Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Name: Sotirios

Surname: Karageorgopoulos

Username: [me2137@unipi.gr](mailto:me2137@unipi.gr)

Password: 1k2A34567!

Business Report

SAS and MSc Information Systems and Services

University of Piraeus

Academic Specialization in

Data Management and Data Analytics

Milestone Project

Table of Contents

[Images 2](#_Toc125714444)

[Introduction 3](#_Toc125714445)

[Findings 3](#_Toc125714446)

[Conclusions 8](#_Toc125714447)

[Recommendations 9](#_Toc125714448)

[Appendix 10](#_Toc125714449)

# Images

# Introduction

This business report presents various findings from data exploration and analysis of a retail shop. More specifically, the demographic and behavioral characteristics of the customers were explored to find the target audience. Afterward, the sales and promotions were examined to provide the necessary information for the financial status of the retail shop. Moreover, the customer profiles were analyzed with data analytics techniques to discover the categories of customers and which products are most important. In addition, research was done to promote a product to a sample of customers that want to buy this product using discount coupons. Finally, recommendations about the necessary organization acts were provided to become more efficient and effective in their operation.

# Findings

The total value of each invoice has been examined, taking into account all types of invoices. The results indicate that the average value of invoices is $964,93. In addition, the invoice with the lowest and maximum value is $5,7 and $8.603,77, correspondingly. The total number of invoices in the period 2010-2011 was 24.214.

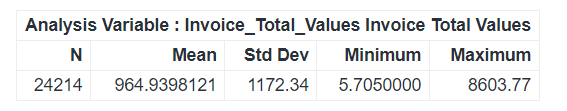


Figure 1 Statistics of invoices total values.

The customers with gender male are 70,16% of the sample, whereas the customers with gender female are 29,84%. Thus, most of the customers are male.

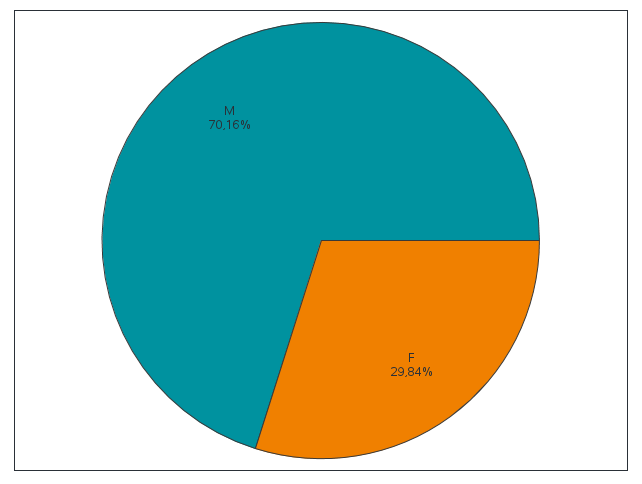


Figure 2 Pie chart with the gender of customers.

Most customers are in the region of Sao Paolo (3.476 customers), Rio de Janeiro (825 customers), and Minas Gerais (822 customers). The rest regions do not have a satisfactory number of customers. The low standard of living is an important factor. Moreover, the population of these regions is less than Southeast regions of Brazil like Sao Paolo, Rio de Janeiro, and Minas Gerais. Thus, the marketing department should investigate whether any promotional activity needs to be done in these areas.

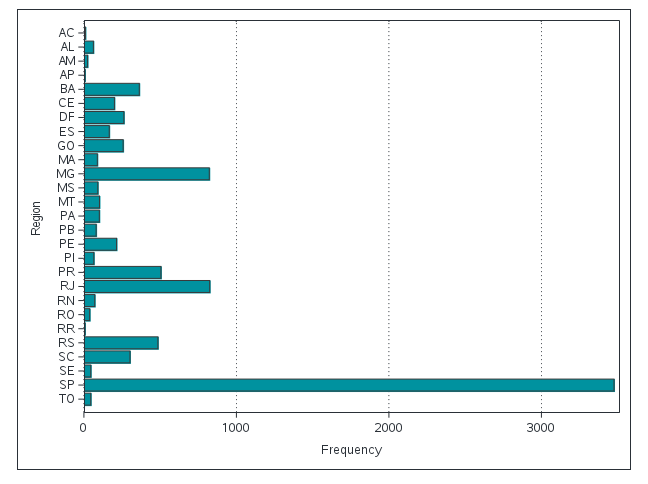


Figure 3 Customers of the retail shop in each region.

The age of most customers is between 28 and 45 years. The median age is 36 years and the average age is 36 years and 9 months. The results indicate that the retail shop is addressed to young and middle-aged people, whereas elderly customers are the minority. The age of most customers is between 28 and 45 years. The median age is 36 years and the average age is 36 years and 9 months. The results indicate that the retail shop is addressed to young and middle-aged people, whereas elderly customers are the minority. More specifically, young and middle-aged customers have a total percentage of 84,26%, whereas mature and elderly customers have a total percentage of 15,77%. In addition, young and middle-aged customers have spent $16.973.871,62, whereas mature and elderly customers have spent $3.255.011,45. It is worth mentioning that the mature and elderly groups are the minority of customers but make more expensive purchases. The young and middle-aged customers have visited the retail shop 18.096 visits, whereas mature and elderly customers have visited the retail shop 3.353 visits. Finally, all age groups of customers have bought all kinds of products from the retail shop. Consequently, the marketing department will have to decide if it wants to attract elderly customers because they spend lots of money on their purchases.

Εικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα

Figure 4 The number of customers for each age group.

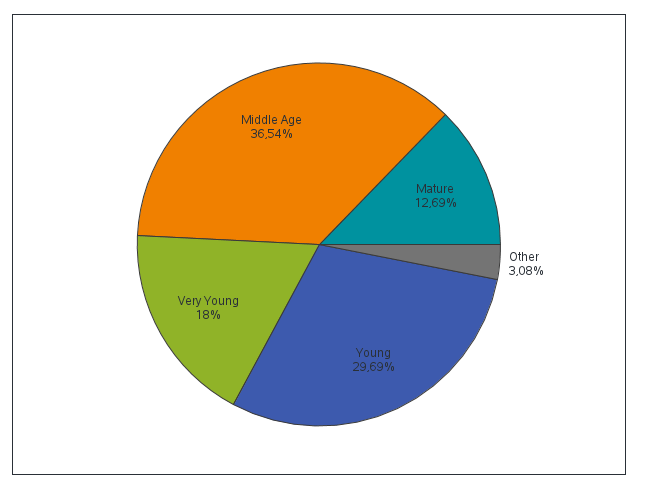


Figure 5 The percentages of customers per age group.

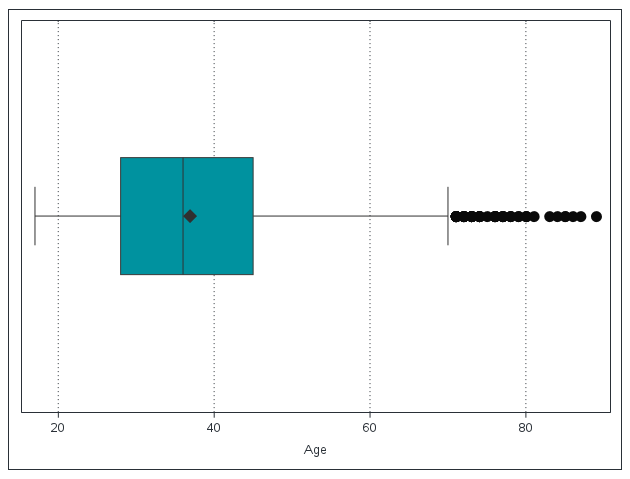


Figure 6 Boxplot with ages of customers.

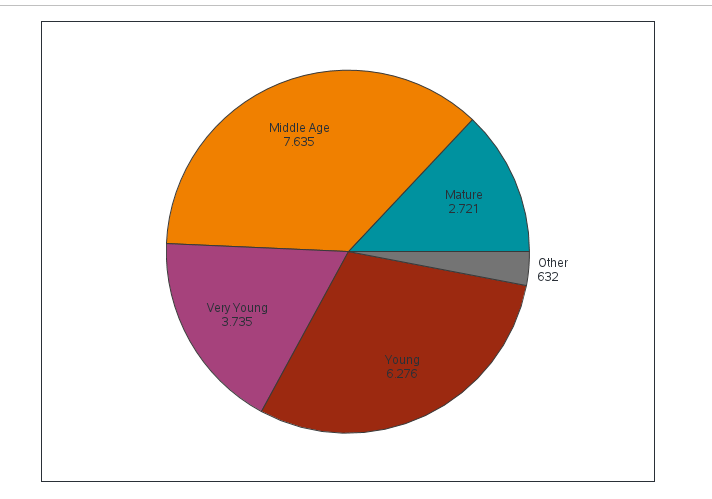


Figure 7 The visits of customers per age group.

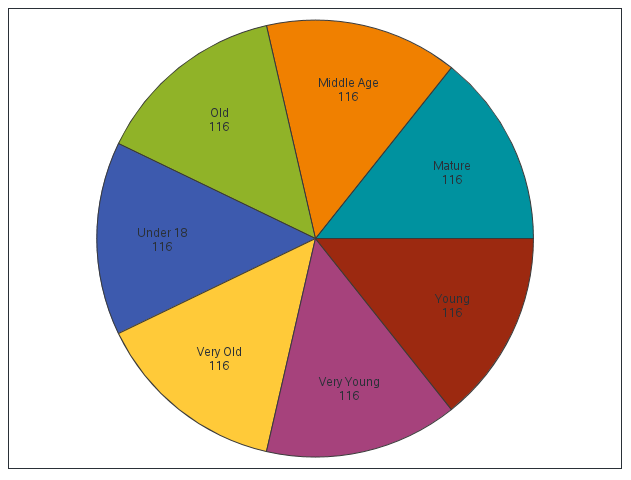


Figure 8 The number of SKUs purchased by age group.

In the period 2010-2011, the sales reached 19,6 million dollars, whereas the returns reached 3,27 million dollars, according to the bar chart of Figure 9. Thus, the benefits are 15,88 million dollars approximately. The customers buy two products per invoice. The invoice with the maximum number of products sold contains 303 products, whereas the invoice with the minimum number of products sold contains only one product.

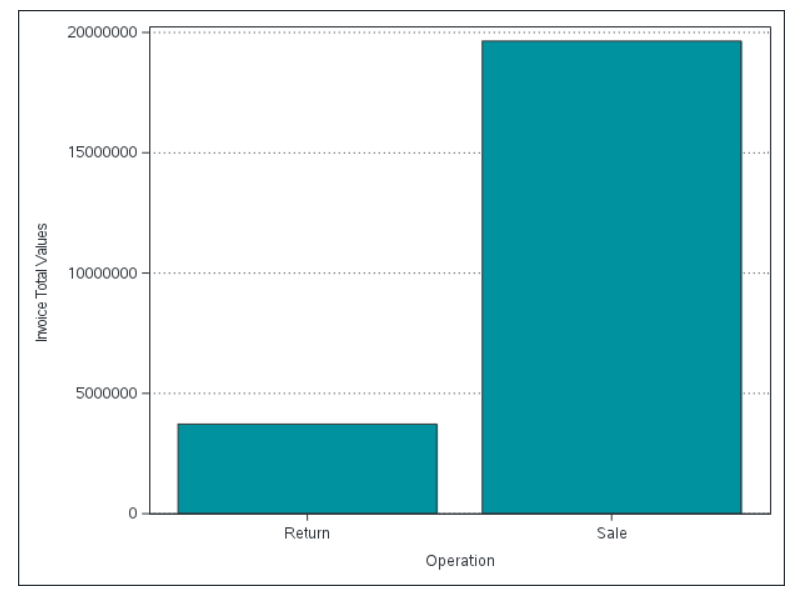


Figure 9 The monetary value of sales and returns.

For each day, the average basket size ranged between one to two products, according to the histogram in Figure 10. 57,77% of days have an average basket size of value two. The same is observed in the scatter plot of Figure 11, in which the majority of points are located near value 2.

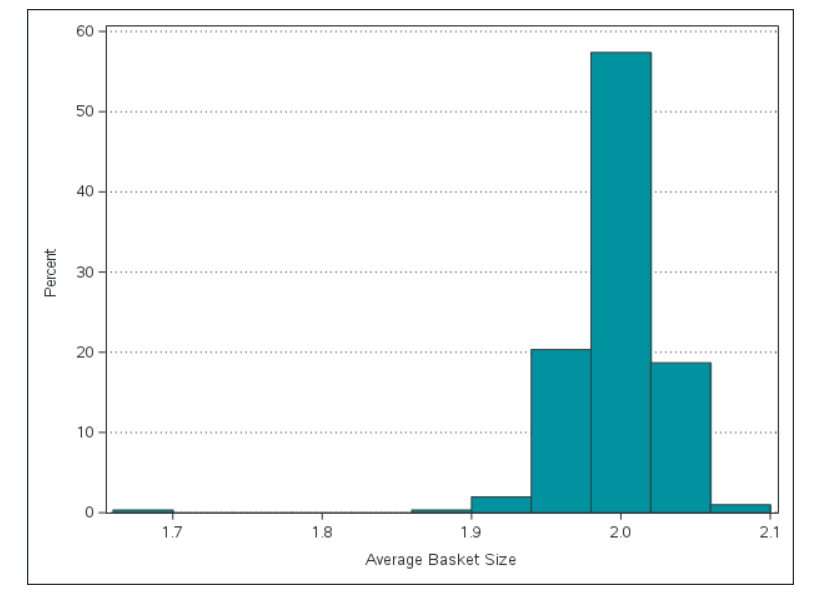


Figure 10 The histogram of average basket size per day.

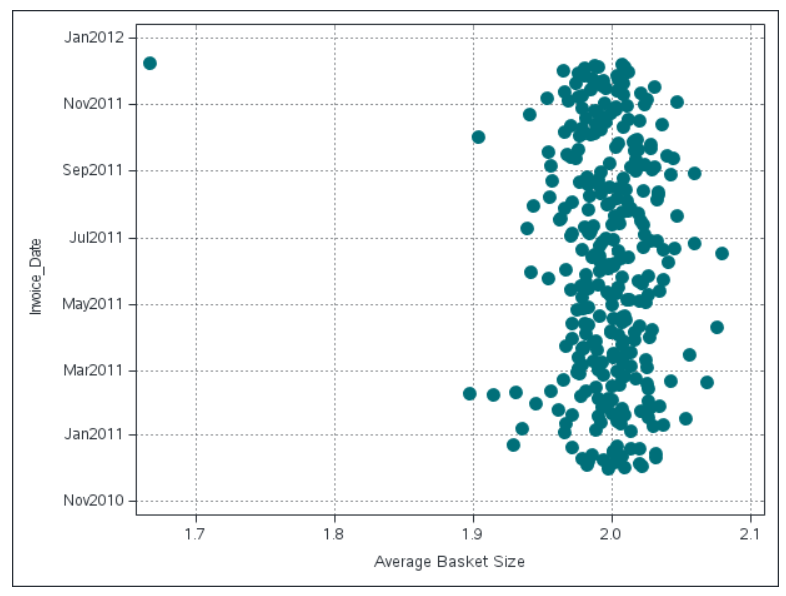


Figure 11 The scatter plot of average basket size per day.

The products that have the most sales are Eyewear (2.906.846,55$), Watches ($2.024.325,73), and Lanterns ($1.886.076,90). The products that have the fewest sales are Woods ($590.351,76), Safety ($441.357,80), and Putters ($428.003,04). In addition, the product line with the most sales is Personal Accessor ($8.897.252,57), whereas the product line with the fewest sales is Golf Equipment ($428.003,04). These results show that the products that belong to the Golf Equipment product line should be promoted.

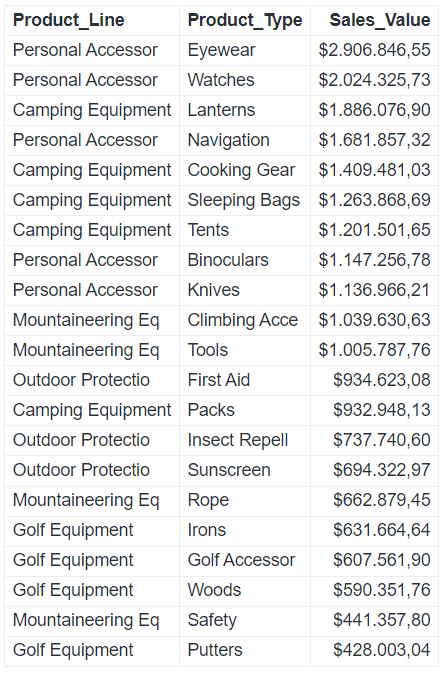


Figure 12 The monetary values of sales per product line and product type.

Εικόνα που περιέχει πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα

Figure 13 The monetary value of sales per product line.

The region with the most sales is Sao Paolo ($8.100.532,58), whereas the region with the fewest sales is Amapa ($7.620,04). The results show that the southeast regions of Brazil contribute more to the retail profits, like Sao Paolo, Rio de Janeiro, and Rio Grande do Sul. In Sao Paolo, the majority of customers are male. Thus, the marketing department should attract more female customers in Sao Paolo.

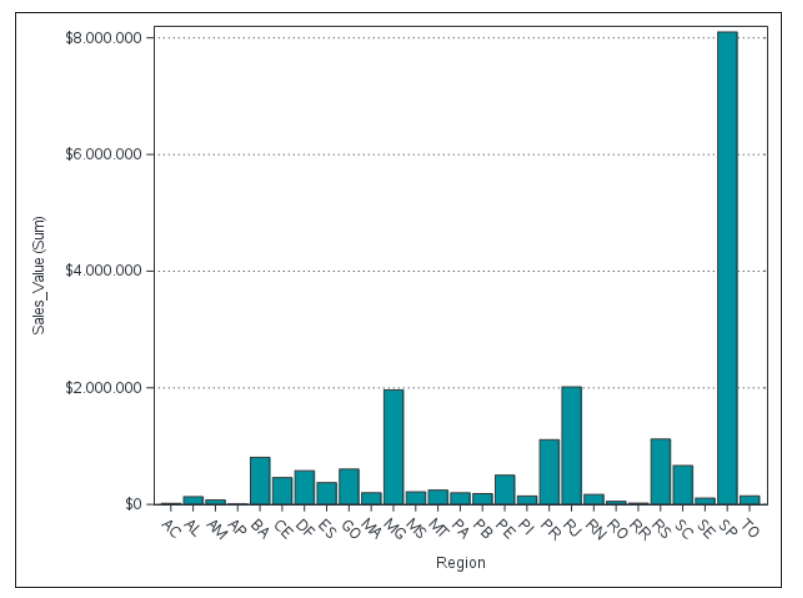


Figure 14 The monetary value of sales per region in bar chart.

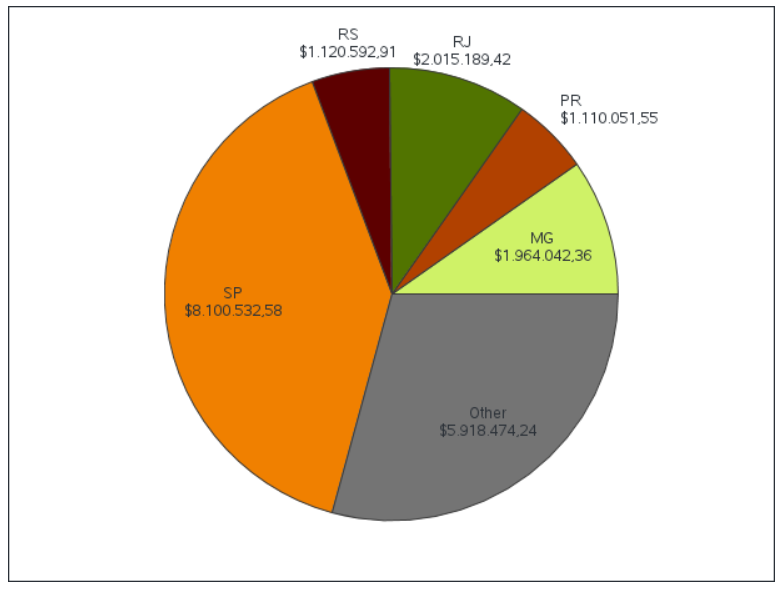


Figure 15 The monetary value of sales per region in pie chart.

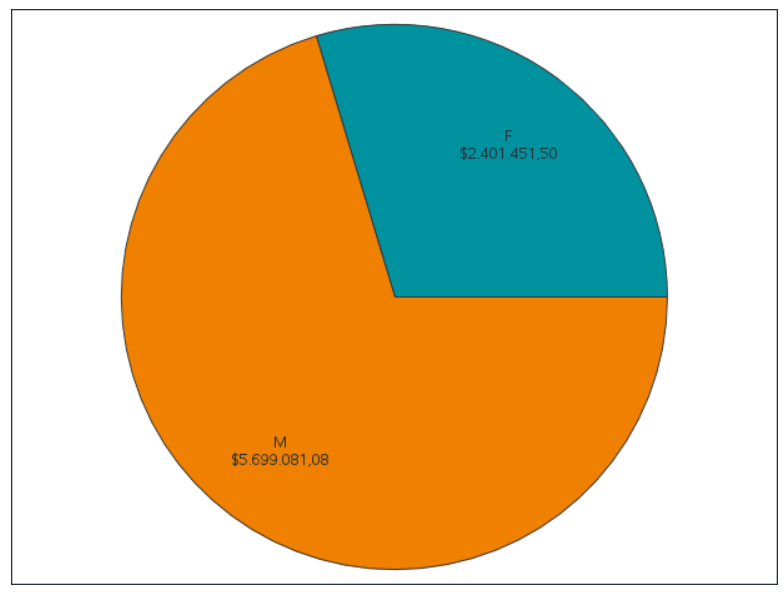


Figure 16 The monetary value of sales per gender in Sao Paolo region.

The number of products sold is 220.061 (62,5%) without promotion, whereas the number of products sold is 132.022 (37,5%) with promotion. In addition, the percentages for products with percentages of 10%, 20%, and 30% are equal, as shown in Figure 19. Thus, the customers consider that discounts are not a key factor in choosing the product, or the discounts are not notably satisfying to buy them.

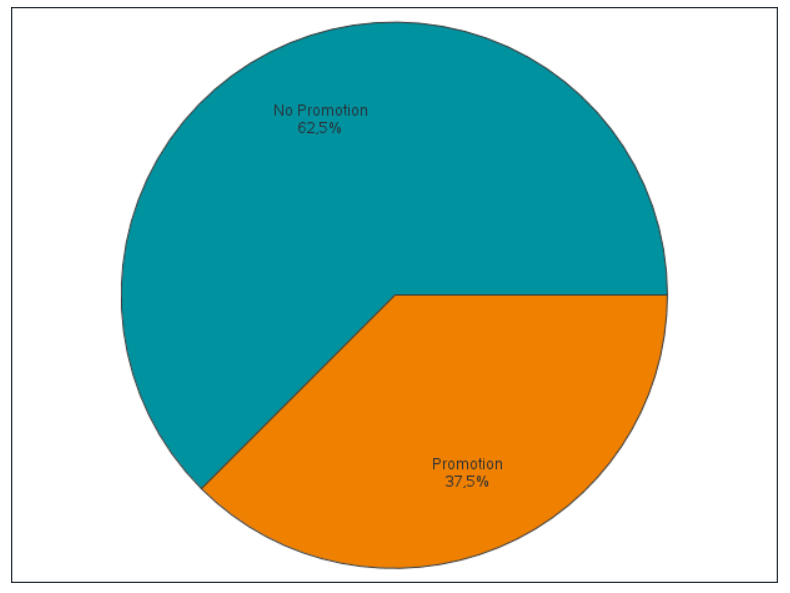


Figure 17 The percentage of products that are promoted and no promoted in pie chart.

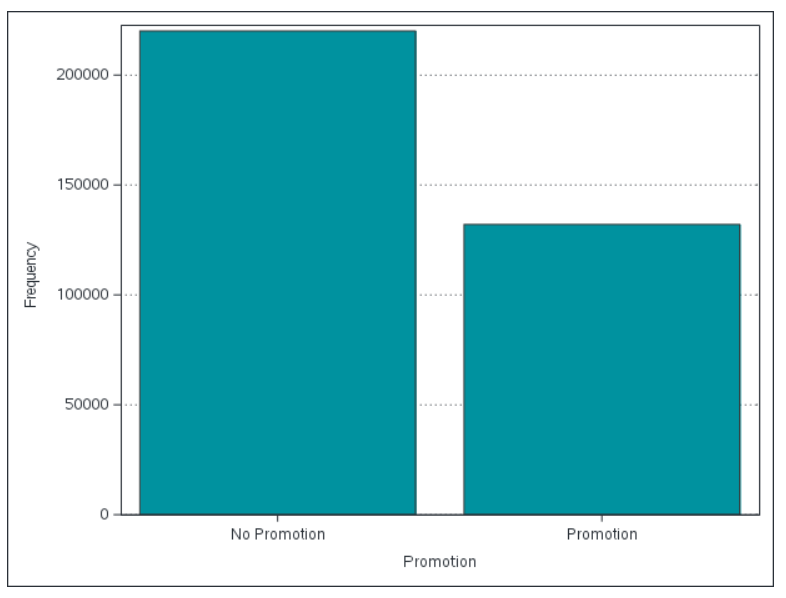


Figure 18 The number of products that are promoted and no promoted in bar chart.

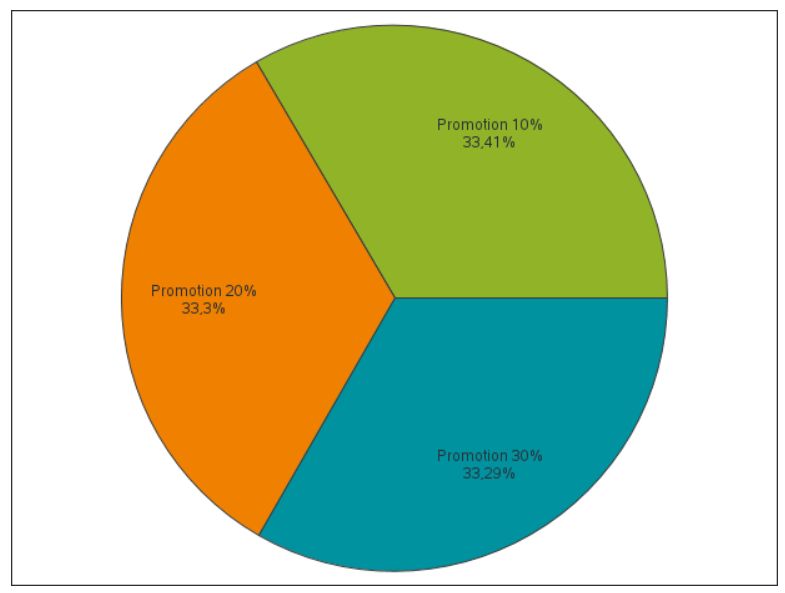


Figure 19 The percentage of promoted products per category.

The sales with a high monetary value are on Thursday ($4.270.000), whereas the sales with the lowest monetary value are on Saturday ($0). The monetary value of sales on Saturday is $0 because the retail shop is closed on Saturdays. In figure 20, it is observed that all SKUs are sold, so there is no difference between different SKUs and sales.

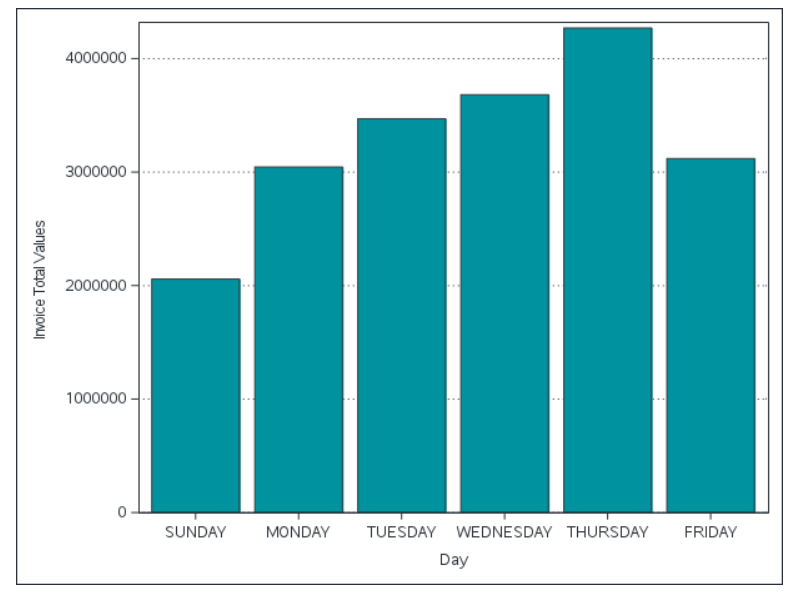


Figure 20 The sales per day of the week.

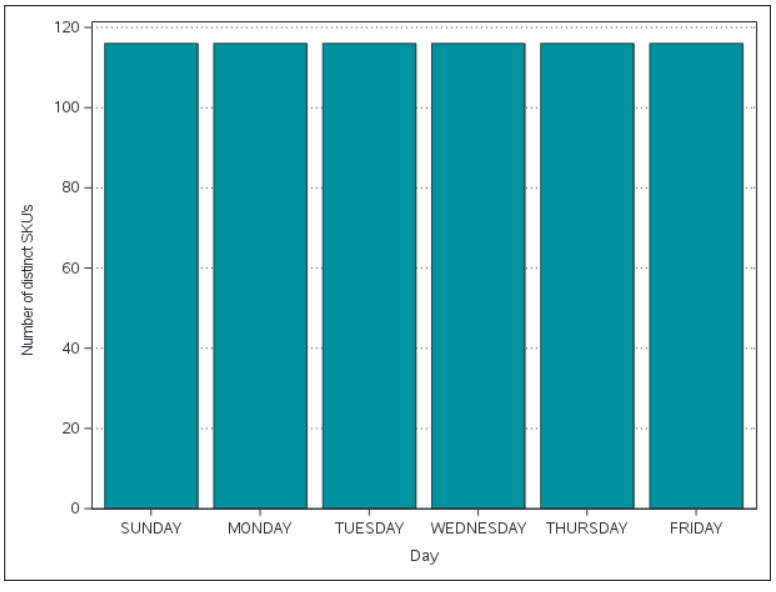


Figure 21 Different SKUs per day of the week.

The most products that have been sold are of the company Dragon SA. More specifically, the products that have been sold are 66.677, which consists the 15,92% of products sold. On the other hand, Easy Creator was the company with the fewest products sold. The Easy Creator products that have been sold are 34,988, which consists the 8,35% of products sold. The percentages of the products sold by the remaining companies are between 9 and 11 %. The Toktai and Chen products that have been sold are 60.790, which consists the 14.52% of the total products sold.

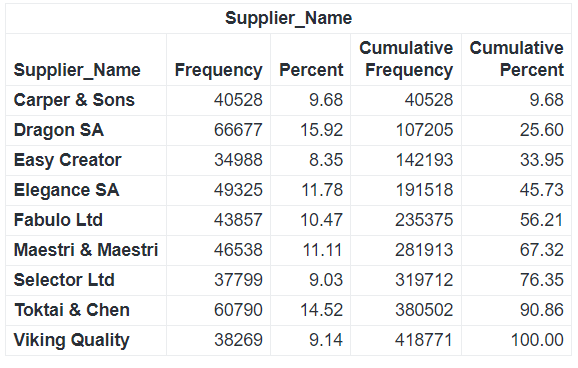


Figure 22 The frequency report of products sold.

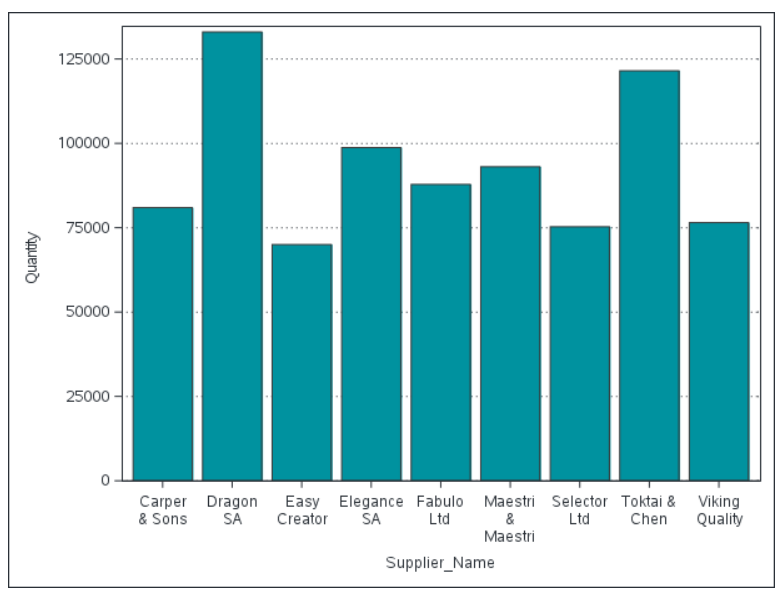


Figure 23 The quantity of products sold in bar chart

The company with the highest sales value is Dragon SA, at $3.809.087,42 which consists the 16,3% of sales. Indeed, the company with the lowest sales value is Selector Ltd, at $1.913.426,29 which consists the 8,2% of sales. The sales of the rest companies are between 2 to 3.1 million dollars.

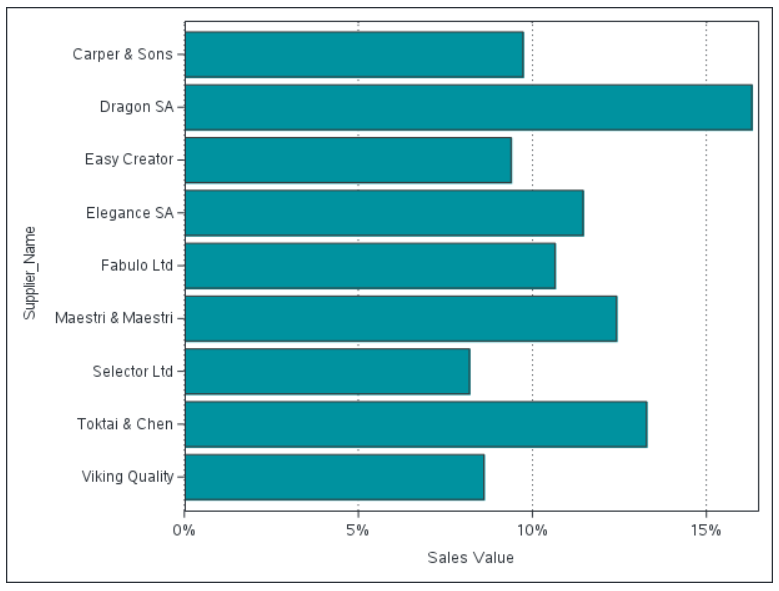


Figure 24 The percentage of sales per company.

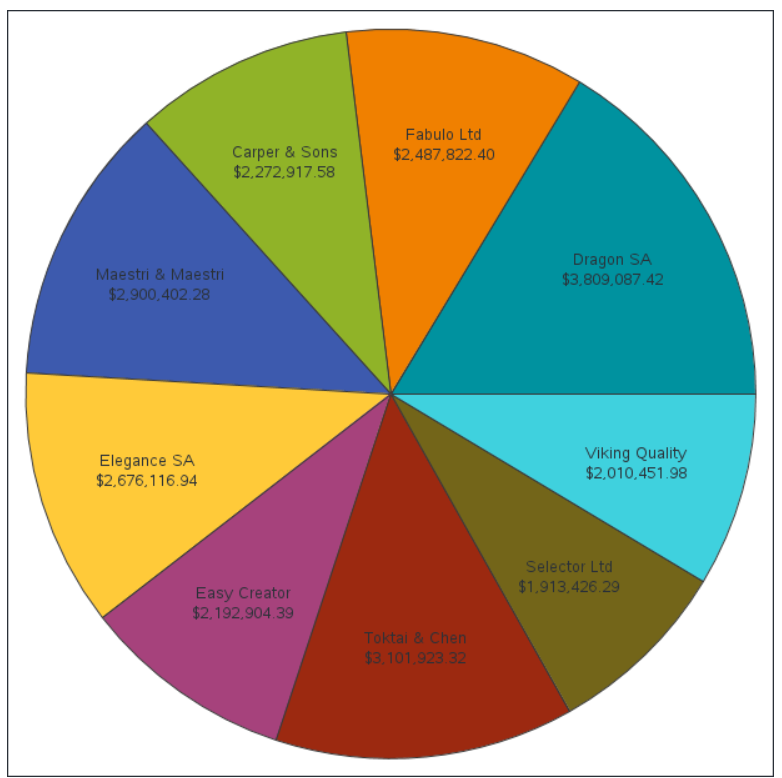


Figure 25 The value of sales per company.

The products of **Carper & Sons** that have been made in India, have the **highest** sales value at **$790,209.95**, whereas the products that have been made in Turkey have the **lowest** sales values at **$247,867.87**.

The products of **Dragon SA** that have been made in Spain have the **highest** sales values at **$900,410.39**, whereas the products that have been made in Turkey have the **lowest** sales values at **$643,193.34**.

The products of **Easy Creator** that have been made in China have the **highest** sales values at **$802,474.13**, whereas the products that have been made in Turkey have the **lowest** sales values at **$132,038.42**.

The products of **Elegance SA** that have been made in the China have the **highest** sales value at **$1,033,999.05,** whereas the products that have been made in India have the **lowest** sales value at **$96,527.05**.

The products of **Fabulo Ltd** that have been made in India have the **highest** sales values at **$931,715.44**, whereas the products that have been made in Spain have the **lowest** sales value at **$168,546.28**.

The products of **Maestri & Maestri** that have been made in Spain have the **highest** sales value at **$1,044,866.08**, whereas the products that have been made in China have the **lowest** sales value at **$67,752.30**.

The products of **Selector Ltd** that have been made in Spain have the **highest** sales value at **$890,860.32**, whereas the products that have been made in China have the **lowest** sales value at **$91,108.40**.

The products of **Toktai & Chen** that have been made in China have the **highest** sales value at **$849,079.28**, whereas the products that have been made in Turkey have the **lowest** sales value at **$402,765.73**.

The products of **Viking Quality** that have been made in the USA have the **highest** sales value at **$664,172.77**, whereas the products that have been made in India have the **lowest** sales value at **$313,966.67**.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Figure 26 The table with sales per products origin country and company.

Consequently, Dragon SA and Toktai & Chen contribute to sales more than the rest companies. Selector Ltd is the company with the fewest sales, so the products of Selector Ltd should be promoted by the marketing department. The products made in Turkey have the fewest sales, which indicates that customers do not trust Turkish products. On the other hand, the products made in Spain and China have the most sales.

RFM analysis was performed, and the customers were segmented into 3 groups based on their profile, i.e., best customers, worst customers, and churners. The most representative customer of the best customers has spent $3.242,10, made the last purchase 1 month ago, and has bought 58 times from the shop. In addition, the most representative customer of the worst customers has spent $277,45, made the last purchase 3 months ago, and has bought 5 times from the shop. Finally, the most representative customer of the churners has spent $2.364,05, made the last purchase 5 months ago, and has bought 42 times from the shop.

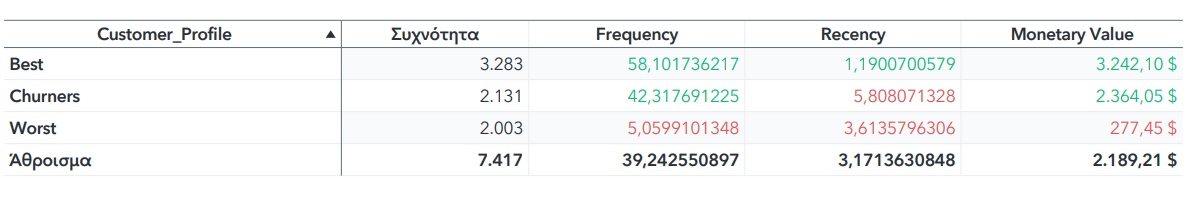


Figure 27 Cross Tabulation of customers' profiles, and the RFM values.

The most significant groups are the worst customers and churners. More specifically, the best customers will continue to buy from the shop, whereas the worst customers and churners are not satisfied. Thus, it is necessary to investigate the characteristics of customers that belong to these groups because they are the most of our registered customers, i.e., 55,7% of the total customers.

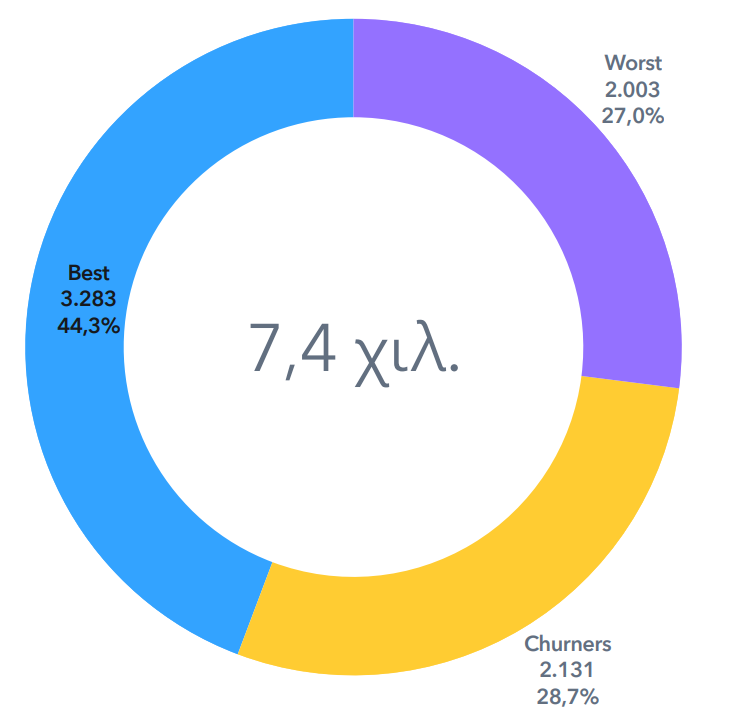


Figure 28 The pie chart of customers' profiles.

The majority of churners are middle-aged customers, i.e. their ages ranged between 36 to 50 years. In addition, the gender of most churners is Male, with a percentage of 71,8%. Finally, all churners are from Brazil. Thus, it is important to be done promotional activities that target middle-aged men and take place in Brazil.

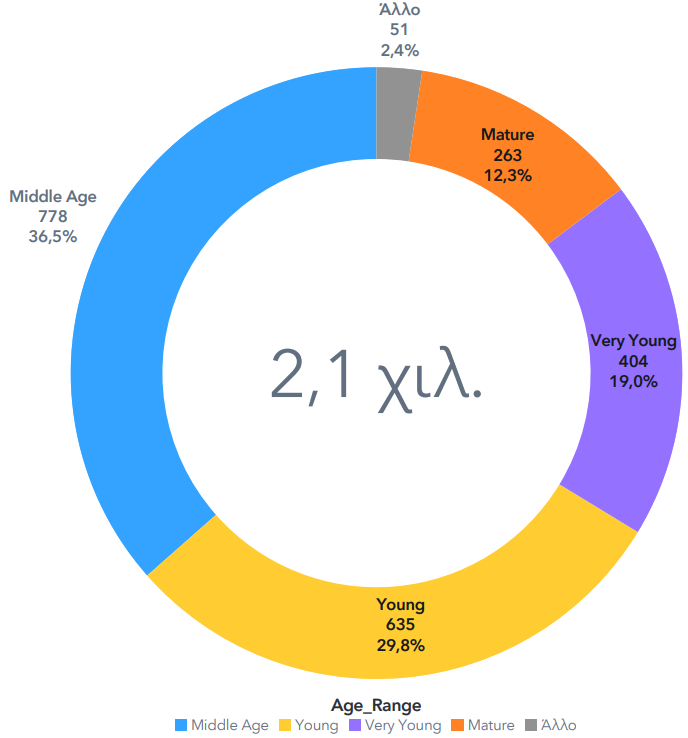


Figure 29 The age range of churners in pie chart.

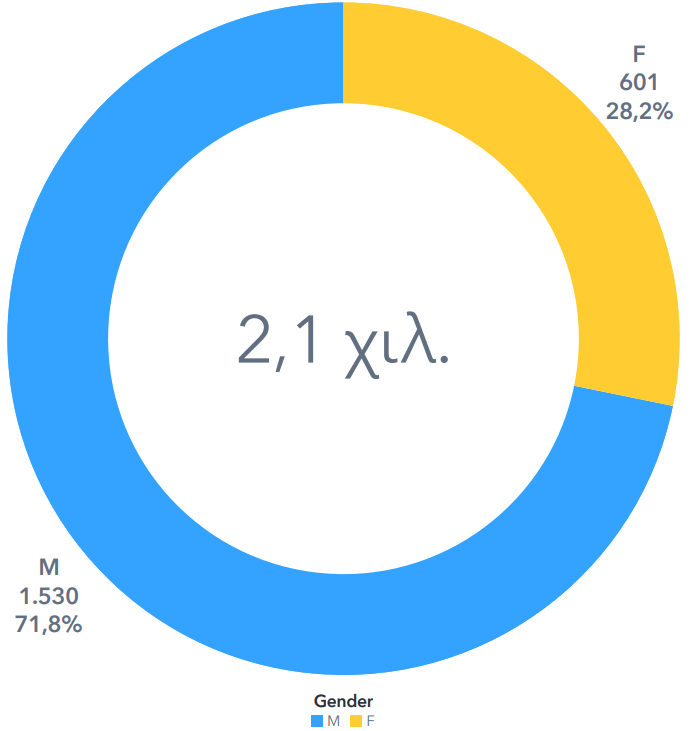


Figure 30 The gender of churners in pie chart.

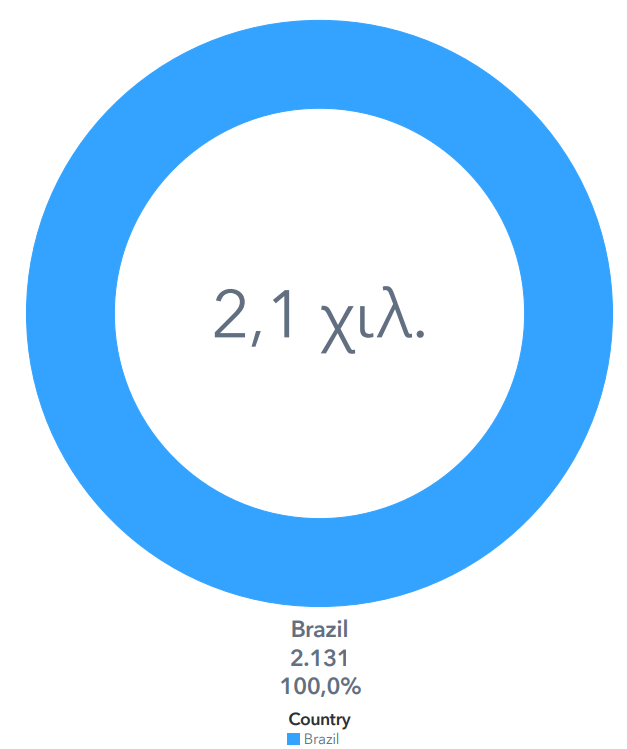


Figure 31 The origin country of churners.

The majority of worst customers are middle-aged customers, i.e. their ages ranged between 36 to 50 years. In addition, the gender of most churners is Male, with a percentage of 71,8%. Finally, worst customers are from Brazil. Thus, it is important to be done promotional activities that target middle-aged men and take place in Brazil.

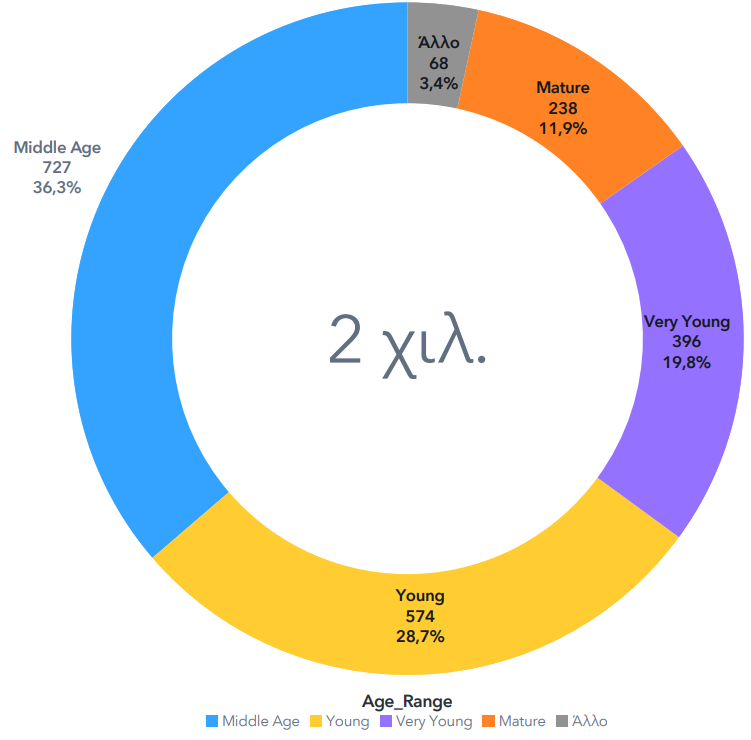


Figure 32 The age range of worst customers in pie chart.

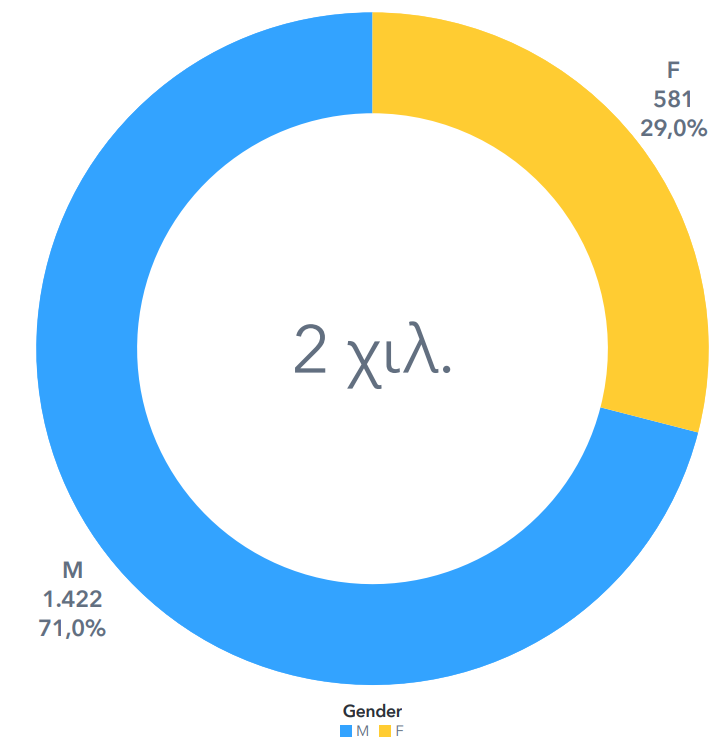


Figure 33 The gender of worst customers.

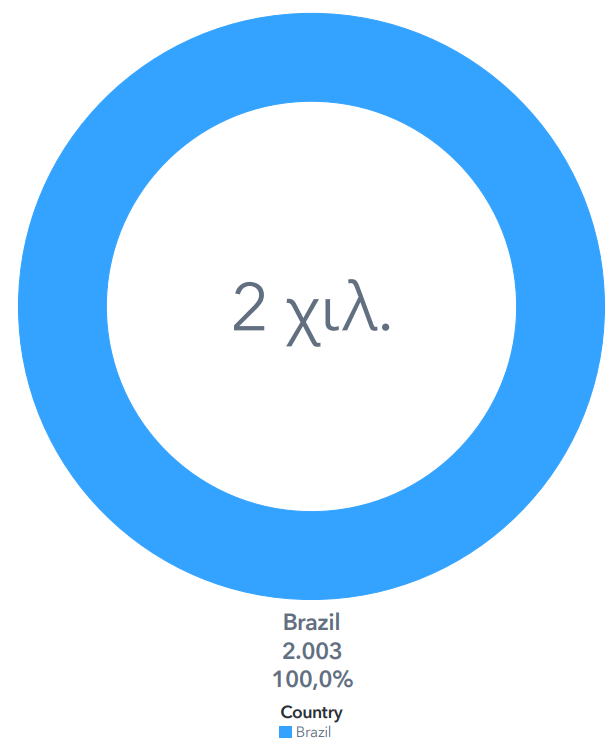


Figure 34 The origin country of worst customers.

If the proportion of buyers and non-buyers for product category 97 was 5% (1200 customers) - 95% (22800 customers) sampling should be done on the entire number of customers who have bought from the store (buyers) and an equal number of customers who have not bought from the store (non-buyers). This is necessary to do so that there is no bias in the results of the model over the customers who do not buy from the store, as they constitute the majority of the sample. Finally, the sample will consist of 1200 customers who bought from the store and 1200 customers who did not buy from the store, i.e. 2400 observations. The variable Customer\_Code has one missing value, whereas the rest variables do not have missing variables. This observation was removed from the dataset to affect the results of the analysis.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Figure 35 The number of missing and not missing values.

The buyers are the 30,2% (631 customers) of the dataset and the non-buyers are the 69,8% (1460 customers) of the dataset.

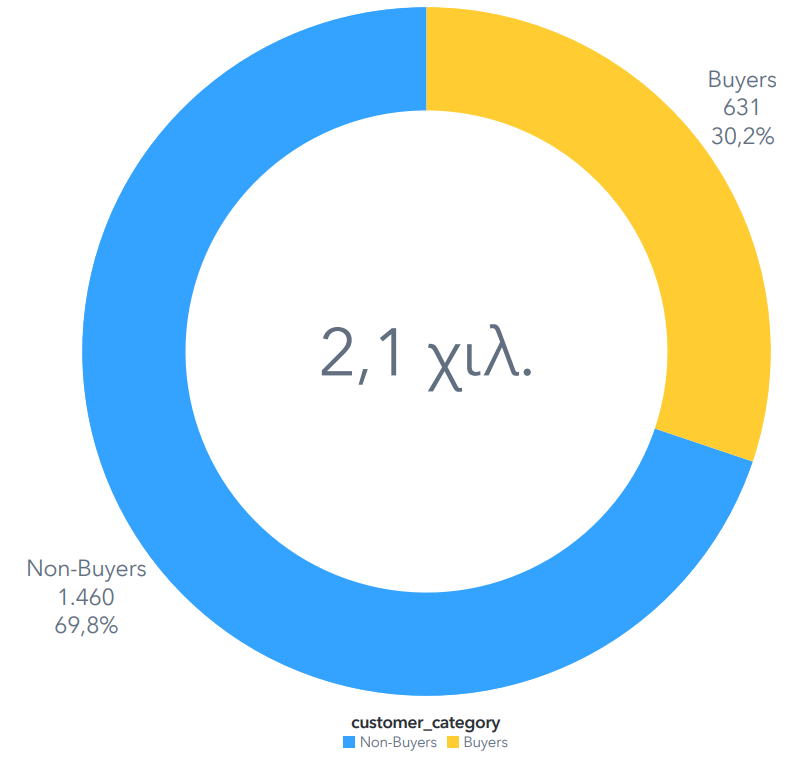


Figure 36 The proportion of buyers and non-buyers in dataset for product category 97.

The percentage of buyers over 40 years is 6% (87 customers) and the percentage of non-buyers over 40 years is 94% (1354 customers).

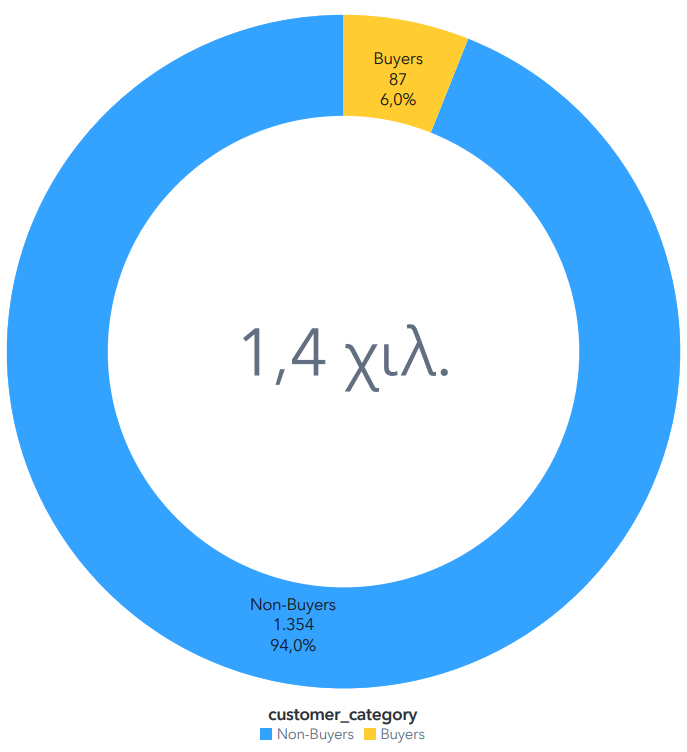


Figure 37 The buyers and non-buyers over 40 years for product category 97.

The average age of buyers is 32 years and non-buyers is 51 years. The age range of buyers is between 18 and 57, whereas the age range of non-buyers is between 26 and 67. It is observed that non-buyers are older than the buyers, so the product is addressed mainly by young and middle-aged.

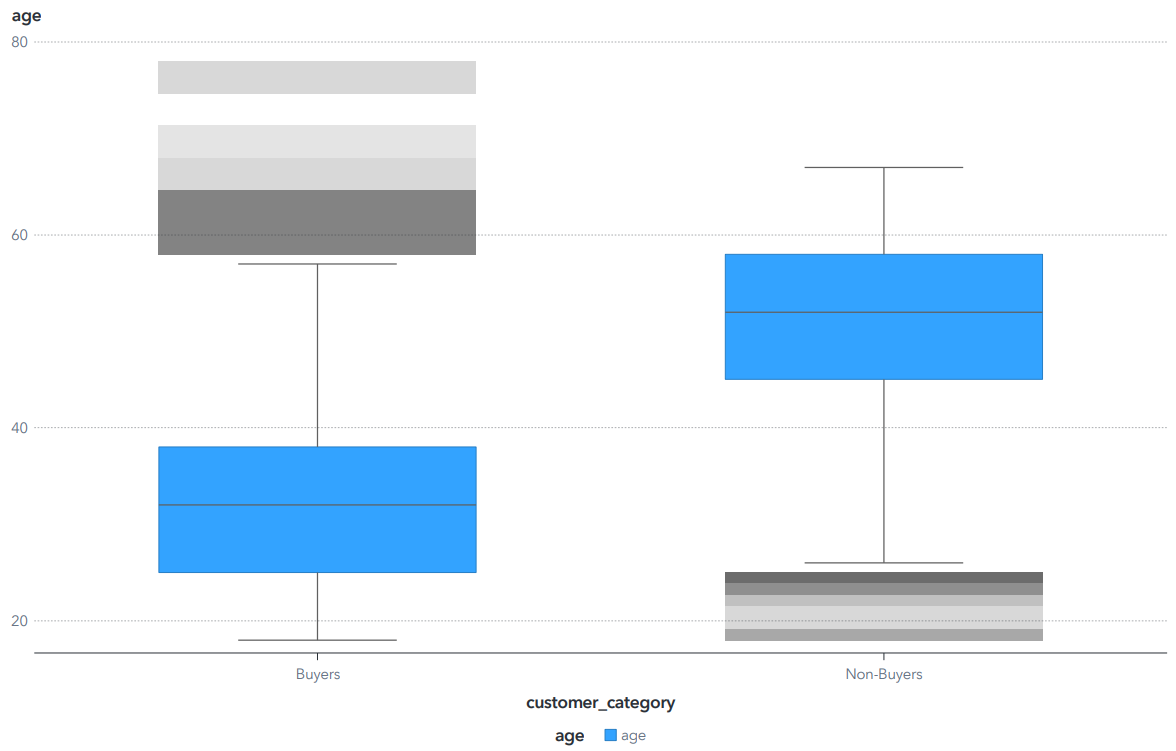


Figure 38 The boxplot for buyers and non-buyers of product 97.

The minimum probability that a customer must have to be considered as the buyer is calculated as follows:

Thus, the minimum probability that a customer must have to be considered as the buyer is at least 16,6%.

The terminal leaves of maximal tree are 6 according to “Pruning Error Plot”. It has been selected validation error so that the decision tree model does not have overfitting. The misclassification rate is 0.0829 in 6 leaves.



Figure 39 Pruning Error Plot.

The KS cutoff reference line is drawn at the value of 1-specificity where the greatest difference between sensitivity and 1-specificity is observed for the validate partition. Thus, the cutoff value of selected model is 0.1, where the 1-specificity value is 0.082 and the sensitivity value is 0.915.

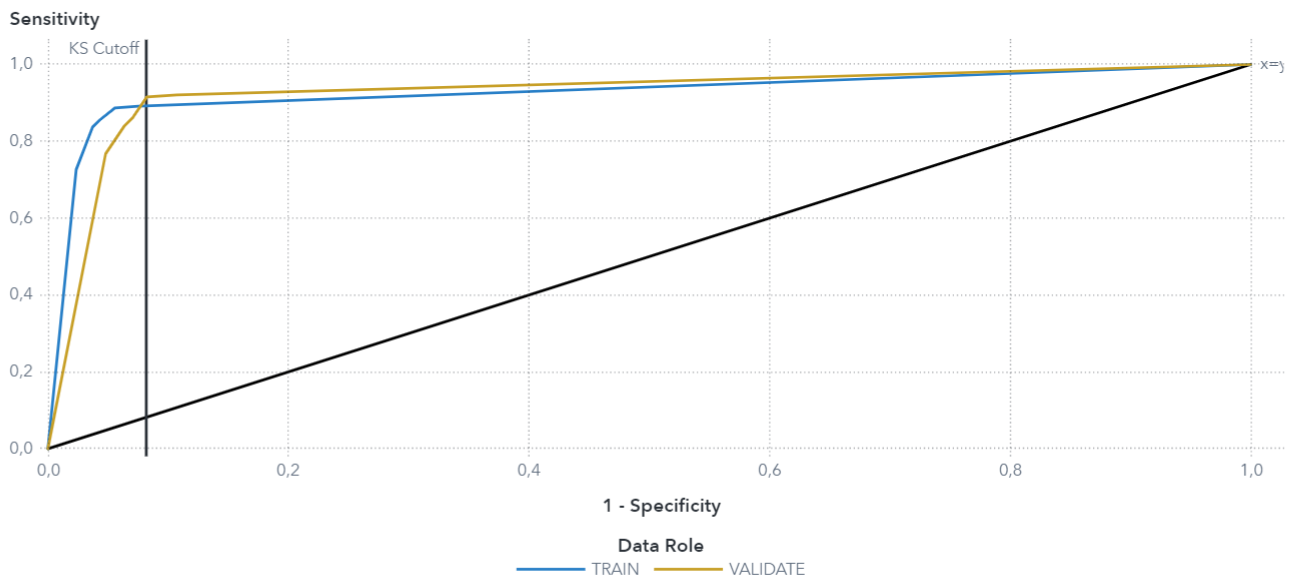


Figure 40 The ROC curve of decision tree model.

The accuracy of the decision tree model for validation partition is 91,71% for cutoff 0.1 which indicates that it is a good predictive model.

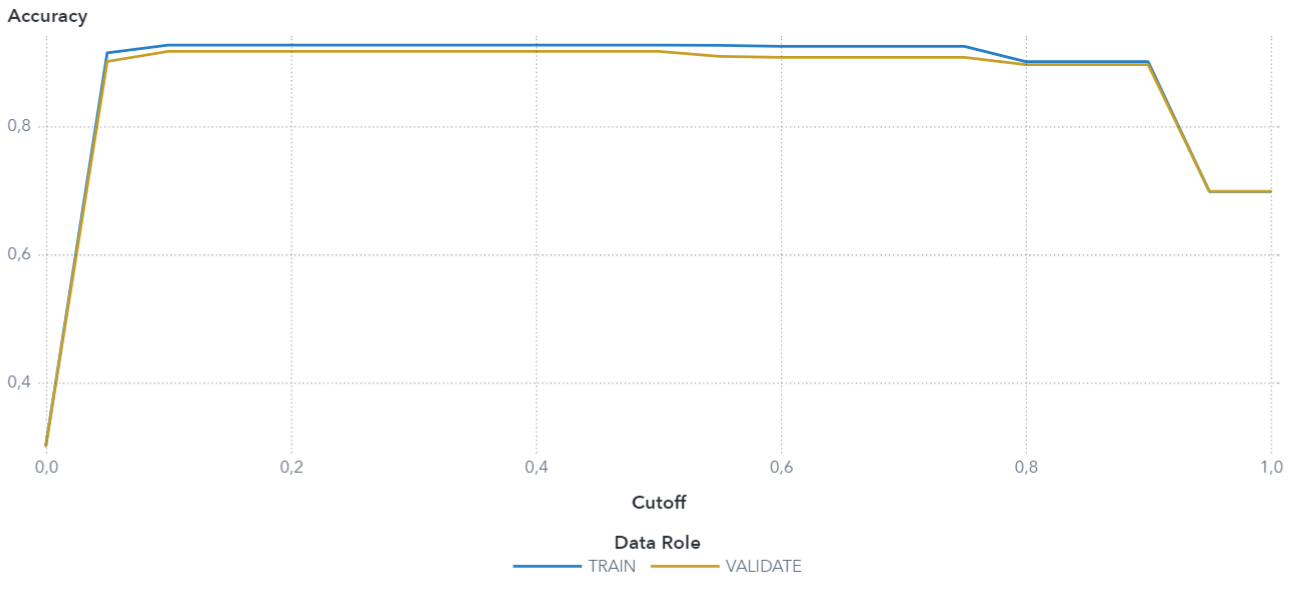


Figure 41 Accuracy for the different cutoff values.

Similarly, the F1-score of the decision tree model for validation partition is 0.87% for cutoff 0.1 which indicates that it is a good predictive model.

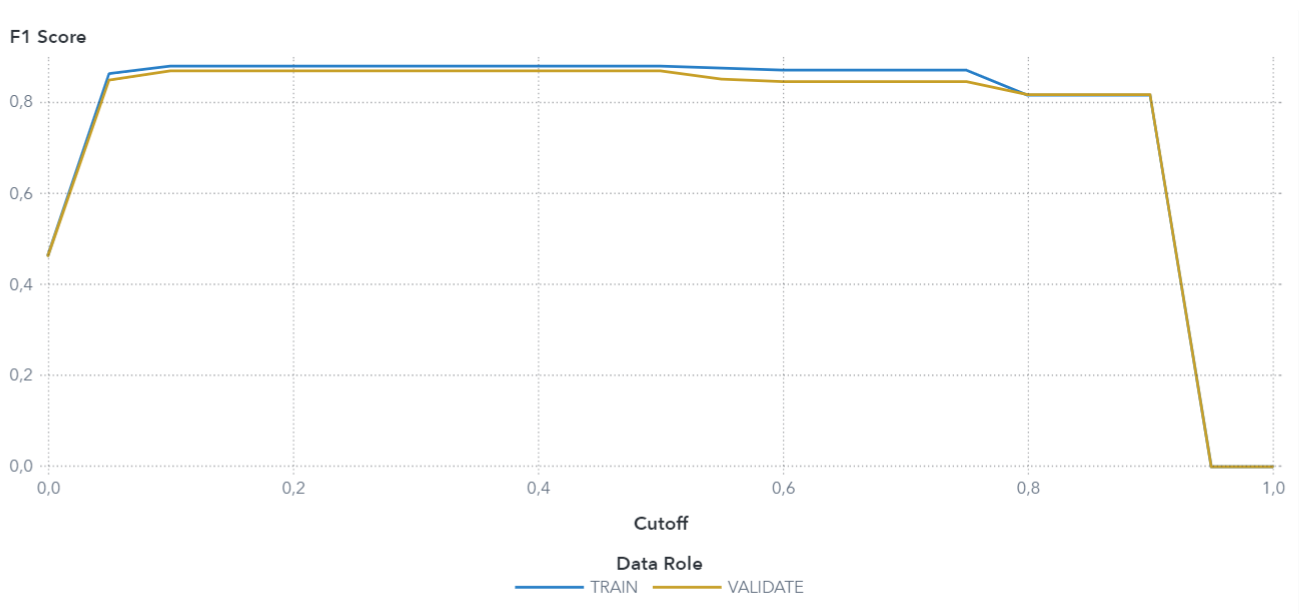


Figure 42 F1 score for the different cutoff values.

The phenomenon that is presented with the line of the training data set is called overfitting. Overfitting exists in a machine learning model which learns the details and the noise in the training data to such an extent that it dents the performance. So, the overfitted model in production will be inaccurate in its predictions.

The decision tree model is the following:

1. age < 41 or age is missing
2. monetary < 133 or monetary is missing
3. frequency < 14 or frequency is missing
4. recency < 7 or recency is missing
5. D = 1

# Appendix