
WONDERFUL
WINES of the
WORLD

Wonderful Wines of the World

Business Case 1: Customer Segmentation

Marjorie Kinney	*m20210647*
Bruno Mendes	*m20210627*
Lucas Neves	*m20211020*
Farina Pontejos	*m20210649*

Business Cases for Data Science

NOVA Information Management School

March 2022

Table of Contents

Introduction

This project aims to perform segmentation analysis on a portion of the customer database of Wonderful Wines of the World (WWW), a company that specializes in selling specialty wines.

The company sells wines through several avenues: its website, physical stores in the United States, and through phone ordering. Until recently, the company has mass-marketed to all its customers by providing all 350 000 customers with the catalog, and using no other techniques such as loyalty programs or cross-marketing. By using segmentation on its customers, the company will be able to group customers based on shared characteristics and provide a more tailored approach to marketing. This not only reduces costs, it will also enable the company to deliver a more targeted marketing strategy.

Business Understanding

After a difficult 2020 with closed tasting rooms and restaurants, new trends emerged in the wine industry that impact expectations and possible marketing strategies. More specifically, recent trends include the improved demand from premium customers, which increased 21% from 2021 to 2022, the growth of e-commerce to retain Covid clientele, and the increased pressure put on producers from climate change events like drought, fire, low soil moisture and record low reservoir levels.

Data Understanding

We used various statistical and visualization techniques to explore and better understand the nature of the available data. In examining the data, we observed that the last row of the spreadsheet contained an aggregation of the previous rows and therefore removed it. Using histograms and box plots, we are able to visualize the distributions of the different variables. We observed several that exhibited a skewed distribution, as can be seen in the appendix. We did not observe any missing values or duplicated rows.

Data Preparation

We divided the available features into different groups based on the insights that they provided:

Value and Engagement Segmentation Features, related to the economic value that the customer brings to the company. **Wine Preference Segmentation Features**, related to the customers' wine preferences, and based on what proportion of their purchases account for the different wine categories. We used these for the clustering algorithms.

Demographic Features, which represent the demographic characteristics of the customers. **Other Features**, which describe the customers' behavior in ways not captured by the segmentation features. We used these to describe the obtained clusters.

Feature Selection

Looking at the Pearson and Spearman correlation across the Value Segmentation features, we observed that the Perdeal and Monetary variables are highly correlated to the other variables and removed them. In this sense, we made sure that we are not working with redundant data that may affect our clustering. We did not remove any of the Wine Preference variables.

Outlier Detection

We observed highly skewed distributions on some of the variables available. We calculated the interquartile range to determine the presence of outliers and found that too many rows would be removed using this metric. Instead we considered the box plots of each variable to manually determine a cutoff value. See figure in appendix. Moreover, we also used DBSCAN to complement our analysis. After combining both methods, we removed a total of 34 observations that we did not use in the modeling process in order to improve our results. Although, these observations were then reintroduced and labeled using a predictive algorithm so that we didn't lose any customer information in the segmentation.

Transformations

To account for relevancy, we started by normalizing the value segmentation features. This procedure made all the features equally important in the clustering process, regardless of their current scales. To do this, we used one of the most common normalization techniques – MinMax scaling. We have chosen MinMax scaling because it is effective and its major downsides such as sensitivity to outliers have already been addressed.

Regarding the wine preference features, since they were already in the same scale (%), these were simply converted to decimal values. As for the Demographic and Other features, we did not perform any transformation seeing that they are already numeric and will only be used to characterize our clusters.

Modeling

We tested several clustering algorithms on each set of features separately, and in the end combined the respective results to form a holistic view of the customers. The clustering performance was mostly evaluated on the basis of the R2 metric, which is a score that represents how well the data points are explained by the clustering solution. A higher R2 score means a better fit.

For the **value and engagement variables**, clustering with the hierarchical-clustering algorithm produced the best results with an R^2 of 0.957, a bit higher than the k-means solution which reported an R^2 of 0.947. We also tried using density-based algorithms like DBSCAN and Mean-Shift, and also constructed a Self-Organizing Map of the data but none produced comparable results. The following figure shows the scaled mean value for each characteristic, for each cluster:

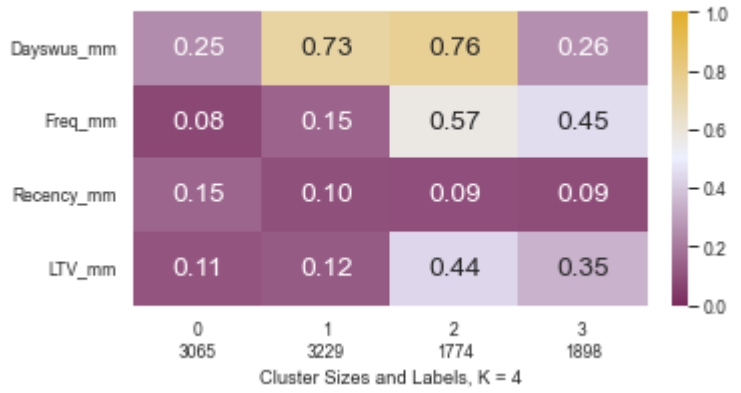

Figure: Cluster Means of Value Segmentation using Hierarchical Clustering

The cluster analysis is discussed in the **Evaluation** section of this report.

Four clusters were found to be optimal for this data for both the k-means and hierarchical clustering algorithms. This optimum was determined based on factors such as the uniformity of cluster size and metrics[1] such as R^2, silhouette scores[2] and euclidean distances.

For the **wine variables**, clustering with k-means produced the best results with an R^2 of 0.949, higher than the hierarchical clustering solution which reported an R^2 of 0.898 for a similar number of clusters. Clustering with DBSCAN did not yield any interpretable results as it only identified a single cluster. Similarly, the Self-Organizing Map produced sub-optimal results. The following figure shows the mean proportion value for each wine type, for each cluster:

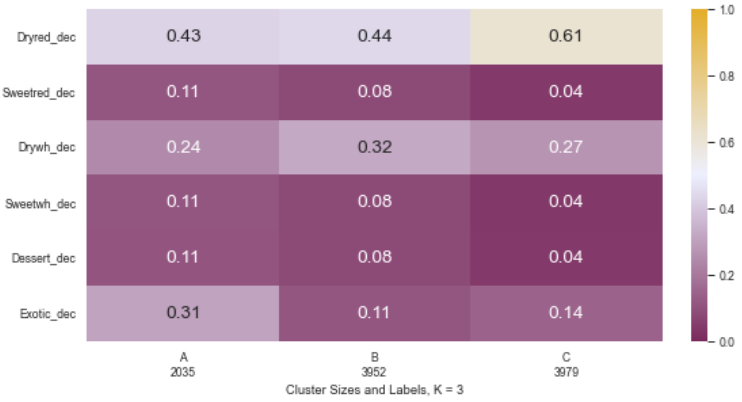

Figure: Cluster Means of Wine Segmentation Using K-Means Clustering

The cluster analysis is discussed in the **Evaluation** section of this report.

In this case, three clusters were found to be optimal for this data for the k-means solution. Just as before, this optimum was determined based on factors such as the uniformity of

[1] Dendrogram of Value Segmentation Clusters
[2] Silhouette Scores, Value Segmentation

cluster size and metrics such as R^2, silhouette scores[3] and euclidean distance. The hierarchical clustering solution's optimum was two clusters, which is part of the reason why we did not obtain good results.

Afterward, we combined the results to create 12 customer segments. Each customer has a label for their value/engagement, and a label for their wine preferences.

Evaluation

Each of the clusters represent a grouping of customers that on average share similar characteristics. They can be described as follows:

Value and Engagement Clusters:

0: (Young) On average, this cluster represents clients that have not been with WWW for long, and do not have a high lifetime value. They do not purchase frequently and spend little compared to the other customers. When they do purchase, they tend to buy products on sale. They are younger and have the lowest income. On average, they have not purchased recently. Without intervention, these customers will likely become Cluster 1 customers.

1: (Hibernation) On average, customers in cluster 1 are quite similar to cluster 0, but have been with WWW for a long time. They are slightly more valuable than cluster 0 in all value and engagement characteristics, but are not as valuable as the customers in clusters 2 and 3.

2: (Platinum) This cluster represents clients that have been with WWW the longest and have the highest lifetime value. They purchase the most frequently and spend more money than customers in all other clusters. They are the oldest customers and have the highest income of all the clusters. Promotions and sales are not typically associated with their purchases.

3: (Potential Platinum) On average, customers in cluster 3 are quite similar to cluster 2, but have not been with WWW for as long. Although they have a relatively high lifetime value, it is lower than customers in cluster 2. The same applies to how much they spend, and how frequently they do so. These customers in time could become Cluster 2 customers.

Wine Preference Clusters:

A: (Exotic Wine Lovers) This cluster represents clients who buy the highest proportion of exotic wines. They also spend a higher proportion of their money on sweet red, sweet white, and dessert wines than the other clusters.

B: (Dry White Wine Lovers) This cluster represents clients who have the highest preference for dry white wine. They are the least likely to purchase exotic wines.

C: (Dry Red Wine Lovers) This cluster represents clients who spend the highest proportion of their money on dry red wine, typically spending almost 20% more on dry reds than their counterparts in other clusters. They spend the lowest proportion on sweet and dessert wines.

[3] Silhouette Scores

Deployment

We established different strategies for each cluster. Since each customer is assigned to a value/engagement cluster and a wine preferences cluster, the marketing strategy for each customer should be a combination of the suggestions below:

Young - Give a good first impression of the company. Offer wine pairing tips and suggest wines they may not have previously tried. Offer discounts and promotions for these customers.

Hibernation - Re-engage them and offer new products that have been released since they last purchased. Offer free shipping and/or discount if they purchase in the next X weeks. Consider reaching out to these customers to see what would entice them to return.

Platinum - Provide special rewards for these customers to help show appreciation and allow them to feel valued, such as custom bottles or exclusive wine tasting opportunities. They have already proven a willingness to spend money, so discount pricing may not be the most enticing reward. Instead, focus on value added services such as personalized recommendations based on their own previous purchases, or wine purchases of people in the same wine cluster they belong to. Offer an exclusive newsletter to the Platinum tier so they hear about new releases first before any other customers.

Potential Platinum - Build their loyalty by maintaining contact as much as possible. Invite them to events and keep them engaged. Entice them with the exclusive perks that Platinum tier offers to encourage them to purchase more to qualify.

Exotic Wine Lovers - Continually search for novelty and ensure there is a large enough selection of exotic wines available.

Dry White Wine Lovers - Consider creating a catalog specific to dry white wines to distribute to these customers.

Dry Red Wine Lovers - Consider creating a catalog specific to dry red wines to distribute to these customers.

Automatic Classification & Outlier Prediction

We built an algorithm to reclassify data points identified as noise/outliers through each of the clusters. This algorithm can also be used to classify new customers, as well as identify and predict noise in the data.

This process uses a decision tree that splits the data through the various clusters, based on features with highly predictive ability. On average, with our current data, we are able to assign customers to the correct clusters for Wine an estimated ~86% of the time and ~90% for the Value clusters.

Appendix

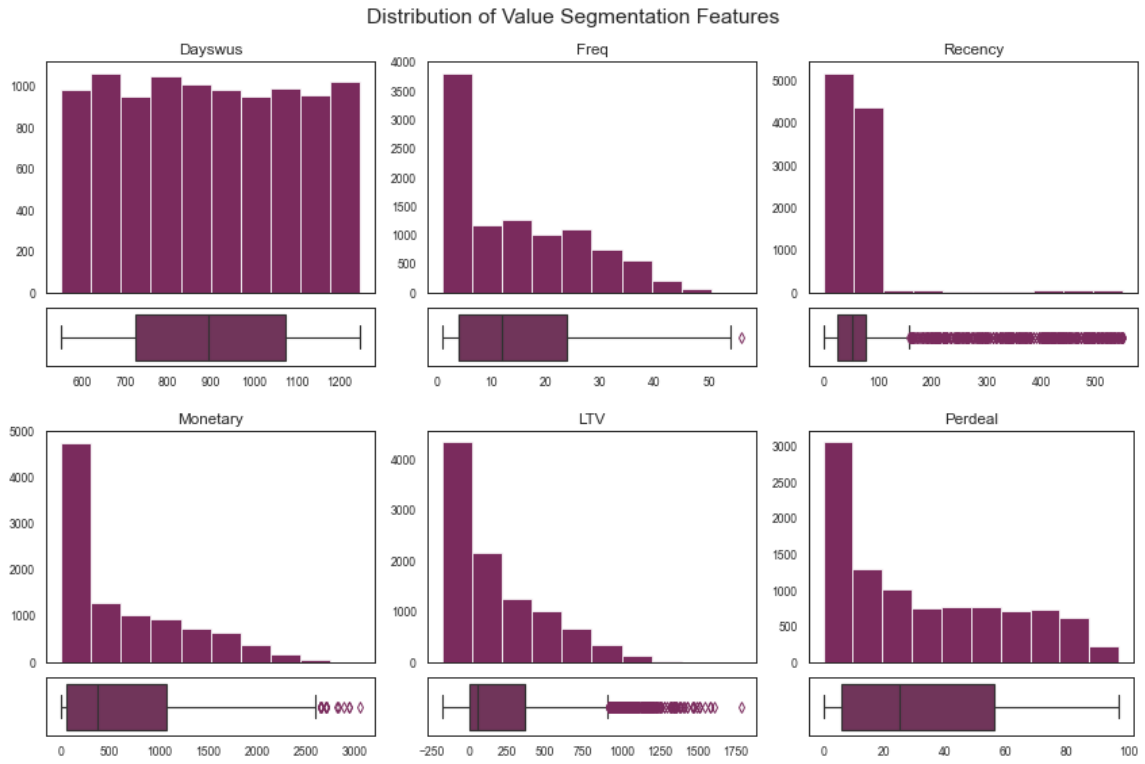

Figure 1: Distribution of Value Segmentation Variables Before Preprocessing

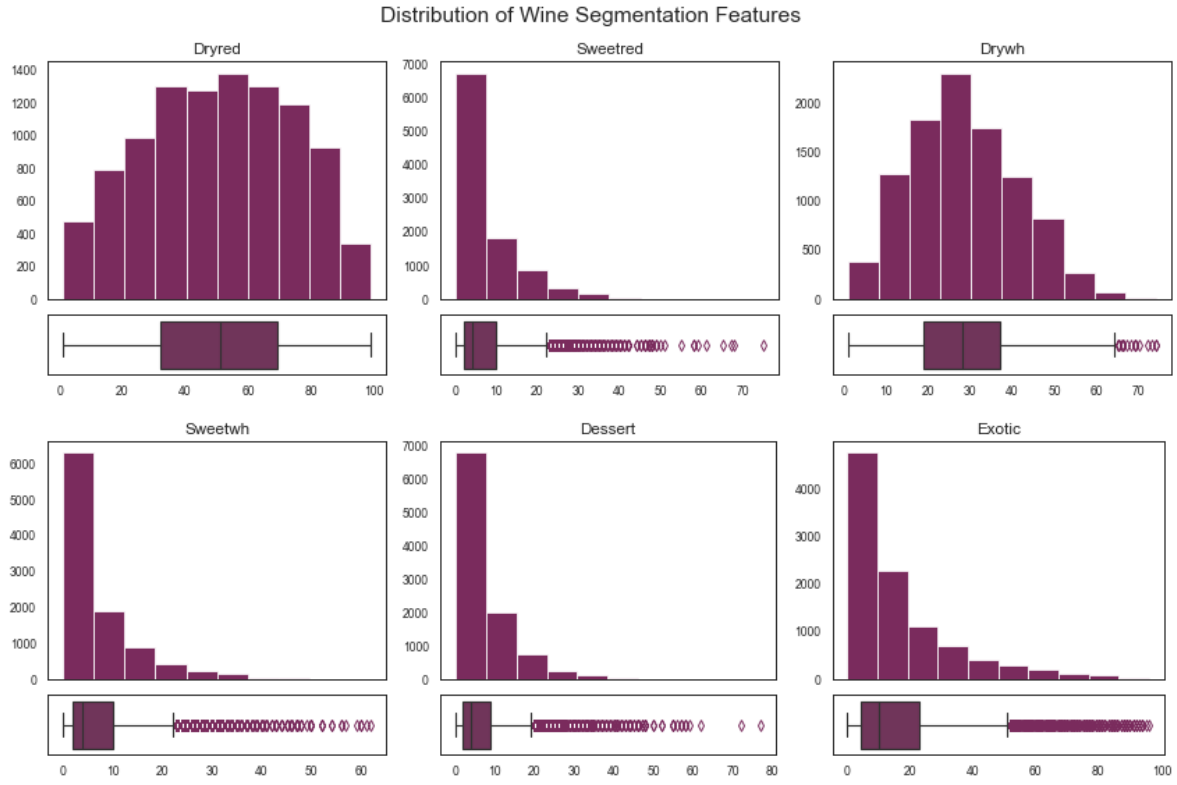

Figure 2: Distribution of Wine Segmentation Variables Before Preprocessing

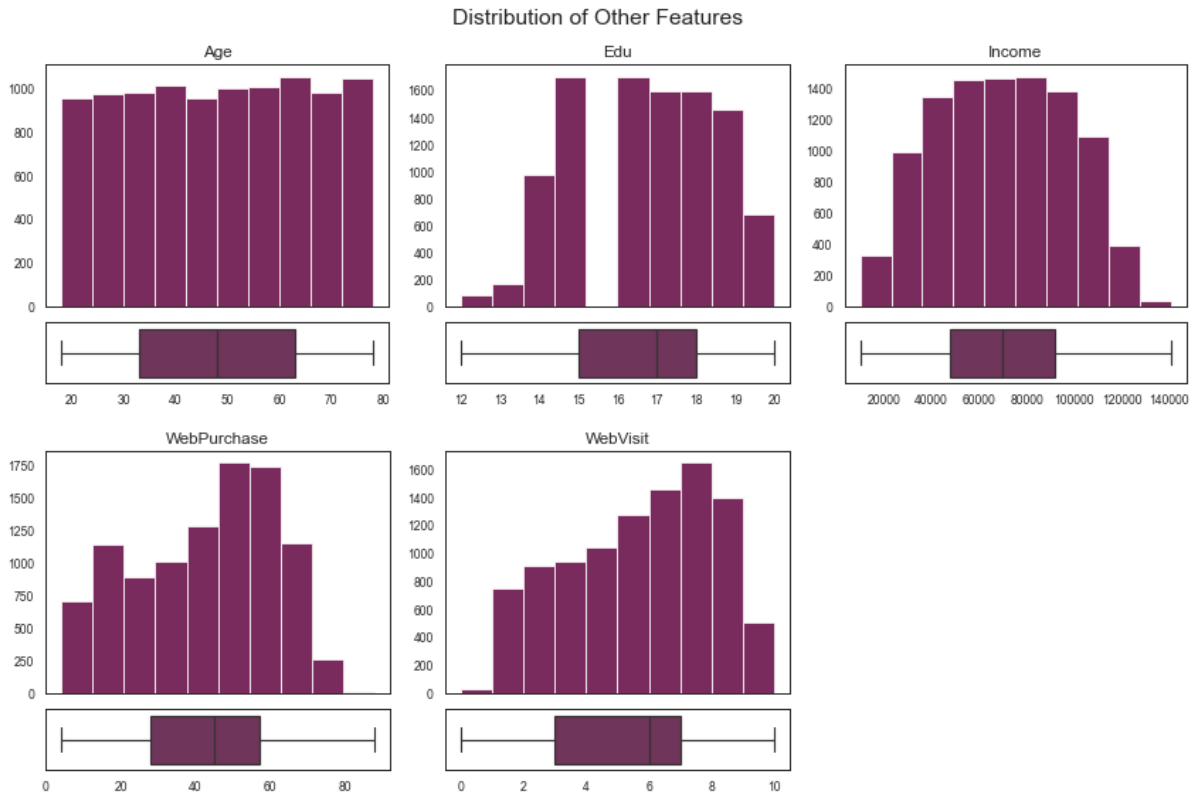

Figure 3: Distribution of Demographic Variables

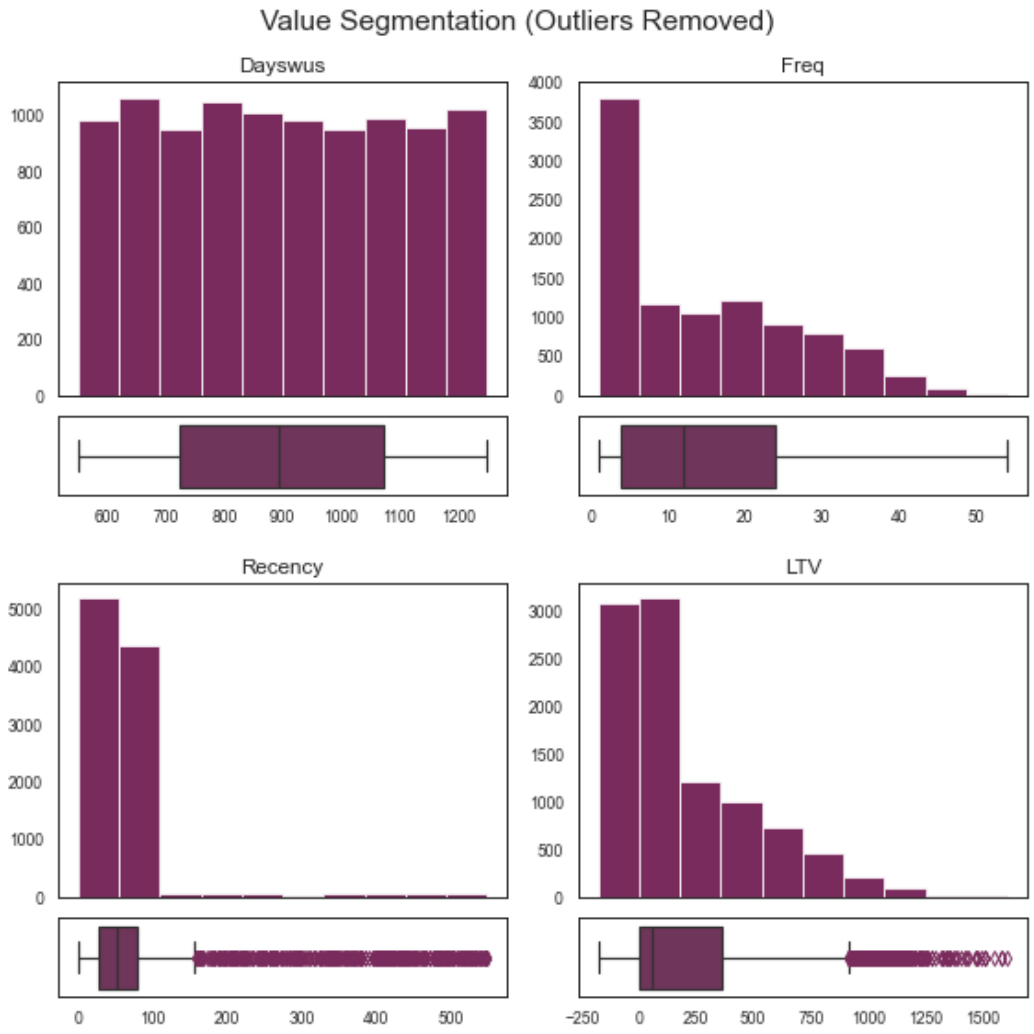

Figure 4: Distribution of Value Segmentation Variables After Preprocessing

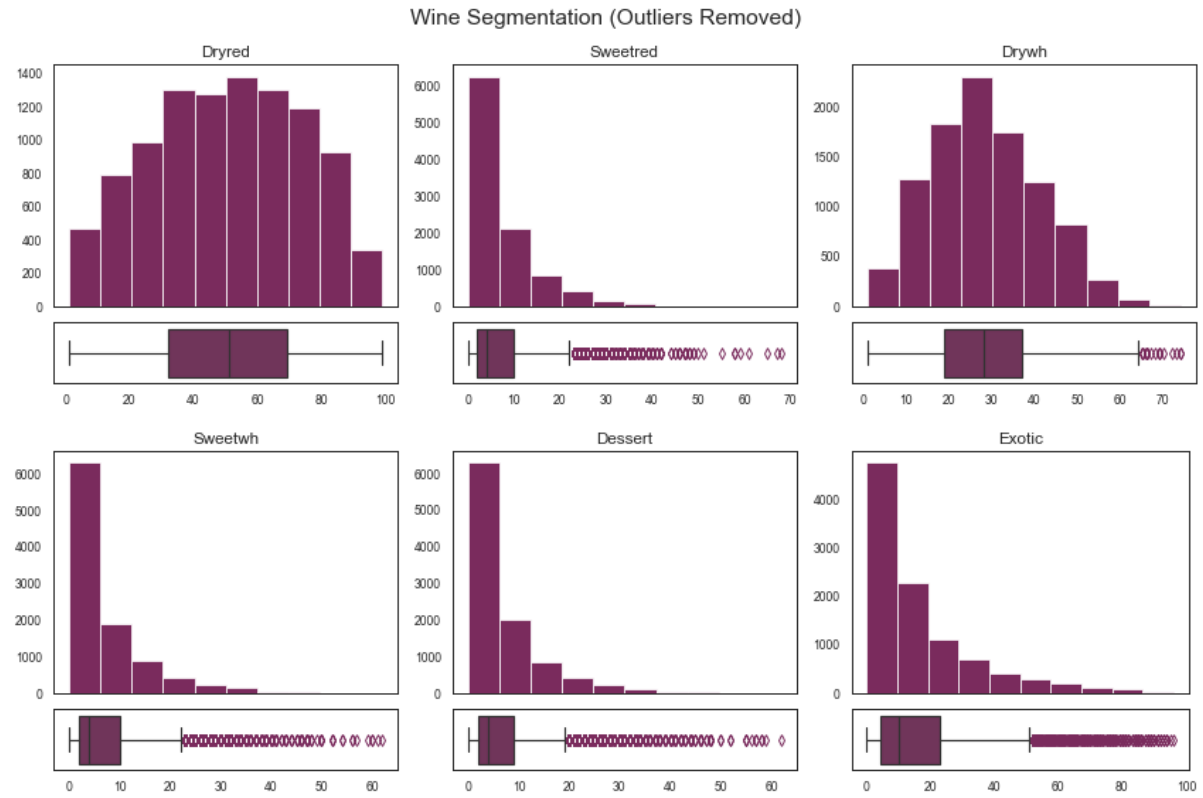

Figure 5: Distribution of Wine Segmentation Variables After Preprocessing

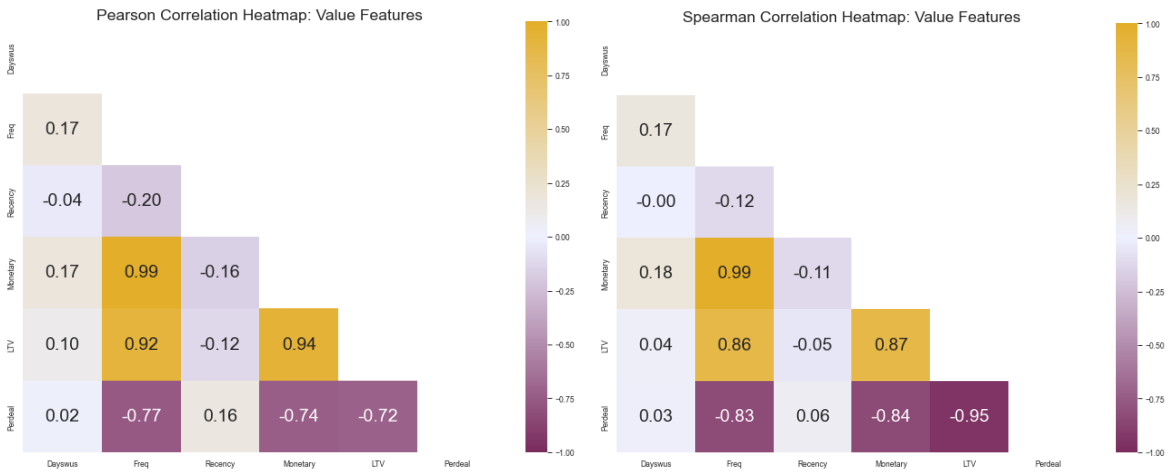

Figure 6: Value Features Variable Correlations

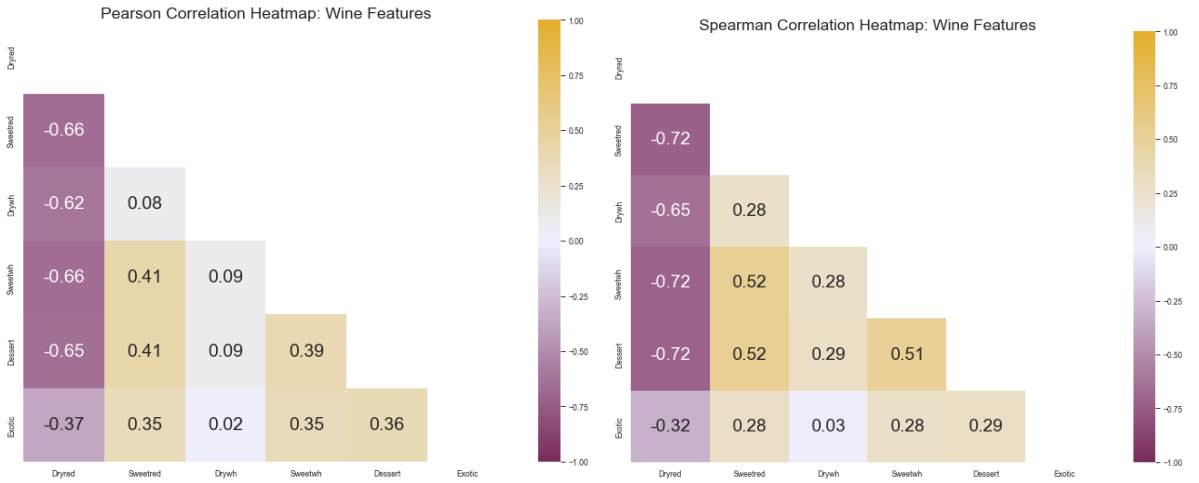

Figure 7: Wine Features Variable Correlations

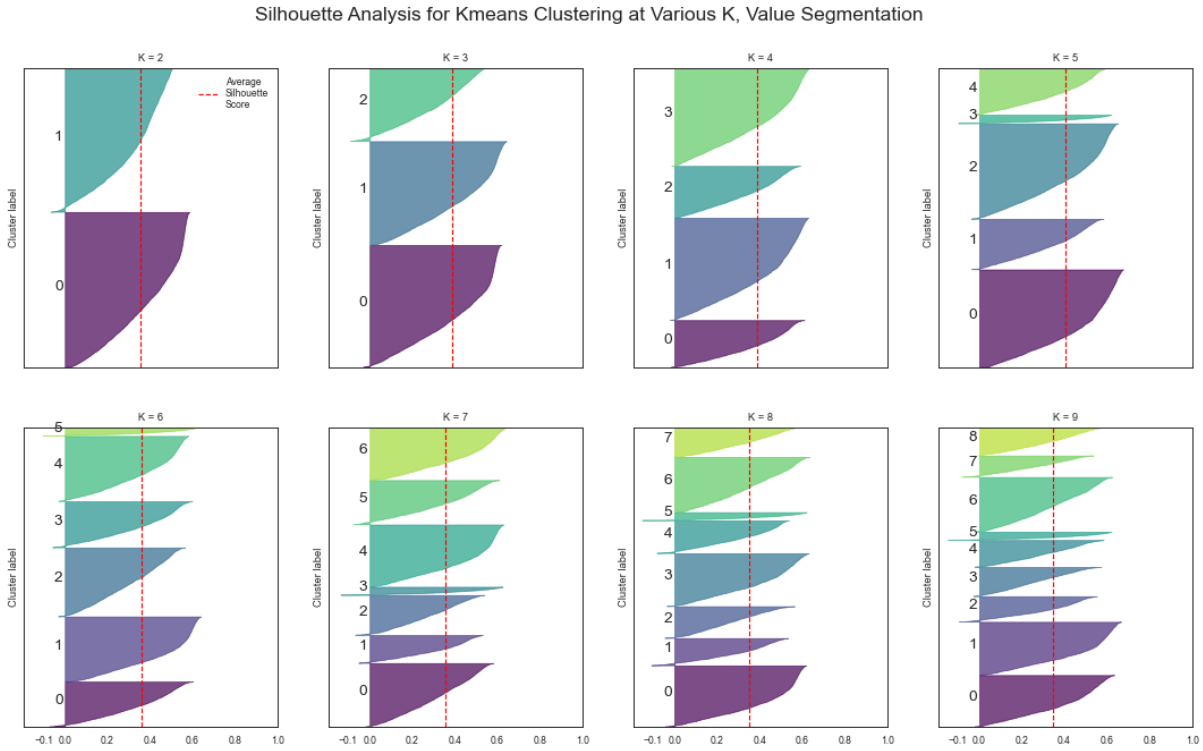

Figure 8: Silhouette Plots for Value Features

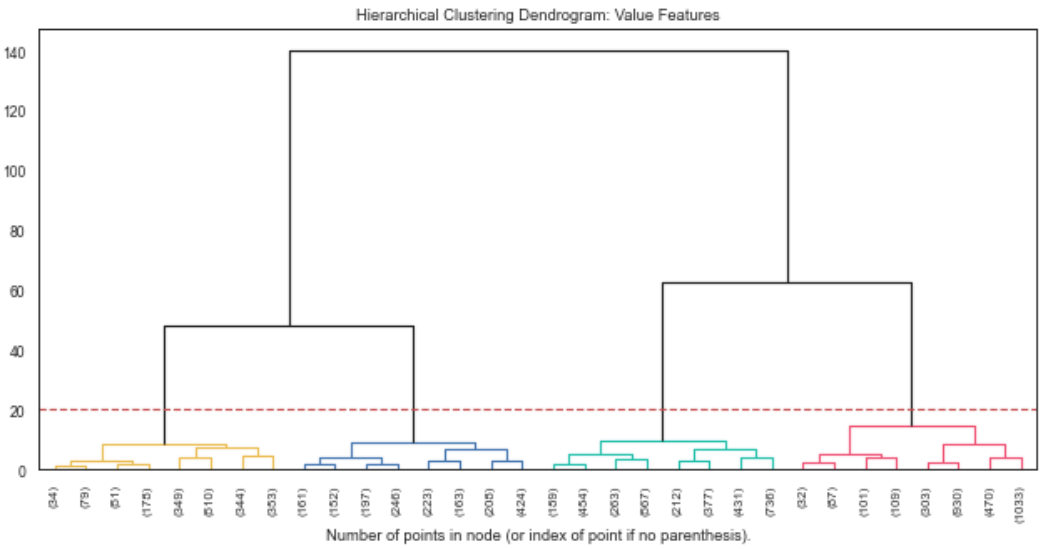

Figure 9: Dendrogram for Value Features

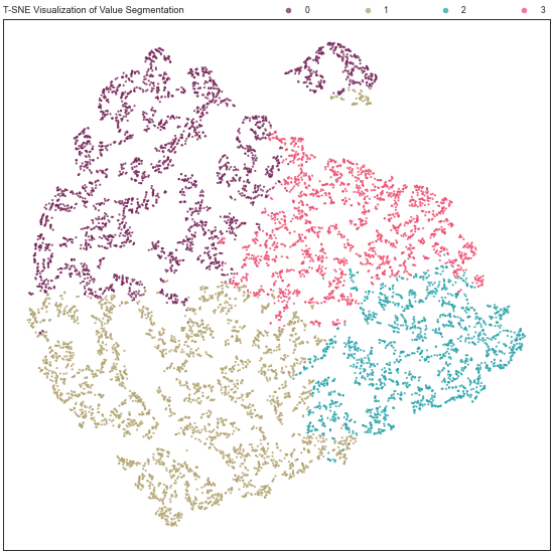

Figure 10: T-SNE Visualization for Value Features

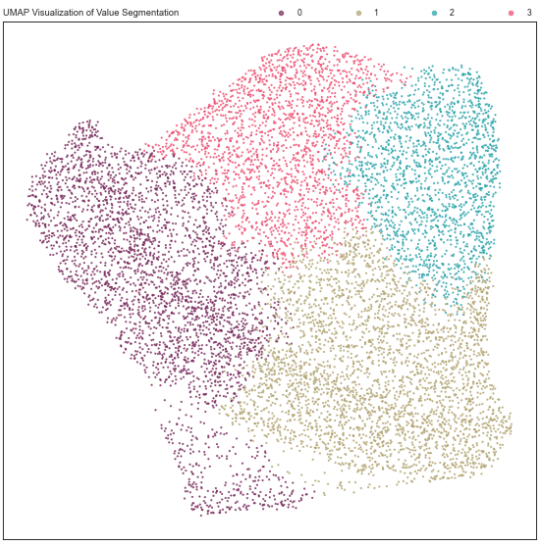

Figure 11: UMAP Visualization for Value Features

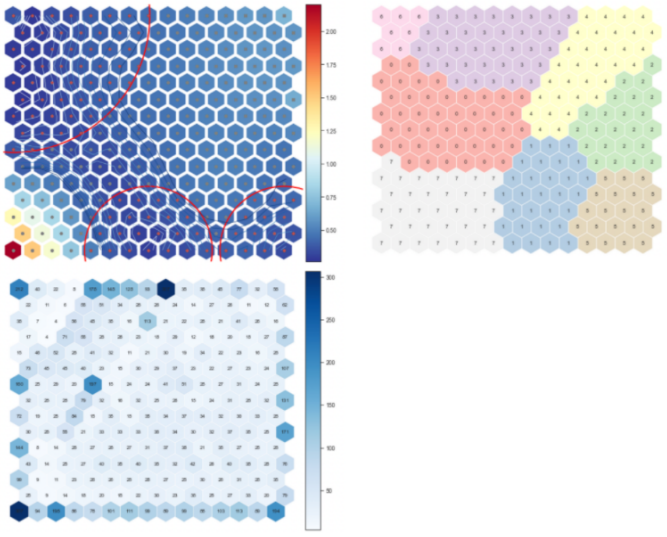

Figure 12: Self-Organizing Map for Value Features

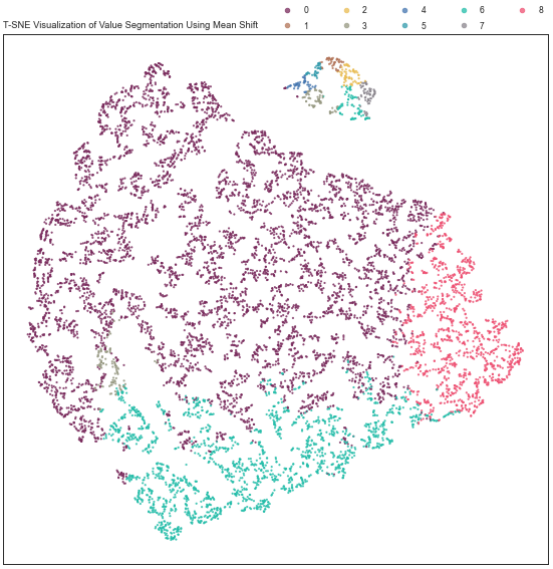

Figure 13: T-SNE Visualization for Value Features Using Mean-Shift Algorithm

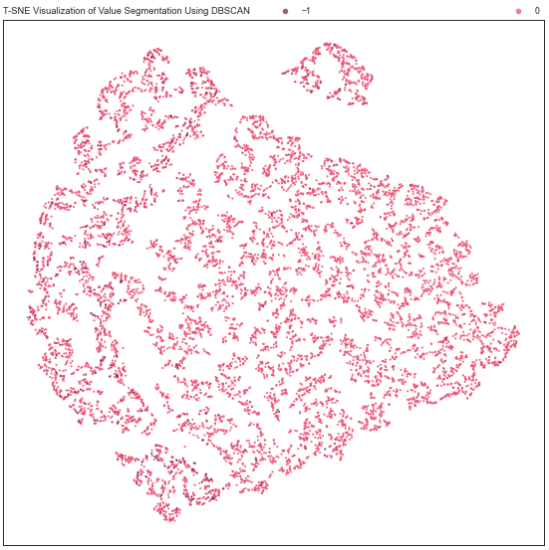

Figure 14: T-SNE Visualization for Value Features Using DBSCAN

Figure 15: Heatmap for Various Wine Features Cluster Numbers

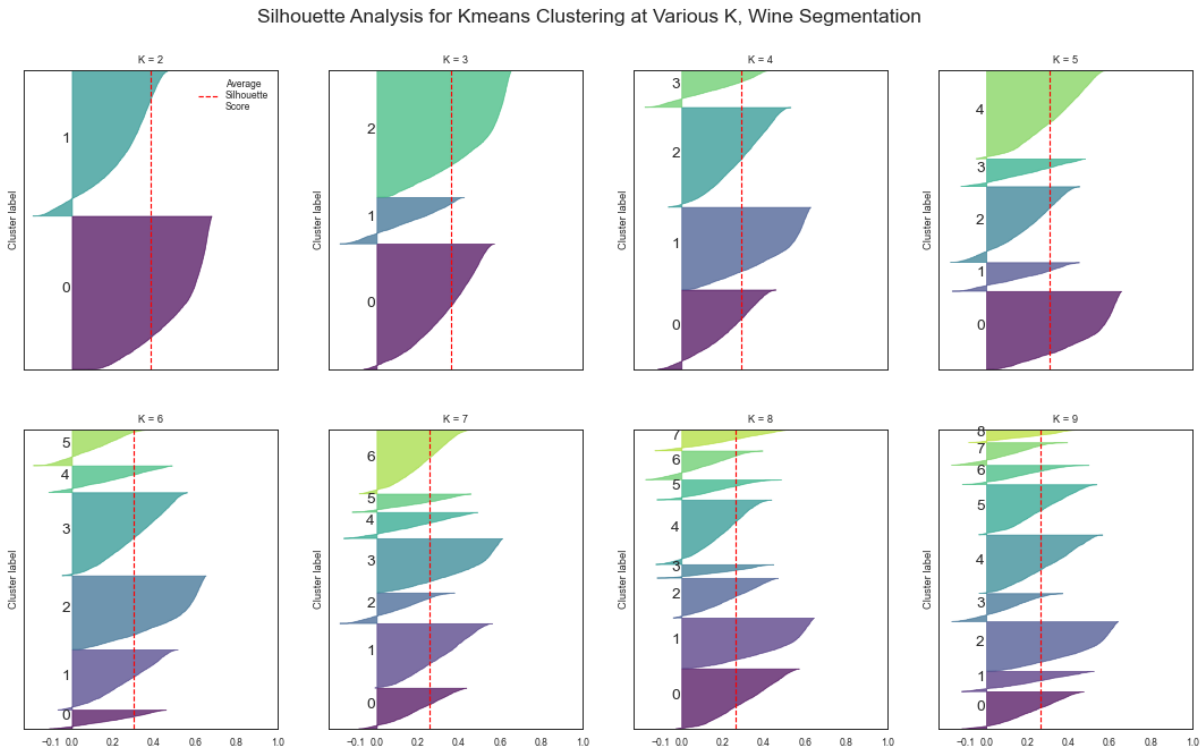

Figure 16: Silhouette Plots for Wine Features

Figure 17: Dendrogram for Wine Features

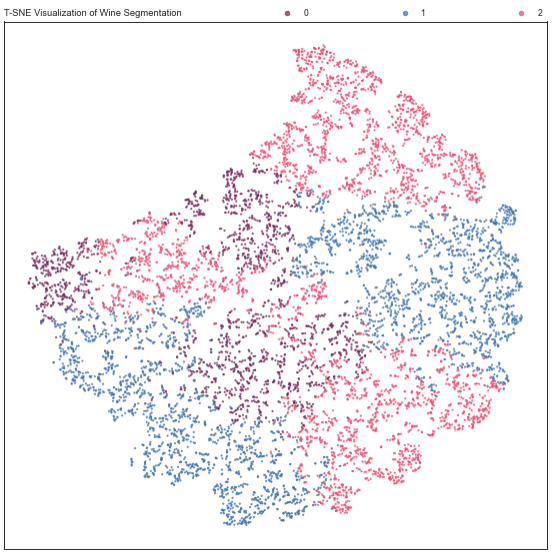

Figure 18: T-SNE Visualization for Wine Features

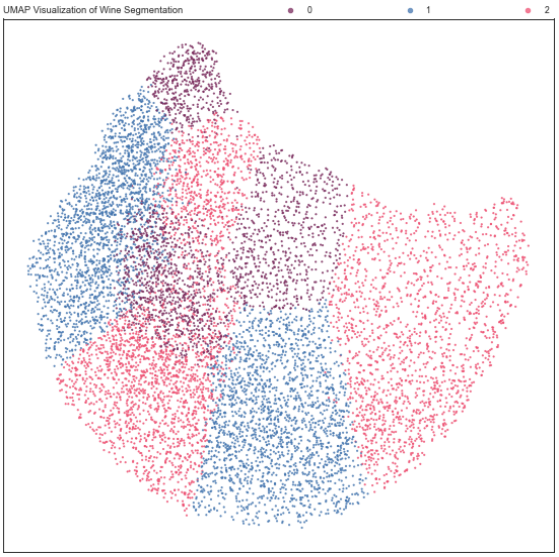

Figure 19: UMAP Visualization for Wine Features

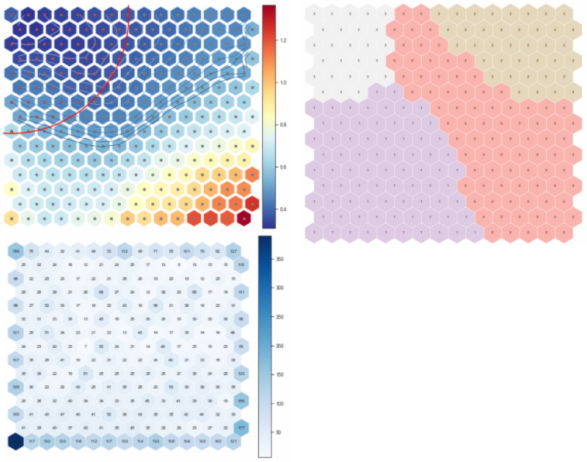

Figure 20: Self-Organizing Map for Wine Features

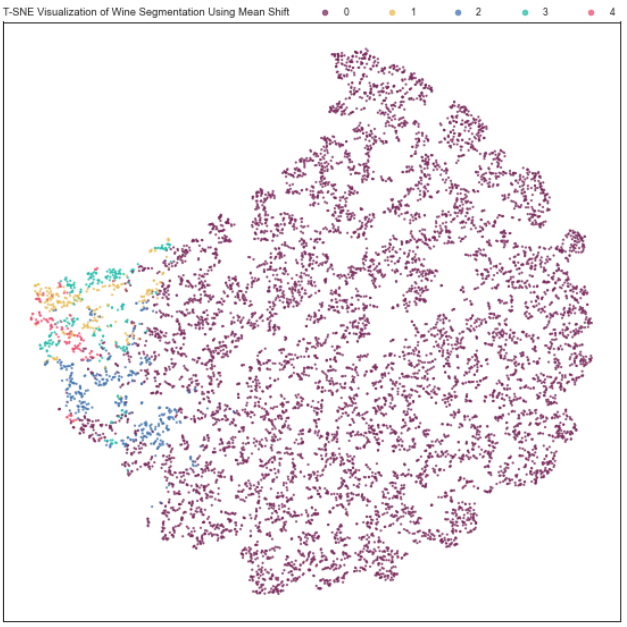

Figure 21: T-SNE Visualization for Wine Features Using Mean-Shift Algorithm

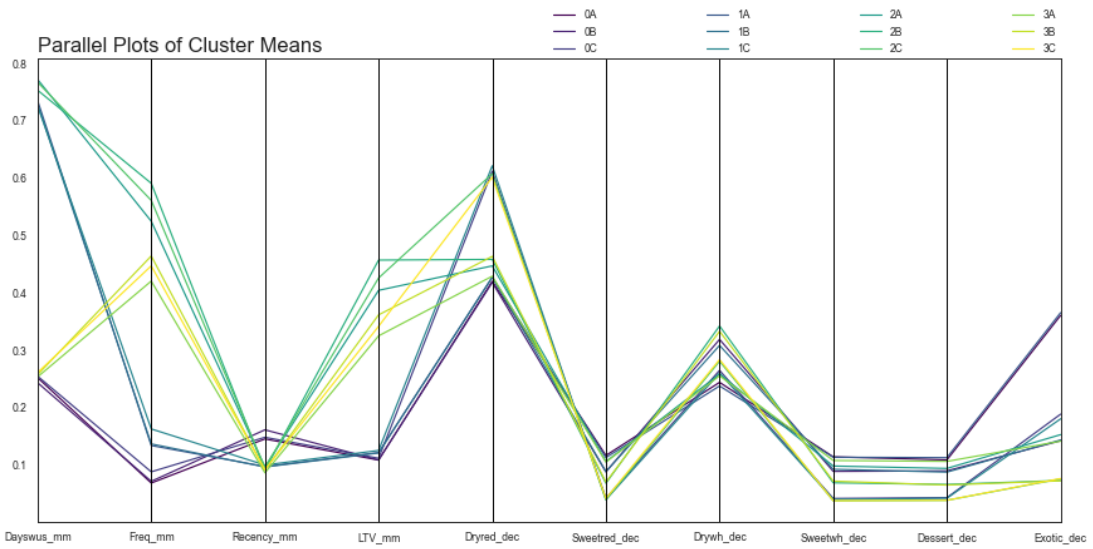

Figure 22: Parallel Plots of Cluster Means

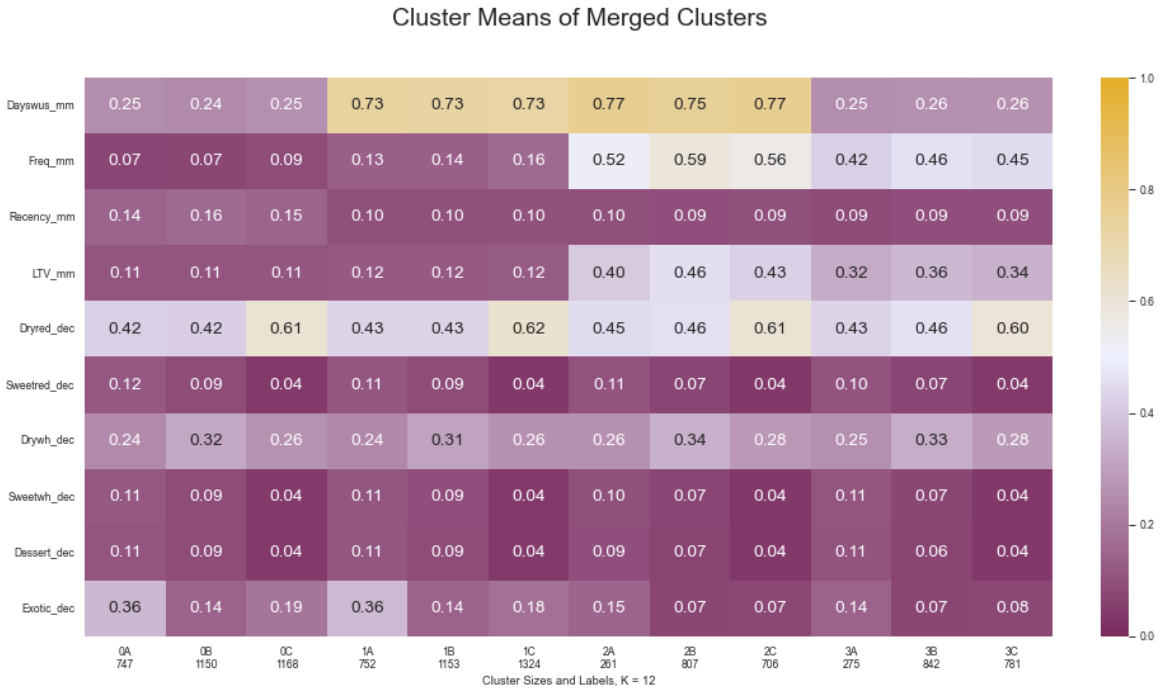

Figure 23: Cluster Means of Merged Clusters

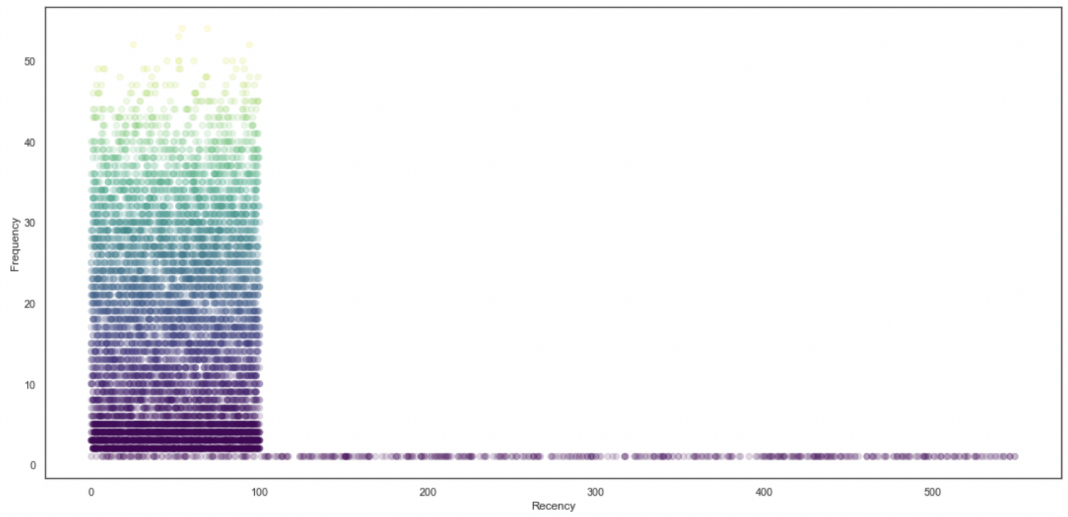

Figure 24: Recency vs Frequency

Plotting Recency vs Frequency, we can see a sharp demarcation at Recency = 100 days. At Recency > 100, the Frequency = 1, meaning they did not repurchase beyond 100 days.

There are 419 data points to the right of this number.

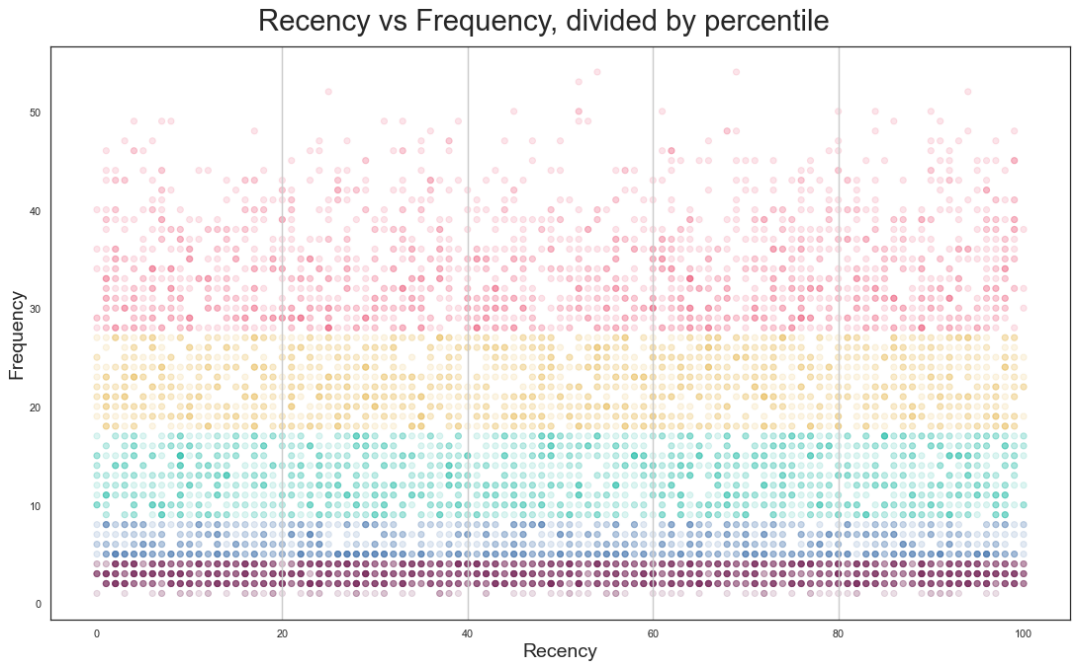

Figure 25: Recency vs Frequency with Recency>100 Removed

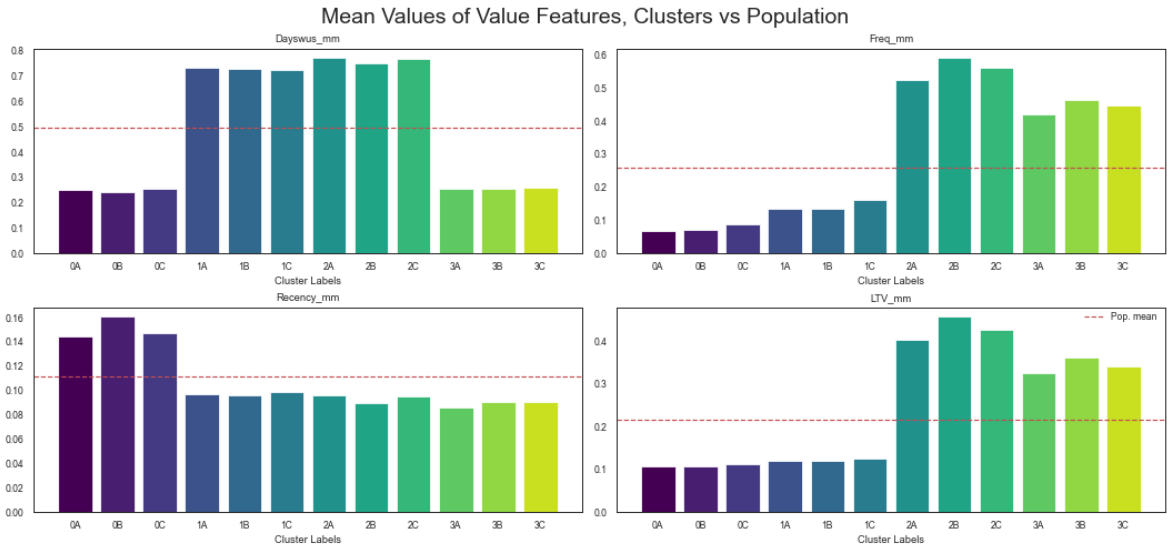

Figure 26: Mean Values of Clusters vs Population for Value Features

Figure 26: Mean Values of Clusters vs Population for Wine Features

Figure 26: Mean Values of Clusters vs Population for Other Variables