NOVA IMS
Information Management School

100% DELICIOUS

Spice Alley: A data-driven marketing campaign

Data Science and Machine Learning

Celso Cassamá, Guilherme Silva, Murilo Oliveira, Nuno Marques

Table of Contents

Introduction

This project aimed to produce a comprehensive report identifying the main segments the customers of Spice Alley restaurant, clustering them into groups of clients that share similar characteristics, with the aim of obtaining homogeneous customer groups across the same cluster, and heterogeneous between clusters.

These groups will then serve as the basis of a preliminary marketing plan to guide future initiatives. This project will allow Spice Alley to better understand their customers, make informed decisions, and potentially allow them to boost the effectiveness of their marketing, sales and product creation campaigns, tailoring them to each group. This will hopefully result in an improved and customized customer experience, reducing expenditure and increasing revenues and profit.

Methodology

The CRoss Industry Standard Process for Data Mining (CRISP-DM) methodology served as the basis for this project, dividing it into the following sections: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

Business Understanding

The Business Understanding stage focused on understanding the business needs that lead to the project request.

The information gathered from the project requirements provided indicated that Spice Alley wanted to identify patterns in customer behavior and improve the effectiveness of their marketing campaigns. The focus of the marketing campaign would be increasing revenue and profitability, as well as allowing their customers to receive campaigns better suited to their needs and tastes, giving them a more tailored and personalized experience.

Data Understanding

The Data Understanding stage focused on better understanding the provided, identifying data quality, and gaining initial insights into underlying patterns and structures.

In order to do so, several actions were performed including loading and merging data, identifying duplicate entries, categorizing the data types of each variable, a first glance at the data, simple statistical analysis, visualizing variable distribution and correlation between variables and general exploratory tasks.

Basic Exploration

The combined datasets had 7000 rows, excluding 31 duplicated observations, across 26 categories. It contained different data types, such as float64, int64, and objects. Some columns contained missing values, on *Education* (14), *Recency* (23), and *MntDrinks* (28). Regarding summary statistics, a comparison between the minimum and maximum, the mean, the median, and the standard deviation of each variable was performed. Some of the initial insights gathered were: for the numerical variables, the Income had a minimum (2493,8) and maximum (237639,725), the mean (77988,96), median (77190), and standard deviation of (35409,8), such values suggest the presence of outliers, due to very a distant maximum value from either the mean or the median values, as well as, a sizable standard deviation. Similar results were found in other variables too, such as on all the monetary ones (*MntMeat&Fish, MntEntries, MntVegan&Vegetarian, MntDrinks, MntDesserts, and MntAdditionalRequest*), *NumOfferPurchases*, and *NumTakeAwayPurchases (see Annexes – Table 4 and 5)*. Skewness wise, a moderate value (between |0.5| and |1.0|) was noticed in the *Income, Kid_Younger6, Children_6to18, Recency, NumAppPurchases, NumStorePurchases, NumAppVisitsMonth* variables, and high skewness (higher than |1.0|) in the *MntMeat&Fish, MntEntries, MntVegan&Vegetarian, MntDrinks, MntDesserts, MntAdditionalRequests, NumOfferPurchases, NumTakeAwayPurchases, Complain, Response_Cmp1, Response_Cmp2,*

Response_Cmp3, Response_Cmp4, Response_Cmp5, which indicate that the distributions were not Gaussian distributions that could impact the performance of some algorithms. At last, and yet regarding numerical variables, the kurtosis has higher than 3 (kurtosis from Gaussian distribution) on the following variables: MntEntries, MntVegan&Vegetarian, MntDrinks, MntDesserts, NumOfferPurchases, NumAppVisitsMonth which again indicate the presence of outliers. On the categorical variables: a pattern was observed in the Name variable regarding the title (Mr., Miss, and Mrs.) before the actual name, Education has 9 levels (Graduation, PhD, Master, HighSchool, Basic, master, graduation, phd, highschool) although we can perceive that should be only 5, the Marital_Status has 10 levels (Married, Together, Single, Divorced, Widow, married, together, single, divorced, widow) but again only 5 different were different in human standards, and Date_Adherence has the following format "2020-09-19 00:00:00" but is of the object type, meaning that, we will need to transform it to perform.

Visual Exploration

The subsequent step involved undertaking a visual exploration to ascertain the distribution of variables and establish fundamental correlations. The objective was to maximize variables of interest to businesses, such as Monetary spending and Purchases, while also considering potential constraints, such as available Income and having children, which are widely recognized as additional expenses in family cost-of-living. To provide a comprehensive overview and uncover potentially valuable insights that may not be immediately apparent, all variables were correlated. The visual exploration results are available in Annexes – Figures 5 to 20.

Data Preparation

After acquiring the data, a plan was devised for its utilization in the modeling phase. It was established that by the end of the Data Preparation phase, the data would be refined to ensure cleanliness, meaning the absence of duplicate observations, missing values, and outliers. This would include data transformation such as creating new variables, correcting inconsistencies (e.g., Monetary spending without any purchase is not a valid observation), reclassifying variables (e.g., PhD and phd refer to the same information), binning of numerical variables, and creating dummy variables for categorical variables. Moreover, the data was transformed to resemble a normal distribution to avoid reducing the effectiveness of the model. Additionally, scaling the data was performed due to the negative impact of different scales on the model's measurement of distances. Finally, the data was reduced by excluding unary variables and checking for and avoiding multicollinearity.

Data Cleaning

Duplicates

Duplicate observations (31) were dropped, resulting in a dataset containing 7000 observations.

Outliers

Observations with an absolute z-score above 3 on *MntVegan&Vegetarian, Income*, *NumTakeAwayPurchases*, and *NumAppVisitsMonth* were dropped, resulting in the removal of 288 observations, representing approximately 2.6% of the data. Although some values in *MntVegan&Vegetarian* could have been retained with a more tailored approach, we opted for a more generalized framework to address the issue, which would be better equipped to handle changes in client profiles, while still maintaining an acceptable data drop rate of less than 3%.

Missing Values

Education, Recency, and *MntDrinks* were identified as having missing values, which were filled using the following criteria: Education was filled with the mode value, as it had the highest number of Graduation levels in the data (~50%). Recency had a very homogeneous distribution, so we used the mean to fill in the missing values. Finally, *KNNImputer* was utilized to fill in the missing values for MntDrinks, using correlated variables including *'MntMeat&Fish'*, *'MntEntries'*, *'MntVegan&Vegetarian'*, *'MntDesserts'*, and *'MntAdditionalRequests'*, with the mean of the three nearest neighbors.

Data Transformation

New variables

Variables were added to the analysis to gain additional insights, build stronger segmentation, or display information in a more comprehensible way. These variables included: Gender - created using the title in the Name feature; Age - calculated using the Birthyear as a reference; *daysAsCardClient* - representing the number of days since the customer subscribed to the Spice Alley card; *MntTotal* - a variable summing up all monetary amounts spent; *Mnt_pday_card* - representing the average daily spending per customer; *Response_Campaigns* - the sum of all *Response_Cmp*[1-5]; *Total_Kids* - the sum of kids from both *Kid_Younger6* and *Children_6to18*; *has_Kids* - a flag indicating whether a customer has kids; Monetary Ratios - six variables, *Pct_Meat&Fish*, *Pct_Desserts*, *Pct_Entries*, *Pct_Drinks*, *Pct_Vegan&Vegetarian*, and *Pct_AdditionalRequests*, which represent the percentage of each monetary category in the total monetary spend (e.g. *Pct_Meat&Fish=Pct_Meat&Fish*/*MntTotal*); *NumTotalPurchases* - the sum of all purchases through the App, TakeAway, or Store; and Behavior Ratios - three variables, *Pct_Store, Pct_App*, and *Pct_TakeAway*, which divide the number of purchases through each channel by the total number of purchases.

Incoherencies

An attempt was made to investigate the existence of real-world inconsistencies, specifically regarding customers who had monetary spending without corresponding purchases, and customers who used offers for purchases but did not make any purchases through available channels (App, TakeAway, or Store). No such inconsistencies were found in either case, although a query was performed out of curiosity prior to removing outliers, which yielded 38 and 33 observations, respectively. Moreover, a discrepancy was discovered during the conversion of *Date_Adherence* to datetime on 2/29/2022, and a decision was made to modify it to 2022-03-01.

Reclassification

First, normalization of capitalization was performed on the *Education* and *Marital_Status* variables. Then, levels on each variable that were considered similar were joined. Specifically, the *Basic* level was merged into *HighSchool* on the *Education* variable, and *Togheter* was merged into *Married* in the *Marital_Status* variable.

Skewness correction

Non-normal distributions were observed during the visual exploration of several variables, including *MntMeat&Fish*, *MntVegan&Vegetarian*, *MntEntries*, *MntDrinks*, *MntDesserts*, *MntAdditionalRequests*, *NumOfferPurchases*, and *NumTakeAwayPurchases*. To achieve a more normal-like distribution, a logarithm base 10+1 transformation was applied to these variables.

Binning and Dummy Variable Creation

A decision was made to create 5 age bins after a histogram distribution showed different groups of customers due to a significant drop in customers in certain ages, as suggested by image x. Dummy variables were then created for the *Education* and *Marital_Status* levels, which were string-based variables, since distances to strings are a misleading measurement.

Scaling

The last step in the data preprocessing was to ensure that every feature was in the same range, for which scaling was applied. Initially, the *MinMaxScaler* was used, but later it was decided to use *StandardScaler*, which appeared to yield better results.

Data Reduction

Before modeling, the features to be used were selected and those to be dropped were determined. The features that were replaced by others, such as Birthyear, Date_Adherence, Education, Marital_Status, and Name, were dropped. The unary variable Complain, which had 98.9% of flag 0, was also dropped.

To assist with feature selection from a customer value perspective and maximize the MntTotal output, a routine based on the RandomForest was created. Afterward, Principal Component Analyses were applied to reduce the data on the following variables: Income, Recency, MntMeat&Fish, MntEntries, MntVegan&Vegetarian, MntDrinks, MntDesserts, MntAdditionalRequests, NumOfferPurchases, NumAppPurchases, NumTakeAwayPurchases, NumStorePurchases, NumAppVisitsMonth, daysAsCardClient, Married, Single, Widow, NumPurchasesTotal, Response_Campaigns, Kid_Younger6, age_(17.943, 29.4], age_(29.4, 40.8], age_(40.8, 52.2], age_(52.2, 63.6], and age_(63.6, 75.0].

Modeling

Implementation

PCA was used before implementation of the clustering models, as previously stated. A threshold of 80% explained variance was established and the number of principal components was decided based on this value. The resulting dataframe was then used for the implementation of the clustering methods. The variables selected for PCA differed between K-Means and K-Prototypes, as only numerical variables were selected for the K-Prototypes implementation.

In order to implement the clustering method, it was necessary to define the number of clusters to be used. That was achieved using a combination of the Elbow Method and the total sum of squares of the distance of both data points and their clusters and between cluster centers. K-Means was used as a baseline for applying these methods.

The Elbow Method consists of plotting the inertia of the clusters against the total number of clusters and choosing the point where the curve begins to flatten. That point is then defined as the appropriate number of clusters. After obtaining the number of clusters k, both the sum of squares of the distances between and within clusters was calculated, for k-1 and k+1, in order to ensure the number of clusters is appropriate.

After settling on the number of clusters, K-Means clustering was performed with 4 clusters and a maximum of 100 iterations on the previously selected variables. The resulting labels were then appended to the original dataset and a table containing the information for each cluster, with numerical variables being described by their mean and categorical ones. The euclidean distances between each pair of cluster centers were also computed.

K-Prototypes implementation followed the same pipeline, with the main difference being the separation of categorical and numerical columns. The other processes, including the decision of the number of clusters, the verification of said number, clustering and addition of cluster labels to the dataset, remained the same. The results for each cluster were also summarized in a table, using the same data treatment for numerical and categorical variables as in the K-Means implementation. Euclidean distances between cluster centers were also computed.

Model Comparison

Based on the results presented, it appears that the K-Prototypes algorithm was superior to K-Means due to its higher Calinski-Harabasz Index. The index measures the relationship between the variance between groups and the variance within groups, thus indicating that the groups formed by K-Prototypes were more separated and distinct.

The K-Prototypes algorithm was chosen over K-Means because of a higher Calinski-Harabasz Index observed in the presented results. This index, which measures the relationship between the variance between groups and the variance within groups, indicates that the formed groups were more separated and distinct. However, it is worth noting that the Silhouette Score, which assesses the quality of clustering based on intra-cluster and inter-cluster distance, showed a slightly higher score for K-Means compared to K-Prototypes, although the difference was very small (0.151 vs. 0.140). The advantage of K-Prototypes in

handling mixed variable types, such as categorical and numerical variables, could be another reason for its choice, as it creates a larger distance between group centroids.

Evaluation

The deployment of the clustering technique detailed on the Modeling section, provided us the following four distinct clusters (Table 1- Summary table for category values by cluster) As it is observable in images x and y, there are significant differences between these clusters on a consumer behavior perspective, as well as, in customer lifetime value.

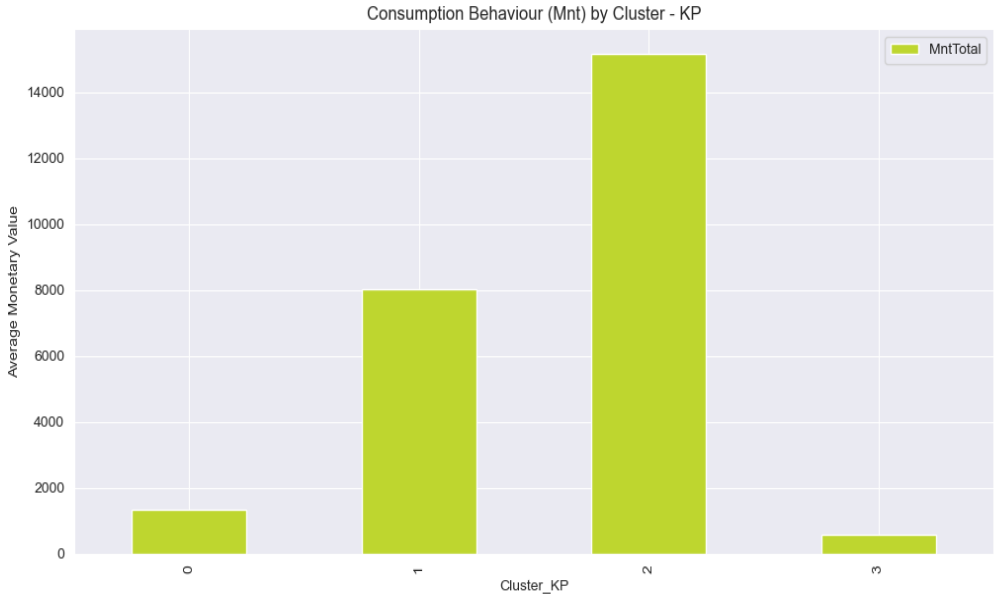

Figure 1- Average Monetary Value spent by cluster.

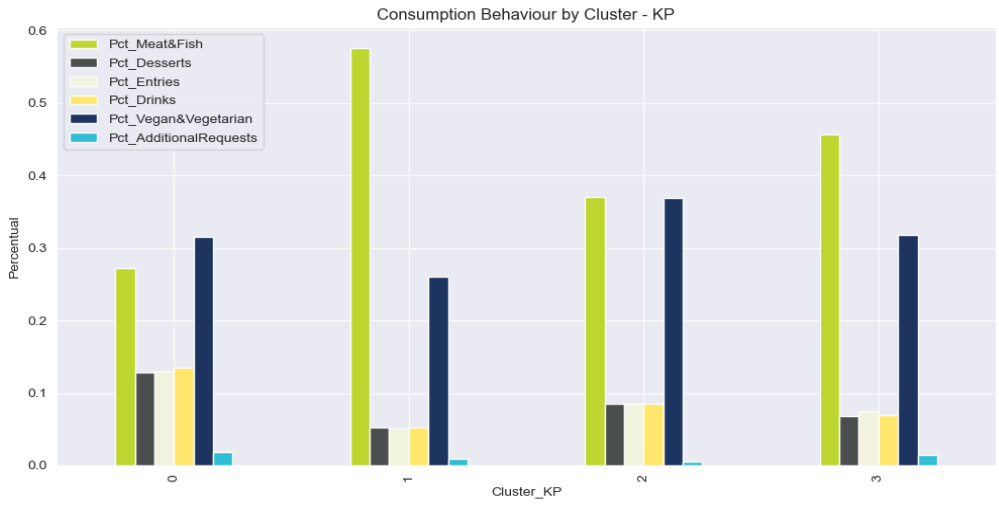

Figure 2- Percentage spent in each product category by cluster.

Figure 3- Percentage of purchases made in each Trade Channel by cluster.

The identified clusters can be described as such:

Cluster 0 alias "Family Pack" (Low Monetary Value)

This cluster is comprised of 1529 individuals, with the following characteristics:

- Lowest average Income
- 2nd lowest average spend
- Highest % of kids younger than 6 in the household
- Highest spent % in Desserts, Entries and Drinks
- Most purchases via app

Cluster 1 alias "Loyalty" (Medium-High Monetary Value)
This cluster is comprised by 1646 individuals, with the following characteristics:

- 2nd highest average *Income*
- Highest % on Offer Purchasing
- Highest % spent on *Meat&Fish*
- Lowest% spent on *Vegan&Vegetarian*
- Most purchases via app
- Highest average antiquity as company's card adherent

Cluster 2 alias "Indulgence Seekers" (High Monetary Value)

This cluster is comprised by 2280 individuals, with the following characteristics:

- Highest average Income and Monetary value spent
- Lowest % of kids in the household
- Highly Educated
- Highest Campaign Response %

- Highest spent % in Vegan&Vegetarian
- Most purchases via store
- Highest % on TakeAway Purchases between the four identified clusters

Cluster 3 alias “Value-Conscious” (Low Monetary Value)

This cluster is comprised by 1357 individuals, with the following characteristics:

- 2nd lowest average Income
- 2nd highest % of kids younger than 6
- Lowest average Monetary value spent.
- Lowest campaign response %
- Lowest spent % in Desserts, Entries and Drinks
- Most purchases via app

Deployment

Regarding the Marketing Plan, it should be directed and tailored for each cluster. Here are our recommendations:

Cluster 0 - "Family Pack":

Based on the characteristics of this cluster, the marketing plan should focus on providing family-friendly deals, such as meal packages that include dessert and drinks. As most of the purchases are made via the app, it is crucial to ensure that the app is user-friendly and easy to navigate. Additionally, the marketing plan should emphasize the convenience of using the app to make purchases, especially for busy parents who have young children. Finally, the marketing team can consider offering discounts or loyalty rewards to encourage repeat purchases.

Cluster 1 - "Loyalty":

This cluster has the second-highest average income and the highest percentage of offer purchasing, making it a key target group for the company. The marketing plan should focus on providing personalized promotions and offers that are focused on maximizing the consumption of meat/fish dishes, that are the main preference of this cluster. As a consequence, vegan and vegetarian dishes should not be part of a marketing strategy regarding this cluster. As most of the purchases are made via the app, the marketing team should ensure that the app provides a seamless user experience and that promotions are prominently displayed. Finally, the marketing plan should highlight the company's commitment to quality and loyalty plans, which are likely to appeal to this cluster's discerning and loyal customers. This cluster, having around 80% of Cluster 2 Income, spends only 50% as much, which signals a high growth potential that Space Alley should focus on.

Cluster 2 - "Indulgence Seekers":

This cluster has the highest average monetary value spent and the highest campaign response rate, making it an important target group for the company. The marketing plan should focus on emphasizing the quality and exclusivity of the products, as well as their focus on main courses, both vegan&vegetarian and meat/fish options. Since this cluster tends to make purchases via the store, the marketing team should ensure that the physical stores are well-maintained, with an attractive display of products and an inviting atmosphere. As this cluster also has the highest % of purchases via TakeAway, Space Alley should also make an effort to ensure a reliable (on time and in full) service level on this service.

Cluster 3 - "Value-Conscious":

This cluster has the lowest average monetary value spent and the lowest campaign response rate, making it a challenging target group for the company. The marketing plan should focus on emphasizing the affordability of the products and offering discounts or promotions that are attractive to this cluster. As most of the purchases are made via the app, the marketing team should ensure that the app is easy to use and that promotions are prominently displayed. Additionally, the marketing plan should emphasize the convenience of using the app, especially for time-pressed parents with young children. Finally, the marketing team can consider partnering with other companies or organizations that are popular with this cluster to increase brand visibility and appeal.

Conclusion

In conclusion, understanding the different clusters of customers is crucial for Space Alley's marketing strategy. By analyzing the purchase behavior and characteristics of each cluster, the company can tailor its marketing plan to appeal to each group. The Family Pack cluster should be targeted with family-friendly deals and an emphasis on convenience, while the Loyalty cluster should be provided with personalized promotions and a focus on meat/fish dishes. The Indulgence Seekers cluster should be targeted with exclusivity and an emphasis on both vegan&vegetarian and meat/fish dishes, and the Value-Conscious cluster should be targeted with affordability and discounts. By effectively targeting each cluster, Space Alley can maximize its marketing efforts and increase customer satisfaction and loyalty.

Bibliography

[1]K. B. Salem and A. B. Abdelaziz, “Principal Component Analysis (PCA),” *Springer eBooks*, pp. 1–4, Jan. 2020, doi: 10.1007/978-3-030-03243-2_649-1.

[2] Z. Huang, “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values,” *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, Sep. 1998, doi: 10.1023/a:1009769707641.

Annexes

PCA

Principal Component Analysis (PCA) is a statistical technique used in data analysis and machine learning in order to reduce the dimensionality of the dataset. The technique involves transforming a high-dimensional dataset into a lower-dimensional space while retaining most of the variance of the data contained within it. This is achieved by creating a new set of orthogonal variables, principal components (PCs), which are created by performing linear combinations of the original variables that capture a certain amount of variance in the data. Each subsequent PC will explain less variance, which leads to the need to find a balance between the number of new variables and the amount of variance described. [1]

PCA is primarily applied when dimensionality reduction is considered important as a part of the data preprocessing phase. It is a widely used technique for simplifying complex datasets, improving the efficiency of running certain machine learning algorithms and reducing the noise in the dataset. Furthermore, it also helps in handling the effect of redundant variables and variables which have low variance.

An example of the application of PCA can be seen in Figure 4- Left: Principal components derived from height and weight. Right: Projection of the datapoints against the first principal component. In *Left*, the principal components derived from the dataset can be seen, rotating the original data so that the first principal component contains the maximum variance. *Right* contains only the first PC, including the projections of the data points against this new variable. The variance lost can be seen by the distance from each data point to the PC. This example describes the tradeoffs involved in using PCA for dimensionality reductions, compromising on data variance in order to reduce the dimensionality of the data.

In conclusion, PCA is a useful technique that has many applications in data driven fields. It enables the reduction of the dimensionality of data, one of the foremost problems when handling big data, the handling of noisy or otherwise low-quality data and redundant information, resulting in data that is more manageable and efficient.

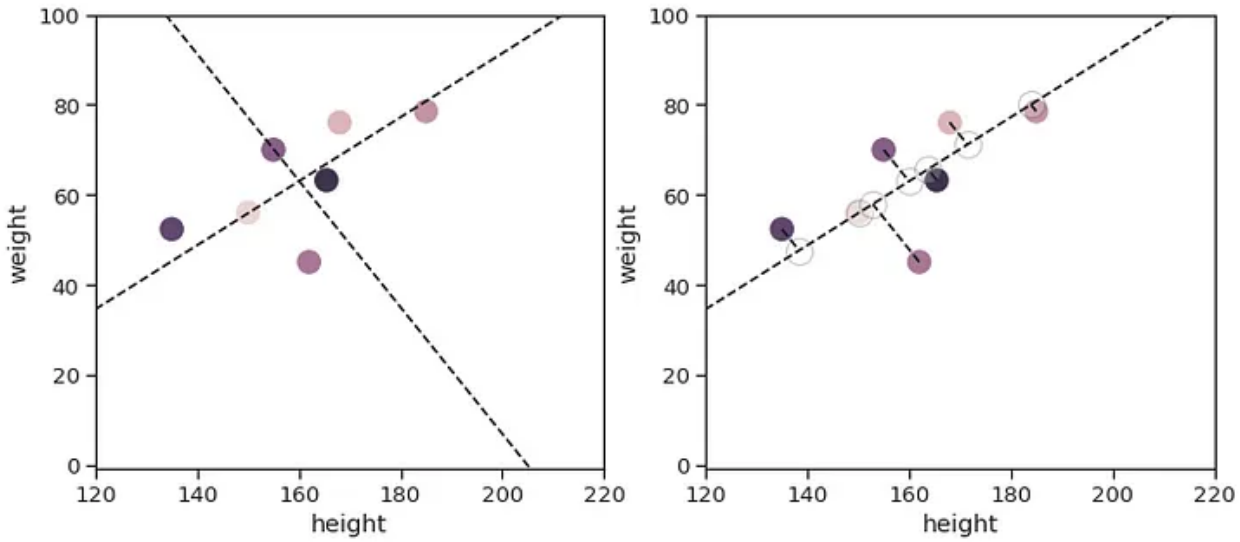

Figure 4- Left: Principal components derived from height and weight. Right: Projection of the datapoints against the first principal component.

K-Prototypes

K-Prototypes is a type of clustering algorithm that handles both numerical, as in K-Means, and categorical data, as in K-Modes. It was first developed by Z. Huang and integrates both K-Means and K-Modes, enabling the creation of clusters in mixed datasets. [2]

While certain algorithms, such as hierarchical clustering, are able to handle mixed data, the computational cost associated with these often make them inefficient or even impossible to implement in large datasets. On the other hand, algorithms which are able to handle large amounts of data, such as K-Means, are limited by the type of data contained in the dataset. This can sometimes be solved by converting categorical data into numeric, however results are often not meaningful, especially when categories are not ordered.

As previously stated, K-prototypes works by combining K-Means' ability of handling large numerical datasets and K-Modes' ability of handling large amounts of categorical data. It hinges on the dissimilarity measure between sets of data, a measure of the differences between points in each category, with smaller values signifying more similar datasets. Clusters are then formed iteratively, by trying to reduce the sum of dissimilarities between points and cluster centers. The centers are calculated by combining the means of the numerical variables and the modes of the categorical ones, which are then weighted in order to avoid favoring one type of data.

In summary, K-Prototypes clustering is an encompassing clustering technique, enabling the usage of large mixed datasets. It combines both K-Means and K-Modes, resulting in an algorithm that solves most of the shortcomings of each of its components.

Figures

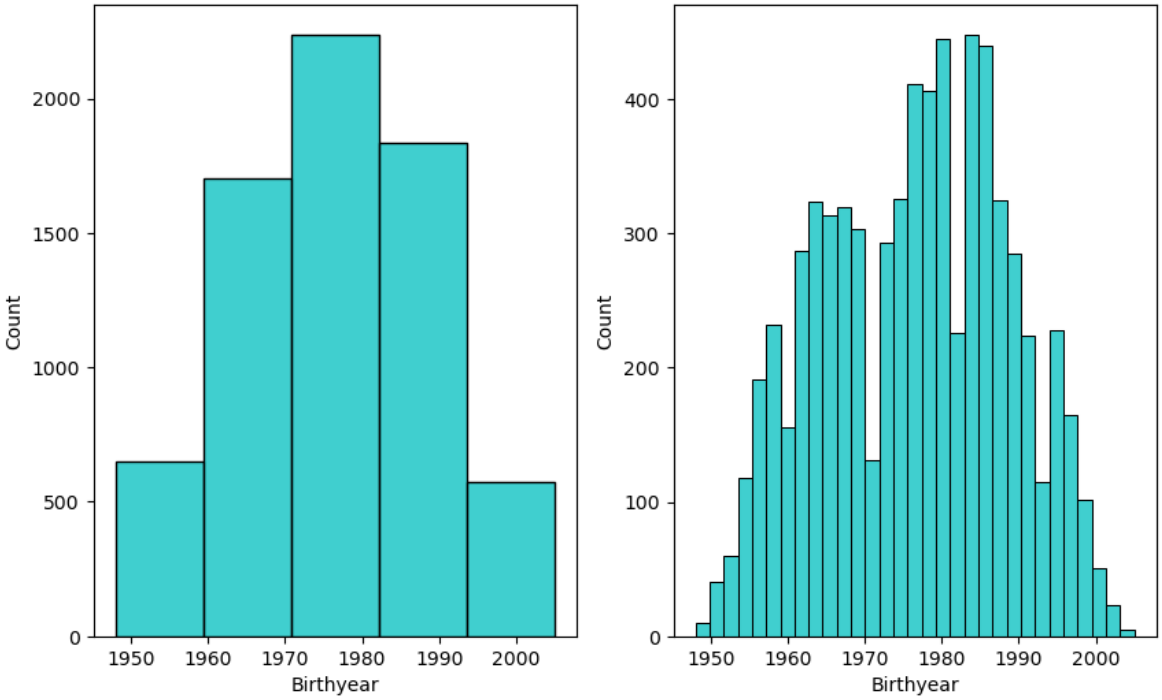

Figure 5- Histogram of Birthyear. Left: 5 bins. Right: 31 bins.

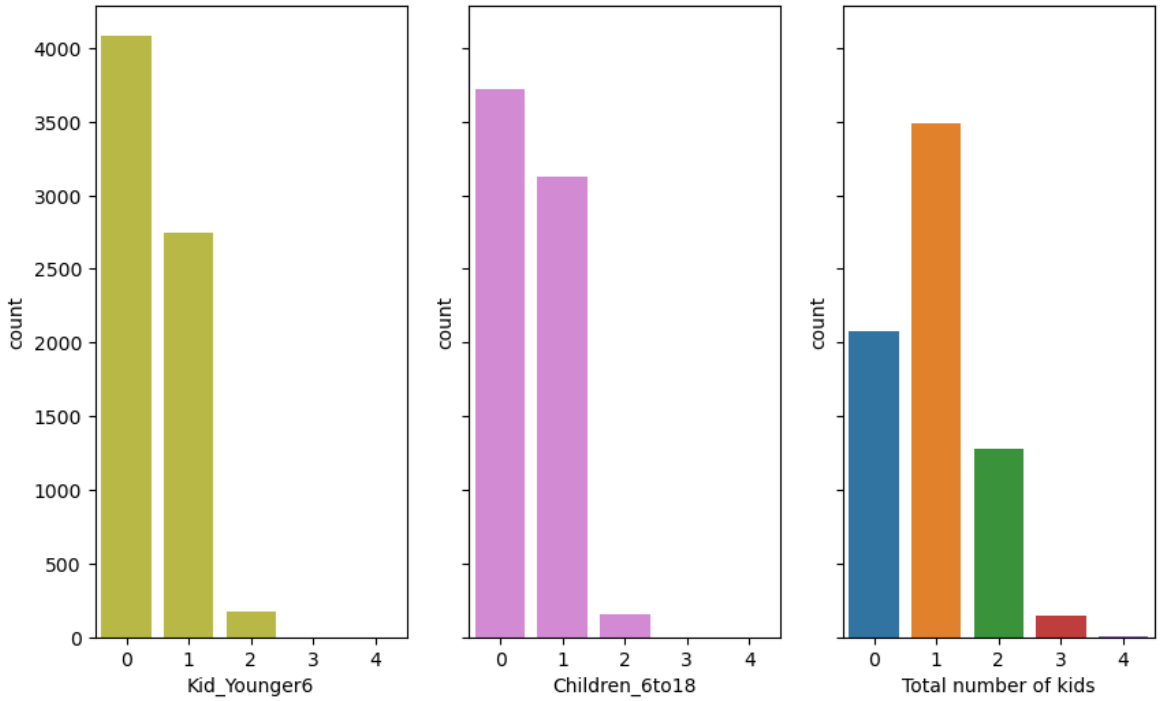

Figure 6- Left: Histogram of Kid_Younger6. Center: Histogram of Children_6to18. Right: Histogram of the total number of kids.

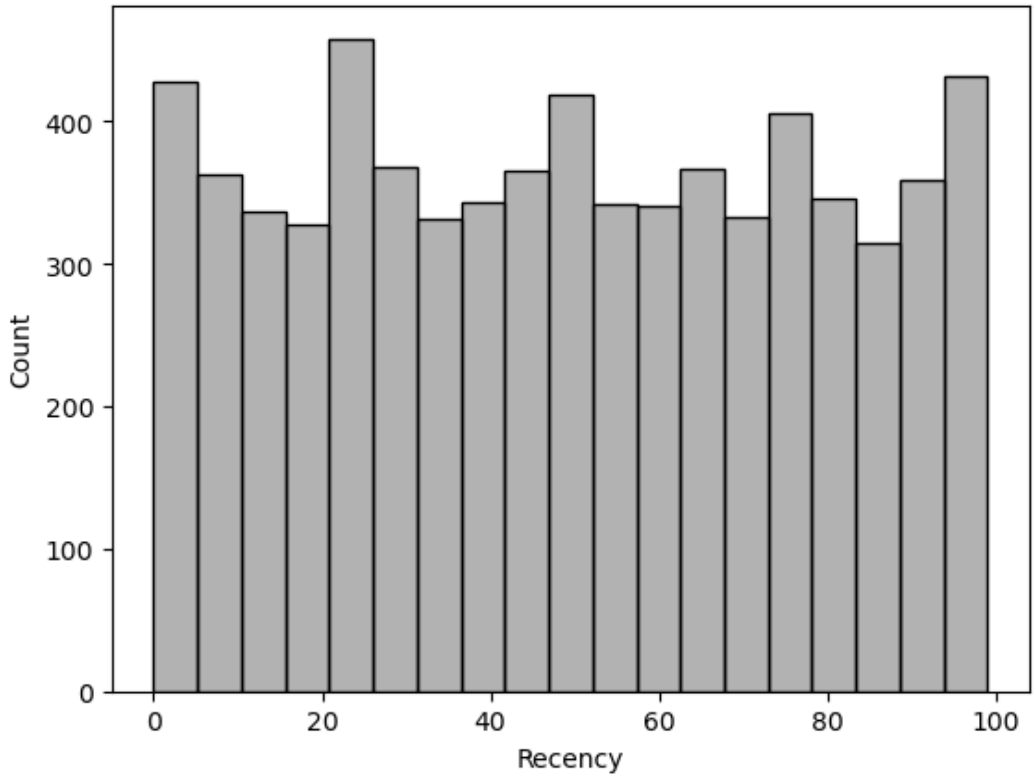

Figure 7- Histogram of Recency.

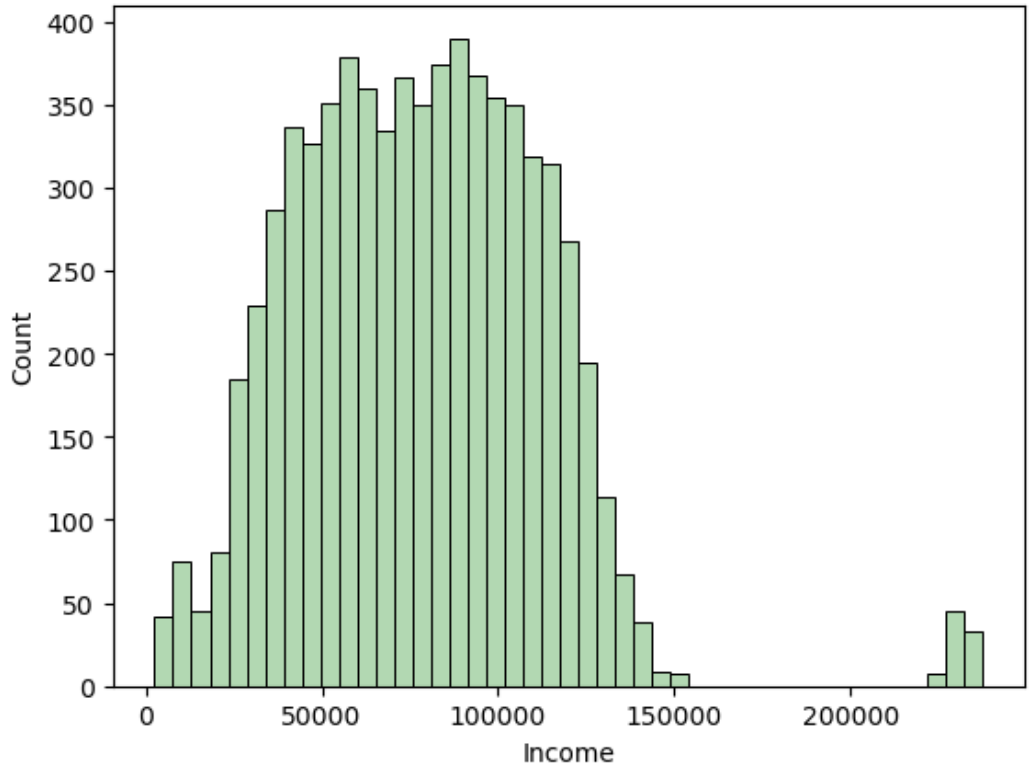

Figure 8- Histogram of income

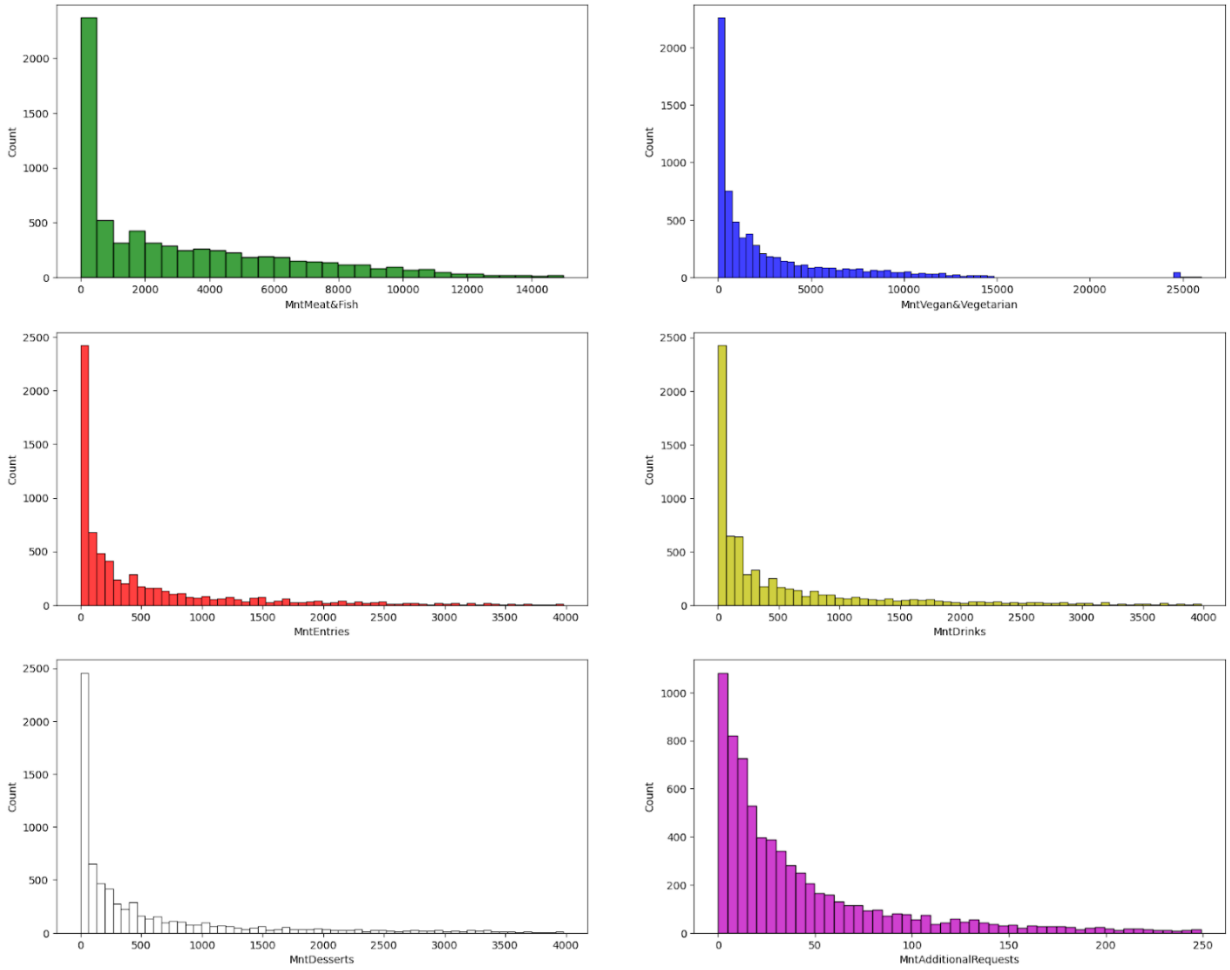

Figure 9- Histogram of all Mnt variables.

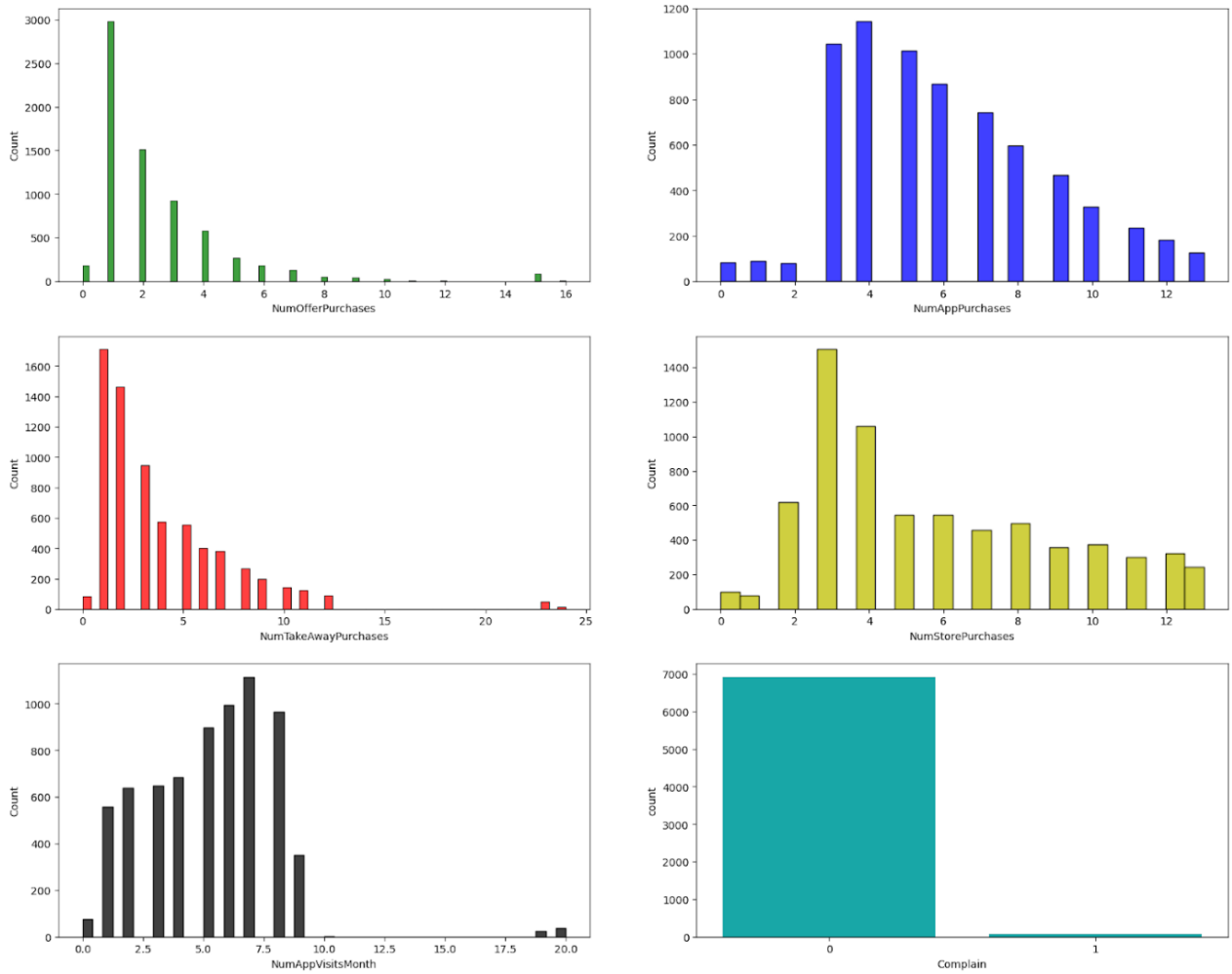

Figure 10- Histograms of all Purchases, NumAppVisitsMonth and Complai.

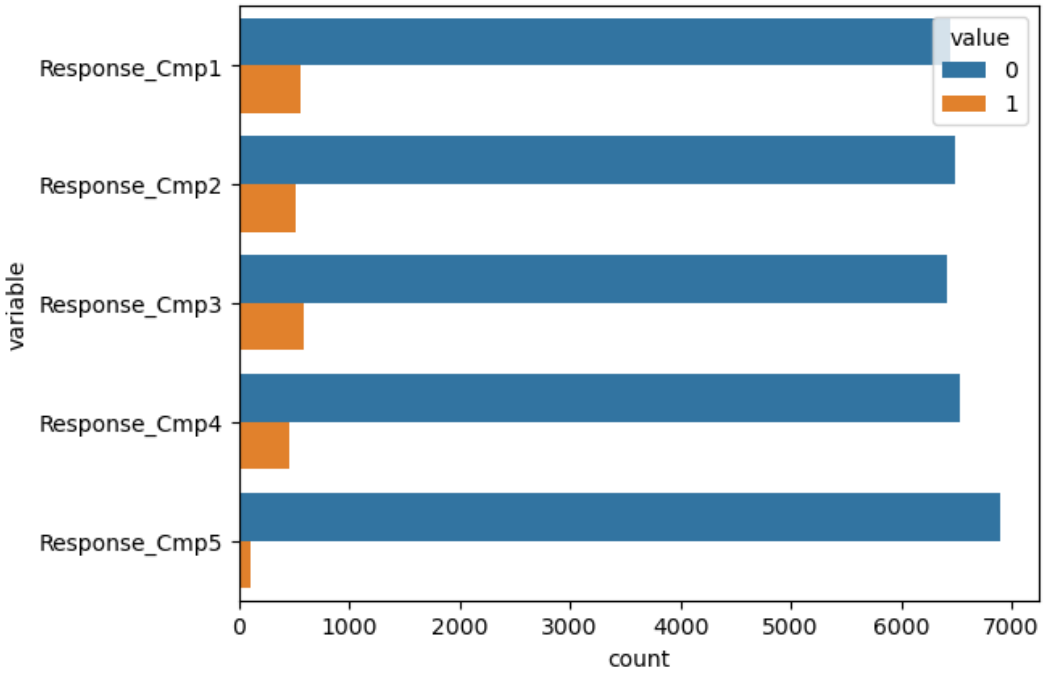

Figure 11- Countplot of all the Response_Cmp[1-5].

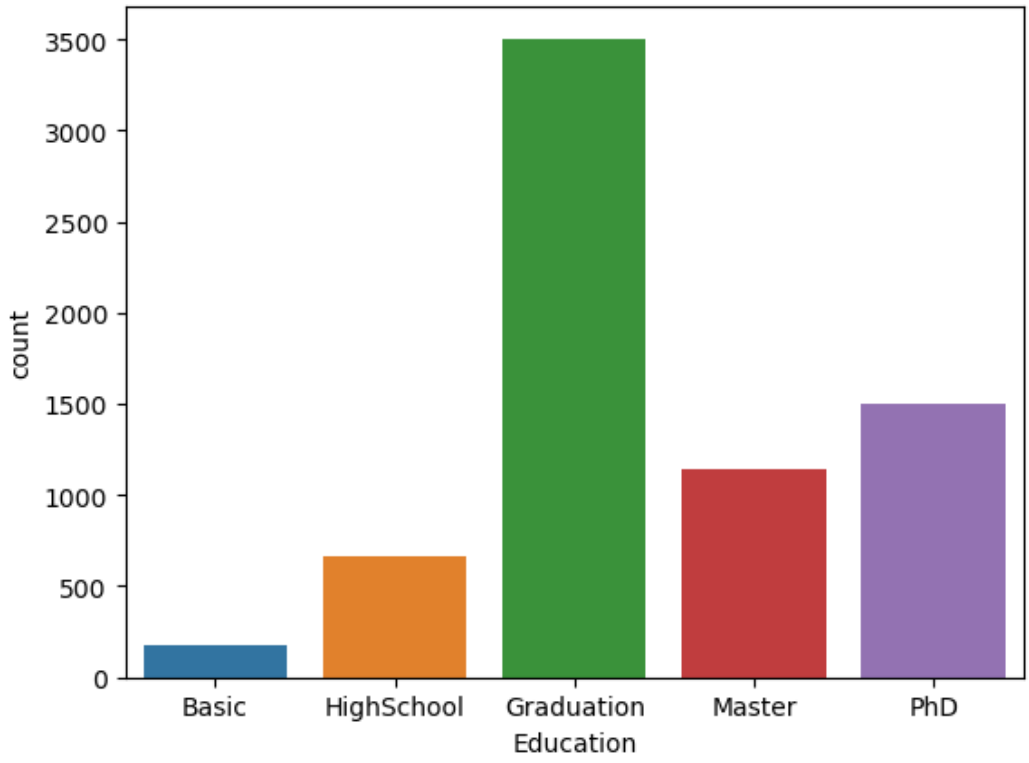

Figure 12- Histogram of Education.

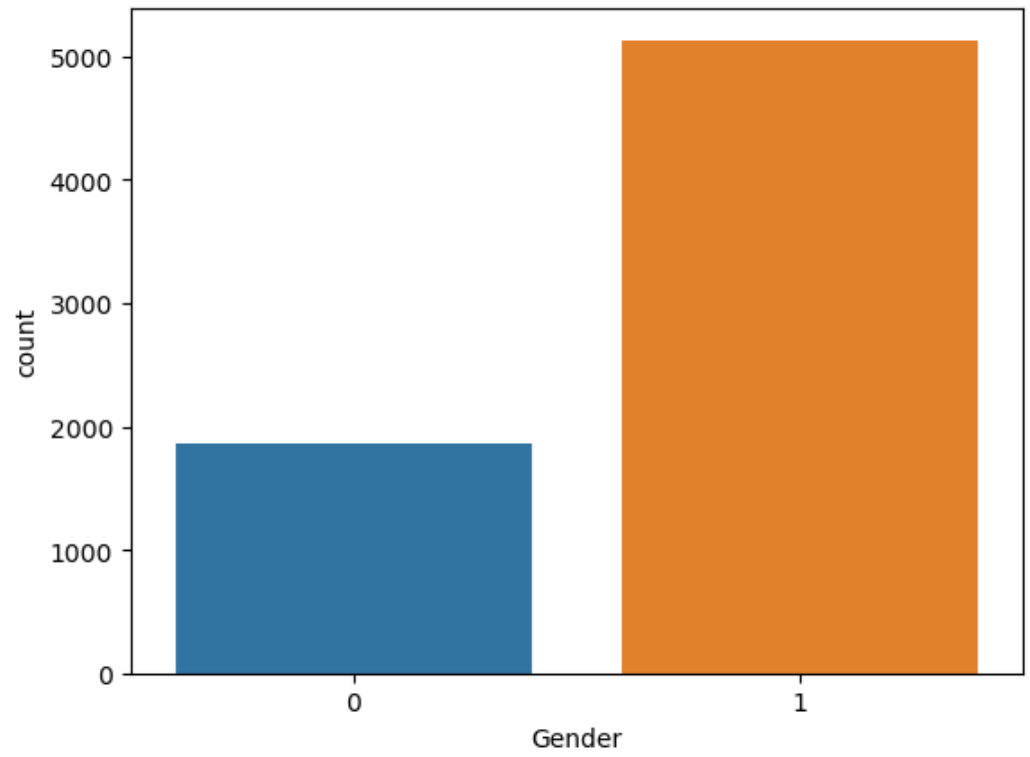

Figure 13- Countplot of Gender.

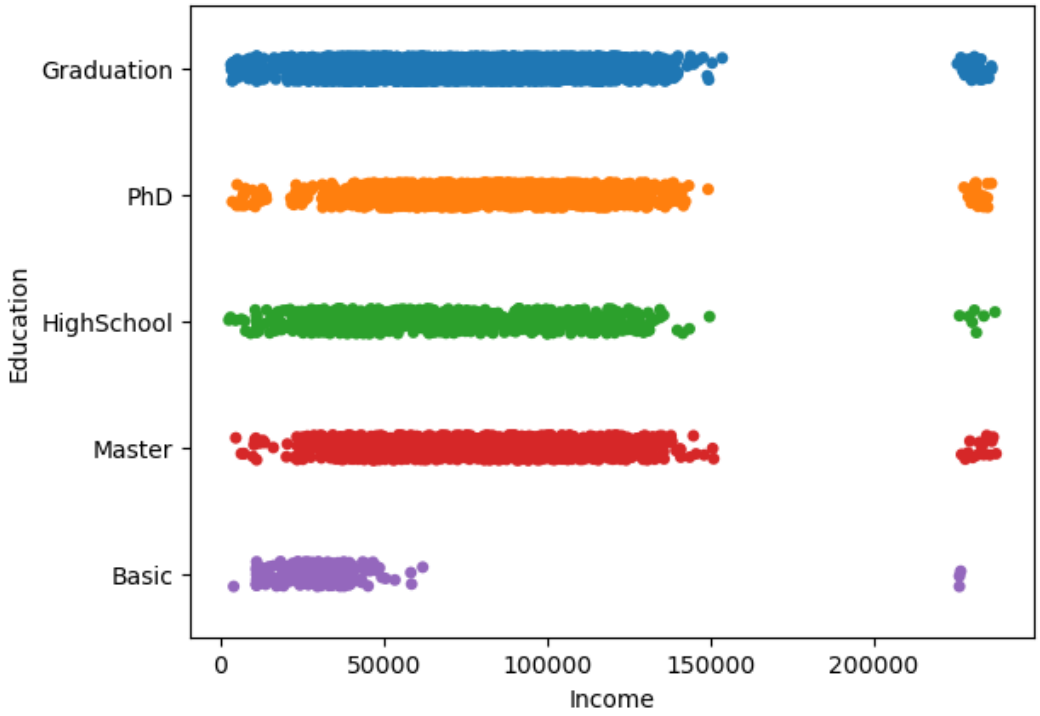

Figure 14- Income vs Education.

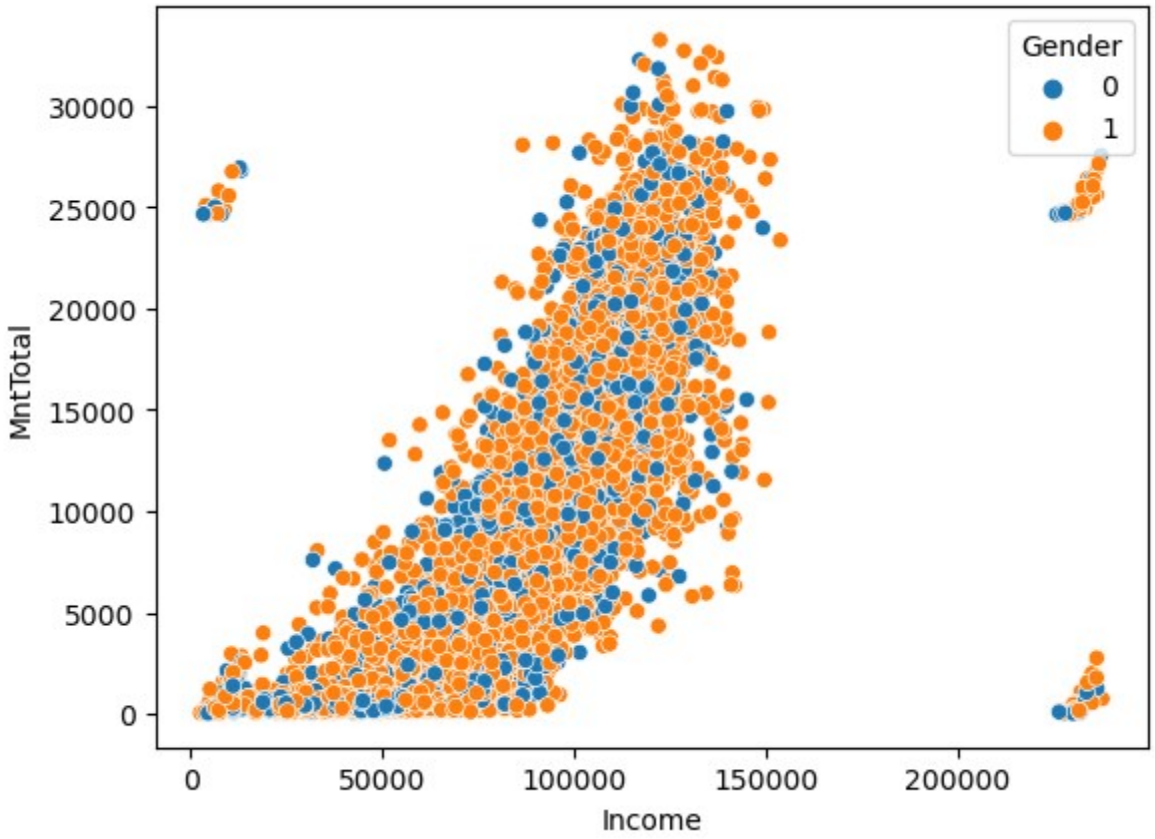

Figure 15- Income vs MntTotal.

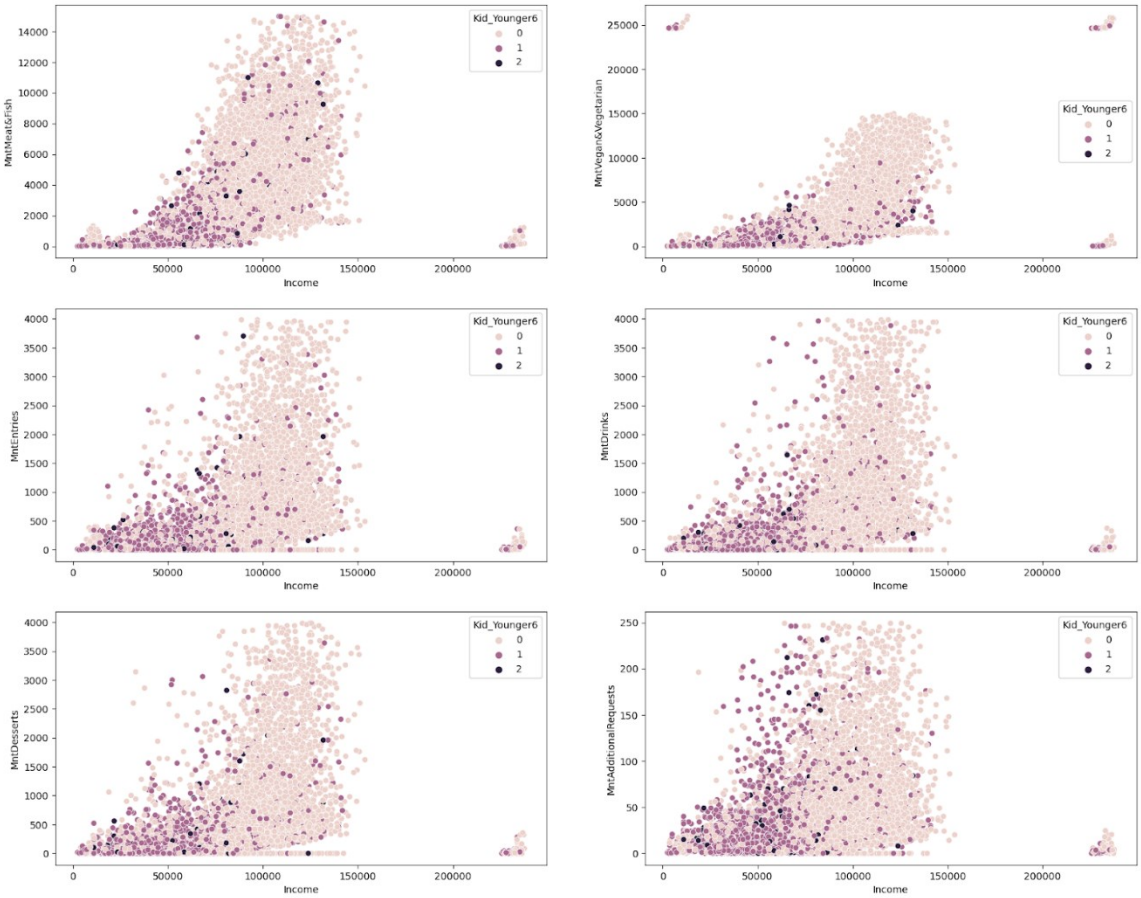

Figure 16- Income vs all Monetary vs Kid_Younger6.

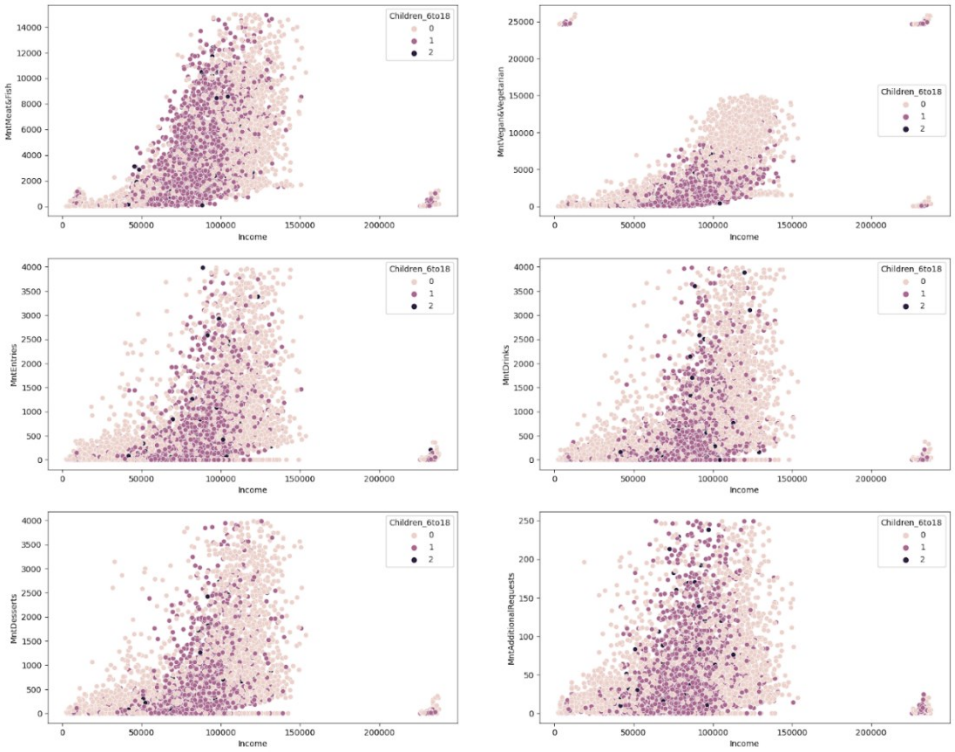

Figure 17-Income vs all Monetary vs Children_6to18.

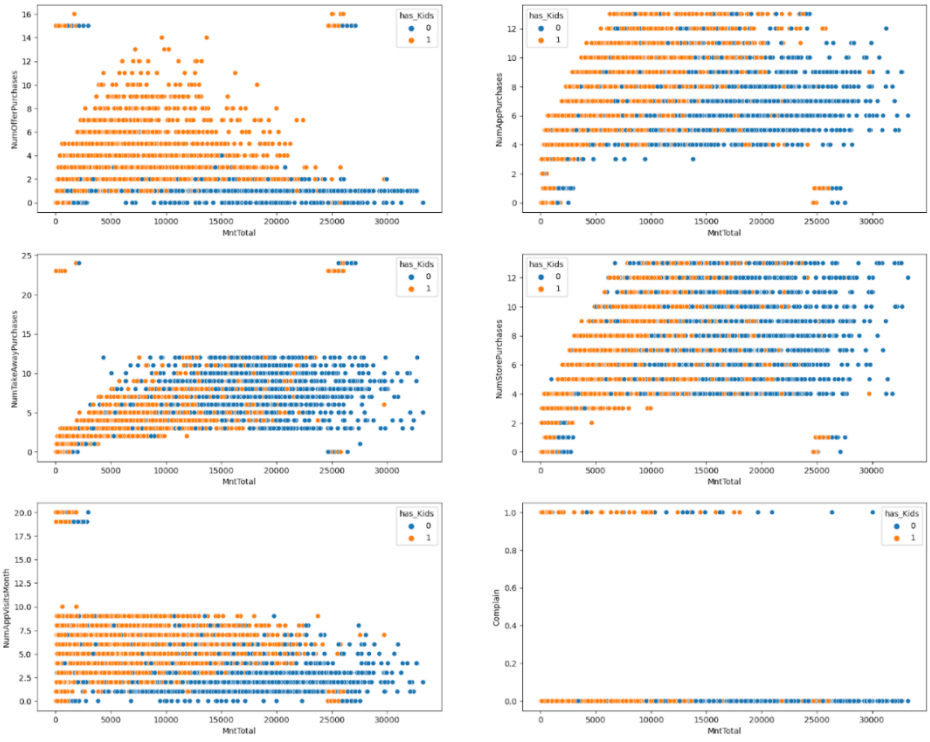

Figure 18- MntTotal vs Purchases, NumAppVisits, Complain vs has_Kids.

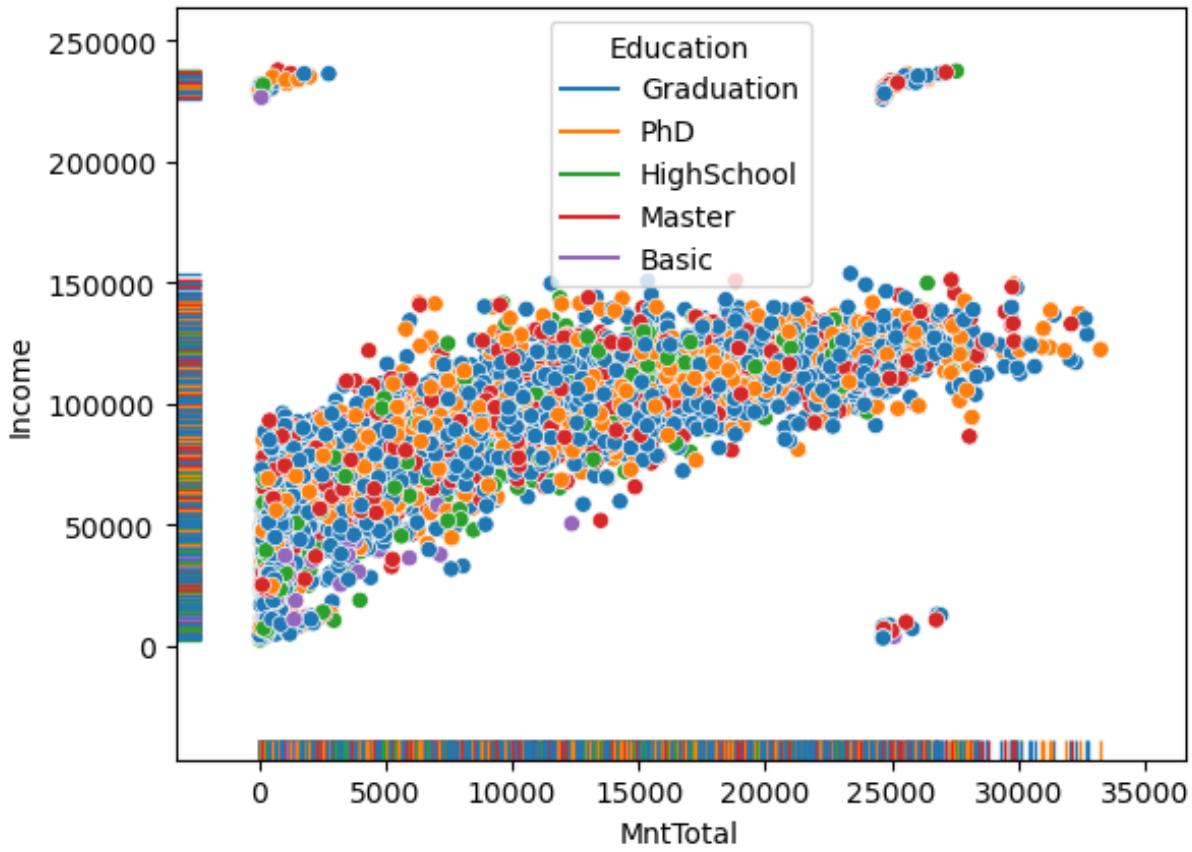

Figure 19-MntTotal vs Income vs Education.

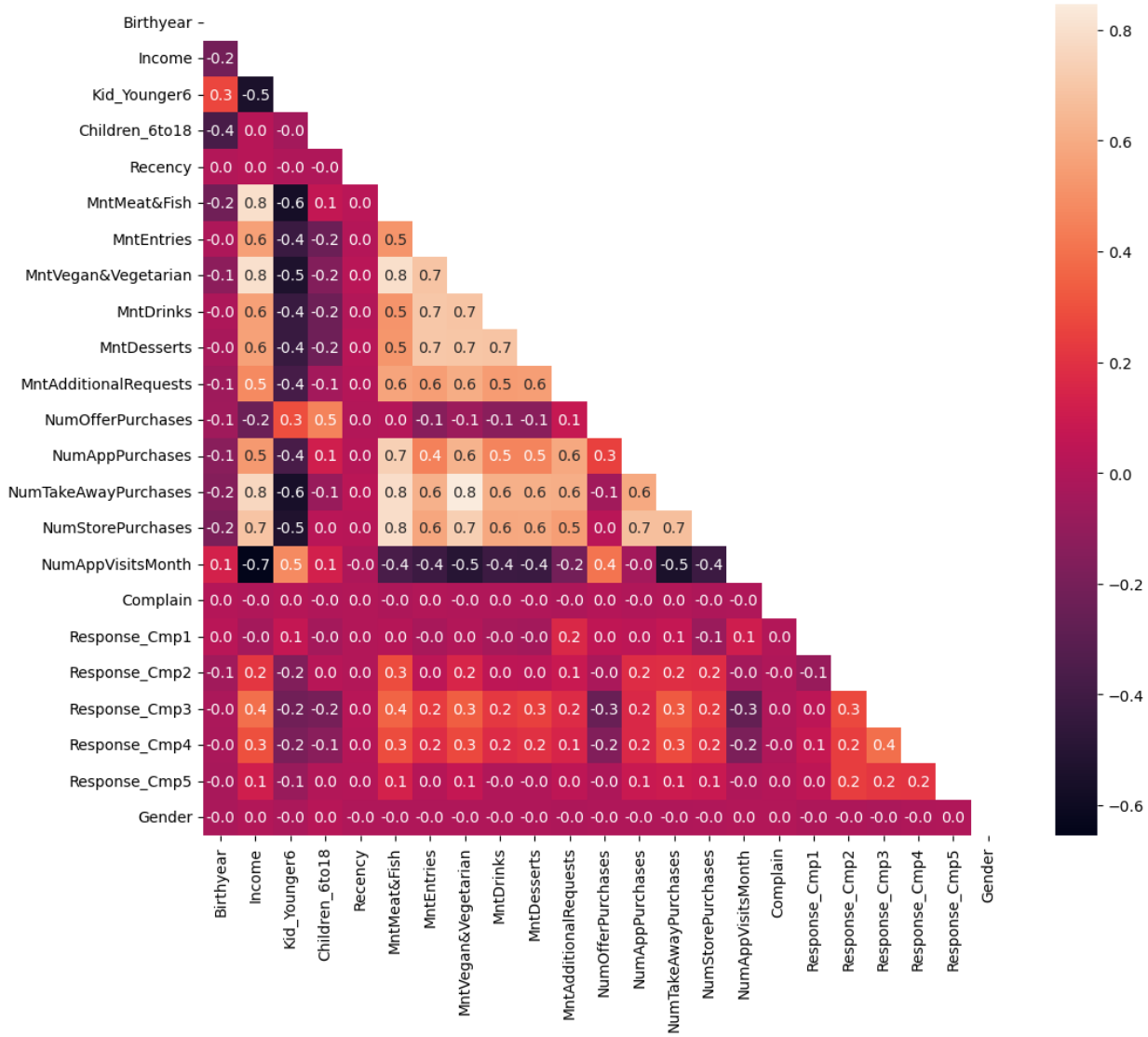

Figure 20- Correlation plot.

Tables

Table 1- Summary table for category values by cluster

	0	1	2	3
Income (mean)	47749.222	81212.408	108400.182	51379.293
Kid_Younger6 (mean)	0.784	0.377	0.053	0.788
Children_6to18 (mean)	0.382	0.803	0.294	0.57
NumPurchasesTotal (mean)	9.557	19.662	21.968	7.723
Age (mean)	42.734	48.75	48.087	45.562
Male (sum)	1108.0	1226.0	1656.0	1007.0
Female (sum)	421.0	420.0	624.0	350.0
daysAsCardClient (mean)	674.652	738.566	606.946	523.943
Graduation (sum)	813.0	796.0	1186.0	624.0
HighSchool (sum)	226.0	118.0	201.0	102.0
Master (sum)	183.0	300.0	383.0	250.0
PhD (sum)	178.0	429.0	510.0	338.0
Married (sum)	1003.0	1067.0	1490.0	890.0
Single (sum)	380.0	342.0	474.0	306.0
Widow (sum)	31.0	70.0	97.0	37.0
MntTotal (mean)	1328.754	8041.721	15164.037	586.074
Response_Campaigns (mean)	0.095	0.342	0.571	0.084
Total_Kids (mean)	1.166	1.18	0.346	1.357
Pct_Meat&Fish (mean)	0.272	0.576	0.371	0.456
Pct_Desserts (mean)	0.129	0.052	0.084	0.068
Pct_Entries (mean)	0.13	0.051	0.085	0.074
Pct_AdditionalRequests (mean)	0.019	0.009	0.005	0.014
Pct_Drinks (mean)	0.135	0.052	0.086	0.069
Pct_Vegan&Vegetarian (mean)	0.316	0.26	0.369	0.318

Pct_Store (mean)	0.353	0.362	0.387	0.38
Pct_App (mean)	0.472	0.443	0.323	0.454
Pct_TakeAway (mean)	0.174	0.195	0.291	0.165
NumOfferPurchases (mean)	0.503	0.681	0.345	0.389

Table 2- Euclidean distance between cluster centers for K-Means.

K-MEANS - cluster_centers_				
	0	1	2	3
0	0.000	5.966	3.266	6.203
1	5.966	0.000	4.257	2.554
2	3.266	4.257	0.000	4.706
3	6.203	2.554	4.706	0.000

Table 3- Euclidean distance between cluster centers for K-Prototypes.

K-PROTOTYPE - cluster_centers_				
	0	1	2	3
0	0.000	8.714	15.769	3.035
1	8.714	0.000	7.676	11.138
2	15.769	7.676	0.000	18.070
3	3.035	11.138	18.070	0.000

Table 4 - Numerical variables summary statistics

	count	mean	std	min	25%	50%	75%	max
Birthyear	7000.0	1976.451429	11.996270	1948.0	1967.00	1977.0	1985.00	2005.000
Income	7000.0	77988.962407	35409.810253	2493.8	51586.25	77190.0	102016.25	237639.725
Kid_Younger6	7000.0	0.440571	0.543477	0.0	0.00	0.0	1.00	2.000
Children_6to18	7000.0	0.490571	0.542174	0.0	0.00	0.0	1.00	2.000

Recency	6977.0	49.235058	28.922688	0.0	24.00	49.0	74.00	99.000
MntMeat&Fish	7000.0	3079.523800	3370.377166	0.0	250.00	1820.0	5070.00	14980.000
MntEntries	7000.0	534.749429	787.846684	0.0	40.00	180.0	680.00	3980.000
MntVegan&Vegetarian	7000.0	2785.050786	3908.718244	0.0	240.00	1110.0	3795.00	25974.000
MntDrinks	6972.0	545.657544	805.149088	0.0	40.00	180.0	700.00	3980.000
MntDesserts	7000.0	540.656029	802.221866	0.0	40.00	180.0	680.00	3980.000
MntAdditionalRequests	7000.0	42.556186	49.650747	0.0	9.00	24.0	57.00	249.000
NumOfferPurchases	7000.0	2.448429	2.306968	0.0	1.00	2.0	3.00	16.000
NumAppPurchases	7000.0	6.015714	2.745537	0.0	4.00	6.0	8.00	13.000
NumTakeAwayPurchases	7000.0	3.834571	3.331142	0.0	1.00	3.0	5.00	24.000
NumStorePurchases	7000.0	5.790571	3.295708	0.0	3.00	5.0	8.00	13.000
NumAppVisitsMonth	7000.0	5.278286	2.748596	0.0	3.00	5.0	7.00	20.000
Complain	7000.0	0.010286	0.100903	0.0	0.00	0.0	0.00	1.000
Response_Cmp1	7000.0	0.079143	0.269981	0.0	0.00	0.0	0.00	1.000
Response_Cmp2	7000.0	0.073286	0.260624	0.0	0.00	0.0	0.00	1.000
Response_Cmp3	7000.0	0.083000	0.275902	0.0	0.00	0.0	0.00	1.000
Response_Cmp4	7000.0	0.065857	0.248050	0.0	0.00	0.0	0.00	1.000
Response_Cmp5	7000.0	0.014286	0.118675	0.0	0.00	0.0	0.00	1.000
Gender	7000.0	0.733286	0.442273	0.0	0.00	1.0	1.00	1.000
MntTotal	6972.0	7527.322433	7650.867424	16.8	811.75	4644.5	12763.25	33256.000
Total_Kids	7000.0	0.931143	0.753599	0.0	0.00	1.0	1.00	4.000

Table 5 - Categorical variables summary statistics

	Name	Education	Marital_Status	Date_Adherence
count	7000	6986	7000	7000
unique	6241	9	10	701
top	Mr. Stewart Grant	Graduation	Married	19/09/2020 00:00
freq	3	3497	2830	2