφορίες για τις μεταβλητές του συνόλου δεδομένων

Με την εντολή `df_raw.head(10)` βλέπουμε ενδεικτικά τις δέκα πρώτες συναλλαγές του dataset όπως απεικονίζονται στην Εικόνα 7.

```
# Review first 10 rows of dataset
df_raw.head(10)
```

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom
5	489434	22064	PINK DOUGHNUT TRINKET POT	24	2009-12-01 07:45:00	1.65	13085.0	United Kingdom
6	489434	21871	SAVE THE PLANET MUG	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom
7	489434	21523	FANCY FONT HOME SWEET HOME DOORMAT	10	2009-12-01 07:45:00	5.95	13085.0	United Kingdom
8	489435	22350	CAT BOWL	12	2009-12-01 07:46:00	2.55	13085.0	United Kingdom
9	489435	22349	DOG BOWL , CHASING BALL DESIGN	12	2009-12-01 07:46:00	3.75	13085.0	United Kingdom

Εικόνα 7. Απεικόνιση των πρώτων δέκα συναλλαγών του συνόλου δεδομένων

3.2.2 Διερευνητική ανάλυση δεδομένων και καθαρισμός δεδομένων

3.2.2.1 Αφαίρεση διπλών καταχωρίσεων

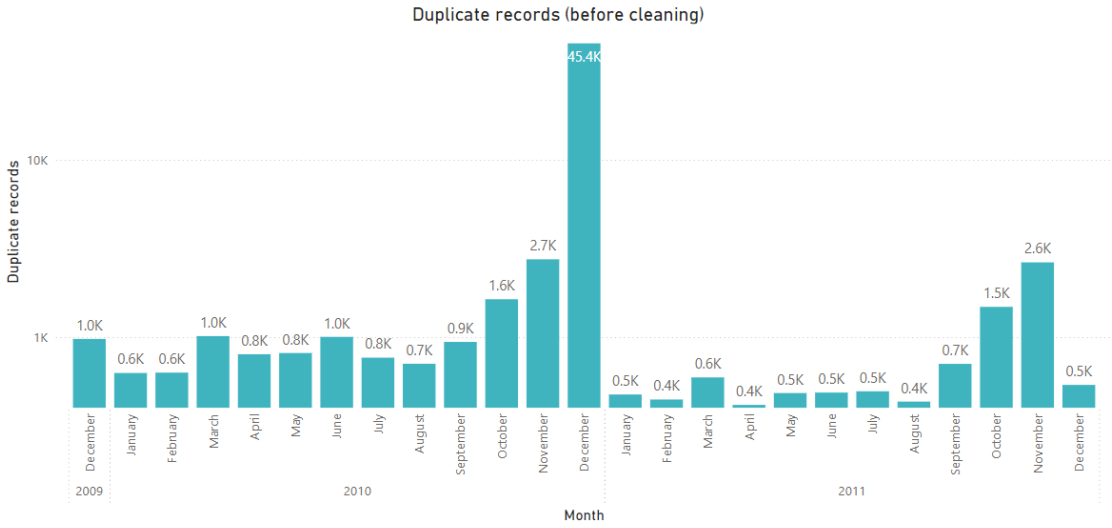
Το πρώτο βήμα της διερευνητικής ανάλυσης που πραγματοποιήθηκε ήταν ο έλεγχος για την ύπαρξη διπλών καταχωρίσεων. Όπως φαίνεται στην Εικόνα 8, με τη χρήση της εντολής `df_raw.sort_values(by="InvoiceDate").duplicated().sum()` παρατηρούμε ότι ο αριθμός των διπλών καταχωρίσεων είναι 34 335.

```
# Check for duplicates
print("Number of duplicate rows: ", df_raw.sort_values(by = "InvoiceDate").duplicated().sum())
```

```
Number of duplicate rows:  34335
```

Εικόνα 8. Αριθμός διπλών καταχωρίσεων που προέκυψε από την ενοποίηση των δύο υποσυνόλων

Στην Εικόνα 9 απεικονίζεται η κατανομή των διπλών καταχωρίσεων όπου διαπιστώνεται ότι το μεγαλύτερο μέρος αυτών παρατηρείται κατά το μήνα Δεκέμβριο 2010. Το γεγονός αυτό πιθανότατα προκλήθηκε κατά την ενοποίηση των δύο excel sheets, τα οποία περιείχαν κοινές συναλλαγές που αφορούσαν στον συγκεκριμένο μήνα.



Εικόνα 9. Κατανομή διπλών καταχωρίσεων στους μήνες του συνόλου δεδομένων (πριν την αφαίρεση μέρους τους)

Για να εξακριβώσουμε την ισχύ της παραπάνω υπόθεσης, υπολογίζουμε την τελευταία ημερομηνία του `df1_raw` (Year 2009-2010) και την πρώτη ημερομηνία του `df2_raw` (Year 2010-2011). Όπως φαίνεται στην Εικόνα 10, υπάρχουν εγγραφές που αφορούν το χρονικό διάστημα 01/12/2010 έως 09/12/2010 (9 ημέρες), οι οποίες συναντώνται και στους δύο πίνακες.

```
# We review the final date of the *df1_raw* and the first date of the *df2_raw*.

df1_raw_final_date = df1_raw["InvoiceDate"].max()
df2_raw_start_date = df2_raw["InvoiceDate"].min()

print("Final date of the 'Year 2009-2010' dataset:", df1_raw_final_date)
print("Start date of the 'Year 2010-2011' dataset:", df2_raw_start_date)
```

```
Final date of the 'Year 2009-2010' dataset: 2010-12-09 20:01:00
Start date of the 'Year 2010-2011' dataset: 2010-12-01 08:26:00
```

Εικόνα 10. Διερεύνηση των δύο υποσυνόλων δεδομένων για εντοπισμό overlap στις ημερομηνίες

Ταυτόχρονα μπορούμε να διαπιστώσουμε το παραπάνω γεγονός και οπτικά παρατηρώντας την Εικόνα 11, εκτυπώνοντας από τον πίνακα `df1_raw` τις συναλλαγές που πραγματοποιήθηκαν κατά το συγκεκριμένο χρονικό διάστημα, και από τον πίνακα `df2_raw` τις πρώτες εγγραφές του.

```
# First few rows from 12.2010 from the sheet 'Year 2009-2010'
df1_raw[df1_raw["InvoiceDate"] >= dt.datetime(2010,12,1)].sort_values(by = "InvoiceDate").head()
```

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
502938	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
502939	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
502940	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
502941	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
502942	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

```
# First rows from the sheet 'Year 2010-2011'
df2_raw.head()
```

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	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

Εικόνα 11. Απεικόνιση των πρώτων πέντε συναλλαγών που πραγματοποιήθηκαν κατά το διάστημα 01/12/2010 - 09/12/2010 στα δύο υποσύνολα δεδομένων

Μετά τις παραπάνω παρατηρήσεις, αφαιρούμε τις διπλές καταχωρίσεις μέσω της εντολής `pd.concat()`, η οποία έχει ως αποτέλεσμα τη δημιουργία του πίνακα `df`, ο οποίος περιέχει 1 044 848 συναλλαγές, όπως φαίνεται στην Εικόνα 12.

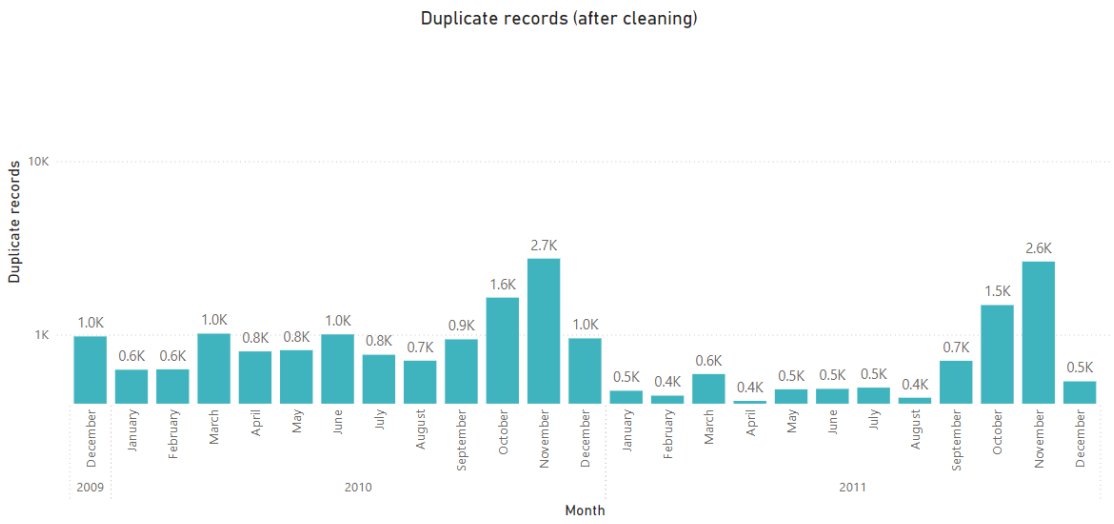
```
# Remove overlapped records, create df
df = pd.concat([df1_raw[df1_raw["InvoiceDate"] < df2_raw_start_date], df2_raw])
```

```
df.shape
```

```
(1044848, 8)
```

Εικόνα 12. Αφαίρεση των διπλών καταχωρίσεων που παρατηρήθηκαν τις πρώτες 9 ημέρες του Δεκεμβρίου 2010

Σε αυτό το σημείο ο αριθμός των διπλών καταχωρίσεων είναι 11 812, δηλαδή περίπου 66% λιγότερες σε σχέση με εκείνες του αρχικού dataset. Στην Εικόνα 13 απεικονίζονται οι διπλές καταχωρίσεις μετά τον καθαρισμό που προαναφέρθηκε.



Εικόνα 13. Κατανομή διπλών καταχωρίσεων στους μήνες του συνόλου δεδομένων (μετά την αφαίρεση μέρους τους)

Σε πραγματικές συνθήκες θα ζητούσαμε από τους υπεύθυνους συλλογής των δεδομένων του dataset να μας εξηγήσουν τί σημαίνουν οι διπλές αυτές καταχωρίσεις. Συγκεκριμένα, δεδομένου ότι η ποσότητα του κάθε προϊόντος σε κάθε μία από αυτές τις διπλές καταχωρίσεις είναι 1, μπορούμε να σχηματίσουμε την υπόθεση ότι για αυτές τις συναλλαγές το σύστημα καταχωρίσεων της επιχείρησης αντί να αθροίσει τις ποσότητες του κάθε προϊόντος για κάθε παραγγελία, τις καταχώρισε ξεχωριστά σε ισόποσες εγγραφές. Εφόσον δεν υπάρχει κάποιο ανασταλτικό επιχείρημα που να καταρρίπτει αυτή την υπόθεση, θα συμπεριλάβουμε το υπόλοιπο 34% των διπλών καταχωρίσεων στην ανάλυση που ακολουθεί.

3.2.2.2 Ελλείπουσες τιμές

Χρησιμοποιώντας την εντολή `df.isnull().sum()` της Python παρατηρούμε τον αριθμό των ελλειπουσών τιμών για κάθε μεταβλητή. Οι μόνες μεταβλητές που έχουν missing values είναι η μεταβλητή Description (4 275) και η μεταβλητή Customer ID (235 287). Για να αποκτήσουμε εικόνα του σχετικού μεγέθους των ελλειπουσών τιμών της κάθε μεταβλητής χρησιμοποιούμε την εντολή `missing/len(df) * 100`, όπου `missing` είναι το αποτέλεσμα της προηγούμενης εντολής που εκτελέσαμε. Οι εν λόγω εντολές παρουσιάζονται στην Εικόνα 14 που ακολουθεί.

```
# Identify missing values
missing = df.isnull().sum()
print(missing)
```

```
Invoice             0
StockCode           0
Description      4275
Quantity            0
InvoiceDate         0
Price               0
Customer ID    235287
Country             0
dtype: int64
```

```
# Percentage of missing values in each column
missing/len(df) * 100
```

```
Invoice         0.000000
StockCode       0.000000
Description     0.409150
Quantity        0.000000
InvoiceDate     0.000000
Price           0.000000
Customer ID    22.518778
Country         0.000000
dtype: float64
```

Εικόνα 14. Προσδιορισμός ελλειπουσών τιμών ανά μεταβλητή

Προτού ξεκινήσουμε την διαμόρφωση του συνόλου δεδομένων για τις ανάγκες της παρούσας ανάλυσης θα πρέπει να διερευνηθούν οι μεταβλητές του και να αντληθούν χρήσιμες πληροφορίες. Στην επόμενη ενότητα ακολουθεί η διερευνητική ανάλυση του συνόλου δεδομένων.

3.2.2.3 Ανάλυση των μεταβλητών του συνόλου δεδομένων

- **Στήλη Invoice**

Όπως παρατηρούμε στην Εικόνα 15, ο μοναδικός αριθμός των τιμολογίων είναι 53 628.

```
# Unique number of invoices
len(df["Invoice"].unique())
```

```
53628
```

Εικόνα 15. Μοναδικός αριθμός τιμολογίων

Μέσω της περιγραφής που έχει δώσει ο ίδιος ο εκδότης του dataset πληροφορούμαστε ότι ορισμένα από τα τιμολόγια αφορούν ακυρώσεις, και συγκεκριμένα όσα ξεκινούν με το γράμμα “C”. Θα επικεντρωθούμε αρχικά στις ακυρώσεις, και έπειτα στο πώς αυτές συνδέονται με τις

μεταβλητές Quantity και Price. Για τον σκοπό αυτό δημιουργήσαμε ένα DataFrame[6] με τις συναλλαγές που αφορούν ακυρώσεις τιμολογίων όπως φαίνεται στην Εικόνα 16. Ο αριθμός των ακυρώσεων είναι 19 165 και αντιστοιχεί στο 1,8% του συνολικού αριθμού των συναλλαγών του dataset.

```
# Convert column Invoice's data type to string
df["Invoice"] = df["Invoice"].astype(str)

# Create a DataFrame with the records of cancelled invoices
df_cancelled_inv = df[df["Invoice"].str.contains("C")]
```

```
# Review dimensions of cancelled invoices DataFrame
df_cancelled_inv.shape
```

```
(19165, 8)
```

Εικόνα 16. Συναλλαγές που αφορούν ακυρώσεις τιμολογίων

Από τις ακυρώσεις τιμολογίων μόνο μία συναλλαγή παρουσιάζει θετική τιμή στη στήλη Quantity, ενώ οι υπόλοιπες έχουν αρνητική τιμή. Η εν λόγω συναλλαγή έχει ελλείπουσα τιμή στη στήλη Customer ID, όπως φαίνεται και στην Εικόνα 17.

```
# Invoices that start with "C" and Quantity > 0
df_cancelled_inv[df_cancelled_inv["Quantity"] > 0]
```

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
76799	C496350	M	Manual	1	2010-02-01 08:24:00	373.57	NaN	United Kingdom

Εικόνα 17. Συναλλαγές με ακυρωτικά τιμολόγια και θετική τιμή στη στήλη Quantity

Εκτός από τα τιμολόγια που ξεκινούν από "C" υπάρχουν και κάποια που ξεκινούν από "A". Τα τιμολόγια αυτά είναι μόνο 6 και έχουν missing value στη στήλη Customer ID. Τα 5 από αυτά έχουν αρνητική τιμή στην στήλη Price και μοναδιαία τιμή στην στήλη Quantity. Παρατηρώντας το περιεχόμενο της στήλης Description για κάθε ένα από αυτά τα 6 τιμολόγια ("Adjust bad debt") ερμηνεύουμε πως πρόκειται για ενέργειες λογιστικής φύσεως και συνεπώς ότι δεν σχετίζονται με τις πωλήσεις του ηλεκτρονικού καταστήματος. Επιπλέον παρατηρούμε

[6] Η δομή δεδομένων DataFrame είναι η βασική δομή δεδομένων της βιβλιοθήκης pandas. Η βιβλιοθήκη pandas είναι το βασικό μέσο επεξεργασίας δεδομένων σε μορφή πινάκων της Python.

ότι υπάρχει μία συσχέτιση μεταξύ των τιμολογίων “A563185”, “A563186”, και “A563187”. Στην Εικόνα 18 παρατίθενται λεπτομέρειες για τα εν λόγω 6 τιμολόγια.

```
# Invoices that start with "A"
df[df["Invoice"].str.contains("A")]
```

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
179403	A506401	B	Adjust bad debt	1	2010-04-29 13:36:00	-53594.36	NaN	United Kingdom
276274	A516228	B	Adjust bad debt	1	2010-07-19 11:24:00	-44031.79	NaN	United Kingdom
403472	A528059	B	Adjust bad debt	1	2010-10-20 12:04:00	-38925.87	NaN	United Kingdom
299982	A563185	B	Adjust bad debt	1	2011-08-12 14:50:00	11062.06	NaN	United Kingdom
299983	A563186	B	Adjust bad debt	1	2011-08-12 14:51:00	-11062.06	NaN	United Kingdom
299984	A563187	B	Adjust bad debt	1	2011-08-12 14:52:00	-11062.06	NaN	United Kingdom

```
df[df["Invoice"].str.contains("A")].shape
```

```
(6, 8)
```

Εικόνα 18. Απεικόνιση εγγραφών με τιμολόγια που ξεκινούν από "A"

Από τα 19 165 ακυρωμένα τιμολόγια τα 719 (3%) έχουν missing value στην στήλη Customer ID (Εικόνα 19).

```
# Invoice starts with "C" and Customer ID is blank
df_cancelled_inv["Customer ID"].isnull().sum()
```

```
719
```

Εικόνα 19. Ακυρωμένα τιμολόγια με ελλείπουσα τιμή στη στήλη CustomerID

- **Στήλη Quantity**

Η στήλη Quantity αναφέρεται είτε στην ποσότητα των προϊόντων που αγοράστηκαν σε κάθε συναλλαγή (Quantity > 0) είτε, όπως είδαμε και στην ανάλυση της στήλης Invoice, σε επιστροφές προϊόντων (Quantity < 0). Παρατηρούμε ότι δεν υπάρχουν ελλείπουσες τιμές στην στήλη Quantity και συνεπώς δεν χρειάζεται να προβούμε σε κάποια διαγραφή για την ανάλυση που θα ακολουθήσει.

Οι εγγραφές με αρνητική τιμή στην στήλη Quantity είναι συνολικά 22 557 (Εικόνα 20). Ο αριθμός αυτός ξεπερνάει τις εγγραφές των τιμολογίων που ξεκινούν από “C” και έχουν αρνητική τιμή (19 164). Αυτό υποδηλώνει ότι η αρνητική τιμή στην ποσότητα δεν σημαίνει απαραίτητα μόνο την επιστροφή κάποιου προϊόντος, αλλά ενδέχεται να είναι συνδεδεμένη και με άλλα είδη συναλλαγών, όπως ακυρώσεις πωλήσεων ή ελαττωματικά προϊόντα.

```
# Number of records where Quantity < 0
len(df[(df["Quantity"] < 0)])
```

```
22557
```

Εικόνα 20. Αριθμός εγγραφών με αρνητική τιμή στη στήλη Quantity

- **Στήλη Customer ID**

Η στήλη Customer ID αναφέρεται στον μοναδικό αριθμό πελάτη (αναγνωριστικό πελάτη). Όπως φαίνεται στην Εικόνα 21, η στήλη αυτή περιέχει 235 287 missing values, δηλαδή περίπου το 22,5% των εγγραφών. Το ποσοστό είναι σημαντικά μεγάλο για να αγνοηθεί, και ανάλογα με την ανάλυση που θα πραγματοποιηθεί στα δεδομένα απαιτείται και διαφορετικός χειρισμός των missing values.

```
# Number of records where Customer ID is blank
len(df[(df["Customer ID"].isnull())])
```

```
235287
```

Εικόνα 21. Αριθμός εγγραφών με ελλείπουσα τιμή στη στήλη CustomerID

Σε πραγματικές συνθήκες θα ζητούσαμε από τους υπεύθυνους συλλογής των δεδομένων του dataset να μας εξηγήσουν γιατί οι συγκεκριμένες εγγραφές δεν έχουν Customer ID. Αφού κάτι τέτοιο δεν είναι εφικτό, μπορούμε να υποθέσουμε ότι οι συγκεκριμένοι πελάτες δεν πραγματοποίησαν ποτέ εγγραφή στο online κατάστημα και επομένως δεν τους αποδόθηκε αναγνωριστικό πελάτη.

Ο αριθμός των μοναδικών Customer ID του dataset που εξετάζουμε είναι 5 943, στον οποίο αριθμό φτάνουμε με χρήση της εντολής της Python που παρουσιάζεται στην Εικόνα 22.

```
# Number of unique customers
unique_customers = len(df["Customer ID"].unique())
print("There are {} unique customers in the dataset.".format(unique_customers))
```

```
There are 5943 unique customers in the dataset.
```

Εικόνα 22. Αριθμός μοναδικών τιμών στη στήλη CustomerID

Ο αριθμός αυτός αντιπροσωπεύει 5 942 εγγεγραμμένους πελάτες συν όλους εκείνους οι οποίοι δεν πραγματοποίησαν ποτέ εγγραφή και των οποίων το Customer ID συμβολίζεται με τον κενό χαρακτήρα.

Οι ελλείπουσες τιμές της στήλης Customer ID θα αφαιρεθούν στην επόμενη ενότητα για τους σκοπούς της ανάλυσης RFM με στόχο την τμηματοποίηση των πελατών.

- **Στήλη Country**

Οι συναλλαγές του dataset προέρχονται από 42 χώρες. Ενώ δεν παρατηρούνται missing values στην στήλη Country, υπάρχουν 756 εγγραφές στις οποίες η χώρα είναι “Unspecified”. Η λίστα με τις χώρες των συναλλαγών παρουσιάζεται στην Εικόνα 23.

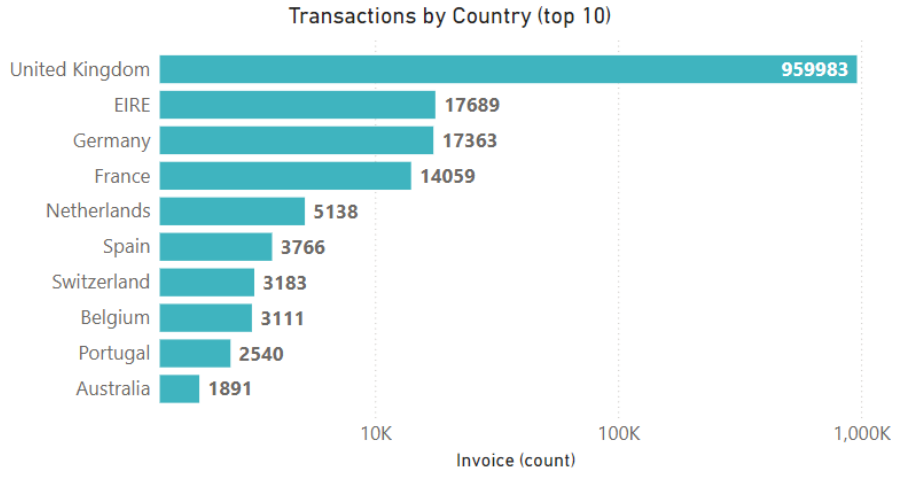
```
# List of countries
countries_list = df["Country"].unique()
number_of_countries = len(df["Country"].unique())
print("The number of countries in the dataset is {} and they are listed below:\n\n".format(number_of_countries), countries_list)
```

```
The number of countries in the dataset is 43 and they are listed below:

 ['United Kingdom' 'France' 'USA' 'Belgium' 'Australia' 'EIRE' 'Germany'
 'Portugal' 'Japan' 'Denmark' 'Nigeria' 'Netherlands' 'Poland' 'Spain'
 'Channel Islands' 'Italy' 'Cyprus' 'Greece' 'Norway' 'Austria' 'Sweden'
 'United Arab Emirates' 'Finland' 'Switzerland' 'Unspecified' 'Malta'
 'Bahrain' 'RSA' 'Bermuda' 'Hong Kong' 'Singapore' 'Thailand' 'Israel'
 'Lithuania' 'West Indies' 'Lebanon' 'Korea' 'Brazil' 'Canada' 'Iceland'
 'Saudi Arabia' 'Czech Republic' 'European Community']
```

Εικόνα 23. Λίστα με τις χώρες των συναλλαγών του dataset

Το μεγαλύτερο μερίδιο των συναλλαγών πραγματοποιήθηκε στο Ηνωμένο Βασίλειο (91.87%), ύστερα στην Ιρλανδία (1.69%) και στη Γερμανία (1.66%). Στο παρακάτω γράφημα παρουσιάζονται οι 10 χώρες με τον μεγαλύτερο αριθμό συναλλαγών (Εικόνα 24).



Εικόνα 24. Χώρες με τον μεγαλύτερο αριθμό συναλλαγών (κορυφαίες 10)

- **Στήλη StockCode**

Η στήλη StockCode αναφέρεται στους κωδικούς προϊόντων που αγοράστηκαν. Η στήλη StockCode δεν περιέχει ελλείπουσες τιμές.

Κάτι το οποίο θα είχε αξία να ερευνηθεί είναι ποιοι κωδικοί έχουν τις περισσότερες πωλήσεις κατά το χρονικό διάστημα που εξετάζεται. Προς αυτόν τον σκοπό εκτελέστηκε η εντολή `df.unique()` της Python, σύμφωνα με την οποία οι μοναδικές τιμές της στήλης StockCode είναι 5 305 (Εικόνα 25).

```
# Number of unique StockCode items
number_of_StockCodes = pd.unique(df["StockCode"])
print("The number of unique StockCodes is", len(number_of_StockCodes))
```

```
The number of unique StockCodes is 5305
```

Εικόνα 25. Αριθμός μοναδικών τιμών στη στήλη StockCode

Παράλληλα εκτελέσαμε την αντίστοιχη εντολή στο Power BI χρησιμοποιώντας την έκφραση `DISTINCTCOUNTNOBLANKS` της γλώσσας DAX, όπου διαπιστώσαμε ότι ο μοναδικός αριθμός προϊόντων διαφέρει (5 131 σε σχέση με 5 305). Είναι κρίσιμο να μάθουμε γιατί είναι διαφορετικά τα αποτελέσματα που προκύπτουν από τα δύο εργαλεία, και να φέρουμε τις τιμές σε συμφωνία, γιατί αλλιώς οι επερχόμενες αναλύσεις μέσω Python και Power BI θα αποκλίνουν και θα αποφέρουν πιθανώς διαφορετικά συμπεράσματα και συνεπώς αμφισημία.

Δεδομένου ότι η Python αναφέρει περισσότερους μοναδικούς κωδικούς είναι λογικό να υποθέσουμε ότι η Python διάβασε κάποιες τιμές της στήλης ως μοναδικές, κάτι που δεν έκανε το Power BI, διαφορετικά δεν θα υπήρχε αυτή η ασυμφωνία. Με σκοπό να ερευνήσουμε το αν είναι σωστή αυτή η υποψία πρέπει αρχικά να φέρουμε τις τιμές της στήλης σε ένα κοινό επίπεδο ώστε να είναι συγκρίσιμες μεταξύ τους. Συγκεκριμένα, μέσω Python μετατρέπουμε το data type όλων των τιμών της στήλης StockCode σε τύπο string και έπειτα όλους τους χαρακτήρες των strings σε κεφαλαία γράμματα: με αυτό τον τρόπο όλες οι τιμές μπορούν να συγκριθούν μεταξύ τους. Σε αυτό το σημείο μετράμε εκ νέου τις μοναδικές τιμές της στήλης και το αποτέλεσμα που προκύπτει είναι 5 132 (Εικόνα 26), δηλαδή μία παραπάνω από αυτές του Power BI.

```
len(df['StockCode'].astype(str).str.upper().sort_values().unique())
```

```
5132
```

Εικόνα 26. Αριθμός μοναδικών StockCode μετά την αλλαγή του data type σε string και την μετατροπή των χαρακτήρων σε κεφαλαία γράμματα

Για να εντοπίσουμε τη μία τιμή που εμφανίζεται ως διαφορά, παραθέσαμε τη λίστα των μοναδικών τιμών που προέκυψαν από την Python και από το PowerBI σε δύο διαφορετικά αρχεία. Στη συνέχεια ταξινομήσαμε αλφαβητικά τις τιμές που περιέχονται στα δύο αρχεία και αντιπαραβάλλοντάς τες εντοπίσαμε ότι η εγγραφή "47503J" εμφανιζόταν με δύο τρόπους στα αποτελέσματα της Python: "47503J " και "47503J" (στην πρώτη περίπτωση μετά τον χαρακτήρα J υπάρχει ο κενός χαρακτήρας). Στη συνέχεια χρησιμοποιήσαμε την εντολή `str.strip()` για να αφαιρέσουμε το κενό που δημιουργήθηκε στην ονομασία και στο τέλος επιβεβαιώσαμε ότι οι μοναδικές τιμές της στήλης StockCode είναι 5 131 και για τα δύο εργαλεία (Εικόνα 27).

```
(df['StockCode'].astype(str).str.upper().sort_values().unique() == '47503J').sum()
```

```
1
```

```
(df['StockCode'].astype(str).str.upper().sort_values().unique() == '47503J ').sum()
```

```
1
```

```
unique_stockcodes = df['StockCode'].astype(str).str.upper().str.strip().sort_values().unique()
len(unique_stockcodes)
```

```
5131
```

Εικόνα 27. Εντοπισμός κενού χαρακτήρα σε μία εκ των τιμών '47503J' της στήλης StockCode με αποτέλεσμα την διπλή καταμέτρησή της και έπειτα, αφαίρεση του κενού χαρακτήρα προς διόρθωση

Σημειώνουμε ότι δεν χρειάστηκε να κάνουμε όλα τα παραπάνω βήματα στο Power BI για να καταλήξουμε στον ίδιο αριθμό, καθώς η εντολή `DISTINCTCOUNTNOBLANKS` έκανε έναν αριθμό από υποθέσεις, οικεία βουλήσει, αγνοώντας για παράδειγμα τη διαφορά ανάμεσα στα πεζά και τα κεφαλαία γράμματα ή τους προπορευόμενους και λήγοντες κενούς χαρακτήρες στις ονομασίες των StockCode. Αν και στην συγκεκριμένη περίπτωση μπορεί να θεωρηθεί πλεονέκτημα, δεν σημαίνει ότι σε οποιαδήποτε ανάλυση θα καταλήγαμε στο ίδιο συμπέρασμα. Στο συγκεκριμένο dataset έχουμε να κάνουμε με μοναδικούς αριθμούς προϊόντων που είναι ξεκάθαροι και πεπερασμένοι, ενώ σε κάποιο άλλο dataset μια μικρή διαφορά στην ονομασία των προϊόντων θα μπορούσε να υποδηλώνει δύο αντικείμενα διακριτά. Κάθε dataset είναι ξεχωριστό και η εις βάθος κατανόηση των στοιχείων του είναι κρίσιμη προτού ο αναλυτής προβεί στον χειρισμό του.

Στην συνέχεια δημιουργήθηκε ένα νέο DataFrame, το `df2`, το οποίο περιλαμβάνει τις νέες τιμές της στήλης StockCode έπειτα από τις τροποποιήσεις που εφαρμόστηκαν πιο πάνω (Εικόνα 28).

```
# New version of df. This is the base version for the next steps of the analysis.
df['StockCode'] = df['StockCode'].astype(str).str.upper().str.strip()
df2 = df.copy()
```

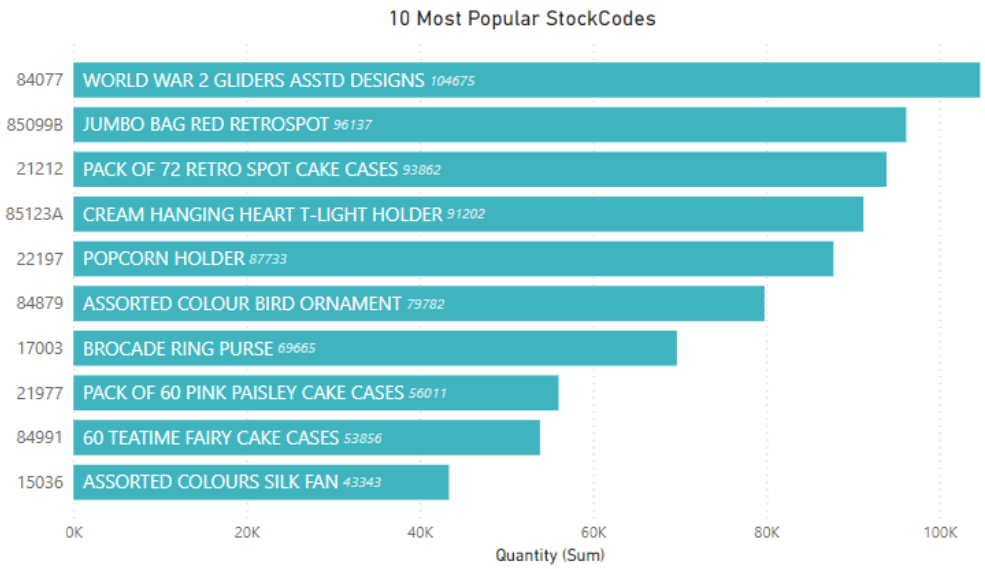
Εικόνα 28. Δημιουργία του αντιγράφου df2 που περιλαμβάνει τις τροποποιημένες τιμές της στήλης StockCode

Κρίνουμε πως ο υπολογισμός των 10 πιο δημοφιλών προϊόντων του dataset θα είχε αξία για τον ιδιοκτήτη του καταστήματος και συνεπώς στην συνέχεια προβαίνουμε ακριβώς σε αυτή την ανάλυση. Αρχικά πρέπει να αποκλείσουμε τα προϊόντα που έχουν μηδενική τιμή στην στήλη Price, καθώς τα συγκεκριμένα είναι προϊόντα που δεν πωλήθηκαν αλλά προσφέρθηκαν ως δώρα ή προσφορές από το ηλεκτρονικό κατάστημα, και συνεπώς δεν υποδηλώνουν τις προτιμήσεις των πελατών. Το ίδιο θα κάναμε και στην περίπτωση που είχαμε ελλείπουσες τιμές στη στήλη Price, κάτι το οποίο δεν συμβαίνει στο παρόν dataset. Στην Εικόνα 29 η συλλογή των προϊόντων με μη μηδενική τιμή συμβολίζεται με το DataFrame `df_nnz_prices`. Εκτελώντας τις υπόλοιπες εντολές της Εικόνας 29 φτάνουμε στη λίστα με τα δέκα πιο ευπώλητα προϊόντα, η οποία περιλαμβάνεται και αυτή στην ίδια εικόνα. Ακολουθώντας την ίδια ανάλυση μέσω Power BI καταλήγουμε στο ραβδόγραμμα της Εικόνας 30. Παρατηρούμε ότι οι δύο ξεχωριστές αναλύσεις συμπίπτουν.

```
# We group by StockCode and sum the values of Quantity in order to review the ten most popular StockCodes of the online shop.
df_fixed_stockcodes = df_nnz_prices.copy()
most_preferred_products = df_fixed_stockcodes.groupby(["StockCode"])["Quantity"].sum().sort_values(ascending=False)[:10]
print(most_preferred_products)
most_preferred_products = pd.DataFrame(most_preferred_products)
```

```
StockCode
84077     104675
85099B     96137
21212      93862
85123A     91202
22197      87733
84879      79782
17003      69665
21977      56011
84991      53856
15036      43343
Name: Quantity, dtype: int64
```

Εικόνα 29. Λίστα με τους κωδικούς και τις ποσότητες των πιο ευπώλητων προϊόντων του ηλεκτρονικού καταστήματος



Εικόνα 30. Ραβδόγραμμα με τα δέκα πιο δημοφιλή προϊόντα του ηλεκτρονικού καταστήματος

- **Στήλη InvoiceDate**

Η στήλη InvoiceDate αναφέρεται στην ημερομηνία και ώρα έκδοσης κάθε τιμολογίου. Όπως αναφέρθηκε και στην αρχική περιγραφή του dataset οι συναλλαγές εκτείνονται από 01/12/2009 έως 09/12/2011. Με σκοπό να ερευνηθεί ποιοι μήνες είχαν τις περισσότερες πωλήσεις δημιουργήθηκαν δύο νέες στήλες: η στήλη *Month* και η στήλη *TotalAmount*, οι οποίες προσαρτήθηκαν στο dataset. Η στήλη Month δημιουργήθηκε με την χρήση της βιβλιοθήκης `datetime` της Python, ενώ η στήλη TotalAmount προέκυψε από το γινόμενο των στηλών Price και Quantity (Εικόνα 31).

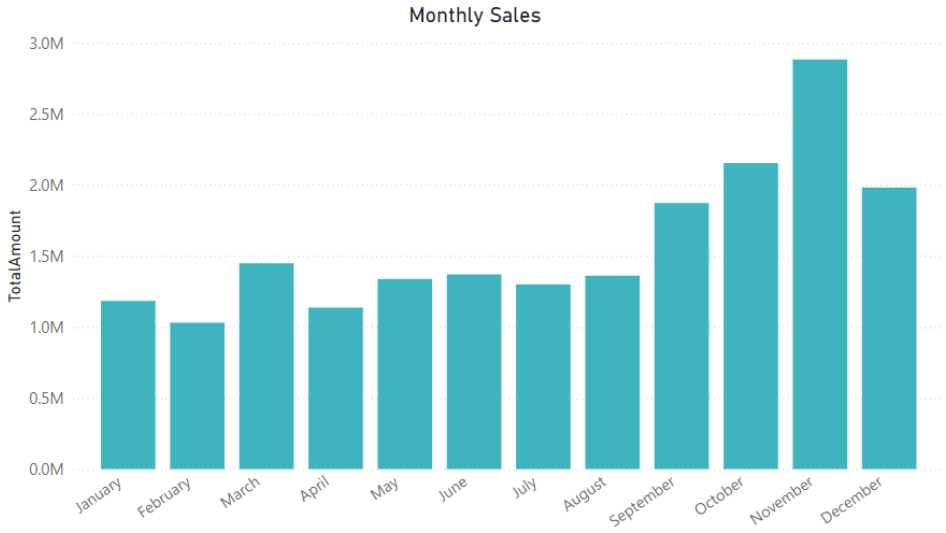
```
# Convert InvoiceDate column to datetime format
df3["InvoiceDate"] = pd.to_datetime(df3["InvoiceDate"])

# Create Month column
df3["Month"] = df3["InvoiceDate"].dt.strftime("%B")

# Create TotalAmount column
df3["TotalAmount"] = df3["Price"] * df3["Quantity"]
```

Εικόνα 31. Δημιουργία στηλών Month και TotalAmount

Στο ραβδόγραμμα της Εικόνας 32 απεικονίζονται οι πωλήσεις ανά μήνα. Παρατηρήθηκε ότι οι μήνες Οκτώβριος, Νοέμβριος και Δεκέμβριος ήταν οι μήνες με τις περισσότερες πωλήσεις.

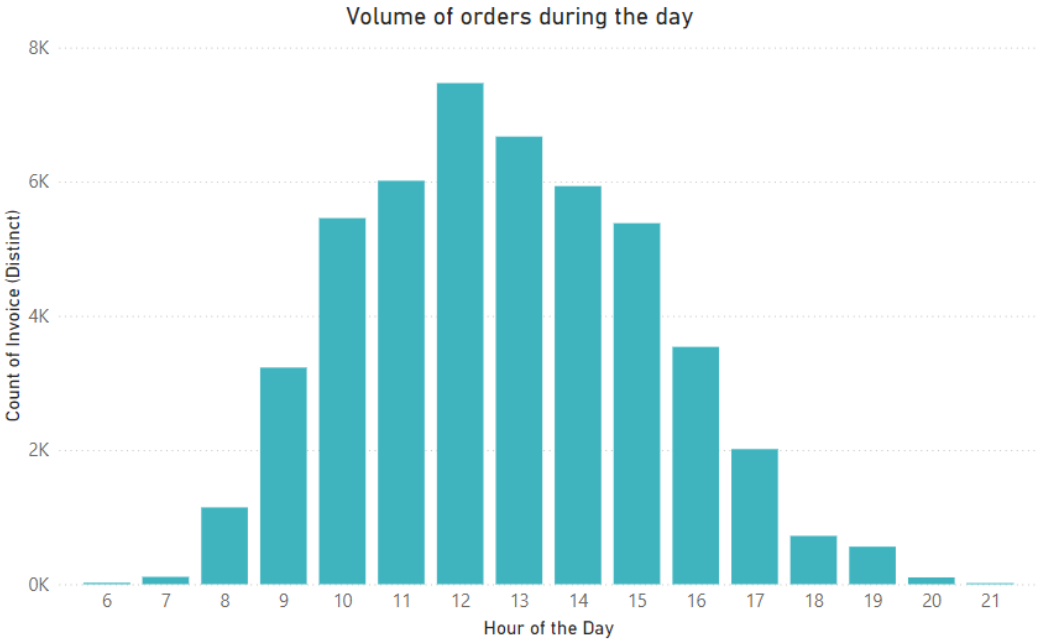


Εικόνα 32. Πωλήσεις ανά μήνα

Επιπλέον δημιουργήθηκε η στήλη *Hour* ώστε να ερευνηθεί ποιες ώρες κατά την διάρκεια της ημέρας έγιναν οι περισσότερες παραγγελίες (Εικόνα 33). Σύμφωνα με το γράφημα που ακολουθεί ο μεγαλύτερος όγκος των παραγγελιών πραγματοποιείται ανάμεσα στις 10 π.μ. και στις 3 μ.μ (Εικόνα 34).

```
# Create Hour column
df3["Hour"] = df3["InvoiceDate"].dt.strftime("%H")
```

Εικόνα 33. Δημιουργία στήλης Hour



Εικόνα 34. Όγκος παραγγελιών ανά ώρα της ημέρας

Για τους σκοπούς της ανάλυσης που θα ακολουθήσει θα αφαιρέσουμε τα στοιχεία της ώρας από την στήλη InvoiceDate, της οποίας οι τιμές θα εμφανίζονται στο εξής στη μορφή «έτος-μήνας-ημέρα».

- **Συνδυασμοί Customer ID και StockCode**

Όπως αναφέρθηκε στην περιγραφή του dataset, συχνά παρατηρείται ότι ο ίδιος πελάτης αγόρασε ένα προϊόν σε διαφορετικές ποσότητες ενώ επίσης μπορεί να επέστρεψε έναν αριθμό τους. Θεωρητικά αναμένουμε πως ο αριθμός που προκύπτει από τη διαφορά των πωληθείσων ποσοτήτων και των επιστραμμένων είναι μεγαλύτερος ή ίσος του μηδενός (αφού κανείς δεν μπορεί να επιστρέψει μεγαλύτερο αριθμό από προϊόντα από όσα αγόρασε). Παρόλα αυτά στην πράξη γνωρίζουμε ότι αυτό είναι μια υπόθεση, της οποίας η αλήθεια θα πρέπει να εξεταστεί. Συνεπώς σκοπός μας είναι για κάθε πελάτη να συμψηφίσουμε τις ποσότητες που αγόρασε και τις ποσότητες που ακύρωσε και, εάν η τελική διαφορά ανά προϊόν βρεθεί με αρνητικό πρόσημο, να διαγραφούν από το dataset όλες οι εγγραφές που αναφέρονται στον συγκεκριμένο συνδυασμό πελάτη και κωδικού προϊόντος.

3.2.3 Ενέργειες μεταχείρισης αρχικού συνόλου δεδομένων εν συνόψει

Στην παρακάτω λίστα παρατίθενται συνοπτικά οι αλλαγές που αναφέρθηκαν παραπάνω που είτε έχουν γίνει είτε θα γίνουν για τον καθαρισμό και την προετοιμασία ενός ορθού dataset για το επόμενο βήμα, το οποίο θα είναι η σύσταση του dataset που θα χρησιμοποιηθεί για την ανάλυση RFM:

- Αναφορικά με τα πρωταρχικά δύο υποσύνολα δεδομένων: διαγραφή των εγγραφών των τελευταίων εννέα ημερών από το πρώτο υποσύνολο, και έπειτα συγχώνευση των δύο υποσυνόλων σε ένα ενιαίο
- Μετατροπή των τιμών της στήλης Invoice σε τύπο string
- Μετατροπή των τιμών της στήλης StockCode σε τύπο string, κεφαλαία, και αφαίρεση των περιμετρικών κενών χαρακτήρων
- Διαγραφή ελλειπουσών τιμών από την στήλη Customer ID
- Η διαγραφή των ελλειπουσών τιμών από την στήλη Customer ID ταυτόχρονα συμπαρασύρει και όλες τις εγγραφές οι οποίες έχουν αρνητική τιμή στην στήλη Price, τις οποίες κατά συνέπεια δεν χρειάζεται να διαγράψουμε

- Για κάθε μοναδικό συνδυασμό Customer ID και StockCode άθροιση των ποσοτήτων των επιμέρους τιμολογίων στα οποία αναφέρεται ο κάθε κωδικός προϊόντος StockCode, και μετέπειτα διαγραφή όλων εκείνων των εγγραφών για τις οποίες το προκύπτον άθροισμα είναι αρνητικό.
- Αφαίρεση της ώρας από την στήλη InvoiceDate

Τα παραπάνω βήματα παρουσιάζονται αναλυτικά στο notebook με την ονομασία 01.cleandata[7] το οποίο είναι διαθέσιμο στην ιστοσελίδα GitHub.

3.3 Η μέθοδος RFM

3.3.1 Υπολογισμός των μετρικών Recency, Frequency, Monetary

Για την τμηματοποίηση των πελατών σε ομάδες με όμοια χαρακτηριστικά, διαμορφώνεται το σύνολο δεδομένων RFM. Για τον υπολογισμό της μετρικής Recency θα πρέπει αρχικά να σημειωθεί πότε πραγματοποιήθηκε η τελευταία συναλλαγή του dataset (09/12/2011). Ως ημερομηνία αναφοράς ορίζεται η αμέσως επόμενη ημέρα, δηλαδή η ημερομηνία 10/12/2011 (Εικόνα 35). Η μεταβλητή Recency προκύπτει από τον αριθμό των ημερών που μεσολαβούν ανάμεσα στην ημερομηνία της πιο πρόσφατης αγοράς που πραγματοποίησε ο κάθε πελάτης και στην ημερομηνία αναφοράς.

```
# Set date 12/10/20211, which is one day after the last transaction date, as today_date
today_date = dt.datetime(2011,12,10)
```

Εικόνα 35. Προσδιορισμός της ημερομηνίας αναφοράς

Η μετρική Frequency ορίζεται ως ο συνολικός αριθμός μοναδικών τίτλων προϊόντων που αγόρασε ένας πελάτης, δηλαδή ο συνολικός αριθμός εγγραφών κάθε πελάτη στο συγκεκριμένο dataset. Επιπλέον, για τον υπολογισμό της μετρικής Monetary, είναι απαραίτητη η δημιουργία της στήλης TotalValue η οποία προκύπτει από το γινόμενο των στηλών Price και Quantity για κάθε συναλλαγή. Η μετρική Monetary προκύπτει από το άθροισμα των τιμών της στήλης TotalValue για κάθε πελάτη.[8]

[7] https://github.com/ArgyroMp/msc-thesis/blob/master/Notebooks/01.cleandata.ipynb

[8] Η ανάλυση που ακολουθεί διεξήχθη και θέτοντας ως Frequency την συνολική ποσότητα προϊόντων που αγόρασε ένας πελάτης. Σε αυτή την περίπτωση οι μετρικές Frequency και Monetary δεν ήταν γραμμικά ανεξάρτητες καθώς και οι δύο περιλάμβαναν τον όρο της συνολικής ποσότητας προϊόντων. Για αυτό το λόγο διαιρέσαμε την μετρική Monetary με την συνολική ποσότητα προϊόντων, όπως είναι ένας δεύτερος τρόπος ανάλυσης RFM που αναφέρεται στην βιβλιογραφία (Wei et al., 2010), ώστε στο τέλος να προκύψει η μέση χρηματική αξία που ξόδεψε ο κάθε πελάτης. Στην δική μας περίπτωση ο διαχωρισμός των πελατών με αυτό τον τρόπο δεν απέδωσε αποτελέσματα τόσο σαφή όσο η παραδοσιακή μέθοδος, την οποία ακολουθήσαμε.

Έχοντας ως βάση τις στήλες InvoiceDate, Invoice και TotalValue και με τον συνδυασμό των εντολών της Python `groupby()` και `agg()`, υπολογίζουμε τις μετρικές Recency, Frequency και Monetary για κάθε πελάτη (Εικόνα 36).

```
# Group by Customer ID and calculate RFM values
rfm = customer_data.groupby(["Customer ID"], as_index=False).agg(
                              {"InvoiceDate": lambda x :(today_date-x.max()).days,
                                 "Invoice": lambda x : x.count(),
                                 "TotalValue":lambda x : x.sum()})
```

```
# Rename columns accordingly
rfm.columns = ["Customer ID", "Recency", "Frequency", "Monetary"]
rfm.head()
```

	Customer ID	Recency	Frequency	Monetary
0	12346	432	34	368.36
1	12347	3	222	4921.53
2	12348	76	51	2019.40
3	12349	19	179	4419.49
4	12350	311	17	334.40

Εικόνα 36. Υπολογισμός των μετρικών Recency, Frequency και Monetary στη γλώσσα Python

Τα παραπάνω βήματα υλοποιήθηκαν αντίστοιχα και στο Power BI όπου με την χρήση των εντολών `DATEDIFF`, `COUNT` και `SUM` της γλώσσας DAX υπολογίστηκαν ως measures οι τρεις μετρικές για κάθε πελάτη (Εικόνα 37).

```
R value = DATEDIFF([LastTransactionDate],
                   DATE (2011, 12, 10),
                   DAY)
```

```
F Value = COUNT('02_rfmdata'[Invoice])
```

```
M value = SUM('02_rfmdata'[TotalValue])
```

Εικόνα 37. Υπολογισμός των μετρικών Recency, Frequency και Monetary στη γλώσσα DAX του Power BI

3.3.2 Τμηματοποίηση πελατών

Ο υπολογισμός των παραπάνω μετρικών για κάθε πελάτη δίνει την δυνατότητα στην επιχείρηση να έχει πρόσβαση σε χρήσιμες πληροφορίες, ο όγκος των οποίων παρόλα αυτά τις καθιστά δύσκολα διαχειρίσιμες. Για να αντιμετωπιστεί αυτό το πρόβλημα, σύμφωνα με την RFM ανάλυση, για κάθε μία από τις μετρικές Recency, Frequency και Monetary, κάθε καταναλωτής λαμβάνει μία βαθμολογία (score) από τον αριθμό 1 έως τον αριθμό 5 (το 5 αντιπροσωπεύει την καλύτερη βαθμολογία). Για να γίνει αυτό, το σύνολο των καταναλωτών χωρίζεται σε 5 ισοπληθείς κατηγορίες. Το 20% των πελατών με μεγαλύτερη συχνότητα συναλλαγών λαμβάνει score 5 για την στήλη Frequency, το 20% των πελατών με τις αμέσως μεγαλύτερες συχνότητες λαμβάνει score 4, και η λογική αυτή επαναλαμβάνεται μέχρι το score 1. Η ίδια διαδικασία επαναλαμβάνεται για τη στήλη Monetary, ενώ για τη στήλη Recency μεγαλύτερα score λαμβάνουν μικρότερες τιμές εγγύτητας αγοράς προς την ημερομηνία αναφοράς. Στην συνέχεια για κάθε πελάτη οι βαθμολογίες που έχει λάβει σε κάθε κατηγορία συνενώνονται και δημιουργούν το τριψήφιο RFM score του πελάτη.

Ο παραπάνω τρόπος εξαγωγής των RFM scores είναι ο προκαθορισμένος τρόπος που αναφέρεται στην βιβλιογραφία. Καθώς όμως όλα τα dataset δεν έχουν τα ίδια χαρακτηριστικά μεταξύ τους δεν επιδέχονται απαραίτητα της ίδιας αντιμετώπισης γενικής χρήσεως. Για παράδειγμα θα μπορούσε να υφίσταται σύνολο δεδομένων για το οποίο είναι αδύνατος ο χωρισμός των πελατών σε 5 ισόποσες κατηγορίες για κάποια από τις τρεις μετρικές, στην οποία περίπτωση θα έπρεπε να χρησιμοποιηθεί κάποιος άλλος τρόπος ομαδοποίησης σε 5 κατηγορίες, όπως η ομαδοποίηση μέσω clustering. Σε αυτό το πρόβλημα θα είχαμε καταλήξει με το παρόν σύνολο δεδομένων εάν για την μετρική Frequency αντί για τον μοναδικό αριθμό τίτλων προϊόντων είχε χρησιμοποιηθεί ο μοναδικός αριθμός τιμολογίων. Επιπρόσθετα, δεν είναι παράλογο για μία εταιρία να επιθυμεί να προσαρμόσει την αντιμετώπιση των πελατών της στα ιδιαίτερα και ιδιάζοντα χαρακτηριστικά των συνόλων δεδομένων της, με αποτέλεσμα να επιθυμεί να τους ομαδοποιήσει με λεπτότερο τρόπο ή διαφορετικά κριτήρια από το χωρισμό τους σε πέντε ισόποσες ομάδες. Για αυτό τον λόγο παρακάτω παρουσιάζουμε δύο διαφορετικές μεθόδους εξαγωγής των RFM scores (οι οποίες μπορεί να απευθύνονται σε αναλυτές διαφορετικού επιπέδου γνώσεων προγραμματισμού), και μία μέθοδο τμηματοποίησης πελατών, η οποία σε αντίθεση με τις άλλες δύο εφαρμόζεται απευθείας στον τρισδιάστατο RFM χώρο.

Στις επόμενες δύο ενότητες παρουσιάζονται δύο διαφορετικές μέθοδοι υπολογισμού των RFM scores τα οποία στην συνέχεια θα χρησιμοποιηθούν για τον καθορισμό των τμημάτων (segments) των πελατών της επιχείρησης. Ο συνηθέστερος εν λόγω τρόπος παραγωγής τμημάτων βασίζεται στην εφαρμογή στατικών κανόνων οι οποίοι ομαδοποιούν πελάτες με συναφή RFM scores στα λεγόμενα τμήματα. Ο αριθμός των κανόνων (και συνεπώς η αναλυτικότητα τους) και η συνολική ή μερική κάλυψη του αριθμού των πελατών αποτελεί ανάκλαση της εκάστοτε στρατηγικής προσέγγισης της εταιρίας που εφαρμόζει την RFM ανάλυση, δηλαδή των πόρων της και των επιθυμητών οραματιζόμενων αποτελεσμάτων της διοίκησής της. Στην Εικόνα 1 της ενότητας 2.2.3 παρουσιάστηκε μία συνήθης τμηματοποίηση σε έντεκα τμήματα, των οποίων οι κανόνες εξαγωγής με βάση ένα RFM score εμφανίζονται στις τρεις τελευταίες στήλες του εικονιζόμενου πίνακα (Cuce & Tiryaki, 2022).

3.3.2.1 Μέσω των RFM scores και πεμπτημορίων

Στην Εικόνα 38 παρουσιάζεται ο τρόπος υπολογισμού των R και F scores μέσω Power BI με την χρήση της συνάρτησης `PERCENTILE.INC()`. Το M score υπολογίζεται κατά τον ίδιο τρόπο με το F score. Στο τέλος τα 3 scores συνενώνονται για την δημιουργία του RFM score κάθε πελάτη.

```
R Score = SWITCH(
  TRUE(),
   [Recency] <= PERCENTILE.INC('RFMtable'[Recency], 0.20), "5",
   [Recency] <= PERCENTILE.INC('RFMtable'[Recency], 0.40), "4",
   [Recency] <= PERCENTILE.INC('RFMtable'[Recency], 0.60), "3",
   [Recency] <= PERCENTILE.INC('RFMtable'[Recency], 0.80), "2",
   "1"
    )

F Score = SWITCH(
  TRUE(),
   [Frequency] <= PERCENTILE.INC('RFMTable'[Frequency], 0.20), "1",
   [Frequency] <= PERCENTILE.INC('RFMTable'[Frequency], 0.40), "2",
   [Frequency] <= PERCENTILE.INC('RFMTable'[Frequency], 0.60), "3",
   [Frequency] <= PERCENTILE.INC('RFMTable'[Frequency], 0.80), "4",
   "5"
    )
```

Εικόνα 38. Υπολογισμός R score και F score με τη χρήση της συνάρτησης PERCENTILE.INC() της γλώσσας DAX στο Power BI

Ο πίνακας της Εικόνας 39 απεικονίζει τα RFM scores των 10 πρώτων πελατών ταξινομημένων κατά αύξουσα σειρά με βάση το Customer ID τους.

Customer ID	Recency	Frequency	Monetary	R Score	F Score	M Score	RFM Score
12346	432	34	368.36	1	2	2	122
12347	3	222	4921.53	5	5	5	555
12348	76	51	2019.40	3	3	4	334
12349	19	179	4419.49	5	4	5	545
12350	311	17	334.40	2	1	2	212
12351	376	21	300.93	2	2	2	222
12352	37	107	1889.21	4	4	4	444
12353	205	24	406.76	2	2	2	222
12354	233	58	1079.40	2	3	3	233
12355	215	35	947.61	2	2	3	223

Εικόνα 39. Τα RFM scores των 10 πρώτων πελατών ταξινομημένων κατά αύξουσα σειρά με βάση το Customer ID τους (μέθοδος percentiles)

3.3.2.2 Μέσω των RFM scores και συσταδοποίησης ανά στήλη

Στην Εικόνα 40 παρουσιάζεται η συνάρτηση στην γλώσσα Python που μπορεί να χρησιμοποιηθεί για να ομαδοποιήσει τα δεδομένα της στήλης `rfm_column_name` του συνόλου δεδομένων `rfm` σε `n_clusters`.

```
In [20]: def my_iterative_kmeans(rfm, rfm_column_name, num_runs, n_clusters):

             uids = []
             rfms = []

             # python produces 0:4 labels, we want 1:5
             labels_mapping = {0:1, 1:2, 2:3, 3:4, 4:5}

             for i in range(num_runs):
                 # The following two lines are the essence of Kmeans
                 # After appropriate renamings these two lines constitute the code that should run in Power BI
                 kmeans = KMeans(n_clusters=n_clusters).fit(rfm[[rfm_column_name]])
                 rfm[rfm_column_name+'_clustered'] = kmeans.labels_

                 # https://stackoverflow.com/questions/44888415/how-to-set-k-means-clustering-labels-from-highest-to-lowest-with-python
                 idx = np.argsort(kmeans.cluster_centers_.sum(axis=1))
                 lut = np.zeros_like(idx)
                 lut[idx] = np.arange(n_clusters)

                 # Map from 0-4 to 1-5
                 lut[kmeans.labels_] = [labels_mapping[i] for i in lut[kmeans.labels_]]
                 rfm[rfm_column_name+'_clustered'] = [labels_mapping[i] for i in rfm[rfm_column_name+'_clustered']]

                 uid = [sum(lut[kmeans.labels_] == 1),
                        sum(lut[kmeans.labels_] == 2),
                        sum(lut[kmeans.labels_] == 3),
                        sum(lut[kmeans.labels_] == 4),
                        sum(lut[kmeans.labels_] == 5)],

                 uids.append(''.join(str(uid)))
                 rfms.append(rfm)

             return uids,rfms
```

Εικόνα 40. Συνάρτηση που ομαδοποιεί τα δεδομένα της στήλης rfm_column_name του συνόλου δεδομένων rfm σε n_clusters

Το αποτέλεσμα του clustering μέσω k-means είναι πιθανοτικό, το οποίο σημαίνει ότι δεν είναι απαραίτητο ότι κάθε εκτέλεσή του παράγει το ίδιο αποτέλεσμα. Για αυτό τον λόγο, διεξάγουμε clustering μέσω k-means `num_runs` = 1000 φορές, ώστε να δεχθούμε ως ορθό αποτέλεσμα των RFM scores το πιο συχνό αποτέλεσμα συσταδοποίησης. Σαν επόμενο βήμα θα μπορούσαμε να εισάγουμε αυτό το αποτέλεσμα στο Power BI και να προχωρήσουμε και στα δύο εργαλεία στην περαιτέρω ανάλυση με την λογική RFM, ήτοι στην ομαδοποίηση των πελατών στον τρισδιάστατο RFM χώρο. Παρατηρήστε ότι θα μπορούσαμε να κάνουμε ακριβώς την ίδια ανάλυση εκτελώντας κώδικα Python μέσω Power BI και να εισάγουμε το αποτέλεσμα πίσω στο Jupyter notebook. Αυτός ο τρόπος όμως θα ήταν πιο χρονοβόρος και θα απαιτούσε από αυτόν που χρησιμοποιούσε αποκλειστικά το Power BI ως εργαλείο γνώσεις Python.

Ο πίνακας της Εικόνας 41 απεικονίζει τα RFM scores των 10 πρώτων πελατών ταξινομημένων κατά αύξουσα σειρά με βάση το Customer ID τους.

```
In [361]: rfm4.head(10)
```

Out[361]:

	Customer ID	Recency	Frequency	Monetary	Recency_clustered	Frequency_clustered	Monetary_clustered	score
0	12346	432.0	34.0	368.36	2	1	1	211
1	12347	3.0	222.0	4921.53	5	2	1	521
2	12348	76.0	51.0	2019.40	5	1	1	511
3	12349	19.0	179.0	4419.49	5	1	1	511
4	12350	311.0	17.0	334.40	3	1	1	311
5	12351	376.0	21.0	300.93	2	1	1	211
6	12352	37.0	107.0	1889.21	5	1	1	511
7	12353	205.0	24.0	406.76	3	1	1	311
8	12354	233.0	58.0	1079.40	3	1	1	311
9	12355	215.0	35.0	947.61	3	1	1	311

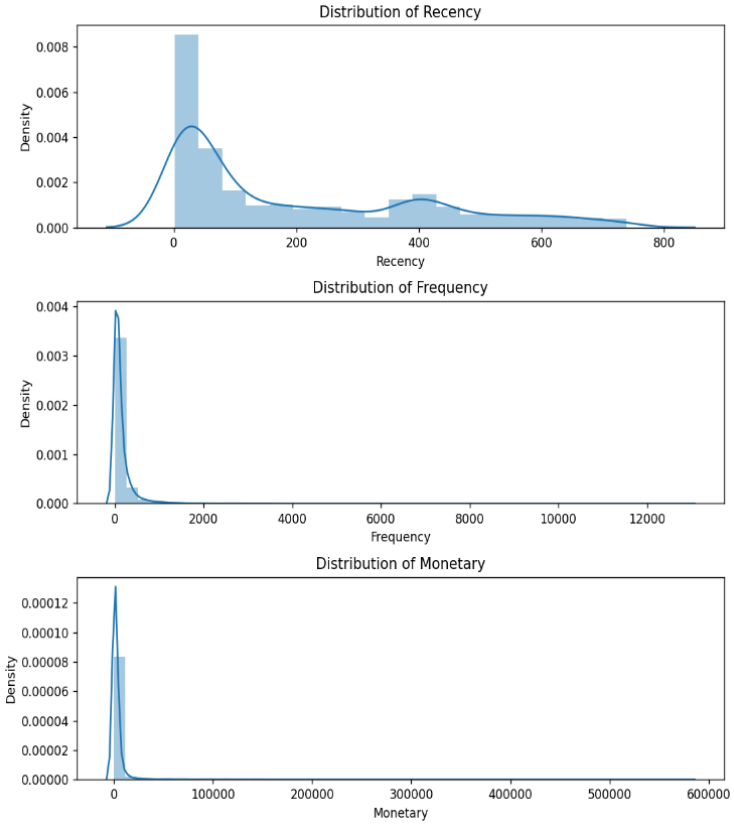
Εικόνα 41. Τα RFM scores των 10 πρώτων πελατών ταξινομημένων κατά αύξουσα σειρά με βάση το Customer ID τους (μέθοδος column k-means)

3.3.2.3 Μέσω συσταδοποίησης στον τρισδιάστατο RFM χώρο

Στην παρούσα ενότητα παρουσιάζεται η τρίτη μέθοδος τμηματοποίησης των πελατών της επιχείρησης, η οποία δεν απαιτεί τη δημιουργία βαθμολογιών αλλά πραγματοποιείται με την εφαρμογή του αλγορίθμου K-means απευθείας στα σημεία του τρισδιάστατου χώρου που έχει ως διαστάσεις του τις μετρικές Recency, Frequency και Monetary.

Από τη βιβλιογραφία (Melnykov & Zhu, 2019) είναι γνωστό ότι η κατανομή των δεδομένων επηρεάζει σημαντικά τα αποτελέσματα συσταδοποίησης από την εφαρμογή του αλγορίθμου

K-means. Ο αλγόριθμος παράγει καλύτερα αποτελέσματα όταν τα δεδομένα στα οποία εφαρμόζεται ικανοποιούν τις δύο συνθήκες που παρουσιάζονται στη συνέχεια. Η πρώτη συνθήκη αναφέρεται στον έλεγχο της λοξότητας (skewness) της κατανομής που τα δεδομένα ακολουθούν. Καθώς τα δεδομένα και των 3 μεταβλητών είναι ασύμμετρα κατανεμημένα (κατανομές λοξές δεξιά), αυτά λογαριθμοποιήθηκαν ώστε να μειωθεί η λοξότητά τους (να τείνει όσο γίνεται στο 0). Η δεύτερη συνθήκη που πρέπει να ικανοποιείται αφορά την έκφραση των δεδομένων με όρους κοινής κλίμακας (standardisation). Η κανονικοποίηση είναι η διαδικασία με την οποία επαναπροσδιορίζονται οι τιμές μιας μεταβλητής με βάση την τυπική απόκλισή τους. Η προκύπτουσα τυπική απόκλιση μετά την κανονικοποίηση είναι 1. Από τα ιστογράμματα της Εικόνας 42 παρατηρείται ότι τα δεδομένα των στηλών είναι εκφρασμένα σε διαφορετικές μονάδες μέτρησης, και συνεπώς για την προσαρμογή τους σε κοινή κλίμακα έγινε χρήση του `PowerTransformer` της βιβλιοθήκης `sklearn`. Στην Εικόνα 43 παρουσιάζεται ο αλγόριθμος που εφαρμόζεται για την εκτέλεση των παραπάνω βημάτων εκφρασμένος σε ψευδοκώδικα.



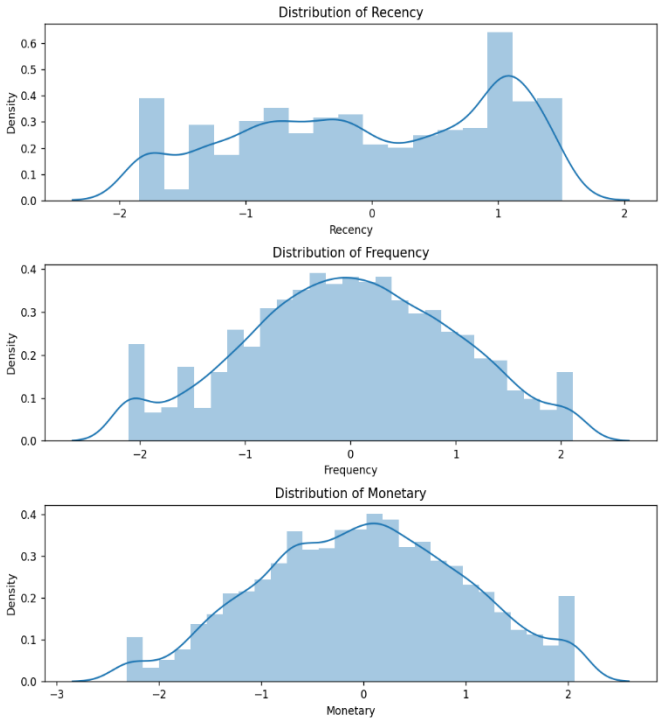
Εικόνα 42. Κατανομές των τιμών των μεταβλητών R, F και M

Algorithm 1: `Sanitise K-means input`

Require: Columns `Recency`, `Frequency`, `Monetary`
Ensure: Input standardised and unskewed
1: **for** Column $c \in$ {Recency, Frequency, Monetary} **do**
2: **if** c is skewed **then**
3: apply log transformation to c
4: **end if**
5: standardise c via power transform
6: **end for**
7: **return** standardised and unskewed inputs

Εικόνα 43. Ψευδοκώδικας για την λογαριθμοποίηση και κανονικοποίηση των δεδομένων των στηλών R, F και M

Στην Εικόνα 44 παρουσιάζονται οι κατανομές των δεδομένων έπειτα από τους προαναφερθέντες μετασχηματισμούς.



Εικόνα 44. Κατανομές των τιμών των μεταβλητών R, F και M (μετά την λογαριθμοποίηση και κανονικοποίησή τους)

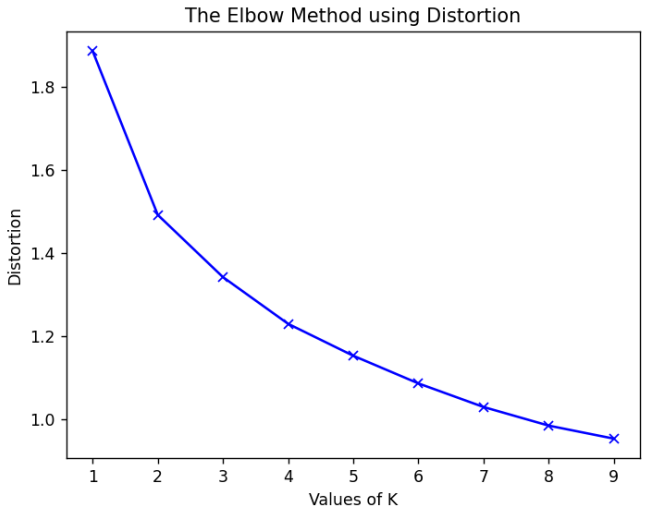
Αφού κανονικοποιηθούν τα δεδομένα, το επόμενο βήμα είναι η επιλογή του βέλτιστου αριθμού συστάδων των δεδομένων στον τρισδιάστατο RFM χώρο. Προς αυτή την κατεύθυνση εφαρμόζουμε τη μέθοδο Elbow, κατά την οποία τρέχουμε τον αλγόριθμο k-means επανειλημμένα για k=1,…,10 και για κάθε τιμή του k υπολογίζουμε την τιμή WCSS (within-

cluster sum of squares). Αυτό το βήμα παρουσιάζεται επίσης εκφρασμένο σε ψευδοκώδικα στην Εικόνα 45. Η τιμή WCSS δείχνει ουσιαστικά την «παραμόρφωση» (distortion) που παρατηρείται στην ομοιομορφία των σημείων που σχηματίζουν την κάθε συστάδα: είναι εύλογο να επιθυμούμε την μικρότερη δυνατή «παραμόρφωση». Η μέθοδος Elbow υποδεικνύει ότι το βέλτιστο k εντοπίζεται στο σημείο από το οποίο η «παραμόρφωση» ξεκινά να μειώνεται με γραμμικό τρόπο (Εικόνα 46). Στην συγκεκριμένη περίπτωση παρατηρούμε πως αυτή η συνθήκη ισχύει στη γειτονιά της τιμής k = 4.

Algorithm 2: `Cluster distortion from RFM values for varying` K

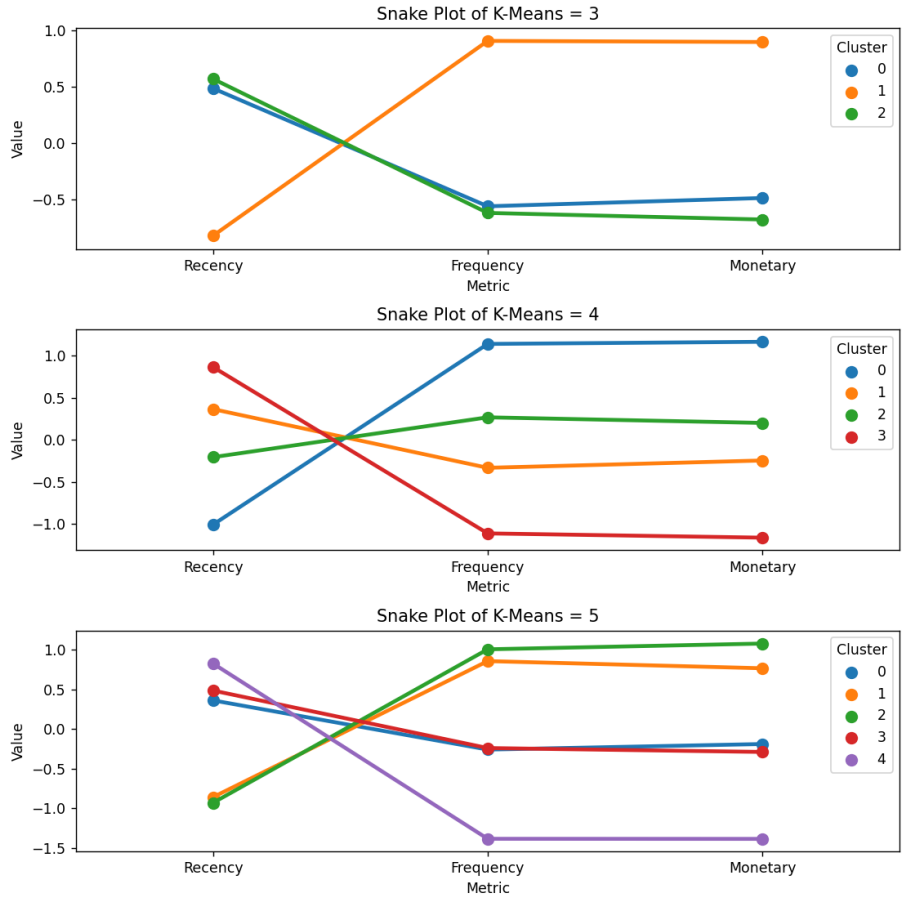
Require: Sanitised RFM values D
Ensure: Cluster distortion for varying K
1: DISTORTION $\leftarrow \varnothing$
2: **for** $k \in \{1, 10\}$ **do**
3: $km \leftarrow$ `kmeans`(D, k)
4: distortion $\leftarrow 0$
5: **for** $d \in D$ **do**
6: distortion $\leftarrow$ distortion + `euclidean_distance`$(d,$ `km.centroid_of`$(d))$
7: **end for**
8: append distortion to DISTORTION
9: **end for**
10: **return** DISTORTION

Εικόνα 45. Ψευδοκώδικας για τον υπολογισμό του WCSS (distortion) για k=1,...,10



Εικόνα 46. Εφαρμογή της μεθόδου Elbow

Για την εξακρίβωση της ακριβούς ελάχιστης τιμής του αριθμού συστάδων k στρεφόμαστε στη χρήση των snake plots της Εικόνας 47, όπου για k=3, k=4, και k=5, παρουσιάζονται οι μέσες τιμές των μετρικών Recency, Frequency, και Monetary των σημείων που ανήκουν σε κάθε συστάδα, ανά συστάδα. Στα εν λόγω διαγράμματα παρατηρούμε ότι για k=3 και k=5 υπάρχουν συστάδες που με μικρές αποκλίσεις εμφανίζουν τις ίδιες μέσες τιμές μεταξύ τους (πχ για k=3 οι συστάδες 0 και 2, και για k=5 ανά δύο οι συστάδες 1 και 2, και οι 0 και 3). Αυτό το γεγονός υπονοεί πως, κατά μέσο όρο, τα σημεία αυτών των συστάδων ισαπέχουν από τις δύο συστάδες, και συνεπώς ότι η αντιστοίχιση των σημείων στις συστάδες είναι αμφίσημη. Αντιθέτως, για k=4 οι αντίστοιχες μέσες τιμές είναι διακριτές μεταξύ τους ανά συστάδα. Κατά συνέπεια επιλέγουμε να τμηματοποιήσουμε το σύνολο των πελατών σε τέσσερα μέρη.



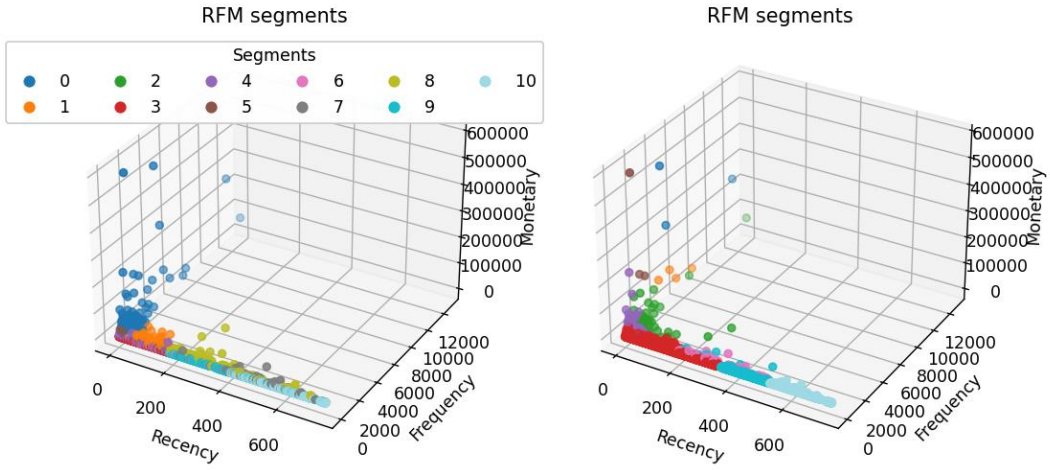
Εικόνα 47. Snake plots για k=3, k=4, και k=5

Κεφάλαιο 4: Αποτελέσματα-Συζήτηση

4.1 Αποτελέσματα κατηγοριοποίησης πελατών μέσω percentiles και column k-means

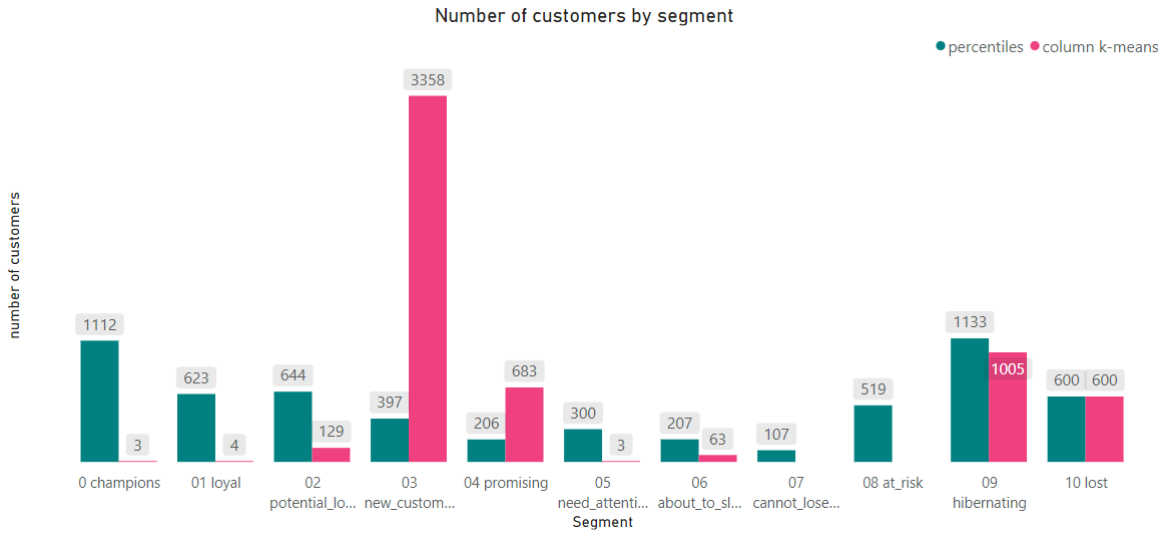
Στην Εικόνα 48 παρουσιάζεται στον τρισδιάστατο RFM χώρο η κατηγοριοποίηση των πελατών σε 11 και 9 τμήματα αντίστοιχα με βάση το RFM score τους, έπειτα από την εφαρμογή των δύο διαφορετικών μεθόδων υπολογισμού τους (percentiles και column k-means).

Το κυριότερο που παρατηρούμε στην Εικόνα 48 είναι η διαφορά στον χωρικό διαχωρισμό των τμημάτων των δύο μεθόδων. Η μέθοδος column k-means έχει παράξει τμήματα τα οποία είναι περισσότερο ευδιάκριτα σε σχέση με τη μέθοδο percentiles, και με μεγαλύτερη συνοχή μεταξύ τους (τα τμήματα αυτά δεν εμφανίζουν πελάτες άλλων τμημάτων μέσα τους, σε αντίθεση με τη μέθοδο percentiles). Αυτό ίσως οφείλεται στο γεγονός ότι η μέθοδος k-means δεν έχει τον περιορισμό της ομαδοποίησης σε ομάδες με ίσο πλήθος πελατών, όπως η μέθοδος percentiles, δίνοντας έτσι βάση στην χωρική συνάφεια των πελατών και όχι στον απλό ποσοτικό διαχωρισμό τους.



Εικόνα 48. Τμηματοποίηση με βάση τη μέθοδο percentiles (αριστερά) και με βάση τη μέθοδο column k-means (δεξιά). Τα ονόματα των τμημάτων δεν αναφέρονται για λόγους οικονομίας χώρου. Για τα ονόματα βλ. Εικόνα 1.

Αποτελέσματα αυτού του γεγονότος αποτελούν (α) η διαφορά στον αριθμό των μη κενών scores των δύο μεθόδων, όπου η μεν μέθοδος percentiles κατατάσσει τους πελάτες σε 115 διαφορετικά scores, ενώ η μέθοδος column k-means τους κατατάσσει σε 40, εκ των 5x5x5=125 μέγιστων δυνατών scores, και (β) οι αριθμοί των πελατών που περιέχει κάθε τμήμα ανά μέθοδο υπολογισμού, όπως αυτοί απεικονίζονται στην Εικόνα 49. Σε αυτό το διάγραμμα βλέπουμε ότι οι αριθμοί πελατών ανά τμήμα στη μέθοδο percentiles εμφανίζουν μικρότερη διακύμανση από αυτούς της μεθόδου k-means, η οποία εν προκειμένω έχει κρίνει πως περίπου το 60% των πελατών του dataset σε σχέση με τους άλλους πελάτες είναι νέοι πελάτες (ανήκουν στο segment id: 03). Αυτό επίσης σημαίνει ότι η μέθοδος column k-means δεν είναι υποχρεωμένη να κατατάξει πελάτες σε κάθε κατηγορία, όπως για παράδειγμα παρατηρούμε ότι δεν εντάσσει κανέναν πελάτη στα segment id: 07 (“cannot lose them but losing”) και segment id: 08 (“at risk”).



Εικόνα 49. Αριθμός πελατών ανά segment και ανά μέθοδο (percentiles vs. column k-means)

Κάτι που αξίζει να σημειωθεί για το συγκεκριμένο σύνολο δεδομένων ως επιπρόσθετο αποτέλεσμα της διαφοράς των δύο μεθόδων ως προς το μέγεθος των πεμπτημορίων/συστάδων είναι η κατανομή των πελατών στα ακραία τμήματα. Παρατηρήστε στην Εικόνα 49 πως τα δύο πιο επικερδή τμήματα για την επιχείρηση, δηλαδή οι “champions” και οι “loyal” αποτελούνται από ελάχιστους πελάτες από τη μεν μέθοδο column k-means (3 και 4 αντίστοιχα), ενώ για τη μέθοδο percentiles αυτές οι δύο κατηγορίες αποτελούν σχεδόν το ένα τρίτο των συνολικών πελατών της επιχείρησης. Στον αντίποδα, όσον αφορά στους πελάτες που είναι πλέον αδιάφοροι για την επιχείρηση, δηλαδή οι “hibernating” και “lost”, οι δύο μέθοδοι βρίσκονται

σε συμφωνία ως προς το μέγεθός τους. Από τον πίνακα της Εικόνας 50 παρατηρούμε επιπλέον ότι για τα δύο τελευταία segments οι δύο μέθοδοι συμφωνούν σε ένα μεγάλο ποσοστό και ως προς το ποιοι πελάτες κατατάσσονται σε αυτά. Συγκεκριμένα, οι δύο μέθοδοι κατατάσσουν τους ίδιους 429 πελάτες στο τμήμα "hibernating", αριθμός που αντιστοιχεί στο 37,86% των πελατών "hibernating" της μεθόδου percentiles και στο 42,69% της μεθόδου column k-means. Αντίστοιχα οι δύο μέθοδοι κατατάσσουν τους ίδιους 380 πελάτες στο τμήμα "lost", αριθμός που αντιστοιχεί στο 63,33% των πελατών "lost" και για τις δύο μεθόδους (αφού και οι δύο συμπεριέλαβαν 600 πελάτες στο συγκεκριμένο τμήμα). Όσον αφορά το τμήμα με τον μεγαλύτερο αριθμό πελατών της μεθόδου column k-means, 397 πελάτες αυτού του τμήματος, ήτοι σχεδόν το 12% των συνολικών πελατών του τμήματος, βρέθηκαν να ανήκουν αυτούσιοι στο αντίστοιχο τμήμα της μεθόδου percentiles, οι οποίοι αντιστοιχούν στο σύνολο των πελατών αυτού του τμήματος. Αντίστροφα, όλοι οι πελάτες που ο column k-means θεώρησε ως "champions" θεωρήθηκαν ως "champions" και από την μέθοδο percentiles, για την οποία όμως αποτελούν μοναχά λιγότερο από το 1% των πελατών του συγκεκριμένου τμήματος (0,27%). Οι δύο μέθοδοι βρίσκονται σε διαφωνία για τους πελάτες όλων των υπόλοιπων τμημάτων πέραν αυτών που αναφέρθηκαν.

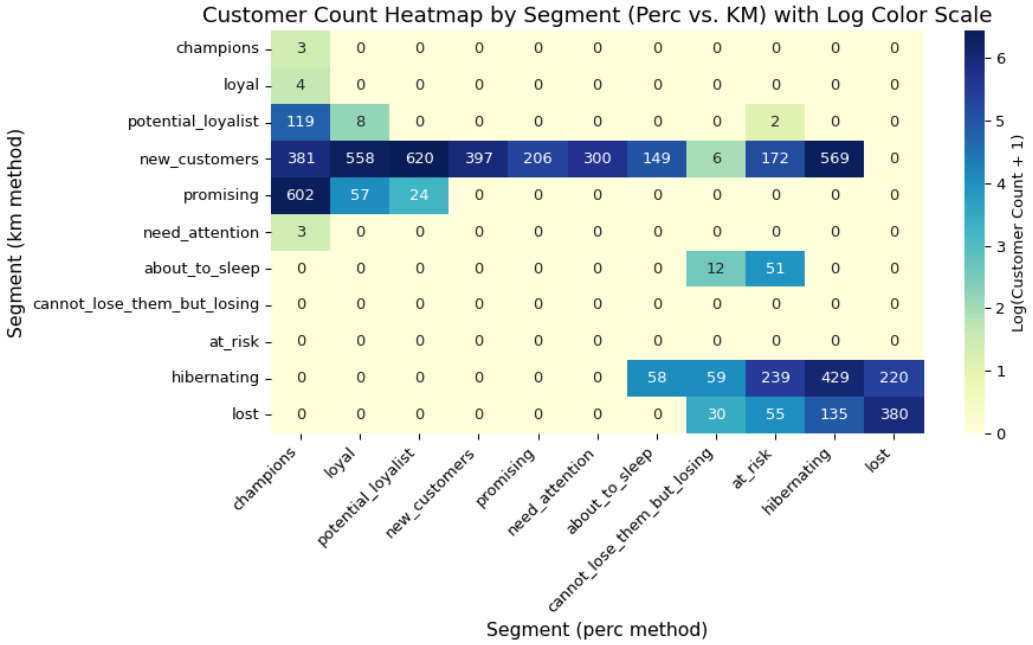
Out[53]:

	segment name	Common Count	Percentage_percentiles	Percentage_column_k_means
0	champions	3	0.27	100.00
1	loyal	0	0.00	0.00
2	potential_loyalist	0	0.00	0.00
3	new_customers	397	100.00	11.82
4	promising	0	0.00	0.00
5	need_attention	0	0.00	0.00
6	about_to_sleep	0	0.00	0.00
7	cannot_lose_them_but_losing	0	0.00	0.00
8	at_risk	0	0.00	0.00
9	hibernating	429	37.86	42.69
10	lost	380	63.33	63.33

Εικόνα 50. Ποσοστά συμφωνίας των μεθόδων percentiles και column k-means ως προς την κατανομή των πελατών στα segments

Στο heatmap της Εικόνα 51 εμφανίζεται η πλήρης διασταυρωμένη κατανομή και κατάταξη πελατών σε segments ανά μέθοδο. Αυτός ο πίνακας μας δείχνει για παράδειγμα (α) ότι η μέθοδος column k-means κατέταξε από κοινού με τη μέθοδο percentiles 3 πελάτες στο τμήμα "champions", αλλά η πρώτη κατέταξε 4 πελάτες στο τμήμα "loyal", 119 στο τμήμα "potential

loyalists", 381 στο τμήμα "new customers", 602 στο τμήμα "promising" και 3 στους "need attention", ενώ η δεύτερη τους κατέταξε όλους αυτούς στο τμήμα "champions". Η μεγαλύτερη διαφορά σε κατάταξη παρατηρείται για 569 πελάτες που η μεν μέθοδος column k-means τους κατατάσσει σε "new customers" ενώ η μέθοδος percentiles τους κατατάσσει σε "hibernating".



Εικόνα 51. Heatmap που απεικονίζει την πλήρη διασταυρωμένη κατανομή και κατάταξη πελατών σε segments ανά μέθοδο (percentiles & column k-means)

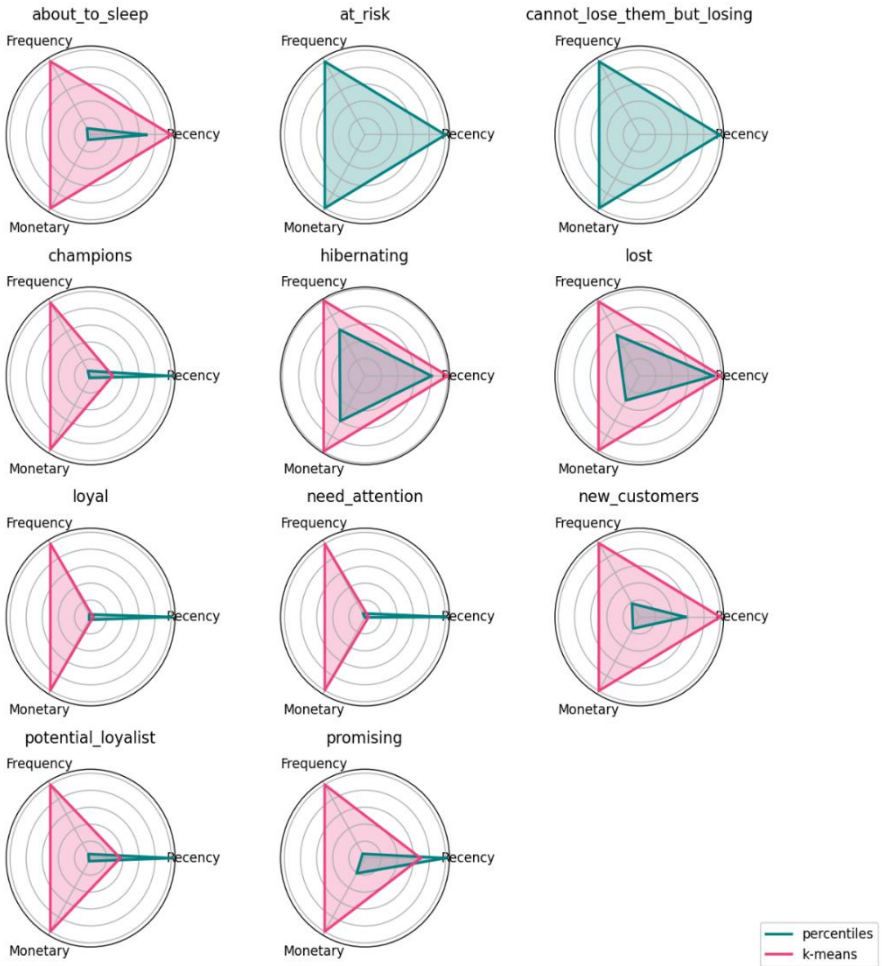
Αν η επιχείρηση αποφασίσει να υιοθετήσει τη μέθοδο **column k-means** για την τμηματοποίηση των πελατών της αυτό πιθανότατα σημαίνει ότι διαθέτει τους πόρους και είναι πρόθυμη να εφαρμόσει μία εξαιρετικά απαιτητική εκστρατεία μάρκετινγκ για την μεγαλύτερη ομάδα πελατών της, τους "new customers" (57,42% του συνόλου των πελατών). Οι νέοι πελάτες χαρακτηρίζονται από πρόσφατες αγορές (υψηλό R), χαμηλή συχνότητα αγορών (χαμηλό F), και μέτριου ύψους χρηματικές δαπάνες (μέτριο M). Στόχος της επιχείρησης είναι να προβιβάσει τους πελάτες αυτού του τμήματος σε "promising" ή και "potential loyalists" και προς επίτευξη του εν λόγω σκοπού απαιτούνται ενέργειες μάρκετινγκ οι οποίες θα την κάνουν να ξεχωρίσει στα μάτια τους σε σχέση με τον ανταγωνισμό. Για να αυξήσει την συχνότητα των αγορών τους μπορεί να τους παρέχει εκπτωτικά κουπόνια για μελλοντικές προσφορές ή μικρά δώρα με την ολοκλήρωση της πρώτης τους αγοράς. Επιπλέον η συμμετοχή τους σε μια έρευνα ικανοποίησης μέσω τηλεφώνου ή ιστοσελίδας όπου θα μπορέσουν να αξιολογήσουν τα προϊόντα, μπορεί να τους κινητοποιήσει σημαντικά προς μελλοντικές αγορές και να τους κάνει

να «συνδεθούν» με το brand της επιχείρησης. Όσον αφορά τους "hibernating" πελάτες της που αποτελούν το 2ο μεγαλύτερο segment σύμφωνα με τη μέθοδο column k-means (17,19% των πελατών), συνήθως δεν συνίσταται να επενδύονται πόροι και προσπάθειες προσέγγισης της συγκεκριμένης ομάδας πελατών, καθώς έχει δείξει απροθυμία σε τέτοιου είδους προσπάθειες στο παρελθόν. Αυτοί οι πελάτες είναι οι πελάτες με χαμηλή συχνότητα αγορών, χαμηλές δαπάνες και μη πρόσφατες αγορές. Τέλος, ιδιαίτερο χειρισμό απαιτεί το 3ο σε μέγεθος segment (11,68% των πελατών), οι πελάτες που χαρακτηρίζονται ως "promising". Οι πελάτες αυτού του τμήματος παρά το γεγονός ότι χαρακτηρίζονται από πρόσφατες αγορές και υψηλές χρηματικές δαπάνες, δεν πραγματοποιούν αγορές σε τακτά χρονικά διαστήματα. Για να αυξηθεί λοιπόν η συχνότητα των αγορών τους (F), η επιχείρηση μπορεί να τους δελεάσει παρέχοντας προτάσεις για παρεμφερή προϊόντα, στέλνοντάς τους μικρά δώρα με κάθε αγορά και ενθαρρύνοντάς τους να κάνουν αξιολόγηση των προϊόντων/υπηρεσιών της επιχείρησης.

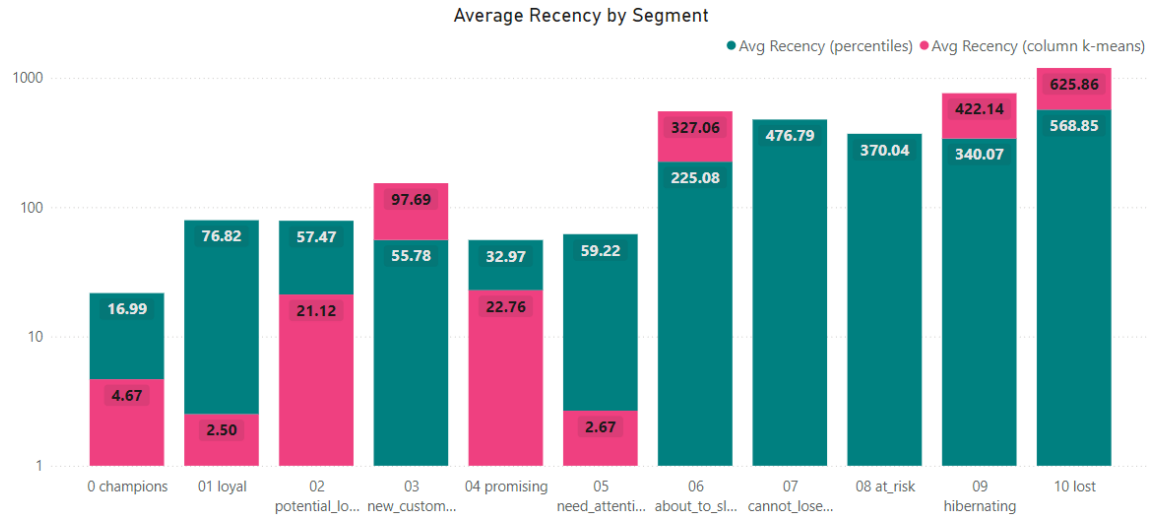
Σε αντίθεση, αν η επιχείρηση αποφασίσει να υιοθετήσει τη μέθοδο **percentiles** για την τμηματοποίηση των πελατών της και τον σχεδιασμό της στρατηγικής μάρκετινγκ, θα πρέπει να στοχεύσει ίσως σε περισσότερα segments και με διαφορετικά είδη ενεργειών, καθώς οι πελάτες είναι πιο ομοιόμορφα κατανεμημένοι στα διάφορα τμήματα. Το μεγαλύτερο σε πλήθος segment που προκύπτει από την εν λόγω μέθοδο είναι οι "hibernating" πελάτες (19,37% των πελατών), ενώ ακολουθούν οι "champions" (19,02% των πελατών) και οι "potential loyalists" (11,01% των πελατών). Όπως προαναφέρθηκε, δεν συνηθίζεται να δαπανώνται πόροι και προσπάθειες για την προσέγγιση των "hibernating" πελατών καθώς παραμένουν απρόθυμοι να επιστρέψουν στην επιχείρηση. Αξίζει όμως να δοθεί έμφαση στην κατηγορία "champions" επιβραβεύοντάς τους καθώς οι πελάτες αυτοί είναι υπεύθυνοι για ένα μεγάλο μερίδιο των εσόδων της αλλά και για την διαφήμιση των προϊόντων της μέσω θετικών σχολίων (positive word of mouth). Η επιχείρηση μπορεί να τους προσφέρει αποκλειστικές προσφορές και προτεραιότητα στη διάθεση νέων προϊόντων. Η επιχείρηση θα πρέπει επίσης να εστιάσει στην προσέγγιση των potential loyalists και συγκεκριμένα στην παρότρυνσή τους να δαπανήσουν περισσότερα χρήματα στα προϊόντα της επιχείρησης, καθώς οι μεταβλητές R και F είναι ήδη υψηλές σε αυτή την κατηγορία των πελατών. Οι προσπάθειες μάρκετινγκ θα πρέπει να επικεντρωθούν σε προτάσεις για συμπληρωματικά ή επιπλέον προϊόντα αλλά και ενέργειες που θα τους κάνουν να νιώσουν πολύτιμοι και να επομένως να αυξήσουν την αφοσίωσή τους στην επιχείρηση.

Στα radar charts της Εικόνας 52 απεικονίζονται οι ανά τμήμα μέσες τιμές των μετρικών Recency, Frequency και Monetary των δύο μεθόδων ως κλάσματα των εκάστοτε μέγιστων

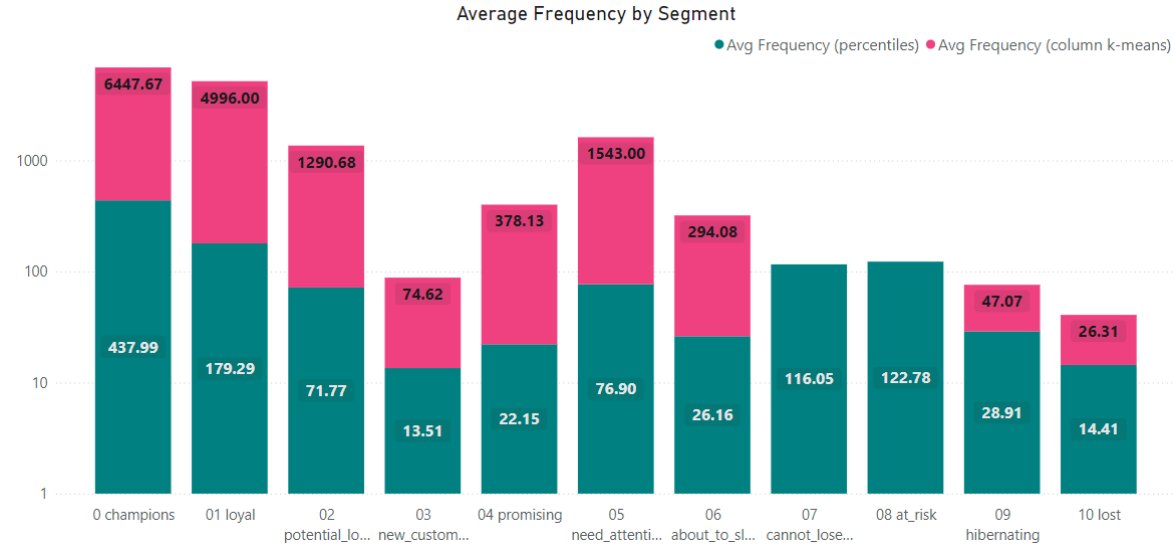
τιμών. Με άλλα λόγια για κάθε τμήμα υπολογίστηκαν οι μέσες τιμές των μετρικών για κάθε μέθοδο και η εκάστοτε μέγιστη τέθηκε ως το 100% επί του οποίου απεικονίζεται η τιμή της μεθόδου που κατέγραψε το ελάχιστο. Εκτός από τις κατηγορίες “hibernating” και “lost” παρατηρούμε και εδώ, με αυτόν τον τρόπο, τις αναντιστοιχίες στην κατανομή των πελατών της κάθε μεθόδου, ποσοτικοποιημένες ανά μετρική ενδιαφέροντος. Στις Εικόνες 53, 54 και 55 παρατίθενται σε ραβδογράμματα οι μέσες τιμές των εν λόγω μετρικών ανά τμήμα και ανά μέθοδο, σε λογαριθμική κλίμακα, για μεγαλύτερη λεπτομέρεια. Αυτό που παρατηρούμε και στα δύο είδη διαγραμμάτων είναι ότι οι μέσες τιμές των μετρικών Frequency και Monetary είναι με συνέπεια μεγαλύτερες για κάθε segment που έχει προκύψει από τον αλγόριθμο column k-means σε σχέση με αυτές της μεθόδου percentiles. Αντιθέτως, όσον αφορά στη μετρική Recency οι μέσες τιμές ανά μέθοδο δεν εμφανίζουν κάποια ξεκάθαρη σχέση μεταξύ τους. Τα διαγράμματα της Εικόνα 56 εμβαθύνουν σε μεγαλύτερη λεπτομέρεια στην κατανομή των τιμών των τριών μετρικών R, F και M, απεικονίζοντας με τη μορφή διαγραμμάτων boxplots τα αποτελέσματα των δύο μεθόδων ανά segment.



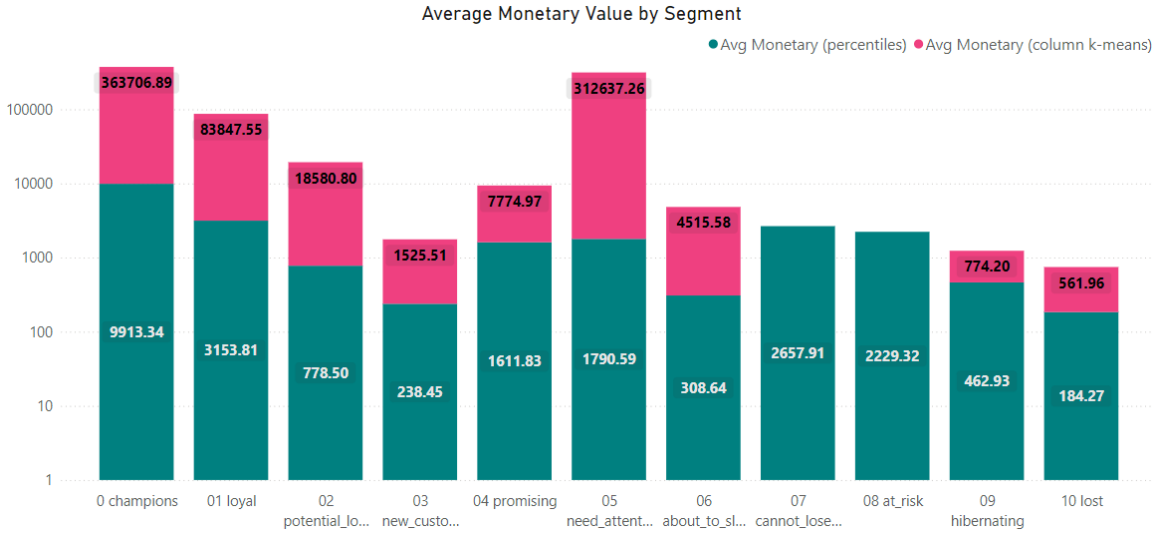
Εικόνα 52. Radar charts με τις ανά τμήμα μέσες τιμές των μετρικών R, F και M των δύο μεθόδων ως κλάσματα των εκάστοτε μέγιστων τιμών



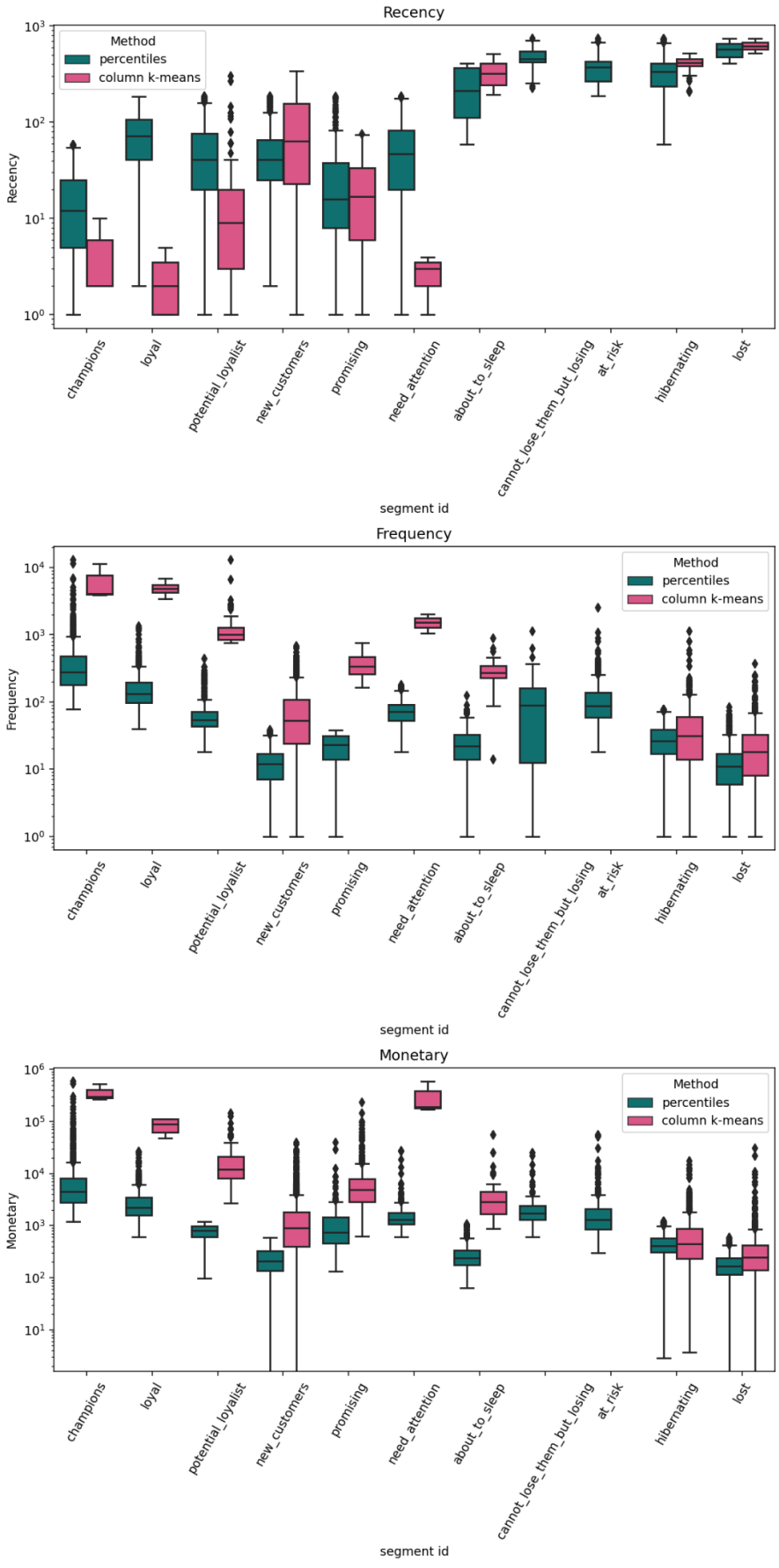
Εικόνα 55. Bar chart μέσων τιμών της μετρικής R ανά τμήμα και ανά μέθοδο (σε λογαριθμική κλίμακα)



Εικόνα 54. Bar chart μέσων τιμών της μετρικής F ανά τμήμα και ανά μέθοδο (σε λογαριθμική κλίμακα)



Εικόνα 53. Bar chart μέσων τιμών της μετρικής M ανά τμήμα και ανά μέθοδο (σε λογαριθμική κλίμακα)



Εικόνα 56. Boxplots που απεικονίζουν την κατανομή των τιμών των R, F και M ανά segment και ανά μέθοδο

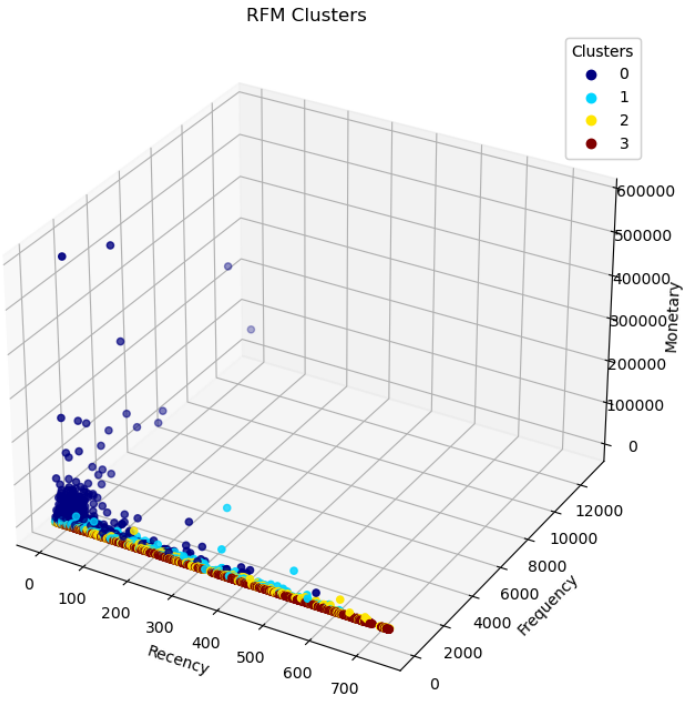
4.2 Αποτελέσματα κατηγοριοποίησης πελατών μέσω K-means απευθείας στον τρισδιάστατο RFM χώρο

Στην Εικόνα 58 απεικονίζονται οι μέσες τιμές των τριών μετρικών Recency, Frequency και Monetary ανά συστάδα για την επιλεγμένη τιμή k=4, και η πληθικότητα της κάθε συστάδας σε πελάτες. Στην Εικόνα 57 απεικονίζεται η κατανομή των πελατών στις 4 συστάδες.

Out[42]:

	Recency	Frequency	Monetary	
	mean	mean	mean	count
Cluster				
0	34.0	388.0	8802.0	1401
1	129.0	105.0	1615.0	1545
2	249.0	46.0	912.0	1558
3	396.0	16.0	258.0	1344

Εικόνα 58. Μέσες τιμές των μετρικών R, F και M και αριθμός πελατών σε κάθε συστάδα για k=4



Εικόνα 57. Κατανομή πελατών στις 4 συστάδες

Όπως και για k=5 παρατηρούμε πως, ξεκινώντας από τη συστάδα 0, κάθε διαδοχική συστάδα έχει μειωμένη συχνότητα και λιγότερες χρηματικές δαπάνες σε σχέση με την προηγούμενή

της, και πιο πρόσφατες αγορές. Στην προκειμένη, για k=4, η διάρθρωση των συστάδων δεν εμφανίζει κάποια ευκρινή συσχέτιση με την διάρθρωση των segments που είδαμε στην προηγούμενη ενότητα. Αυτό ίσως είναι αναμενόμενο δεδομένου του γεγονότος πως οι τέσσερις συστάδες εμφανίζουν μία γραμμική κλιμάκωση αξίας πελατών (στο cluster 3 περιλαμβάνονται πελάτες με την μικρότερη αξία για την επιχείρηση, στο cluster 2 με λίγο μεγαλύτερη, στο cluster 1 με μεγαλύτερη αξία από αυτούς στο cluster 2, και τέλος στο cluster 0 εμφανίζονται οι πελάτες με τη μέγιστη αξία), ενώ η κατηγοριοποίηση σε segments ενέχει κάποια λεπτά νοήματα τα οποία δεν βρίσκουν όλα τα ομόλογά τους στη συσταδοποίηση του παρόντος κεφαλαίου (για παράδειγμα, ποια θα μπορούσαν να είναι τα αντίστοιχα clusters των segments “at risk”, “hibernating” ή “loyal”;). Μπορούμε να δούμε αυτή την δυσκολία μετάφρασης στις επόμενες δύο εικόνες, 59 και 60, οι οποίες δείχνουν τη σύνθεση των segments των δύο μεθόδων της προηγούμενης ενότητας ως προς τα clusters της παρούσας ενότητας. Εδώ βλέπουμε πως τα πράγματα είναι σχεδόν ξεκάθαρα για τα segments που βρίσκονται στις άκρες της αξίας, δηλαδή για τα segments “champions” και “lost”, τα οποία αποτελούνται σε ποσοστό άνω του 80% από τα αντίστοιχα clusters μέγιστης και ελάχιστης αξίας, όμως τα segments ενδιάμεσων αξιών αποτελούνται σε διάφορα ποσοστά από σχεδόν όλα τα clusters, και ειδικά τα segments που προέκυψαν από τη μέθοδο percentiles.

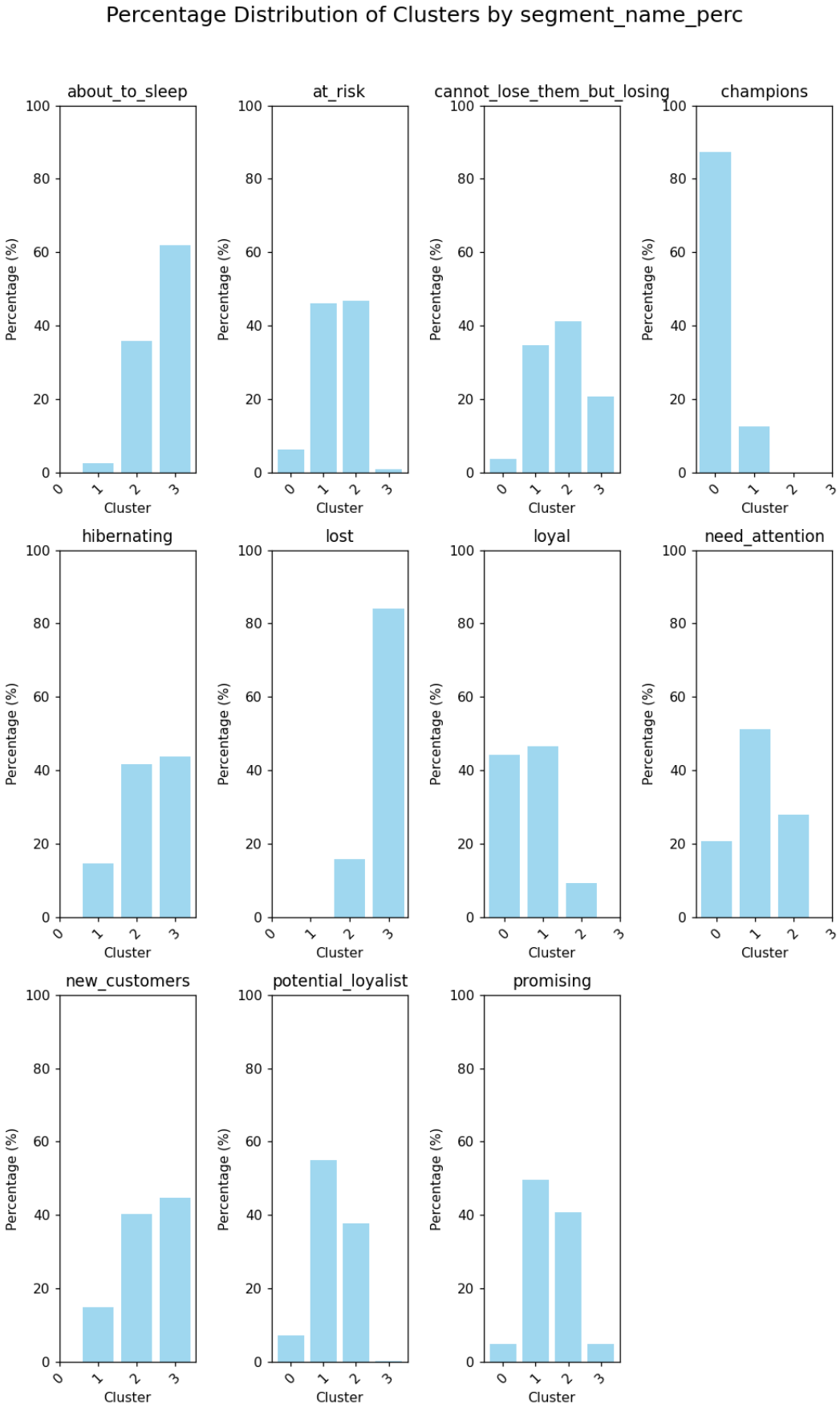
Η αδυναμία απευθείας μετάφρασης των segments των δύο πρώτων μεθόδων σε clusters της τελευταίας δημιουργεί την ανάγκη ενός κοινού συστήματος ομαδοποίησης. Για τον σκοπό της σύγκρισης των τριών μεθόδων κρίναμε σκόπιμη την ομαδοποίηση των 11 segments των δύο πρώτων μεθόδων (percentiles και column k-means) σε 4 κατηγορίες πελατών. Αυτές τις ονομάσαμε ομάδες αξίας ώστε να μπορούν να συγκριθούν με τα 4 clusters που προέκυψαν από τη μέθοδο K-means, με βάση την αξία των πελατών για την επιχείρηση. Ο Πίνακας 1 δείχνει την ομαδοποίηση των segments σε ομάδες αξίας σε φθίνουσα σειρά.

Πίνακας 1. Ομαδοποίηση των segments σε ομάδες αξίας

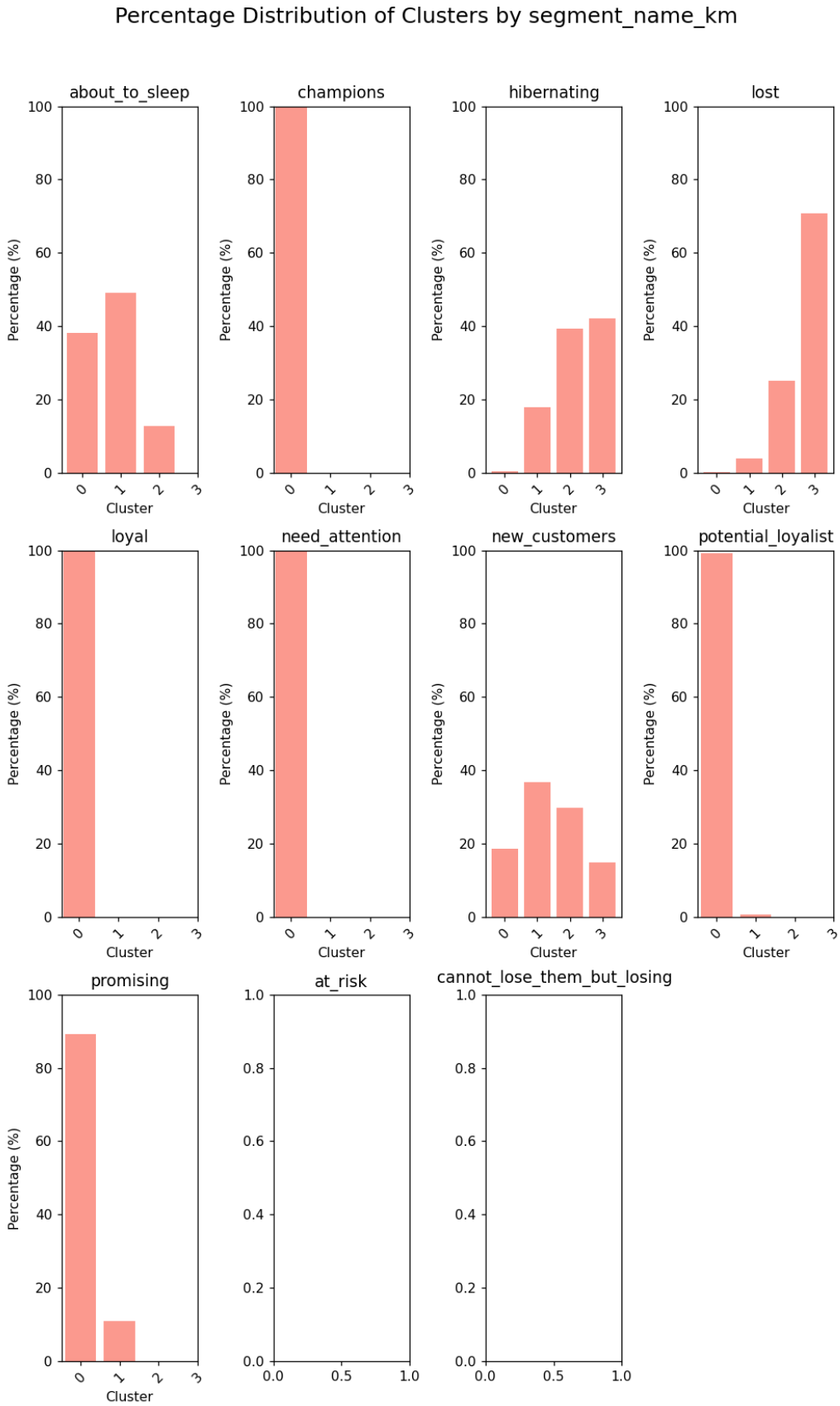
Ομάδα αξίας	**Τμήματα πελατών**			
high_value	Champions	Loyal		
mid_high_value	Potential loyalists	New customers	Promising	Need attention
mid_low_value	About to sleep	Can’t lose them but losing	At risk	
low_value	Hibernating	Lost		

Η πρώτη ομάδα αξίας περιλαμβάνει τους “champions” και τους “loyal”, δηλαδή τους πιο πολύτιμους πελάτες για την επιχείρηση, οι οποίοι είναι υπεύθυνοι για ένα μεγάλο μερίδιο της κερδοφορίας της. Η δεύτερη ομάδα αξίας περιλαμβάνει τους “potential loyalists”, τους “new customers”, τους “promising” και τους “need attention”, πελάτες που έχουν την αμέσως μεγαλύτερη αξία για την επιχείρηση και τους οποίους είναι σημαντικό να κρατήσει στο πελατολόγιό της μέσα από συνεχείς προσπάθειες ενθάρρυνσής τους για αύξηση των αγορών τους. Είναι σημαντικό η προσέγγιση μάρκετινγκ που ακολουθείται για αυτή την κατηγορία πελατών να στοχεύει στη δημιουργία, ή εφόσον ήδη υφίσταται, στην όξυνση του αισθήματος σύνδεσης που νιώθουν οι πελάτες σε σχέση με την επιχείρηση. Στην τρίτη ομάδα περιλαμβάνονται οι πελάτες που έχουν μικρότερη αξία για την επιχείρηση από ότι αυτοί της δεύτερης, δηλαδή οι “about to sleep”, οι “can’t lose them but losing” και οι “at risk”. Η προσέγγιση αυτής της κατηγορίας πελατών είναι πιο δύσκολη για την επιχείρηση, η οποία οφείλει να διερευνήσει τους παράγοντες που κάνουν τους πελάτες της να απέχουν και έπειτα να εφαρμόσει τις τακτικές μάρκετινγκ που θα τους παρακινήσουν να επιστρέψουν σε εκείνη. Τέλος η τέταρτη κατηγορία πελατών περιέχει τους “hibernating” και “lost”, τους πελάτες που ουσιαστικά δεν ανήκουν πια στο ενεργό πελατολόγιο της επιχείρησης και που οποιαδήποτε προσπάθεια εκ νέου προσέγγισής τους θα απέβαινε μάταιη.

Με την ομαδοποίηση των segments στις παραπάνω τέσσερις ομάδες αξίας, οι οποίες έχουν ανάλογη κλιμάκωση με τα τέσσερα clusters που προέκυψαν από την τρίτη μέθοδο, είναι τώρα δυνατή η σύγκριση των τριών μεθόδων μεταξύ τους. Αυτή ακολουθεί στην επόμενη ενότητα.



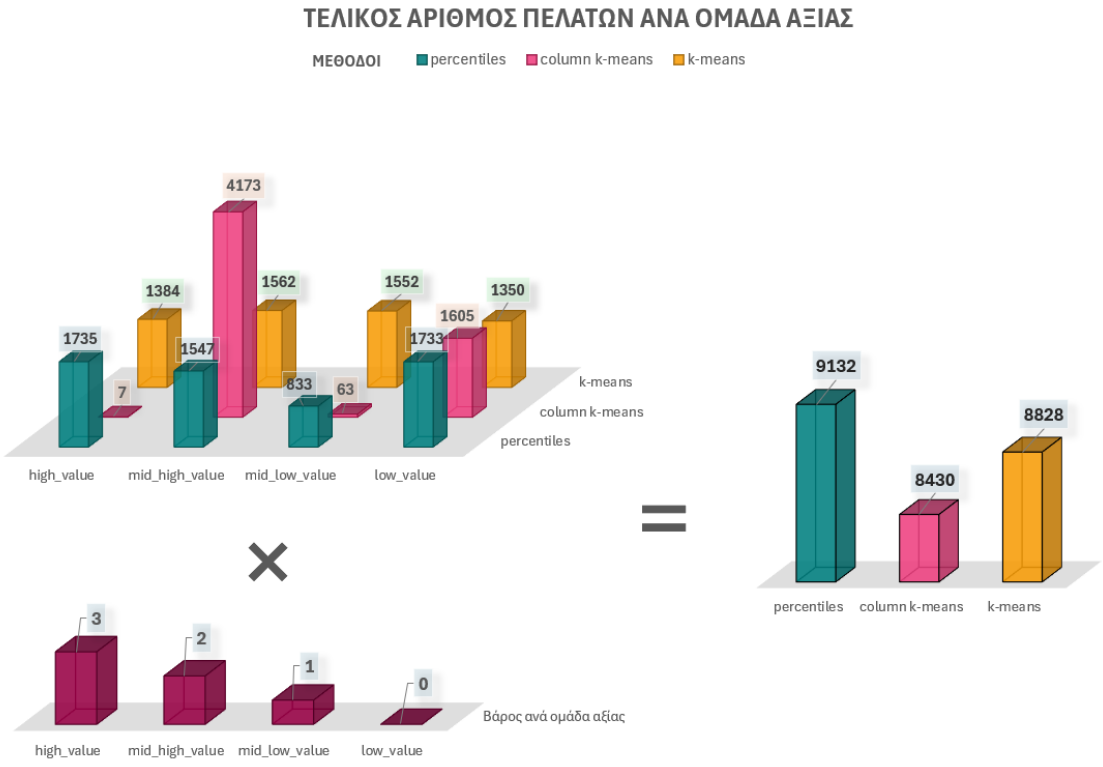
Εικόνα 59. Σύνθεση των segments της μεθόδου percentiles ως προς τα clusters της μεθόδου K-means



Εικόνα 60. Σύνθεση των segments της μεθόδου column k-means ως προς τα clusters της μεθόδου K-means

4.3 Σύγκριση αποτελεσμάτων των τριών μεθόδων

Η Εικόνα 61 αποτελείται από τρία σχήματα. Το σχήμα στην πάνω αριστερά πλευρά απεικονίζει τον αριθμό πελατών ανά ομάδα αξίας για κάθε μία από τις τρεις μεθόδους που παρουσιάστηκαν στην εργασία. Το ζητούμενο σε αυτή την ενότητα είναι η σύγκριση των τριών μεθόδων με βάση τον αριθμό των πελατών ανά ομάδα αξίας. Προς αυτό τον σκοπό, για την εξαγωγή της τελικής ιεραρχίας, δεδομένης της ίδιας κλιμάκωσης της αξίας των πελατών των τεσσάρων ομάδων, χρησιμοποιούμε βάρη με τιμή ανάλογη της θέσης της κάθε ομάδας στην ιεραρχία αξίας που αυτές σχηματίζουν με τον εξής τρόπο: πολλαπλασιάζουμε το βάρος της θέσης της κάθε ομάδας με τον αριθμό των πελατών της, και αθροίζουμε τα τέσσερα γινόμενα για κάθε μέθοδο. Αυτή η μεθοδολογία κατάταξης αποτελεί μία προσαρμοσμένη και συστηματική σταθμισμένη παραλλαγή της μεθοδολογίας κατάταξης RFM μεθόδων που βρίσκεται στην εργασία (Λαζαρίδου, 2023). Το κάτω αριστερά σχήμα στην Εικόνα 61 απεικονίζει τις τιμές των βαρών που χρησιμοποιήθηκαν, και το σχήμα στα δεξιά τις τιμές που προέκυψαν από την παραπάνω μεθοδολογία.



Εικόνα 61. Τελικός αριθμός πελατών ανά ομάδα αξίας

Σύμφωνα με τα αποτελέσματα η μέθοδος percentiles κατατάσσεται πρώτη, ακολουθούμενη από τη μέθοδο k-means, ενώ στην τελευταία θέση κατατάσσεται η μέθοδος column k-means. Η σειρά κατάταξης είναι αναμενόμενη δεδομένης της πληθικότητας των ομάδων αξίας high_value των τριών μεθόδων (1735, 1384, 7 - πρώτη στήλη του σχήματος πάνω αριστερά στην Εικόνα 61).

Κεφάλαιο 5: Συμπεράσματα

5.1 Συμπεράσματα έρευνας

Στην εργασία αναπτύχθηκαν τρεις μέθοδοι τμηματοποίησης πελατών ενός ηλεκτρονικού καταστήματος με είδη δώρων, βασισμένες σε ένα σύνολο δεδομένων χρονικού διαστήματος δύο ετών: η τυπική μέθοδος RFM (percentiles – ενότητα 3.3.2.1), η μέθοδος συσταδοποίησης ανά στήλη column k-means (ενότητα 3.3.2.2), και ο αλγόριθμος συσταδοποίησης K-means στον τρισδιάστατο RFM χώρο (ενότητα 3.3.2.3). Η σύγκριση των τριών μεθόδων έγινε στην κοινή βάση τμηματοποίησης πελατών σε τέσσερις ομάδες φθίνουσας αξίας – και συνεπώς κερδοφορίας – για την επιχείρηση. Σύμφωνα με τα αποτελέσματα η μέθοδος percentiles κατατάχθηκε πρώτη, ακολουθούμενη από τον αλγόριθμο K-means, ενώ η column k-means κατατάχθηκε τελευταία. Ο Πίνακας 2 συνοψίζει την τμηματοποίηση των πελατών της επιχείρησης για το εν λόγω χρονικό διάστημα με βάση την προτεινόμενη μέθοδο, τη μέθοδο percentiles.

Πίνακας 2. Ποσοστά πελατών ανά ομάδα αξίας κατανεμημένα σε φθίνουσα σειρά

Ομάδα αξίας 1		**Ομάδα αξίας 2**				**Ομάδα αξίας 3**			**Ομάδα αξίας 4**	
29.67%		**26.45%**				**14.24%**			**29.63%**	
Champions	**Loyal**	**Potential loyalists**	**New customers**	**Promising**	**Need attention**	**About to sleep**	**Can't lose them**	**At risk**	**Hibernating**	**Lost**
19.02%	**10.65%**	**11.01%**	**6.79%**	**3.52%**	**5.13%**	**3.54%**	**1.83%**	**8.87%**	**19.37%**	**10.26%**

Κατά βάση, τα αποτελέσματα της σειράς κατάταξης εξηγούνται από τη διαφορά στον αριθμό των πελατών υψηλής αξίας που προέκυψαν από κάθε μέθοδο. Συγκεκριμένα, η μέθοδος column k-means (#3) κατέταξε πολύ λιγότερους πελάτες στα τμήματα υψηλής αξίας ("champions", "loyal") σε σύγκριση με την percentiles (#1 – 7 έναντι 1 735), αν και οι δύο μέθοδοι συμφώνησαν ως προς τον αριθμό των πελατών των τμημάτων χαμηλής αξίας ("hibernating και "lost" – 1 733 και 1 605). Η διαφορά στην κατανομή πελατών σε τμήματα ήταν σημαντική: ενδεικτικό παράδειγμα αποτελεί η ένταξη περίπου του 60% των πελατών της μεθόδου column k-means στο τμήμα "new customers", σε αντίθεση με την μέθοδο percentiles,

κατά την οποία οι αριθμοί πελατών ανά τμήμα εμφάνισαν συνολικά μικρότερη διακύμανση. Παρόλα αυτά, για αυτή τη μέθοδο, περισσότεροι πελάτες κατανεμήθηκαν σε ομάδες υψηλής ή χαμηλής αξίας (3 468 πελάτες – 60% του συνόλου) παρά στις ομάδες μεσαίας αξίας (2 380 πελάτες – 40% του συνόλου), ενώ για τον αλγόριθμο K-means (#2) κάθε ομάδα είχε σχεδόν το ίδιο μέγεθος.

5.2 Συμπεράσματα σε σχέση με τα αποτελέσματα της βιβλιογραφίας

Η παρούσα εργασία αποτελεί επέκταση της μεταπτυχιακής εργασίας Λαζαρίδου (2023). Η εν λόγω εργασία εξέτασε τις συναλλαγές της ίδιας επιχείρησης που εξετάστηκαν στην παρούσα εργασία, αλλά για το διάστημα 12.2010 έως 12.2011. Σε αυτήν αρχικά εφαρμόστηκε η ανάλυση RFM, από όπου προέκυψε η τμηματοποίηση των πελατών σε έξι τμήματα. Στη συνέχεια, ως ξεχωριστή μέθοδος τμηματοποίησης, με την χρήση του αλγορίθμου K-means παρήχθησαν τρεις συστάδες πελατών (clusters). Για την επιλογή εκείνης της μεθόδου τμηματοποίησης που βελτιώνει την κερδοφορία της επιχείρησης, ελέγχθηκε ποια από τις δύο μεθόδους συγκεντρώνει το μεγαλύτερο ποσοστό πελατών σε τμήματα με μέτρια ή υψηλή αξία για την επιχείρηση. Η εργασία κατέληξε στην πρόταση της τμηματοποίησης των πελατών με βάση τα έξι τμήματα που προέκυψαν από την απλή ανάλυση RFM.

Στην παρούσα έρευνα εφαρμόστηκε αντίστοιχα η ανάλυση RFM για το διάστημα 12.2009 έως 12.2011 (δηλαδή για το διπλάσιο χρονικό διάστημα), με τους πελάτες να τμηματοποιούνται σε έντεκα αντί για έξι τμήματα, όπως προτάθηκε στην έρευνα των Cuce & Tïryakï (2022). Ακόμη, από την εφαρμογή του αλγορίθμου K-means, με βάση το συγκεκριμένο σύνολο δεδομένων που χρησιμοποιήσαμε, και τα αποτελέσματα της μεθόδου elbow, προέκυψαν τέσσερις συστάδες πελατών (clusters). Επιπλέον, παρουσιάστηκε και μία τρίτη μέθοδος, η μέθοδος συσταδοποίησης ανά στήλη (*column k-means*), κατά την οποία εφαρμόστηκε ο αλγόριθμος K-means σε καθεμία από τις στήλες Recency, Frequency και Monetary για την παραγωγή των RFM scores των πελατών. Για την επιλογή της μεθόδου τμηματοποίησης που οδηγεί στη βελτίωση της κερδοφορίας της επιχείρησης, η εργασία της Λαζαρίδου επέλεξε ως κριτήριο κατάταξης των δύο μεθόδων τον αριθμό των πελατών του αποκλειστικά πιο επικερδούς τμήματος, απορρίπτοντας από τη σύγκριση τη συμβολή των υπολοίπων. Σε αντίθεση, στην παρούσα εργασία χρησιμοποιήθηκαν ειδικά βάρη ανά ομάδα αξίας ανάλογα με το ύψος της αξίας των πελατών τους για την επιχείρηση. Η σύγκριση των τριών μεθόδων έδειξε ότι η παραδοσιακή μέθοδος RFM – η οποία αναφέρεται, χάριν ευκολίας στην εργασία ως *percentiles*

–, όπως και στην περίπτωση της Λαζαρίδου, κρίνεται πρώτη στην σειρά κατάταξης από τις άλλες δύο.

Με βάση τα παραπάνω θεωρούμε ότι τα αποτελέσματα της ανάλυσής μας είναι συνεπή με τα αποτελέσματα της βιβλιογραφίας. Για αυτό το λόγο είναι ίσως εύλογο να υποθέσουμε ότι η φύση της επιχειρηματικής δραστηριότητας της συγκεκριμένης επιχείρησης μπορεί να κάνει τους πελάτες να διαρθρώνονται διαχρονικά βέλτιστα με τη μέθοδο percentiles. Σε κάθε περίπτωση η επιχείρηση θα πρέπει να διερευνά περιοδικά την κατάσταση της αγοράς λόγω του αυξανόμενου όγκου πληροφοριών που την επηρεάζει, καθώς η ίδια ανάλυση ενδέχεται στο μέλλον να καταλήξει στην πρόκριση μιας από τις άλλες δύο μεθόδους ή κάποιας που δεν έχει εξεταστεί στην παρούσα εργασία.

5.3 Προτάσεις προς τους υπεύθυνους λήψης των αποφάσεων

Σύμφωνα με τα αποτελέσματα που προέκυψαν από την ανάλυση που προηγήθηκε, η παρούσα εργασία προτείνει στην επιχείρηση τη μέθοδο percentiles για την βελτιστοποίηση της κερδοφορίας της (ενότητα 3.3.2.1). Η επιχείρηση μπορεί να επιλέξει να σχεδιάσει τις στρατηγικές μάρκετινγκ σύμφωνα με τις τέσσερις ευρύτερες ομάδες αξίας που παρουσιάστηκαν στην ανάλυση (ενότητα 4.2) ή σύμφωνα με τα έντεκα τμήματα της μεθόδου percentiles (ενότητα 4.1). Εφόσον διαθέτει επαρκείς πόρους – χρηματικούς και χρονικούς - μπορεί να απευθυνθεί ξεχωριστά σε κάθε μία από τις έντεκα εδραιωμένες στην βιβλιογραφία κατηγορίες πελατών, σχεδιάζοντας εξειδικευμένες στρατηγικές μάρκετινγκ για την προσέγγιση της κάθε μίας. Σε διαφορετική περίπτωση, ελλείψει πόρων και ευρηματικότητας, η επιχείρηση μπορεί να χρησιμοποιήσει την τμηματοποίηση σε πιο ευρείες ομάδες πελατών (ομάδες αξίας) για τον σχεδιασμό των στρατηγικών μάρκετινγκ. Η ανάπτυξη του σχεδίου μάρκετινγκ σε αυτή την περίπτωση θα έχει πιο γενικές κατευθύνσεις, καθώς θα θεωρείται ότι τμήματα όπως, για παράδειγμα, οι “potential loyalists” έχουν μικρό βαθμό διάκρισης από τους “promising” πελάτες ή τους “need attention”.

Στην πρώτη ομάδα αξίας περιλαμβάνονται οι “champions” και “loyal” πελάτες της επιχείρησης με ποσοστό 29,67% του συνόλου. Οι “champions” δεν είναι μόνο υπεύθυνοι για ένα μεγάλο μερίδιο των εσόδων της επιχείρησης, αλλά είναι και οι πρώτοι που θα υιοθετήσουν τα νέα προϊόντα που εκείνη θα λανσάρει. Οι στρατηγικές μάρκετινγκ θα πρέπει να είναι ξεχωριστές για αυτή την κατηγορία πελατών και να επικεντρώνονται στην επιβράβευσή τους.

Δεν μπορούν, για παράδειγμα, να προσεγγίζονται με την παραδοσιακή αποστολή των newsletters ή των φυλλαδίων προσφορών. Η επιχείρηση θα πρέπει να τους παρέχει αποκλειστικές προσφορές, προτεραιότητα στην εξυπηρέτηση και sneak peek σε νέα προϊόντα. Αντίστοιχα, για να μπορέσει να προβιβάσει τους "loyal" πελάτες στην ομάδα "champions", η επιχείρηση μπορεί να τους προωθήσει υψηλότερης αξίας προϊόντα (up-selling strategy), καθώς η συγκεκριμένη ομάδα πελατών παραδοσιακά ανταποκρίνεται θετικά στις καμπάνιες μάρκετινγκ. Συνολικά για τα δύο segments της πρώτης ομάδας αξίας θα πρέπει να δοθεί έμφαση στην ανάπτυξη εξατομικευμένων ενεργειών προσέγγισης, καθώς οι πελάτες αυτοί γνωρίζουν και οι ίδιοι ότι είναι ξεχωριστοί για την επιχείρηση και επομένως προσδοκούν την βέλτιστη δυνατή εξυπηρέτηση από αυτήν.

Η δεύτερη ομάδα αξίας περιλαμβάνει τους "potential loyalists", τους "new customers", τους "promising" και τους "need attention" πελάτες της επιχείρησης – ένα ετερογενές πλήθος πελατών – σε ποσοστό 26,45% του συνολικού πελατολογίου. Για τους "potential loyalists" η επιχείρηση θα πρέπει να εστιάσει στην βελτίωση του αισθήματος σύνδεσής τους με αυτήν, προσφέροντάς τους την ευκαιρία να ενταχθούν σε loyalty προγράμματα που μπορούν να συνδέονται για παράδειγμα είτε με συλλογή πόντων για μελλοντική εξαργύρωση είτε με την προσφορά μικρών εκπτώσεων κάθε φορά που εκείνοι θα προτείνουν το ηλεκτρονικό κατάστημα σε φίλους τους στα social media. Με αυτούς τους τρόπους ενθαρρύνεται η συγκεκριμένη ομάδα πελατών να επαναλάβει σε πιο άμεσο χρονικό διάστημα αγορές που ενδεχομένως να πραγματοποιούσε σε δεύτερο χρόνο. Όσον αφορά τους "new customers", καθώς είναι πελάτες που έκαναν πρόσφατα την αγορά «γνωριμίας» τους με την επιχείρηση, προτείνεται η αποστολή emails για την πληροφόρησή τους για την γκάμα προϊόντων που προσφέρονται καθώς και η προσφορά ενός μικρού δώρου κατά την ολοκλήρωση της πρώτης παραγγελίας τους ώστε να αποκτήσουν μία καλή πρώτη εντύπωση. Οι "promising" πελάτες της επιχείρησης θα μπορούσαν να προσεγγιστούν με παρόμοιο τρόπο με τους "new customers" με τη διαφορά ότι αυτό το τμήμα είναι πιο καλά πληροφορημένο για τα προϊόντα που προσφέρονται καθώς έχει αγοράσει ξανά στο παρελθόν από την επιχείρηση αλλά με μικρή συχνότητα και συνεπώς, χρειάζεται ίσως ένα κίνητρο για αυξήσει τις αγορές του. Αυτό το κίνητρο θα μπορούσε να είναι η προσφορά εκπτώσεων που πρέπει να εξαργυρωθούν σε σύντομο χρονικό διάστημα (limited-time discounts) ή η πρόταση οικονομικών συμπληρωματικών προϊόντων με κάθε αγορά. Για το τελευταίο τμήμα της δεύτερης ομάδας αξίας, τους "need attention", οι οποίοι υπήρξαν στο παρελθόν καλοί πελάτες αλλά δεν έχουν πρόσφατα πραγματοποιήσει αγορές από την επιχείρηση, προτείνεται η αποστολή

εξατομικευμένων emails που θα τους υπενθυμίσουν παλαιότερες αγορές τους και θα τους προτείνονται νέα προϊόντα. Μία καλή πρακτική θα ήταν επίσης και η προσφορά μιας ειδικής έκπτωσης ή δωρεάν αποστολής προϊόντων με την επανέναρξη των αγορών τους από το ηλεκτρονικό κατάστημα. Συνολικά, για τα τμήματα που ανήκουν στην δεύτερη ομάδα αξίας, η επιχείρηση θα πρέπει να εστιάσει στην προσφορά μικρών εκπτώσεων ή δώρων για να ενισχύσει το ήδη υπάρχον θετικό αίσθημα και τις καλές εντυπώσεις που έχουν οι πελάτες για την επιχείρηση οδηγώντας τους έτσι να αυξήσουν τις αγορές τους από το κατάστημά της.

Προχωρώντας στην τρίτη ομάδα αξίας, η οποία περιλαμβάνει τους "about to sleep", τους "can't lose them but losing" και τους "at risk" πελάτες, η επιχείρηση θα πρέπει να επικεντρωθεί αρχικά στην έρευνα των παραγόντων που οδήγησαν αυτά τα τμήματα πελατών να απομακρυνθούν από εκείνη. Τα τρία αυτά τμήματα αποτελούν το 10,82% του συνολικού πελατολογίου της και το κοινό τους χαρακτηριστικό είναι περιέχουν πελάτες οι οποίοι στο παρελθόν υπήρξαν είχαν μεγάλη αξία για την επιχείρηση, ενώ έχουν πλέον μειώσει σημαντικά τις αγορές τους από αυτήν. Για τους "about to sleep" προτείνεται η εφαρμογή τακτικών που θα τους δημιουργήσουν το αίσθημα του «επείγοντος» ώστε να παρακινηθούν να αγοράσουν ξανά άμεσα, όπως η πληροφόρησή τους για προϊόντα με περιορισμένη διαθεσιμότητα ή η παροχή προϊόντων σε bundles για περιορισμένο χρονικό διάστημα. Για τους "can't lose them but losing" και τους "at risk" πελάτες, ενέργειες όπως η αποστολή "we miss you" emails προτείνεται σε συνδυασμό με την προσφορά μιας δυνατής έκπτωσης ώστε να αναζωπυρώσει το ενδιαφέρον και τη δέσμευση που αισθάνονταν προηγουμένως με την επιχείρηση ώστε τελικά να επανέλθουν σε αυτή. Προτείνεται επίσης η υπενθύμισή τους σχετικά με τους συνολικούς πόντους που έχουν συγκεντρώσει – καθώς ανήκουν συνήθως σε loyalty προγράμματα – και τα πιθανά οφέλη που θα αποκομίσουν με την επανέναρξη των αγορών τους. Είναι κρίσιμο, λοιπόν, για τα τμήματα της τρίτης ομάδας αξίας η επιχείρηση να ακολουθήσει τακτικές που επιδεικνύουν την εκτίμησή της για τις προηγούμενες αλληλεπιδράσεις των πελατών, ενθαρρύνοντάς τους να γίνουν ξανά ενεργοί.

Η τέταρτη και τελευταία ομάδα αξίας πελατών αποτελείται από τους "hibernating" και τους "lost" πελάτες με ποσοστά 19,37% και 10,26% του συνολικού πελατολογίου αντίστοιχα. Για τα δύο αυτά τμήματα δεν προτείνεται να δαπανηθούν πόροι για την προσέγγιση των πελατών τους, καθώς η εμπειρία έχει δείξει ότι πρόκειται για πελάτες που εμφανίζονται απρόθυμοι στις καμπάνιες μάρκετινγκ που έχουν προηγηθεί. Παρόλα αυτά, θα ήταν χρήσιμο για την επιχείρηση να πληροφορηθεί για τις αιτίες που οδήγησαν στην παύση των αγορών από αυτούς τους πελάτες, ώστε αν πρόκειται για δικές της αστοχίες να μπορεί να τις εντοπίσει και να τις

αποφύγει στο μέλλον. Προτείνεται, συνεπώς, η αποστολή email στους πελάτες αυτών των τμημάτων για την συμπλήρωση ενός πολύ σύντομου ερωτηματολογίου με την δυνατότητα ελεύθερου κειμένου για όσους μπορεί να επιθυμούν να εκφραστούν πιο αναλυτικά σε σχέση με την εμπειρία τους. Σαφώς, θα πρέπει να αναμένεται ότι πολύ μικρό ποσοστό των πελατών θα ανταποκριθούν ακόμη και σε αυτό το κάλεσμα.

Κεφάλαιο 6: Περιορισμοί και προτάσεις για μελλοντική έρευνα

6.1 Περιορισμοί έρευνας

Η παρούσα εργασία εξέτασε **ένα** σύνολο δεδομένων που αφορά σε **ένα** κατάστημα **μίας** επιχείρησης με είδη δώρων. Η αποτελεσματικότητα της RFM ανάλυσης μπορεί να διέφερε σημαντικά αν εφαρμοζόταν σε δεδομένα επιχειρήσεων άλλων βιομηχανιών ή σε δεδομένα διαφορετικού καταστήματος της ίδιας επιχείρησης ή ακόμα και σε σύνολο δεδομένων διαφορετικής χρονικής περιόδου (π.χ. προ covid-19 εποχής, μετά covid-19 εποχής).

Συγκεκριμένα πραγματοποιήθηκε τμηματοποίηση των πελατών βάσει της ημερομηνίας πραγματοποίησης συναλλαγών, του αριθμού των συναλλαγών και των συνολικού χρηματικού ποσού που αυτοί δαπάνησαν. Συνεπώς η ανάλυση που προηγήθηκε βασίστηκε αποκλειστικά σε αυτές τις τρεις μεταβλητές: αυτό σημαίνει πως δεν έλαβε υπ' όψιν άλλα σημαντικά χαρακτηριστικά, όπως δημογραφικά στοιχεία, γεωγραφικές μεταβλητές, την αξία διάρκειας ζωής του πελάτη (CLV) ή άλλα ποιοτικά χαρακτηριστικά που μπορούν να ποσοτικοποιηθούν.

Ένας άλλος περιορισμός της έρευνας αφορά τον αριθμό των τμημάτων (segments) που επιλέχθηκε για την κατηγοριοποίηση των πελατών με βάση τα RFM scores. Η ίδια ανάλυση θα μπορούσε να εξάγει διαφορετικά αποτελέσματα εάν αντί για έντεκα επιλεγόταν ένας διαφορετικός αριθμός τμημάτων. Αυτό θα μπορούσε να οδηγήσει και στην πρόκριση μιας διαφορετικής μεθόδου - πέραν της percentiles - ως κατάλληλης για την υλοποίηση της βέλτιστης τμηματοποίησης των πελατών.

Όσον αφορά στο εργαλείο που χρησιμοποιήθηκε για την οπτικοποίηση ορισμένων αποτελεσμάτων, δηλαδή το Power BI, αναφέρονται οι παρακάτω περιορισμοί:

- Παρά το γεγονός ότι προσφέρει την δυνατότητα ενσωμάτωσης με την Python για προηγμένες αναλύσεις, τα ενσωματωμένα εργαλεία του για συσταδοποίηση (π.χ. αλγόριθμος K-means) είναι περιορισμένα σε σχέση με εξειδικευμένα εργαλεία της επιστήμης των δεδομένων, όπως είναι η Python, η R ή το MATLAB. Κατά συνέπεια, αν και λειτούργησε αποτελεσματικά για την υλοποίηση της παραδοσιακής μεθόδου RFM, δεν κατέστη δυνατό να υλοποιηθεί και η μέθοδος K-means μέσω του Power BI.

- Το Power BI μπορεί να είναι φιλικό προς τον χρήστη, παρόλα αυτά μπορεί να περιορίσει κάποιον μη εξοικειωμένο με τη γλώσσα DAX (Data Analysis Expressions), η γνώση της οποίας είναι απαραίτητη για την πραγματοποίηση πολύπλοκων υπολογισμών.
- Ενώ το Power BI προσφέρει μια εκτενή σειρά επιλογών οπτικοποίησης, οι ιδιαίτερα προσαρμοσμένες ή εξειδικευμένες απεικονίσεις (π.χ. τα boxplots και τα radar charts της ενότητας 4.1) μπορεί να απαιτούν την χρήση εξωτερικών εργαλείων.

6.2 Προτάσεις για μελλοντική έρευνα

Στο στάδιο της περιγραφικής και διερευνητικής ανάλυσης του συνόλου δεδομένων παρατηρήθηκε ότι κατά τους μήνες Οκτώβριο, Νοέμβριο και Δεκέμβριο οι πωλήσεις υπήρξαν αυξημένες σε σύγκριση με τους υπόλοιπους μήνες. Θα ήταν ενδιαφέρον να εξεταστούν τα δεδομένα περισσότερων οικονομικών ετών για να διαπιστωθεί αν παρατηρείται το φαινόμενο της εποχικότητας (seasonality). Σε αυτή την περίπτωση, επιχείρηση θα όφειλε να λάβει υπ' όψιν της την εξωτερική αυτή μεταβλητή και να εξετάσει αν η εν λόγω αύξηση των πωλήσεων προέρχεται από τους ήδη πιστούς της πελάτες ή από νέους πελάτες, καθώς αυτή η πληροφορία είναι κρίσιμη για την ανάπτυξη των στρατηγικών μάρκετινγκ που θα εφαρμόσει για την προσέγγισή τους.

Όπως αναφέρθηκε στην ενότητα των περιορισμών της παρούσας έρευνας, ο αριθμός των τμημάτων που χρησιμοποιήθηκαν υπήρξε καθοριστικός για τα αποτελέσματα της ανάλυσης. Σε μελλοντική έρευνα προτείνεται να χρησιμοποιηθεί διαφορετικός αριθμός τμημάτων (ενδεχομένως μικρότερος των έντεκα) για την κατάταξη των RFM scores που προκύπτουν από τις μεθόδους percentiles και column k-means, ώστε να ελεγχθεί ποια μέθοδος φέρει καλύτερα αποτελέσματα για την επιχείρηση.

Η παρούσα εργασία εφάρμοσε την ανάλυση RFM στην παραδοσιακή της μορφή. Μία πρόταση για μελλοντική έρευνα αποτελεί η εξέταση προηγμένων παραλλαγών του μοντέλου RFM, ενσωματώνοντας διαστάσεις όπως η αξία διάρκειας ζωής του πελάτη (CLV) ή μετρικές όπως ο χρόνος που μεσολάβησε από την αρχική αγορά που πραγματοποίησε ο πελάτης από την επιχείρηση (Time since first purchase) και η τυπική απόκλιση των χρόνων μεταξύ των επισκέψεων του πελάτη (Periodicity).

Τέλος, αναφορικά με την χρήση του Power BI προτείνεται η ανάπτυξη διαφορετικών ταμπλό (dashboards) ή αναφορών (reports) που να απευθύνονται σε χρήστες με διαφορετικούς ρόλους

στην επιχείρηση. Οι εργαζόμενοι του τμήματος μάρκετινγκ, οι εργαζόμενοι του οικονομικού τμήματος και τα διοικητικά στελέχη που είναι υπεύθυνα για την λήψη των αποφάσεων αντιλαμβάνονται τις πληροφορίες με διαφορετικό τρόπο, δίνοντας μεγαλύτερη έμφαση στις ιδιαίτερες ανάγκες και προτεραιότητες του τμήματός τους. Για αυτόν το λόγο ενδέχεται να πρέπει να συμπεριληφθούν και άλλα δεδομένα σχετικά με τον ρόλο των τελικών χρηστών των παραγόμενων dashboards.

Βιβλιογραφία

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, 1142

Bačić, D., & Fadlalla, A. (2016). Business information visualization intellectual contributions: An integrative framework of visualization capabilities and dimensions of visual intelligence. *Decision Support Systems*, 89, 77-86.

Bacila, M. F., Radulescu, A., & Marar, I. L. (2012). RFM based segmentation: An analysis of a telecom company's customers. In *The Proceedings of the International Conference" Marketing-from Information to Decision"* (p. 52). Babes Bolyai University.

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. (1969). A New Product Growth for Model Consumer Durables. *Management Science*.

Bauer, H. H., & Hammerschmidt, M. (2005). Customer-based corporate valuation: Integrating the concepts of customer equity and shareholder value. *Management Decision*, 43(3), 331-348.

Blattberg, R. C., Kim, B. D., Neslin, S. A. (2008). RFM analysis. *Database Marketing: Analyzing and Managing Customers*, 323-337.

Brynjolfsson, E., & McAfee, A. (2017). The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. W.W. Norton & Company.

Cavusgil, S. T., Knight, G., & Riesenberger, J. (2004). International Business: The New Realities. Pearson.

Chandon, P., Morwitz, V. G., & Reinartz, W. J. (2005). Do intentions really predict behavior? Self-generated validity effects in survey research. *Journal of marketing*, 69(2), 1-14.

Chang, H. H., & Tsay, S. F. (2004). Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation. *Journal of Information Management*, Vol. 11 (4) (2004), pp. 161–203.

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2018). RFM Ranking – An Effective Approach to Customer Segmentation. Journal of King Saud University - Computer and Information Sciences.

Cuce, Aylanur & Tiryaki, Eda. (2022). Data Analytics in Customer Segmentation and RFM Method.

Demby, E. H. (1974). Psychographics and from whence it came. Journal of Consumer Research.

Devarapalli, Dharmaiah & Virajitha, Ayinavilli & Geddam, Sai Veera Venkata Satya Sunanda & Sunanda, Satya & Sravya, Amudalapalli & Keerthi, Boddu & Devi, Allada. (2022). Analysis of RFM Customer Segmentation Using Clustering Algorithms. Journal of Mechanical Engineering. 7. 6375-6381.

Ding, Chris & He, Xiaofeng. (2002). Cluster Merging and Splitting in Hierarchical Clustering Algorithms. Proceedings - IEEE International Conference on Data Mining, ICDM. 139- 146.

Dolnicar, S. (2004). Beyond “Commonsense Segmentation”: A Systematics of Segmentation Approaches in Tourism. Journal of Travel Research.

Dursun, A., & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. Tourism Management Perspectives, 18, 153 160.

Elsner, R., Krafft, M., & Huchzemeier, A. (2003). Optimizing Rhenania's mail-order business through dynamic multilevel modeling (DMLM). Interfaces, 33(1), 50–66.

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

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD’96). AAAI Press, 226–231

Fadaei, A., & Khasteh, S. H. (2019). Enhanced K-means re-clustering over dynamic networks. Expert Systems with Applications, 132, 126-140.

Fader, P. S. (2009). Customer Centricity: Focus on the Right Customers for Strategic Advantage. Wharton Digital Press.

Fitzpatrick, M. (2001). Statistical analysis for direct marketers—In plain English. Data Management, 64(4), 54–56.

Gustriansyah, Rendra & Suhandi, Nazori & Antony, Fery. (2020). Clustering optimization in RFM analysis Based on k-Means. Indonesian Journal of Electrical Engineering and Computer Science. 18. 470.

Heldt, R., Silveira, C. S., & Luce, F. B. (2021). Predicting customer value per product: From RFM to RFM/P. Journal of Business Research, 127, 444–453.

Hu, Y. -H., & Yeh, T. -W. (2014). Discovering valuable frequent patterns based on RFM analysis without customer identification information. Knowledge-Based Systems, 61, 76–88.

Hughes, A. M. (1994). Strategic Database Marketing. McGraw-Hill.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. ACM Computing Surveys.

Janvrin, D. J., Raschke, R. L., & Dilla, W. N. (2014). Making sense of complex data using interactive data visualization. Journal of Accounting Education, 32(4), 31-48.

Jiang, Tianyi & Tuzhilin, Alexander. (2009). Improving Personalization Solutions through Optimal Segmentation of Customer Bases. IEEE Transactions on Knowledge and Data Engineering. 21(3), pp. 305-320.

Jin, X., Han, J. (2011). K-Means Clustering. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA.

John, J.M., Shobayo, O., Ogunleye, B. (2023). An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market. Analytics. 2. 809-823.

Kao, Y.-T., Wu, H.-H., Chen, H.-K., & Chang, E.-C. (2011). A case study of applying LRFM model and clustering techniques to evaluate customer values. Journal of Statistics and Management Systems, 14(2), 267–276.

Kotler, P. (1980). Marketing Management: Analysis, Planning, and Control. Prentice-Hall.

Lin Lang, Shuang Zhou, Minjuan Zhong, Guang Sun, Bin Pan, Peng Guo (2022) A Big Data Based Dynamic Weight Approach for RFM Segmentation, Computers, Materials and Continua, Volume 74, Issue 2, Pages 3503-3513, ISSN 1546-2218

Liu, D.-R. and Shih, Y.-Y. (2005). Integrating AHP and datamining for product recommendation based on customer lifetime value, Information & Management, 42(3), pp. 387–400.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.

Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9(1), 381-386.

Maheswari, K. (2019). Finding best possible number of clusters using k-means algorithm. International Journal of Engineering and Advanced Technology, 9(1S4), 533-538.

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

Maryani, I., & Riana, D. (2017). Clustering and profiling of customers using RFM for customer relationship management recommendations. 2017 5th International Conference on Cyber and IT Service Management (CITSM).

Melnykov, V., & Zhu, X. (2019). An extension of the K-means algorithm to clustering skewed data. Computational Statistics, 34, 373-394.

Miglautsch, J. R. (2000). Thoughts on RFM scoring. Journal of Database Marketing, 8(1), 67–72.

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill Science/Engineering/Math.

Mohammed, M., Khan, M.B., & Bashier, E.B.M. (2016). Machine Learning: Algorithms and Applications (1st ed.). CRC Press.

Monalisa, S., & Kurnia, F. (2019). Analysis of DBSCAN and K-means algorithm for evaluating outlier on RFM model of customer behaviour. Telkomnika (Telecommunication Computing Electronics and Control), 17(1), 110-117

Olson, D. L., Cao, Q., Gu, C., & Lee, D. (2009). Comparison of customer response models. Service Business, 3, 117–130.

Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: a case study. Marketing Intelligence & Planning, 35(4), 544–559.

Phill, Lindsey. (2024). Unsupervised Learning: A Comprehensive Exploration of Algorithms, Applications, and Challenges.

Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. Expert Systems with Applications, 97, 205–227.

Reinartz, W.J. and Kumar, V. (2000), “On the profitability of long-life customers in a noncontractual setting: An empirical investigation and implications for marketing”, Journal of Marketing, American Marketing Association, Vol. 64 No. 4, pp. 17–35.

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. PLOS ONE, 14(1), e0210236.

Saabith, A. S., Vinothraj, T., & Fareez, M. (2022). Business Intelligence Tools-Systematic Review. Int. J. Res. Eng. Sci. ISSN, 10(10), 394-408.

Sandhya, N., & Charanjeet, K. R. (2016). A review on machine learning techniques. International Journal on Recent and Innovation Trends in Computing and Communication, 4(3), 451-458.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science, 2(3).

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.

Sheng, T. K., & Subramanian, P. (2019, January). Proposition of rank-based stepwise interactive visualization for customer segmentation in e-commerce. In Proceedings of the 2nd International Conference on Software Engineering and Information Management (pp. 244-248).

Shinde, V., Ransing, N., Ransing, S., Chitranshi, S., Shinde, A. S. (2023). An Implementation on User Centered Website Using Customer Segmentation. International Journal for Research in Applied Science & Engineering Technology (IJRASET), May, 11(V), pp. 5212-5217.

Shindler, M., Wong, A., & Meyerson, A. (2011). Fast and accurate k-means for large datasets. Advances in neural information processing systems, 24.

Shutaywi, M., & Kachouie, N. N. (2021). Silhouette analysis for performance evaluation in machine learning with applications to clustering. Entropy, 23(6), 759.

Sinaga, K.P., & Yang, M. (2020). Unsupervised K-Means Clustering Algorithm. IEEE Access, 8, 80716-80727.

Smith, W. R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. Journal of Marketing.

Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction. MIT Press. ISBN 9780262039246

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). 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.

Tripathi, Anuj & Bagga, Teena & Aggarwal, Rashmi. (2020). Strategic Impact of Business Intelligence : A Review of Literature. Prabandhan: Indian Journal of Management. 13. 35.

Wedel, M., & Kamakura, W. A. (2000). Market Segmentation: Conceptual and Methodological Foundations. Springer.

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.

Wells, W. D. (1975). Psychographics: A critical review. Journal of Marketing Research.

Wind, Y. J., & Bell, D. R. (2008). Market segmentation. In The marketing book (pp. 260-282). Routledge.

Wu, J., Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Yang, L., & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm. Mathematical Problems in Engineering, 2020, 1–7.

Yeh, I.-C., Yang, K.-J., & Ting, T.-M. (2009). Knowledge discovery on RFM model using Bernoulli sequence. Expert Systems with Applications, 36(3), 5866–5871.

Ανδρουτσόπουλος, Ι. (2019). Τεχνητή νοημοσύνη και μηχανική μάθηση. Εκδόσεις Καστανιώτη.

Λαζαρίδου, Β. (2023). Ανάλυση καταναλωτικής συμπεριφοράς και εφαρμογή κατάλληλης στρατηγικής μάρκετινγκ με τη χρήση της ανάλυσης RFM και του μοντέλου μηχανικής μάθησης kmeans clustering