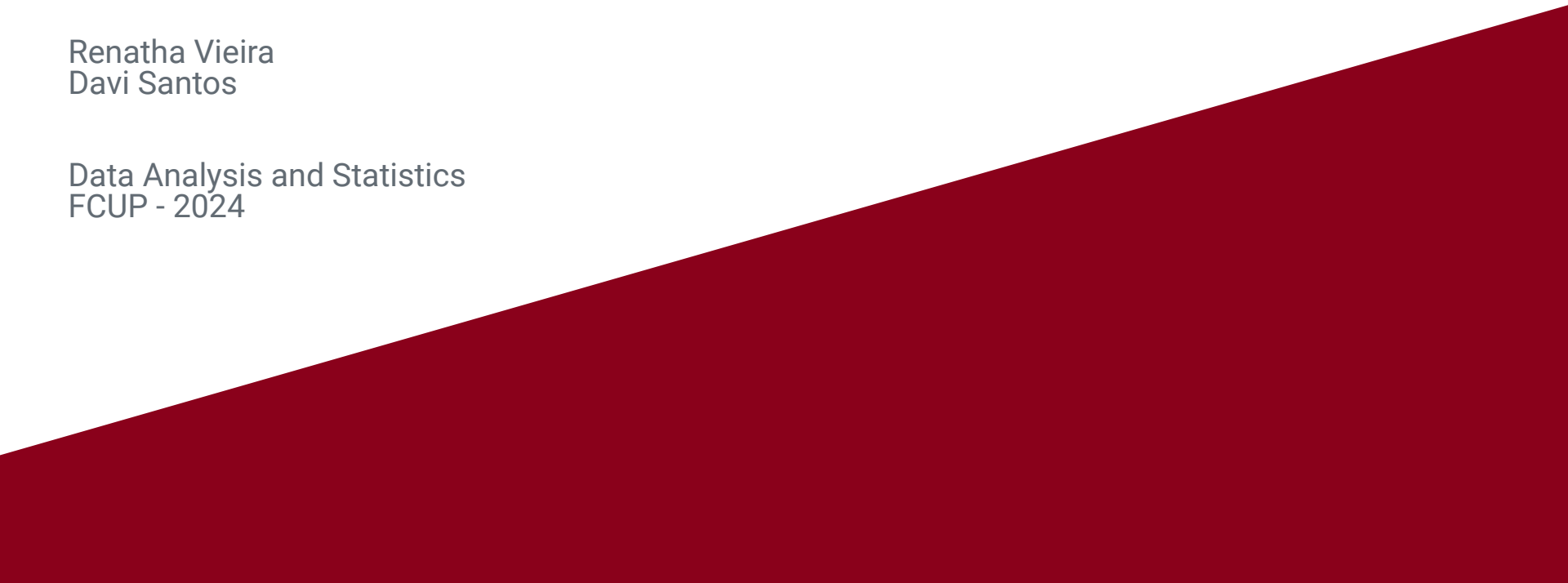


# Data Analysis with Marketing Analytics of IFood

Renatha Vieira  
Davi Santos

Data Analysis and Statistics  
FCUP - 2024



# Dataset understanding

- This dataset was originally presented to potential data analysts as a trial dataset at iFood
  - is the largest food delivery business in Brazil and Latin America and one of Brazil major successful technology startups
- The dataset includes more than 2,000 rows:
  - demographic attributes;
  - sales information;
  - the acceptance of six different marketing campaigns;
  - the amounts spent on different food and goods via delivery;
  - customer complaints;

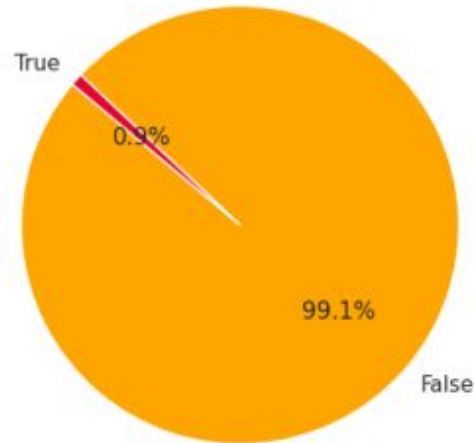
## 24 variables

- Numerical Discrete:
  - AcceptedCmpOverall;
  - Customer\_Days
  - Kidhome
  - Teenhome
  - Income
  - MntFishProducts, MntMeatProducts, MntFruit, MntSweetProducts, MntWines, MntGoldProds, MntTotal, MntRegularProds;
  - NumDealsPurchases, NumCatalogPurchases, NumStorePurchases, NumWebPurchases;
  - NumWebVisitsMonth
  - Recency
  - Age
- Categorical Nominal:
  - Marital Status;
- Categorical Ordinal:
  - Education level;
- Binary:
  - Complain;
  - Campaign 1-6

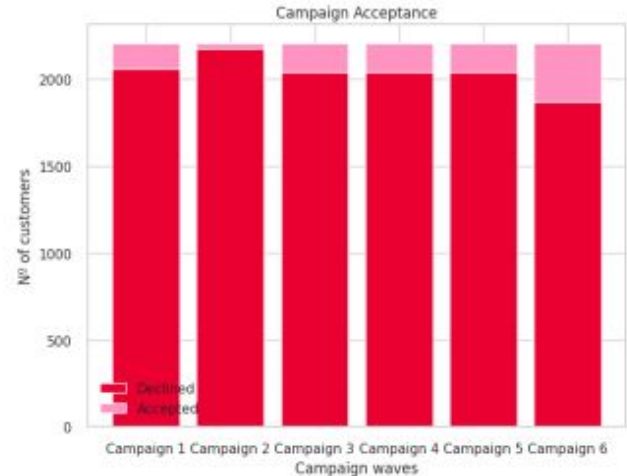
# Univariate Analysis – Categorical Variables

- 1% of customers complained;
- Low campaign acceptance
- Campaign 2 higher rejection;
- Campaign 6 higher acceptance;

Percentage of customers that complained



(a) Percentage of Complaining

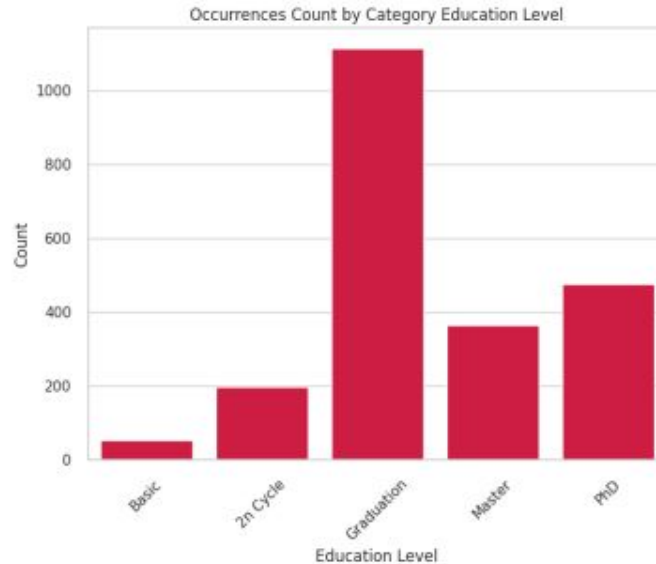


(b) Campaign Acceptance

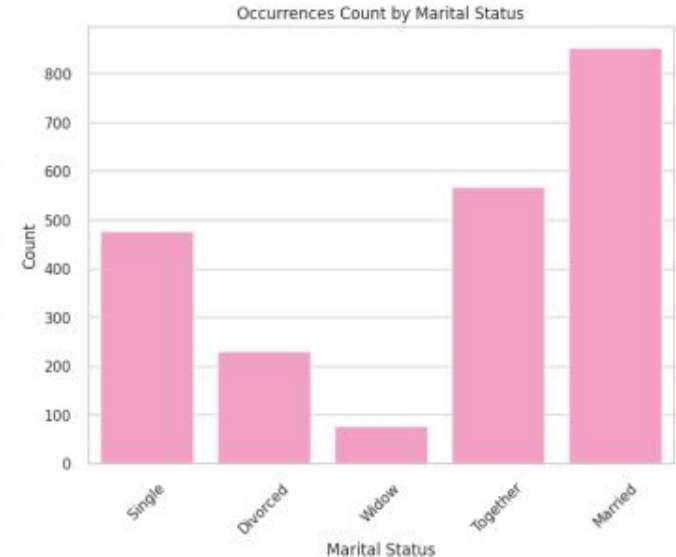
Fig 1. Univariate Analysis of Consumer Behavior

# Univariate Analysis – Categorical Variables

- The majority of customers have at least a Graduation;
- Also the majority of customers are couples (together or married);



(a) Count by Education Level



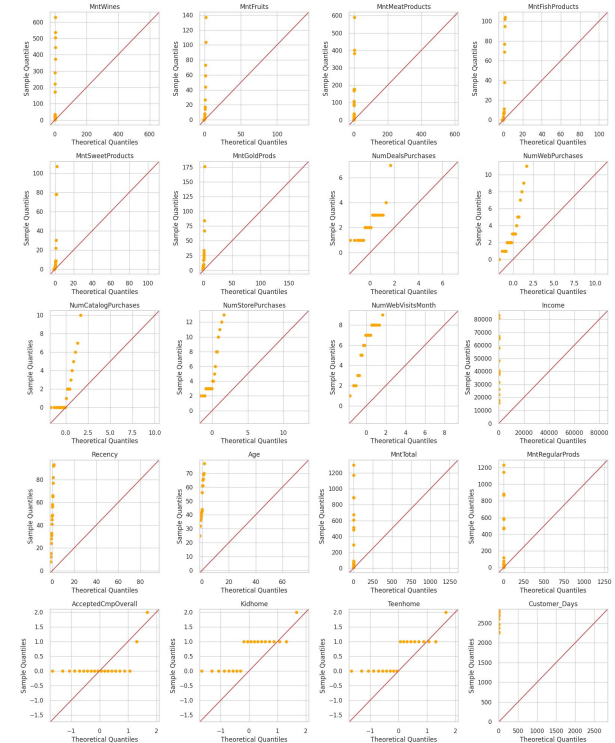
(b) Count by Marital Status

Fig 2. Univariate Analysis of Consumer's Demographics Data

# Univariate Analysis – Numerical Variables

- Shapiro-Wilk tests were conducted for all variables: None of the variables follow a normal distribution

Fig 3. QQ-Plot with the reference to the diagonal of Normal distribution



# Univariate Analysis – Numerical Variables

- High occurrence of outliers in the amount spent values.
  - Different customer profiles

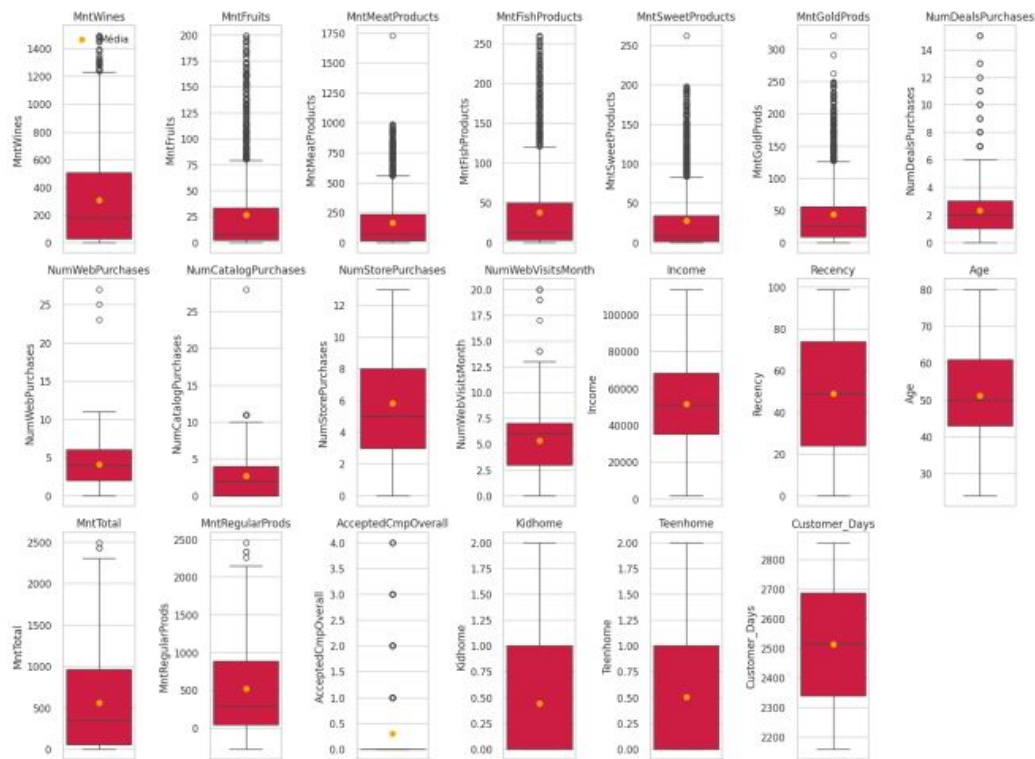


Fig 4. Boxplots of numerical variables

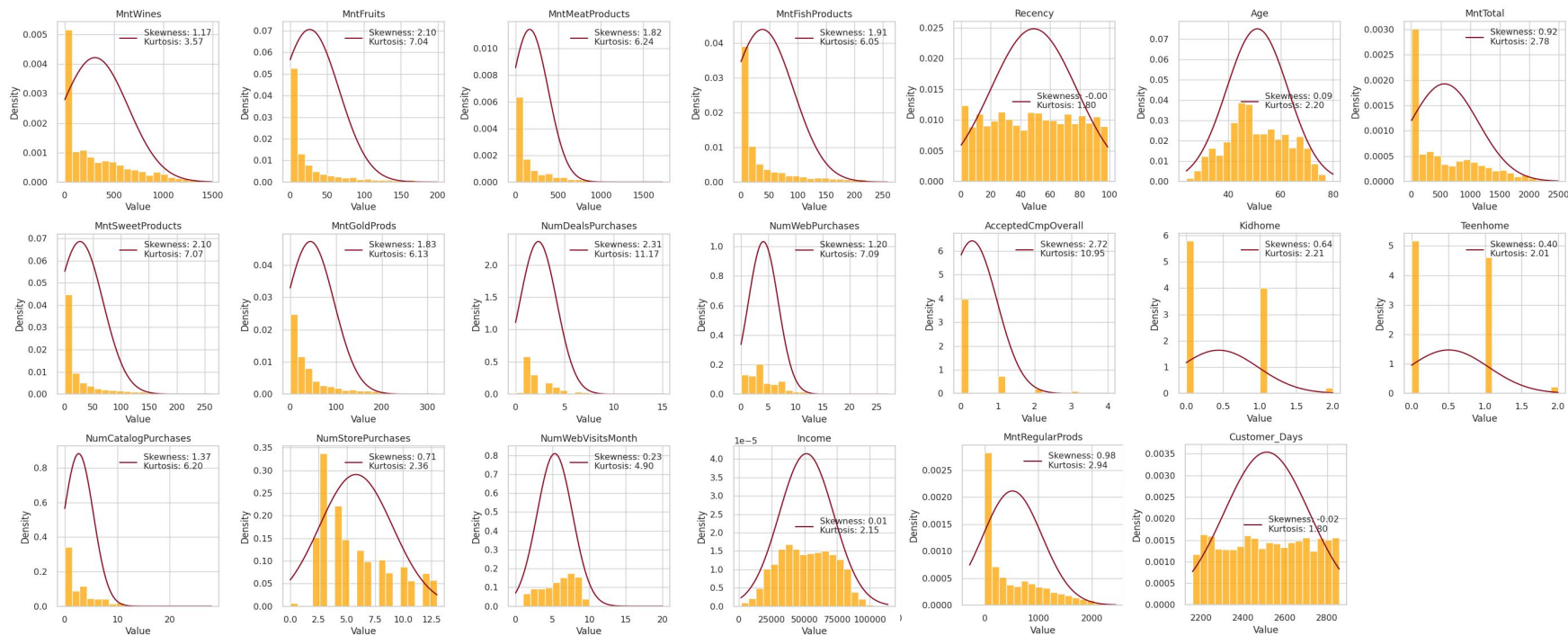
# Univariate Analysis – Numerical Variables

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

Table 1. Summary statistics of numerical variables

- Amount of products purchased:  $CV > 100\%$  = high variation of amount of products purchase behaviour;
- Monthly purchases by channels: Store is the most important purchase channel
- The mean Age of customers is 51 years;
- Kidhome/Teenhome: at least 50% of the customers don't have kids

# Univariate Analysis – Numerical Variables



Kurtosis and Skewness of numerical variables



# Bivariate Analysis – Categorical x Categorical

The Chi-square independence test indicated association between the categories:

Campaign 6 acceptance:

- Marital status: Single customers had higher acceptance;
- Education level : Higher is the education level, higher was the campaign acceptance;

Campaign 4 acceptance:

- Education level : Higher is the education level, higher was the campaign acceptance.  
Those customers with basic education, had almost total rejection.

# Bivariate Analysis – Categorical x Categorical

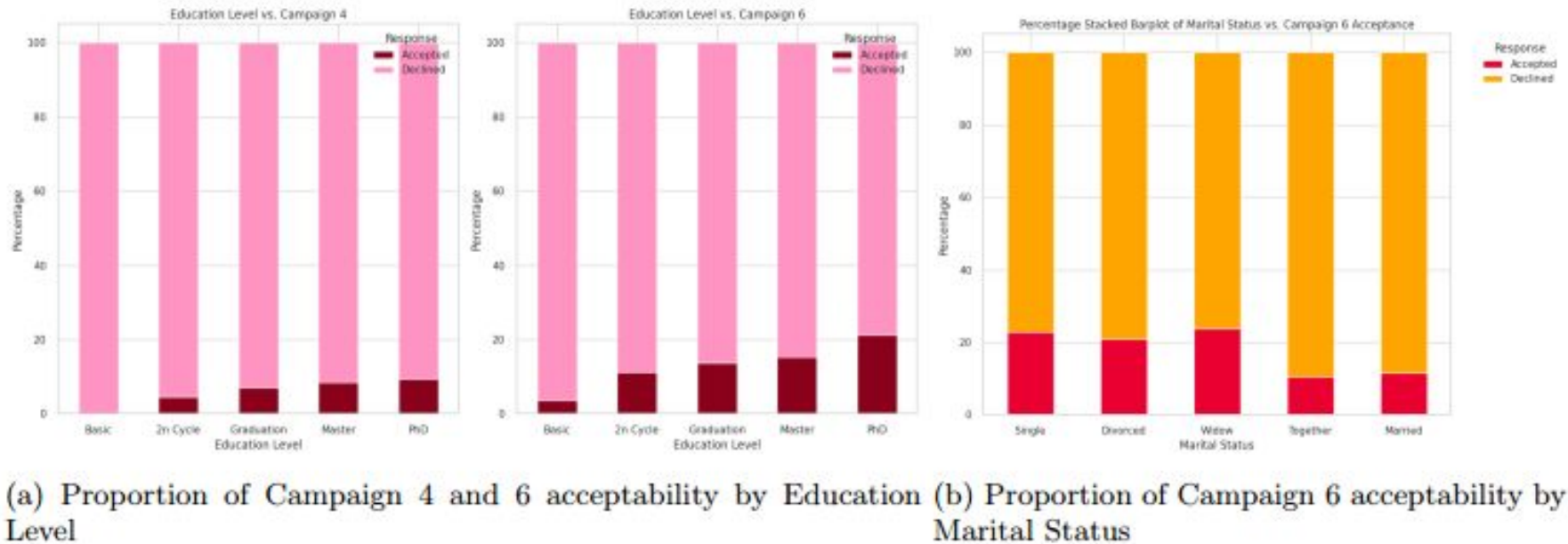


Fig 4. Proportion of Associated Categorical Variable

# Bivariate Analysis – Numerical x Categorical (N classes)

Kruskal-Wallis test to conduct bivariate analysis between the level of education, marital status, and the other numerical variables - non normality, 5 classes per variable.

Education level: statistically significant difference in the medians of these variables among different levels of education:

- Amount spent in products;
- Number of purchases done and the channels to do so;
- Customer Days
- N° teen and kid at home

# Bivariate Analysis – Numerical x Categorical (N classes)

Few examples:

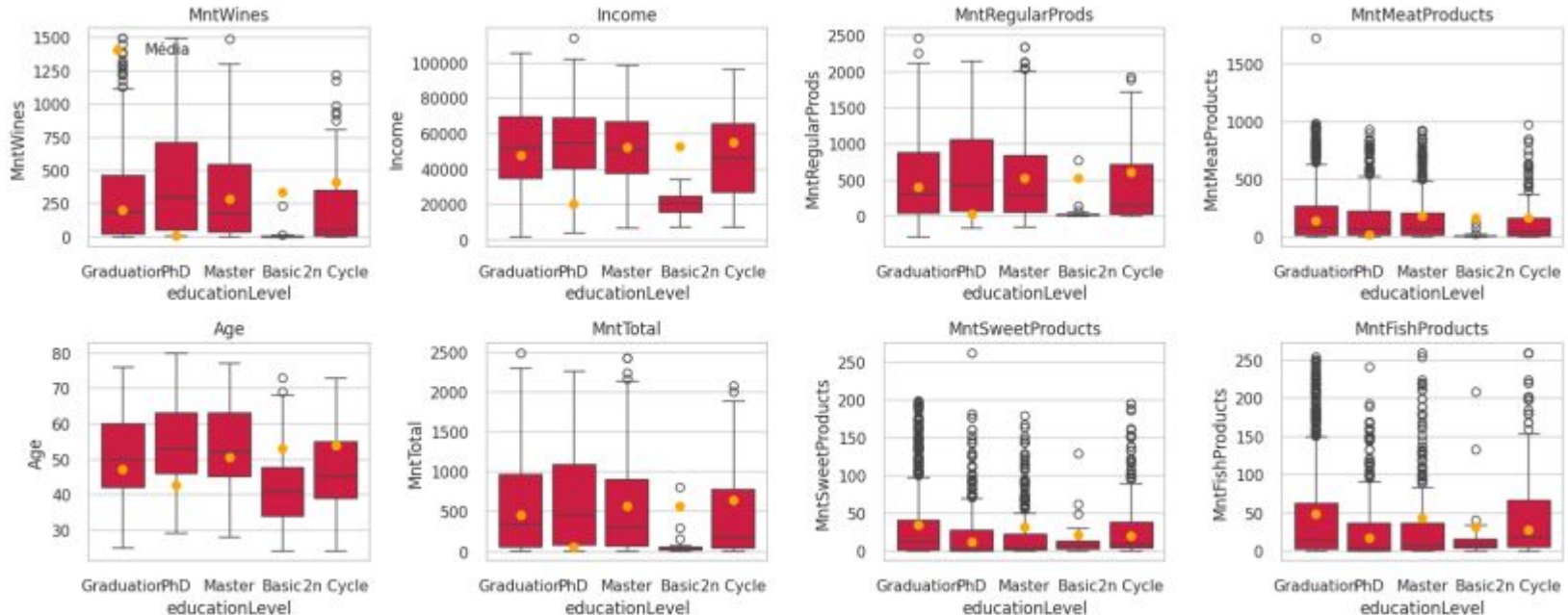


Fig 5. Numerical Variables by Level of Education

# Bivariate Analysis – Numerical x Binary

We conducted the Mann-Whitney U test to analyze the relationship between our categorical binary variables and numerical data.

Market campaigns analysis: spotlight only the numerical variables that demonstrated the strongest association with each campaign.

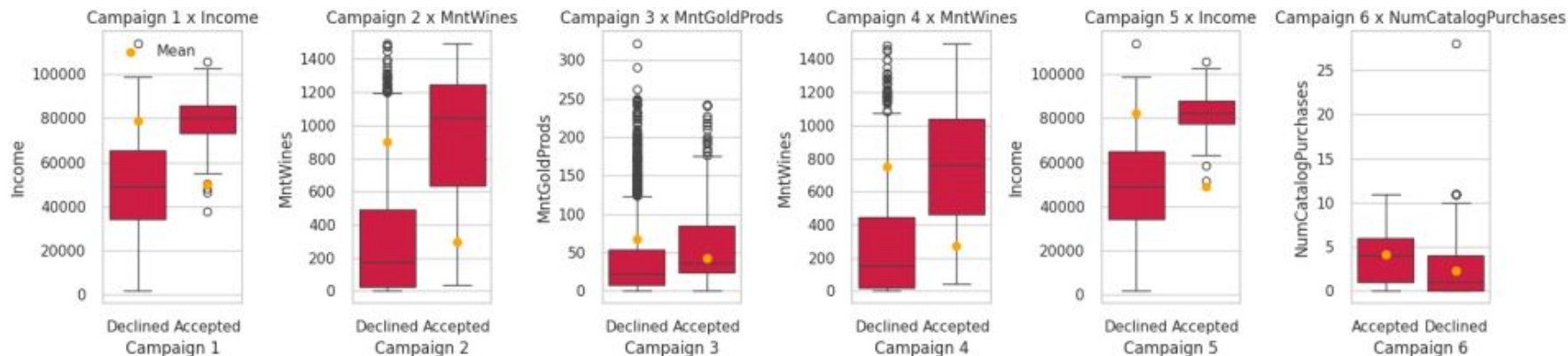
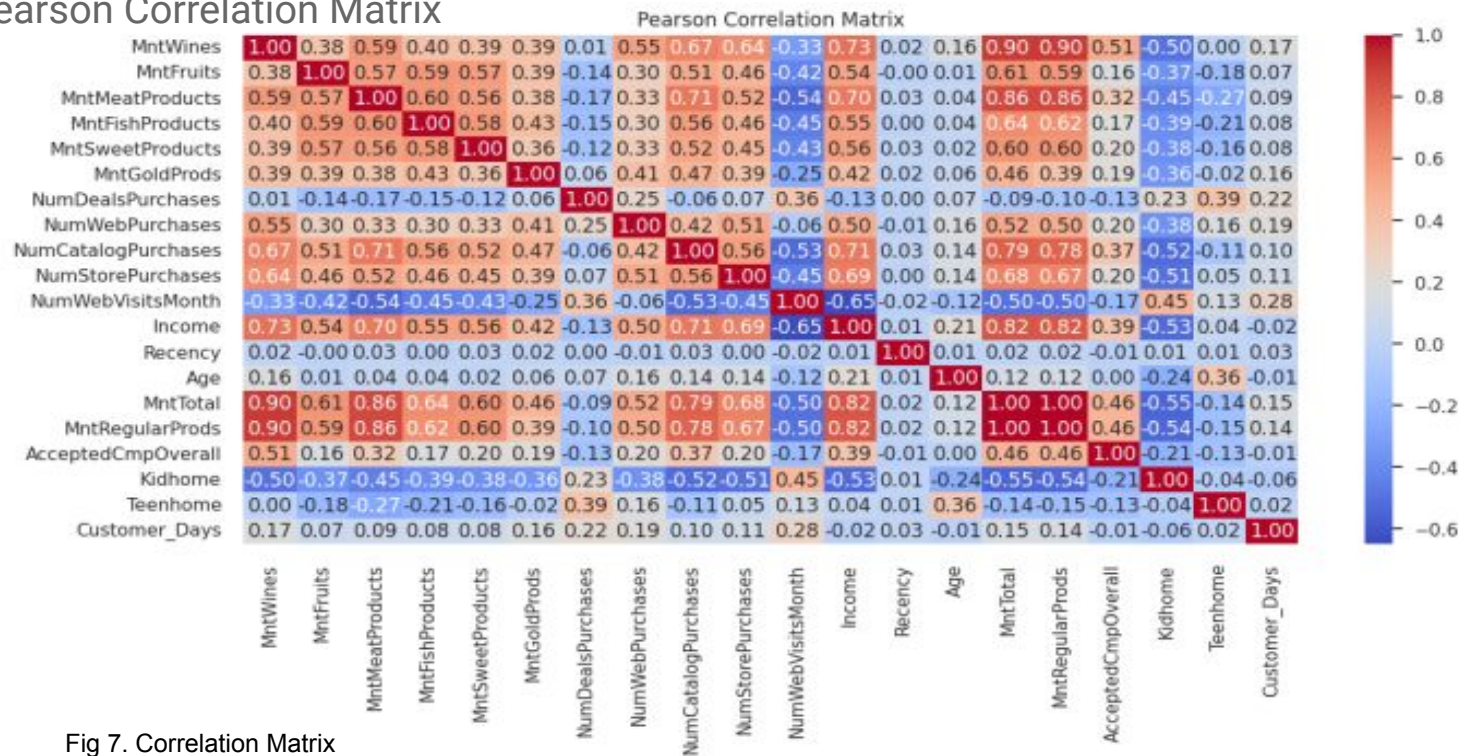


Fig 6. Box-plot of Campaign acceptance versus the numerical variable with stronger association

# Bivariate Analysis – Numerical x Numerical

Pearson Correlation Matrix



- + **Income:**
  - + Amount spent;
  - + Campaign Acceptance
  - Web visits;
  - Kidhome;
  - Purchase in Deals;
- 
- + **Kidhome:**
  - Amount spent;
  - + Web visits;
  - + Purchase in Deals;

Fig 7. Correlation Matrix

# Bivariate Analysis – Numerical x Numerical

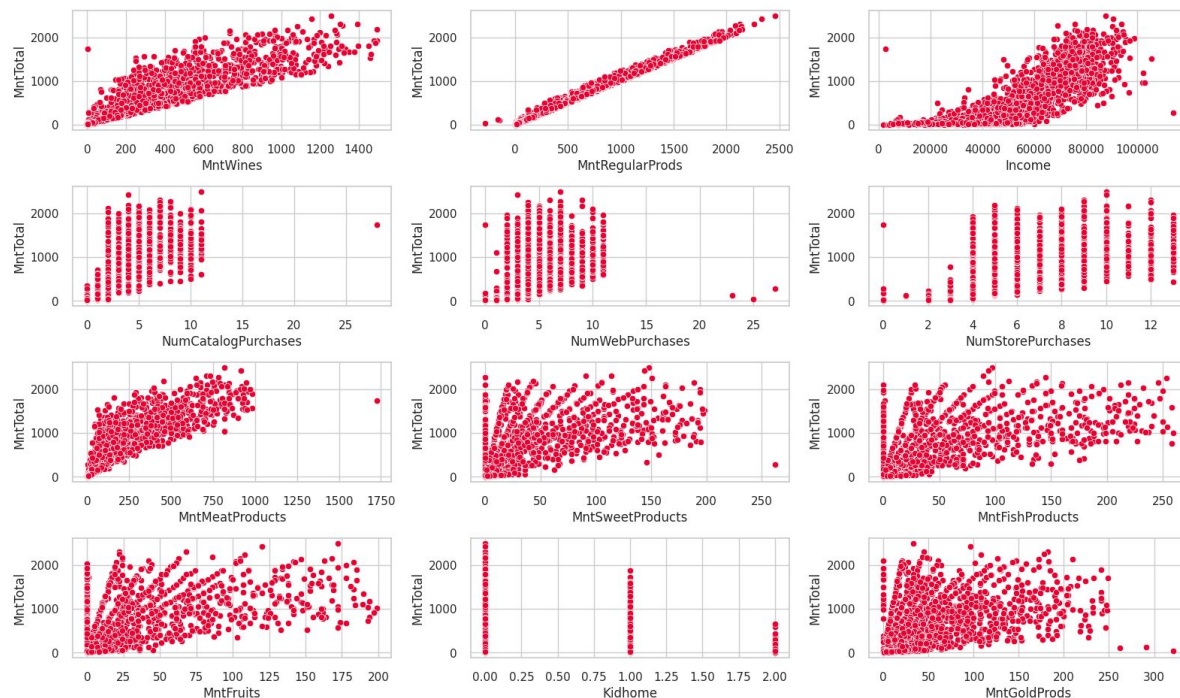


Fig 8. Scatter Plots of Relationships between Total Amount Spent and the Most Correlated Variables



# Multivariate Analysis – PCA

Significant correlations between the numerical variables;

From the total variance of 20 variables, the first 5 components explains 70,23% of variance.

- Kaiser's criterion ( eigenvalues > 1)
- Pearson's criterion (80% exp. variance)
- Cattell's criterion ( Elbow rule)

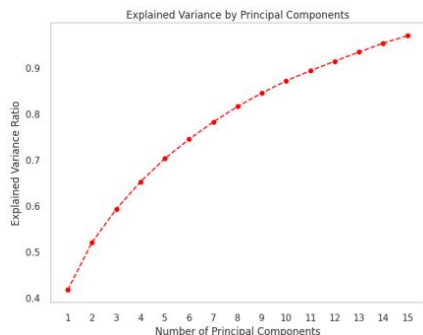


Fig. 9 Explained Variance and Principal Components.

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

Table 2. Principal components, eigenvalues and explained variance.



# Multivariate Analysis – PCA

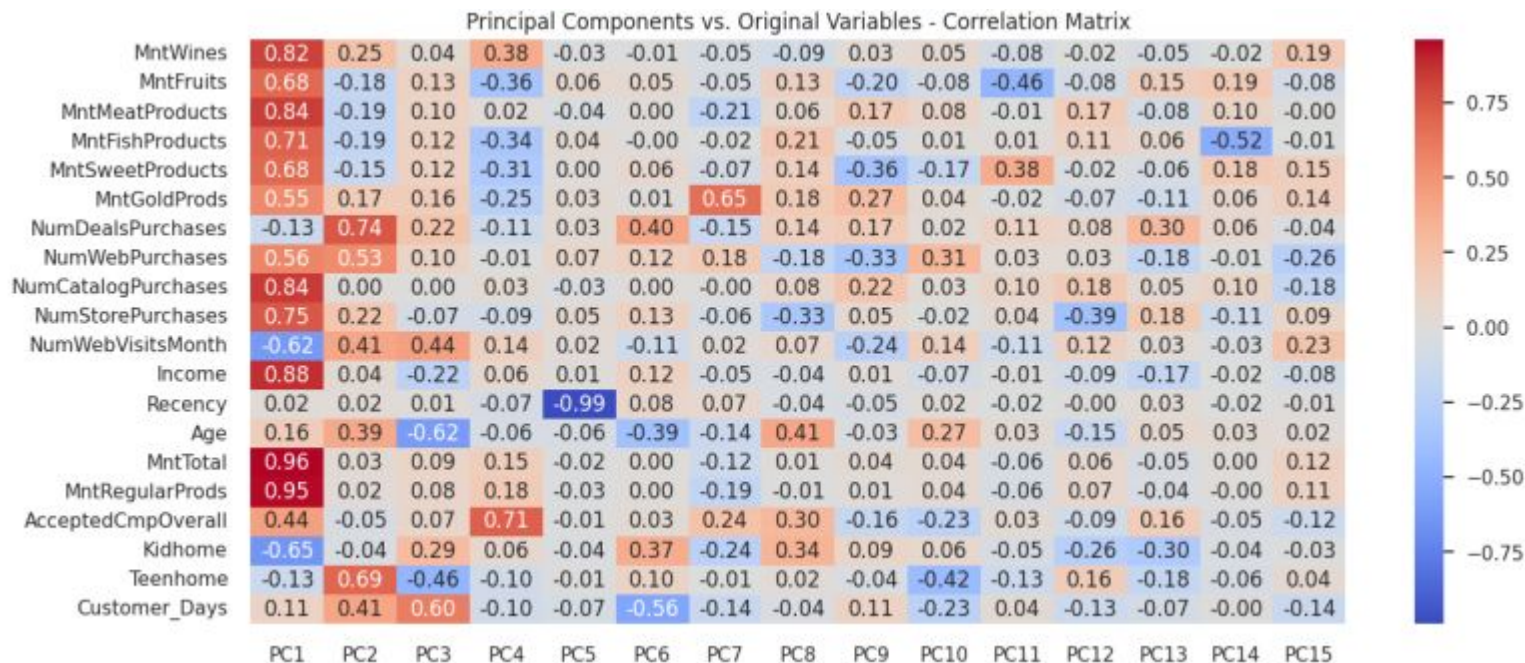


Fig. 10 Correlation matrix of Principal components and original variables.

# Multivariate Analysis – PCA

## 1st Principal Component:

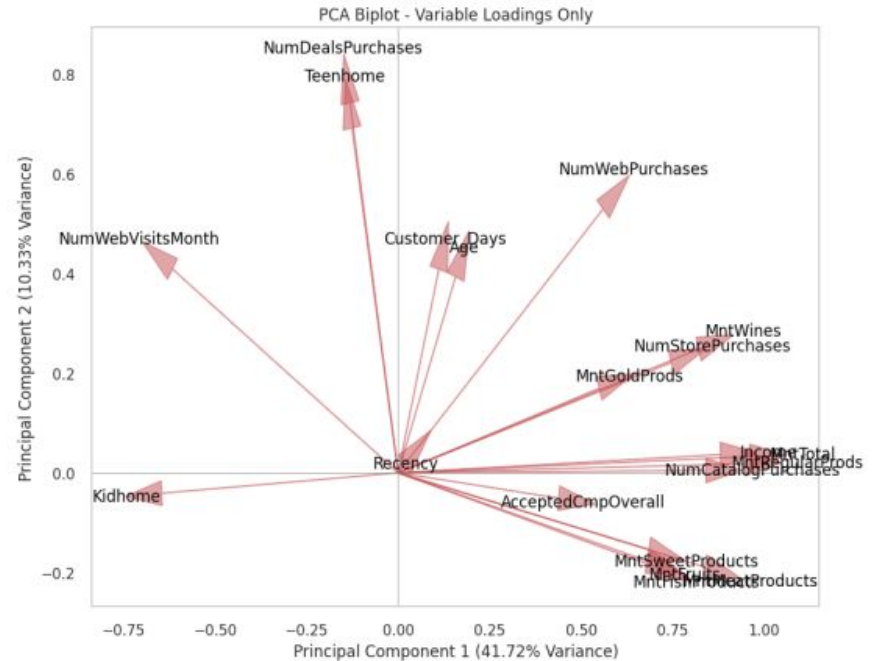
- + Income;
- + Amount spent in the delivery platform;
- + Campaigns acceptance;
- Kids at home;
- Website visits;

⇒ Customers with more money, spend more and have higher campaign acceptance. Possibly they have less kids at home and does less visits to the platform website.

## 2nd Principal Component -

- + Teens at home;
- + The customers have mature age;
- + Customer Days;
- + Purchases in deals;
- + Usage of web platform;

⇒ Customers with higher age, have teens at home and possibly does more deal purchases.



(a) Biplot of PC1 and PC2.

Fig 11. Principal components 1 and 2

# Multivariate Analysis – PCA

The 3rd Principal Component:

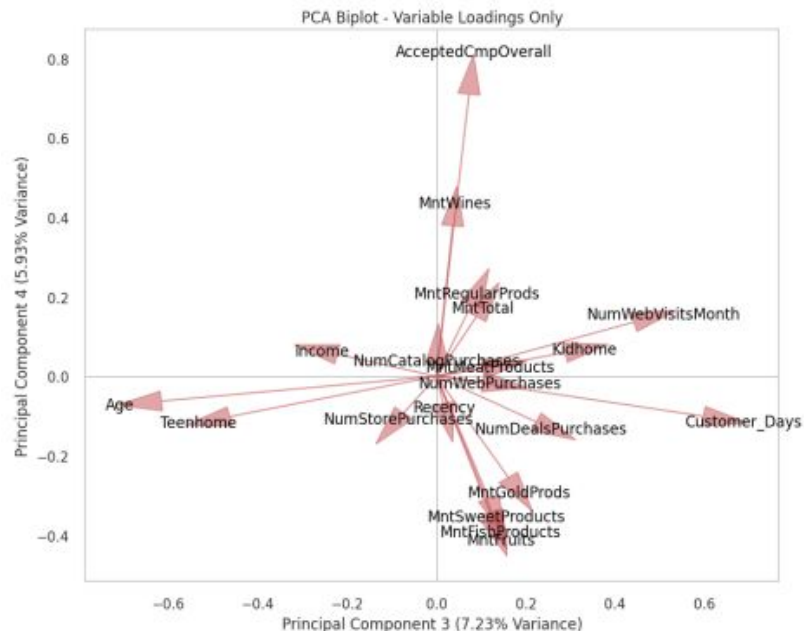
- + Customer Days;
- + NumWebVisitsMonth;
- + Kids at home;
- + Deal purchases;
- Age and Teens at home;

⇒ Customers with kids at home does more visits to the website and purchases in deals. These customers are younger.

The 4th Principal Component:

- + Campaign acceptance;
- + Wine products;
- Traditional products;

⇒ 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.



(b) Biplot of PC3 and PC4.

Fig 12. Principal components 3 and 4

# Multivariate Analysis – Clustering

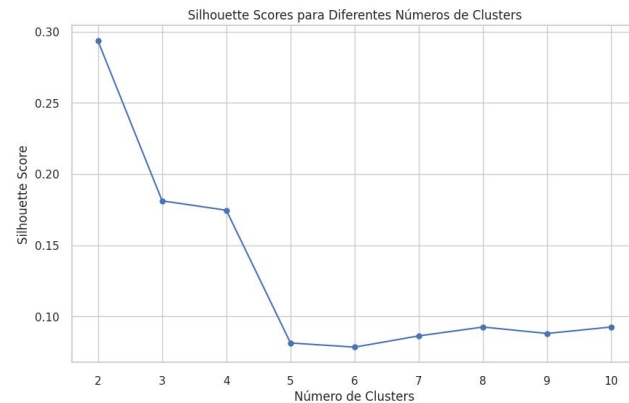
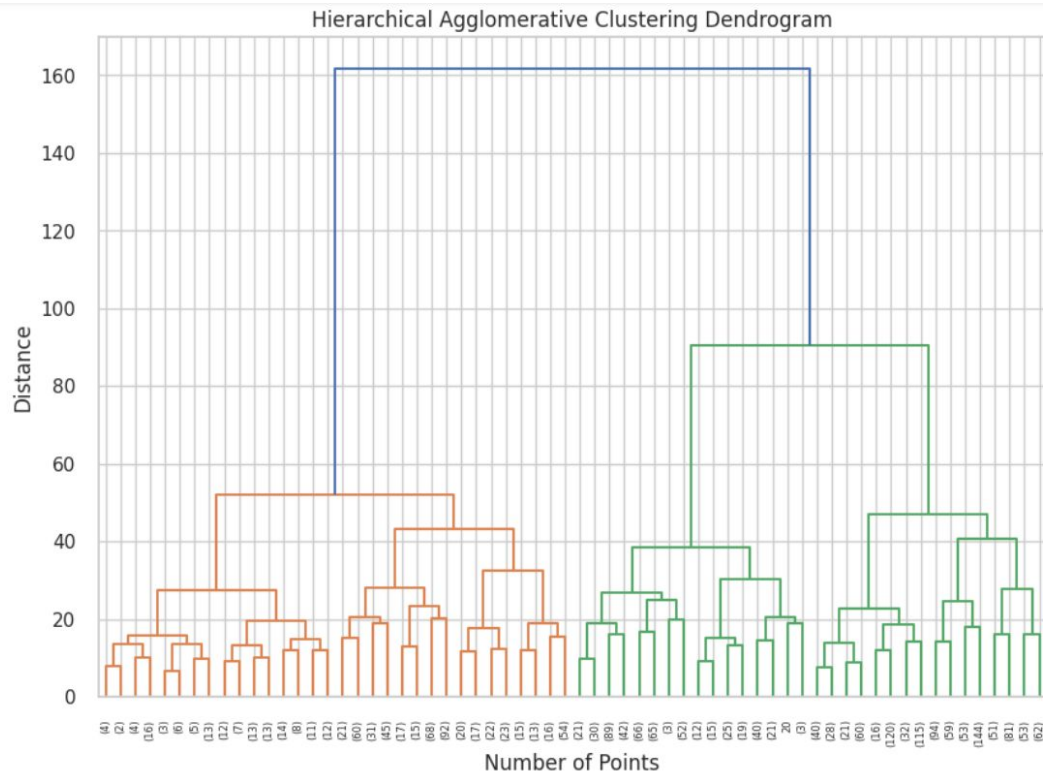


Fig 14. Silhouette index - Cluster determination.

Fig 13, Agglomerative Hierarchical Clustering Dendrogram: Custom  
Segmentation Analysis

# Multivariate Analysis – Clustering

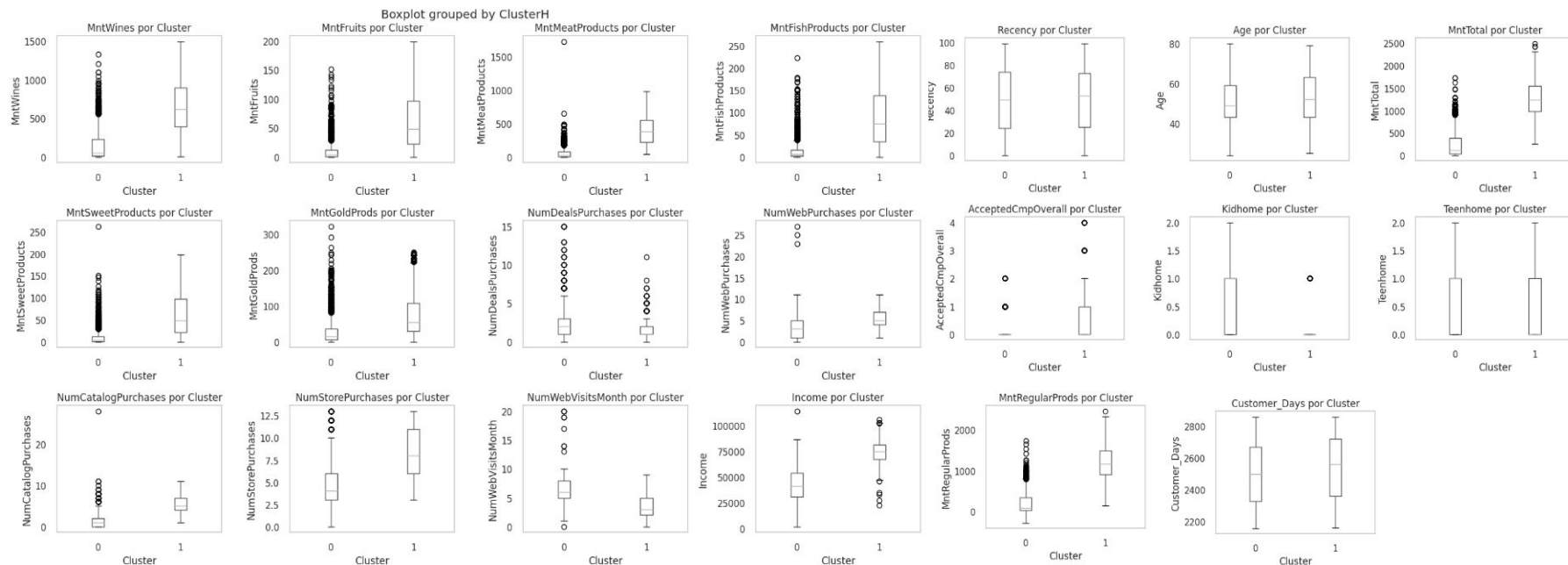


Fig 15. Cluster-wise Distribution: Box Plots

# Conclusion

