WINE QUALITY PREDICTION IN R

Wine is an alcoholic beverage made from grapes, generally Vitis vinifera, fermented without the addition of sugars, acids, enzymes, water, or other nutrients. Yeast consumes the sugar in the grapes and converts it to ethanol and carbon dioxide. Different varieties of grapes and strains of yeasts produce different styles of wine. These variations result from the complex interactions between the biochemical development of the grape, the reactions involved in fermentation, the terroir, and the production process.

Our goal was to find a regression model of Wine quality with the various physicochemical variables. Physicohemical Properties: -> fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily) (tartaric acid - g / dm^3) -> volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste (acetic acid - g / dm^3) -> citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines (g / dm^3) -> residual sugar: the amount of sugar remaining after fermentation stops (g / dm^3) ->chlorides: the amount of salt in the wine (sodium chloride - g / dm^3 -> free sulfur dioxide: he free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion (mg / dm^3) -> total sulfur dioxide: amount of free and bound forms of S02 (mg / dm^3) -> density: the density of water is close to that of water depending on the percent alcohol and sugar content (g / cm^3) -> pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic) -> sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels (potassium sulphate - g / dm3) -> alcohol: the percent alcohol content of the wine (% by volume) Output variable (based on sensory data): -> quality (score between 0 and 10) .

First, before doing any analysis between the variables its necessary to plot the distribution of each of the variable. Based on the distribution shape, i.e. Normal, Positive Skew or Negative Skew, this will help us to plot different variables against each other. Also for many variables, there are extreme outliers present in this dataset. For those, we need to remove the extreme outliers for a more robust analysis. The project first imports the necessary libraries, including quantmod, forecast, ggplot2,dplyr,gridExtra,GGallymemisc,pander and corrplot

CODE :

```
library("ggplot2")
library("dplyr")
library("gridExtra")
library(Simpsons)
library(GGally)
library(memisc)
library(pander)
library(corrplot)

wine <- read.csv('wineQualityReds.csv')

#Converting Wine quality into a ordered factor
wine$quality <- factor(wine$quality, ordered = T)
wine$rating <- ifelse(wine$quality < 5, 'bad', ifelse(
  wine$quality < 7, 'average', 'good'))
wine$rating <- ordered(wine$rating,
                       levels = c('bad', 'average', 'good'))

wine$X = factor(wine$X)
#Structure of the Dataframe

str(wine)
summary(wine)

#Univariate plots #Quality and rating
ggplot(data = wine, aes(x = quality)) +
  stat_count(width = 1, color = 'black',fill = I('orange'))

ggplot(data = wine, aes(x = rating)) +
  stat_count(width = 1, color = 'black',fill = I('blue'))
```

```
summary(wine$fixed.acidity)  #Median = 7.9 but some outliers dragged the mean upto 8.32
summary(wine$volatile.acidity)
summary(wine$citric.acid)
summary(wine$residual.sugar)
summary(wine$chlorides)
summary(wine$free.sulfur.dioxide)
summary(wine$total.sulfur.dioxide)
summary(wine$density)
summary(wine$pH)
summary(wine$sulphates)
summary(wine$alcohol)

grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,p11, ncol = 4)

#Bivariate analysis #Correlation table
c <- cor(
  wine %>%
    # first we remove unwanted columns
    dplyr::select(-X) %>%
    dplyr::select(-rating) %>%
    mutate(
      # now we translate quality to a number
      quality = as.numeric(quality)
    )
)

pandoc.table(c)

#Fixed acidity : Doesn't seem to have much effect
ggplot(data = wine, aes(x = quality, y = fixed.acidity)) +
  geom_boxplot()
```

```
#Volatile Acidity : Seems to have negative effect. With increase, quality seems to go down
ggplot(data=wine, aes(x = quality, y = volatile.acidity)) +
  geom_boxplot()
#Citric acid (Better wines tend to have higher citric acid)
ggplot(data=wine, aes(x=quality, y=citric.acid)) +
  geom_boxplot()

#Residual Sugar(Almost has no effect to quality. This is contrary to previous assumption)
ggplot(data=wine, aes(x=quality, y=residual.sugar)) +
  geom_boxplot()

#Chlorides
ggplot(data=wine, aes(x=quality, y=chlorides)) +
  geom_boxplot()

#Free SO2(We see too little and we get a poor wine and too much : we get an average wine)
ggplot(data=wine, aes(x=quality, y=free.sulfur.dioxide)) +
  geom_boxplot()

#Total SO2(Just like free SO2)
ggplot(data=wine, aes(x=quality, y=total.sulfur.dioxide)) +
  geom_boxplot()

#Density(Better wines tend to have lower densities but is it due to alcohol content?)
ggplot(data=wine, aes(x=quality, y=density)) +
  geom_boxplot()

#pH(Better wines seems to be more acidic. Now let's see contribution of each acid on pH)
ggplot(data=wine, aes(x=quality, y=pH)) +
  geom_boxplot()

#Contribution of each acid to pH(We see all of them has negative correlation on pH except
```

```
#volatile acidity. But how's that possible! Is it possible that there is a Simson's effect?)

ggplot(data = wine, aes(x = fixed.acidity, y = pH)) +
  geom_point() +
  scale_x_log10(breaks=seq(5,15,1)) +
  xlab("log10(fixed.acidity)") +
  geom_smooth(method="lm")

ggplot(data = wine, aes(x = volatile.acidity, y = pH)) +
  geom_point() +
  scale_x_log10(breaks=seq(.1,1,.1)) +
  xlab("log10(volatile.acidity)") +
  geom_smooth(method="lm")

ggplot(data = subset(wine, citric.acid > 0), aes(x = citric.acid, y = pH)) +
  geom_point() +
  scale_x_log10() +
  xlab("log10(citric.acid)") +
  geom_smooth(method="lm")

#Sulphates(better wines seems to have higher sulphates. Although medium wines have many
outliers)

ggplot(data=wine, aes(x=quality, y=sulphates)) +
  geom_boxplot()

#Alcohol(Better wines have higher alcohol)
ggplot(data=wine, aes(x=quality, y=alcohol)) +
  geom_boxplot()
#Linear model test(From R squared value, it seems alcohol contributes only 22% to the quality
variance)
alcoholQualityLM <- lm(as.numeric(quality) ~ alcohol,
                       data = wine)
```

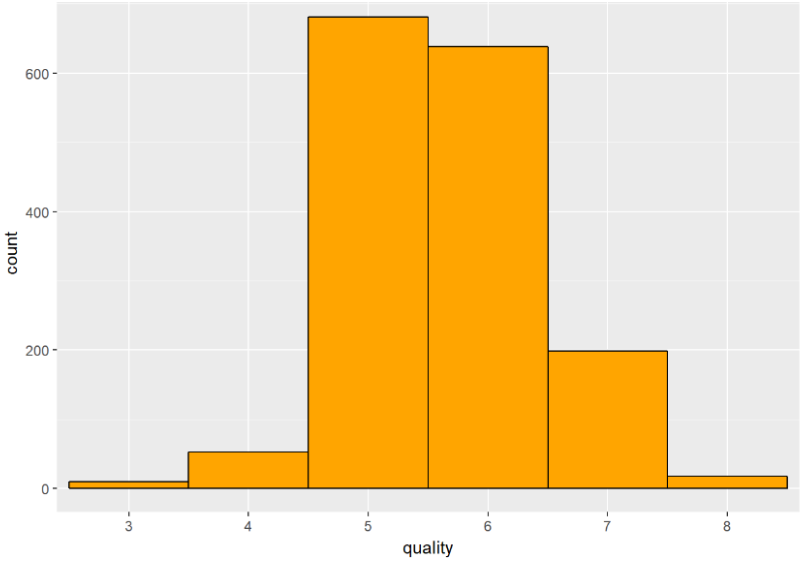
```r
summary(alcoholQualityLM)
df = data.frame(wine$quality )
df$predictions <- predict(alcoholQualityLM, wine)
df$error <- (df$predictions - as.numeric(wine$quality))/as.numeric(wine$quality)

ggplot(data=df, aes(x=wine.quality, y=error)) +
  geom_boxplot()

#Putting a Cor test together

simple_cor_test <- function(x, y) {
  return(cor.test(x, as.numeric(y))$estimate)
}

correlations <- c(
  simple_cor_test(wine$fixed.acidity, wine$quality),
  simple_cor_test(wine$volatile.acidity, wine$quality),
  simple_cor_test(wine$citric.acid, wine$quality),
  simple_cor_test(log10(wine$residual.sugar), wine$quality),
  simple_cor_test(log10(wine$chlorides), wine$quality),
  simple_cor_test(wine$free.sulfur.dioxide, wine$quality),
  simple_cor_test(wine$total.sulfur.dioxide, wine$quality),
  simple_cor_test(wine$density, wine$quality),
  simple_cor_test(wine$pH, wine$quality),
  simple_cor_test(log10(wine$sulphates), wine$quality),
  simple_cor_test(wine$alcohol, wine$quality))
names(correlations) <- c('fixed.acidity', 'volatile.acidity', 'citric.acid',
                         'log10.residual.sugar',
                         'log10.chlordies', 'free.sulfur.dioxide',
                         'total.sulfur.dioxide', 'density', 'pH',
                         'log10.sulphates', 'alcohol')
```

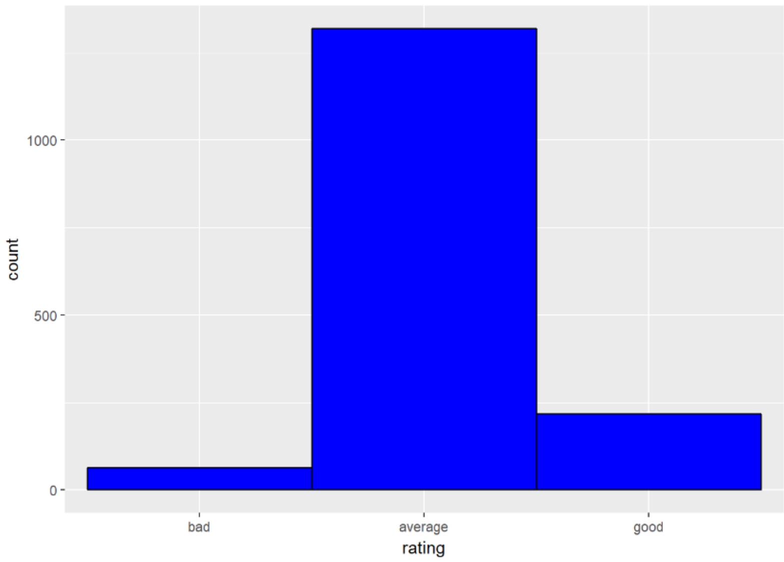
```
correlations
#Making the linear model


set.seed(1221)
training_data <- sample_frac(wine, .6)
test_data <- wine[ !wine$X %in% training_data$X, ]
m1 <- lm(as.numeric(quality) ~ alcohol, data = training_data)
m2 <- update(m1, ~ . + sulphates)
m3 <- update(m2, ~ . + volatile.acidity)
m4 <- update(m3, ~ . + citric.acid)
m5 <- update(m4, ~ . + fixed.acidity)
m6 <- update(m2, ~ . + pH)
mtable(m1,m2,m3,m4,m5,m6)
df <- data.frame(
  test_data$quality,
  predict(m5, test_data) - as.numeric(test_data$quality)
)
names(df) <- c("quality", "error")
ggplot(data=df, aes(x=quality,y=error)) +
  geom_point()
#Final plots
ggplot(data=wine, aes(y=alcohol, x=quality)) +
  geom_boxplot() +
  xlab("alcohol concentration (% by volume)") +
  ggtitle("Influence of alcohol on wine quality")


ggplot(data = wine,
       aes(y = sulphates, x = alcohol,
           color = quality)) +
  geom_point() +
  scale_y_continuous(limits=c(0.3,1.5)) +
  ylab("potassium sulphate (g/dm3)") +
```

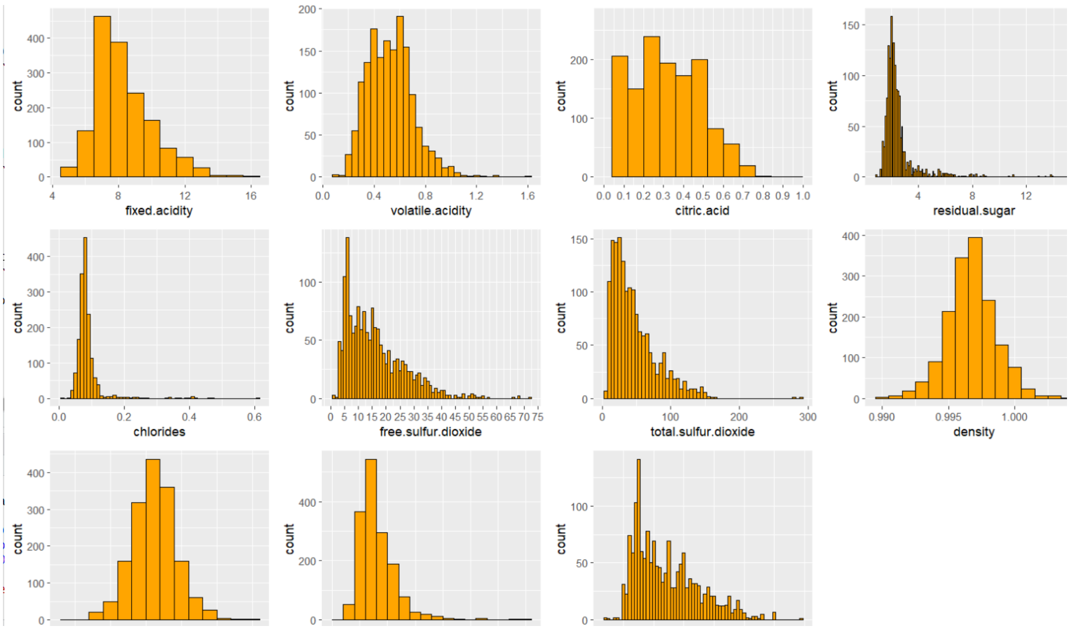
```
  xlab("alcohol (% by volume)") +
  scale_color_brewer() +
  ggtitle("Alcohol and sulphates over wine quality")
df <- data.frame(
  test_data$quality,
  predict(m5, test_data) - as.numeric(test_data$quality)
)
names(df) <- c("quality", "error")
ggplot(data=df, aes(x=quality,y=error)) +
  geom_point() +
  ggtitle("Linear model errors over expected quality")
```

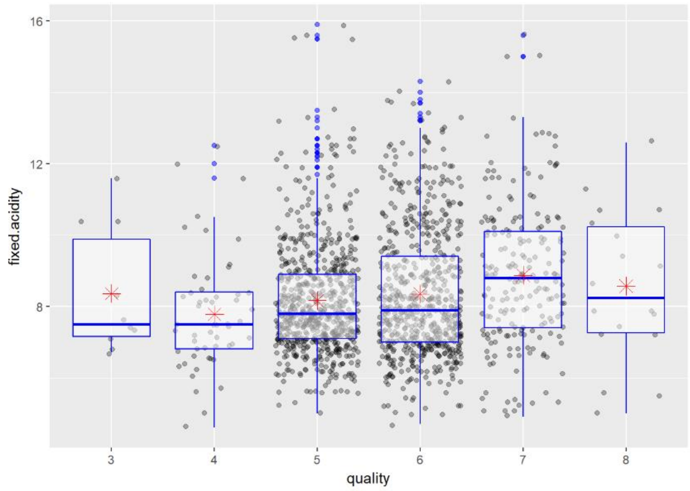
Summary and structure :

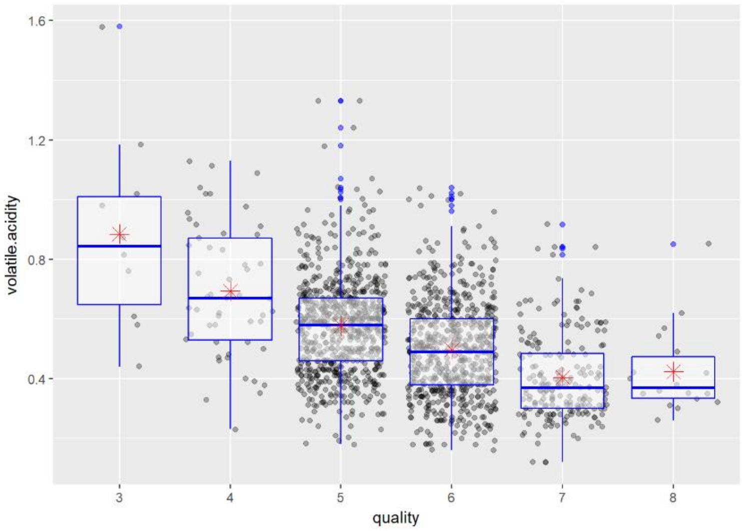
```
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : Ord.factor w/ 6 levels "3"<"4"<"5"<"6"<..: 3 3 3 4 3 3 3 5 5 3 ...
 $ rating              : Ord.factor w/ 3 levels "bad"<"average"<..: 2 2 2 2 2 2 2 3 3 2 ...
> summary(wine)
       X          fixed.acidity   volatile.acidity  citric.acid     residual.sugar
 1      :   1   Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
 2      :   1   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
 3      :   1   Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
 4      :   1   Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
 5      :   1   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
 6      :   1   Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
 (Other):1593
   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density             pH
 Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901   Min.   :2.740
 1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210
 Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968   Median :3.310
 Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967   Mean   :3.311
 3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400
 Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037   Max.   :4.010

   sulphates         alcohol        quality     rating
 Min.   :0.3300   Min.   : 8.40   3: 10    bad    :  63
 1st Qu.:0.5500   1st Qu.: 9.50   4: 53    average:1319
 Median :0.6200   Median :10.20   5:681    good   : 217
 Mean   :0.6581   Mean   :10.42   6:638
 3rd Qu.:0.7300   3rd Qu.:11.10   7:199
 Max.   :2.0000   Max.   :14.90   8: 18
```

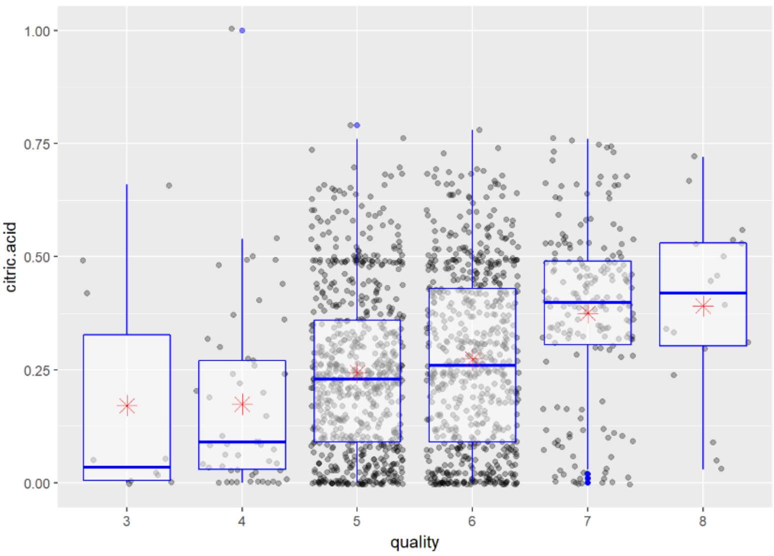

count
0
200
400
600
3
4
5
6
7
8
quality


count
0
500
1000
bad
average
good
rating

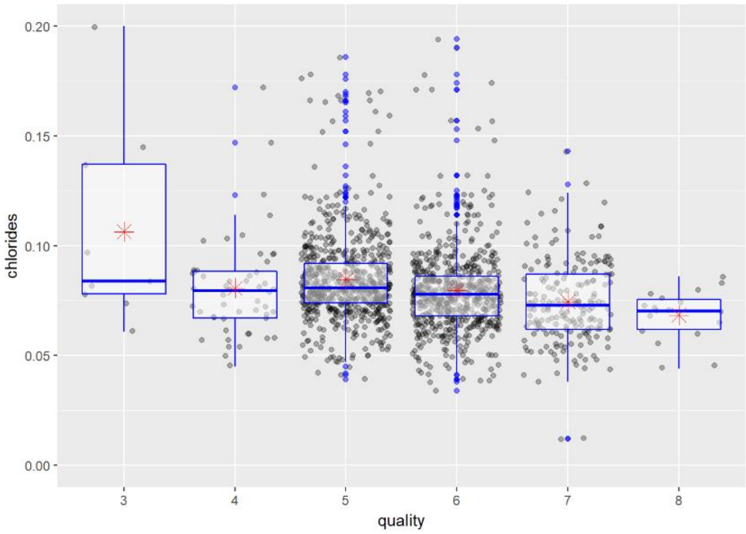

count
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density



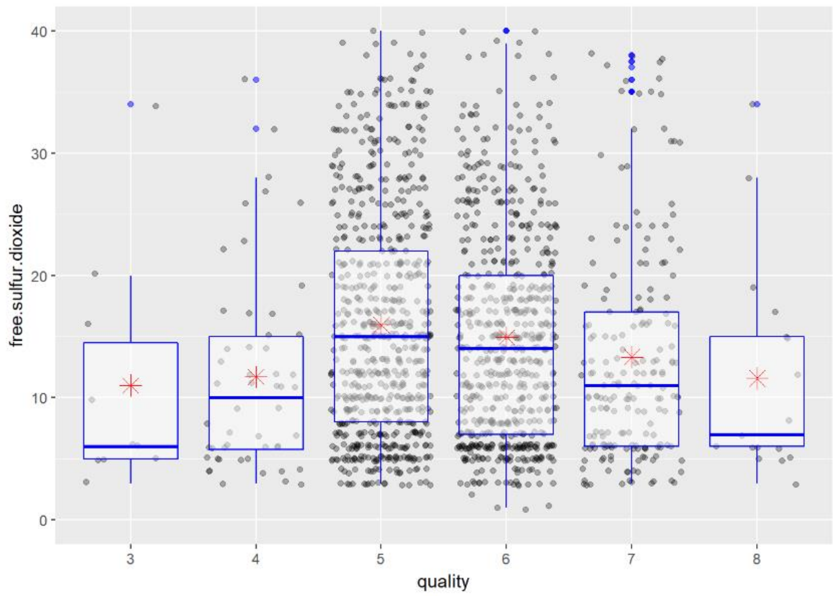




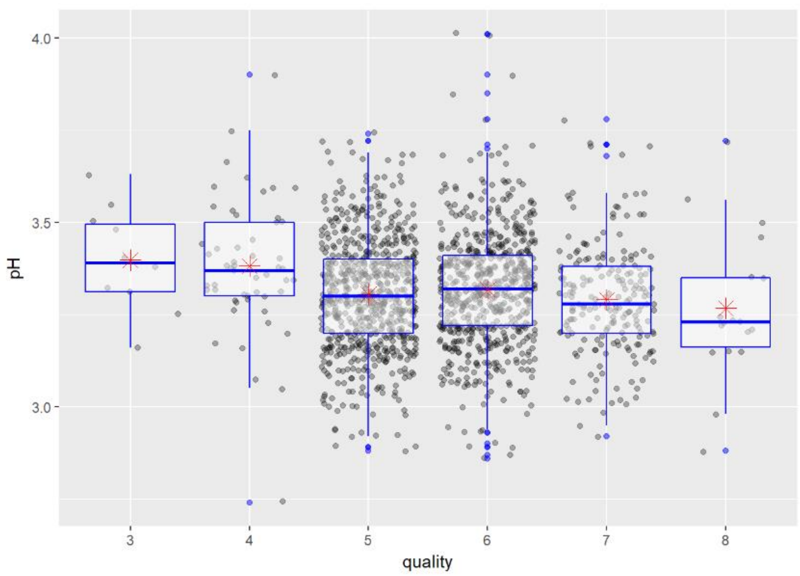
Citric acid seems to have a positive correlation with Wine Quality. Better wines have higher Citric Acid.



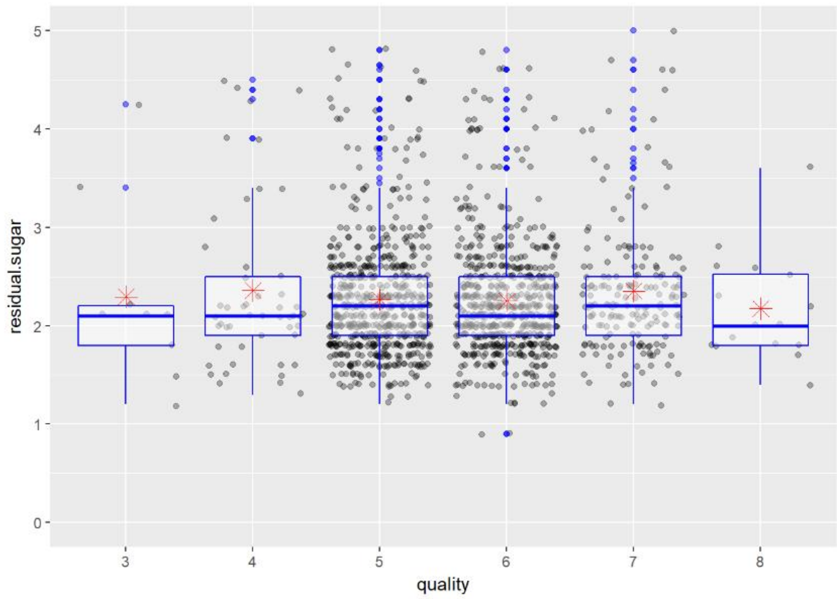
0.20
0.15
0.10
0.05
0.00
chlorides
3
4
5
6
7
8
quality

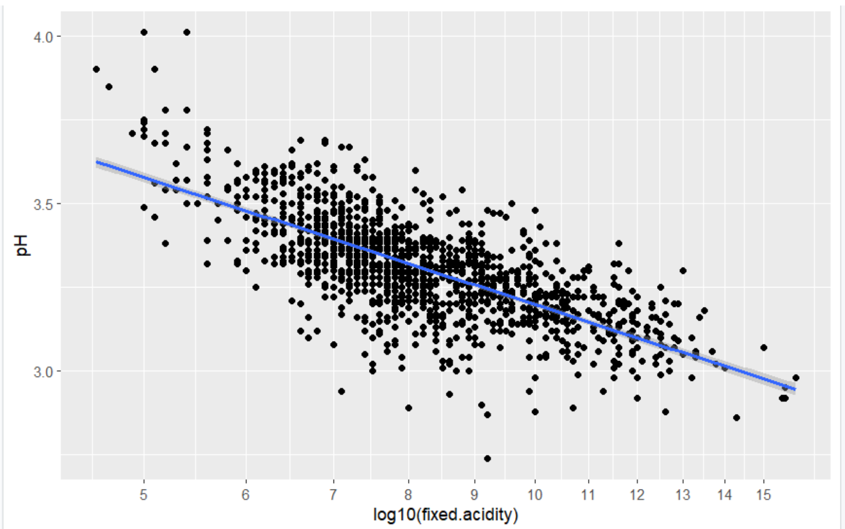


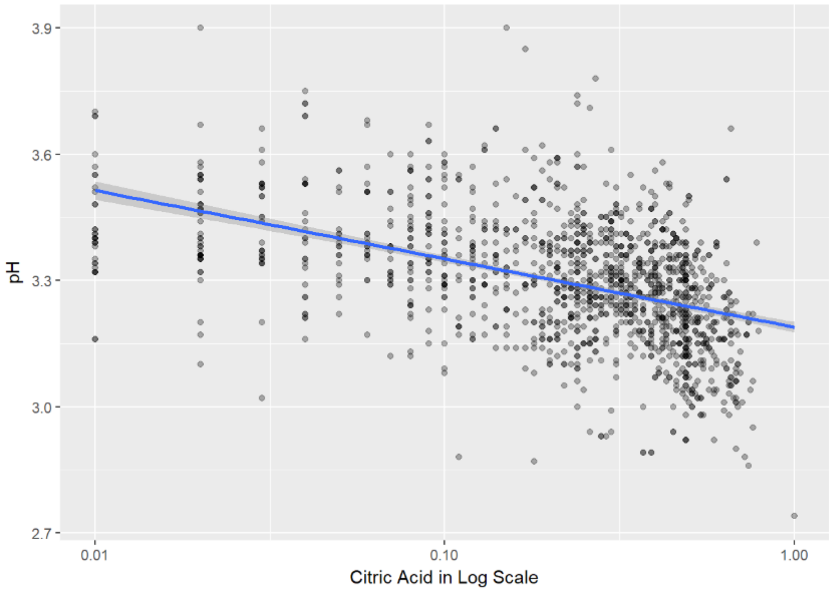
40
30
20
10
0
free.sulfur.dioxide
3
4
5
6
7
8
quality

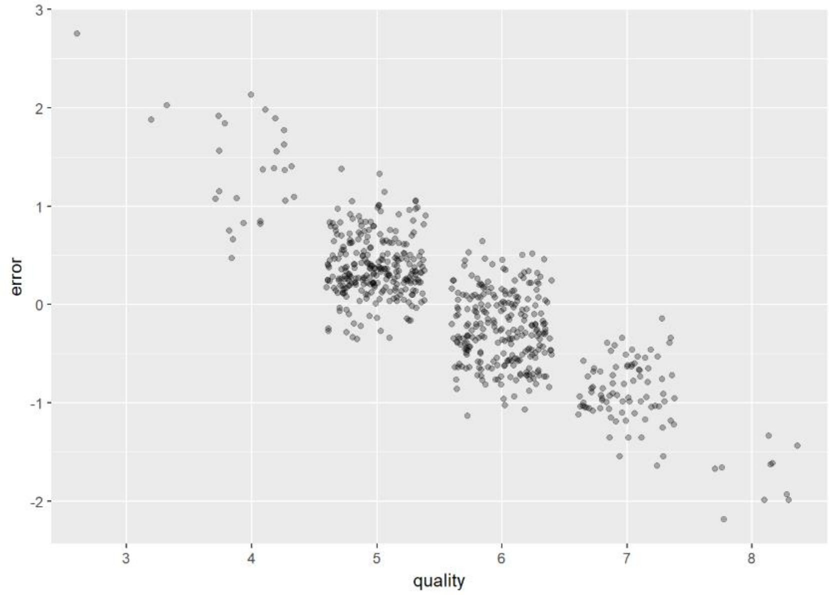


4.0
3.5
3.0
pH
3
4
5
6
7
8
quality

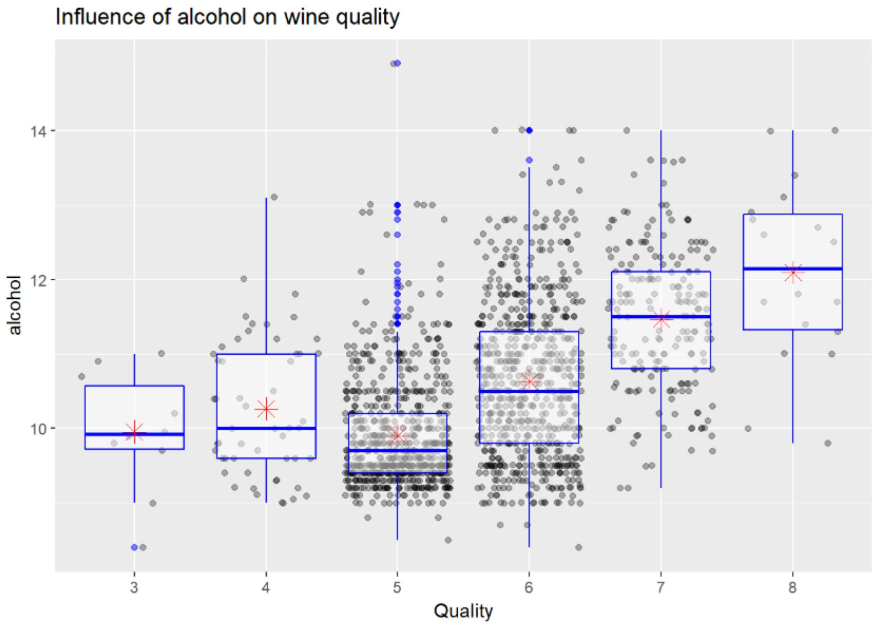

residual.sugar
quality


pH
log10(fixed.acidity)


pH
Citric Acid in Log Scale



Result :



This plot tells us that Alcohol percentage has played a big role in determining the quality of Wines. The higher the alcohol percentage, the better the wine quality. In this dataset, even though most of the data pertains to average quality wine, we can see from the above plot that the mean and median coincides for all the boxes implying that for a particular Quality it is very normally distributed. So a very high value of the median in the best quality wines imply that almost all points have a high percentage of alcohol. But previously from our linear model test, we saw from the R Squared value that alcohol alone contributes to about 22% in the variance of the wine quality. So alcohol is not the only factor which is responsible for the improvement in Wine Quality.

In this plot, we see that the best quality wines have high values for both Alcohol percentage and Sulphate concentration implying that High alcohol contents and high sulphate concentrations together seem to produce better wines. Although there is a very slight downwards slope maybe because in best quality wines, percentage of alcohol is slightly greater than the concentration of Sulphates.