DATA ANALYSIS PROJECT I

RED WINE QUALITY

GROUP 6

15332-MALEESHA NILUMINDA

15378-SHANALIE RANASINGHE

15669-GANESHI UMAYANGANA

ABSTRACT

Maintaining and growing a customer base in the current competitive industry depends heavily on the quality of the product. The market is overflowing with different varieties of wine, and the wine industry is no different. "How do we pick the best one?" is a question that comes up frequently. This research focuses on the chemical characteristics of red wine in order to answer this specific topic.

We performed a thorough analysis of the Red Wine quality dataset from the UCI Machine Learning Repository using R software. The objective is to identify the critical chemical elements influencing the quality of red wine. The results are intended to assist customers and winemakers in making knowledgeable decisions about the selection and production of wine.

"A bottle of wine contains more philosophy than all the books in the world" - Louis Pasteur

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Wine can be both very simple and incredibly complex. It's an alcoholic drink made by fermenting grape juice. Most wine as we know, is made with grapes, but it can technically be made from other fruits too, such as apples, blueberries, and strawberries.

Why have grapes become the standard? There are two main reasons. Grapes contain acids; malic, tartaric, and citric acids that preserve the wine, allowing it to be aged for decades or even centuries. Secondly, grapes have a much higher sugar content than other fruits, which allows them to ferment so successfully and produce complex wines.

Wine is the widely consumed beverage globally, and its values are considered important in society. The quality of wine is always important for its consumers, and mainly for producers in the present competitive market to raise their revenue.

Historically, wine quality used to be determined by testing at the end if the production; to reach the level, one already spends lots of time and money. Every person has their own opinion about taste, so identifying a quality based on a person's taste is challenging.

With the development of technology, manufacturers started to rely on various devices testing in development phases. So, they can have a better idea about wine quality. This helped in accumulating lots of data with various parameters such as quality of different chemicals and temperature used during the production, and the quality of wine produced. One can adjust the variables that directly affect the wine's quality during this process. This provides the producer with a greater understanding of how to adjust various development process parameters to optimize the wine's quality. Additionally, this might produce wines with a variety of flavors and, ultimately, might create a new brand. Therefore, it is crucial to analyze the fundamental factors that affect wine quality. In this work, we have demonstrated how machine learning (ML) may be used to predict wine quality by determining the optimal parameter that determines wine quality.

2. DESCRIPTION OF THE QUESTIONS

Our main goal in this exploratory analysis is to better understand the complex relationship that exists between wine's physiochemical characteristics and its perceived quality, which is represented by the qualitative variable "Quality" and is scaled from 0 to 10. Acknowledging the industry's critical need for wine quality evaluation, we aim to accomplish the following main goals:

1. Identification of Impactful Physiochemical Properties
2. Determination of Optimal Physiochemical Levels
3. Development of a Predictive Model for Quality Assessment

While acknowledging that personal preferences may occasionally deviate from the objective standard of quality, our goal in doing this analysis is to promote a more objective approach to wine quality prediction. Our goal is to give the wine industry a strong foundation for improving the quality of their goods and enabling better informed decision-making by breaking down quality evaluation into quantifiable and explicable variables.

3. DESCRIPTION OF THE DATASET

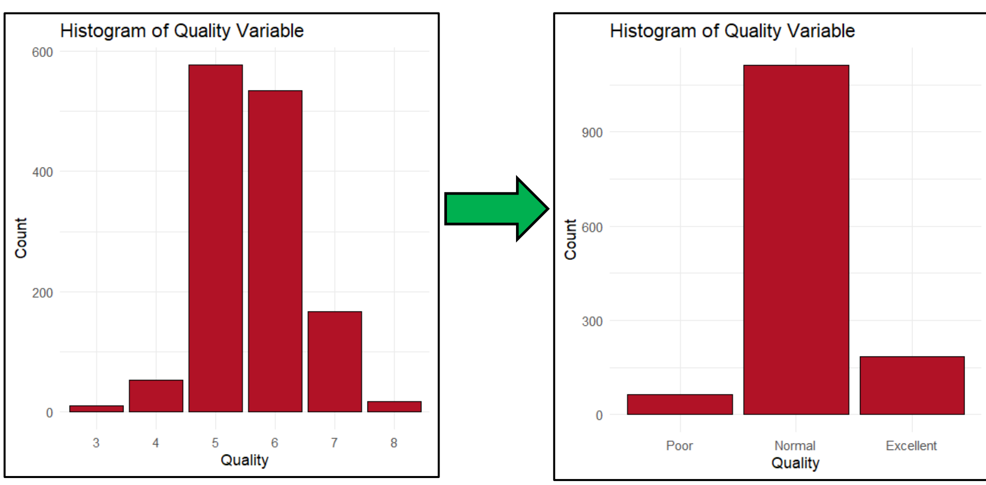
This Red Wine Quality Dataset was taken from Kaggle. It contains 12 different properties of wine. One of them is "Quality" which is the response variable for the study and the remaining variables are based on physicochemical factors. There are 1599 observations produced in the Vinho-Verde region of Portugal.

Variable	Description	Data Type
Quality	Based on sensory data (a score between 0-10)	Categorical
Fixed acidity	These are non-volatile acids that do not evaporate readily(g/dm3)	Numeric
Volatile acidity	are high acetic acid in wine which leads to an unpleasant vinegar taste(g/dm3)	Numeric
Citric acid	Acts as a preservative to increase acidity (small quantities add freshness and flavor to wines) (g/dm3)	Numeric
Residual sugar	The amount of sugar remaining after fermentation stops. The key is to have a perfect balance between — sweetness and sourness (wines > 45g/dm3 are sweet)	Numeric
Chlorides	The amount of salt in the wine(g/dm3)	Numeric
Free sulfur dioxide	So2 is used for the prevention of wine by oxidation and microbial spoilage(mg/dm3)	Numeric
Total sulfur dioxide	Is the amount of free and bound forms of SO2(mg/dm3)	Numeric
Density	The density of wine is close to that of water depending on the present alcohol and sugar content. (g/dm3)	Numeric
pH	the level of acidity-free Sulfur Dioxide: it prevents microbial. Basic wine is on a scale from 0(very acidic) to 14(very basic)	Numeric
Sulphates	A wine additive that contributes to SO2 levels and acts as an antimicrobial and antioxidant. Which preserves freshness and protects wine from oxidation and bacteria(g/dm3)	Numeric
Alcohol	Percent of alcohol present in wine (% by volume)	Numeric

Table 1 - Variable Description

4. MAIN RESULTS OF THE DESCRIPTIVE ANALYSIS

1. Response variable – Quality of wine



Excellent	13.5394%
Normal	81.8249%
Poor	4.6358%

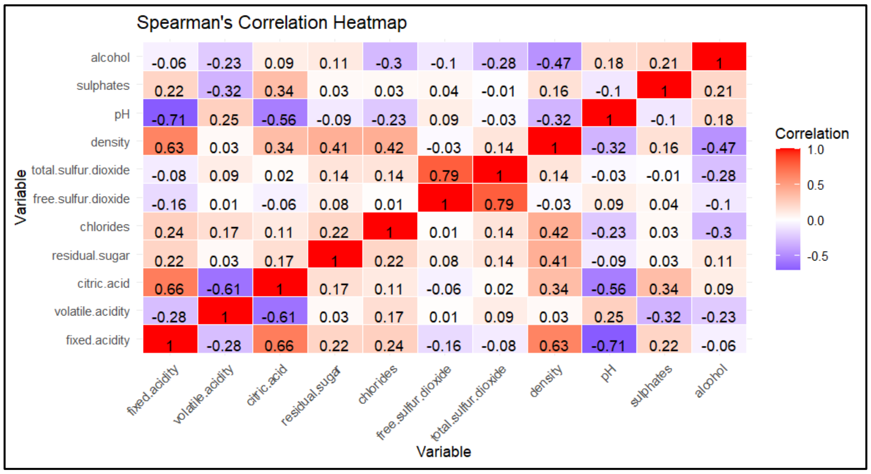
Figure 4.1- Histogram of Wine Quality

The "quality" variable assigns a rating, ranging from 0 to 10, to each red wine sample as determined by an expert in the field. Their categories have an order. So quality is an ordinal categorical variable.

From the histogram, the unique data in column "quality" is 3–8 which is a rating of wine's quality. Here most of the observations had fallen under 5 and 6, while a very low count was observed for 3 and 8. This result is determined by sulfate level, pH level, acid level, and also personal preference.

Hence for the convenience of further analysis, the quality variable was recorded as Poor (1-4), Normal (5-6), and Excellent (7-10). Using the new percentage-based values for model development helps reduce bias compared to using the original rating system. Based on the updated percentages, most wines were still categorized as "Normal," with fewer falling into the "Excellent" or "Poor" ratings.

2. Correlation Plot of variables



Variable	Correlation
fixed.acidity	0.1331152
volatile.acidity	-0.3300082
citric.acid	0.2276544
residual.sugar	0.0481449
chlorides	-0.1410433
free.sulfur.dioxide	-0.0300366
total.sulfur.dioxide	-0.0998603
density	-0.1279824
pH	-0.1042078
sulphates	0.3106701
alcohol	0.3399701

Figure 4.2 - Correlation plot of the dataset & and Spearman's results

The predictor factors exhibit notable positive and negative correlations.

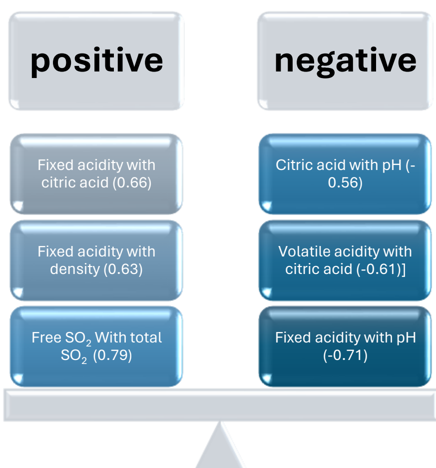


Figure 4.3 - Correlation between predictors

Correlations between predictor variables and response variables

❖ Total SO2, free SO2, and residual sugar all have a very weak, nearly nonexistent relationship with wine quality.

❖ The amount of alcohol exhibits the strongest link with wine quality, followed by volatile acidity and sulfates, which all exhibit significant correlations.

❖ The correlation table shows that higher-quality wines have lower density and lower chlorine levels.

3. Acidity of Wine

In examining the acidity components of the wines, including Fixed Acidity, Citric Acidity, and Volatile Acidity distributions are positively skewed. This means there are outliers, and those outliers of the distribution are further out towards the right.

Fixed acidity is the set of natural acids in wine that remain in a liquid when it's boiled. Box plot of Fixed Acidity by Quality (figure 4.5) illustrates that high quality wines tend to have higher Fixed Acidity compared to other two; normal and poor, box plots. The variability among the quality categories is moderate, as evidenced by overlapping IQRs. Nevertheless, a subtle but consistent increase in median Fixed Acidity is observed with improving wine quality.

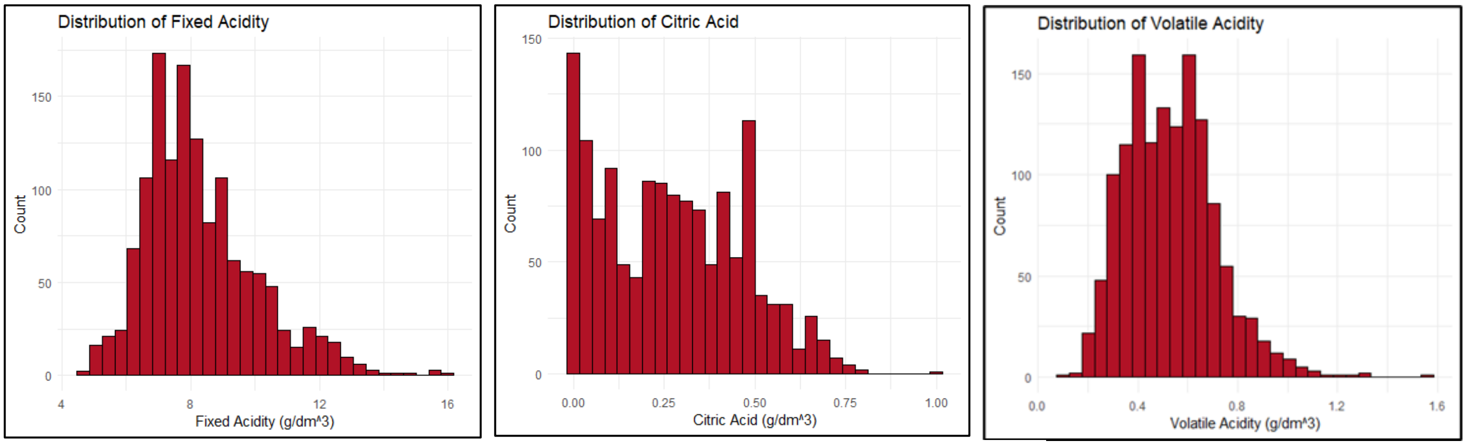


Figure 4.4 - Histogram of Fixed, Citric and Volatile acids

Citric acid is a weak organic acid that is less commonly found in wine. It is often added to wines after fermentation to increase their total acidity. Citric acid found in small quantities, citric acid can add 'freshness' and flavor to wines. Figure 4.5 reveals that high quality wines demonstrate a discernible increase in citric acidity compared to lower quality categories. As quality increases there is a significant increase in the median of citric acid.

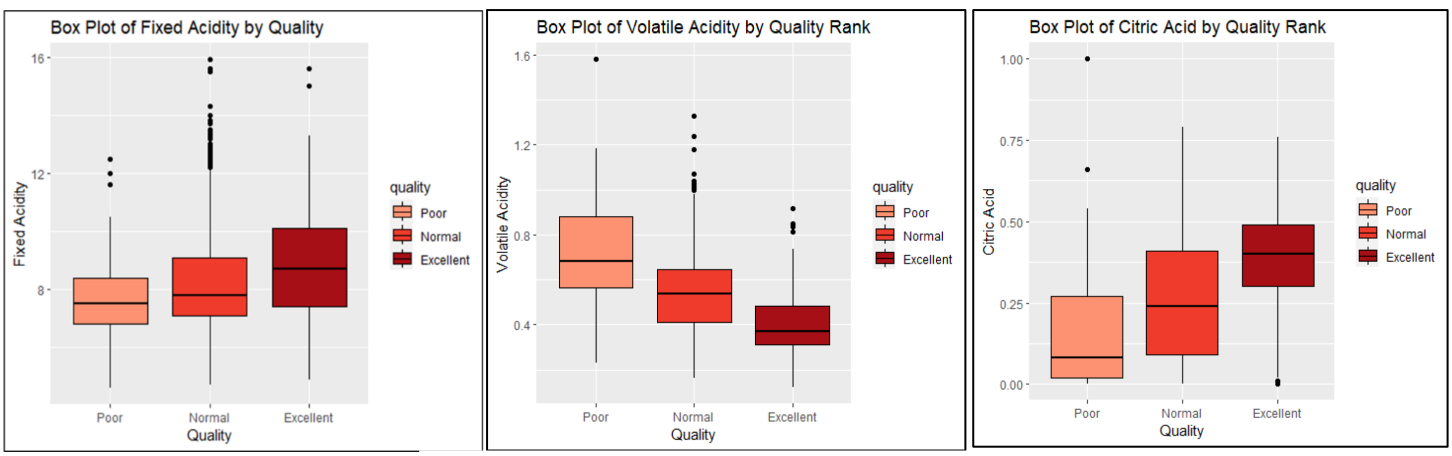


Figure 4.5- Boxplot of Fixed, Citric and Volatile acids

Volatile acidity (VA) is a measure of the wine's gaseous acids that contribute to the smell and taste of vinegar in wine. High quality wine has lower VA compared to lower quality wines. The box plot of VA by quality rank (figure 4.5) shows a clear trend of decreasing VA as quality improves.

Volatile Acidity Vs Citric Acid

The plot of Volatile Acid vs Citric Acid shows a negative correlation. The reason for negative correlation is, in winemaking, citric acid concentrations can increase the concentration of diacetyl. Citric-sugar co-metabolism can also increase the formation of volatile acid in wine. However, excessive levels of volatile acid can negatively affect the wine aroma.

Fixed Acidity Vs Citric Acid

Citric acid is a fixed acid found in small quantities in wine therefore Fixed Acidity Vs Citric acid (figure 4.6) shows a positive correlation.

Fixed acidity, volatile acidity, and citric acid have outliers. If those outliers are eliminated distribution of the variables may be taken to be symmetric.

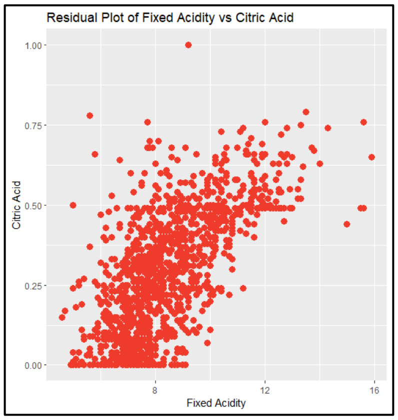


Figure 4.6- Fixed Acidity Vs Citric

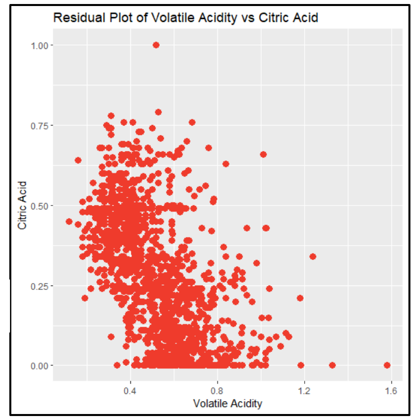


Figure 4.7- Volatile Acidity Vs Citric Acid

4. pH

When measuring pH, there are readings that extend from 0-14. Anything below 7.0 is considered acidic, while every reading above 7.0 is known as basic or alkaline. The acidity in wine will affect its freshness and microbial stability while also acting as a preserving agent.

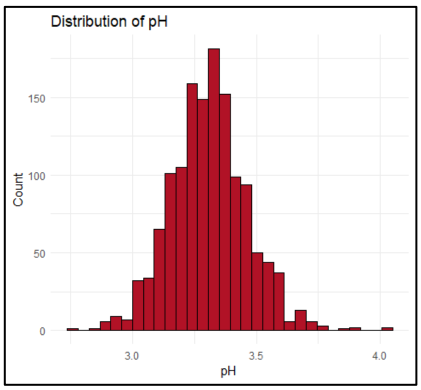


Figure 4.9 - Histogram of pH

Distribution of pH has a normal distribution. Here most of the wine data lies in between the 3.2 to 3.4 pH range.

Box plot of pH by quality rank shows that when quality increases the pH level decreases. High quality wine has 3.29 of average pH value.

Figure 4. 8 - Boxplot of pH

The wine pH affects red wine color. The flavor changes with pH as well. Low pH wines are fresh, bright, crisp, and possibly sour. High pH wines are round, soft, fatty, and possibly flabby. When it comes to flavor, we focus on acids rather than pH.

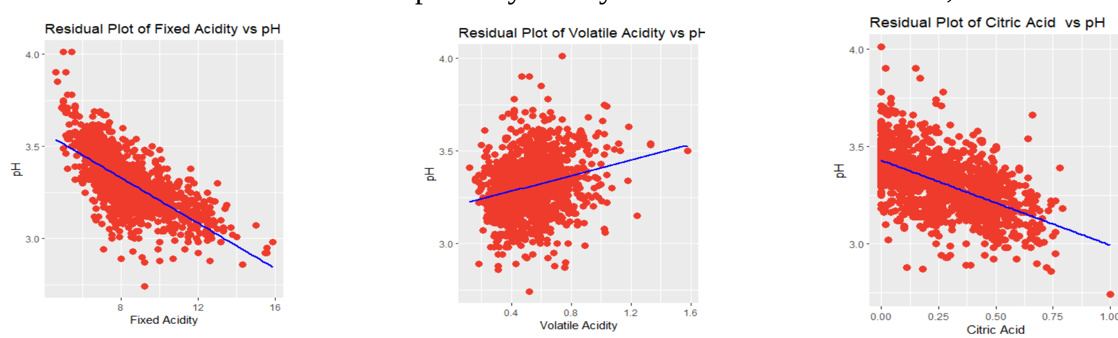


Figure 4. 10 - Scatter plots of pH

In examine fixed acidity vs pH and citric acid vs pH, there is a negative relationship (when acidity increases, pH decrease) because of the pH scale. When acid level increases pH level should be decrease, but positive relationship is observed in volatile acidity and pH. This unusual behavior may be due to the limited number of observations in the dataset. Higher citric and fixed acidity levels may lead to a sour taste, negatively impacting wine quality, while a moderate presence of volatile acidity, like acetic acid, can enhance complexity and flavor, resulting in a positive association with quality. (figure 4.9)

Table 2 - Summary of Acid Types

Fixed acidity 7g/l-10g/l	Volatile acidity 0.3g/l-0.5g/l	Citric acid 0.3g/l-0.5g/l	pH 3.2-3.4
Max fixed acidity :15.6g/l	Max volatile acidity: 0.9g/l	Usual range:0g/l-0.1g/l	Usual range:3.3-3.6

5. Sulfites and Sulphur Dioxide (Total/Free)

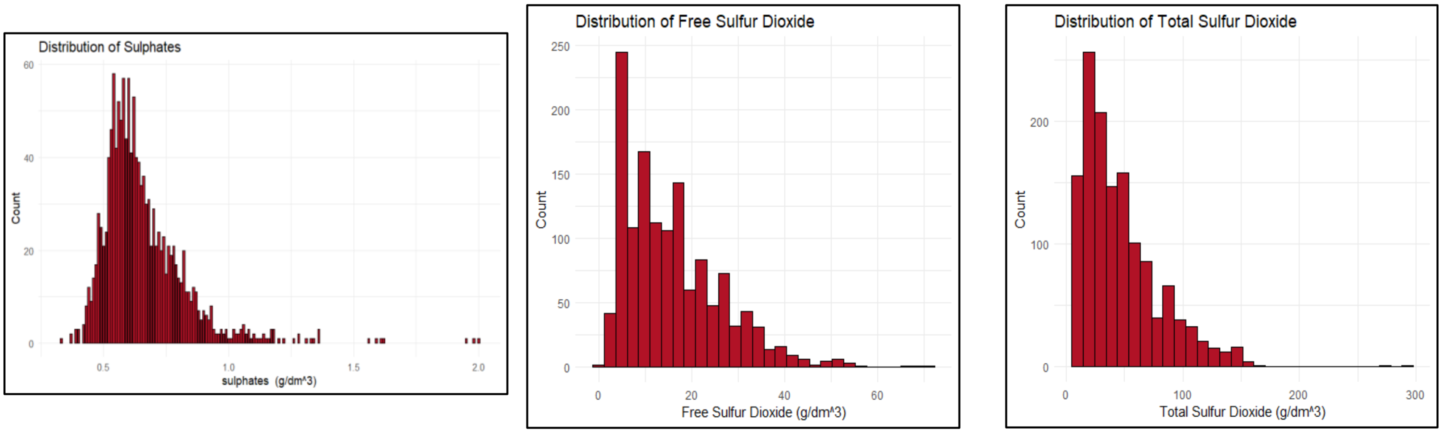


Figure 4.11 - Histograms of Sulphate and Sulfur Dioxide

By considering the histograms of sulfate, free SO_2, and total SO_2 all three variables are positively skewed suggesting the presence of outlying observations. Also, the sulphate variable has some outliers.

Figure 4.12 - Boxplot of Sulphate and Sulfur Dioxide

Sulphur dioxide (SO2) has an important role in the wine industry as an antioxidant, antioxidizes, and antiseptic additive. However, since SO2 is also responsible for allergic reactions, it is of great interest it replaces it with alternative additives or technologies.

In figure 4.12 Higher sulphate connections are found in high-quality wine, according to the boxplot figure that shows quality level vs. sulphate. Because the lowest and maximum values of an excellent box plot are higher than those of a poor or normal box plot. However, there is no evident correlation between total SO2 and free SO2. Comparing normal wine to the other two categories, it has higher mean total SO2 and mean free SO2 concentration.

Summary result of high-quality wine:

Sulphate: 0.65 - 0.82 (g/dm^3)	Free SO_2 : 6.00 – 17.00 (mg/dm^3)	Total SO_2 : 17.00 – 42.25 (mg/dm^3)
Maximum limit : 1.36 (g/dm^3)	Maximum limit: 54.00 (mg/dm^3)	Maximum limit: 289.00 (mg/dm^3)

Table 3 - Summary of Sulfites

Free sulfur dioxide has a few outliers, but these are very different from the rest.

6. Chlorides

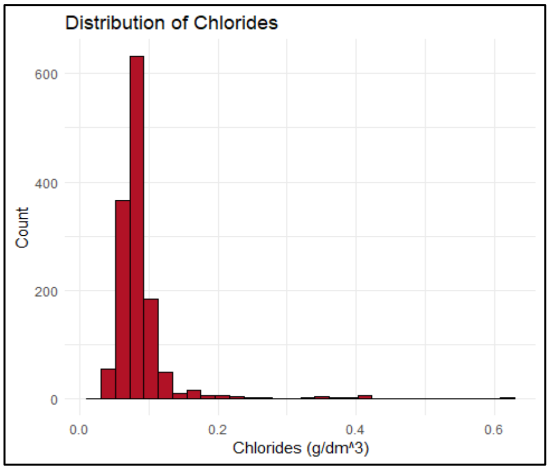


Figure 4.13 - Histogram of Chlorides

Chloride content makes the wine a little salty Both the terroir and the grape variety affect the quantity of chloride in wine, and measurement is crucial since this specific ion has a significant effect on wine flavor and, at large concentrations, gives the wine an unwanted salty taste.

The histogram indicates that the amount of chloride does not follow a normal distribution. most wine lying between 0.07 and 0.09 grams per liter. Concentration of chloride. The distribution is positively skewed, indicating that higher-quality wines have lower chloride concentrations. Additionally, we see a significant number of outliers in the Normal quality wines, which calls for additional investigation to address those outliers.

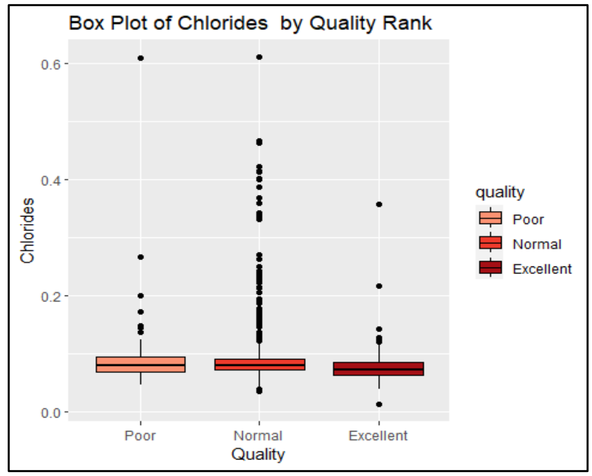


Figure 4.14 - Boxplot of Chloride

7. Density

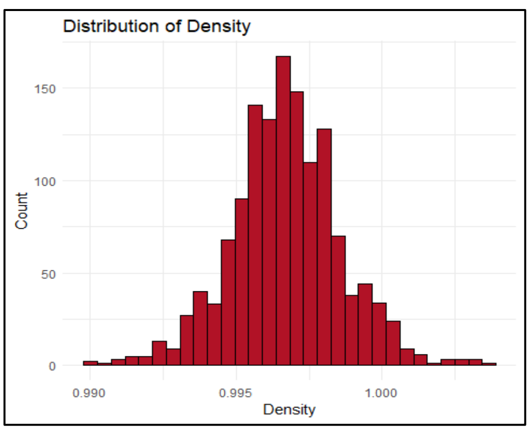


Figure 4.16 - Histogram of Density

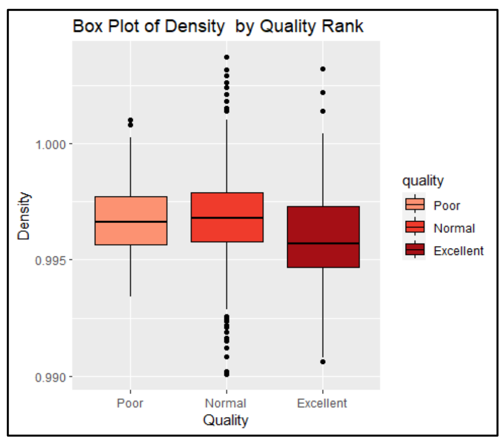


Figure 4.15 - Boxplot of Density

Distribution of wine density follows an approximately normal pattern. Analyzing the Density by quality Rank (figure 4.16 & 4.15) reveals that there is no direct association between density and wine quality. Lowest density observed for high quality wine (mean density for high quality wine is approximately 0.996), this inverse relationship doesn't extend uniformly to poor and normal categories. Remarkably, the mean and median densities for poor and normal wines are nearly identical, indicating that density alone may not be a decisive factor distinguishing these quality categories.

By taking a closer look at figure 4.18, it's apparent that the density of wine demonstrates an approximate negative correlation with alcohol content and slight positive correlation with residual sugar content. This suggests that the addition of sugar, alcohol, and other supplementary ingredients, aimed at enhancing the quality of the wine, may contribute to a decrease in the overall density of the liquid.

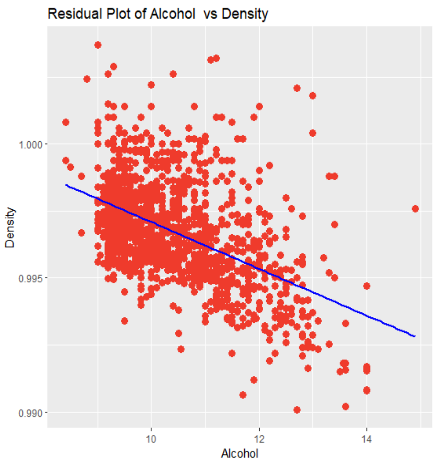


Figure 4.18 -Scatterplot of Alcohol Vs Density

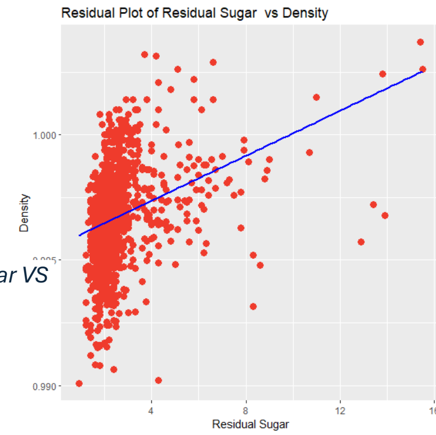


Figure 4.18 - Scatterplot of residual Sugar VS density

Density has a few outliers, but these are very different from the rest.

8. Alcohol level

Alcohol level affects a wine's body, texture, and taste. Wines with higher alcohol content are rounder, suppler, and sometimes denser or chewier than lower-alcohol wines. They also have a fuller, richer body and a slightly bitter taste.

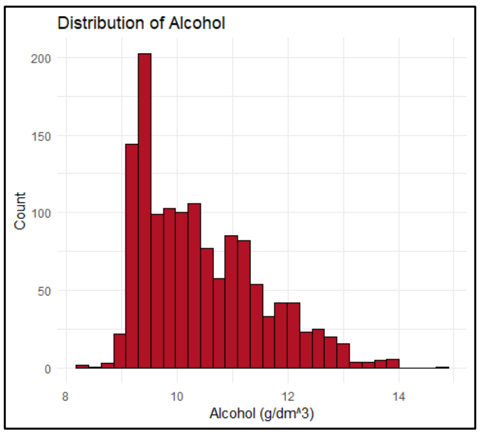


Figure 4.19 - Histogram of Alcohol

Alcohol content can also affect a wine's sweetness. Low-alcohol wines are typically sweeter due to the sugar leftover from the fermenting process.

Alcohol can also contribute to a wine's aromas. As alcohol levels increase during primary fermentation, compounds such as methyl butanoate, ethyl butanoate, and ethyl acetate are produced.

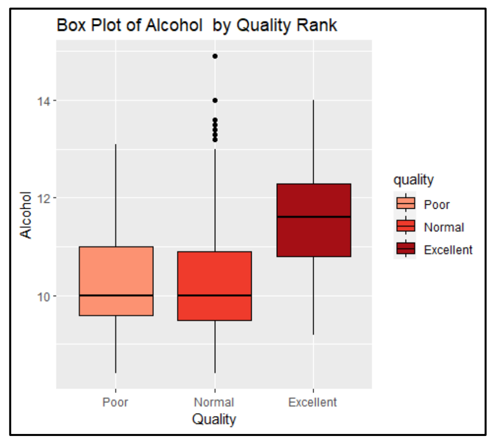


Figure 4.20 - Boxplot of Alcohol Vs Quality

Figure 4.19 shows that the distribution of alcohol is positively skewed. The majority of alcohol ranged between 9.5% to 11.1% and minimum alcohol level is 8.4%. That indicates that all wine samples in our dataset are alcoholic.

Figure 4.2 illustrates that high quality wine has an average of approximately 11.55% alcohol and the alcohol level of high-quality wine is greater than lower quality categories.

9. Residual Sugar

Residual sugar concentration is a measure of the amount of sugar solids in a given volume of wine following the end of fermentation and any sugar addition when making a sweet wine. Residual sugar concentration is expressed in grams per liter (g/L) or as a percentage of weight to volume.

The histogram indicates that the data exhibits a non-normal distribution with a steadily decreasing right side (right skewed) when the values on the x-axis represent all of the residual sugar concentration values. Most red wines have between 1.9 and 2.6 g/l. Excellent quality wines typically include a little bit more residual sugar than the other two categories.

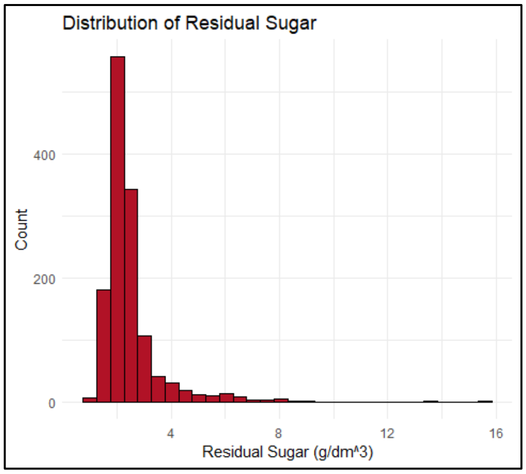


Figure 4.21 - Histogram of Residual Sugar

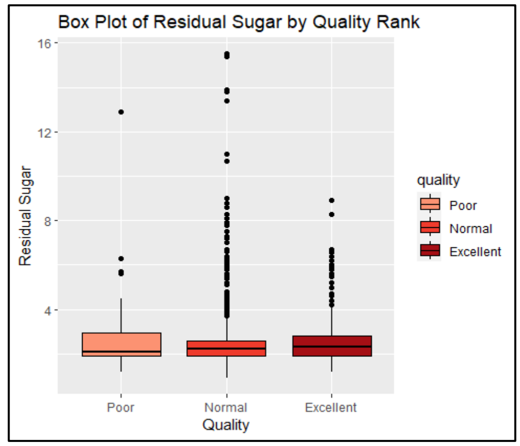


Figure 23 - Boxplot of Residual Sugar

mean_residualugar_poor	mean_residualugar_normal	mean_residualugar_excellent
<dbl>	<dbl>	<dbl>
2.68	2.48	2.70

Table 4 -Residual Sugar means relates to quality categories.

Without additional research, it is unlikely to determine the relationship between residual sugar and wine quality because the mean residual sugar level for normal quality differs greatly from the other two.

Since it is a known fact that sugar is the source of alcohol, obtain a scatter plot and look for any connections between the residual sugar content and alcohol level.

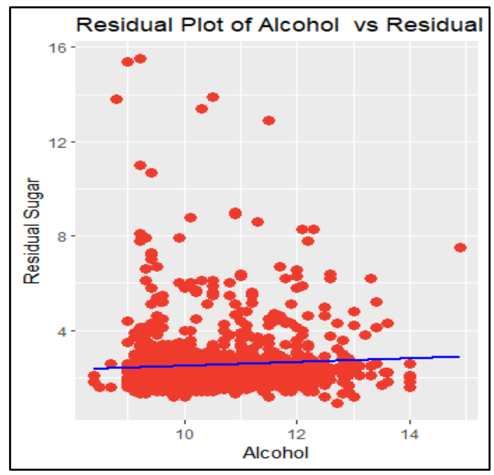


Figure 4.22 - Scatterplot of Residual Sugar Vs Alcohol

However, this plot does not demonstrate a strong association, and the matrix scatterplot likewise demonstrates a very low correlation; however, without further study, it is not possible to conclude that these predictors do not have a relationship.

10. Partial Least Square

To find any clusters among the observations and any strongly correlated predictors, partial least square regression was run on the data.

Figure 4.24, It is clear from the plot of Scores(X), that the observation set doesn't include any significant clusters. One cluster contains all of the observations. Therefore, there are no clusters in the observations on the wine's quality.

Here, the plot of loadings of XY (figure 4.25) suggests that while some predictors and responses have high correlations, other predictors are also orthogonal to the response. This instance of connection among observations highlights the dataset's absence of clustering once more.

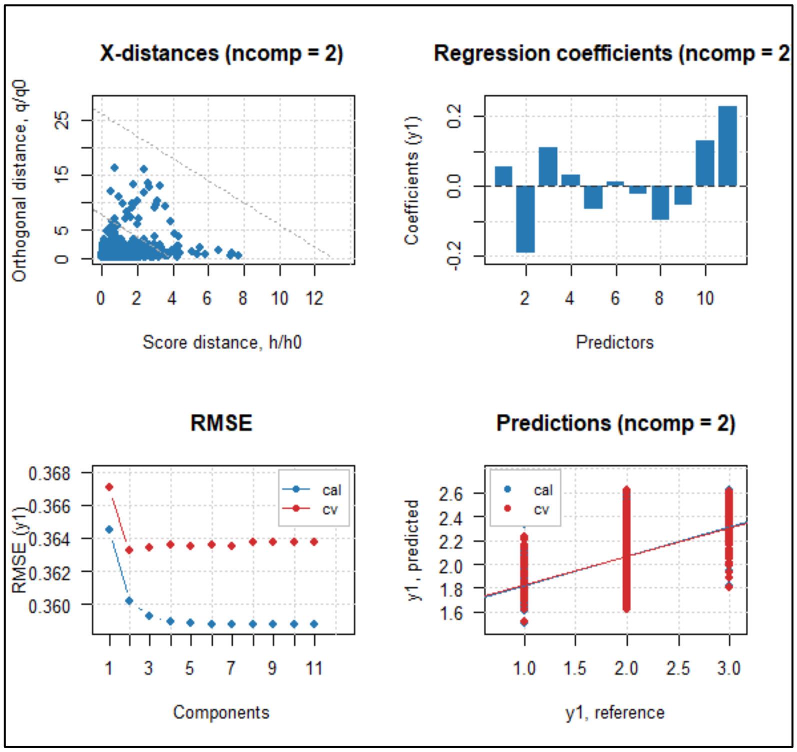


Figure 4.25 - Plots relates to PLS Analysis_1

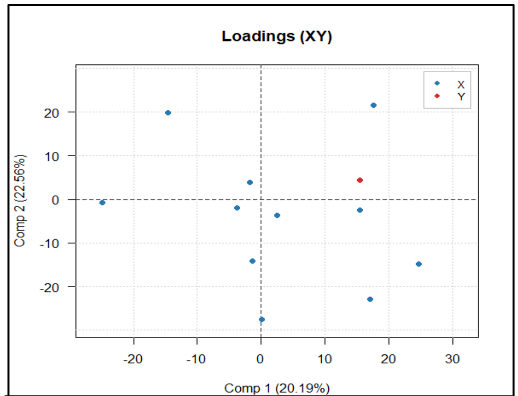
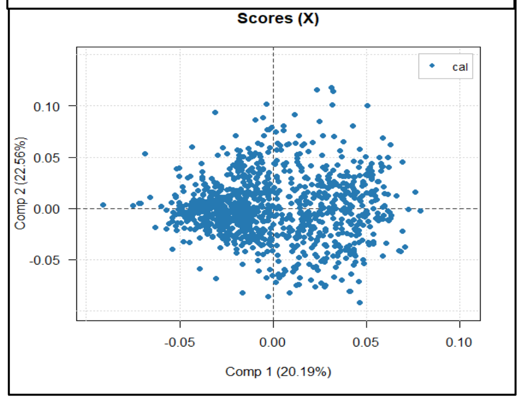


Figure 4.24- Plots relates to PLS Analysis_2



11. Tri-variate Analysis with highly correlated explanatory variables.

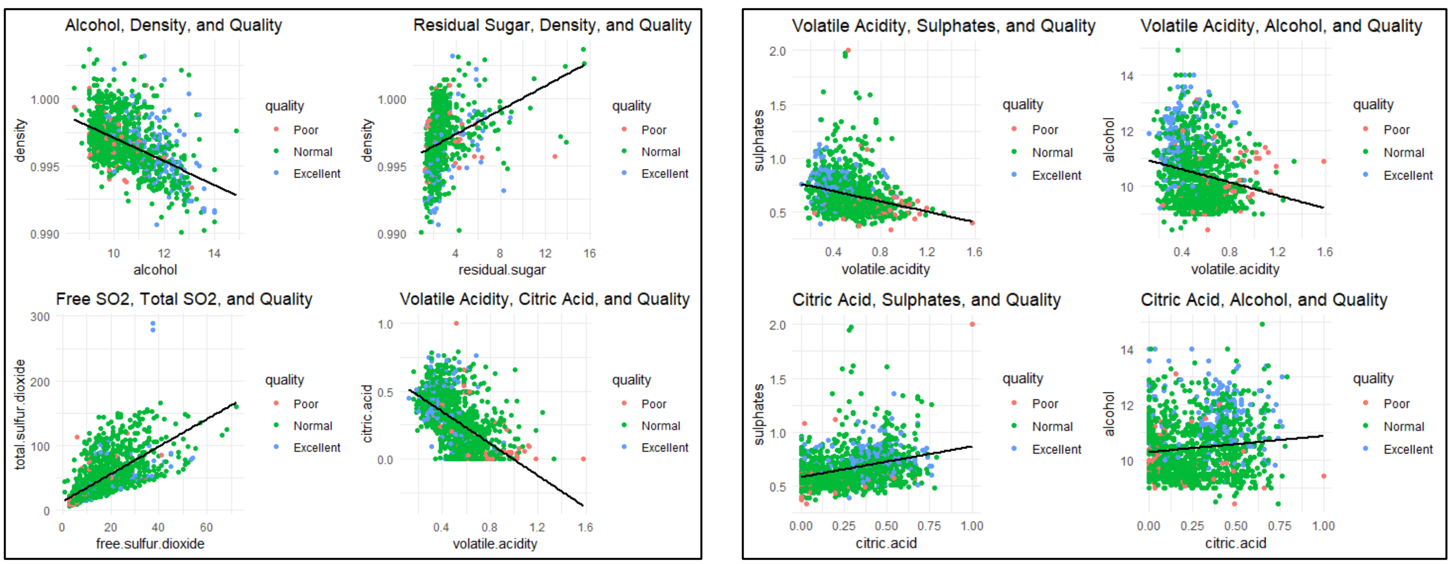


Figure 4. 26 - Tri-variate Analysis

Based on the results of the correlation test, alcohol, sulphates, volatile acidity, and citric acid have the strongest relationships with quality. Strong but less substantial correlations exist between a few other variables and quality. They are density, total sulfur dioxide, chlorides, and fixed acidity. Also, the density of water is close to that of water depending on the percentage alcohol and sugar content.

Using figure 4.26 we can see that these variables have strong positive and negative relationships.

12. SUGGESTIONS FOR A QUALITY ADVANCED ANALYSIS

Since we have ordinal categorical response (wine quality) with correlated predictors, the following techniques are suggested:

- Ordinal logistic regression – as a proportional odds model
- Multiple logistic regression – as a benchmark model
- Ridge regression
- Polynomial model
- Lasso and Elastic-Net regression
- K- nearest neighborhood regression
- Decision trees
- Random Forest Classifier

13. APPENDIX

```r
data =read.csv("Red_Wine_Data.csv")

# get the summary result of the data set
summary(data)
sapply(data,function(x) sum(is.na(x)))
sum(is.na(data))
# So there haven't any missing values_______________________

#checking foe duplicate data in the data set
dup = sum(duplicated(data)==TRUE)
dup
my_data1 = unique(data)
t=my_data1$quality
nrow(my_data1)
# So 240 observations are duplicate observations we remove that observations.
Then we have
# Then we have only 1359 observations.________________________________
#reprocessing the quality variable as a categorical, using factor.
my_data1$quality = as.factor(my_data1$quality)
a=table(my_data1$quality)
a
my_data= my_data1
levels(my_data$quality)=c("Poor","Poor","Normal","Normal","Excellent","Excellent")
b=table(my_data$quality)
b
# Check the data types of each column
sapply(my_data, class)
# install.packages("ggplot2")
library(ggplot2)

# Create a histogram of the "quality" variable
ggplot(my_data1, aes(x = quality)) +
  geom_bar(fill = "#b11226", color = "black", position = "dodge") +
  theme_minimal() +
  labs(title = "Histogram of Quality Variable",
       x = "Quality",
       y = "Count")
ggplot(my_data, aes(x = quality)) +
  geom_bar(fill = "#b11226", color = "black", position = "dodge") +
  theme_minimal() +
  labs(title = "Histogram of Quality Variable",
       x = "Quality",
       y = "Count")
```

```r
# install.packages(c("ggplot2", "dplyr"))
library(ggplot2)
library(dplyr)
# Select only numeric variables
numeric_data <- my_data %>%
  select_if(is.numeric)

# Calculate Spearman's correlation coefficients
cor_matrix <- cor(numeric_data, method = "spearman")

# Reshape the correlation matrix for plotting
cor_long <- as.data.frame(as.table(cor_matrix))
colnames(cor_long) <- c("Variable1", "Variable2", "Correlation")

# Create a heatmap using ggplot2 with correlation values annotated
ggplot(cor_long, aes(x = Variable1, y = Variable2, fill = Correlation)) +
  geom_tile(color = "white") +
  geom_text(aes(label = round(Correlation, 2)), vjust = 1) +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Spearman's Correlation Heatmap",
       x = "Variable",
       y = "Variable")
# Calculate point-biserial correlation between "quality" and each explonatory variables
# Install and load the knitr package
# install.packages("knitr")
library(knitr)

# Your existing code
r <- character(0)
cor <- numeric(0)

for (i in 1:(ncol(my_data)-1)) {
  r[i] <- names(my_data)[i]
  cor[i] <- cor(as.numeric(my_data$quality), my_data[[i]], method = "spearman")
}

# Create a data frame from vectors r and cor
result_table <- data.frame(Variable = r, Correlation = cor)

# Print the result as a formatted table
kable(result_table, format = "markdown")
```

```r
###############Histograms of the explanatory variable

ggplot(my_data, aes(x = fixed.acidity)) +
  geom_histogram(fill = "#b11226", color = "black", bins = 30) +
  theme_minimal() +
  labs(title = "Distribution of Fixed Acidity",
       x = "Fixed Acidity (g/dm^3)",
       y = "Count")
ggplot(my_data, aes(x = volatile.acidity)) +
  geom_histogram(fill = "#b11226", color = "black", bins = 30) +
  theme_minimal() +
  labs(title = "Distribution of Volatile Acidity",
       x = "Volatile Acidity (g/dm^3)",
       y = "Count")
ggplot(my_data, aes(x = citric.acid)) +
  geom_histogram(fill = "#b11226", color = "black", bins = 30) +
  theme_minimal() +
  labs(title = "Distribution of Citric Acid",
       x = "Citric Acid (g/dm^3)",
       y = "Count")
ggplot(my_data, aes(x = sulphates)) +
  geom_bar(fill = "#b11226", color = "black", position = "dodge") +
  theme_minimal() +
  labs(title = "Distribution of Sulphates",
       x = "sulphates  (g/dm^3)",
       y = "Count")
ggplot(my_data, aes(x = free.sulfur.dioxide )) +
  geom_histogram(fill = "#b11226", color = "black", position = "dodge") +
  theme_minimal() +
  labs(title = "Distribution of Free Sulfur Dioxide",
       x = "Free Sulfur Dioxide (g/dm^3)",
       y = "Count")
ggplot(my_data, aes(x = total.sulfur.dioxide)) +
  geom_histogram(fill = "#b11226", color = "black", bins = 30) +
  theme_minimal() +
  labs(title = "Distribution of Total Sulfur Dioxide",
       x = "Total Sulfur Dioxide (g/dm^3)",
       y = "Count")
ggplot(my_data, aes(x = pH)) +
  geom_histogram(fill = "#b11226", color = "black", bins = 30) +
  theme_minimal() +
  labs(title = "Distribution of pH",
       x = "pH ",
       y = "Count")
ggplot(my_data, aes(x = density )) +
  geom_histogram(fill = "#b11226", color = "black", bins = 30) +
  theme_minimal() +
  labs(title = "Distribution of Density ",
       x = "Density",
       y = "Count")
ggplot(my_data, aes(x = residual.sugar)) +
  geom_histogram(fill = "#b11226", color = "black", bins = 30) +
  theme_minimal() +
  labs(title = "Distribution of Residual Sugar ",
       x = "Residual Sugar (g/dm^3)",
       y = "Count")
ggplot(my_data, aes(x = chlorides)) +
  geom_histogram(fill = "#b11226", color = "black", bins = 30) +
  theme_minimal() +
  labs(title = "Distribution of Chlorides ",
       x = "Chlorides (g/dm^3)",
       y = "Count")
ggplot(my_data, aes(x = alcohol)) +
  geom_histogram(fill = "#b11226", color = "black", bins = 30) +
  theme_minimal() +
  labs(title = "Distribution of Alcohol ",
       x = "Alcohol (g/dm^3)",
       y = "Count")
```

```r
########Boxplots############

ggplot(my_data, aes(x = quality, y = fixed.acidity, fill = quality)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Poor" = "#FC9272", "Normal" = "#EF3B2C",
"Excellent" = "#A50F15")) +
  labs(title = "Box Plot of Fixed Acidity by Quality",
       x = "Quality",
       y = "Fixed Acidity")
ggplot(my_data, aes(x = quality, y = volatile.acidity, fill = quality)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Poor" = "#FC9272", "Normal" = "#EF3B2C",
"Excellent" = "#A50F15")) +
  labs(title = "Box Plot of Volatile Acidity by Quality Rank",
       x = "Quality",
       y = "Volatile Acidity")
ggplot(my_data, aes(x = quality, y = citric.acid, fill = quality)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Poor" = "#FC9272", "Normal" = "#EF3B2C",
"Excellent" = "#A50F15")) +
  labs(title = "Box Plot of Citric Acid by Quality Rank",
       x = "Quality",
       y = "Citric Acid")
ggplot(my_data, aes(x = quality, y = pH, fill = quality)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Poor" = "#FC9272", "Normal" = "#EF3B2C",
"Excellent" = "#A50F15")) +
  labs(title = "Box Plot of pH by Quality Rank",
       x = "Quality",
       y = "pH")
ggplot(my_data, aes(x = quality, y =  sulphates, fill = quality)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Poor" = "#FC9272", "Normal" = "#EF3B2C",
"Excellent" = "#A50F15")) +
  labs(title = "Box Plot of Sulphates by Quality Rank",
       x = "Quality",
       y = "Sulphates")
ggplot(my_data, aes(x = quality, y =free.sulfur.dioxide, fill = quality)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Poor" = "#FC9272", "Normal" = "#EF3B2C",
"Excellent" = "#A50F15")) +
  labs(title = "Box Plot of Free SO2  by Quality Rank",
       x = "Quality",
       y = "Free Sulfur Dioxide")
ggplot(my_data, aes(x = quality, y =total.sulfur.dioxide, fill = quality)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Poor" = "#FC9272", "Normal" = "#EF3B2C",
"Excellent" = "#A50F15")) +
  labs(title = "Box Plot of Total SO2  by Quality Rank",
       x = "Quality",
       y = "Total Sulfur Dioxide")
ggplot(my_data, aes(x = quality, y =chlorides, fill = quality)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Poor" = "#FC9272", "Normal" = "#EF3B2C",
"Excellent" = "#A50F15")) +
  labs(title = "Box Plot of Chlorides  by Quality Rank",
       x = "Quality",
       y = "Chlorides")
ggplot(my_data, aes(x = quality, y =density, fill = quality)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Poor" = "#FC9272", "Normal" = "#EF3B2C",
"Excellent" = "#A50F15")) +
  labs(title = "Box Plot of Density  by Quality Rank",
       x = "Quality",
       y = "Density")
ggplot(my_data, aes(x = quality, y =alcohol, fill = quality)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Poor" = "#FC9272", "Normal" = "#EF3B2C",
"Excellent" = "#A50F15")) +
  labs(title = "Box Plot of Alcohol  by Quality Rank",
       x = "Quality",
       y = "Alcohol")
ggplot(my_data, aes(x = quality, y =residual.sugar, fill = quality)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("Poor" = "#FC9272", "Normal" = "#EF3B2C",
"Excellent" = "#A50F15")) +
  labs(title = "Box Plot of Residual Sugar by Quality Rank",
       x = "Quality",
       y = "Residual Sugar")
```

```r
ggplot(my_data, aes(x = volatile.acidity, y = citric.acid)) +
  geom_point(color = "#EF3B2C", size = 3) +
  labs(title = "Residual Plot of Volatile Acidity vs Citric Acid",
       x = "Volatile Acidity",
       y = "Citric Acid")
ggplot(my_data, aes(x = fixed.acidity, y = pH )) +
  geom_point(color = "#EF3B2C", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Residual Plot of Fixed Acidity vs pH",
       x = "Fixed Acidity",
       y = "pH")
ggplot(my_data, aes(x = volatile.acidity, y = pH )) +
  geom_point(color = "#EF3B2C", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Residual Plot of Volatile Acidity vs pH",
       x = "Volatile Acidity",
       y = "pH")
ggplot(my_data, aes(x = citric.acid , y = pH )) +
  geom_point(color = "#EF3B2C", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Residual Plot of Citric Acid  vs pH",
       x = "Citric Acid",
       y = "pH")
ggplot(my_data, aes(x = residual.sugar , y = density )) +
  geom_point(color = "#EF3B2C", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Residual Plot of Residual Sugar  vs Density",
       x = "Residual Sugar",
       y = "Density")
ggplot(my_data, aes(x = alcohol , y = density )) +
  geom_point(color = "#EF3B2C", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Residual Plot of Alcohol  vs Density",
       x = "Alcohol",
       y = "Density")
ggplot(my_data, aes(x = alcohol , y = residual.sugar )) +
  geom_point(color = "#EF3B2C", size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Residual Plot of Alcohol  vs Residual Sugar",
       x = "Alcohol",
       y = "Residual Sugar")
#Trivariate plot
# Load required libraries

library(mdatools)

# Trivariate Plot Function
trivariate_plot <- function(data, x_var, y_var, z_var, title) {
  ggplot(data, aes_string(x = x_var, y = y_var, color = z_var)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, color = "black", size = 1) +
    labs(title = title, x = x_var, y = y_var) +
    theme_minimal()
}

# Generate and arrange trivariate plots
plot1 <- trivariate_plot(my_data, "alcohol", "density", "quality", "Alcohol,
Density, and Quality")
plot2 <- trivariate_plot(my_data, "residual.sugar", "density", "quality",
"Residual Sugar, Density, and Quality")
plot3 <- trivariate_plot(my_data, "free.sulfur.dioxide",
"total.sulfur.dioxide", "quality", "Free SO2, Total SO2, and Quality")
plot4 <- trivariate_plot(my_data, "volatile.acidity", "citric.acid",
"quality", "Volatile Acidity, Citric Acid, and Quality")

plot5 <- trivariate_plot(my_data, "volatile.acidity", "sulphates", "quality",
"Volatile Acidity, Sulphates, and Quality")
plot6 <- trivariate_plot(my_data, "volatile.acidity", "alcohol", "quality",
"Volatile Acidity, Alcohol, and Quality")
plot7 <- trivariate_plot(my_data, "citric.acid", "sulphates", "quality",
"Citric Acid, Sulphates, and Quality")
plot8 <- trivariate_plot(my_data, "citric.acid", "alcohol", "quality",
"Citric Acid, Alcohol, and Quality")
# Plot arrangement
library(gridExtra)

grid.arrange(plot1, plot2, plot3, plot4, nrow = 2)

grid.arrange(plot5, plot6, plot7, plot8, nrow = 2)
```

```r
########################
#Splitting data set
set.seed(100)
library(caTools)
split <- sample.split(my_data, SplitRatio = 0.8)#80% for training and 20% for testing
split
train <- subset(my_data, split == "TRUE")
test <- subset(my_data, split == "FALSE")
########### observing any patterns in the dataset #####################

#PC model
numeric_data = train[, sapply(train, is.numeric)]
pca_result = prcomp(numeric_data, scale. = TRUE)

# Display summary of PCA results
summary(pca_result)
pc_scores = pca_result$x[,1:2]
plot(pc_scores[,1],pc_scores[,2],xlab = "Principal Component 1 (PC1)",ylab =
"Principal Component 2 (PC2)",mar = c(4, 4, 2, 2))
 #pls model


 k=as.numeric(t)
 #k
 Quality_Ordinal = ifelse(k<=4,1 ,ifelse(k>4 & k<=6,2,3))
 dataset_new = cbind(my_data1,Quality_Ordinal)
 #view(dataset_new)
 dim(dataset_new)
split <- sample.split(dataset_new, SplitRatio = 0.8)#80% for training and 20%
for testing
split
train1 <- subset(dataset_new, split == "TRUE")
test1 <- subset(dataset_new, split == "FALSE")

xc= train1[,1:11]
yc =train1[,13]
#xc
#yc
dim(xc)
 xt= test1[,1:11]
 yt =test1[,13]
 dim(xt)
yc <- as.numeric(as.character(yc))

# PLS model
library(mdatools)
model <- pls(xc, yc, scale = TRUE, cv = 1, info = "Wine Quality Prediction")

# Print the summary of the PLS model
summary(model)

plotXScores(model,show=1,labels=F)
plotXYLoadings(model,show=1,labels =F)
```

R Codes

https://drive.google.com/file/d/128vOmPHlDcD6jrV-jbxIxXqWLP6AZT1f/view?usp=sharing

14. REFERENCES

https://www.researchgate.net/publication/350110244_Prediction_of_Wine_Quality_Using_Machine_Learning_Algorithms

https://winefolly.com/deep-dive/vinho-verde-the-perfect-poolside-wine-from-portugal/

https://rstudio-pubs-static.s3.amazonaws.com/57835_c4ace81da9dc45438ad0c286bcbb4224.html

https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/common-chemical-reagents/citric-acid#:~:text=Increasing%20citric%20acid%20concentrations%20will,if%20present%20at%20excessive%20levels.

https://winemakermag.com/technique/ph-acid-relationship-in-winemaking#:~:text=The%20common%20pH%20range%20in,the%20pH%20of%20a%20wine.

https://winemakermag.com/technique/501-measuring-residual-sugar-techniques#:~:text=Residual%20Sugar%20Concentration,-Residual%20sugar%20concentration&text=Dry%20wines%20are%20typically%20in,the%205.0%E2%80%9315%20percent%20range.

https://whicherridge.com.au/blog/what-is-residual-sugar-in-wine/

https://www.scielo.br/j/cta/a/HQsrPrPMNZYgRzSKtrjHyHh/?format=pdf#:~:text=1%2D%20Chloride%20concentration%20in%20the,the%20highest%20levels%20of%20chlorides.

https://medium.com/@spynyahya/analyzing-red-wine-quality-69aadb08a303

https://rpubs.com/aratakagan0412/wine_quality_analysis

https://www.oiv.int/public/medias/7840/oiv-collective-expertise-document-so2-and-wine-a-review.pdf