MAT3024 Regression Analysis Assignment

2024-07-17

Group Members

Chong Kai Yuan (21081609)
Ethan Lee Jia Hua (21089750)
Daniel Chin Wei Jian (20095204)
Darren Yap Yee Shern (21001235)
Harreesh Dev Chandrasager (20083200)
Nur' Aliah Syamimi Binti Muhammad Sazali (21081765)

Introduction & Problem Statement

Predicting Wine Quality Using Regression Analysis

Wine quality assessment is a crucial aspect of the wine industry. The ability to predict wine quality based on various chemical properties can significantly enhance decision-making processes in viticulture. Chemical analysis has long been an established method for determining wine quality, with factors such as acidity, sugar content, and alcohol levels known to influence taste and overall consumer satisfaction (Marianthi Basalekou et al., 2023).

Consumers rely heavily on quality ratings when selecting wines. Accurate predictions of wine quality enable producers to better meet consumer expectations, resulting in higher satisfaction and repeat purchases (Corduas et al., 2013). This study aims to identify and quantify the impact of different physicochemical properties on the quality ratings of white Vinho Verde wines.

Determining Key Physicochemical Factors Influencing Wine Quality

Our primary objective is to create a regression model that highlights the most significant factors contributing to wine quality. The analysis will focus on identifying the key physicochemical properties that influence wine quality and quantifying their impact. As such, our task will help winemakers optimize their production processes and improve the quality of their wines.

Data Processing

Load the use of all relevant packages used under this assignment

```
library(car)
library(MASS)
library(dplyr)
library(olsrr)
library(leaps)
library(psych)
library(moments)
library(ggplot2)
library(pastecs)
library(corrplot)
library(tidyverse)
```

```
library(AICcmodavg)
library(rmdformats)
library(kableExtra)
library(ggThemeAssist)
options(scipen=10)
```

Import dataset into R Studio

```
Dataset<-read.csv("winequality-white.csv", sep =";")
str(Dataset)

## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049
0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

Dataset Description: Wine Quality Dataset

The dataset provided is considered for *vinho verde*, a unique product from Minho from the northwest region in Portugal and was collected from May (2004) to February (2007). Note that both red and white wine datasets were available, but the white wine dataset was chosen for analysis.

Figure 1: *Vinho Verde White Wine*

Variables Involved under considered dataset:

1. Sulphates (g(potassium sulphate)/dm^3)
2. Chlorides (g(sodium chloride)/dm^3)
3. Volatile acidity (g(acetic acid)/dm^3)
4. Fixed acidity (g(tartaric acid)/dm^3)
5. Total sulfur dioxide (mg/dm^3)
6. Free sulfur dioxide (mg/dm^3)
7. Residual sugar (g/dm^3)
8. Citric acid (g/dm^3)
9. Density (g/cm^3)
10. Alcohol (vol.%)
11. pH level
12. Quality

This Dataset consists of 4898 observations based on 12 variables

Data source: Wine Quality Dataset

Preliminary Regression Metrics and Diagnostic Measures

Here we provide a summary of all the key regressionary metrics used during our analysis.

1. Akaike Information Criterion (AIC) : measures relative and estimated quality of different statistical models; a lower value indicates a better-fitting model

2. Bayesian Information Criterion (BIC) : similar to AIC but with stricter penalty in regards to the number of parameters being fitted in the model; a lower value indicates a better-fitting model

3. R-squared (R^2) : statistical measure representing total proportion of variance explained by the regressor variables; values range from 0 to 1 with higher values indicating a higher quality model

4. Residual Standard Error (s) : measures the standard deviation of the residuals in a regression model; a lower value indicates smaller residuals and thus indicates a better fitted model

5. Variation Inflation Factor (VIF) : indicates how much the variance of the regression coefficients are inflated due to multicollinearity issues; values above 5 are considered as problematic and indicates the predictor is highly collinear

6. Ols Mallows Cp (Cp) : measures the quality of the fitted model while also considering its overall complexity; values close to the number of predictors plus one suggests a good model

Data Analysis

Descriptive Statistics

```
round(stat.desc(Dataset),2)
```

```
##              fixed.acidity volatile.acidity citric.acid residual.sugar
## nbr.val            4898.00          4898.00     4898.00        4898.00
## nbr.null              0.00             0.00       19.00           0.00
## nbr.na                0.00             0.00        0.00           0.00
## min                   3.80             0.08        0.00           0.60
## max                  14.20             1.10        1.66          65.80
## range                10.40             1.02        1.66          65.20
## sum                33574.75          1362.83     1636.87       31305.15
## median                6.80             0.26        0.32           5.20
## mean                  6.85             0.28        0.33           6.39
## SE.mean               0.01             0.00        0.00           0.07
## CI.mean.0.95          0.02             0.00        0.00           0.14
## var                   0.71             0.01        0.01          25.73
## std.dev               0.84             0.10        0.12           5.07
## coef.var              0.12             0.36        0.36           0.79
##              chlorides free.sulfur.dioxide total.sulfur.dioxide density
## nbr.val        4898.00             4898.00               4898.00 4898.00
## nbr.null          0.00                0.00                  0.00    0.00
## nbr.na            0.00                0.00                  0.00    0.00
## min               0.01                2.00                  9.00    0.99
## max               0.35              289.00                440.00    1.04
## range             0.34              287.00                431.00    0.05
## sum             224.19           172939.00             677690.50 4868.75
## median            0.04               34.00                134.00    0.99
## mean              0.05               35.31                138.36    0.99
## SE.mean           0.00                0.24                  0.61    0.00
## CI.mean.0.95      0.00                0.48                  1.19    0.00
## var               0.00              289.24               1806.09    0.00
## std.dev           0.02               17.01                 42.50    0.00
## coef.var          0.48                0.48                  0.31    0.00
##                     pH sulphates  alcohol  quality
## nbr.val       4898.00   4898.00  4898.00  4898.00
## nbr.null         0.00      0.00     0.00     0.00
## nbr.na           0.00      0.00     0.00     0.00
## min              2.72      0.22     8.00     3.00
## max              3.82      1.08    14.20     9.00
## range            1.10      0.86     6.20     6.00
## sum          15616.13   2399.27 51498.88 28790.00
## median           3.18      0.47    10.40     6.00
## mean             3.19      0.49    10.51     5.88
## SE.mean          0.00      0.00     0.02     0.01
## CI.mean.0.95     0.00      0.00     0.03     0.02
## var              0.02      0.01     1.51     0.78
## std.dev          0.15      0.11     1.23     0.89
## coef.var         0.05      0.23     0.12     0.15
```

INTERPRETATION

From the complete summary of the dataset, several key observations emerge regarding specific variables:

1. Citric Acid: Among all regressors, citric acid exhibited the smallest minimum value across all observations.

2. Total Sulfur Dioxide: This variable had the highest maximum value, the widest range, and the largest standard error of the mean.

Interestingly, when applying linear modeling techniques to our updated full model, both citric acid and total sulfur dioxide were found to be statistically insignificant. This lack of significance may be partially attributable to the distinctive properties they exhibit, as mentioned above.

Visualizing the dependent variable

```
# construct histogram

hist_data <- hist(Dataset$quality, breaks = seq(3, 10, by = 1), plot = FALSE, right =
FALSE)

hist_df <- data.frame(midpoints = hist_data$mids, counts = hist_data$counts)


ggplot(Dataset, aes(x = quality)) +
  geom_histogram(breaks = seq(3, 9, by = 1), fill = "grey", color = "white", closed =
"left", boundary = 3) +
  geom_line(data = hist_df, aes(x = midpoints, y = counts), color = "red", size = 1) +
  geom_point(data = hist_df, aes(x = midpoints, y = counts), color = "red", size = 2) +
  labs(title = "Figure 2: Histogram of Wine Quality",
       x = "Quality of Wine",
       y = "Frequency") +
  theme(plot.title = element_text(hjust=0.5)) +
  scale_x_continuous(breaks = seq(3, 9, by = 1)) # Customize x-axis labels to show end
values
```

```r
skewness_value<-skewness(Dataset$quality)
skewness_value
```

```
## [1] 0.1557487
```

INTERPRETATION

- The histogram shows a roughly bell-shaped curve, which suggests that the data might be approximately normally distributed. Many statistical methods, including linear regression, assume that the residuals of the model (differences between observed and predicted values) are normally distributed. Thus, A bell-shaped histogram for quality suggests that the response variable itself is normally distributed, which is a good starting point for meeting this assumption.
- The mean quality score is 5.88 as identified when using the stat.desc() function, this is consistent with the peak of the histogram around 6. This suggests that the mean is a good measure of central tendency for this data.

Fitting the Full Model

Here, we observe the model using the entirety of the wine dataset in terms of its AIC, BIC, Residual standard error and its R^2. Additionally, we investigate the whether each of the predictor variables appear to be statistically significant in predicting our dependent variable

```r
# fitting full model

full_model<-lm(quality~.,data=Dataset)
summary(full_model)

# extracting and finding relevant metrics

Residual_standard_error_full_model<-summary(full_model)$sigma
R_squared_full_model<-summary(full_model)$r.squared
full_model_AIC<-AIC(full_model)
full_model_BIC<-BIC(full_model)

# constructing table

table_1 <- data.frame(
  Metrics = c("AIC", "BIC", "R-squared", "Residual standard error"),
  full_model = c(full_model_AIC, full_model_BIC, R_squared_full_model,
Residual_standard_error_full_model))

# formatting table

table_1$full_model <- format(table_1$full_model, digits = 5)

# displaying table

knitr::kable(table_1, caption = "Table 1: Metrics from the Full Model") %>%
  kable_styling(bootstrap_options = "striped", full_width = F)
```

```
##
## Call:
```

```
## lm(formula = quality ~ ., data = Dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8348 -0.4934 -0.0379  0.4637  3.1143
##
## Coefficients:
##                         Estimate   Std. Error t value Pr(>|t|)
## (Intercept)           150.1928425   18.8041772   7.987 1.71e-15 ***
## fixed.acidity           0.0655200    0.0208737   3.139  0.00171 **
## volatile.acidity       -1.8631771    0.1137933 -16.373  < 2e-16 ***
## citric.acid             0.0220902    0.0957696   0.231  0.81759
## residual.sugar          0.0814828    0.0075273  10.825  < 2e-16 ***
## chlorides              -0.2472765    0.5465423  -0.452  0.65097
## free.sulfur.dioxide     0.0037328    0.0008441   4.422 9.99e-06 ***
## total.sulfur.dioxide   -0.0002857    0.0003781  -0.756  0.44979
## density              -150.2841806   19.0745080  -7.879 4.04e-15 ***
## pH                      0.6863437    0.1053791   6.513 8.10e-11 ***
## sulphates               0.6314765    0.1003856   6.291 3.44e-10 ***
## alcohol                 0.1934757    0.0242214   7.988 1.70e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7514 on 4886 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2803
## F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

Table 1: Metrics from the Full Model

Metrics	full_model
AIC	11113.48027
BIC	11197.93584
R-squared	0.28187
Residual standard error	0.75136

Observation of Full Model

From the output of the summary of the dataset, we can observe that three predictor variables can be highlighted and appears to not be statistically significant. Those variables being:

1. Citric Acid (g/dm^3)
2. Total sulfur dioxide (mg/dm^3)
3. Chlorides ($g(sodium\ chloride)/dm^3$)

From here we can start making the hypothesis that when performing the variable selection process to find our “best” regression model, it is more likely than not that one or several of these variables will likely not be included under the fitted model.

Investigating Problematic Data Points

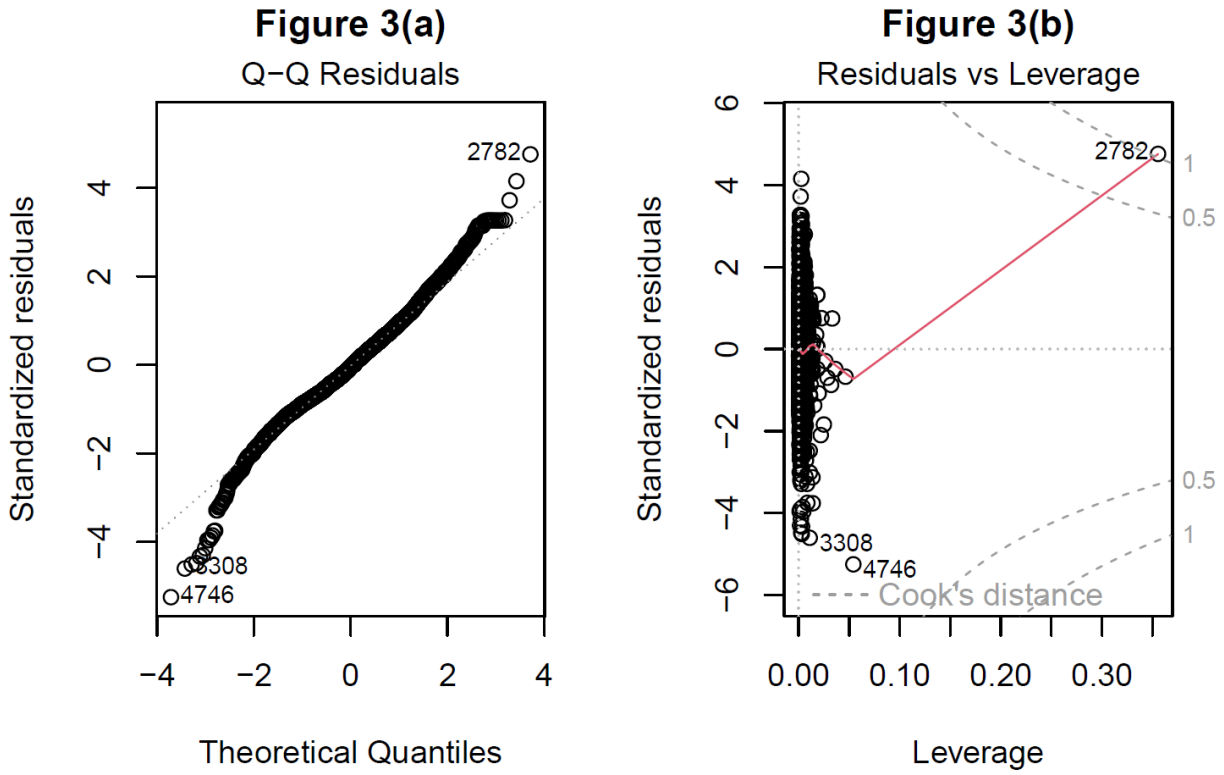
Before we proceed into building our regression models, it is crucial to ensure the reliability and accuracy of our current dataset by investigating any potentially problematic observation points. In here, we decided to take into account the following metrics:

1) Leverage (h_{ii}) - may affect regression coefficients
2) Cook's Distance (D_i) - to observe influential points
3) Studentized Residuals (r_i) - to observe potential outliers

Observation points that exceed tolerated thresholds in these diagnostic measures will then be considered as problematic. Consequently, an updated full model will be created after removing these problematic points, ensuring a more robust and accurate predictive model.

Visualizing Model using Diagnostic Plots

```
par(mfrow=c(1,2))
plot(full_model,which=2, main="Figure 3(a)") # QQ Residuals
plot(full_model,which=5, main="Figure 3(b)") # Residuals vs Leverage
```



INTERPRETATION

QQ Residuals: The lines curve at the ends, implying that under the full model, the residuals have heavier tails than a normal distribution, this acts as an indicator for the presence of outliers in the dataset.

Residuals vs Leverage: Presence of observations clearly outside the Cook's distance contour lines, indicating presence of influential points within the Dataset.

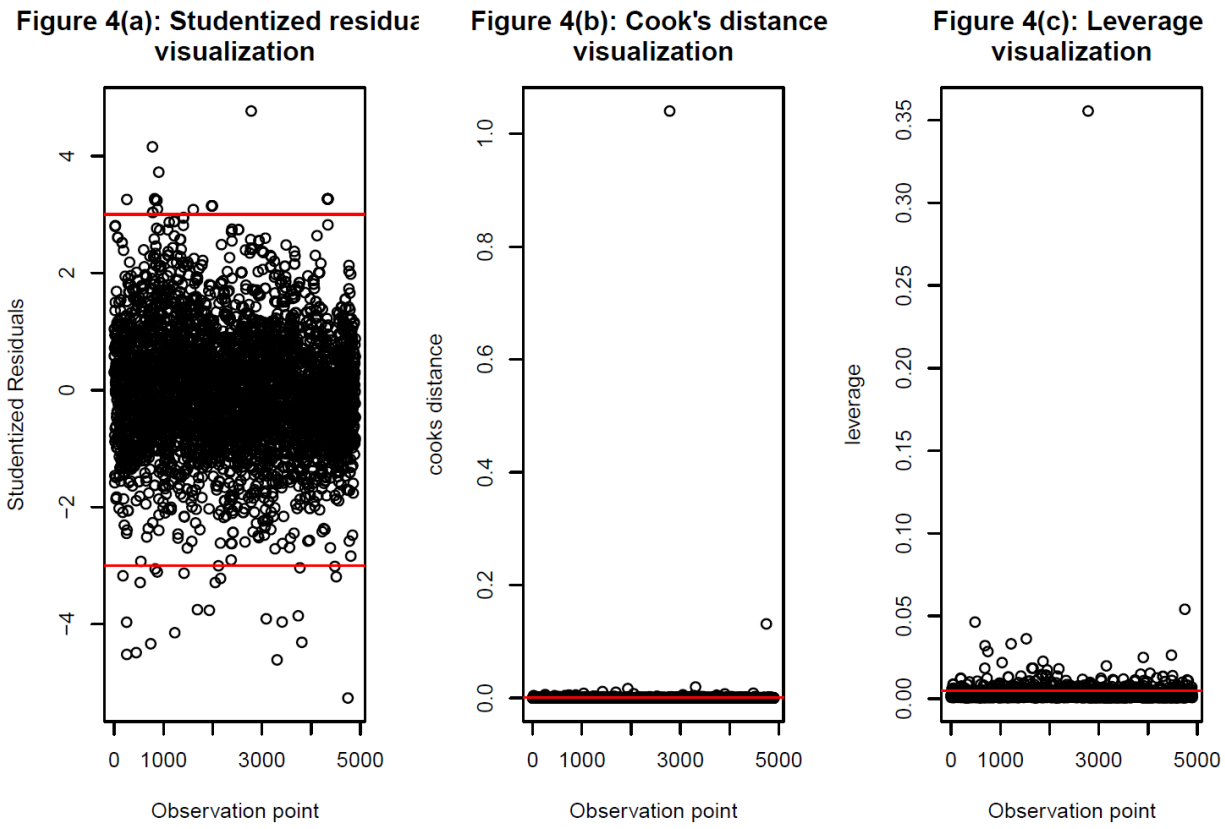
Now, we will make use of the following thresholds in deciding which observations points are considered problematic which should be removed from our dataset.

***Table 2**: Diagnostics values and their associated thresholds*

Diagnostic Value Tested	Tolerated Threshold
Studentized Residuals (ri)	< 3
Cook's Distance (Di)	$< 4/n$
Leverage (hii)	$< 2p/n$

Under different applications, the threshold for the cook's distance may also be < 1, however, we decided to use the threshold of $< 4/n$ instead being that in this way, the number of observations in the dataset is taken into account when performing our investigation of problematic observation points.

```
# studentized residuals

studentized_residuals_full_model<-studres(full_model)

# cooks distance

cd_full_model<-cooks.distance(full_model)

# leverage

hat_values_full_model<-hatvalues(full_model)

# plot for visualization

par(mfrow=c(1,3))

plot(studentized_residuals_full_model, main="Figure 4(a): Studentized residuals
\nvisualization", ylab="Studentized Residuals",xlab="Observation point")
abline(h=3,col="red")
abline(h=-3,col="red")

plot(cd_full_model, main="Figure 4(b): Cook's distance \nvisualization", ylab="cooks
distance",xlab="Observation point")
abline(h = 4/(nrow(Dataset)), col = "red")

plot(hat_values_full_model, main="Figure 4(c): Leverage \nvisualization",ylab="leverage",
xlab="Observation point")
abline(h= 2 * (length(full_model$coefficients)) / nrow(Dataset),col ="red")
```




```
# indicate all flagged points that need to be removed

high_studentized_residual_points<-which(studentized_residuals_full_model>(abs(3)))
high_leverage_points<-
which(hat_values_full_model>(2*(length(full_model$coefficients))/nrow(Dataset)))
high_cooks_distance_points<-which(cd_full_model>(4/nrow(Dataset)))

all_potential_issues<-
unique(c(high_studentized_residual_points,high_leverage_points,high_cooks_distance_points
))

# create updated full model

updated_Dataset<-Dataset[-all_potential_issues, ]
```

Fitting Updated Full Model

```
# fitting updated full model and observing observations

updated_full_model<-lm(quality~.,data=updated_Dataset)
str(updated_Dataset)

## 'data.frame':    4470 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049
0.044 ...
```

```
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

```r
# inbestigating updated full model

summary(updated_full_model)

# extracting and finding relevant metrics

Residual_standard_error_updated_full_model<-summary(updated_full_model)$sigma
R_squared_updated_full_model<-summary(updated_full_model)$r.squared
updated_full_model_AIC<-AIC(updated_full_model)
updated_full_model_BIC<-BIC(updated_full_model)

# constructing the table

table_2 <- data.frame(
  Metrics = c("AIC", "BIC", "R-squared", "Residual standard error"),
  full_model = c(full_model_AIC, full_model_BIC, R_squared_full_model,
Residual_standard_error_full_model),
  updated_full_model = c(updated_full_model_AIC, updated_full_model_BIC,
R_squared_updated_full_model, Residual_standard_error_updated_full_model))

# formatting table

table_2$full_model <- format(table_2$full_model, digits = 5)
table_2$updated_full_model <- format(table_2$updated_full_model, digits = 5)

# display updated table

knitr::kable(table_2, caption = "Table 3: Comparison of Original vs Updated Full Model")
%>%
  kable_styling(bootstrap_options = "striped", full_width = F)
```

```
##
## Call:
## lm(formula = quality ~ ., data = updated_Dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.98635 -0.47870 -0.04531  0.44472  2.18522
##
## Coefficients:
##                         Estimate   Std. Error t value Pr(>|t|)
## (Intercept)           178.1641778   23.0787764   7.720 1.43e-14 ***
## fixed.acidity           0.1110092    0.0228811   4.852 1.27e-06 ***
## volatile.acidity       -1.7635996    0.1185578 -14.875  < 2e-16 ***
## citric.acid             0.0642252    0.0982086   0.654  0.51317
## residual.sugar          0.0871529    0.0085888  10.147  < 2e-16 ***
## chlorides              -2.9133938    0.9265630  -3.144  0.00168 **
```

```
## free.sulfur.dioxide     0.0040293    0.0008528   4.725 2.38e-06 ***
## total.sulfur.dioxide   -0.0003297    0.0003679  -0.896  0.37022
## density              -178.8304217   23.3912821  -7.645 2.54e-14 ***
## pH                      0.8673713    0.1077638   8.049 1.06e-15 ***
## sulphates               0.6937781    0.0996787   6.960 3.89e-12 ***
## alcohol                 0.1474818    0.0287033   5.138 2.89e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6528 on 4458 degrees of freedom
## Multiple R-squared:  0.3274, Adjusted R-squared:  0.3258
## F-statistic: 197.3 on 11 and 4458 DF,  p-value: < 2.2e-16
```

Table 3: Table 3: Comparison of Original vs Updated Full Model

Metrics	full_model	updated_full_model
AIC	11113.48027	8886.90538
BIC	11197.93584	8970.17224
R-squared	0.28187	0.32741
Residual standard error	0.75136	0.65283

INTERPRETATION

By inspection, the new dataset has 4470 observations which shows that a total of 428 observation points have been removed. At first glance, we can see that after removing the problematic points, the updated model evidently exhibit better model properties. It has a lower AIC and BIC value, which indicates that the model has an improved fit and better performance relative to the previous model. Additionally, it has a smaller residual standard error, which shows that the predictions of the model are more accurate while having less variability. The R-squared value is also higher, indicating that a greater proportion of the variance in the dependent variable is explained by the independent variables in the model which is ideal for a regression model.

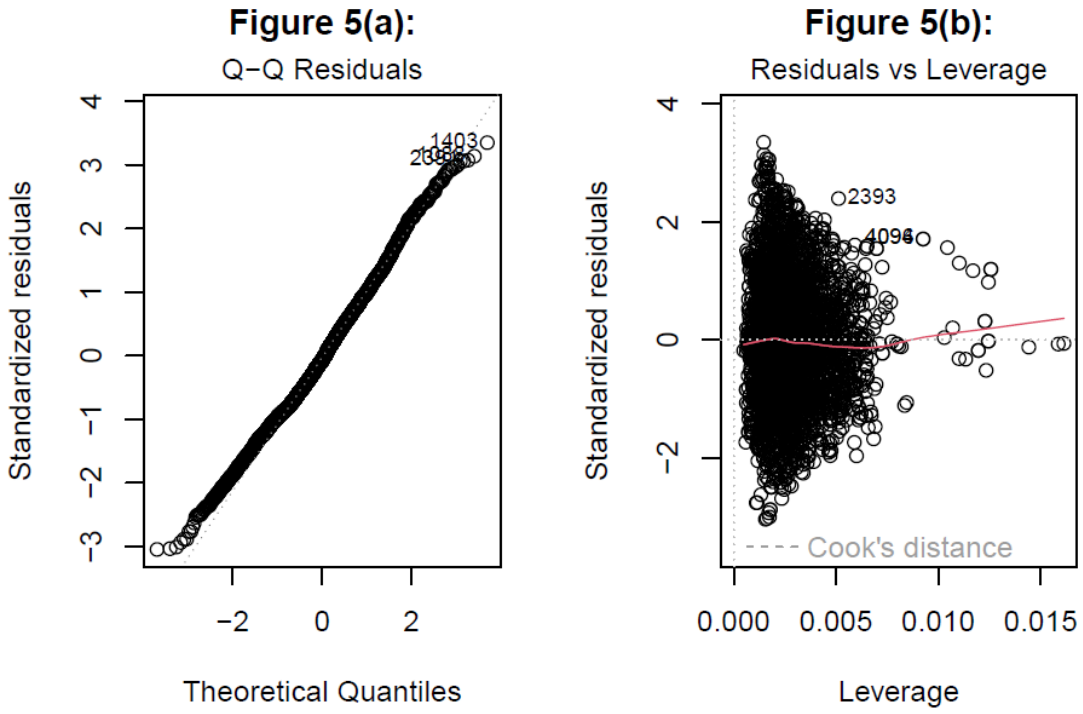
Additionally, from the output of the summary of the updated dataset, we can see that now only two of the predictor variables appears to not be statistically significant. Those variables being:

1. Citric Acid (g/dm^3)
2. Total sulfur dioxide (mg/dm^3)

This strengthens our previous assumption that these variables have a higher chance of being excluded from the “best” regression model later when performing the variable selection process.

Inspecting Updated Model using Diagnostic Plots

```
par(mfrow=c(1,2))
plot(updated_full_model,which=2, main="Figure 5(a):") # QQ Residuals
plot(updated_full_model,which=5, main="Figure 5(b):") # Residuals vs Leverage
```



INTERPRETATION

QQ Residuals: Seem to follow along the 45 degree reference line. This shows that under the new model, the residuals are better suited by assuming a normal distribution, which is an ideal condition for regression analysis.

Residuals vs Leverage: There appear to be no points that lie outside the designated Cook's distance contour lines, indicating that there are no overly influential points under the new model.

Finding "Best" Regression Model

Investigating Correlation

```
# finding correlation values between all variables

correlation_matrix<-round(cor(updated_Dataset),2)

# visualize correlation in a diagram
par(mfrow=c(1,2))

corrplot::corrplot(correlation_matrix, title="Figure 6(a) : Correlation Plot \nbetween
Regressors",mar=c(0,0,3,0))
corrplot(correlation_matrix, method = "number", number.cex = 0.45,title="Figure 6(b) :
Correlation Plot \nbetween Regressors",mar=c(0,0,3,0))
```

Figure 6(a) : Correlation Plot between Regressors

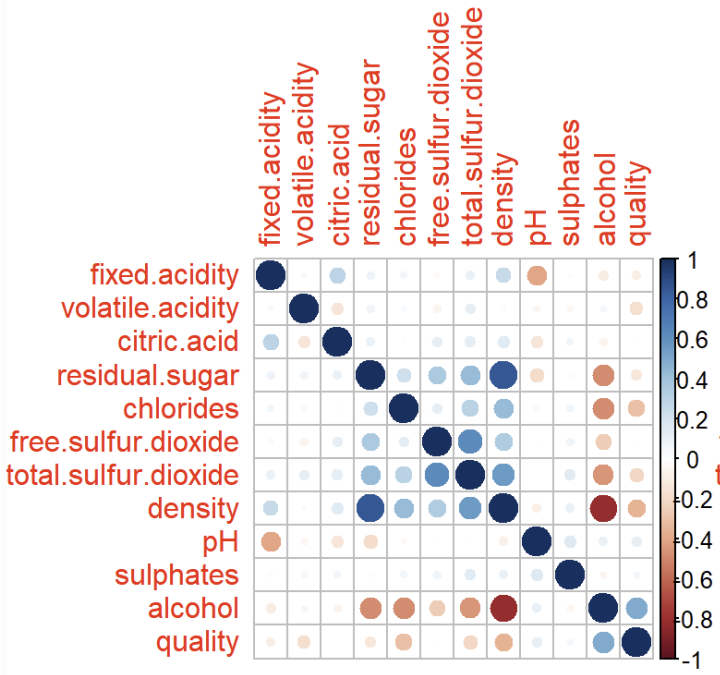
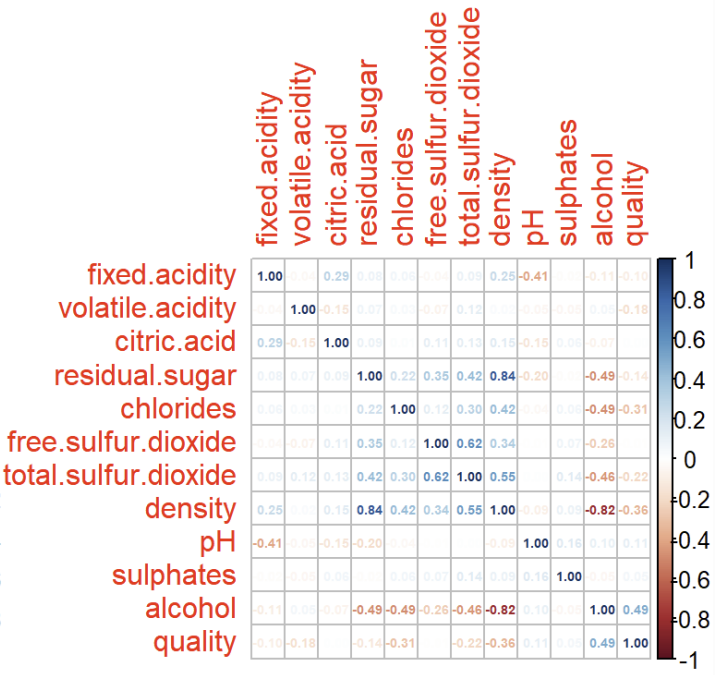


Figure 6(b) : Correlation Plot between Regressors



INTERPRETATION

We can separate the interpretation into 2 parts:

1) Dependent variable with the regressor variables:

By referring to the last row under the correlation plot, we can identify the correlation between the dependent variable against all 11 other independent variables under this dataset. We notice out of all variables, there are two variables that stand out, that being alcohol (vol.%) and also density (g/cm^3). The exact correlation values are 0.49 and -0.36 respectively and a higher value indicates a stronger linear relationship and the potential to be a valuable predictor of y.

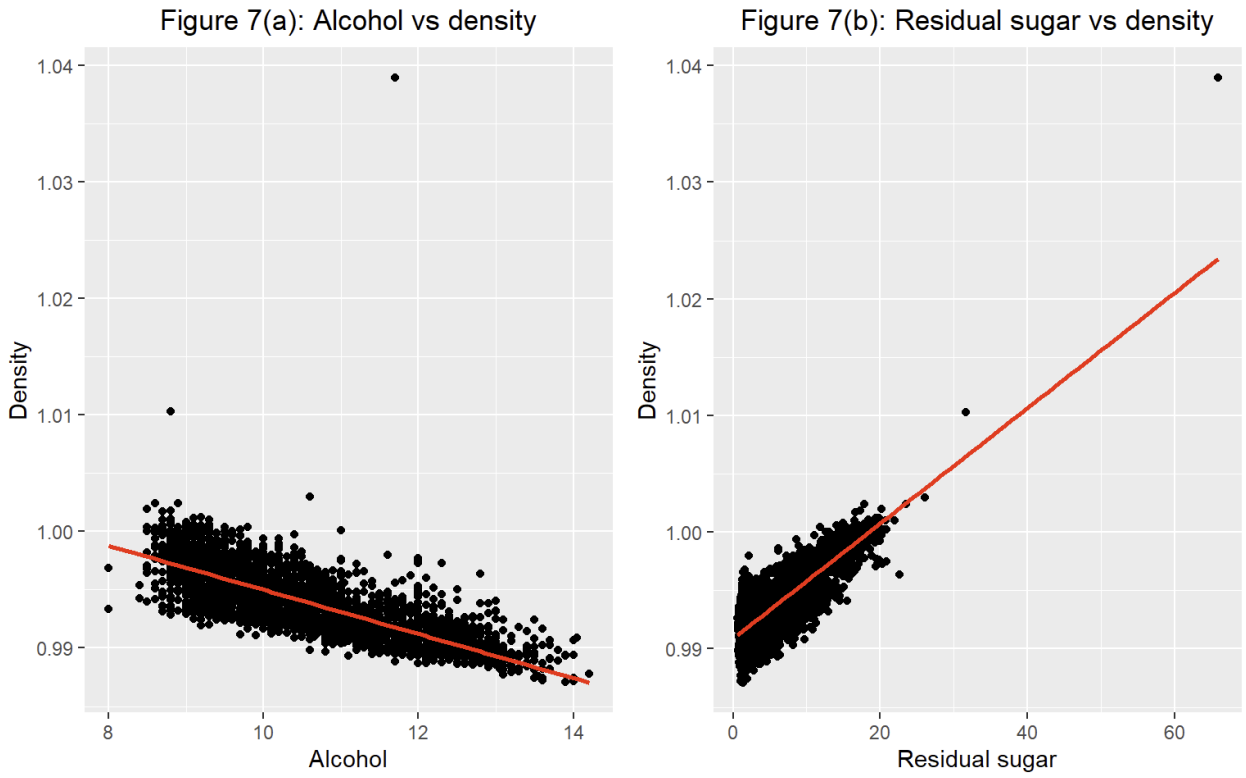
2) Regressor variables with each other (multicollinearity)

By referring to the upper or lower non-diagonal entries under the correlation plot, we observe that there appears to be very little correlation between the x variables, some of which are even orthogonal (no linear relationship between the regressors) and thus indicating that issues of multicollinearity is minimal under this dataset. In other words, this means that it is likely each predictor provides unique and valuable information in predicting the dependent variable. However, there are two exceptions when observing the correlation plot:

- Alcohol & Density : correlation value of -0.82
- Density & Residual sugar : correlation value of 0.84

```
plota = ggplot(Dataset, aes(x = alcohol, y = density)) +
  geom_point() +            # Add points
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Figure 7(a): Alcohol vs density",
       x = "Alcohol",
       y = "Density") + theme(plot.title = element_text(hjust=0.5))
```

```r
plotb = ggplot(Dataset, aes(x = residual.sugar, y = density)) +
  geom_point() +            # Add points
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Figure 7(b): Residual sugar vs density",
       x = "Residual sugar",
       y = "Density") + theme(plot.title = element_text(hjust=0.5))

require(gridExtra)
grid.arrange(plota,plotb, ncol=2)
```



Both are relatively high in terms of their correlation, thus further investigation into their multicollinearity properties should be investigated using other diagnostics. In this case, we will make use of the "variance inflation factor". Note that we will use the threshold as VIF values > 5 being considered to have high multicollinearity problems.

Investigating Multicollinearity for Full Model

```r
# investigating multicollinearity under model

cat("================================================================================
==============","\n","\n")
cat("( VIF: updated full model )","\n","\n")

vif_updated_full_model<-vif(updated_full_model)
vif_updated_full_model
```

```
cat("================================================================================
===============","\n","\n")

##
================================================================================
==========
##
## ( VIF: updated full model )
##
##        fixed.acidity     volatile.acidity          citric.acid
##             3.502695             1.129971             1.157082
##       residual.sugar            chlorides  free.sulfur.dioxide
##            18.945214             1.395005             1.836549
## total.sulfur.dioxide              density                   pH
##             2.373692            47.964676             2.604507
##            sulphates              alcohol
##             1.168026            12.698705
##
================================================================================
==========
##
```

```
# pinpoint key predictors with high vif values

vif_alcohol_full_model <- vif_updated_full_model['alcohol']
vif_density_full_model <- vif_updated_full_model['density']
vif_residual_sugar_full_model <- vif_updated_full_model['residual.sugar']

# constructing table

table_3 <- data.frame(
vif_full_model = c(vif_alcohol_full_model, vif_residual_sugar_full_model,
vif_density_full_model ))

#display table

knitr::kable(table_3, caption = "Table 4: High VIF values under the Updated Full Model ")
%>%
  kable_styling(bootstrap_options = "striped", full_width = F)
```

Table 4: High VIF values under the Updated Full Model

	vif_full_model
alcohol	12.69870
residual.sugar	18.94521
density	47.96468

INTERPRETATION

Under the updated full model, these three variables exhibit high multicollinearity indications. Thus when building our regression models later, we will need to take into account these values and make appropriate adjustments if necessary.

Fitting Simple Linear Model

With the updated dataset, we are now ready to start finding our "best" regression model. We will employ the use of a forward step procedure where we start with including one regressor variable in our model, and work our way up towards the best regression model by adding one regressor variable at a time.

In order to determine the first variable that should be included under the simple linear regression model, we will make use of the variable with the highest correlation value to our dependent variable, i.e. Alcohol (vol.%). Note that from here on out, "best_model_x" will refer to a the fitted model in respect to "x" number of variables, for example best_model_1 refers to 1 predictor being fitted in the model, or basically the best simple linear regression model.

```
# fitting the "best" simple linear model

best_model_1<-lm(quality~alcohol,data = updated_Dataset)
summary(best_model_1)

# extracting and finding relevant metrics

Residual_standard_error_best_model_1<-summary(best_model_1)$sigma
R_squared__best_model_1<-summary(best_model_1)$r.squared
best_model_1_AIC<-AIC(best_model_1)
best_model_1_BIC<-BIC(best_model_1)

# displaying results

table_4 <- data.frame(
  Metrics = c("AIC", "BIC", "R-squared", "Residual standard error"),
  updated_full_model = c(updated_full_model_AIC, updated_full_model_BIC,
R_squared_updated_full_model, Residual_standard_error_updated_full_model),
  best_model_1 = c(best_model_1_AIC, best_model_1_BIC, R_squared__best_model_1,
Residual_standard_error_best_model_1))

# formatting table

table_4$best_model_1 <- format(table_4$best_model_1, digits = 5)
table_4$updated_full_model <- format(table_4$updated_full_model, digits = 5)

# display table of simple model and full model

knitr::kable(table_4, caption = "Table 5: Comparison of Full vs Simple Linear Model ")
%>%
  kable_styling(bootstrap_options = "striped", full_width = F)
```

```
##
## Call:
## lm(formula = quality ~ alcohol, data = updated_Dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.20806 -0.52864 -0.01572  0.46514  2.14457
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.521459   0.090578   27.84   <2e-16 ***
## alcohol     0.320574   0.008558   37.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6936 on 4468 degrees of freedom
## Multiple R-squared:  0.239,  Adjusted R-squared:  0.2388
## F-statistic:  1403 on 1 and 4468 DF,  p-value: < 2.2e-16
```

Table 5: Comparison of Full vs Simple Linear Model

Metrics	updated_full_model	best_model_1
AIC	8886.90538	9419.08213
BIC	8970.17224	9438.29757
R-squared	0.32741	0.23898
Residual standard error	0.65283	0.69364

INTERPRETATION

When comparing the metrics above with the simple linear model against the new full model, it is apparent that the simple linear model have values that are less ideal, such as a higher AIC and Residual standard error. However, this is to be expected as only one regressor has been added, in theory, as we add more explanatory variables, the metrics will eventually get more and more ideal until it reaches its optimum value stance where we can conclude the best model has been found.

Fitting Multiple Linear Model

To decide on the next variables added, we decided to see which of the remaining regressor variables when added to the current model has the best statistical significance ,i.e. the lowest p-value. Then by setting our stopping criterion based on a p-value of 0.05, any variables when added to the model results in a p-value greater than 0.05 will be deemed as non significant and the best model will have been determined.

```
cortable = as.data.frame(cor(updated_Dataset))
cortablefiltered = subset(cortable, select = quality)
```

```r
numbering = seq(12)
names = names(cortable)

data1 = data.frame(cortablefiltered, numbering, names)
data1 = data1[-12,]
data1$quality = abs(data1$quality)

pairs = c(NULL)
naming = names[-12]
naming = naming[-11]
nametest = naming

a = 12
b= 10

for (i in 1:a) {
  pvalue = c(NULL)
  AIC = c(NULL)
  count = c(NULL)
  stop = FALSE

  framefull = data1 %>% arrange(desc(quality))
  pairs[1] = framefull[1,3]
  cat("\nquality ~", pairs[i], "+ \n")

  for (j in 1:b) {
    formula = as.formula(paste("quality ~",pairs[i]," + ", nametest[j]))
    pvalue[j] = (summary(lm(formula, data = updated_Dataset))$coefficients)[i+2,4]
    AIC[j] = summary(glm(formula, data = updated_Dataset))$aic
    cat("\n\n")
    print(summary(lm(formula, data = updated_Dataset))$coefficients)
    count[j] = j
  }

  framefull2 = data.frame(pvalue, nametest, count)
  framefull2 = framefull2 %>% arrange(pvalue)
  nametest = nametest[-framefull2[1,3]]

  pairs[i+1] = paste(c(pairs[i], framefull2[1,2]), collapse = " + ")
  b = b - 1
  cat("\n", i," iteration done\n")

  cat("\nCurrent AIC: ", summary(glm(as.formula(paste("quality ~",pairs[i])), data =
updated_Dataset))$aic, "\n\n")

  print(AIC)

  cat("\nP-values for each of next variables\n")
  print(pvalue)
  print("============================================================")

  if (pvalue[framefull2[1,3]]>0.05){
    stop = TRUE
    formula2 = paste("quality ~", pairs[i])
```

```
    print(summary(glm(formula2, data = updated_Dataset)))
    break
  }
  if (stop){break}
}
```

```
##
## quality ~ alcohol +
##
##
##                   Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept)    2.89341627 0.134079289 21.579890  2.489220e-98
## alcohol        0.31711142 0.008595353 36.893358 2.484199e-260
## fixed.acidity -0.04902708 0.013046307 -3.757928  1.735128e-04
##
##
##                     Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)       2.9531655 0.09196732  32.11103 9.602901e-204
## alcohol           0.3278254 0.00832984  39.35554 4.175346e-291
## volatile.acidity -1.8709469 0.11534011 -16.22113  1.531971e-57
##
##
##              Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) 2.4312823 0.098370455 24.715575 1.617611e-126
## alcohol     0.3220096 0.008575948 37.547984 2.053794e-268
## citric.acid 0.2278051 0.097206895  2.343508  1.914697e-02
##
##
##                   Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept)    1.96701077 0.111189405 17.690631  9.646230e-68
## alcohol        0.36095935 0.009738901 37.063664 1.993538e-262
## residual.sugar 0.02020446 0.002385842  8.468485  3.331427e-17
##
##
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  3.0686182 0.12840453 23.898053 6.973219e-119
## alcohol      0.2920822 0.00976276 29.917999 2.035546e-179
## chlorides   -5.6945346 0.95083910 -5.988957  2.277688e-09
##
##
##                       Estimate   Std. Error  t value      Pr(>|t|)
## (Intercept)         2.06969415 0.1019149214 20.30806  8.978000e-88
## alcohol             0.34208060 0.0087838140 38.94443 6.735226e-286
## free.sulfur.dioxide 0.00641216 0.0006862682  9.34352  1.433473e-20
##
##
##                          Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)          2.4529188782 0.1254825164 19.5478936  1.023184e-81
## alcohol              0.3241192472 0.0096655641 33.5334021 4.293517e-220
## total.sulfur.dioxide 0.0002261661 0.0002865465  0.7892824  4.299889e-01
##
##
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) -31.54804 6.29284861 -5.013316  5.557128e-07
## alcohol       0.38600 0.01479154 26.095992 7.608848e-140
## density      33.58283 6.20231923  5.414559  6.466238e-08
```

```
##
##
##              Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept) 1.5119267 0.235704118  6.414511  1.559320e-10
## alcohol     0.3164725 0.008584471 36.865689 5.434529e-260
## pH          0.3300271 0.071165754  4.637443  3.628207e-06
##
##
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 2.2369583 0.10424707 21.458237  2.681675e-97
## alcohol     0.3229069 0.00854159 37.804072 1.350621e-271
## sulphates   0.5339578 0.09780526  5.459397  5.036464e-08
##
##  1  iteration done
##
## Current AIC:  9419.082
##
##  [1] 9406.973 9165.243 9415.590 9349.889 9385.334 9334.565 9420.459 9391.841
##  [9] 9399.613 9391.356
##
## P-values for each of next variables
##  [1] 1.735128e-04 1.531971e-57 1.914697e-02 3.331427e-17 2.277688e-09
##  [6] 1.433473e-20 4.299889e-01 6.466238e-08 3.628207e-06 5.036464e-08
## [1] "========================================================="
##
## quality ~ alcohol + volatile.acidity +
##
##
##                     Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)       3.37591042 0.13353091  25.281864 6.436872e-132
## alcohol           0.32398039 0.00835978  38.754655 1.704027e-283
## volatile.acidity -1.88602142 0.11516038 -16.377347  1.374014e-58
## fixed.acidity    -0.05526277 0.01267849  -4.358783  1.337280e-05
##
##
##                      Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)       2.949746808 0.100978100  29.211748 8.025553e-172
## alcohol           0.327869408 0.008348053  39.274954 4.490153e-290
## volatile.acidity -1.869574125 0.116560479 -16.039520  2.468220e-56
## citric.acid       0.007836203 0.095522685   0.082035  9.346226e-01
##
##
##                     Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept)       2.30173419 0.109225979  21.07314  4.635369e-94
## alcohol           0.37822859 0.009471332  39.93404 1.808240e-298
## volatile.acidity -2.01377992 0.114667955 -17.56184  8.120558e-67
## residual.sugar    0.02493961 0.002323451  10.73387  1.484097e-26
##
##
##                     Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)       3.3983335 0.126643230  26.833914 3.479904e-147
## alcohol           0.3040305 0.009528518  31.907432 1.955228e-201
## volatile.acidity -1.8323275 0.115268213 -15.896208  2.168681e-55
## chlorides        -4.7258041 0.927138291  -5.097194  3.589718e-07
##
```

```
##
##                         Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)          2.533678977 0.103460704  24.489288 2.204552e-124
## alcohol              0.346880754 0.008554964  40.547307 2.437112e-306
## volatile.acidity    -1.809422818 0.114634296 -15.784306  1.169336e-54
## free.sulfur.dioxide  0.005752526 0.000669272   8.595199  1.133681e-17
##
##
##                           Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)           2.6718338611 0.1225133814  21.808506 2.782159e-100
## alcohol               0.3434056863 0.0094537181  36.324934 2.284160e-253
## volatile.acidity     -1.9360844611 0.1167170300 -16.587849  5.156729e-60
## total.sulfur.dioxide  0.0009779298 0.0002818065   3.470217  5.249575e-04
##
##
##                      Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)      -41.8619220 6.12979034  -6.829258  9.682299e-12
## alcohol            0.4142627 0.01443364  28.701202 2.072109e-166
## volatile.acidity  -1.9580837 0.11528630 -16.984530  9.606749e-63
## density           44.1947408 6.04426706   7.311844  3.106579e-13
##
##
##                     Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)       2.1285584 0.232464000   9.156508  7.979901e-20
## alcohol           0.3244013 0.008364044  38.785224 7.016623e-284
## volatile.acidity -1.8459216 0.115343186 -16.003734  4.253663e-56
## pH                0.2676855 0.069326174   3.861247  1.144054e-04
##
##
##                     Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)       2.7032547 0.105500293  25.623196 3.225193e-135
## alcohol           0.3297132 0.008318704  39.635164 1.181577e-294
## volatile.acidity -1.8430143 0.115204183 -15.997807  4.654453e-56
## sulphates         0.4569428 0.095250186   4.797291  1.660683e-06
##
##  2  iteration done
##
## Current AIC:  9165.243
##
## [1] 9148.268 9167.237 9053.387 9141.314 9093.905 9155.206 9114.050 9152.346
## [9] 9144.268
##
## P-values for each of next variables
## [1] 1.337280e-05 9.346226e-01 1.484097e-26 3.589718e-07 1.133681e-17
## [6] 5.249575e-04 3.106579e-13 1.144054e-04 1.660683e-06
## [1] "========================================================="
##
## quality ~ alcohol + volatile.acidity + residual.sugar +
##
##
##                     Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept)       2.74925461 0.143771416  19.12240  2.074964e-78
## alcohol           0.37480495 0.009475546  39.55497 1.266795e-293
## volatile.acidity -2.03215226 0.114454359 -17.75513  3.311630e-68
## residual.sugar    0.02530221 0.002319056  10.91057  2.267096e-27
```

```
## fixed.acidity    -0.05973964 0.012520909  -4.77119  1.889860e-06
##
##
##                      Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)       2.33203979 0.115033794  20.272650  1.742616e-87
## alcohol           0.37812093 0.009472511  39.917708 2.987491e-298
## volatile.acidity -2.02867218 0.116034178 -17.483402  2.953601e-66
## residual.sugar    0.02510752 0.002332109  10.766011  1.056840e-26
## citric.acid      -0.07951761 0.094665399  -0.839986  4.009612e-01
##
##
##                      Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)       2.72831924 0.140122066  19.471018  4.103434e-81
## alcohol           0.35521768 0.010575802  33.587776 1.022300e-220
## volatile.acidity -1.97561291 0.114652305 -17.231341  1.803848e-64
## residual.sugar    0.02460455 0.002318669  10.611498  5.362782e-26
## chlorides        -4.43562175 0.916174748  -4.841458  1.332450e-06
##
##
##                         Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)          2.106437058 0.1137007819  18.526144  7.010613e-74
## alcohol              0.383812966 0.0094824370  40.476195 2.050960e-305
## volatile.acidity    -1.947486373 0.1147805155 -16.967047  1.272022e-62
## residual.sugar       0.021000522 0.0024083442   8.719901  3.867215e-18
## free.sulfur.dioxide  0.004089122 0.0006905886   5.921213  3.435491e-09
##
##
##                           Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)           2.2293543827 0.1286586761  17.327665  3.770532e-65
## alcohol               0.3819004760 0.0100797589  37.887858 1.266076e-272
## volatile.acidity     -2.0307260324 0.1157659974 -17.541645  1.133726e-66
## residual.sugar        0.0243537456 0.0023877104  10.199623  3.656643e-24
## total.sulfur.dioxide  0.0003047922 0.0002863208   1.064513  2.871540e-01
##
##
##                       Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)       56.04161520 12.542367656   4.468185 8.083808e-06
## alcohol            0.31034393  0.018448853  16.821855 1.286081e-61
## volatile.acidity  -2.01231466  0.114446251 -17.583054 5.729616e-67
## residual.sugar     0.04308934  0.004829039   8.922965 6.504028e-19
## density          -53.46350897 12.477393766  -4.284830 1.867294e-05
##
##
##                      Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)       0.99723915 0.248946012   4.005845  6.280163e-05
## alcohol           0.37774453 0.009436955  40.028221 1.152222e-299
## volatile.acidity -1.98929335 0.114324613 -17.400394  1.150646e-65
## residual.sugar    0.02725599 0.002348821  11.604116  1.078346e-30
## pH                0.40382652 0.069311662   5.826242  6.067429e-09
##
##
##                      Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)       2.00976459 0.121683614  16.516312  1.581326e-59
## alcohol           0.38150573 0.009461560  40.321654 1.989762e-303
## volatile.acidity -1.98622706 0.114426441 -17.358113  2.295272e-65
```

```
## residual.sugar    0.02552721 0.002318808  11.008766  7.881223e-28
## sulphates         0.50578081 0.094098395   5.375021  8.047703e-08
##
##  3  iteration done
##
## Current AIC:  9053.387
##
## [1] 9032.655 9054.680 9031.982 9020.424 9054.252 9037.044 9021.532 9026.557
##
## P-values for each of next variables
## [1] 0.000001889859870 0.400961226510513 0.000001332449888 0.000000003435491
## [5] 0.287153981531523 0.000018672937600 0.000000006067429 0.000000080477034
## [1] "============================================================"
##
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
##
##
##                         Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)          2.522101416 0.1490253920  16.923971  2.533915e-62
## alcohol              0.380382463 0.0094974116  40.051172 5.983687e-300
## volatile.acidity    -1.968120924 0.1146563602 -17.165388  5.250393e-64
## residual.sugar       0.021568930 0.0024072611   8.959946  4.681829e-19
## free.sulfur.dioxide  0.003838619 0.0006916922   5.549606  3.028867e-08
## fixed.acidity       -0.053890075 0.0125237717  -4.303023  1.720898e-05
##
##
##                         Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)          2.148171675 0.1186281428  18.108449  8.830487e-71
## alcohol              0.383731358 0.0094821165  40.468956 2.598299e-305
## volatile.acidity    -1.968399266 0.1160209271 -16.965898  1.296848e-62
## residual.sugar       0.021192812 0.0024132511   8.781851  2.254304e-18
## free.sulfur.dioxide  0.004144843 0.0006920263   5.989430  2.271249e-09
## citric.acid         -0.116488442 0.0944996790  -1.232686  2.177578e-01
##
##
##                         Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)          2.534654975 0.1433181439  17.685514  1.053620e-67
## alcohol              0.360695606 0.0105752163  34.107634 8.478602e-227
## volatile.acidity    -1.908862395 0.1147600089 -16.633516  2.522123e-60
## residual.sugar       0.020648985 0.0024032735   8.592025  1.165090e-17
## free.sulfur.dioxide  0.004104297 0.0006888328   5.958336  2.744261e-09
## chlorides           -4.460136449 0.9126646898  -4.886939  1.060061e-06
##
##
##                           Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)           2.2765821872 0.1282766226  17.747444  3.767682e-68
## alcohol               0.3737171748 0.0101125408  36.955814 4.452257e-261
## volatile.acidity     -1.8697124164 0.1178759870 -15.861690  3.656345e-55
## residual.sugar        0.0215790310 0.0024149223   8.935704  5.808811e-19
## free.sulfur.dioxide   0.0054740270 0.0008433231   6.491020  9.454901e-11
## total.sulfur.dioxide -0.0009950132 0.0003483246  -2.856569  4.302258e-03
##
##
##                          Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)          55.111742386 12.4963139555   4.410240 1.056860e-05
```

```
## alcohol               0.316818059  0.0184124856  17.206697 2.691739e-64
## volatile.acidity     -1.946515227  0.1145629448 -16.990793 8.707039e-63
## residual.sugar        0.038929856  0.0048624823   8.006169 1.496850e-15
## free.sulfur.dioxide   0.004059882  0.0006893126   5.889753 4.151771e-09
## density             -52.731321694 12.4312080416  -4.241850 2.261757e-05
##
##
##                        Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)          0.89366447 0.2488055543   3.591819  3.319147e-04
## alcohol              0.38302718 0.0094529024  40.519532 5.805015e-306
## volatile.acidity    -1.92843882 0.1144627017 -16.847749  8.531664e-62
## residual.sugar       0.02340842 0.0024405147   9.591591  1.398570e-21
## free.sulfur.dioxide  0.00384641 0.0006897841   5.576252  2.602556e-08
## pH                   0.37902090 0.0692223352   5.475413  4.604406e-08
##
##
##                         Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)          1.847388828 0.1247303072  14.811066  1.730408e-48
## alcohol              0.386519124 0.0094727507  40.803261 1.274626e-309
## volatile.acidity    -1.925882797 0.1145569043 -16.811582  1.515218e-61
## residual.sugar       0.021779144 0.0024070025   9.048243  2.124276e-19
## free.sulfur.dioxide  0.003846276 0.0006904696   5.570522  2.689007e-08
## sulphates            0.468842808 0.0940176682   4.986752  6.373251e-07
##
##  4  iteration done
##
## Current AIC:  9020.424
##
## [1] 9003.921 9020.903 8998.573 9014.260 9004.443 8992.504 8997.592
##
## P-values for each of next variables
## [1] 0.00001720897861 0.21775781028881 0.00000106006138 0.00430225810900
## [5] 0.00002261756801 0.00000004604406 0.00000063732512
## [1] "============================================================"
##
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide + pH +
##
##
##                         Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)          1.358952862 0.3217528101   4.223593  2.452303e-05
## alcohol              0.381187234 0.0094828916  40.197363 8.183524e-302
## volatile.acidity    -1.943911062 0.1146102015 -16.961065  1.402411e-62
## residual.sugar       0.023291264 0.0024399106   9.545950  2.155617e-21
## free.sulfur.dioxide  0.003746392 0.0006908555   5.422830  6.176113e-08
## pH                   0.308794699 0.0757408517   4.076990  4.641670e-05
## fixed.acidity       -0.031190990 0.0136856185  -2.279107  2.270771e-02
##
##
##                         Estimate   Std. Error     t value      Pr(>|t|)
## (Intercept)          0.922775577 0.2577359509   3.5803138  3.468344e-04
## alcohol              0.383007541 0.0094538711  40.5133025 7.138728e-306
## volatile.acidity    -1.936074453 0.1158211253 -16.7160736  6.876099e-61
## residual.sugar       0.023448122 0.0024424557   9.6002242  1.288553e-21
## free.sulfur.dioxide  0.003869021 0.0006918172   5.5925486  2.371198e-08
## pH                   0.374544922 0.0699948454   5.3510358  9.183356e-08
```

```
## citric.acid         -0.041278873 0.0952512243  -0.4333684  6.647681e-01
##
##
##                         Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)          1.314150073 0.2621579277   5.012818  5.571619e-07
## alcohol              0.359571003 0.0105420089  34.108395 8.409758e-227
## volatile.acidity    -1.889046332 0.1144342017 -16.507708  1.811932e-59
## residual.sugar       0.023079348 0.0024349535   9.478353  4.075963e-21
## free.sulfur.dioxide  0.003859031 0.0006879628   5.609361  2.153452e-08
## pH                   0.383346954 0.0690445675   5.552167  2.985161e-08
## chlorides           -4.523775424 0.9097030771  -4.972804  6.846833e-07
##
##
##                          Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)           1.029496511 0.2518112178   4.088366  4.420656e-05
## alcohol               0.371151932 0.0100863224  36.797548 3.979523e-259
## volatile.acidity     -1.836275121 0.1175991962 -15.614691  1.476534e-53
## residual.sugar        0.024212822 0.0024495398   9.884641  8.310249e-23
## free.sulfur.dioxide   0.005457031 0.0008403180   6.494007  9.271191e-11
## pH                    0.398899888 0.0693977723   5.748022  9.630879e-09
## total.sulfur.dioxide -0.001166330 0.0003483585  -3.348072  8.204991e-04
##
##
##                          Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)          81.830637762 12.9785722781   6.305057 3.159729e-10
## alcohol               0.279887309  0.0190273604  14.709729 7.254141e-48
## volatile.acidity     -1.920036312  0.1139877770 -16.844230 9.030878e-62
## residual.sugar        0.051808367  0.0051611773  10.038091 1.835641e-23
## free.sulfur.dioxide   0.003713448  0.0006872048   5.403700 6.868016e-08
## pH                    0.516551285  0.0723710520   7.137540 1.103829e-12
## density             -80.956399620 12.9793218033  -6.237337 4.863112e-10
##
##
##                         Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)          0.822089449 0.248935002   3.302426  9.660684e-04
## alcohol              0.385417314 0.009452741  40.773076 3.198775e-309
## volatile.acidity    -1.912448907 0.114315483 -16.729570  5.556138e-61
## residual.sugar       0.023774590 0.002437585   9.753338  2.973466e-22
## free.sulfur.dioxide  0.003670158 0.000689799   5.320620  1.084799e-07
## pH                   0.332754855 0.069973656   4.755430  2.042663e-06
## sulphates            0.397473654 0.094984036   4.184636  2.911168e-05
##
##  5  iteration done
##
## Current AIC:  8992.504
##
## [1] 8989.304 8994.316 8969.805 8983.291 8955.707 8976.999
##
## P-values for each of next variables
## [1] 0.0227077114663887 0.6647680804733489 0.0000006846833440 0.0008204990645198
## [5] 0.0000000004863112 0.0000291116778756
## [1] "============================================================"
##
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide + pH +
density +
```

```
##
##
##                          Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)          148.093538027 20.5265184103   7.214742 6.318350e-13
## alcohol                0.199079575  0.0271612803   7.329536 2.727508e-13
## volatile.acidity      -1.867945074  0.1144663346 -16.318729 3.416403e-58
## residual.sugar         0.075876211  0.0077450163   9.796779 1.953937e-22
## free.sulfur.dioxide    0.003893723  0.0006873181   5.665096 1.561805e-08
## pH                     0.836022609  0.1054096634   7.931176 2.722856e-15
## density             -148.590937542 20.7836757977  -7.149406 1.013520e-12
## fixed.acidity          0.090860894  0.0218324604   4.161734 3.217901e-05
##
##
##                          Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)          86.489302507 13.461978109   6.424710 1.459339e-10
## alcohol               0.273896430  0.019574583  13.992453 1.427307e-43
## volatile.acidity     -1.895822641  0.115486740 -16.415934 7.574162e-59
## residual.sugar        0.053351433  0.005295162  10.075505 1.266205e-23
## free.sulfur.dioxide   0.003635402  0.000689762   5.270517 1.424604e-07
## pH                    0.538526032  0.074308295   7.247186 4.989311e-13
## density             -85.706639932 13.481571887  -6.357318 2.258646e-10
## citric.acid           0.128235716  0.098510732   1.301744 1.930712e-01
##
##
##                          Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)          74.068149837 13.093948597   5.656670 1.639834e-08
## alcohol               0.270766415  0.019124292  14.158245 1.511097e-44
## volatile.acidity     -1.888201655  0.114053083 -16.555464 8.595401e-60
## residual.sugar        0.048689231  0.005208206   9.348560 1.368691e-20
## free.sulfur.dioxide   0.003737243  0.000686021   5.447710 5.376749e-08
## pH                    0.506356891  0.072286750   7.004837 2.842438e-12
## density             -72.843164898 13.107385780  -5.557414 2.897465e-08
## chlorides            -3.752548467  0.917231400  -4.091169 4.367782e-05
##
##
##                           Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)           75.5754966546 13.6675453473   5.529559 3.393627e-08
## alcohol                0.2825023100  0.0191092923  14.783505 2.560028e-48
## volatile.acidity      -1.8785088577  0.1174775076 -15.990370 5.226930e-56
## residual.sugar         0.0499598629  0.0053139437   9.401655 8.354978e-21
## free.sulfur.dioxide    0.0044610069  0.0008572924   5.203600 2.042355e-07
## pH                     0.5149153374  0.0723706178   7.114978 1.297981e-12
## density              -74.6375719697 13.6820266098  -5.455155 5.157687e-08
## total.sulfur.dioxide  -0.0005338286  0.0003660875  -1.458199 1.448560e-01
##
##
##                          Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)          100.830570297 13.3236994927   7.567761 4.584983e-14
## alcohol                0.258999685  0.0192825634  13.431808 2.371489e-40
## volatile.acidity      -1.894930367  0.1136369717 -16.675298 1.309187e-60
## residual.sugar         0.059041080  0.0052854891  11.170410 1.358778e-28
## free.sulfur.dioxide    0.003427191  0.0006863237   4.993550 6.154175e-07
## pH                     0.482107900  0.0723332693   6.665092 2.966599e-11
## density             -100.064420652 13.3288536177  -7.507354 7.243001e-14
## sulphates              0.574779379  0.0973099097   5.906689 3.749962e-09
```

```
##
##  6  iteration done
##
## Current AIC:  8955.707
##
## [1] 8940.390 8956.010 8940.971 8955.578 8922.892
##
## P-values for each of next variables
## [1] 0.000032179006410 0.193071247830624 0.000043677823329 0.144855983104451
## [5] 0.000000003749962
## [1] "============================================================"
##
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide + pH +
density + sulphates +
##
##
##                          Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)          197.856655207 21.5912816086   9.163729 7.475892e-20
## alcohol                0.141194150  0.0282239297   5.002640 5.872465e-07
## volatile.acidity      -1.816347859  0.1140748882 -15.922416 1.464862e-55
## residual.sugar         0.094357890  0.0081333020  11.601425 1.112773e-30
## free.sulfur.dioxide    0.003614482  0.0006847052   5.278888 1.361441e-07
## pH                     0.921297387  0.1055244992   8.730649 3.523426e-18
## density             -199.034226703 21.8643899611  -9.103123 1.295431e-19
## sulphates              0.705193831  0.0996325944   7.077943 1.691636e-12
## fixed.acidity          0.127132630  0.0223098687   5.698493 1.286526e-08
##
##
##                          Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)          105.116761513 13.7795385997   7.628467 2.885889e-14
## alcohol                0.253476859  0.0198070969  12.797275 7.498672e-37
## volatile.acidity      -1.872425514  0.1151220583 -16.264698 7.872272e-58
## residual.sugar         0.060457876  0.0054115965  11.171911 1.336871e-28
## free.sulfur.dioxide    0.003355287  0.0006888189   4.871072 1.148387e-06
## pH                     0.502708331  0.0742790097   6.767838 1.476578e-11
## density             -104.435684204 13.8025038524  -7.566430 4.631756e-14
## sulphates              0.573006712  0.0973155001   5.888134 4.192529e-09
## citric.acid            0.119595872  0.0981521115   1.218475 2.231080e-01
##
##
##                         Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)          93.075901799 13.4317180668   6.929560 4.824933e-12
## alcohol               0.249772672  0.0193772304  12.890009 2.361449e-37
## volatile.acidity     -1.862852865  0.1136987752 -16.384107 1.242473e-58
## residual.sugar        0.055924013  0.0053297303  10.492841 1.840235e-25
## free.sulfur.dioxide   0.003450354  0.0006851154   5.036165 4.936957e-07
## pH                    0.471765285  0.0722469320   6.529901 7.316945e-11
## density             -91.957356281 13.4490042610  -6.837484 9.149198e-12
## sulphates             0.576313100  0.0971360489   5.933051 3.198641e-09
## chlorides            -3.773277227  0.9137428849  -4.129474 3.702447e-05
##
##
##                           Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)           93.4692723728 13.9372151813   6.706453 2.242930e-11
## alcohol                0.2618580347  0.0193433994  13.537333 6.000429e-41
```

```
## volatile.acidity      -1.8434993928  0.1171650684 -15.734207 2.486728e-54
## residual.sugar         0.0568930699  0.0054179402  10.500867 1.693749e-25
## free.sulfur.dioxide    0.0043405348  0.0008541872   5.081480 3.898150e-07
## pH                     0.4795155950  0.0723296730   6.629583 3.766839e-11
## density              -92.6267291862 13.9547142430  -6.637666 3.567937e-11
## sulphates              0.5845074998  0.0974364748   5.998857 2.144355e-09
## total.sulfur.dioxide  -0.0006556745  0.0003652259  -1.795257 7.268039e-02
##
##  7  iteration done
##
## Current AIC:  8922.892
##
## [1] 8892.471 8923.404 8907.837 8921.663
##
## P-values for each of next variables
## [1] 0.00000001286526 0.22310803958521 0.00003702446865 0.07268038605135
## [1] "============================================================"
##
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide + pH +
density + sulphates + fixed.acidity +
##
##
##                          Estimate    Std. Error     t value     Pr(>|t|)
## (Intercept)          199.442918587 21.6946202114   9.1931971 5.715813e-20
## alcohol                0.139071000  0.0283654706   4.9028272 9.782399e-07
## volatile.acidity      -1.803287550  0.1153885008 -15.6279659 1.214570e-53
## residual.sugar         0.094846592  0.0081594908  11.6240823 8.605486e-31
## free.sulfur.dioxide    0.003567939  0.0006875163   5.1896068 2.200997e-07
## pH                     0.929219386  0.1060515786   8.7619571 2.682810e-18
## density             -200.651777153 21.9704846827  -9.1327879 9.903216e-20
## sulphates              0.702664944  0.0996938757   7.0482258 2.090304e-12
## fixed.acidity          0.125736496  0.0223876873   5.6163236 2.069160e-08
## citric.acid            0.073991450  0.0981542646   0.7538282 4.509922e-01
##
##
##                          Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)          182.306130477 22.1029986397   8.248027 2.097525e-16
## alcohol                0.145369017  0.0282241701   5.150515 2.709724e-07
## volatile.acidity      -1.798683790  0.1140874024 -15.765840 1.549327e-54
## residual.sugar         0.088465355  0.0083284284  10.622095 4.804610e-26
## free.sulfur.dioxide    0.003614544  0.0006839878   5.284516 1.320512e-07
## pH                     0.870438790  0.1065912949   8.166134 4.105862e-16
## density             -183.016739211 22.4010783635  -8.169997 3.978384e-16
## sulphates              0.693724252  0.0995919585   6.965665 3.746133e-12
## fixed.acidity          0.114772255  0.0226148389   5.075086 4.030790e-07
## chlorides             -2.976533003  0.9246428689  -3.219116 1.295077e-03
##
##
##                           Estimate    Std. Error    t value     Pr(>|t|)
## (Intercept)           191.2206288324 22.5545461936   8.478141 3.071730e-17
## alcohol                 0.1457828681  0.0285817890   5.100551 3.527191e-07
## volatile.acidity       -1.7889781990  0.1172024116 -15.264005 2.579614e-51
## residual.sugar          0.0922456297  0.0083939757  10.989504 9.711712e-28
## free.sulfur.dioxide     0.0041308665  0.0008522517   4.847003 1.295996e-06
## pH                      0.9087908988  0.1062373934   8.554341 1.608200e-17
```

```
## density              -192.3058112880 22.8422412120  -8.418868 5.062424e-17
## sulphates               0.7074694417  0.0996572921   7.099023 1.455185e-12
## fixed.acidity           0.1239404862  0.0225292452   5.501316 3.980621e-08
## total.sulfur.dioxide   -0.0003740801  0.0003676149  -1.017587 3.089295e-01
##
##  8  iteration done
##
## Current AIC:  8892.471
##
## [1] 8893.902 8884.097 8893.433
##
## P-values for each of next variables
## [1] 0.450992200 0.001295077 0.308929475
## [1] "============================================================"
##
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide + pH +
density + sulphates + fixed.acidity + chlorides +
##
##
##                           Estimate    Std. Error     t value     Pr(>|t|)
## (Intercept)          183.737942060 22.2244059621   8.2673950 1.787906e-16
## alcohol                0.143586408  0.0283716576   5.0609101 4.340549e-07
## volatile.acidity      -1.788066865  0.1153686565 -15.4987231 8.254152e-53
## residual.sugar         0.088915333  0.0083604520  10.6352303 4.188674e-26
## free.sulfur.dioxide    0.003576202  0.0006868149   5.2069378 2.006251e-07
## pH                     0.877373886  0.1071818358   8.1858449 3.495190e-16
## density             -184.478080237 22.5258357615  -8.1896220 3.388808e-16
## sulphates              0.691733215  0.0996503926   6.9416005 4.435280e-12
## fixed.acidity          0.113721534  0.0226795787   5.0142701 5.530142e-07
## chlorides             -2.952595216  0.9255093822  -3.1902380 1.431411e-03
## citric.acid            0.060953589  0.0981385616   0.6210972 5.345674e-01
##
##
##                            Estimate    Std. Error     t value     Pr(>|t|)
## (Intercept)           176.8088006498 22.9840523934   7.6926731 1.761998e-14
## alcohol                 0.1492518651  0.0285736065   5.2234171 1.836578e-07
## volatile.acidity       -1.7754346441  0.1171608523 -15.1538215 1.278453e-50
## residual.sugar          0.0867272282  0.0085635824  10.1274472 7.546697e-24
## free.sulfur.dioxide     0.0040573165  0.0008516999   4.7637867 1.960290e-06
## pH                      0.8603454164  0.1072200419   8.0241101 1.296576e-15
## density              -177.4459776399 23.2937924915  -7.6177367 3.133071e-14
## sulphates               0.6958176112  0.0996235191   6.9844713 3.281928e-12
## fixed.acidity           0.1121883392  0.0228085237   4.9187024 9.025575e-07
## chlorides              -2.9396450250  0.9256336500  -3.1758191 1.504327e-03
## total.sulfur.dioxide   -0.0003207542  0.0003676247  -0.8725045 3.829802e-01
##
##  9  iteration done
##
## Current AIC:  8884.097
##
## [1] 8885.711 8885.334
##
## P-values for each of next variables
## [1] 0.5345674 0.3829802
## [1] "============================================================"
```

```
##
## Call:
## glm(formula = formula2, data = updated_Dataset)
##
## Coefficients:
##                        Estimate  Std. Error t value Pr(>|t|)
## (Intercept)          182.306130   22.102999   8.248  < 2e-16 ***
## alcohol                0.145369    0.028224   5.151 2.71e-07 ***
## volatile.acidity      -1.798684    0.114087 -15.766  < 2e-16 ***
## residual.sugar         0.088465    0.008328  10.622  < 2e-16 ***
## free.sulfur.dioxide    0.003615    0.000684   5.285 1.32e-07 ***
## pH                     0.870439    0.106591   8.166 4.11e-16 ***
## density             -183.016739   22.401078  -8.170 3.98e-16 ***
## sulphates              0.693724    0.099592   6.966 3.75e-12 ***
## fixed.acidity          0.114772    0.022615   5.075 4.03e-07 ***
## chlorides             -2.976533    0.924643  -3.219   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.4261068)
##
##     Null deviance: 2824.8  on 4469  degrees of freedom
## Residual deviance: 1900.4  on 4460  degrees of freedom
## AIC: 8884.1
##
## Number of Fisher Scoring iterations: 2
```

INTERPRETATION

By inspection, we can conclude that the our current best fitted model consists of the following variables in the specified order:

1. Alcohol (vol.%)
2. Volatile acidity (g(acetic acid)/dm^3)
3. Residual sugar (g/dm^3)
4. Free sulfur dioxide (mg/dm^3)
5. pH level
6. Density (g/cm^3)
7. Sulphates (g(potassium sulphate)/dm^3)
8. Fixed acidity (g(tartaric acid)/dm^3)
9. Chlorides (g(sodium chloride)/dm^3)

For further investigation and verification, we will also see how the other values (AIC, BIC, R-squared, Residual standard error) change as each new variable is fitted, if it is consistent with our findings, we should observe a trend where each of the values become more ideal with every variable included at each step.

```
# fitting each of the models as best_model_x

# Create an empty list to store the models

best_models <- list()
```

```r
# Loop through each element in pairs

for (i in 1:9) {
formula <- as.formula(paste("quality ~", pairs[i]))
best_models[[i]] <- lm(formula, data = updated_Dataset)

assign(paste0("best_model_", i), best_models[[i]])
}

# finding the relevant values for each of the models

model_names <- paste0("best_model_", 1:9)

# Loop through each model

for (i in 1:9) {
model <- get(model_names[i])

assign(paste0("Residual_standard_error_", model_names[i]), summary(model)$sigma)
assign(paste0("R_squared_", model_names[i]), summary(model)$r.squared)
assign(paste0(model_names[i], "_AIC"), AIC(model))
assign(paste0(model_names[i], "_BIC"), BIC(model))}

table_5 <- data.frame(
  Metrics = c("AIC", "BIC", "R-squared", "Residual standard error"))

# Loop through each model to extract metrics and store them in the data frame

for (i in 1:9) {
model <- get(model_names[i])

aic_value <- AIC(model)
bic_value <- BIC(model)
r_squared_value <- summary(model)$r.squared
residual_standard_error <- summary(model)$sigma

table_5[paste0("best_model_", i)] <- c(aic_value, bic_value, r_squared_value,
residual_standard_error)}

# formatting the table

for (i in 1:9) {
  column_name <- paste0("best_model_", i)
  table_5[[column_name]] <- format(table_5[[column_name]], digits = 5)}

# displaying the table

knitr::kable(table_5, caption = "Table 6: Comparison of all Fitted Models") %>%
  kable_styling(bootstrap_options = "striped", full_width = F)
```

Table 6: Comparison of all Fitted Models

Metrics	best_model_1	best_model_2	best_model_3	best_model_4	best_model_5	best_model_6	best_model_7	best_model_8	best_model_9
AIC	9419.08213	9165.24346	9053.38678	9020.42385	8992.50376	8955.70723	8922.89163	8892.47115	8884.09726
BIC	9438.29757	9190.86403	9085.41250	9058.85471	9037.33977	9006.94838	8980.53793	8956.52259	8954.55384
R-squared	0.23898	0.28131	0.29939	0.30485	0.30948	0.31545	0.32076	0.32567	0.32723
Residual standard error	0.69364	0.67415	0.66569	0.66317	0.66103	0.65824	0.65575	0.65345	0.65277

INTERPRETATION

We can observe that for every new variable added, the metric values become more ideal for every model until best_model_9. To further investigate the quality of our model, we check on the Mallows Cp value.

Checking Mallows Cp

```
# checking Mallows Cp for the current best model fitted

ols_best_model_9<-ols_mallows_cp(best_model_9,updated_full_model)
ols_best_model_9

## [1] 9.18884
```

INTERPRETATION

Under the model, we can observe that the Mallows Cp value of 9.18884 which is seemingly lower than th ideal value of $k + 1 = p = 12$ which indicates that the the current model might be underfitting the data.

Reinvestigating Multicollinearity Properties

Despite all the metrics pointing towards the conclusion that we have fitted the best model, it is also essential to consider multicollinearity in addition to traditional model evaluation metrics such as R-squared, AIC, BIC, and residual standard error. Hence we need to evaluate the presence of multicollinear issues for each model starting from the model including two regressor variables and so forth. Again, we will make use of the VIF values to investigate multicollinearity.

```
# seeing vif for each model

cat("================================================================================
==============","\n","\n")
cat("( VIF: 2 Predictor Model )","\n","\n")
car::vif(mod=best_model_2)%>%sort()
cat("\n")

cat("================================================================================
==============","\n","\n")
cat("( VIF: 3 Predictor Model )","\n","\n")
car::vif(mod=best_model_3)%>%sort()
cat("\n")
```

```
cat("================================================================================
===============","\n","\n")
cat("( VIF: 4 Predictor Model )","\n","\n")
car::vif(mod=best_model_4)%>%sort()
cat("\n")


cat("================================================================================
===============","\n","\n")
cat("( VIF: 5 Predictor Model )","\n","\n")
car::vif(mod=best_model_5)%>%sort()
cat("\n")


cat("================================================================================
===============","\n","\n")
cat("( VIF: 6 Predictor Model )","\n","\n")
car::vif(mod=best_model_6)%>%sort()
cat("\n")

cat("================================================================================
===============")
```

```
##
================================================================================
==========
##
## ( VIF: 2 Predictor Model )
##
##          alcohol volatile.acidity
##         1.002888         1.002888
##
##
================================================================================
==========
##
## ( VIF: 3 Predictor Model )
##
## volatile.acidity          alcohol   residual.sugar
##         1.016578         1.329738         1.333360
##
##
================================================================================
==========
##
## ( VIF: 4 Predictor Model )
##
##    volatile.acidity free.sulfur.dioxide              alcohol      residual.sugar
##             1.026343            1.166973             1.343023             1.443502
##
##
================================================================================
==========
```

```
##
## ( VIF: 5 Predictor Model )
##
##    volatile.acidity                  pH free.sulfur.dioxide             alcohol
##            1.027292            1.048168            1.171813            1.343333
##      residual.sugar
##            1.491945
##
##
================================================================================
==========
##
## ( VIF: 6 Predictor Model )
##
##    volatile.acidity                  pH free.sulfur.dioxide             alcohol
##            1.027436            1.155421            1.172941            5.488864
##      residual.sugar             density
##            6.729140           14.526068
##
##
================================================================================
==========
```

INTERPRETATION

We have observed that each model starting from 2 predictor variables to 5 predictor variables have VIF values < 5 which shows no potential multicollinearity problems. However, it is only when the 6th predictor, density, was added where we observe that there exists VIF values that exceed 5, specifically

- Density : 14.526068
- Residual Sugar : 6.729140
- Alcohol : 5.488864

This observation is consistent with when we investigated multicollinearity for the updated full model where density had a high correlation value with both alcohol alongside residual sugar. It is with reasonable assumption that density is the root predictor in which is resulting in the presence of multicollinear issues and thus, provides justification into removing this variable from our current best model.

Fitting "Best" Model without Density Variable

Now, we will observe how the different metrics have changed after removing the density predictor variable.

```
# fitting best model without density

updated_model<-lm(quality~ alcohol + volatile.acidity + residual.sugar +
free.sulfur.dioxide + pH + sulphates + fixed.acidity + chlorides, data=updated_Dataset)
summary(updated_model)

##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + pH + sulphates + fixed.acidity + chlorides,
##     data = updated_Dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.00525 -0.48472 -0.03287  0.43924  2.16545
##
## Coefficients:
##                       Estimate Std. Error t value    Pr(>|t|)
## (Intercept)           1.744579   0.330347   5.281 0.000000135 ***
## alcohol               0.359512   0.010545  34.092     < 2e-16 ***
## volatile.acidity     -1.887133   0.114406 -16.495     < 2e-16 ***
## residual.sugar        0.023340   0.002430   9.603     < 2e-16 ***
## free.sulfur.dioxide   0.003563   0.000689   5.171 0.000000243 ***
## pH                    0.258770   0.076429   3.386    0.000716 ***
## sulphates             0.428830   0.094858   4.521 0.000006321 ***
## fixed.acidity        -0.033216   0.013639  -2.435    0.014917 *
## chlorides            -4.654505   0.908164  -5.125 0.000000310 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6576 on 4461 degrees of freedom
## Multiple R-squared:  0.3172, Adjusted R-squared:  0.3159
## F-statistic:   259 on 8 and 4461 DF,  p-value: < 2.2e-16

# recheck multicollinearity

cat("================================================================================
==============","\n","\n")
cat("( VIF: Updated Best model without density variable )","\n","\n")
car::vif(mod=updated_model)%>%sort()
cat("\n")

cat("================================================================================
==============")

##
================================================================================
==========
##
## ( VIF: Updated Best model without density variable )
##
```

```
##    volatile.acidity           sulphates free.sulfur.dioxide       fixed.acidity
##            1.037130            1.042603            1.181440            1.226738
##                  pH           chlorides      residual.sugar             alcohol
##            1.291279            1.320928            1.495393            1.689441
##
##
================================================================================
==========
```

INTERPRETATION

After removing the density variable, it aligns with our assumption where the model should no longer have any indication of multicollinearity issues as evidenced with all the VIF values being below the threshold of 5 and the maximum value being only 1.689441 in respect to the alcohol variable. However, we should reinvestigate how this removal has affected the other metrics which can help in determining the overall quality of the model.

```
# extracting the relevant metrics

Residual_standard_error_updated_model<-summary(updated_model)$sigma
R_squared_updated_model<-summary(updated_model)$r.squared
updated_model_AIC<-AIC(updated_model)
updated_model_BIC<-BIC(updated_model)
ols_updated_model<-ols_mallows_cp(updated_model,updated_full_model)

# constructing the table

table_6 <- data.frame(
  Metrics = c("AIC", "BIC", "R-squared", "Residual standard error", "Mallows Cp"),
  initial_best_model = c(best_model_9_AIC, best_model_9_BIC, R_squared_best_model_9,
Residual_standard_error_best_model_9,ols_best_model_9),
 updated_best_model = c(updated_model_AIC, updated_model_BIC, R_squared_updated_model,
Residual_standard_error_updated_model,ols_updated_model))

# formatting the table

table_6$initial_best_model <- format(table_6$initial_best_model, digits = 5)
table_6$updated_best_model <- format(table_6$updated_best_model, digits = 5)

# displaying the table

knitr::kable(table_6, caption = "Table 7: Comparison of Best Model vs Model without
Density Variable ") %>%
```

Table 7: Comparison of Best Model vs Model without Density Variable

Metrics	initial_best_model	updated_best_model
AIC	8884.09726	8948.50010
BIC	8954.55384	9012.55154
R-squared	0.32723	0.31717
Residual standard error	0.65277	0.65756
Mallows Cp	9.18884	73.92555

INTERPRETATION

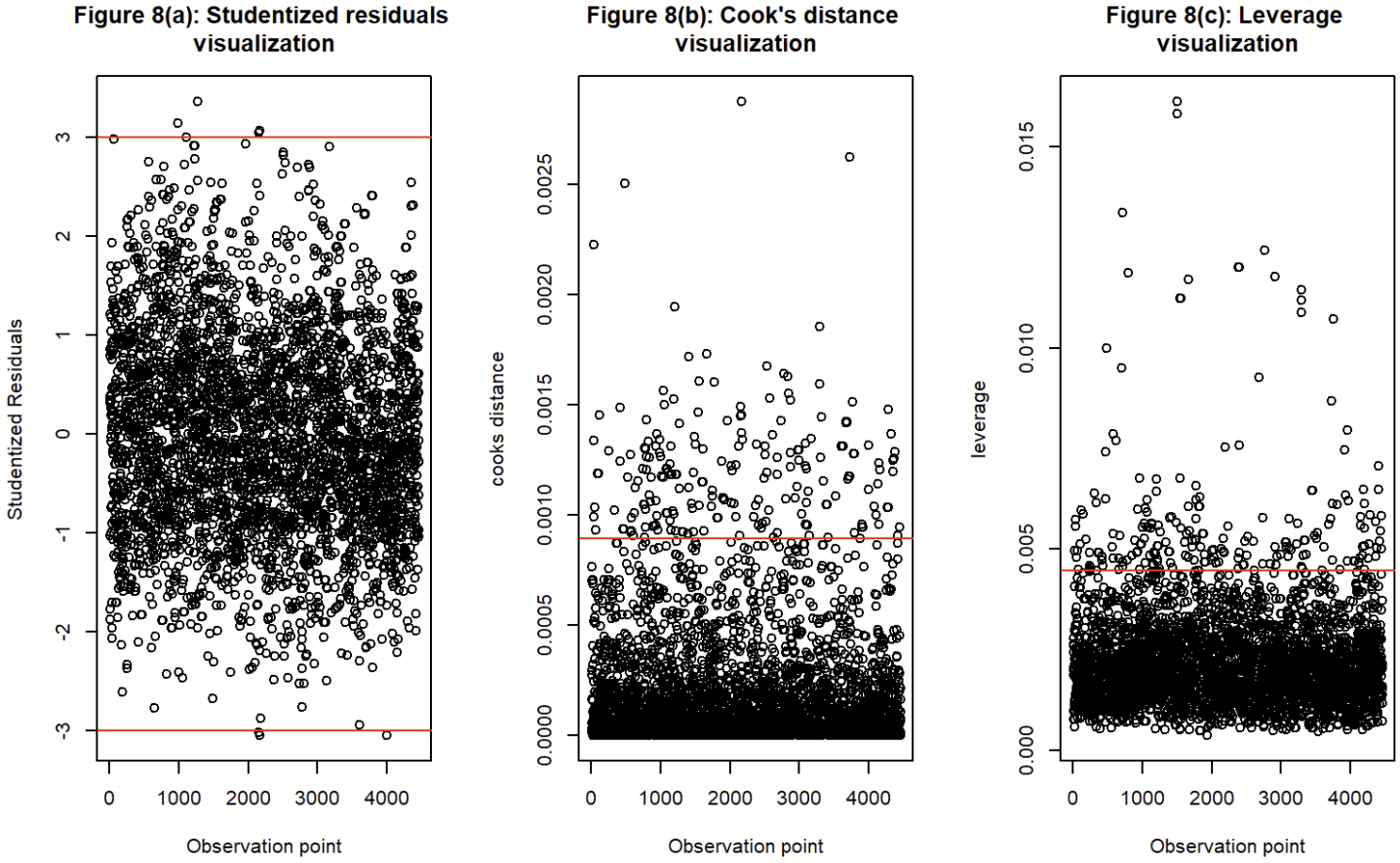
From our analysis, we observe that the removal of the density variable has actually led to these metric values exhibiting less ideal values instead. In particular, we can observe that the Mallows Cp is much higher than the number of predictors in the model which indicates that under this new model, it displays biasness and does not fit the data well. In such cases, the benefits of improved predictive accuracy and better model performance may outweigh the drawbacks of multicollinearity. Hence, while multicollinearity can complicate the interpretation of individual coefficients, the overall enhancement in model quality justifies retaining the collinear variable

Investigating Problematic Data Points

To further improve the model we should reinvestigate problematic points under the best model. It is important to note that even after removing some influential points during the initial fitting of the updated full model, it is possible to encounter new influential points under the subsequent models as we fit different models with varying numbers and combinations of predictors. Each model may highlight different data points as problematic, necessitating a thorough review and adjustment to ensure the final model's robustness and accuracy.

```
# studentized residuals

studentized_residuals_best_model_9<-studres(best_model_9)

# cooks distance

cd_best_model_9<-cooks.distance(best_model_9)

# leverage

hat_values_best_model_9<-hatvalues(best_model_9)

# plot for visualization

par(mfrow=c(1,3))
```

```r
plot(studentized_residuals_best_model_9, main="Figure 8(a): Studentized residuals
\nvisualization", ylab="Studentized Residuals",xlab="Observation point")
abline(h=3,col="red")
abline(h=-3,col="red")

plot(cd_best_model_9, main="Figure 8(b): Cook's distance \nvisualization", ylab="cooks
distance",xlab="Observation point")
abline(h = 4/(nrow(updated_Dataset)), col = "red")

plot(hat_values_best_model_9, main="Figure 8(c): Leverage
\nvisualization",ylab="leverage", xlab="Observation point")
abline(h= 2 * (length(best_model_9$coefficients)) / nrow(updated_Dataset),col ="red")
```




```r
high_studentized_residual_points_2<-which(studentized_residuals_best_model_9>(abs(3)))
high_leverage_points_2<-
which(hat_values_best_model_9>(2*(length(best_model_9$coefficients))/nrow(updated_Dataset
)))
high_cooks_distance_points_2<-which(cd_best_model_9>(4/nrow(updated_Dataset)))

all_potential_issues_2<-
unique(c(high_studentized_residual_points_2,high_leverage_points_2,high_cooks_distance_po
ints_2))

# create finalized dataset

finalized_Dataset<-updated_Dataset[-all_potential_issues_2, ]

# fitting finalized best model
```

```r
finalized_best_model<-lm(quality~alcohol + volatile.acidity + residual.sugar +
free.sulfur.dioxide + pH + density + sulphates + fixed.acidity +
chlorides,data=finalized_Dataset)
finalized_full_model<-lm(quality~.,data=finalized_Dataset)

summary(finalized_best_model)

# comparing initial best model with finalized model

Residual_standard_error_finalized_best_model<-summary(finalized_best_model)$sigma
R_squared_finalized_best_model<-summary(finalized_best_model)$r.squared
finalized_best_model_AIC<-AIC(finalized_best_model)
finalized_best_model_BIC<-BIC(finalized_best_model)
ols_finalized_best_model<-ols_mallows_cp(finalized_best_model,finalized_full_model)

# constructing the table

table_7 <- data.frame(
  Metrics = c("AIC", "BIC", "R-squared", "Residual standard error", "Mallows Cp"),
  initial_best_model = c(best_model_9_AIC, best_model_9_BIC, R_squared_best_model_9,
Residual_standard_error_best_model_9,ols_best_model_9),

 finalized_best_model = c(finalized_best_model_AIC, finalized_best_model_BIC,
R_squared_finalized_best_model,
Residual_standard_error_finalized_best_model,ols_finalized_best_model))

# formatting the table

table_7$initial_best_model <- format(table_7$initial_best_model, digits = 5)
table_7$finalized_best_model <- format(table_7$finalized_best_model, digits = 5)

# displaying the table

knitr::kable(table_7, caption = "Table 8: Comparison of Initial vs Finalized Best Model
") %>%
  kable_styling(bootstrap_options = "striped", full_width = F)
```

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + pH + density + sulphates + fixed.acidity +
##     chlorides, data = finalized_Dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82796 -0.44728 -0.03295  0.43411  1.81323
##
## Coefficients:
##                         Estimate   Std. Error t value           Pr(>|t|)
## (Intercept)          205.6980507   22.6036111   9.100            < 2e-16 ***
## alcohol                0.1103980    0.0286650   3.851           0.000119 ***
## volatile.acidity      -1.6708200    0.1179816 -14.162            < 2e-16 ***
## residual.sugar         0.0923997    0.0084717  10.907            < 2e-16 ***
## free.sulfur.dioxide    0.0043974    0.0006704   6.560 0.0000000000607736 ***
```

```
## pH                    1.1097245    0.1071075  10.361           < 2e-16 ***
## density            -207.2098617   22.9063769  -9.046           < 2e-16 ***
## sulphates             0.7512165    0.0974882   7.706 0.0000000000000163 ***
## fixed.acidity         0.1446669    0.0225925   6.403 0.0000000001695433 ***
## chlorides            -4.7363543    1.1195804  -4.230 0.0000238403825246 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5863 on 4015 degrees of freedom
## Multiple R-squared:  0.3746, Adjusted R-squared:  0.3732
## F-statistic: 267.2 on 9 and 4015 DF,  p-value: < 2.2e-16
```

Table 8: Comparison of Initial vs Finalized Best Model

Metrics	initial_best_model	finalized_best_model
AIC	8884.09726	7135.99331
BIC	8954.55384	7205.29640
R-squared	0.32723	0.37456
Residual standard error	0.65277	0.58627
Mallows Cp	9.18884	11.72388

INTERPRETATION

Initially, our best model had a Mallows' Cp value of 9.18. After removing the influential points and recalculating, the Cp value changed to 11.72 which is much closer to the ideal value of 12 which indicates that model adequately explains the variance in the response variable. Additionally, when examining other metrics, the updated model shows a clear improvement: the AIC and BIC values are lower, indicating a better balance between model fit and complexity; and additionally the R-squared value is higher alongside a smaller residual standard error value. These improvements in key performance indicators justify that the updated model, is overall a better fitted model.

Verification of Results

Now that we have identified our best model, we can make further investigation and verification using inbuilt functions designed for model selection. Specifically, we will utilize forward stepwise selection, backward stepwise selection, all-subsets regression (also known as all possible subsets or best subsets regression) and the best subsets regression. By comparing our manually selected best model with those identified by these automated procedures, we can validate our findings and ensure that we have indeed chosen the most accurate and efficient model.

```
# fitting null model necessary for the following model selection techniques

nullmodel<-lm(quality~1,data=updated_Dataset)
summary(nullmodel)
```

```
##
## Call:
## lm(formula = quality ~ 1, data = updated_Dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.892 -0.892  0.108  0.108  2.108
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.89195    0.01189   495.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.795 on 4469 degrees of freedom
```

All Subsets Regression

```
# Using all subsets selection

best_step_model<-step(nullmodel,scope=list(lower=nullmodel,
upper=updated_full_model,direction = "both"))
summary(best_step_model)
```

```
## Start:  AIC=-2049.49
## quality ~ 1
##
##                       Df Sum of Sq    RSS     AIC
## + alcohol              1    675.07 2149.7 -3268.2
## + density              1    363.49 2461.3 -2663.2
## + chlorides            1    264.87 2559.9 -2487.6
## + total.sulfur.dioxide 1    134.29 2690.5 -2265.2
## + volatile.acidity     1     90.73 2734.1 -2193.4
## + residual.sugar       1     58.39 2766.4 -2140.9
## + pH                   1     34.45 2790.4 -2102.3
## + fixed.acidity        1     28.87 2795.9 -2093.4
## + sulphates            1      6.10 2818.7 -2057.2
## <none>                             2824.8 -2049.5
## + free.sulfur.dioxide  1      0.38 2824.4 -2048.1
## + citric.acid          1      0.06 2824.8 -2047.6
##
## Step:  AIC=-3268.23
```

```
## quality ~ alcohol
##
##                         Df Sum of Sq    RSS     AIC
## + volatile.acidity       1    119.58 2030.2 -3522.1
## + free.sulfur.dioxide    1     41.21 2108.5 -3352.7
## + residual.sugar         1     33.97 2115.8 -3337.4
## + chlorides              1     17.12 2132.6 -3302.0
## + sulphates              1     14.25 2135.5 -3296.0
## + density                1     14.02 2135.7 -3295.5
## + pH                     1     10.30 2139.4 -3287.7
## + fixed.acidity          1      6.77 2143.0 -3280.3
## + citric.acid            1      2.64 2147.1 -3271.7
## <none>                               2149.7 -3268.2
## + total.sulfur.dioxide   1      0.30 2149.4 -3266.9
## - alcohol                1    675.07 2824.8 -2049.5
##
## Step:  AIC=-3522.07
## quality ~ alcohol + volatile.acidity
##
##                         Df Sum of Sq    RSS     AIC
## + residual.sugar         1     51.06 1979.1 -3633.9
## + free.sulfur.dioxide    1     33.04 1997.1 -3593.4
## + density                1     24.02 2006.1 -3573.3
## + chlorides              1     11.74 2018.4 -3546.0
## + sulphates              1     10.41 2019.8 -3543.0
## + fixed.acidity          1      8.60 2021.5 -3539.0
## + pH                     1      6.75 2023.4 -3535.0
## + total.sulfur.dioxide   1      5.46 2024.7 -3532.1
## <none>                               2030.2 -3522.1
## + citric.acid            1      0.00 2030.2 -3520.1
## - volatile.acidity       1    119.58 2149.7 -3268.2
## - alcohol                1    703.92 2734.1 -2193.4
##
## Step:  AIC=-3633.92
## quality ~ alcohol + volatile.acidity + residual.sugar
##
##                         Df Sum of Sq    RSS     AIC
## + free.sulfur.dioxide    1     15.42 1963.7 -3666.9
## + pH                     1     14.93 1964.2 -3665.8
## + sulphates              1     12.72 1966.4 -3660.8
## + chlorides              1     10.34 1968.8 -3655.3
## + fixed.acidity          1     10.04 1969.1 -3654.7
## + density                1      8.10 1971.0 -3650.3
## <none>                               1979.1 -3633.9
## + total.sulfur.dioxide   1      0.50 1978.6 -3633.1
## + citric.acid            1      0.31 1978.8 -3632.6
## - residual.sugar         1     51.06 2030.2 -3522.1
## - volatile.acidity       1    136.67 2115.8 -3337.4
## - alcohol                1    706.70 2685.8 -2271.1
##
## Step:  AIC=-3666.89
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide
##
##                         Df Sum of Sq    RSS     AIC
## + pH                     1     13.10 1950.6 -3694.8
```

```
## + sulphates             1     10.88 1952.8 -3689.7
## + chlorides             1     10.45 1953.2 -3688.7
## + fixed.acidity         1      8.11 1955.6 -3683.4
## + density               1      7.88 1955.8 -3682.9
## + total.sulfur.dioxide  1      3.58 1960.1 -3673.1
## <none>                              1963.7 -3666.9
## + citric.acid           1      0.67 1963.0 -3666.4
## - free.sulfur.dioxide   1     15.42 1979.1 -3633.9
## - residual.sugar        1     33.44 1997.1 -3593.4
## - volatile.acidity      1    126.61 2090.3 -3389.6
## - alcohol               1    720.52 2684.2 -2271.7
##
## Step:  AIC=-3694.81
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     pH
##
##                         Df Sum of Sq    RSS     AIC
## + density               1     16.86 1933.7 -3731.6
## + chlorides             1     10.75 1939.8 -3717.5
## + sulphates             1      7.62 1943.0 -3710.3
## + total.sulfur.dioxide  1      4.89 1945.7 -3704.0
## + fixed.acidity         1      2.27 1948.3 -3698.0
## <none>                              1950.6 -3694.8
## + citric.acid           1      0.08 1950.5 -3693.0
## - pH                    1     13.10 1963.7 -3666.9
## - free.sulfur.dioxide   1     13.59 1964.2 -3665.8
## - residual.sugar        1     40.20 1990.8 -3605.6
## - volatile.acidity      1    124.03 2074.6 -3421.2
## - alcohol               1    717.41 2668.0 -2296.8
##
## Step:  AIC=-3731.6
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     pH + density
##
##                         Df Sum of Sq    RSS     AIC
## + sulphates             1    15.003 1918.7 -3764.4
## + fixed.acidity         1     7.477 1926.2 -3746.9
## + chlorides             1     7.227 1926.5 -3746.3
## + total.sulfur.dioxide  1     0.921 1932.8 -3731.7
## <none>                              1933.7 -3731.6
## + citric.acid           1     0.734 1933.0 -3731.3
## - free.sulfur.dioxide   1    12.652 1946.4 -3704.5
## - density               1    16.856 1950.6 -3694.8
## - pH                    1    22.073 1955.8 -3682.9
## - residual.sugar        1    43.659 1977.4 -3633.8
## - alcohol               1    93.751 2027.5 -3522.0
## - volatile.acidity      1   122.933 2056.7 -3458.1
##
## Step:  AIC=-3764.42
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     pH + density + sulphates
##
##                         Df Sum of Sq    RSS     AIC
## + fixed.acidity         1    13.866 1904.8 -3794.8
## + chlorides             1     7.307 1911.4 -3779.5
```

```
## + total.sulfur.dioxide  1     1.385 1917.3 -3765.6
## <none>                              1918.7 -3764.4
## + citric.acid           1     0.638 1918.1 -3763.9
## - free.sulfur.dioxide   1    10.723 1929.4 -3741.5
## - sulphates             1    15.003 1933.7 -3731.6
## - pH                    1    19.103 1937.8 -3722.1
## - density               1    24.236 1943.0 -3710.3
## - residual.sugar        1    53.656 1972.4 -3643.1
## - alcohol               1    77.580 1996.3 -3589.2
## - volatile.acidity      1   119.572 2038.3 -3496.2
##
## Step:  AIC=-3794.84
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     pH + density + sulphates + fixed.acidity
##
##                         Df Sum of Sq    RSS     AIC
## + chlorides              1     4.416 1900.4 -3803.2
## <none>                              1904.8 -3794.8
## + total.sulfur.dioxide  1     0.442 1904.4 -3793.9
## + citric.acid           1     0.243 1904.6 -3793.4
## - alcohol               1    10.686 1915.5 -3771.8
## - free.sulfur.dioxide   1    11.899 1916.8 -3769.0
## - fixed.acidity         1    13.866 1918.7 -3764.4
## - sulphates             1    21.392 1926.2 -3746.9
## - pH                    1    32.548 1937.4 -3721.1
## - density               1    35.384 1940.2 -3714.6
## - residual.sugar        1    57.471 1962.3 -3664.0
## - volatile.acidity      1   108.255 2013.1 -3549.8
##
## Step:  AIC=-3803.21
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     pH + density + sulphates + fixed.acidity + chlorides
##
##                         Df Sum of Sq    RSS     AIC
## <none>                              1900.4 -3803.2
## + total.sulfur.dioxide  1     0.324 1900.1 -3802.0
## + citric.acid           1     0.164 1900.3 -3801.6
## - chlorides             1     4.416 1904.8 -3794.8
## - fixed.acidity         1    10.975 1911.4 -3779.5
## - alcohol               1    11.304 1911.7 -3778.7
## - free.sulfur.dioxide   1    11.900 1912.3 -3777.3
## - sulphates             1    20.675 1921.1 -3756.8
## - pH                    1    28.415 1928.8 -3738.9
## - density               1    28.442 1928.9 -3738.8
## - residual.sugar        1    48.077 1948.5 -3693.5
## - volatile.acidity      1   105.914 2006.3 -3562.8
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + pH + density + sulphates + fixed.acidity +
##     chlorides, data = updated_Dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99007 -0.48040 -0.04752  0.44836  2.19096
```

```
##
## Coefficients:
##                         Estimate  Std. Error t value Pr(>|t|)
## (Intercept)          182.306130   22.102999   8.248  < 2e-16 ***
## alcohol                0.145369    0.028224   5.151 2.71e-07 ***
## volatile.acidity      -1.798684    0.114087 -15.766  < 2e-16 ***
## residual.sugar         0.088465    0.008328  10.622  < 2e-16 ***
## free.sulfur.dioxide    0.003615    0.000684   5.285 1.32e-07 ***
## pH                     0.870439    0.106591   8.166 4.11e-16 ***
## density             -183.016739   22.401078  -8.170 3.98e-16 ***
## sulphates              0.693724    0.099592   6.966 3.75e-12 ***
## fixed.acidity          0.114772    0.022615   5.075 4.03e-07 ***
## chlorides             -2.976533    0.924643  -3.219   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6528 on 4460 degrees of freedom
## Multiple R-squared:  0.3272, Adjusted R-squared:  0.3259
## F-statistic:   241 on 9 and 4460 DF,  p-value: < 2.2e-16
```

Forward Stepwise Selection

```
# Using forward stepwise selction

forward_model<-
step(nullmodel,scope=list(lower=nullmodel,upper=updated_full_model,direction="forward"))
summary(forward_model)
```

```
## Start:  AIC=-2049.49
## quality ~ 1
##
##                        Df Sum of Sq    RSS     AIC
## + alcohol               1    675.07 2149.7 -3268.2
## + density               1    363.49 2461.3 -2663.2
## + chlorides             1    264.87 2559.9 -2487.6
## + total.sulfur.dioxide  1    134.29 2690.5 -2265.2
## + volatile.acidity      1     90.73 2734.1 -2193.4
## + residual.sugar        1     58.39 2766.4 -2140.9
## + pH                    1     34.45 2790.4 -2102.3
## + fixed.acidity         1     28.87 2795.9 -2093.4
## + sulphates             1      6.10 2818.7 -2057.2
## <none>                               2824.8 -2049.5
## + free.sulfur.dioxide   1      0.38 2824.4 -2048.1
## + citric.acid           1      0.06 2824.8 -2047.6
##
## Step:  AIC=-3268.23
## quality ~ alcohol
##
##                        Df Sum of Sq    RSS     AIC
## + volatile.acidity      1    119.58 2030.2 -3522.1
## + free.sulfur.dioxide   1     41.21 2108.5 -3352.7
## + residual.sugar        1     33.97 2115.8 -3337.4
## + chlorides             1     17.12 2132.6 -3302.0
## + sulphates             1     14.25 2135.5 -3296.0
## + density               1     14.02 2135.7 -3295.5
## + pH                    1     10.30 2139.4 -3287.7
```

```
## + fixed.acidity         1      6.77 2143.0 -3280.3
## + citric.acid           1      2.64 2147.1 -3271.7
## <none>                              2149.7 -3268.2
## + total.sulfur.dioxide  1      0.30 2149.4 -3266.9
## - alcohol               1    675.07 2824.8 -2049.5
##
## Step:  AIC=-3522.07
## quality ~ alcohol + volatile.acidity
##
##                         Df Sum of Sq    RSS     AIC
## + residual.sugar         1     51.06 1979.1 -3633.9
## + free.sulfur.dioxide    1     33.04 1997.1 -3593.4
## + density                1     24.02 2006.1 -3573.3
## + chlorides              1     11.74 2018.4 -3546.0
## + sulphates              1     10.41 2019.8 -3543.0
## + fixed.acidity          1      8.60 2021.5 -3539.0
## + pH                     1      6.75 2023.4 -3535.0
## + total.sulfur.dioxide   1      5.46 2024.7 -3532.1
## <none>                              2030.2 -3522.1
## + citric.acid            1      0.00 2030.2 -3520.1
## - volatile.acidity       1    119.58 2149.7 -3268.2
## - alcohol                1    703.92 2734.1 -2193.4
##
## Step:  AIC=-3633.92
## quality ~ alcohol + volatile.acidity + residual.sugar
##
##                         Df Sum of Sq    RSS     AIC
## + free.sulfur.dioxide    1     15.42 1963.7 -3666.9
## + pH                     1     14.93 1964.2 -3665.8
## + sulphates              1     12.72 1966.4 -3660.8
## + chlorides              1     10.34 1968.8 -3655.3
## + fixed.acidity          1     10.04 1969.1 -3654.7
## + density                1      8.10 1971.0 -3650.3
## <none>                              1979.1 -3633.9
## + total.sulfur.dioxide   1      0.50 1978.6 -3633.1
## + citric.acid            1      0.31 1978.8 -3632.6
## - residual.sugar         1     51.06 2030.2 -3522.1
## - volatile.acidity       1    136.67 2115.8 -3337.4
## - alcohol                1    706.70 2685.8 -2271.1
##
## Step:  AIC=-3666.89
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide
##
##                         Df Sum of Sq    RSS     AIC
## + pH                     1     13.10 1950.6 -3694.8
## + sulphates              1     10.88 1952.8 -3689.7
## + chlorides              1     10.45 1953.2 -3688.7
## + fixed.acidity          1      8.11 1955.6 -3683.4
## + density                1      7.88 1955.8 -3682.9
## + total.sulfur.dioxide   1      3.58 1960.1 -3673.1
## <none>                              1963.7 -3666.9
## + citric.acid            1      0.67 1963.0 -3666.4
## - free.sulfur.dioxide    1     15.42 1979.1 -3633.9
## - residual.sugar         1     33.44 1997.1 -3593.4
## - volatile.acidity       1    126.61 2090.3 -3389.6
```

```
## - alcohol               1    720.52 2684.2 -2271.7
##
## Step:  AIC=-3694.81
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     pH
##
##                        Df Sum of Sq    RSS     AIC
## + density               1     16.86 1933.7 -3731.6
## + chlorides             1     10.75 1939.8 -3717.5
## + sulphates             1      7.62 1943.0 -3710.3
## + total.sulfur.dioxide  1      4.89 1945.7 -3704.0
## + fixed.acidity         1      2.27 1948.3 -3698.0
## <none>                              1950.6 -3694.8
## + citric.acid           1      0.08 1950.5 -3693.0
## - pH                    1     13.10 1963.7 -3666.9
## - free.sulfur.dioxide   1     13.59 1964.2 -3665.8
## - residual.sugar        1     40.20 1990.8 -3605.6
## - volatile.acidity      1    124.03 2074.6 -3421.2
## - alcohol               1    717.41 2668.0 -2296.8
##
## Step:  AIC=-3731.6
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     pH + density
##
##                        Df Sum of Sq    RSS     AIC
## + sulphates             1    15.003 1918.7 -3764.4
## + fixed.acidity         1     7.477 1926.2 -3746.9
## + chlorides             1     7.227 1926.5 -3746.3
## + total.sulfur.dioxide  1     0.921 1932.8 -3731.7
## <none>                              1933.7 -3731.6
## + citric.acid           1     0.734 1933.0 -3731.3
## - free.sulfur.dioxide   1    12.652 1946.4 -3704.5
## - density               1    16.856 1950.6 -3694.8
## - pH                    1    22.073 1955.8 -3682.9
## - residual.sugar        1    43.659 1977.4 -3633.8
## - alcohol               1    93.751 2027.5 -3522.0
## - volatile.acidity      1   122.933 2056.7 -3458.1
##
## Step:  AIC=-3764.42
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     pH + density + sulphates
##
##                        Df Sum of Sq    RSS     AIC
## + fixed.acidity         1    13.866 1904.8 -3794.8
## + chlorides             1     7.307 1911.4 -3779.5
## + total.sulfur.dioxide  1     1.385 1917.3 -3765.6
## <none>                              1918.7 -3764.4
## + citric.acid           1     0.638 1918.1 -3763.9
## - free.sulfur.dioxide   1    10.723 1929.4 -3741.5
## - sulphates             1    15.003 1933.7 -3731.6
## - pH                    1    19.103 1937.8 -3722.1
## - density               1    24.236 1943.0 -3710.3
## - residual.sugar        1    53.656 1972.4 -3643.1
## - alcohol               1    77.580 1996.3 -3589.2
## - volatile.acidity      1   119.572 2038.3 -3496.2
```

```
##
## Step:  AIC=-3794.84
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     pH + density + sulphates + fixed.acidity
##
##                        Df Sum of Sq    RSS     AIC
## + chlorides             1     4.416 1900.4 -3803.2
## <none>                              1904.8 -3794.8
## + total.sulfur.dioxide  1     0.442 1904.4 -3793.9
## + citric.acid           1     0.243 1904.6 -3793.4
## - alcohol               1    10.686 1915.5 -3771.8
## - free.sulfur.dioxide   1    11.899 1916.8 -3769.0
## - fixed.acidity         1    13.866 1918.7 -3764.4
## - sulphates             1    21.392 1926.2 -3746.9
## - pH                    1    32.548 1937.4 -3721.1
## - density               1    35.384 1940.2 -3714.6
## - residual.sugar        1    57.471 1962.3 -3664.0
## - volatile.acidity      1   108.255 2013.1 -3549.8
##
## Step:  AIC=-3803.21
## quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
##     pH + density + sulphates + fixed.acidity + chlorides
##
##                        Df Sum of Sq    RSS     AIC
## <none>                              1900.4 -3803.2
## + total.sulfur.dioxide  1     0.324 1900.1 -3802.0
## + citric.acid           1     0.164 1900.3 -3801.6
## - chlorides             1     4.416 1904.8 -3794.8
## - fixed.acidity         1    10.975 1911.4 -3779.5
## - alcohol               1    11.304 1911.7 -3778.7
## - free.sulfur.dioxide   1    11.900 1912.3 -3777.3
## - sulphates             1    20.675 1921.1 -3756.8
## - pH                    1    28.415 1928.8 -3738.9
## - density               1    28.442 1928.9 -3738.8
## - residual.sugar        1    48.077 1948.5 -3693.5
## - volatile.acidity      1   105.914 2006.3 -3562.8
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + pH + density + sulphates + fixed.acidity +
##     chlorides, data = updated_Dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99007 -0.48040 -0.04752  0.44836  2.19096
##
## Coefficients:
##                         Estimate  Std. Error t value Pr(>|t|)
## (Intercept)          182.306130   22.102999   8.248  < 2e-16 ***
## alcohol                0.145369    0.028224   5.151 2.71e-07 ***
## volatile.acidity      -1.798684    0.114087 -15.766  < 2e-16 ***
## residual.sugar         0.088465    0.008328  10.622  < 2e-16 ***
## free.sulfur.dioxide    0.003615    0.000684   5.285 1.32e-07 ***
## pH                     0.870439    0.106591   8.166 4.11e-16 ***
## density             -183.016739   22.401078  -8.170 3.98e-16 ***
```

```
## sulphates              0.693724    0.099592   6.966 3.75e-12 ***
## fixed.acidity          0.114772    0.022615   5.075 4.03e-07 ***
## chlorides             -2.976533    0.924643  -3.219   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6528 on 4460 degrees of freedom
## Multiple R-squared:  0.3272, Adjusted R-squared:  0.3259
## F-statistic:   241 on 9 and 4460 DF,  p-value: < 2.2e-16
```

Backward Stepwise Selection

```
# Using forward stepwise selction

backward_model<-step(updated_full_model,direction = "backward")
summary(backward_model)
```

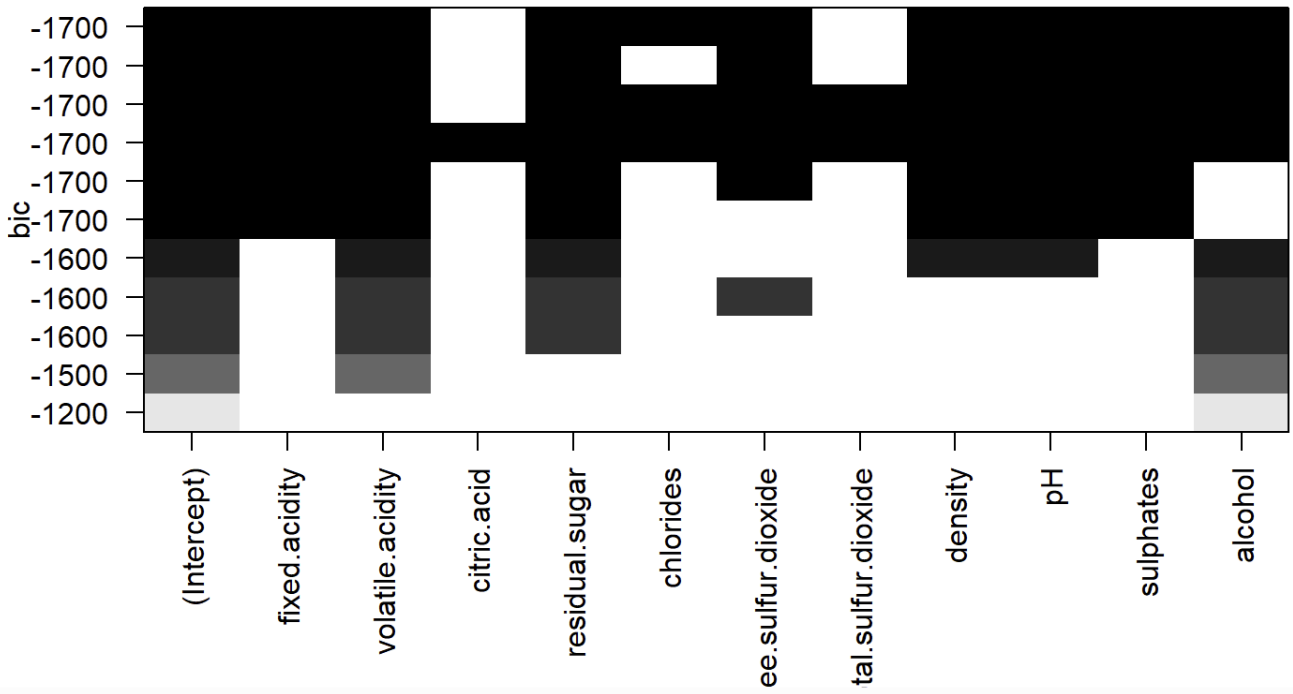
```
## Start:  AIC=-3800.41
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS     AIC
## - citric.acid           1     0.182 1900.1 -3802.0
## - total.sulfur.dioxide  1     0.342 1900.3 -3801.6
## <none>                              1899.9 -3800.4
## - chlorides             1     4.214 1904.1 -3792.5
## - free.sulfur.dioxide   1     9.513 1909.4 -3780.1
## - fixed.acidity         1    10.031 1910.0 -3778.9
## - alcohol               1    11.251 1911.2 -3776.0
## - sulphates             1    20.646 1920.6 -3754.1
## - density               1    24.910 1924.8 -3744.2
## - pH                    1    27.610 1927.5 -3737.9
## - residual.sugar        1    43.883 1943.8 -3700.3
## - volatile.acidity      1    94.306 1994.2 -3585.9
##
## Step:  AIC=-3801.98
## quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS     AIC
## - total.sulfur.dioxide  1     0.324 1900.4 -3803.2
## <none>                              1900.1 -3802.0
## - chlorides             1     4.298 1904.4 -3793.9
## - free.sulfur.dioxide   1     9.670 1909.8 -3781.3
## - fixed.acidity         1    10.310 1910.4 -3779.8
## - alcohol               1    11.627 1911.7 -3776.7
## - sulphates             1    20.788 1920.9 -3755.3
## - density               1    24.728 1924.8 -3746.2
## - pH                    1    27.437 1927.5 -3739.9
## - residual.sugar        1    43.706 1943.8 -3702.3
## - volatile.acidity      1    97.856 1998.0 -3579.5
##
## Step:  AIC=-3803.21
## quality ~ fixed.acidity + volatile.acidity + residual.sugar +
```

```
##     chlorides + free.sulfur.dioxide + density + pH + sulphates +
##     alcohol
##
##                       Df Sum of Sq    RSS     AIC
## <none>                              1900.4 -3803.2
## - chlorides            1     4.416 1904.8 -3794.8
## - fixed.acidity        1    10.975 1911.4 -3779.5
## - alcohol              1    11.304 1911.7 -3778.7
## - free.sulfur.dioxide  1    11.900 1912.3 -3777.3
## - sulphates            1    20.675 1921.1 -3756.8
## - pH                   1    28.415 1928.8 -3738.9
## - density              1    28.442 1928.9 -3738.8
## - residual.sugar       1    48.077 1948.5 -3693.5
## - volatile.acidity     1   105.914 2006.3 -3562.8
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + density + pH + sulphates +
##     alcohol, data = updated_Dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99007 -0.48040 -0.04752  0.44836  2.19096
##
## Coefficients:
##                        Estimate  Std. Error t value Pr(>|t|)
## (Intercept)          182.306130   22.102999   8.248  < 2e-16 ***
## fixed.acidity          0.114772    0.022615   5.075 4.03e-07 ***
## volatile.acidity      -1.798684    0.114087 -15.766  < 2e-16 ***
## residual.sugar         0.088465    0.008328  10.622  < 2e-16 ***
## chlorides             -2.976533    0.924643  -3.219   0.0013 **
## free.sulfur.dioxide    0.003615    0.000684   5.285 1.32e-07 ***
## density             -183.016739   22.401078  -8.170 3.98e-16 ***
## pH                     0.870439    0.106591   8.166 4.11e-16 ***
## sulphates              0.693724    0.099592   6.966 3.75e-12 ***
## alcohol                0.145369    0.028224   5.151 2.71e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6528 on 4460 degrees of freedom
## Multiple R-squared:  0.3272, Adjusted R-squared:  0.3259
## F-statistic:   241 on 9 and 4460 DF,  p-value: < 2.2e-16
```

Best Subsets Regression

```
# Using Best Subsets Regression

best_subsets_model<-regsubsets(quality~.,data=updated_Dataset, nbest=1,nvmax=11)
plot(best_subsets_model, main= "Figure 9: BIC against Regressors")
```

Figure 9: BIC against Regressors



INTERPRETATION

After conducting best subsets regression and stepwise regression, we found that all results interestingly enough produced the same best model as our previously identified one. Both methods systematically evaluated different combinations of predictors and, despite their different approaches, consistently pointed to the same model as the optimal choice. This convergence of results allows for a greater confidence in the accuracy and effectiveness of our final best model.

Justify the Inclusion of Variable

As such, we can say that the best model fitted according to our steps is shown as follows:

$$y = 205.698 + 0.110x_1 - 1.671x_2 + 0.092x_3 + 0.004x_4 + 1.110x_5 - 207.210x_6 + 0.751x_7 + 0.145x_8 - 4.736x_9$$

$Where:$

$x_1 = alcohol$

$x_2 = volatile\ acidity$

$x_3 = residual\ sugar$

$x_4 = free\ sulfur\ dioxide$

$x_5 = pH\ level$

$x_6 = density$

$x_7 = sulphates$

$x_8 = fixed\ acidity$

$x_9 = chlorides$

1. **Volatile and Fixed Acids**
   - Responsible for the wine's freshness and vitality
   - Good acid balance - zesty, crisp character.
   - Good structure and aging.

   Exclusion of Citric acid : Not significant - Weak organic acid (Not much effect as compared to other fixed acids) - Microbial instability - Growth of unwanted microbes

   (Jamescharleswine, 2023)

2. **Residual Sugars**
   - Natural grape sugars leftover.
   - Affects a wine's sweetness.
   - Sweet Wine: 45 g/l.

   (Wu, 2020)

3. **Alcohol Content**
   - Subtle impression of sweetness.
   - Hint of bitterness
   - Less Content – Flat and Dull

   (LiquidLINE Blog – Things That Affect the Wine Quality, n.d.-b)

4. **Chlorides**
   - Saltiness of a wine

   (Application Note #105 -Chloride in Wine by Titration, n.d.)

5. **Density**
   - Lower density is preffered
   - Higher density - imbalance composition

6. **Total and Free Sulphur Dioxide**
   - Preservative
   - Continuously monitor sulfur dioxide levels until wine is bottled.
   - Act as a cushion for FSO2.
   - FSO2 is lost, chemical equilibrium in the wine shifts so that some of the S02 can be released to its free state.

   (Admin, 2018) & (Comfort, 2008)

7. **pH Level**
   - Low pH wines will taste tart and crisp
   - Higher pH wines are more susceptible to bacterial growth
   - 3.0 to 3.4 for white wines.
   - 3.3 to 3.6 for red wines.

   (Edison, 2022)

8. **Sulphates**
   - Potassium Metabisulfite (derivatives of SO2)
   - Antioxidant to maintain the wine's colour, flavour, and aroma.
   - Also a result of the lightly oxidizing environment represented by wooden casks.
   - Low sulfate levels could attribute to wine being stored in stainless steel containers.

   (Contributors, 2021)

Conclusion and Recommendation

Based on the results obtained, the following may contribute to higher wine quality:

a) High alcohol level (over 10)

By harvesting grapes later in the season, particularly 1-2 months after the regular harvest time will allow them to accumulate more sugars which will result in higher alcohol content. This is a result of the yeast molecules having more sugar available to convert to alcohol during fermentation (Puckette, 2016). Other than that, winemakers can utilize cryoextraction, which involves freezing grapes and removing water in the form of ice to increase alcohol content in wine. This method does so by further concentrating sugars (Tatum, 2016).

b) Low density (under 0.998)

In order to achieve low density wines, winemakers should maintain optimal fermentation temperatures (32°C to 35°C) to enable the fermentation process to be completed efficiency. This will reduce residual sugars which will lead to a lower wine density (Liszkowska & Berlowska, 2021). Besides that, winemakers can blend the wine with another wine with a lower alcohol content or even water, although it is strictly regulated and rarely used, to reduce alcohol content which will in turn reduce density (Sullivan, 2023).

c) Increased pH (over 3)

In order to raise the alkalinity of wine, winemakers can add chemical compounds, such as potassium carbonate and calcium carbonate. They help neutralize acidity by reacting with the tartaric acid in wine to precipitate calcium tartrate or potassium bitartrate out of the solution (Pambianchi, n.d.). Malolactic fermentation (MLF) can also help increase the pH of wines. This method is a secondary fermentation process where lactic acid bacteria convert malic acid into lactic acid, which is a weaker acid which reduces the acidity of the wine. This also helps to add complexity and softness to the wine, especially red wine and some white wines, like Chardonnay (Lonvaud-Funel, 2022).

By implementing the above techniques, Vinho Verde will be able to meet consumer expectations on a more prestigious level and maintain their reputation as the top wine brand in Portugal.

References

- Admin. (2018, February 27). *Total Sulfur Dioxide – Why it Matters, Too!* Midwest Grape and Wine Industry Institute. https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too/
- *Application Note #105 -Chloride in Wine by Titration*. (n.d.). https://mantech-inc.com/wp-content/uploads/2014/07/105-Chloride-in-Wine-by-Titration.pdf
- Comfort, S. (2008, December 5). *About Acidity and Adding Acid to Must/Wine | MoreWine*. Morewinemaking.com. https://morewinemaking.com/articles/Acidifying_must
- Contributors, W. E. (2021, June 22). *What to Know About Sulfites in Wine*. WebMD. https://www.webmd.com/diet/what-to-know-sulfites-in-wine
- Corduas, M., Cinquanta, L., & Ievoli, C. (2013). The importance of wine attributes for purchase decisions: A study of Italian consumers' perception. *Food Quality and Preference*, *28*(2), 407–418. https://doi.org/10.1016/j.foodqual.2012.11.007
- Edison, T. (2022, January 26). *Is Wine Acidic or Alkaline? [Guide to pH in Winemaking]*. Wineturtle.com. https://wineturtle.com/wine-acidic-alkaline-basic/
- Jamescharleswine. (2023, September 26). *The role of acids, sugars, and tannins in wine quality*. James Charles Winery. https://jamescharleswine.com/the-role-of-acids-sugars-and-tannins-in-wine-quality/#:~:text=Role%3A%20Acids%20are%20essential%20in,tartaric%2C%20malic%2C%20and%20citric
- *LiquidLINE Blog – Things that Affect the Wine Quality*. (n.d.). OPSIS LiquidLINE. https://www.liquidline.se/blog/things-that-affect-the-wine-quality/#:~:text=The%20alcohol%20in%20wine%20gives,much%20and%20not%20too%20little.
- Liszkowska, W., & Berlowska, J. (2021). Yeast Fermentation at Low Temperatures: Adaptation to Changing Environmental Conditions and Formation of Volatile Compounds. Molecules, 26(4). https://doi.org/10.3390/molecules26041035

- Lonvaud-Funel, A. (2022). Malolactic fermentation and its effects on wine quality and safety. Elsevier EBooks, 105–139. https://doi.org/10.1016/b978-0-08-102065-4.00008-0
- Marianthi Basalekou, Panagiotis Tataridis, Georgakis, K., & Christos Tsintonis. (2023). Measuring Wine Quality and Typicity. *Beverages*, *9*(2), 41–41. https://doi.org/10.3390/beverages9020041
- Pambianchi, D. (n.d.). Monitoring & Adjusting pH. WineMakerMag.com. https://winemakermag.com/technique/1650-monitoring-adjusting-ph
- Puckette, M. (2016, October 3). Late Harvest Wines and Why They're Awesome. Wine Folly. https://winefolly.com/tips/late-harvest-wines-and-why-theyre-awesome/
- Sullivan, S. P. (2023, May 4). Basics: The Why, When and How of Wine Blending | Wine Enthusiast. Www.wineenthusiast.com. https://www.wineenthusiast.com/culture/wine/wine-blending/
- Tatum, M. (2024, May 16). What is Cryoextraction? (with pictures). DelightedCooking. https://www.delightedcooking.com/what-is-cryoextraction.htm
- Wu, S. (2020, July 16). *What is residual sugar in wine? – Ask Decanter*. Decanter. https://www.decanter.com/learn/residual-sugar-46007/#:~:text=The%20amount%20of%20residual%20sugar,be%20consumed%20by%20the%20yeast