**A Classification of Wine**

CSC 367- Brian Artz, Jacob Gasyna, Steven Zhdanov

**I. Executive Summary**

The data analysis which classified “Vinho Verde” white wine into two categories, “Bad wine” or “Good wine”, contributed to a general understanding of what determines a wine’s quality. The analysis set out to find answers for five questions formulated by the research group:

1. Are the attributes able to sufficiently determine a quality rating for Vinho wines?
2. Are there any observable associations among the 12 attributes? Which attributes are closely associated with one another?
3. Is the alcohol content of a given wine strongly associated with a Vinho wine’s overall quality?
4. How consistent are the quality ratings of Vinho wines with similar attributes?
5. What are the common characteristics among highly rated (6 and above, “Good wine”) and lower rated (5 and below, “Bad wine”) Vinho wines, respectively?

The Vinho wine attributes recorded in the dataset, which did not contain any missing values (though there were significant outliers in several categories) were able to consistently provide a quality rating for every wine which was under evaluation. In other words, every wine contained within the dataset was given a quality rating. The attributes which evidently determined a Vinho wine’s quality rating were observed to be closely associated with one another. In particular, alcohol content was negatively associated with density meaning wines with higher percentages of alcohol were also less dense; Density was observed to be positively associated with the levels of residual sugar in a wine, meaning as sugar levels in wine rose so did its density; Most importantly, the alcohol content of wine was observed to be strongly associated with its quality rating which means the Vinho wines in this dataset with higher percentages of alcohol were given better quality ratings. One can reasonably infer from these associations that good wine has a high percentage of alcohol, low levels of sugar, and are typically not very dense. The measures of quality were moderately consistent among similar wines contained within the dataset, as per the 75% accuracy rating of the predictive model.

One anomaly observed within the results of the analysis was the significance of “Volatile Acidity” on wine quality. This attribute did not associate with any of the others, but became the second-best predictor for Vinho wine quality ratings (after Alcohol content), in particular for “Bad wine”. Upon research of Volatile Acidity and its effects on wine, we found that it is responsible for how rapidly a wine ferments. In other words, high levels of volatile acidity indicate the wine is especially prone to the fermentation process and is likely to have a sour taste (much like vinegar), which would imply that it tastes bad and will receive a low quality rating.

To conclude, “Good wine” has a higher alcohol content, a low density, low sugar level, and a low instance of volatile acidity. “Bad wine” has a lower alcohol content, greater density, high sugar level, higher levels of acids, and are more likely to rapidly succumb to the fermentation process. Due to the limitations set by our dataset, these results are generalizable only in so far as with white Vinho Verde wines.

**II. Data Description**

The data set contains 4898 records and contains 12 attributes. The data set originally contained a red wine data set, but we chose to remove it from the study because the results were inconsistent and unnecessarily complicated our exploratory analysis. The attributes the data set contain are:

* Fixed Acidity
* Volatile acidity
* Citric acid
* Residual sugar
* Chlorides
* Free sulfur dioxide
* Total sulfur dioxide
* Density
* pH
* Sulphates
* Alcohol
* Quality

Fixed acidity is all of the nonvolatile acids and is calculated by subtracting the volatile acidity from the total acidity. The volatile acidity is produced during the fermentation and malolactic stages of vinification. It is what adds to the aroma and the taste of the wine. Citric acid is an acid produced in the metabolism in the grapes. Residual sugar are the sugars in the grapes left over after the wine has been fermented. Chlorides add to the salty taste in wines and typical is based on the type of grapes used. Sulfur dioxide free and total is what prevents a wine from oxidation and microbial growth. Density typically dictates how sweet a wine is which will be more dense compared to more bitter red wines. pH in wine is a way to measure the ripeness of a wine. Low pH will have a more crisp and tart taste. Sulfates are what keeps the consistency of a wine. Alcohol is produced through the fermentation process. Quality is the rating given to the specific wine on a 1-10 scale.

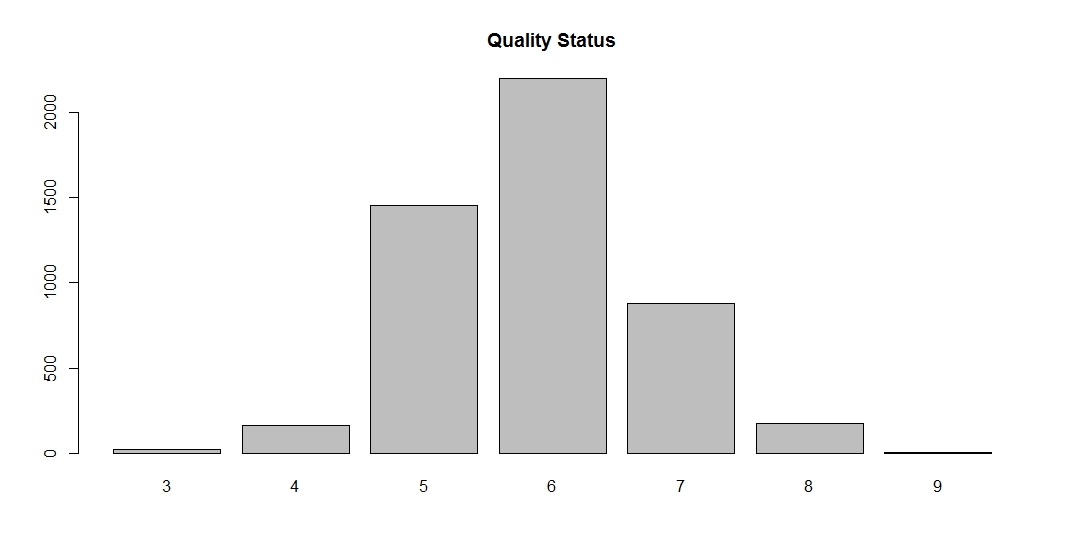
All of the variables are all quantitative except for quality which is a categorical variable by ordinal nature. The variables were acquired through physicochemical tests so the quantitative variables were measured accurately. The quality variable however is a subject variable which is important to keep in mind because one person’s rating for an average wine can be different from another. The goal of the data set was to see if wine producers can get hard numbers for producing good wines.

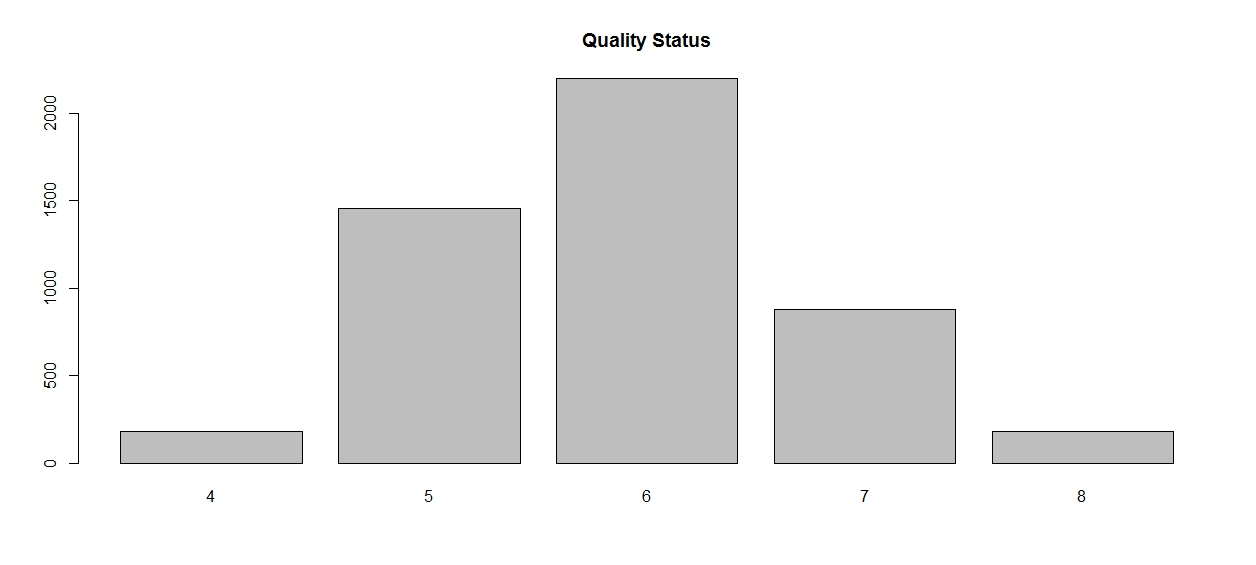
Our main focus on this data set will be to analyze the wines based on quality and see if we can find any general trends or commonalities between the different ratings of wines.

**III. Exploratory Analysis**

After analysis of the data we were able to find out many things about the data set. One of the first things that we figured out was that the data set contained no missing variables, which greatly simplifies our future classification and analysis because values do not need to be removed, randomized or imputed. We did however discover that our dataset contains an incredibly large amount of outliers in multiple variables, but we decided that no further action is required due to the fact that it does not seem like any of the data had been recorded or entered improperly.

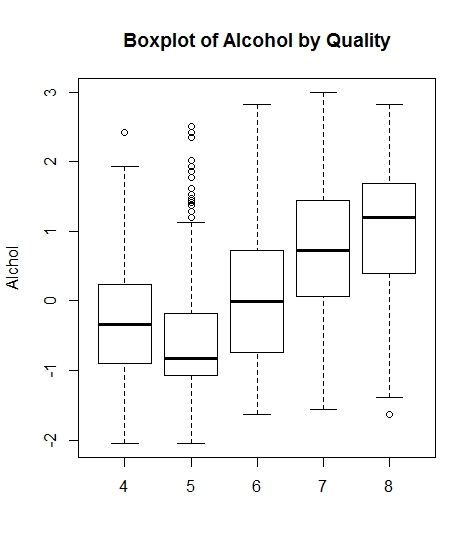
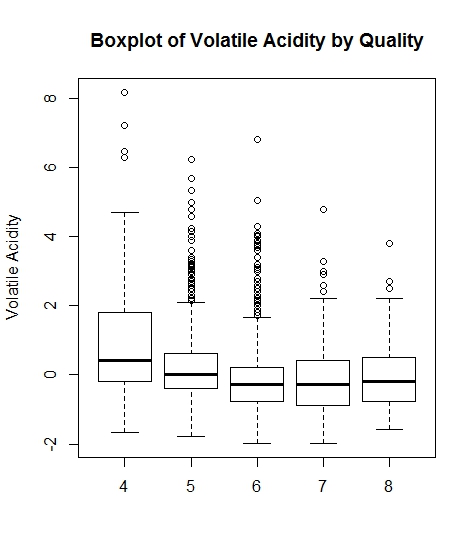
In addition to the presence of outliers in our dataset, the descriptive statistics show the high variance in scaling between variables. This led to the decision which standardized all variables but one: the quality rating. A separate analysis on the quality ratings revealed the ratings for a quality of 9 and a quality of 3 are quite minimal, and no ratings for a quality of 1 or 10. This led to the merging of quality 3 with quality 4; Quality 9 was merged with quality 8. The frequency of the data appears to be centered around a quality rating of 5 and 6, with the overall distribution of the data expressed as right-skewed.

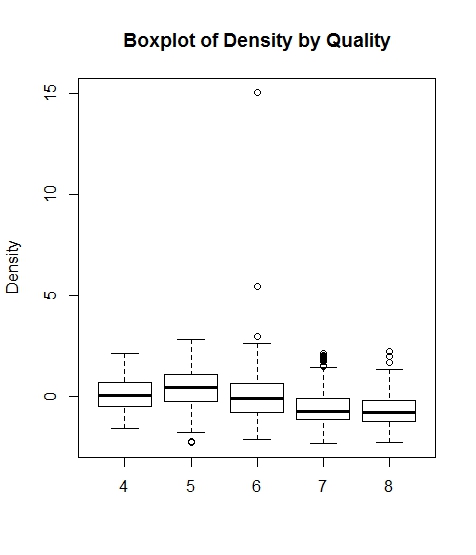
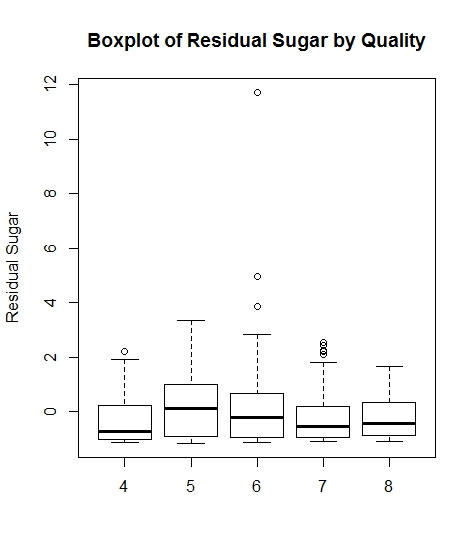




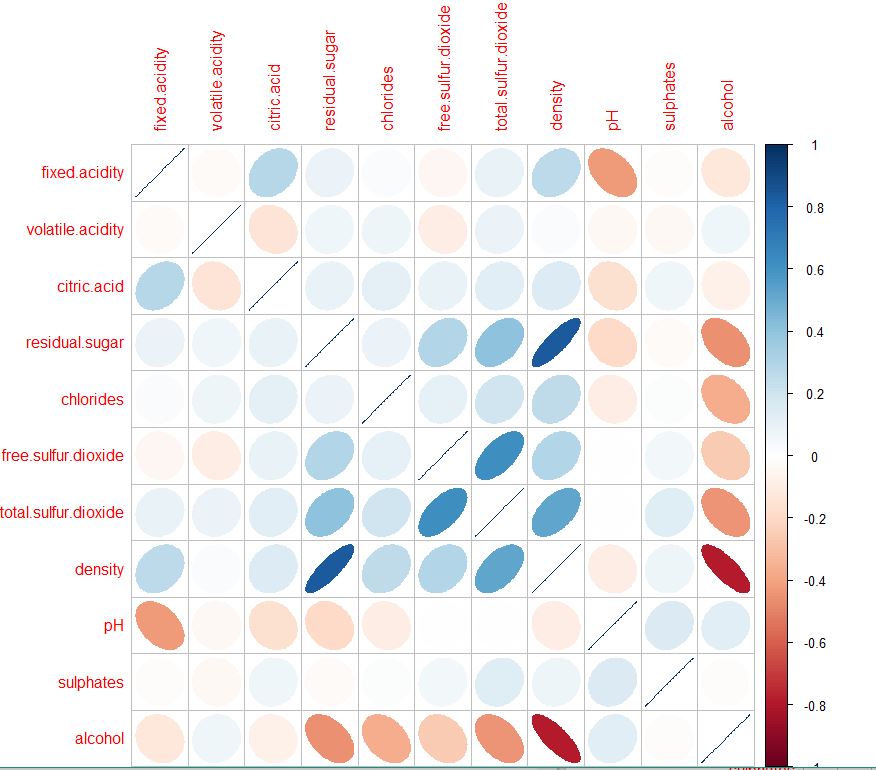
We then proceeded to use the variable quality as the metric for analyzing the distribution of qualitative variables, and deduced the following conclusions:

* The variables Fixed Acidity, Citric Acid, Residual Sugars, Free Sulfur Dioxide, Total Sulfur Dioxide, PH, Sulphates, Chlorides, and Density all show fairly similar medians and distributions across all qualities,
* The variables Alcohol and Volatile Acidity show a fairly substantial difference in medians and range across many of the qualities rating, Alcohol showing a stronger difference than Volatile Acidity.





Finally, we analyzed the linear associations among attributes of this dataset to determine the potential for classification data mining techniques. A scatterplot matrix of the standardized data was created to visualize the linear relationships, or lack thereof, among variables. The matrix clearly shows there are several attributes which are not linearly related. However it was hard to distinguish which variables were showing linear relationships, so a correlation matrix was used to better visualize the associations between variables.

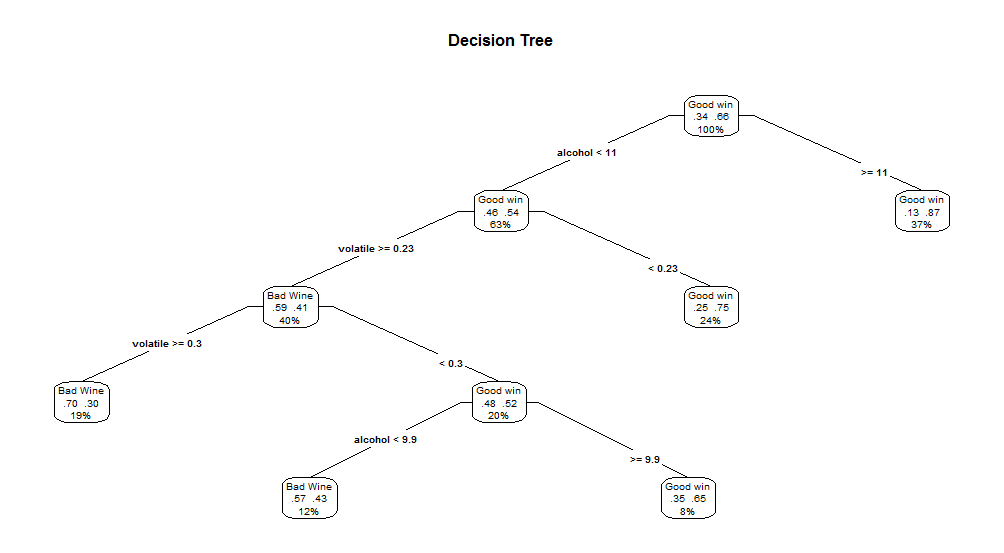


The variables which exemplified strong linear associations were found within density/citric acid, total sulfur dioxide/chlorides, residual sugar/density, alcohol/density, alcohol/residual sugar, and alcohol/total sulfur dioxide. The associations we view of particular importance in our analysis lie among those with alcohol, due to the identification of alcohol percentage as a strong predictor for quality rating. Most importantly are the associations of alcohol with density,strong negative, and residual sugar with density,strong positive. Density is at it lowest with high alcohol percentages (and low residual sugars) and highest with high residual sugar levels (and low alcohol percentage).

The route for Principal Component Analysis post-Exploratory-Analysis was considered, but ultimately ruled out after a KMO test resulted in a value below far below 0.6. This coupled with the other facts that we learned during this analysis suggested that we use decision tree classification.

**IV. Classification & Analysis**

Before creating our decision, we decided to convert data in the quality variable into binary data, to make the tree easier to read as well hopefully being more precise and conclusive. We decided that every wine that was ranked 6 or higher would be considered a good wine, and all the wines which was ranked 5 or lower would be considered a bad wine.



The decision tree that was created turned out less complex than we thought it would; it turns out that only 2 of the variables in our data are required for predicting whether or not a wine is considered good. The basic rules that can be deduced from the decision tree are that the wines with high Alcohol content and low Volatile Acidity are most likely going to be good wines, while wines with high Volatile acidity and low Alcohol content are most likely to be bad wines.

What is interesting is that even though there are only two variables which contribute to the prediction of what kind of wine is good or bad, many of the prediction have a very high level of confidence. For example a wine that has a an Alcohol content of over 11 has an 87% chance of being a good wine, and if the wine has an Alcohol content of under 11, and a Volatile Acidity greater than .3 has a 70% chance of being bad wine. Although wine quality is fairly subjective, it is clear that there are in fact certain variables which define what is a good wine and what is a bad wine.

**V. Validation & Testing**

We decided that cross-validation would be the most optimal solution to validating our model. The reason is because it will allow us to use the data we’ve already gathered and analyzed from the exploratory analysis to be placed into our model and then divided into subsets (the majority being our training set). So this is why we ended up using a confusion matrix with the binary variables of good and bad wine. Good wines being wine quality 6+ and bad wines being binned as 5-. The training set was ⅔ of the data set while the testing set was ⅓ .

* accurracy.png

When running the confusion matrix, we see that a good proportion of good wines are being labeled accurately, however there is a significant amount of good wines being mislabeled as bad wines. After doing a accuracy prediction calculation, we find that the model is 74.25% accurate. The model however, is 61.57% precise. So what we can infer from these results is that there is a good chance that if a wine taster was given a random wine, they will be able to identify it correctly as a good or bad wine ¾ of the time. This is pretty good provided that wine quality ratings are subjective. This means that although there is a subjective component to rating a wine, good wines have a common characteristic that is identifiable by wine enthusiasts.

The limitation to this model however would be the size of bins. There is a significantly higher amount of good wines versus bad wines, which might skew the accuracy prediction towards a higher rate. So a solution might be change the bad wines bin from 1-5 to 1-6, but the problem once against comes down to subjectivity. A quality of 6 to us was an above average wine.

**VI. Discussion of Results**

The results of the decision tree above show that alcohol content and volatile acidity are the two most important and influential properties of Vinho wine for its quality rating given by wine tasters. Perhaps the reason why alcohol content has been the most significant predictor is due to the indication of optimal ripeness and completion of fermentation which accompanies a higher alcohol percentage in wine. In addition, understanding how the levels of volatile acidity affect the quality of the Vinho wines is important. Volatile acidity has a direct effect on the taste of wine. It is responsible for rapid fermentation, and is accompanied by higher levels of sugar in the wine itself. The process of fermentation, in the case of higher levels of volatile acidity will lead to a wine tasting like vinegar. Taste testers, producers, and casual drinkers alike will not appreciate this sour flavor in their wines so it comes as no surprise that volatile acidity is a strong predictor for poor quality ratings.

A few key items to keep in mind when processing the results of the analysis:

1. The data collected and used in this study applies solely to Portuguese “Vinho Verde” **white** wines. While there are numerous variants of the Vinho white wines, the results of the study can not be applied to any other white wine vintners. In addition, the results of this study will certainly not apply to any red wine, Vinho reds or otherwise, because the data for Vinho red wines or any other red wine was not analyzed.
2. Quality ratings are subjective by nature. The assigned quality to any given Vinho white is not to be understood as an absolute rating. To be clear, each wine evaluated in this study may not have received the same quality rating if it had been rated by another wine taster. The dataset did not include any information about the wine tasters (experience, number of tasters, criteria for ratings, etc.), so we must rely on the credibility of the data source (University of Minho, Portugal) to make our analysis. The quality ratings did appear to be consistent as the normal distribution of the data showed. However, several attributes did contain significant outliers which had to be standardized for the analysis to progress.
3. The bins in which this analysis applied to the quality ratings were also subjective. This research group agreed upon appropriate ratings for what a “Good” wine could possibly be rated (6-10) and, conversely, what the range for a “Bad” wine should be (1-5). It is entirely possible that our judgement could have been too broad, and that wine experts would think any wine given the rating of 6 out of 10 could certainly never be considered “Good”.

**VII. Final Thoughts**

Our research concluded that good Vinho Verde white wines contain higher contents of alcohol, lower levels of sugar, are less dense, and have low instances of volatile acidity; The bad Vinho white wines have lower alcohol percentages, are much more dense, have higher levels of sugar and acids, and are at a higher risk of rapid fermentation. While these results are interesting and generalizable for the dataset, they do not really give us a complete understanding of good vs. bad wine. Our group are by no means wine experts, most of us have only just reached the point in our lives where we’re legally allowed to consume alcohol. Therefore, it would be interesting to conduct a new study which seeks to determine the qualities of good and bad wine. Ideally, the results could be more generalizable and the data collected could be more complete (i.e. more wine variants, but still separate the white from red wines.) Wines could come from all different vintners from around the globe, harvested in different conditions, fermented at different rates, and assigned quality ratings by an army of experienced wine tasters. It would be of much greater use if the results of this study could be more precise in its predictions, meaning the researchers could run a classification and/or clustering analysis with the intention of receiving multivariate outcomes.

**VIII. Sources**

(1) Bannon, Anne Louise. "The Perils of Volatile Acidity." *WineMaker Magazine*. N.p., n.d. Web..

(2) Cortez, Paulo, Antonio Cerdeira, Fernando Almeida, Telmo Matos, and Jose Reis. "Modeling Wine Preferences By Data Mining From Physicochemical Properties." *Decision Support Systems* 47.4 (2009): 275-76. Web.

(3) Gregutt, Paul. "Wine Q & A." *The Seattle Times*. The Seattle Times, n.d. Web. 17 Nov. 2016.

(4) "UCI Machine Learning Repository: Wine Data Set." *UCI Machine Learning Repository: Wine Data Set*. N.p., n.d. Web.

(5) "Vinho Verde." *Wikipedia*. Wikimedia Foundation, n.d. Web.