Wonderful Wines of the World

Customer Segmentation

Data Mining I

Group 27

Nuno Neves 20230413
Daniel Santos 20230299
Mariana Ribeiro 20230303

INDEX

LIST OF FIGURES

ABSTRACT

The project's original goal was to define and improve Wonderful Wines of the World's marketing initiatives based on a variety of client traits, wine preferences, and wine-related behavior. Our team used a variety of Python approaches and algorithms to create a descriptive model in order to do this.

To achieve the desired outcomes, our initial step involved importing, investigating, examining, and modifying the raw dataset. The data preparation phase, adjustments were made based on insights gained through data visualization. Subsequently, our focus shifted towards modelling the datasets, wherein we categorized them into two key objectives: the value segment and wine preferences.

In pursuit of optimal results, we implemented clustering solutions, specifically employing the k-means technique. This approach proved to be most effective for our requirements. After carefully analyzing the data from the generated profiles, it is now possible to create personalized marketing campaigns for a specific target based on various consumer segmentations and behaviors.

I. INTRODUCTION

With eleven years of experience in the wine business, Wonderful Wines of the World (WWW) specializes in sourcing from unique and small wineries across the world. Their main objective is to present these distinctive wines to consumers, hoping to wow them with outstanding choices that encapsulate the spirit of their individual growing regions.

To accomplish this goal, ten actual physical locations thoughtfully placed in significant American cities, a user-friendly website, and both physical and electronic catalogs are used by WWW to make these wines more accessible. To provide a smooth and convenient buying experience, customers can choose to buy these wines via the website, phone orders, or the specialized mobile app.

The business targets particular clients who are interested in wine and have sufficient purchasing ability to treat themselves to a genuine wine experience as entertainment. Identifying each consumer as distinct in order to market their particular wines.

To understand further opportunities in cross-selling markets and to be able to produce direct marketing campaigns targeting the right customers for their behavior as wine appreciators, the company needs to start focusing on the customer experience and better understand their buying behavior, who are the most valuable customers, and what wines are bought together.

Marketing is the division of a company that helps with customer acquisition and deal completion. Because it increases sales and revenue, builds brand awareness, and draws in and keeps customers, effective marketing is essential to a company's success. To remain competitive in their sector, businesses must also regularly review and modify their marketing plans.

Segmenting your audience to improve communication is one approach to run marketing efforts that work. The technique of breaking a target market up into smaller groups according to shared traits is known as segmentation. Segmentation is crucial in marketing because it allows companies to customize their marketing campaigns aimed at particular customer demographics. Additionally, segmentation can assist companies in crafting more tailored marketing messages that have a higher chance of drawing in and keeping clients.

For this reason, the project's goal is to segment a portion of the Wonderful Wines of the World client database to create more marketing campaigns. For the purpose to address WWW needs, the research focuses on two distinct segmentations using data mining techniques and clustering algorithms. The value of the client group is identified in the initial segmentation. A description of the purchasing behavior is included in the second segmentation.

II. METHODOLOGY

1. Data Exploration and Understanding

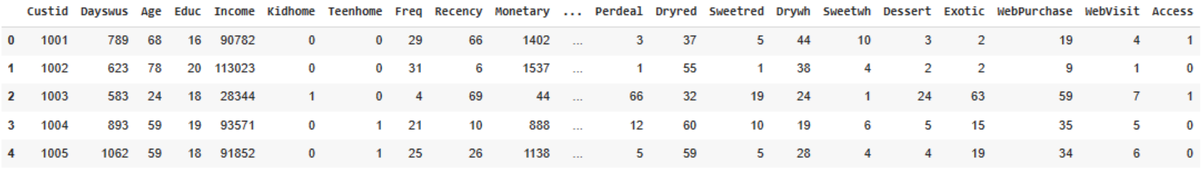
1.1. Description and Attributes

Wonderful Wines of the World (WWW) has supplied a dataset featuring 10,000 customers sourced from its current active database. These customers are defined as active based on their engagement, having made at least one purchase within the last 18 months.

The current dataset has 21 variables serving as key indicators and characteristics of the active customers. To facilitate the initial exploration and comprehension of the dataset, the table in appendix A presents all these variables along with their description.

This detailed information aims to provide a foundation for understanding and analysing the dataset, offering insights into customer description and characteristics and into customer behaviours and preferences, relevant factors critical to our project. This differentiation will be of upmost importance to focus on both and different objectives, the customers value segmentation and wine buying behaviour, addressed further in the project.

We started by assuming that all variables' data type were int64, however once we visualized the first 5 rows of the dataset, as shown below, we concluded that Kidhome and Teenhome variables are Boolean, by having 2 unique values, and so they were transformed. Having all the other variables more than 2 unique values, they stayed as int65.



	Custid	Dayswus	Age	Educ	Income	Kidhome	Teenhome	Freq	Recency	Monetary	...	Perdeal	Dryred	Sweetred	Drywh	Sweetwh	Dessert	Exotic	WebPurchase	WebVisit	Access
0	1001	789	68	16	90782	0	0	29	66	1402	...	3	37	5	44	10	3	2	19	4	1
1	1002	623	78	20	113023	0	0	31	6	1537	...	1	55	1	38	4	2	2	9	1	0
2	1003	583	24	18	28344	1	0	4	69	44	...	66	32	19	24	1	24	63	59	7	1
3	1004	893	59	19	93571	0	1	21	10	888	...	12	60	10	19	6	5	15	35	5	0
4	1005	1062	59	18	91852	0	1	25	26	1138	...	5	59	5	28	4	4	19	34	6	0

Figure 1 - First 5 rows of WWW Customer Dataset

We performed an analysis to assess several aspects of the dataset. Firstly, we examined the dataset for unique value to understand the diversity and distribution within each variable. Next, we investigated the presence of null values, identifying any missing values that might impact the integrity of our analysis. Additionally, we checked for duplicated rows to ensure data accuracy and prevent any potential biases. With this analysis we concluded that there were no null values nor duplicate rows.

Lastly, to better understand WWW dataset, we assembled a descriptive statistics table for all variables, present in appendix B. This multifaceted analysis aims to uncover insights and patterns within the dataset, setting the stage for more in-depth exploration and informed decision-making in our project.

By analysing these outputs, we can conclude several aspects like the fact that we have customers from a diverse range of ages and that there is a big range in Income variable, being the highest value more than 10 times higher than the minimum value. We also can assume that the preferred wine by the customers is the dry red, followed by the dry white. Lastly, we can also conclude that web purchases represent less than the physical ones and that, on average, customers tend to visit the WWW digital platforms 5 times per month.

Finally, we set the Custid as the data frame index, since it is an identifier of customers, and divided the features in non-metric, Kidhome and Teenhome, and metric ones, all the rest. This strategic segregation facilitates more accurate and meaningful application of clustering algorithms in the analysis.

1.2. Visualization – Histograms

To effectively identify and address outliers and inconsistencies, we proceeded with a visual exploration, starting with histograms, representing the distribution of each feature, and providing insights into the overall shape of the data.

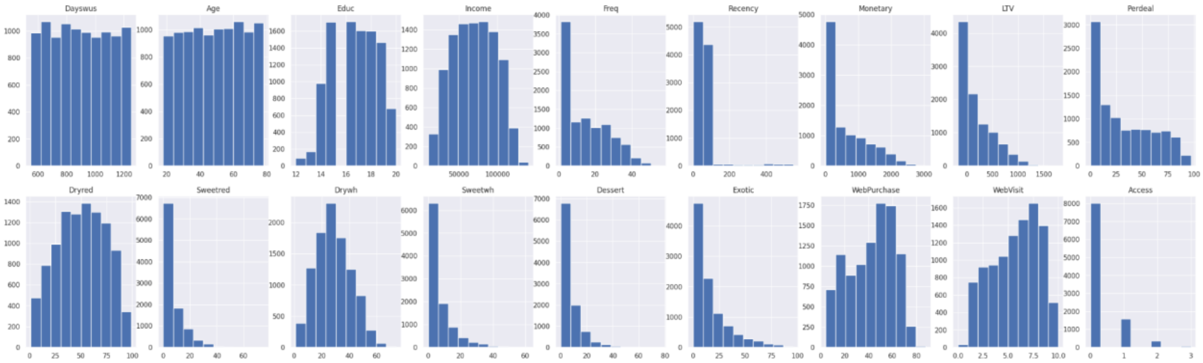


Figure 2 - Variables' Histograms

The examination of our data reveals distinct distributions among various variables. Variables such as Dayswus and Age exhibit a uniform or flat distribution, while Income, Dryred, and Drywh demonstrate a more symmetrical, normal distribution. On the other hand, most variables display a right-skewed pattern, except for WebPurchase and WebVisit, which present a left-skewed distribution.

This observation is further supported by the descriptive statistics table, where the means of the variables are larger than their respective medians. This discrepancy indicates the right-skewed nature of the distributions, contrary to left-skewed distributions where the median would be greater than the mean. Understanding these distribution characteristics is crucial for the identification and handling of outliers, laying the groundwork for subsequent analysis in our project.

Continuing our exploration, we focused on a deep understanding of the features Access, number of accessories a customer has purchased in the past 18 months, and Educ, educational background of a client in terms of years of education. To gain richer insights into the data, we created separate plots for each variable, employing distinct binning strategies.

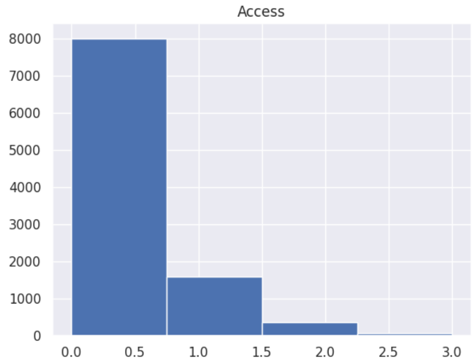


Figure 3 - Access Histogram

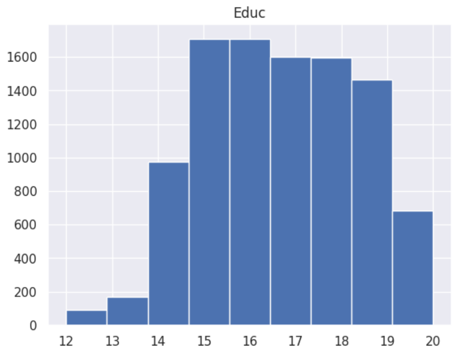


Figure 4 - Educ Histogram

It is possible to assume that the company faces a considerable challenge in marketing accessories to its clients, as evidenced by the fact that 8000 customers have not made recent accessory purchase. This observation aligns with the left-skewed distribution of the Access variable.

Additionally, it's worth noting that the variable Educ is predominantly concentrated in the range of 15 to 19 years of education, suggesting that many WWW customers possess a high level of educational attainment.

Continuing our data exploration, we delved deeper into the characteristics of the features Teenhome and Kidhome, both being Boolean variables. The below charts provide a clear representation of the count of each attribute, offering a visual overview of the distribution of customers based on the presence of teens or kids in their households.

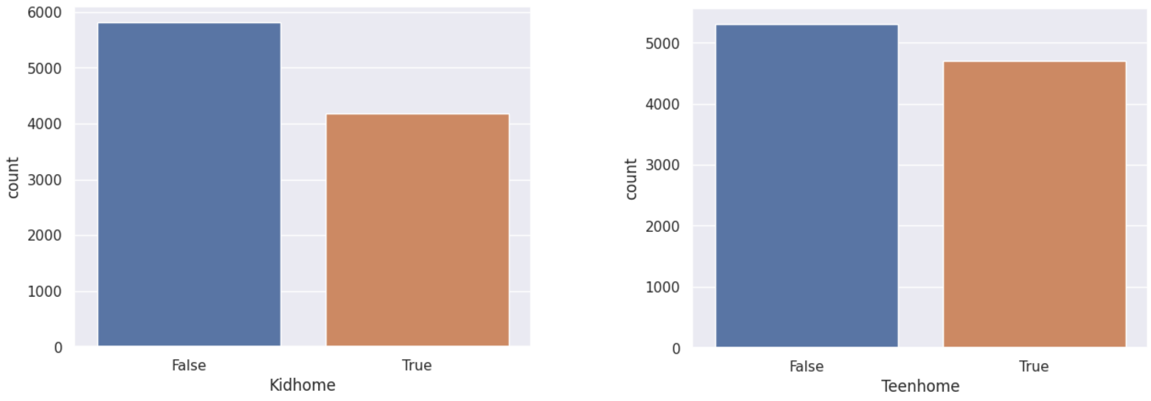


Figure 5 - Kidhome and Teenhome Histograms

To better capture their intended meaning, we adjusted the data type of Kidhome and Teenhome variables to Boolean. This modification was necessary due to the limitations of certain clustering algorithms, which struggle to normalize Boolean values effectively as numeric inputs. Consequently, we opted to exclude these variables from the set of features used in the solution.

1.3. Visualization – Box Plots

We utilized box plot charts to integrate the insights gleaned from earlier steps with the visual information presented in the box plots. This approach enables us to efficiently identify and address outliers, leveraging the graphical representation to enhance our understanding and handling of data points that deviate significantly from the norm.

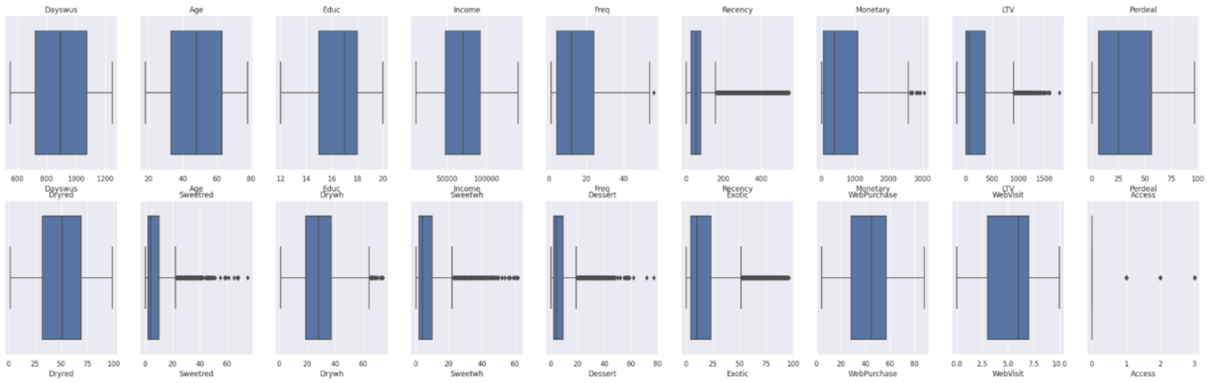


Figure 6 - Variables' Box Plots

1.4. Visualization - Correlation Matrix

To comprehend the interrelationships within the WWW dataset, we opted to conduct an analysis through the creation of a Correlation Matrix plot. The primary goal was to mitigate redundancy by excluding variables exhibiting high correlation, whether positive or negative. This approach was taken to prevent duplicative impacts on the ultimate results. Additionally, the process aided in reducing the overall dimensionality of the problem. We identified and removed correlated variables using the Pearson Correlation Coefficient as our chosen metric for this purpose.

Examining the graph, it becomes evident that the most strongly correlated variables include Age, Income, Freq, Monetary, LTV, WebPurchase, and WebVisit.

From the observation of the Correlation Matrix, we can deduce a high correlation between Age and Income, indicating that as age increases, so does the customer's income. Additionally, we observe a significant correlation between Income and Monetary, and both are correlated with Freq, in particular Monetary. This implies that higher greater expenditure on wine in the past 18 months correspond to increased buying frequency. Considering these interconnections among features, we have decided to eliminate Age, Freq, LTV, Monetary and WebPurchase from our analysis.

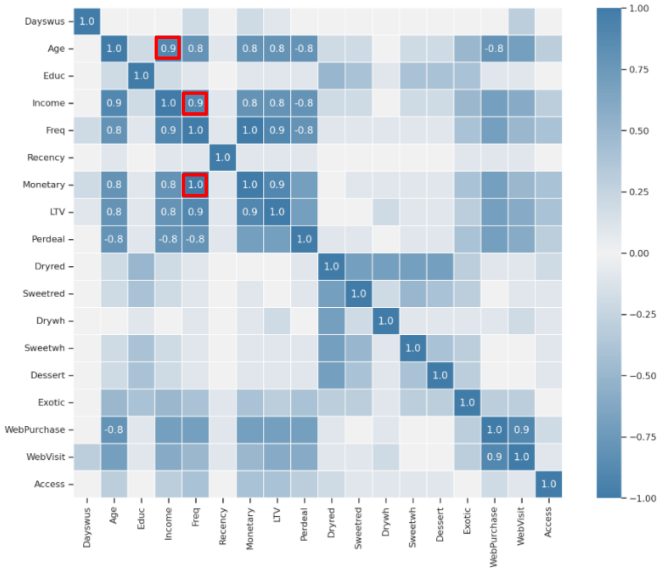


Figure 7 - Correlation Matrix

1.5. Visualization - Inconsistency Check

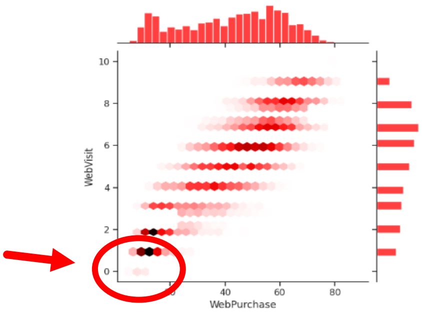
Following the collection of the primary data regarding the variables and the knowledge of distributions and correlations, it is necessary to determine whether the variables' values make sense and are consistent with one another. This is known as feature consistency.

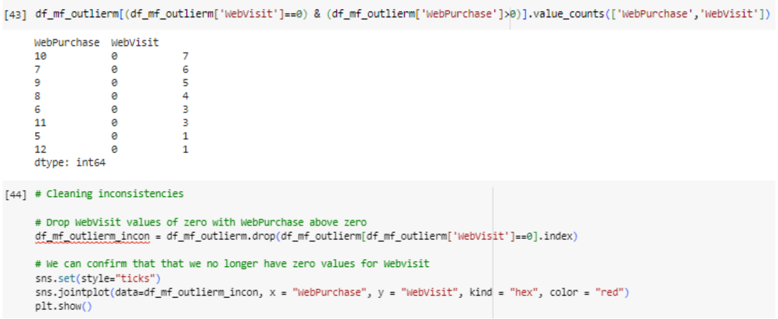
With this in mind, we draw the conclusion that it is not possible to include clients who have made online purchases (WebPurchases) in the dataset without any web visits. Based on the preceding comparison, we've come to the conclusion that neither can have more recent days than the days the customer is in the dataset and is a WWW customer (Dayswus).

We also checked to see if there was a discrepancy of more than six years between the age and educational attainment of the clients, which would indicate an inaccurate data point that ought to be ignored. In order to wrap up the examination of discrepancies, we have verified that the minimum age required to purchase wine is 18 years old, and we have looked for features that were meaningless if their values were equal to zero.

To check these incoherences, we plot the following charts and proceed to the execution of the following formulas:

1) Cleaning inconsistencies on WebVisit & WebPurchase

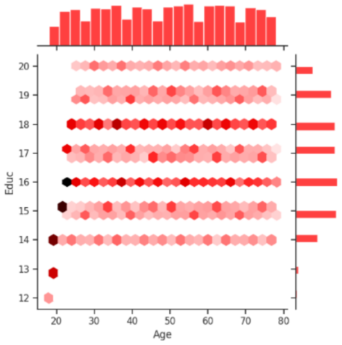


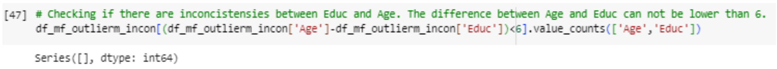


```
[43] df_mf_outlierm[(df_mf_outlierm['WebVisit']==0) & (df_mf_outlierm['WebPurchase']>0)].value_counts(['WebPurchase','WebVisit'])

     WebPurchase  WebVisit
     10           0           7
     7            0           6
     9            0           5
     8            0           4
     6            0           3
     11           0           3
     5            0           1
     12           0           1
     dtype: int64

[44] # Cleaning inconsistencies

     # Drop WebVisit values of zero with WebPurchase above zero
     df_mf_outlierm_incon = df_mf_outlierm.drop(df_mf_outlierm[df_mf_outlierm['WebVisit']==0].index)

     # We can confirm that that we no longer have zero values for Webvisit
     sns.set(style="ticks")
     sns.jointplot(data=df_mf_outlierm_incon, x = "WebPurchase", y = "WebVisit", kind = "hex", color = "red")
     plt.show()
```

Figure 8 - WebVisit and WebPurchase Cleaning Inconsistencies

2) Checking for inconsistencies on Educ & Age

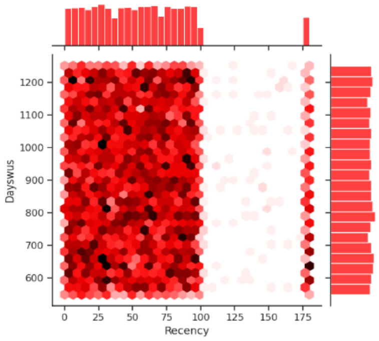


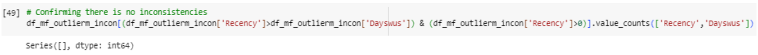


```
[47] # Checking if there are inconcistensies between Educ and Age. The difference between Age and Educ can not be lower than 6.
     df_mf_outlierm_incon[(df_mf_outlierm_incon['Age']-df_mf_outlierm_incon['Educ'])<6].value_counts(['Age','Educ'])

     Series([], dtype: int64)
```

Figure 9 - Educ and Age Cleaning Inconsistencies

3) Checking for inconsistencies on Dayswus & Recency





```
[49] # Confirming there is no inconsistencies
     df_mf_outlierm_incon[(df_mf_outlierm_incon['Recency']>df_mf_outlierm_incon['Dayswus']) & (df_mf_outlierm_incon['Recency']>0)].value_counts(['Recency','Dayswus'])

     Series([], dtype: int64)
```

Figure 10 - Dayswus and Recency Cleaning Inconsistencies

2. Data Preparation

Once the dataset, variables, and their unique features have been evaluated using descriptive statistics, the Data Preparation stage becomes essential to fulfilling the objectives of the overall data mining research. During this stage, the dataset is put through a certain number of transformation and cleanup operations that are meant to improve the quality of the data and maximize the results' computational efficiency. The analysis carried out in the visualization step informs these tasks.

We changed variable data types, eliminated and replaced outliers, assessed redundancy and relevance, and then used data scaling and dimensional reduction in order to realize benefits and efficiency.

Note: Custid feature was defined as an index of the dataset, once it is a unique key representing each customer.

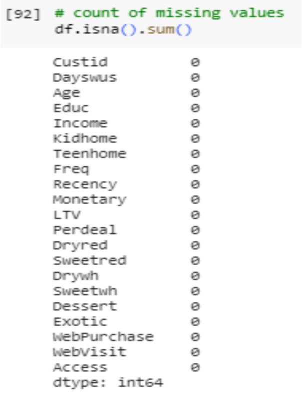
2.1. Variable data types

As mentioned before, the data type of the features Kidhome and Teenhome, which were declared as integers, was determined to be incorrect earlier in the Data Exploration phase via the dtypes method and unique function. It was deemed reasonable and correct to specify these features as Booleans since they can only have two possible values, 0 and 1.

2.2. Missing Values

The count of NaNs was utilized to identify any potential missing values in the dataset and replacement of the values was considered to ensure that any potential empty string values were not interpreted as NANs.

To appropriately proceed with data imputation by central tendency measures (median for metric features and mode for non-metric features) or K-Nearest Neighbors imputer, however, no missing values were found in either approach so it wasn't necessary to apply methods.



```
[92] # count of missing values
     df.isna().sum()

     Custid          0
     Dayswus         0
     Age             0
     Educ            0
     Income          0
     Kidhome         0
     Teenhome        0
     Freq            0
     Recency         0
     Monetary        0
     LTV             0
     Perdeal         0
     Dryred          0
     Sweetred        0
     Drywh           0
     Sweetwh         0
     Dessert         0
     Exotic          0
     WebPurchase     0
     WebVisit        0
     Access          0
     dtype: int64
```

Figure 11 - Dataset Missing Values

2.3. Outliers

As was previously noted, assessing the presence and management of outliers was an important stage because they might add bias and affect the ongoing analysis. Histograms and box plots were first used to find outliers. The effectiveness of the four outlier removal techniques—manual, IQR Method, KNN and manual plus substitution— was then reevaluated, along with the percentage of data that was kept. Finding the best outlier elimination technique for our analysis required this process.

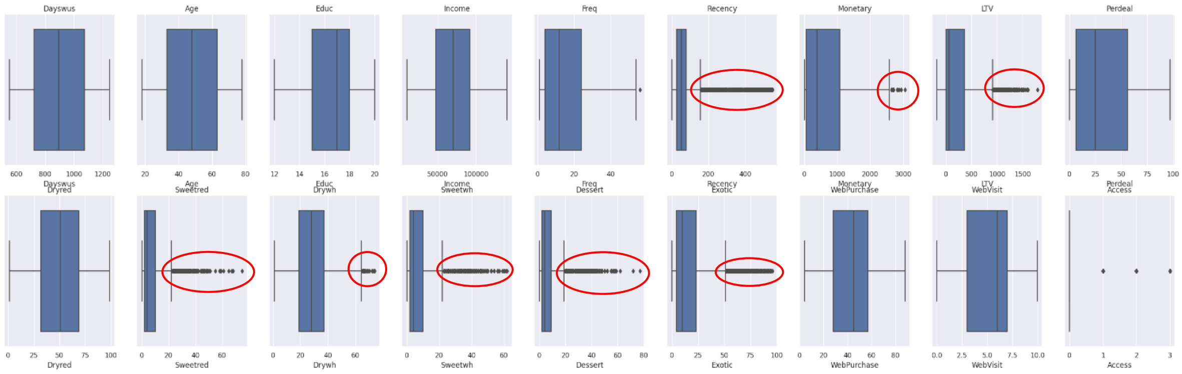


Figure 12 - Outliers

2.3.1 Manual Removal

As seen in the accompanying box plot analysis, the manual strategy used just visual inspection of the data frame to identify and choose outliers for removal. The plot's red boxes indicate the presence of outliers, and the criteria used to decide their removal are listed below:

- Recency: Greater than 350
- Monetary: Greater than 2750
- LTV: Greater than 1500 on LTV
- Sweetred: Greater than 50
- Drywh: Greater than 70
- Sweetwh: Greater than 50
- Dessert: Greater than 45
- Exotic: Greater than 75

It is significant to note that the removal of those outliers did not cause a significant loss of data—the loss amounted to just 3,8% of the observations in the entire dataset.

2.3.2 IQR Method

The Interquartile Range (IQR) method involves a detailed analysis of box plots that includes all features' components: Q1 (25%), Q3 (75%), Median (50%), and the interquartile range (Q3-Q1). The computation of outliers designated for elimination depends on this thorough study. Lower and Upper boundaries are obtained by the computation, and numbers outside of these ranges are categorized as outliers that need to be removed. The lower_bound (Q1 - 1.5 * IQR) and upper_bound (Q3 + 1.5 * IQR) are the exact definitions of these boundaries. After using this process, it was discovered that **20%** of the dataset's total observations were lost, **suggesting that this approach might not be the best one**.

2.3.3 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a machine learning algorithm used for classification and regression tasks. It operates on the principle of proximity, where an instance is classified based on the majority class of its nearest neighbors in the feature space. In the context of outlier detection, KNN can be leveraged to identify data points that deviate significantly from their neighbors, making it a valuable tool for improving data quality.

After applying the KNN method in our Data Mining I project, noteworthy outcomes have been achieved:

- Number of Outliers Removed: 789
- Percentage of Outliers Removed: 4.98%
- Percentage of Data Kept After Removing Outliers: 95.02%

These results signify the successful identification and removal of outliers, amounting to nearly 5% of the dataset. The retained data, comprising 95.02%, reflects the effective application of the KNN method in enhancing the quality of our dataset for subsequent analysis. Although the KNN approach proves to be a valuable technique for outlier detection, we opted for the Manual method.

2.3.3 Manual & Substitution

After taking into account the findings of the significant loss of data when utilizing the IQR methodology and KNN, in comparison to the manual strategy of presenting acceptable values for lost data, the two following approaches were chosen:

Manual approach:

- Monetary: Greater than 2500
- Sweetred: Greater than 35
- Drywh: Greater than 65
- Sweetwh: Greater than 35
- Dessert: Greater than 35

Substitution by imputing a threshold value for the outlier values within the next conditions:

- Recency: Greater than 150
- LTV: Greater than 900

By combining these two methods, we were able to represent the dataset significantly more smoothly and with an acceptable data loss of only about 4%. The box plots below show the outcomes of this procedure, which was chosen as the ultimate approach to remove outliers. The green boxes represent the corrections made, the red boxes represent the outliers that remained, and the grey boxes are explained below.

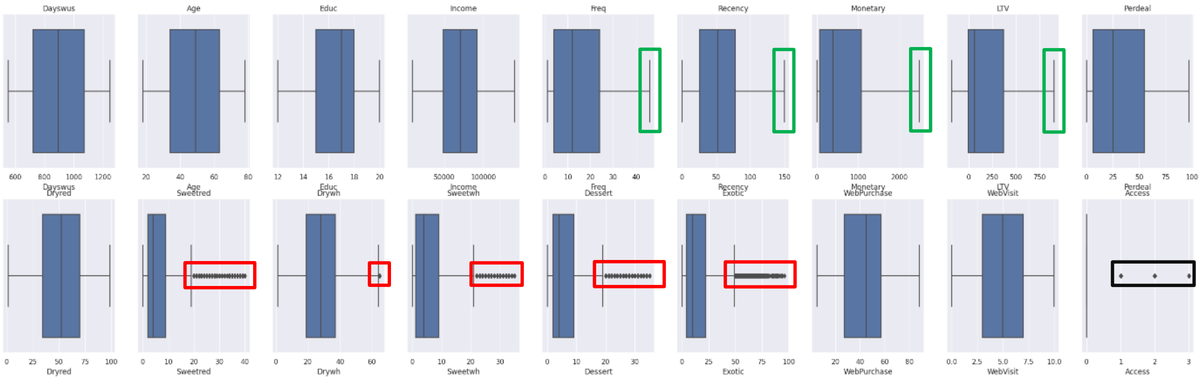
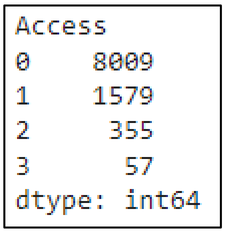


Figure 13 - Manual and Substitution Approach Outcome

Following the removal of outliers, we reassess descriptive statistics to gauge the impact of the deleted observations. As observed, our outlier treatment method effectively addressed all outliers in Frq, Recency, Monetary and LTV. To preserve the true variability representation, especially for features related to wine types and considering the segmentation of wine purchase behavior, a decision was made to refrain from further outlier removal for these features. For instance, the "Exotic" feature retained its outliers since removing them all would obscure valuable information about the purchase behavior of the samples, making it challenging to identify the meaningful "breaking point" of the outliers.

The variable "Access" is noticeably absent from the preceding list, despite having a notable number of outliers. The distribution of "Access" is characterized by its association with the quantity of accessories purchased alongside other items. Given this nature, we deem it necessary to retain the complete set of variables. In this context, we would only label a value as an outlier if it reaches a count of 10 for the "Access" variable.

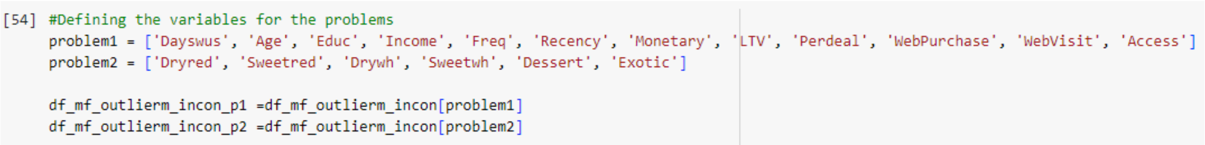


```
Access
0     8009
1     1579
2      355
3       57
dtype: int64
```

2.4. Feature Engineering

2.4.1 Data separation

The primary features for each problem were chosen, and the original data frame was split in half so that different approaches could be used for the various problems encountered to handle future data about both problems. Finally, we arrived at the following resolution:



```
[54] #Defining the variables for the problems
     problem1 = ['Dayswus', 'Age', 'Educ', 'Income', 'Freq', 'Recency', 'Monetary', 'LTV', 'Perdeal', 'WebPurchase', 'WebVisit', 'Access']
     problem2 = ['Dryred', 'Sweetred', 'Drywh', 'Sweetwh', 'Dessert', 'Exotic']

     df_mf_outlierm_incon_p1 =df_mf_outlierm_incon[problem1]
     df_mf_outlierm_incon_p2 =df_mf_outlierm_incon[problem2]
```

Figure 14 - Outliers Definition

2.4.2 Incoherence Check

The validity and consistency of the observations being studied are crucial and essential for them to be employed to further the company's business goals. In order to handle the necessary changes on the illogical elements, such as in point 1.5 Inconsistency Check, when we compare the WebPurchase versus Webpage Visit. None of the remaining discrepancies needed the dataset to be cleaned up or transformed.

After addressing the coherence of WebVisit and WebPurchase, we move on to the last format changes:

The dataset contained observations that purportedly represented online transactions made by customers without ever visiting the store's website, which is inconsistent. Thirteen observations were removed to remedy this problem; these clients had WebVisit values of zero and WebPurchase values greater than zero.

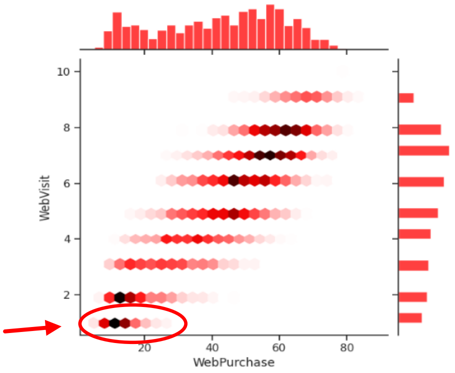


Figure 15 - WebVisit and WebPurchase Joint Plot

When features like Dayswus, Recency, Age, and Educ were examined, none of the zero values were discovered, indicating that there was no incoherence. Age and education were compared to see whether any observations were made about students starting school earlier than the age of six, which they weren't.

2.4.3 Creating new variables

Throughout this phase, characteristics were created to improve the data set and facilitate subsequent analysis in the development process. The decision to create the following features was intended to derive them from transformations of existing variables. However, it was determined that applying these new features would not bring additional benefits in line with the objectives of this study. The proposed new features included:

- **Perdeal_number:** Freq*(Perdeal/100) - where the perdeal number is presented in % and we will normalize it to a specific number.
- **WebPurchase_value:** Monetary*(WebPurchase/100) - where the WebPurchase value we will normalize it to a specific number.
- **WebPurchase_number:** Freq*(WebPurchase/100) - where the WebPurchase number is presented in % and we will normalize it to a specific number.
- **All wine types to values** (Dryred, Sweetred, Drywh, Sweetwh, Exotic): WhineType*Monetary - where the wine type value we will normalize it to a specific number.

For the previous metrics, we believe that the only new feature that would be interesting to analyze would be the average sale value per each client purchase and the average price of each wine type per client.

- **Avg Purchase Value per client:** Monetary/Freq
- **Avg Price Values for each wine type:** (%Wine * Monetary Value)/Freq

2.4.4 Correlations

By using the correlation matrix, variables that are highly associated can be found and eliminated from the main data frame. Most of the variance and interpretability of the data are preserved by this method. We proceed with a feature removal criterion that is comparable to the insights and evaluations from the visualization step, where correlation values more than |0.8| were considered high.

It was decided to move forward with feature elimination following a thorough assessment of the general association between the metric features. This crucial stage is carried out in accordance with the division of features pertaining to issues 1 and 2. It is the ultimate visualization. The predetermined data frames that were previously established are used to apply this method to each of the problems that have been discovered.

Correlation Matrix – Problem 1:

Based on the following image (highly associated features are shown in red), the features that were eliminated were:

- **Age**: vs Income - where Income translates into business as Age is higher, income is higher too.
- **Freq**: vs Income – where Freq is directly correlated with the increase of income.
- **Monetary**: vs Income, LTV- where Monetary is directly correlated with the increase of Income and LTV.
- **LTV**: vs Income – where it increases when Monetary increases too.
- **WebPurchase**: vs Webvisit – as it exhibits a strong correlation with WebVisit, the decision to choose between the two is influenced by the fact that WebPurchase demonstrates the highest correlation values compared to the other existing features. (purple shape showing the contrast between colors of WebPurchase and WebVisit).

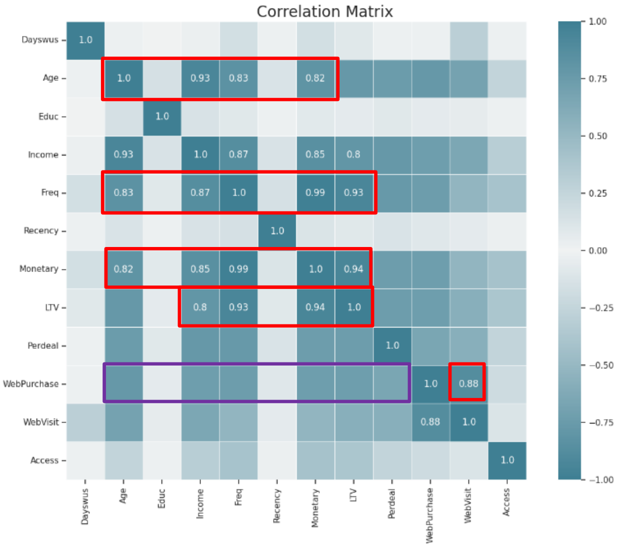
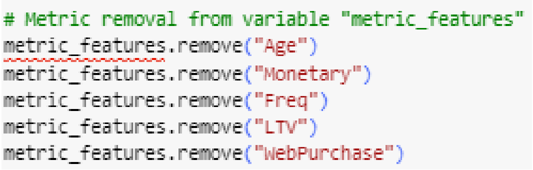


Figure 16 - Correlation Matrix Problem 1

Correlation Matrix – Problem 2:

Following the removal and validation process, the metric_features were appropriately updated to eliminate the features that were dropped from the initially defined metric data frame. Correlation extraction is unnecessary.



```
# Metric removal from variable "metric_features"
metric_features.remove("Age")
metric_features.remove("Monetary")
metric_features.remove("Freq")
metric_features.remove("LTV")
metric_features.remove("WebPurchase")
```

Figure 18 - Metric Feature Removal Code

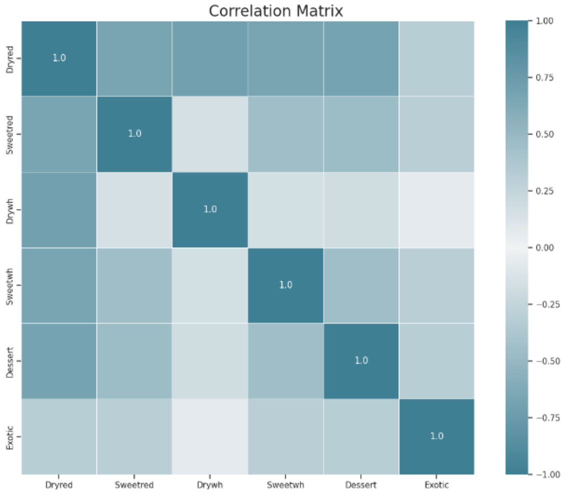


Figure 17 - Correlation Matrix Problem 2

In simpler terms, we have confirmed that there are no strong correlations (beyond |0.8|) among the features relevant to problem 2, which focuses on wine segmentation. After thoroughly examining all correlations, we have derived the final correlation matrix for the overall data frame.

2.4.5 Scaling

Data scaling is a crucial step in data preparation, particularly when the distance between data points significantly impacts cluster determination. Scaling ensures that all variables are on a consistent scale, preventing any one variable from disproportionately influencing results. To achieve this, we initially employed the Min-Max Scaler, restricting metric feature values to a range of 0 to 1. We opted to continue with the Min-Max Scaler due to its effectiveness, especially in cases of skewed distributions observed in some features within our dataset.

3. Modelling

The next step, after preparing the data, is to apply clustering algorithms to segment our customers. The segmentation will follow the two problems identified in the previous phase, Customer Value Segmentation and Wine Segmentation. In both perspectives, we started by applying R2, where we identified from the outset that the method, we were going to use would be K-Means. However, Ward Dendogram also proved to be a viable alternative.

For both problems, we started by analyzing K-Means, as this was the most reliable method and after that we chose to use it, Hierarchical Clustering, DBSCAN and only in problem 1 RFM Clustering.

3.1. Problem 1 - Costumer Value Segmentation

Considering the variables defined for problem 1 (Dayswus, Educ, Income, Recency, Perdeal, WebVisit, Access), the aim would now be to segment customers.

R2 Scores

To do this, we began by calculating the R2 scores for each cluster solution in this way we can spot the number of k's where the R2 metric variance decrease and the best method to apply.

Here we saw that the two methods with the highest R2 were K-Means and Hierarchical Clustering. Besides that, we can even see that the number of clusters should be 3 or 4.

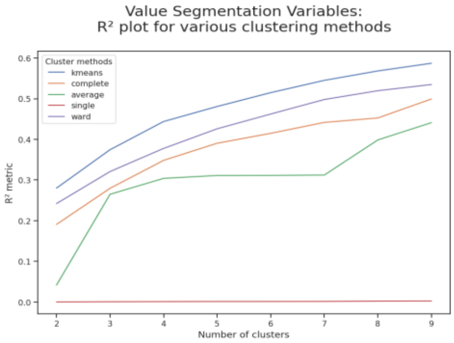


Figure 19 - R2 Plot for Clustering Methods for Problem 1

K-Means

In order to choose the right number of clusters to apply the K-Means Method, we decided to calculate the Elbow Curve and the Silhouette Curve. And as we can see in the graphs, the elbow curve starts to add one more cluster and the number of observations starts to get smaller when we go from 4 to 5 clusters. That's why we decided to use 4 clusters in the K-Means method.

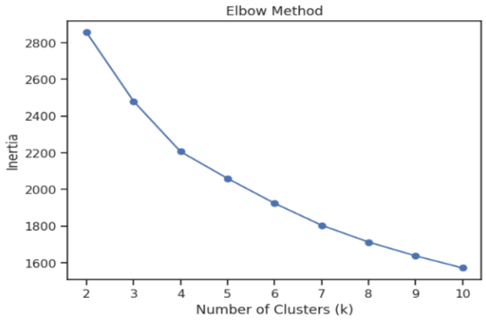


Figure 20 - Elbow Curve for Problem 1

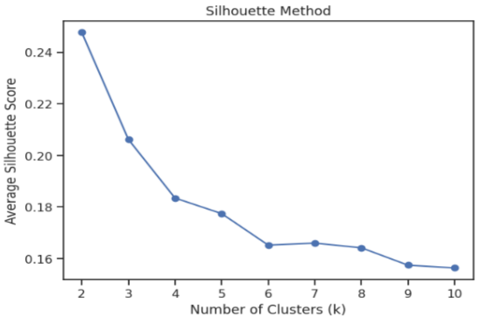


Figure 21 - Silhouette Method for Problem 1

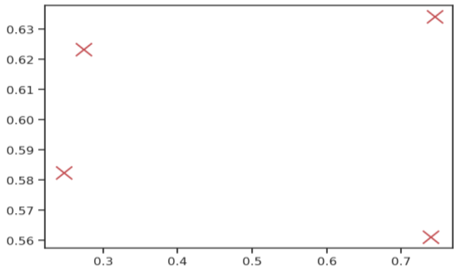


Figure 22 - K-Means Method – Centroids for Problem 1

We initialized k-means several times with different seeds to find the initialization that gives us the smallest sum of square roots, i.e. the initialization that minimizes the intra-cluster distances. After running the algorithm, we were able to plot the centroids of the clusters.

By analysing the centroids, we can tell that they are distant from each other, and the centroids of the clusters can be identified visually.

Hierarchical Clustering

After calculating K-Means and identifying the clusters, we decided to try other methods to see if the solution we had identified was correct.

To begin with, we decided to use Hierarchical Clustering, which gave us 5 clusters as a solution. We didn't think it was 100% incorrect, as the distance between 6 and 4 clusters was very small, hence the possible result.

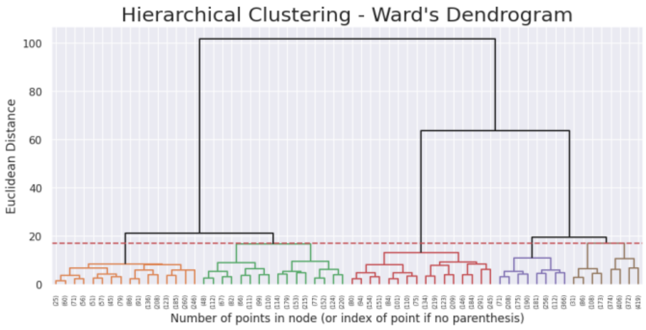


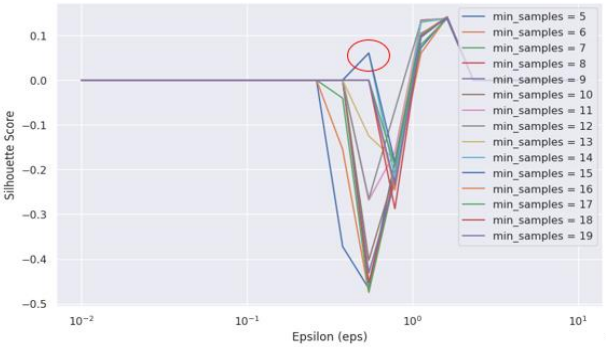
Figure 23 - Hierarchical Clustering for Problem 1

DBSCAN

The following Method we decided to test was DBSCAN algorithm, and a key point to produce good results in DBSCAN is the hyper parameterization, where it is intended to produce the best inputs for the number eps (maximum distance between two points in the same cluster) and the minimum of samples to check (as min_samples).

In order to check the best parameters, we designed the following graph, which will analyse the eps and min samples that correspond to the different Silhouette scores.

The appropriate value for Silhouette Score is greater than 0, but the maximum score for Silhouette gives us the worst eps and



min_samples, so we decided to use eps=0.19, min_samples=7. Which gave us the following results, 3 clusters, and 0,31% of noise rows. Although, the results are aligned with the previous ones and we obtained a good % of noise rows, we have decided to not give so much importance to this method and pay more attention to the others.

3.2. Problem 2 - Wine Segmentation

Taking into account the variables defined for problem 2 (Dryred, Sweetred, Drywh, Sweetwh, Desset, and Exotic), the aim would now be to segment customers.

In accomplishing the project's second objective, which involves segmenting the Wonderful Wines of the World (WWW) customer database according to purchasing behaviour, we will employ a similar approach to determine the number of clusters as applied in the first problem. However, this won't mean that the analysis and results will be the same, since we are using a different set of variables.

R2 Scores

To do this, we also began by calculating the R2 scores for each cluster solution in this way we can spot the number of k's where the R2 metric variance decrease and the best method to apply based on the higher R2 possible.

As with problem 1, the best methods to use are K-Means and Ward Dendrogram. Besides that, we can even see that the number of clusters should be 3 or 4, although in this case the number 3 is apparently the best number.

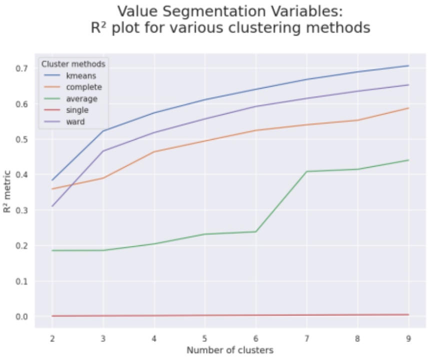


Figure 25 - R2 Plot for Clustering Methods for Problem 2

K-Means

Even though from R2 we realized that 3 clusters would be the best decision, we decided to calculate the Elbow Curve and Silhouette Curve. And we were able to prove that 3 clusters are indeed the best option because the elbow curve starts to add one more cluster and the number of observations starts to get smaller when we go from 3 to 4 clusters and with 3 clusters, we have the higher score Silhouette Curve. That's why we decided to use 3 clusters in the K-Means method.

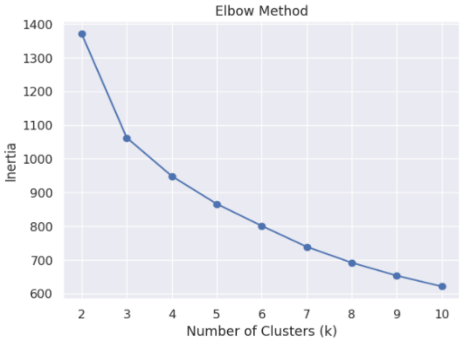


Figure 26 - Elbow Curve for Problem 2

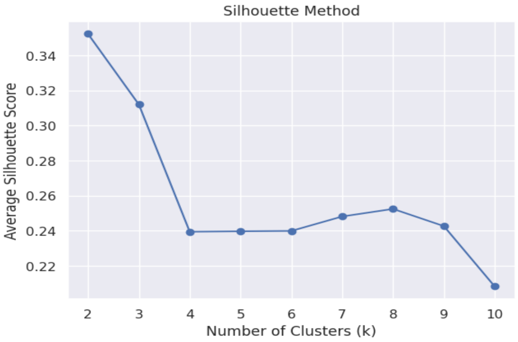


Figure 27 - Silhouette Method for Problem 2

We did the same as in problem 1 initialized k-means several times with different seeds to find the initialization that gives us the smallest sum of square roots, i.e. the initialization that minimizes the intra-cluster distances. After running the algorithm, we were able to plot the centroids of the clusters.

By analysing the centroids, we can tell that they are distant from each other, and the centroids of the clusters can be identified visually.

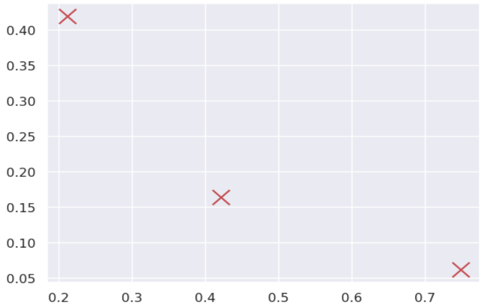


Figure 28 - K-Means Method – Centroids for Problem 2

Nevertheless, we decided to compare the K-Means solution with 4 clusters and found that two of them were extremely close to each other and that the best solution would be 3 clusters.

Hierarchical Clustering

After K-Means we also decided to test Hierarchical Clustering, and by analysing the result of the Dendrogram we can see that the best number of clusters is 3 (as we did find in K-Means). Once again, we can see that the best adjuster is k-means, and the number of clusters should be 3, as we increase the number of k's the R2 decreases in variance.

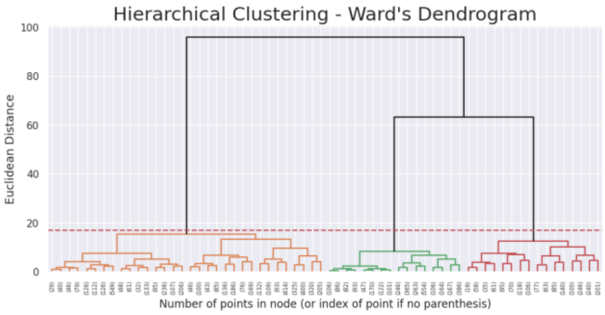


Figure 29 - Hierarchical Clustering for Problem 2

DBSCAN

The following Method we decided to test was DBSCAN algorithm, and as we did in Problem 1, in order to check the best parameters, we designed the following graph, which will analyse the eps and min samples that correspond to the different Silhouette scores.

The appropriate values for Silhouette Score are greater than 0, but the maximum score for Silhouette gives us the worst eps and min_samples, so we decided to use eps=0.19, min_samples=7. Which gave us the following results: Number of estimated clusters = 4, and 4,86% of noise rows. As we did in problem 1, we have decided to not give so much importance to this method and pay more attention to the others.

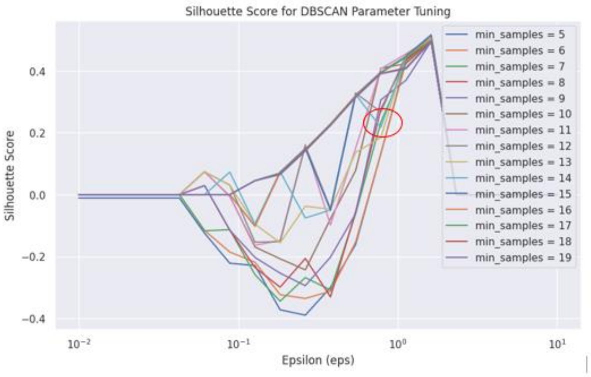


Figure 30 - Silhouette Score for DBSCAN Parameter for Problem 2

III. Results

To present the results, we have again combined the two perspectives into a single view with all the characteristics. To do this, we merged the results of each perspective selected, obtaining a third solution resulting from this combination. To follow up this fusion, we used the hierarchical method which pointed to a final clustering solution.

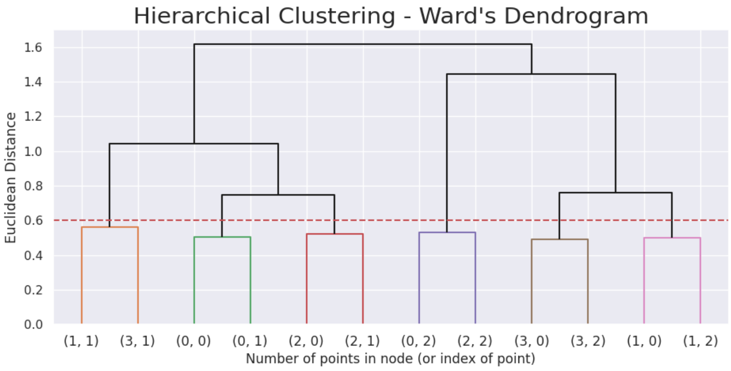


Figure 31 - Hierarchical Clustering - Ward's Dendogram

1. Profiling

Now we are going to observe the associated behaviours and features for each individual clustering solutions, and as well for the combined solution.

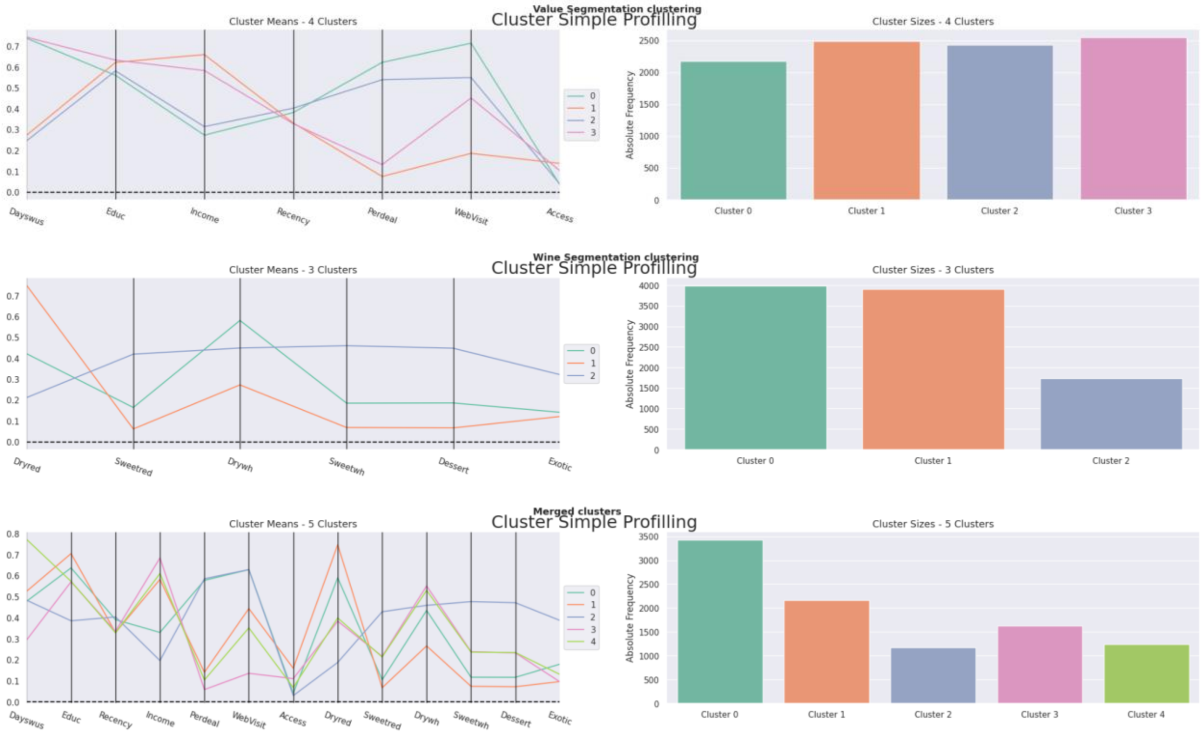


Figure 32 - Cluster Simple Profilling

Based on the above observations, we delve into a comprehensive exploration of how different behaviours manifest themselves in various groups.

In the detailed overview of value segments, it is essential to highlight the fundamental role played by customers associated with cluster 1, who are distinguished by their substantial monetary value. At the same time, cluster 0 emerges as a cluster that includes customers with a high number of days of use, indicating a propensity for frequent visits to the website, but at the same time are the cluster with less money. Turning our attention to wine purchasing behaviours, a captivating narrative unfolds as we examine the preferred wines within each group, and the preferences are Dryred and Drywh.

Take the merged solution in consideration, individuals with the highest income value show a distinct inclination towards Dry Wines, whether of the red or white variety, summarized by the distinctive characteristics of clusters 0 and 1.

To enhance our understanding, the following heatmaps. This visual aid not only reinforces the knowledge discussed earlier, but also provides a deeper and more granular understanding.

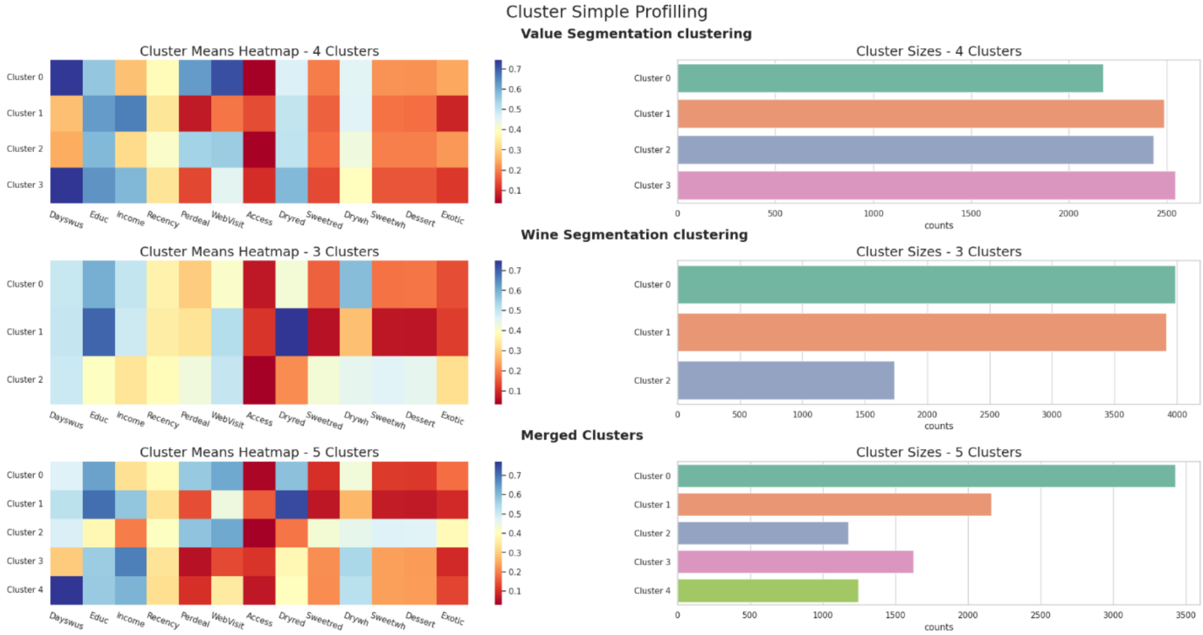


Figure 33 - Clusters Heatmaps

To validate the exact implementation of the clustering solutions for problem 1 and problem 2, we examined the t-SNE plots, as the dimensionality reduction seems to have been carried out effectively. The images presented below show a discernible and distinct picture of the clusters within the plotted points, affirming the accuracy of the implementation.

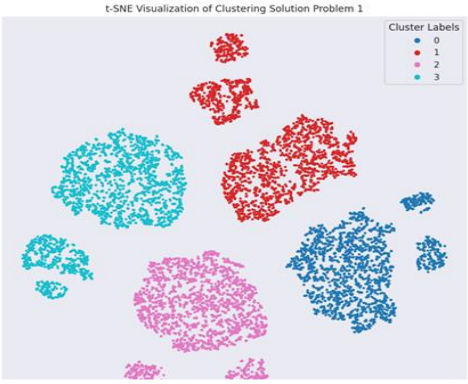


Figure 34 - t-SNE Visualization of Clustering Solution Problem 1

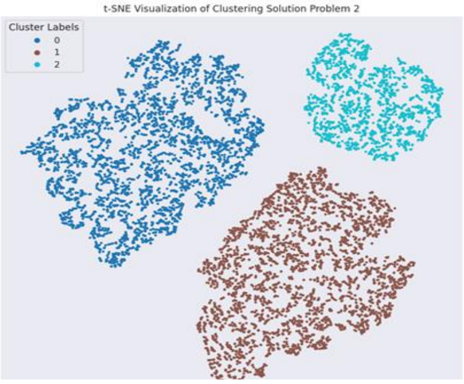


Figure 35 - t-SNE Visualization of Clustering Solution Problem 2

IV. CONCLUSION

Dividing customers into distinct groups according to their characteristics and behavior presents a valuable strategy for streamlining marketing efforts, leading to both time and cost savings. With this objective in mind, the insights gained from this project on WWW customers should serve as a roadmap for enhancing efficiency in future marketing campaigns. The group has provided recommendations for refining each individual solution and exploring additional opportunities within the combined clustering approach. These suggestions offer avenues for further development and improvement in the optimization of marketing strategies.

Main insights from Problem 1

- **C0** is characterized by having the lowest INCOME, the highest PERDEAL, and the highest WEBVISIT among our customers. As this customer is particularly sensitive to price changes, they fall within the target audience for promotional campaigns. It is worth noting that online promotional campaigns conducted through the website may be more effective in reaching and engaging this customer segment.

- **C1** is a customer who has recently begun making purchases on WWW, having the highest INCOME and the lowest PERDEAL. While not fitting the profile of the most loyal customers, this individual can be perceived as among the most valuable due to their substantial INCOME and minimal purchases of discounted products. However, this customer does not align with the target audience for promotional campaigns.

- **C2** stands as our least valuable customer, having recently begun making purchases on WWW, with a low INCOME and a high PERDEAL. This customer, characterized by sensitivity to campaigns and discounts, along with C0, may be included in the campaign's target group, whether through the website or catalogs. Nevertheless, it's essential to note that this customer category should not be the primary focus of the campaign.

- **C3** holds the distinction of being our most loyal and valuable customer. With the lengthiest purchasing history on the WWW, this customer possesses a high INCOME and the lowest PERDEAL. Such customers are integral to the company's success and should be preserved, receiving special attention and prompt resolution of any issues. Given their non-sensitivity to price changes, this customer category should be excluded from the target group of any promotional campaign.

Main insights from Problem 2

- **C0 and C1** represent the clusters of dry wine, with emphasis on Dryred (C1) and DryWh (C0). Given their susceptibility to advertising of this kind, these customer categories would react to a dry wine campaign.

- **C2** - the generalist buyer - is the cluster formed by the less loyal wine type customers. There is no dominant wine type, their basket is mostly homogeneous across all wine types. For these clients, multipacks would be a good option, as well as assorted half-sized bottle gift boxes or testers. To create more adapted strategies to cover their interests, crossing information from both clustering solutions would be crucial, as will be approached right away.

Main insights of Merged solutions

C0 – The low budget older clients:

They have a high Perdeal, high Webvisit and the 2nd highest Educ, while having the 2nd lowest Income. Given the strong association between education and age, we can assume that these customers are one of the older ones we have. Their preference are Dry wines, particularly the Dryred. The company could think about creating a dry wine campaign for these types of customers, combining their will to buy products in discount and their will to buy dry wines.

C1 – Modern vintage clients:

Characterized by the highest Education (Educ), a high income, low Perdeal and being the biggest buyer of the dry red wine, we can include this client on the most valuable clients list. This client is not sensitive to price changes, as he other things. This client loves to have a good experience while buying, so the company should give him a good in-store experience, and, while buying online, the company could offer him a wine related freebee/souvenir.

C2 –Youngest clients:

They have the lowest scores for Income and Education and the highest values for Perdeal and Webvisit. Given the strong association between education and age, the customer in this cluster should be the youngest as they have the lowest educational values. This customer still doesn't have a preferred wine, as they buy all types of wines, in exception of dryred. This type of client values a campaign that give a discount on multiple types of wine, across dry, sweet, exotic, etc (could be a basket with 3 different wine types).

C3 – The new clients:

Having the lowest Daysus value, we can consider them as the new clients in the block. They have a considerably high Education (Educ), the highest Income, the lowest perdeal, low Web Visits and a particular love for dry wine, in particular white (drywh). With this type of customer being the new clients, we can conclude that the company is moving to a place where they are a premium choice. WWW needs to be cautious with first impressions transmitted to the client, as they need to engage and secure them. This customer wants a good in-store experience, where the company can think about offering simple welcoming gifts, one-time special vouchers, and an opportunity to join special online subscription services could be some good commercial hooks. Additionally, a loyalty program, with advantages as they buy more, could be implemented.

Strategy: welcoming samples, gift, or voucher; online subscription plan; loyalty programs.

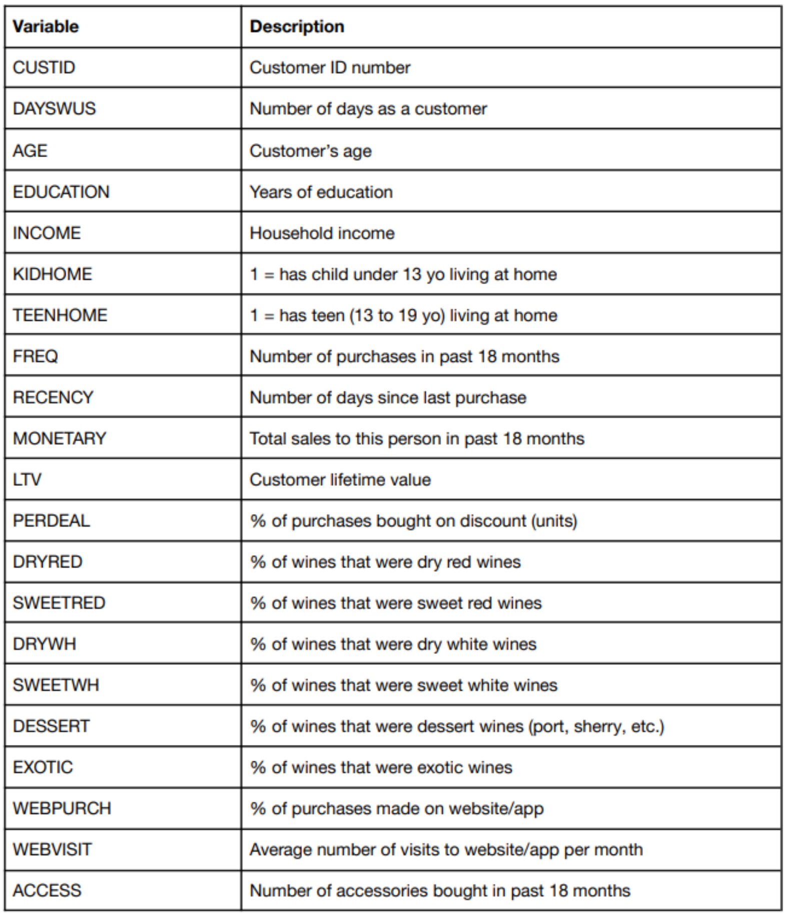
Cluster 4 - The most loyal clients:

Identifiable for the highest Dayswus, 3rd highest Income, and 2nd lowest Perdeal, this can indicate that they are not sensitive to price changes. Their purchases are scattered across physical stores and website and, for this motive, having in mind the evidence of dry wines preference (white dominant), WWW can build some strategy around dry wine for these types of customers. With these being the most loyal customers, and having a high income, WWW should be cautious with their churn rate by giving them personalized support and special treatment.

APPENDIX

Appendix A

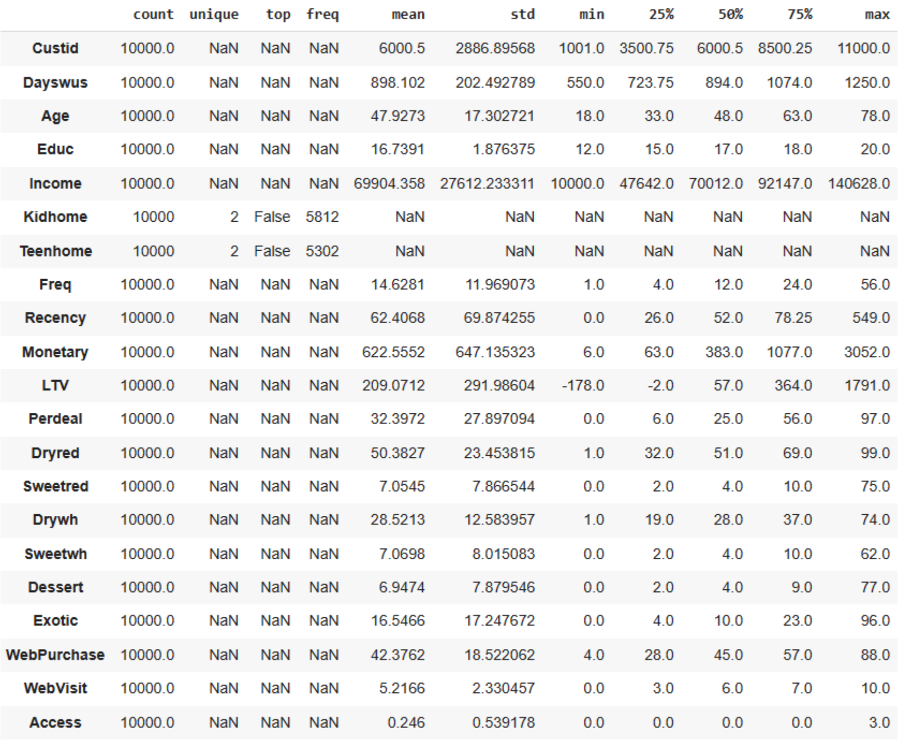
Variables’ Description



Variable	Description
CUSTID	Customer ID number
DAYSWUS	Number of days as a customer
AGE	Customer’s age
EDUCATION	Years of education
INCOME	Household income
KIDHOME	1 = has child under 13 yo living at home
TEENHOME	1 = has teen (13 to 19 yo) living at home
FREQ	Number of purchases in past 18 months
RECENCY	Number of days since last purchase
MONETARY	Total sales to this person in past 18 months
LTV	Customer lifetime value
PERDEAL	% of purchases bought on discount (units)
DRYRED	% of wines that were dry red wines
SWEETRED	% of wines that were sweet red wines
DRYWH	% of wines that were dry white wines
SWEETWH	% of wines that were sweet white wines
DESSERT	% of wines that were dessert wines (port, sherry, etc.)
EXOTIC	% of wines that were exotic wines
WEBPURCH	% of purchases made on website/app
WEBVISIT	Average number of visits to website/app per month
ACCESS	Number of accessories bought in past 18 months

Appendix B

Variables' Descriptive Statistics



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Custid	10000.0	NaN	NaN	NaN	6000.5	2886.89568	1001.0	3500.75	6000.5	8500.25	11000.0
Dayswus	10000.0	NaN	NaN	NaN	898.102	202.492789	550.0	723.75	894.0	1074.0	1250.0
Age	10000.0	NaN	NaN	NaN	47.9273	17.302721	18.0	33.0	48.0	63.0	78.0
Educ	10000.0	NaN	NaN	NaN	16.7391	1.876375	12.0	15.0	17.0	18.0	20.0
Income	10000.0	NaN	NaN	NaN	69904.358	27612.233311	10000.0	47642.0	70012.0	92147.0	140628.0
Kidhome	10000	2	False	5812	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Teenhome	10000	2	False	5302	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Freq	10000.0	NaN	NaN	NaN	14.6281	11.969073	1.0	4.0	12.0	24.0	56.0
Recency	10000.0	NaN	NaN	NaN	62.4068	69.874255	0.0	26.0	52.0	78.25	549.0
Monetary	10000.0	NaN	NaN	NaN	622.5552	647.135323	6.0	63.0	383.0	1077.0	3052.0
LTV	10000.0	NaN	NaN	NaN	209.0712	291.98604	-178.0	-2.0	57.0	364.0	1791.0
Perdeal	10000.0	NaN	NaN	NaN	32.3972	27.897094	0.0	6.0	25.0	56.0	97.0
Dryred	10000.0	NaN	NaN	NaN	50.3827	23.453815	1.0	32.0	51.0	69.0	99.0
Sweetred	10000.0	NaN	NaN	NaN	7.0545	7.866544	0.0	2.0	4.0	10.0	75.0
Drywh	10000.0	NaN	NaN	NaN	28.5213	12.583957	1.0	19.0	28.0	37.0	74.0
Sweetwh	10000.0	NaN	NaN	NaN	7.0698	8.015083	0.0	2.0	4.0	10.0	62.0
Dessert	10000.0	NaN	NaN	NaN	6.9474	7.879546	0.0	2.0	4.0	9.0	77.0
Exotic	10000.0	NaN	NaN	NaN	16.5466	17.247672	0.0	4.0	10.0	23.0	96.0
WebPurchase	10000.0	NaN	NaN	NaN	42.3762	18.522062	4.0	28.0	45.0	57.0	88.0
WebVisit	10000.0	NaN	NaN	NaN	5.2166	2.330457	0.0	3.0	6.0	7.0	10.0
Access	10000.0	NaN	NaN	NaN	0.246	0.539178	0.0	0.0	0.0	0.0	3.0