Univariate, Bivariate and Multivariate Data Analysis with Marketing Analytics of Ifood

Davi Barrel Santos, Renatha Vieira

Faculdade de Ciências da Universidade do Porto

1 Introduction

The dataset for this study was obtained from Kaggle, specifically the Marketing Data dataset by Jack Daoud. This dataset was originally presented to potential data analysts as a trial dataset at iFood, which is the largest food delivery business in Brazil and Latin America and one of Brazil major successful technology startups, equivalent to DoorDash in the United States. The dataset can be retrieved from the link, Kaggle Marketing Data.

This dataset includes more than 2,000 rows of customer data, featuring various demographic attributes along with sales information related to the acceptance of six different marketing campaigns, the amounts spent on different food and goods via delivery, and any customer complaints filed. Each marketing campaign entry, labeled Campaign 1 through Campaign 6, is marked as "Accepted" if the customer accepted the offer and "Rejected" otherwise. A 'True' under 'Complain' indicates that the customer filed a complaint in the last two years.

Throughout this assessment, our aim was to uncover consumer behavior trends and evaluate the effectiveness of marketing efforts through bivariate analysis. By examining the relationships between different customer attributes and sales metrics, our goal was to identify links between various customer characteristics and sales performance indicators. These insights are intended to inform strategic decision-making processes, ultimately leading to enhanced marketing strategies and improved customer satisfaction.

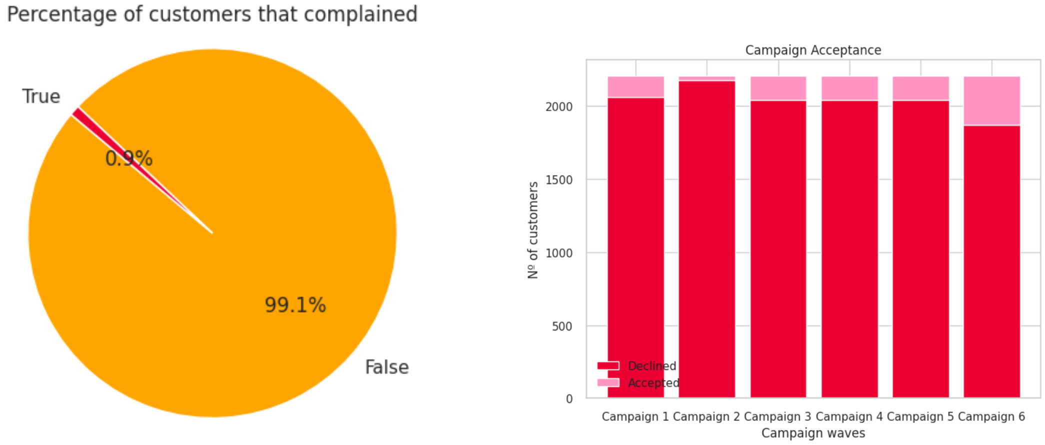
Key columns in the dataset include:

- **Campaign 1-6**: Indicates whether the customer accepted ("Accepted") or did not accept ("Rejected") the offer in each of the six marketing campaigns. - Binary;
- **AcceptedCmpOverall**: The total amount of accepted campaign per customer - Numerical Discrete;
- **Complain**: 'True' if the customer filed a complaint in the last two years. - Binary;
- **Customer_Days**: Represents the number of days since customer enrollment with the company. - Numerical Discrete;
- **Education**: Denotes the customer's level of education. - Categorical Ordinal;
- **Marital**: Indicates the customer's marital status. - Categorical Nominal;
- **Kidhome**: Represents the number of small children in the customer's household. - Numerical Discrete;
- **Teenhome**: Indicates the number of teenage children in the customer's household. - Numerical Discrete;
- **Income**: Represents the customer's yearly household income. - Numerical Discrete;
- **MntFishProducts, MntMeatProducts, MntFruit, MntSweetProducts, MntWines, MntGoldProds, MntTotal, MntRegularProds**: Denote the amount spent on various product categories in the last two years. - Numerical Discrete;
- **NumDealsPurchases, NumCatalogPurchases, NumStorePurchases, NumWebPurchases**: Represent the number of purchases made with discounts, through catalogs, directly in stores, and via the company's website, respectively. - Numerical Discrete;
- **NumWebVisitsMonth**: Indicates the number of visits to the company's website in the last month. - Numerical Discrete;
- **Recency**: Denotes the number of days since the customer's last purchase. - Numerical Discrete;
- **Age**: Indicates the user's age. - Numerical Discrete;

2 Univariate Analysis

2.1 Categorical Variables

When analyzing responses to the campaigns, it was found that all of them were predominantly declined, with the sixth campaign being the most accepted among customers. Despite various promotional efforts, the overall acceptance rate remained low. Additionally, it was found that the percentage of customers who lodged complaints was less than 1% in the last 2 years, indicating a generally satisfactory level of service.

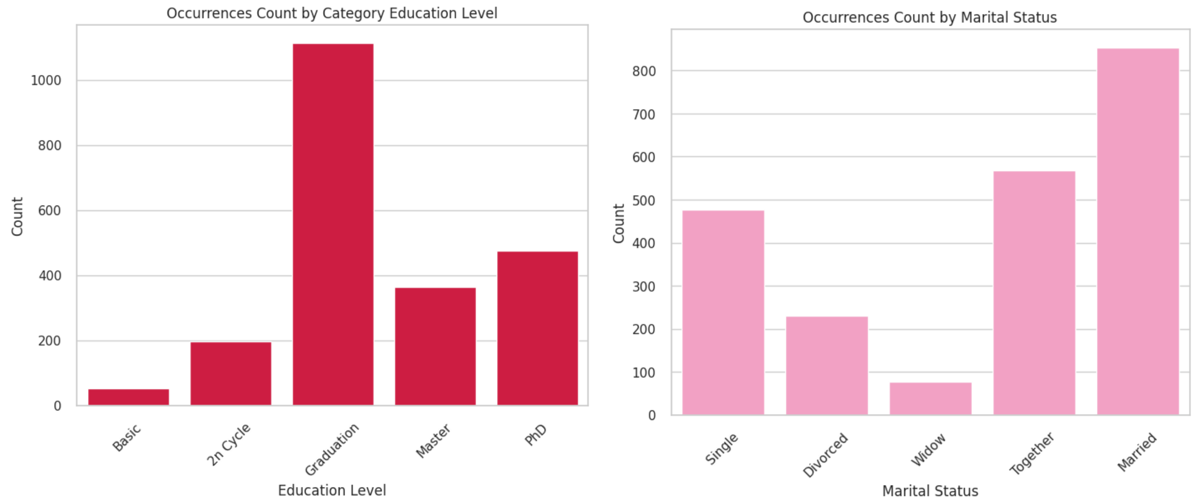


(a) Percentage of Complaing

(b) Campaign Acceptance

Fig. 1: Univariate Analysis of Consumer Behavior

Shifting focus to user demographics, an analysis of education levels revealed an interesting pattern. The majority of users were found to have a bachelor's degree, followed by those with a doctorate, master's degree, high school education, and finally, elementary education. This suggests that the platform predominantly attracts an educated user base, with higher education levels being more prevalent among its customers. Furthermore, it was observed that the overwhelming majority of users are married or in a stable relationship.



(a) Count by Education Level

(b) Count by Marital Status

Fig. 2: Univariate Analysis of Consumer's Demographics Data

2.2 Numerical Variables

For the numerical variables, histograms were plotted to assess their distributions. The majority of the distributions were found to be asymmetric and skewed to the left, indicating that the data is concentrated around the lower values with a longer tail towards the higher values. Additionally, Shapiro-Wilk tests were conducted for all variables, revealing that none of the variables follow a normal distribution.

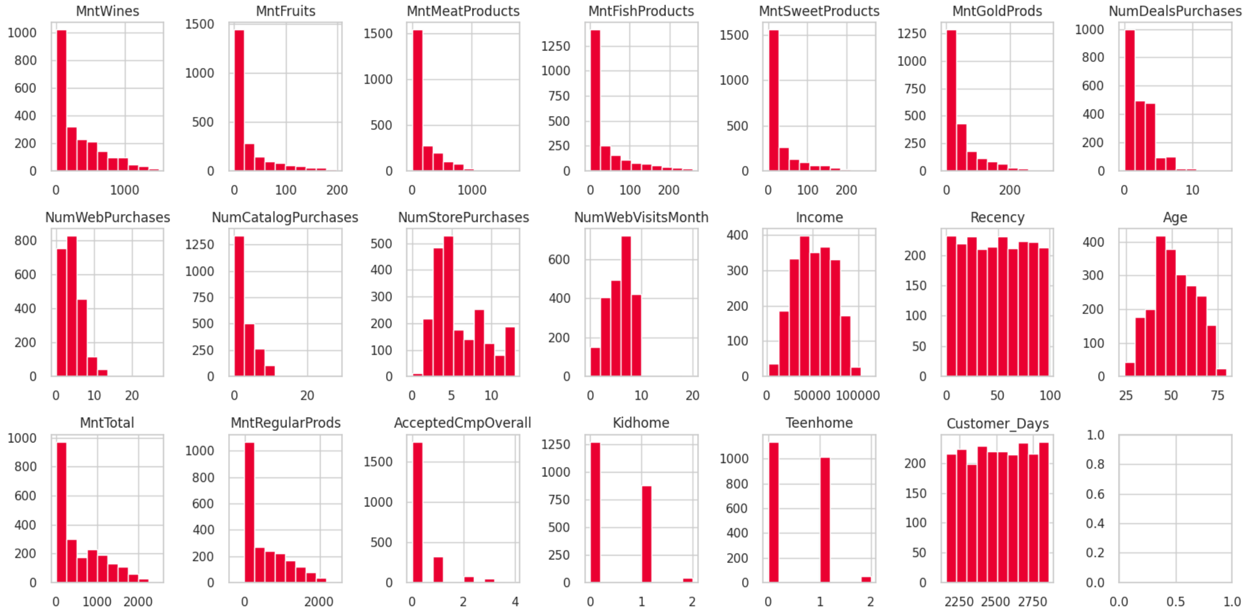


Fig. 3: Histograms of numerical variables

Box plots were generated for each variable, revealing the presence of outliers in almost all variables, with the exceptions of 'NumStorePurchases', 'Income', 'Recency', 'Age', 'Kidhome', and 'Teenhome'. Upon examination of the graphs, it was observed that the interquartile ranges (IQRs) for most variables were relatively small, suggesting that the data points are concentrated within a narrow range. This observation is further supported by the proximity of the first and third quartiles in many variables.

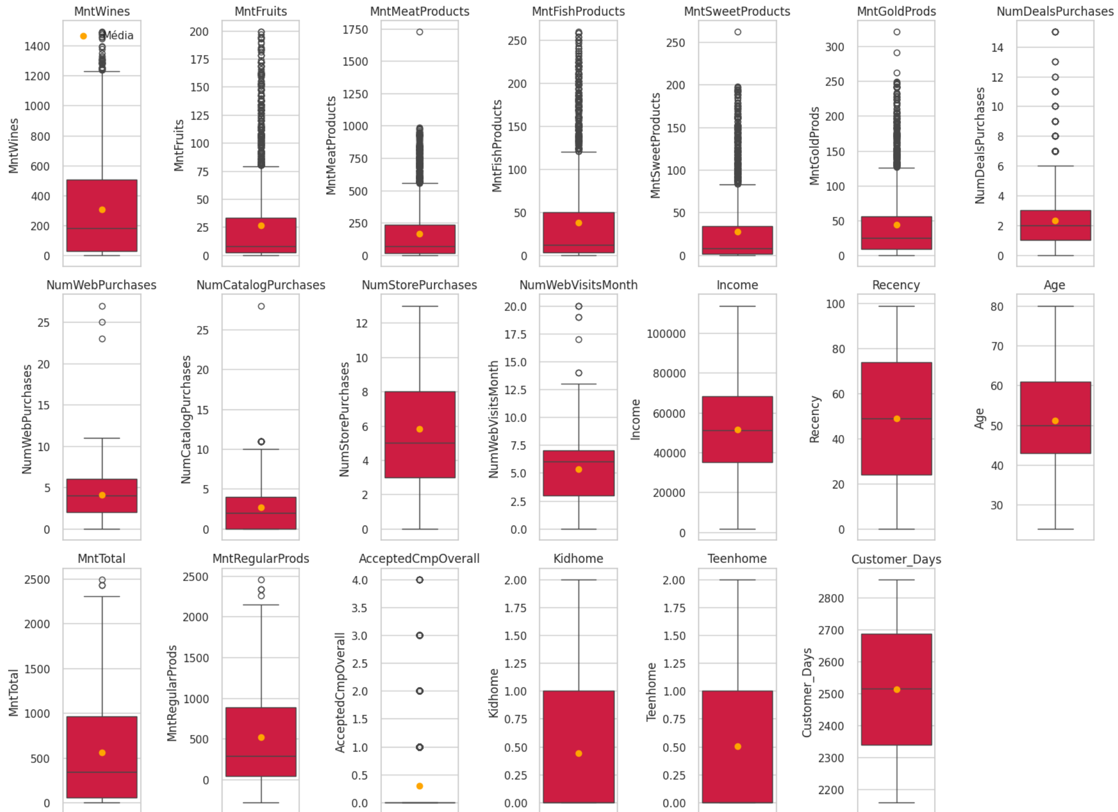


Fig. 4: Boxplots of numerical variables

Summary statistics including Mean, Trimmed Mean (useful in the presence of outliers), Mode, 1st Quartile, Median, 3rd Quartile, IQR, Variance, Standard Deviation, and Coefficient of Variation were computed for each variable and are presented in the table below:

Index	Mean	Trimmed Mean	Mode	1st Quartile	Median	3rd Quartile	IQR	Variance	Standard Deviation	Coefficient of Variation (%)
MntWines	306.165	274.828	2.0	24.0	178.0	507.0	483.0	113902.091	337.494	110.233
MntFruits	26.403	20.809	0.0	2.0	8.0	33.0	31.0	1582.805	39.784	150.68
MntMeatProducts	165.312	138.323	7.0	16.0	68.0	232.0	216.0	47430.091	217.785	131.742
MntFishProducts	37.756	30.523	0.0	3.0	12.0	50.0	47.0	3005.741	54.825	145.209
MntSweetProducts	27.128	21.385	0.0	1.0	8.0	34.0	33.0	1691.715	41.13	151.615
MntGoldProds	44.057	37.787	3.0	9.0	25.0	56.0	47.0	2676.636	51.736	117.43
NumDealsPurchases	2.318	2.096	1.0	1.0	2.0	3.0	2.0	3.557	1.886	81.363
NumWebPurchases	4.101	3.932	2.0	2.0	4.0	6.0	4.0	7.493	2.737	66.74
NumCatalogPurchases	2.645	2.388	0.0	0.0	2.0	4.0	4.0	7.832	2.799	105.822
NumStorePurchases	5.824	5.663	3.0	3.0	5.0	8.0	5.0	10.509	3.242	55.666
NumWebVisitsMonth	5.337	5.346	7.0	3.0	6.0	7.0	4.0	5.825	2.414	45.231
Income	51622.095	51630.889	7500.0	35196.0	51287.0	68281.0	33085.0	429031013.055	20713.064	40.124
Recency	49.009	49.001	56.0	24.0	49.0	74.0	50.0	837.067	28.932	59.034
Age	51.096	51.071	44.0	43.0	50.0	61.0	18.0	137.026	11.706	22.91
MntTotal	562.765	517.364	39.0	56.0	343.0	964.0	908.0	331703.325	575.937	102.341
MntRegularProds	518.707	472.892	16.0	42.0	288.0	884.0	842.0	306746.774	553.847	106.775
AcceptedCmpOverall	0.299	0.188	0.0	0.0	0.0	0.0	0.0	0.463	0.68	227.425
Kidhome	0.442	0.413	0.0	0.0	0.0	1.0	1.0	0.289	0.537	121.493
Teenhome	0.507	0.482	0.0	0.0	0.0	1.0	1.0	0.296	0.544	107.298
Customer_Days	2512.718	2513.052	2826.0	2339.0	2515.0	2688.0	349.0	41032.031	202.564	8.062

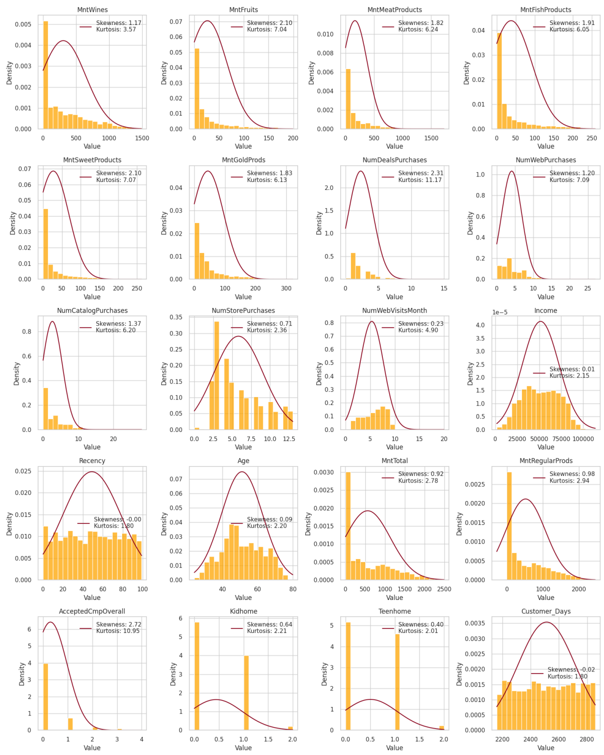
Fig. 5: Summary statistics of numerical variables

As also highlighted by the boxplots, the Coefficient of Variation (CV) exposes that specially the features related with amount of products purchased in the last two years per customer (MntMeatProducts, MntFishProducts, ...) has high variation, most of them above 100% of variation. The meaning is that along the customers profile we have a high variation of amount of products purchase behaviour. Although the distribution of the customers are concentrated in lower values of amount spent, there are outliers that indicates the existence of customer with high amount spent in products, diverging from the majority.

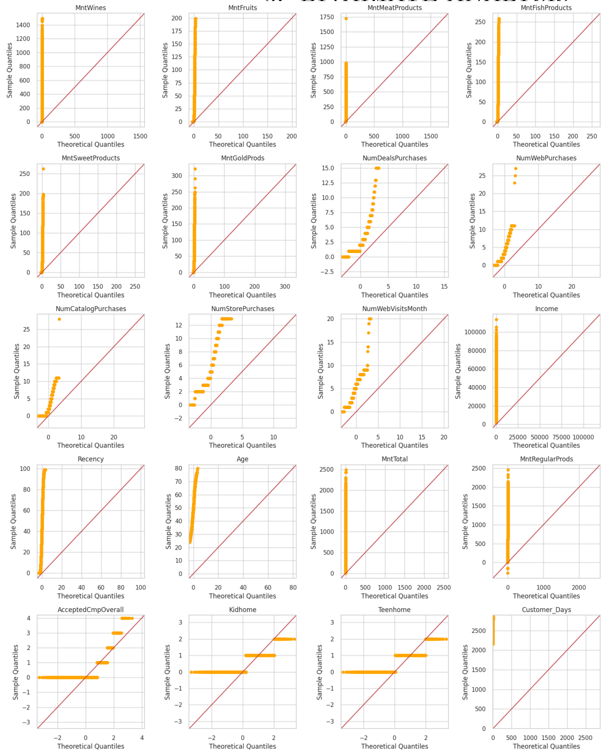
Regarding the monthly purchases by channels (NumDealsPurchase, NumWebPurchases...), the channel with higher mean is the NumStorePurchases (5.8). The mean NumWebVisitsMonth is also high (5.3 monthly visits per customer), but we need to consider that the purchase behaviour relate is represented by NumWebPurchases, which has the mean of 4.1 monthly purchases per customer. Converting to business understanding we can interpret that the Store is the most important purchase channel and it is followed by the NumWebPurchases.

The mean Age of customers is 51 years, indicating mature customers. It is also reinforced by the median for the Age feature, that is 50 meaning that at least 50% of the customers has 50 years or more. About the Kidhome/Teenhome, the statistics shows that at least 50% of the customers don't have kids, and 75% have at maximun 1 teen/kid home.

The distribution of selected variables in a dataset was also investigated, highlighting distinct tail and peak characteristics through kurtosis analysis. Upon exploring the kurtosis of the variables, we observed a variety of patterns reflecting the diversity and complexity of the data. For example, we identified several variables, including 'MntFishProducts', 'MntMeatProducts', 'MntFruit', 'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases', 'NumCatalogPurchases', 'NumWebPurchases', and 'NumWebVisitsMonth', that exhibited a kurtosis greater than 3, indicating heavier tails and more pronounced peaks. This observation suggests a higher probability of extreme values compared to the mean for these variables. In contrast, variables such as 'NumStorePurchases', 'Income', 'Recency', 'Age', 'Kidhome', 'Teenhome', and 'Customer_Days' demonstrated a kurtosis less than 3, indicating lighter tails and flatter peaks, typical characteristics of platykurtic distributions. Additionally, we found that the majority of variables exhibit positive skewness, suggesting that most values are concentrated to the left of the mean, with a small proportion of extremely high values pulling the mean to the right. This observation indicates the presence of outliers or extreme values at the upper end of the distribution, while most of the data is concentrated at lower values. This interpretation is supported by graphical analysis of the distributions, as can be observed below:



(a) Kurtosis and Skewness



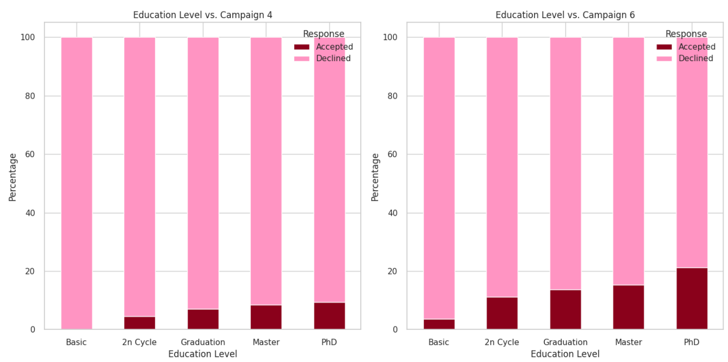
(b) QQ-Plot with the reference to the diagonal of Normal distribution.

Fig. 6: Kurtosis, Skewness and Sample Quantiles of numerical variables

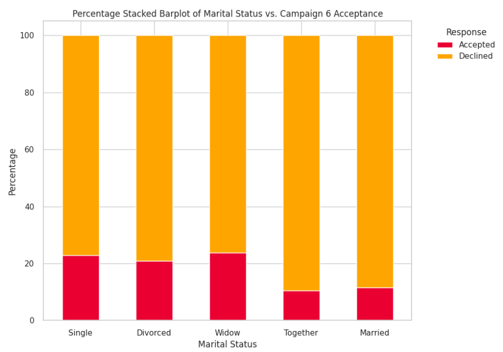
3 Bivariate Analysis

3.1 Contingency tables

Our bivariate analysis of categorical variables involved exploring relationships between customer demographics and their engagement with promotional campaigns. We utilized contingency tables to visually represent these associations and conducted independence tests to assess their significance.



(a) Proportion of Campaign 4 and 6 acceptability by Education Level



(b) Proportion of Campaign 6 acceptability by Marital Status

Fig. 7: Proportion of Associated Categorical Variables

Interestingly, we found that marital status was notably associated with the acceptance of Campaign 6, indicating that customers' relationship status may influence their response to specific marketing efforts. Furthermore, education level showed a significant relationship with the acceptance of Campaigns 4 and 6, suggesting that customers with higher education levels may be more receptive to certain campaign messages.

Moreover, our analysis revealed intriguing trends suggesting that individuals in less serious or no relationships demonstrated higher engagement with Campaign 6. This highlights the importance of considering not only demographic factors but also the relational context of customers when designing targeted marketing strategies.

We also conducted tests between the campaigns, and all tests resulted in wide critical regions, leading to the rejection of independence across the board. This can be attributed to the substantial number of rejections observed for all six marketing proposals evaluated.

3.2 Categorical vs. Numerical Variables

We chose the Kruskal-Wallis test to conduct bivariate analysis between the level of education, marital status, and the other numerical variables. This choice was made due to the non-normal distribution observed in all variables, as well as the presence of 5 categories in each categorical variable.

The Kruskal-Wallis analysis demonstrated a relationship between different levels of education and various numerical variables, including 'MntWines', 'Income', 'MntRegularProds', 'MntMeatProducts', 'Age', 'MntTotal', 'MntSweetProducts', 'MntFishProducts', 'NumStorePurchases', 'MntFruits', 'MntGoldProds', 'NumWebPurchases', 'NumCatalogPurchases', 'Teenhome', 'NumWebVisitsMonth', 'Kidhome', and 'Customer_Days'. This suggests a statistically significant difference in the medians of these variables among different levels of education, implying that education level can significantly influence spending behavior in these areas.

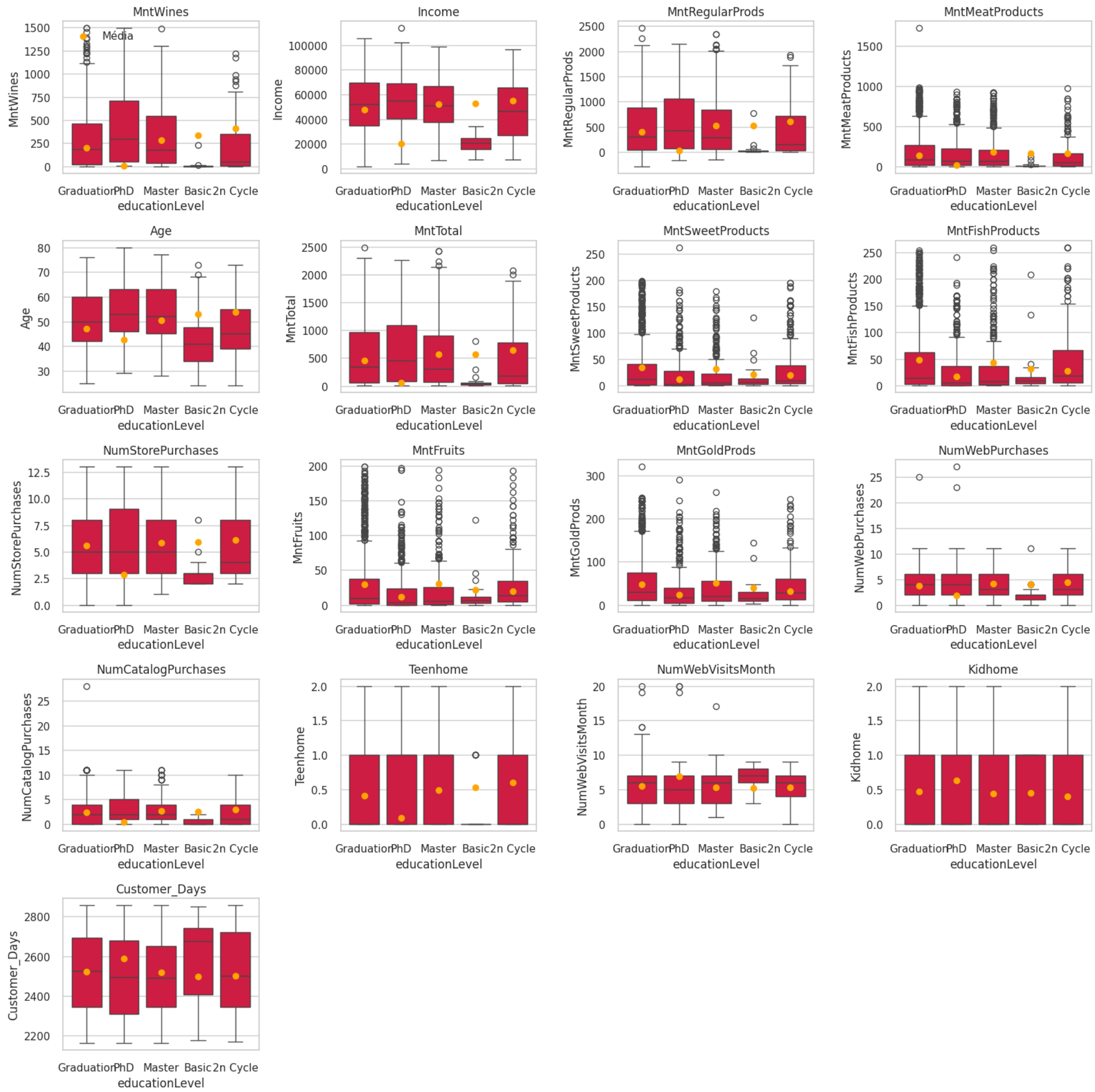


Fig. 8: Numerical Variables by Level of Education

Additionally, the Kruskal-Wallis test revealed an association between customers' marital status and the variables 'Age', 'Teenhome', and 'Kidhome'. This finding suggests that marital status may exert a significant influence on customers' age and the number of teenagers and children in their households.

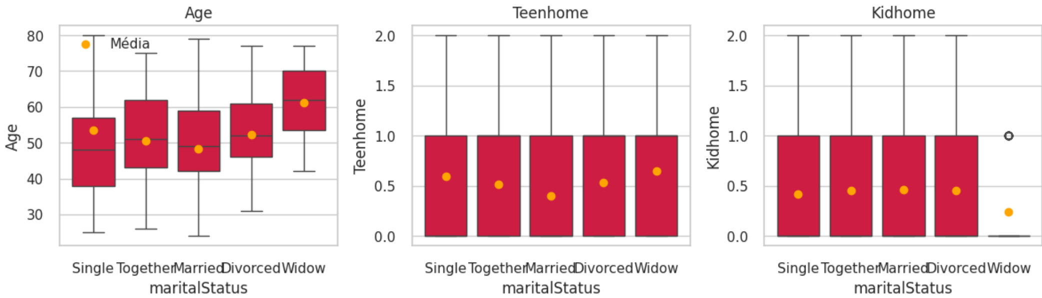


Fig. 9: Numerical Variables by Marital Status

We employed the Mann-Whitney U test to analyze the relationship between our categorical binary variables and numerical data. These binary variables include campaign acceptance (accepted or declined) across various campaigns (Campaign 1 to Campaign 6) and whether the customer lodged a complaint (True or False). Our analysis revealed that there was no significant association between customer complaints and numerical variables, considering an alpha = 0,05.

For the market campaigns analysis, we opted to spotlight only the numerical variables that demonstrated the strongest association with each campaign.

- For Campaigns 1 and 5, customer income emerged as the numerical variable with the strongest association. We observed a higher acceptance rate among customers with higher incomes;
- For Campaigns 2 and 4, the 'MntWines' variable exhibited the highest statistical association. While the distribution of this variable was fairly even, the majority of customers who spent more on wines showed greater acceptance, whereas those who declined tended to have spent less on wines.;
- For the Campaign 3 the variable 'MntGoldProds' had higher association, with both the customers with higher amount spent in gold products being those with higher acceptance;
- For Campaign 6, the 'NumCatalogPurchases' variable showed the strongest association. We observed higher acceptance rates among customers who made a greater number of purchases from the catalog.

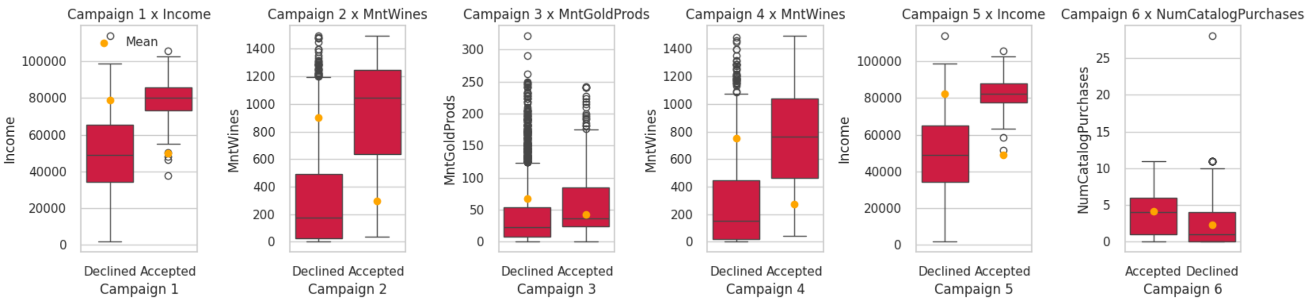


Fig. 10: Numerical Variables by Campaign acceptance

3.3 Correlations Between Numerical Variables

We began our bivariate analysis of numerical variables by creating a heatmap using the Spearman correlation coefficient, due to the non-normality of the variables. This approach allowed us to identify which variables had strong correlations, guiding our subsequent analysis to these specific areas.

We observed that 'MntRegularProds' and 'MntTotal', representing the amount spent on regular products and the total amount spent, respectively, exhibit a strong positive correlation with each other and also with 'MntWines', 'Income', 'NumCatalogPurchases', 'NumWebPurchases', 'NumStorePurchases', and 'MntMeatProducts'.

These correlations indicate that customers who spend more on regular products tend to have a higher total expenditure and also consume more wine and meat, suggesting a pattern of higher-quality and more expensive consumption.

Additionally, both 'MntRegularProds' and 'MntTotal' show a negative correlation with the number of children at home. This result suggests that families with more children may have budget constraints that lead them to spend less on regular products and, consequently, on the total in this delivery service platform.

Moreover, the purchases made in 'Deals' is the only purchase channel that shows a positive correlation with the number of website visits by customers. This indicates that customers who make purchases during deals are the ones most likely to visit the company's website.

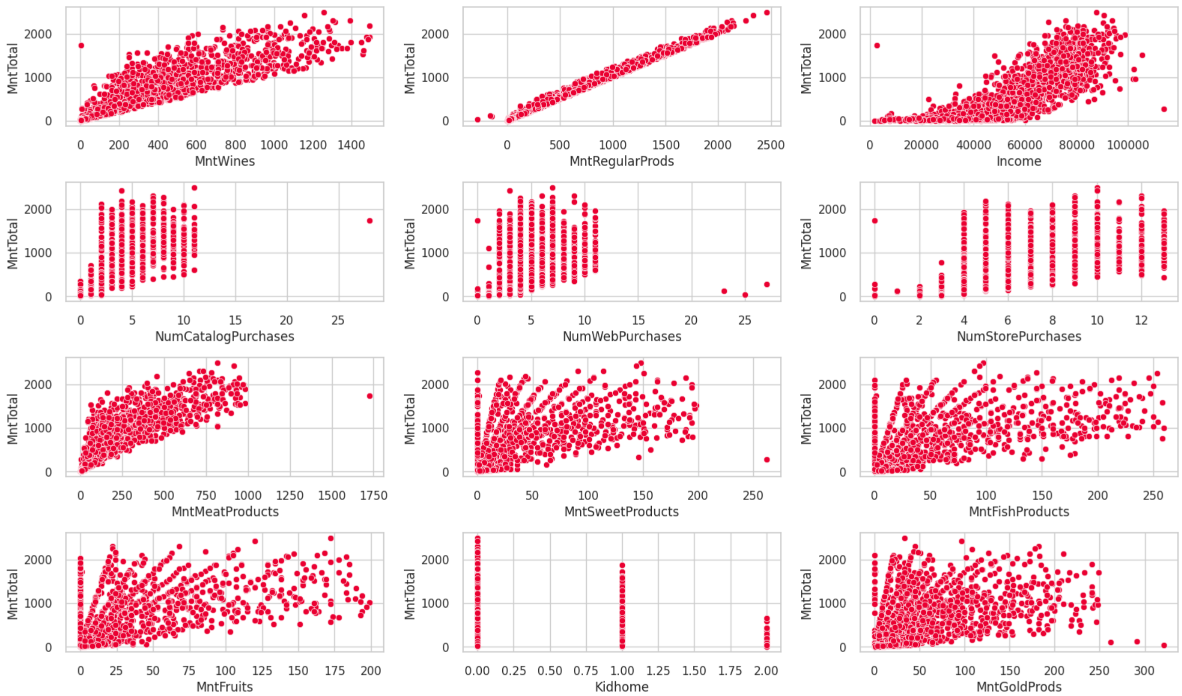


Fig. 11: "Scatter Plots of Relationships between Total Amount Spent and the Most Correlated Variables

Our analysis revealed a strong positive correlation between annual household income ('Income') and the amount spent on wines ('MntWines'), regular products ('MntRegularProds'), total amount spent ('MntTotal'), catalog purchases ('NumCatalogPurchases'), store purchases ('NumStorePurchases'), and meats ('MntMeatProducts'). This pattern corroborates the insights observed above.

Additionally, a moderate negative correlation was observed between income and the number of web visits per month ('NumWebVisitsMonth') and the presence of children at home ('Kidhome'). This suggests that wealthier families may visit websites less frequently, potentially due to preferences for other shopping channels or less time available for online browsing. Furthermore, these families tend to have fewer children, which may be associated with a more flexible budget for luxury purchases and fewer typical constraints of households with children.

4 Multivariate Analysis

4.1 Principal Component Analysis (PCA)

Considering the significant correlations revealed by Pearson correlation statistics, the application of normalized Principal Component Analysis (PCA) was deemed necessary. Standardization of numerical features precedes the analysis in normalized PCA. The subsequent table showcases the outcomes of the PCA analysis, presenting eigenvalues and cumulative explained variances for each principal component.

Table 1: Eigenvalues and Explained Variance in Normalized PCA Analysis

	Eigenvalues	Explained Var. cum.
PC1	8.34786	0.417204
PC2	2.06728	0.520521
PC3	1.44591	0.592783
PC4	1.18588	0.65205
PC5	1.00731	0.702393
PC6	0.839801	0.744364
PC7	0.751808	0.781938
PC8	0.67489	0.815667
PC9	0.577163	0.844512
PC10	0.535887	0.871294
PC11	0.434521	0.89301
PC12	0.413737	0.913688
PC13	0.399533	0.933655
PC14	0.387283	0.953011
PC15	0.320776	0.969042

The explained variance ratio for the 15th principal component are listed in the table 1. To decide how many principal components to keep in the analysis we will list the relation of criteria and number of components to keep considering our results:

- Pearson's criterion: Keep a number q of components such that they explain at least 80% of the total dispersion. In this sense, the 8 first principal components explains 81.45% of the data variance;
- Cattel's criterion: Consider the elbow's rule to decide. The 4 or 5 first principal components explains 4 (65.20%) and 5 (70.23%) of th data variance;
- Kaiser criterion: Keep the principal components that has eigenvalues >1. The 5 first principal components explains 70.23% of the data variance;

We decided to keep the first 5 principal components, since they explains about 70% of the variance and the addition of new dimensions did not explained that much about the customer profiles understanding.

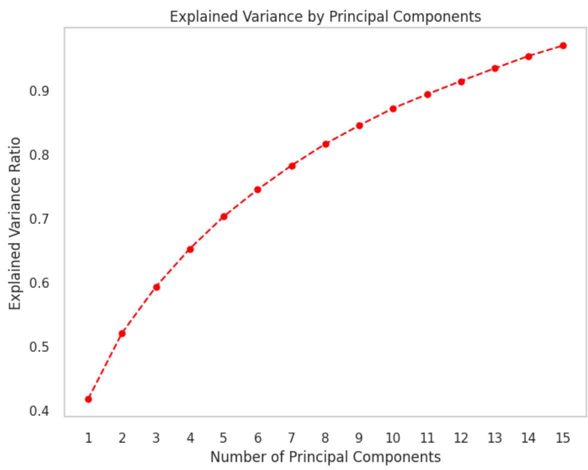


Fig. 12: "Cumulative Sum of Explained Variance for the first 15 Principal Components.

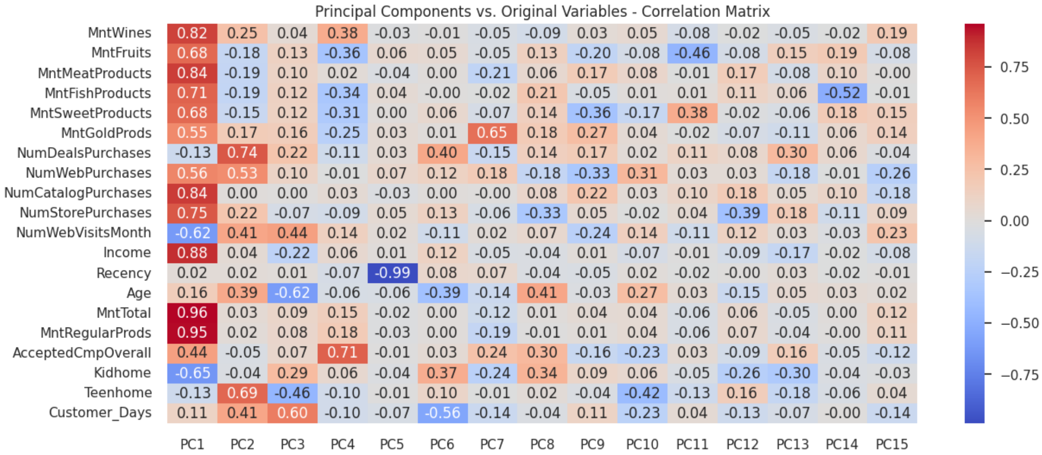
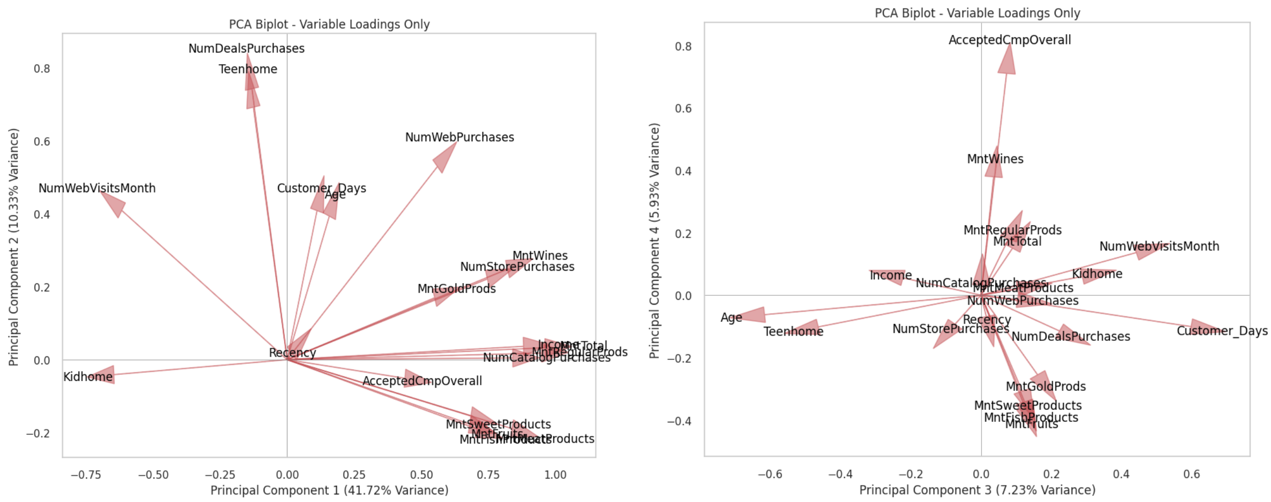


Fig. 13: The correlation of PC with original variables



(a) Biplot of PC1 and PC2.

(b) Biplot of PC3 and PC4.

Fig. 14: Interpretation of Principal Components and Original Variables: The correlation loadings with original variables.

The 1st PC has strong correlations (> $\tilde{0}.7$) with the amount spent in products, the number of purchases done and the income in the positive axis. It contrasts with the moderate correlation (< $\tilde{-}0.6$) of NumWebVisitsMonth and Kidhome. This representations indicate that for PC1 these variables are opposite. Possibly represents that those customers with more kids at home does more web visits in the platform, spend less money in the delivery platform and hve lower income. In the other way, the clients that spend more money have higher campaign acceptance and higher income. Just the 1st component explains 47% of the data variance.

The 2nd PC has strong correlation with the NumDealsPurchases and Teenhome in the positive axis and moderate correlation with the NumbWebVisitsMonth, Age and CustomerDays. From this dimension we can understand that the positive axis represents the customer profile that has teens at home; the customers have mature age, having also more time in the delivery platform. This customer profile seems to perform purchases in deals opportunities making usage of web platform, since the NumbWebVisitsMonth is also moderately correlated in the positive axis.

The 3rd PC contrasts the Customer Days and NumWebVisitsMonth with the Age and Teens at home. This dimension seems to represent the customer behaviour in the platform and its social profile. The custormers with more time in the platform makes use of its website with higher frequency, they have kids at home, are younger and make more usage of deals. In contrast the clients with higher age have teens at home, they have higher income and perform purchases by the store.

The 4th PC is representing the market campaing acceptance in the positive axis, with the amount spend with wines being the only kind of product being represented in the same direction. The amount spend in the other type of products, like sweet, fish, fruits and gold products are all represented in the negative axis. We can interpret that the campaign acceptance is correlated with the customer profile that spend money in wines. This contrast with those customers that spent on the other product types.

The 5th PC represents mostly the Recency in the negative axis. The Age and customer days in platform are in the same direction, but they have a really low correlation with the 5th component.

The Cos2 statistic that indicates the quality of representation by the components exposed lower values. Although there is a high correlation between the components and the variable loadings, the low values of Cos2 can indicate that the variables has complex relations that was possibly not well represented in a single dimension.

4.2 Clustering

Within the scope of customer segmentation study, an exploratory analysis was conducted to identify distinct groups among consumers based on their purchasing patterns and demographic characteristics. Employing hierarchical clustering method, two significantly different clusters were discerned among the customers, whose distinction was validated both visually on the dendrogram and by the silhouette score metric. Detailed analysis of the cluster characteristics revealed substantial differences in purchasing behaviors among them.

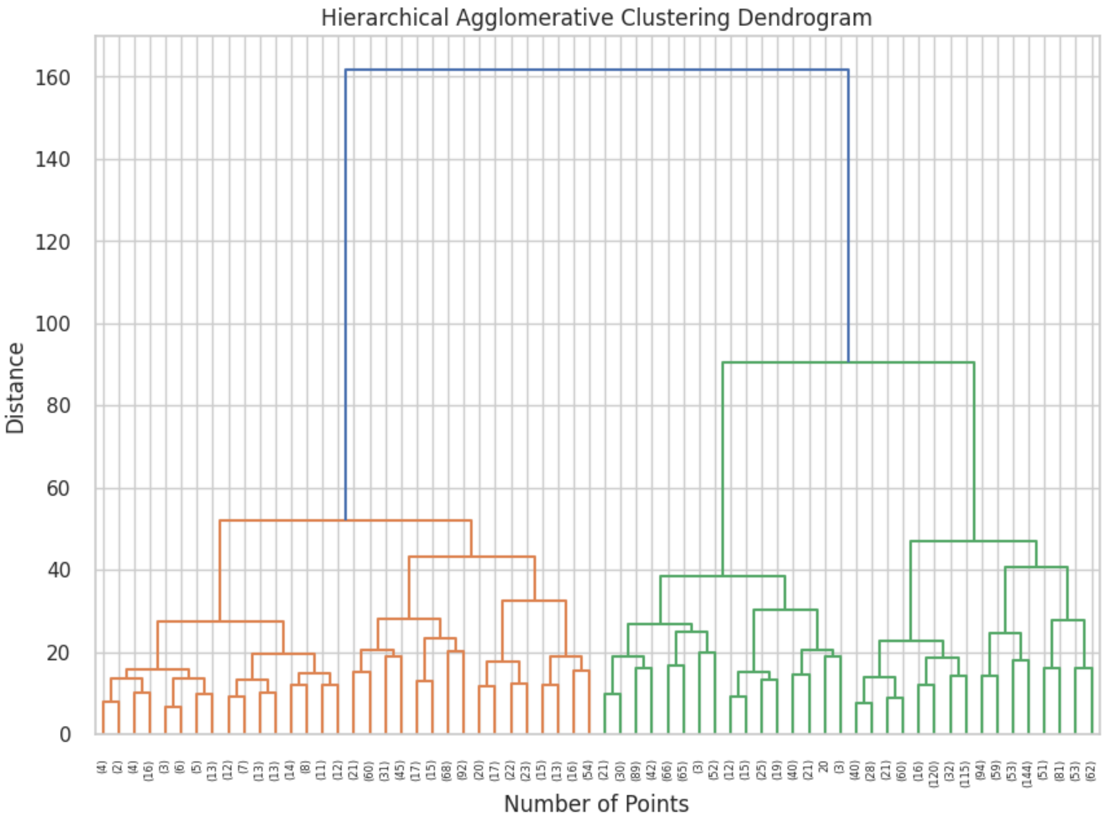


Fig. 15: Agglomerative Hierarchical Clustering Dendrogram: Customer Segmentation Analysis

Cluster 0 has a significant number of outliers compared to Cluster 1. This may indicate a more heterogeneous segmentation of customers in Cluster 0, with some customers exhibiting atypical spending patterns compared to the majority of the group. This group demonstrated tendencies towards moderate spending in various product categories, including wines, fruits, meats, fish, sweets, and gold products. Additionally, consumers in Cluster 0 tended to have an average or below-average income, as well as a shorter relationship time with the company. Notably, a significant presence of families with children at home was observed in this cluster, which aligns with the trend of lower frequency of spending on food delivery services such as iFood.

On the other hand, Cluster 1 represented consumers with a higher receptivity to marketing campaigns, albeit this receptivity being relatively modest. Consumers in this cluster showed a greater propensity to experiment with new products and purchasing channels, and were more engaged with the company's promotional offerings. Despite the greater diversity of behaviors, consumers in Cluster 1 exhibited a more varied income and a longer relationship time with the company compared to Cluster 0, suggesting a higher spending capacity and a longer history of interactions with the company. This group also showed a prevalence of families with children at home and a higher income, indicating greater freedom to use food delivery services such as delivery apps.

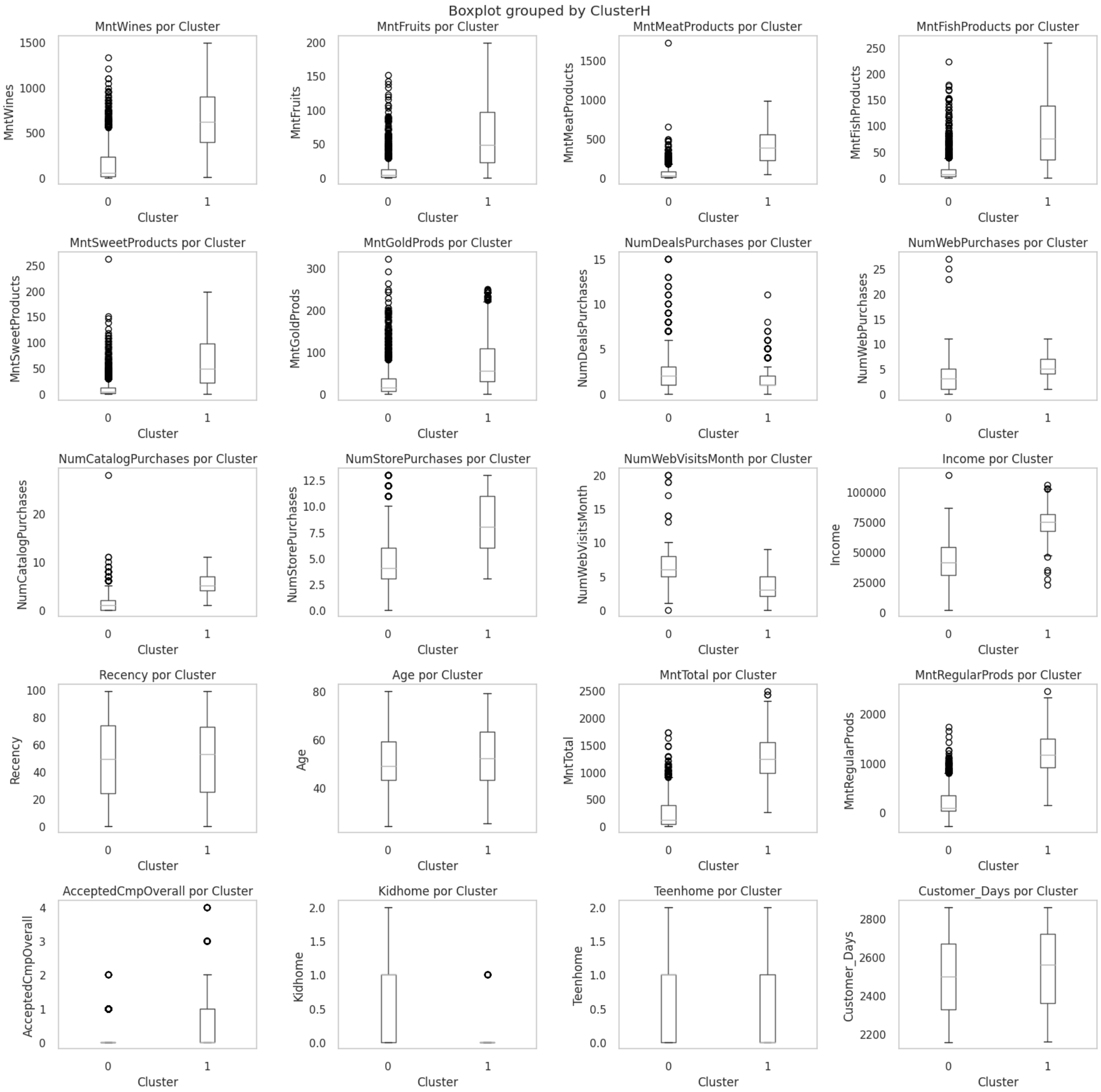


Fig. 16: Cluster-wise Distribution: Box Plots

These analyses provide valuable insights for understanding the different consumer profiles and may inform the development of more targeted and effective marketing strategies to meet the specific needs of each customer segment.

5 Conclusion

This research offers a comprehensive examination of consumer behavior and the effectiveness of marketing strategies using data from iFood. While the sixth marketing campaign boasted the highest acceptance rate, a prevailing trend of rejection was observed across campaigns, underscoring the necessity for more targeted marketing efforts to bolster acceptance rates. Notably, the analysis revealed high customer satisfaction, with less than 1% of customers lodging complaints in the past two years, indicating a generally satisfactory level of service.

In dissecting spending behavior, clear correlations between income levels and expenditures across various product categories were identified. More affluent customers demonstrated higher expenditures on wines, meat, regular products, and overall purchases, suggesting opportunities for promoting premium products to this segment. Moreover, bivariate analysis of marketing campaigns uncovered correlations between demographic variables such as education level and marital status, and campaign acceptance. Additionally, household composition, particularly the presence of children, emerged as a significant influencer of spending behavior, with families with more children tending to spend less on certain product categories and overall.

The discernible correlations among variables prompted the adoption of normalized Principal Component Analysis (PCA) to delve deeper into the dataset. Utilizing various selection criteria, including Pearson's, Cattel's, and Kaiser's, we opted to retain the first five principal components, explaining approximately 70% of the variance. The PCA analysis echoed the contrast between different consumer types observed in the clustering results.

Hierarchical clustering analysis further delineated two distinct consumer segments based on purchasing patterns and demographic characteristics. Cluster 0 exhibited moderate spending behavior across product categories, characterized by a significant presence of families and lower-income individuals. In contrast, Cluster 1 comprised consumers with higher receptivity to marketing campaigns, indicative of a more engaged and financially capable demographic.

The integration of PCA and clustering analyses furnishes a nuanced understanding of consumer segments, facilitating the development of tailored marketing strategies and service offerings. Leveraging these insights can optimize marketing endeavors, elevate customer satisfaction, and propel business growth in the dynamic realm of food delivery services.

6 Appendix

A Additional Graphs

A.1 Biariate Analisis

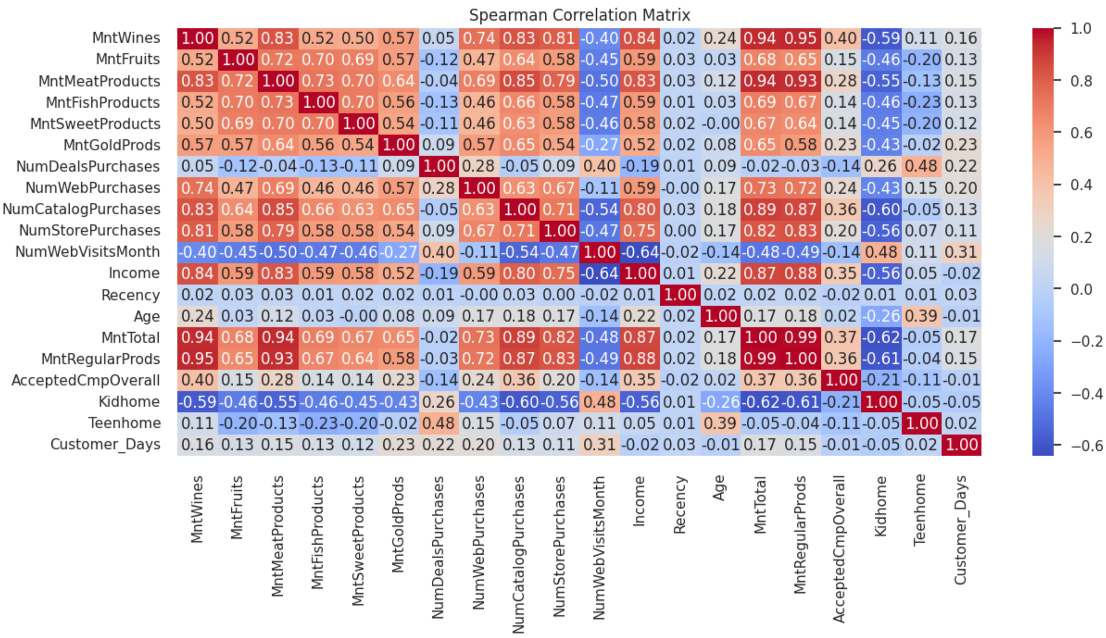


Fig. 17: Heatmap of Spearman Correlation Coefficient of Numerical Variables

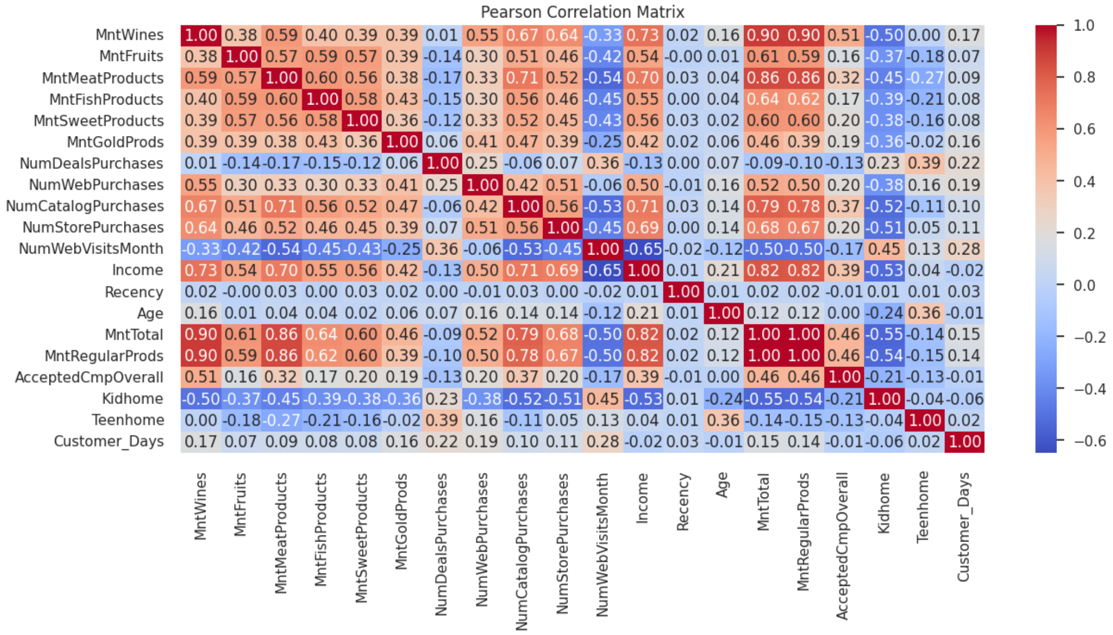


Fig. 18: Heatmap of Pearson Correlation Coefficient of Numerical Variables

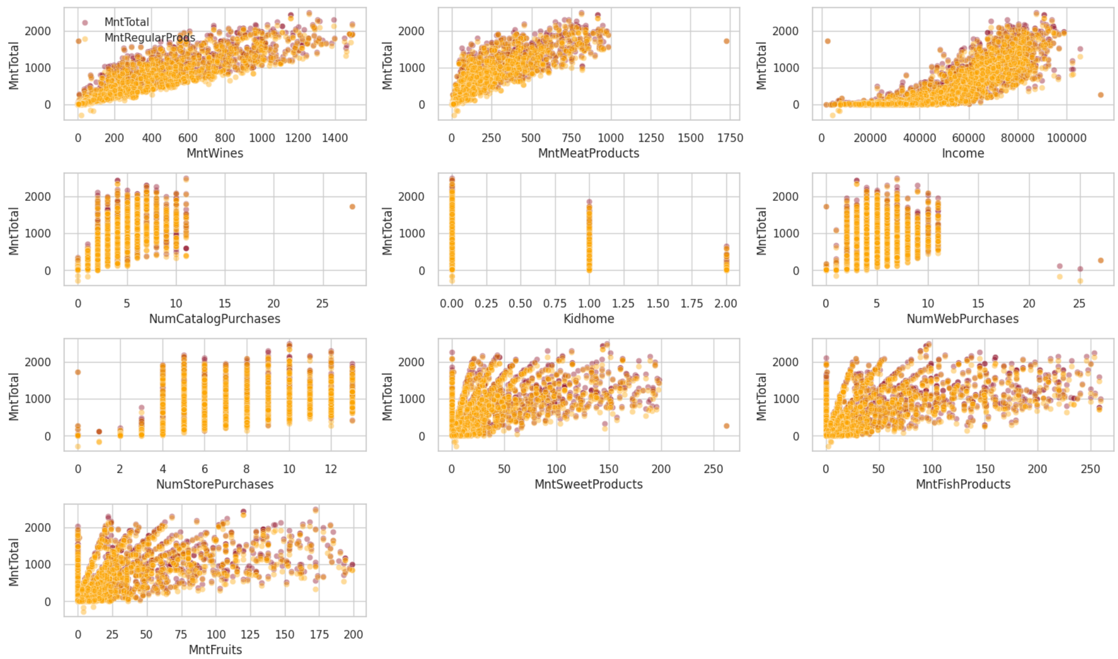


Fig. 19: Scatter Plots of Relationships between 'MntRegularProd', 'MntTotal' and the Most Correlated Variables

A.1.1 Correlations Between Numerical Variables

A.2 Multivariate Analysis

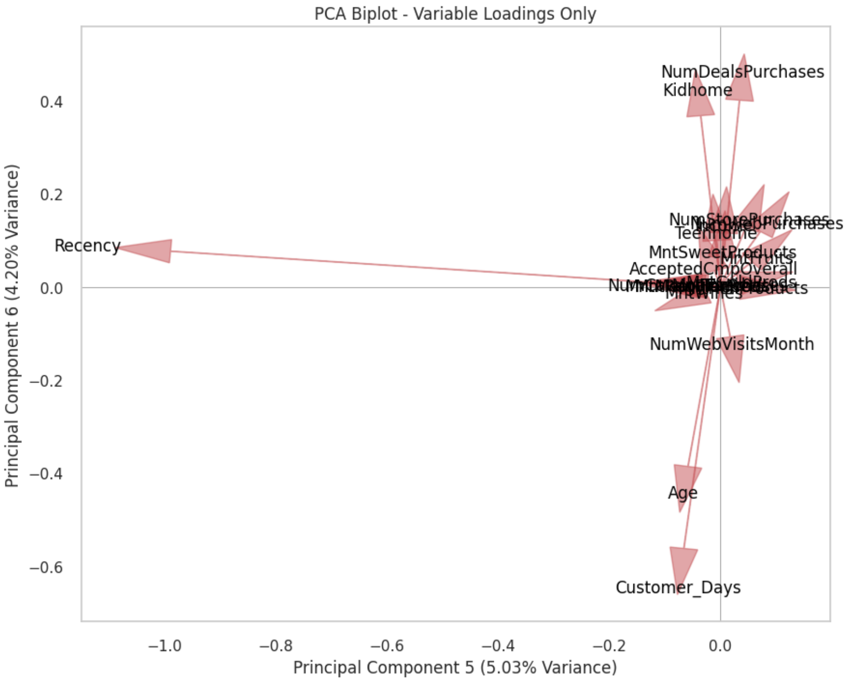


Fig. 20: Biplot of PC5 and PC6: The correlation loadings with original variables.

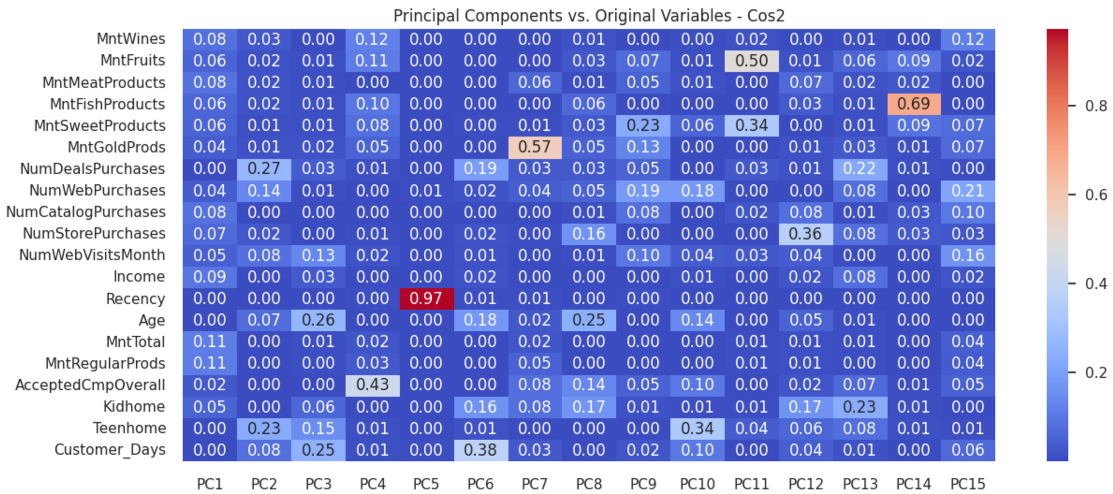


Fig. 21: Matrix with the quality of representation of variables in the principal components - Cos^2

A.2.1 PCA

A.2.2 Clustering

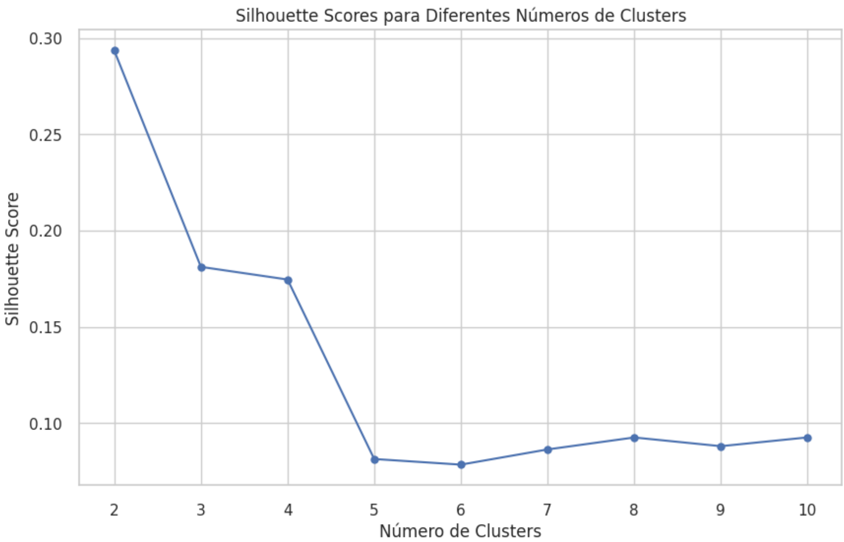


Fig. 22: Silhouette Scores for Different Numbers of Clusters