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.

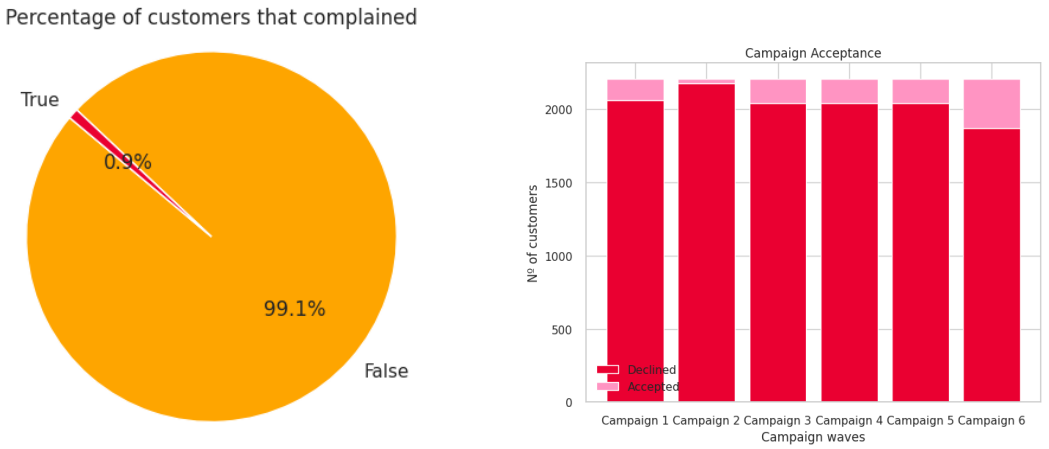
Key columns in the dataset include:

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

2 Univariate Analysis

2.1 Categorical Variables

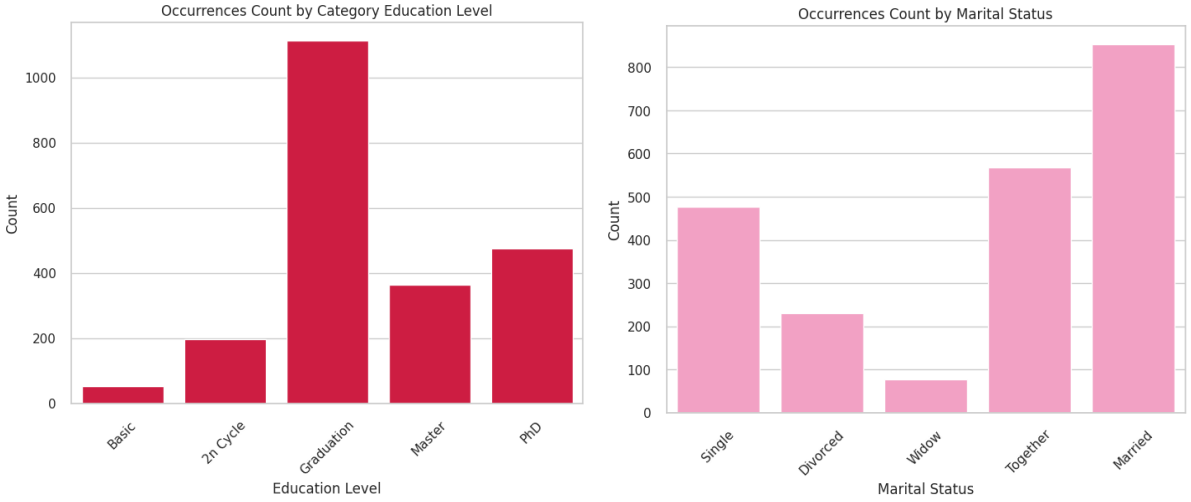
When analyzing responses to the campaigns, it was found that all of them were predominantly declined, with the sixth campaign being the most accepted among customers. Despite various promotional efforts, the overall acceptance rate remained low. Additionally, it was found that the percentage of customers 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 towards the higher values with a longer tail towards the lower values. Additionally, Shapiro-Wilk tests were conducted for all variables, revealing that none of the variables follow a normal distribution.

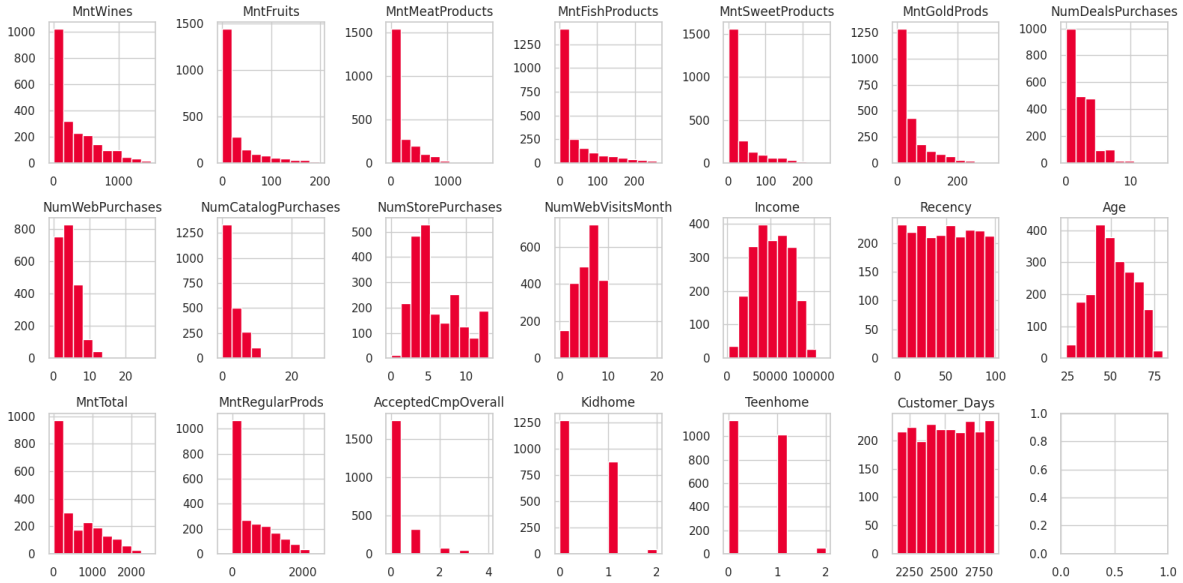


Fig. 3: Histograms of numerical variables

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.

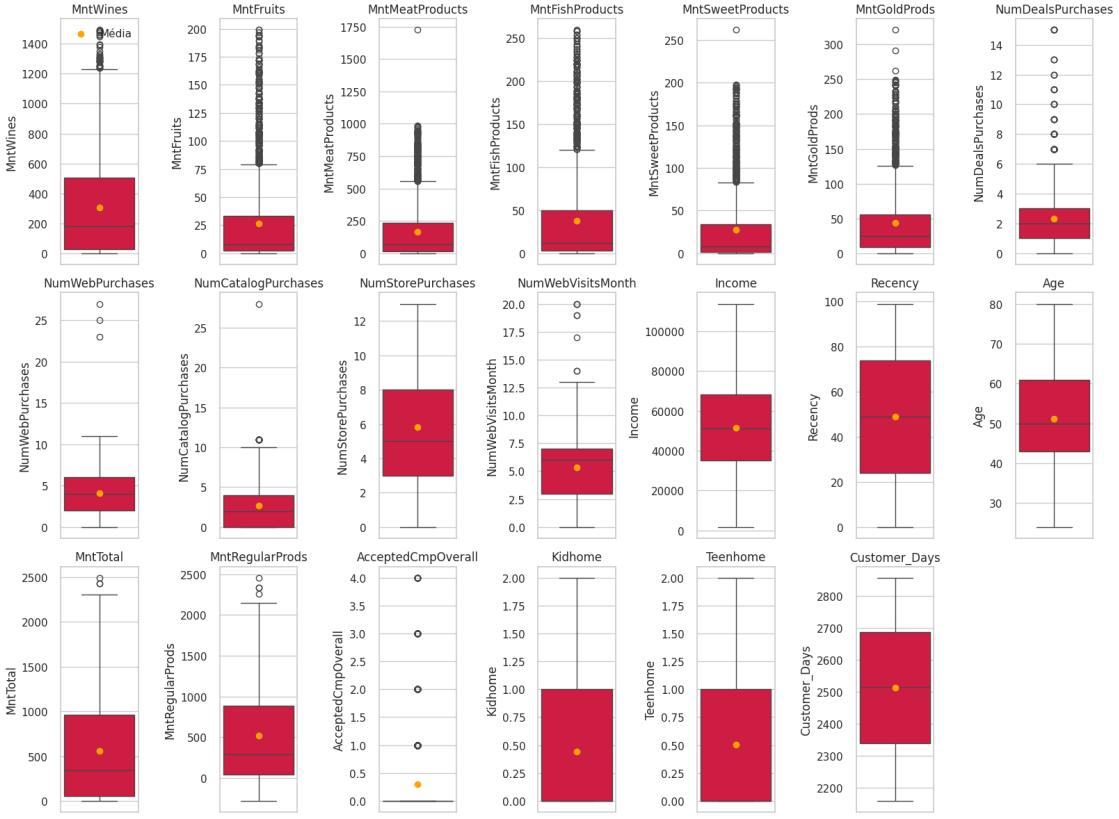


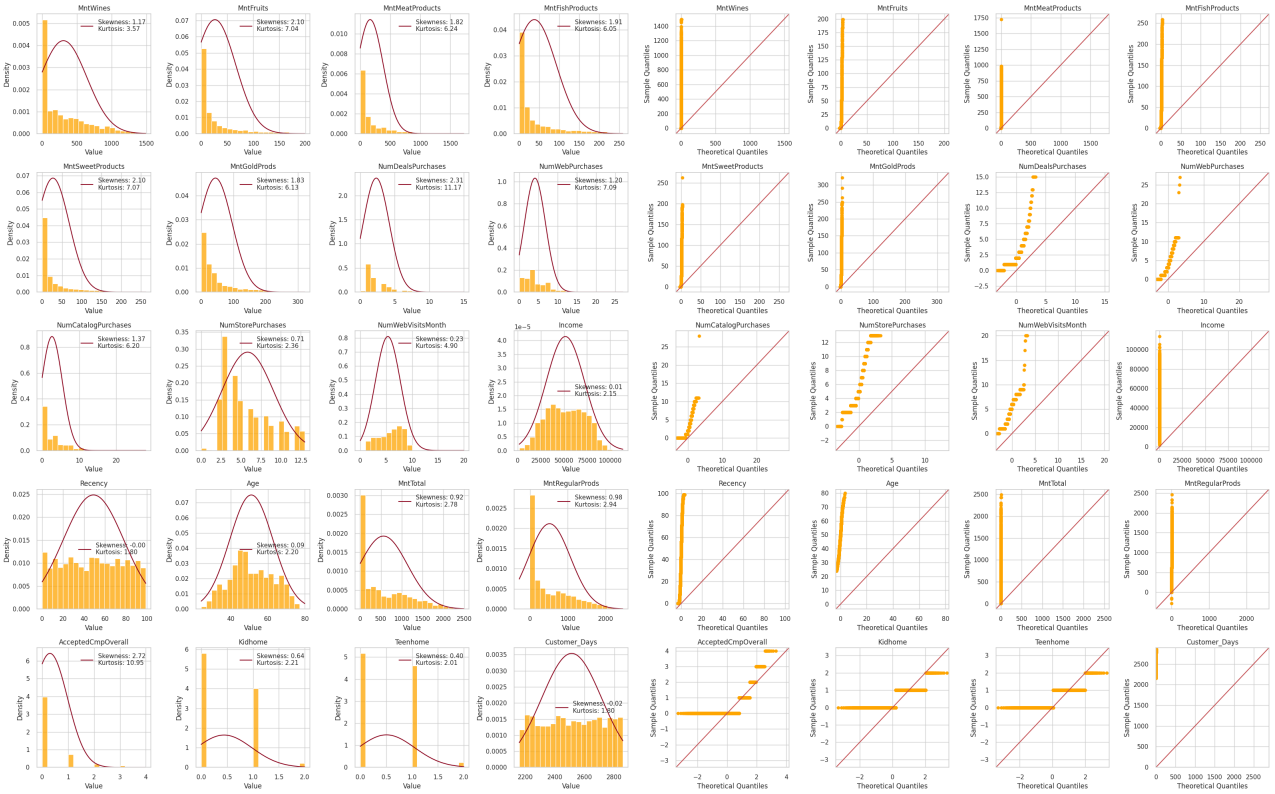
Fig. 4: Boxplots of numerical variables

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

The analysis of kurtosis revealed a positive skewness in the distribution of the variables overall, confirming the trend observed, where the median tends to be smaller than the mean due to the presence of outliers.



(a) Kurtosis and Skewness (b) Sample Quantiles

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

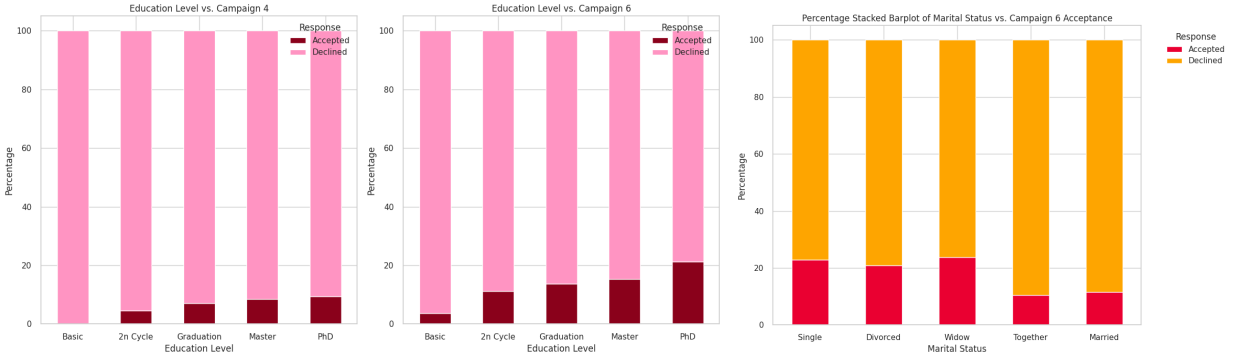
3 Bivariate Analysis

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.

Interestingly, we found that marital status was notably correlated 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.



(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

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.

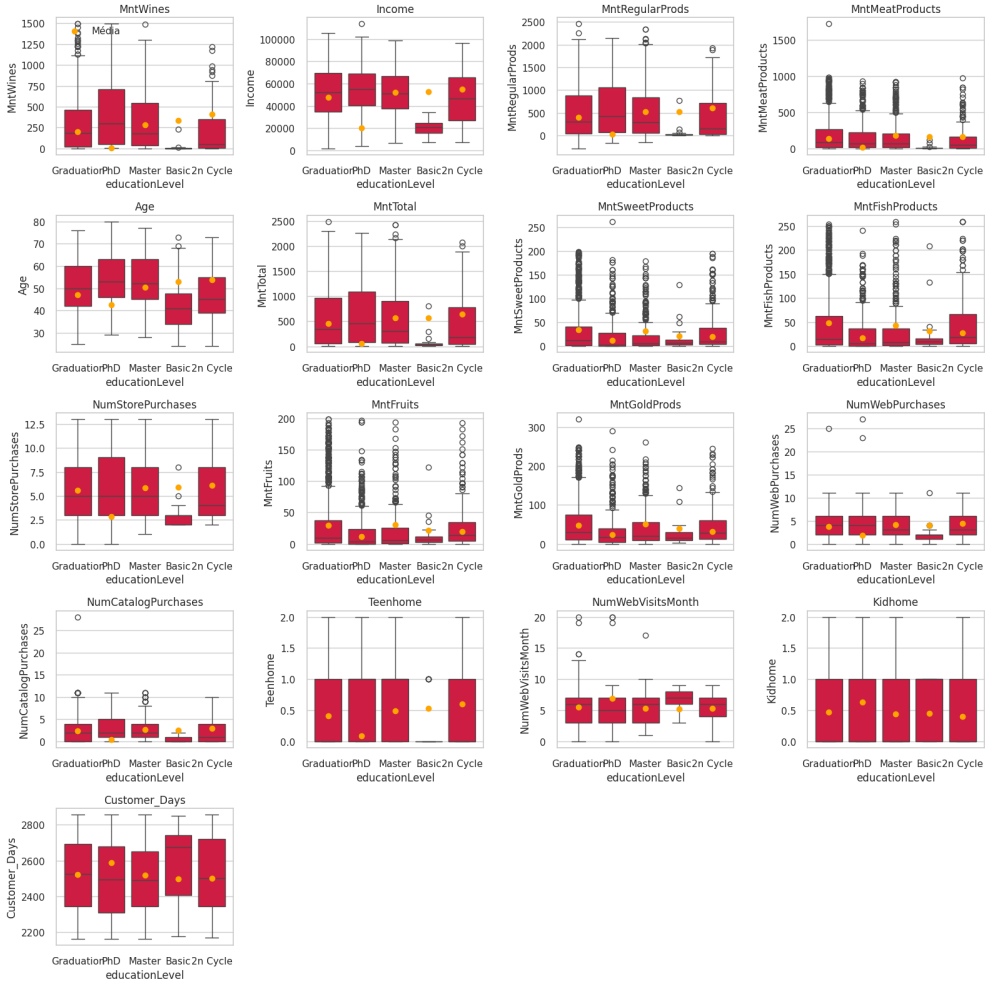


Fig. 8: Numerical Variables by Level of Education

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.

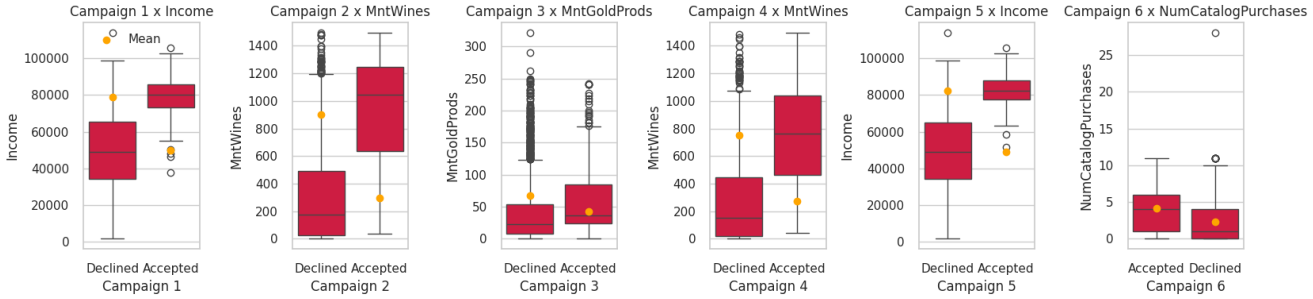


Fig. 10: Numerical Variables by Campaign acceptance

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.

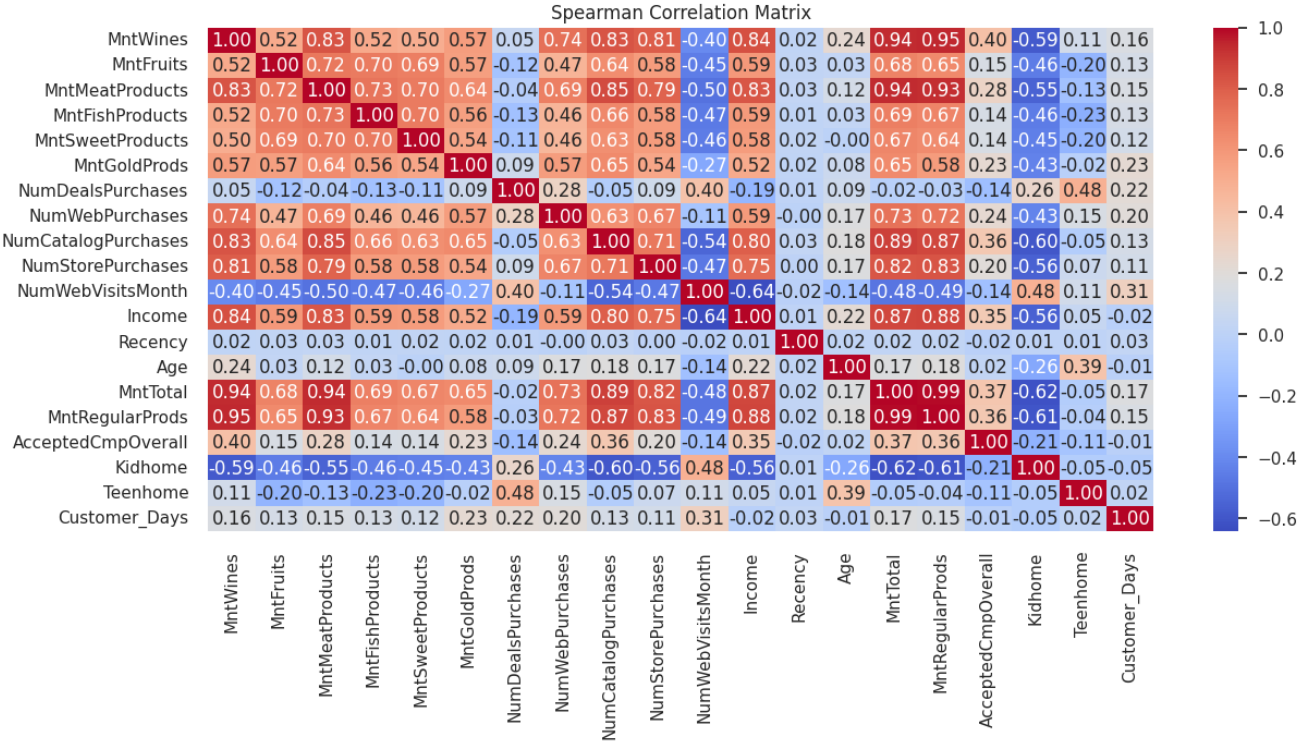


Fig. 11: Heatmap of Spearman Correlation Coefficient of Numerical Variables

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 during '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.

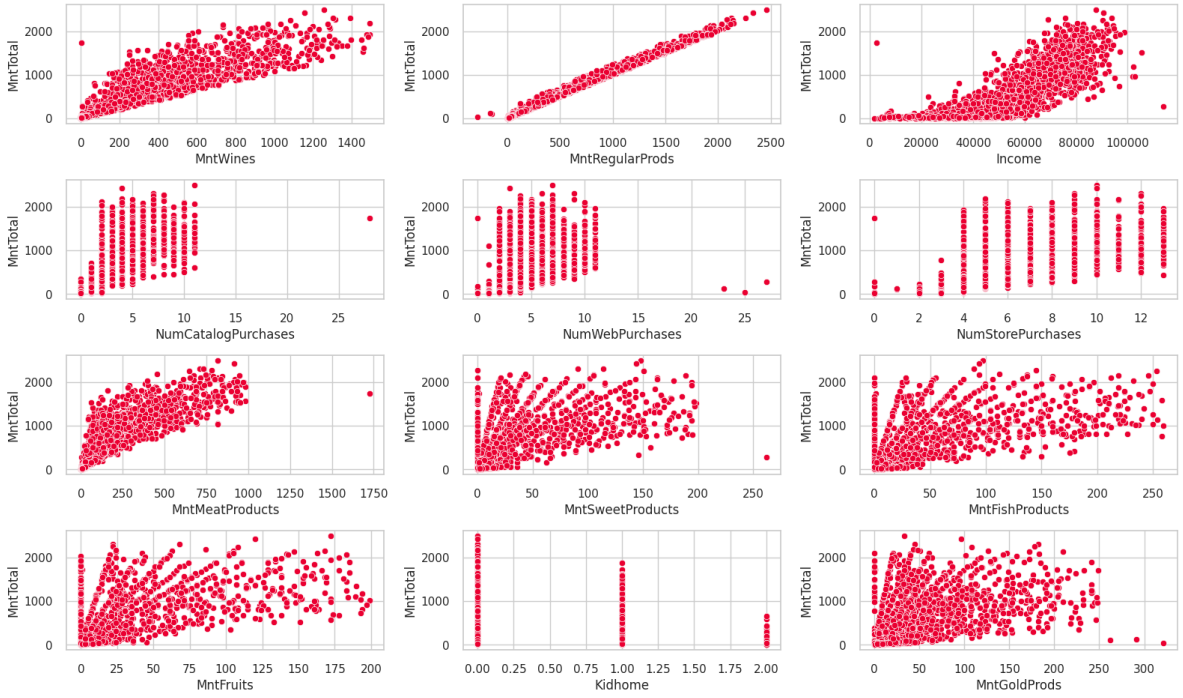


Fig. 12: "Scatter Plots of Relationships between Total Amount Spent and the Most Correlated Variables

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 Conclusion

This research provides a thorough examination of consumer behavior and the efficacy of marketing strategies using the dataset from iFood. The analysis revealed that although the sixth marketing campaign experienced the highest acceptance rate, the general tendency across campaigns was rejection. This highlights the need for more targeted marketing to increase acceptance rates. The study also showed that customer satisfaction is high, with less than 1% of customers lodging complaints in the last two years. This indicates a generally satisfactory level of service.

In terms of spending behavior, the analysis identified clear correlations between income levels and expenditures across various product categories. More affluent customers were found to spend more on wines, meat, regular products, and their total purchases. This insight suggests that there are opportunities for promoting premium products to this segment.

The bivariate analysis of marketing campaigns revealed correlations between demographic variables, such as education level and marital status, and campaign acceptance. Additionally, the study found that household composition, particularly the presence of children, influences spending behavior. Families with more children tend to spend less on certain product categories and also spend less overall, indicating the need for marketing strategies tailored to the financial dynamics of larger families.

Overall, this analysis offers valuable insights into the dynamics of consumer behavior and the effectiveness of marketing strategies, providing actionable recommendations that could help enhance marketing efforts and drive business growth.