**Introduction to Data Science**

**ISCG 8026**

**Semester 1, 2020**

****

**Assignment-Part B**

**White Wine Data Analysis**

**Priya Payyadi Surendran**

**ID:1533656**

**Supervisor: Miss Natt Nuntalid**

**Declaration:” Wherever this assignment draws on the work of others, such sources are clearly acknowledged.”**

**ABSTRACT**

Wine industry depicts a spurt in growth as there is a rise in social drinking. Wine appreciation by wine tasters decides the price of wine to an extent. Physicochemical tests of wine considers the factors like acidity, density, sugar level and presence of other chemicals. This is also a key factor in assessing the wine quality and wine certification. This report analyses the physicochemical factors which determines the white wine quality through exploratory data analysis. The dataset consists of approximately 5000 Portuguese’s “Vinho Verde” white wine samples . If chemical properties of wine can be linked to human quality of tasting, it will control the quality assurance and wine certification process and it would be of great interest to the wine market. Finally, the analysed data is interpreted for future researches.

**Table of Contents**

[**SECTION 1** 5](#_Toc43078408)

[**1.** **Introduction** 5](#_Toc43078409)

[**1.1 Research background** 5](#_Toc43078410)

[**1.2 Data analysis topic** 5](#_Toc43078411)

[**1.3 Literature review on related data analysis methods** 6](#_Toc43078412)

[**SECTION 2** 6](#_Toc43078413)

[**2. Methodology** 7](#_Toc43078414)

[**2.1 Flowchart of data analysis process** 7](#_Toc43078415)

[**2.2. Explanation of each module in data analysis** 8](#_Toc43078416)

[**2.3 Difficulties and Solutions** 9](#_Toc43078417)

[**SECTION 3** 10](#_Toc43078418)

[**3. Experiments and Results** 10](#_Toc43078419)

[**3.1 Dataset Introduction** 10](#_Toc43078420)

[**3.2 Experimental Setup** 11](#_Toc43078421)

[**3.3 The procedure of experimentation** 11](#_Toc43078422)

[**3.4 Result Interpretation** 25](#_Toc43078423)

[**SECTION 4** 26](#_Toc43078424)

[**4. Conclusion** 26](#_Toc43078425)

[**Appendix A** 27](#_Toc43078426)

[**Appendix B** 27](#_Toc43078427)

[**References** 28](#_Toc43078428)

# **SECTION 1**

## **Introduction**

### **1.1 Research background**

The first white wine was produced in present-day Iran approximately 7500 years back. In ancient Greece, Hippocrates, the famous Greek physician prescribed it as medicines to his patients. The romans considered wine as delicious and was included in the wealthiest feasts. Wine and viticulture declined with the fall of the roman empire. Later, the catholic church secretly started viticulture and was a time of great reputation in making of white wine. Again, wine became popular among people but in 1943, the wine market took a down curve. White wine was created around the world by mid-sixteenth century, and it endures to be the most prevalent wine used up till date (The White Wine Primer, 2020). White wine can be made from any variety of grapes like chardonnay, Sauvignon, Riesling etc. The white wine is made by removing the skin of the grapes during fermentation process. White wines are not white in colour but straw or yellow coloured. Types of white wine includes dry, sweet and fortified, sparkling wine. White wine has become one of the most prevalent fermented beverages in the world.

The intention of the research is to evaluate qualities of white wine variants, establishing some correlation between quality parameters of wine and determined elements.

### **1.2 Data analysis topic**

This report examines a data set of nearly 5000 instances of white wine from the UCI repository and consists of 6000 instances for combined white and red wine types of Portuguese “Vinho Verde”. The scope of the analysis is to understand the relationship of many features which impacts the quality rating of wine. It is a complex and subjective task to judge a wine. The inputs involved objective tests like density, alcohol content, sugar content etc and the output is on sensory data i.e. evaluations made by wine experts. A minimum of 3 experts gave each wine in this dataset a quality rating from 0(poor) to 10(excellent). Balance and depth are two standards used by experts to judge wine quality. Balance is the relationship between the fundamental traits of the wine : sweetness, sourness, bitterness and alcohol. Each trait shouldn’t overpower the other. Wine has depth when it doesn’t taste flat or plain. Alcohol is produced as a part of fermentation process . Our dataset contains 12 attributes including density, acidity, pH, alcohol etc. Goal of the report is to ascertain which attribute is associated with high quality rating and which are defects of a wine.

### **1.3 Literature review on related data analysis methods**

At present, many industries uses product quality certifications in order to promote their products. The mentioned process is assessed by human experts and is time consuming and expensive (Chen , Rhodes , Crawford, & Hambuchen, 2014).

In (Gupta, 2017), to determine wine quality in two ways on both red and white wine dataset, machine learning techniques are used . First dependency of target attribute on independent attribute is determined by using linear regression and important attributes are selected. Neural network and support vector machines are used to predict the values of dependant attributes. The paper proves that if only selected attributes are considered instead of entire attributes, better prediction can be made.

In (M. Gambetta, Cozzolino, Bastian, & W. Jeffery , 2016),the composition aspects of Chardonnay produce from berries of different regions are examined. For 3 diverse regions in South Australia, different sensory attributes were recognised through descriptive analysis which helped for finding attributes relating to high and low wine quality. Correlation between different attributes like free volatiles, fatty acids in berries of different level of qualities and their compositions were identified. High berry concentrations of zinc, pH levels and behenic acid were related to higher wine quality.

In (R.Kowalski & Frank, 1984), to model the relationship between subjective sensory evaluation and objective chemical measurements, a method called multivariate regression was used to Pinto Noir wine samples. Wine classification according to their geographical origin, descriptive and predictive models were used. The position of inorganic attributes were studied in these forecast models.

In (Cortez, Cerdeira, Almeida, Matos, & Reis, 2009), a data mining approach to find taste preferences . Wine samples of white and red wine samples of Portugal is considered and 3 distinct regression techniques were applied and out of which support vector machine showed promising results than neural networks and multiple regression methods. This can help oenologists for wine evaluation and improving wine production.

# **SECTION 2**

## **2. Methodology**

### **2.1 Flowchart of data analysis process**

A close up of a device

Description automatically generated

Figure 1: Data Analysis Flowchart

### **2.2. Explanation of each module in data analysis**

* Define your Questions: The question is to ascertain attributes of wine which can determine high quality wine and thus these selected attributes can be given importance to speed up the quality certification and to improve the wine quality.
* Clear Measurement priorities: Physicochemical factors like acidity, density, sugar level and presence of other chemicals are measured, and sensory data like quality of wine assessed by experts and a quality rating from 0(poor) to 10(excellent) by them.
* Data Collection: Data is collected according to the framed question and is collected from already existing source (Cortez, 2020). Two datasets are available for each red wine and white wine samples of the Portuguese “Vinho Verde” .
* Data Cleansing: The data collected was almost tidy. There were no missing values. The duplicate values of approximately 940 rows were removed. A new factored variable called rating was created considering the quality of the wine and was rated in to 3 groups namely bad, average, good. Another variable called label was created to group the alcohol concentration in the wine to 3 groups light, medium and strong.
* Data Analysis:
* Univariate Data Analysis (Univariate Analysis, 2020): In this analysis, preliminary exploration of white wine data set is analysed. To describe pattern in univariate variables, central tendency methods like mean, mode, median , dispersion methods like range, variance, maximum, minimum, quartiles and standard deviations and visualization methods like box plot, histogram are used on all the attribute of the wine dataset. Frequency distribution and percentage of sensory output like quality, rating and alcohol concentration is calculated. Bar chart is used to visualize these sensory data.
* Bivariate Data Analysis: In this module, frequency distribution and percentage of variables quality and rating is calculated. Boxplots between output factor quality and other attributes created to study the relation. To find the relation between all the different attributes, a correlation matrix is plotted. Scatter plot and linear regression method are done on the correlated variables .
* Multivariate Data Analysis: To analyses certain kind of patterns of an average and good quality wine has, three meaningful and correlated data variables are used.
* Result Interpretation: Major attributes of wine are ascertained which can determine high quality wine and thus these selected attributes can be given importance to speed up the quality certification and to improve the wine quality.

### **2.3 Difficulties and Solutions**

The major difficulty was in collecting data with all the requirements and understanding each attribute. The selected data set contains only physicochemical and sensory variable due to privacy issues and no data about grape type, brand of wine, market price or selling price of wine. After classifying , the classes were imbalanced because most of the wine samples are average rather than poor or excellent. There were many physicochemical variables and was not certain if all input variables are appropriate.

For the imbalanced classes to detect poor or excellent wines outlier detection algorithms can be used. Correlation matrix helped me find the major important variables affecting the quality of wine. Including or being able to access the variables like type of grape, brand of wine will help to determine the high-quality wine in future.

# **SECTION 3**

## **3. Experiments and Results**

### **3.1 Dataset Introduction**

The dataset consists of white wine samples with 12 attributes and around 4898 instances collected from the Portuguese “Vinho Verde” wine in 2009.It contains 11 physicochemical attributes as input and one sensory attribute which is the quality output.

The input attributes are:

* Fixed Acidity(g/dm3): It is the Tartaric acid from the samples and fixed acidity means non-volatile and will not evaporate easily.
* Volatile Acidity(g/dm3): It is presence of acetic acid which cause vinegar like taste which is unpleasant if present in high amount.
* Citric Acid (g/dm3): If present in low quantity can cause wine to taste fresh and flavoured.
* Residual Sugar(g/dm3): This is the sugar level present after the fermentation. Minimum quantity present will always be one gram per litre and sugar level above 45g/L is considered sweet.
* Chlorides(g/dm3): The quantity of sodium chloride or salt .
* Free Sulfur Dioxide(mg/dm3): It helps to prevent oxidation as well as microbial growth of wine.
* Total Sulfur Dioxide(mg/dm3): It is the total of bound and free sulfur dioxide. It becomes apparent in the taste and smell of wine when there is higher level of sulfur dioxide.
* Density(g/cm3): Wine’s density depends on quantity of alcohol and sugar and can become close to that of water.
* pH: Acidic level of wine can be determined by checking this value. Most of the wine’s pH is between 3-4 on the scale. The most acidic will be 0 and least acidic will be 14 on the pH scale.
* Sulphates(g/dm3): It is mainly potassium sulphate and is added as an additive to boost sulfur dioxide level to improves wine’s antioxidant and antimicrobial process.
* Alcohol(percentage by volume): The content of alcohol in wine is produced during fermentation process.

The output attribute is:

* Quality: A score between 0 to 10 given by 3 wine experts.

### **3.2 Experimental Setup**

The experiment is to analyse white wine data using univariate and bivariate data analysis. In this experiment, r language is chosen, and R studio is used with enhanced library functions to visualize the data analysis. The library packages imported includes:

* library(readr) to read CSV File
* library(ggplot2) plotting package based on grammar of graphics for Data visualization
* library (gridExtra) for laying multiple plots on a page
* library(GGally) for scatter plot matrix visualization
* library(dplyr) for efficient manipulation of dataset
* library(reshape2) to flexibly reshape data
* library(RColorBrewer) provides colour palettes for graphing
* library(e1071)for calculating skewness and kurtosis
* library(corrplot) for graphically displaying correlation matrix

### **3.3 The procedure of experimentation**

**3.3.1 Data Import**

Dataset is loaded into R Studio and assigned it as “Wine\_List”.

A close up of a keyboard

Description automatically generated

Figure 2:First view of the dataset

Looking at the summary of the dataset to get a better understanding of the attributes

A close up of text on a white background

Description automatically generated

Figure 3: Summary of dataset

The data import and summarization procedure is shown in Appendix A.

**3.3.2 Data Cleansing and transformation**

The dataset is already tidy with null missing values. There were around 937 duplicate rows which were removed. The unit representation of free sulfur dioxide and total sulfur dioxide was converted to g/dm3 from mg/dm3 to make all the unit representation same. Two new variables names as rating and label. Rating is to group the wine samples in relation to the quality. Quality value below 5 is assigned bad, below 7 is average and 8 above is good. The label variable is to group the wine sample according to the alcohol content. Alcohol content up to 9 is light, up to 12 is medium and above 12 is strong. The data cleansing procedure is shown in Appendix B

**3.3.3 Data Analysis**

**Univariate Data Analysis**

* Frequency: The frequency of the quality, newly created variable rating and label is calculated.

|  |
| --- |
|  |

* Percentage: The percentage of quality, rating and alcohol content of the wine samples were calculated.

|  |
| --- |
| A screenshot of a cell phone  Description automatically generated |

It can be observed that most of the quality of the wine samples are average and the content of alcohol in the wine samples are medium i.e. between 9 to 12.

* Location, Spread and Shape
* Some wines samples do not have citric acid content.
* Mean of quality is 5.878, max and min quality are 9 and 3 respectively with a median of 6.
* Around 75% of the wines have a residual sugar value below 10 grams/litre
* The percentage of alcohol in wine samples varies from 8 and 14.20 percent.
* pH with a median of 3.180 and varies from 2.720 to 3.820.
* There is a huge range of sulfur dioxide among the wine samples.

The methods and values used for analysing location, spread and shape is shown in the below table.

A screenshot of a cell phone

Description automatically generated

* Bar Chart : Bar chart of quality, rating and alcohol content in wine samples is plotted.

A screenshot of a cell phone

Description automatically generated

Figure 4: Bar chart of Quality, Rating & Alcohol content

Quality distribution shows unimodal normal distribution with 6 as centre. Most of the white wine samples are having average rating and the alcohol content is medium in most of the samples.

* Histogram and boxplot of all the input variables

A screenshot of a video game

Description automatically generated

Figure 5: Histogram & boxplot of Fixed Acidity

Perfect normal distribution with some outliers at both ends. Fixed acidity peaks at 7 and most values range between 5 and 9.It may contribute to pH of wine.

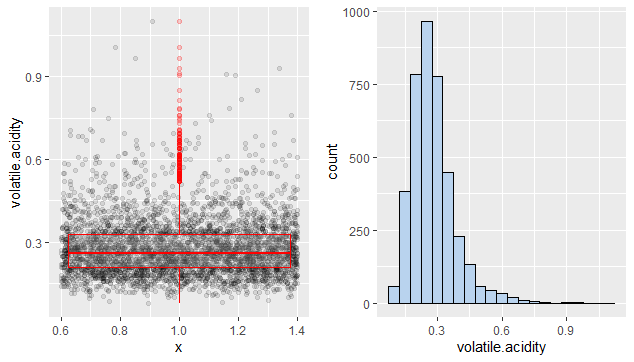


Figure 6: Histogram and Boxplot of Volatile Acidity

Volatile acid distribution is right skewed meaning most of the wine samples are containing less amount of acetic acid which improves wine quality. Higher the amount of acetic acid, bitter the taste.

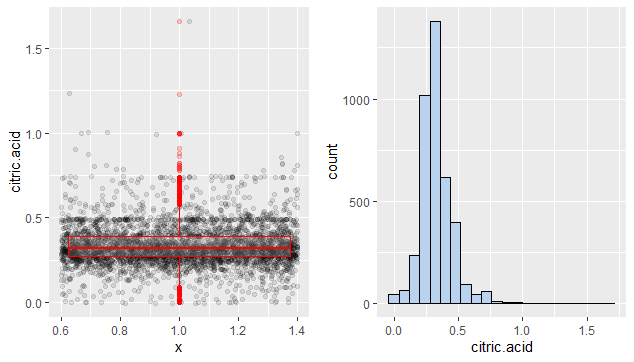


Figure 7: Histogram and Boxplot of Citric Acid

It peaks at 0.3 and since it makes wine fresh, it adds up to the quality of wine.

A screenshot of a computer

Description automatically generated

Figure 8: Histogram and Boxplot of Residual Sugar

The residual sugar is extremely left skewed. This can be fixed by changing into log distribution form. The log10 scale representation shows that residual sugar is bimodal and is peaked at 3 and 10.

A close up of a device

Description automatically generated

Figure 9:Histogram of Log10 Residual Sugar

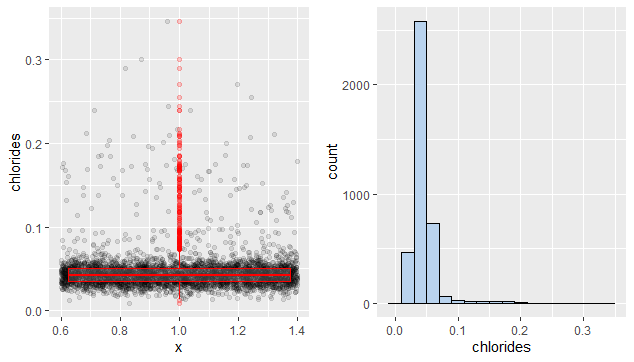


Figure 10: Histogram and Boxplot of Chlorides

A screenshot of a cell phone

Description automatically generated

Figure 11: Histogram of log10 Chlorides

Majority of value is between 0 and 0.3 and peaks at 0.5. Lesser quantity of chlorides is preferred in wine.

A picture containing room

Description automatically generated

Figure 12: Histogram and Boxplot of Free Sulfur Dioxide

The level of free sulfur dioxide must be below 50ppm and 3/4th of the wine sample contains free sulfur content of 46 which is good for quality.

A screenshot of a cell phone

Description automatically generated

Figure 13: Histogram and Boxplot of Total Sulfur Dioxide

This distribution of is normal with few outliers and it peaks at 125.

The density peaks at 0.95 and majority of it lies between 0.9 and 1.

A screenshot of a cell phone

Description automatically generated

Figure 14: Histogram and Boxplot of Density

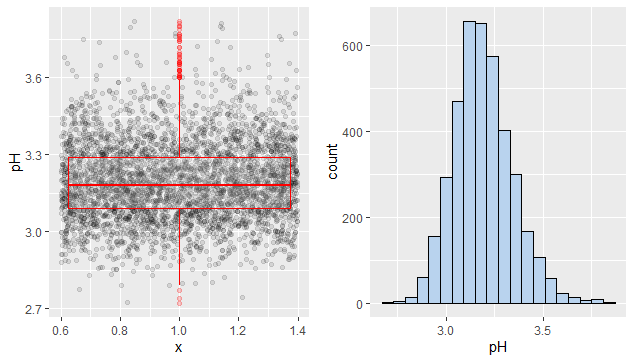


Figure 15:Histogram and Boxplot of pH

Wines are acidic in nature and pH distribution is normal with peak value at 3.2.

A close up of a map

Description automatically generated

Figure 16: Histogram and Boxplot of Alcohol

Alcohol distribution is multimodal distribution and peak value is between 9 and 9.5 and as well as at 10 and 11.Most of the wine samples contain medium alcohol content.

Density, Sugar content , sulfur dioxide and alcohol will help to improve the quality of the white wine.

**Bivariate Data Analysis**

* Frequency : Frequency of quality rating and the alcohol level of the wine samples are calculated.

|  |
| --- |
| A screen shot of a person  Description automatically generated |

It is evident that major wine samples are having average quality with medium alcohol content.

* Correlation: Correlation matrix helps to find meaningful relation between attributes. Attributes with correlation greater than 0.4 is considered meaningful in this analysis.

A close up of text on a white background

Description automatically generated

Figure 17: Correlation Matrix

The main correlated variables are density-alcohol, density-total sulfur dioxide, residual sugar-alcohol , density-residual sugar. Citric acid , volatile acidity and sulphates have no identified correlation with any of other variables. Highest correlation is between density and residual sugar which is positive and density and alcohol which is negative. The sugar is used during fermentation process to produce alcohol and hence the negative correlation and an inverse relation between residual sugar and alcohol.

A screenshot of a cell phone

Description automatically generated

Figure 18: All plot matrix between every attribute

* Boxplot: Since quality attribute is the key area of interest, all attributes are plotted against quality using boxplot.

A screenshot of a map

Description automatically generated

Figure 19:Boxplot between Quality and Alcohol

Good quality wines have strong alcohol content.

A screenshot of a social media post

Description automatically generated

Figure 20: Boxplot between density and quality rating

Good quality wines have lower density.

A screenshot of a cell phone

Description automatically generated

Figure 21:Boxplot between residual sugar and quality

No relation between residual sugar and quality of wine as discussed in univariate data analysis.

A screenshot of a cell phone

Description automatically generated

Figure 22:Boxplot between sulfur dioxide and quality

* Scatter Plot and linear regression: The main attributes under consideration are density, residual sugar, alcohol , total sulfur dioxide.

A close up of a map

Description automatically generated

Figure 23:Scatter plot and linear regression of density and alcohol

Alcohol and density are negatively corelated hence the inverse relation.

A close up of a map

Description automatically generated

Figure 24:Scatter plot and linear regression of residual sugar and density

Residual sugar is strongly and positively correlated with density which in turn affects the quality. Residual sugar also affects alcohol since alcohol and density are negatively correlated. Sugar is consumed to produce alcohol during fermentation. So, as the sugar decreases, alcohol content increases in wine.

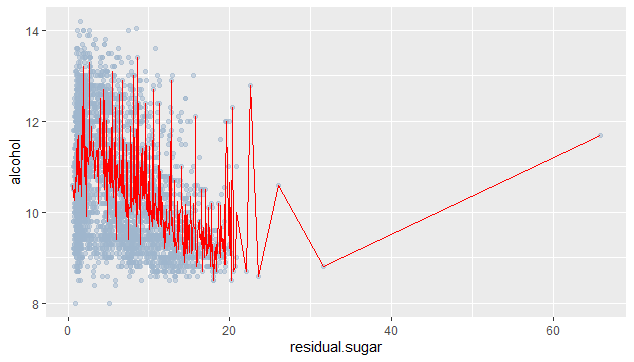


Figure 25:Scatter plot of residual sugar and alcohol

A close up of a map

Description automatically generated

Figure 26:Scatter plot and linear regression of total sulfur dioxide and density

Large quantity of Sulfur dioxide increases the density of the wine as well as will be evident in smell and taste.

**Multivariate Data Analysis**

There is a strong relation between alcohol , density and wine quality.

A close up of a map

Description automatically generated

Figure 27:Multivariate Data Analysis

Higher quality wines have less density and strong alcohol content.

### **3.4 Result Interpretation**

Major physicochemical attributes affecting wine quality are density, alcohol, sulfur dioxide. Volatile acid like acetic acid should be less since it causes bitterness to wine. There should be an optimum quantity of citric acid as it add freshness to wine. Lesser level of chlorides is preferred in wine. Wine density should be less and high amount of sulfur dioxide will increase the density. Maximum sulfur dioxide content should be up to 50. Alcohol in wine should be strong to improve the quality. As alcohol content increase, it decrease the density and the sugar content of wine.

# **SECTION 4**

## **4. Conclusion**

The wine quality is not only restricted to chemical components but the experts who rated the white wine quality in the dataset preferred wines with more alcohol, less sugar, less density, less volatile acid, sulfur dioxide and chlorides. Bivariate Analysis helped me to understand in depth the relation between different attributes. Boxplots of attributes by quality helped me to understand the major related attributes and as predicted in univariate data analysis, sugar and quality had no relation . From this analysis common behaviour of high-quality wine can be extracted.

Major limitation was the quality scale of 10.The quality rating was given between 3 to 9 and most of the rating was between 5 to 7 and made most of the samples as middle quality. This restricted the data analysis as there were less of distributed quality values and most of the wines are of average quality. Current sample consists of wine from Portugal region. It will be interesting to analyse wine samples from various other regions to eradicate the biased quality analysis. Majority of the attributes have weak correlation with the quality with an exception of alcohol. Including more variables like grape type, price of the wine will also provide a strong feature for analysing the quality. Quality is integer and if we convert it to factor, certain plotting cannot be performed.

A linear model for quality prediction didn’t perform as expected as the process of finding the quality of wine is subjective. The experts can be biased in qualifying the wine samples. In future a different set of models like logistic or multiple regression model can be used to find the best features which contributes to the high-quality wine.

# **Appendix A**

A screenshot of a cell phone

Description automatically generated

# **Appendix B**

A screenshot of a cell phone

Description automatically generated

# **References**

*A Beginner’s Guide To Data Science*. (2020, June 12). Retrieved from Towards Data Science: https://towardsdatascience.com/a-beginners-guide-to-data-science-55edd0288973

Chen , B., Rhodes , C., Crawford, A., & Hambuchen, L. (2014). Wineinformatics: Applying Data Mining on Wine Sensory Reviews Processed by the Computational Wine Wheel. *2014 IEEE International Conference on Data Mining Workshop.* Shenzhen, China: IEEE.

Cortez, P. (2020, June 11). Retrieved from http://www3.dsi.uminho.pt/pcortez

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 547-553.

*Exploratory data Analysis*. (2020, June 12). Retrieved from R for Data Science: https://r4ds.had.co.nz/exploratory-data-analysis.html

Gupta, Y. (2017). Selection of important features and predicting wine quality using machine learning techniques. *ICSCC 2017-18.* Kurukshethra, India.

M. Gambetta, J., Cozzolino, D., Bastian, S., & W. Jeffery , D. (2016). Towards the Creation of a Wine Quality Prediction Index: Correlation of Chardonnay Juice and Wine Compositions from Different Regions and Quality Levels. *Food Analytical Methods, 9*, 2842-2855.

R.Kowalski, B., & Frank, I. (1984). prediction of wine quality and geographic origin from chemical measurements by parital least-squares regression modeling. *Analytica Chimica Acta*, 241-251.

*The White Wine Primer*. (2020, June 10). Retrieved from Gentleman's Gazette: https://www.gentlemansgazette.com/white-wine-primer/

*Univariate Analysis*. (2020, june 10). Retrieved from Science Direct: https://www.sciencedirect.com/topics/economics-econometrics-and-finance/univariate-analysis