UNIVERSIDADE NOVA DE LISBOA

NOVA INFORMATION MANAGEMENT SCHOOL

Group Project: Spice Alley

DATA SCIENCE AND MACHINE LEARNING

POST GRADUATION IN ENTERPRISE DATA SCIENCE AND ANALYTICS

PROFESSORS:	**GROUP DSML 202223 21**
Carina Albuquerque	André Oliveira nº. 20222156
Ricardo Santos	Diogo Fernandes nº. 20220507
	Gonçalo Eloy nº. 20222162
	Gonçalo Matos nº. 20221194
	Rafael Chamusca nº. 20222174

2023

Index

1. *Introduction & Methodology*

This project aims to segment the customers of Spice Alley through clustering algorithms. Identifying the main customers segments allows Spice Alley a better understanding of its customers' needs and behaviors. The project follows the CRISP approach, with some flexibility, and essentially tries to answer the following questions: What are our customer segments? Which features allow us to distinguish our customers and their consumption behavior? What patterns and tendencies better allow us to segment our customers?

This project's goal is to answer these questions using a data driven methodology and uncover new opportunities for expanding the business while being better equipped to target existing customer groups with product offerings, channels and marketing campaigns.

2. *Business Understanding*

Spice Alley is a well-known player in the restaurant industry. The restaurant is well recognized for its innovation and ability to heighten customer experience. Spice Alley has a wide variety of product offerings and channels to better cater its value proposition to different customer groups. Spice Alley seeks to create great experiences to share with your family, partner or to enjoy alone.

3. *Understanding & Exploring Data*

Spice Alley provided 3 data sets to our team. One with sociodemographic aspects of customers, one with business indicators and one with of the 5 marketing campaigns used by the restaurant on existing customers. We merged these datasets into one with 7000 rows and 26 variables (dropped 8 duplicated rows). Regarding univariate analysis, we characterized each variable distribution through visualizations, skewness and kurtosis values. Our main findings are presented in **Table 1**. Concerning multi-variate analysis, correlation analysis does not show any perfect correlation between variables (**see Figure 1**). However, the "Income" variable shows significant correlation with most variables and all food preference variables heavily correlate to each other. Correlation analysis on a channel perspective shows that "NumStorePurchases" and "NumTakeAwayPurchases" heavily correlate with the "Drinks" and "Desserts" variables. "Age", "Recency" and "NumAppVistisMonth" don't show significant correlations with the other variables (see **Figure 2**). No relevant natural clusters are visible. Indicating the need to apply Machine Learning algorithms to find patterns and potentially segment customers according to their characteristics and historical data gathered from Spice Valley.

4. Data Preparation

The data provided were prepared according to the 4 perspectives established as our approach, which includes **Error! Reference source not found.**Customer behavior, Customer Value, Channel preference and Demographic Characterization (see **Table 2).**

4.1. Data Cleaning

Data preprocessing plays a crucial role in the accuracy of the final model(s). Based on kurtosis and distribution analysis we identified that the variables NumpAppVisitsMonth, MntVegan&Vegetarian, NumOfferPurchases, NumTakeAwayPurchases and Income may contain outliers or extreme values. Their description can be found in Table 1.

4.2. Missing Values (MV)

The variables "Education", "Recency" and "MntDrinks" have, respectively, 14, 23 and 28 MV. We applied the mode to fill the values in "Education", the median to fill the values in "Recency", and KNN to fill the values in "MntDrinks" (three highly correlated variables were used). The detailed analysis of MVs is in **Table 3.**

4.3. Data Transformation and Feature Engineering

The data transformation and feature engineering are shown in Error! Reference source not found.. Lastly, we have one incongruence, specifically a customer with frequency = 0 but that spent in total more than 20 thousand monetary units. To preserve as much data as we could, on a specific group of 26 customers we set the frequency of previously 24 to 1, since the amount spent could be reasonably spent on only one visit and could just be a random error in purchase registry. The other two customers we dropped since their amount spent surpassed 20 thousand monetary units.

4.4. Data Reduction

We applied PCA in the data set with 31 variables. The first 3 components explain around 50% of the variance and the first 10 approximately 80%. After the third, the components show an accentuated drop (see Figure 3). As a first approach, and acknowledging the limitations of the analysis, we used 3 components to make it easier to visualize and make sense of the results (see Figures 4, 5 6 and 7). It should be noted that 2 components are related with the demographic characteristics and 1 with monetary and frequency attributes, based on the correlation between these variables and the components. Considering the shape, concentration and large distance between data points we decided to use Density-based Spatial Clustering of Applications with Noise (DBSCAN) to cluster our components. The application of the model highlighted some outliers but returned 6 very well-defined clusters (see Figures 8, 9 and 10).

Nonetheless, the clusters obtained through PCA have a weak translation to the original variables, limiting the interpretation and contextualization within the business context. In summary, DBSCAN results only allow us to conclude that there are 6 different groups of customers, but we can't say in each way they differentiate from each other. It is possible that the resulting clusters based on the transformed data do not directly correspond to the original variables. Since our goal is to support the marketing campaigns' targeting and infer an overall understanding of relevant and distinct groups of customers, as well as their consumption behaviors, we will implement a K-means centered approach.

We selected the Education_bins, Marital_Status, Income_bins and Have_kids for this perspective and encoded the data to apply k-modes. We applied Elbow method and silhouette score to help identify optimal number of clusters. The highest silhouette score was 0.235 for k = 4. The elbow method pointed possible optimal number of clusters k=2, k=3 and k=4 (see figure X).

5. Modelling

The Spice Alley customers were segmented according to the 4 perspectives. Our main goal is to provide actionable insights and answer 4 main questions: which are the costumers more valuable for the company; how they behave and which characteristics they possess; through which channel they prefer to purchase; and what are their purchasing habits/behaviors. The techniques used in the modelling were chosen accordingly.

The techniques used in the modelling were chosen accordingly the 4 perspectives chosen as approach.

5.1. Customer Value Perspective

In this perspective we selected the features "Recency", "Freq" and "Mnt_Total" grouped in a subset. A cubic root transformation was applied on the variable "Mnt_Total", to bring it closer to a normal distribution. We then applied Minmax Scaling to transform all variables to the same scale. To estimate the number of clusters we used the Silhouette Score, the Elbow and dendrogram (see Figures 11, 12 and 13). All silhouettes' scores are similar, being the one with four clusters the largest. Also, every cluster score is above the average score. The elbow method and the dendrogram also suggest that four is an appropriate number of clusters.

In this perspective we applied K-means algorithm with k=4, giving us two pairs of clusters that differ in recency: Two clusters with high frequency and high total amount spent but one with high recency and the other with low recency; Two clusters with low frequency, low total amount spent but one with high recency and the other with lower. These findings enable us to target high and low spending, more and less frequent customers that might be churning, allowing the company to allocate more or less resources considering the revenue witch group can bring to the company. On the other hand, it also enables the company to reward its most loyal, frequent and high spending customers. This model was also used with two clusters, which didn't enable us to draw conclusions about recency, only grouping customers into high/low spending and more/less frequent. And with three clusters, which only enable us to discriminate based on recency between high spending, high frequency customers.

To expand our knowledge, and to provide even more information to the marketing team we applied a RFM analysis. This method considers the Frequency, Recency and Monetary. For each variable it's attributed a socre from 1(worst) to 5(best). There are many aproches to divide the customers into each score, for example creating equal count groups or dividing into 20% quantiles. Data with high skweness, as was our case, poses a serius disadvantage to this aproches since it can produce non reliant groups and therefore conclusions (Kabasakal, İ., 2020). To minimize this risk we adopted the same aproach as Anitha and Patil (2019), using kmeans in wich of the three variables to get 5 scores/clusters. We then grouped the combinations in ten groups regarding the frequency and recency scores to achive an easier vizualization (see figure 14).

5.2. Customer Behavior

In this perspective, we selected the variables "MntMeat&Fish%", "MntDrinks%", "MntEntries%" and "MntVegan&Vegetarian%", which represent the proportions of the amount's costumers spent per food type, excluding the "MntAdditionalRequests". "MntAdditionalRequests" was not included since it has a relatively small expression, and therefore amounts the costumers spent on, was rather small. Thus, it was not considered in the subsequent modeling, as it not contributed significantly to it. Similarly, to the above-mentioned perspectives, Silhouette Score and Elbow Method, were applied with the addition of the Hierachical Method, to better help identify optimal number of clusters (see Figures 14, 15 and 16).

After analysing the different perspectives, we decide to use 3 clusters.

In this perspective we used K-means clustering algorithm with a K-means++ initialization. The distributions on entries, drinks and desserts percentages were right skewed, and that was influencing the results of our models, so, in order to avoid that we applied a log1p() function. After applying the function, we ended up with a more normalized distribution on those variables (see Figures 17 and 18).

When we apply 3 cluster the results are similar, but we end up losing the cluster 1 of the 4 clusters appliance. We have a cluster 0 which is characterized by a higher proportion of spending on meat and fish. Cluster 1 has similar distributions between every metric, while Cluster 2 exhibits a higher proportion towards entries, drinks and desserts (see Figure 19).

K =3: We obtained 3 clusters with similar size. We have a group that spends lower percentages on both (Between 0% and 40% of meat and fish and 0% and 50% of vegan and vegetarians), we have a second group that spend high percentages on both (between 0% and 50% on meat and fish and 20% to 100% on vegan and vegetarian. Finally, a third group that spends more meat and fish (more than 40% and between 0% and 55% of vegan and vegetarian approximately). When applied to drinks and entries the numbers are much more spread out, having a small group spends small percentages on both, other group that spends a bit more of them and a then a group with the remaining.

Applying a K-medoids with a K-medoids++ initialization and setting the number of clusters to 3 and 4, we obtained similar results to the previously applied K-means (see Figure 20).

The Gaussian Mixture Model was utilized with 4 clusters, however the results were not as satisfactory when compared to the ones obtained through K-means. It was also possible to identify that the outliers are significant since their existence alters the results of the GMM, which is more sensitive to outliers (see Figure 21).

5.3. Channel Preference

In this perspective we selected the features “NumAppPurchases%”, “NumTakeAwayPurchases%” and “NumStorePurchases%”. Important to note introducing our model, these exist outliers seem to be "loyalists" to a specific channel, as they have almost 100% expenditure in one channel and an extremely low percentage use of other channels. We applied Silhouette Score, Hierarchical and Elbow Method to check the optimal number of k (see results in Figure X). In this perspective we applied K-Means clustering algorithm with a K-means++ initialization. As inputs for our model, we chose “NumAppPurchases%”, “NumTakeAwayPurchases%” and “NumStorePurchases%”. To identify the possible number of clusters we utilized the elbow method (see Figure 21 and 22). We tested our model with 2,3 and 4 clusters and the highest silhouette score was for 2 clusters. Our model with 2 clusters indicates a cluster with a clear distinction between preferences App and Takeaway users and another with balanced channel preferences. The 2-cluster model shows a silhouette score of 0,42, indicating the clusters are reasonably well-separated, but some overlap between them should be expected. Our 3 cluster model showed the following results: Cluster 0 (Customers with a preference for the App channel that still frequent the restaurant but use the takeaway option in low numbers), Cluster 1 (Customers with a preference for the takeaway channel but still use the app and frequent the restaurant), Cluster 2 (Customers that have a clear preference for frequenting the restaurant that still moderately use the app while utilizing the takeaway option in a lesser degree)(see figure 23 and 24). The third3cluster model reports a silhouette score of 0.4, indicating relatively well separated and concise clusters. Focusing on our third cluster model within Spice Alley’s business context, the results allow us to target app users, takeaway users and on-site users with marketing campaigns as well as recognizing the different customer groups as well as their preferred channel.

In this perspective we selected the features “NumAppPurchases%”, “NumTakeAwayPurchases%” and “NumStorePurchases%”. Important to note introducing our model, these exist outliers seem to be "loyalists" to a specific channel, as they have almost 100% expenditure in one channel and an extremely low percentage use of other channels. We applied Silhouette Score, Hierarchical and Elbow Method to check the optimal number of k (see results in Figure X). In this perspective we applied K-Means clustering algorithm with a K-means++ initialization. As inputs for our model, we chose “NumAppPurchases%”, “NumTakeAwayPurchases%” and “NumStorePurchases%”. To identify the possible number of clusters we utilized the elbow method (see Figure 21 and 22). We tested our model with 2,3 and 4 clusters and the highest silhouette score was for 2 clusters. Our model with 2 clusters indicates a cluster with a clear distinction between preferences App and Takeaway users and another with balanced channel preferences. The 2-cluster model shows a silhouette score of 0,42, indicating the clusters are reasonably well-separated,

but some overlap between them should be expected. Our 3 cluster model showed the following results: Cluster 0 (Customers with a preference for the App channel that still frequent the restaurant but use the takeaway option in low numbers), Cluster 1 (Customers with a preference for the takeaway channel but still use the app and frequent the restaurant), Cluster 2 (Customers that have a clear preference for frequenting the restaurant that still moderately use the app while utilizing the takeaway option in a lesser degree)(see figure 23 and 24). The 3 cluster model reports a silhouette score of 0.4, indicating relatively well separated and concise clusters. Focusing on our 3 cluster model within Spice Alley's business context, the results allow us to target app users, takeaway users and on-site users with marketing campaigns as well as recognizing the different customer groups as well as their preferred channel.In relation to the first part of the second question *Which features allow us to distinguish our customers?* Through this perspective, all three input variables were relevant for our analysis

5.4. Demographic Characterization

After testing we selected the variables "Education_bins", "Marital_Status", "Income_bins" and "Have_kids" for this perspective in view of optimizing our model's performance. We applied the Elbow method and silhouette score to help identify optimal number of clusters. The elbow method pointed to possible optimal numbers of clusters at 2,3 and 4 (see figure 2). We chose 3 clusters because from a holistic view of the project it presents the better results to improve the marketing campaigns. We then applied a K-Modes clustering algorithm (9.2. Annex II) and obtained the following results: Cluster 0 (Customers with high education, together, low income, with kids), Cluster 1 (Customers with low education, single, high income, without kids) and Cluster 2 (Customers with low education, together, high and medium income, with kids).

6. Marketing Plan

Considering our clustering analysis, we have identified potential recommendations that can be applied to drive Spice Alley's business context in retaining and growing their customer base. For this purpose, we divide customer groups into 3 main categories: Lurker customers, High value Clients with a churning risk, and Loyal Clients. In this context we will utilize Customer Value clusters to inform on the importance Spice Alley's future marketing decisions.

Lurkers Characterized by Low frequency, low total expenditure and high recency, this group of customers presents an opportunity for Spice Alley's growth. With a targeted marketing approach, these customers could be upgraded to High value clients at churning risk and given enough time to Loyal customers. This group represents a considerable part of Spice Alley's customer population standing at 1678 in total, the ability for Spice Alley to cater to these customers' needs and preferences should be a priority in their growth efforts.

High value clients at churning risk: Characterized by High frequency, High total expenditure and High recency. We see that a significant number of customers is in this group. With this we can identify a potential marketing campaign to retain these customers. This group of high value customers represents a total amount of 3394 individuals and Spice Alley should focus their marketing efforts on this group while catering to their preferences. The marketing campaigns should be aligned to these individuals by appealing to their consumer behaviour and preferred channel in ways to decrease their recency.

Loyal Customers Characterized by high frequency, high total expenditure and low recency as well as dividing their expenditure across most of Spice Alley's product offerings, this is the most valuable group of customers in a business context. Future marketing efforts should always focus on maintaining these customers' low recency, high expenditure and frequency by providing customer loyalty programs and bundling pricing strategies. Monitoring this group's churn into High value clients at churning risk should be a priority, representing a possible marketing KPI.

7. Conclusions

In summary, our team took a data-driven approach to segmenting Spice Alley's customer population on 4 different perspectives: Customer Value, Customer Behavior, Customer Channels and Customer Demographics. By adopting this approach, we were able to supply Spice Alley with insights into how their customers differ in demographics, the value they present to the business (frequency, total amount spent, recency), customer behavior (identifying customers' consumption patterns) and preferred channels (app, restaurant and takeaway). Based on each of these groups of features, our team was able to create segments of customers that will inform Spice Alleys' marketing and strategic decision making.

Utilizing our previous Customer Value segments obtained through applying k-means algorithm to provided datasets, we divided Spice Alley's customer population into three groups: Lurker customers, High value Clients with a churning risk, and Loyal Clients. By dividing customers on their value to the business, we were able to create recommendations for Spice Alley's targeted marketing efforts to promote growth of new customers and retention of valuable customers.

8. Annex I – Tables & Figures

8.1. Tables

Table 1. - Data Types, description of measure of shapes and extreme values

Data Types	Variables	Distribution Observations
Categorical Nominal	Name, Data_Adherence, Marital_Status	
Categorical Ordinal	Education	
Categorical Binary	Kid_Younger6, Children_to18	
Numerical	Income	Variable shows a skewness positive value of 0.845, indicating a tail to the right. A kurtosis of 2.649. 86 costumers have extreme values (>200k). Standard deviation of 35409.810 and mean of 77988.962.
Numerical	Recency	Variable shows a skewness of 0.018. A kurtosis of – 1.198. High standard deviation (28.92) indicating we are likely to find outliers or a big spread in this variable. Std of 28.923 and mean of 49.235
Numerical	MntMeat&Fish	Variable shows a skewness value of 1.149 surpassing the threshold of 1, indicating a characteristic of a normal distribution with a positive value, indicating a tail to the right. Kurtosis of 0.532. Does not show outliers or extreme values. Std of 3370.377 and mean of 3079.524
Numerical	MntEntries	Variable shows a skewness value of 2.087 surpassing the threshold of 1 and with a positive value, indicating a heavy tail to the right and a kurtosis value of 4.096 suggesting a moderate peak and tails. Don’t seem to have outliers and extreme values. A std of 787.846 and a mean of 534.749

Numerical	MntVegan&Vegetarian	Variable shows a skewness value of 2.487 surpassing the threshold of 1 and with a positive value, meaning tail to the right. A kurtosis value of 8.432 suggests that are extremely peaked and an extremely heavy tail. Some extreme values (>200k). A std of 3908.718 and a mean of 2785.051. it seems that Spice Alley customer population has a defined group of vegans, vegetarian and pescatarian customers, since certain customers almost only purchase vegan and vegetarian dishes coupled with a low expenditure on fish and meat dishes (280). This group of customers seems to prefer utilizing the takeaway service.
Numerical	MntDrinks	Variable shows a skewness value of 2.05 surpassing the threshold of 1 and with a positive value, indicating a heavy tail to the right. A kurtosis value of 3.84 suggests a moderate peak and tails. The variable does not seem show outliers or extreme values, but its distribution is skewed to the right. A std of 805.149 and a mean of 545.658
Numerical	MntDesserts	Variable shows a skewness value of 2.058 surpassing the threshold of 1 and showing a positive value, indicating a tail to the right. Kurtosis of 3.813 suggests a moderate peak and tails. A std of 802.222 and a mean of 540.656 Don’t seem to have extreme values.
Numerical	MntAdditionalRequests	Variable shows a skewness value of 1.83 surpassing the threshold of 1 with positive value, indicating a relatively heavy tail to the right. A kurtosis of 3.084. A std of 49.651 and a mean of 42.556.
Numerical	NumStorePurchases	Variable show a skewness of 0.063. A kurtosis of –0.694. A std 3.296 and a mean of 5.790. Some lower counts from 0 to 2.
Numerical	NumAppVisitsMonth	A Kurtosis value of 4.99 suggests a moderate peak and tails. A skewness of 1.01. Some lower counts from 0 to 2. A std of 2.749 and a mean of 5.278. Some extreme values (>15) which means possible outliers. On average, when comparing customers with less than 19 app visits with customers with at least 19 visits (64), the latter has a low number of purchases (app = 1, takeaway = 0, instore = 0) and high number of “NumOfferPurchases” (8). This indica

		tes that there exists a group of people who only buys at discount and check the app regularly
Numerical	Complain[1]	Very little complaints
Numerical Discrete	Birthyear	Variable show a high std (12.00) may imply a high spread and variability in the data.
Numerical Discrete	NumOfferPurchases	Variable shows a skewness value of 2.86 indicating a heavy tail to the right of the distribution. A Kurtosis value of 11.00 suggests the distribution to have an extreme tail and peaks— might have some extreme values since max is 16 and mean is around 2.
Numerical Discrete	NumTakeAwayPurchases	Variable shows a skewness value of 2.25 indicating a heavy tail to the right. A Kurtosis value of 8.59 suggests an extreme tail and peaks — max (24) really high compared to the mean (4) which means possible outliers. Some extreme values (>20), maybe customers that exclusively take the food away – requires further exploring
Numerical Discrete	Responses (1,2 3, 4, 5)	Low number of customers accepting campaign offers

[1] Complain, Response_Cmp1, Response_Cmp2, Response_Cmp3, Response_Cmp4, Response_Cmp5 are coded as numeric, but they are categorical variables and will be explored in their respective section.

Table 2. Perspectives

Perspective	Goal	Variables Used
Customer behavior/Food Preference	Divide customers taking into consideration their product usage – which type of dishes they order more often	MntMeat&Fish MntEntries, MntVegan&Vegetarian, MntDrinks MntDesserts, MntAdditionalRequests
Customer Value	Divide customers taking into consideration their value for the company (how much and often they spend, how much each time)	Recency, Mnt_Total, Freq
Channel Preference	Divide customers taking into consideration their channel usage – which method they use more often (In restaurant, home delivery or Takeaway).	NumAppPurchases, NumTakeAwayPurchases, NumInStorePurchases
Demographic Characterization	Divide customers based on their characteristics and see if any significant groups emerge	Education_bins, Marital_Status, Income_bins, Have_kids

Table 3. Missing Values

Variable	Applied Processes
Education	14 MV seems to be missing completely at random (MCAR) since these seem to be independent from the rest of “Education's” values as well as from other variables. We decided to full these values with the mode because of the low number of MV and the fact that the variable is categorical
Recency	23 MV seems to be missing completely at random (MCAR) since the variable is a time delta between the date of calculation and date of last purchase. Since “Recency” is not correlated with other variables (KNN out of picture) we decided to fill these values with the median (mean rounded would be the same value)
MntDrinks	28 MV likely MCAR since nor other values nor other variables predict whether a value will be missing. We decided to utilize a KNN imputer with “MntDrinks”, “MntEntries”, “MntVegan&Vegetarian” and “MntDesserts” since these have shown to be the most correlated variables (0.7). (9.3. Theory Annex)

Table 4. Data Transformation and Feature Engineering

Variable	Applied transformations
Gender	From the “name” variable we inferred the gender of customers by linking the title ‘Mr.’ and ‘Mrs’ to, respectively, Male and Female.
Age	To create this variable, we subtracted the current date from each birthday of Spice Alley customers.
Antiquity	Initially, Date_Adherence had 16 strings (‘2/29/2022’ converted to ‘3/01/2022’) and was converted to date time to create Antiquity (how many years the customer has the company card).
Length	Utilizing both “Antiquity” and “Recency”, we created the “Length**”** variable (representing the length of time between the customer’s first purchase and most recent one).
Frequency	The total number of multichannel purchases — this variable was created by summing all variables relative to purchase data.
Total Spent (Mnt_total)	Through the sum of Mnt- variables Total Spent (Mnt_total) was created.
Avg_Ticket	Created with Freq and Mnt_total — number of monetary units spent in each purchase.
Education and Marital Status	Both variables were categorized in 2 bins, respectively, Low & High, and Together & Single.
Income_bins	The variable “Income” was categorized into 3 bins (Low, Medium & High). The High bin has a large range, but we decided to keep the income outliers since from a business perspective we are not interested in discriminating between people who earn around 93k and 237k, and it didn’t affect the model.
Have_kids	This variable was created to check how many children each household has from both initial variables relating to children.
Food preference proportions	Food preference variables were transformed into a percentage of the whole (MntMeat&Fish, MntDrinks, MntEntries, MntVegan&Vegetarian, MntDesserts) to ascertain how much each customer spend in each option.
Frequency Channel variables	The proportion of purchases made through specific channel in relation to total purchases made only through the NumAppPurchases, NumTakeAwayPurchases, NumStorePurchases (NumOfferPurchases not included).

Table 5. Data Cleaning

NumAppVisitsMonth	On average, when comparing customers with less than 19 app visits with customers with at least 19 visits (64), the latter have a low number of purchases (app = 1, takeaway = 0, instore = 0) and high number of “NumOfferPurchases” (8). This indicates that there exists a group of people who only buy at discount and check the app regularly
MntVegan&Vegetarian	it seems that Spice Alley customer population has a defined group of vegans, vegetarian and possible pescatarian customers, since certain customers almost only purchase vegan and vegetarian dishes coupled with a low expenditure on fish and meat dishes (280). This group of customers seem to prefer utilizing the takeaway service.
NumOfferPurchases	Of 5 customers, 4 are the same customers present in the previous group, there's only one extra that only purchases when there’s an offer
NumTakeAwayPurchases	Of 61 customers, 53 are the same customers present in the Vegan&Vegetarian. These 61 customers almost use takeaway services exclusively
Income	86 customers who have an income above 200k

8.2. Figures

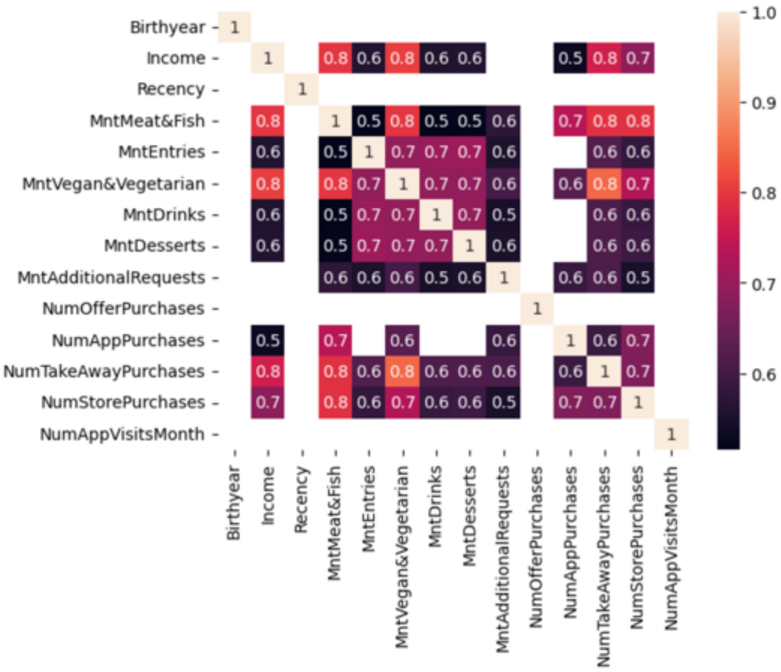


Figure 1. Spearman Correlation Heatmap

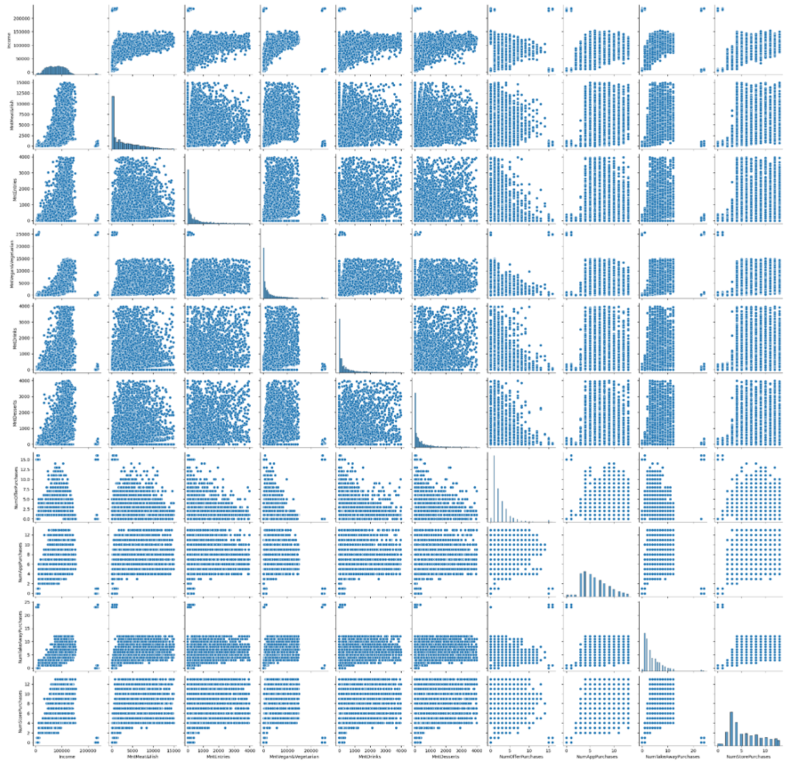


Figure 2. Scatterplot Matrix of Consumer Purchasing Behaviors and Income.

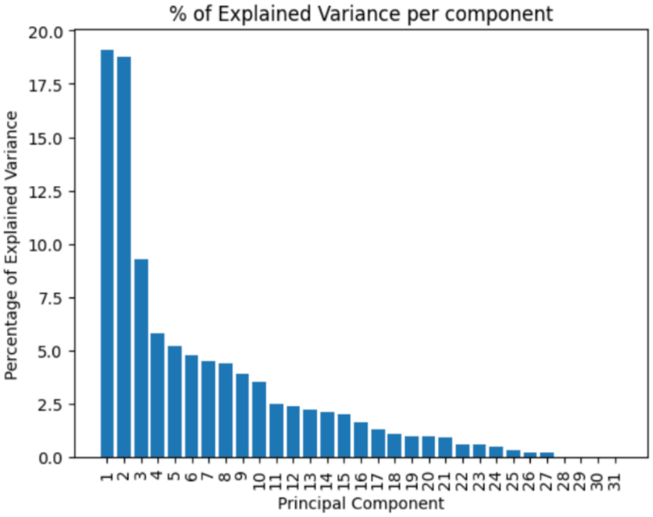


Figure 3. Percentage of Explained Variance per component.

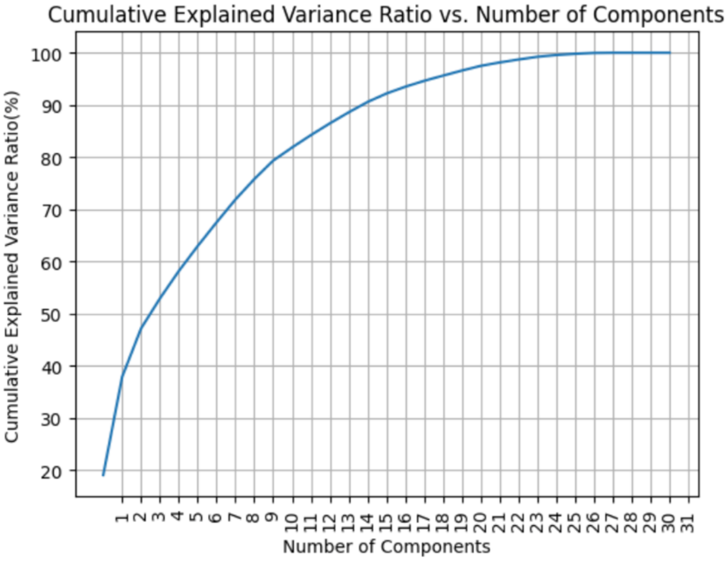


Figure 4. Cumulative Explained Variance Ratio vs Number of Components.

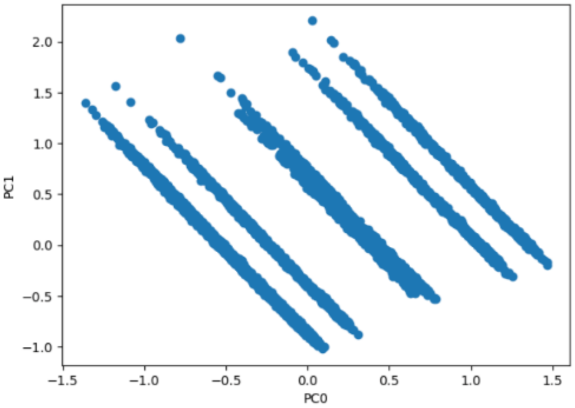


Figure 5. PCA Results Visualizations

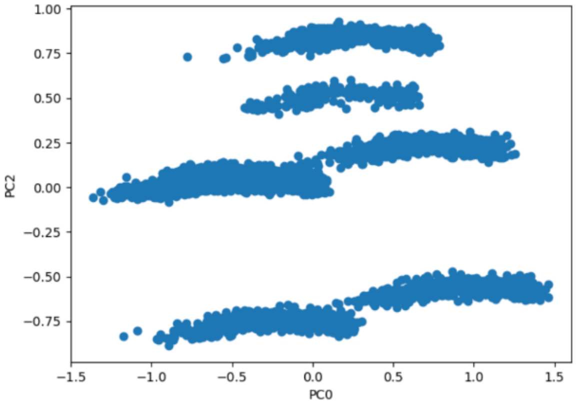


Figure 6. PCA Results Visualizations

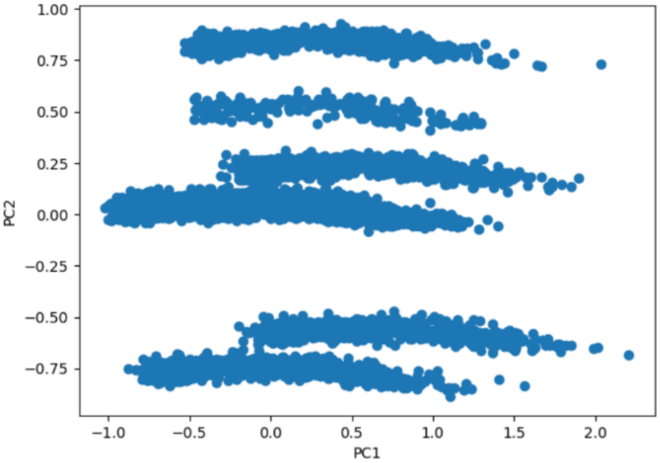
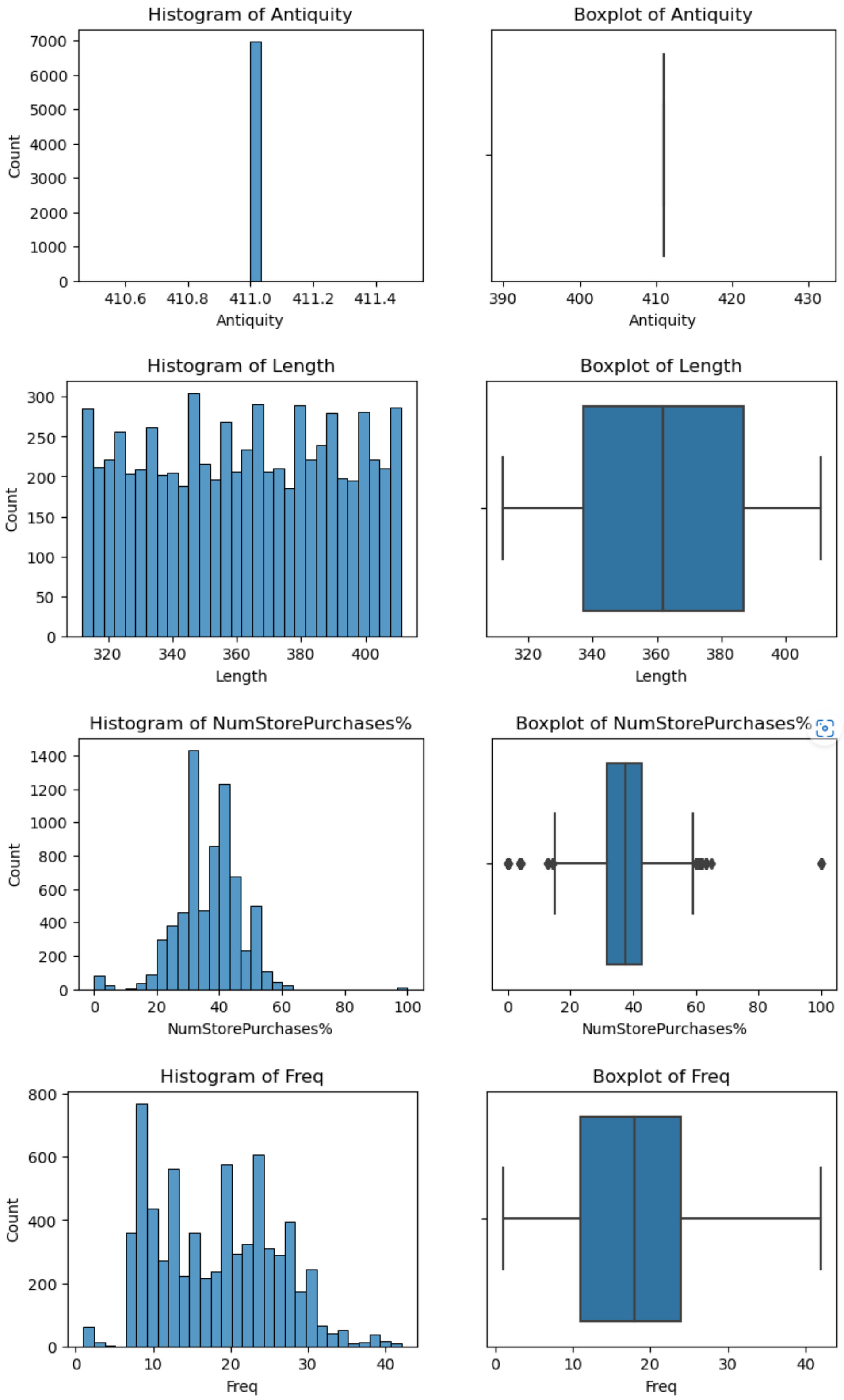
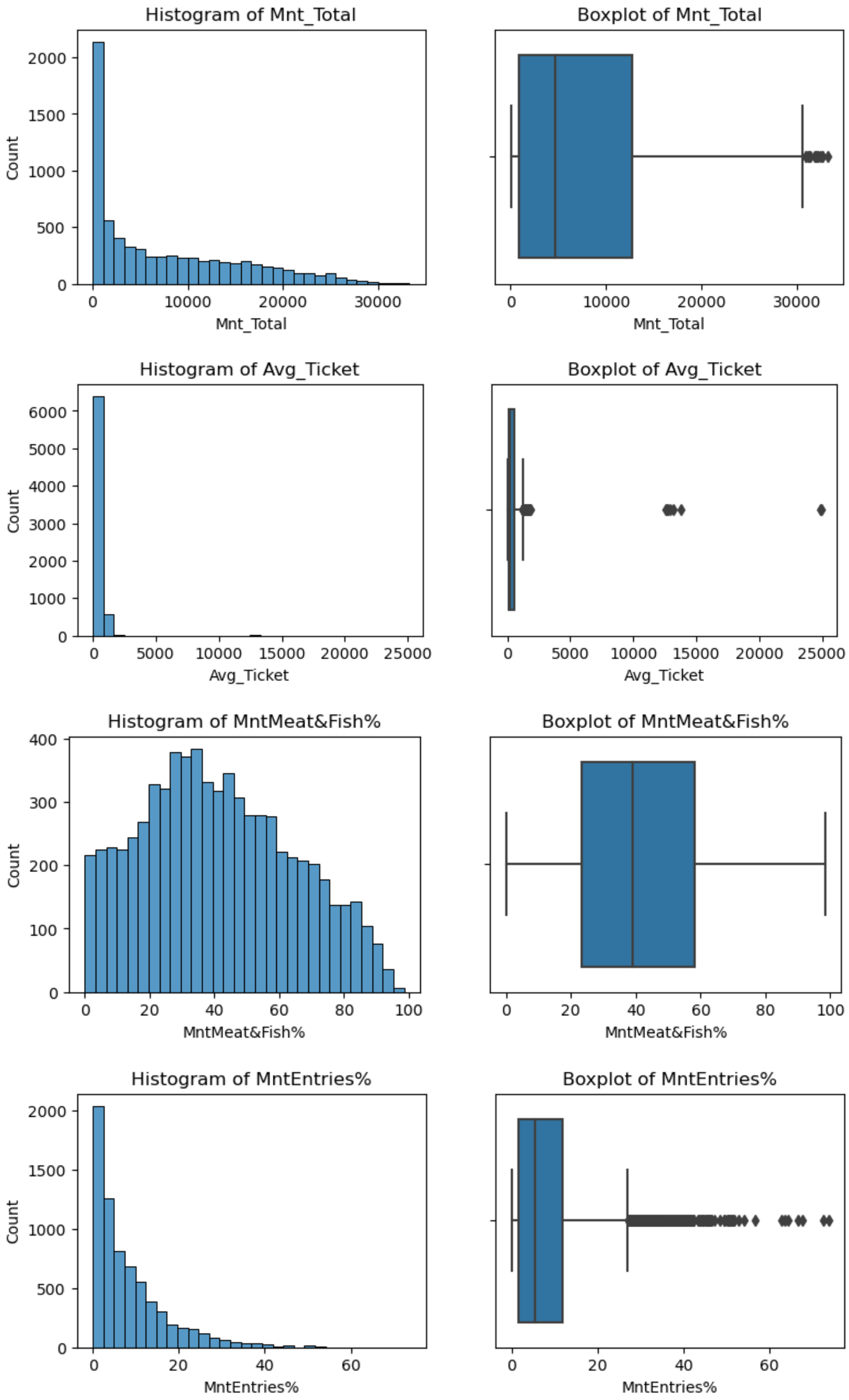
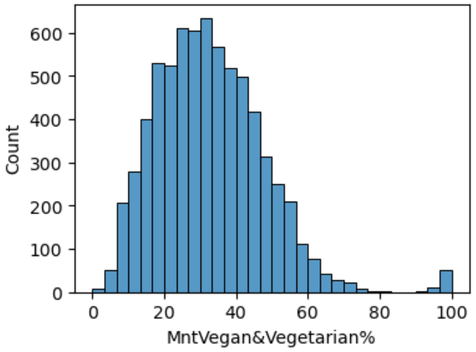
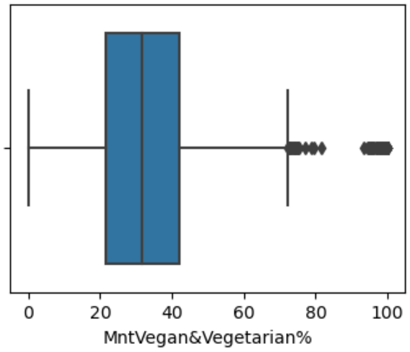


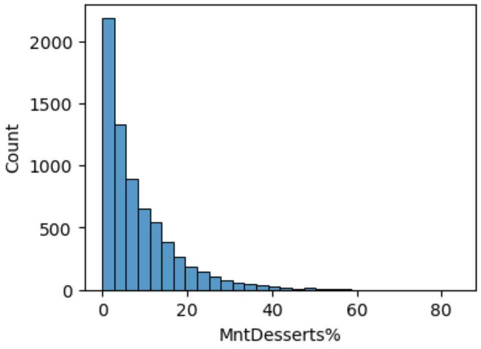
Figure 7. PCA Results Visualization

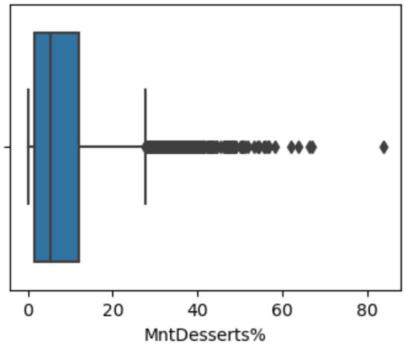

Histogram of Antiquity
Count
7000
6000
5000
4000
3000
2000
1000
0
410.6
410.8
411.0
411.2
411.4
Antiquity
Boxplot of Antiquity
390
400
410
420
430
Antiquity
Histogram of Length
Count
300
250
200
150
100
50
0
320
340
360
380
400
Length
Boxplot of Length
320
340
360
380
400
Length
Histogram of NumStorePurchases%
Count
1400
1200
1000
800
600
400
200
0
0
20
40
60
80
100
NumStorePurchases%
Boxplot of NumStorePurchases%
0
20
40
60
80
100
NumStorePurchases%
Histogram of Freq
Count
800
600
400
200
0
0
10
20
30
40
Freq
Boxplot of Freq
0
10
20
30
40
Freq

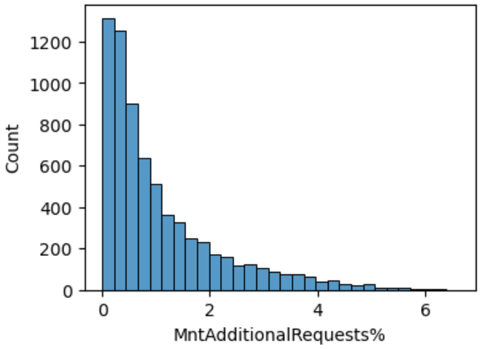

Histogram of Mnt_Total
Count
2000
1500
1000
500
0
0
10000
20000
30000
Mnt_Total
Boxplot of Mnt_Total
0
10000
20000
30000
Mnt_Total
Histogram of Avg_Ticket
Count
6000
5000
4000
3000
2000
1000
0
0
5000
10000
15000
20000
25000
Avg_Ticket
Boxplot of Avg_Ticket
0
5000
10000
15000
20000
25000
Avg_Ticket
Histogram of MntMeat&Fish%
Count
400
300
200
100
0
0
20
40
60
80
100
MntMeat&Fish%
Boxplot of MntMeat&Fish%
0
20
40
60
80
100
MntMeat&Fish%
Histogram of MntEntries%
Count
2000
1500
1000
500
0
0
20
40
60
MntEntries%
Boxplot of MntEntries%
0
20
40
60
MntEntries%

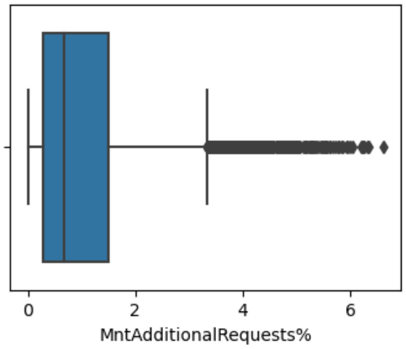

Histogram of MntVegan&Vegetarian%

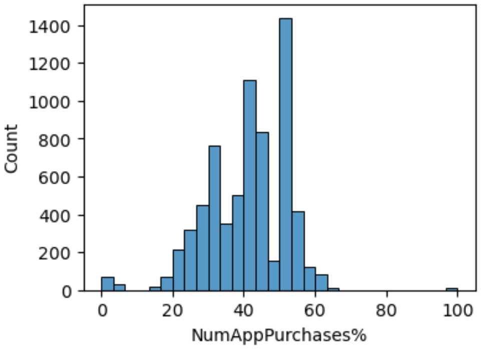

Boxplot of MntVegan&Vegetarian%

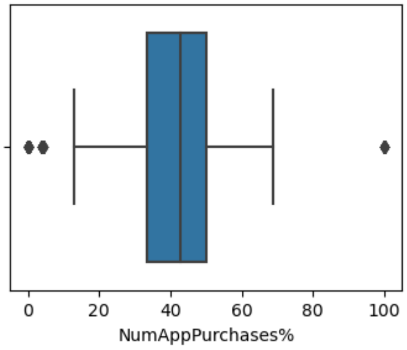

Histogram of MntDesserts%

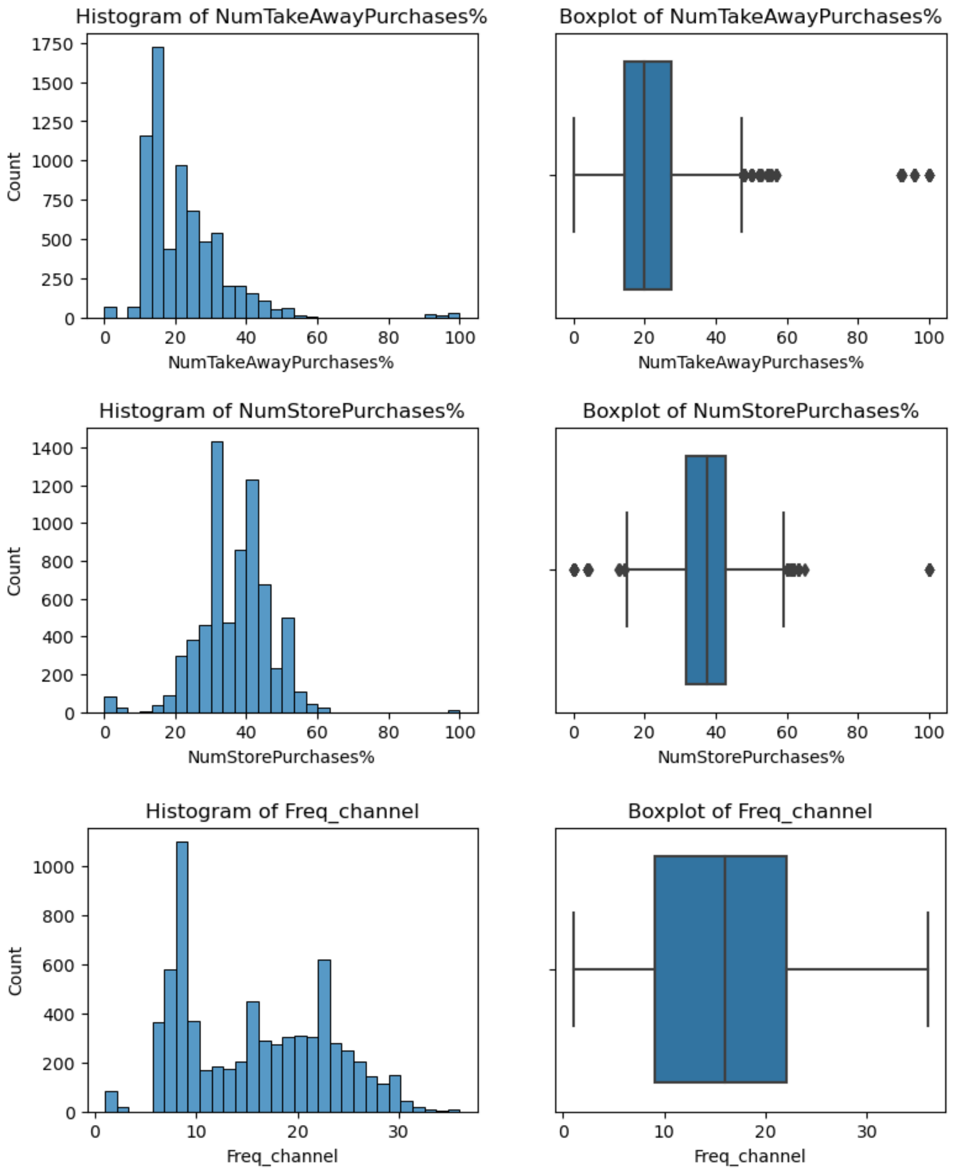

Boxplot of MntDesserts%


Histogram of MntAdditionalRequests%


Boxplot of MntAdditionalRequests%


Histogram of NumAppPurchases%


Boxplot of NumAppPurchases%



Figures 7-22. Histogram and boxplots of new and transformed variables.

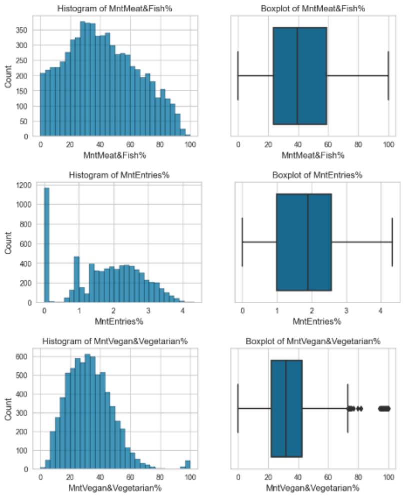


Figure 23. Distributions after log transformation

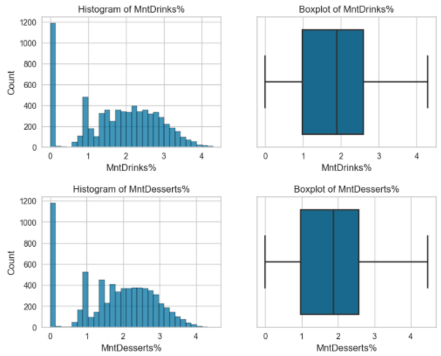


Figure 24. Distributions after log transformation

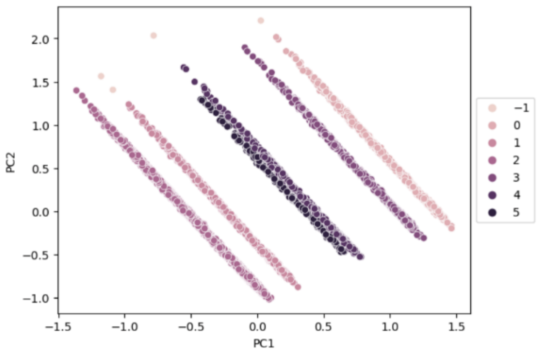


Figure 25. DBSCAN Results

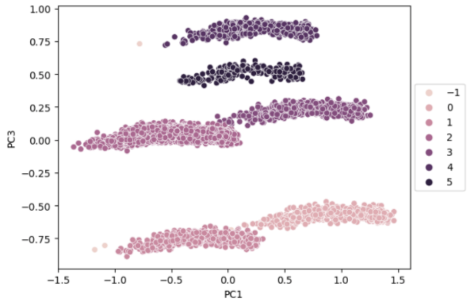


Figure 26. DBSCAN Results

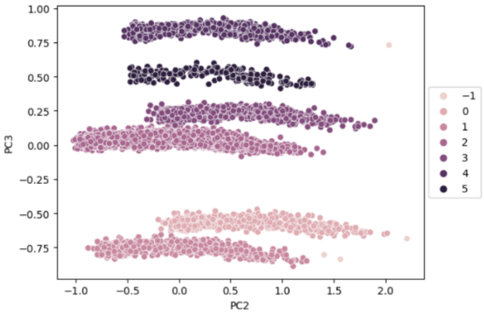
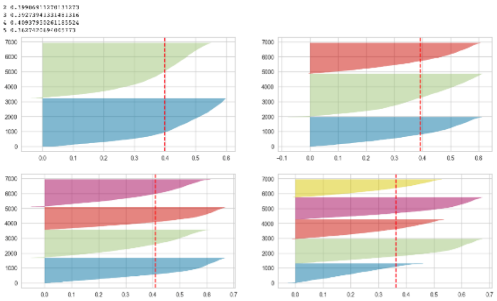


Figure 27. DBSCAN Results



```
2 0.19906911270131273
3 0.39273941331481316
4 0.40937930261185524
5 0.16274206940065773
```

Figure 28. Silhouette Score Customer Value Perspective

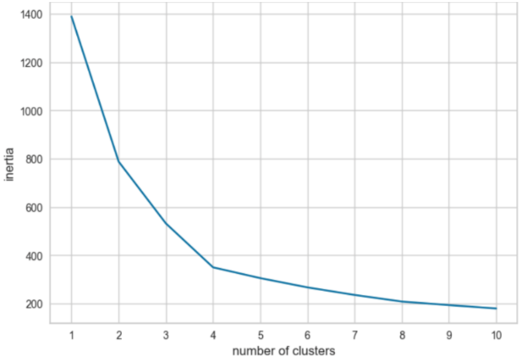


Figure 29. Elbow Customer Value Perspective

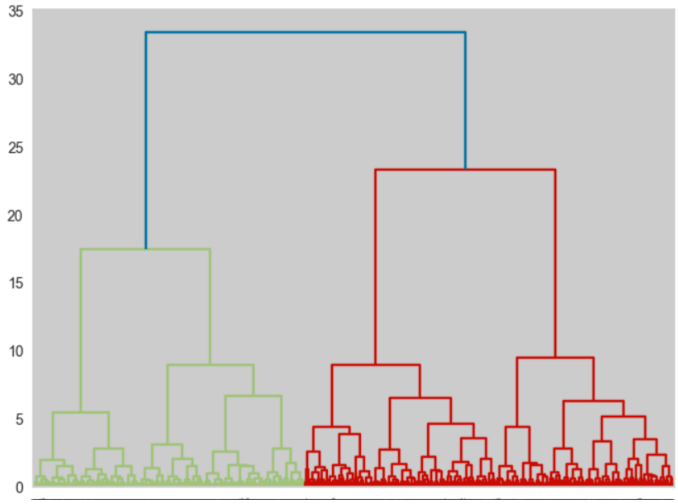


Figure 30. Dendrogram Customer Value Perspective

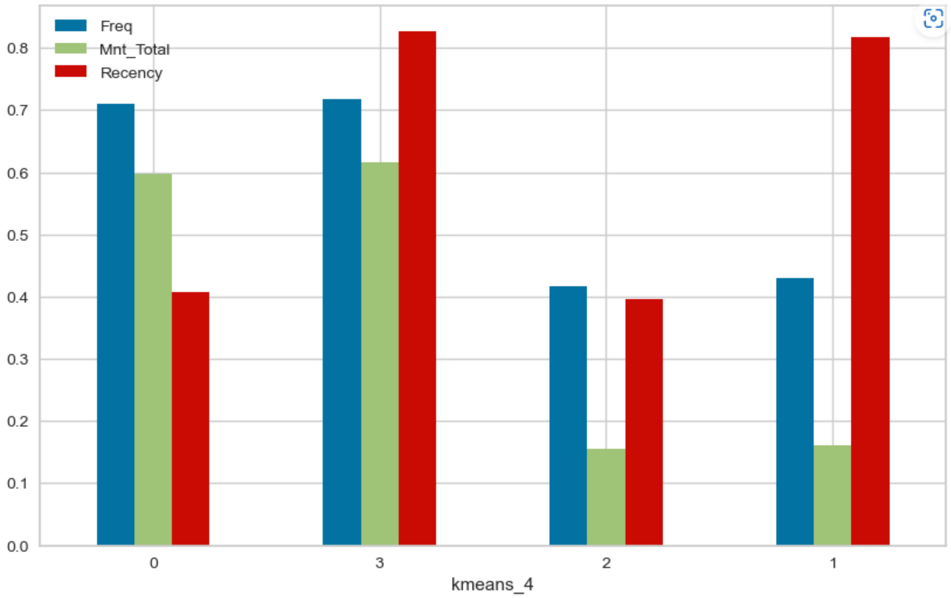
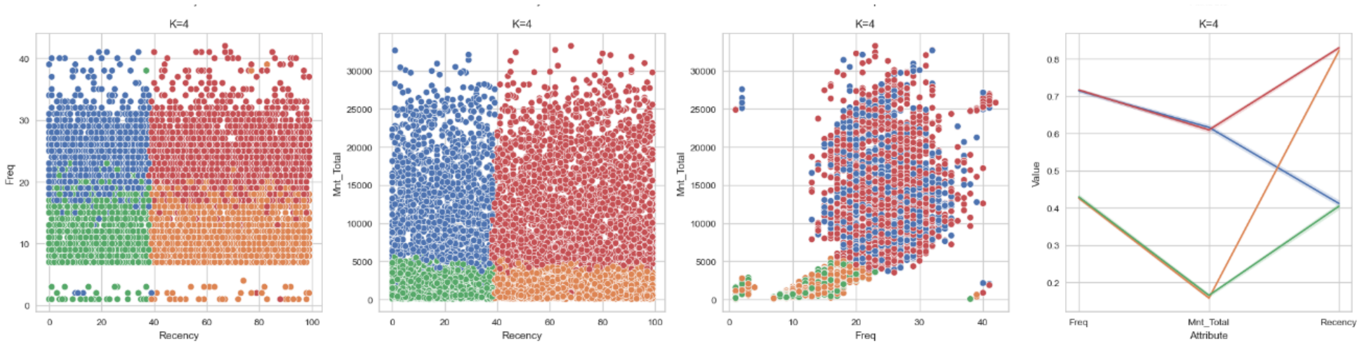
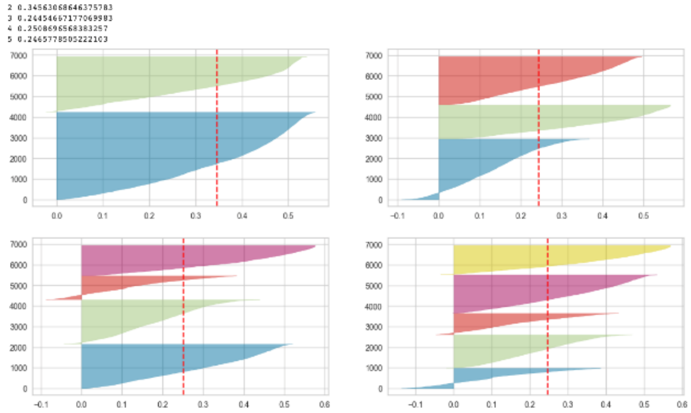


Figure 31. Customer Value Perspective K-means clusters for k=4





```
2 0.34563068646375783
3 0.24454667177069983
4 0.25086965683832 57
5 0.2465778505222103
```

Figure 32. Silhouette Score Customer Behavior Perspective

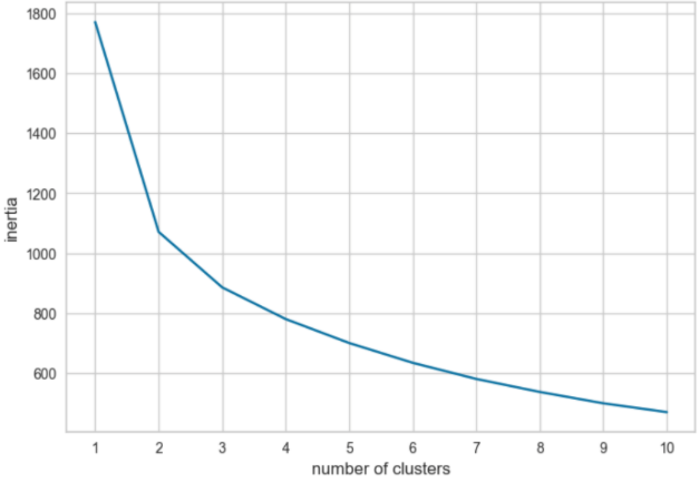


Figure 33. Elbow method Customer Behavior Perspective

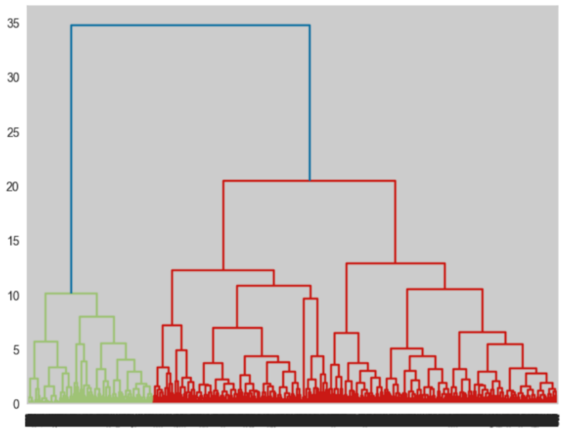


Figure 34. Dendrogram Customer Behavior Perspective

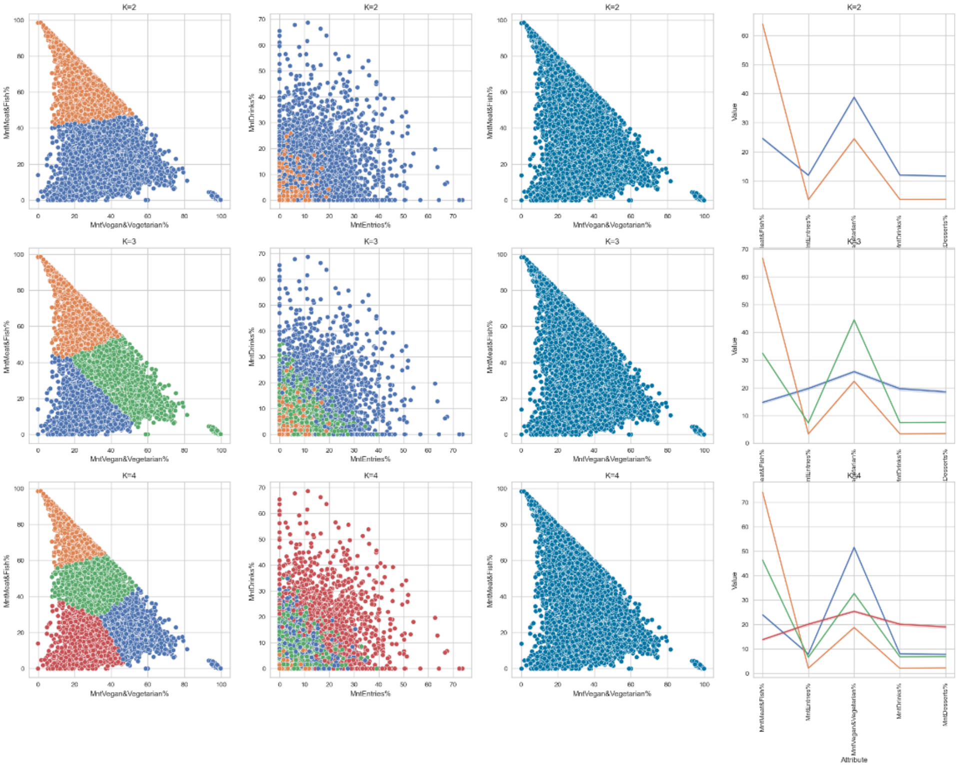


Figure 35. Customer Behaviour K-means for k = (2,3,4)

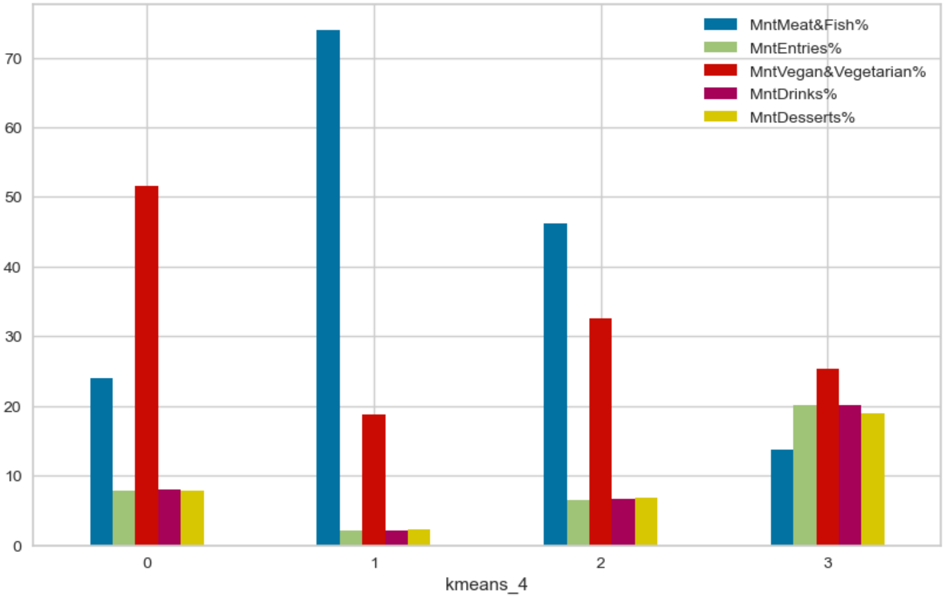


Figure 36. Customer Behaviour K-medoids for k = 4 and k = 3

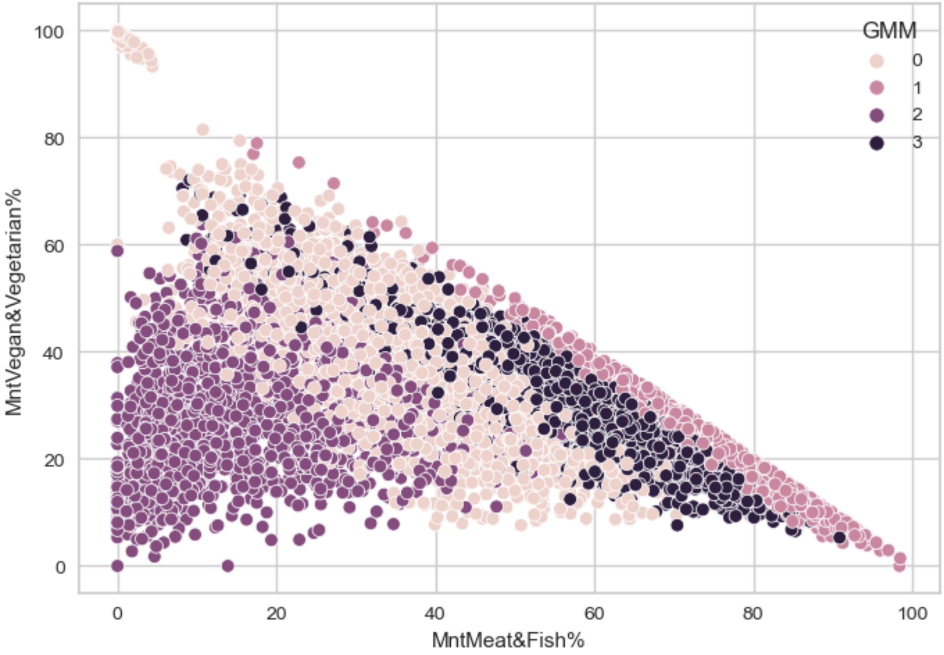


Figure 37. GMM for Customer Behaviour

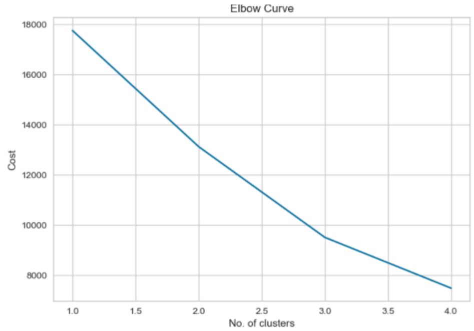
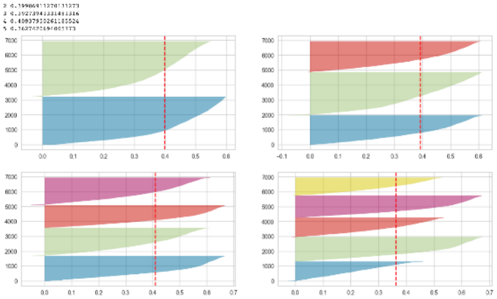
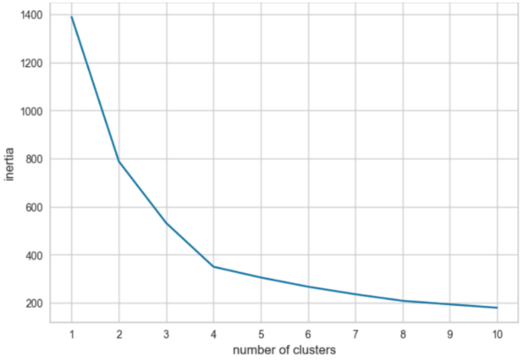


Figure 38. Elbow method Demographic Perspective



```
2 0.39906911270111273
3 0.39273941331481316
4 0.40937930261185524
5 0.3627420694065773
```

Figure 39. Silhouette Score Customer Channel Perspective



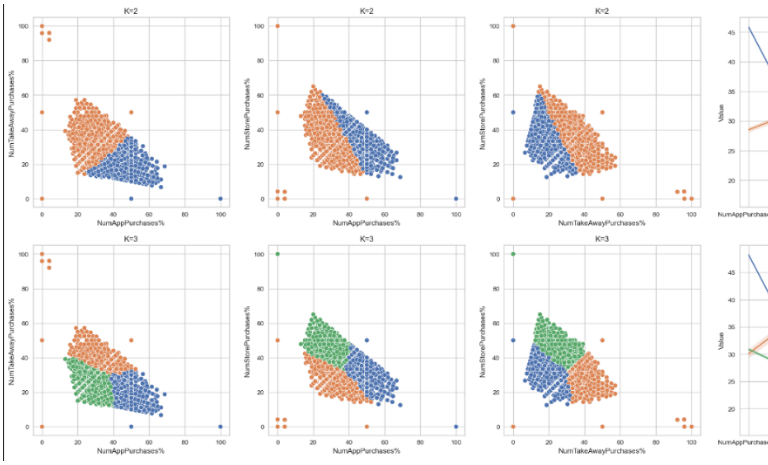


Figure 40. Elbow Customer Channel Perspective

Figure 41. Customer Channel K-means clusters for k=2 and k=3

9. Annex II – Theory

9.1. PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms the original variables into a new set of uncorrelated variables called principal components. The principal components are linear combinations of the original variables, so their interpretation can be different from the original variables. The first principal component is the direction that explains the largest amount of variance in the data, and each subsequent component explains the maximum amount of remaining variance in orthogonal directions to the previous ones. PCA is widely used when we have too many variables in a dataset. It can help us to reduce the number of variables in a dataset while retaining the maximum amount of information. This can be helpful in reducing the computational complexity of data analysis and improving the performance of machine learning models. Despite his usefulness in terms of dimensional reduction, we have to be aware that this model is sensitive to outliers in the data, which can affect the performance of the analysis and the accuracy of the results and also that it can be difficult to interpret the principal components, especially if they are combinations of several original variables.

9.2. K-Modes

Description: Choose the number of clusters (k) you want to create, Randomly select k instances from the dataset to serve as the initial cluster centroids, For each data point, calculate the distance to each cluster centroid using a dissimilarity measure (such as the Hamming distance for categorical data), Assign each data point to the closest cluster based on the dissimilarity measure,

- For each cluster, calculate the mode for each categorical feature in the data points assigned to that cluster, Repeat steps 3-6 until the clusters stop changing or a maximum number of iterations is reached.

The process of K-Modes clustering algorithm is very similar to the K-Means clustering algorithm. One of the differences is the distance parameter. K-Modes clustering is a variant of K-Means used for categorical data. Instead of the mean, K-Modes use the mode to determine the centroids of the clusters. It iteratively assigns observations to clusters based on the mode of their categorical features and updates the centroids accordingly. The k-modes algorithm needs to be run several times with different initializations in order to find a good solution (Cao, Liang, Bai, 2009). The clustering algorithm require numerical data to run so we encoded our subset. After several tests with different numbers for k and initializations, we selected Cao initialization because of the method to initiate the centroids. In opposition to Huang, that starts by randomly selecting an object from the subset as first centroid and select the next one based on distance, Cao is based on a hierarchical partitioning algorithm, and considers distance and density. Also, the results were better with Cao initialization.

9.3. K-Nearest Neighbors Algorithm

K-nearest neighbors (KNN) is a popular machine learning technique used for classification of unlabeled observations based on feature similarity measurement. It finds the K closest training examples in the dataset using a distance function, such as the Euclidean distance which is the distance knn employs by default. After calculating the distance, the new data point is assigned to the category with the highest number of neighbors, determining the model's predictions (Keikhosrokiani, P. 2022) . The selection of an appropriate value of k, the number of considered neighbours in the algorithm, is crucial as it affects the performance of the algorithm. We used the KNN imputer algorithm to fill our missing values in the variable "MntDrinks".An important parameter to define is the value of k, since a larger value of k can help reduce the impact of variance caused by random errors, but it also runs the risk of overlooking small yet significant patterns in the data (Zhang, Z. 2016). In our application we applied a GridSearchCV to find the best value, through hyperparameter optimization. A 5-fold cross validation was applied meaning the KNN imputer is trained and evaluated 5 times each time using a different fold as the validation set, from which a scoring metric is retrieved, the negative mean squared error. In this case the best scoring metric was determined to be n_neighbours = 1.

9.4. Silhouette Score

The silhouette score, also referred to as silhouette coefficient is a metric utilized to evaluate the quality of clustering. The Silhouete score ranges from -1 and 1 where higher silhouette scores indicate more coherent clusters. To calculate this coherence, it is necessary to calculation cluster cohesion and cluster separation. Cluster cohesion refers to the average distance between an instance and all other data points while cluster separation refers to the average distance between an instance and all other data points in the nearest cluster. A silhouette score of -1 indicates that the data points within clusters are dissimilar to each other and close to points in other clusters, suggesting that the data points could have been potentially assigned to the wrong clusters. A silhouette score of 0 indicates that the data points have similar distances to points in other clusters, which suggests that the clusters are not well separated or may be overlapping. A silhouette score of 1 indicates that the data points are like each other and distant from points in neighboring clusters, suggesting that these clusters are well defined and well-separate (Belyadi, H. and Haghighat, A. 2021)

9.5. RFM Score

RFM is a marketing technique that allows you to segment your audience according to their relevance to your business. This is done based on three criteria: Recency, Frequency and Monetary Value. RFM analysis is one of the most efficient methods to identify high-value targets and increase conversion rates.
To calculate the RFM score, each metric is assigned a numerical value based on the customer's behavior. For example, a customer who made a purchase yesterday would have a higher recency score than a customer who made a purchase six months ago. Similarly, a customer who makes purchases every week

would have a higher frequency score than a customer who makes purchases every six months. Finally, a customer who spends $1000 on purchases would have a higher monetary score than a customer who spends $100.

10. References

Belyadi, H. and Haghighat, A. (2021) “Unsupervised machine learning: Clustering Algorithms,” Machine Learning Guide for Oil and Gas Using Python, pp. 125–168. Available at: https://doi.org/10.1016/b978-0-12-821929-4.00002-0.

Connectif, (2022) What Are RFM Scores and How To Calculate Them, Available at: https://connectif.ai/en/what-are-rfm-scores-and-how-to-calculate-them/Connectif, (2022) What Are RFM Scores and How To Calculate Them, Available at: https://connectif.ai/en/what-are-rfm-scores-and-how-to-calculate-them/

Keikhosrokiani, P. (2022) Big Data Analytics for healthcare: Datasets, techniques, life cycles, management, and applications. Amsterdam: Academic Press.

P. Anitha, M.M. Patil, “RFM model for Customer Purchase Behavior using K-Means Algorithm”, Journal of King Saud University-Computer and Information Sciences, In Press, 2019

Zhang, Z. (2016) Introduction to machine learning: K-Nearest Neighbors, Annals of translational medicine. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4916348/

Team, G.L. (2022) Hyperparameter tuning with GRIDSEARCHCV, Great Learning Blog: Free Resources what Matters to shape your Career! Available at: https://www.mygreatlearning.com/blog/gridsearchcv/#:~:text=GridSearchCV%20is%20a%20technique%20for,parameter%20values%2C%20predictions%20are%20made. (Accessed: April 10,

Cao, Fuyuan; Liang, Jiye; Bai, Liang, A new initialization method for categorical data clustering, Expert Systems with Applications, vol. 36, Issue 7, September 2009, pages 10223-10228

Kabasakal, İ. (2020). Customer Segmentation Based On Recency Frequency Monetary Model: A Case Study in E-Retailing . Bilişim Teknolojileri Dergisi , 13 (1) , 47-56 . DOI: 10.17671/gazibtd.570866

Sharma, Neha; Ashok, Samrat; Gaud, Nirmal, K-modes Clustering Algorithm for Categorical Data, International Journal of Computer Applications ((0975 – 8887) vol 127 – No.17, Oct