Project6

Irene Cheung

1/28/2019

Introduction

We are going to investigate the Red Wine dataset on physicochemical properties and quality ratings. By analyzing a dataset with 1,599 red wine samples from the north of Portugal. Each wine sample comes with a quality rating from 1 to 10, and results from several physical chemical tests,like: alcohol content, acidity level and residual sugar. There are 11 columns describing their chemical properties, and a column for quality ratings.

Description of attributes

1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

2 - volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

3 - citric acid: found in small quantities, citric acid can add ‘freshness’ and flavor to wines

4 - residual sugar: the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet

5 - chlorides: the amount of salt in the wine

6 - free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

7 - total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine

8 - density: the density of water is close to that of water depending on the percent alcohol and sugar content

9 - pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant

11 - alcohol: the percent alcohol content of the wine

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

Objective

To understand their individual characteristics AND investigate which chemical propeties are the most influence on the quality of red wines.

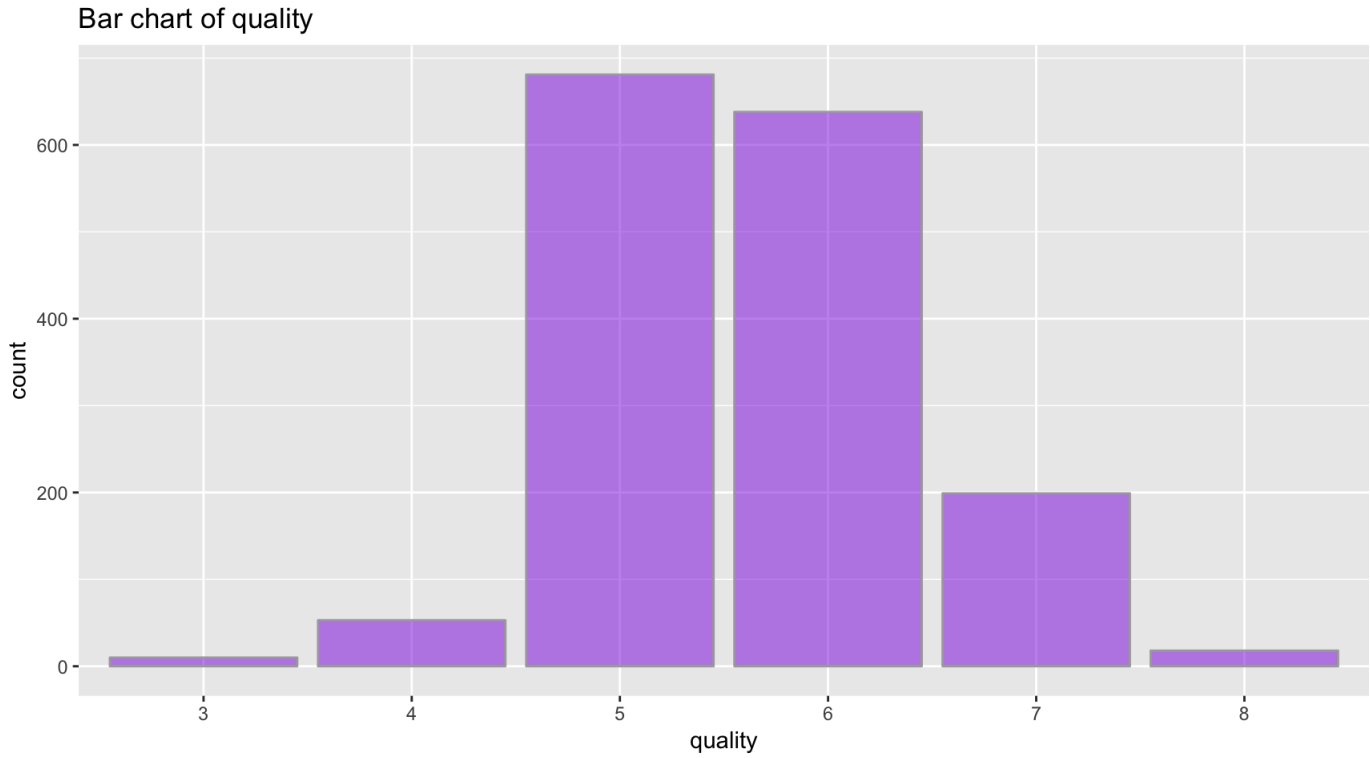
```
# knitr: Suppress code/messages/warnings
knitr::opts_chunk$set( echo=FALSE,warning=FALSE,message=FALSE)
```

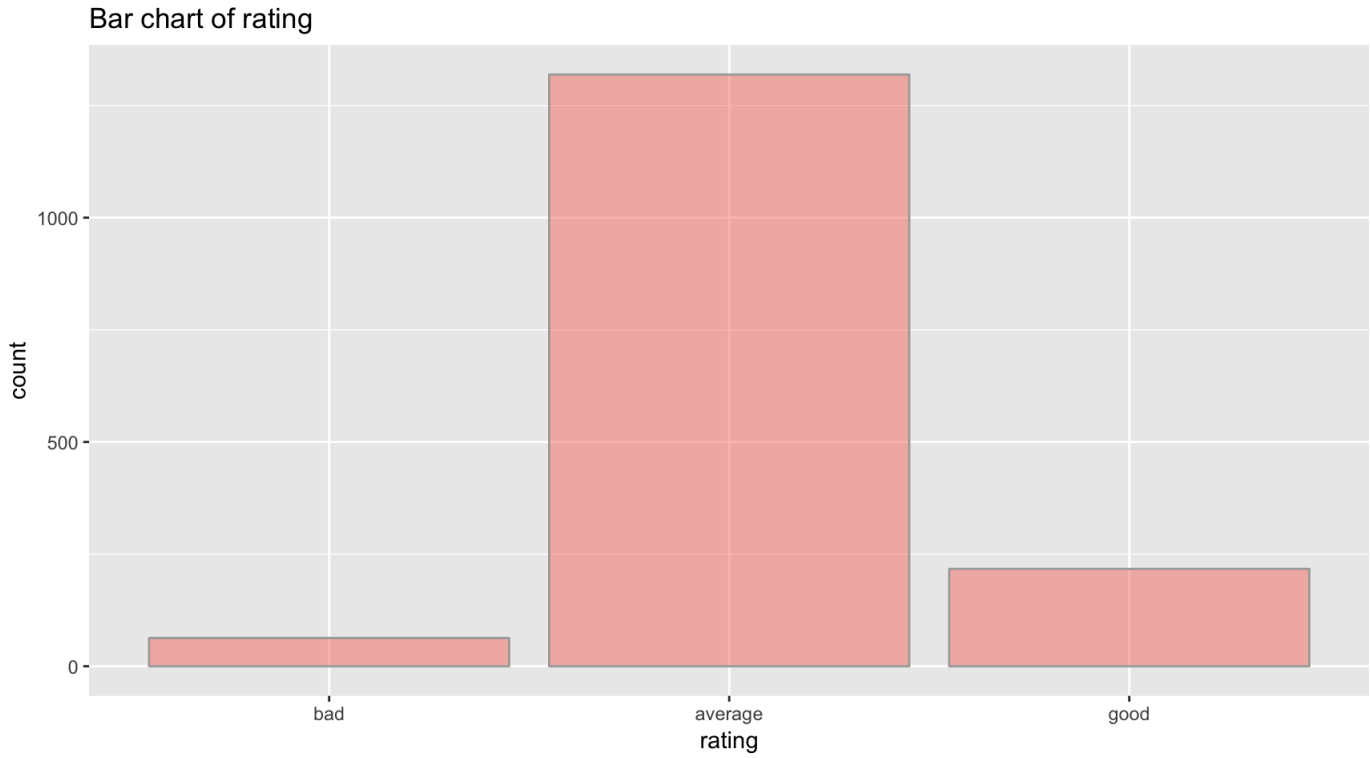
```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 1599 obs. of  14 variables:
##  $ X1                 : num  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity      : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity   : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ 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 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   X1 = col_double(),
##   ..   fixed.acidity = col_double(),
##   ..   volatile.acidity = col_double(),
##   ..   citric.acid = col_double(),
##   ..   residual.sugar = col_double(),
##   ..   chlorides = col_double(),
##   ..   free.sulfur.dioxide = col_double(),
##   ..   total.sulfur.dioxide = col_double(),
##   ..   density = col_double(),
##   ..   pH = col_double(),
##   ..   sulphates = col_double(),
##   ..   alcohol = col_double(),
##   ..   quality = col_double()
##   .. )
```

```
##        X1         fixed.acidity   volatile.acidity  citric.acid
##  Min.   :   1.0   Min.   : 4.60   Min.   :0.1200   Min.   :0.000
##  1st Qu.: 400.5   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090
##  Median : 800.0   Median : 7.90   Median :0.5200   Median :0.260
##  Mean   : 800.0   Mean   : 8.32   Mean   :0.5278   Mean   :0.271
##  3rd Qu.:1199.5   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420
##  Max.   :1599.0   Max.   :15.90   Max.   :1.5800   Max.   :1.000
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
##  1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
##  Median : 2.200   Median :0.07900   Median :14.00
##  Mean   : 2.539   Mean   :0.08747   Mean   :15.87
##  3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
##  Max.   :15.500   Max.   :0.61100   Max.   :72.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300
##  1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500
##  Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200
##  Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581
##  3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300
##  Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000
##     alcohol      quality     rating
##  Min.   : 8.40   3: 10   bad    :  63
##  1st Qu.: 9.50   4: 53   average:1319
##  Median :10.20   5:681   good   : 217
##  Mean   :10.42   6:638
##  3rd Qu.:11.10   7:199
##  Max.   :14.90   8: 18
```

Univariate Plots

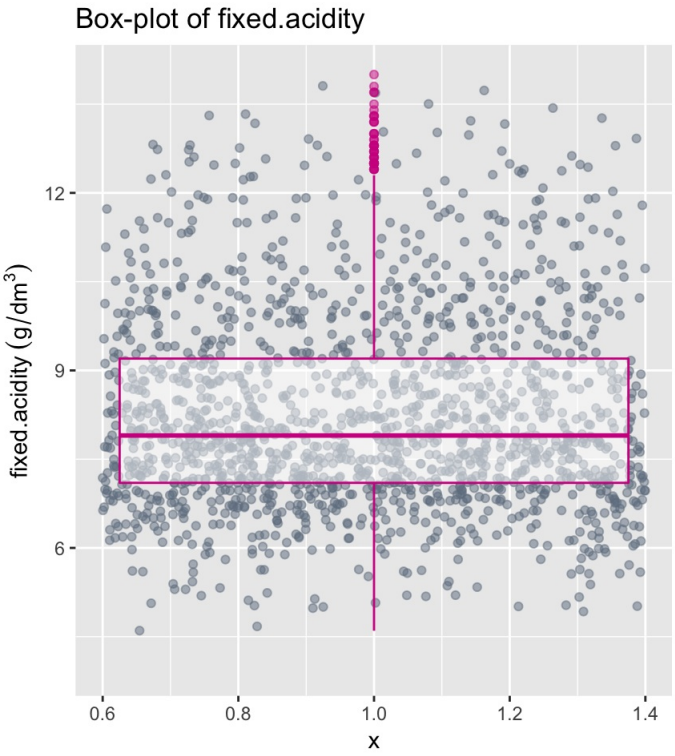
In this part, I am going to plot the distribution of each attributes, like Normal, Positive Skew or Negative Skew.There are outliers present in this dataset, I will remove those extreme outliers for robust analysis.

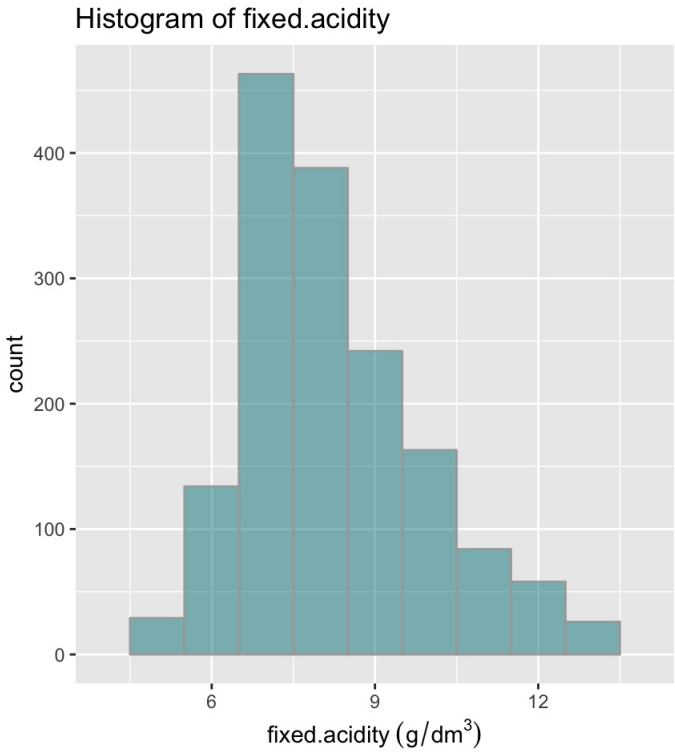




Based on above plots, I found that most of the wines are average quality wines. So it might be difficult to get an accurate model of the Wine Quality. Let's look at the other plots.

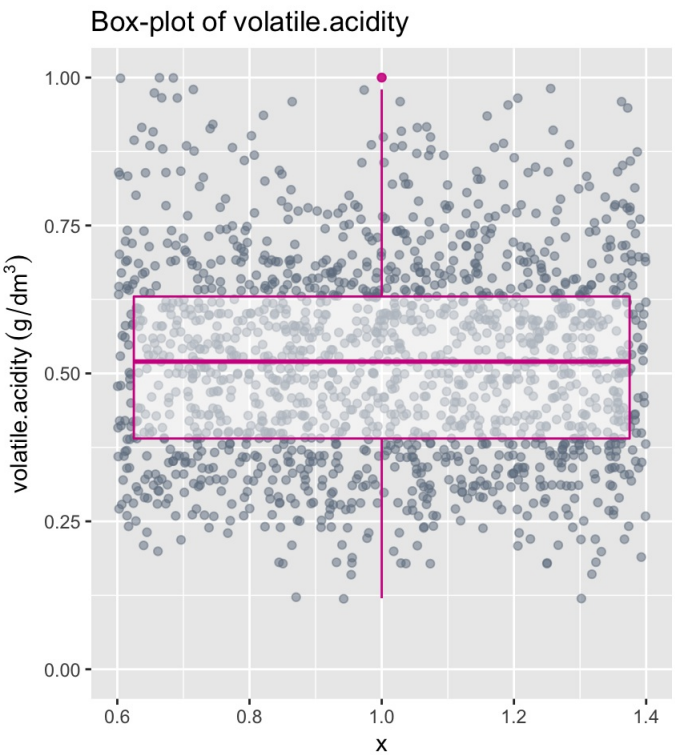
For fixed.acidity

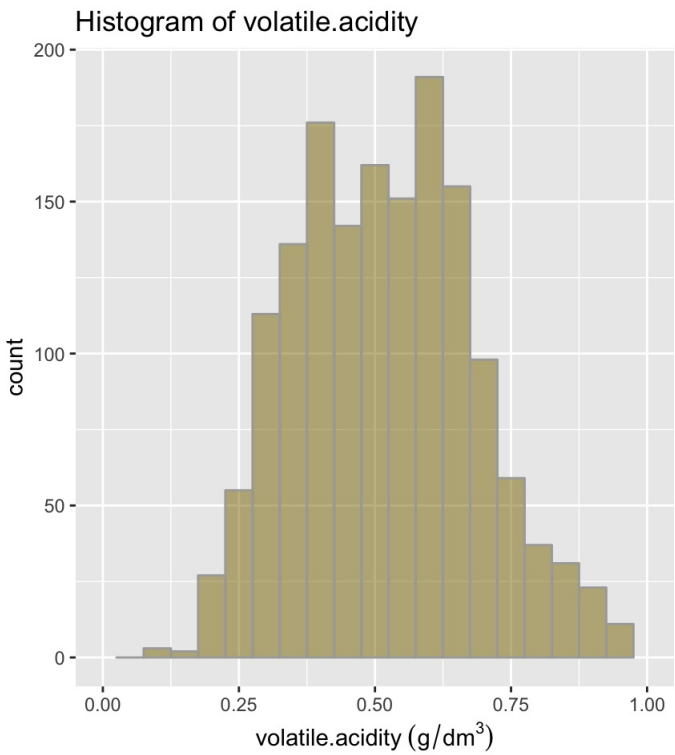




The distribution of Fixed Acidity is skewed right with median and mean are 7.90 and 8.32, respectively. There are some outliers, the image has been rescaled to get rid of the high outliers.

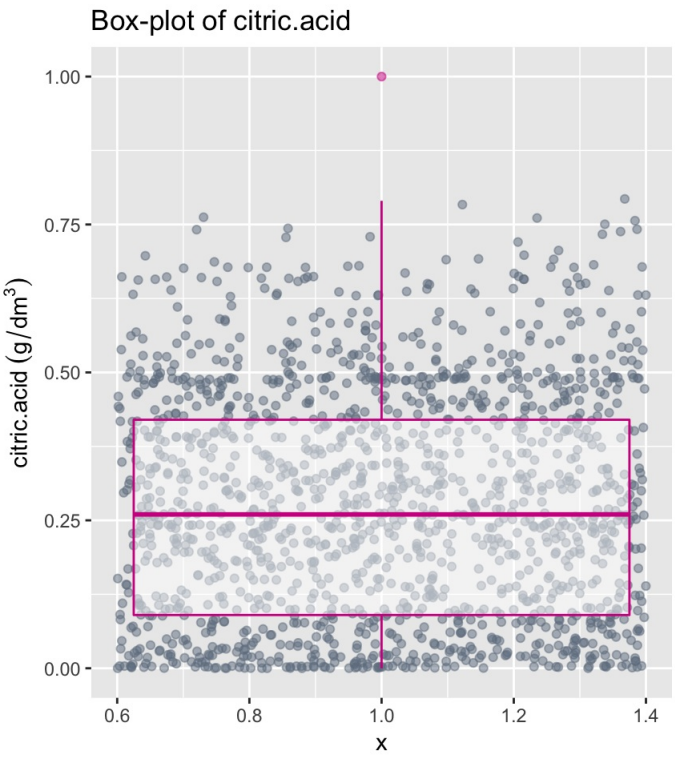
For volatile.acidity

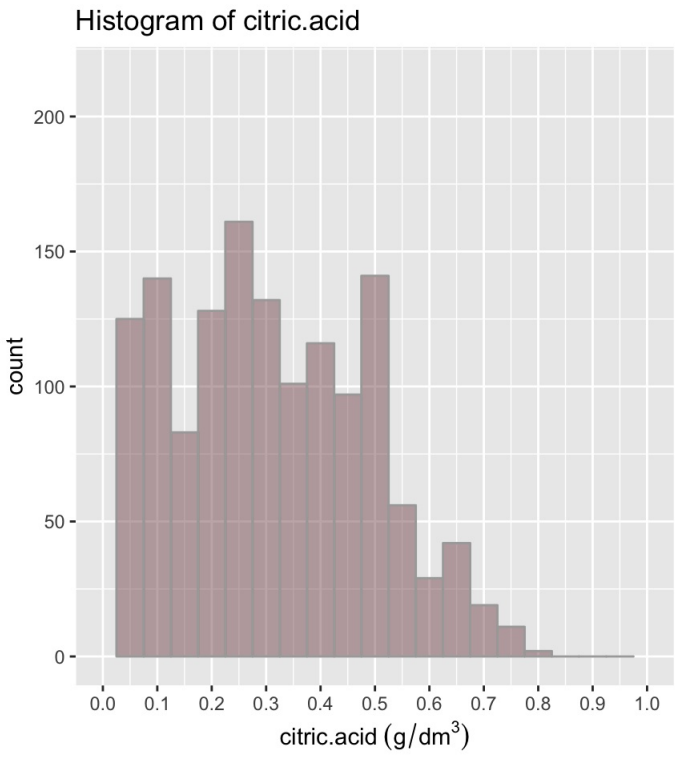




The distribution of Volatile acidity is slightly like bimodal distribution with two peaks around 0.4 and 0.6.

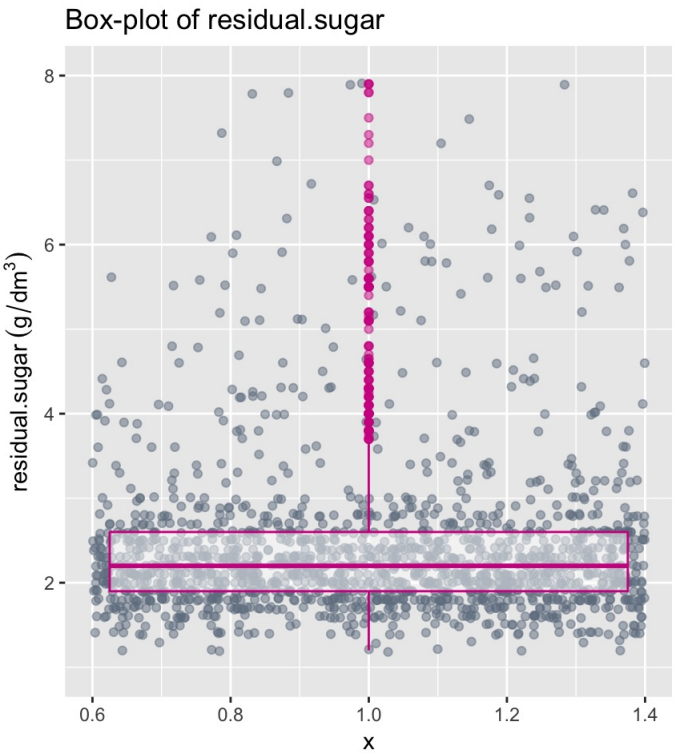
For citric.acid

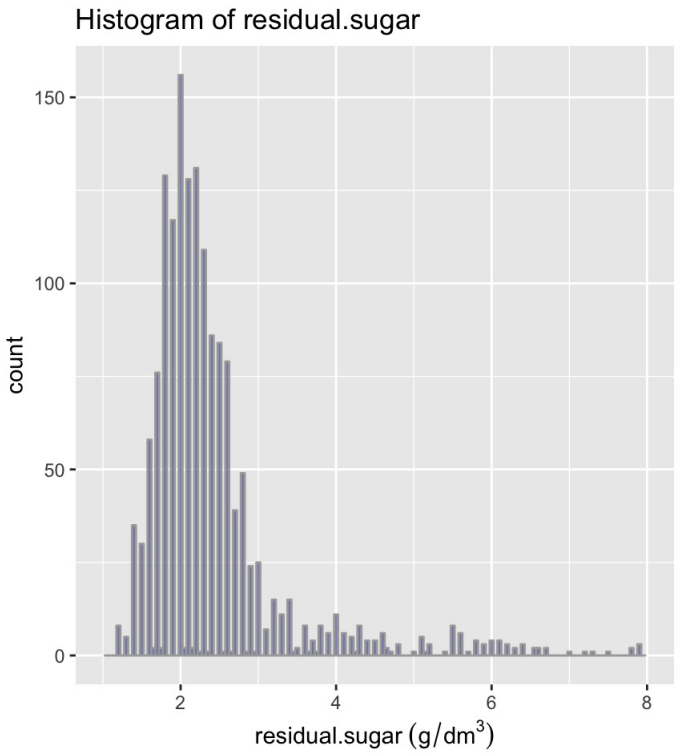




The distribution of citric.acid is slightly skewed right with median and mean are 0.26 and 0.271, respectively. Not much outliers.

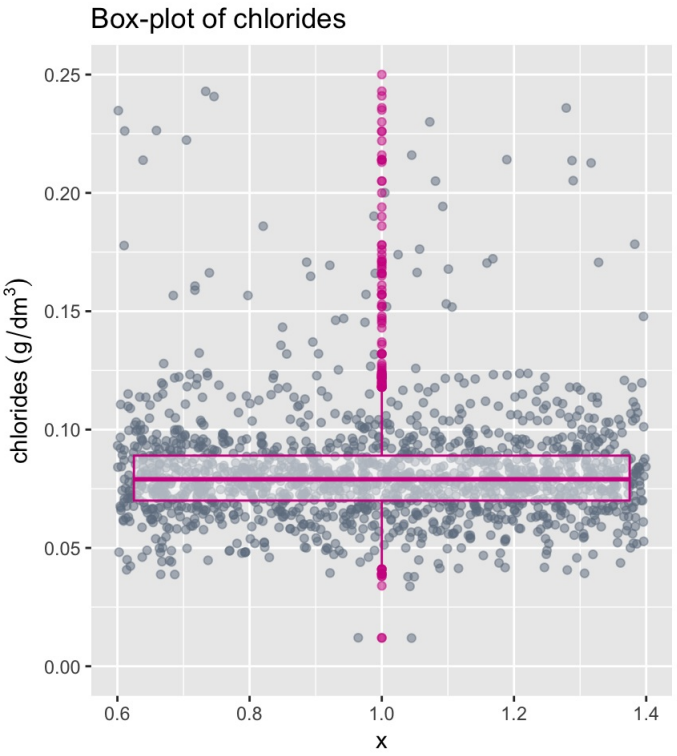
For residual.sugar

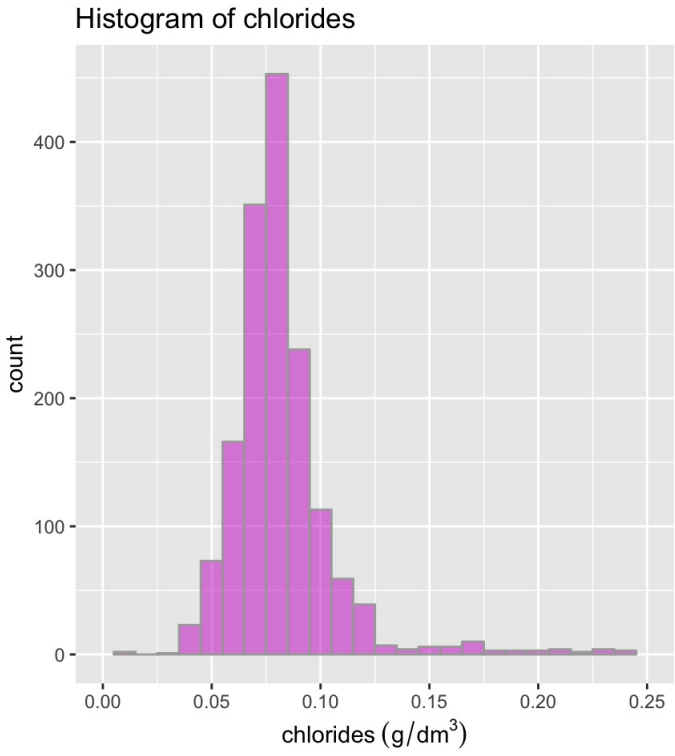




The distribution of Residual Sugar is skewed right with median at 2.2. Many outliers present at the higher ranges.

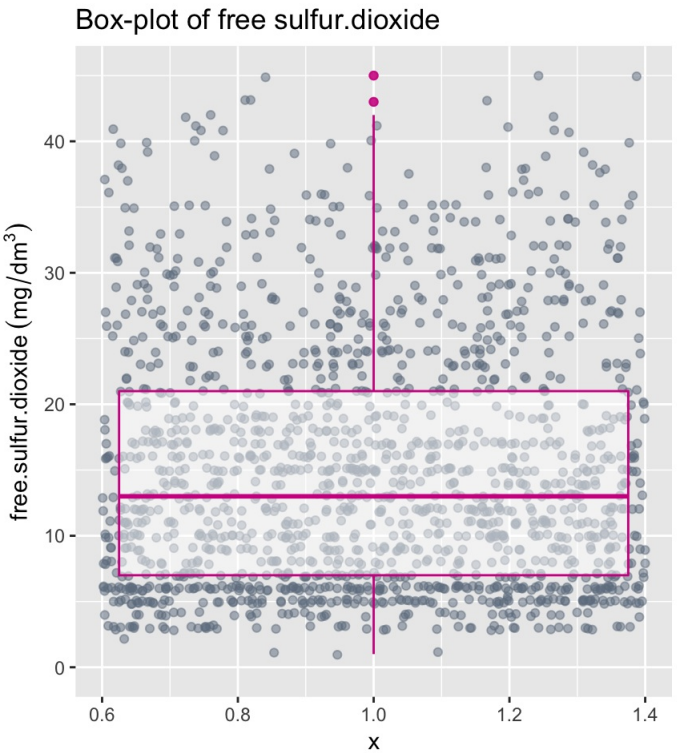
For chlorides

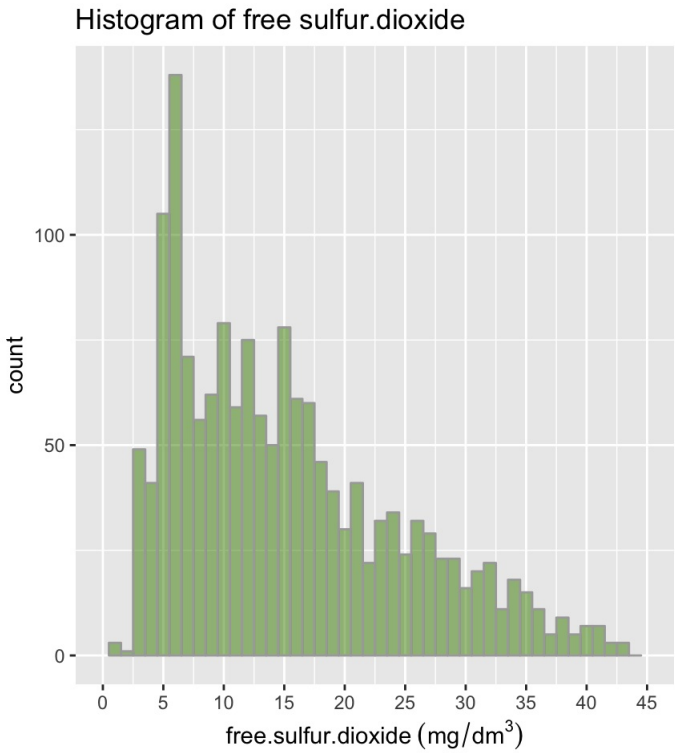




For Chlorides, it is skewed right with median at 0.079. We got lots of outliers at the higher ranges.

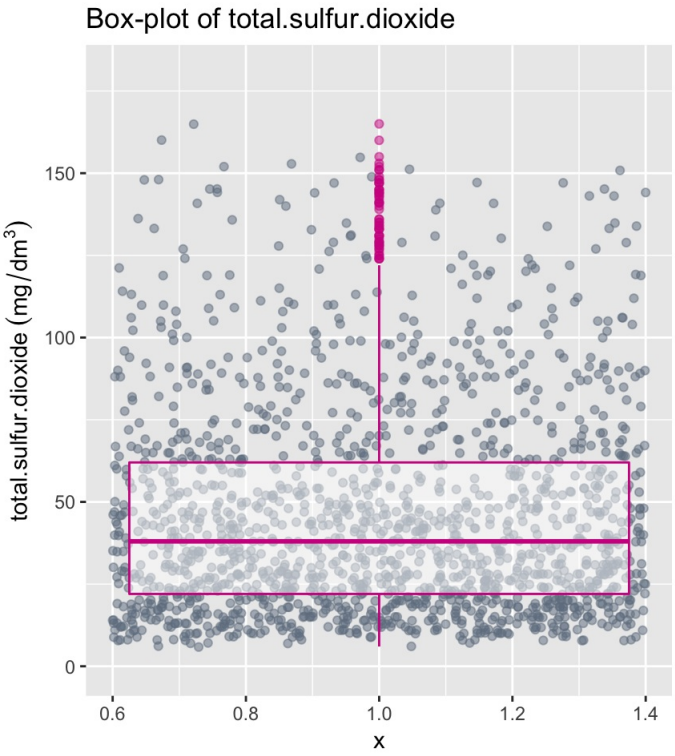
For free sulfur.dioxide

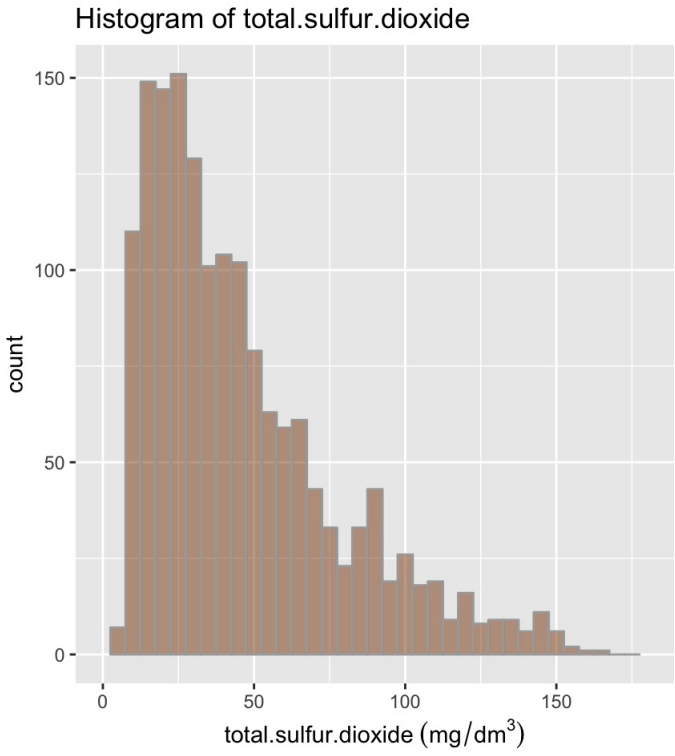




For Free Sulphur Dioxide, it also follows skewed right with long tailed. There is a peak at 7 and some outliers in higher range.

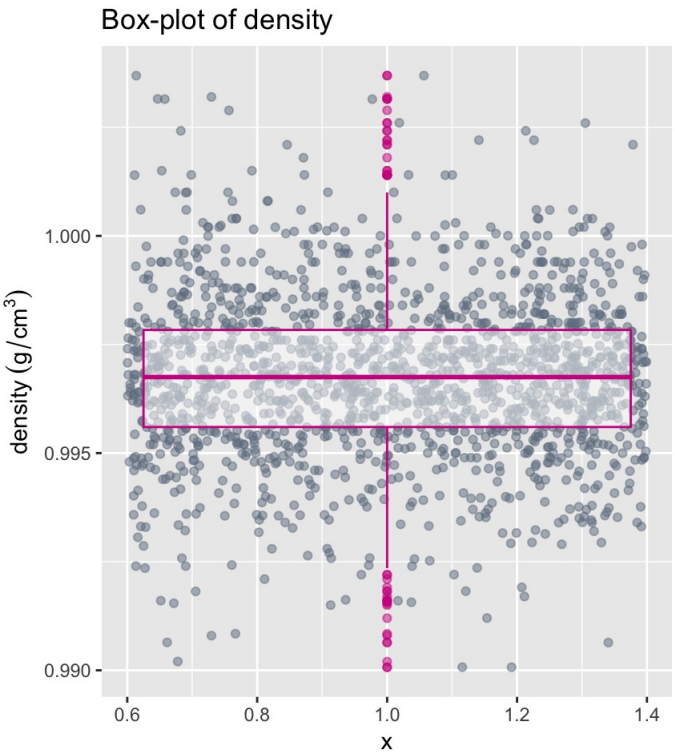
For total sulfur.dioxide

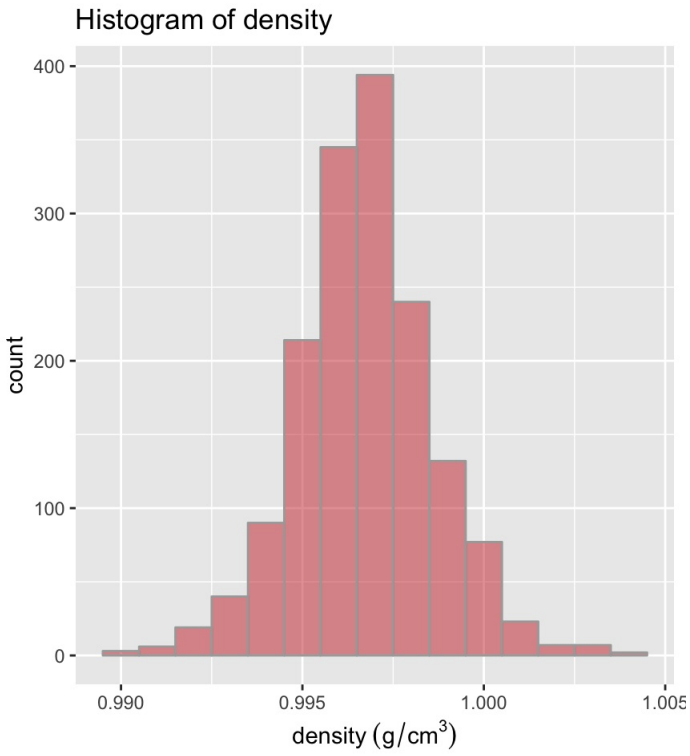




For Total Sulphur Dioxide, it also follows skewed right with long tailed. There are some outliers in higher range.

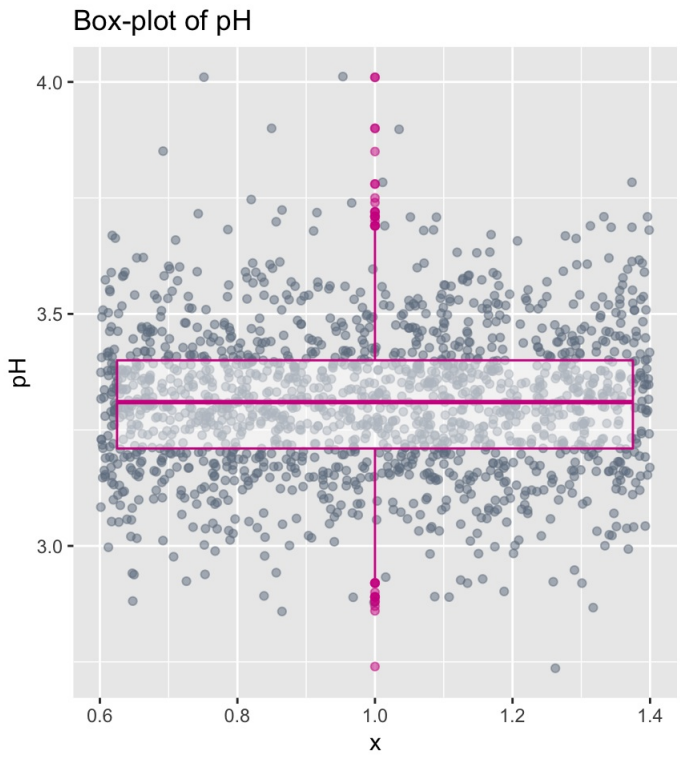
For density

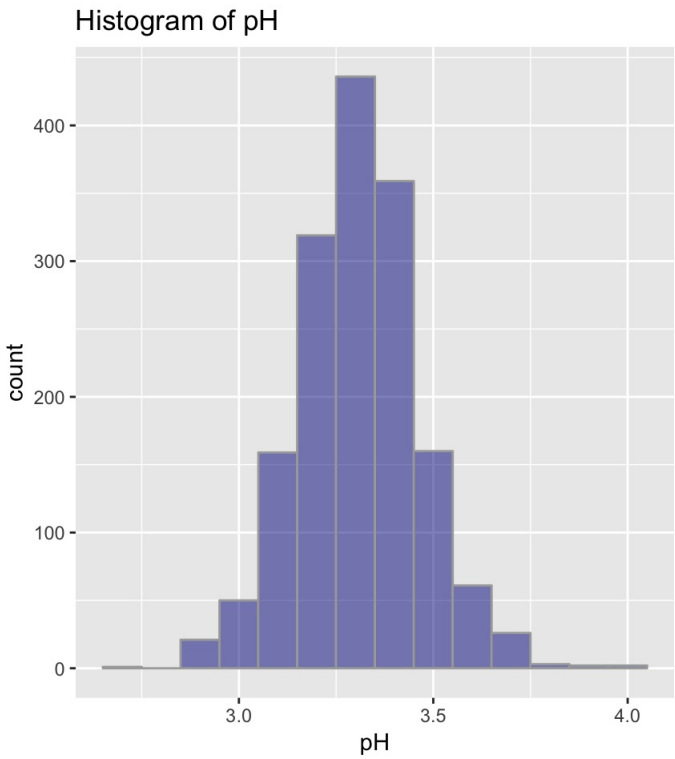




For Density variable, it occured Normal Distribution with several outliers.

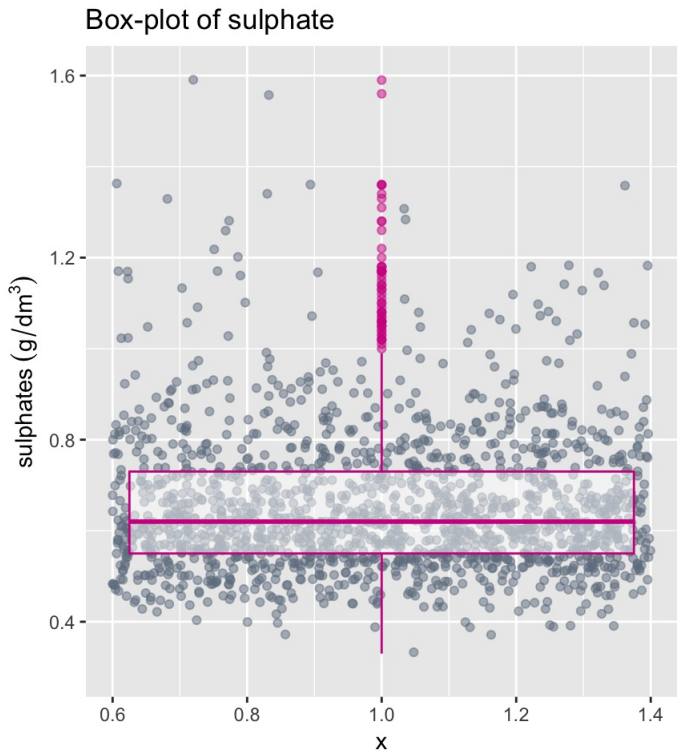
For pH

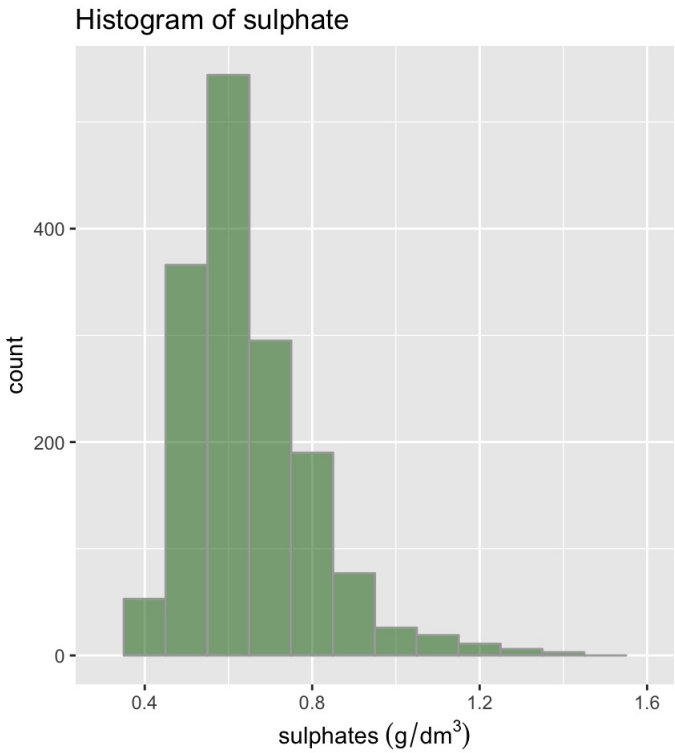




For pH variable, it also occured Normal Distribution with several outliers.

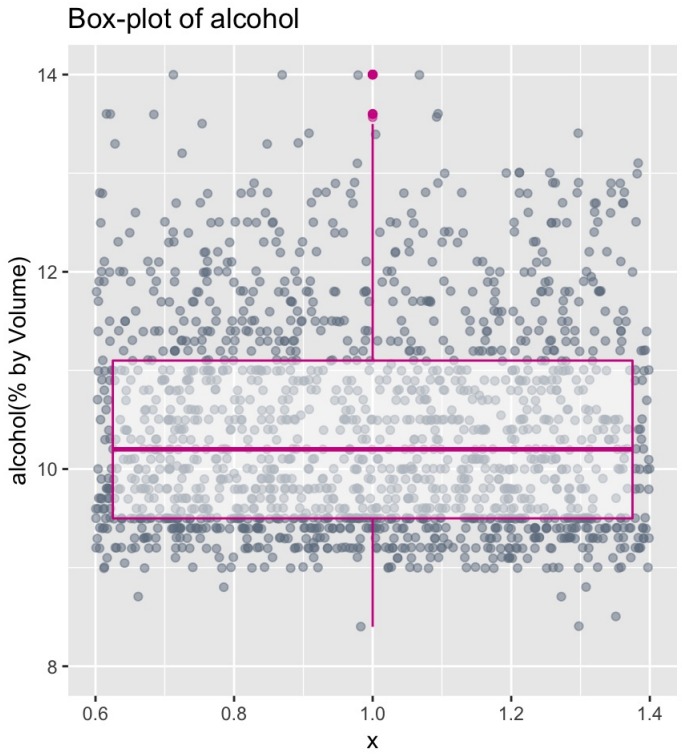
For sulphate

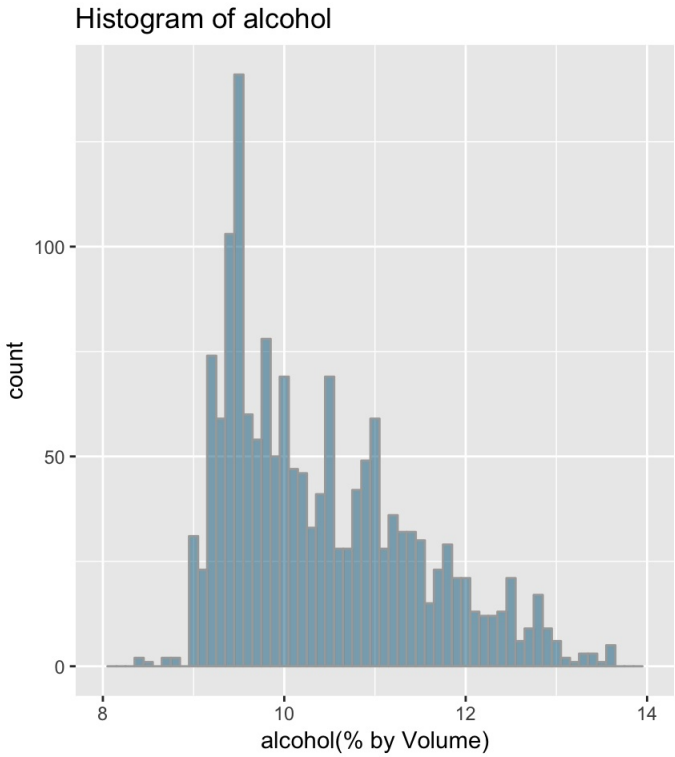




For Total Sulphate, it follows skewed right distribution. There are some outliers in higher range.

For Alcohol





Alcohol also follows a skewed right distribution but the skewness is less than Chlorides or Residual Sugars.

Correlation table

```
##
## ------------------------------------------------------------------------
##                        fixed.acidity   volatile.acidity   citric.acid
## ------------------------- --------------- ------------------ -------------
##     **fixed.acidity**             1              -0.2561         **0.6717**
##
##    **volatile.acidity**         -0.2561             1            **-0.5525**
##
##      **citric.acid**           **0.6717**       **-0.5525**          1
##
##     **residual.sugar**          0.1148           0.001918          0.1436
##
##       **chlorides**             0.09371           0.0613           0.2038
##
##  **free.sulfur.dioxide**        -0.1538          -0.0105          -0.06098
##
##  **total.sulfur.dioxide**       -0.1132          0.07647           0.03553
##
##        **density**              **0.668**         0.02203         **0.3649**
##
##           **pH**               **-0.683**         0.2349          **-0.5419**
##
##       **sulphates**              0.183            -0.261          **0.3128**
##
##        **alcohol**             -0.06167          -0.2023           0.1099
##
##        **quality**              0.1241          **-0.3906**         0.2264
## ------------------------------------------------------------------------
##
## Table: Table continues below
##
##
## ------------------------------------------------------------------------
##                        residual.sugar   chlorides    free.sulfur.dioxide
## ------------------------- --------------- ------------ --------------------
##     **fixed.acidity**           0.1148        0.09371           -0.1538
##
##    **volatile.acidity**        0.001918        0.0613           -0.0105
##
##      **citric.acid**           0.1436          0.2038          -0.06098
##
```

residual.sugar 1 0.05561 0.187
##
chlorides 0.05561 1 0.005562
##
free.sulfur.dioxide 0.187 0.005562 1
##
total.sulfur.dioxide 0.203 0.0474 **0.6677**
##
density **0.3553** 0.2006 -0.02195
##
pH -0.08565 -0.265 0.07038
##
sulphates 0.005527 **0.3713** 0.05166
##
alcohol 0.04208 -0.2211 -0.06941
##
quality 0.01373 -0.1289 -0.05066

Table: Table continues below
##
##
---
total.sulfur.dioxide density pH
---
fixed.acidity -0.1132 **0.668** **-0.683**
##
volatile.acidity 0.07647 0.02203 0.2349
##
citric.acid 0.03553 **0.3649** **-0.5419**
##
residual.sugar 0.203 **0.3553** -0.08565
##
chlorides 0.0474 0.2006 -0.265
##
free.sulfur.dioxide **0.6677** -0.02195 0.07038
##
total.sulfur.dioxide 1 0.07127 -0.06649
##
density 0.07127 1 **-0.3417**
##
pH -0.06649 **-0.3417** 1
##
sulphates 0.04295 0.1485 -0.1966
##
alcohol -0.2057 **-0.4962** 0.2056
##
quality -0.1851 -0.1749 -0.05773

Table: Table continues below
##
##
---
sulphates alcohol quality
---
fixed.acidity 0.183 -0.06167 0.1241
##
volatile.acidity -0.261 -0.2023 **-0.3906**
##
citric.acid **0.3128** 0.1099 0.2264
##
residual.sugar 0.005527 0.04208 0.01373
##
chlorides **0.3713** -0.2211 -0.1289
##
free.sulfur.dioxide 0.05166 -0.06941 -0.05066
##
total.sulfur.dioxide 0.04295 -0.2057 -0.1851
##
density 0.1485 **-0.4962** -0.1749
##
pH 0.1966 0.2056 0.05773

```
##          **pH**              -0.1966       0.2056       -0.05773
##
##       **sulphates**             1           0.09359       0.2514
##
##        **alcohol**            0.09359          1         **0.4762**
##
##        **quality**             0.2514     **0.4762**         1
## ---------------------------------------------------------------------
```

• Fixed acidity is strongly positive correlated with critic acid and density, while negative correlated with pH.

• Volatile acidity is strongly negative correlated with critic acid and quality.

• Citric acid is strongly positive correlated with density and sulphates, while while negative correlated with pH.

• Residual sugar is strongly positive correlated with density.

• Chlorides is strongly positive correlated with sulphates.

• Free sulfur dioxide is strongly positive correlated with total sulfur dioxide.

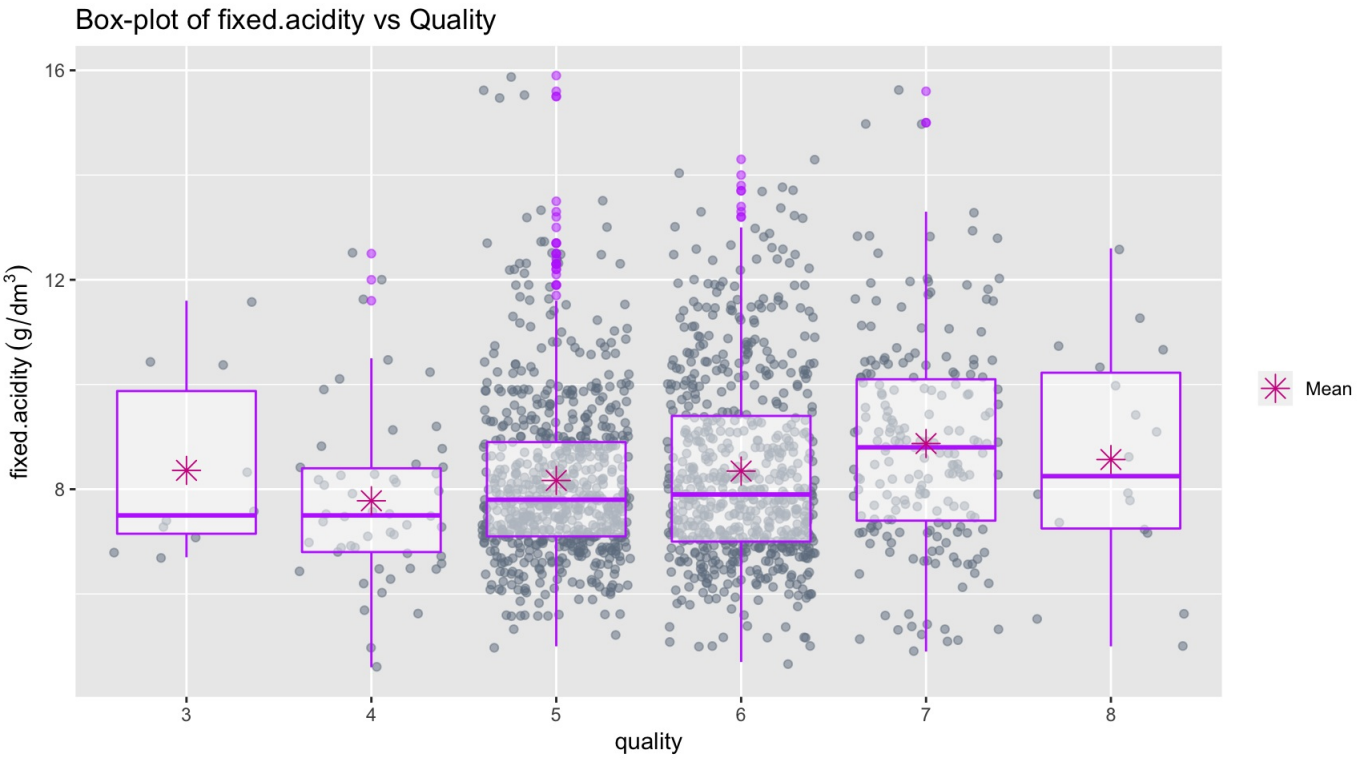
• Density is strongly negative correlated with pH and alcohol.

• Alcohol is strongly negative correlated with density.

• Quality is strongly positive correlated with alcohol.

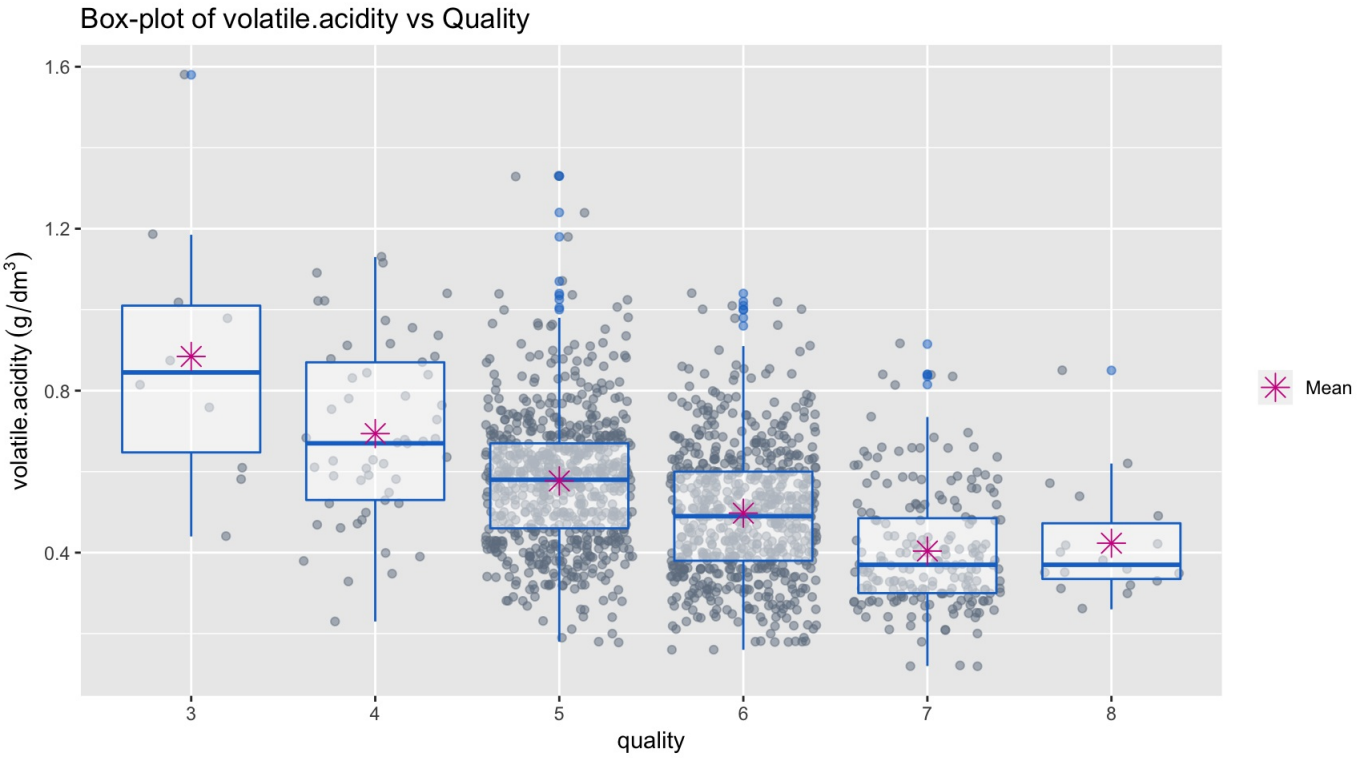
Boxplot Plots

For fixed acidity



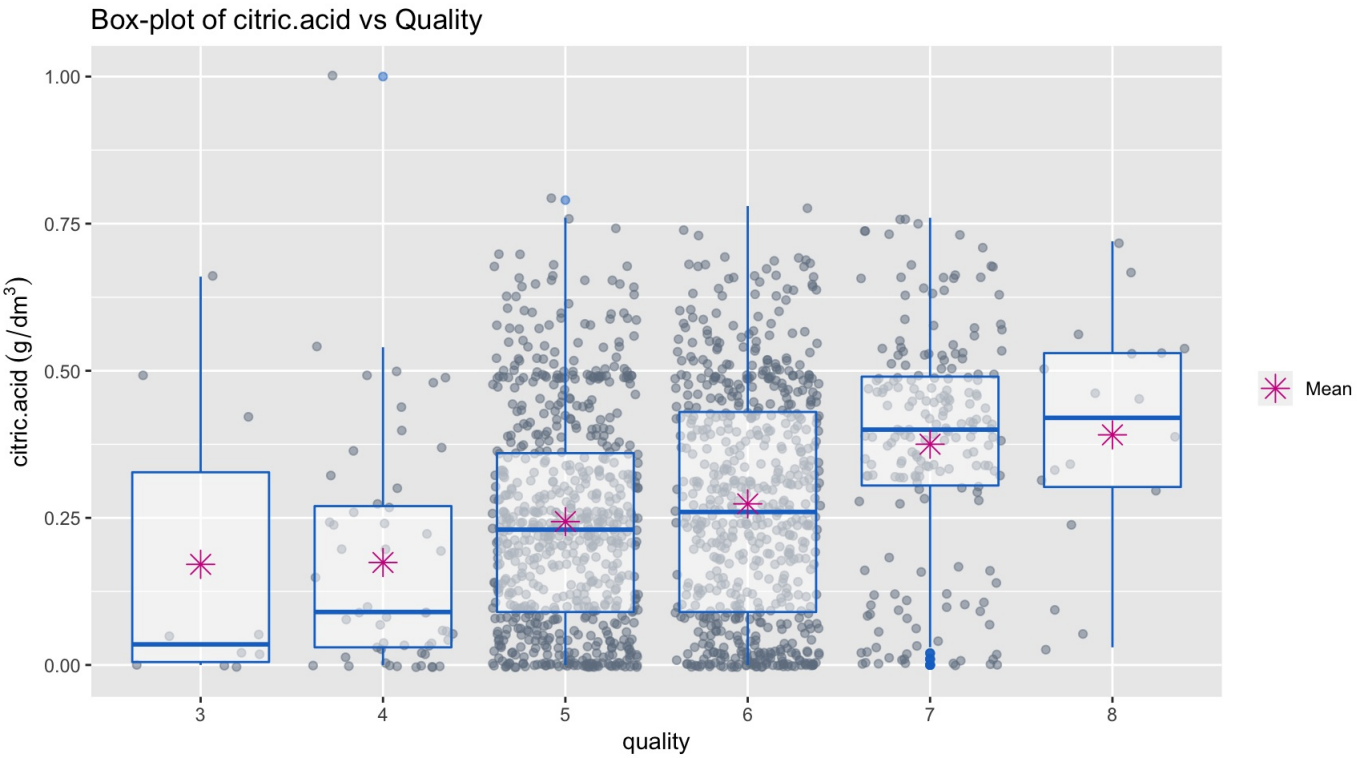
Fixed Acidity does not find statistically significant on Quality. The mean and median of fixed acidity almost remains unchanged with increase in quality.

For volatile acidity



Volatile Acidity has negative effect on Quality. Lesser Volatile Acidity, the poor in quality of wine.

For citric.acid

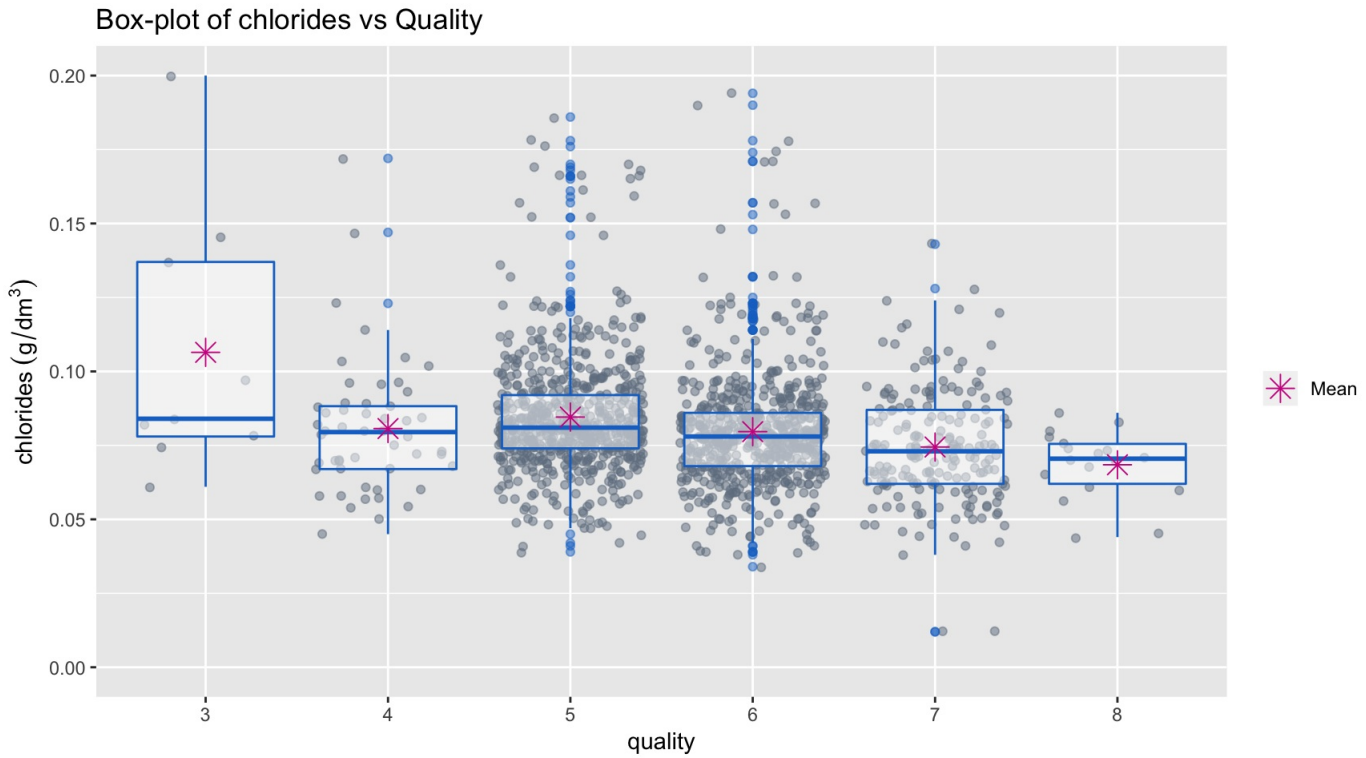


Citric acid has positive effect on Quality. The more Citric acid, the better in quality of wine.

For residual.sugar

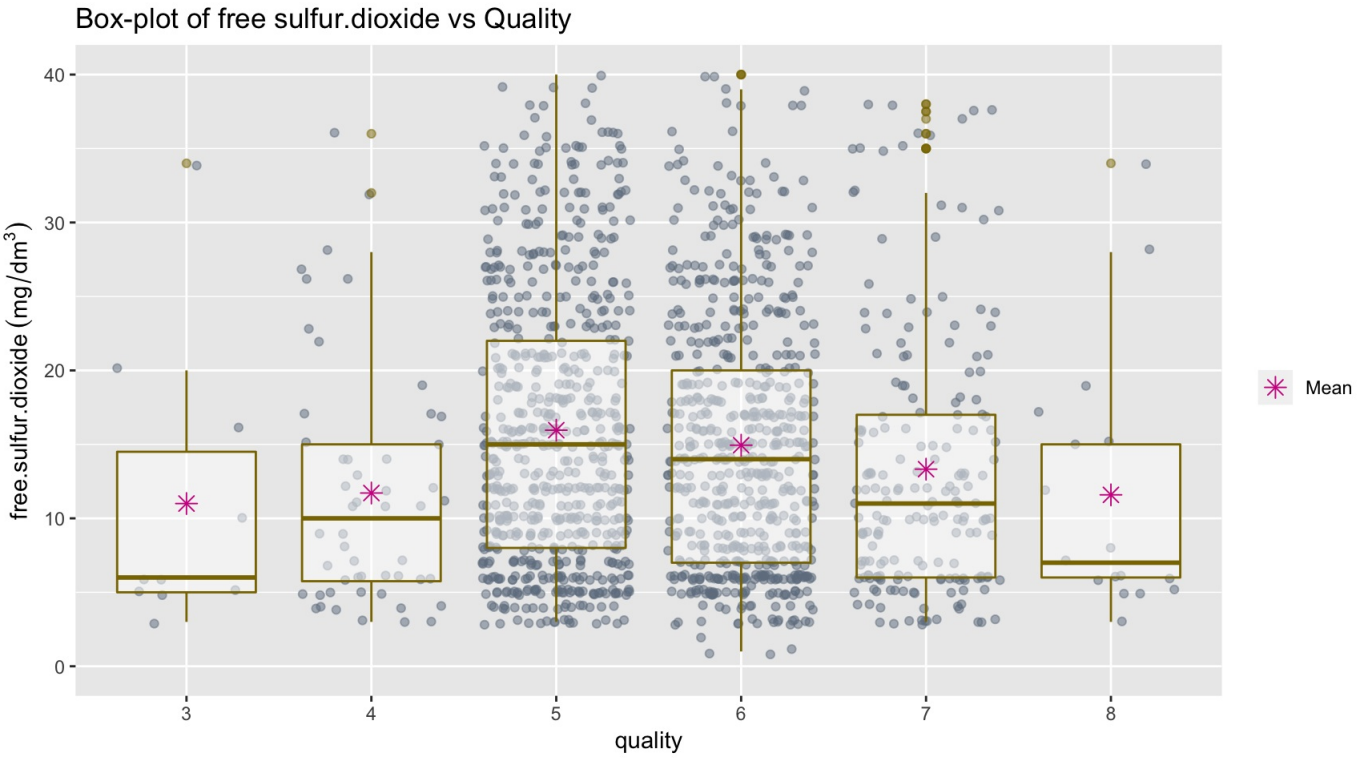
Residual.sugar does not find statistically significant on Quality. The mean and median of Residual.sugar almost remains unchanged regarding to quality.

For chlorides



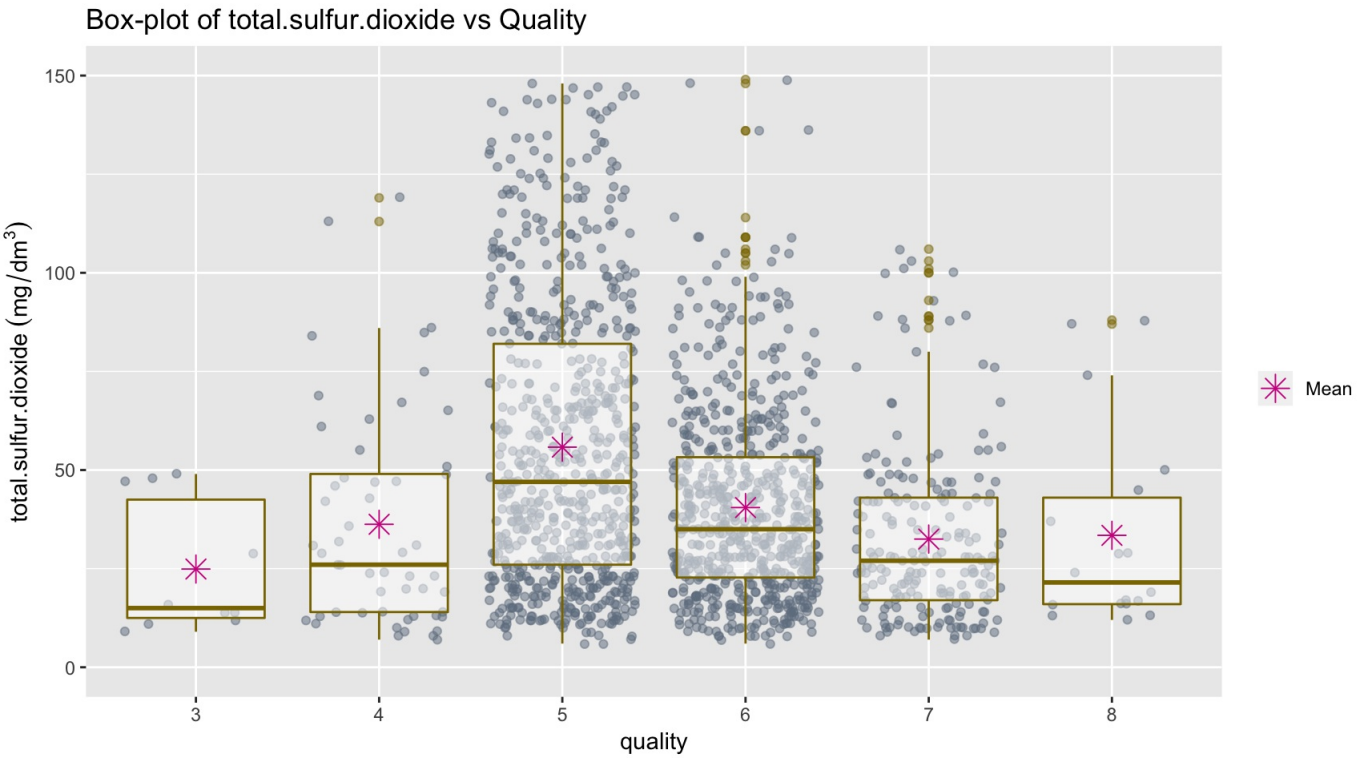
Chlorides is only a week correlated with quality, it seems slightly that the lower Chloride, the better quality of wines.

For free sulfur.dioxide



We found that moderate concentration of free sulfur dioxide can find in moderate quality of wine. While the poor and better quality of wines have found in lower concentration of free sulfur dioxide.

For total sulfur.dioxide

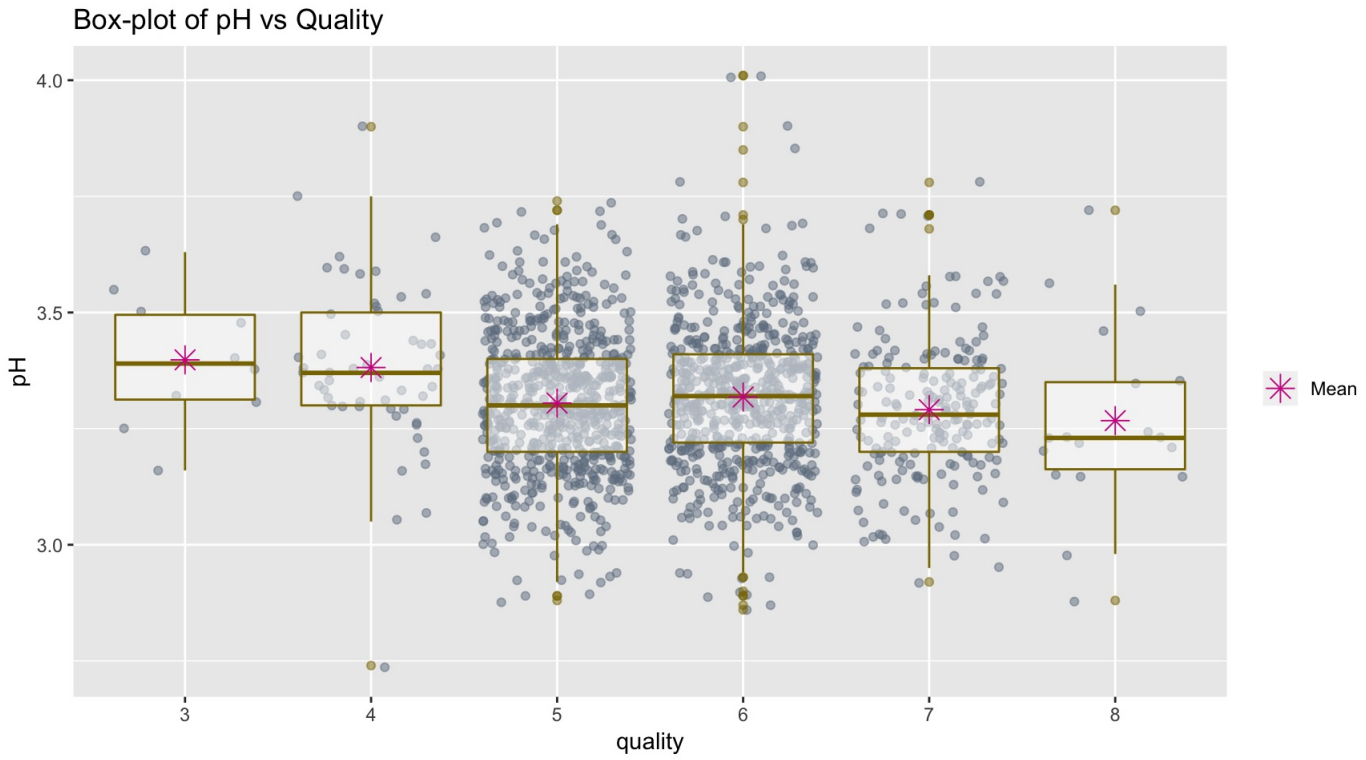


We found similar pattern in free sulfur dioxide, moderate concentration of total.sulfur.dioxide can find in moderate quality of wine. While the poor and better quality of wines have found in lower concentration of total.sulfur.dioxide.

For density

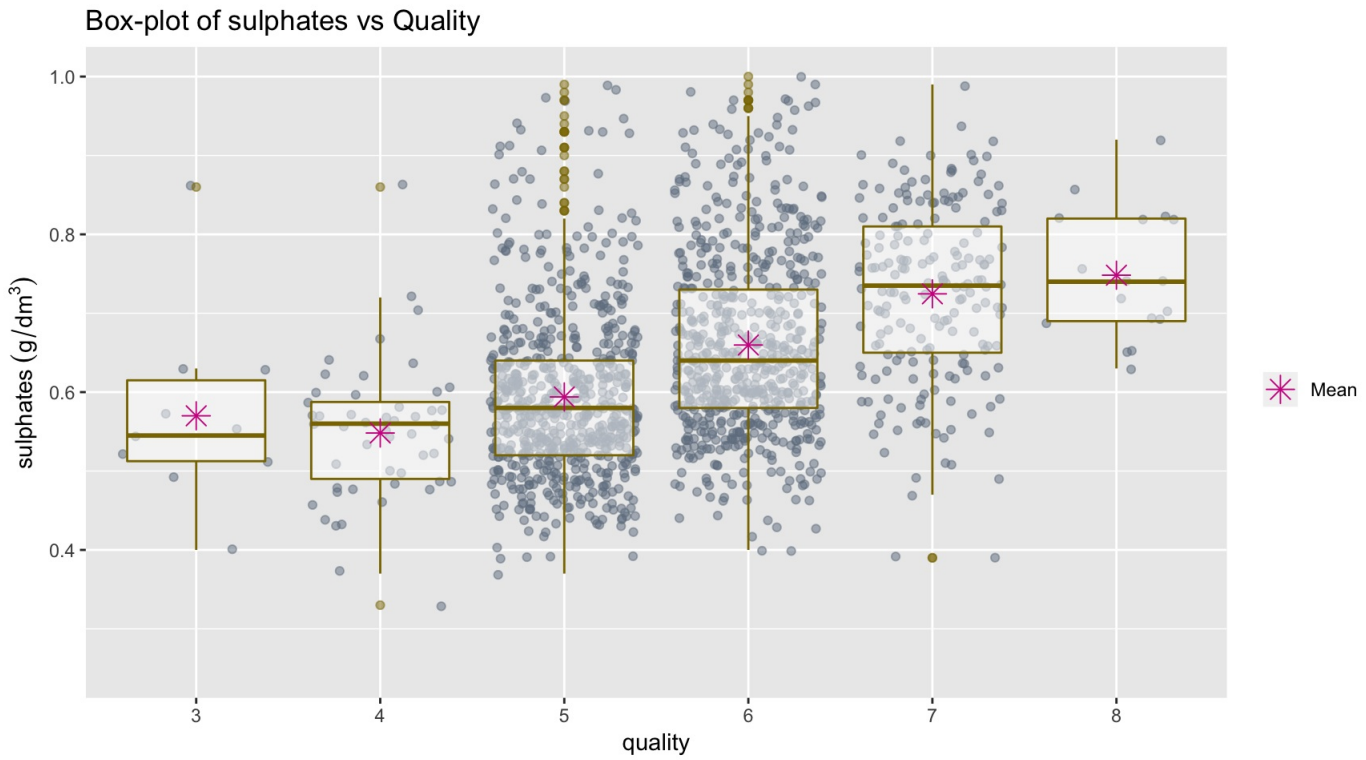
We found that better wines seems to have lower densities.

For pH



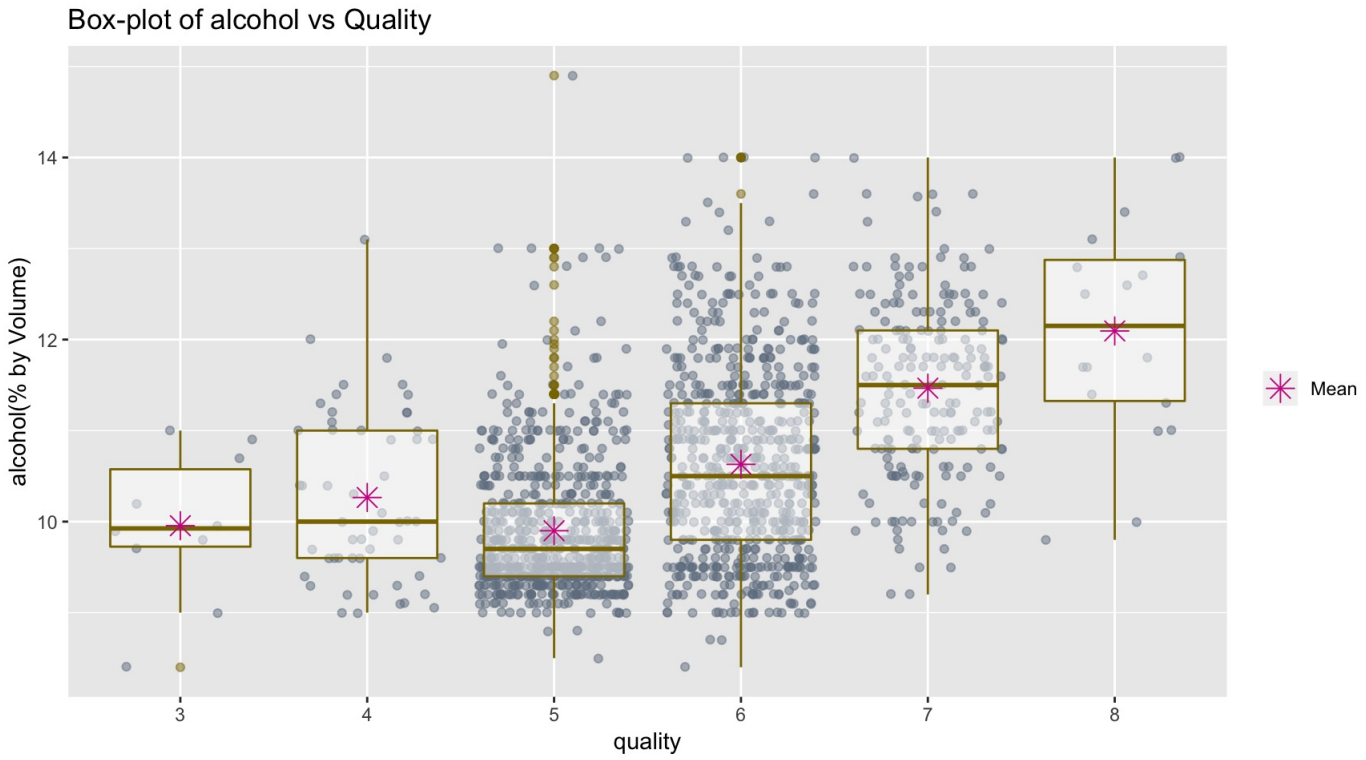
It seems better quality of wines have lower pH, i.e: more acidic.

For sulphates



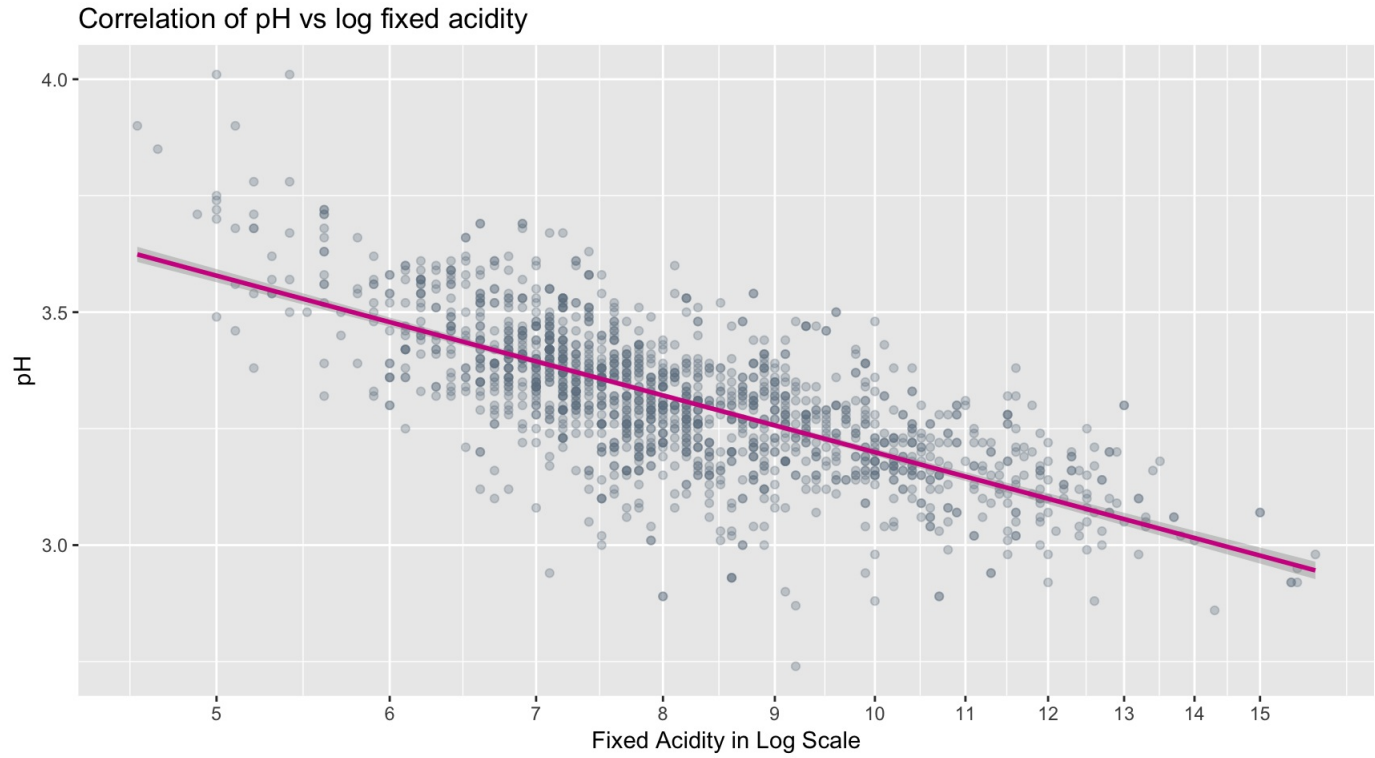
Sulphates has positive effect on Quality. The more sulphates, the better in quality of wine.

For alcohol



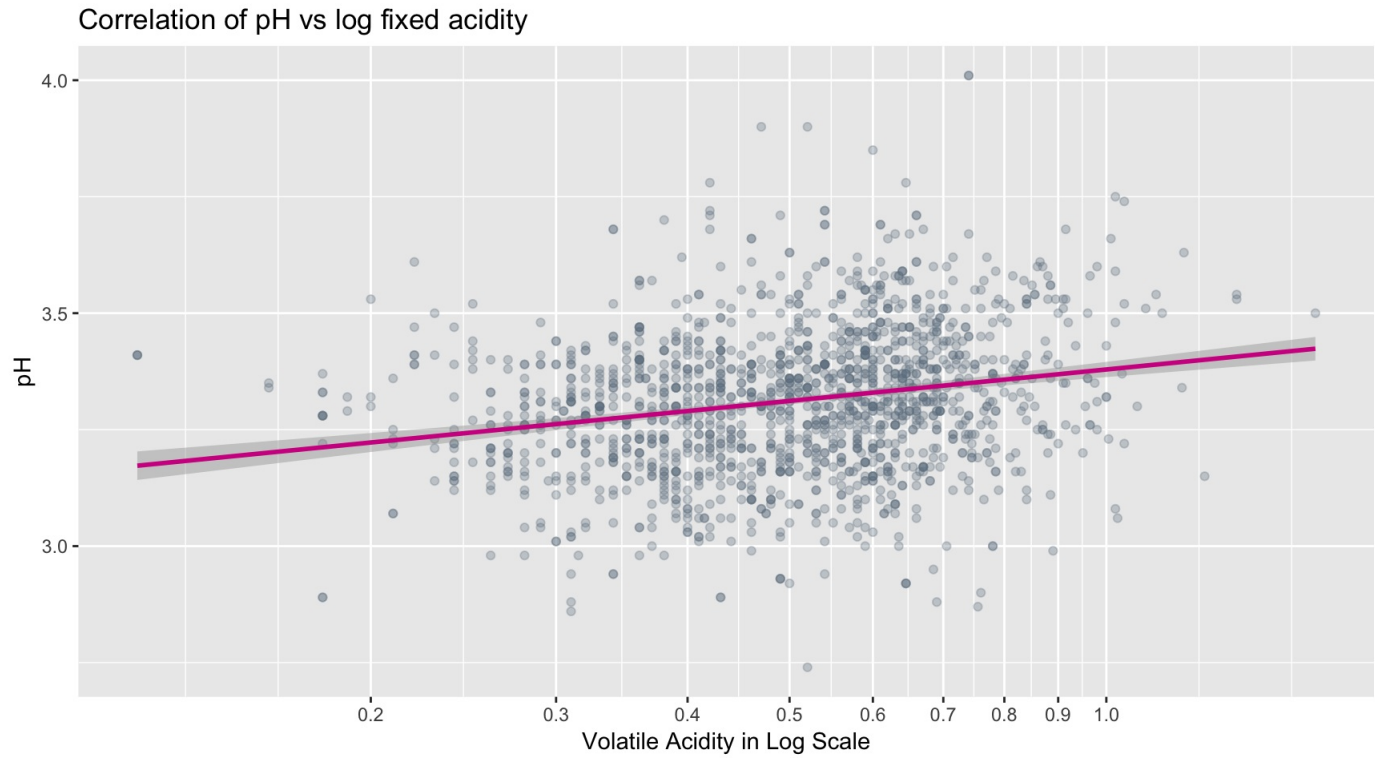
Alcohol has positive effect on Quality. The more alcohol, the better in quality of wine.

Correlation of pH vs log fixed acidity



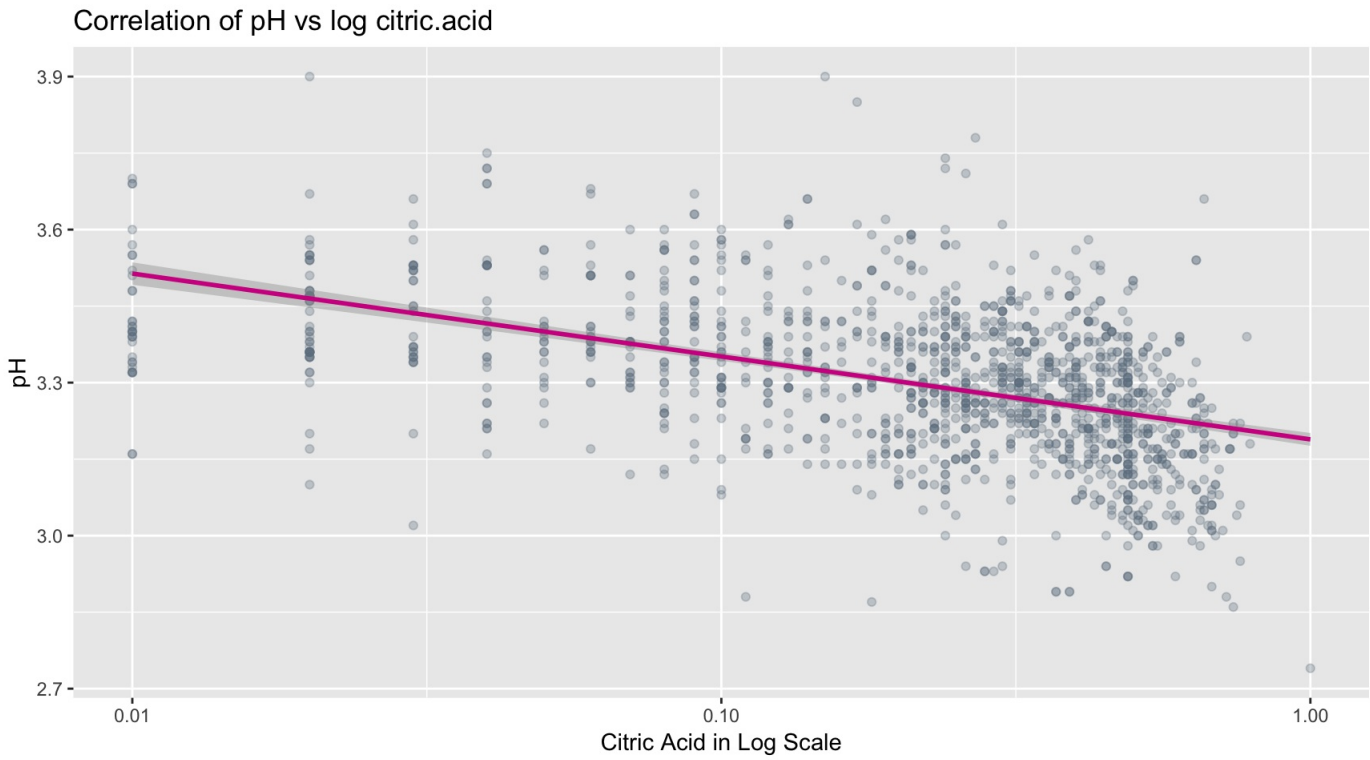
We found that negative correlation with pH and log fixed acidity. The lower pH (ie:more acidic) the more fixed acidity.

Correlation of pH vs log volatile.acidity



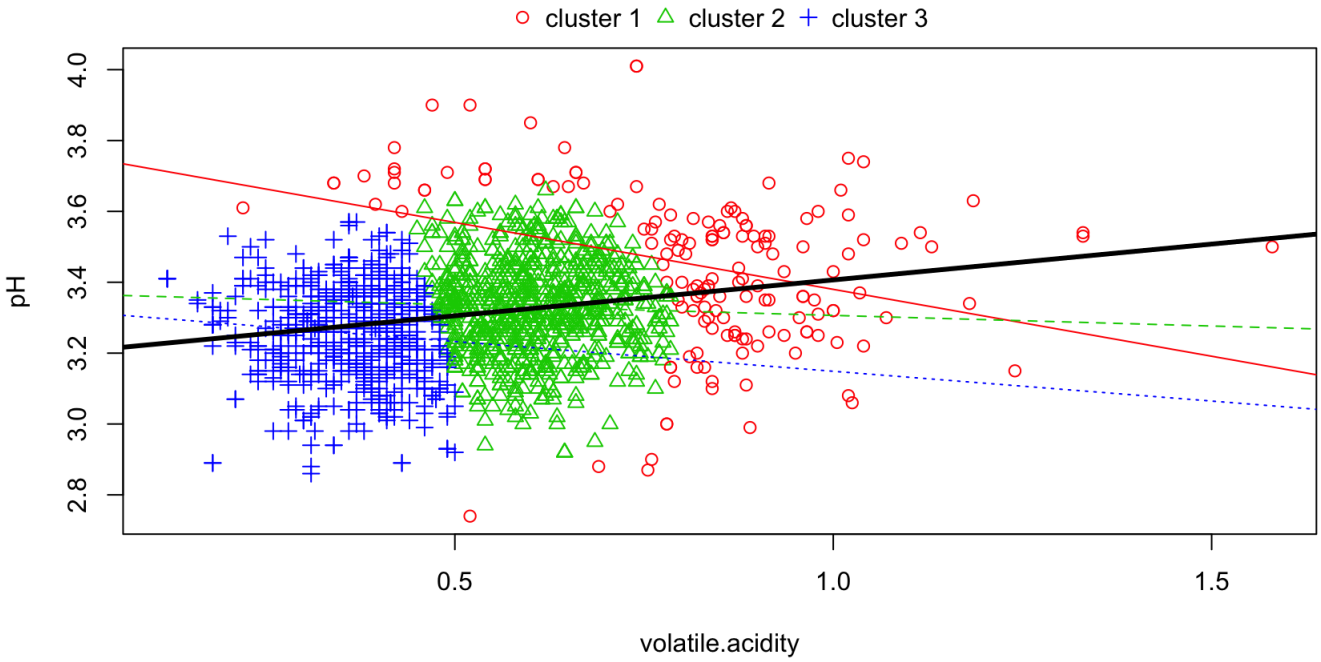
We found that positive correlation with pH and log volatile acidity. The lower pH (ie:more acidic), lesser fixed acidity.

Correlation of pH vs log citric.acid



We found that negative correlation with pH and log citric.acid. The lower pH (ie:more acidic) the more citric.acid.

From above three correlation plots, we found that negative correlation with pH vs log fixed acidity and log citric.acid. Whereas, positive correlation with pH and log volatile acidity.So they may have Simpson's Paradox effect on it.



Simpson's paradox is present which is responsible for the trend reversal of pH vs Volatile Acid. We first clustered the data into 3 segments and calculated the regression coefficient. We found that there is indeed a sign reversal. It is due to a lurking variable is present.

As it is not included as an explanatory or response variable in the analysis but can affect the interpretation of relationships between variables. Therefore it is falsely identify a strong positive relationship between pH and volatile acidicity and finally hide the true relationship.

Regression model of alcohol

```
##
## Call:
## lm(formula = as.numeric(quality) ~ alcohol, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8442 -0.4112 -0.1690  0.5166  2.5888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.12503    0.17471  -0.716    0.474
## alcohol      0.36084    0.01668  21.639   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7104 on 1597 degrees of freedom
## Multiple R-squared:  0.2267, Adjusted R-squared:  0.2263
## F-statistic: 468.3 on 1 and 1597 DF,  p-value: < 2.2e-16
```

From regression model, r-square is 0.23, which means alcohol is only contribute 22% of the Wine quality. Therefore, there are other variables to contribute there. So I will figure them out as below.

```
##        fixed.acidity     volatile.acidity          citric.acid
##           0.12405165          -0.39055778           0.22637251
## log10.residual.sugar      log10.chlordies  free.sulfur.dioxide
##           0.02353331          -0.17613996          -0.05065606
## total.sulfur.dioxide              density                   pH
##          -0.18510029          -0.17491923          -0.05773139
##      log10.sulphates              alcohol
##           0.30864193           0.47616632
```

From the above result, we found that alcohol, log10.sulphates, volatile.acidity and citric.acid have strong correlation with wine quality.

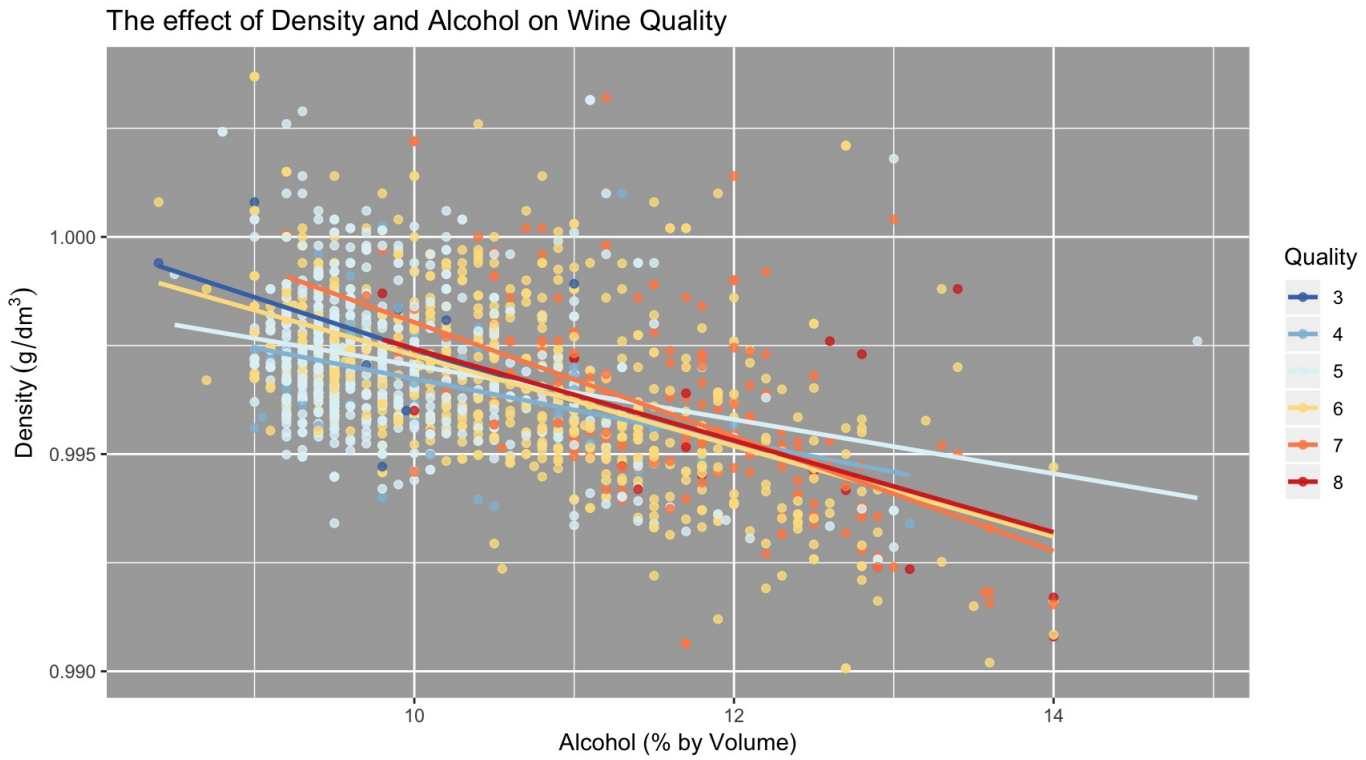
Observation of Bivariate Plots

1. Fixed Acidity is not statistically significant on wine quality.
2. Volatile Acidity has a strong negative effect on wine quality.
3. Citric.acid has a positive effect on wine quality. Better quality of wines have higher concentration of Citric Acid.
4. Very weak corrleation between residual.sugar and quality of wines.
5. A weak correlation between log10.chlordies and quality of wines., but lower Chloride seems to produce better quality of wines.
6. Very weak corrleation between free.sulfur.dioxide and quality of wines.
7. A weak correlation between total.sulfur.dioxide and quality of wines., but lower total.sulfur.dioxide seems to produce better quality of wines.
8. Better quality of wines seem to have lower densities.
9. Better quality of wines seem to have lower pH (ie: more acidic).
10. Strong positive corrleation between log10.sulphates and quality of wines.
11. Strong positive corrleation between alcohol and quality of wines.
12. Simpson’s paradox is present which is responsible for the trend reversal of pH vs Volatile Acid.

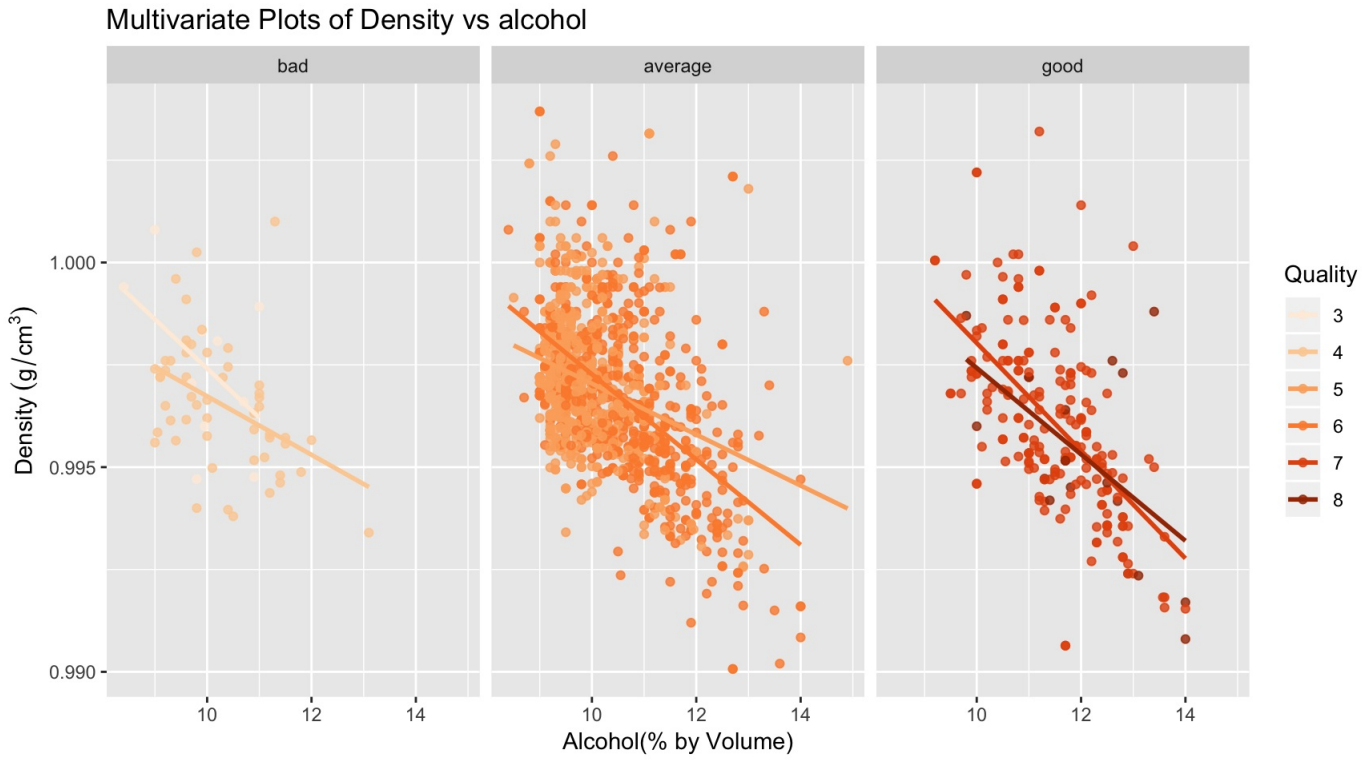
Multivariate Plots

From the above result, we found that alcohol is strongly significant on wine quality. Although alcohol is only contribute 22% of the Wine quality. So we try to investigate other contributors on wine quality.

Effect on quality of wine with density and alcohol

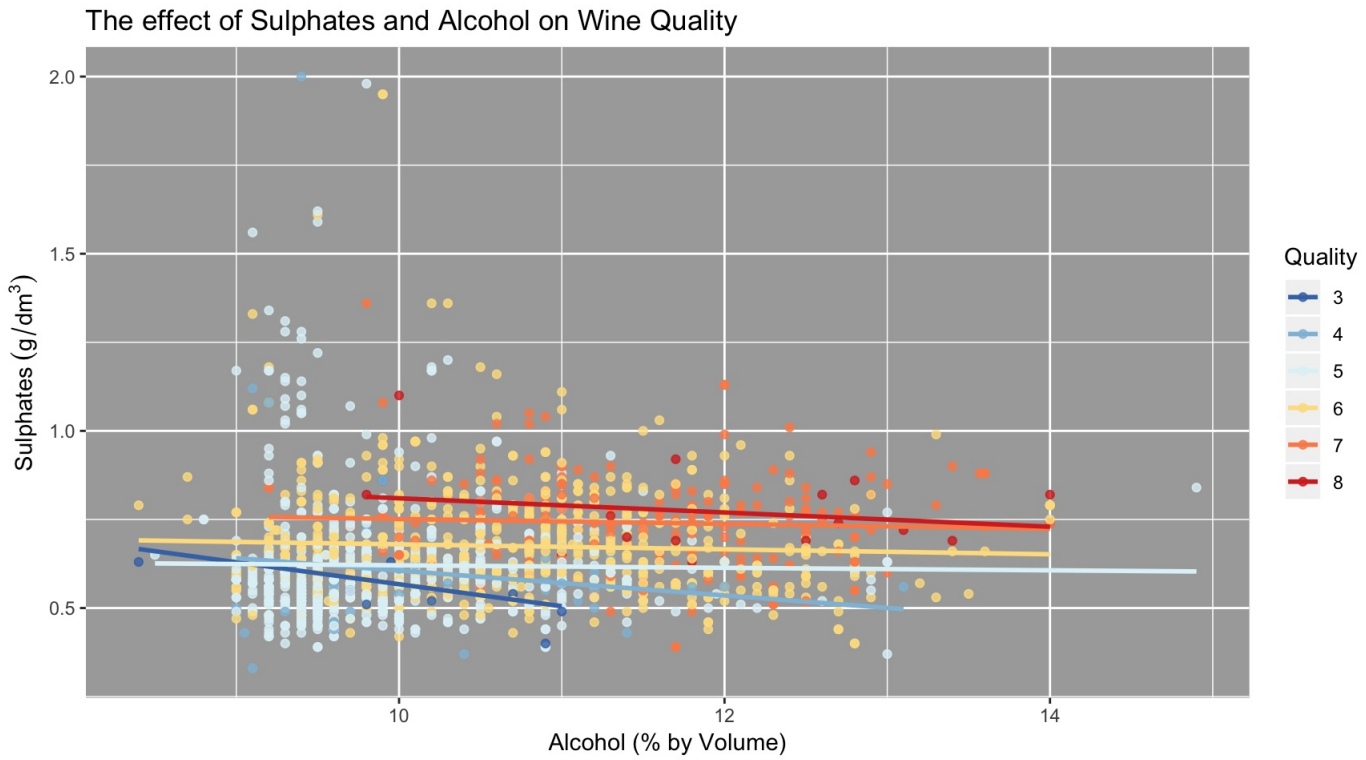


The density has a negative effect on alcohol.

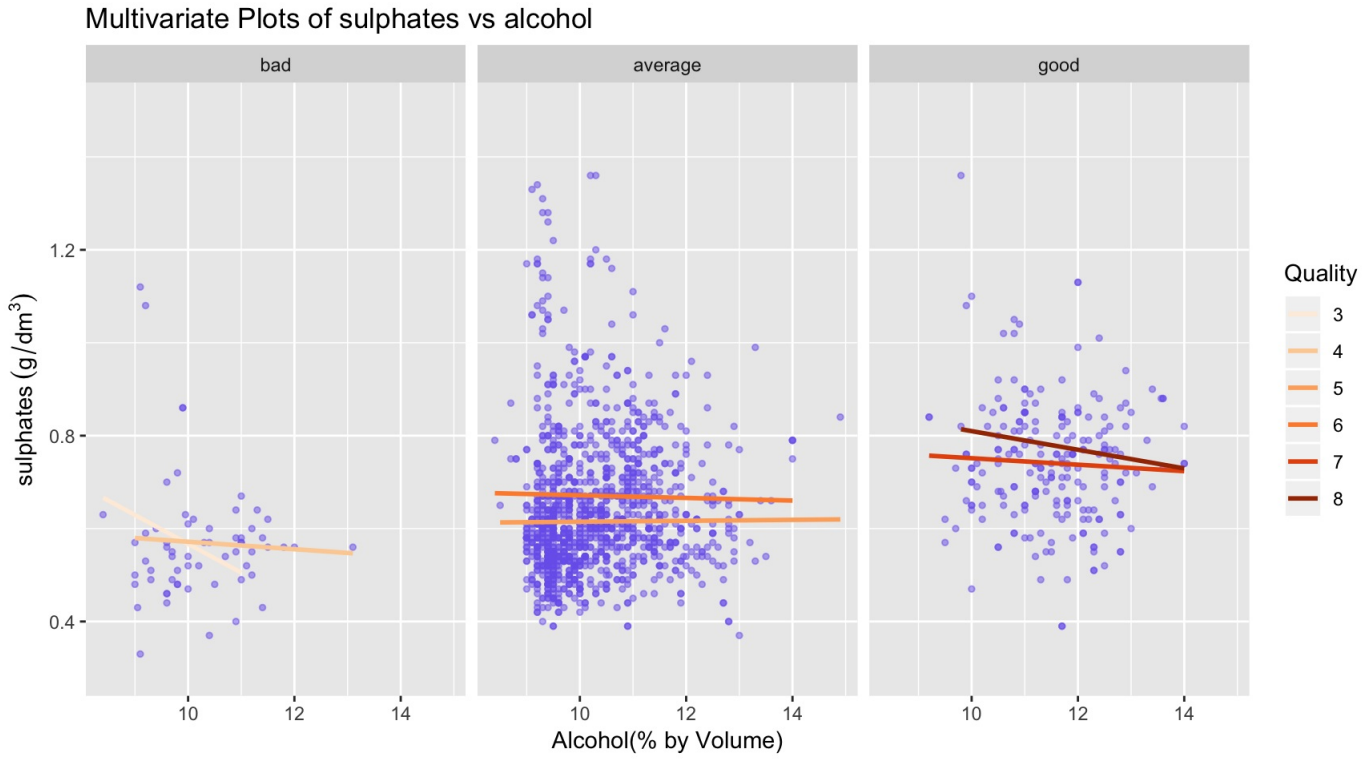


By keeping Alcohol constant, Density does not play siginficant effect on quality of wine. Therefore the density has a negative effect on wine quality was due to alcohol.

Effect on quality of wine with sulphates and alcohol

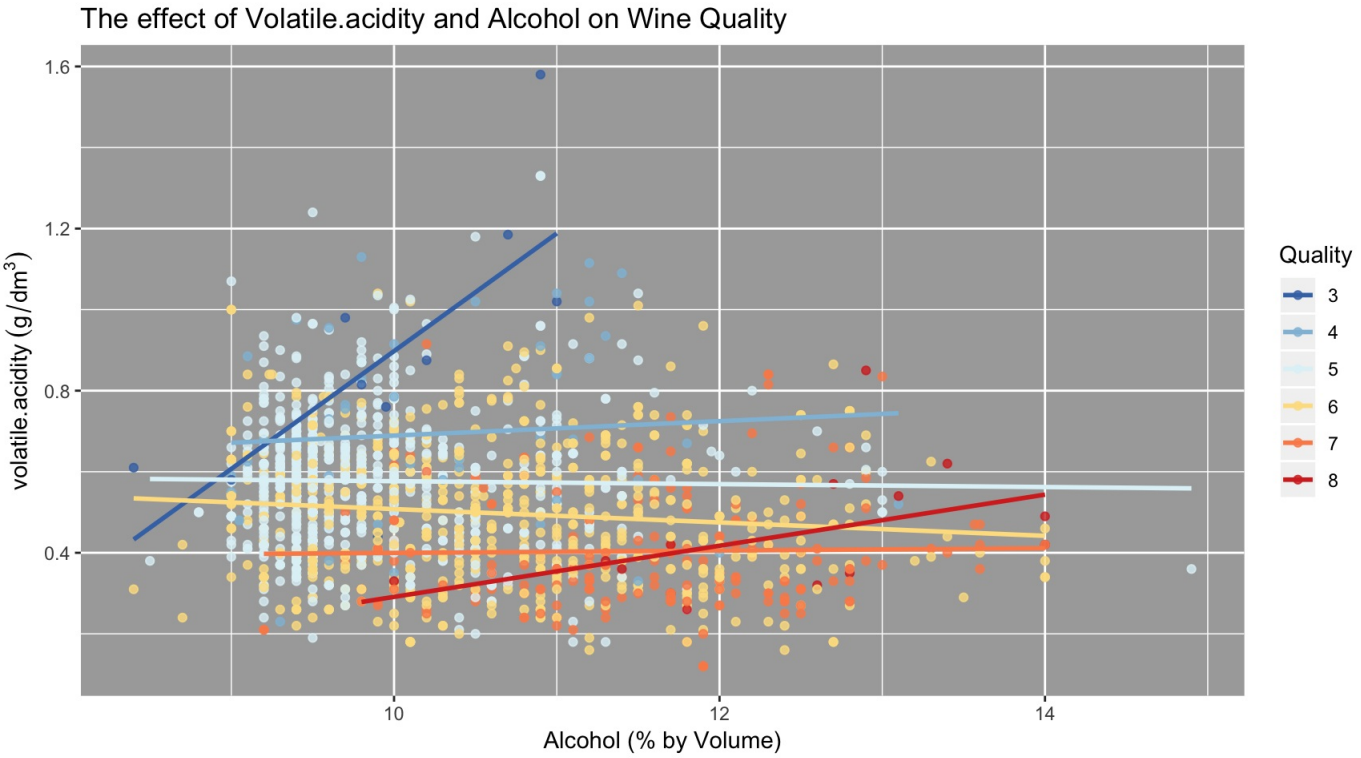


In this plot, sulphates has not played a significant effect on alcohol.

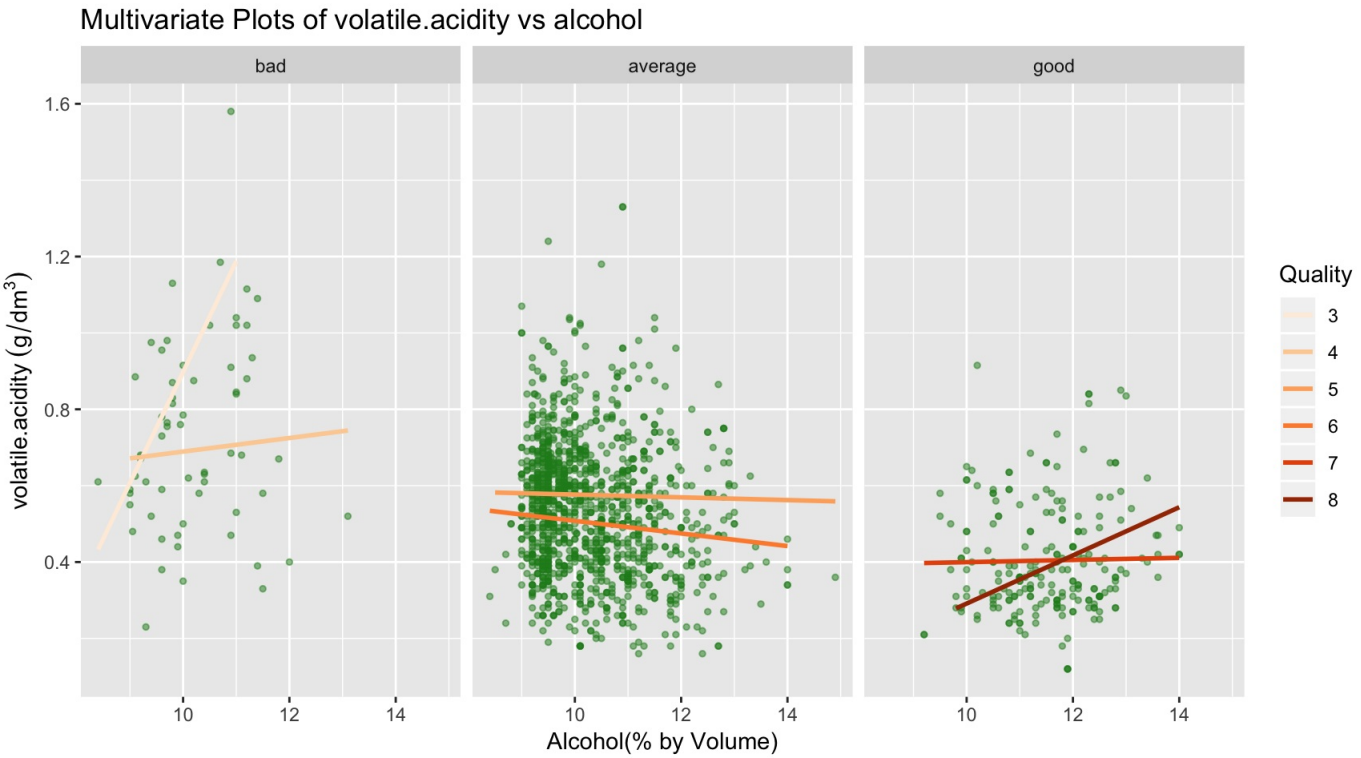


By keeping alcohol constant, we found that the higher sulphates, the better quality of wine.

Effect on quality of wine with volatile.acidity and alcohol

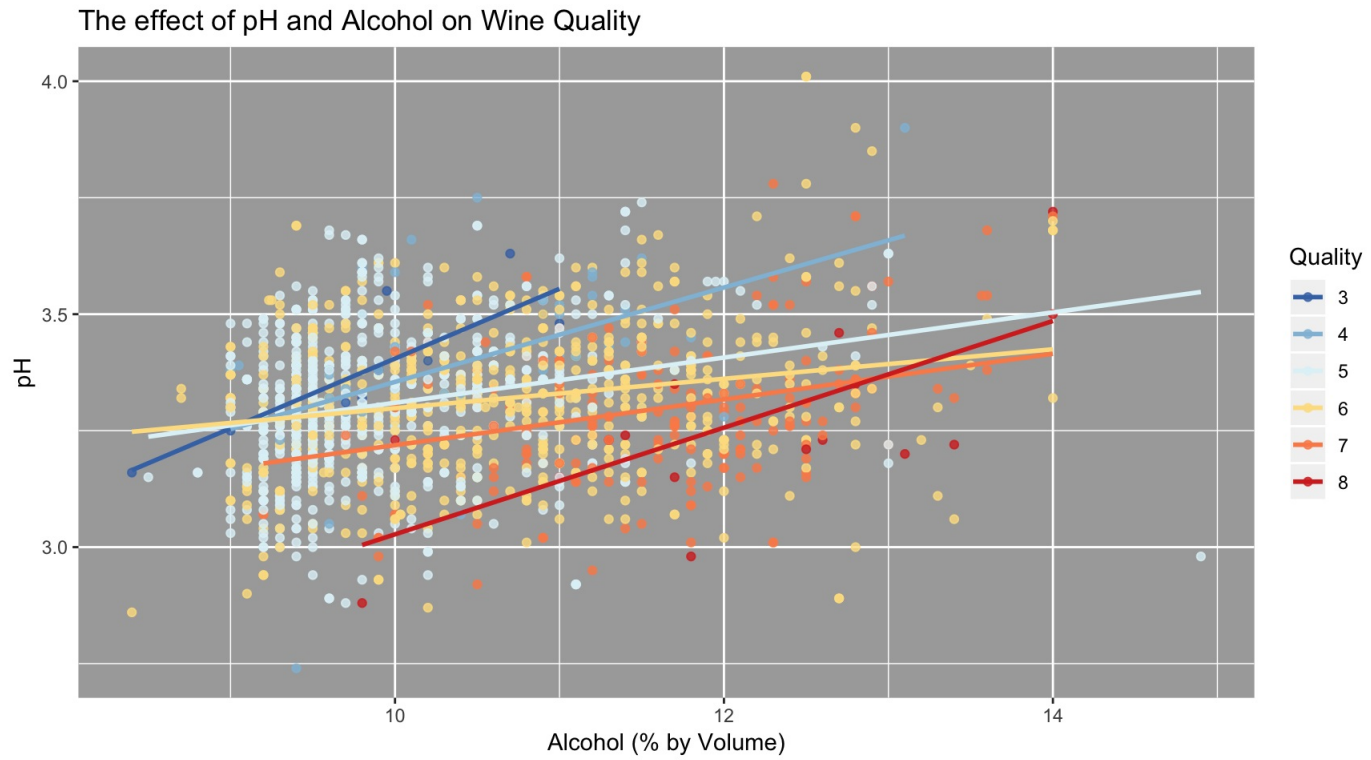


In this plot, volatile.acidity has only played a significant effect on high AND low quality of wine, as their volatile.acidity increase with alcohol content increase.

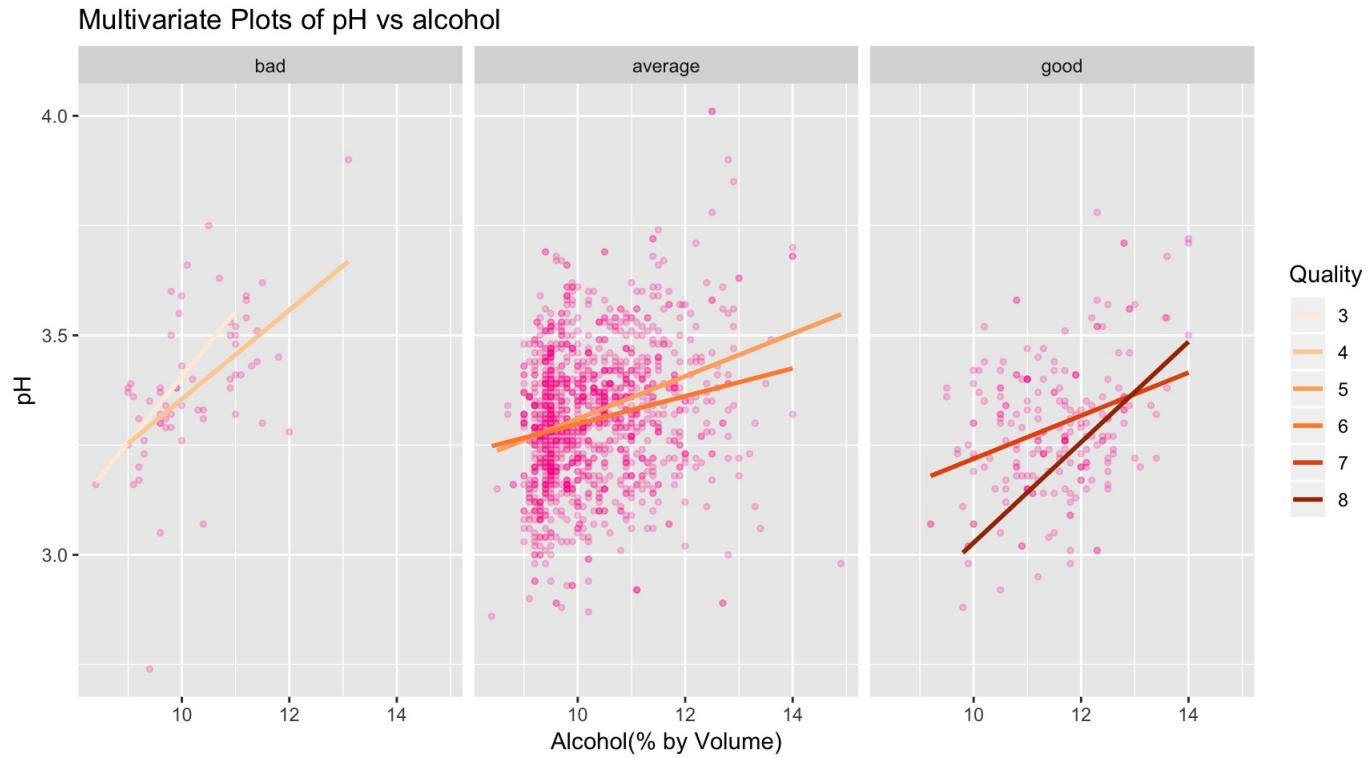


For Volatile acid which has opposite effect that the lower concentration of volatile acid with higher concentration of alcohol, better the quality of wine.

Effect on quality of wine with pH and alcohol

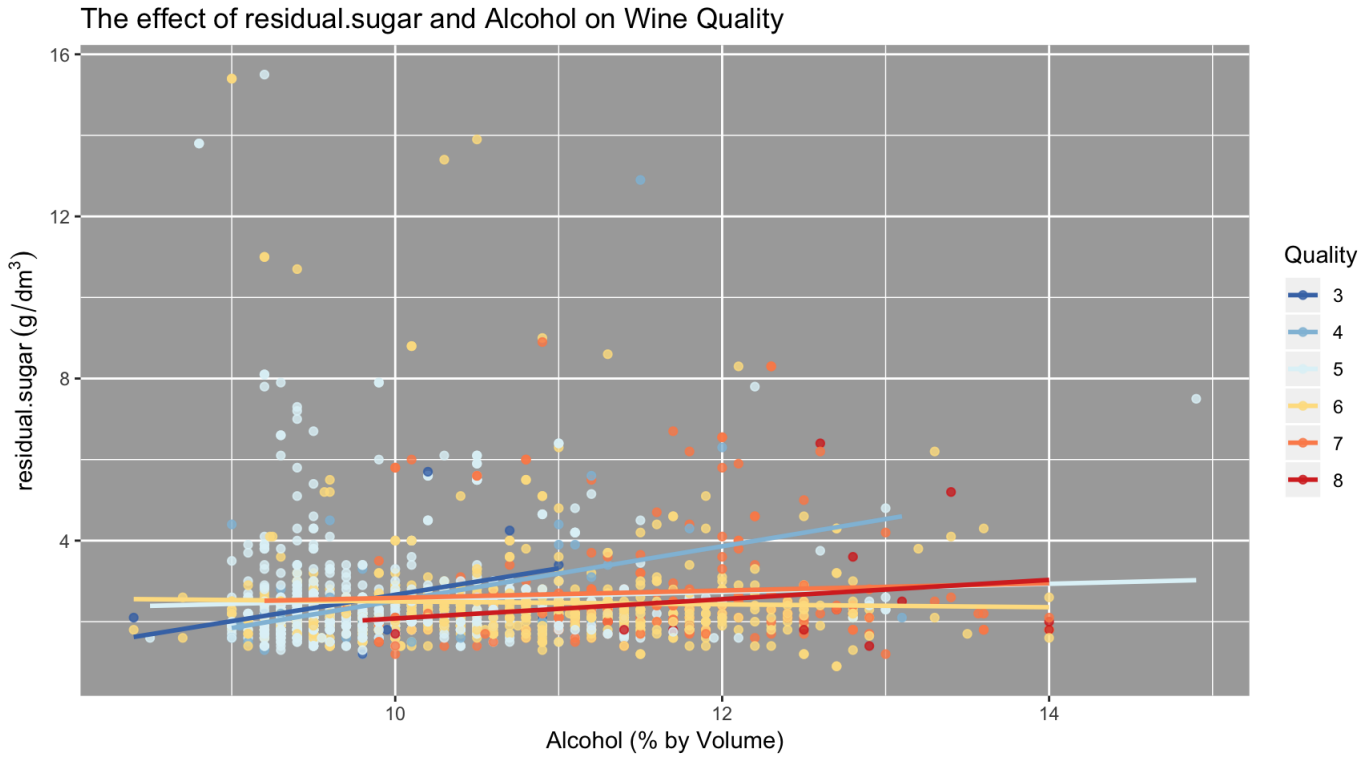


pH has positive effect on alcohol.

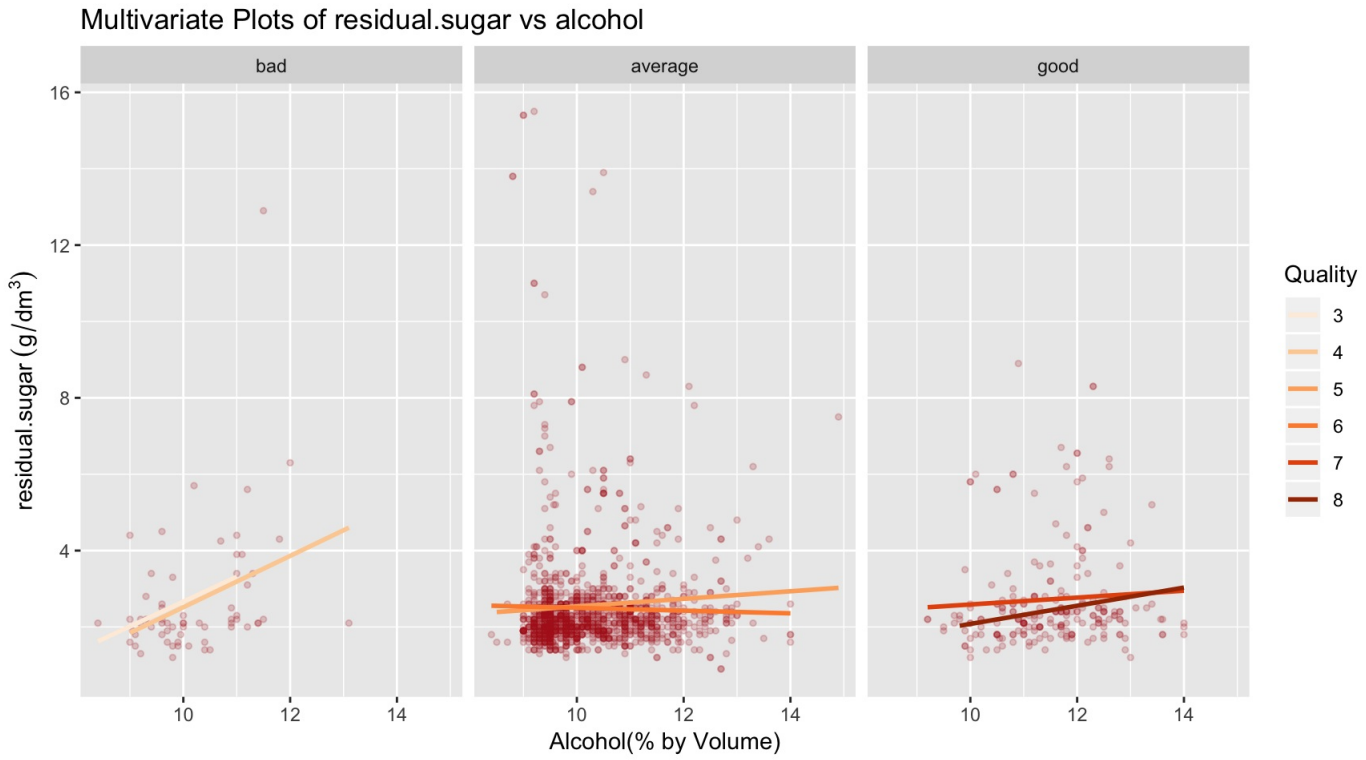


The lower pH and high content of Alcohol seems to produce better quality of wines.

Effect on quality of wine with residual.sugar and alcohol

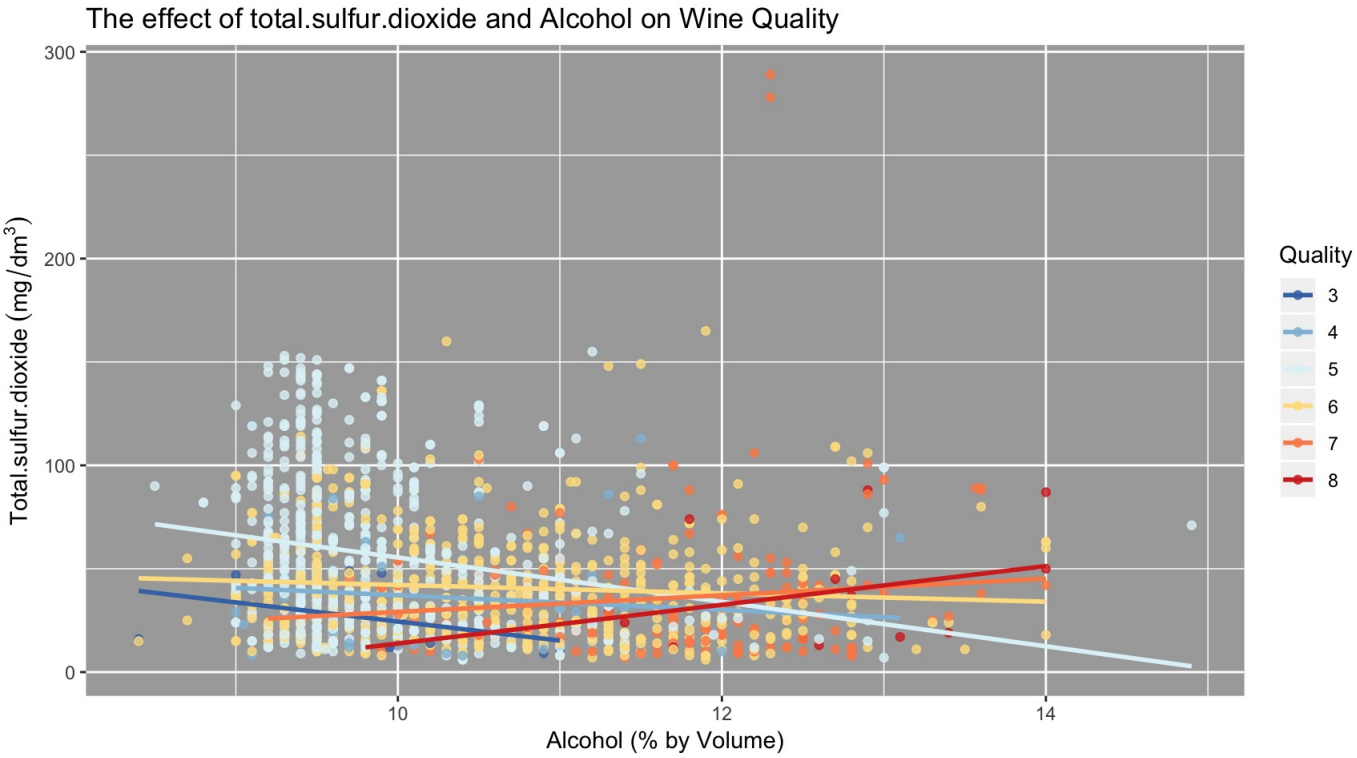


In this plot, residual sugar is slightly positive effect on alcohol, particularly in low AND high quality of wine.

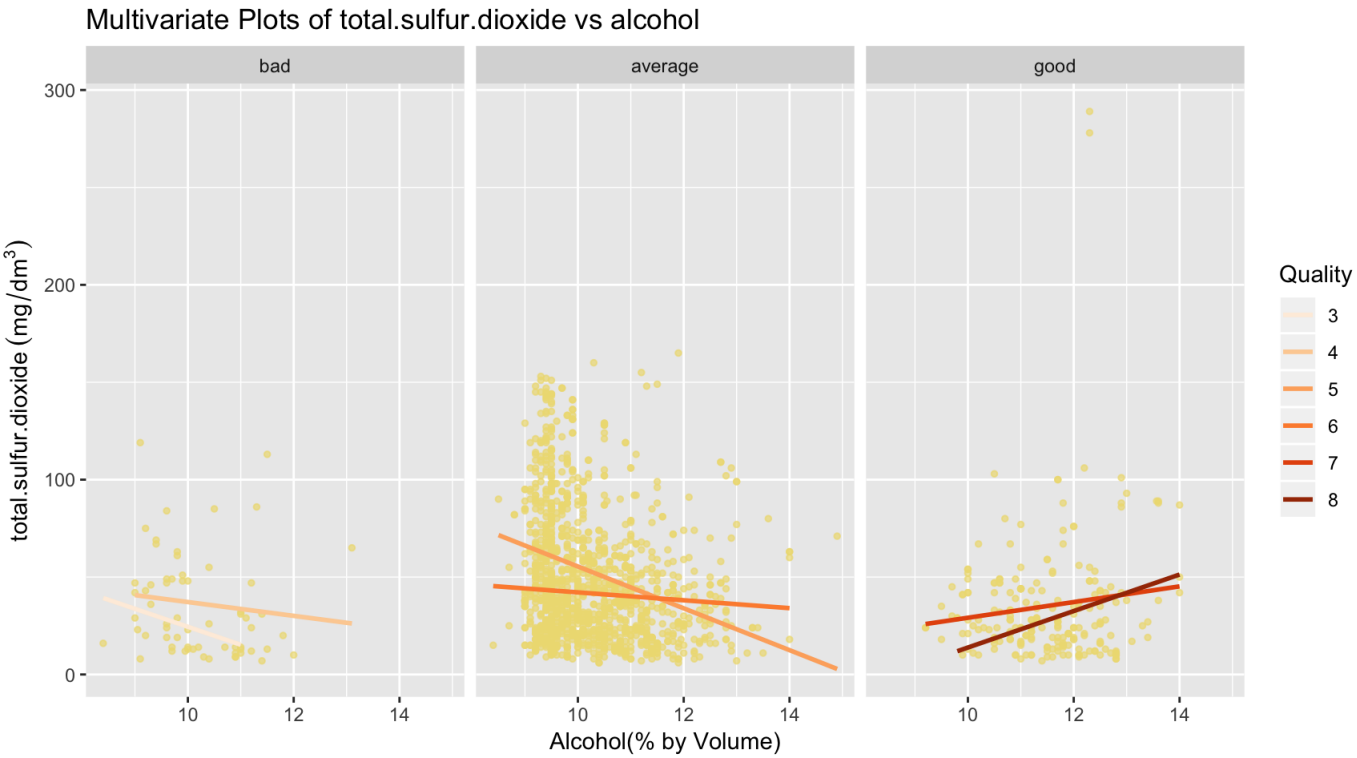


Same as above, Residual sugar is slightly positive effect on alcohol, particularly in low AND high quality of wine.

Effect on quality of wine with total.sulfur.dioxide and alcohol

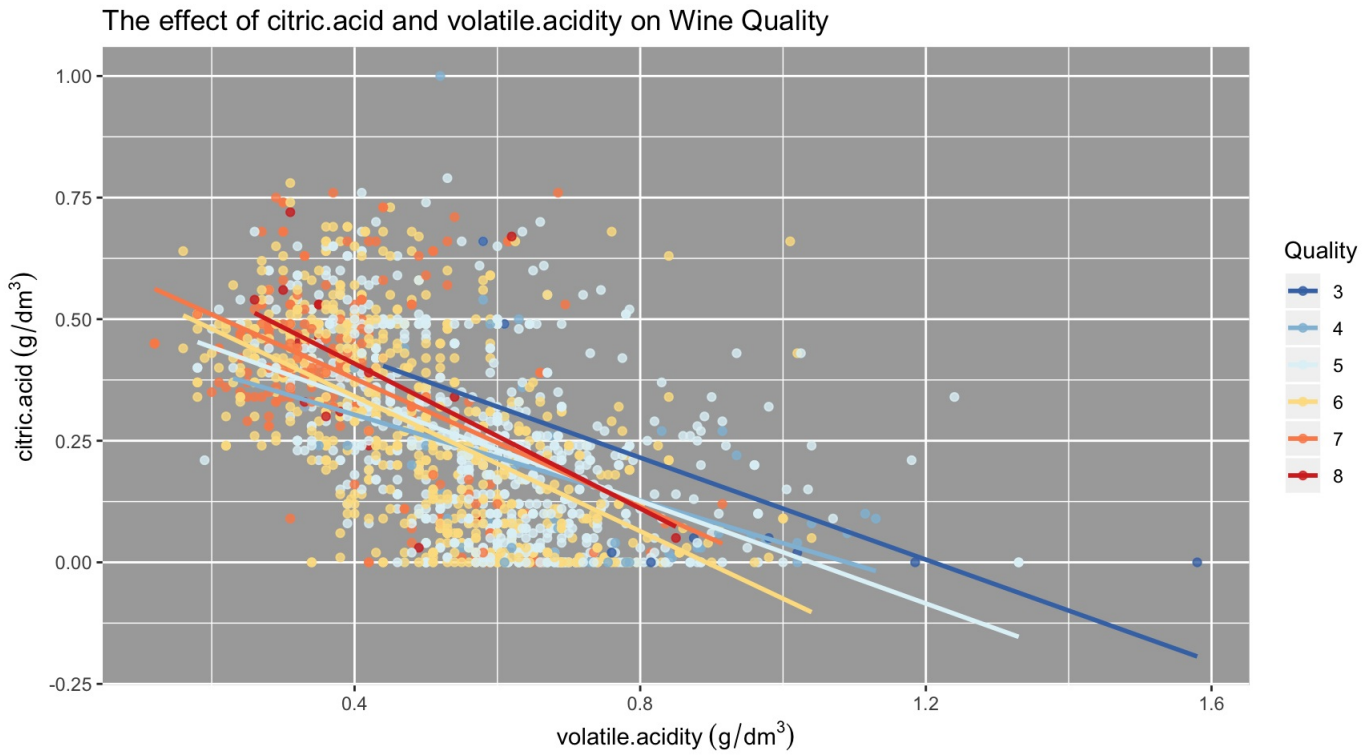


In this plot, total.sulfur.dioxide is slightly negaive effect on alcohol, particularly in low AND medium quality of wine. In contrast, total.sulfur.dioxide is slightly positive effect on alcohol in high quality of wine.

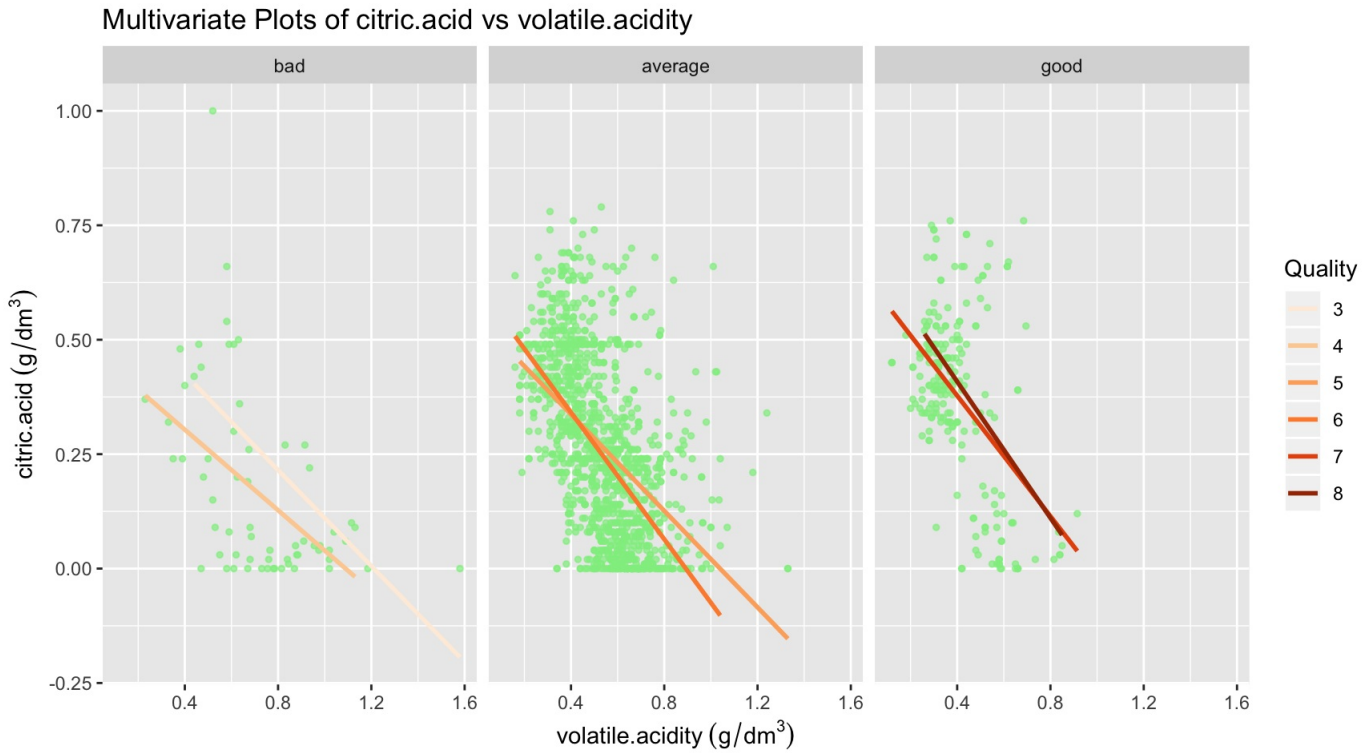


The lower Sulphur Dioxide seems to produce better quality of wine even though some outliers of high Sulphur Dioxide in better wine.

Effect on quality of wine with citric.acid and volatile.acidity

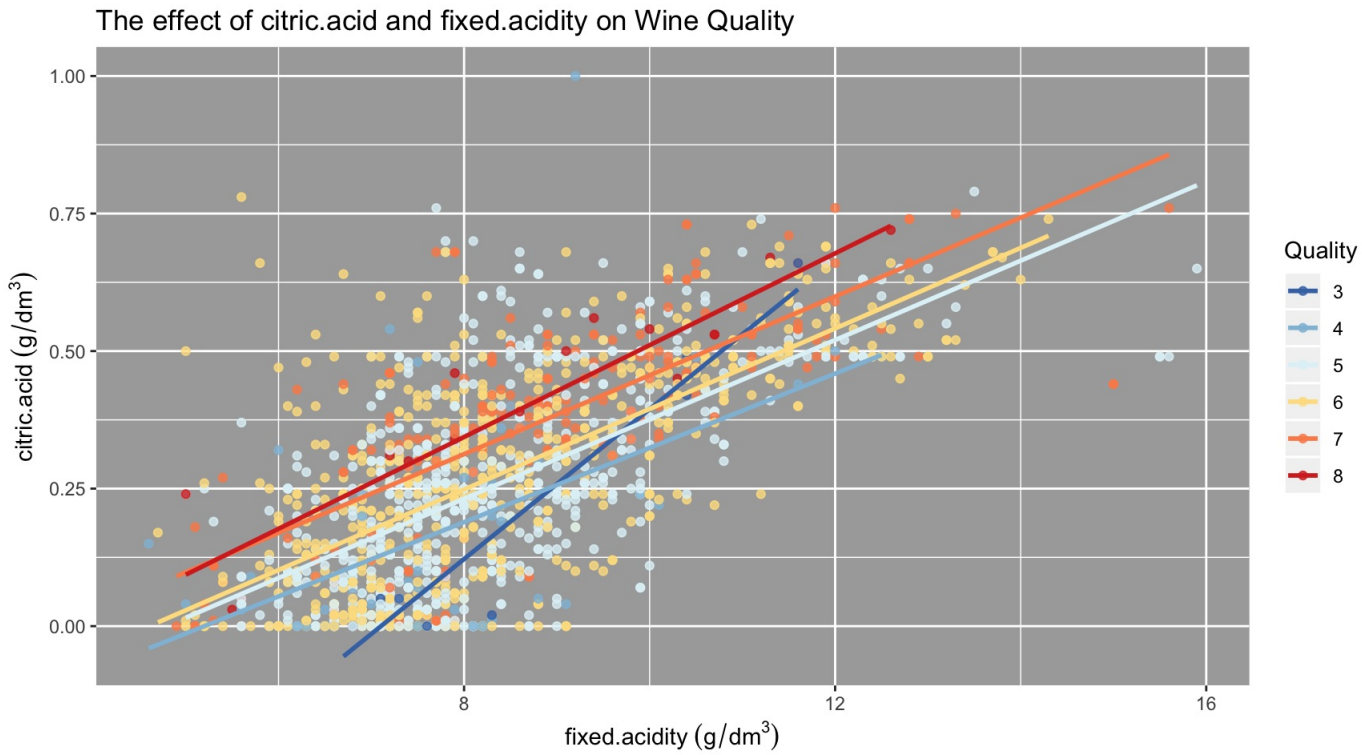


In this plot, citric.acid is played strong negaive effect on voltaile acidity in both quality of wine.

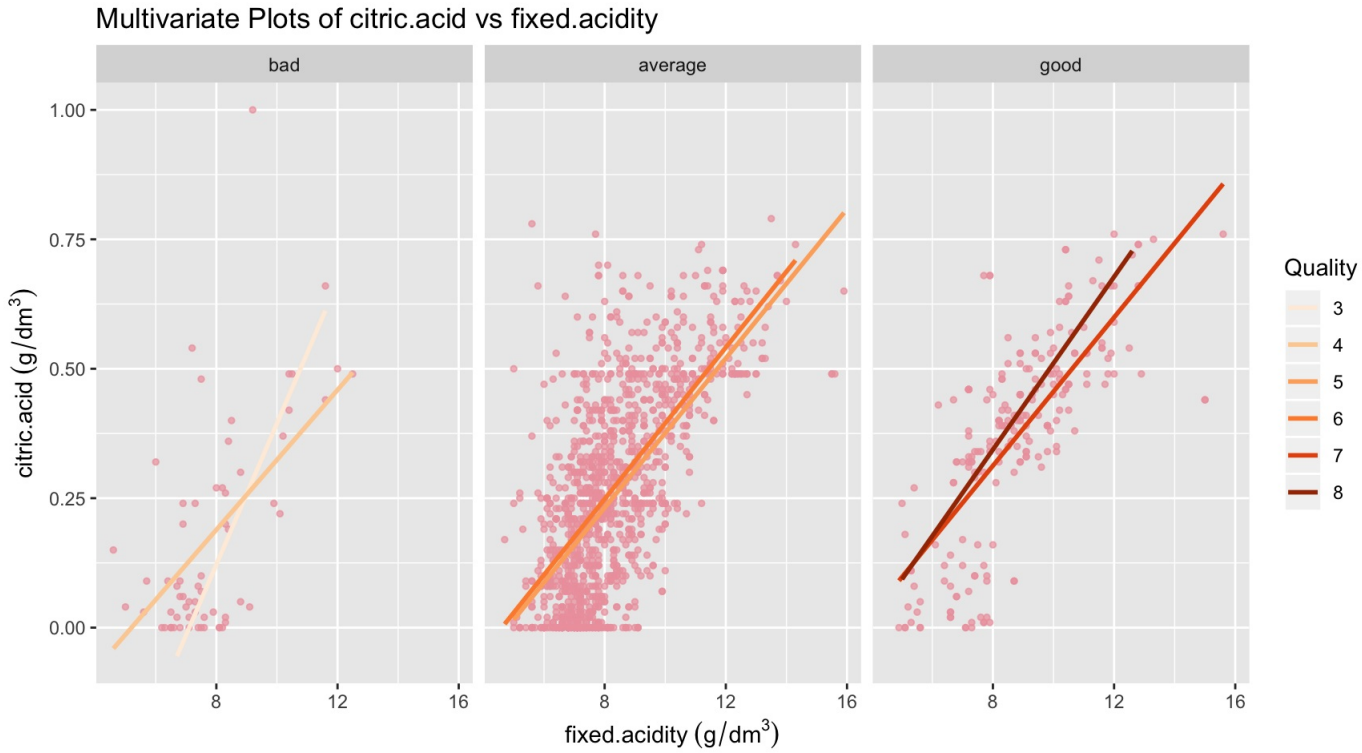


Higher Citric Acid with low Volatile Acid seems to produce better Wines.

Effect on quality of wine with citric.acid and fixed.acidity

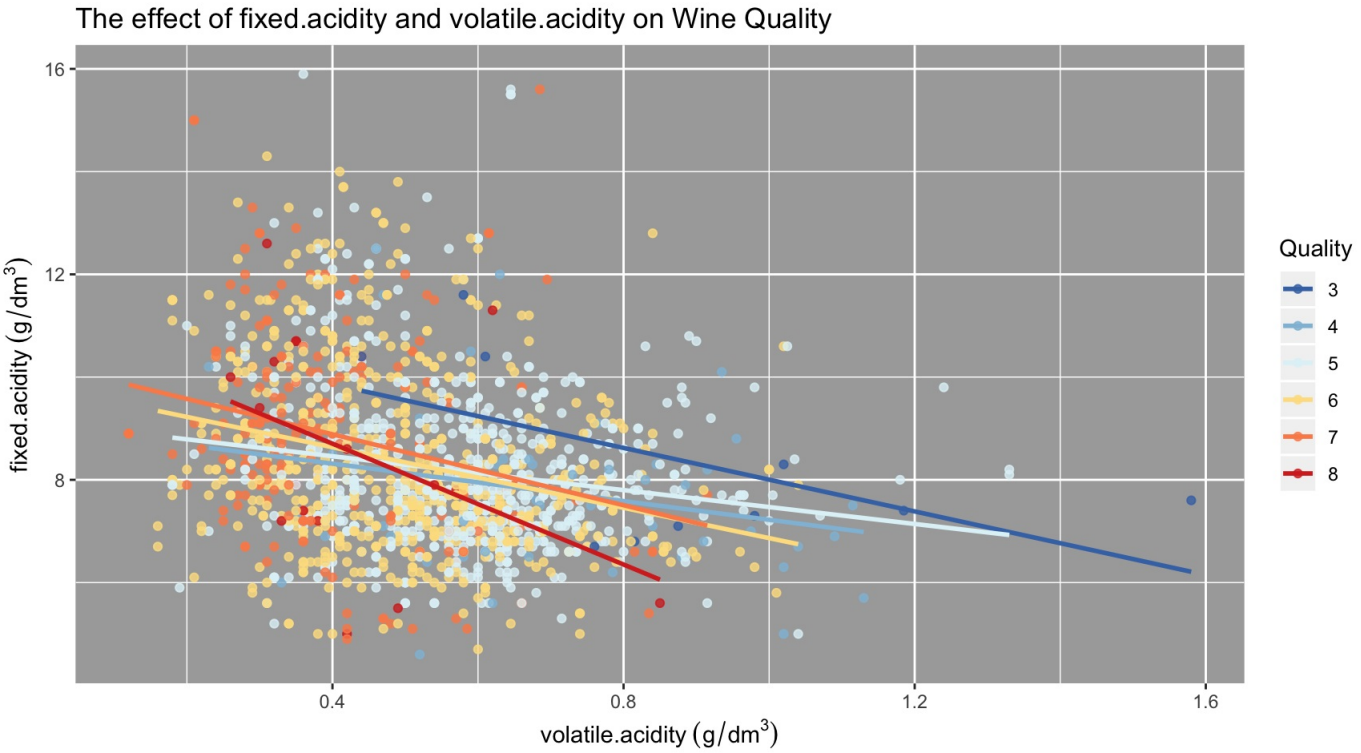


In this plot, citric.acid is played strong positive effect on fixed acidity in both quality of wine.

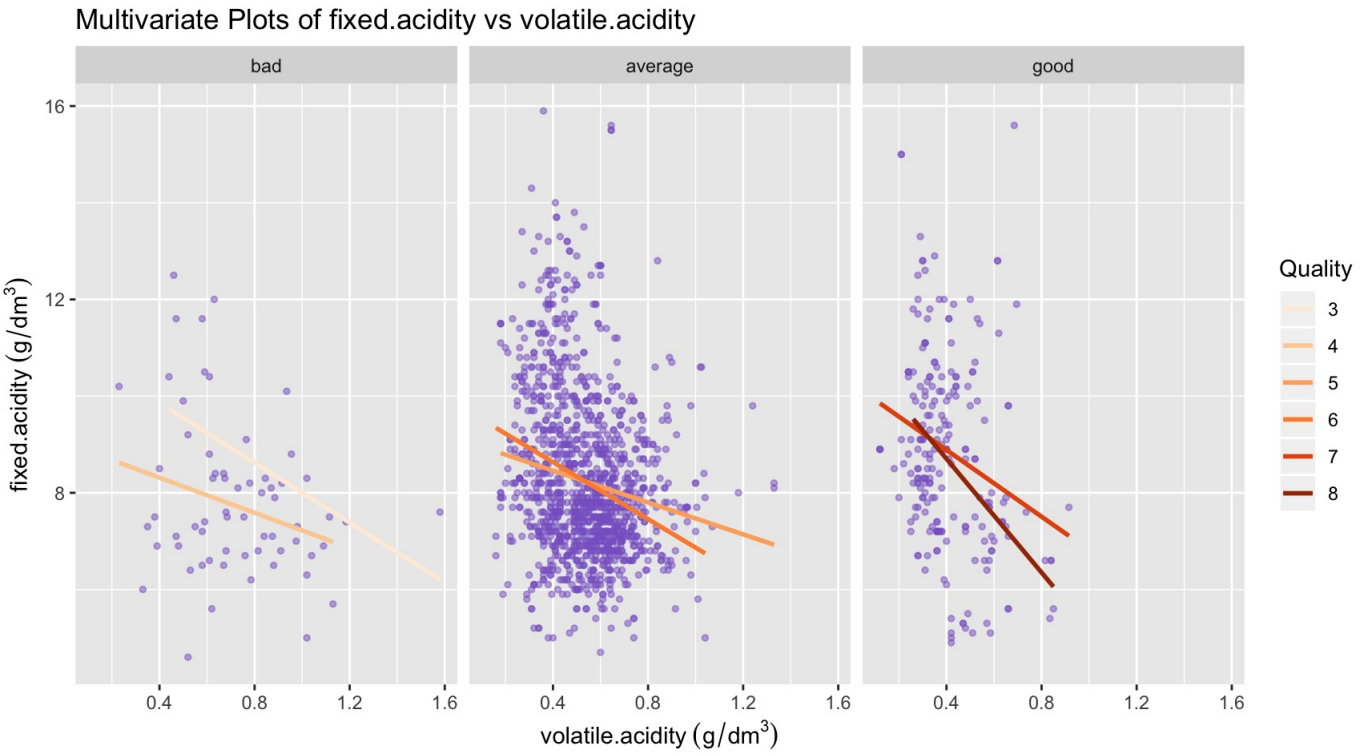


Higher Citric Acid with low fixed acidity seems to produce better Wines.

Effect on quality of wine with fixed.acid and volatile.acidity



In this plot, fixed.acidity is played negative effect on volatile acidity in both quality of wine.



Lower fixed Acidity with higher volatile acidity seems to produce better Wines.

Analysis of the Multivariate Plots

Observations

-High Alcohol and Sulaphate content seems to produce better quality of wines.

-Alough citric Acid has weak correlation on wines, it still makes an improvement on wine quality.

Multiple Regression Model

First combination of all the variables that increase the quality with their increasing levels. Next combination is density and fixed.acidity is implied for predicting the quality. Next use volatile.acidity, as it has the highest negative correlation coefficient with the quality variable. And the last combination of pH, total.sulfur.dioxide and free.sulfur.dioxide.

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol * sulphates * citric.acid * fixed.acidity,
##     data = wine)
## m2: lm(formula = quality ~ alcohol + sulphates + citric.acid + fixed.acidity +
##     density + alcohol:sulphates + alcohol:citric.acid + sulphates:citric.acid +
##     alcohol:fixed.acidity + sulphates:fixed.acidity + citric.acid:fixed.acidity +
##     fixed.acidity:density + alcohol:sulphates:citric.acid + alcohol:sulphates:fixed.acidity +
##     alcohol:citric.acid:fixed.acidity + sulphates:citric.acid:fixed.acidity +
##     alcohol:sulphates:citric.acid:fixed.acidity, data = wine)
## m3: lm(formula = quality ~ alcohol + sulphates + citric.acid + fixed.acidity +
##     density + volatile.acidity + alcohol:sulphates + alcohol:citric.acid +
##     sulphates:citric.acid + alcohol:fixed.acidity + sulphates:fixed.acidity +
##     citric.acid:fixed.acidity + fixed.acidity:density + alcohol:sulphates:citric.acid +
##     alcohol:sulphates:fixed.acidity + alcohol:citric.acid:fixed.acidity +
##     sulphates:citric.acid:fixed.acidity + alcohol:sulphates:citric.acid:fixed.acidity,
##     data = wine)
## m4: lm(formula = quality ~ alcohol + sulphates + citric.acid + fixed.acidity +
##     density + volatile.acidity + pH + total.sulfur.dioxide +
##     free.sulfur.dioxide + alcohol:sulphates + alcohol:citric.acid +
##     sulphates:citric.acid + alcohol:fixed.acidity + sulphates:fixed.acidity +
##     citric.acid:fixed.acidity + fixed.acidity:density + pH:total.sulfur.dioxide +
##     pH:free.sulfur.dioxide + total.sulfur.dioxide:free.sulfur.dioxide +
##     alcohol:sulphates:citric.acid + alcohol:sulphates:fixed.acidity +
##     alcohol:citric.acid:fixed.acidity + sulphates:citric.acid:fixed.acidity +
##     pH:total.sulfur.dioxide:free.sulfur.dioxide + alcohol:sulphates:citric.acid:fixed.acidity,
##     data = wine)
##
## ================================================================================================
##                                                        m1        m2        m3        m4
## ------------------------------------------------------------------------------------------------
##   (Intercept)                                           4.004   -18.407   -30.327   -49.730
##
##   alcohol                                              -0.248    -0.031    -0.006     0.166
##
##   sulphates                                            -1.259     1.733     1.544     1.326
##
##   citric.acid                                          24.816    31.954    33.083    36.242
##
##   fixed.acidity                                         0.078    10.120     8.308     7.842
##
##   alcohol x sulphates                                   0.400     0.151     0.199     0.126
##
##   alcohol x citric.acid                                -1.790    -2.483    -2.711    -3.038
##
##   sulphates x citric.acid                             -55.541   -64.136   -62.580   -61.781
##
##   alcohol x fixed.acidity                               0.004    -0.029    -0.014    -0.039
##
##   sulphates x fixed.acidity                            -0.679    -1.025    -0.781    -0.833
##
##   citric.acid x fixed.acidity                          -3.359    -4.130    -4.076    -4.447
##
##   alcohol x sulphates x citric.acid                     4.484     5.231     5.177     5.190
##
##   alcohol x sulphates x fixed.acidity                   0.052     0.082     0.048     0.067
##
##   alcohol x citric.acid x fixed.acidity                 0.273     0.349     0.342     0.385
##
##   sulphates x citric.acid x fixed.acidity               6.979     7.874     7.330     7.300
##
##   alcohol x sulphates x citric.acid x fixed.acidity    -0.597    -0.675    -0.621    -0.634
##
##   density                                                       19.745    31.965    55.838
##
##   fixed.acidity x density                                       -9.685    -7.953    -7.355
##
##   volatile.acidity                                                        -1.106    -0.985
##
##   pH                                                                                -1.566
##
##   total.sulfur.dioxide                                                              -0.072
```

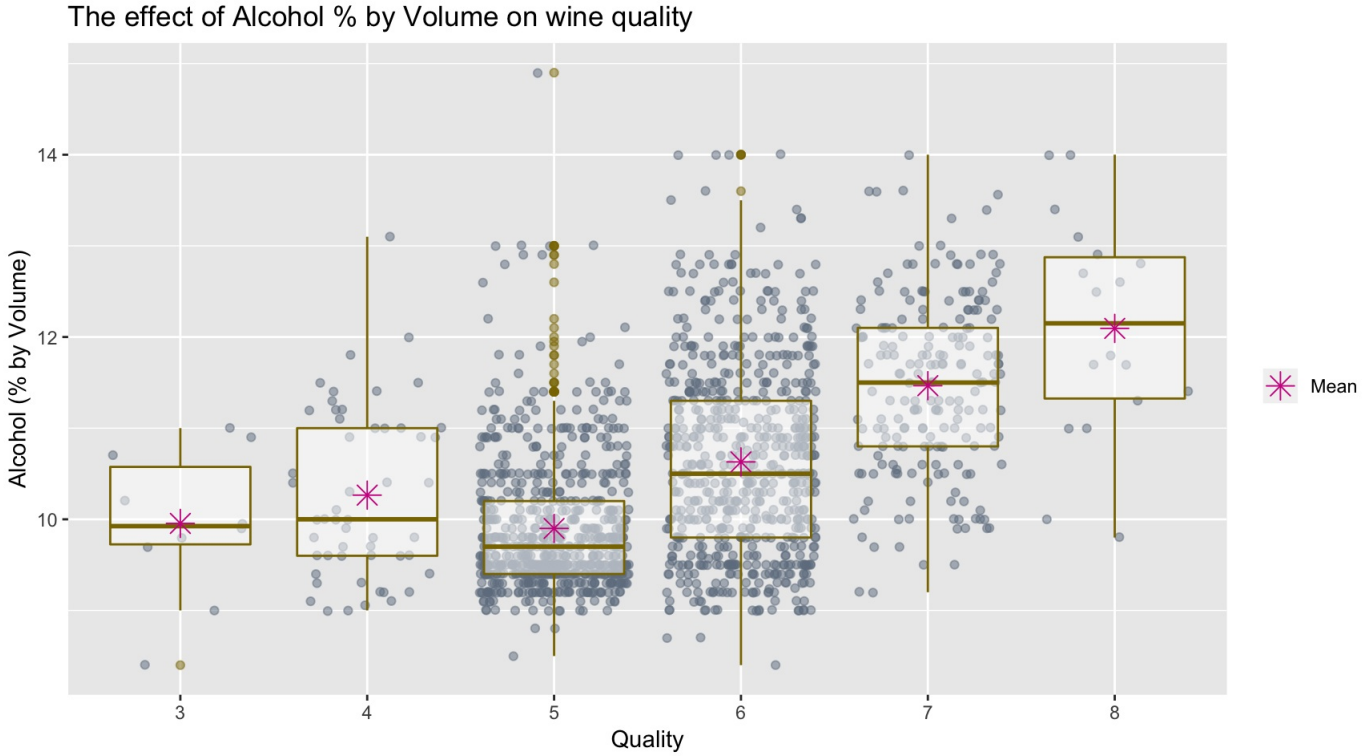
```
##
##   free.sulfur.dioxide                                                                  -0.207
##
##   pH x total.sulfur.dioxide                                                             0.021
##
##   pH x free.sulfur.dioxide                                                              0.065
##
##   total.sulfur.dioxide x free.sulfur.dioxide                                            0.003
##
##   pH x total.sulfur.dioxide x free.sulfur.dioxide                                      -0.001
##
## ------------------------------------------------------------------------------------------------
##   R-squared
##   adj. R-squared
##   sigma
##   F
##   p
##   Log-likelihood
##   Deviance                                                0.000     0.000     0.000     0.000
##   AIC
##   BIC
##   N                                                    1599      1599      1599      1599
## ================================================================================================
```

Multiple regression model can maximum to explain ~39% of existing 4 cases in the dataset. Additional dataset with more data of extreme quality cases (both high and low-quality) should help improve the results of this model. Moreover, more sophisticated prediction models should be able to provide more accurate predictions for the quality of wine based on its chemical characteristics.

Final Plots and Summary

Alcohol and Sulphates are played a significant contribution on wine quality.

Plot 1

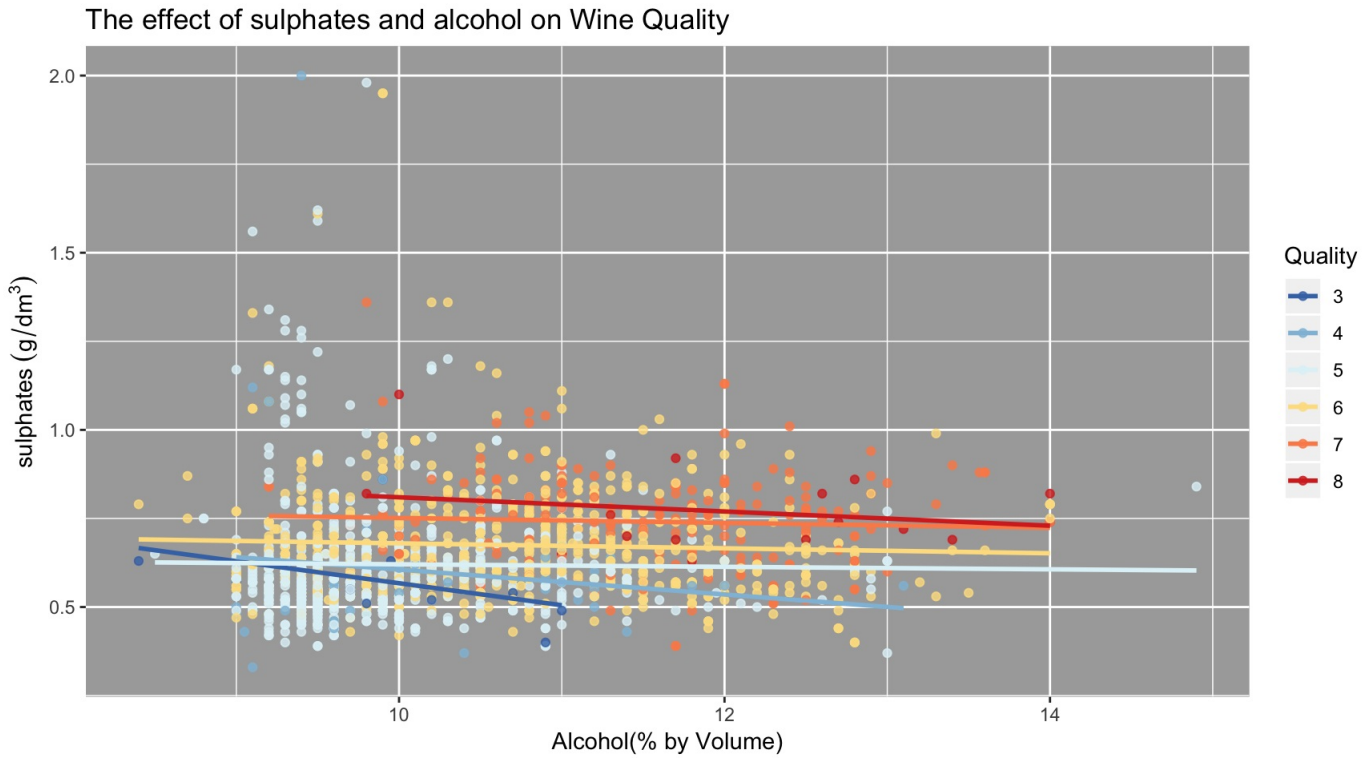


The plot shows that Alcohol content has played a significant contribution on the quality of Wines. The higher the content of alcohol, the better the quality of wine.

In this dataset, most of the data consists average quality wine. We found that the mean and median coincides for a particular Quality, which means it is normally distributed. In addition, the highest value of the median in the best quality wines imply that almost all points have a high content of alcohol.

From linear model, we found that R Squared in alcohol is contribute to ~22% in the variance of the wine quality. Therefore, alcohol is not the sole contributor for the improvement in Wine Quality.

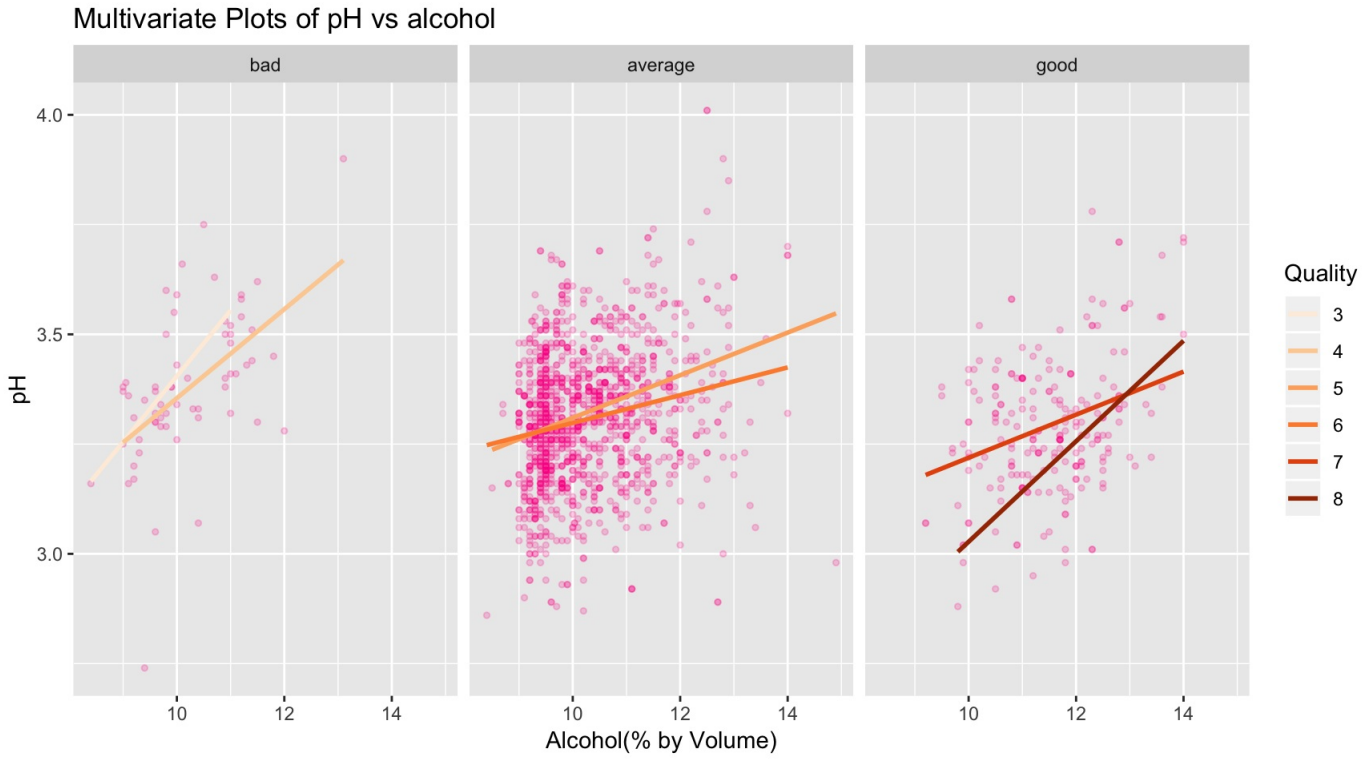
Plot 2



In this plot, we found that the best quality of wines have high values in both Alcohol percentage and Sulphate concentration implying that high alcohol contents and high sulphate concentrations together appear to have better wines. The slop is slightly downward, it may due to the higher quality of wines, percentage of alcohol is slightly greater than the concentration of Sulphate.

Plot 3

Effect on quality of wine with pH and alcohol



We found that the lower pH and high content of Alcohol have to produce better quality of wines. ### Conclusions

The main contributos for the quality of wine are: High level contributor:alcohol, fixed acidity, citric acid & sulphates Moderate level contributor: sulfur dioxides Low level contributor: density, volatile acidity, pH, and sugar

Lesson learned

Simpson’s paradox is present in this dataset which is responsible for the trend reversal of pH vs Volatile Acid. I first clustered the data into 3 segments and calculated the regression coefficient. I found that there is indeed a sign reversal. It is due to a lurking variable is present. As it is not included as an explanatory or response variable in the analysis but can affect the interpretation of relationships between variables. Therefore it is falsely identify a strong positive relationship between pH and volatile acidicity and finally hide the true relationship.

Reflection

-In this project I have used exploratory data analysis (EDA) approach to analyzing red wine dataset and discover which chemical properties are influence the quality of red wine. I tried several visualization methods such as histogram, scatter plots and box-plots. It is a good practice to understand the data first and try to gather as many insights from it. Like find out their patterns, identify any anomalies, to test hypothesis and make assumptions which can help to make statistical analysis.

-The difficulty to analysis this dataset as several of the attributes may be correlated, thus we need to apply some sort of feature selection. Alternatively, this is good approach for beginner to using this dataset for EDA, as there is no missing attributes.

-For future work, I would like to create a model so as we can predict the quality of our wine.

Reference

https://s3.amazonaws.com/udacity-hosted-downloads/ud651/wineQualityInfo.txt

https://www.britannica.com/topic/Simpsons-paradox