BUSINESS CASES FOR DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Business Cases 1 – Wonderful World of Wines
Customer Segmentation

Group R

Andreia Bastos, number: 20210604

João Silva, number: 20211014

Pauline Richard, number: 20211019

Tiago Quaresma, number: 20210766

March, 2022

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

Index of Figures

1. INTRODUCTION

The wine company, Wonderful Wine of The World (simplified to WWW), is a recent company selling its products in USA, including wines and derivative accessories. Its client's database has been created 4 years ago and gathered 350,000 people, buying through the physical stores, the website or via a phone call (by looking at the catalog). WWW is for now using only mass-marketing strategy.

1.1. Business Objectives

The company wants now to take full advantage of its database by using it as an accurate segmentation tool and offer more focused marketing programs. Through a clustering approach, we will help WWW to identify the values and purchasing behaviors of their customers.

1.2. Assess Situation

For this project, we are provided with a sample of 10,000 customers from the active database, meaning customers that have been active in the last 18 months. The dataset *WonderfulWinesoftheWorld.xlsx* contains 18 numerical variables, ordered by the ID of the customers. All customers have the minimum required age of 18 years old.

2. PREDICTIVE ANALYTICS PROCESS

2.1. Data understanding

For starters, to obtain information about the Data Frame, we used the *info* command. This showed us that all eighteen variables were stored as floats and there was only one missing value in the 'Custid' column. After a quick analysis, we noticed that the correspondent row was the average of each column. Since we had no use for it, we ended up dropping it. Afterwards, we checked for duplicates but found none.

2.2. Data Visualization

In this step, we computed a few graphics to identify patterns in the different features. We checked the pairwise relationships (**Figure 1** of the appendix) between the numerical variables with a scatterplot. We verified that there were no linear relationships in data which of course would later influence the choice of our correlation coefficient method. We also used boxplots (**Figure 2** of the appendix) to visualize outliers, where it is possible to see that are few outliers to be dealt with, apart from the wine range and *Frequency* variables.

.

According to the table above, some of the outliers are gone, while others are still there. Notice that *FirstPolYear* maximum and the *BirthYear* minimum were solved but, we still have a huge gap between 0.75 and the max on the *ClaimsRate* column. Other big gaps indicate the existence of outliers, so we moved on to that step.

2.3.Data Preparation

Outliers Detection

To detect outliers we applied two methods, the IQR (Interquartile range) and Density-based clustering algorithm (DBSCAN). When preforming the IQR, the percentage of data kept after removing outliers was 76.89%, whereas when we applied the DBSCAN algorithm (with the parameters MinPts and epsilon defined as 36 and 3) the percentage of data kept was 98.83%. DBSCAN was chosen for two reasons: first, it is a global approach compared to IQR (the IQR selects each variabel's outliers); also, for structure purpose we aim to remove less than 3% of our initial dataset. (See **Figure 3** on the appendix)

Variables' Correlation

As mentioned before, the relationship between variables is not linear and therefore we can not use Pearson's correlation coefficient. Hence, we decided to use the Kendall and Spearman correlation, but ended up choosing the Kendall correlation since it is more robust and efficient than spearman (smaller gross error sensivity (GES) and smaller asymptotic variance (AV)).
As seen in the heatmap (**Figure 4** of the appendix) we can see that we have a considerable amount of higly correlated variables and decided to drop the following:

- ***Income*** - highly correlated with Age
- ***WebVisit*** - highly correlated with WebPurchase
- ***LTV*** - highly correlated with Monetary and Frequency (we will only use it to perform the RFM analysis)
- ***Recency*** - *not* relevant for the segmentation

However, we will keep the other correlated variables such as ***Monetary*** or ***Perdeal*** since they provide important information for our segmentation.

Coherence Check

We ran a coherence check to make sure that the data is logically consistent and can be reliably used for our analysis. We noticed that almost all of the data seemed coherent except for the age of our younger customers. Some of them are 18 years old , however the legal age to drink in USA is 21 so they shouldn't be able to buy alcohol related products and appear in the database. However, this observation represents more than 3% of our dataset (around 400), so we cannot delete it and we are making the assumption that the legal drinking age is 18.

Scaling the Data

It is an important step to scale the dataset for models that rely on the distance between the different observations, and also to get an easier visualization. For this dataset we decided to use the MinMaxscaler [0,1]. Since we removed the outliers, the MinMaxScaler is a good option because it is highly influenced by the maximum and minimum values, therefore it is not going to be biased.

2.4. Modeling

Before starting the clustering, we applied the RFM analysis. We started by sorting the database according to recency and dividing it into 5 quintiles (5 equal segments). Then doing the same for the variable's frequency and monetary. Then we computed the *RFM_score,* which represents the segment's total sum for each customer. Afterwards, we classified the customers according to it. If the value is higher or equal to twelve, they are classified as 'Can't Lose Them' and Champions, respectively. If it is equal to eleven, ten or nine, then they belong with the 'Loyal/Committed', 'Potential' and 'Promising', respectively. If, however the RFM_score is equal or lower than eight (the worst

classification) they are classified as 'Requires Attention' and 'Demands Activation', respectively. (**Figure 5**)
Afterwards the Hierarchical Clustering was applied on RFM, resulting in a total of four clusters. However, when comparing this approach with our final clusters, it was a littlr different and the partition between the clusters wasn't very clear, therefore we did not pursue this any further.

Final Clustering
For this stage we chose 2 clustering algorithms, K-means and SOM, both combined with the Hierarchical Clustering.
Before applying the clustering algorihtms, we divided our dataset into 2 views:

- **Behaviour View,** composed of the profile and purchasing history of WWW' customers
- **Wine Type View**, gathering the product lines of WWW and consumption information

In order to understand the results given by each algorithm for the views, we used the function "r^2", which returns the R^2 metric for each clustering solution.

K-means
For this algorithm, the optimal number of clusters was chosen with the help of the elbow method(**Figures 6.1 and 7.1** on the appendix) and the dendogram(**Figures 6.2 and 7.2**) resulting from the hierarchical clustering. For the Customer's Behaviour view we obtained a 4-cluster solution, and a 3-cluster solution for the Wine Type view.

SOM
In both views we used a R2 plot for various hierarchical methods, complemented by a Ward's Dendogram, to get the most efficient number of clusters. The Customer's Behaviour view has an optimal number of 3 clusters as seen in **figures 8.1 and 8.2 of** the appendix. For the Wine Type view, we also obtained a 3-cluster solution that can be seen in **figures 9.1 and 9.2**
Finally, the result that gave us the most satisfactory clusters was **the Hierarchical Clustering with SOM (Self Organizing Map).**

3. RESULTS EVALUATION

Looking at the different views and the final merged clusters (with an optimal number of clusters of 3), we can identify 3 distinct customers' segmentation

Cluster 0: Economic wine Explorers (Demands Attention)
Customer Profile - The youngest customers profile (25-40 years old), with the lowest education years but still a good education level (an average of 16 years).
Purchasing history and Value - They have the lowest Monetary and Frequency values. It is the segmentation that shows the lowest customers' value for WWW but can be explained by the fact they are not consistent and regular in their purchase (1 purchase every 2 or 3 months only). However, they are still spending an average of 31 euros per purchase, and it is the segmentation that uses Web purchase options the most. Finally, this profile is the most sensitive to promotions' offer with Cluster 0 (highest perDeal), and it should be used for the marketing strategy.
Product Mix - Looking at the Wine Type view, we can see compared to the 2 other segmentations, it is the one most oriented to sweetened wines, which can be explained by the tendency of sweet wines to be cheaper than dry wines. However, like the 2 other clusters, it shows a slight preference for dry wines. They are also the ones more attracted to derived products and new types of wines, with the highest percentage in both exotic wines (25%) and dessert (12%).

Cluster 1: Champions (Loyal)
Customer profile - The oldest customers in terms of age (60-80 years old), with the longest days since being a customer, and a high education level.
Purchasing history and Value - Highest Monetary and Frequency values but lowest PerDeal rate. Meaning customers of this segmentation are the most profitable for the company, spending more money than the two other clusters (minimum of one purchase per month and at an average price of 49€). It can be explained since they come more frequently to the shops (loyal) and are buying full price (no interest in promotional deals). Most of the purchases are made offline (physical shops or via the catalog), which can be explained by the age range of the customers, favoring more traditional channels.
Product Mix - This segmentation is interested in buying mostly Dry wines. They are slightly more interested in Dry Red (45%) than Dry White (34%) and show occasional interest in sweet and exotic variety of wines.

Cluster 2: **Dry Red Lovers** (Promising)
Customer profile - Most educated customers with a middle age range (40-55 years old).
Purchasing history and Value - They have low Monetary and Frequency values but are customers showing more loyalty than the ones of Cluster 0, with an average of 1 purchase every 1 or 2 months. They have a high good PerDeal rate, meaning they are also sensitive to promotion deals and almost half of their expenses are done via the website (49,42%).
Product Mix - This segmentation distincts itself from cluster 0 in its consumption habit. They show no interest in sweet wines, mostly buy Dry Red Wines (70%) and some Dry White or Exotic Wines, for an average price of 36,5€ per purchase.

In the appendix we can see the final cluster distribution for each view (**Figures 10 and 11**) and also the final merged cluster solution in **Figure 12** – these visualizations were obtained using the *cluster_profiles* function from the Data Mining course.

To obtain a more detailed visualization of our final data distributions per cluster we also created a bar plot for each cluster's variables as seen in **Figure 13** of the appendix

4. CLUSTER ANALYSIS

Suggestion
For our marketing approach, we must take into consideration both aspects of the segmentation just mentioned above: the purchasing profile of our customer but especially its value for the company (their willingness to spend money on our products).
Thanks to the Clustering analysis, we evaluated the most profitable of the customers and we will advise prioritizing the marketing costs on those ones and minimizing the money spent on the others. We want to make sure that the money invested will be translated as a greater financial return for WWW.

Marketing Strategy
Segmentation 1: Economic wine Explorers
Considering this type of customer and all the factors in which it stands out, we can conclude that it is an adventurous profile, which is always looking for the best opportunity and is e-shopping friendly.
Therefore, our Marketing approach for this profile would be to promote the products mostly using online communication, with innovative suggestions and discount offers. We want to attract the customer into buying WWW's products rather than the competitor's (and therefore increase its Loyalty, Frequency and Monetary value).

Here are some ideas with a minimal marketing cost (since those clients are still the least profitable):

- Recipes on the website & giveaways for the best cocktails;
- Mystery box subscription program with varied selection of wines (monthly, bimonthly, quarterly);
- Discount Coupon for the next purchase with a validation date of 1 or 2 months;
- Social network ads as a communication channel.

Segmentation 2: Champions

For this segmentation we group the best client in terms of value and purchases. Our marketing strategy is aiming to retain and reward their loyalty by offering them a more Premium experience. It can be done through a "winery club membership" subscription program. It would be accessible starting from a minimum number of purchases of 14 in the last 18 months. As they are not regular internet users, the subscription should be done when the customer visits the shop with the help of a collaborator doing all the steps online.

It would give special benefits such as:

- Pre-sale for special edition wines;
- Tasting sessions. It would also be a good way to increase their purchases by advertising them other product ranges;
- Customization services such as for the label of their bottle for instance.
- Special service such as an attributed adviser in the shop or also to help them placing online orders
- Receiving a monthly pack of diversified brands and qualities, with a value fixed by the customer without having to place an order. At the end of a year's subscription, the customer receives a free order with the most popular wines of the year.

Segmentation 3: **Dry Red Lovers**

This cluster consists of people who buy less frequently, so cant yet be part of the membership club, but are promising and could be more easily loyal to us than the first segmentation. Moreover, we know they are online friendly, incentive to discount, and more selective regarding wines' quality (mostly purchase Dry Red Wines). Therefore, for the marketing strategy, we advise to:

- Create a Newsletter with content about wines to know, producers, grape varieties, recipes to accompany wines etc. It will allow WWW to create a closer relationship with those customers. The newsletter can also have links to the website about the current offer and sales, and thus encourage this cluster to shop online.
- Discount on the quantities of Dry Wines bought. For example, 1 Dry Red bought the second one has 20% off.
- Discounts for every pack of 12 or 24 - bigger the quantity bigger the discount.
- Partnership with museum, link to mini serie or documentaries etc that gives information about the provenance of our wines.

Communication channel to prioritize would be social networks, websites, and monthly newsletters.

5. CONCLUSIONS

When dealing with a project like ours, many approaches could have been taken. Regarding the outlier detection, even though we tried a univariate and a multivariate approach, we could have also worked with Support Vector Machines (using a one-class SVM). Besides the K-means and SOM, we could have applied the K-medoids algorithm (it is less sensitive to outliers), Hidden Markov Models or even GMM.

Regarding the dataset, although it was meant to be simple, we felt like we could use some extra or more detailed data, for example, more details regarding the purchasing history of each client as we feel it would lead us to better and more complete clusters and therefore better marketing suggestions

for the company. In that case the marketing strategy would be the development of an activation pack that offers better conditions or discounts to returners.
Even with room for improvement, the clusters and segmentation obtained with the chosen algorithms were satisfactory overall, leading us to a clear and succinct final segmentation.

6. REFERENCES

Croux, C. and Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. Statistical Methods and Applications, 19, 497-515.

7. APPENDIX

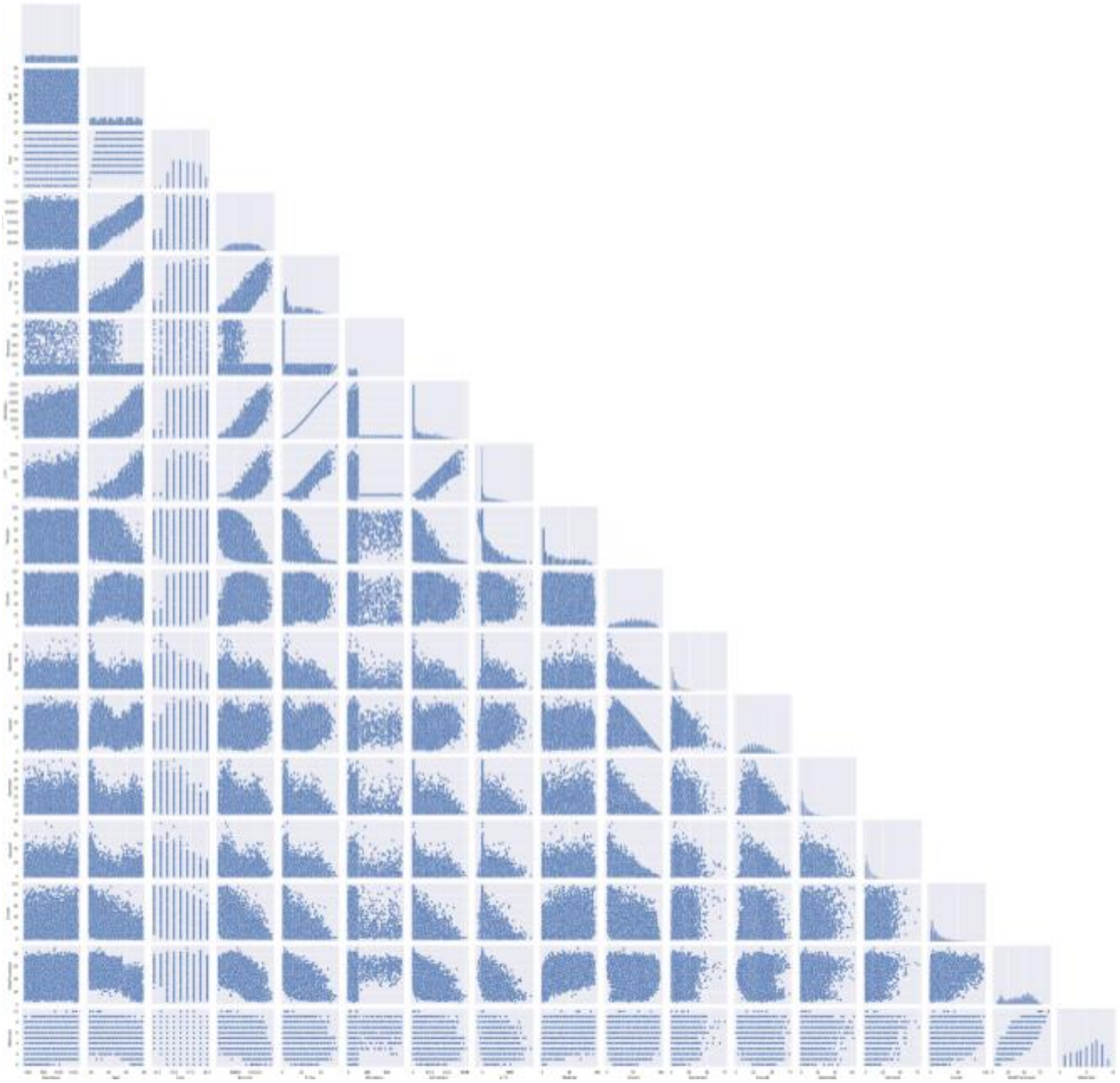


Figure 1: Pairwise Relationship of numerical variables .

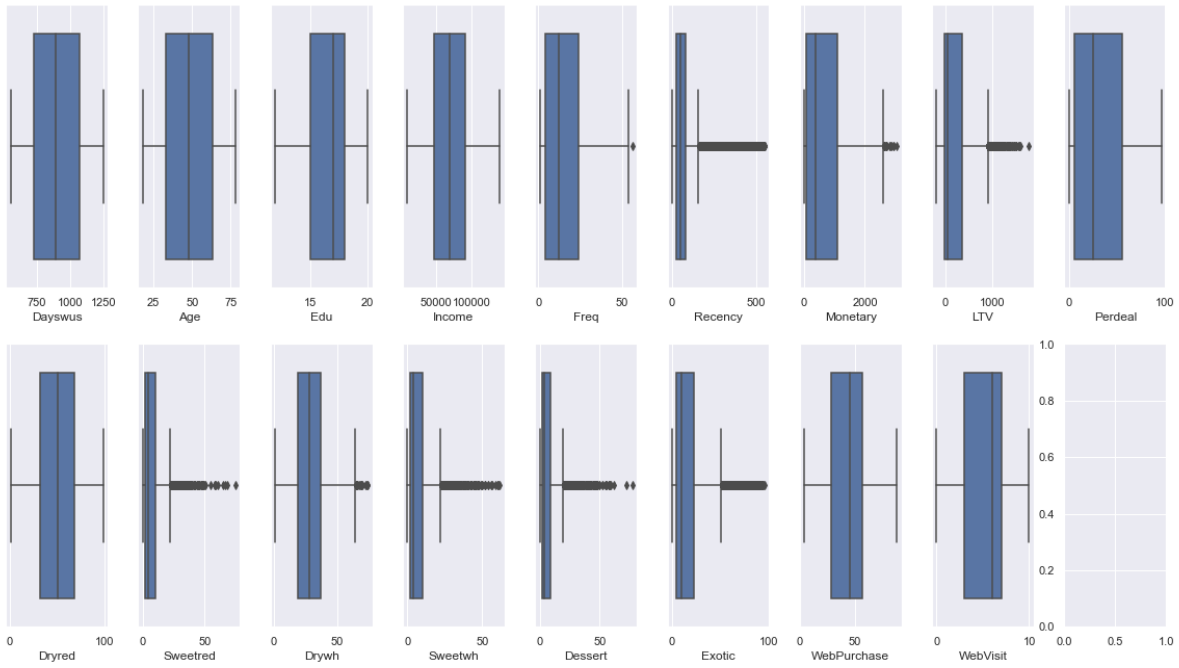


Figure 2: Numerical Variables' Box plots.

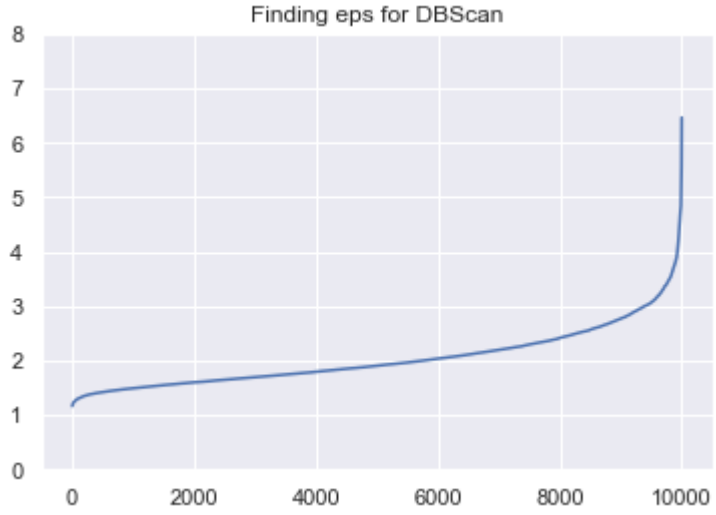


Figure 3: Determination of optimal value for Eps .

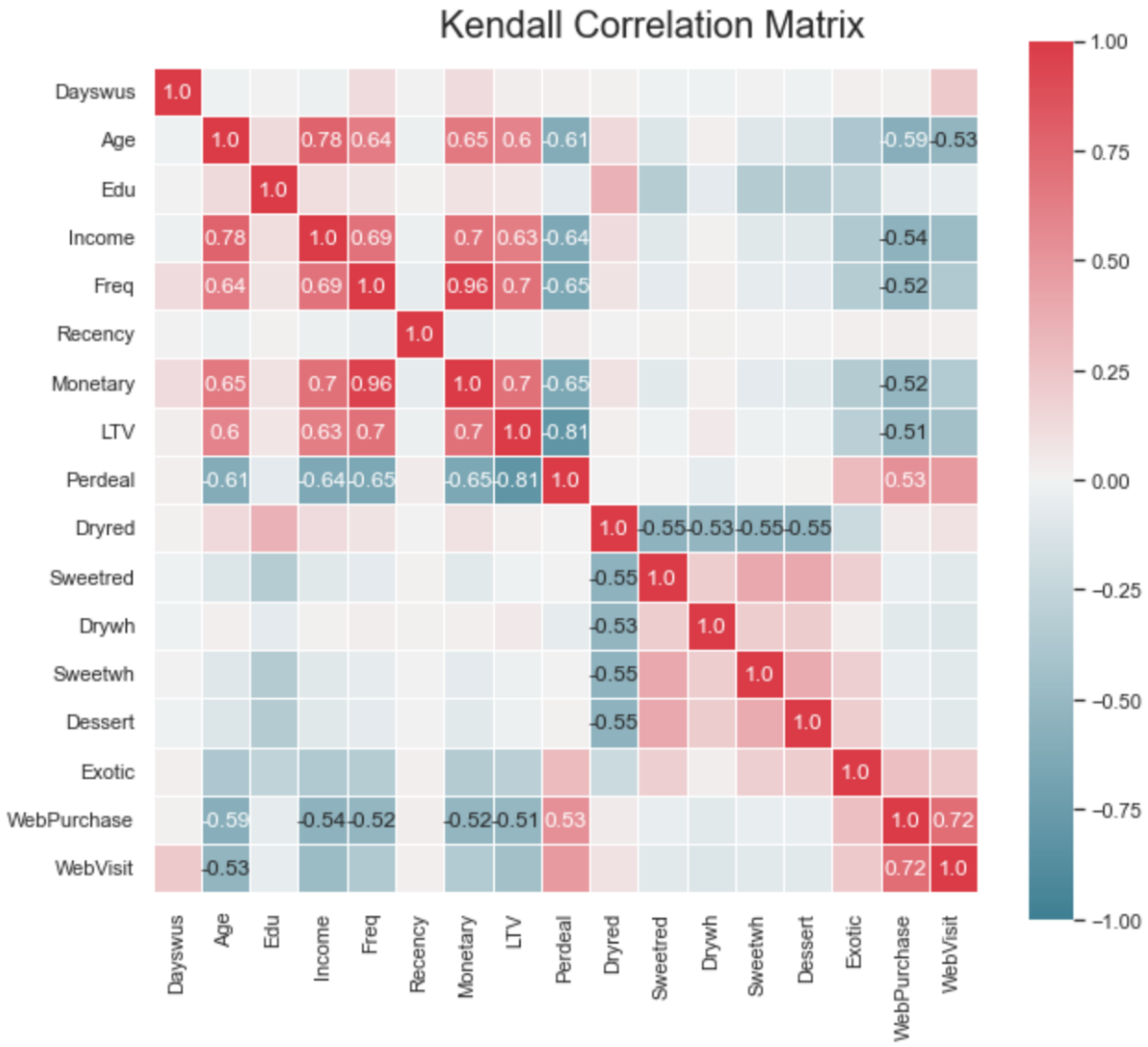


Figure 4: Kendall correlation Matrix.

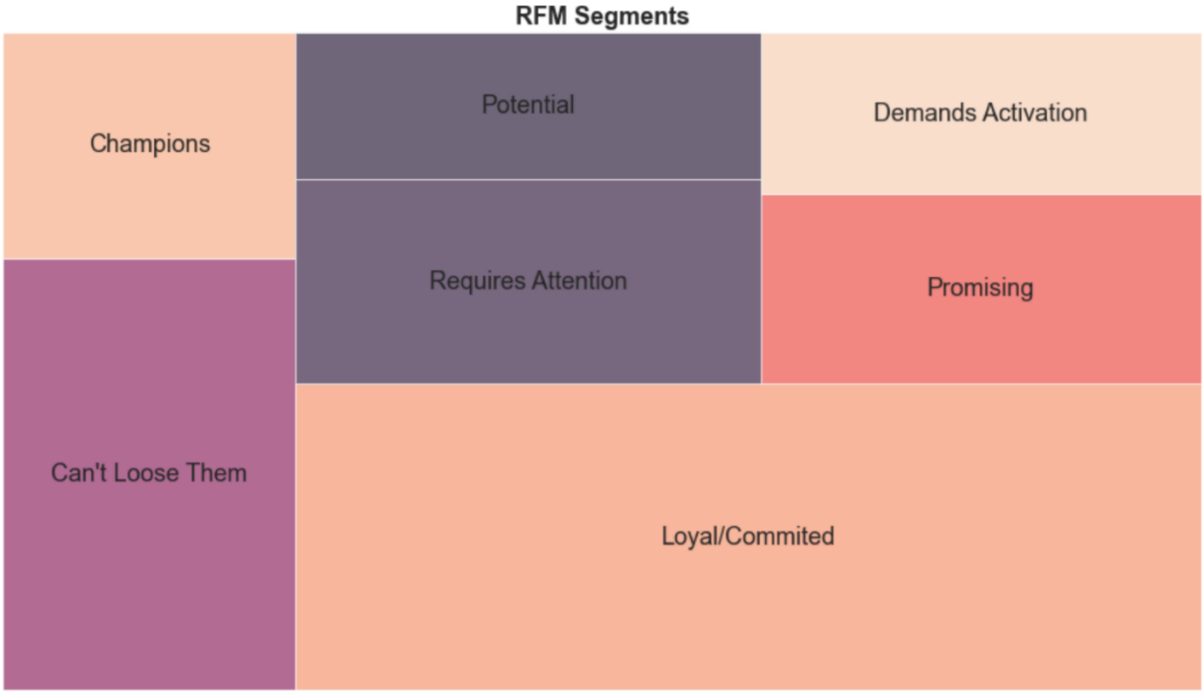


Figure 5: RFM final segmentation .

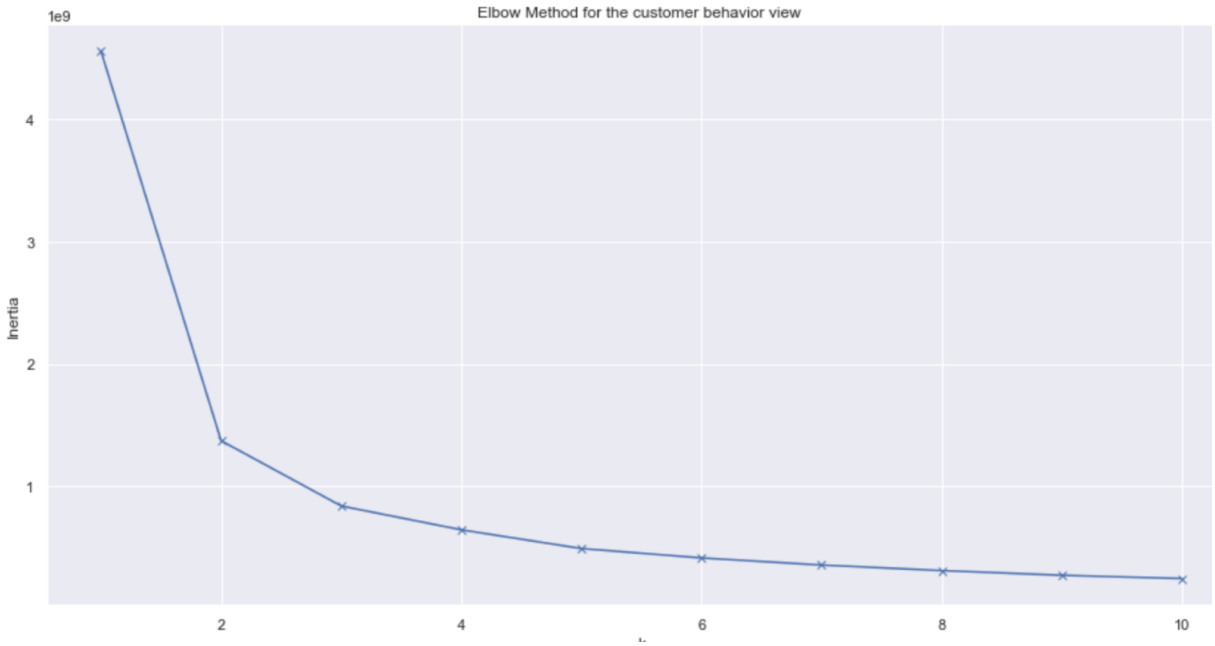


Figure 6.1: Best clustering solution.

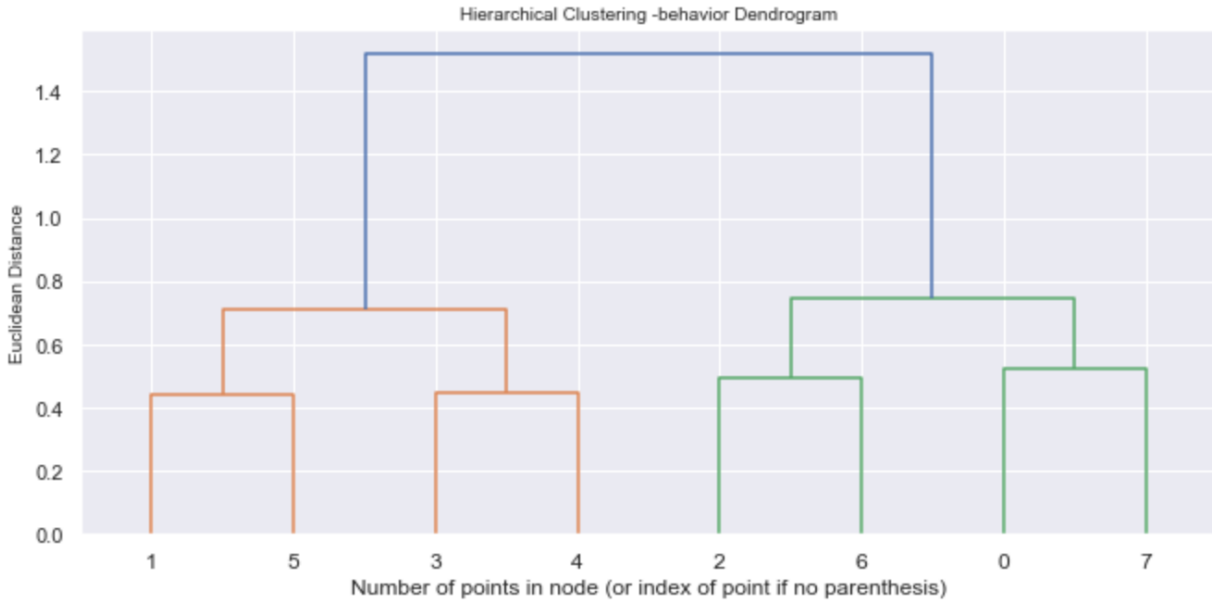


Figure 6.2: k-means dendogram for Behavior view.

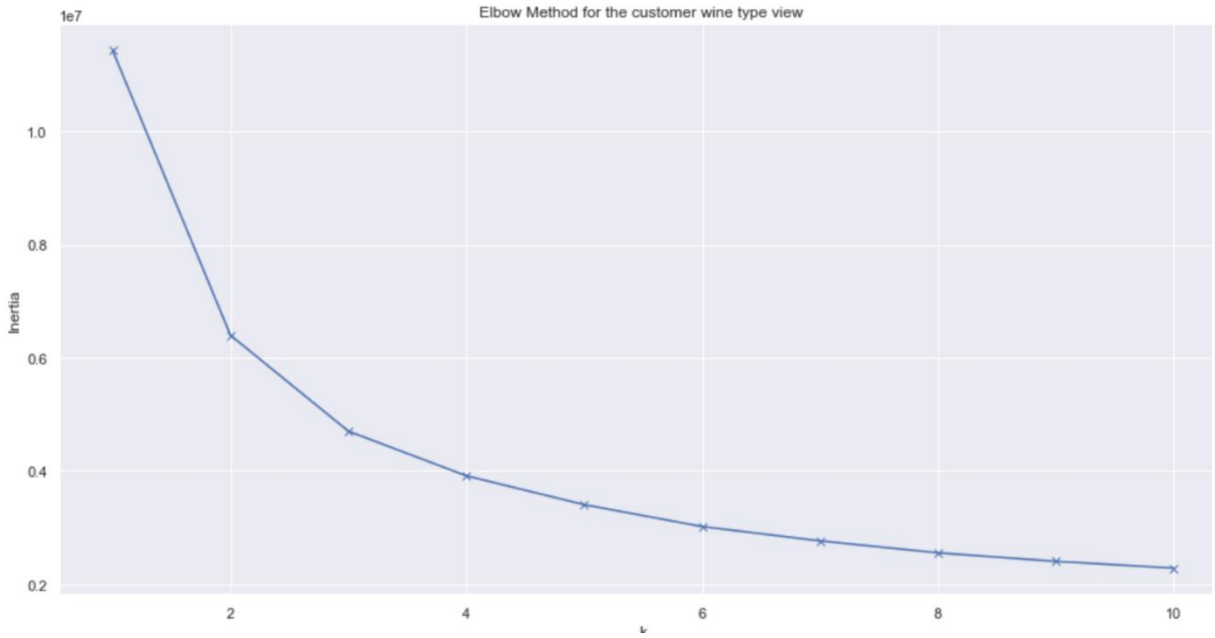


Figure 7.1: Elbow method for Wine Type.

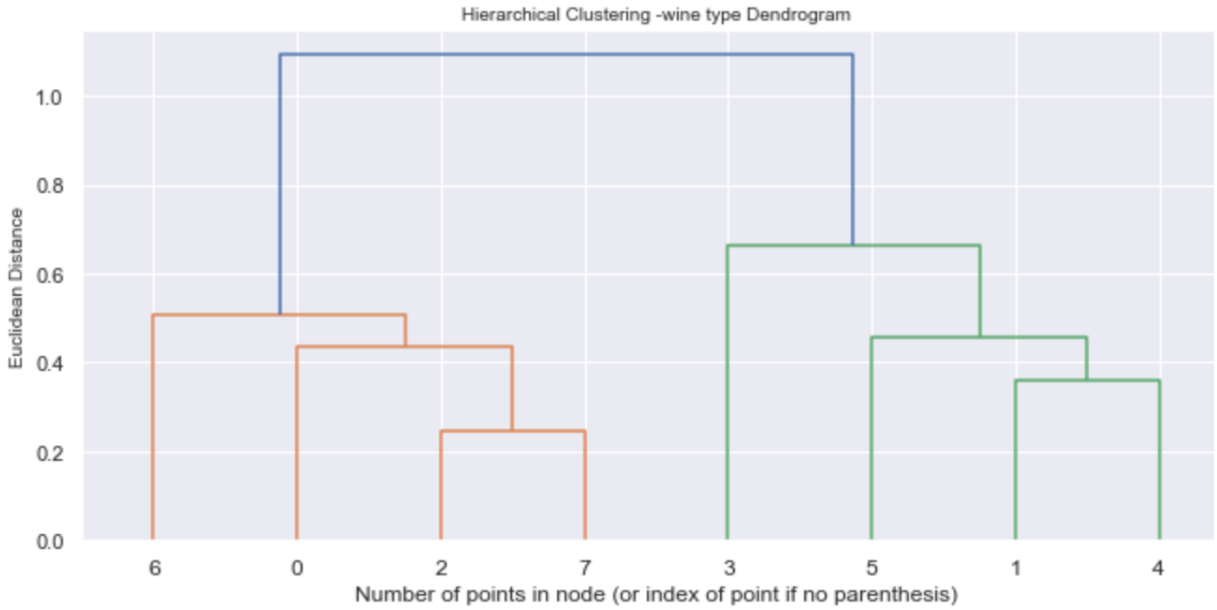


Figure 7.2: k-means dendogram for Wine Type view.

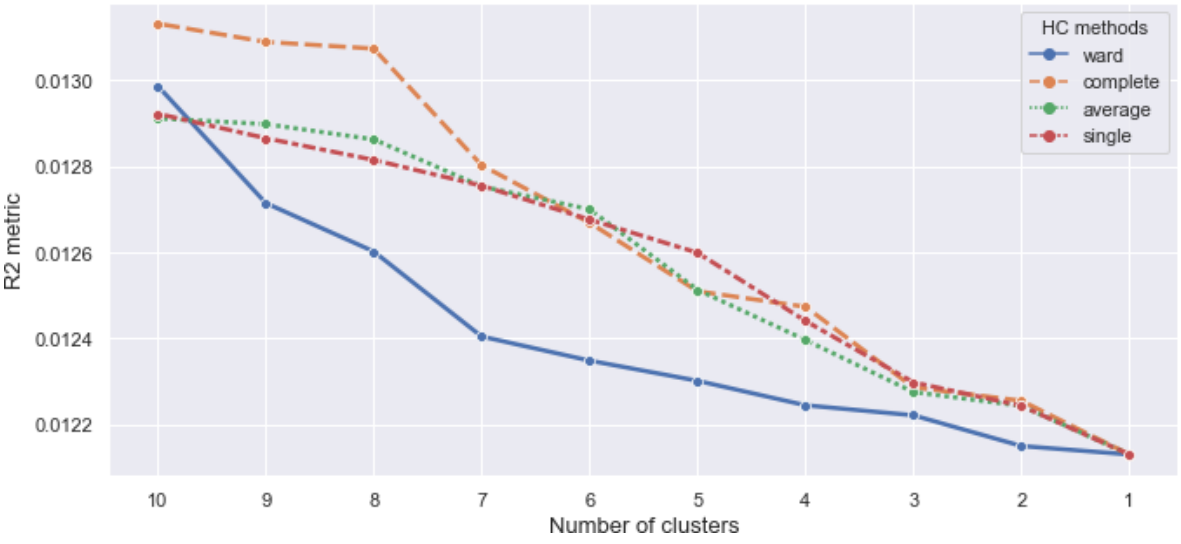


Figure 8.1: R2 plot for Behaviour view.

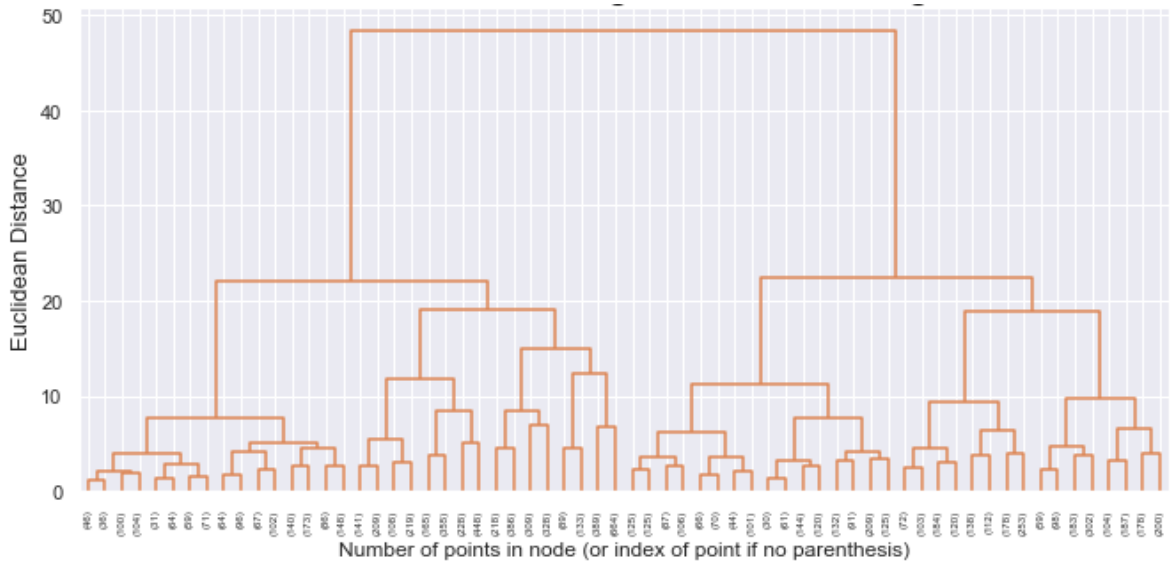


Figure 8.2: Ward's Dendogram for Behaviour view

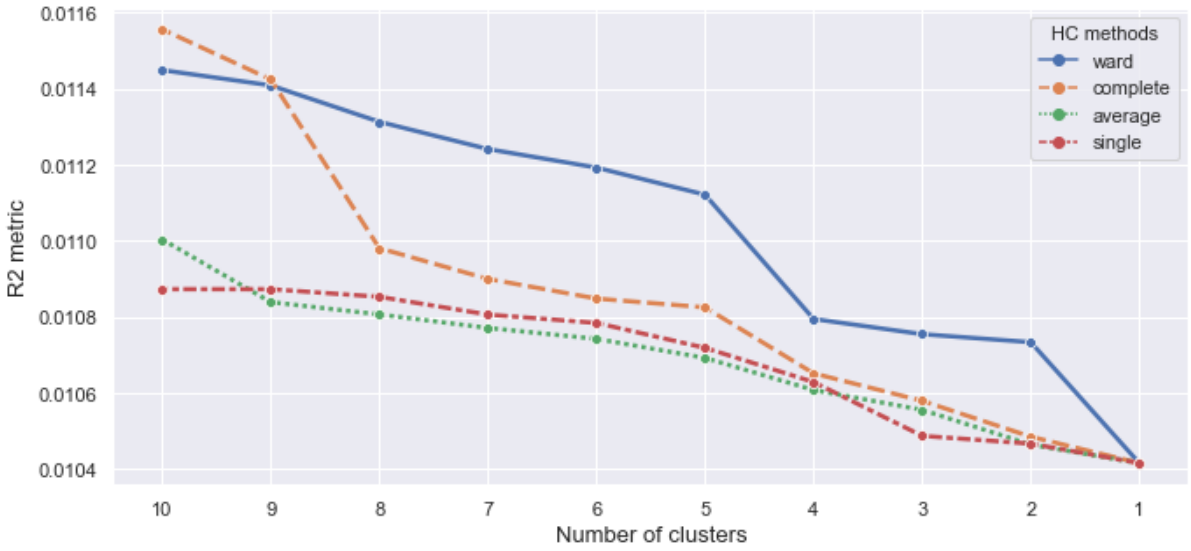


Figure 9.1: R2 plot for Wine Type view.

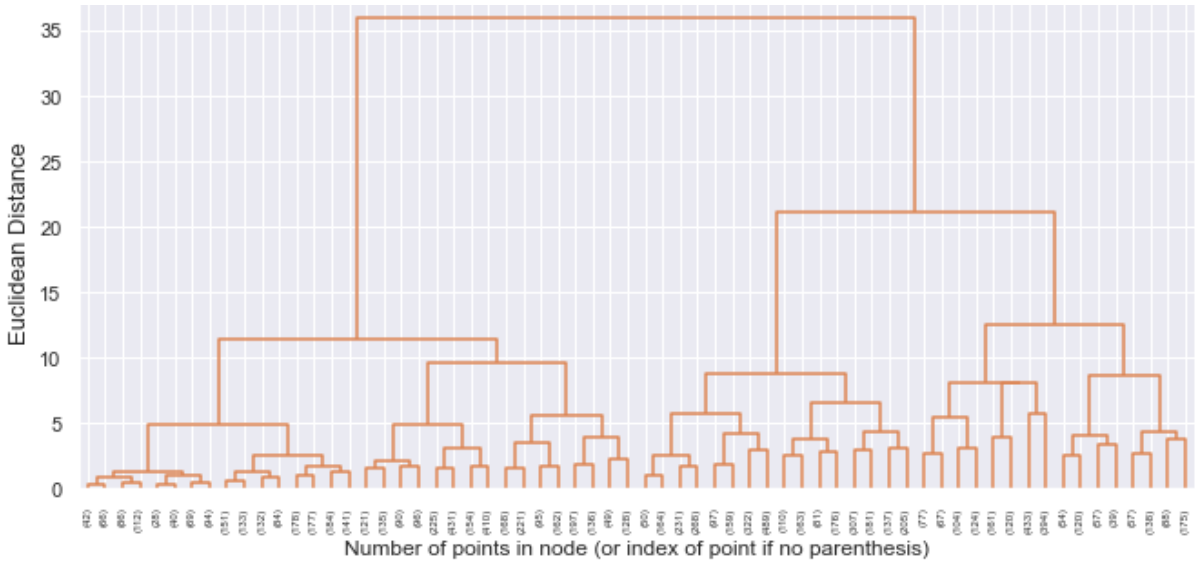


Figure 9.2: Ward's Dendogram for Wine Type view.

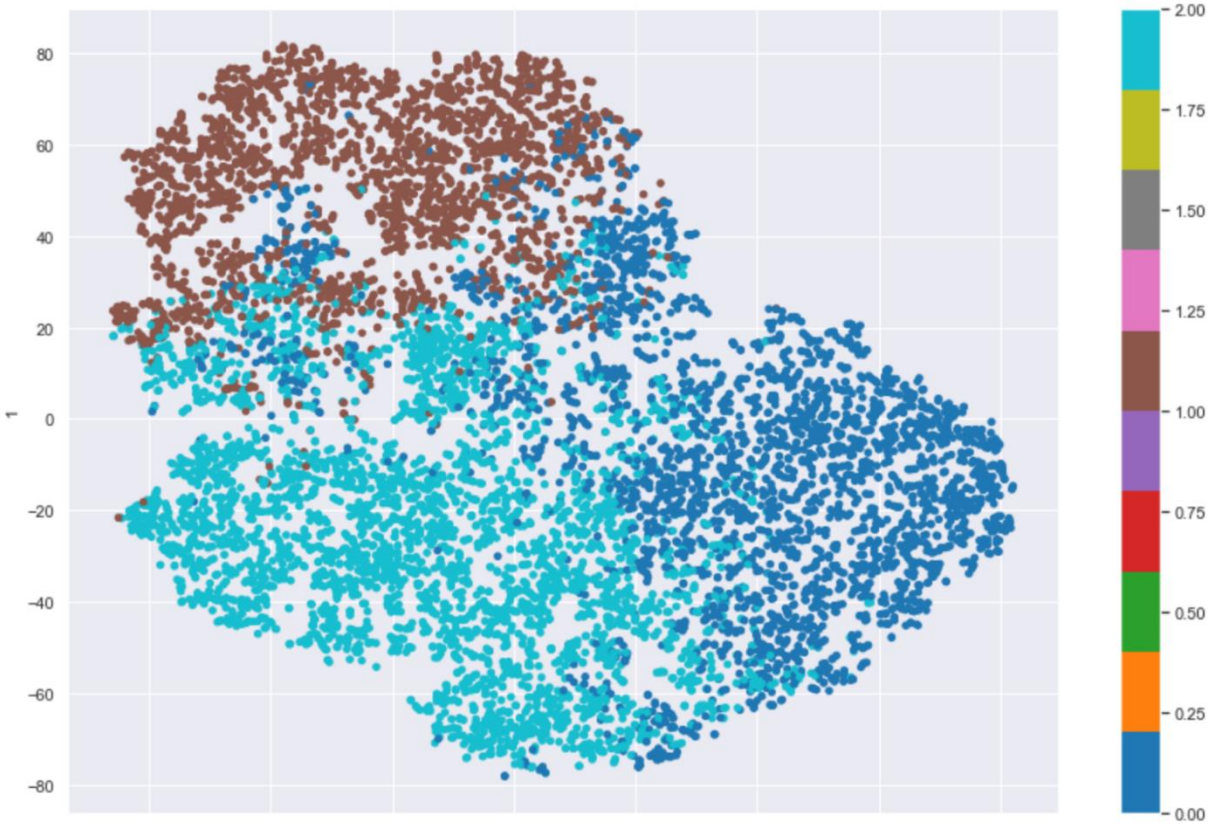
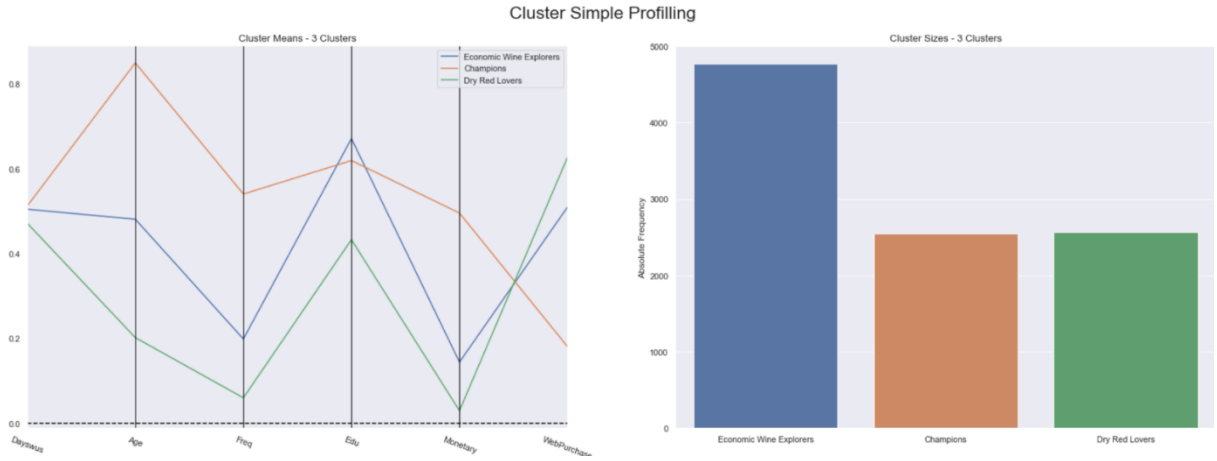


Figure 9.3: t-SNE of the chosen cluster solution



. Figure 10: Behavior Cluster

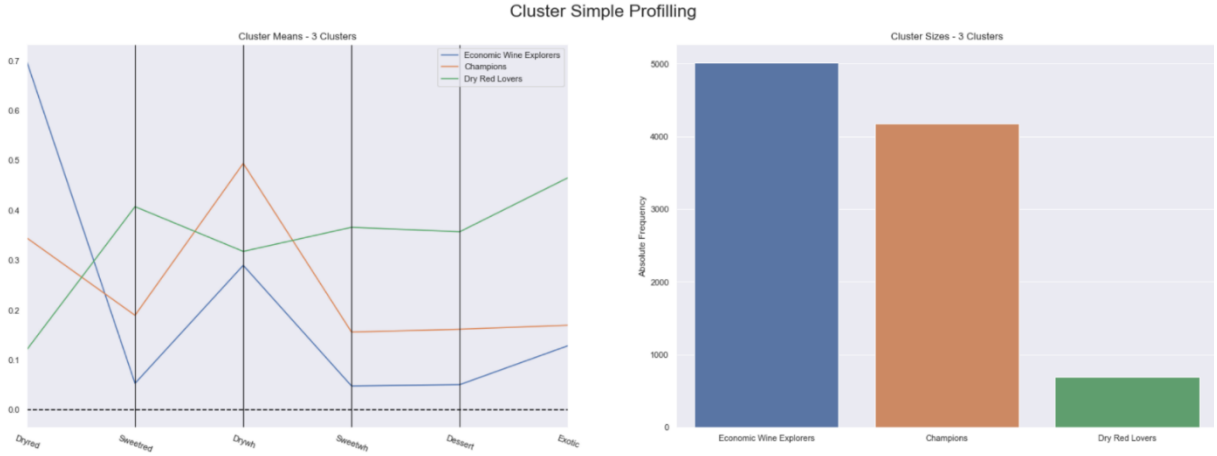


Figure 11: Wine Cluster

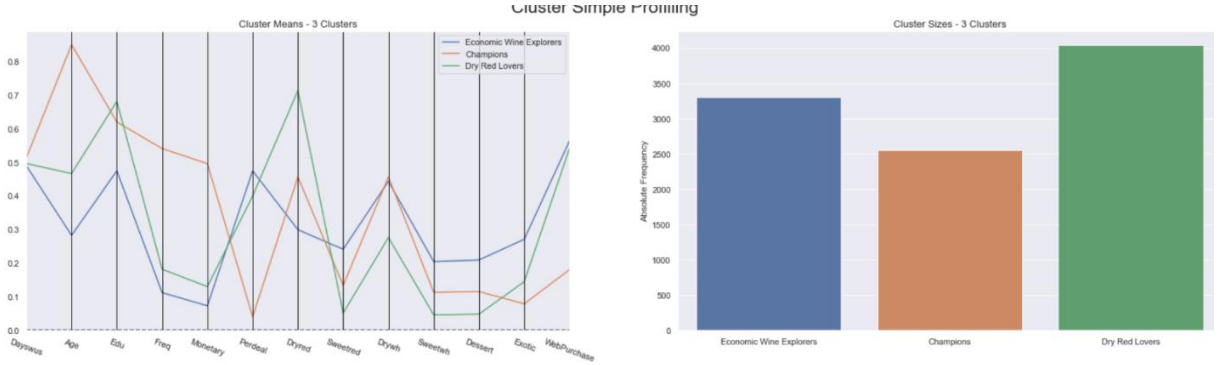


Figure 12: Merged Cluster

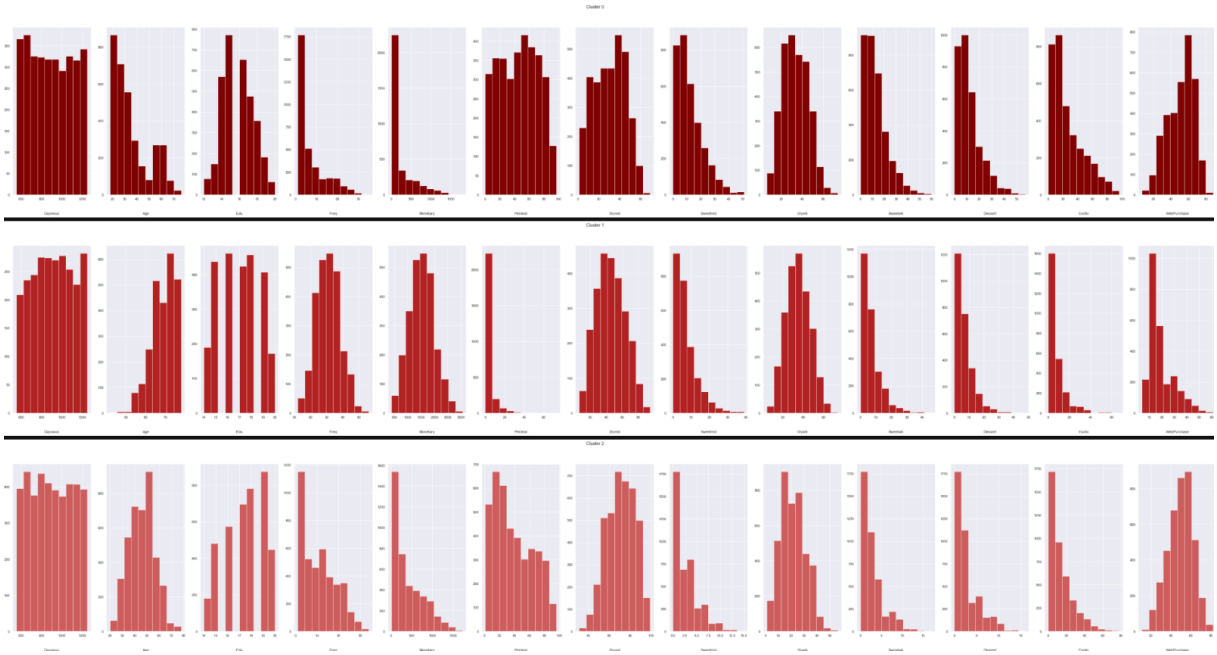


Figure 13: Data distribution per cluster