DATA SCIENCE | FINAL PROJECT

Στόχος είναι η πραγματοποίηση μιας εργασίας ανάλυσης δεδομένων σχετικά με τις πωλήσεις μίας φανταστικής εταιρείας λιανικής, εστιάζοντας στις βασικές δεξιότητες του data science και του Project management.

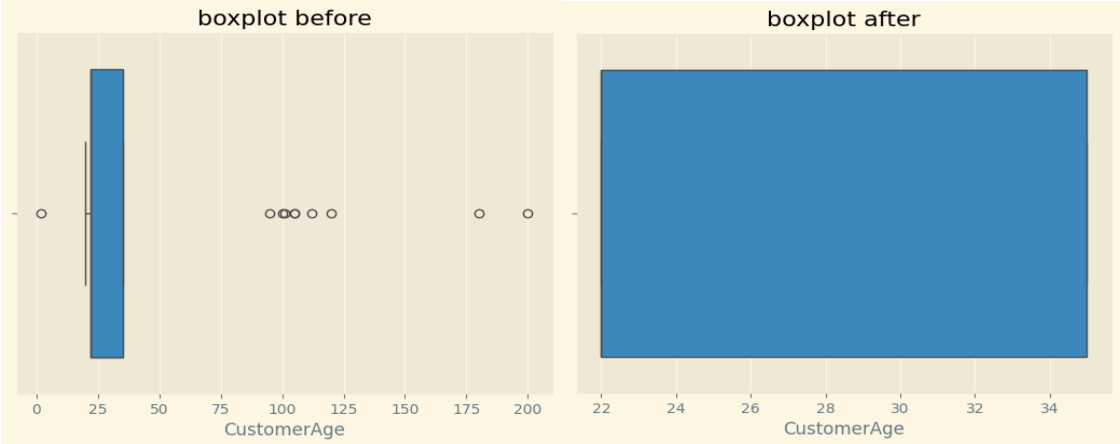
1. Data cleaning and preparation

- **SALES DATE**
  Η στήλη SALES DATE περιέχει ημερομηνίες πωλήσεων σε τύπο δεδομένων *text* και σε μορφή YYYY-MM-DD. Υπήρχαν όμως αρκετές εγγραφές με λάθος μορφή ημερομηνίας. Διορθώθηκαν οι εγγραφές με την εντολή *pd.todatetime* και δημιουργήθηκαν άλλες τρείς στήλες για καλύτερη στατιστική ανάλυση των πωλήσεων. Οι στήλες αυτές ήταν α) *SalesYear* με την χρονολογία της πώλησης, β) *SalesMonth* με τον μήνα της πώλησης και ο συνδυασμός των δυο γ) *Salesym*.

- **PRODUCT CATEGORY**
  Η στήλη PRODUCT CATEGORY περιέχει μια κατηγορική μεταβλητή για την κατηγορία προϊόντων της πώλησης. (κάθε πώληση έχει μόνο μία κατηγορία προϊόντος). Οι κατηγορίες αυτές είναι *Clothing, Electronics, Home Appliances*. Τα δεδομένα ήταν σε μορφή *text* εκτός από μία εγγραφή που είχε καταχώρηση *Female*. Μετά από αναζήτηση της εγγραφής ανακαλύφθηκε πως είχε γίνει αντίστροφη καταχώρηση με την στήλη *CUSTOMER GENDER* καθώς σε εκείνο το κελί υπήρχε η καταχώρηση Clothing. Θεωρήσαμε πως ήταν τυχαίο λάθος και αποφασίσαμε την εναλλαγή των δυο καταχωρήσεων για να μην υπάρχουν λάθη.

- **CUSTOMER LOCATION**
  Η στήλη CUSTOMER LOCATION περιέχει και αυτή μια κατηγορική μεταβλητή σε μορφή *text* με την πληροφορία της χώρας διαμονής του πελάτη. Οι πιθανές χώρες είναι η Ιαπωνία, Αυστραλία, Ινδία, Καναδάς, Η.Π.Α., Ηνωμένο Βασίλειο. Σε αυτή τη στήλη εντοπίσαμε τις μοναδικές *Nan* καταχωρήσεις και αποφασίσαμε τη διαγραφή τους. Σε αυτή την απόφαση μας οδήγησε το μικρό τους πλήθος και πως η στατιστική μας ανάλυση θα βασιζόταν στα δημογραφικά στοιχεία των πελατών.

- **CUSTOMER AGE**
  Η στήλη CUSTOMER AGE ήταν και αυτή μια κατηγορική μεταβλητή σε μορφή αριθμού *int.* Από την καταμέτρηση των εγγραφών και του boxplot$_{(g1)}$ εντοπίσαμε παρουσία outliers. Οι εγγραφές αυτές ήταν εκτός ρεαλιστικών ορίων καθώς πλησίαζαν ή και ξεπερνούσαν κατά πολύ τα 100 χρόνια και έφταναν μέχρι τα 200, καθώς και μια καταχώρηση 2 ετών. Με την μέθοδο IQR εντοπίσαμε τα όρια και εξαιρέσαμε ό,τι ήταν εκτός ορίων. Μετά μετρήσαμε τις εγγραφές μας, μία εγγραφή ήταν στα 20 έτη και όλες οι υπόλοιπες είχαν μοιραστεί στις επιλογές 22 και 35 έτη.



g1 boxplot before and after outlier removal

Αποφασίσαμε την μετατροπή του 20 σε 22 καθώς είναι πολύ κοντά και η συμπερίληψη αυτής της μίας εγγραφής σε ένα μεγαλύτερο σύνολο θα βοηθούσε πολύ την ανάλυση μας να γίνει καλύτερα κατανοητή. Τέλος αλλάξαμε τον τύπο των δεδομένων σε *text* καθώς είναι κατηγορική μεταβλητή και θα δούλευαν καλύτερα έτσι τα γραφήματα και οι αλγόριθμοι πρόβλεψης, ομαδοποίησης.

- **CUSTOMER GENDER**

  Στο CUSTOMER GENDER οι καταχωρήσεις ιδανικά θα είχαν μόνο τρείς επιλογές, *Male, Female και Non-binary*. Εντοπίσαμε όμως 23 εγγραφές χωρίς πληροφορία για το φύλλο(Unknown) και 1 εγγραφή που δεν είχε απαντήσει (Did not Answer). Αποφασίσαμε να συμπεριλάβουμε την εγγραφή Did not Answer μέσα στις Unknown καθώς και οι δυο περιείχαν μηδενική πληροφορία για το φύλλο. Οι εγγραφές Non-binary λόγω του μικρού τους πλήθους, μόνον 4 στις 1000, δεν θα βοηθούσαν την στατιστική μας ανάλυση, για λόγους σαφήνειας λοιπόν αποφασίσαμε να τις συμπεριλάβουμε και αυτές στις Unknown. Έτσι δημιουργήθηκε ένα σύνολο 28 εγγραφών που θεωρούμε πως είναι αρκετά μεγάλο για να το εξαιρέσουμε από την ανάλυση μας καθώς περιέχει πληροφορίες για τις άλλες μεταβλητές μας είχαμε απολέσει σημαντικό όγκο δεδομένων.

- **SALES AMOUNT**

  Η SALES AMOUNT είναι η στήλη που περιέχει τις πωλήσεις σε δολλάρια και είναι η μόνη συνεχής αριθμητική μεταβλητή μας. Εντοπίσαμε εγγραφές με αλφαριθμητικές εγγραφές που αποφασίσαμε να τις αποκλείσουμε από την ανάλυση μας και μια εγγραφή σχεδόν δύο εκατομμυρίων που αποτελούσε θόρυβο. Η καταχώρηση αυτή επηρέαζε σε τεράστιο βαθμό τις μετρικές μας καθώς το μέγεθος όλων των υπόλοιπων καταχωρήσεων ήταν κάτω από 2000. Η εγγραφή αποφασίστηκε να εξαιρεθεί της ανάλυσης.

- **RATINGS**

  Η στήλη RATINGS περιέχει την αξιολόγηση της παραγγελίας από τους πελάτες σε κλίμακα 1-5 (1 το μικρότερο και 5 το μεγαλύτερο) σε μορφή *text*. Σε αυτές τις καταχωρήσεις εντοπίσαμε δύο καταχωρήσεις που είχαν γίνει λεκτικά (π.χ. one αντί για 1). Εδώ αποφασίσαμε την μετατροπή τους στο αντίστοιχο αριθμητικό. Επίσης υπήρχε μια εγγραφή εκτός κλίμακας (10) και καθώς δεν είχαμε επιπλέον πληροφορίες για το αν αποτελεί τυπογραφικό λάθος αποφασίσαμε την εξαίρεση της.

-Το αρχικό dataframe αποτελούνταν από 1000 εγγραφές και μετά τον καθαρισμό και την προετοιμασία έμειναν 980.΄Οσες εγγραφές εξαιρέθηκαν από τα δεδομένα μας αποθηκεύτηκαν σε ξεχωριστό dataframe.

2. Basic statistics of our dataset

STATS	Sales Date	Sales Year	Sales Month	Sales ym	Sales Amount	Customer Age	Customer Gender	Customer Location	Product Category	Product Ratings
COUNT	980	980	980	980	980	980	980	980	980	980
UNIQUE			12	24		3	3	6	3	
MIN	2/1/2022	2022			22					
MAX	31/12/2023	2023			1994					
MODE	16/7/2023	2023	May	Dec_23	609	22	Female	USA	Electronics	3
FREQ	6	525	91	52	14	505	477	205	347	214
MEAN			Jun		980					
MEDIAN		2023	Jun		987					
STD					538					

Με τις εντολές describe, median, mode, value counts υπολογίσαμε κάποια βασικά στατιστικά του dataset μας. Περιληπτικά

Sales Date: Υπάρχουν εγγραφές από την δεύτερη μέρα του 2022 μέχρι και το τέλος του 2023. Η ημερομηνία με τις περισσότερες εμφανίσεις(έξι) ήταν η 16η Ιουλίου του 2023.

Sales Year: Στις χρόνιες είχαμε μόνο δυο επιλογές, 2022 και 2023 με το 2023 να εμφανίζεται 525 φορές.

Sales Month: Στους μήνες υπήρχαν και οι δώδεκα επιλογές με πιο συχνή τον Μάιο με 91 καταχωρήσεις.

Sales ym: Εδώ εμφανίστηκαν και οι 24 δυνατές επιλογές με τον Δεκέμβριο του 2023 να εμφανίζεται 52 φορές.

Product Category. Εδώ τα Electronics είχαν τις περισσότερες εμφανίσεις 347 εμφανίσεις.

Customer Age: Όπως αναλύσαμε και στον καθαρισμό προηγουμένως, στη ηλικία του πελάτη υπάρχουν μόνο δύο κατηγορίες, 22 και 35 έτη με την πρώτη να εμφανίζεται 505 φορές.

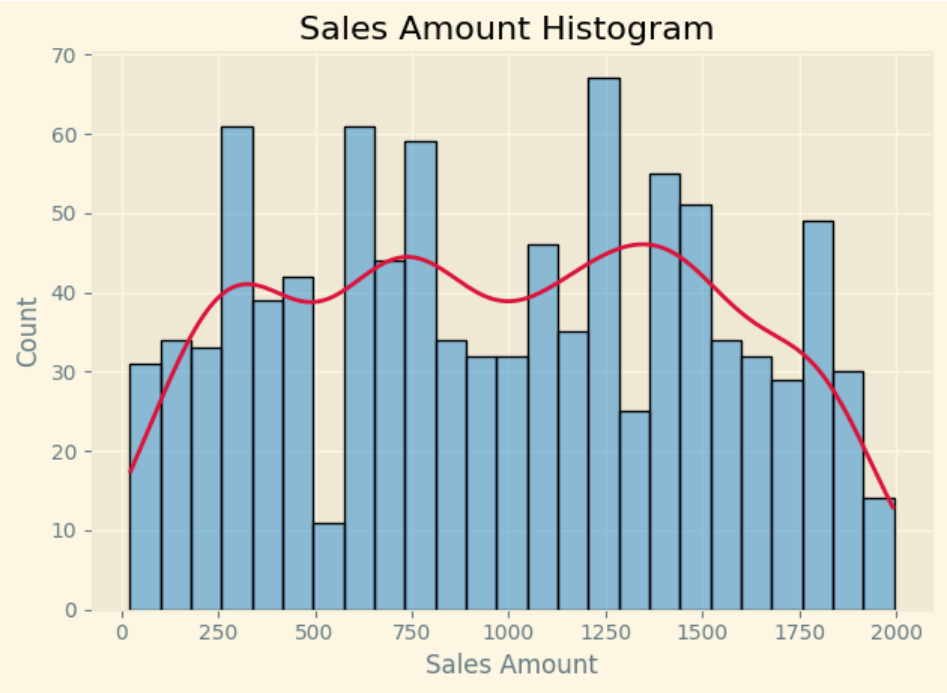
Customer Gender: Στο φύλλο είχαμε τρείς κατηγορίες με την επιλογή Female να εμφανίζεται 477 φορές.

Customer Location: Έξι τοποθεσίες με τις χώρες Ιαπωνία, Αυστραλία, Ινδία, Καναδάς, Ηνωμένο Βασίλειο και Η.Π.Α. με την τελευταία να εμφανίζεται τις περισσότερες φορές (205).

Product Ratings: Από τις 5 πιθανές βαθμολογήσεις, η επιλογή 3 εμφανίστηκε τις περισσότερες φορές με 214 καταχωρήσεις.

Sales Amount: Ελάχιστη καταχώρηση(MIN): 22, Μέγιστη καταχώρηση(MAX): 1994, Τυπική απόκλιση(STD): 558, Μέσος όρος(MEAN): 980, Διάμεσος(MEDIAN): 987, Επικρατούσα τιμή(MODE): 609(14 εμφανίσεις).

Με βάση αυτές τις μετρικές, διασπορά μεγαλύτερη του 25% του εύρους, θα περιμένουμε αρκετή διασπορά στις καταχωρήσεις των πωλήσεων και όχι συγκέντρωση γύρω από τον Μέσο όρο. Η μικρή απόσταση του Μέσου όρου και της Διαμέσου υποδηλώνουν πως η κατανομή θα είναι κάπως συμμετρική αλλά καθώς η Επικρατούσα τιμή είναι μικρότερη των δυο ίσως υπάρχει μια θετική ασυμμετρία με διόγκωση των δεδομένων αριστερά της κορυφής. Το histogram μας δείχνει πολλά σημεία με υψηλή συχνότητα οδηγώντας μας στο συμπέρασμα πως η κατανομή δεν είναι ομοιόμορφη και οι αριθμοί πωλήσεων είναι διασπαρμένοι σε όλο το εύρος τους(graph1) .



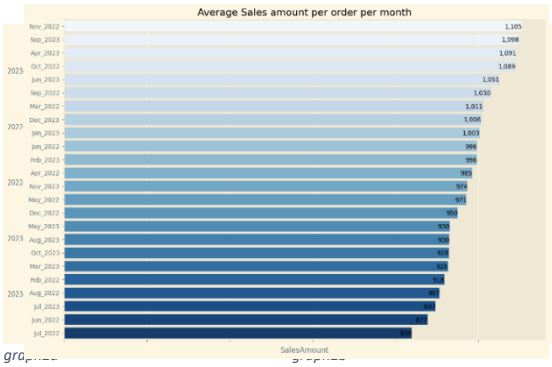
graph1

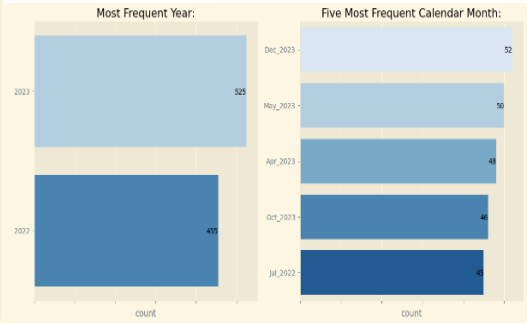
Με βάση τα στατιστικά θα μπορούσαμε να περιγράψουμε τον μέσο πελάτη της εταιρίας ως μια Γυναίκα 22 χρονών από την Αμερική, που πιο συχνά αγοράζει Electronics αξίας 980$ και δίνει αξιολόγηση 3.
Η μεγαλύτερη πώληση πραγματοποιήθηκε την 27η Σεπτεμβρίου του 2023 από μία 35χρονη από την Αυστραλία. Αγόρασε Electronics είδη αξίας 1994$ και βαθμολόγησε τα προϊόντα με 4.

Στην συνέχεια προχωρήσαμε σε ανάλυση των πωλήσεων πάνω σε κάθε μια από τις άλλες μεταβλητές.

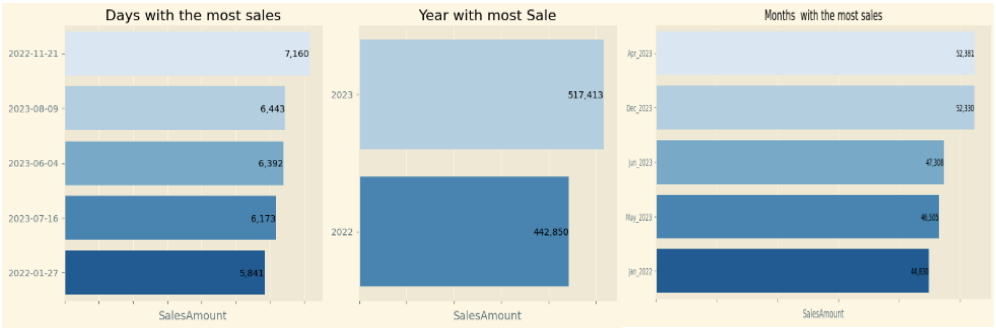
- **DATES**

Τα δεδομένα πωλήσεών μας αποκαλύπτουν μερικά ενδιαφέροντα πράγματα.
Ενώ η 16η Ιουλίου 2023(graph2a) κατέχει το ρεκόρ για την ημερομηνία με τις περισσότερες μοναδικές παραγγελίες (6) και σημαντικό αριθμό πωλήσεων (6.173), οι υψηλότερες συνολικές πωλήσεις για μια ημέρα ανήκουν στην 21η Νοεμβρίου 2023 (7.160 $)(graph3a). Από την άλλη πλευρά, η 28η Ιουλίου 2022 σημείωσε τη λιγότερη δραστηριότητα με μόλις 28$. Ο μήνας που οι περισσότεροι πελάτες προτίμησαν να πραγματοποιήσουν παραγγελία ήταν ο Μάιος με 91 παραγγελίες(graph2b). Όσον αφορά τους μήνες, ο Δεκέμβριος του 2023(graph2d) .κατέχει την πρωτιά για τις περισσότερες παραγγελίες (52), ενώ ο Απρίλιος του 2023 κυριαρχεί από πλευράς συνολικών πωλήσεων (52.381$)(graph3c), ξεπερνώντας τον Δεκέμβριο κατά μόλις 51$. Ο Αύγουστος του 2022 κατέγραψε τις χαμηλότερες πωλήσεις για οποιονδήποτε μήνα με 24.476$.





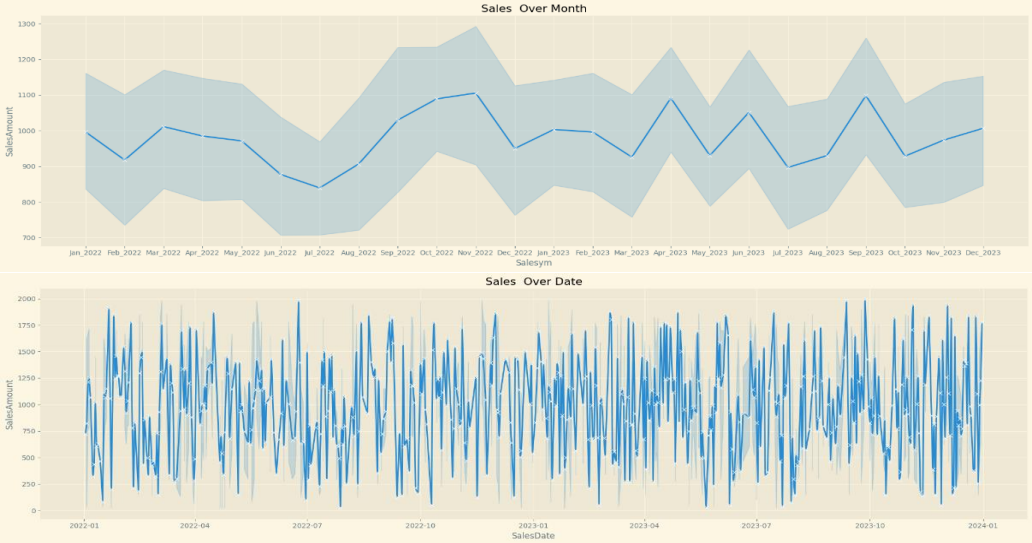
graph4 graph2c graph2d



graph3a graphs3b graph3c

Ο μέσος πελάτης τείνει να αγοράζει προϊόντα αξίας 980$ ανά παραγγελία, με τον Νοέμβριο του 2023 να έχει τις υψηλότερες μέσες πωλήσεις ανά παραγγελία με 1105$. (graph4)
Εξετάζοντας τη συνολική εικόνα, το 2023 αποδείχθηκε πολύ πιο δυνατή χρονιά, με περισσότερες παραγγελίες(graph2c) αλλά και πωλήσεις(graph3b)

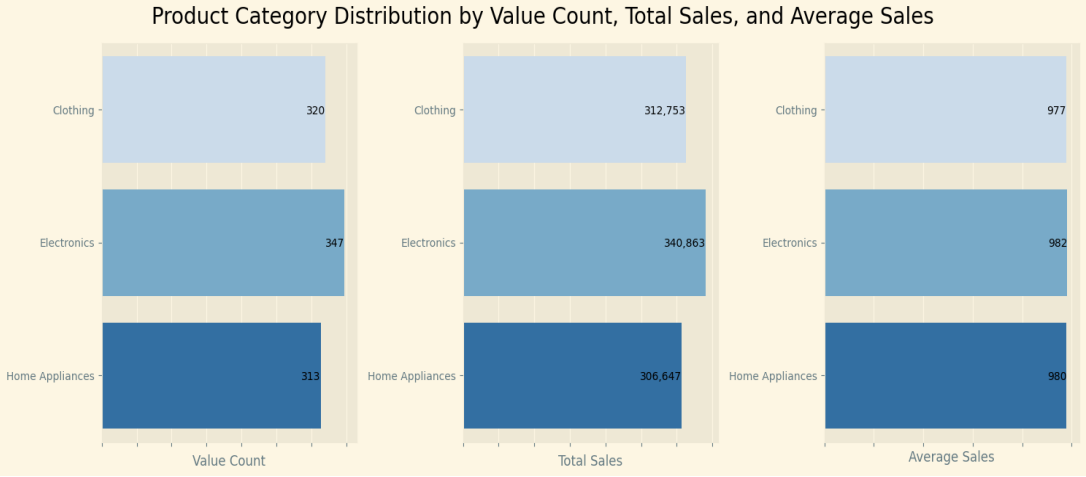
Το time series του Sales Amount δεν δείχνει κάποια σαφή τάση ούτε κατά μήνα, ούτε κατά ημερομηνία. (graph5)



graph5

- **PRODUCT**

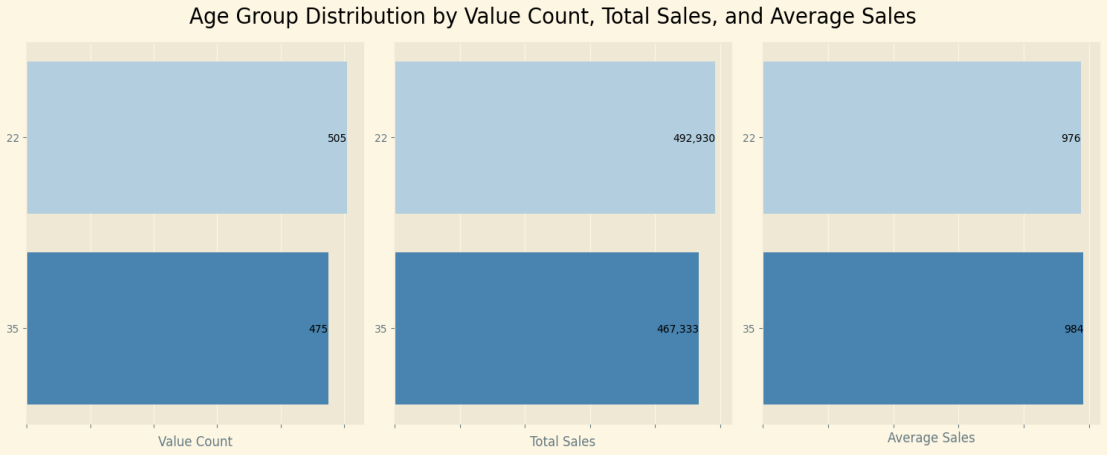
  Τα Electronics είναι η πιο συχνή κατηγορία στα δεδομένα μας, εμφανίζονται 347 φορές και αντιπροσωπεύουν το 35% των πωλήσεων με 340.863 $. Ωστόσο, τα δεδομένα και τα γραφήματα υποδηλώνουν μια σχετικά ομοιόμορφη κατανομή μεταξύ των κατηγοριών παρά το προβάδισμα των ηλεκτρονικών. Οι άλλες δύο κατηγορίες δεν απέχουν πολύ. Αυτό υποστηρίζεται περαιτέρω από το γεγονός ότι ο μέσος όρος πώλησης κάθε κατηγορίας είναι σχεδόν ίσος.(graph6)



graph6

- **AGE**

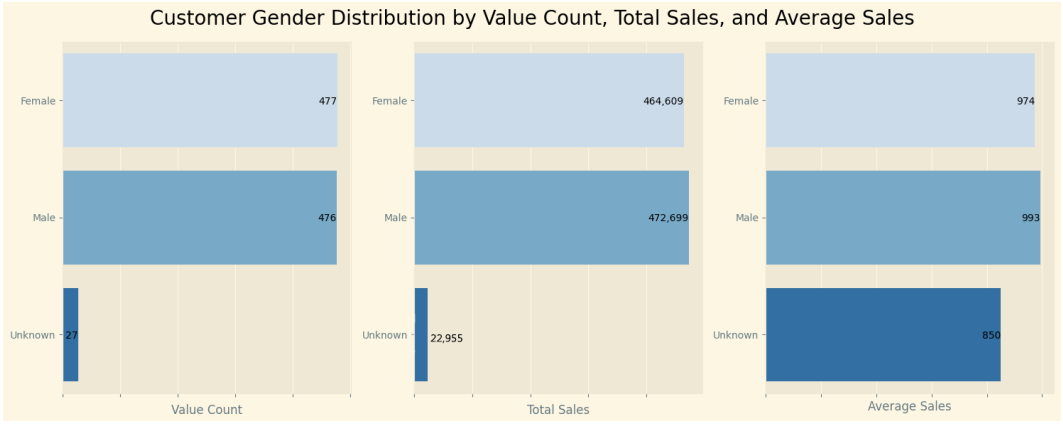
Η ηλικία στα δεδομένα μας είναι μια κατηγορική μεταβλητή με δύο διακριτές κατηγορίες: 22, 35. Η πλειονότητα των πελατών είναι ηλικίας 22 ετών (505/980). Η ηλικιακή ομάδα των 22 ετών έχει τις μεγαλύτερες συνολικές πωλήσεις (492.930$), ελαφρώς υψηλότερες από την ηλικιακή ομάδα των 35 ετών (467.333$). Οι μέσες τιμές για τις ηλικιακές ομάδες 22 (976$) και 35 (984$) είναι πολύ κοντά, γεγονός που υποδηλώνει παρόμοια μέση συμπεριφορά μεταξύ αυτών των δύο κατηγοριών.(graph7)



graph7

- **GENDER**

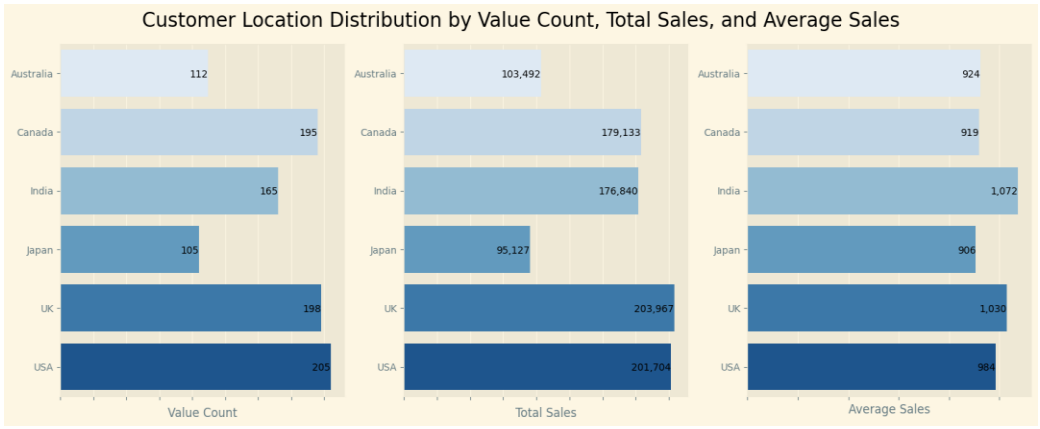
Η πελατεία μας αποτελείται κυρίως από άνδρες και γυναίκες (περίπου 97%), με ένα μικρό ποσοστό που δεν αποκαλύπτει το φύλο του. Οι άνδρες τείνουν να ξοδευούν κατά μέσο όρο περισσότερο (993$), με συνολικές πωλήσεις 472.699$. Οι γυναίκες ακολουθούν με λίγο μικρότερα νούμερα(graph8).



graph8

- **LOCATION**

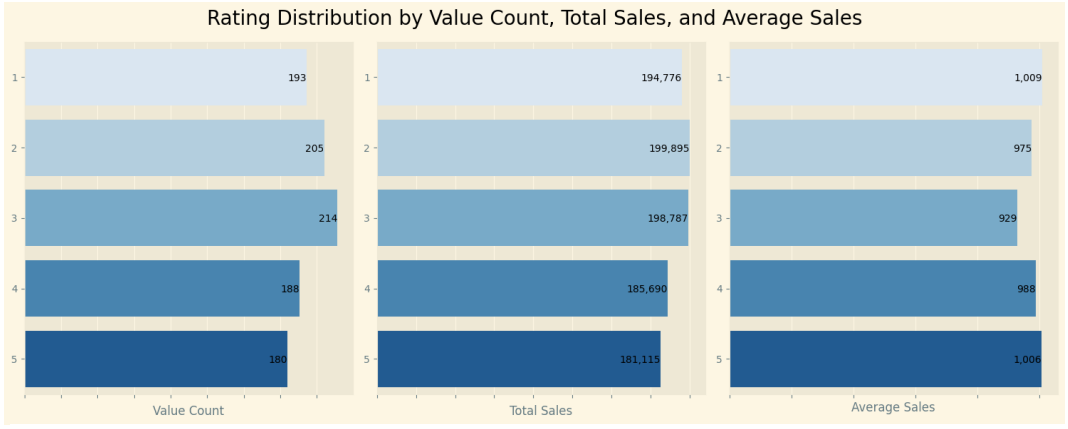
Οι Ηνωμένες Πολιτείες (ΗΠΑ) προηγούνται στον αριθμό των εμφανίσεων στα δεδομένα μας, με 205 εμφανίσεις. Ωστόσο, το Ηνωμένο Βασίλειο (ΗΒ) καταλαμβάνει την πρώτη θέση στις πωλήσεις, κατέχοντας περίπου 21% με 203,967$, ξεπερνώντας ελαφρά τις ΗΠΑ. Ο Καναδάς και η Ινδία ακολουθούν, συμβάλλοντας ο καθένας περίπου στο 19% των πωλήσεων με λίγο κάτω από 180χιλιάδες. Η Αυστραλία και η Ιαπωνία παρουσιάζουν τις λιγότερες πωλήσεις ,100χιλ και κάτω, κάτι που συμφωνεί με τη χαμηλότερη συχνότητα εμφάνισής τους στο σύνολο δεδομένων(graph9).



graph9

- **RATING**

Τα προϊόντα που έχουν βαθμολογηθεί με 2 και 3 έχουν τον μεγαλύτερο όγκο πωλήσεων και εμφανίζονται τις περισσότερες φορές στα δεδομένα μας και έχουν τις περισσότερες συνολικές πωλήσεις. Αυτό θα μπορούσε να υποδηλώνει πιθανά προβλήματα με αυτά τα προϊόντα, οδηγώντας σε χαμηλότερη ικανοποίηση πελατών.
Με βάση την κατανομή που φαίνεται ομοιόμορφη, μπορούμε να δώσουμε προτεραιότητα στις προσπάθειες για την ενίσχυση των προϊόντων με χαμηλότερη βαθμολογία και την προώθηση των προϊόντων με υψηλότερη βαθμολογία.(graph10)



graph10

3. HYPOTHESIS TESTING

Στο Hypothesis testing μας θα χρησιμοποιήσουμε την ANOVA για τις ανεξάρτητες μεταβλητές μας με τρία ή περισσότερα επίπεδα (κατηγορίες ή ομάδες). Η ANOVA σχεδιάστηκε για να συγκρίνει τους μέσους όρους αυτών των πολλαπλών ομάδων. Για τη στήλη Customer Age θα πραγματοποιήσουμε ένα διπλής κατεύθυνσης T test καθώς έχει μόνο δύο κατηγορίες.

Για τις μεταβλήτες Product Category, Customer Gender και Product Ratings, η στατιστική τιμή ANOVA είναι μικρή και χαμηλότερη από την κρίσιμη τιμή F, γεγονός που υποδηλώνει ότι δεν υπάρχει σημαντική διαφορά μεταξύ των μέσων όρων των κατηγοριών του Sales Amount. Επιπλέον, η τιμή p είναι υψηλότερη από την τιμή του επιπέδου σημαντικότητας (alpha=0.05). Αυτό σημαίνει ότι αποτύχαμε να απορρίψουμε την Null hypothesis ότι αυτές οι κατηγορίες δεν επηρεάζουν τον μέσο όρο του Sales Amount. (tableHT)

CustomerAge
Result: Fail to reject the null hypothesis that CustomerAge has no effect on average Sales Amount
---

ProductCategory
Result: Fail to reject the null hypothesis that ProductCategory has no effect on average Sales Amount
---

CustomerGender
Result: Fail to reject the null hypothesis that CustomerGender has no effect on average Sales Amount
---

CustomerLocation
Result: Reject the null hypothesis that CustomerLocation has no effect on average Sales Amount
---

ProductRatings
Result: Fail to reject the null hypothesis that ProductRatings has no effect on average Sales Amount
---

Feature	degrees_between	degrees_within	critical_f	F-statistic	p-value
CustomerAge	1	978	3.851	0.051	0.822
ProductCategory	2	977	3.005	0.007	0.993
CustomerGender	2	977	3.005	0.955	0.385
CustomerLocation	5	974	2.223	2.469	0.031
ProductRatings	4	975	2.381	0.744	0.562

tableHT

Ο συνδυασμός των δύο αποτελεσμάτων υποδηλώνει ότι οι μέσοι όροι των κατηγοριών είναι πολύ παρόμοιοι μεταξύ τους και οποιεσδήποτε παρατηρούμενες διαφορές στους μέσους όρους των ομάδων πιθανόν να οφείλονται σε τυχαιότητα παρά σε πραγματική επίδραση και δεν είναι στατιστικά σημαντικές. Καταλήξαμε στο ίδιο συμπέρασμα για τη μεταβλητή Customer Age μετά την εκτέλεση του t-test δύο δειγμάτων.(table-tt)

T-statistic:	-0.225
P-value:	0.821

table-tt

Η δοκιμή ANOVA για την μεταβλητή Customer Location δείχνει τιμή F υψηλότερη από την κρίσιμη τιμή F (2.223), γεγονός που υποδηλώνει ότι υπάρχει κάποια διακύμανση μεταξύ των μέσων όρων πωλήσεων των έξι τοποθεσιών. Επίσης, η τιμή p (0.031) είναι χαμηλότερη από την τιμή του επιπέδου σημαντικότητας (alpha=0.05). Θα απορρίψουμε τη μηδενική υπόθεση. (graph11)

Με βάση την στατιστική F και την τιμή p, συμπεραίνουμε ότι υπάρχει στατιστικά σημαντική διαφορά μεταξύ των μέσων όρων ορισμένων κατηγοριών. Αυτό σημαίνει ότι η παρατηρούμενη διακύμανση μεταξύ των μέσων όρων πωλήσεων των έξι τοποθεσιών είναι απίθανο να οφείλεται σε τυχαιότητα και ότι υπάρχει επίδραση του Customer Location στον μέσο όρο του Sales Amount. Για να διερευνήσουμε περαιτέρω αυτήν τη σύνδεση, θα πραγματοποιήσαμε ένα διπλής κατεύθυνσης ANOVA.

Ένα διπλής κατεύθυνσης ANOVA μας επιτρέπει να εξετάσουμε πώς η αλληλεπίδραση μεταξύ δύο ανεξάρτητων μεταβλητών επηρεάζει μια εξαρτημένη μεταβλητή. Δυστυχώς, κανένας συνδυασμός χαρακτηριστικών δεν είχε τιμή p μεγαλύτερη από 0.05. (table-da) Οι τρεις συνδυασμοί που είχαν σκορ χαμηλότερο από 0.10 είναι: Customer Age – Product Rating, Product Category – Customer Location και Product Ratings – CustomerLocation.

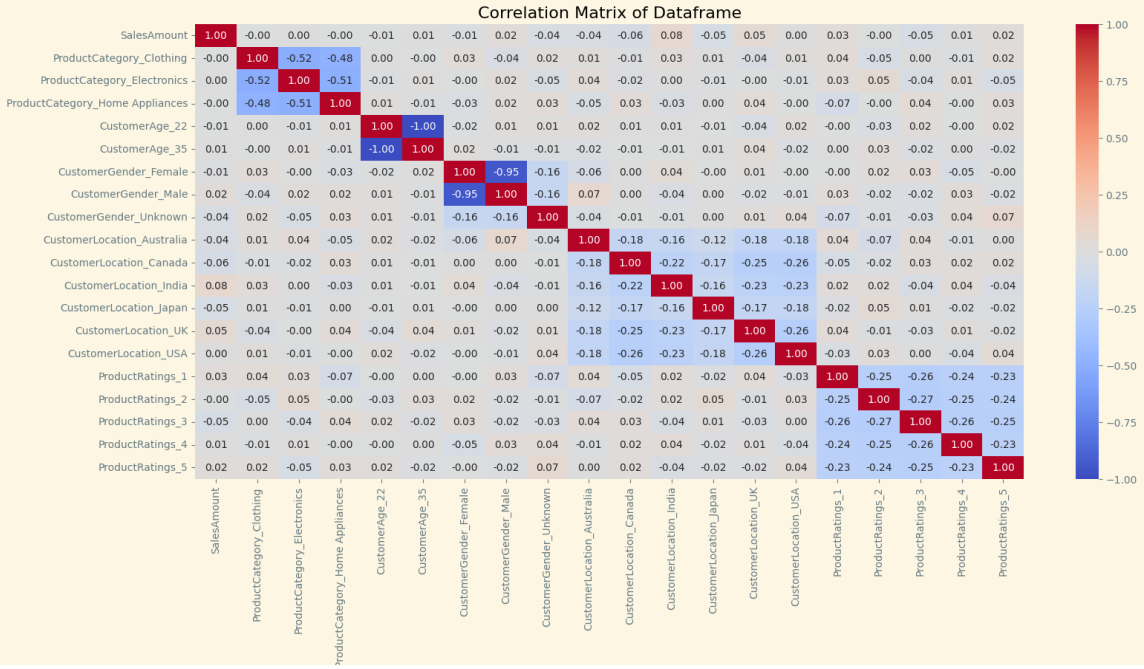
CustomerAge and ProductCategory pr(>f) result:	0.406
CustomerAge and CustomerGender pr(>f) result:	0.742
CustomerAge and ProductRatings pr(>f) result:	**0.099**
CustomerAge and CustomerLocation pr(>f) result:	0.302
ProductCategory and CustomerGender pr(>f) result:	0.189
ProductCategory and ProductRatings pr(>f) result:	0.132
ProductCategory and CustomerLocation pr(>f) result:	**0.077**
CustomerGender and ProductRatings pr(>f) result:	0.591
ProductRatings and CustomerLocation pr(>f) result:	**0.086**
CustomerGender and CustomerLocation pr(>f) result:	0.576

table-da

Ο συνδυασμός με το χαμηλότερο σκορ, Product Category – Customer Location, θα αναλυθεί περαιτέρω σε μεταγενέστερη ενότητα για να δούμε πώς αυτά τα χαρακτηριστικά επηρεάζουν από κοινού τις πωλήσεις.

4. PREDICTIVE MODELING

Η αρχική μας προσέγγιση περιλάμβανε τον υπολογισμό των συσχετίσεων μεταξύ των μεταβλητών αφού πρώτα μετατρέψαμε τις κατηγορικές μεταβλητές σε πολλαπλές δυαδικές μεταβλητές με κάθε νέα μεταβλητή να αντιπροσωπεύει μια μοναδική κατηγορία από την αρχική στήλη. Οι συσχετίσεις μεταξύ των κατηγορικών μεταβλητών και του Sales Amount ήταν όλες πολύ ασθενείς, όλες κάτω από 0.05 και καμία δεν υπερέβαινε το 0.08 (graph12)



graph12

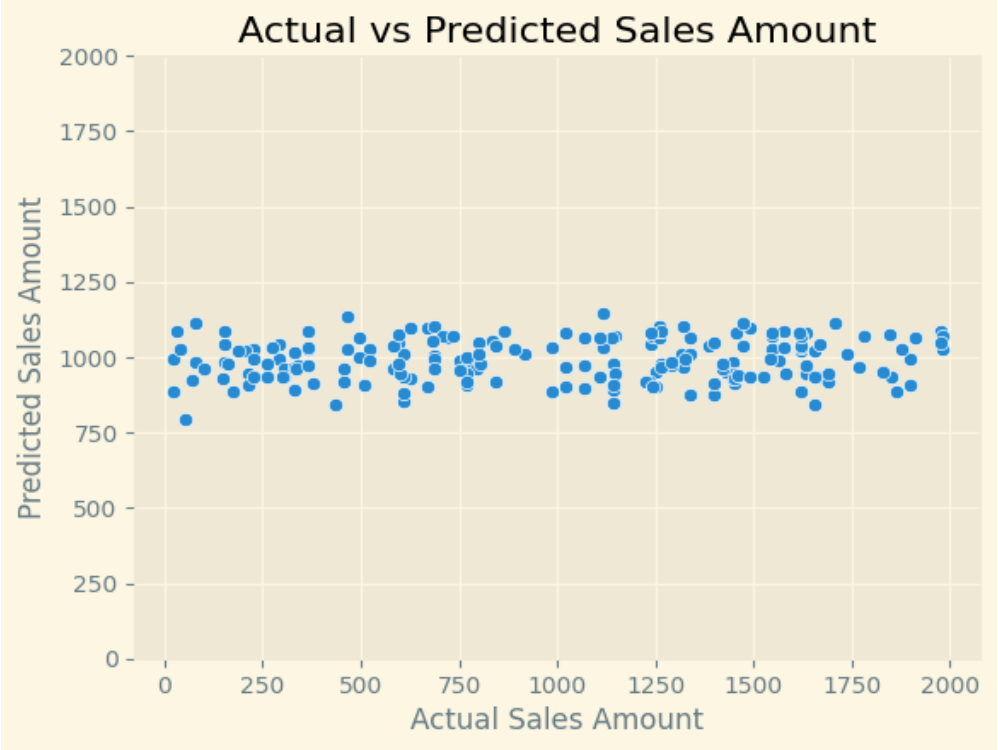
feautures	score	R2	inte	mse	r_mse
['ProductCategory', 'CustomerAge']	-0.015	-0.015	973.788	291.843,239	540.062
['ProductCategory', 'CustomerGender']	-0.019	-0.019	972.484	292.803,968	540.968
['ProductCategory', 'CustomerLocation']	-0.013	-0.013	920.025	291.398,278	539.633
['ProductCategory', 'ProductRatings']	-0.019	-0.019	1.006.060	292.964,171	541.097
['CustomerAge', 'CustomerGender']	-0.017	-0.017	970.343	292.467,982	540.657
['CustomerAge', 'CustomerLocation']	-0.012	-0.012	920.745	291.043,857	539.309
['CustomerAge', 'ProductRatings']	-0.018	-0.018	1.006.033	292.694,911	540.849
['CustomerGender', 'CustomerLocation']	-0.015	-0.015	909.302	291.672,801	539.916
['CustomerGender', 'ProductRatings']	-0.021	-0.021	1.000.786	293.657,960	541.746
['CustomerLocation', 'ProductRatings']	-0.017	-0.017	947.973	292.520,815	540.680
['ProductCategory', 'CustomerAge', 'CustomerGender']	-0.020	-0.020	968.987	293.324,393	541.449
['ProductCategory', 'CustomerAge', 'CustomerLocation']	-0.015	-0.015	917.767	291.934,784	540.134
['ProductCategory', 'CustomerAge', 'ProductRatings']	-0.021	-0.021	1.002.786	293.540,400	541.631
['ProductCategory', 'CustomerGender', 'CustomerLocation']	-0.018	-0.018	907.361	292.567,144	540.744
['ProductCategory', 'CustomerGender', 'ProductRatings']	-0.025	-0.025	998.525	294.541,431	542.561
['ProductCategory', 'CustomerLocation', 'ProductRatings']	-0.021	-0.021	944.038	293.402,075	541.495
['CustomerAge', 'CustomerGender', 'CustomerLocation']	-0.016	-0.016	906.943	292.195,905	540.403
['CustomerAge', 'CustomerGender', 'ProductRatings']	-0.023	-0.023	997.404	294.235,484	542.279
['CustomerAge', 'CustomerLocation', 'ProductRatings']	-0.019	-0.019	945.677	293.108,173	541.228
['CustomerGender', 'CustomerLocation', 'ProductRatings']	-0.022	-0.022	931.340	293.706,996	541.796
['ProductCategory', 'CustomerAge', 'CustomerGender', 'CustomerLocation']	-0.020	-0.020	905.162	293.089,947	541.230
['ProductCategory', 'CustomerAge', 'CustomerGender', 'ProductRatings']	-0.027	-0.027	995.299	295.111,461	543.087
['ProductCategory', 'CustomerAge', 'CustomerLocation', 'ProductRatings']	-0.023	-0.023	941.878	293.986,193	542.040
['ProductCategory', 'CustomerGender', 'CustomerLocation', 'ProductRatings']	-0.025	-0.025	928.704	294.614,053	542.633
['CustomerAge', 'CustomerGender', 'CustomerLocation', 'ProductRatings']	-0.024	-0.024	929.122	294.285,113	542.332
['ProductCategory', 'CustomerAge', 'CustomerGender', 'CustomerLocation', 'ProductRatings']	-0.027	-0.027	926.623	295.189,720	543.167

tablerLR

Στην συνέχεια αφού οι συσχετίσεις δεν μας έδειξαν κάποιες μεταβλητές που θα έπρεπε να προτιμήσουμε για την κατασκευή ενός Linear Regression μοντέλου προχωρήσαμε στην κατασκευή μοντέλου για κάθε πιθανό συνδυασμό μεταβλητών ώστε να εντοπίσουμε τον καλύτερο συνδυασμό. Ωστόσο, αυτά τα μοντέλα απέδωσαν πολύ χαμηλά νούμερα αξιολόγησης, με τις τιμές του R-squared να πέφτουν κάτω από το μηδέν (με μέσο όρο περίπου -0.20).(tableLR) (Σε κάθε μοντέλο χρησιμοποιήθηκε μέθοδος KFold με K=5 και random_state =42)

Αυτή η κακή απόδοση προέρχεται από την έλλειψη σημαντικών σχέσεων μεταξύ των ανεξάρτητων μεταβλητών και της εξαρτημένης μεταβλητής (Sales Amount). Τα συμπεράσματα του Hypothesis testing, σε ό,τι αφορά την σχέση των ανεξάρτητων μεταβλητών με το Sales Amount επιβεβαιώθηκε και από το Correlation matrix.

Ουσιαστικά, τα Linear Regression μοντέλα δεν ήταν σε θέση να εντοπίσουν καμία γραμμική σχέση μεταξύ του Sales Amount και των άλλων χαρακτηριστικών. Κατά συνέπεια, οι προβλέψεις του μοντέλου απλώς κυμαίνονταν γύρω από το μέσο ποσό πωλήσεων(980$).(graph13)



graph13

5. Statistical analysis of Customer Location and Product Category by Sales Amount

Συνολικά, η ανάλυση αποκαλύπτει σημαντικές διαφορές στις αγοραστικές συμπεριφορές μεταξύ των χωρών, καθιστώντας δύσκολο τον εντοπισμό ενός σαφούς προτύπου. Κάθε χώρα παρουσιάζει μοναδικές προτιμήσεις και κατανομές πωλήσεων(table2)

Αυστραλία: Τα Electronics κυριαρχούν στην Αυστραλία με 43% του συνολικού ποσού πωλήσεων και οι οικιακές συσκευές, ενώ έχουν τις λιγότερες παραγγελίες, παρουσιάζουν το υψηλότερο μέσο όρο Sales Amount.

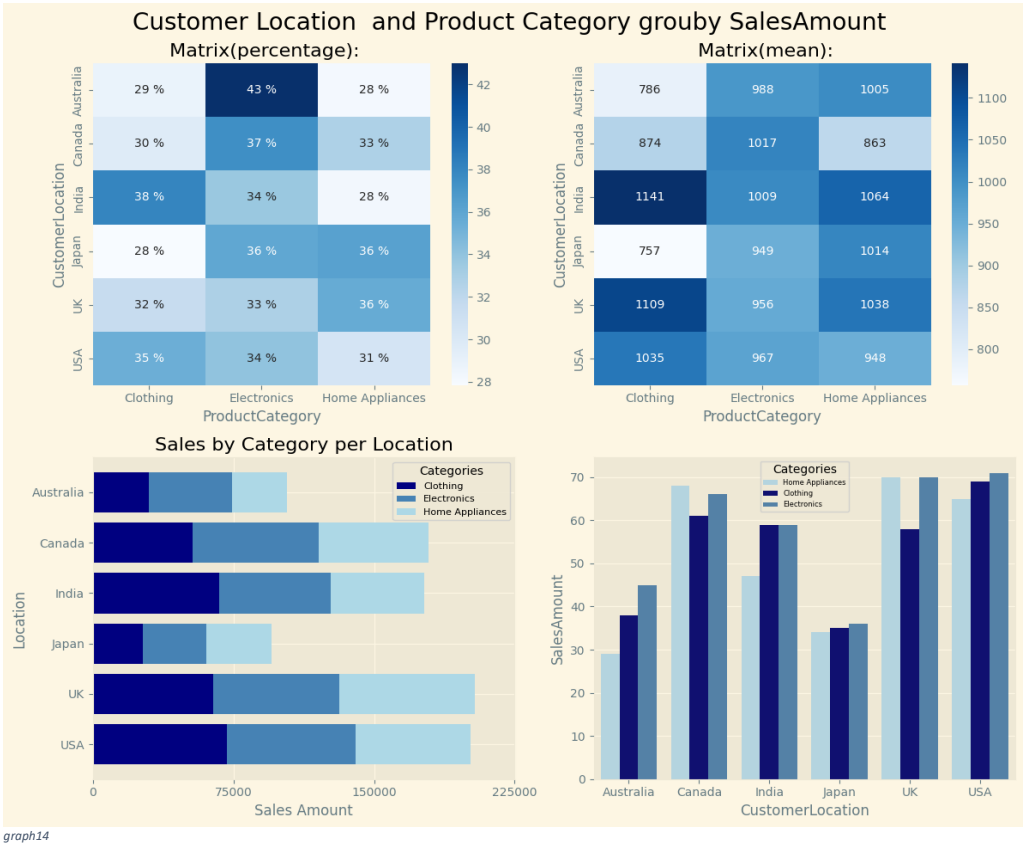
Καναδάς: Πραγματοποιεί περισσότερες παραγγελίες για Home Appliances, αλλά επιτυγχάνει το υψηλότερο ποσό πωλήσεων στα Electronics, τα οποία επίσης έχουν το υψηλότερο μέσο όρο Sales Amount.

Ινδία: Προτιμά τα Clothing προϊόντα, τα οποία έχουν το υψηλότερο ποσοστό Sales Amount και το υψηλότερο μέσο όρο Sales Amount, με όλες τις κατηγορίες να έχουν μέσο όρο Sales Amount άνω των 1000$.

Ιαπωνία: Αντιπροσωπεύει μια μικρότερη αγορά με ίση κατανομή παραγγελιών αλλά χαμηλότερες πωλήσεις στα Clothing προϊόντα, όπως φαίνεται από το χαμηλότερο μέσο όρο Sales Amount.

Ηνωμένο Βασίλειο: Έχει το υψηλότερο συνολικό Sales Amount, με την κατηγορία Clothing να έχει το υψηλότερο μέσο όρο Sales Amount παρά τις λιγότερες παραγγελίες, και τα Home appliances να υπερτερούν των Electronics τόσο στο μέσο όρο Sales Amount όσο και στο ποσοστό του ποσού πωλήσεων.

ΗΠΑ: Δείχνουν προτίμηση για τα ρούχα στο ποσοστό πωλήσεων, αν και πραγματοποιούνται περισσότερες παραγγελίες για ηλεκτρονικά, με τις οικιακές συσκευές να υστερούν σε όλους τους δείκτες.



graph14

Συνοψίζοντας, ενώ υπάρχουν αξιοσημείωτες διαφορές στη συμπεριφορά κάθε χώρας όσον αφορά τις προτιμήσεις κατηγοριών προϊόντων και τις κατανομές πωλήσεων, τα μοναδικά χαρακτηριστικά κάθε αγοράς εμποδίζουν τον εντοπισμό ενός σαφούς συνολικού προτύπου για το ποσό πωλήσεων

CustomerLocation	ProductCategory	sum	mean	count	Percentage
Australia	Clothing	29858	785,7368	38	28,85053917
	Electronics	44481	988,4667	45	42,98013373
	Home Appliances	29153	1005,276	29	28,1693271
Canada	Clothing	53342	874,459	61	29,77787454
	Electronics	67101	1016,682	66	37,45875969
	Home Appliances	58690	863,0882	68	32,76336577
India	Clothing	67324	1141,085	59	38,07057227
	Electronics	59513	1008,695	59	33,65358516
	Home Appliances	50003	1063,894	47	28,27584257
Japan	Clothing	26493	756,9429	35	27,85013719
	Electronics	34155	948,75	36	35,90463275
	Home Appliances	34479	1014,088	34	36,24523006
UK	Clothing	64328	1109,103	58	31,53843514
	Electronics	66951	956,4429	70	32,82442748
	Home Appliances	72688	1038,4	70	35,63713738
USA	Clothing	71408	1034,899	69	35,40237179
	Electronics	68662	967,0704	71	34,04097093
	Home Appliances	61634	948,2154	65	30,55665728

table2

6. CLUSTERING

Επιλέξαμε όλα τα χαρακτηριστικά εκτός από την Product Rating και τις μεταβλητές με ημερομηνίες για το μοντέλο μας. Τρέξαμε τον αλγόριθμο 20 φορές, μία για κάθε τιμή K από 1 έως 20 και καταγράψαμε τις τιμές των inertia και silhouette. Οι τιμες του inertia ήταν υψηλές για τα δεδομένα μας και οι τιμές της μετρικής silhouette ήταν πολύ χαμηλές. Το μοντέλο δεν μπορούσε να αποτυπώσει αποτελεσματικά καμία ουσιαστική δομή μέσα στα δεδομένα μας.

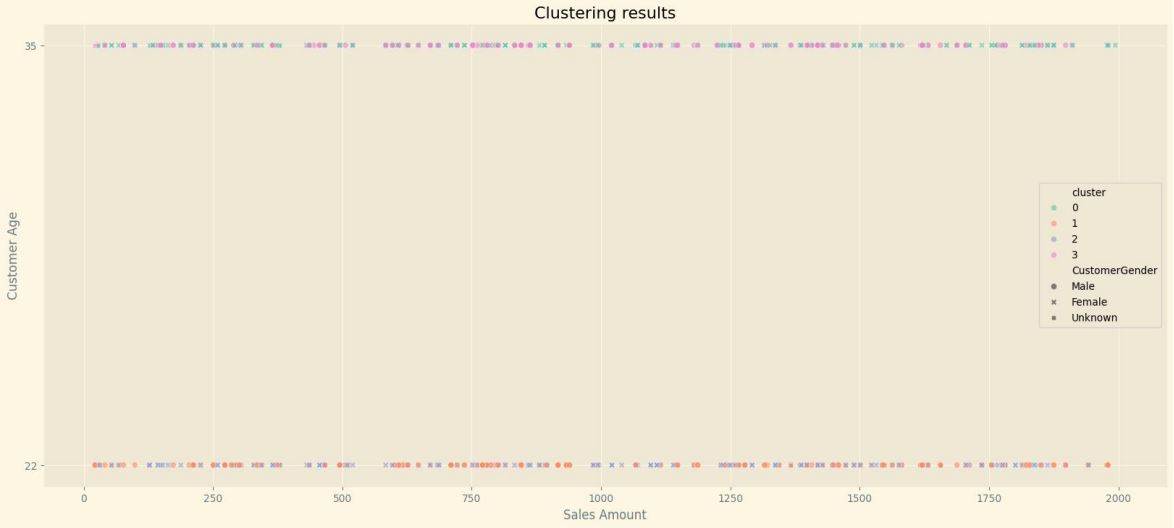
Μετά την ανάλυση των μετρικών, αποφασίσαμε να αναλύσουμε περαιτέρω τα στατιστικά του μοντέλου με 4 ομάδες (clusters) επειδή η μείωση της αδράνειας μετά το K=4 ήταν σημαντικά μικρότερη σε σύγκριση με τις προηγούμενες αυξήσεις.(table3)

K	1	2	3	4	5	6	7	8	9	10
inertia	2537,522	2047,449014	1808,065	1580,819	1492,513	1402,959	1316,765	1233,118	1153,884	1074,578
silhouette	Nan	0,190467956	0,17369	0,223367	0,199511	0,204418	0,21824	0,224269	0,250538	0,273321
drop	Nan	490,0729059	239,3845	227,2458	88,30595	89,55402	86,19407	83,64622	79,23424	79,30584
drop_percentage	Nan	19,31%	11,69%	12,57%	5,59%	6,00%	6,14%	6,35%	6,43%	6,87%

K	11	12	13	14	15	16	17	18	19	20
inertia	999,9832	922,2560955	902,9662	887,3382	865,8994	851,09	832,8271	813,759	794,0584	773,2351
silhouette	0,297606	0,322908438	0,312326	0,300078	0,290906	0,285326	0,275619	0,283069	0,288646	0,281806
drop	74,59521	77,72711162	19,28989	15,62805	21,43877	14,80935	18,26296	19,06806	19,70058	20,82337
drop_percentage	6,94%	7,77%	2,09%	1,73%	2,42%	1,71%	2,15%	2,29%	2,42%	2,62%

table3

Ο αλγόριθμος χώρισε τα δεδομένα μας σε ομάδες που περιείχαν άνδρες ηλικίας 22 ετών, γυναίκες ηλικίας 22 ετών, γυναίκες ηλικίας 35 ετών και άνδρες ηλικίας 35 ετών. Οι καταχωρήσεις με άγνωστο φύλο διαχωρίστηκαν. Αυτή δεν ήταν μια καλή ομαδοποίηση των δεδομένων μας, απλώς χώρισε τα δεδομένα βάσει των δύο στηλών με τις λιγότερες κατηγορίες. Όπως δείχνει η οπτικοποίηση, τα αποτελέσματα της ομαδοποίησης αλληλοκαλύπτονται. (graph15)



graph15

Δεδομένου ότι οι περισσότερες από τις μεταβλητές μας ήταν κατηγορηματικές, η μέθοδος K-means πιθανότατα δεν ήταν η βέλτιστη επιλογή για ομαδοποίηση. Ίσως να εξετάζαμε και εναλλακτικές μεθόδους ομαδοποίησης, όπως οι K-modes ή K-prototypes. Ωστόσο, υπάρχει πιθανότητα ότι και αυτές οι μέθοδοι θα αποδώσουν επίσης χαμηλά λόγω των χαρακτηριστικών των δεδομένων μας.

Αυτό το αποτέλεσμα, σε συνδυασμό με τα άλλα ευρήματά μας, υποδηλώνει ότι τα δεδομένα μας έχουν μεγάλη ποικιλία και όλες οι κατηγορίες συμπεριφέρονται με τρόπο που καθιστά δύσκολη την πρόβλεψη της αγοραστικής συμπεριφοράς βασιζόμενοι μόνο στα δημογραφικά στοιχεία των πελατών ή την κατηγορία προϊόντος. Δεν μπορέσαμε να κατηγοριοποιήσουμε επιτυχώς τους πελάτες με βάση τα χαρακτηριστικά των αγορών τους.