Michael David, Hemen Asfaw

Multivariable Statistics

Multiple Linear Regression Project

The Quality of Red Wine

Table of Contents:

Introduction

Wine is one of the oldest beverages; spanning civilizations and maintaining its prevalence in modern society. The alcoholic beverage that is most commonly known for coming from fermented grapes (Puckette James, 1), is composed of several compounds, which at low concentrations play a significant role in influencing its quality and taste (Markoski, Garavaglia, Olivera, Olivaes, Marcadenti, 1). Historically, wine has opened doors for trades amongst the European Nations allowing them to foster relationships and share their culture (Wynne, 1). Today that still holds, except the export of wine has not been limited to just the neighboring European nations. 50% of the world's largest wine exporters are European nations, among them, Portugal (Statista). In Portugal, the wine industry accounts for 10% of manufacturing enterprises and 25% of total turnover (Fragoso, Vieira, 1). Part of ensuring the continued success of the wine industry and bettering it involves wine certifications and quality assessments (Cortez, Cerdeira, Almeida, Matos, Reis, 2).

Wine certifications stop any form of alteration that could be harmful to consumers and helps maintain the marketplace quality (Cortez, Cerdeira, Almeida, Matos, Reis, 2). The quality evaluations that are conducted through physicochemical and sensory tests are essential to understanding what factors influence the quality of wine most. While physicochemical tests are conducted in laboratory settings, sensory assessments rely on human experts to classify wine (Cortez, Cerdeira, Almeida, Matos, Reis, 2). As of 2024, there are around 13,457 people working in the wine industry in Portugal with around 2,030 companies in the industry (IBIS World). Determining which factors most heavily affect the quality of the wine allows companies to focus their attention and resources on improving these areas with enhanced grape production, better manufacturing, and more research. The benefits of understanding these key factors doesn’t only allow the companies to focus where they most need it, but also keeps the families that depend on revenues from the wine industry afloat. Therefore, for the purpose of this project, we will be analyzing the data collected on the red vinho verde wine from Portugal to assess which physicochemical and sensory properties most significantly influence its quality and discuss how these findings can inform production enhancement and customer awareness. We seek to establish and verify a multiple linear

regression model for wine quality based on the UCI ML repository wine quality dataset. Using linear regression techniques for finding, comparing, evaluating, and analyzing linear regression models, we will use statistical metrics and techniques to best define the relationship between the explanatory variables and our response variable, wine quality.

Methodology

Dataset Overview

Our dataset, collected from May 2004 to February 2007, comes from the UCI ML repository wine quality dataset and consists of 1599 observations spanning 12 variables, of which 11 are possible explanatory variables and one is our target variable. All of our 12 variables are continuous with varying ranges. The variables with their respective definitions are as follows:

1. **Fixed acidity**: Most of the acids involved with wine are either fixed or non-volatile (do not evaporate easily).
2. **Volatile acidity**: The amount of acetic acid in the wine, which at too high levels can cause an unpleasant vinegar taste.
3. **Citric acid**: Found in small amounts, citric acid can add 'freshness' and flavor to wines.
4. **Residual sugar**: The amount of sugar left after fermentation stops. It is rare to find wines with less than 1 gram/liter and wines with more than 45 grams/liter are considered sweet.
5. **Chlorides**: The amount of salt in the wine.
6. **Free sulfur dioxide**: The free form of SO2 exists in equilibrium between molecular SO2 (as dissolved gas) and bisulfite ion; prevents microbial growth and oxidation of wine.
7. **Total sulfur dioxide**: Number of free and bound forms of S02; at low concentrations, SO2 is largely undetectable in wine, but at free SO2 concentrations above 50 ppm, SO2 becomes apparent in the nose and flavor of the wine.
8. **Density**: The density of water is close to that of water depending on the percentage of alcohol and sugar contained.

9. **pH**: Describes how acidic or basic a wine is on a scale of 0 (very acidic) to 14 (very basic); most wines are between 3 and 4 on the pH scale
10. **Sulphates**: A wine additive that can contribute to sulfur dioxide (S02) levels, which acts as an antimicrobial and antioxidant
11. **Alcohol**: The percent alcohol content of the wine
12. **Quality**: target variable (based on sensory data, score between 0 and 10). Indicates how good the wine is at this quality standard.

While our dataset describes physicochemical properties of wine, it does not, however, include all metrics for determining wine quality; factors such as age, grape variety, soil information are excluded from the dataset. This dataset focuses on the chemical composition of the wines surveyed.

Exploratory Data Analysis

One of the initial stages of our project involved examining the distribution of variables. We made use of the R function summary() to display the descriptive statistics of all of our explanatory variables. The descriptive statistics of our explanatory variables are listed in Table 1. It is important to note that none of our variables contain missing values. Following this, we made a correlation matrix (reference figure 1) to see if there was a significant correlation between any of the variables. The heat map shows the darker shades of red and blue (that represent 1 and -1, respectively) show the strongest correlations amongst variables. The diagonal of the plot is dark red as it signifies the correlation that each variable has with itself. From this plot we were able to infer the following:

1. Total sulfur dioxide and free sulfur dioxide have a strong positive correlation (0.67). This is logical as total sulfur dioxide also includes free sulfur dioxide.
2. There is a moderate positive correlation between alcohol and quality (0.48), suggesting that wines with higher alcohol content may be rated as higher quality.
3. Volatility Acidity has a mild negative correlation with quality (-0.39), meaning as volatile acidity increases, the quality decreases.

Variable Selection

In our variable selection phase we made use of the regsubsets() function which returned a subset of the best models. These models had a varying number of predictors each with their respective R^2, Mallow's, CP, and BIC values. We started by narrowing down the top 5 models that had the best R^2 Mallow's, CP, and BIC values. Table 2 has been color coded to show when the same model appears for the R^2, Mallow's CP, and BIC tests. Model 13 places first in the R^2 and Mallow's CP category and appears second in the BIC column. Therefore, we decided to select model 13 for our analysis.

Model 13: *Volatile Acidity + Residual Sugar + Chlorides +Free Sulfur Dioxide + Total Sulfur Dioxide + pH + Sulphates + Alcohol*

$$\hat{y} = 4.430 - 1.013x(1) - 2.018x(2) + 0.005x(3) - 0.003x(4) - 0.4827x(5) + 0.8827x(6) + 0.2893x(7)$$

Multiple R-squared: 0.3595, Adjusted R-squared: 0.3567

Upon examining the correlation plot (see figure 1), we observed a substantial correlation between total sulfur dioxide and free sulfur dioxide. This correlation makes sense, given that free sulfur dioxide is a component of the total sulfur dioxide measurement. Initially, we contemplated excluding free sulfur dioxide from our model, but doing so resulted in a lower adjusted R^2 value compared to our original model. As a result, we opted to maintain the seven-variable model that we had initially developed.

Model Selection and Adequacy Checks

After determining the variables for our model, we moved forward with several model adequacy checks to verify that our model would not be misleading and result in incorrect interpretations. A key assumption we tested was the normal distribution of the residuals, for which we utilized a QQ-plot. The QQ-plot shown in Figure 2 indicates that while the residuals mostly aligned with the y=x line, suggesting normal distribution, there was a discernible deviation to the left side of the plot. This deviation points to a leftward skew in the distribution of the residuals. This means there are more data points on the lower end of the residual scale, which can drag the mean of the residuals to the left, breaching the normality assumption. In order to deal with this, we decided to carry out a robust regression as robust regressions

are particularly useful in situations where the dataset contains outliers and does not follow a normal distribution. Our model after implementing the robust regression is as follows:

$$\hat{y} = 4.021 - 0.913x(1) - 1.726x(2) + 0.005x(3) - 0.004x(4) - 0.418x(5) + 0.908x(6) + 0.301x(7)$$

Multiple R-squared: 0.3837, Adjusted R-squared: 0.381

The QQ-plot shown in figure 3 still shows a left skew in the data. The fact that the robust regression could not solve our issues with the breach of normality lead us to believe that the issues are with normality that are inherent to the dataset and not due to the influence of outliers. However, the robust regression model provided a higher multiple R-squared value compared to the standard regression model, meaning even though there is an issue with normality, the robust model provides a better fit to the data. In order to attempt to fix the problems of normality, we decided to carry out data transformations including the log and square root transformation. Despite our attempts, the skewness of the data was persistent. Therefore, since the model developed exhibits a breach of the normality, as evidenced by the plots, this violation is a notable limitation and should be considered when conducting interpretations.

We then continued our analysis and constructed a fitted values versus residuals plot to assess the assumptions of linearity and homoscedasticity in our regression model. The plot in figure 4 revealed several interesting characteristics: The residuals were grouped into distinct lines, which we believe are due to the discrete nature of the target variable. Although the target variable theoretically ranges from 0 to 10, our exploratory data analysis indicated that the actual observed range is between 3 and 8. This narrower, discrete range explains the six distinct groupings observed in the residual plot. The spread of residuals appeared random and consistent across the range of fitted values, suggesting that the assumption of homoscedasticity (constant variance) was initially met.

We followed up by constructing residual plots for the individual explanatory variables as shown in figures 5 - 11, to test similar assumptions. Upon further examination, specific issues were identified with the variables free sulfur dioxide, total sulfur dioxide, and chlorides (see figures 7, 8, 6, respectively). Two main concerns that arose were: some data points were significantly distant from others but were not

classified as outliers in our outlier tests and the funnel-shaped patterns in residual plots for these variables suggested potential heteroscedasticity, prompting further investigation.

To address potential heteroscedasticity, we applied square root transformations to the variables free sulfur dioxide, total sulfur dioxide, and chlorides. Post-transformation, we reassessed the model using the adjusted R-squared values and the residual plot observations. The new model is as follows:

$$\hat{y} = 4.364 - 0.913x(1) - 1.287x(2) + 0.06x(3) - 0.061x(4) - 0.417x(5) + 0.895x(6) + 0.297x(7)$$

Multiple R-squared: 0.3842

The Multiple R-squared value post-transformation shows very mild improvement of 0.13% from the initial model. Figures 12-14 show the residual plots following the square root transformation which show a slight reduction in the funnel-shaped variance pattern, particularly for free sulfur dioxide, which suggested some improvement of the heteroscedasticity issue. However, the changes in the plots for total sulfur dioxide and chlorides were minimal. Despite the transformations, the adjustments to the R-squared value were not consequential and the difference in the models wasn't as substantial. The need for additional interpretation work due to the transformations, combined with the minor impact on the overall model's explanatory power, led us to decide against adopting the transformed model for further analyses. Given these outcomes, we will continue with our original model without transformations.

Multicollinearity

In our regression model, we assessed multicollinearity using the Variance Inflation Factor (VIF) function in R. Generally, VIF values ranging from 5 to 15 suggest multicollinearity depending on the context. However, our analysis found no evidence of multicollinearity among the explanatory variables, confirming their suitability for inclusion in our model. This absence of multicollinearity reassures us of the reliability of the relationships identified between the variables and wine quality.

Leverage and Influence Points

To assess the influence of individual data points, we employed Cook's Distance, a measure used to estimate the impact of removing a specific data point on the fitted regression coefficients. Our findings indicated no concerning values exceeding the commonly used threshold of 1, suggesting that no single data point had a disproportionate influence on the overall model. However, our analysis using studentized residuals identified three potential leverage points. These points are characterized by having an outsized influence on the parameter estimates, indicating potential anomalies in the dataset that could skew our results.

Based on these insights, it is important to analyze the identified leverage points in more detail. However, we are limited by the fact that we don't have enough background information on the dataset. We don't know where these data points came from or why they exist, which makes it difficult for us to fully understand their meaning and importance.Without this context, it would be unwise and potentially dangerous to simply remove these data points without any specific reason. We might end up losing valuable information that could be crucial to our analysis. Furthermore, our examination using Cook's Distance has confirmed that these points do not have a significant impact on the model. This finding provides additional support for our decision to keep these data points in our analysis.

Conclusion

Our analysis identified several physicochemical and sensory properties that heavily influence wine quality, including volatile acidity, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol. These properties each impact wine quality differently, some have a positive impact while others have negative influences. Understanding these key drivers enables winemakers to refine fermentation, blending, and aging processes to enhance wine quality and better allocate resources towards improving significant attributes.

Our regression model's limitations stem from several key issues. Firstly, the problem of non-normality in residuals was not adequately addressed, necessitating caution in the interpretation of our results. Additionally, the relatively low R^2 suggests limited predictive power, as it indicates that only

38.37% of the variance in wine quality is explained by our model, leaving 61.63% unaccounted for. Moreover, certain data points appeared as outliers in the residual plots, yet they were not confirmed as such by subsequent tests. This inconsistency, coupled with the limited information provided about the dataset, hindered our ability to fully understand these anomalies. Furthermore, the dataset lacks comprehensive metrics such as age, grape variety, and terroir, which are crucial for a more in-depth evaluation of wine quality. These limitations highlight the need for cautious interpretation of the model's findings and underscore the importance of further research, potentially utilizing more detailed data and advanced statistical techniques.

Appendix

(i) Figures and Tables

Table 1

	Min	1st-Q	Median	Mean	3rd-Q	Max	SD
Volatile Acidity	0.12	0.39	0.52	0.5278	0.64	1.58	0.1791
Fixed Acidity	4.6	7.1	7.9	8.32	9.2	15.9	1.741096
Citric acid	0	0.09	0.26	0.271	0.42	1	0.194801
Residual Sugar	0.9	1.9	2.2	2.539	2.6	15.5	1.409928
Chlorides	0.012	0.07	0.079	0.08747	0.90	0.611	0.0471
Free Sulfur Dioxide	1	7	14	15.87	21	72	10.4601
Total Sulfur Dioxide	6	22	38	46.47	62	289	32.8953
Density	0.9901	0.9956	0.9968	0.9967	0.9978	1.0037	0.001887
pH	2.74	3.21	3.31	3.311	3.4	4.01	0.1544
Sulphates	0.33	0.55	0.62	0.6581	0.73	2	0.1695
Alcohol	8.4	9.5	10.20	10.42	11.10	14.9	1.0657
Quality	3	5	6	5.636	6	8	0.807569

Table 2:

Model	R^2 values	Model	Mallow's CP	Model	BIC
13	0.357	13	6.68	11	-654.93
15	0.357	15	7.55	13	-653.27
17	0.357	16	8.29	14	-649.47
16	0.356	17	8.94	9	-648.24
19	0.356	19	9.23	16	-647.04

Figure 1:

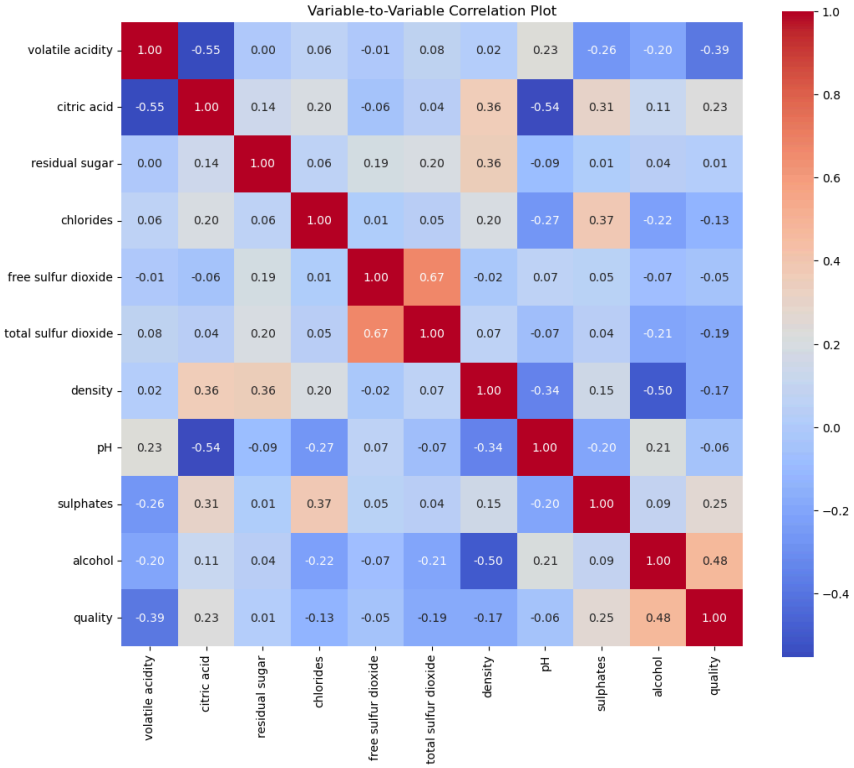


Figure 2:

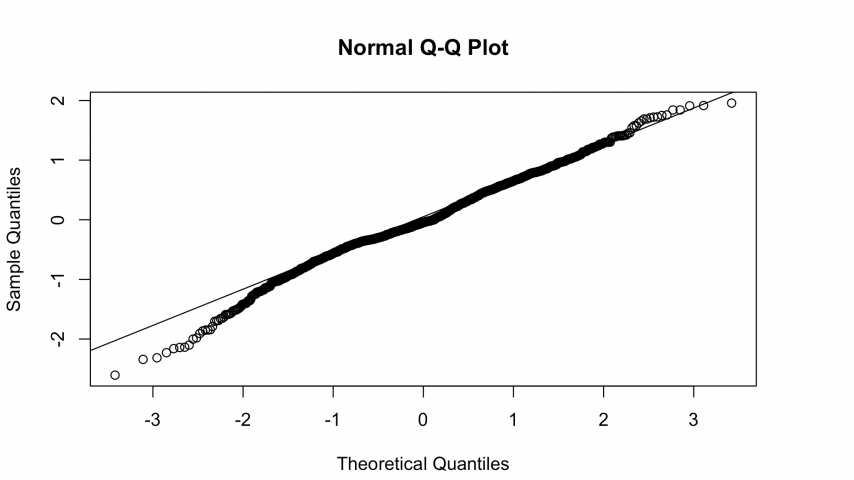


Figure 3:

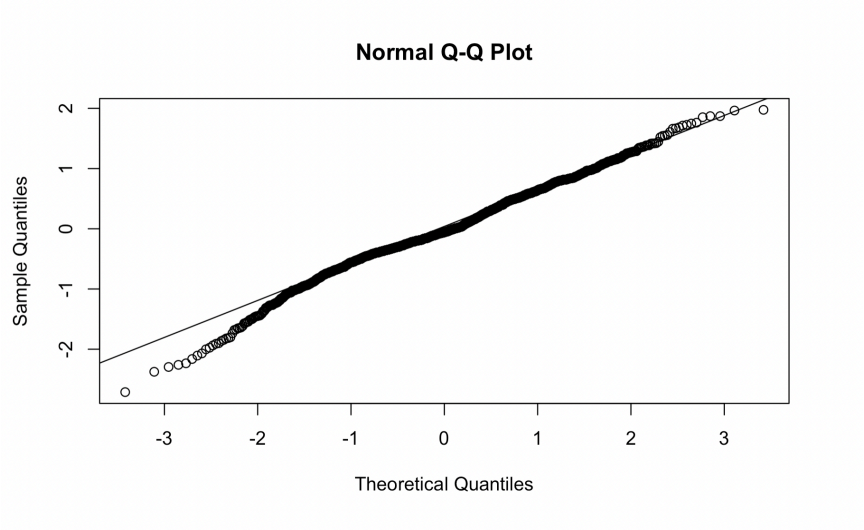


Figure 4:

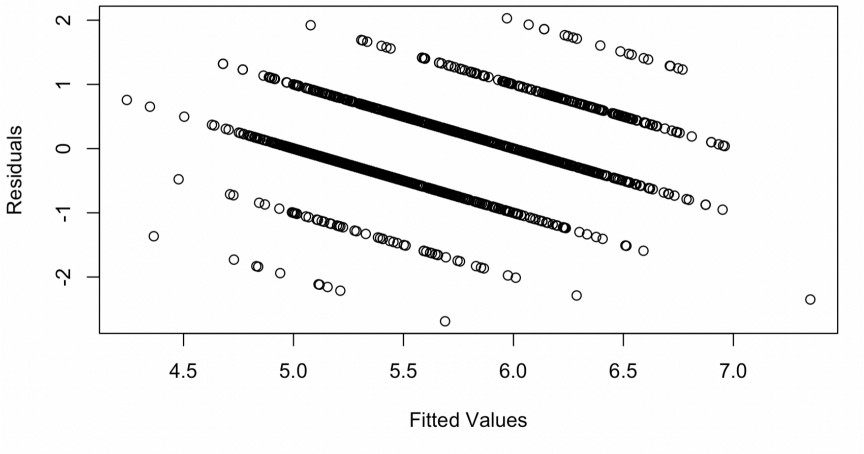


Figure 5:

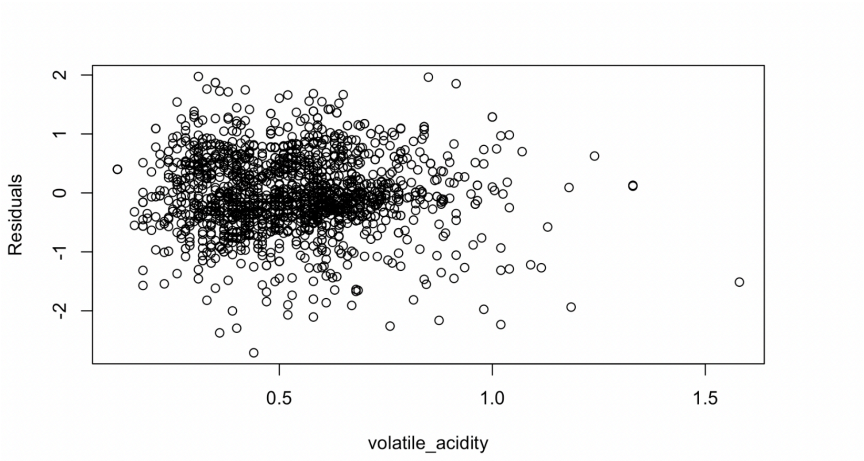


Figure 6:

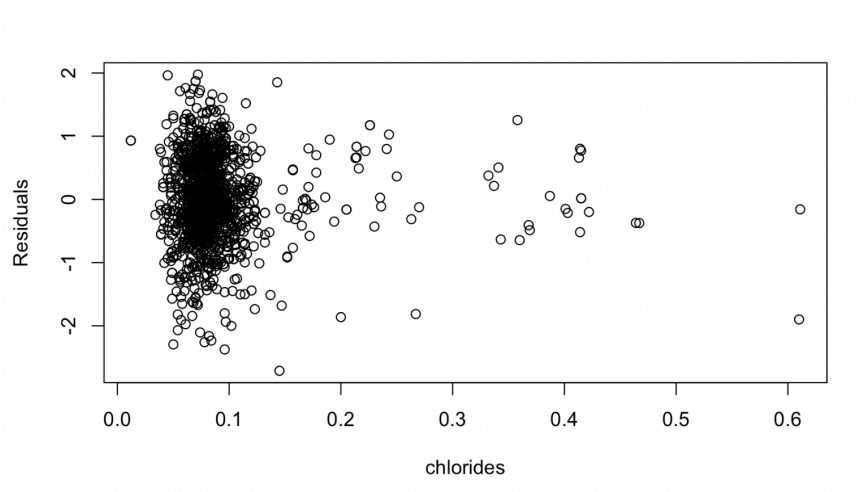


Figure 7:

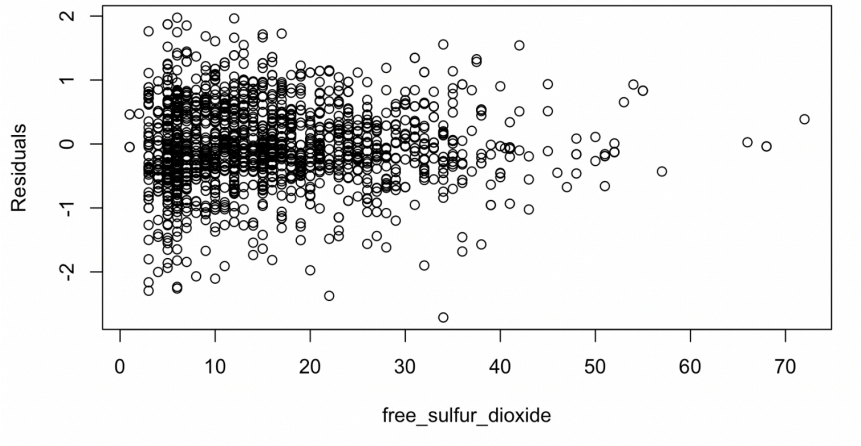


Figure 8:

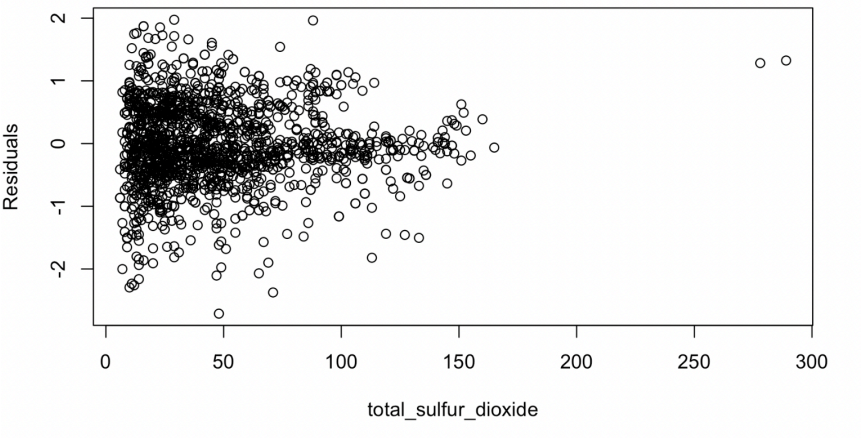


Figure 9:

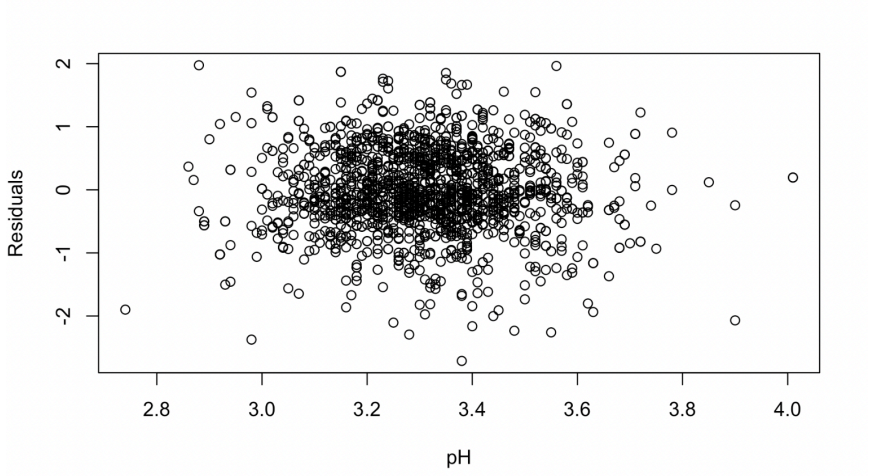


Figure 10:

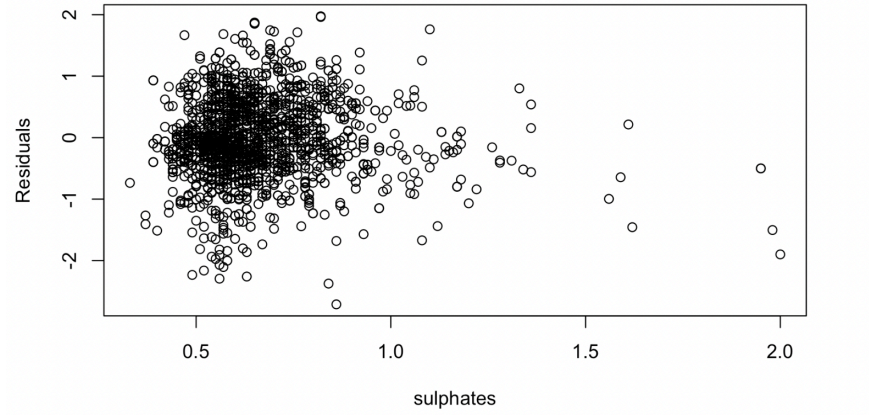


Figure 11:

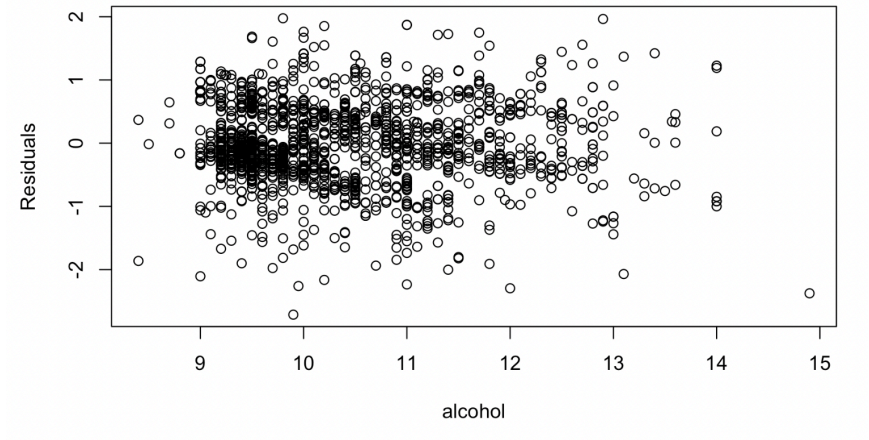


Figure 12

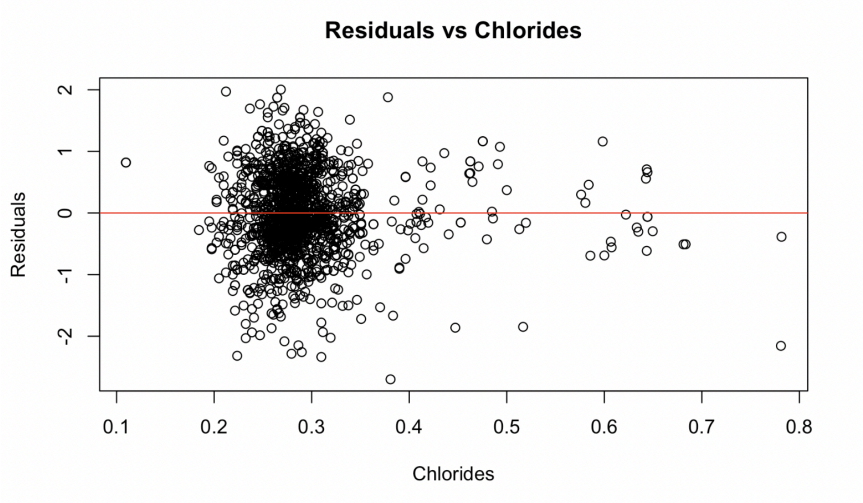
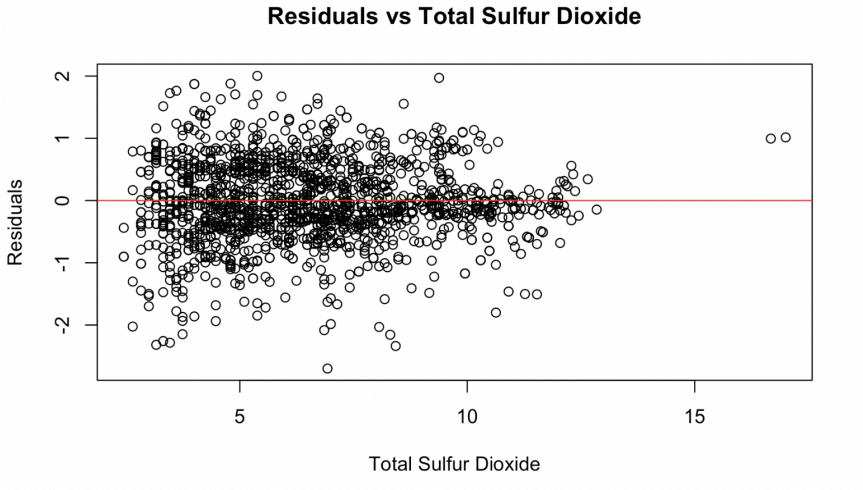


Figure 13:



(ii) Code

```
#Import dependencies ------------------------------------------------------------------
library(readr)
#install.packages("leaps")
library(leaps)
library(robustbase)

#Importing data ------------------------------------------------------------------
df <- read_delim("wine+quality/winequality-red.csv", delim = ";")
#df <- read_csv("MultivariateStatistics/Regression_Project/wine+quality/winequality-red.csv")
View(df)

#EDA ------------------------------------------------------------------------------
summary(df)
hist(df$alcohol, col="blue", xlab="Alcohol Level",
    main="Alcohol Distribution")
hist(df$pH, col="blue", xlab="pH",
    main="pH Distribution")

 #Variable Selection ------------------------------------------------------------------
###All Possible Regressions (best subset)
#run all the possible regression models
#nbest = 2: record the best two models of each model size (1 predictor, 2 predictors, etc.)
res=regsubsets(quality~`fixed acidity` + `volatile acidity` + `citric acid`+ `residual sugar`+
`chlorides`
            + `free sulfur dioxide` + `total sulfur dioxide` + `density`+ `pH` + `sulphates` + `alcohol`
            ,data=df,nbest = 2, nvmax= 11)
best.res=summary(res)
##create a nice matrix to present the results
#round(best.res$rsq,3): round R2 and keep 3 decimal digits
#round(best.res$adjr2,3): round the adjusted R2 and keep 3 decimal digits
results=cbind(best.res$outmat,round(best.res$rsq,3),round(best.res$adjr2,3),
           round(best.res$cp,2),round(best.res$bic,2))
colnames(results)=c("Fixed Acidity","Volatile Acidity","Citric Acid","Residual Sugar","Chlorides"
                ,"Free Sulphur Dioxide", "Total Sulphur Dioxide", "Density", "pH", "Sulphates", "Alcohol"
                ,"R2","R2_adj","Mallow's CP","BIC")
results

#Model Selection  ------------------------------------------------------------------
mod13.rob <- lmrob(quality ~ `volatile acidity` + chlorides + `free sulfur dioxide` + `total sulfur
dioxide`
             + pH + sulphates + alcohol, data = df)
summary(mod13.rob)

#Model Analysis Selection ------------------------------------------------------------
#We will be doing model 13 which is: Volatile acidity, Chlorides, Free super dioxide, Total sulfur
dioxide, pH,
#Sulphates, Alcohol
```

```
mod13 <- lm(quality ~ `volatile acidity` + chlorides + `free sulfur dioxide` + `total sulfur dioxide`
          + pH + sulphates + alcohol, data = df)
summary(mod13)
confint(mod13,level=0.95)
plot(mod13)
#Individual Scatterplots to see Relationship
plot(df$'volatile acidity',df$quality,ylab="quality", xlab="volatile acidity")
plot(df$chlorides,df$quality,ylab="quality", xlab="chlorides")
plot(df$`free sulfur dioxide`,df$quality,ylab="quality", xlab="free sulfur dioxide")
plot(df$`total sulfur dioxide`,df$quality,ylab="quality", xlab="total sulfur dioxide")
plot(df$pH,df$quality,ylab="quality", xlab="pH")
plot(df$sulphates,df$quality,ylab="quality", xlab="sulphates")
plot(df$alcohol,df$quality,ylab="quality", xlab="alcohol")


df_variables <- df[, c("volatile acidity", "chlorides", "free sulfur dioxide", "total sulfur dioxide",
                       "pH", "sulphates", "alcohol")]
cor(df_variables)
vif(mod13)

#Outlier Analysis  ------------------------------------------------------------------
#Indices: 46, 391, 460, 653, 814, 833, 900, 1277, 1470, 1479, 1506
#Index 46  ------
df[46, ]
index_46 <- c(1, 0.52, 0.054, 8, 65, 3.9, 0.56, 13.1)
index_46_matrix_transposed <- t(matrix(index_46, nrow = 1)) #make index_46 a row vector
index_46_matrix <- matrix(index_46, ncol = 1)

Cmat <- summary(mod13)$cov.unscaled #covariance matrix
result <- t(index_46_matrix_transposed) %*% Cmat %*% index_46_matrix #h value for the
specific observation
result #0.01677
#Find diagonal elements of hat matrix
round(sort(lm.influence(mod13)$hat),4) #Look at bottom right: that is H_max
#Index 391 ------
df[391, ]
index_391 <- c(1, 0.85, 0.045, 12, 88, 3.56, 0.82, 12.9)
index_391_matrix_transposed <- t(matrix(index_391, nrow = 1)) #make index_46 a row vector
index_391_matrix <- matrix(index_391, ncol = 1)
Cmat <- summary(mod13)$cov.unscaled #covariance matrix
result <- t(index_391_matrix_transposed) %*% Cmat %*% index_391_matrix #h value for the
specific observation
result #0.01327
#Find diagonal elements of hat matrix
round(sort(lm.influence(mod13)$hat),4) #Look at bottom right: that is H_max
#H_max = 0.0911

#Index 460 ------
df[460, ]
```

```
index_460 <- c(1, 0.58, 0.074, 10, 47, 3.25, 0.57, 9)
index_460_matrix_transposed <- t(matrix(index_460, nrow = 1)) #make index_46 a row vector
index_460_matrix <- matrix(index_460, ncol = 1)
Cmat <- summary(mod13)$cov.unscaled #covariance matrix
result <- t(index_460_matrix_transposed) %*% Cmat %*% index_460_matrix #h value for the
specific observation
result #0.002312923
#Find diagonal elements of hat matrix
round(sort(lm.influence(mod13)$hat),4) #Look at bottom right: that is H_max
#H_max = 0.0911

#Index 653 ------
df[653, ]
index_653 <- c(1, 0.36, 0.096, 22, 71, 2.98, 0.84, 14.9)
index_653_matrix_transposed <- t(matrix(index_653, nrow = 1)) #make index_46 a row vector
index_653_matrix <- matrix(index_653, ncol = 1)
Cmat <- summary(mod13)$cov.unscaled #covariance matrix
result <- t(index_653_matrix_transposed) %*% Cmat %*% index_653_matrix #h value for the
specific observation
result #0.01961045
#Find diagonal elements of hat matrix
# Set max.print to a large value (e.g., 1000)
options(max.print = 10000)
# Calculate the leverage values, round to 4 decimal places, and sort
leverage <- round(sort(lm.influence(mod13)$hat), 4)
# Show the leverage values
leverage
round(sort(lm.influence(mod13)$hat),4) #Look at bottom right: that is H_max
#H_max = 0.0911

#Predictive Analytics  ----------------------------------------------------------------
# Create a new dataset with random values for the explanatory variables
new <- data.frame(`volatile acidity` = c(0.5, 0.6),  # Random values for volatile acidity
                  chlorides = c(0.08, 0.09),        # Random values for chlorides
                  `free sulfur dioxide` = c(20, 25),  # Random values for free sulfur dioxide
                  `total sulfur dioxide` = c(50, 55),# Random values for total sulfur dioxide
                  pH = c(3.5, 3.6),                 # Random values for pH
                  sulphates = c(0.6, 0.7),          # Random values for sulphates
                  alcohol = c(12, 13))              # Random values for alcohol

# Make predictions using the model
predictions <- predict(mod13, newdata = df, interval = "confidence", level = 0.95)

# Show the predictions
predictions

# Create a new data frame with random values for the explanatory variables
new_data <- data.frame(`volatile acidity` = runif(2, min = 0, max = 2),  # Random values for
volatile acidity
```

```
                chlorides = runif(2, min = 0, max = 0.2),        # Random values for chlorides
                `free sulfur dioxide` = runif(2, min = 0, max = 50),  # Random values for free sulfur
dioxide
                `total sulfur dioxide` = runif(2, min = 0, max = 200), # Random values for total sulfur
dioxide
                pH = runif(2, min = 2.5, max = 4),               # Random values for pH
                sulphates = runif(2, min = 0, max = 2),          # Random values for sulphates
                alcohol = runif(2, min = 8, max = 15))           # Random values for alcohol

# Use the model to predict quality
predictions <- predict(mod13, newdata = new_data, interval = "confidence", level = 0.95)

# Show the predictions
predictions
```

Alternate QQ-plots code:

```
red_wine_q <- read_excel("RW.xlsx")

library(robustbase)
#Here is the code for model 13 robust:
mod13.rob <- lmrob(quality ~ volatile_acidity + chlorides + free_sulfur_dioxide +
total_sulfur_dioxide + pH + sulphates + alcohol, data = red_wine_q)
summary(mod13.rob)

raw_mod13_rob = resid(mod13.rob)

qqnorm(raw_mod13_rob)
qqline(raw_mod13_rob)

# Square root transformation
sqrt_chlorides = red_wine_q$chlorides <- sqrt(red_wine_q$chlorides)
sqrt_free_sulfur_dioxide = red_wine_q$free_sulfur_dioxide <-
sqrt(red_wine_q$free_sulfur_dioxide)
sqrt_total_sulfur_dioxide = red_wine_q$total_sulfur_dioxide <-
sqrt(red_wine_q$total_sulfur_dioxide)

model_transformed <- lmrob(quality~ volatile_acidity + sqrt_chlorides + sqrt_free_sulfur_dioxide
+ sqrt_total_sulfur_dioxide + pH + sulphates + alcohol,data=red_wine_q)

# Save the model
mod13.rob_sqrt <- lmrob(quality ~ volatile_acidity + chlorides + free_sulfur_dioxide +
total_sulfur_dioxide + pH + sulphates + alcohol, data = red_wine_q)
summary(mod13.rob_sqrt)

# Calculate residuals from the model
residuals <- residuals(mod13.rob_sqrt)


# Plot residuals against chlorides
plot(red_wine_q$chlorides, residuals, xlab = "Chlorides", ylab = "Residuals", main = "Residuals vs
```

```
Chlorides")
abline(h = 0, col = "red")

# Plot residuals against total sulfer dioxide
plot(red_wine_q$total_sulfur_dioxide, residuals, xlab = "Total Sulfur Dioxide", ylab = "Residuals",
main = "Residuals vs Total Sulfur Dioxide")
abline(h = 0, col = "red")

# Plot residuals against free sulfer dioxide
plot(red_wine_q$free_sulfur_dioxide, residuals, xlab = "Free Sulfur Dioxide", ylab = "Residuals",
main = "Residuals vs Free Sulfur Dioxide")
abline(h = 0, col = "red")
```

Python Code:

```
import pandas as pd

import seaborn as sns

import numpy as np

import matplotlib.pyplot as plt

#wine+quality.zip

#winequality-red.csv

df = pd.read_csv("wine+quality/winequality-red.csv", sep=";") # index_col=0

# Reset index to ensure no duplicate index labels

df.reset_index(drop=True, inplace=True)

df.describe()

sns.pairplot(df, kind='scatter', hue='quality', palette='viridis')

plt.show()

sns.pairplot(df, kind='scatter')

plt.show()

variables = ['volatile acidity', 'citric acid', 'residual sugar', 'chlorides',

        'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH',

        'sulphates', 'alcohol', 'quality']


# Compute the correlation matrix
```

```
corr_matrix = df[variables].corr()


# Create a heatmap for the correlation matrix
plt.figure(figsize=(12, 10))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f", square=True)


plt.title('Variable-to-Variable Correlation Plot')
plt.yticks(rotation=0)


plt.show()
```

References

Fragoso, Rui, and António A.C. Vieira. “Efficiency analysis of the Portuguese wine industry using accounting and operational metrics.” *Results in Engineering*, vol. 14, June 2022, p. 100389, https://doi.org/10.1016/j.rineng.2022.100389.

“IBISWorld - Industry Market Research, Reports, and Statistics.” *IBISWorld Industry Reports*, 16 Feb. 2022, www.ibisworld.com/portugal/industry-statistics/wine-production/665/.

Markoski, Melissa M et al. “Molecular Properties of Red Wine Compounds and Cardiometabolic Benefits.” *Nutrition and metabolic insights* vol. 9 51-7. 2 Aug. 2016, doi:10.4137/NMI.S32909

Puckette James, Madeline “What Is Wine? A Beautiful Explanation.” *Wine Folly*, winefolly.com/deep-dive/what-is-wine/. Accessed 28 Apr. 2024.

“The History of Wine: Flowers, Wine and Gifts.” *Arena Flowers*, 26 Apr. 2024, www.arenaflowers.com/pages/history-of-wine/.

UCI ML Repository: Cortez, P., Cerdeira, A.L., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decis. Support Syst., 47*, 547-553.

“Wine Exports by Country Worldwide 2022.” Statista, 11 Sept. 2023, www.statista.com/statistics/240649/top-wine-exporting-countries-since-2007/.