

**BUSINESS CASES WITH DATA SCIENCE**

**MASTER’S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS**

**Business Case 1 – *Wonderful Wines of the World***

Group C

Celso Christiano Endres Neto, m20200739.

Gabriel Felipe Martins de Souza, m20210598.

Luis Humberto Polaro Vizeu, m20210554.

Rogério Domingos Paulo, m20210597.

March, 2022

TABLE OF coNtenTs

[1. INTRODUCTION 1](#_Toc97055741)

[2. BUSINESS UNDERSTANDING 1](#_Toc97055742)

[2.1. Business Objectives 1](#_Toc97055743)

[2.2. Business Success criteria 1](#_Toc97055744)

[2.3. Determine Data Mining view 1](#_Toc97055745)

[3. ANALYTICS PROCESS 1](#_Toc97055746)

[3.1. Data understanding 1](#_Toc97055747)

[3.1.1 Data Visualization 2](#_Toc97055748)

[3.2. Data preparation 2](#_Toc97055749)

[3.2.1. Missing Values 2](#_Toc97055750)

[3.2.2. Incoherencies 2](#_Toc97055751)

[3.3. Data pre-processing 3](#_Toc97055752)

[3.3.1. Feature Engineering 3](#_Toc97055753)

[3.3.2. Input Space Reduction 3](#_Toc97055754)

[3.3.3. Data Standardization 3](#_Toc97055755)

[4. MODELING 3](#_Toc97055756)

[5. EVALUATION 3](#_Toc97055757)

[5.1. Customer Value Clustering 3](#_Toc97055758)

[5.2. Customer Behavior Clustering 4](#_Toc97055759)

[5.3. Merging Perspective 4](#_Toc97055760)

[5.3.1. RFM Analysis 6](#_Toc97055761)

[6. DEPLOYMENT (Marketing Actions) 6](#_Toc97055762)

[7. CONCLUSIONS 6](#_Toc97055763)

[7.1. Marketing Recommendations 6](#_Toc97055764)

[8. APPENDIX 8](#_Toc97055765)

[**Table 1.** Data Mining Goals and Success Criteria. 1](#_Toc97053434)

[**Table 2.** Clustering solution 5](#_Toc97053435)

[**Figure 13.** Clustering Solution 5](#_Toc97053436)

[**Table 3.** RFM Analysis 6](#_Toc97053437)

[**Table 4.** Marketing Actions 6](#_Toc97053438)

**NOTE: APPENDIX DOES NOT ACCOUNT FOR THE DOCUMENTATION**

# INTRODUCTION

The project aims to analyze a 7-year-old enterprise named Wonderful Wines of the World (WWW) that seeks out small, unique wineries around the world and brings their wines to its customers with the mission of delight with well-made, unique, and interesting wines that would never travel far beyond their points of origin. WWW sells wine through three channels: catalogs, web site and physical stores (10 small stores) and the purchase can be done in the physical stores, telephone or online. Despite the company recently organized a marketing activity which aggressively increased the data volume and also at this moment they keep sending to its customers a newsletter with the updates of wine, the WWW is trying to make use of the database started about 4 years ago because there are no loyalty programs or attempts to identify target markets for cross-selling opportunities.

This project was developed with a sample of 10.000 customers provided by WWW company containing information of the customers who have purchased something in the past 18 months. The report was adjusted based in CRISP-DM methodology (Pete Chapman, 1999).

The GitHub repository where all the present analysis is saved can be accessed through the following link:

# BUSINESS UNDERSTANDING

It’s important at this stage to defined the essential business guidelines to grant a good result of the project and provide the best solution to the WWW. In this regard, the business understanding was based on the current reality of the company presented at the introduction.

## Business Objectives

For the enterprise WWW, the main objectives are below enrolled:

* Use of customer dataset to get familiarity and ability to classify they customers in order to develop marketing strategies by profiling.
* Customer segmentation and track of new customer to allocate on the respect segment.
* Understand the customer value and his importance.

## Business Success criteria

Two main results were defined to guarantee the success of this project which are related to: *a) identify the profile of the customers since the first purchase*; and *b) develop marketing strategies to reach all market segments in order to improve the trading profit*.

## Determine Data Mining view

In this step, based on the business objective above mentioned, below we will describe the intended outputs of the project that enable the achievement of the business goals (Table 1).

**Table 1.** Data Mining Goals and Success Criteria.

|  |  |
| --- | --- |
| **Business Goal** | **Data Mining Goal** |
| Customer profiling | Customer Segmentation |
| Ranking the clients to understand the ROI | Apply the Recency, Frequency and Monetary value (RFM) |

# ANALYTICS PROCESS

## Data understanding

In the initial analyses performed, we can notice that our dataset present 18 columns (variables) with all of them represented by numeric features and 10000 observations (entries). See (**Figure 1 Appendix**).

## Data Visualization

In order to better address the aspects of the data distribution of our dataset, below are some important aspect noticed:

* Analyzing the histograms, we can see that the data distribution does not present large deviations in most of the variables (staying close to the average), except for the variable Recency, which shows some cases further away from the average; (**Figure 2 Appendix**)
* Using the Boxplots visualization, we were able to confirm that the data distribution on most variables is close to normal, again with a few cases further away from the mean and median, and which should be treated with a simpler method for removing outliers. (**Figure 3 Appendix**)
* And finally, using the correlation matrix and a pairwise scatterplot we were able to detect a relationship between some variables, some of which are worth mentioning:
* Freq has a strong positive correlation with Age, Income, Monetary and LTV, on the other hand strong negative correlation with Perdeal, WebPurchase and Webvisit, which is even curious, and seems to mean that customers who buy online have a lower purchase frequency compared to other channels. (**Figure 4 Appendix**);
* Perdeal (which is the % of discounted purchases) has strong negative correlation with Age, Income, Freq, Monetary, LTV, i.e. customers with higher income and age, and also who are more frequent, usually do not have a high % discount on their purchases. On the other hand, Perdeal has strong and positive correlation with online shopping, which shows the numbers of WebPurchase and WebVisit. (**Figure 4 Appendix**);
* LTV has the strongest correlations with Monetary and Freq, as well as other important ones like Age and Income, and what is apparently the biggest detractor from this metric are purchases made at discounted prices and in online media, Perdeal, WebPurchase and Webvisit. (**Figure 4 Appendix**);
* People who buy dryred wine generally do not buy other types of wine to any great extent, and the reverse is also true. And according to the analysis, Dryred wine has a positive moderate correlation with people with more years of education (Edu). (**Figure 4 Appendix**);

## Data preparation

Now we are going to describe the principal activities in order to obtain the final dataset used in the project.

In order to prepare our dataset to make use in the further analyses once we detect the presence of outliers in some of the features in our dataset (**Figure 3 Appendix**), and in order to remove them, firstly was applied the IQR method where we figure out that the amount of observations that should be removed were high and nevertheless, for the purpose of the analysis should be kept the most of the data we could. Therefore, was decided on using a method to filter the outliers manually and to adjust the observations’ threshold according to each variable. (Data Preparation, (Notebook BC1\_GroupC, 2022)).

### Missing Values

No missing Values were detected in the dataset.

### Incoherencies

No incoherencies were detected in the dataset in the scaling of the numerical variables.

## Data pre-processing

### Feature Engineering

In this step were created new features such as: avg\_ticket (Average order value), Redwines (%purchases of any red wine), Whitewines (%purchases of any white wine), Drywines (%purchases of any dry wine), and Sweetwines (%purchases of any sweet). (Data Pre-Processing, (Notebook BC1\_GroupC, 2022)).

### Input Space Reduction

In order to avoid including in the analyze features that present the same information (redundant) but include relevant features to perform the analyzes, through Correlation Matrix (**Figure 5 Appendix**) was noticed that:

* LTV, Freq, Monetary and avg\_ticket present positive high correlation, and WebPurchase with WebVisit also present positive high correlation, probably would add the same information to the model.
* The wines preference seems to be very spread and when using engineered features we end up with a high redundant scenario.

Therefore, for the purpose of the analyzes was performed using one of them and not all of this feature.

### Data Standardization

The data standardization method applied to make sure that the dataset presents the same scale was Z-Score Method. (Data Preparation, (Notebook BC1\_GroupC, 2022))

# MODELING

In the following section, after all steps performed in the previous stages 3 subsets was created from dataset namely:

* Value & Demographic (Age, Edu, Income, Dayswus, LTV, Perdeal, WebPurchase);
* Purchase Behavior (Dryred, Sweetred, Drywh, Sweetwh, Dessert, Exotic); and
* RFM (Recency, Freq, Monetary).

In the process of clustering, two final clusters perspectives were created (**Customer Value & Demographic** and **Customer Behavior**) and among the other cluster’s methods such as (K-means, Complete, Average and Single) the K-means was chosen for clustering of both groups after comparing R2. (**Figure 6 Appendix)** and the number of 3 clusters was decided based on the elbow method. (**Figure 7 Appendix)**. In order to merge the two clusters perspectives, the method performed was Hierarchical Clustering – Wards Dendrogram were the ideal number of clusters suggested was between 2 and 4, however was decided to use 4 clusters. (**Figure 8 Appendix)**

# EVALUATION

At this stage, the two clusters performed are evaluated to ensure the best quality that can achieve the objectives of the WWW enterprise.

## Customer Value Clustering

To construct the customer Value Segment, were considered the variable LTV (Lifetime value of the customer) due to economic importance and after been performed several tests the interpretation of the clusters was not enough to raise important insights (**Figure 9 Appendix**), therefore was considered to perform clustering based on the monetary value and demographic features (**Figure 10 Appendix**) were the result are shown and in the below are insight:

* There were clearly two super opposite clusters in terms of value, which are cluster 0 and cluster 1, blue and orange respectively;
* Cluster 0 (**Class C**): Has the lowest LTV (value), demographically also has the lowest average Age and Income. Has preference for online shopping (WebPurchase) and usually only buys when there is a discount. Represents 43%.
* Cluster 1 (**Class A**): This is the profile with the highest value in the database (highest LTV). This customer profile generally has a higher average Age and Income when compared to other clusters, and is not very adherent to online shopping and does not look for discounts on most of their purchases. It is a profile that may have greater difficulty in engaging with online shopping. However, it may also be interesting to create an exclusive catalog for this group of customers, especially with a strategy focused on retention, since they are the customers who bring the most money to the company. Represents 25%.
* Cluster 2 (**Class B**): Is very close to the average in all aspects. It has an average Age and Income higher than Cluster 0, but apparently the impact on its value (LTV) is more impaired due to online and discounted purchases. An interesting approach for this group may be to work with campaigns to make the customer more aware of other purchase channels, or even create other types of advantages that reduce the need for this customer to search for discounts on their purchases and also could be interesting create an exclusive wine catalog for this profile and test whether customers see the advantage to the point of giving up larger discounts. Represents 32%.
* Adding Freq, Monetary and Recency to the table, we can see that the difference amongst the group of customers is the same. For the cluster 0, the lower LTV is also explained by the low Frequency and Monetary. In the other hand, Cluster 1, which is the most valuable, presents the higher Frequency and Monetary, and consequently its LTV is the highest compared to the other groups. (**Figure 11 Appendix**)

## Customer Behavior Clustering

By using Purchase Behavior Features, were found three heterogeneous clusters of customers. (**Figure 12 Appendix**):

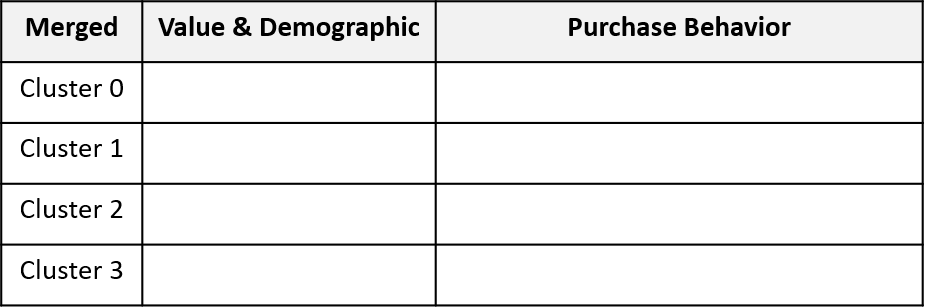
* Cluster 0 (**Dry White**): Represents 42% of customer base. There is a preference for dry white wines.
* Cluster 1 (**Sweet & Exotic**): Represents 15% of customer base. There is a preference for sweet and dessert wines as well as exotic wines.
* Cluster 2 (**Dry Red**): Represents 43% of customer base. There is a preference for dry red wines.

## Merging Perspective

In this step the two perspectives clusters were merged in order to present the final solution to support the business objective and below are the insights presented:

* After merging the clusters, 4-clusters solution was performed revealing a R² of 0.41, which means that about 42% of the variation is explained by the clusters solution in the dataset and this is the final result when combined with the Clustering model 1 where was applied features related to the Value of the Customer (especially LTV) and model 2 based on behavioural features such as the preferences for wines.
* Cluster 0: The largest group (Blue), with approximately 50% of the clients. Age, Income and Education are almost at the same level of the database's average. This profile of customer has a strict preference for dry red wines (Dryred) and buys the other types of wine in lower proportion. Like the Cluster 2 customers, they are heavy Web buyers, demonstrating high preference for online purchases (WebPurchase) and generally the purchases have discounts. This is the second group of customers with the worst LTV. As it represents 50% of the customers, it is extremely important to make sure that we have a specific strategy to turn these customers more profitable. To do that, could be created an exclusive offer of Dry Red selected wines only available in catalogs, or yet, available on the Website but with discounts just in the first purchase.
* Cluster 1: With around 12% of the customers, this profile has Age and Income above the average, and LTV exactly the average of the base (approximately 200), which means that are customers profitable. This group of clients is almost eclectic, but still prefers in average more sweet wines than dry ones. As the incomes are above the average, maybe there's a chance to increase their LTV by reducing the discounts (Perdeal).
* Cluster 2: This group represents around 13% of the customers, and so far, it is the worst customers in terms of value (LTV). They have the lowest Age and Income, and are very focused in Online purchases with a high discount proportion. Still, they really are passionate for sugar, demonstrating a high preference for sweet wines, whether red or white. They are also the nº1 in Dessert and Exotic wines.
* Cluster 3: This group is gold! Representing 25% of the customers, this is the most valuable group in the database, as it has the highest LTV. With also the highest Age and Income average and there’re almost eclectic too, except for their higher preference for Dry White wines (Drywh) and lower appetite for Exotic ones. Most of their purchases are made through the catalog and the stores, since their Web Purchase numbers are very low, still they generally never ask for discount when shopping. A good strategy is to make sure that can be kept in the business, then could be created a premium experience with not only exclusive catalogs, but also physical experiences in the stores to make these customers ambassadors of the brand and guarantee their retention.

**Table 2.** Clustering solution



**Class B**

**Class C**

**Class C**

**Class A**

**Sweet & Exotic**

**Dry White**

**Dry Red**

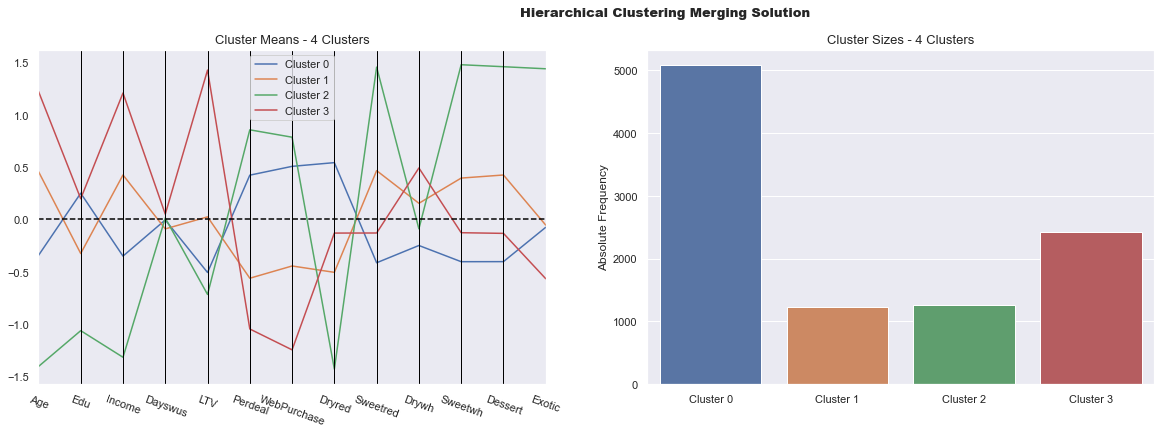
**Sweet & Exotic**

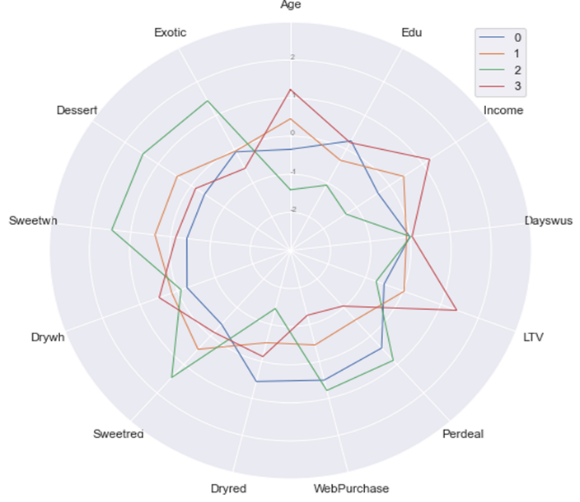
**Dry White**

**Sweet & Exotic**

**Dry White**

**Dry Red**

**Figure 13.** Clustering Solution



**25%**

**12%**

**11%**

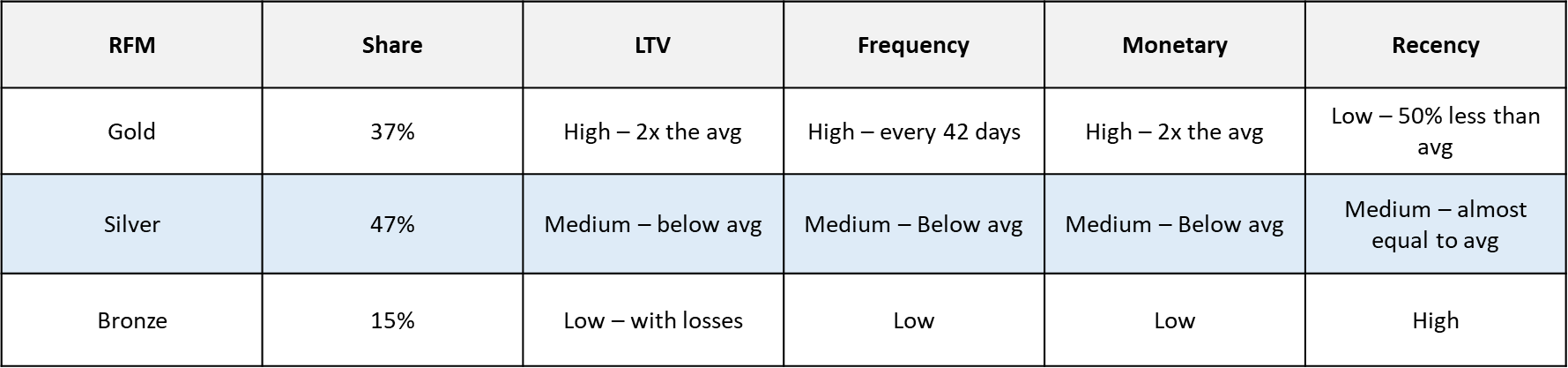
**52%**

### RFM Analysis

From the RFM Level segmentation the groups were described as following:

* **GOLD**: Around 37% of the customers, most valuable group in terms of LTV (average LTV of 488). These customers have the highest average frequency (27.2) twice the average of the base, they return to shopping every 42 days (50% less time compared to the average), and finally, they spend 2 times more money than the average, 6x more than the Silver and around 37x more compared to the Bronze Customers.
* **SILVER**: 47% of the customers lies on this group. In average this group is still profitable, but its LTV of 57.0 is below the average (which is ~208). These customers also return to shopping in less days compared to the average (49.8 versus 61.9), but both its Frequency and Monetary are lower than the average, at least 1.6X and 2.2X, respectively.
* **BRONZE**: 15% of the clients, this group is the worst in terms of LTV (-2.1), which means that these customers at the end of the day cause losses to the business. With a Frequency ~7 times lower than the average and the Monetary in average 17 times below the overall average, these customers probably need to be reactivated, since their Recency is around 150 days.

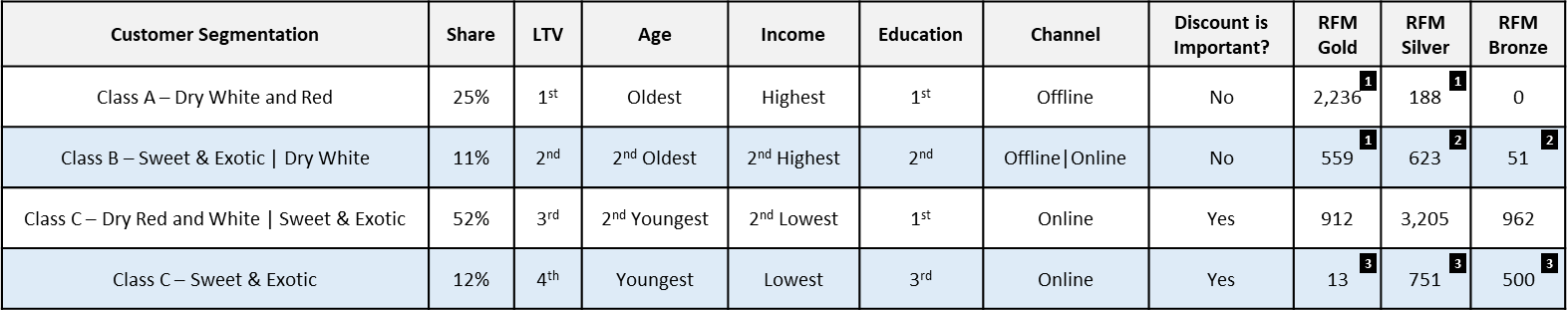
**Table 3.** RFM Analysis



# DEPLOYMENT (Marketing Actions)

Based on the clusters that were found as a result of the segmentation work and in order to achieve the project’s objectives, the team recommends the deployment of the following X marketing actions.

**Table 4.** Marketing Actions



# CONCLUSIONS

## Marketing Recommendations

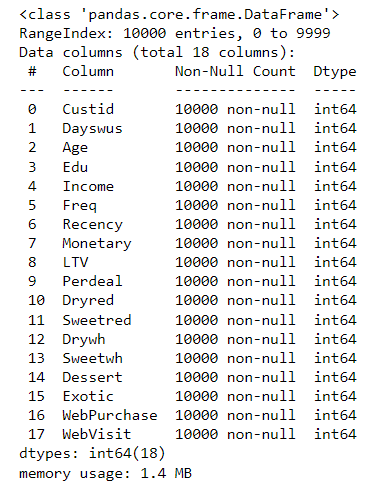
**Action 1**: Design of a loyalty program in order to keep this customer active as longer as possible. The program can be in a format of a wine club, in which a monthly fee is paid (compatible to the current ticket) and a wine basket (compatible to the average quantity) is sent once a month. The wines can be picked based on the customers preferences (in this case dry red and dry white, and more expensive/exclusive wines – also good for inventory management). In case a customer is willing to buy more bottles from a wine that was sent in a previous basket, they can order and it will be delivered with the next month basket. Once every X month the customer can choose a friend to send a similar basket as the one they are receiving for that specific month.

**Action 2**: Online campaign, focusing on offering the specific types of wine which are the preference of this customer segmentation. The objective here is to reduce recency, because that could force a migration to a better customer segmentation (from silver to gold, for example). The advertisement could be sent x days after the last purchase with recommendations of wines that were purchased by other customers from this same cluster.

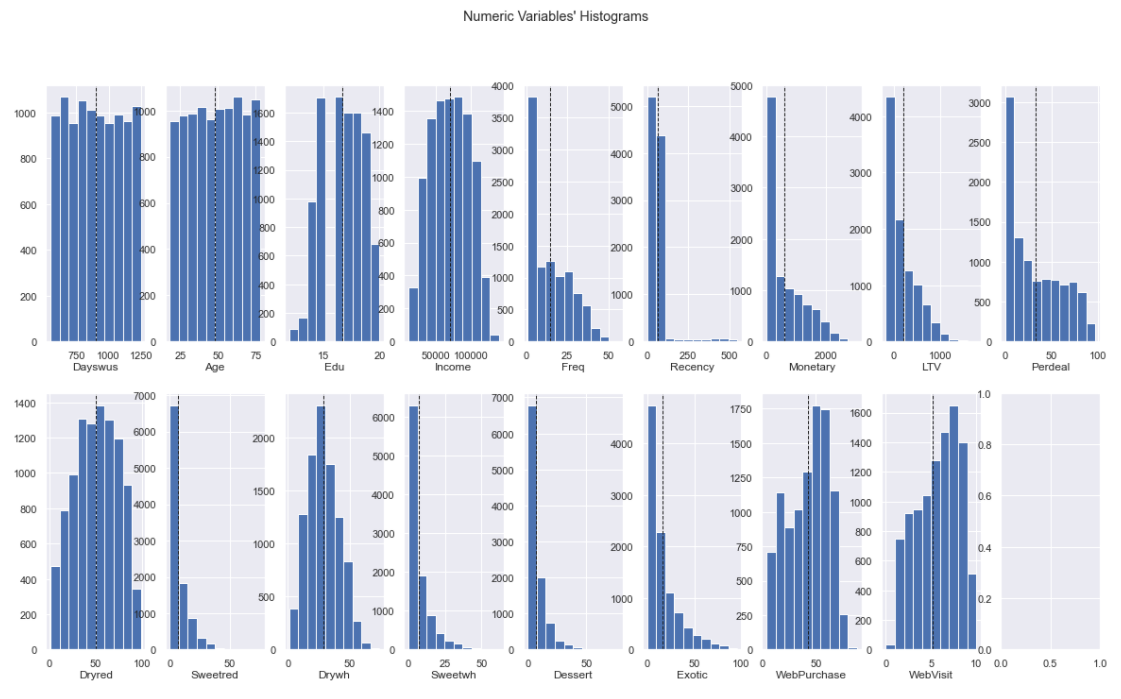
**Action 3**: Online mass campaign targeting the same customer profile as Class C and prioritizing to advertise wines that are not selling well, aiming to inducing inventory turnover by offering great discounts. All types of wine can be covered in this action.

# APPENDIX

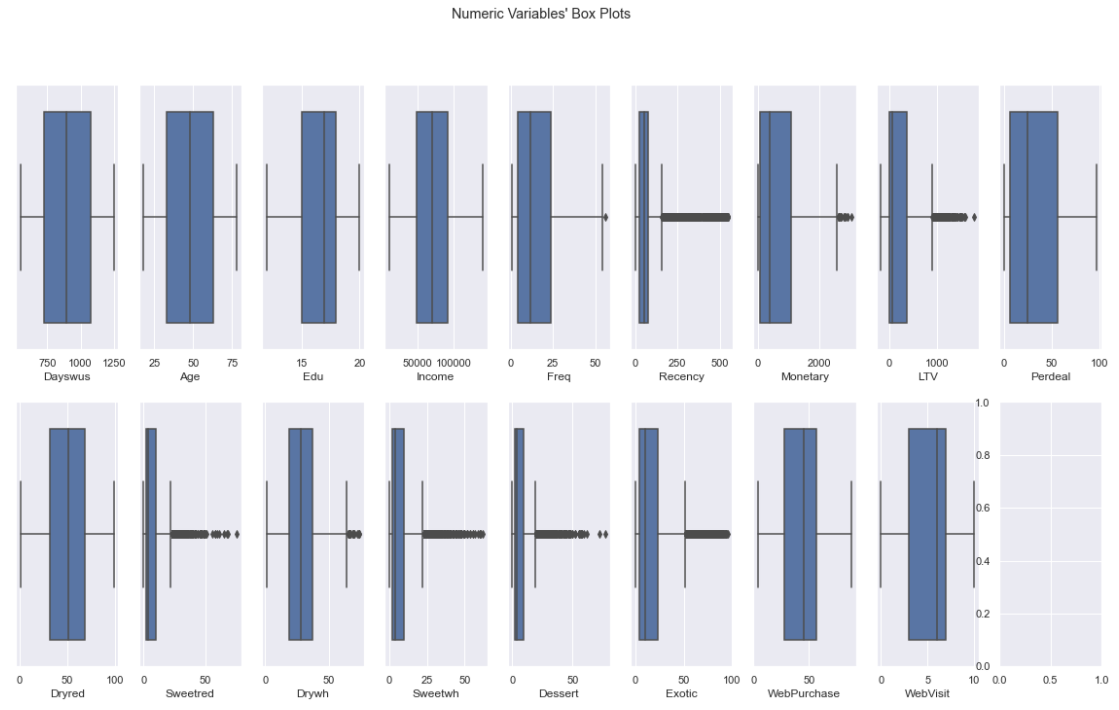
**Figure 1: Variable Information**



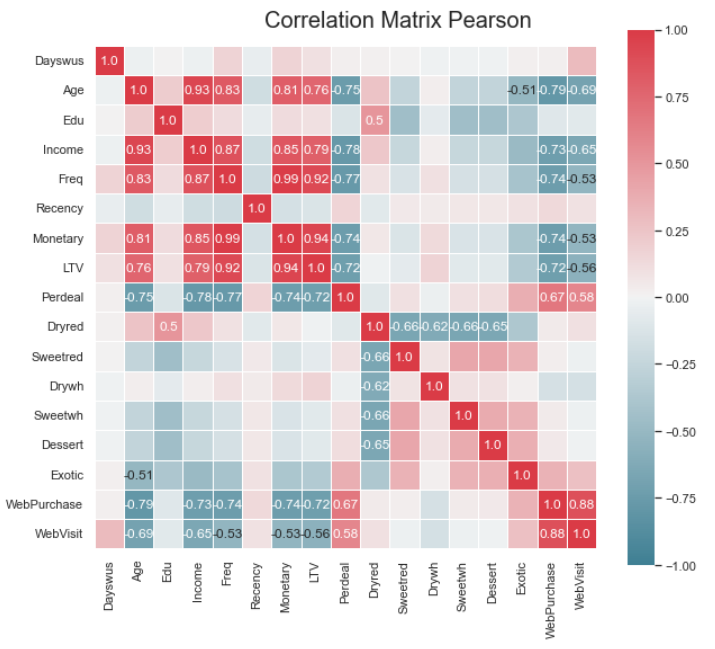
**Figure 2: Histogram**



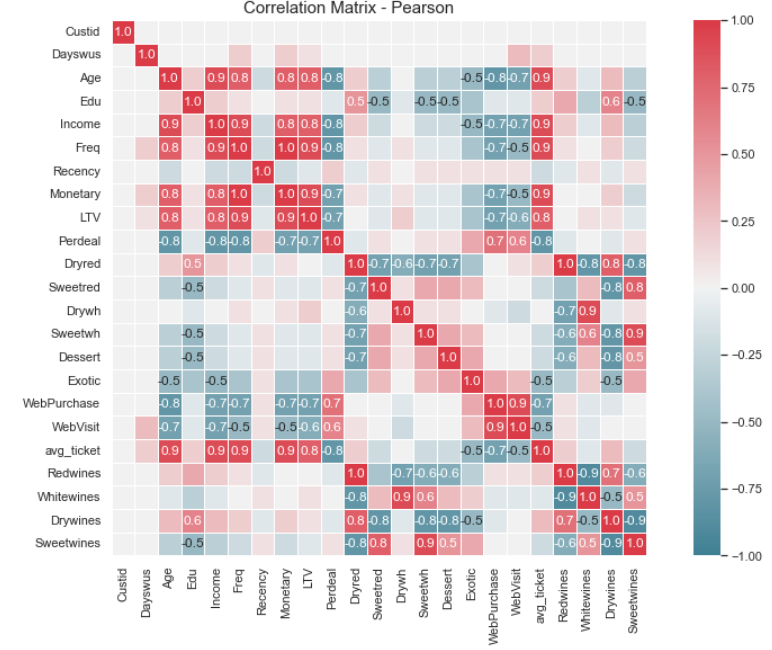
**Figure 3: Box-Plot**



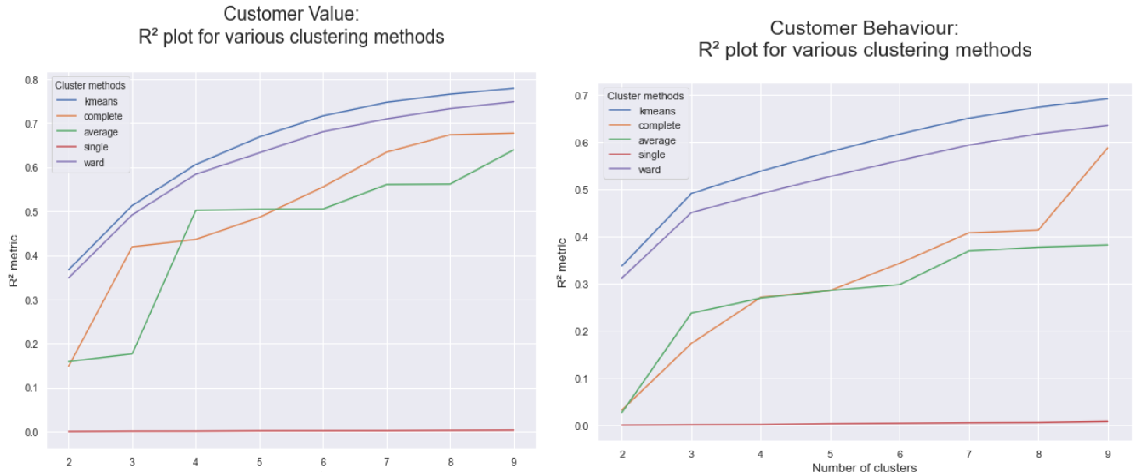
**Figure 4: Correlation Matrix**



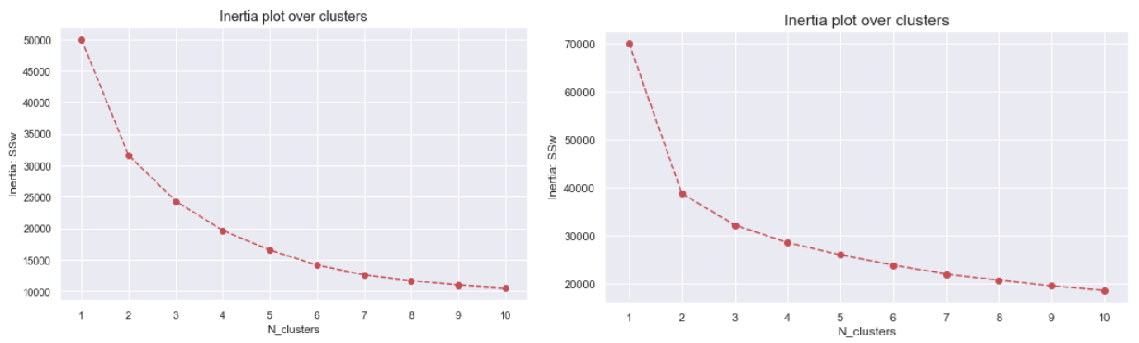
**Figure 5: Correlation Matrix (Redundancy x Relevance)**



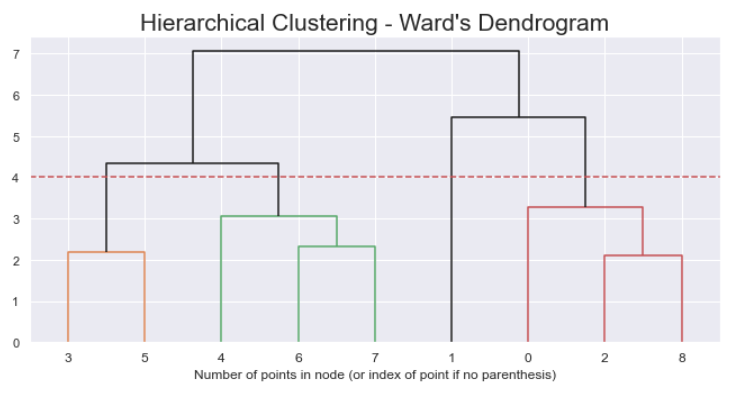
**Figure 6: Customer Value and Customer Behaviour clusters**



**Figure 7: Inertia Plot over Customer Value and Customer Behaviour clusters**



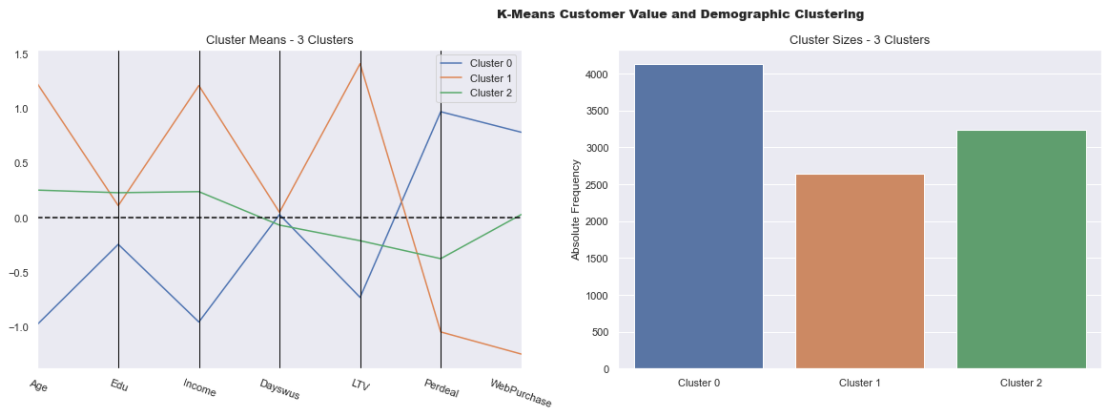
**Figure 8: Hierarchical Clustering – Ward’s Dendrogram**



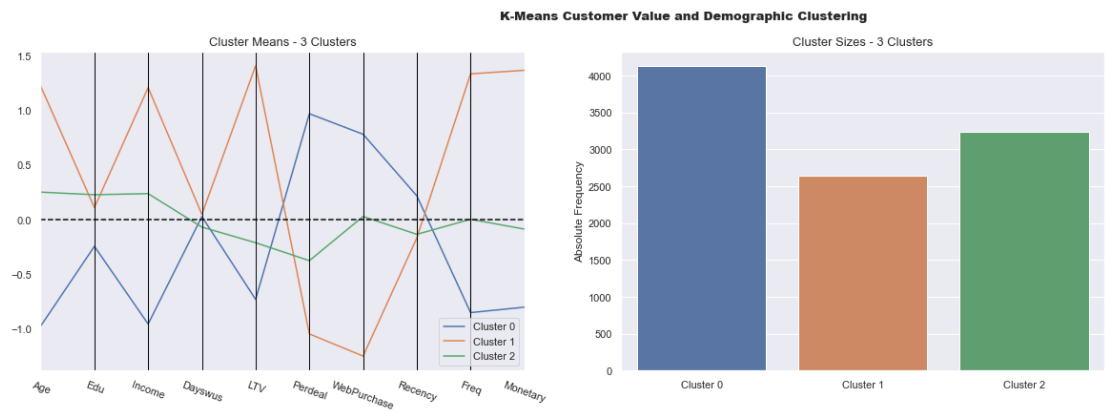
**Figure 9: Customer Value Clustering**



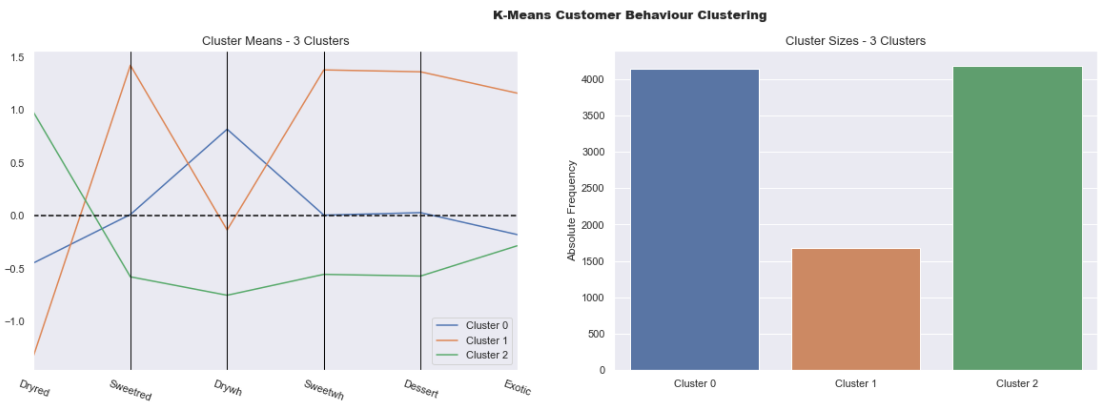
**Figure 10: Customer Value and Demographic Clustering**



**Figure 11: Customer Value and Demographic Clustering (adding Freq, Monetary and Recency)**



**Figure 12: Customer Behaviour Clustering**



**Figure 14: Merging Perspective**

