Red_Wine_DataAnalysis

Sridhar Varanasi

January 20, 2018

0.1 R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com (http://rmarkdown.rstudio.com).

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
## [1] 1599   13
```

```
##  [1] "X"                    "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"          "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "quality"
```

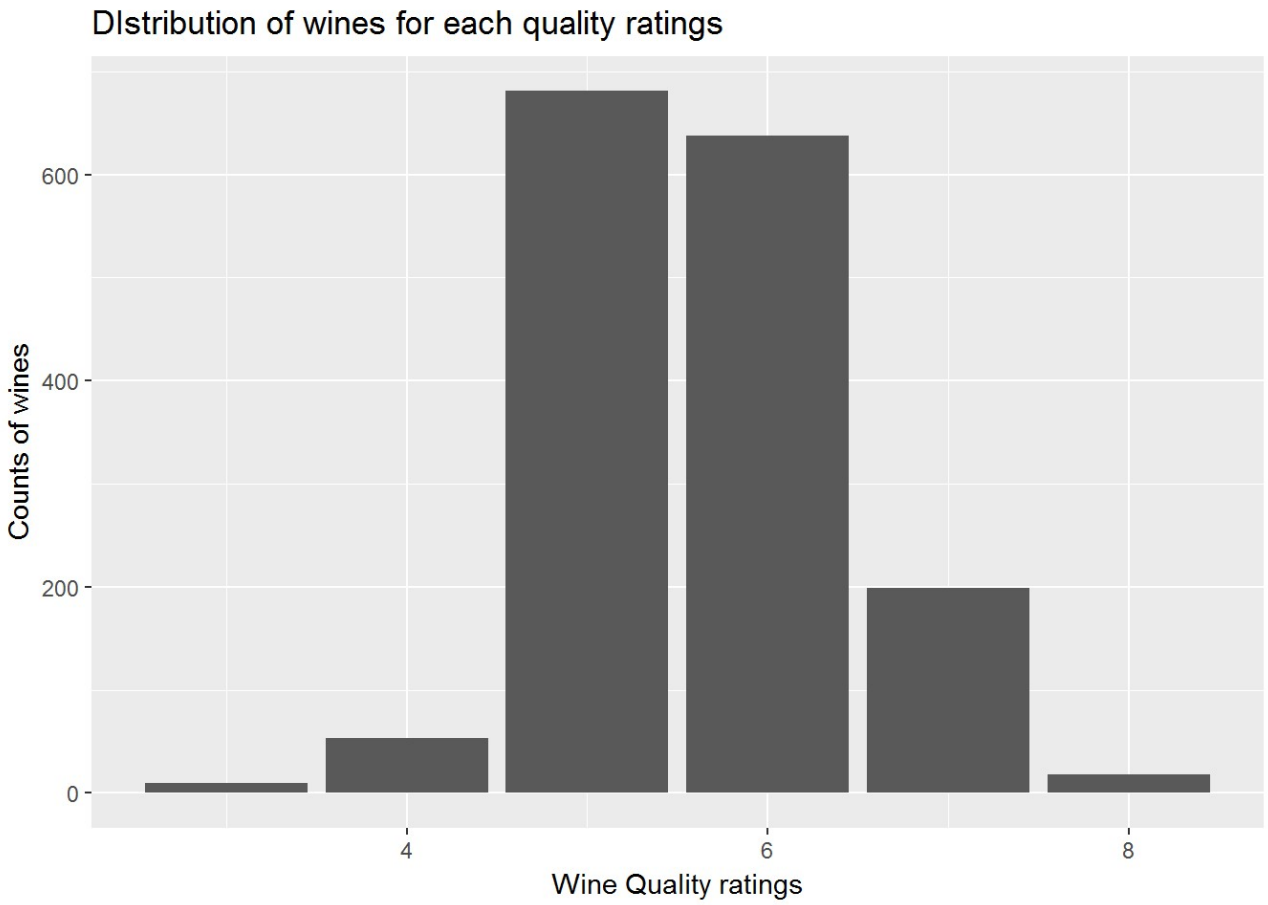
- This data set contains 13 variables and 1599 observations of red wines.
- The first column of this dataset is an identity column which has uniques value to each of the records.
- All the columns are either numeric or of integer types.
- There are variables like acicity , citric acid content, residual sugar, Suplphates ,alcohol content, pH value that determine the quality of red wine.
- The rating for quality of wine is between 0 to 10 . 0 being lowest and 10 being highest.
- In this analysis I will be starting off with doing univariate analysis. Then I wil proceed with Bivariate and Multivariate analysis on the data.
- Let's load the data and check the structure and summary of each variable in the dataset.

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                  : int  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 0.06
5 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 3.3
5 ...
##  $ 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            : int  5 5 5 6 5 5 5 7 7 5 ...
```

```
##        X          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
##  Min.   : 8.40   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.20   Median :6.000
##  Mean   :10.42   Mean   :5.636
##  3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :14.90   Max.   :8.000
```

1 Univariate Analysis

Let' check the distribution of the variables using histograms and boxplots



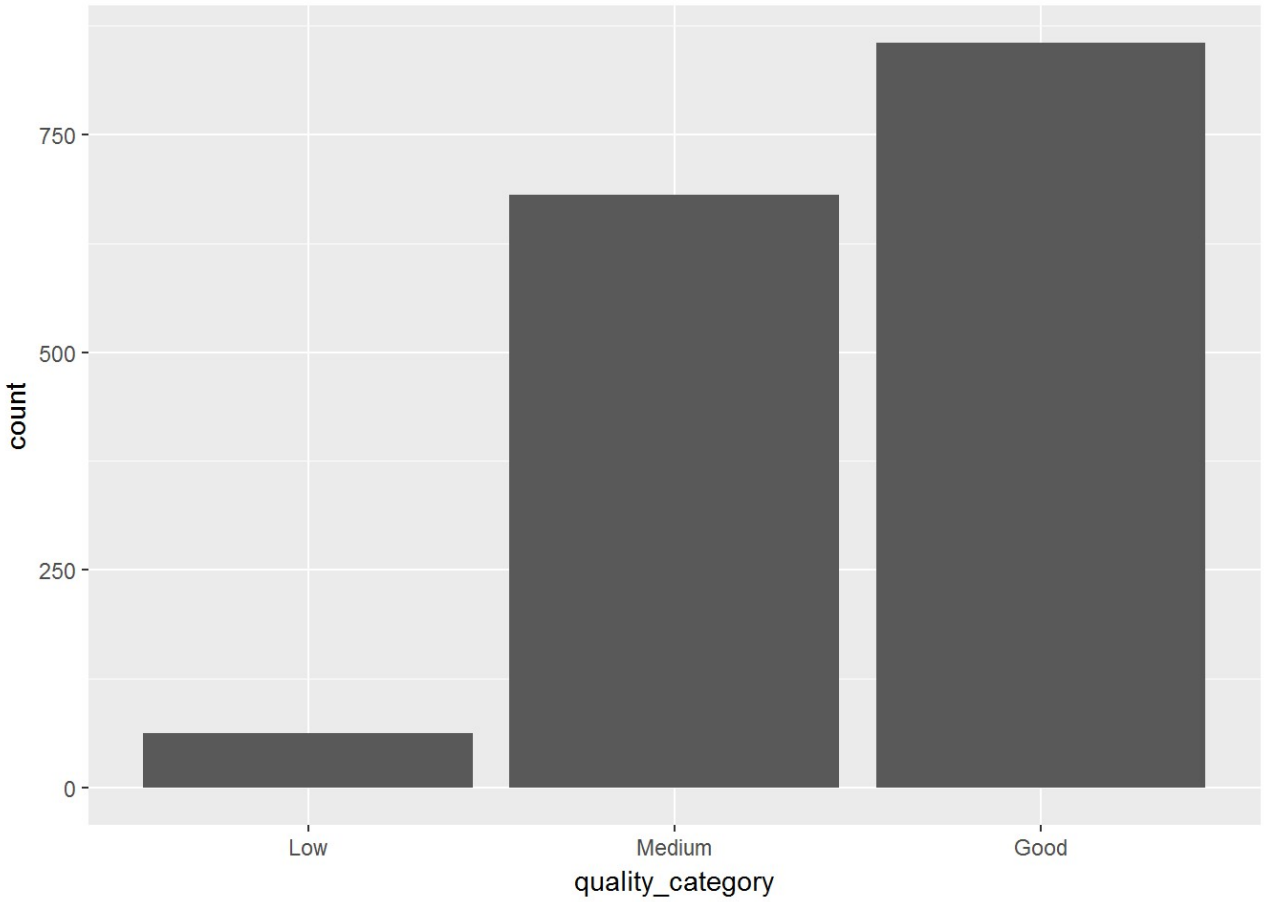

```
##
##   3   4   5   6   7   8
##  10  53 681 638 199  18
```

- The lowest wine rating is 3 in this data set and the highest rating in the dataset is 8.
- Almost 1300 of the 1599 wines in the dataset are having quality ratings as 5 or 6.
- There are 10 and 18 wines in the dataset that have rating of 3 and 8 respectively.

```
##
##    Low Medium   Good
##     63    681    855
```

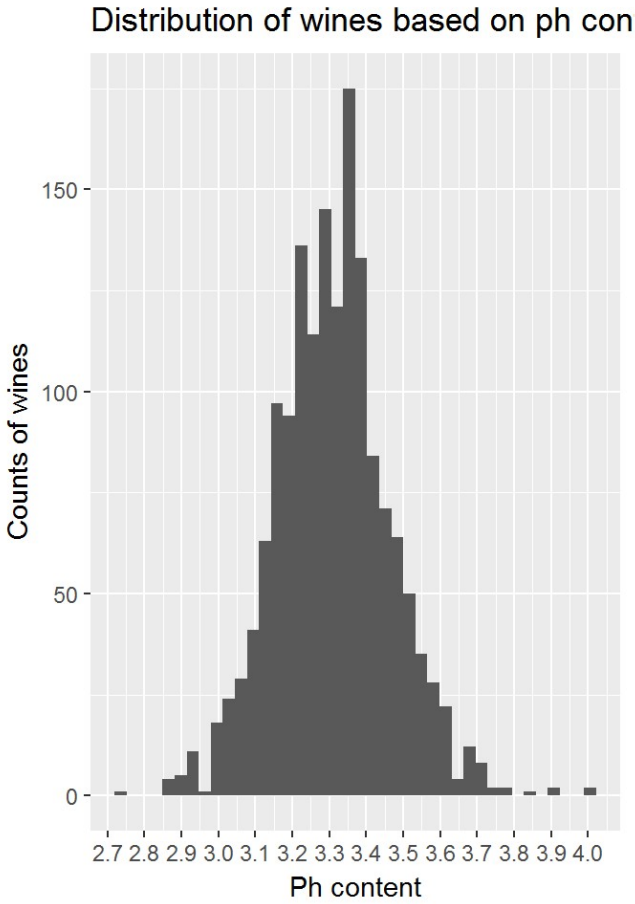
Added a variable quality_category which has 3 categories ‘Low’,‘Medium’ and ‘Good’ based on the quality ratings.

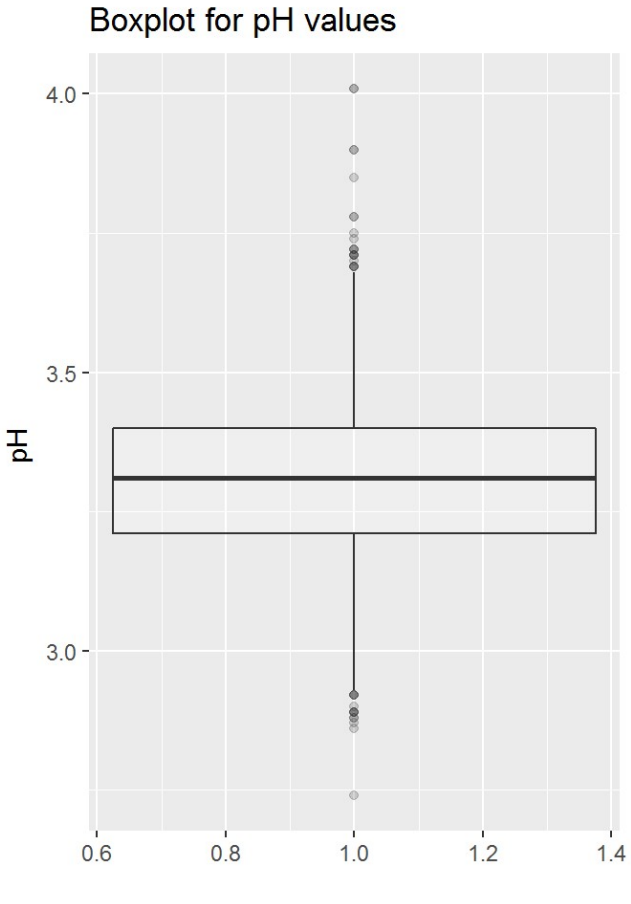
- If the quality ratings is less than 4 then its bad quality , it its between 5 and 6 then Medium else Good quality.



- As seen from above graph most of the wines are unser the Good category followed by Medium and Low.
- We will further analyze relationsheip with other variables for each of the categories in the bi-variate section.

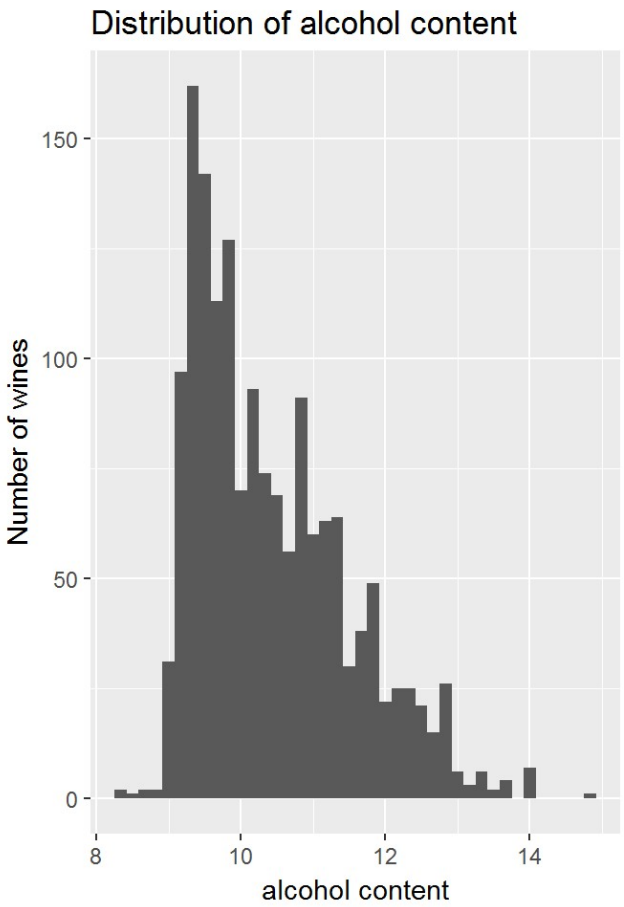
Let’s have a look at the ph content ditribution in the dataset

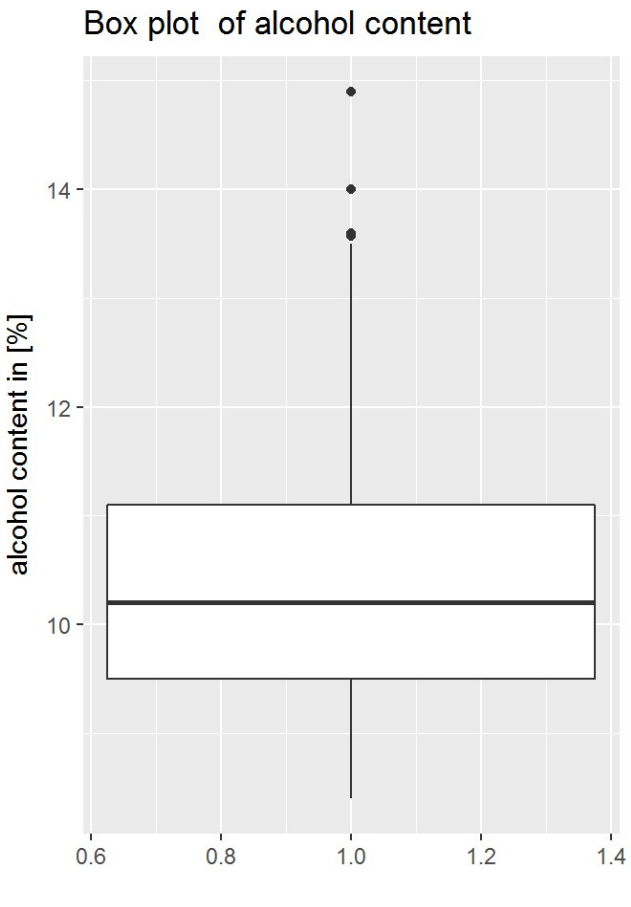




- Most of the wines are having Ph values less than 3.5. Only a few wines are having ph content above 3.5. Overall the distribution looks normal. Let's have a look at the density of wines.
- There are few wines which have pH as low as 2.7 and few wines which have pH value as high as 4.0.
- Also there are a few outliers which can be seen in the boxplot.

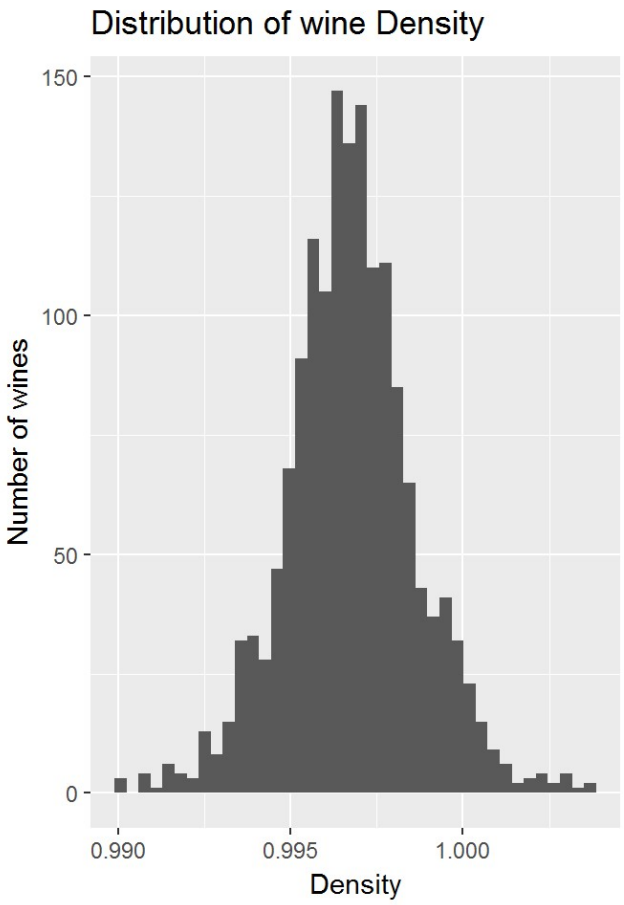
Note:The lower the pH value the more acidic the wine is.

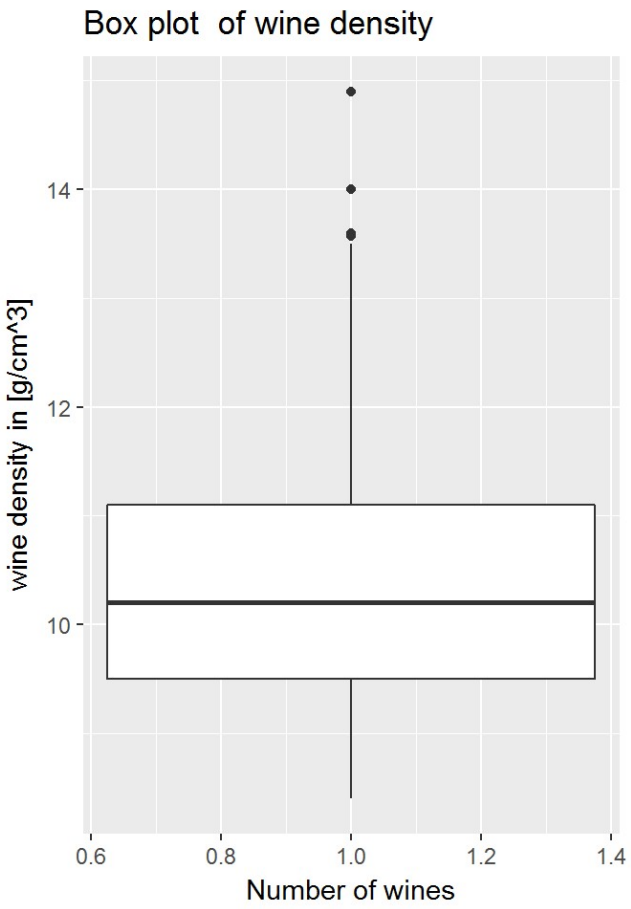





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

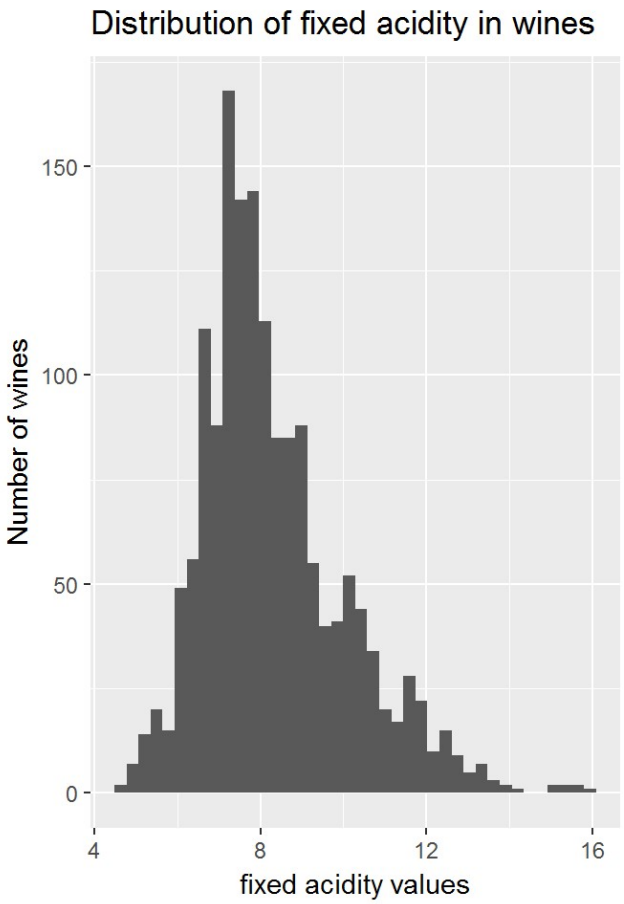
- As seen from the histogram for alcohol content most of the wines have alcohol content less than 12%. Also there are only few outliers for alcohol content.
- The range of alcohol content is between 8.40 to 14.90 with median value of 10.20.
- Let's look at the density avriable of red wine.

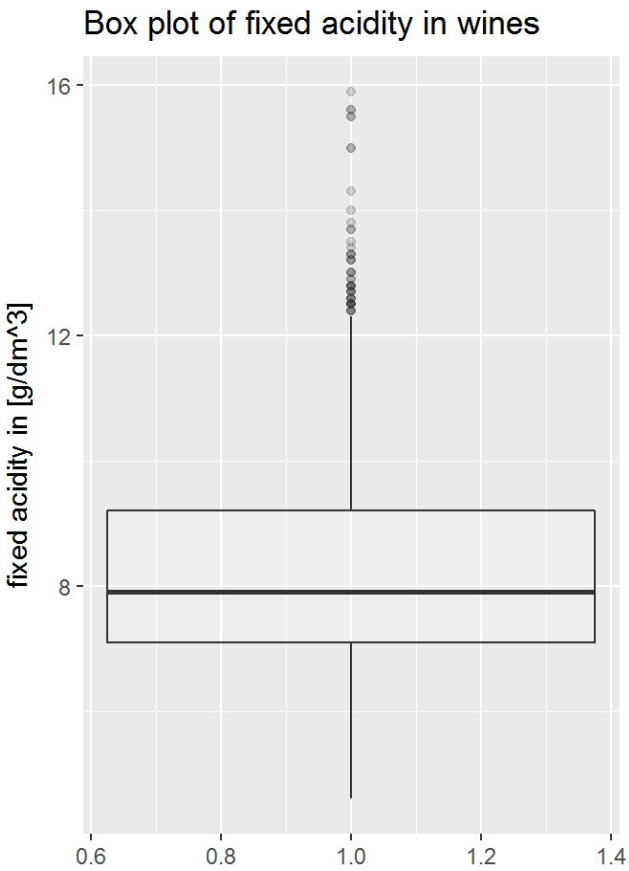





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

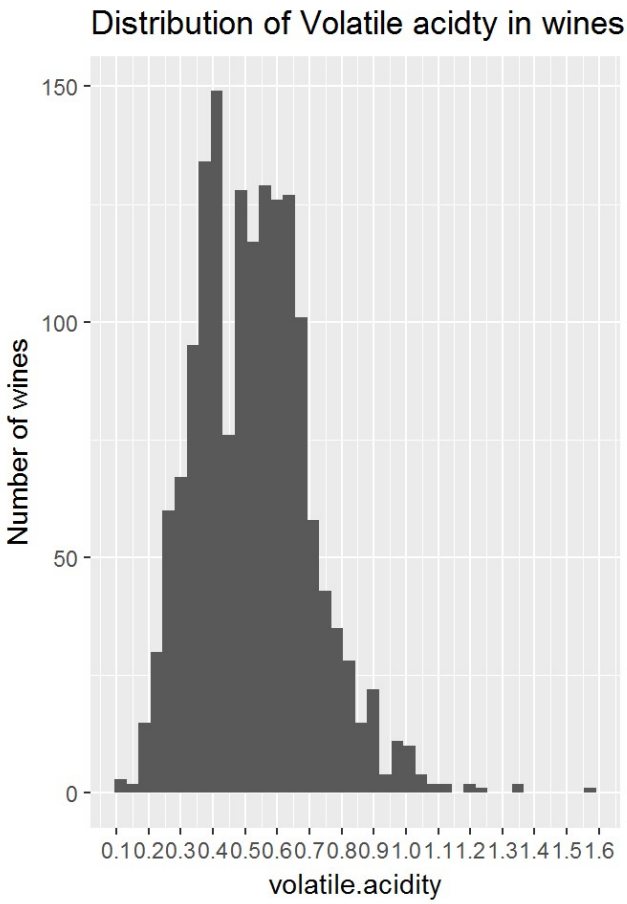
- It can be seen that the data is normally ditributed for the Density of wine. There are few outliers which can be seen from the box plot above.
- The range of density is between 0.99 to 1.003 with a median value of 0.9968.

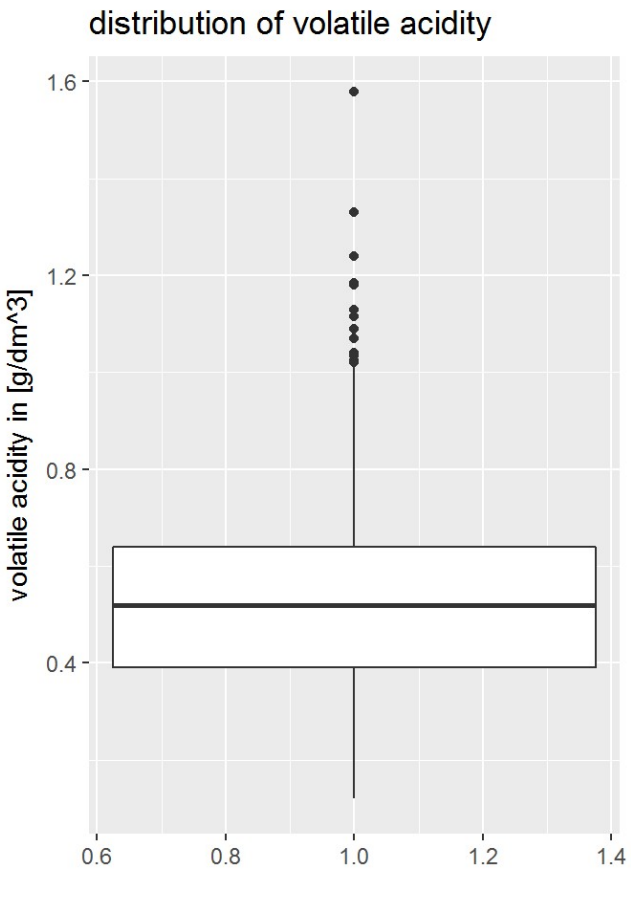





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

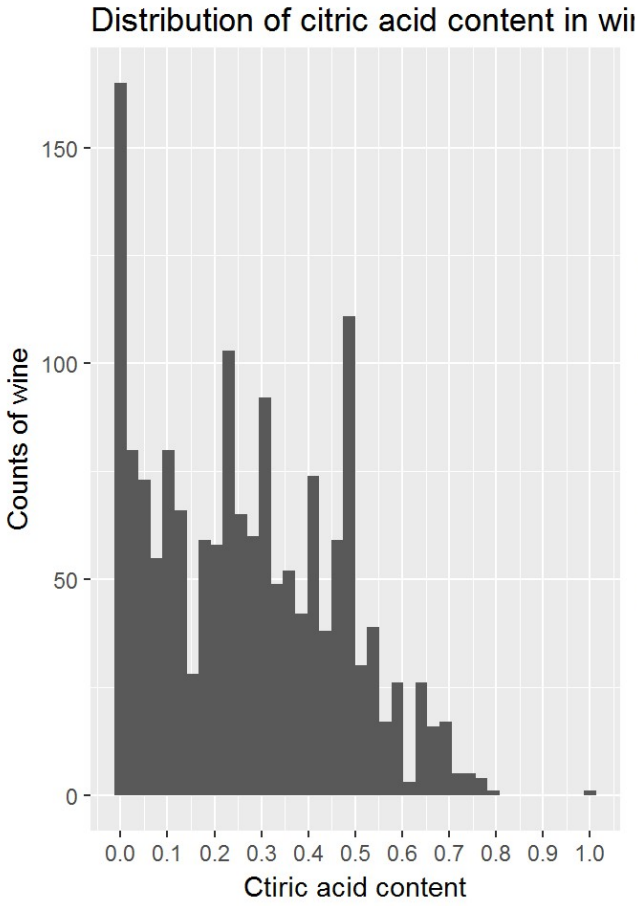
- It can be seen that most of the wines have fixed acidity around 6-10 g / dm^3. There are a few wines in the data whose acidity levels are around 16
- The distribution of the fixed acidity is positively skewed as seen in above
- The minumum vaue for fixed acidity is 4.60 and the maximum is 15.90 with median fixed acidity at 7.90
- There are many outliers which can be seen from the box plot

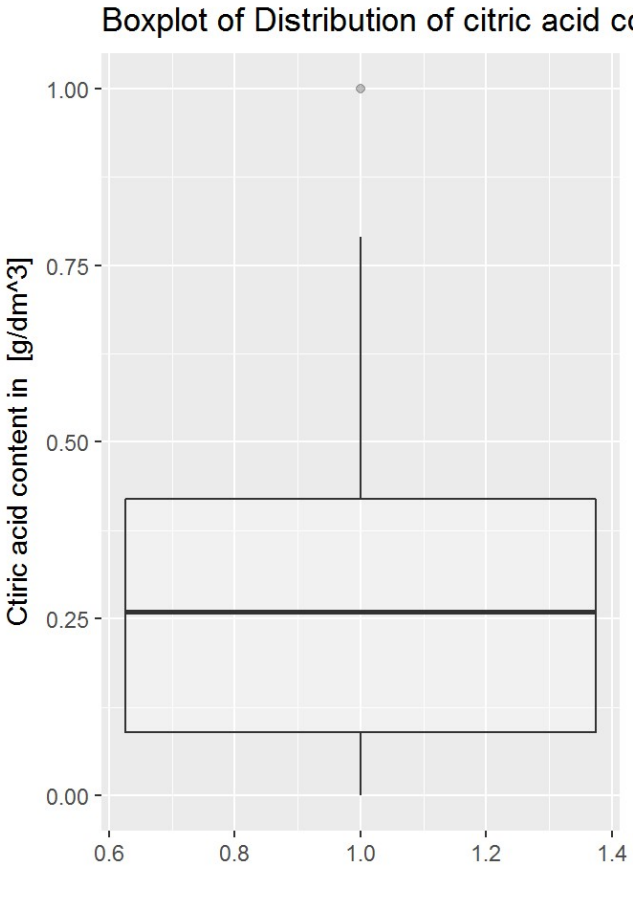





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

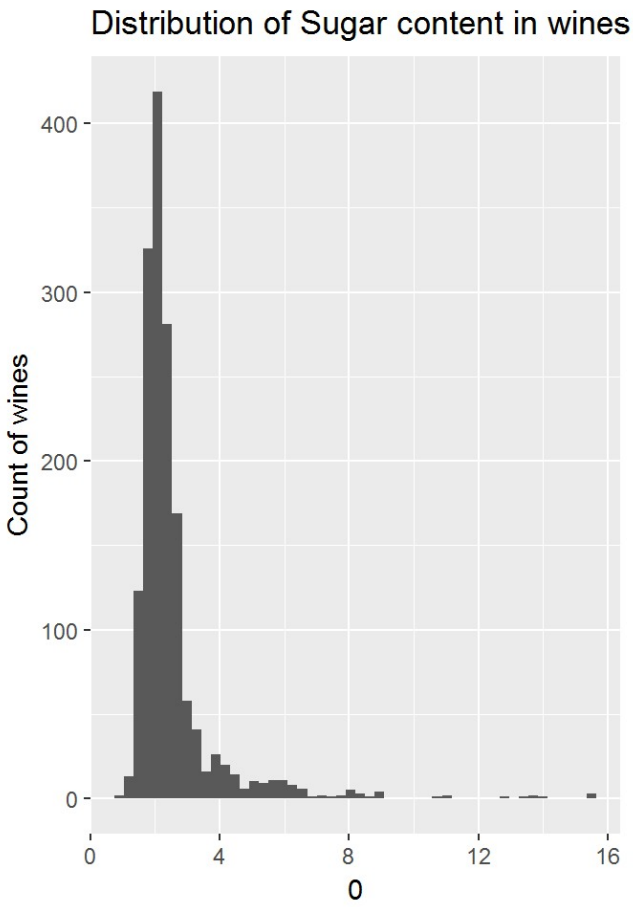
- Based on the above graph it seems that there are more number of wines with volatile acidity levels around 0.3
- There is one more peak at around 0.5-0.6 levels. Also, there are few wines high values of 1.35 and 1.55
- There are few outliers which can be seen in the box plot.Volatile acidity is in the range of 0.12 to 1.58 with a median value of 0.52

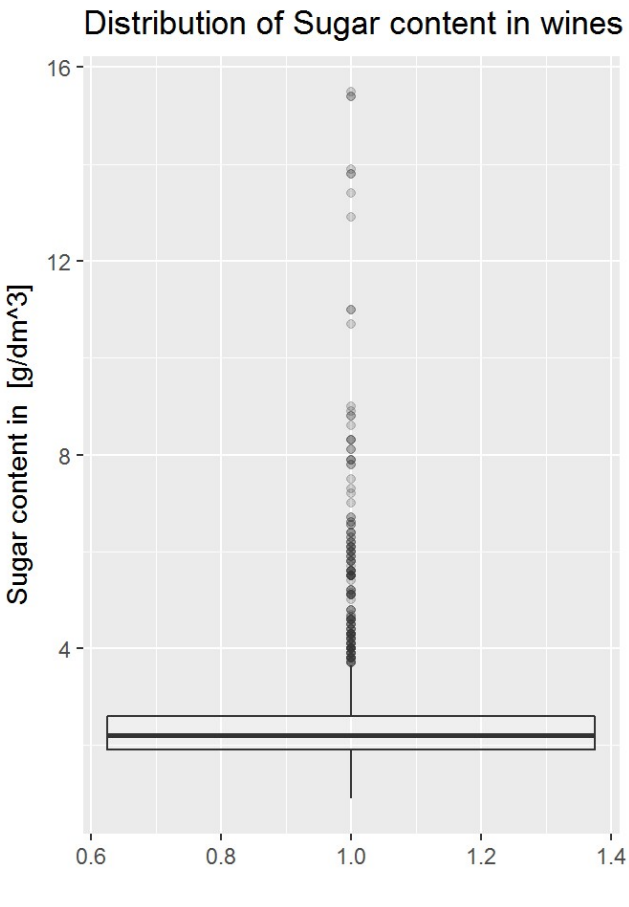





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

- As per the description of the variable citric acid. It is used in small quantities to add freshness and flavor to the wines.
- In this dataset most of the wines are having no citric acid content.
- Amount of citric acid is less than 0.6 for most of the wines. Very few wines have the citric acid content more than 0.6.
- It seems that there is an outlier in the data with high citric acid content.
- Range of Citric acid content for this dataset is between 0 to 1 with a median value of 0.260.

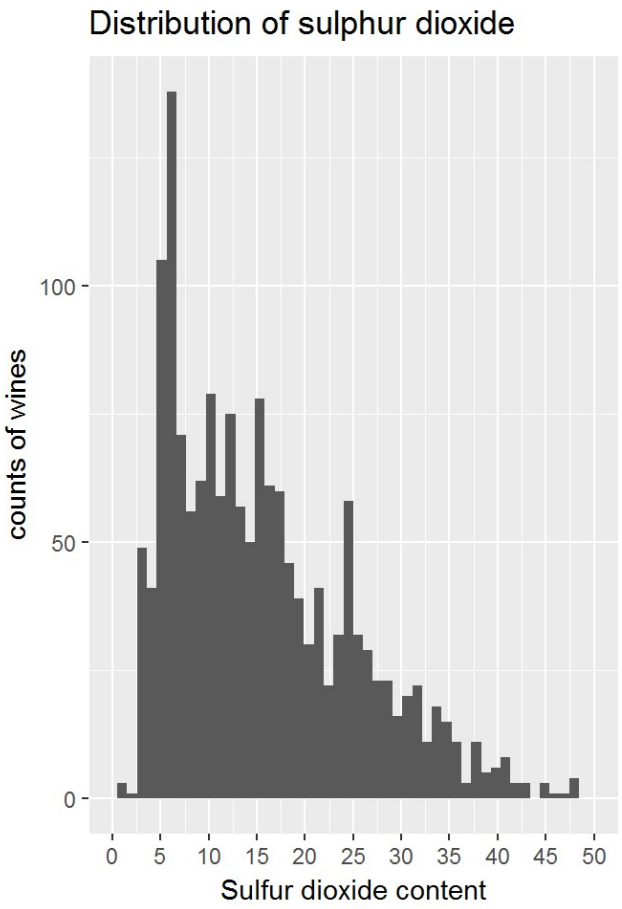


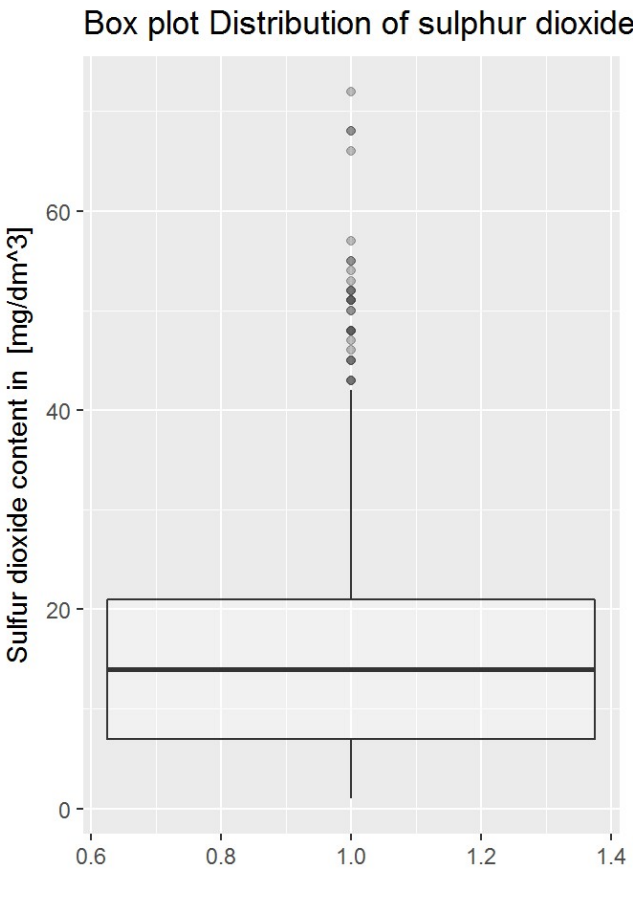



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

```
##
##   2 2.2 1.8 2.1 1.9 2.3 2.4 2.5 2.6 1.7 1.6 2.8 2.7 1.4 1.5   3 2.9 3.2
## 156 131 129 128 117 109  86  84  79  76  58  49  39  35  30  25  24  15
## 3.4 3.3   4 1.2 3.6 3.8 4.3 5.5 3.1 3.9 4.1 4.6
##  15  11  11   8   8   8   8   8   7   6   6   6
```

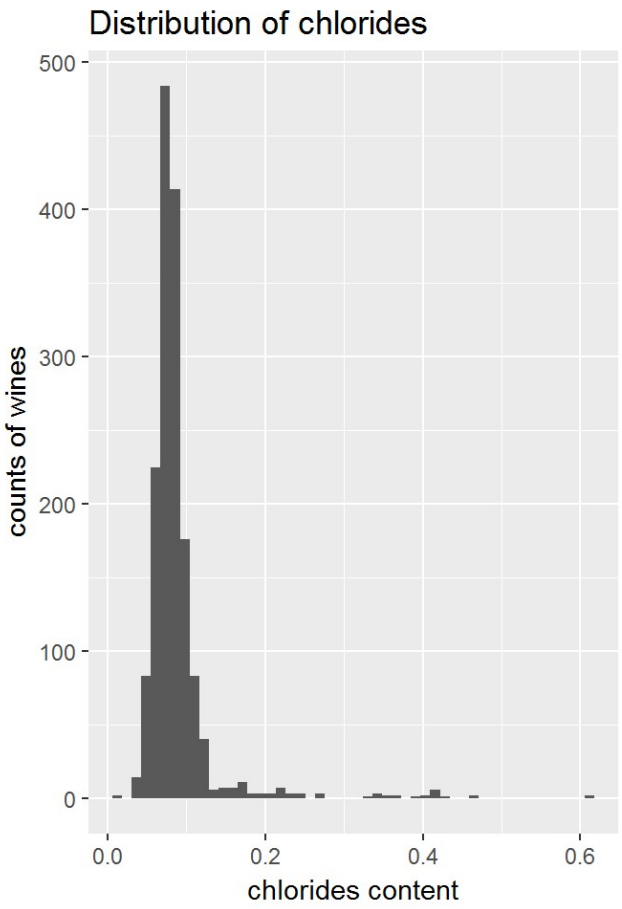
- As the description explains the relevance of sugar content. It says that most of the wines have sugar content between 1 gm/litre to 45 gm/litre. The same can be seen in the distribution also. Most of the wines have sugar content less than 4 gm/litre. Ditribution of sugar content is mostly right skewed.
- There are many outliers for residual sugar in the dataset.
- The range of residual sugar is between 0.9 to 15.50 with a median value of 2.2

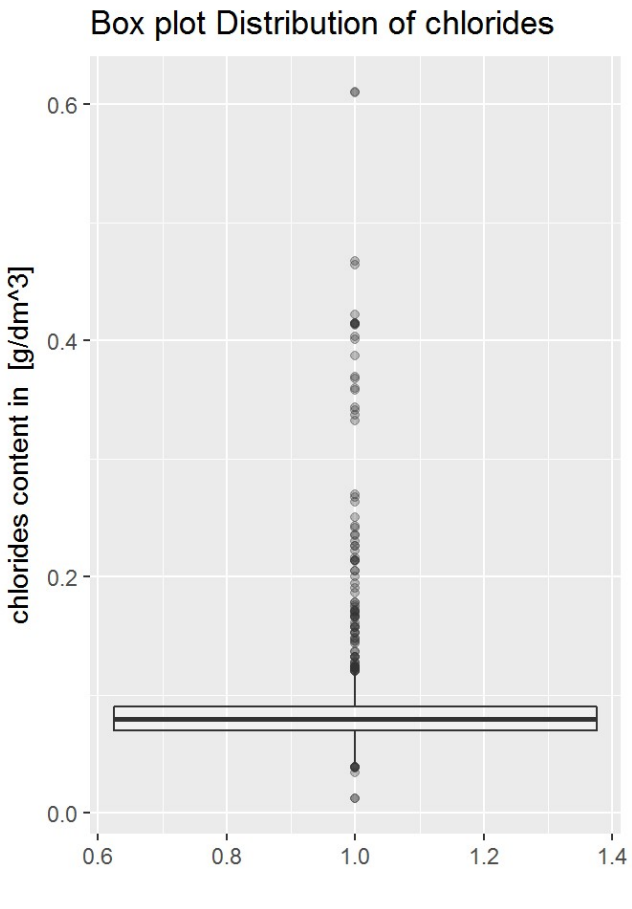





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

- There are More number of wines with sulphur dioxide content less than 20. In the histogram graph I have eleiminated the outliers and have shown ponly the top 99 percentile of data
- Since the higher the content of sulphur dioxide , the better it is for wine as it prevents oxidation of wine and acts as anti microbial
- I think this variable will have impact on the prediction of wine quality
- From the box plot above it can be seen that there are many outliers in the data for this variable
- Also the range for this variable is quite high as it ranges from a min. value of 1 to a maximum value of 72 with a median value of 14






```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

- The chloride content for most of the wines in the dataset is below 0.2 .
- Also from the boxplot it can be seen that there are quite many outliers.
- The range of values for chlorides variable is betwen 0.012 to 0.6110 with a median value of 0.079.

```
##  [1] "X"                   "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"         "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide" "total.sulfur.dioxide" "density"
## [10] "pH"                  "sulphates"            "alcohol"
## [13] "quality"             "quality_category"
```

1.1 What is the structure of your dataset?

- There are 1599 observations with 13 variables in the dataset. There are features like (fixed.acidity,volatile.acidity,citric acid,density, pH, alcohol, sulphate content and quality).
- All the variables here are of numeric type except quality which of integer type.

Few observations about the data

- Most of the wines are of average quality.
- This dataset contains data about only a sample of red wines.
- It can be seen that pH , Density are normally distributed.
- Looking at variables like alcohol content, fixed acidity, volatile acidity, residual sugar content,sulphur dioxide and chloride content are positively skewed.

1.2 What is/are the main feature(s) of interest in your dataset?

- The main feature of interest for this dataset is the wine quality.
- I want to find out what all variables impact the wine quality. From the above univariate analysis I think that pH,sulphates contet and alcohol directly impact the quality of the wines.
- In the multivariate analysis I will try to find out if any more variables can be used to predict the quality of wine.

1.3 What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

- Apart from pH, sulpahates and alcohol content which play a major role in determining the quality of red wine. Variables like acidity(fixed and volatile) and sugar content can also play an important role in predicting the quality rating.
- I can say this by looking at the similarity of distribution of quality and acidity(fixed and volatile) ,residual sugar content.

1.4 Did you create any new variables from existing variables in the dataset?

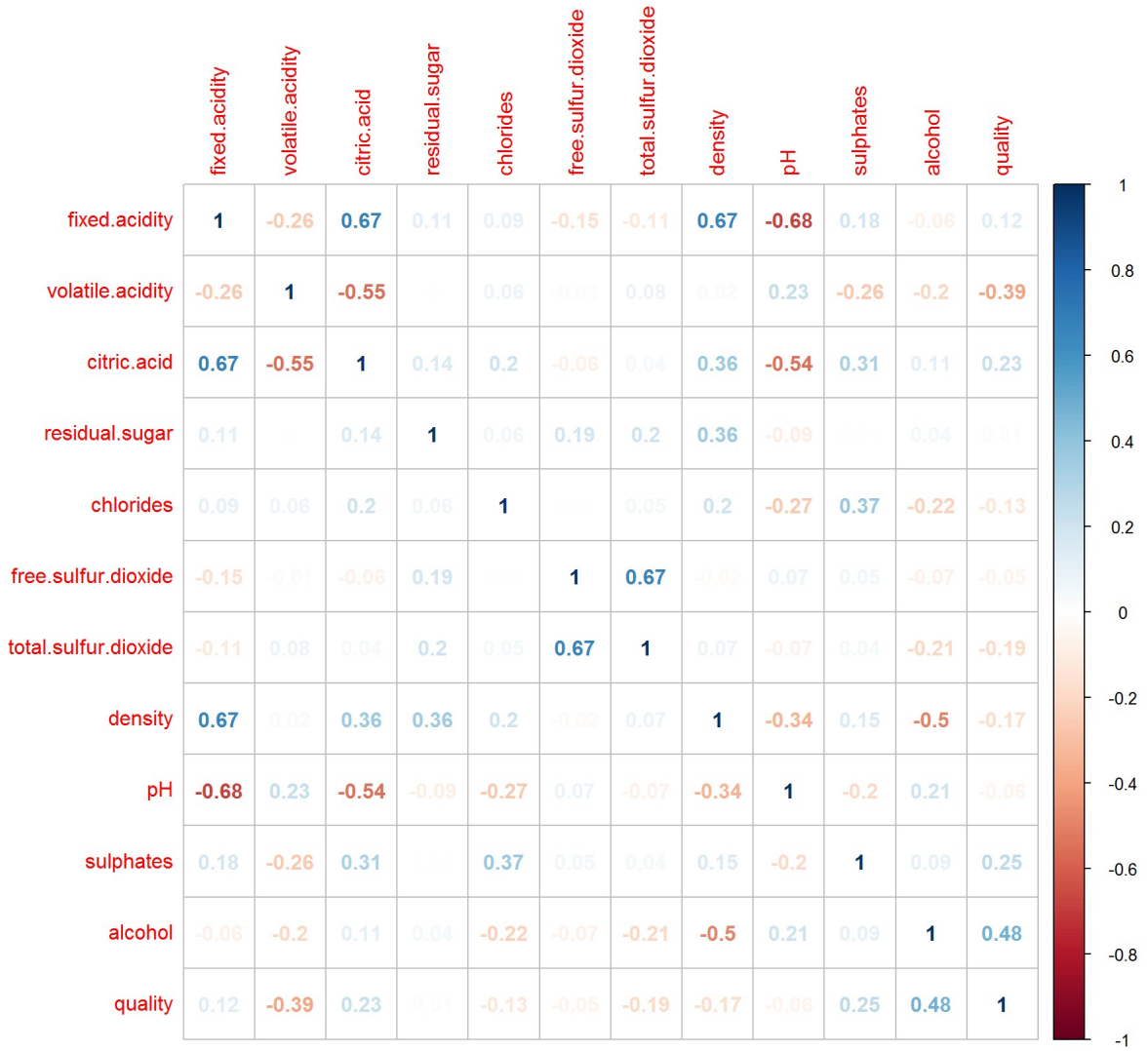
- Yes, I have added one variable to the dataset with the quality_category which has 3 categories Low, Medium and Good. I have split the wine quality based on the below condition:
- If the quality ratings is less than 4 then its bad quality , it its between 5 and 6 then Medium else Good quality.

1.5 Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

- There were few variables like fixed acidity , colatile acidity , residual sugar content which were marginally right skewed.
- I haven't done any specific transformations to the data apart from adding a new variable.

2 Bivariate analysis

We will be further checking on the relationships between each of the variables with quality variable in the Bi variate analysis

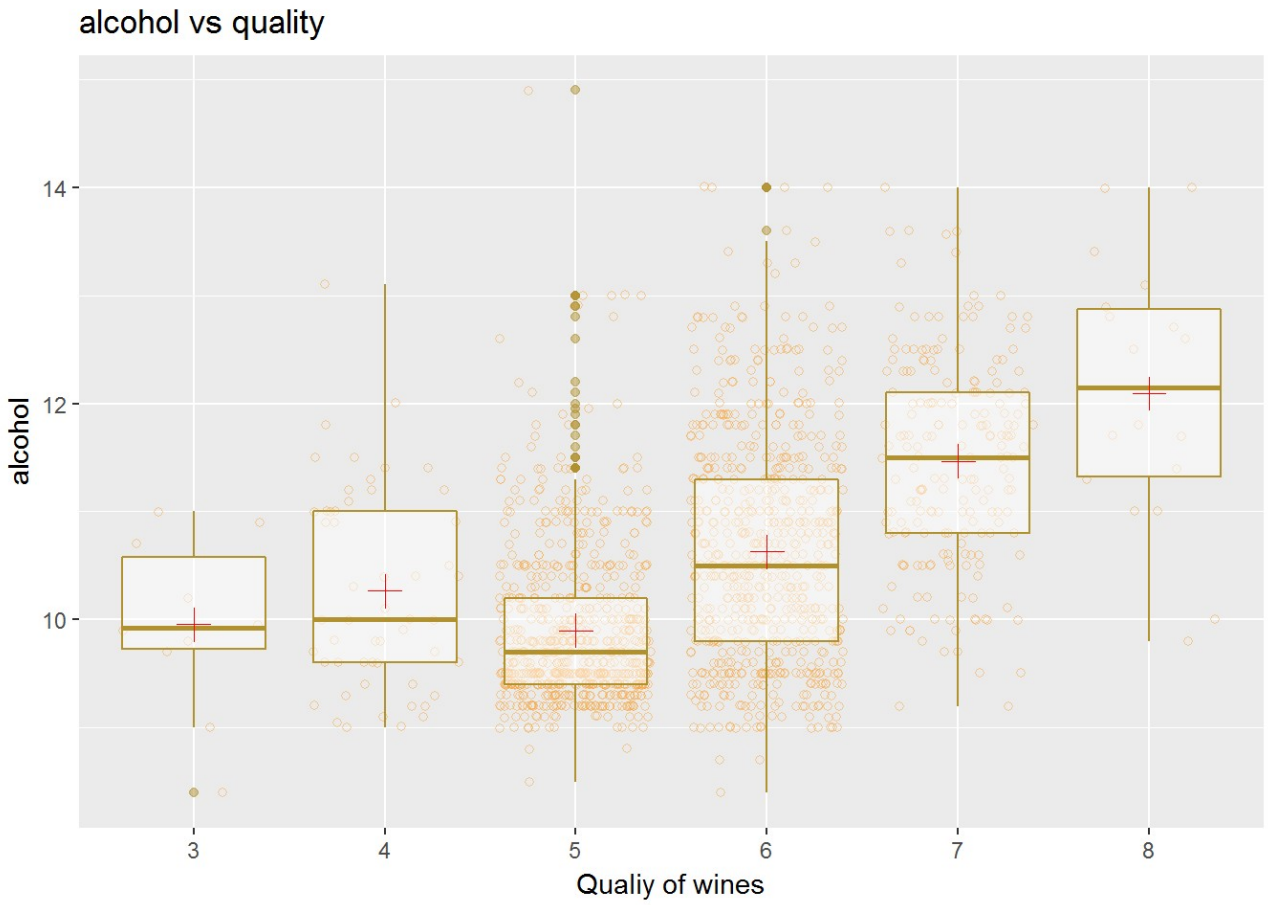


From the above correltion plot following are my observations about correaltion pf other variables with quality variable:

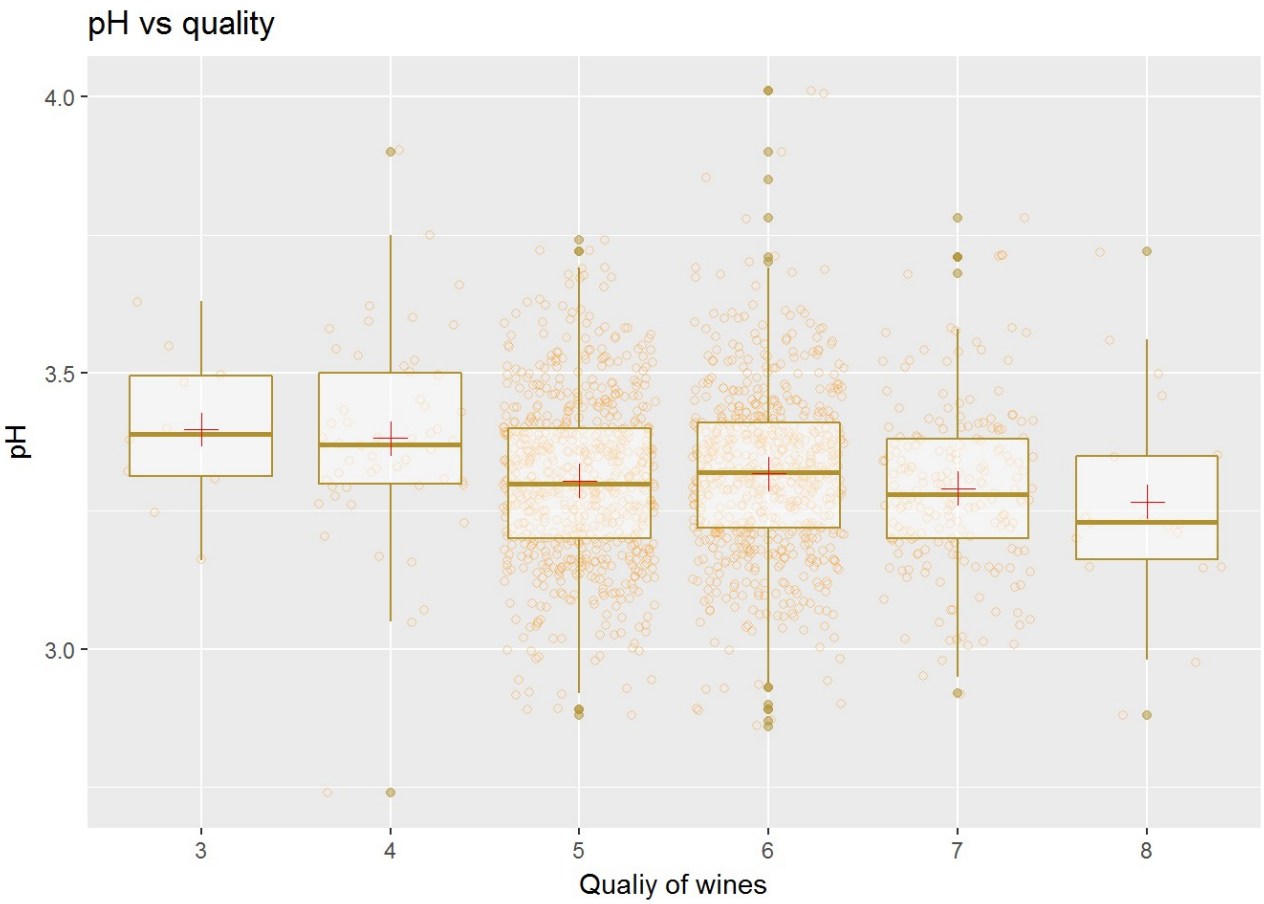
- quality has moderately negative correlation with volatile acidity
- quality has moderately positive correlation with alcohol content
- quality has weak positive correlation with sulphate content
- quality has weak positive correlation with citric acid

Let's check if there is any correlation amongst the other variables themselves:

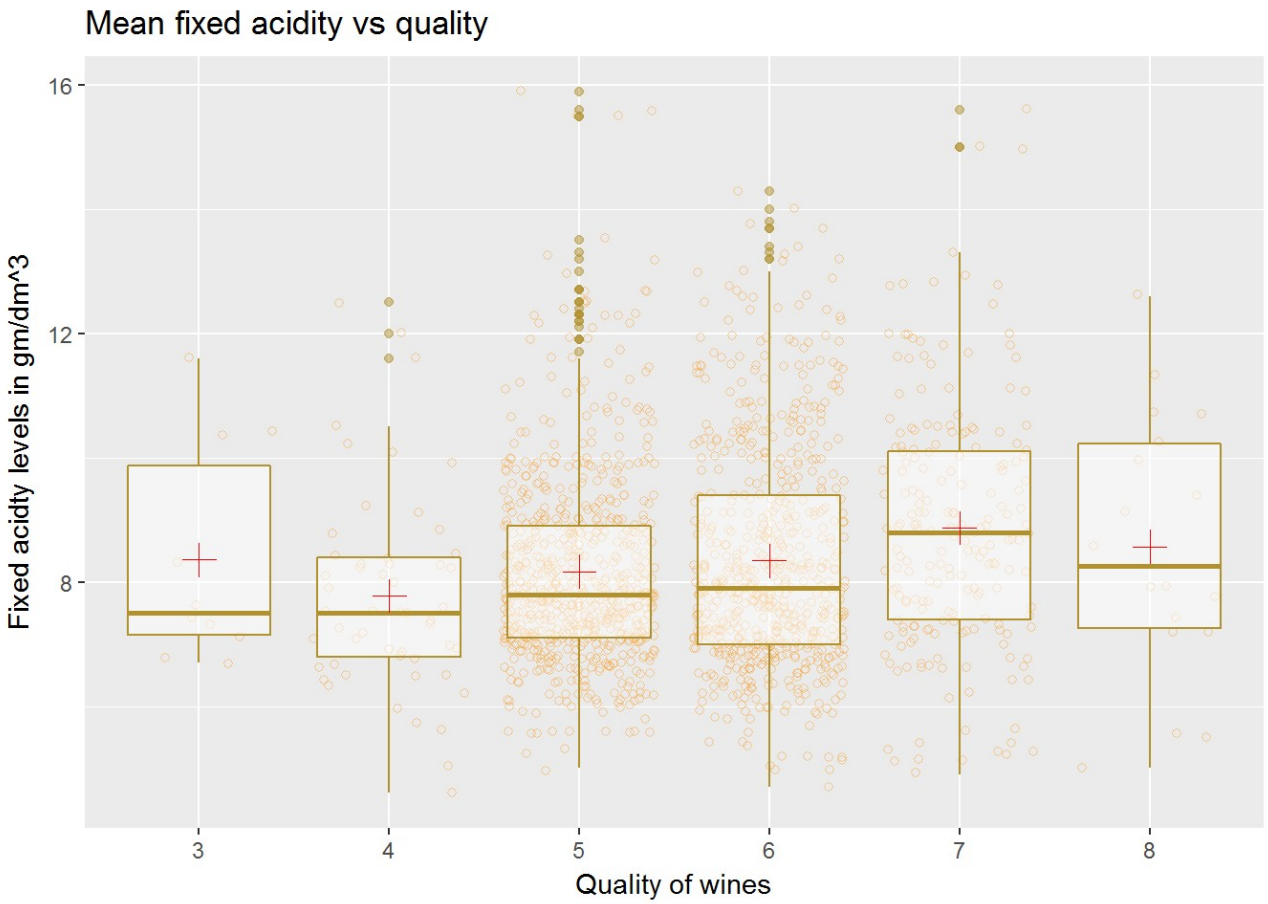
- Fixed acidity and citric acid have nearly strong correlation between them
- Fixed acidity has nearly strong correlation with density
- Fixed acidity has a negative correlation with pH value which is obvious because lesser the pH value higher the acid strength
- Citric acid has a moderately negative correlation with volatile acidity and pH
- Free sulfur dioxide and total sulfur dioxide have a positive correlation with each other.
- Chloride has a moderately positive correlation with sulphate



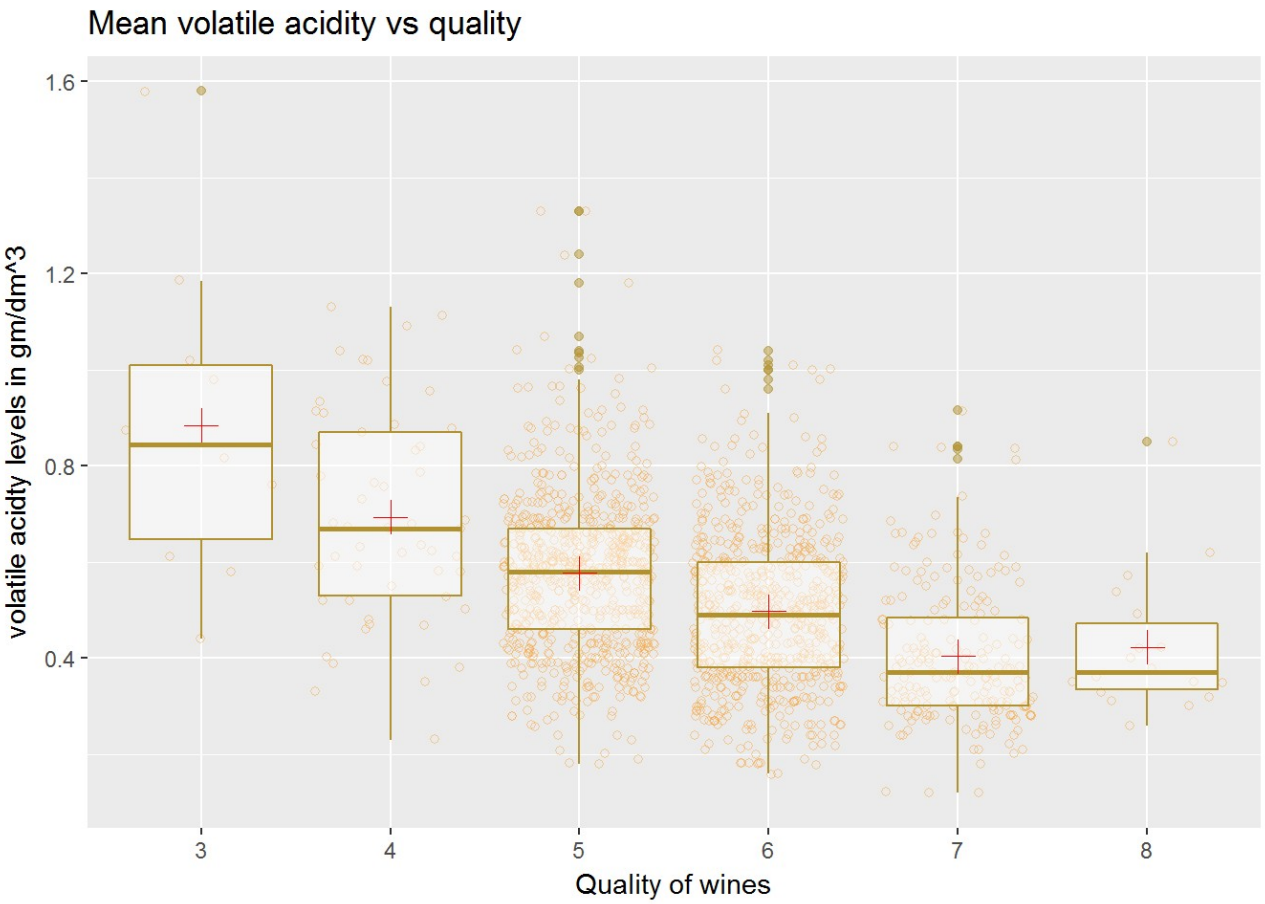
- It can bee seen from that as the mean alcohol content increases the quality of alcohol increase .
- There is a slight dip in the alcohol content for wines with rating of 5 when compared with 4 - rated wines.
- There is a huge difference in the mean alcohol content for alcohols with ratings of 7 and 8. It almost follows a very linear trend for wines with ratings of 6 and above.



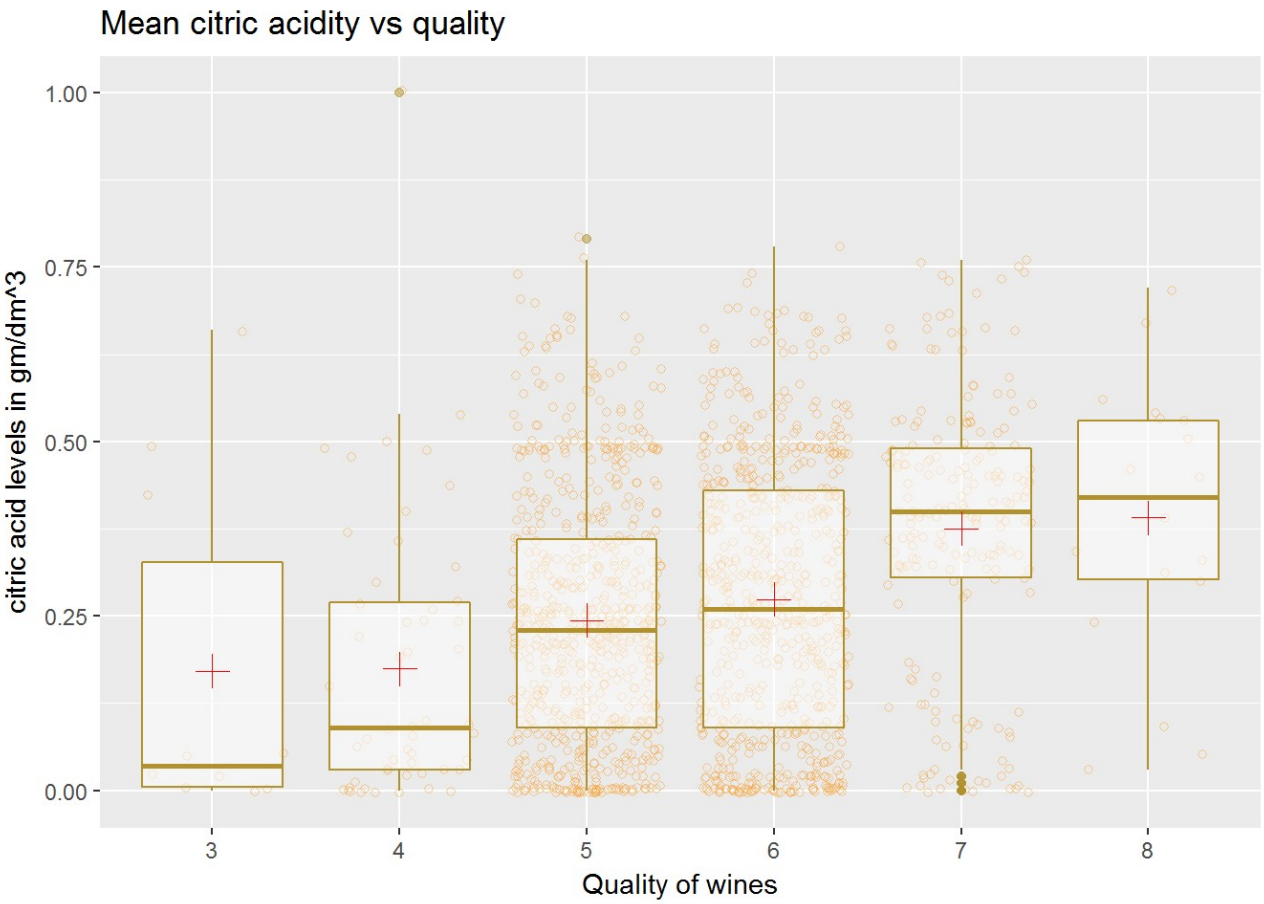
- The pH value seems to be constant for all qualities of wines. There is only a slight difference in the mean pH value. It can be noticed that the pH value is lowest for wines with quality rating of 5 and the wines with quality rating of 3 has the highest pH value.



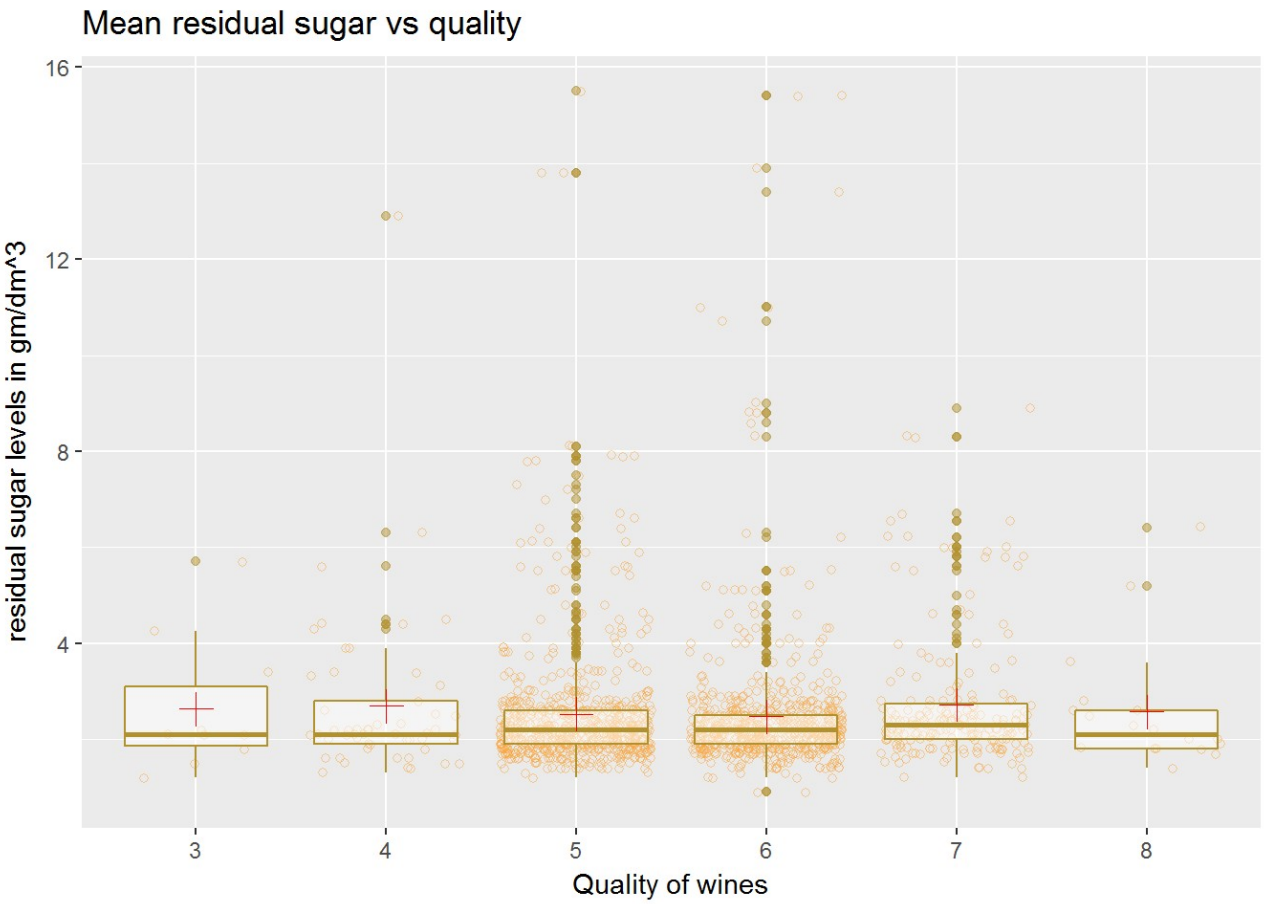
- It can be seen that the mean fixed acidity levels are gradually increasing for wine quality of 4 till 7
- The mean fixed acidity levle of wines with quality rating of 8 is lesser than the wines with quality rating of
- Also, the least mean for fixed acidity is for the quality rating of 3



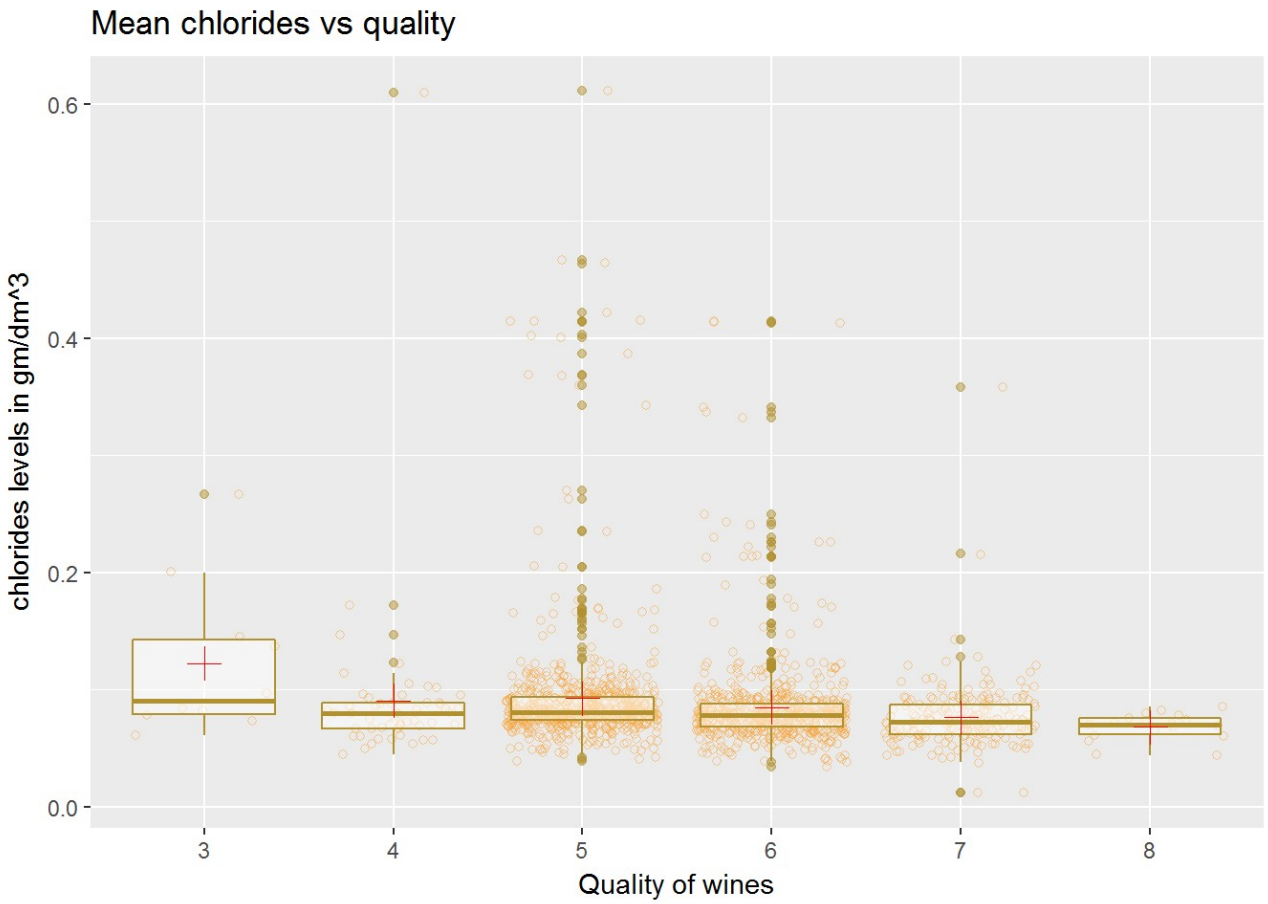
- It can be interepreted from the above graph that as the wine quality increases there is a fall in the mean volatile acid content.
- There is only a slight uptick in the mean volatile acid content for wines with quality rating of 8 as compared to that of wines with rating of 7.
- Let's see how citric acid content affects the quality of wine



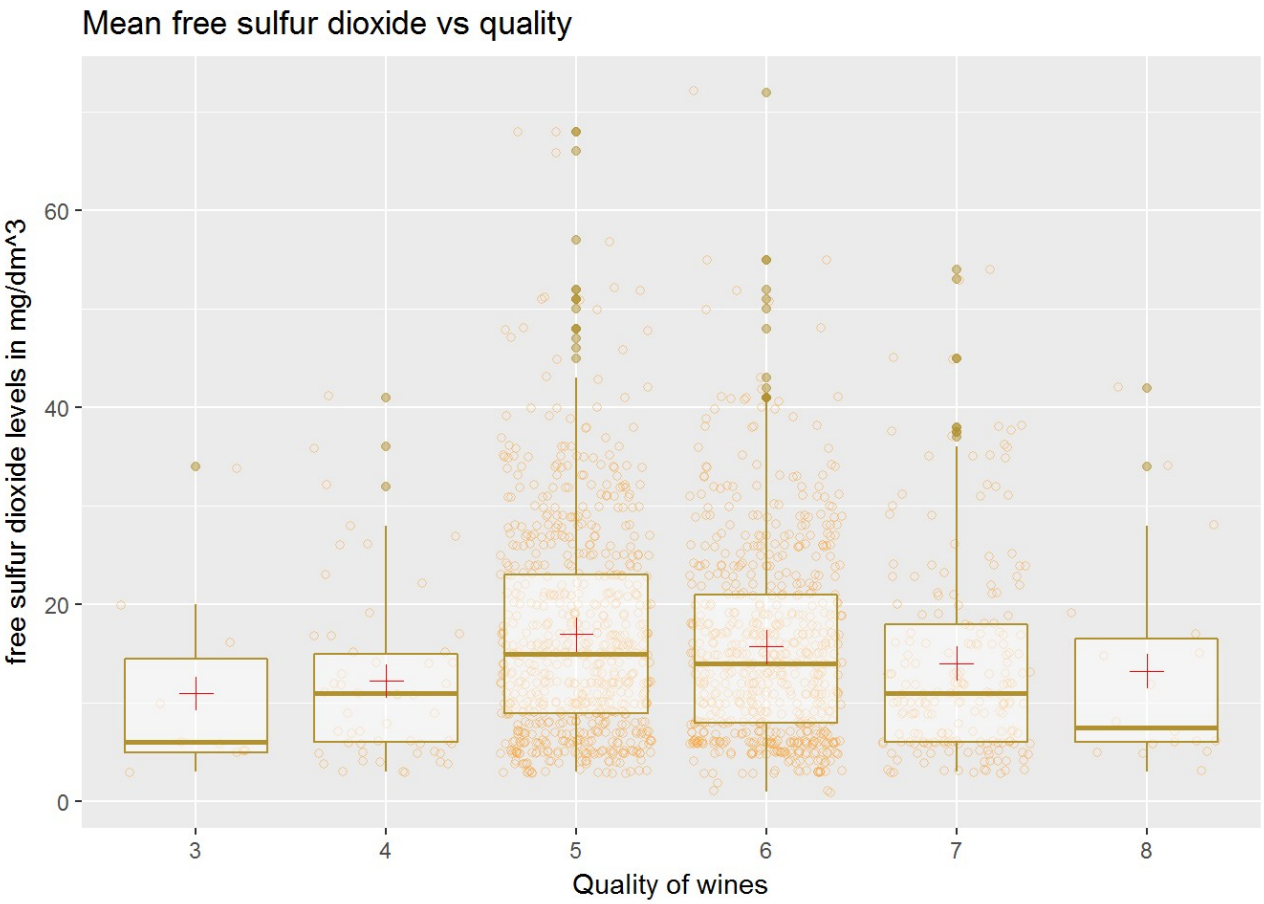
- As seen from the correlation plot where there is a weak positive correlation between the citic acid and quality of wines.
- The same can be noticed in the above plot also, where as the wine quality increases the citric acid content also increases.
- The affect of citric acid is almost linear for wines with quality ratings of 4 through 7.
- There is a very little uptick for mean value of citric acid for wines with quality rating of 8 when compared with wines with quality rating of 7.



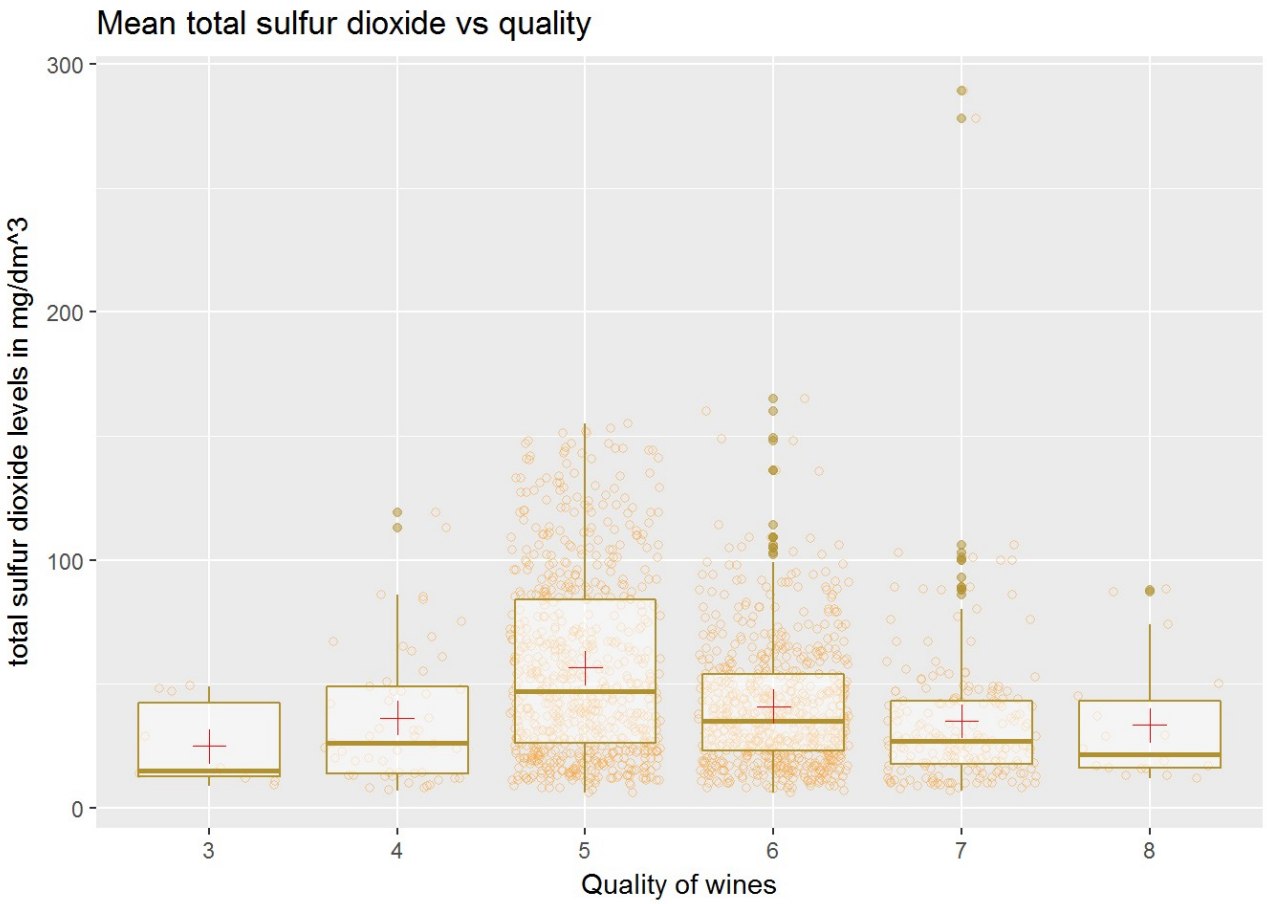
- Mean Residual sugar has remained constant for all quality ratings of wines. It's slightly more in case of wines with quality ratings of 7 when compared with all other wines. It is least for wines with quality ratings of 5.



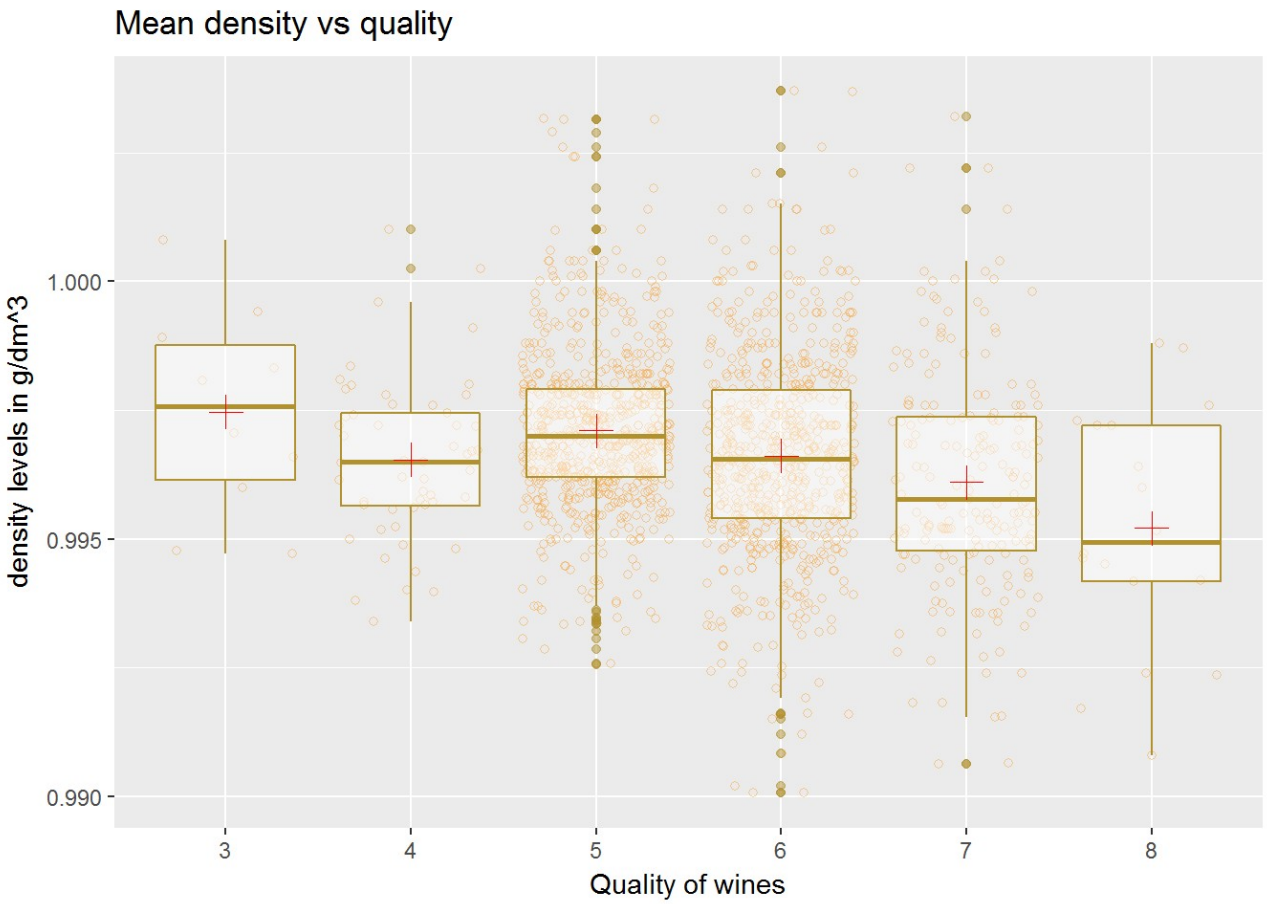
- As seen from the correlation plot there is a weak negative correlation between Quality of wines and Chlorides.In this plot it can be seen that mean value of chlorides is gradually decreasing for as the wine qualities increase.



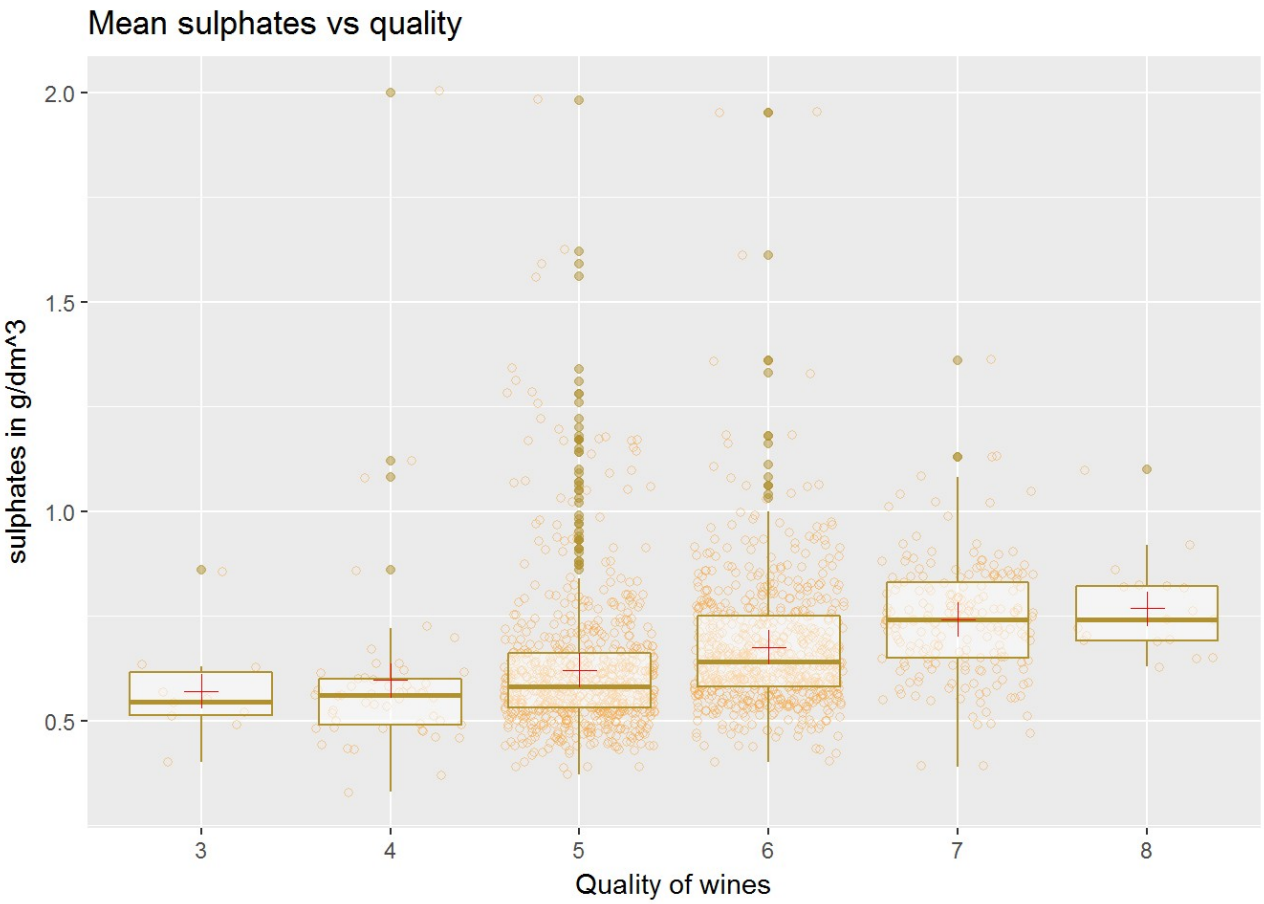
- Mean value of Free sulphur dioxide is highest for the wines with quality ratings of 5 and for all other high rated wines there is a fall in the amount of mean free sulphur dioxide. This value is lowest for the wines with quality rating of 3.



- Mean value of total sulfur dioxide follows the same trend as free sulfur dioxide , it is highest for the wines with quality ratings of 5 and for all other high rated wines there is a fall in the amount of mean total sulphur dioxide.
- This value is lowest for the wines with quality raing of 3.

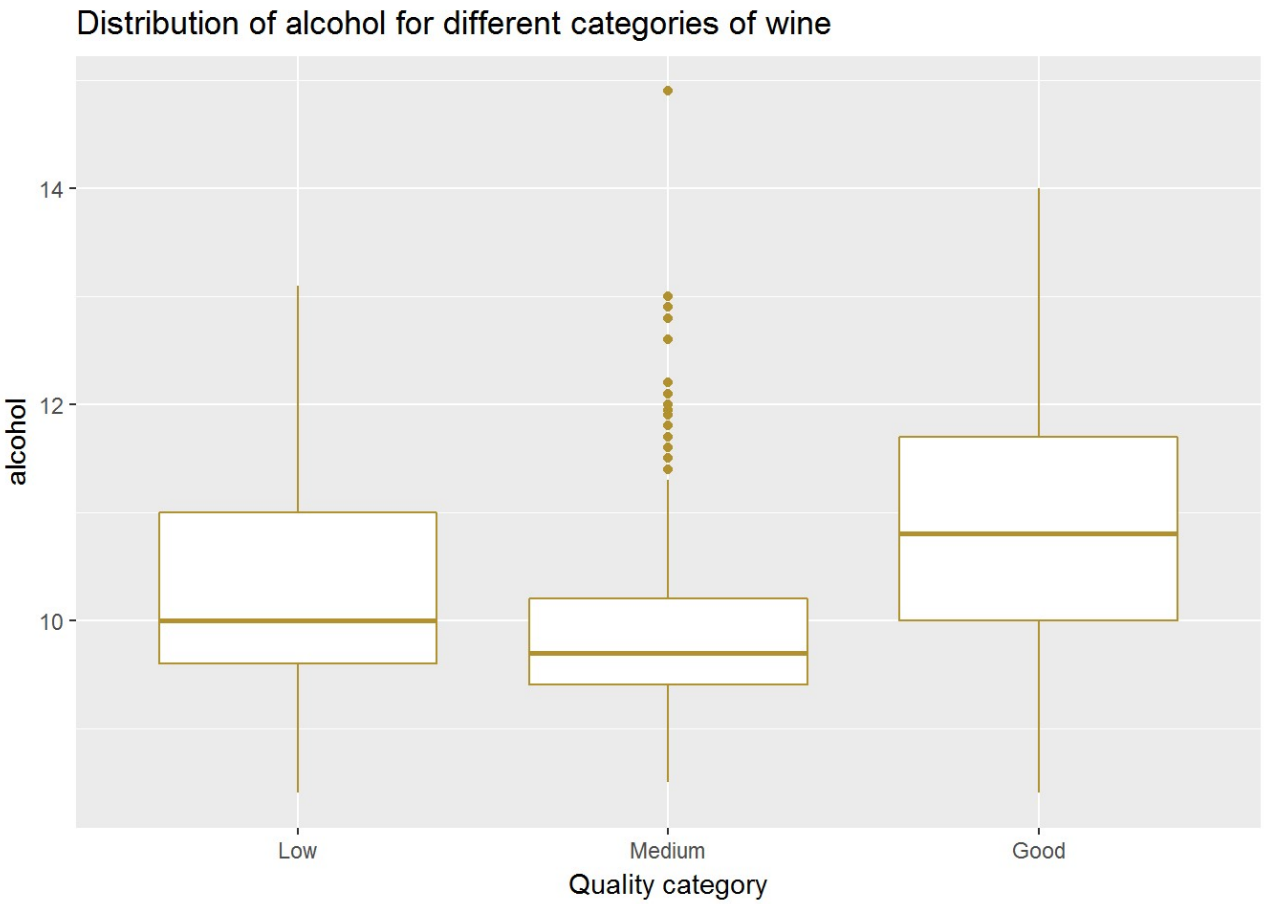


- Density of alcohol is lower than the density of water and as we have seen earlier that higher the quality wines have greater mean alcohol content. Hence in this plot where we compare the mean density wrt. qulaity of wines , there is a gradual fall in the density levels (as the aclohol content is higher) as the quality of wine increases.

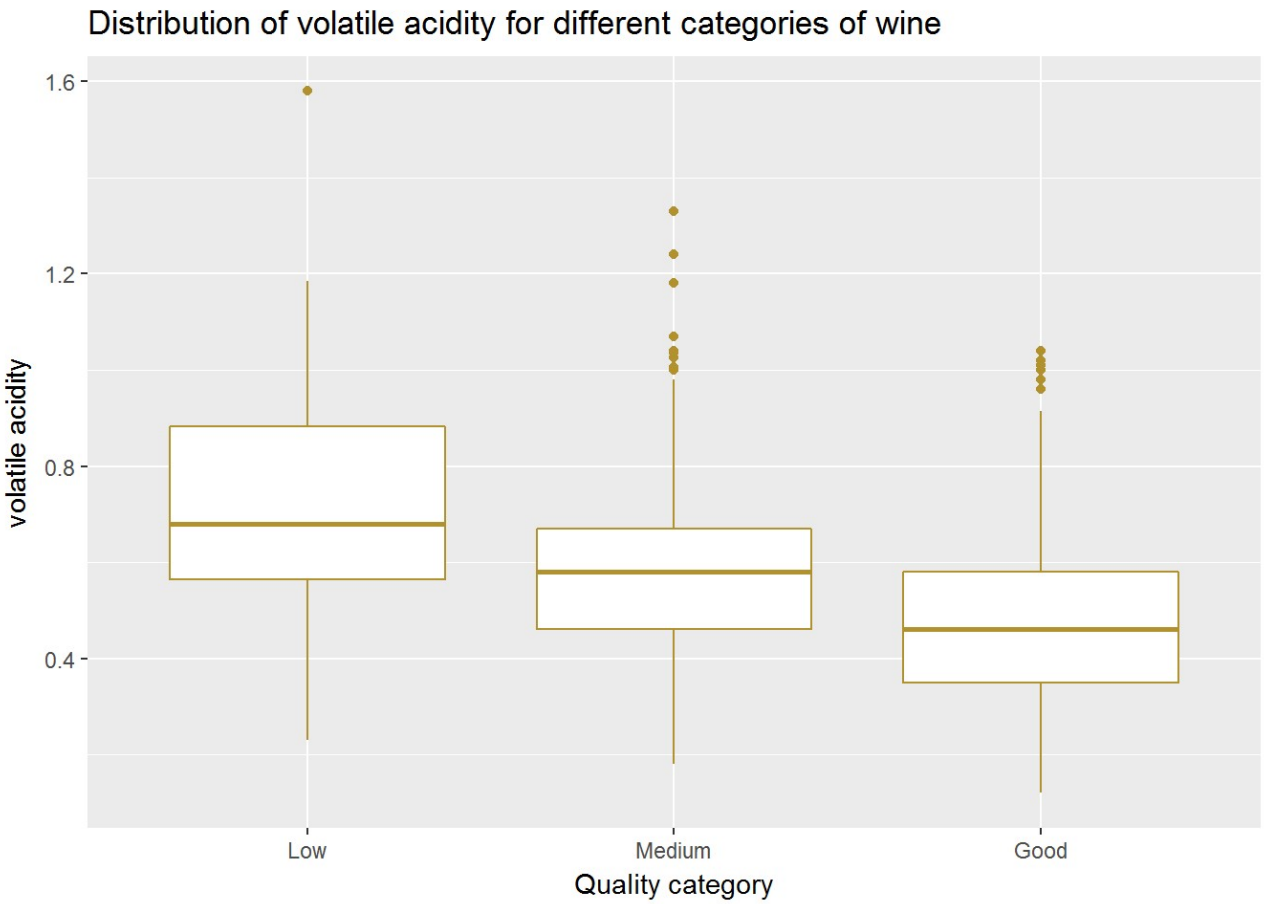


*From the above plot it can be seen that the mean sulphates content is constantly increaing as the quality rating increases. In the correlation plot also it can be seen that there was a weak correlation between these two variables.

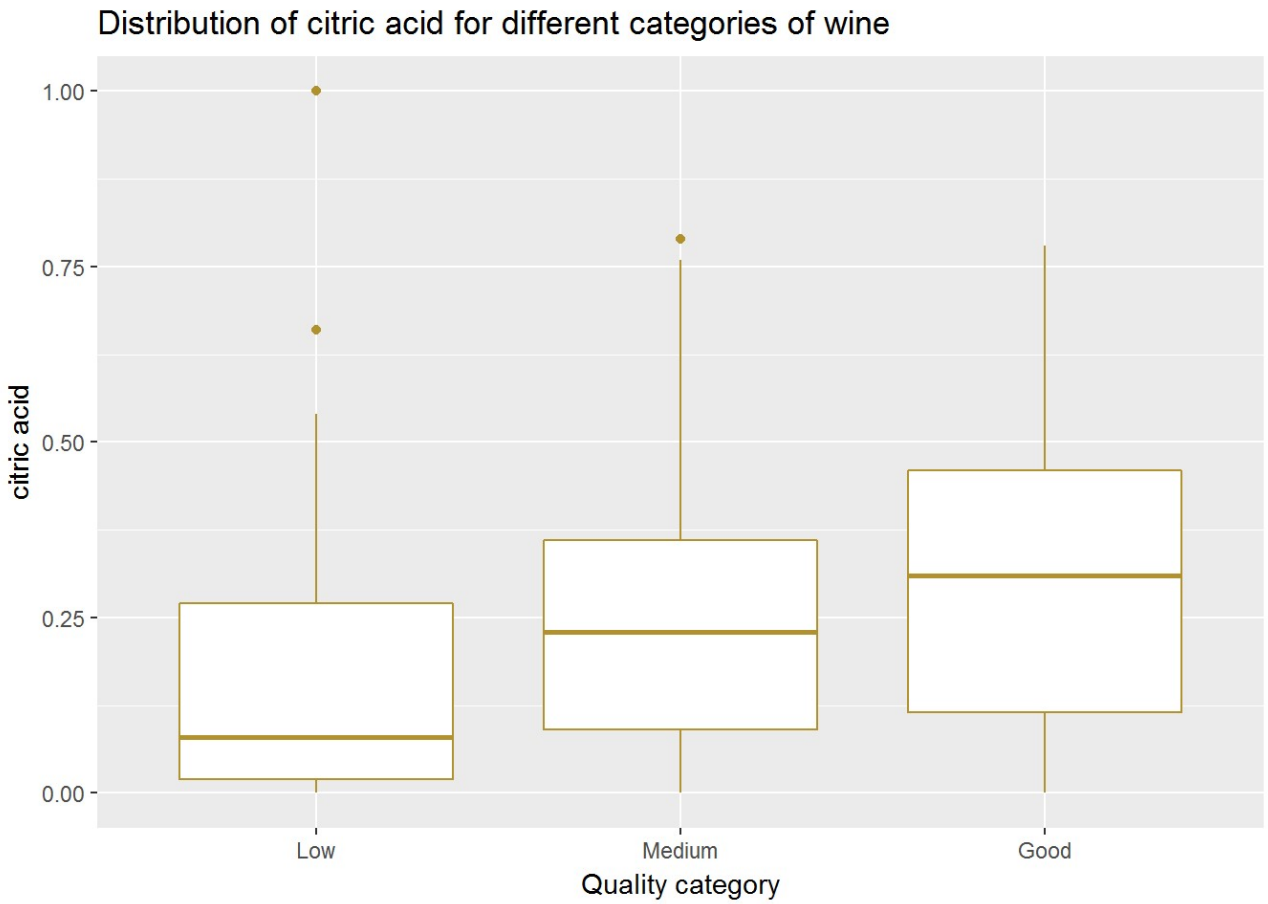
- Sulphate content is lowest for the wines with quality rating of 3 and is highest for the quality ratings of 8.
- Let’s check the volatile acidity, sulphates , citric acid and alcohol content with few box plots. For the box plots I will be using the quality_category which is the categorical variable that I have created.



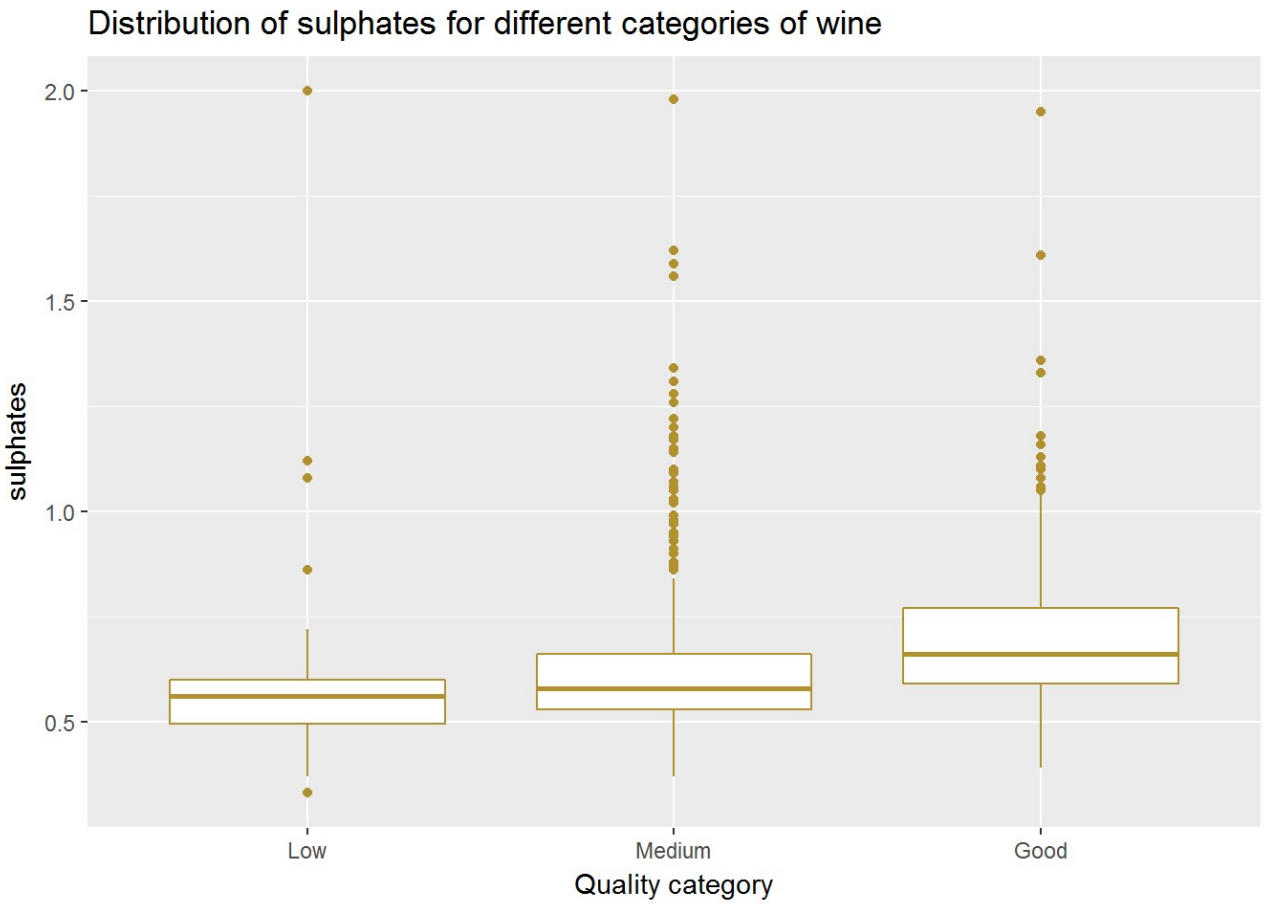
- It can be seen from the above plot that the median alcohol content for low quality wine is higher than the medium quality wine.
- The median alcohol content of Good quality wine is significantly higher than the other two categories.
- There are many outliers for the Medium quality wine in terms of alcohol content.



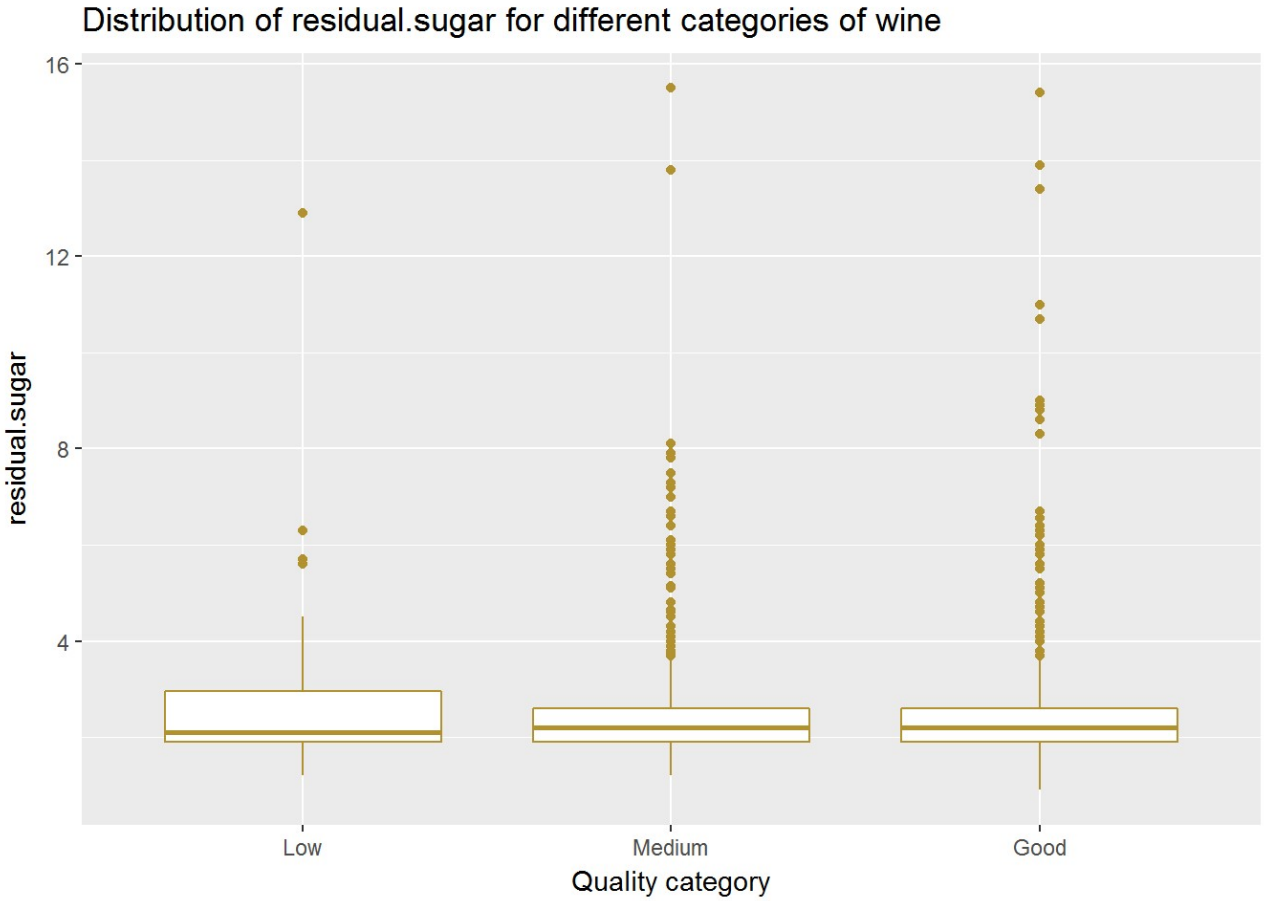
- As expected the median volatile acid content is low as the quality increases. Good quality wines have the least volatile acid content. There are many outliers for the medium category of wines.



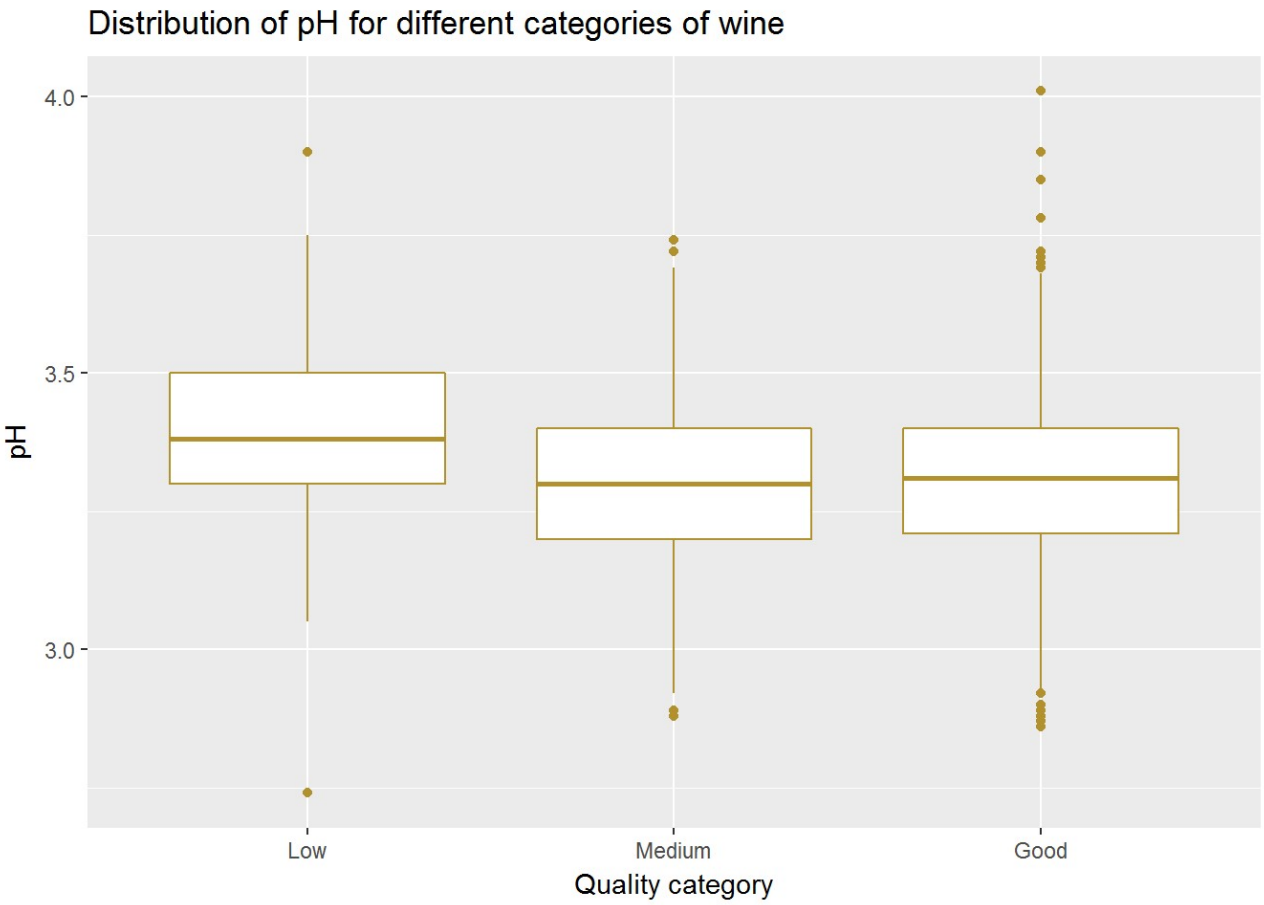
- The distribution of citirc acid is a bit interesting , although there is a mild positive correaltion between the two varibales. In the plot above it can be seen that there is a significant difference in the median citric acid content.
- The good quality wine has the highest median citric acid followed by Medium and Low in that order.
- Also there are only two outliers whcihc an be spotted for citric acid content and those are for the wines in Low quality category.



- Distribution of sulphates also shows that as the quality increase the median value of sulphates also increases in the wine. There are outliers for sulphates for each of the categories.



- From the above plot it can be seen that there is no significant difference in the median residual sugar content for each category.



- pH value of the low quality wines is higher than the other two categories of quality of wines. Medium category has a quality which is slightly less than the Good quality wines.

Let's create a model which will predict Quality of wines based on the alcohol content.

```
##
## Call:
## lm(formula = quality ~ alcohol, data = red_df)
##
## 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)  1.87497    0.17471   10.73   <2e-16 ***
## alcohol      0.36084    0.01668   21.64   <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
```

- As seen the above model has R squared value of 0.2263 which means that this alcohol can describe only 22.63 % of variance in quality.
- Since in our dataset the quality variable is more like a categorical variable which has a discrete value. Linear model creation may not be the correct way to build a model. Hence we have a very low R-squared value for this model.
- A classification model can be built to predict the correct quality value .

2.1 Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

- In this section I did some bi variate analysis and found out that quality is moderately positively correlated with alcohol content.
- I have also noticed that quality is moderately correalted with volatile acidity.
- I have created a linear model to predict the quality using alcohol content and it explains 22.63% variance in quality.
- Red wines with high levels of citric acid seem to have better quality ratings.

2.2 Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

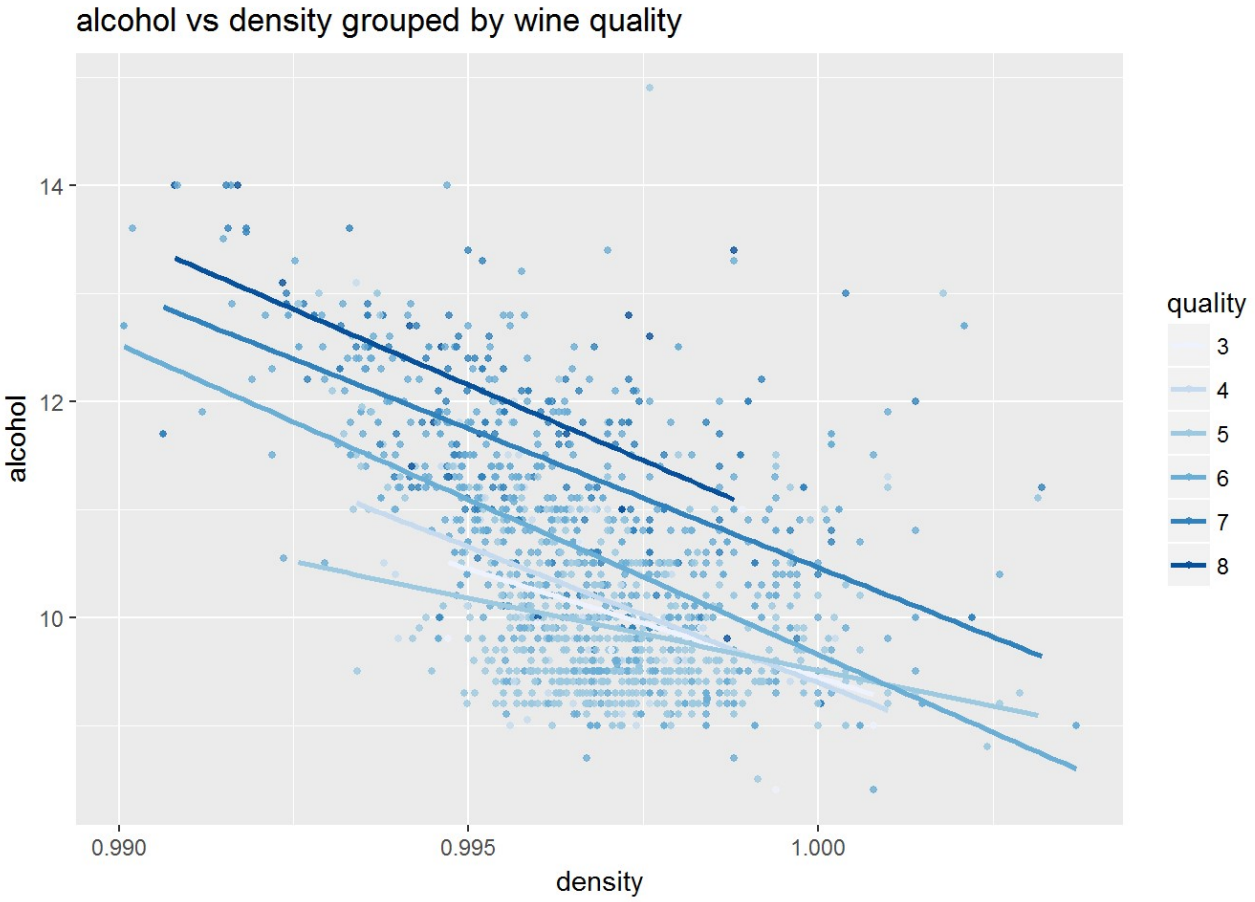
- It can be noted that mean sulphates and mean citric acid have a positive impact on the quality of wines.
- There is an increase in the mean sulphate and citric acid content as the quality of wine increases.
- Also, Density of the wine decreases as the quality of wine rating increases.

2.3 What was the strongest relationship you found?

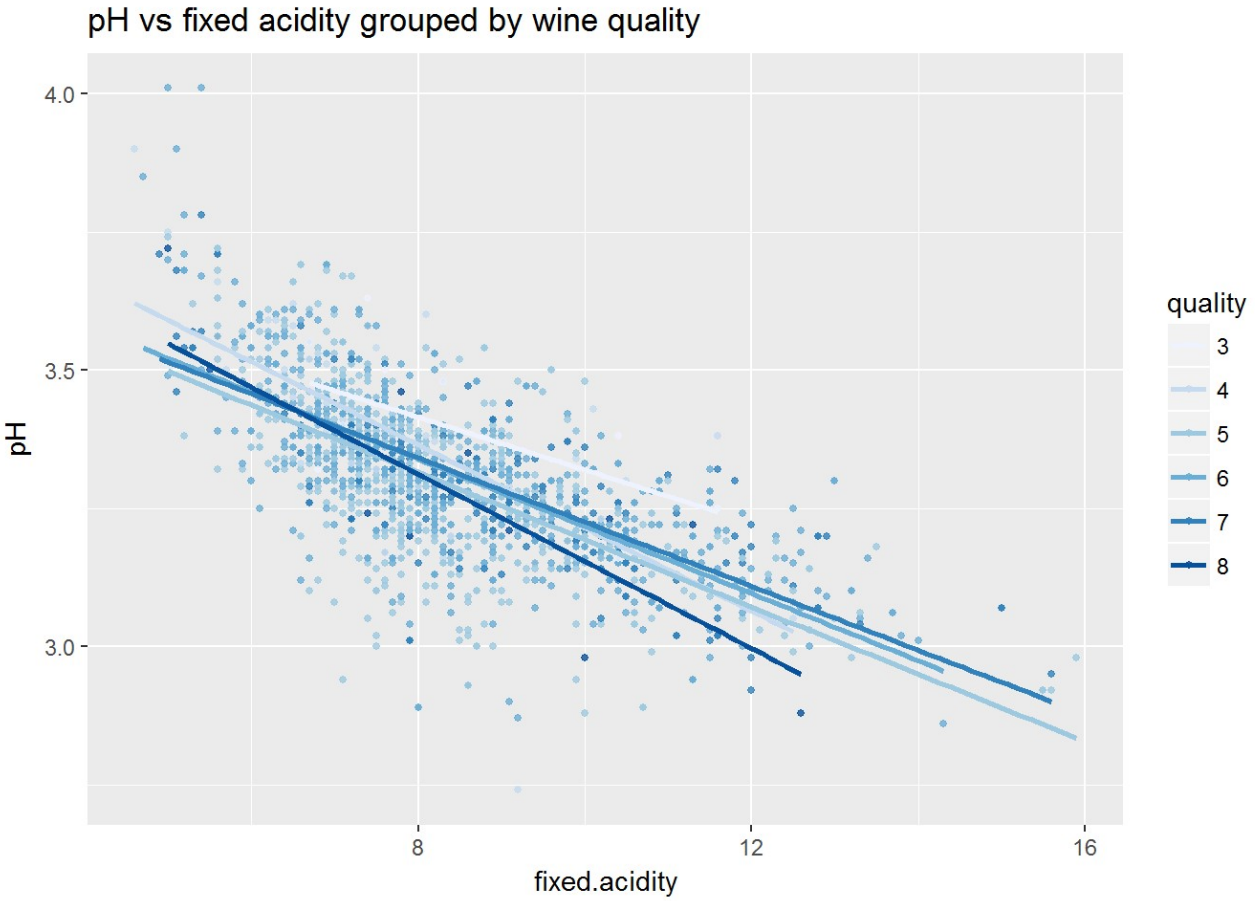
- Strongest poitive correlation was between fixed acidity and Citric acid , density and fixed acidity, and total sulfur dioxide and free sulfur dioxide.
- Strongest negative correlation was between pH and fixed acidity , citric acid and pH and Density and alcohol.

3 Multivariate Plots Section

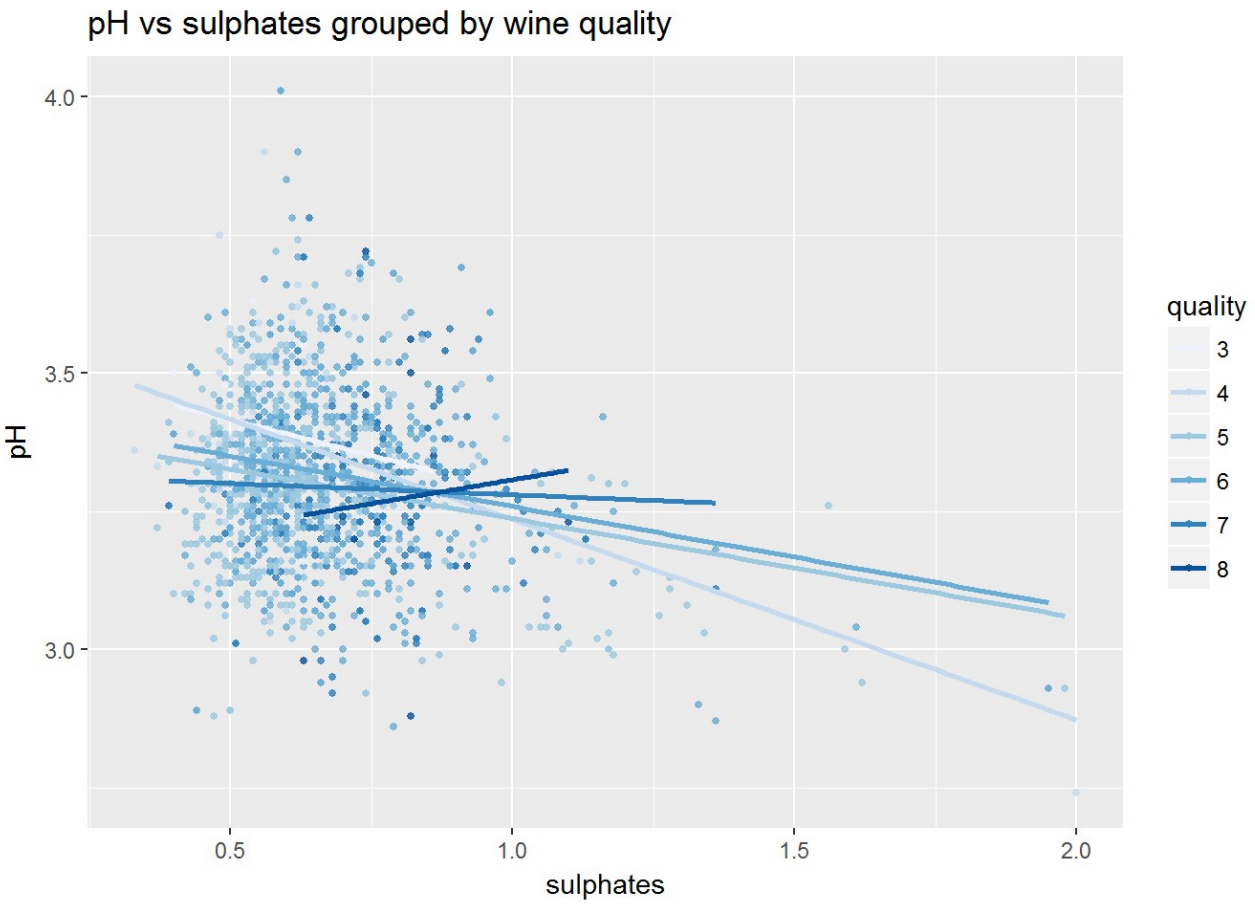
Let's create few multi variate plots



- In the above plot I am comparing alcohol content vs density and each point is denoted by the quality of the alcohol.
- It can be noticed that there is an inverse relationship between alcohol and density by looking at the regression line. It be seen that most of the 5 and 6 quality wines are around the mean density and mean alcohol range.
- As per the correaltion plot higher the alcohol content higher should be the rating. But here in this plot it can be seen that the highest alcohol content is for the wine with rating of 5 and it has an alcohol content of almost 15.
- There are few wines with ratings of 6 and above which have an alcohol content of 14.
- Also, there are few wines with low density and high alcohol content and these wines are having quality rating of 6 and above.
- The highest density wine and the lowest density wine has quality rating of 6.



- There is an inverse relationship betwen pH and Fixed acidity. It is quite obvious by looking at the inverse slope of the regression line.
- Wines with quaity of 6 are having highest pH and hence the lowest acidty levels.
- There are wines with quality ratings of 5 with highest acidty.
- There is a wine of quality 4 which has a very low pH value and has acidity level of around 9.



- pH and Sulphates also have an inverse relationship. There are only a handful wines with sulphates content greater than 1.5 gm/dm^3.
- All Of the wines with sulphate content greater than 1.5 have ratings of 5 and 6.
- Eventhough adding sulphates is benificial for wines in terms of preservation , very few high quality wines have high content of Sulphates.

```
## red_df$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.700   7.150   7.500   8.360   9.875  11.600
## ------------------------------------------------------------
## red_df$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.600   6.800   7.500   7.779   8.400  12.500
## ------------------------------------------------------------
## red_df$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   7.100   7.800   8.167   8.900  15.900
## ------------------------------------------------------------
## red_df$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.700   7.000   7.900   8.347   9.400  14.300
## ------------------------------------------------------------
## red_df$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.900   7.400   8.800   8.872  10.100  15.600
## ------------------------------------------------------------
## red_df$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   7.250   8.250   8.567  10.225  12.600
```

five number summary of fixed acidity for each quality of wine

- Median quantity of fixed acidity is under 8 gm/dm^3 for all qualities of wines except wines with ratings of 7 and

8.

- Maximum volume of fixed acidty is highest for the wines with quality ratings of 5.

```
## red_df$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9947  0.9961  0.9976  0.9975  0.9988  1.0008
## ------------------------------------------------------------
## red_df$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9934  0.9957  0.9965  0.9965  0.9974  1.0010
## ------------------------------------------------------------
## red_df$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9926  0.9962  0.9970  0.9971  0.9979  1.0031
## ------------------------------------------------------------
## red_df$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9954  0.9966  0.9966  0.9979  1.0037
## ------------------------------------------------------------
## red_df$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9906  0.9948  0.9958  0.9961  0.9974  1.0032
## ------------------------------------------------------------
## red_df$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9908  0.9942  0.9949  0.9952  0.9972  0.9988
```

five number summary of density for each quality of wine

- Median density of wines is almost in the same range for all wine qualities.
- Wine with quality rating of 8 has the least maximum density.

```
## red_df$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0050  0.0350  0.1710  0.3275  0.6600
## ------------------------------------------------------------
## red_df$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0300  0.0900  0.1742  0.2700  1.0000
## ------------------------------------------------------------
## red_df$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2300  0.2437  0.3600  0.7900
## ------------------------------------------------------------
## red_df$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2600  0.2738  0.4300  0.7800
## ------------------------------------------------------------
## red_df$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3050  0.4000  0.3752  0.4900  0.7600
## ------------------------------------------------------------
## red_df$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0300  0.3025  0.4200  0.3911  0.5300  0.7200
```

five number summary of citric acid for each quality of wine

- As the qualitty rating increases the meadian citirc acid content also increases.
- But the maximum citric acid content is for the wine with quality rating of 5.

```
## red_df$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5125  0.5450  0.5700  0.6150  0.8600
## ------------------------------------------------------------
## red_df$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.4900  0.5600  0.5964  0.6000  2.0000
## ------------------------------------------------------------
## red_df$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.370   0.530   0.580   0.621   0.660   1.980
## ------------------------------------------------------------
## red_df$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5800  0.6400  0.6753  0.7500  1.9500
## ------------------------------------------------------------
## red_df$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3900  0.6500  0.7400  0.7413  0.8300  1.3600
## ------------------------------------------------------------
## red_df$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6300  0.6900  0.7400  0.7678  0.8200  1.1000
```

five number summary of sulphates for each quality of wine

- As the wine quality increases the median wine quality also increases.
- Wines with rating of 4 has the maximum sulphate content where as the wines with quality rating of 3 has the least sulphate content.

```
## red_df$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.160   3.312   3.390   3.398   3.495   3.630
## ------------------------------------------------------------
## red_df$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.300   3.370   3.382   3.500   3.900
## ------------------------------------------------------------
## red_df$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.880   3.200   3.300   3.305   3.400   3.740
## ------------------------------------------------------------
## red_df$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.860   3.220   3.320   3.318   3.410   4.010
## ------------------------------------------------------------
## red_df$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.920   3.200   3.280   3.291   3.380   3.780
## ------------------------------------------------------------
## red_df$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.880   3.163   3.230   3.267   3.350   3.720
```

five number summary of pH for each quality of wine

- Median pH value of wines decreases as the wine quality increases.
- Maximum pH value is recorded for wine quality of 6.

```
## red_df$quality: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.575  11.000
## ------------------------------------------------------------
## red_df$quality: 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00    9.60   10.00   10.27   11.00   13.10
## ------------------------------------------------------------
## red_df$quality: 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
## ------------------------------------------------------------
## red_df$quality: 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
## ------------------------------------------------------------
## red_df$quality: 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
## ------------------------------------------------------------
## red_df$quality: 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

Five number summary for alcohol contents for each rating of wines

- Eventhough the meadian alcohol content is increasing as the wine quality increases with an exception for wines with quality rating of 4.
- Maximum alcohol content was for the wine with quality rating of 5.

- As the fixed acidty increase in the wines the density of the wines also increases.
- Most of the wines with quality rating of 8 have low fixed acidity and low density.
- Wine with highest and lowestdensity have quality rating of 6.
- Wines with highest fixed acidity has a rating of 5 and lowest fixe acidty has a rating of 5

3.1 Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature (s) of interest?

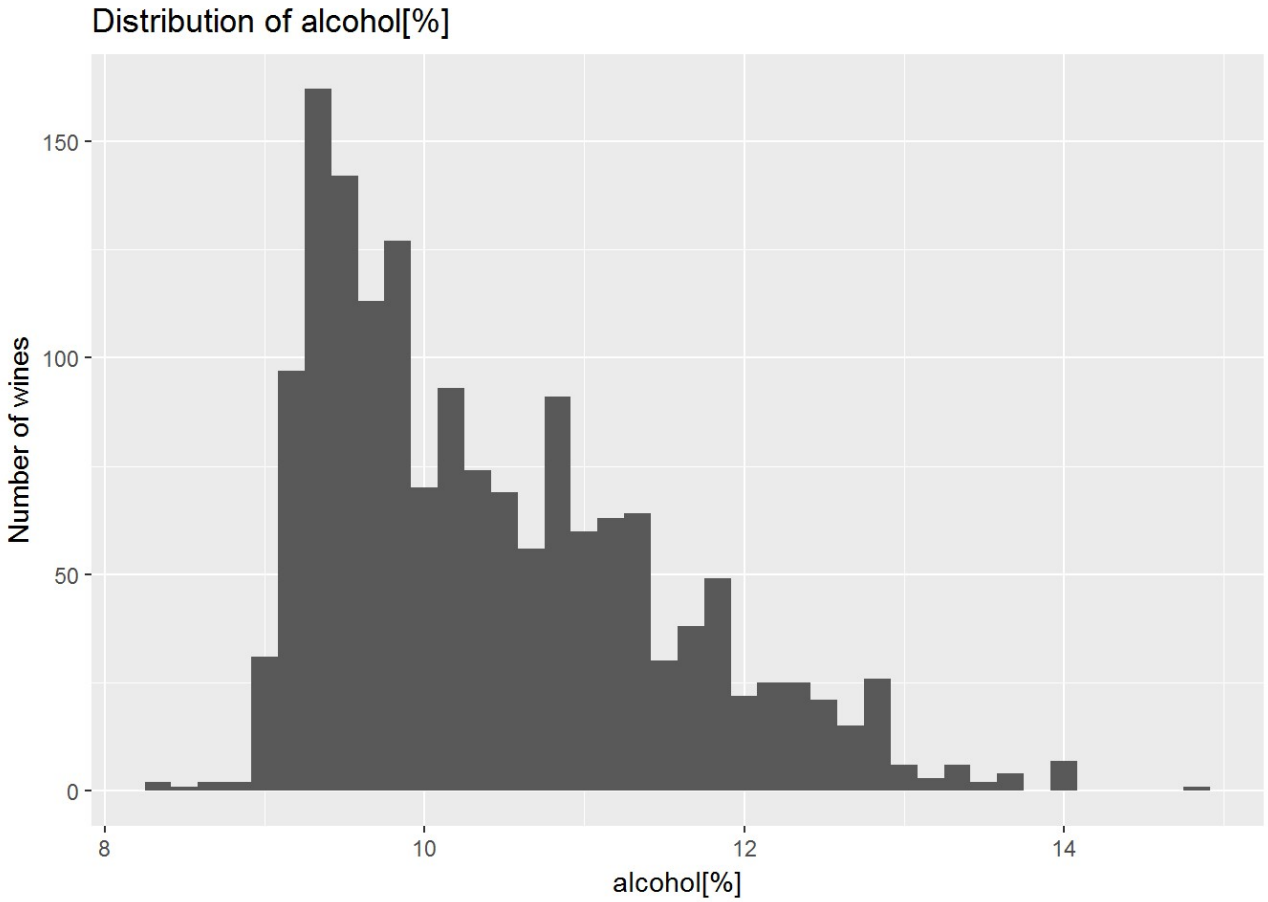
The inverse relationshp between the pH and fixed acidity is very much visible across all categories of qualities of wines. Most of the high quality wines had higher fixed acidity levels.

3.2 Were there any interesting or surprising interactions between features?

Density and alchol had a negative relationship. When I had a regression line in that plot it seems that it is exactly bisecting the wines with quality ratings of 5 and below and wines with ratings of 6 and above. Most of the low quality wines(qines with ratings below 6) are present below the regression line which means they have average density and low in alcohol content. Whereas all the good quality wines are on top of the regression line which means that they have higher alcohol content and average to above average density.

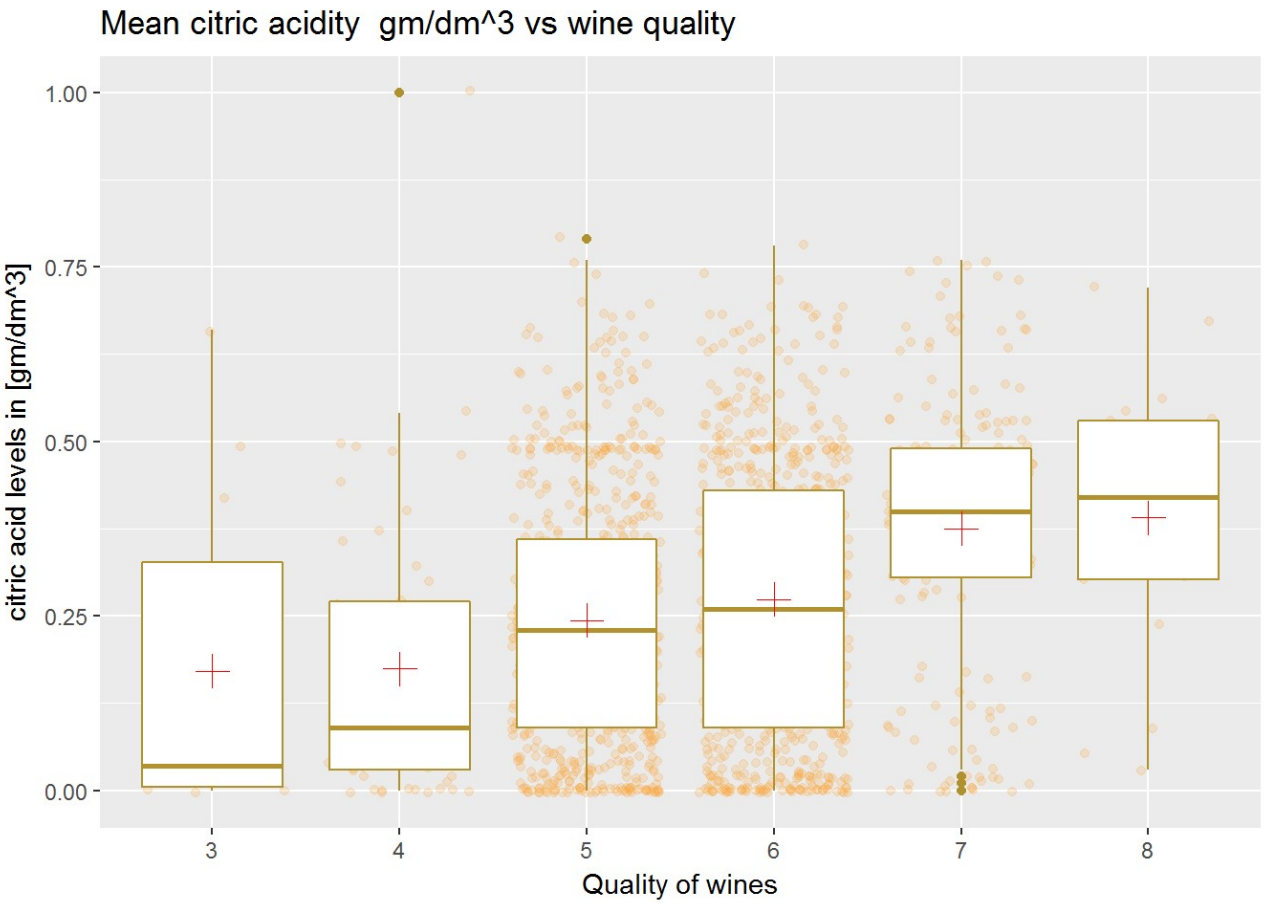
4 Final Plots and Summary

4.1 Plot One



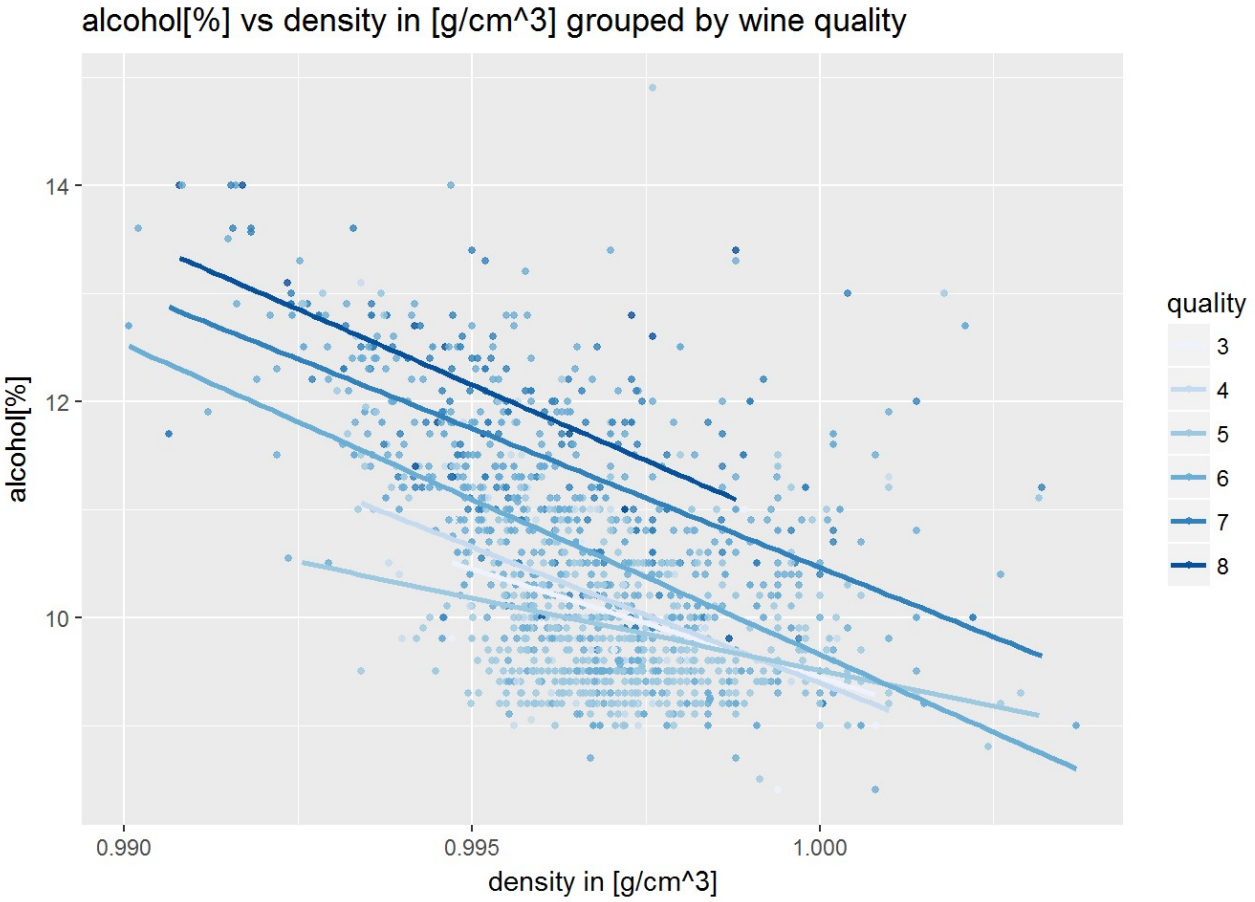
- This plot had my attention because after having prelimnary look at the dataset since wine is an alcohol based drink I was expecting that most of wines would be having high alcohol content.
- But when I look at this distribution it seems that around 50% of the wines have aclohol content of less than 10% and around 75% had less than 12% of alcohol.

4.2 Plot Two



- This plot where I compare the effect of citirc acid content on the quality of wine and it seems it has profound effect on the wine quality.
- It can be seen that the mean citric acid content increases as the quality of wine increases.

4.3 Plot Three



- I choose this graph for final plots because the way the regression line is separated for each of the wine qualities. It can be seen that regression line for wines with quality ratings of 6,7,8 are in parallel to each other. Topmost regression line is of wines with quality rating of 8 then followed by 7 then by 6.
- Most the wines with quality rating of 6 and above are above the regression line and wines with quality rating below 6 are below the regression line

5 Reflection

- The wines dataset has 1599 entries with 13 variables.
- I started my analysis by understanding the spread of each of the variable. I also checked if there are any outliers in wach of the variables by plotting box plots for each of the variables.
- I wanted to check what are the variables impacting the quality of wine. Initially I thought the alcohol content is the major factor influencing the wine quality.
- Higher the alcohol content better the rating of the wine. But as I started to explore the relationship between these two variables a bit more I found that there was a good correlation between these two variables. As the wine quality increases there is an increase in the alcohol conetent also. This relation had an exception only the wines with rating of 5 where there was a slight dip in the mean alcohol content.

- Surprisingly citric acid content had a good impact on the wine quality too. As the quality increased the mean citric acid content also increased and it was quite evident. This may be because it adds freshness to the wine and hence impacts the quality ratings.
- Then I started to look into other variables affecting the wine quality and I found that the sulphate content and pH also had a impact in determining the quality rating of wine.
- As the wine quality increased the mean pH value saw a dip. Mean Sulphate conetent was higher as the quality of wine was increasing. Although the slope was gradual but still there was an improvement as the quality of wine increased.
- The wine quality increases the density saw a dip. This might be because of the higher alcohol content as the wine quality increases.Volatile acidity also increased as the wine quality increased but it was stable for wines with quality ratings of 7 and 8.
- I created an linear model which takes into account alcohol content to predict the quality ratings. This model has a very low R-Squared value.
- Further I felt that a linear model is not very much suitable for this dataset as I am trying to predict a variable which is not continuous eventhough it has a numerical value it is more like a categorical variable.
- Hence a model like Logistic regression , K nearest nieghbour or Random forest would have been more suitable to predict the quality of wines.

6 Limitations of dataset

- Although the dataset has good number of variables for predicting the wine quality. It lacks a bit in number of observations.
- Also, most of the wines have a rating of 5 and 6. There are very few wines at the lower and higher quality ratings. This hinders drawing solid inferences about what are the differentiating factors which help in getting higher or lower quality ratings.
- All the wines realted to thsi dataset are for the wines of variants of the Portuguese “Vinho Verde” wine. It would be better if we had other red wines also included in the dataset in that way I could have analysed a bit more for different types of redwines and how the variables affect the wine quality of the wines.
- It would have been better if both red and white wine were in the same dataset ,so that I could have analysed the impact of different variables on the wine quality for each red and white wine.