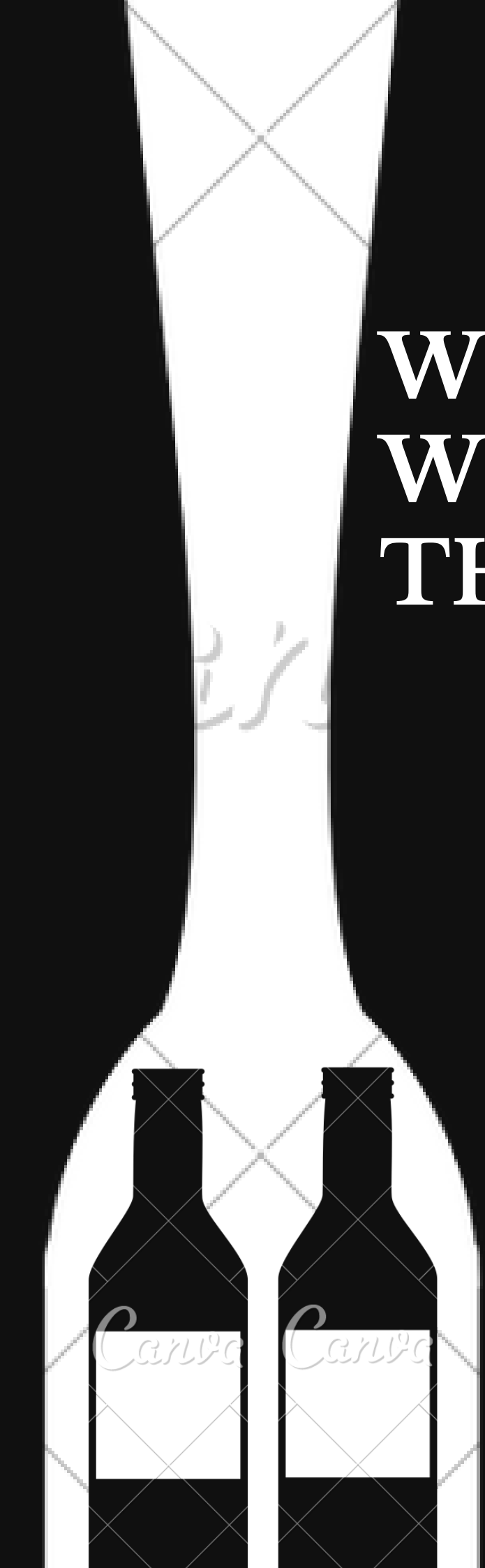

WONDERFUL WINES OF THE WORLD

M CONSULTING

Renan Stoffel 20210594
Mikala Durham 20210645
Sarra Jebali 20210765

Table of contents:

Introduction:

As markets get more prominent and more diversified, competitions get more and more challenging. And companies that want to appeal to a broader market portion can no longer rely on a one-size-fits-all business approach. Nowadays, each business wishing to build a strong clientele is working towards more personalized marketing strategies, which can only be accomplished through learning as much as possible about the existing customers and creating a solid customer segmentation. In this context, this project was made to perform a cluster analysis for Wonderful Wines of the World (WWW), a wine company, with the hopes of defining clear customer clusters that will be later used to develop an effective marketing strategy for the business. To achieve the desired objectives, we will follow the CRISP-DM methodology to analyze and process the WWW database. It will be used to create, describe and explain clusters. Our final goal will be to offer recommendations for the marketing approach based on the customer segmentation obtained.

1 – Business Understanding

Wonderful Wines of the World (WWW) is a 7-year-old enterprise that hired us to perform a customer segmentation using their 10,000 customers sample database. The database brings us information on the customers who have purchased something from WWW in the past 18 months. Our main goal is to understand, prepare and segment the data to find different customer profiles, assigning them to relevant clusters. At the same time, WWW wants to better understand the value of each customer and which wines they are more interested in buying.

2 – Data Understanding

We started by performing some exploration of the data and understanding what each feature represented.

The database is composed of 10,000 observations and 17 features (all numeric variables) related to:

- Customer background: age, education, and income.
- Customer relations with the company: sales information, lifetime value of the customers, frequency of purchases, among others.
- Customer preferences: what kind of wine they usually buy.

3 – Data Preparation

3.1 – Data Preprocessing

No missing values or duplicates were found in the database. Only one null index was found and filled. Also, most of the data was well distributed as presented in the boxplots in figure1 (appendix). However, there were still some multivariate outliers, so we decided to perform a DBscan using the robust scaler to remove 60 observations during clustering. They were re-added after the final labeling.

After applying the robust scaler and DBscan, we started exploring the data itself to decide which variables to keep to start applying the clustering models.

3.2 – Feature Selection

After performing the basic preprocessing steps, we started exploring the data itself and trying to select the features we would need to perform the segmentation in value and customer preferences.

The graph in figure 2 (appendix) presents the correlation matrix, in which we realized that we had a lot of overlapping variables.

As we can see, Age, income, frequency, monetary, and LTV are highly correlated, and, at the same time, Perdeal, WebPurchase, and WebVisit are also highly correlated. It is also noticeable that most wines are highly correlated among themselves, apart from Dryred and Exotic wines. In addition, Recency is a feature that has no correlation whatsoever with the other variables, which made us think it is not so relevant to our study.

Before deciding on which variables to keep, we decided to go a step further and create two new log features based on income and LTV (logIncome and logLTV) to try normalizing the distribution of those variables. Also, we created YearsAsCustomers based on the dayswithus variable.

We also decided to run a PCA analysis just to have additional information on how the features are representing the variance in our data. Knowing that PCA is mostly used when you do not need to interpret the results after the clustering, we did it for visualization and to help inform our feature selection, but we did not keep the PCA variables themselves. The results are seen below:

	PC0	PC1	PC2	PC3
Dayswus	0.040290	-0.051379	0.977909	-0.034100
Age	0.918941	0.150754	-0.074001	-0.078705
Edu	0.170779	0.043030	-0.030285	-0.949248
Income	0.939448	0.147443	-0.063407	-0.086014
Freq	0.936884	0.119370	0.188058	0.067266
Recency	-0.320497	0.946747	0.028300	-0.001388
Monetary	0.922366	0.150965	0.194725	0.085408
LTV	0.886008	0.182777	0.127073	0.136357
Perdeal	-0.840430	-0.112512	0.051793	-0.017120
WebPurchase	-0.832556	-0.151341	0.108213	-0.116417
WebVisit	-0.701046	-0.158106	0.413313	-0.114655
YearsAsCustomer	0.040954	-0.051227	0.977706	-0.033200
logIncome	0.877821	0.126526	-0.076505	-0.184205
logLTV	0.909197	0.104686	0.040178	0.011403

- PCA0 is almost exclusively the combination of income/frequency/age/LTV, which all seem to be overlapping as we have seen in the correlation matrix.
- PCA1 is Recency, which represents a lot of variance in our data, however, we noticed that it has no correlation with any other variable, so it is not relevant for our analysis since this data by itself does not represent how a customer can be valuable or what preferences they have, and would strongly influence any clustering algorithm if included.

- PCA2 is YearsAsCustomer, also representing a significant proportion of our data.
- After PCA3, most information is not as useful since they do not represent a relevant variance in the data.

So given the PCA analysis and the correlation analysis, we will try clustering with a combination of the following:

Value features: logLTV/LTV, YearsAsCustomer, Education, WebVisit.

Preference Features: Dryred, Sweetred, Drywh, Sweetwh, Dessert, Exotic

Although some of the preference features are highly correlated, since we want to analyze customer preferences, it is important that we keep all six variables.

4 - Modeling

With our final variable selection, we moved on to clustering and analysis of our two groups of variables, value, and preference. We tested clustering using hierarchical clustering (HC) and K-means algorithms, as well as a combination of the two (K-means to HC). We found that K-means provided clusters with good separation of centroids. Given that it is one of the least computationally expensive algorithms, and was providing us with the best clusters, we decided to use this algorithm.

Using the elbow method of the inertia line plot for each clustering solution, as well as the silhouette scores, we determined 3 clusters per variable grouping was enough to sufficiently separate the customers, without becoming too cumbersome from the business perspective. It was at this point that the outliers were added back in, using a decision tree classification to predict the three sets of cluster labels, so that these customers could be included in the analysis.

5 - Evaluation

The following are the descriptions of each clustering, as well as the final merged clusters:

Value Clusters:

Cheap Millennials: These are your least valuable customers. They are here for the deal. They use the website almost half of the time they are shopping, the highest of the three clusters. They are the youngest group and are not spending much money. Give them a few decades, they could come around.

Habitual Visitors: These are the fans of the store. They like buying wines on sale, but not as much as your first cluster. They make purchases about two times in three months on average. They spend a decent amount of money when shopping, and use both the website and physical locations. You would probably recognize these customers when they come in.

Wealthy Wine Lovers: These are your best customers! Treat them well! They are older and prefer to come into the store to shop rather than purchase online. They are not afraid to spend some good money when they buy and do not seem to pay much attention to promotions.

Preference Clusters:

The Adventurous: Although having a slight preference for Exotic and Dry White wines, these are the customers who love to try new and different wines, they are very flexible with every kind of wine.

Sophisticated palate: These customers are simple and easier to please: they like their dry red wine. They do not like sweet wines, they might even try some exotic occasionally and dry white, but, by far, their favorite is the Dry Red Wine.

The Sour Palate: Their life is sweet enough. These are the customers who spend the most on dry wines, either white or red wines, they love having their dry wine and do not spend much on sweet and exotic wines.

Merged Clusters:

After merging the value and preference clusters, we were able to distinguish between four different clusters that are categorized into two major groups: customers with no clear preferences and customers with specific sophisticated tastes. Starting with the first group we identified this cluster:

The Pricetag Sommeliers:

These are youngsters who are open to trying every type of wine. Their flexibility when it comes to the choice of drinks gives them the advantage of sommeliers' knowledge and recommendations. Their young age is also witnessed in their website visits and purchases, as they chase deals and discounts. Despite their high focus on price, and low buying frequency, they are not the least valuable segment of customers.

As for the groups with the more particular tastes, they are older customers with an interest in dry wine specifically and are classified as follows:

The Sour Shoppers:

These customers show similar interest in white and red wine as long as it is dry, and a little bit of interest in exotic wines. They don’t buy that often or that frequently and seem to focus on visiting the website and chasing discounts. In fact, more than 42% of their purchases are bought at a discount. This has made them the least valuable client segment.

The Wolves of Wine Street:

These are the customers that are somewhat similar to the previous cluster regarding preferences. Their thing in common with the Sour Shoppers is their appeal to dry wine. They represent the oldest, wealthiest, most valuable customers, who don't seem to care about deals and discounts and are willing to spend any amount of money for the appropriate type of wine, as long as it is fine and dry.

The Cabernet Club:

This group favors dry red wine over any other type. They show a little interest in dry white wine as well, but dry red is on the top of their list by far. They are middle-aged people with relatively high incomes who are not cheap when it comes to buying wine but wouldn't mind finding a good deal and purchasing under a discount. They are not the highest spenders, but still valuable customers.

RFM Validation:

To further validate our merged clusters, we compared our RFM feature against our clusters. By ranking each customer's Recency, Frequency, and Monetary variables by their quartile rankings (1 if the customer falls in the lowest 25%, 4 if they are in the highest), and adding those scores together, we get a variable from the range 3-12 that can be used to measure value in our customers. When looking at the distribution of this variable by our four merged clusters we can see that our Wine Connoisseurs have the highest distribution, which further emphasizes their value as a group. The Cabernet Club group has the next best RFM score, as they know what they like, and are willing to pay a bit more for it. The Sour Shoppers and Pricetag Sommeliers both have low, relatively similar distributions of RFM scores. Given that our clustering was not done with any of the three RFM variables, this gives us good validation of the results. Our most valuable customers are seen in both analyses, while the least valuable only varied slightly, between the bottom two clusters.

6 - Deployment

Below are a few recommendations for different strategies using the customer information we have extracted from the data.

1 – Monthly Wine Club – Target the young, deal-oriented customers by offering monthly subscriptions where they can receive wine deliveries at home and exclusive deals/coupons. An automatic subscription will increase their frequency with passive purchases, and they will be attracted to the deals sent "just for them".

2 – A wine + food pairing event with a guest chef - For the older customers that tend to shop in stores, running interesting events at your store locations is a great way to maintain customer engagement. Wine Connoisseurs are your most valuable group, and are not interested in online deals but enjoy coming into the store. Having interesting events targeting wine drinkers with a more sophisticated palate will keep them engaged in your company.

3 – *If you like this, try this promotion* - Your third main cluster is medium value customers who are open to deals and promotions, so could be open to specific deals for wines related to their preferences. For example a list of popular dry red and white wines for the Sour Shoppers, and a list of popular dry reds for the Cabernet Club, with an occasional exotic wine recommendation.

Conclusion:

Overall, we were able to achieve the purpose of this project, which is to define and distinguish between the different types of customers of wonderful wines of the world. The clusters obtained were segmented first by engagement, which gives the company insights on the most valuable customers and their demographic and behavior. Second, we differentiated between different types of preferences in wine selection and purchase. Finally, the results of these two segmentations were merged to obtain four clusters that allowed us to get to know the type of clients better and offer some marketing recommendations to enhance the current customers' loyalty, and hopefully, attract more customers.

Appendix:

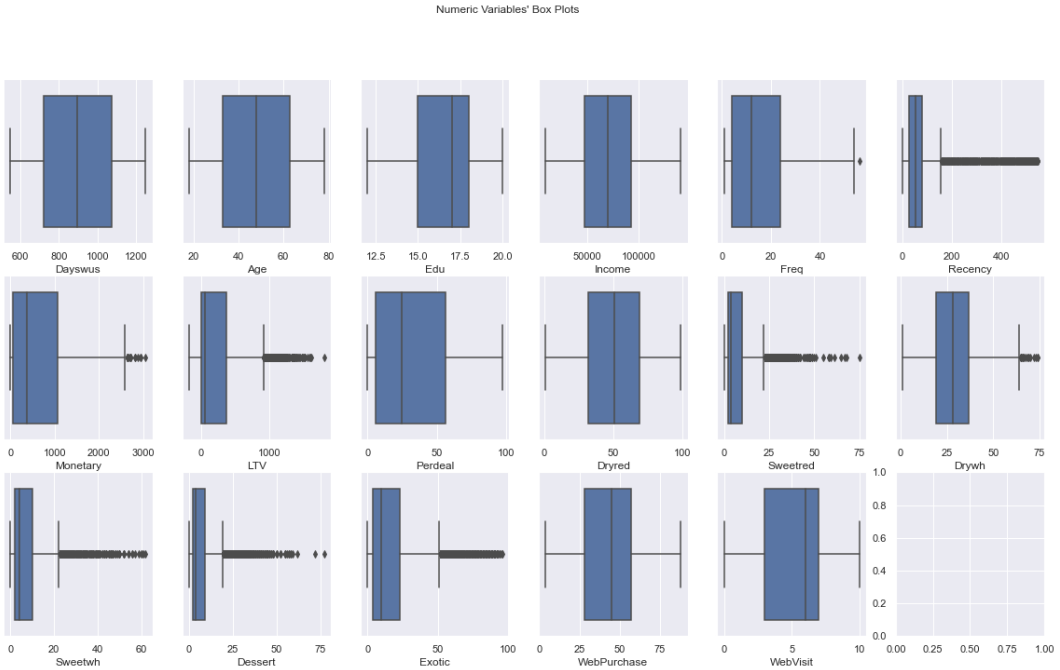

figure1: Variables' boxplots

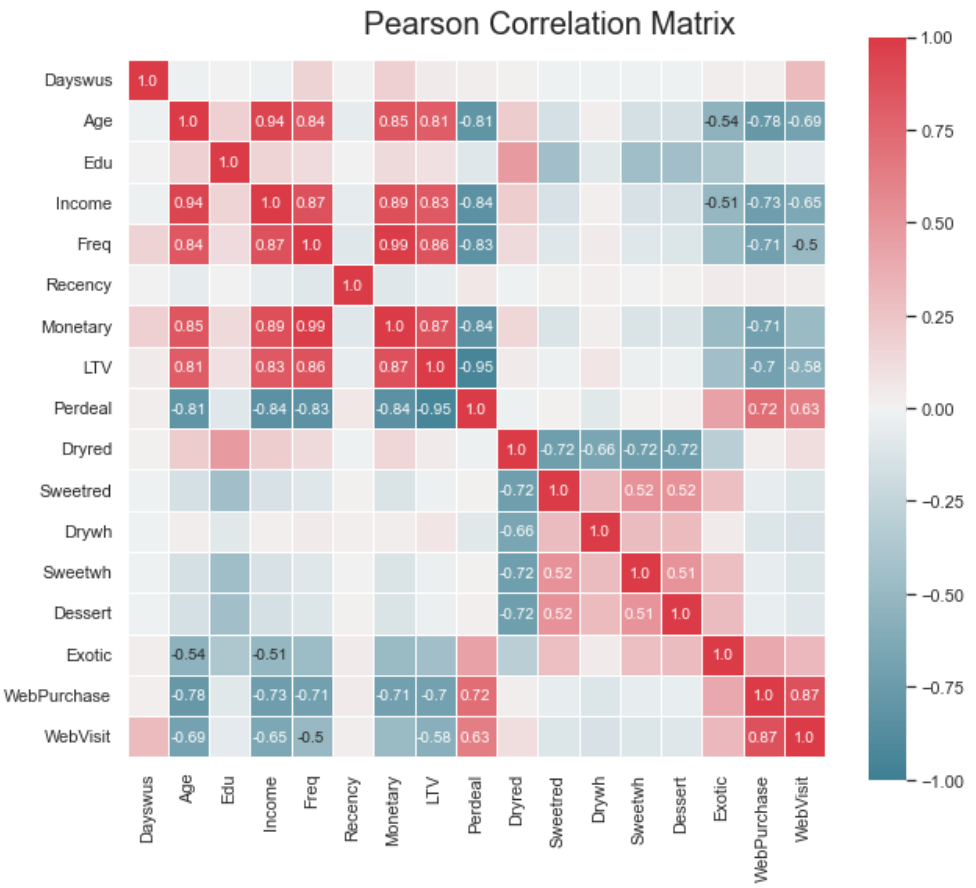

Figure2: Numeric variables correlation matrix