RED WINE QUALITY ANALYSIS by Prathmesh Kumbhare

Introduction:

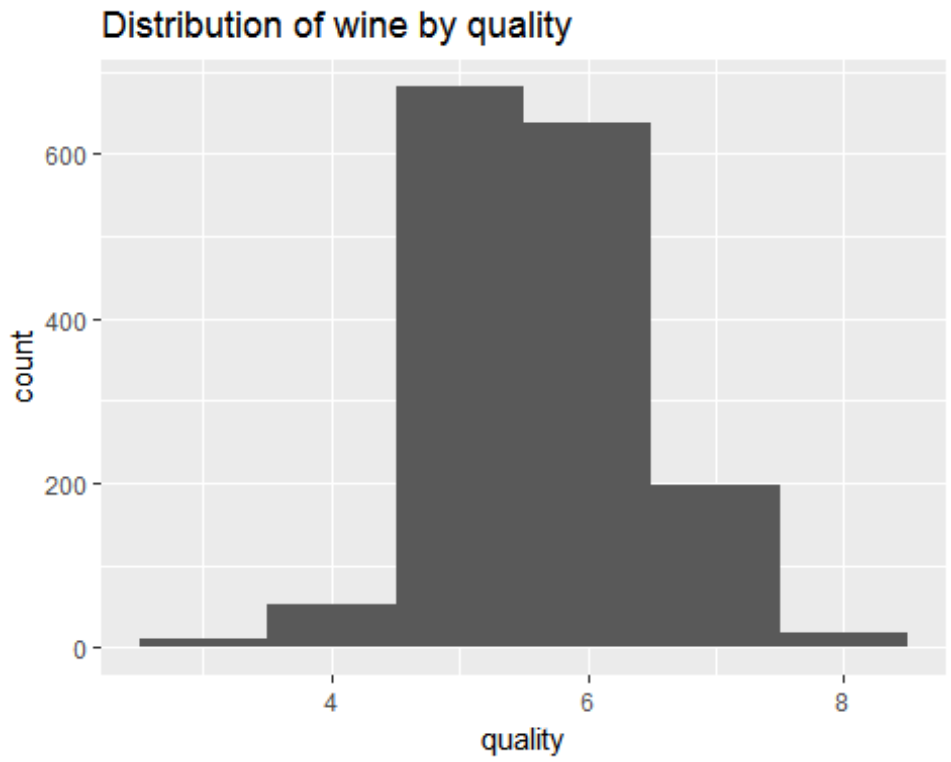
In this project, analysis for quality of red wine is done. Eveyone is aware of the commonly used phrase : "Wine gets better with age". For our study we will be looking at diffrent chemical compositions of wine and how it affects its quality. With the help of statistical tools, we will make predictions on which factors affect the quality. As a first step, all the required libraries are imported and then the data set is loaded to observe its structure. This data set consists of 1599 observations and 13 variables. It seems that the variable 'X' is used to identify each wine while 'quality' is used to rate that wine. Every other variable is the factor used for determining the quality of wine. These factors are all numeric data type and hence there is no need to convert data types further.

```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                  34  0.9978 3.51      0.56     9.4
## 2                  25                  67  0.9968 3.20      0.68     9.8
## 3                  15                  54  0.9970 3.26      0.65     9.8
## 4                  17                  60  0.9980 3.16      0.58     9.8
## 5                  11                  34  0.9978 3.51      0.56     9.4
## 6                  13                  40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5

## '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
0.065 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             : int  5 5 5 6 5 5 5 7 7 5 ...
```

Next, we will see a histogram for 'quality' variable in order to see its distribution in the given data and group them in a new variable called 'taste'. From the plot we can see that maximum number of wines are of medium quality (5-6). So, a new variable 'taste' is created using function 'quality_groups' into 3 groups: 1. bad (0 to 4.5) 2. good (4.6 to 7.0) 3. excellent (7.1 to 10)

Limits for these groups can be changed by passing different parameter values to the function.

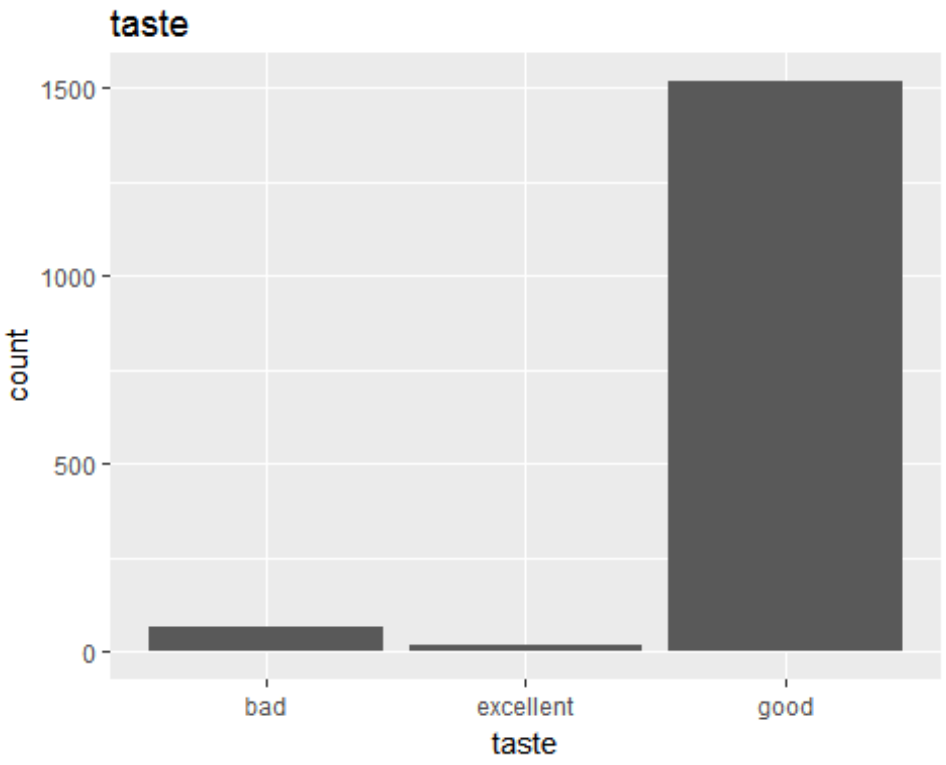



```
## [1] "Mean Value is:  5.63602251407129"

## [1] "Median Value is:  6"

## [1] "Quantile Values are: "

##   0%  25%  50%  75% 100%
##    3    5    6    6    8
```

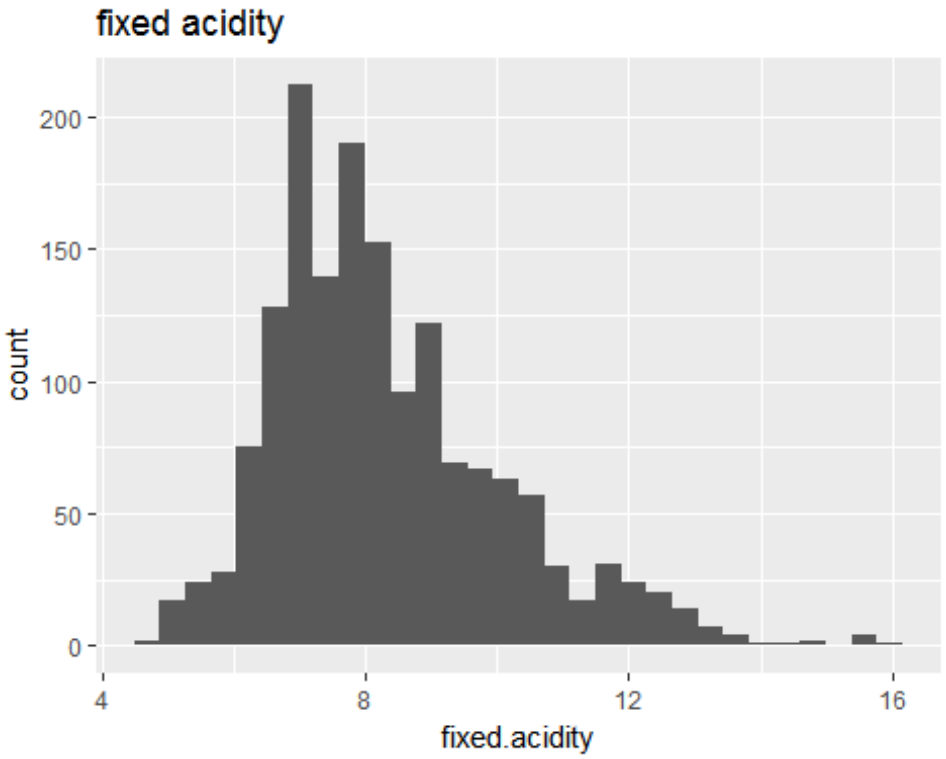
```
##   X fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1 1           7.4             0.70        0.00            1.9     0.076
## 2 2           7.8             0.88        0.00            2.6     0.098
## 3 3           7.8             0.76        0.04            2.3     0.092
## 4 4          11.2             0.28        0.56            1.9     0.075
## 5 5           7.4             0.70        0.00            1.9     0.076
## 6 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                  34  0.9978 3.51      0.56     9.4
## 2                  25                  67  0.9968 3.20      0.68     9.8
## 3                  15                  54  0.9970 3.26      0.65     9.8
## 4                  17                  60  0.9980 3.16      0.58     9.8
## 5                  11                  34  0.9978 3.51      0.56     9.4
## 6                  13                  40  0.9978 3.51      0.56     9.4
##   quality taste
## 1       5  good
## 2       5  good
## 3       5  good
## 4       6  good
## 5       5  good
## 6       5  good
```

Univariate Plots Section

In order to analyize the distribution nature of all the chemical properties mentioned in the data set, histogram plots are created individually and mean,median, quartile and correlation with quality is also provided for clear understanding. Distribution of new variable 'taste' is also plotted in order to get an idea for number of wine samples available in each category.

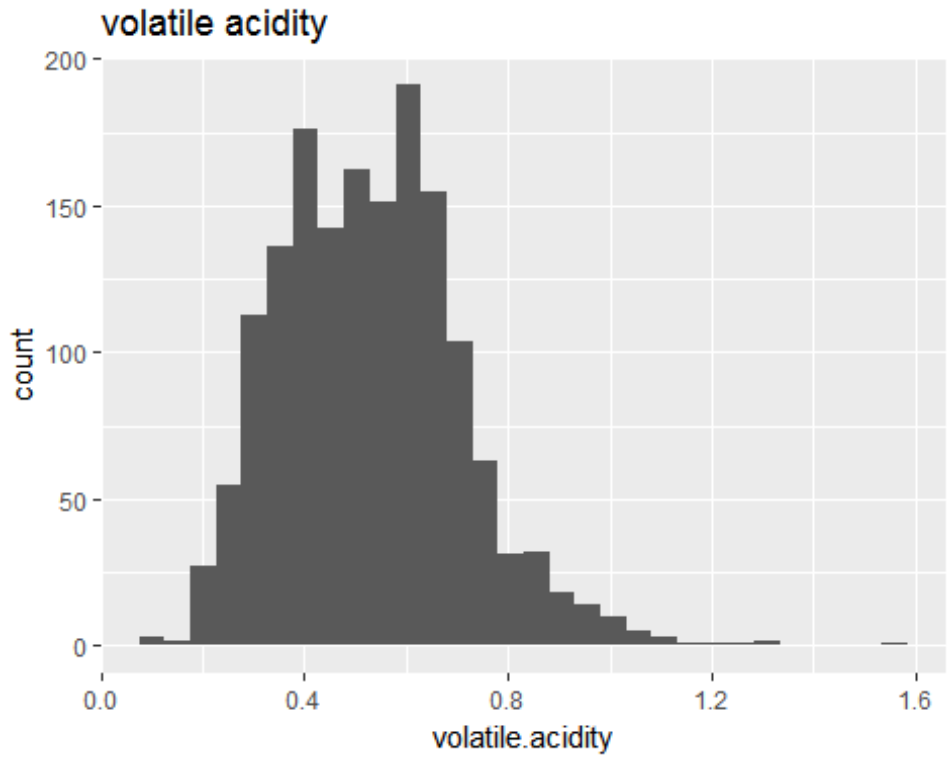


We can see that wines in the group 'good' are very high in number compared to other two ones, which is indicated by mean value of 6.



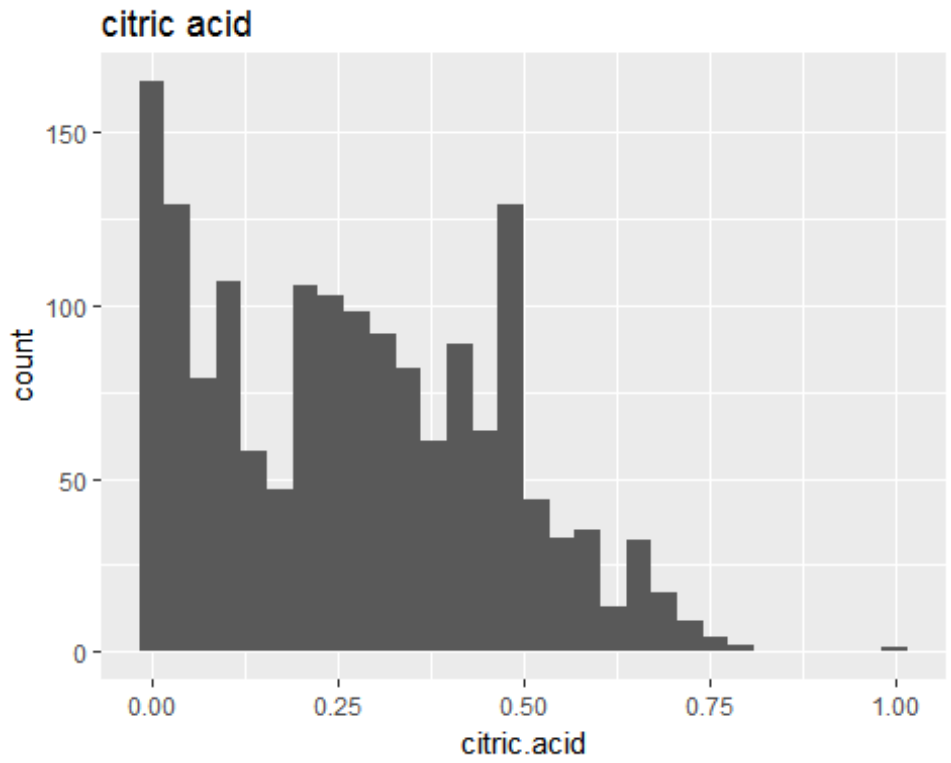
```
## [1] "Mean Value is:  8.31963727329581"

## [1] "Median Value is:  7.9"

## [1] "Correlaion with quality is:  0.124051649113224"

## [1] "Quantile Values are: "

##   0%  25%  50%  75% 100%
##  4.6  7.1  7.9  9.2 15.9
```

Fixed acidity distribution is slightly positively skewed. This factor is not strongly related to quality of wines.



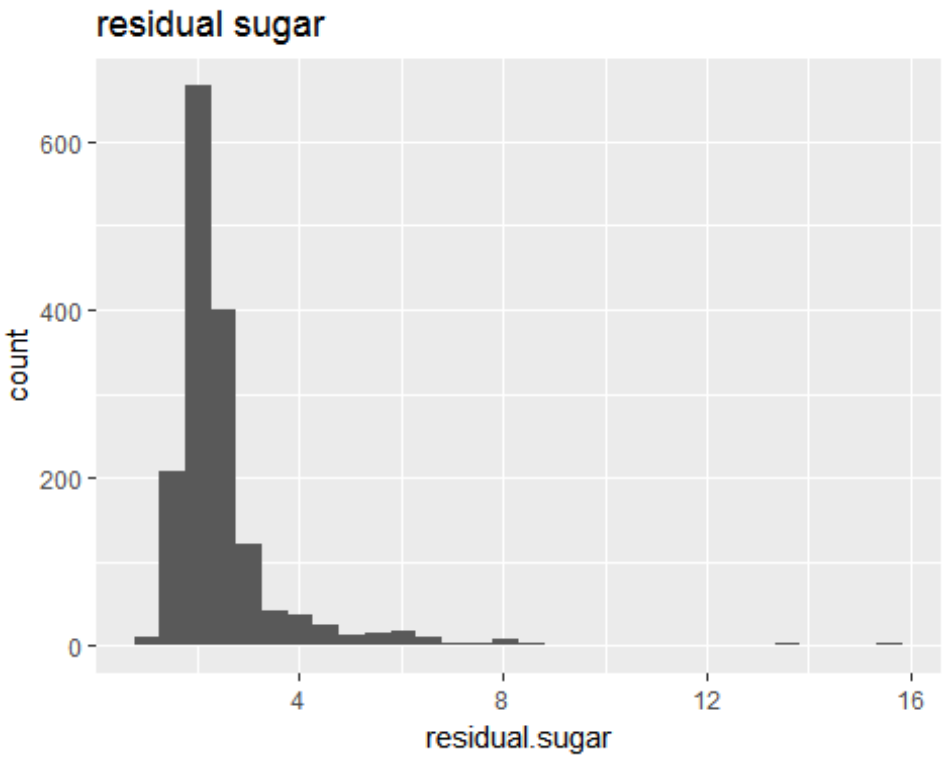

```
## [1] "Mean Value is:  0.527820512820513"

## [1] "Median Value is:  0.52"

## [1] "Correlaion with quality is:  -0.390557780264007"

## [1] "Quantile Values are: "

##   0%  25%  50%  75% 100%
## 0.12 0.39 0.52 0.64 1.58
```

This property is normally distributed in the data and it is moderately related to the quality of wine.

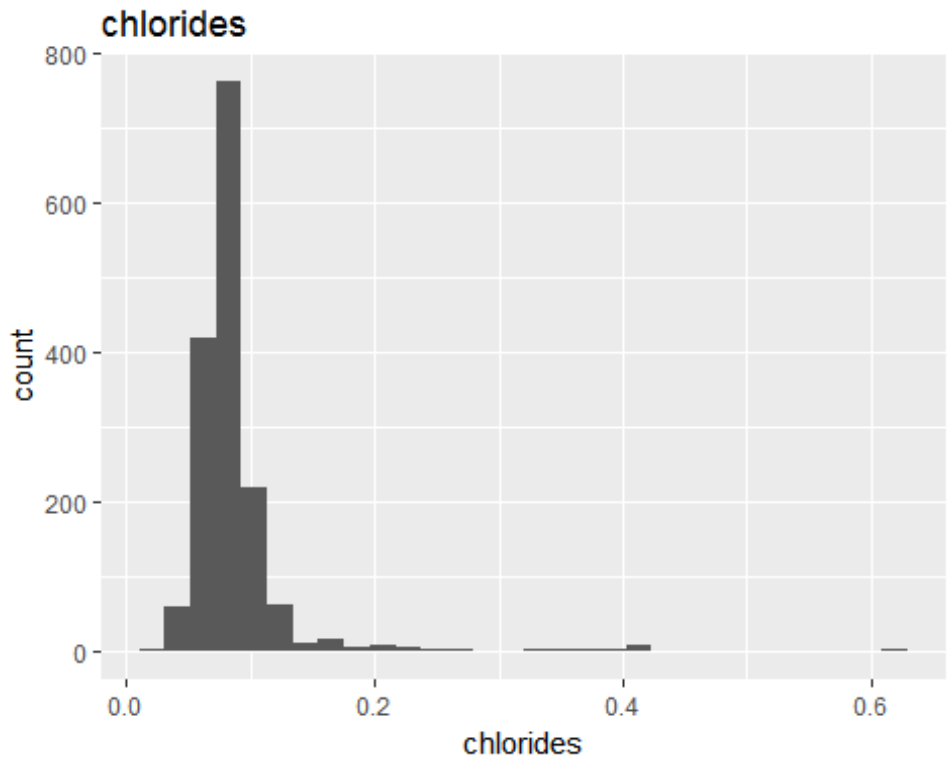



```
## [1] "Mean Value is:  0.270975609756098"

## [1] "Median Value is:  0.26"

## [1] "Correlaion with quality is:  0.226372514318041"

## [1] "Quantile Values are: "

##   0%  25%  50%  75% 100%
## 0.00 0.09 0.26 0.42 1.00
```

Citric acid is another postively skewed quantity with not that strong correlation with wine quality. It is also interesting to see that highest number of samples having value of zero.

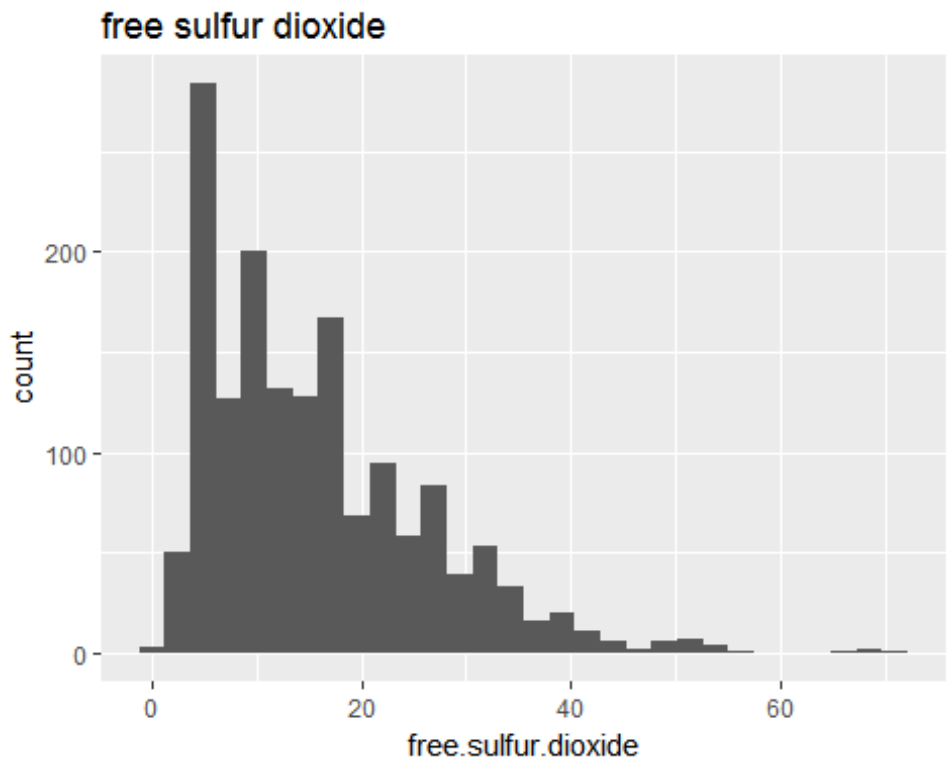



```
## [1] "Mean Value is:  2.53880550343965"

## [1] "Median Value is:  2.2"

## [1] "Correlaion with quality is:  0.0137316373400663"

## [1] "Quantile Values are: "

##   0%  25%  50%  75% 100%
##  0.9  1.9  2.2  2.6 15.5
```

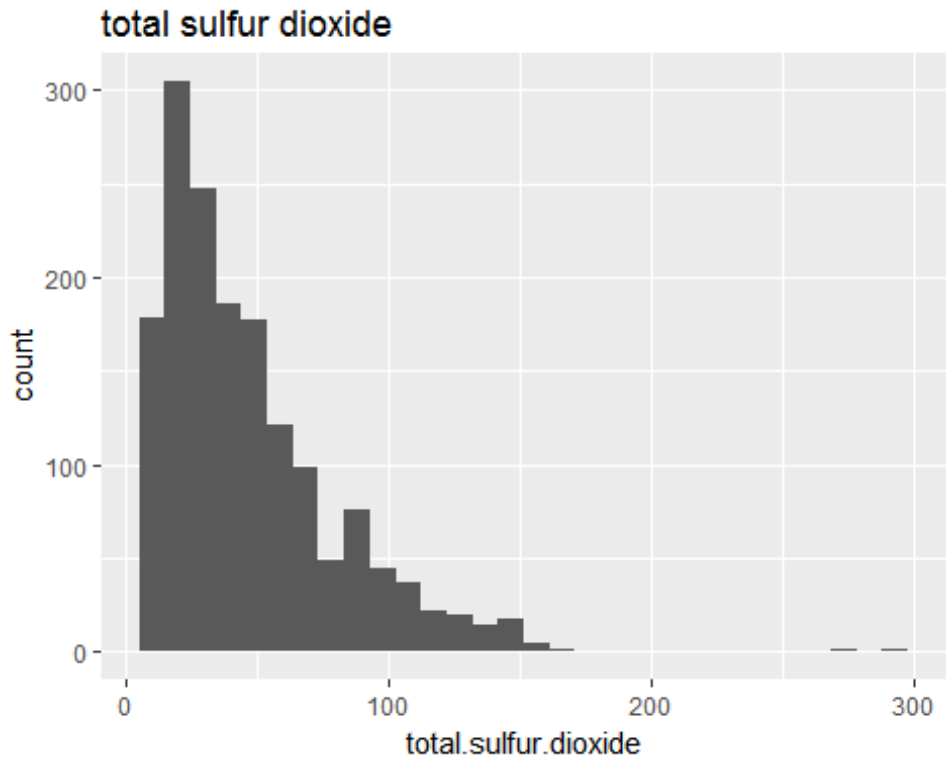



```
## [1] "Mean Value is:  0.0874665415884928"

## [1] "Median Value is:  0.079"

## [1] "Correlaion with quality is:  -0.128906559930053"

## [1] "Quantile Values are: "

##    0%   25%   50%   75%  100%
## 0.012 0.070 0.079 0.090 0.611
```

Both residual sugar and chlorides are long tailed in nature with both of them weakly related to wine quality, especially residual sugar. Some of the values are present towards the higher end.

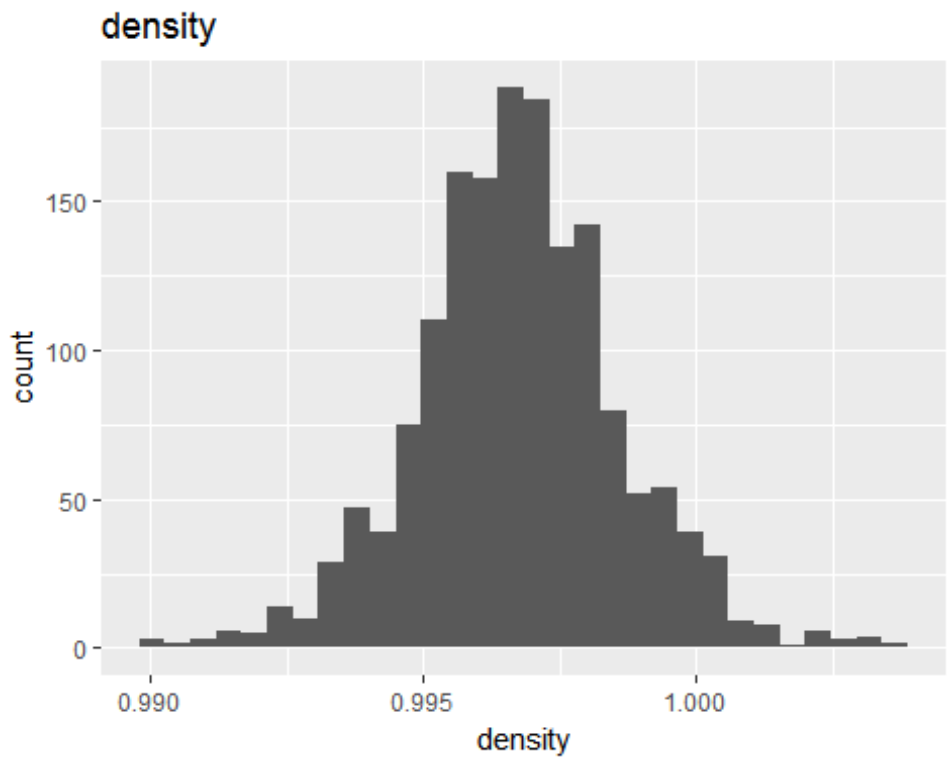



```
## [1] "Mean Value is:  15.8749218261413"

## [1] "Median Value is:  14"

## [1] "Correlaion with quality is:  -0.0506560572442764"

## [1] "Quantile Values are: "

##   0%  25%  50%  75% 100%
##    1    7   14   21   72
```

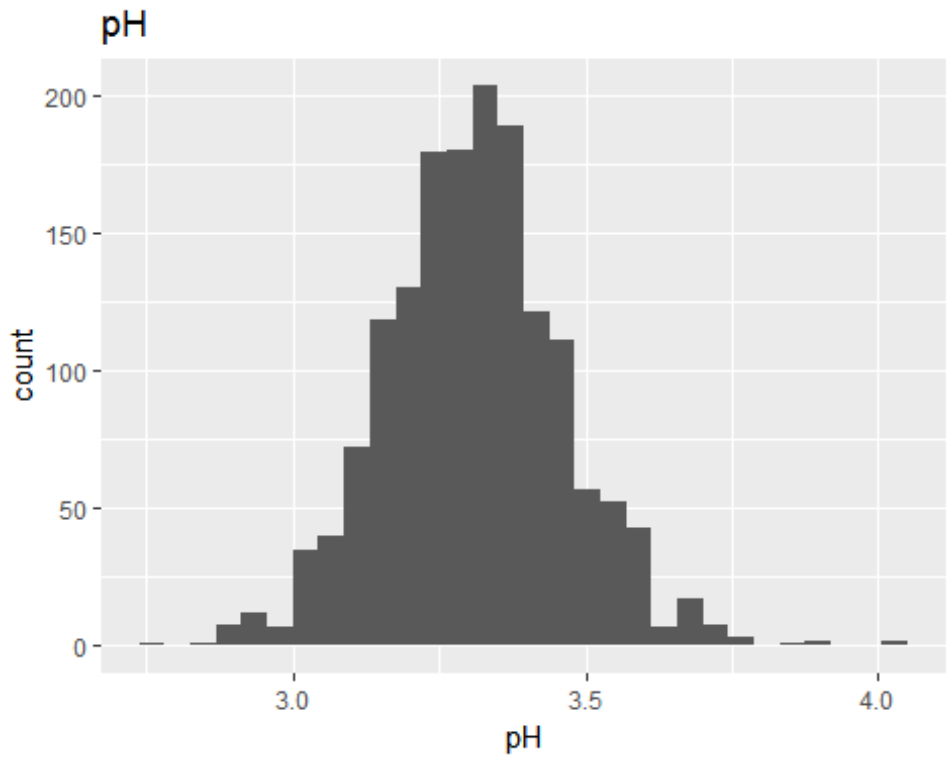



```
## [1] "Mean Value is:  46.4677923702314"

## [1] "Median Value is:  38"

## [1] "Correlaion with quality is:  -0.185100288926538"

## [1] "Quantile Values are: "

##   0%  25%  50%  75% 100%
##    6   22   38   62  289
```

Again, both the suphur dioxide properties (free and total) are positively skewed with neither of them significantly affecting wine quality.

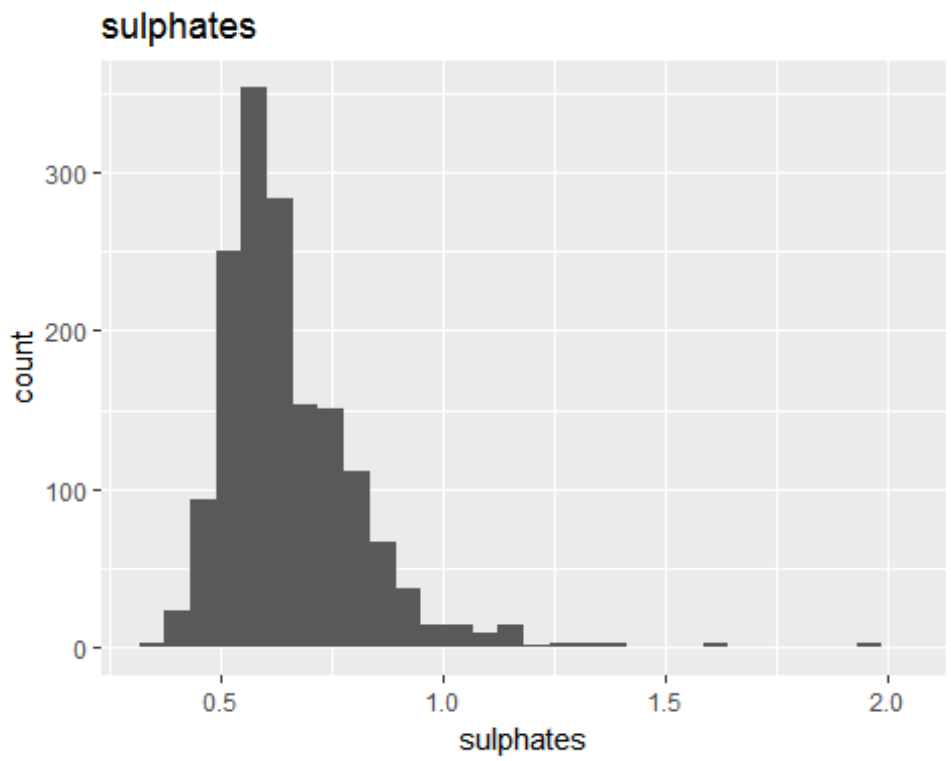



```
## [1] "Mean Value is:  0.996746679174484"

## [1] "Median Value is:  0.99675"

## [1] "Correlaion with quality is:  -0.174919227783349"

## [1] "Quantile Values are: "

##       0%      25%      50%      75%     100%
## 0.990070 0.995600 0.996750 0.997835 1.003690
```

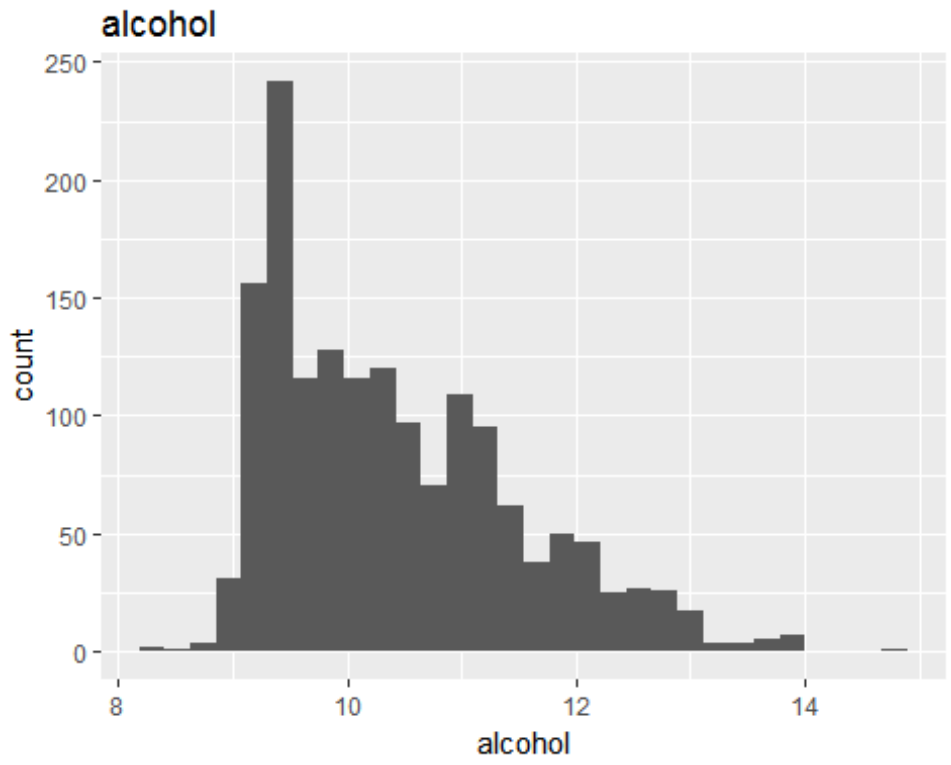



```
## [1] "Mean Value is:  3.31111319574734"

## [1] "Median Value is:  3.31"

## [1] "Correlaion with quality is:  -0.0577313912053821"

## [1] "Quantile Values are: "

##   0%  25%  50%  75% 100%
## 2.74 3.21 3.31 3.40 4.01
```

Density and pH are normally distributed in the given data. Wine quality is not particularly affected by pH. From its mean value we can see that most of the samples are acidic in nature. Density is weak to moderately related to quality and its mean value indicates they mostly consist of water.

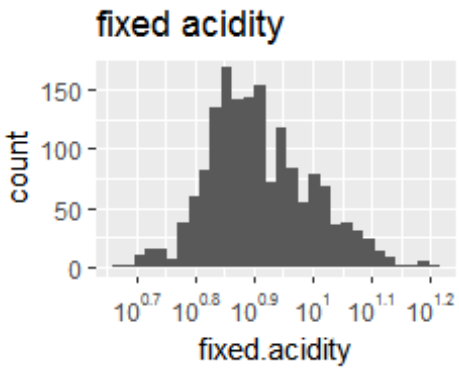


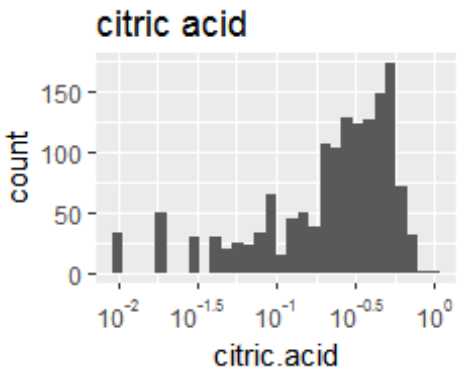

```
## [1] "Mean Value is:  0.658148843026892"

## [1] "Median Value is:  0.62"

## [1] "Correlaion with quality is:  0.251397079069261"

## [1] "Quantile Values are: "

##   0%  25%  50%  75% 100%
## 0.33 0.55 0.62 0.73 2.00
```

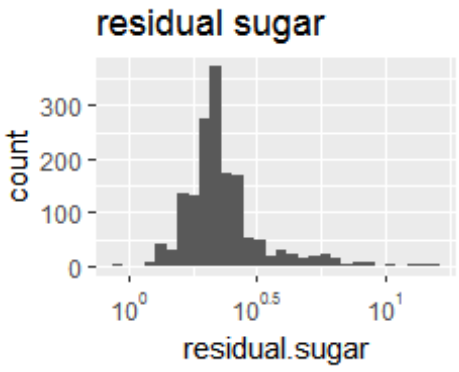


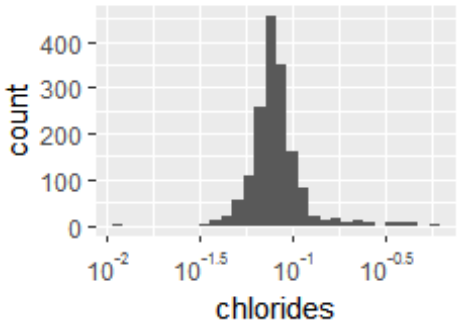

```
## [1] "Mean Value is:  10.4229831144465"

## [1] "Median Value is:  10.2"

## [1] "Correlaion with quality is:  0.476166324001136"

## [1] "Quantile Values are: "

##   0%  25%  50%  75% 100%
##  8.4  9.5 10.2 11.1 14.9
```

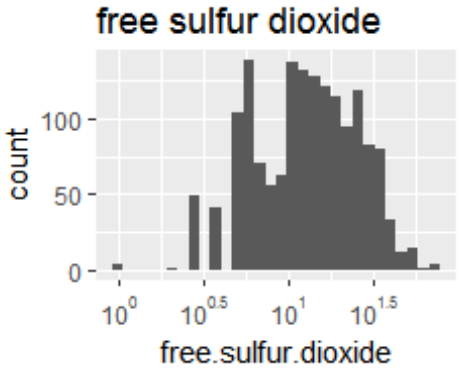
Alcohol and sulphates are both distributed postively skewed and they influence the wine ratings significantly.In order to get further information, long tailed and positively skewed plots are next plotted on log scale. From these plots it is clear that residual sugars and chlorides have outliers present. Also, citric acid and free sulphur dioxide have some extreme values. Observing the scaling of x axis of these plots, we can see that the values do not span across several magnitudes. Hence, log conversion is not much effective conversion.

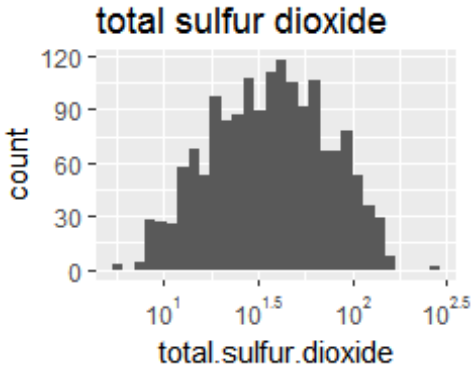

fixed acidity
count
150
100
50
0
$10^{0.7}$ $10^{0.8}$ $10^{0.9}$ 10^{1} $10^{1.1}$ $10^{1.2}$
fixed.acidity

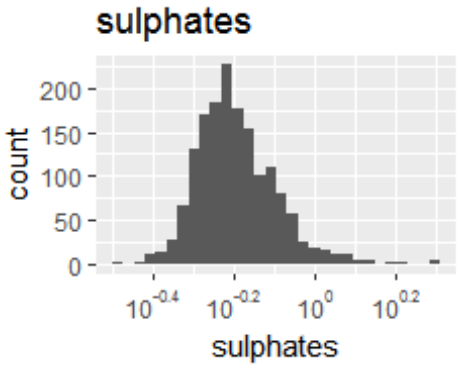

citric acid
count
150
100
50
0
10^{-2} $10^{-1.5}$ 10^{-1} $10^{-0.5}$ 10^{0}
citric.acid

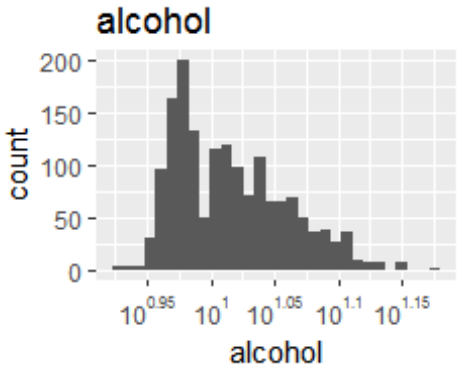

residual sugar
count
300
200
100
0
10^{0} $10^{0.5}$ 10^{1}
residual.sugar


chlorides
count
400
300
200
100
0
10^{-2} $10^{-1.5}$ 10^{-1} $10^{-0.5}$
chlorides


free sulfur dioxide
count
100
50
0
10^{0} $10^{0.5}$ 10^{1} $10^{1.5}$
free.sulfur.dioxide


total sulfur dioxide
count
120
90
60
30
0
10^{1} $10^{1.5}$ 10^{2} $10^{2.5}$
total.sulfur.dioxide


sulphates
count
200
150
100
50
0
$10^{-0.4}$ $10^{-0.2}$ 10^{0} $10^{0.2}$
sulphates


alcohol
count
200
150
100
50
0
$10^{0.95}$ 10^{1} $10^{1.05}$ $10^{1.1}$ $10^{1.15}$
alcohol

Univariate Analysis

What is the structure of your dataset?

Dataset consists of 1599 different compositions of wine described by 12 different factors. X (wine composition number) and quality are integers while rest of the them are numeric type.

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

Quality, alcohol, volatile acidity, sulphates and citric acid are features of interest in the dataset. Reasons for selection of these factors and their detailed analysis will be given in the next section.

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

There should be an additional feature 'age of wine', so that dynamics of other features with time could be analyzed.

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

An additional variable named 'taste' is created which groups the quality of wine into 3 types named bad, good and excellent. The limits for these groups can be changed since they are created using a function.

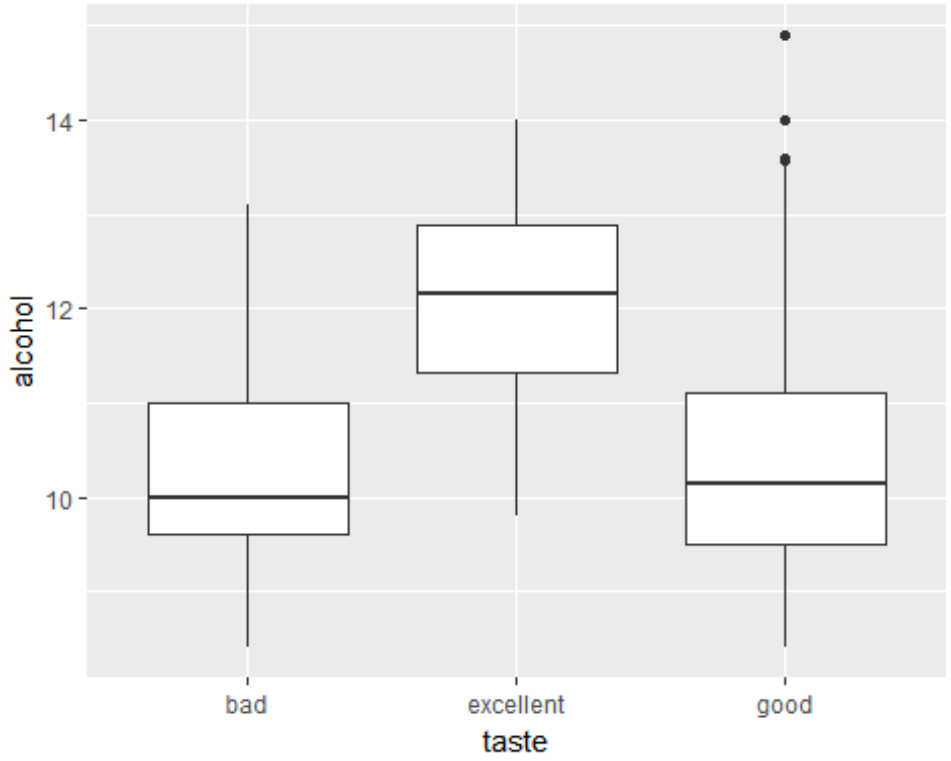
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

Since all the factors are already present in numeric data type, there was no need to change the form of data. Upon plotting these factors, some of the them were not normally distributed. These were plotted on log scale to 'spread' them out. The citric acid factor had 132 values corresponding to zero.

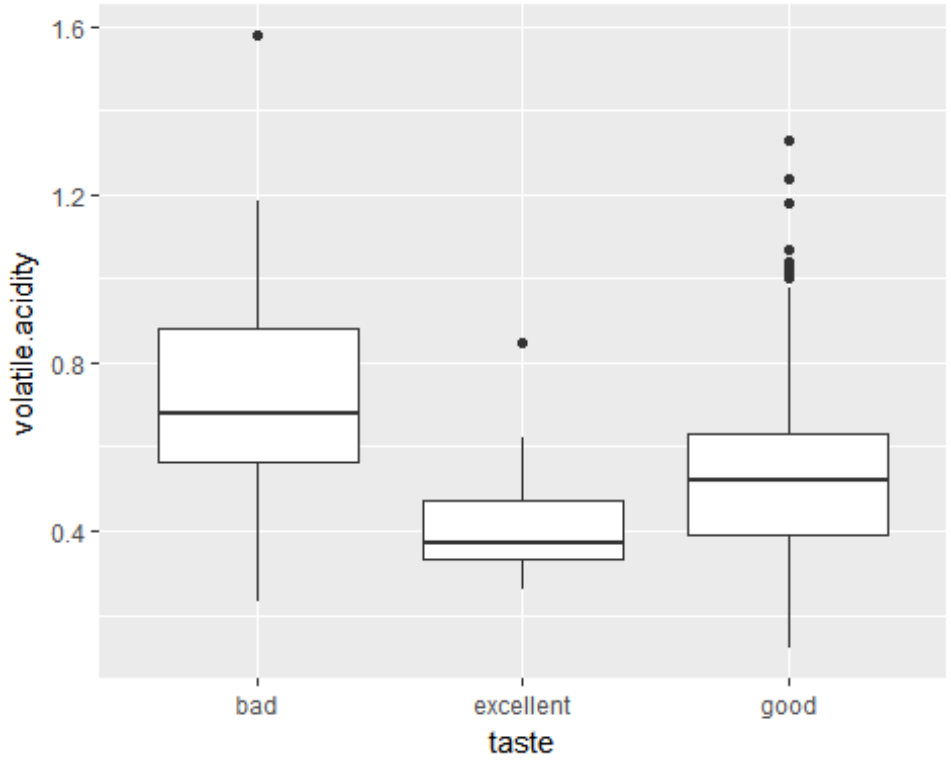
Bivariate Plots Section

Since our most important factor in the dataset is 'quality', correlations for each factors with it is obtained. From these correlation values, realtionship of 'quality' with these 4 factors is explored further:

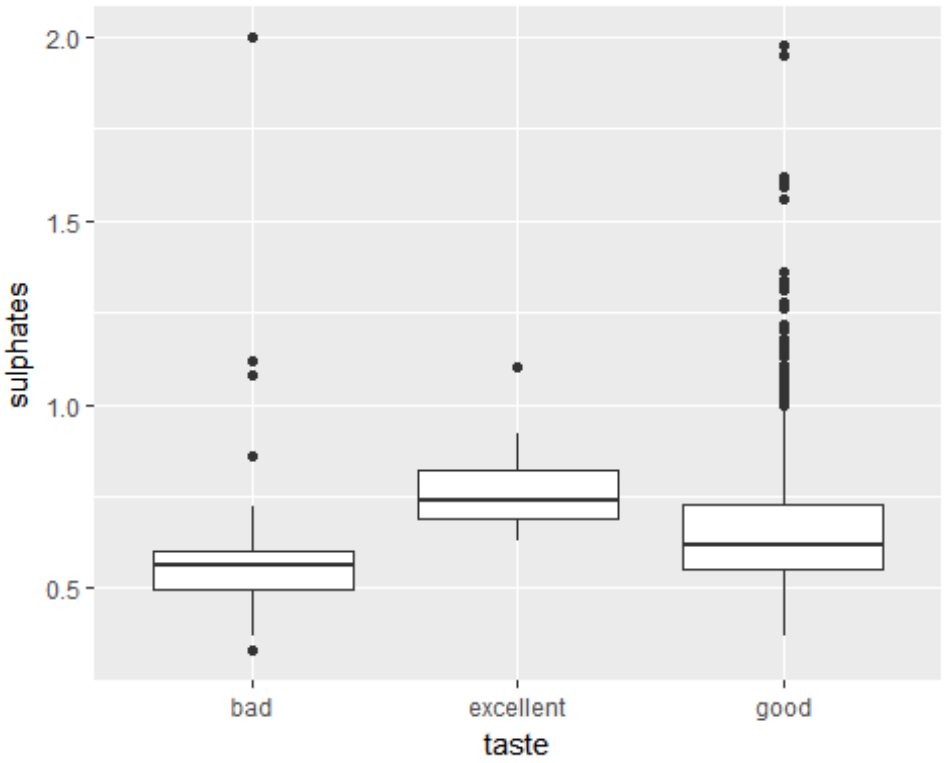
1. Alcohol (cor: 0.476)
2. Volatile acidity (cor: -0.390)
3. Sulphates (cor: 0.251)
4. Citric acid (cor: 0.226)



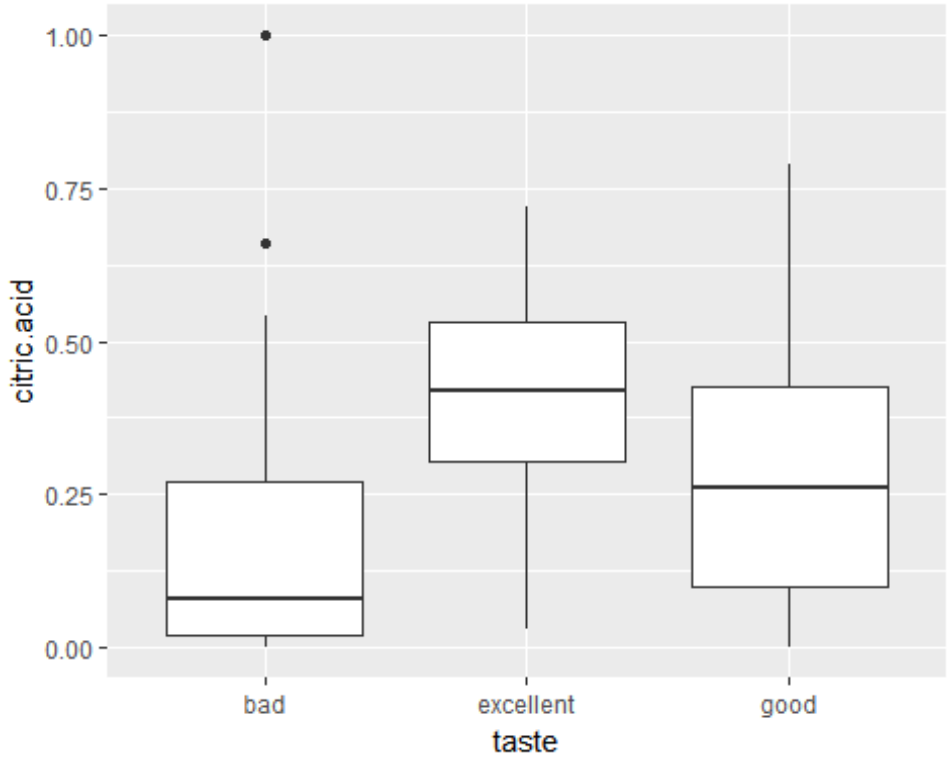
By observing the box plot between 'taste' and 'alcohol', we can see a general trend of increase in alcohol as quality rating increases. Relationships for remaining four factors are analyzed in same way.



This relationship indicates that as the wine quality increases, the amount of volatile acidity decreases.



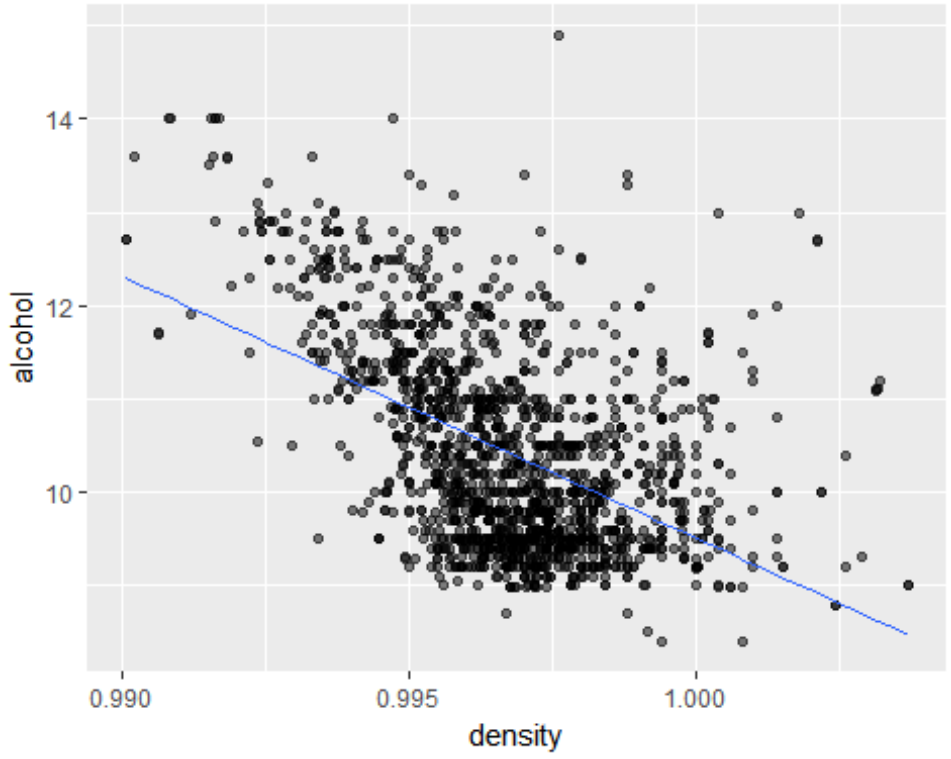
The trend seen in this relation is similar to that between quality and alcohol i.e higher quality wines contains more amount of sulphates. But an important to note that this trend is not as sharp as quality/alcohol one.



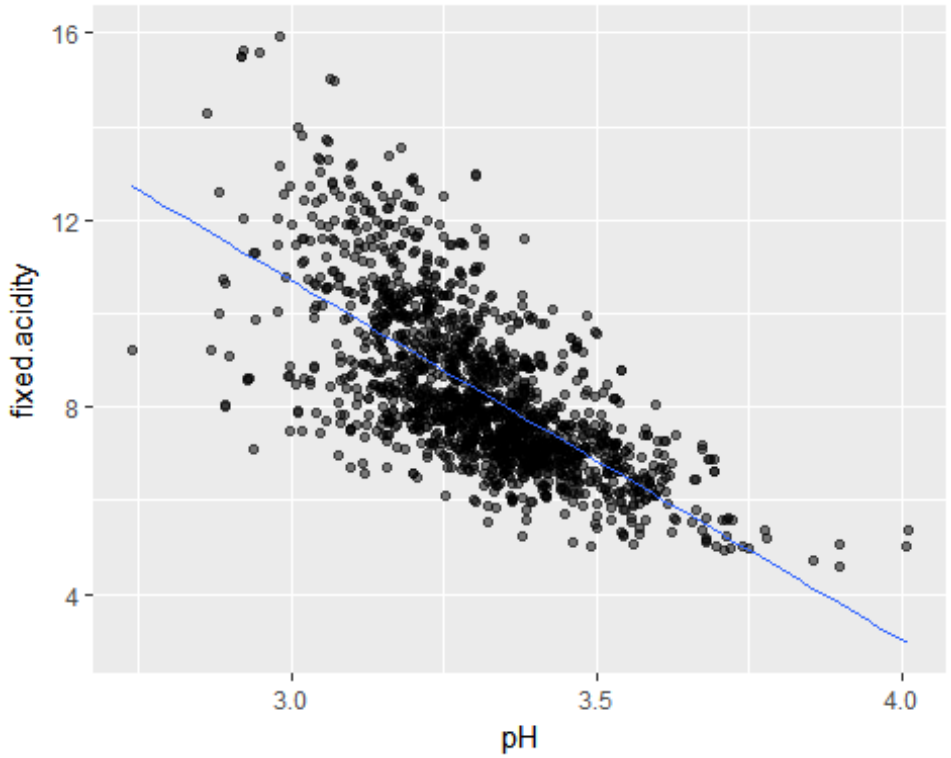
More amount of citric acid indicates more quality of wine. Also, 'excellent' wine contains comparitvely much more amount of citric acid than 'bad' and 'good' wines.

Further, relaions between chemical factors are explored. Two most strongly related variables are:

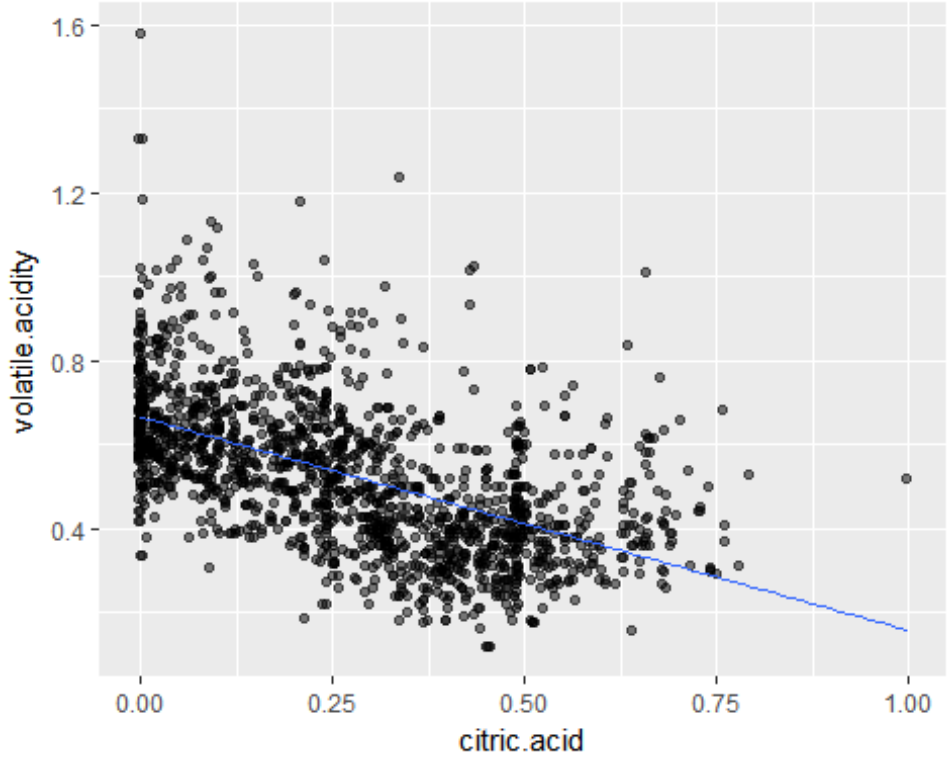
1. Alcohol and density (cor: -0.496)
2. Fixed acidity and pH (cor: -0.682)
3. Volatile acidity and citric acid (cor: -0.552)



Gneneral trend observed indicates that as the alocohol amount decreases, the density of the wine approaches the value of 1.(which is for water)



Higher percent of fixed acidity is seen in wines with low pH value. This trend makes sense as fixed acidity corrresponds to amounts of non volatile acids present in the wine whereas low pH value indicates more acidic nature.



As the quality of wine increases, the relative amount of citric acid increases while that of volatile acidity decreases. Citric acid is used in small amounts to add freshness to the wine and volatile acidity consists of acetic acid which leads to unpleasant taste when used in high quantities.

Bivariate Analysis

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, we started with four most strongly related factors with quality of wine.(Alcohol, Volatile acidity, Sulphates, Citric acid) From these relations we can conclude that an excellent wine has:

1. Higher percent of alcohol
2. Low volatile acids
3. Relatively high amounts of sulphates
4. High citric acid quantity

Next, we studied 3 relations between factors themselves. General comments from these analysis are:

1. More quantity of alcohol leads to density of wine lesser than that of water.
2. Higher fixed acids leads to lower pH value which indicates more acidic nature.
3. As amounts of citric acid increases, less quantity of volatile acids are found in the wines.

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

The relation between volatile acidity and citric acid was most interesting as the only relation I could think of was that volatile acids tend to give unpleasant taste while citric acid provide freshness to the wine. Maybe thats why they are negatively related to each other. Further investigation is neccessary.

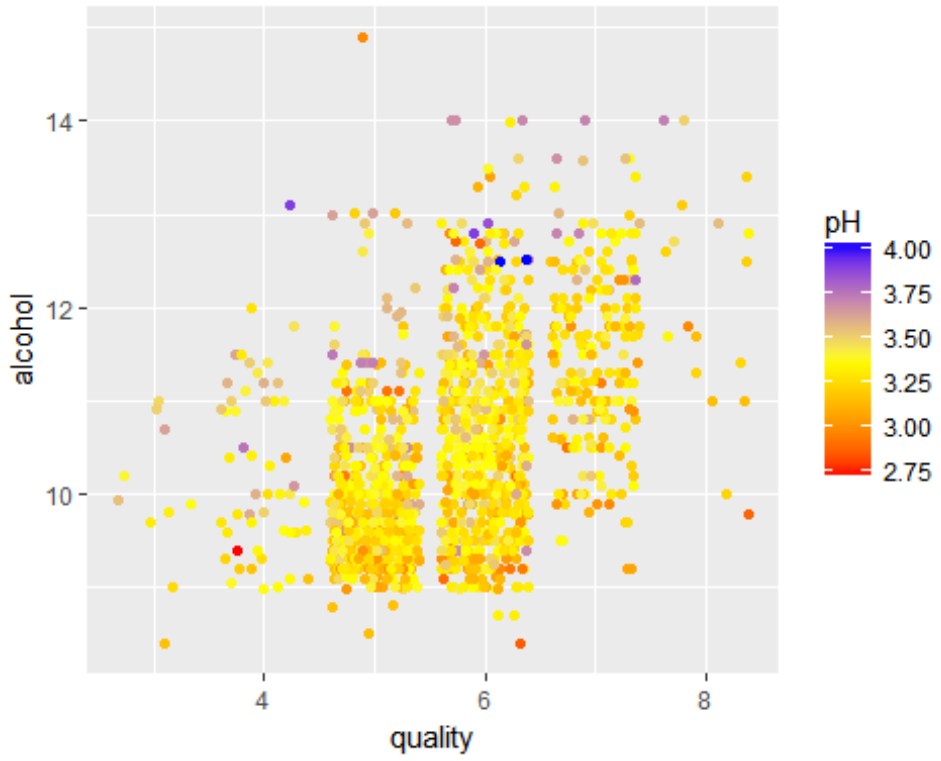
What was the strongest relationship you found?

From the correlation values computed between the variables, these are the three strongest relationships found:

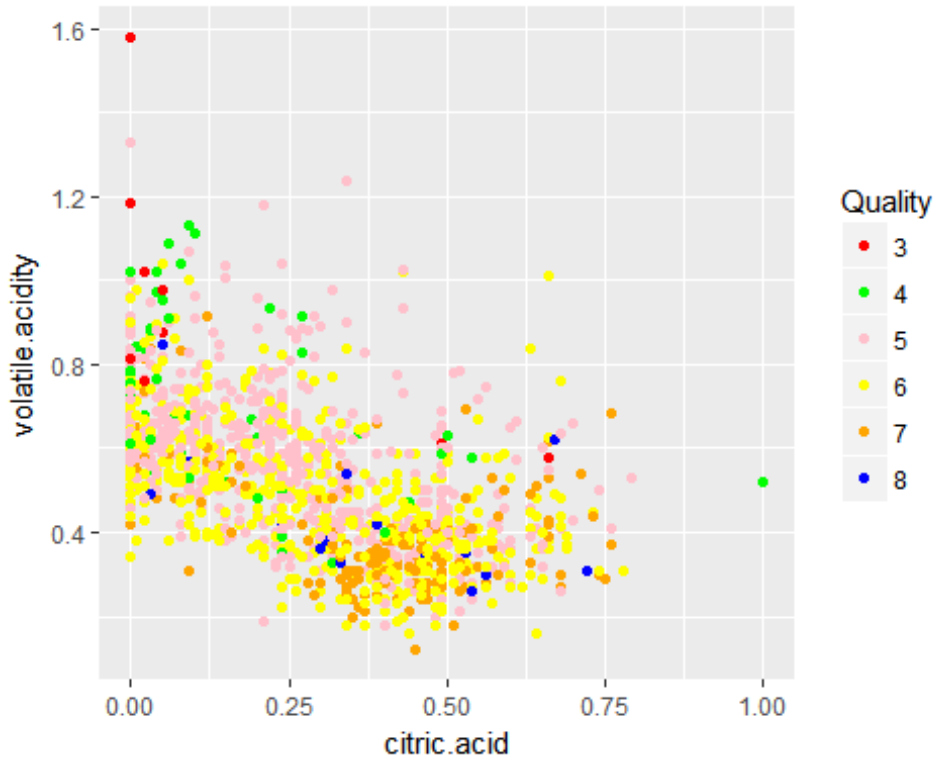
1. Fixed acidity and pH of wine
2. Volatile acidity and citric acid in wine
3. Alcohol and quality of wine

Multivariate Plots Section

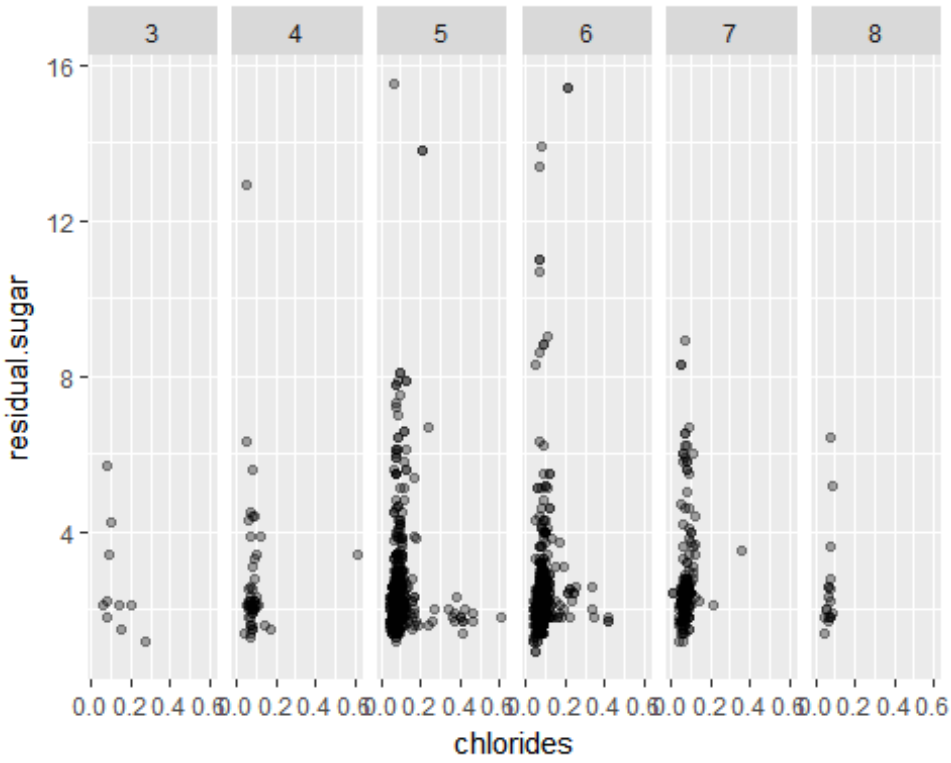
Based on the relations discussed in the bivariate analysis, combinations of those factors are explored further in this section.



It is clear that alcohol content increases as quality of wine increases. This relation was already explored in the earlier part. It is interesting to note that more alcohol contained wines tend towards basic nature while low alcoholic combinations are acidic. This relation is not strongly related as some of the high alcohol wines are acidic.



As suspected in the earlier section, relation between volatile acidity and citric acid is based on the taste it provides. Higher citric acid brings freshness to the wine while volatile acids gives unpleasant vinegar taste when used in more quantity. From this plot we can see that higher quality have more citric acid and less volatile acid present.



Chlorides indicates amount of salt present in the wine while the residual sugar is the quantity of sugar present after fermentation process. From these plots we can observe that in all the qualities of wine a certain ratio of sweet/salty is maintained. As the quality of wine increases, deviation from this ratio is less.

Multivariate Analysis

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?

1. More alcoholic wines (which are higher quality) tends towards basic nature on pH scale. Though this trend is not strongly related.
2. As guessed in the earlier section, the relation explored between volatile acidity and citric acid strengthened the view that higher quality have more freshness and less unpleasant vinegar taste.
3. Relation between residual sugar and chlorides is analyzed.

Were there any interesting or surprising interactions between features?

The relation between residual sugar, chlorides and quality of wine was the most surprising one. According to the initial intuition that red wines are more salty in taste, this relation is analyzed. From the plots it was seen that this assumption was not true and in reality red

wines maintain a certain ratio of sweet/salty taste. Further, this ratio is observed more consistently in the higher quality of wines.

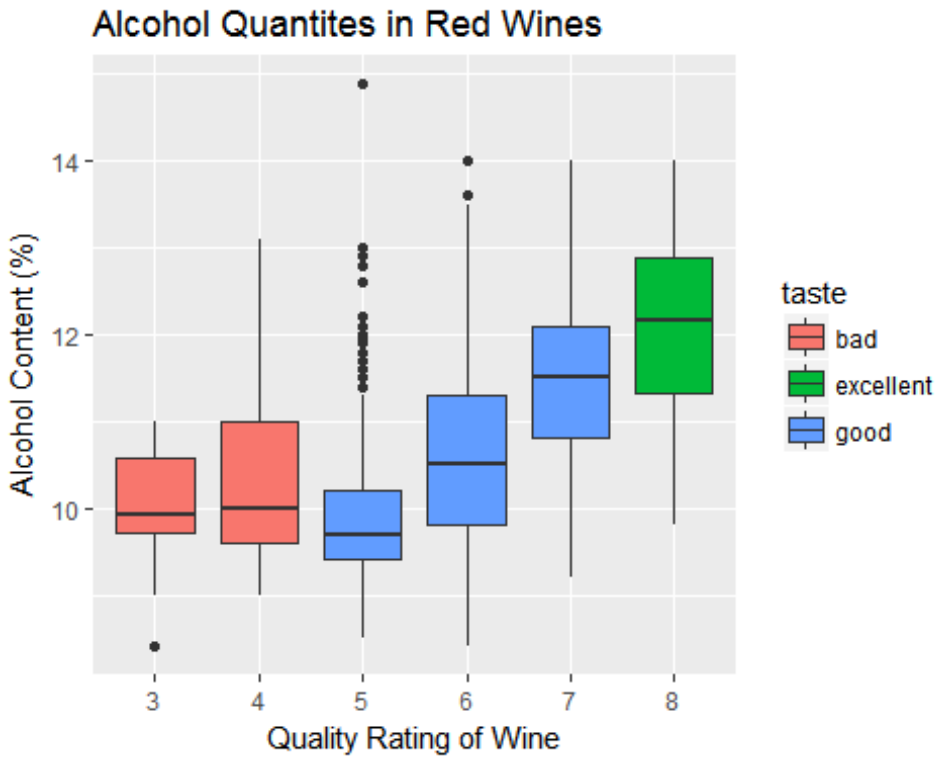
OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

Final Plots and Summary

From the analysis conducted until now these are the three most important relations that are found which help in determining the characteristics of red wines:

1. Alcohol content and quality of wine
2. Volatile acidity and citrus acid in wine
3. Chlorides and residual sugar in wine

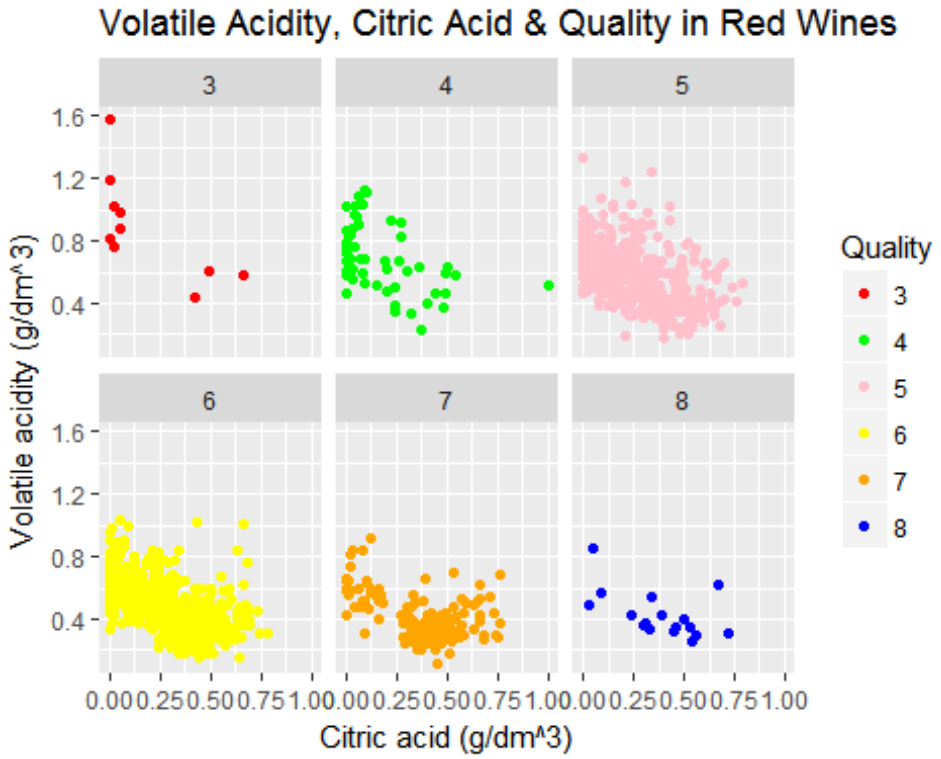
Plot One



Description One

An indication of excellent wine is the higher amounts of alcohol contained in them. Bad wines have lowest quantities of alcohol in them while good wines have slightly higher amounts present in them. From the size of the box plots we can observe that data available for good wines is almost twice as for bad and excellent wines.

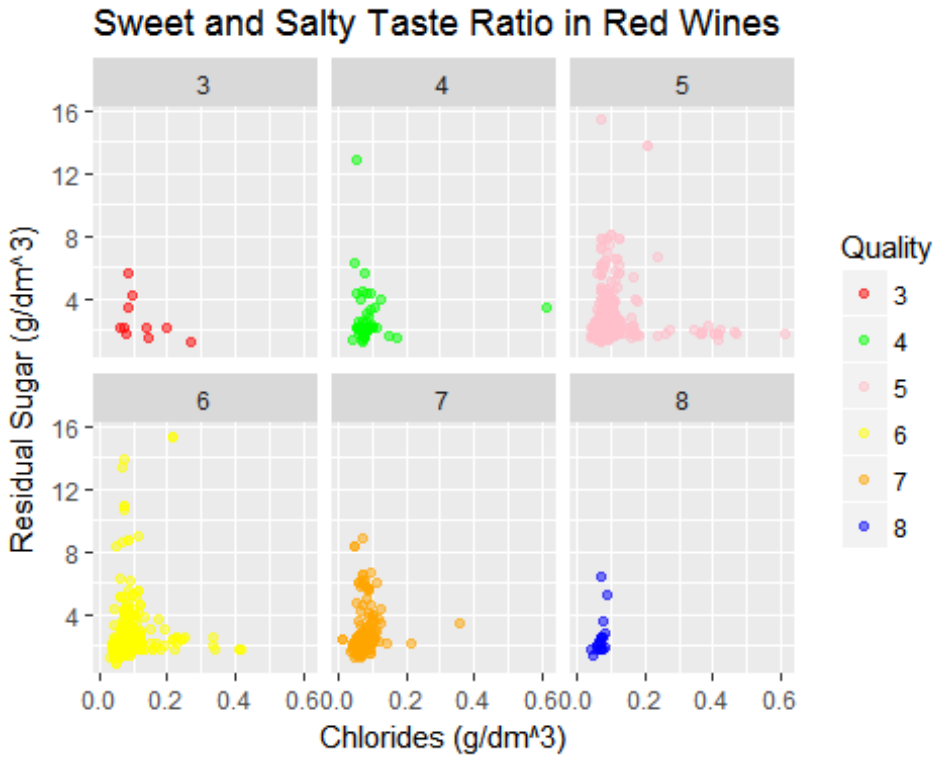
Plot Two



Description Two

Excellent wines can be differentiated from good and bad wines by two important chemical substances. These are volatile acids, which are preferred in less amounts and citric acids, which are available in relatively higher quantites.

Plot Three



Description Three

This plot is the more refined version of the relationship explored in the multivariate analysis section. A specific ratio of sweet/salty taste in maintained in the red wines of all qualities. This is around: (2 units sweet to 0.1 units salty) In order to differentiate excellent wine from others, it is observed that these wines follow this ratio more consistently than other qualities.

Reflection

Exploratory data analysis for this project was done with the aim of undersanding factors behind differentiating the wine by quality. After appropriate tidying up and adddition of variable in the data set, factors which are strongly related to wine quality are determined and are explored further. These analysis led to following statements:

1. Alcohol is one of the main indicators for determining the quality of wine. Superior types of wine contains higher amount of alcohol, though it should not be used as the sole criterion for differentiation.
2. Other factors which play a crucial role in determining best quality of wine are: low amounts of volatile acids and high amounts of sulphates and citric acid.

Further, relations between variables apart from quality of wine are investigated. Some of the conclusions obtianed from these studies are:

1. Adding alcohol results in lesser density of wine than water.
2. Lower pH value of wine indicates higher amounts of non volatile acids. (fixed acidity)
3. Wines containing relatively low amounts of volatile acids and high quantity of citric acid are consider of superior quality.

Muliple factors are compared to form more complex relationships from the given dataset. Some of the trends observed are:

1. More alcoholic wine mixtures, and hence of higher quality tend towards basic nature on the pH scale. This relation is not strongly related and so it must not be used as primary criterion for determining wine quality.
2. High amount of citric acid is found in excellent wine mixtures as it provides freshness. On the other hand, volatile acids in less amount are preferred as it brings unpleasant taste of vinegar.
3. Most interesting relation discovered was the taste of wines in terms of sweet and salty taste. A specific ratio is attempted to maintain in the wines. Higher qualities of wine follow this ratio more consistently than others.

The dataset provided consisted of significant number of observations from which we can formulate statements for determination of wine quality. However, this dataset can be further improved by overcoming these couple of drawbacks:

1. Number of observations for bad, good and excellent wines are not in proportion. Wines which are rated between 5-6 are very high compared to other two classes.
2. There is absence of variable 'time' to indicate the age of the mixture. Analysis of other factors with time could have led to further insights and it could have possible to check the validity of the famous phrase: "Wine gets better with age."

As a final point, rating of wine is a complex analysis because taste preference is different for each individual, but this study gives satisfactory observations for classifying wine by their quality in general.