Introduction to Dataset:

The dataset is a tidy csv file containing 1599 observations and have 13 variables associated to them. I will perform Exploratory Data Analysis (EDA) on a data set which contains red wines with variables on the chemical properties of the wine.

Dataset Structure

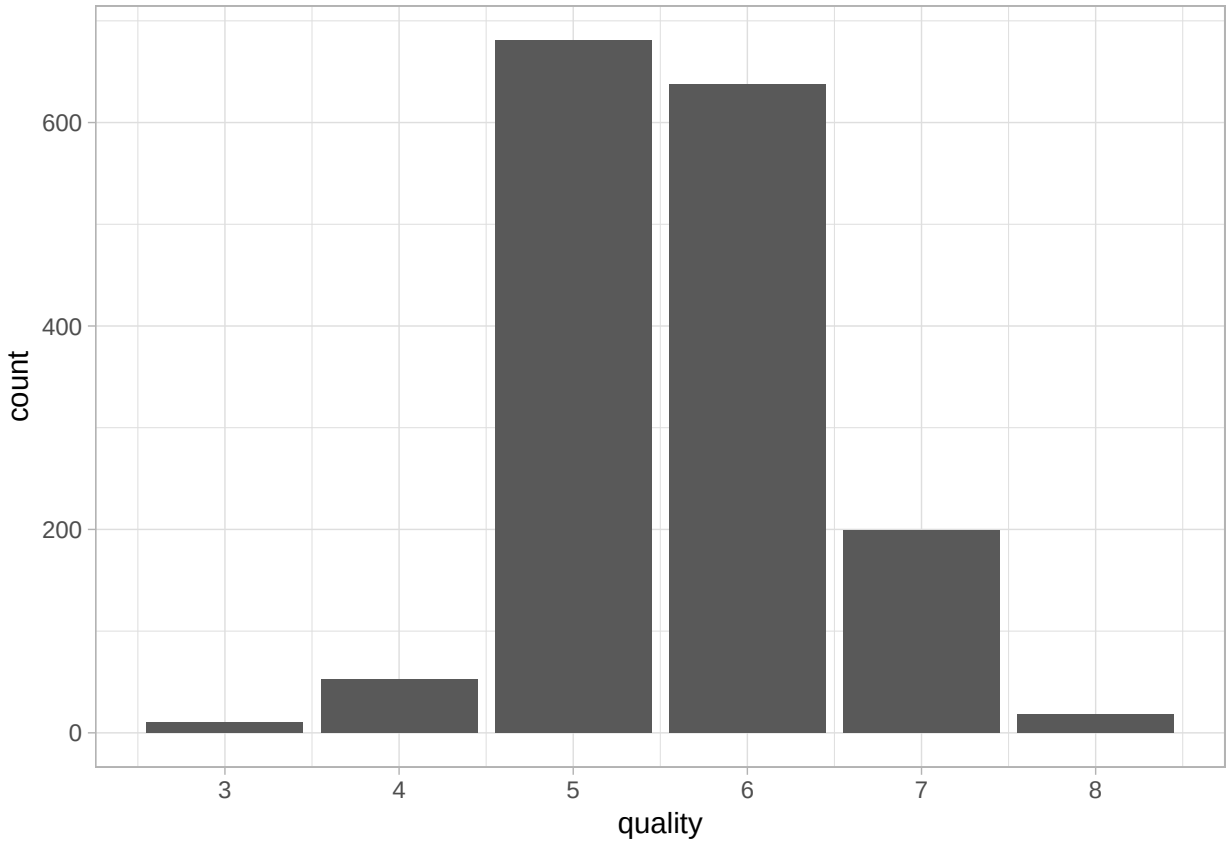
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  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 ...
```

Dataset Summary

```
##        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
```

Closer Look at Quality:

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



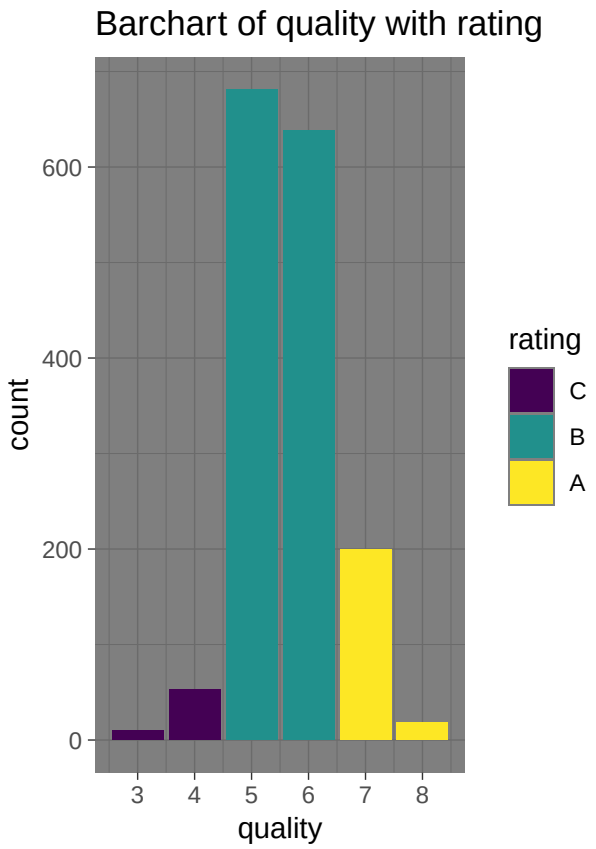

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

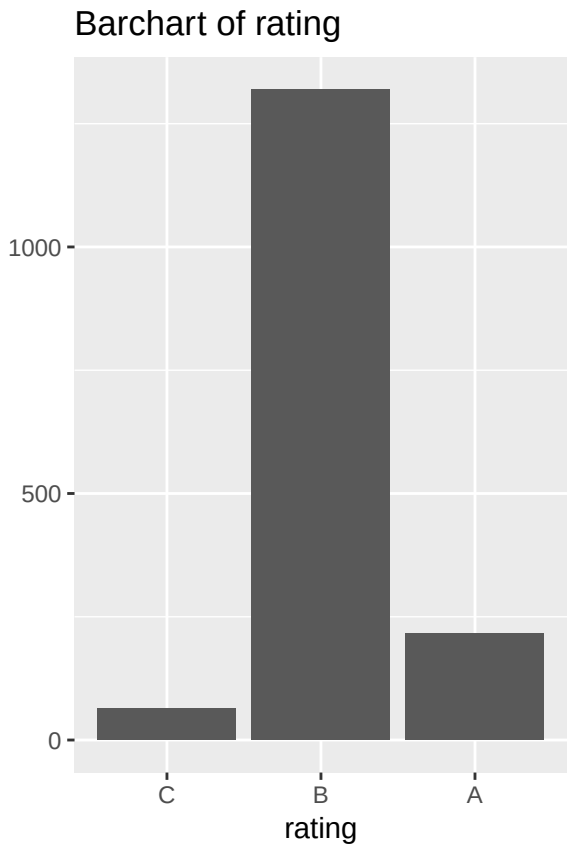
```
## 85%
##   6
```

We can see that 3 is the min rating given to quality & 8 is the max. The distribution is normal with 85% data points below or 6. So I have created a variable called 'rating' based on variable quality

8 to 7 are rated A, 6 to 5 are rated B, 3 to 4 are rated C

```
##    C    B    A
##   63 1319  217
```

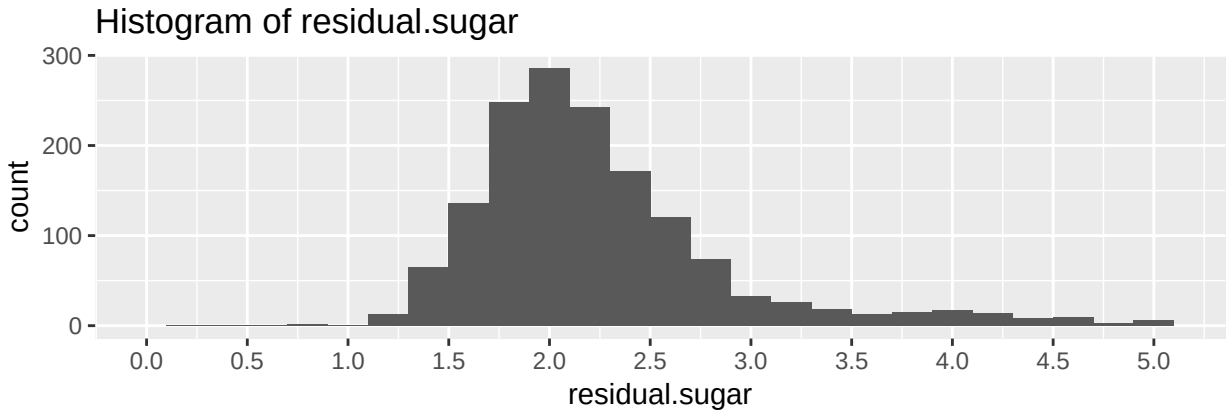


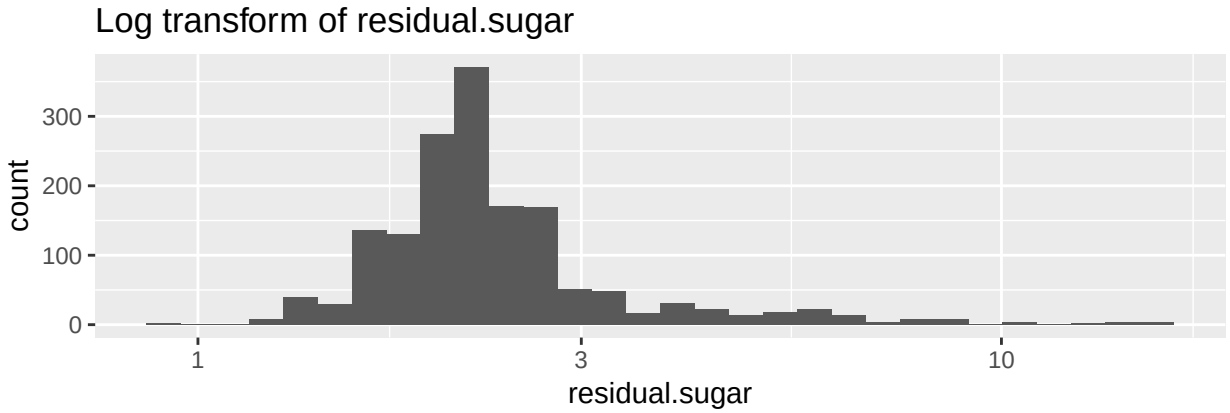


Univariate Plots Section

Residual Sugar (g / dmˆ3)

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



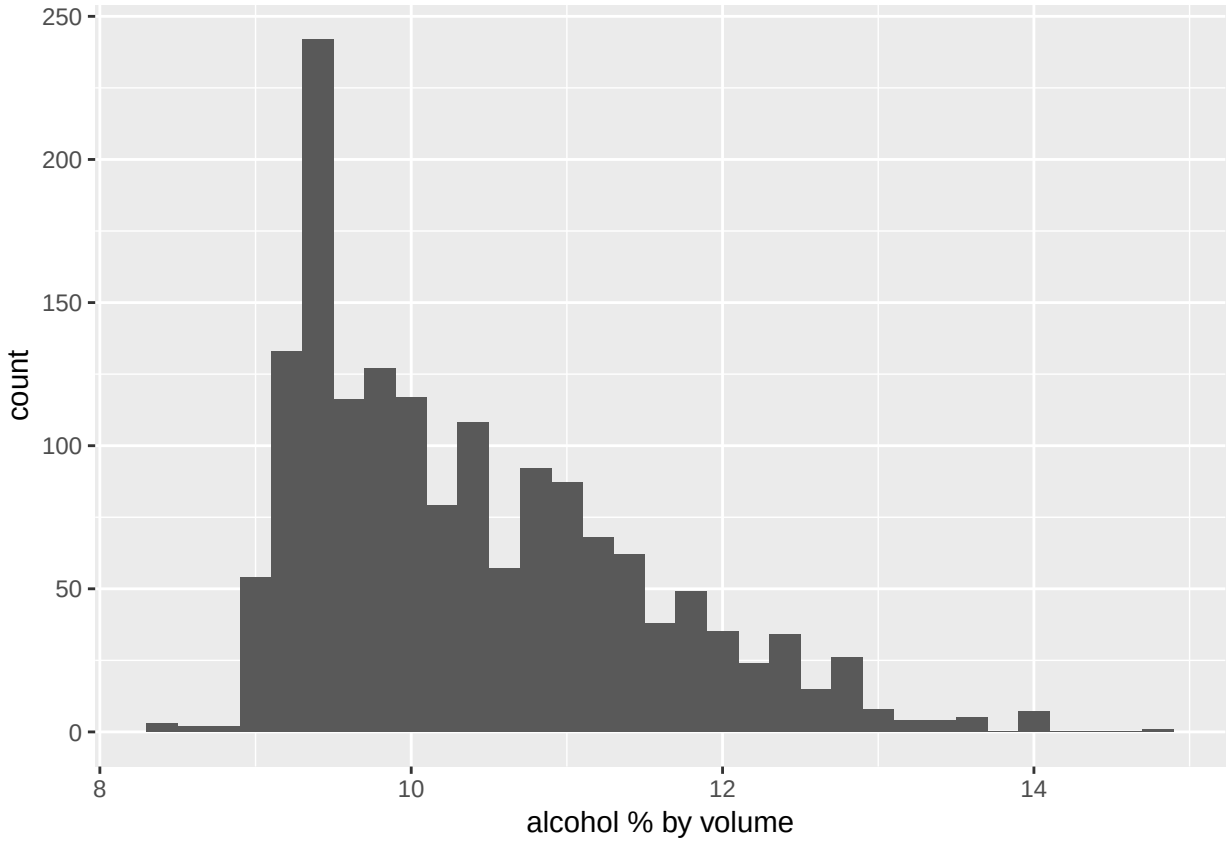



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

It a normal distribution with a long tail,I have removed outliers(top 5%) from the graph .The second graph is the log transformation of residual sugar The mean value is 2.53 and max goes all the way up to 15.50

Alcohol (% by volume)

Alcohol (% by volume): the percent alcohol content of the wine. Wine is an alcoholic breverage . I am hopping to see some trend in this data now and in my further observations

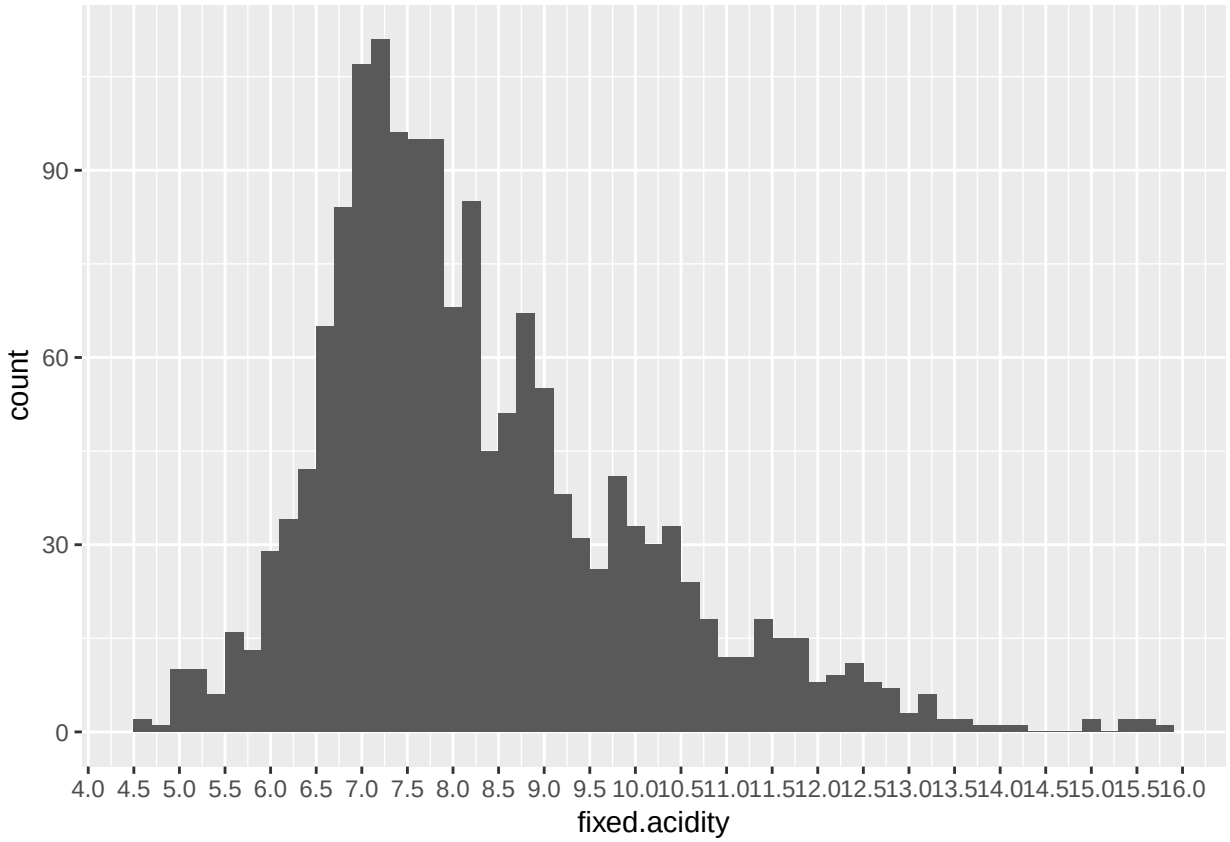



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

The distribution is positively skewed with mean on 10.20 % and having max value of 14.90%

Fixed acidity (tartaric acid - g / dm^3)

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

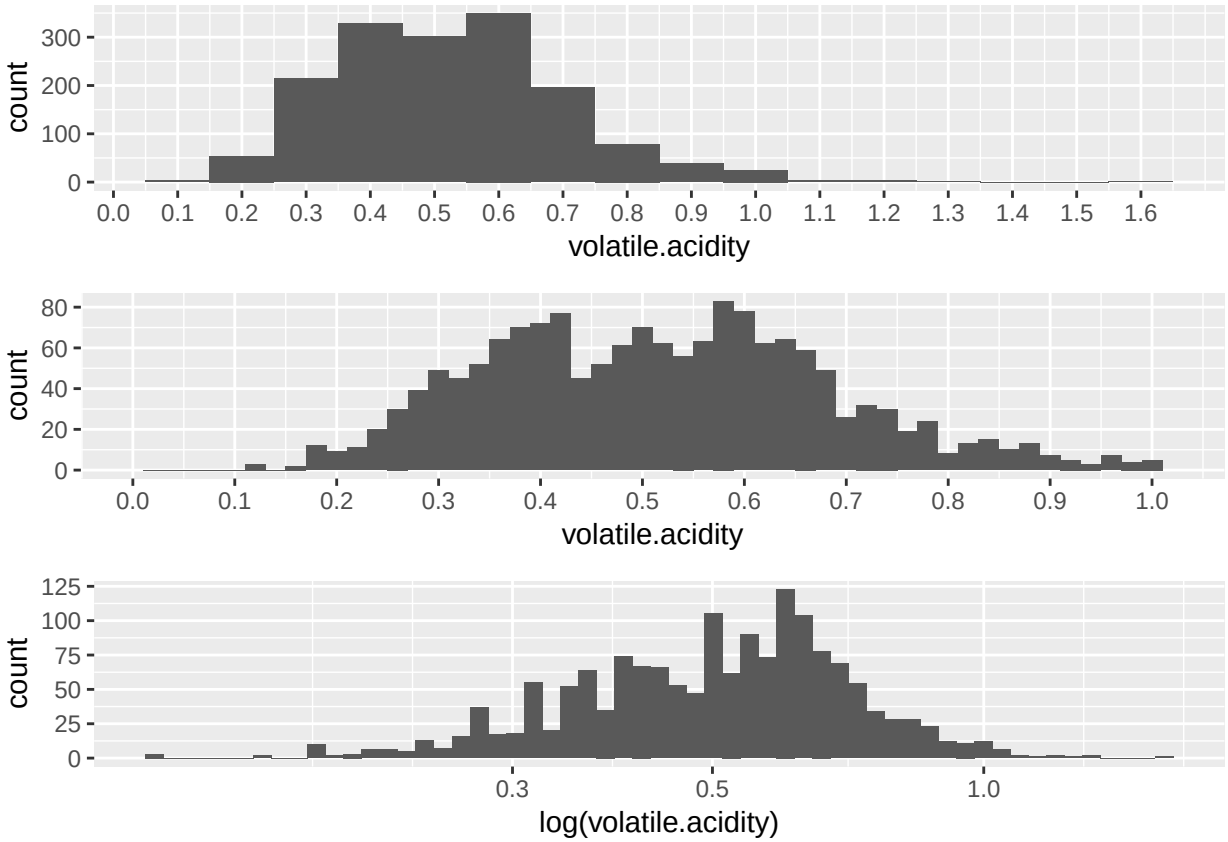



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

Its a Normal Distribution which is peaks at 7 it has a mean of 8.32 and a max value of 15.90

Volatile Qcidity (acetic acid - g / dm^3)

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



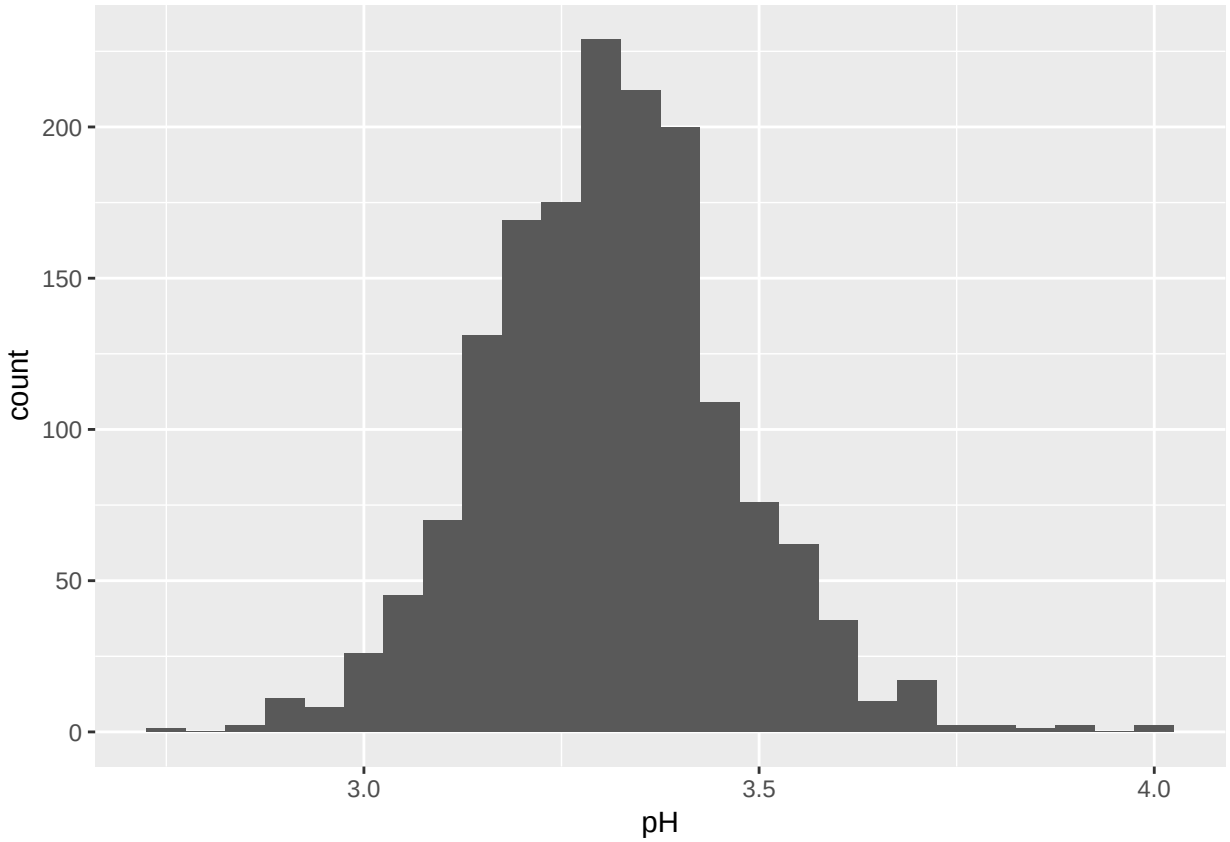
taste

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

At bigger bin width the plot appears to have a long tailm but after reducing the binwidth to 0.02 and removing the outliers,the bimodal nature of the distribution appears which has two peaks at 0.4 and 0.6. In the third graph I have use log transformation to deal with the long tailed nature The mean of the distribution is at 0.52

pH

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

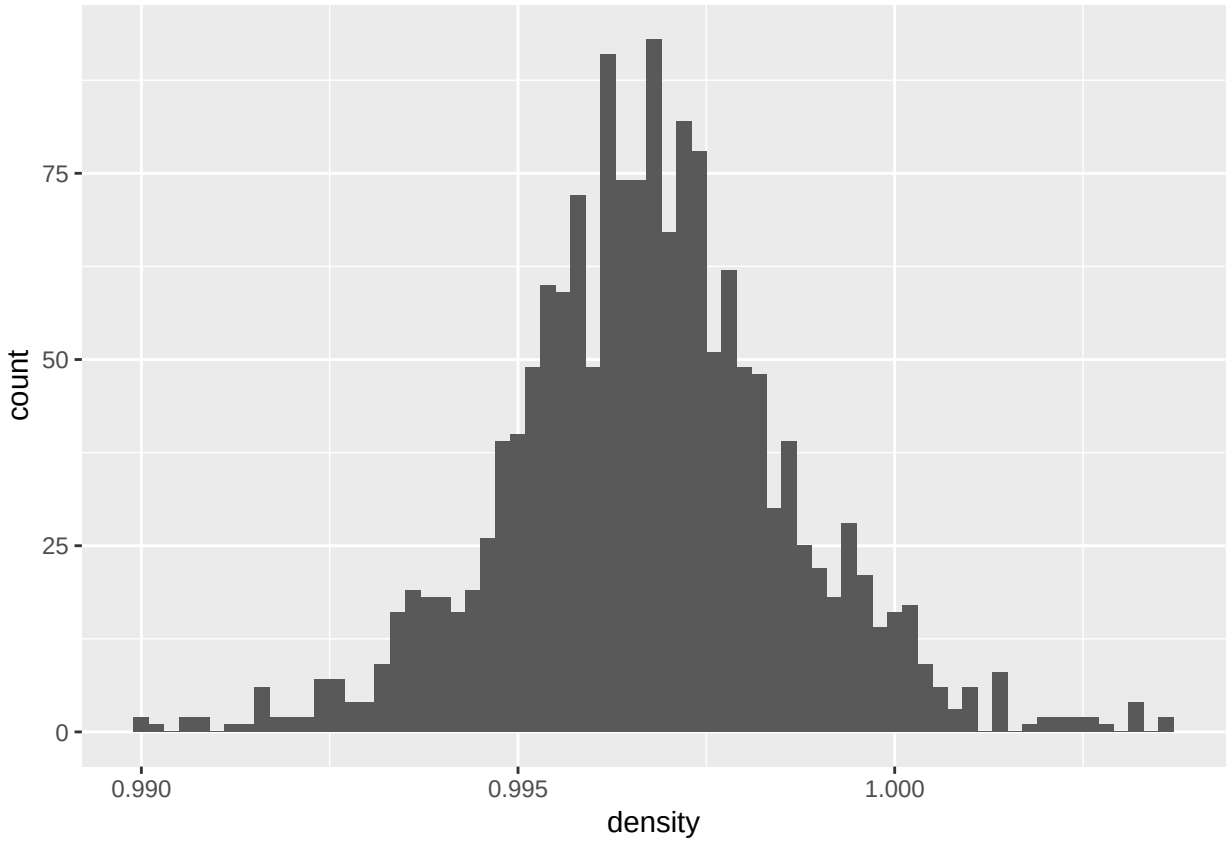



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

The distribution is normal with a little tailing . It has mean of 3.31 and maximum value of 4.01

Density (g / cm^3)

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



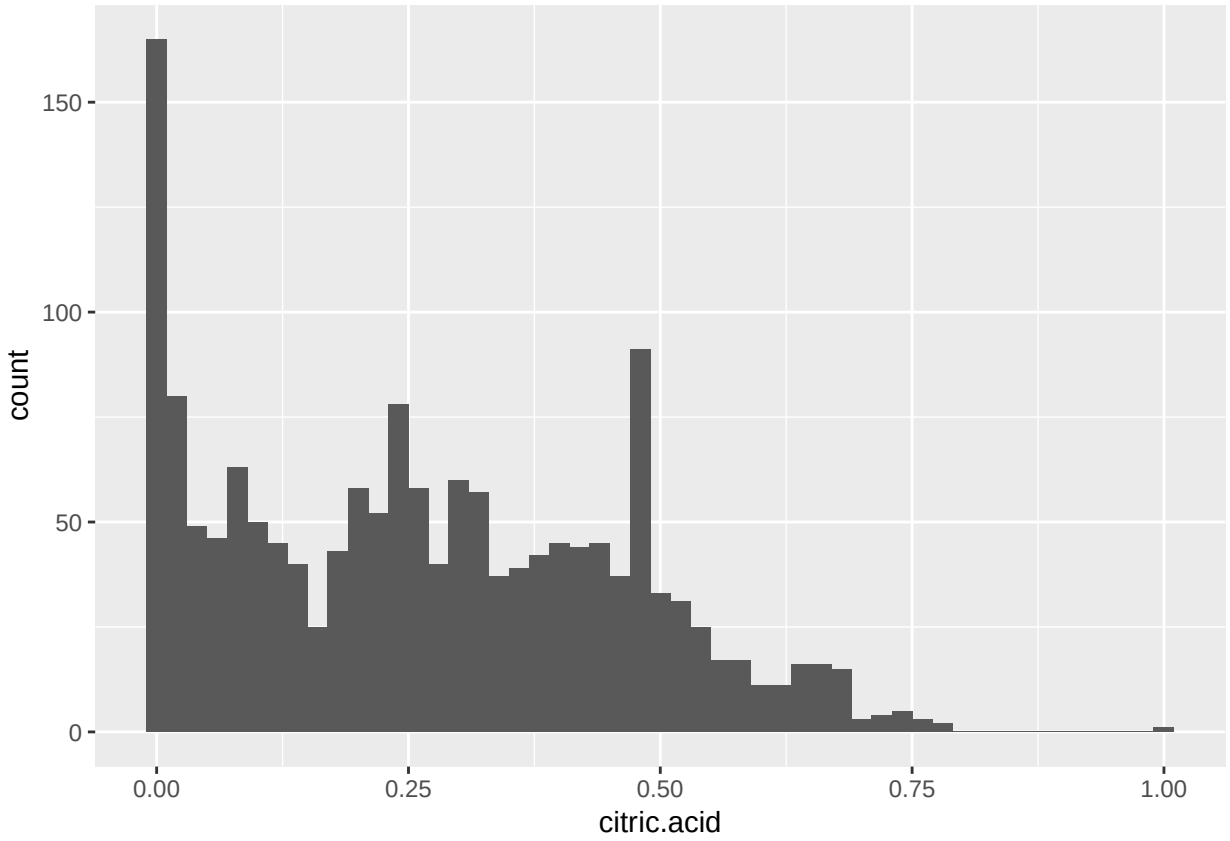

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0037
```

```
## [1] 0.001887334
```

The distribution is normal approaching to 1 which is density of water.The standard deviation is also very small 0.0018

Citric acid (g / dm^3)

citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines

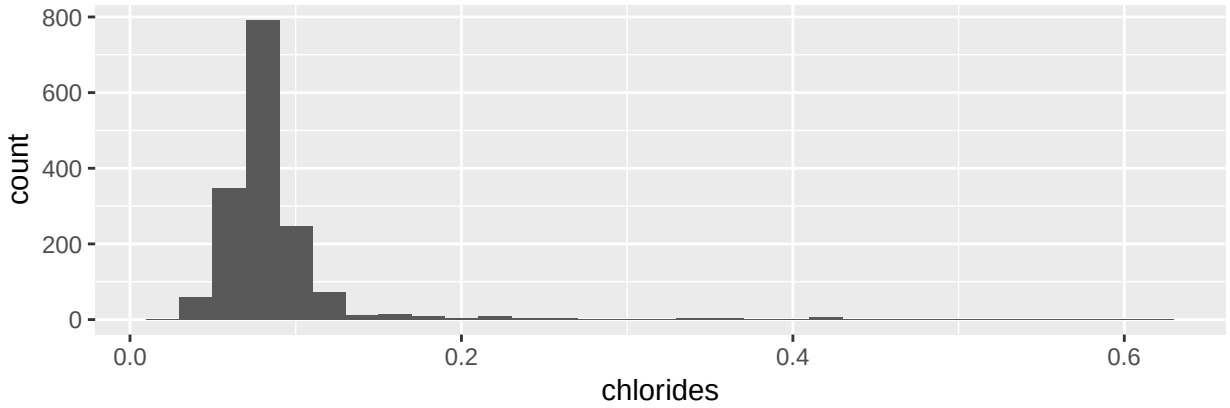


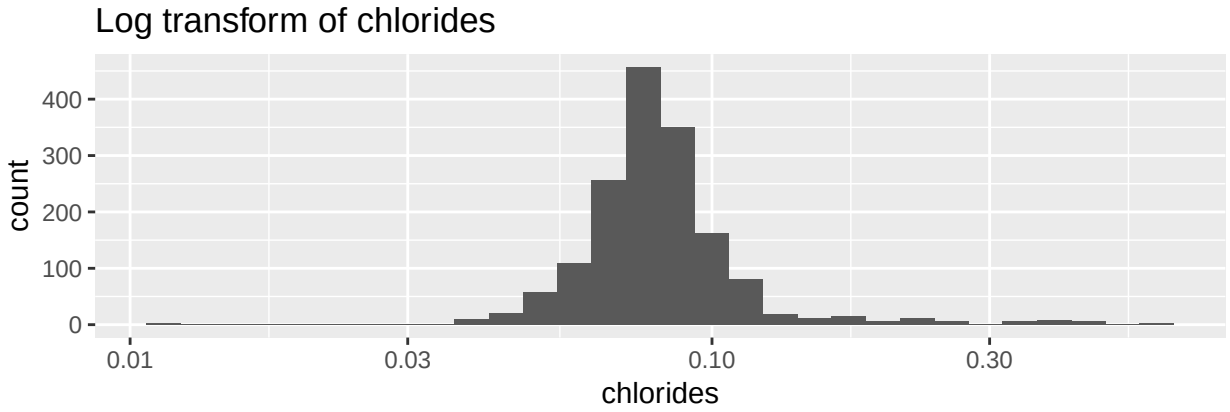

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

We see a positively skewed distribution with mean of 0.27 and max of 1

Chlorides (sodium chloride - g / dm^3)

chlorides: the amount of salt in the wine

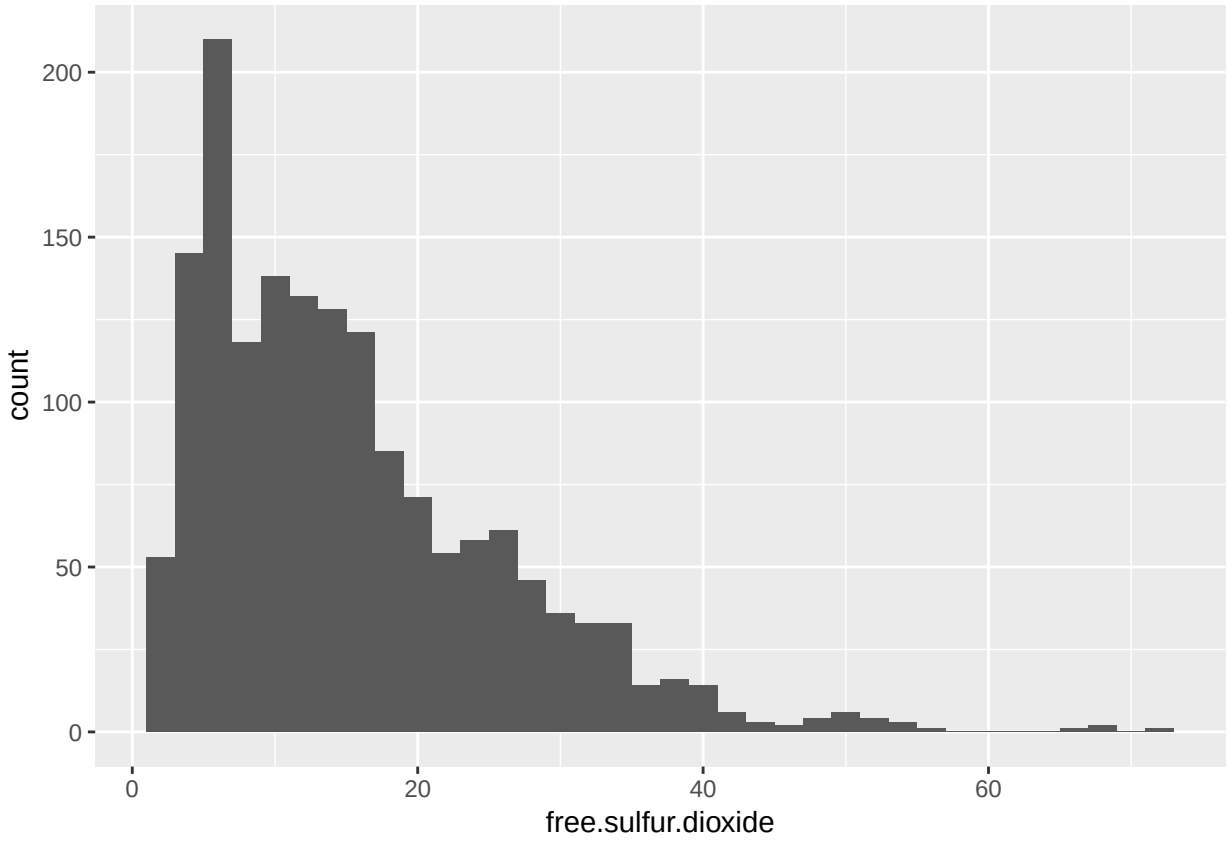





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

Free sulfur dioxide (mg / dm^3)

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

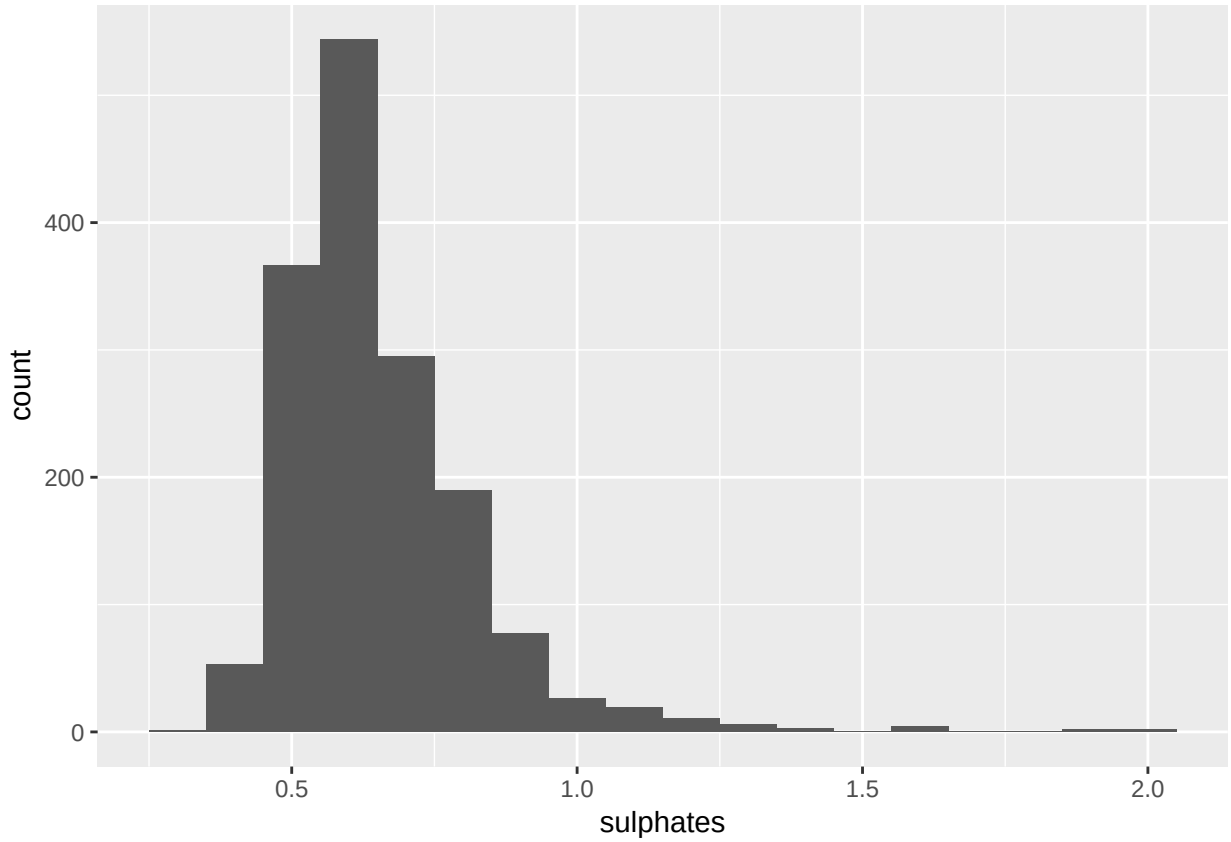



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

The distribution is positive skewed with mean of 15.87

Sulphates (potassium sulphate - g / dm3)

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



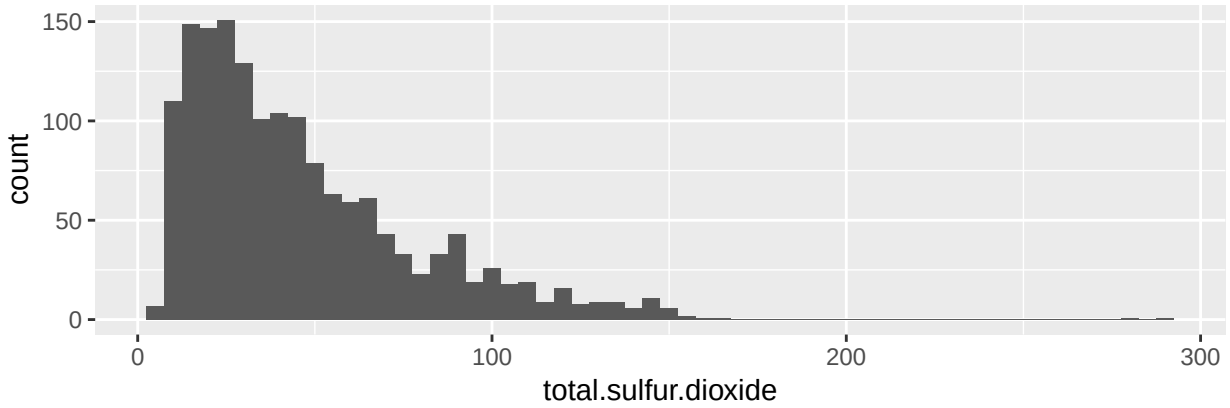
and antioxidant

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

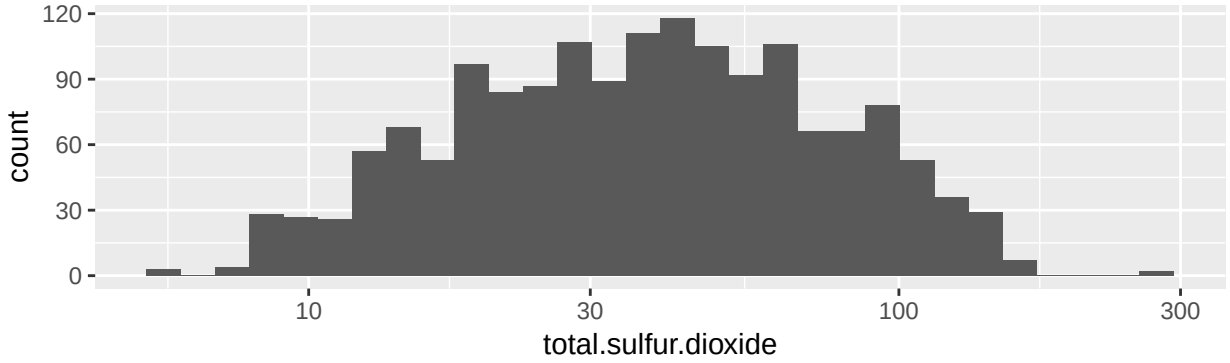
As expected from free sulfur dioxide graph this graph is also positive skewed with long tail.Mean is 0.65 and maximum of 2

Total sulfur dioxide (mg / dmˆ3)

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



Log transform of total SO2




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

Summary of rating A wine

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 4.900   Min.   :0.1200   Min.   :0.0000   Min.   :1.200
##  1st Qu.: 7.400   1st Qu.:0.3000   1st Qu.:0.3000   1st Qu.:2.000
##  Median : 8.700   Median :0.3700   Median :0.4000   Median :2.300
##  Mean   : 8.847   Mean   :0.4055   Mean   :0.3765   Mean   :2.709
##  3rd Qu.:10.100   3rd Qu.:0.4900   3rd Qu.:0.4900   3rd Qu.:2.700
##  Max.   :15.600   Max.   :0.9150   Max.   :0.7600   Max.   :8.900
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 3.00       Min.   :  7.00
##  1st Qu.:0.06200   1st Qu.: 6.00       1st Qu.: 17.00
##  Median :0.07300   Median :11.00       Median : 27.00
##  Mean   :0.07591   Mean   :13.98       Mean   : 34.89
##  3rd Qu.:0.08500   3rd Qu.:18.00       3rd Qu.: 43.00
##  Max.   :0.35800   Max.   :54.00       Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9906   Min.   :2.880   Min.   :0.3900   Min.   : 9.20
##  1st Qu.:0.9947   1st Qu.:3.200   1st Qu.:0.6500   1st Qu.:10.80
##  Median :0.9957   Median :3.270   Median :0.7400   Median :11.60
##  Mean   :0.9960   Mean   :3.289   Mean   :0.7435   Mean   :11.52
##  3rd Qu.:0.9973   3rd Qu.:3.380   3rd Qu.:0.8200   3rd Qu.:12.20
##  Max.   :1.0032   Max.   :3.780   Max.   :1.3600   Max.   :14.00
##     quality      rating
```

```
##  Min.   :7.000   C:  0
##  1st Qu.:7.000   B:  0
##  Median :7.000   A:217
##  Mean   :7.083
##  3rd Qu.:7.000
##  Max.   :8.000
```

Summary of rating C wine

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 4.600   Min.   :0.2300   Min.   :0.0000   Min.   : 1.200
##  1st Qu.: 6.800   1st Qu.:0.5650   1st Qu.:0.0200   1st Qu.: 1.900
##  Median : 7.500   Median :0.6800   Median :0.0800   Median : 2.100
##  Mean   : 7.871   Mean   :0.7242   Mean   :0.1737   Mean   : 2.685
##  3rd Qu.: 8.400   3rd Qu.:0.8825   3rd Qu.:0.2700   3rd Qu.: 2.950
##  Max.   :12.500   Max.   :1.5800   Max.   :1.0000   Max.   :12.900
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.04500   Min.   : 3.00       Min.   :  7.00
##  1st Qu.:0.06850   1st Qu.: 5.00       1st Qu.: 13.50
##  Median :0.08000   Median : 9.00       Median : 26.00
##  Mean   :0.09573   Mean   :12.06       Mean   : 34.44
##  3rd Qu.:0.09450   3rd Qu.:15.50       3rd Qu.: 48.00
##  Max.   :0.61000   Max.   :41.00       Max.   :119.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9934   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9957   1st Qu.:3.300   1st Qu.:0.4950   1st Qu.: 9.60
##  Median :0.9966   Median :3.380   Median :0.5600   Median :10.00
##  Mean   :0.9967   Mean   :3.384   Mean   :0.5922   Mean   :10.22
##  3rd Qu.:0.9977   3rd Qu.:3.500   3rd Qu.:0.6000   3rd Qu.:11.00
##  Max.   :1.0010   Max.   :3.900   Max.   :2.0000   Max.   :13.10
##     quality      rating
##  Min.   :3.000   C:63
##  1st Qu.:4.000   B: 0
##  Median :4.000   A: 0
##  Mean   :3.841
##  3rd Qu.:4.000
##  Max.   :4.000
```

I am looking at the means of the distributions for identifying better variation between A and C rating wines. Some variations are in A—>B terms (percentage is taken over A)

Fixed.acidity - mean reduced by 11%

Volatile.acidity -mean increased by 80%

citric.acidity-mean increased by by 117%

alcohol - mean reduced by 12.7%

(This is just for estimation purposes and setting a way for further analysis.No final conclusion should be drawn from it)

residualsugar and chloride showed a very low variation

Univariate Analysis

What is the structure of your dataset?

Dataset contains 1599 observations with 13 variable (14 is we count the the new ordered factor of rating).The data set is tidy.All values are float expect x and quality which are integer values

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

The main feature of interest in the dataset would be how alcohol concetrations effect the quality of red wine.I am also interested in looking at how pH varies in wines.

What other features in the dataset do you think will help support your

investigation into your feature(s) of interest?

I believe volatile acidity,citric acidity ,pH will also play a deep role . I also suspect that sulphates would have a positive impact on quality

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

Yes i have created a variable rating which is a ordered factor of quality variable which is as follows 3 to 4 are C 5 to 6 are B 7 to 8 are A

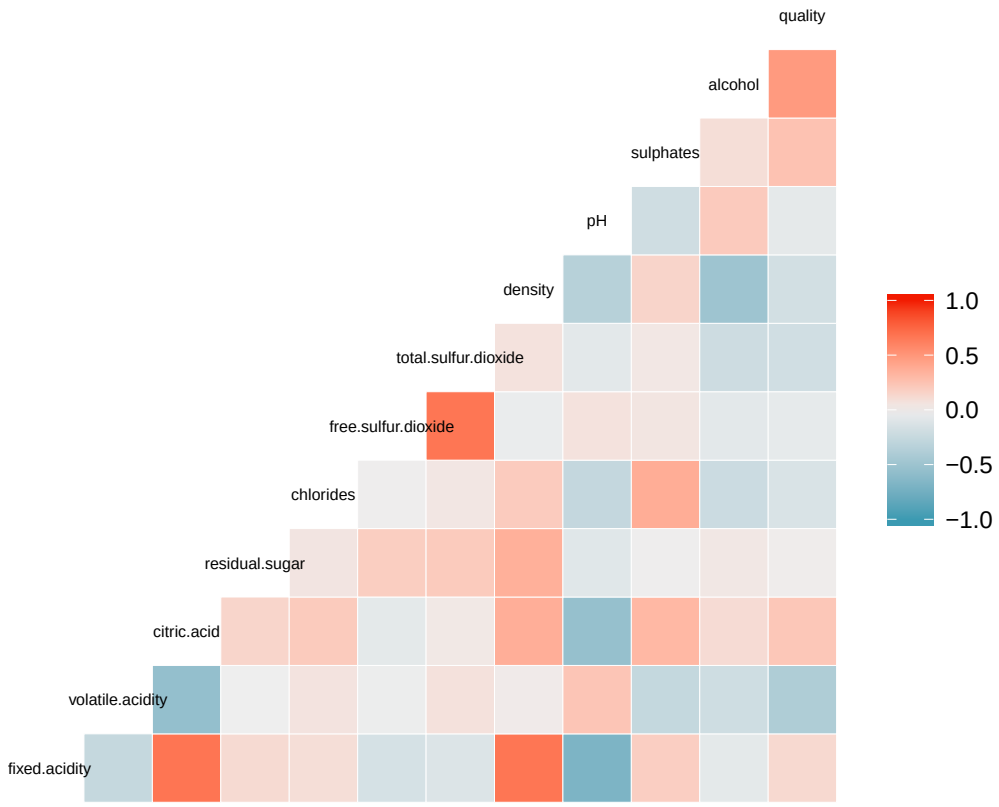
Of the features you investigated, were there any unusual distributions?

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?

No operations were needed to tidy the data, all observations were complete. Data appears to be wrangled There were some unusual distributions and outliers. The outliers seem to be true data values rather than input errors,although in some cases I have taken 99 and 95 qualtiles to make my graphsbetter

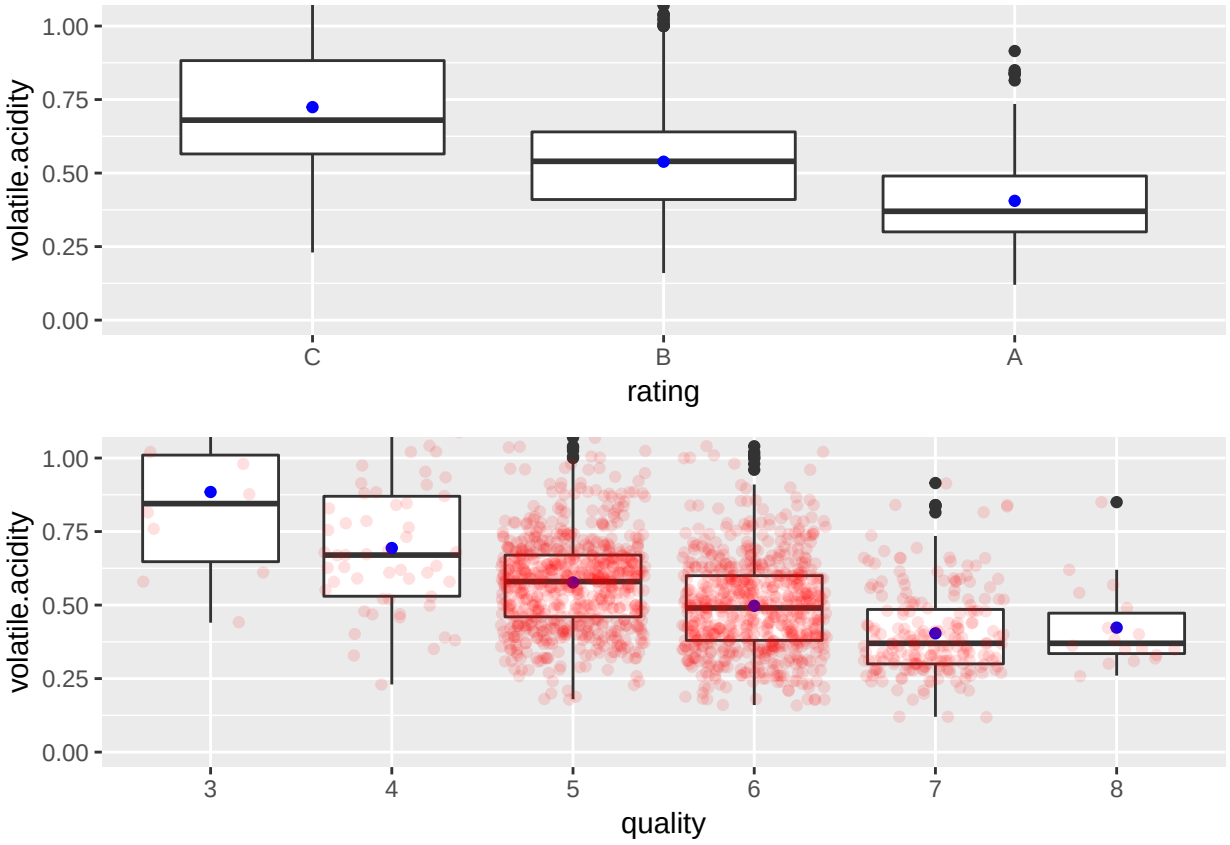
Bivariate Plots Section

Correlation among various variables



We can see the quality has a positive correlation with alcohol and citric acid and suplhates (as suspected in univarent) while it has a strong negative correlation with volatile acidity

Volatile acidity vs Rating and quality



The box plots clearly shows how the volatile acidity decreases as the quality of wine improves.Well its not shocking who want to have a wine tasting like a 'sour' acid .The first boxplot shows the compares between our rating order pair the second one between the quality. The boxplot having quality also depicts the distribution of various wines and we can again see 5 and 6 quality wines have the most share.The blue dot is for the mean and the middle line shows the median

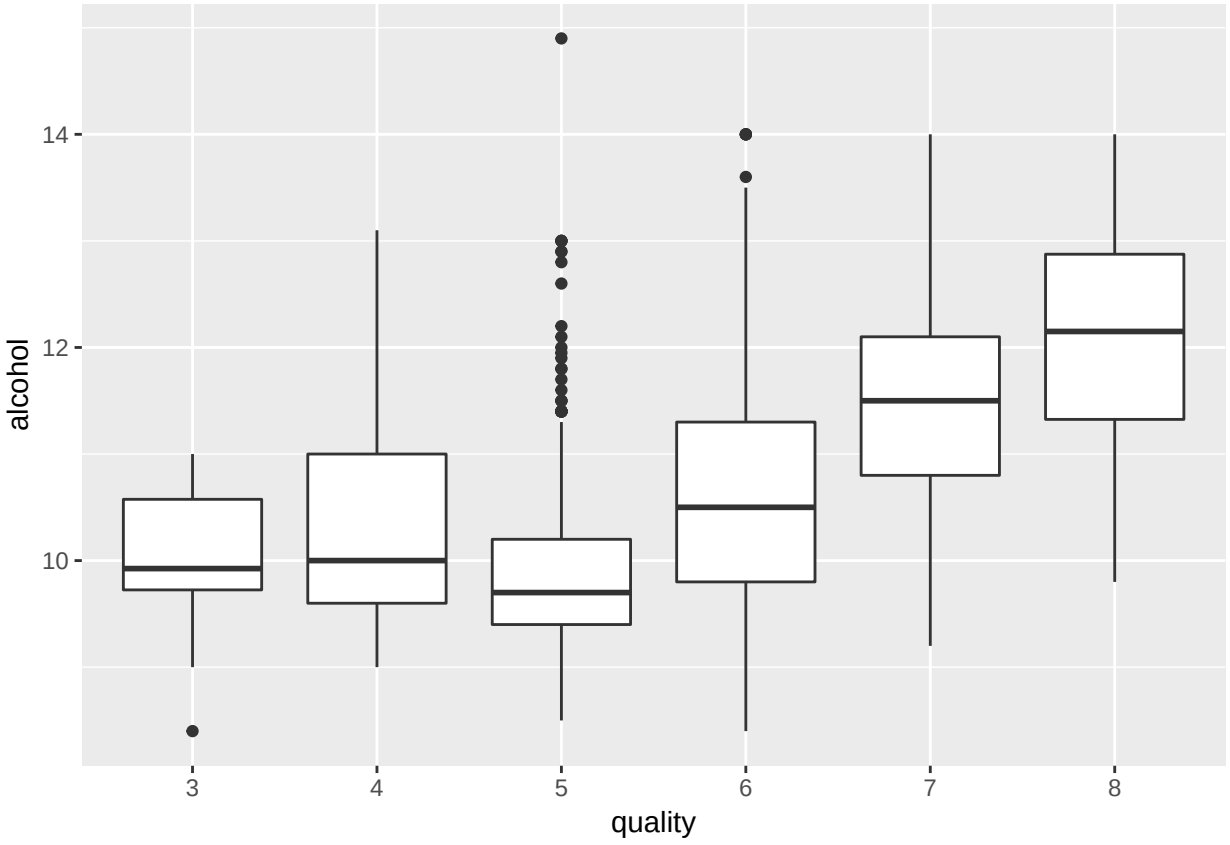
```
##
##  Pearson's product-moment correlation
##
## data:  rwine$volatile.acidity and rwine$quality
## t = -16.954, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4313210 -0.3482032
## sample estimates:
##        cor
## -0.3905578
```

```
##
##  Kendall's rank correlation tau
##
## data:  rwine$volatile.acidity and rwine$quality
## z = -15.498, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
```

```
##        tau
## -0.3007787
```

Both Pearson's and kendall's correlations are negative showing inverse trends like we figured out from the boxplots Further this high value of kendall's rank correlation means high monotonic nature

Quality vs Alcohol

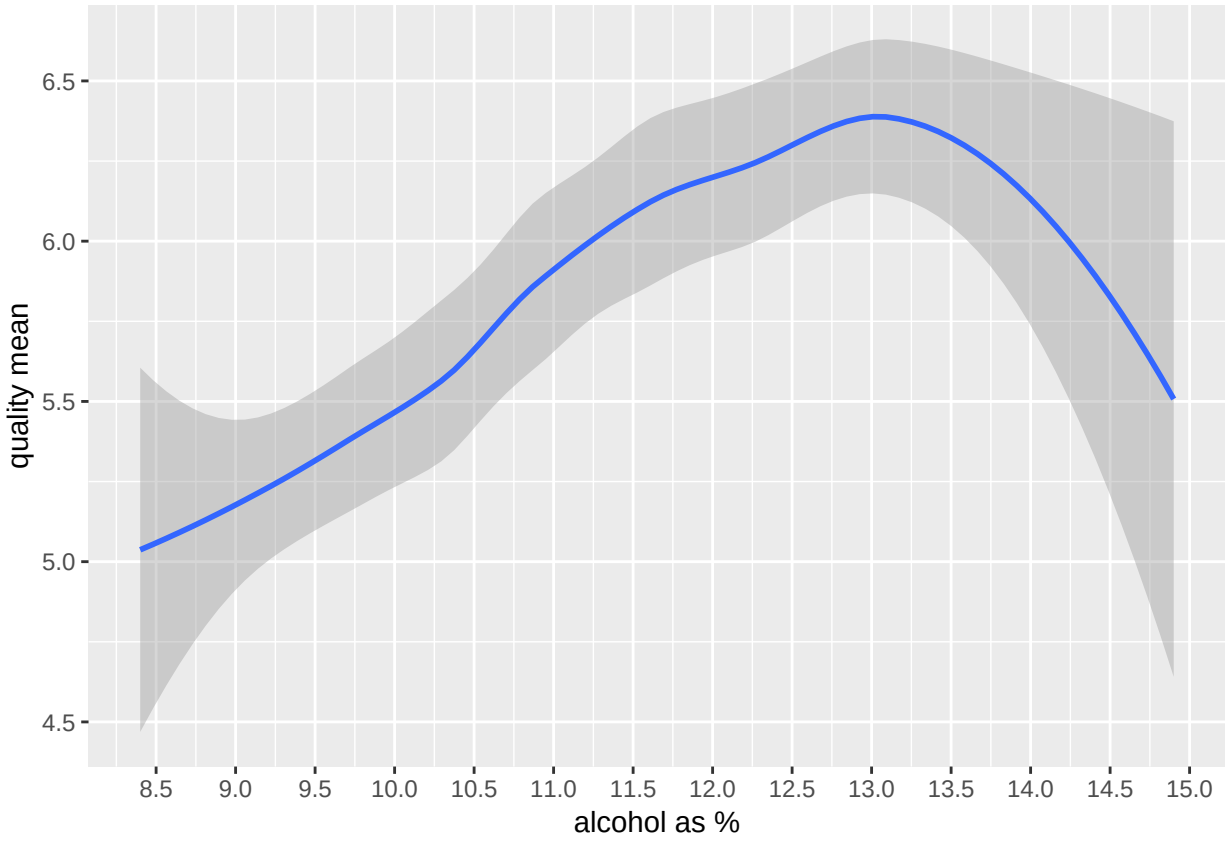



```
##
##  Pearson's product-moment correlation
##
## data:  rwine$alcohol and rwine$quality
## t = 21.639, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4373540 0.5132081
## sample estimates:
##       cor
## 0.4761663
```

The boxplots show an indication that higher quality wines have higher alcohol content.This trend is shown by all the quality grades from 3 to 8 except quality 5. The pearson's R gives us a very high correlation of 0.47

So do if we keep on adding more alcohol will it give us a better wine?

Not necessarily, here is an example of how a more alcohol will ruin a top graded wine or have no effect on its quality



The above line plot indicates as sort of linear increase till 13% alcohol concetration, followed by a steep downwards trend. The graph has be smoothed to remove variances and noise.

```
##
##  Pearson's product-moment correlation
##
## data:  above13$quality and above13$alcohol
## t = -0.39861, df = 21, p-value = 0.6942
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4816540  0.3376049
## sample estimates:
##         cor
## -0.08665653
```

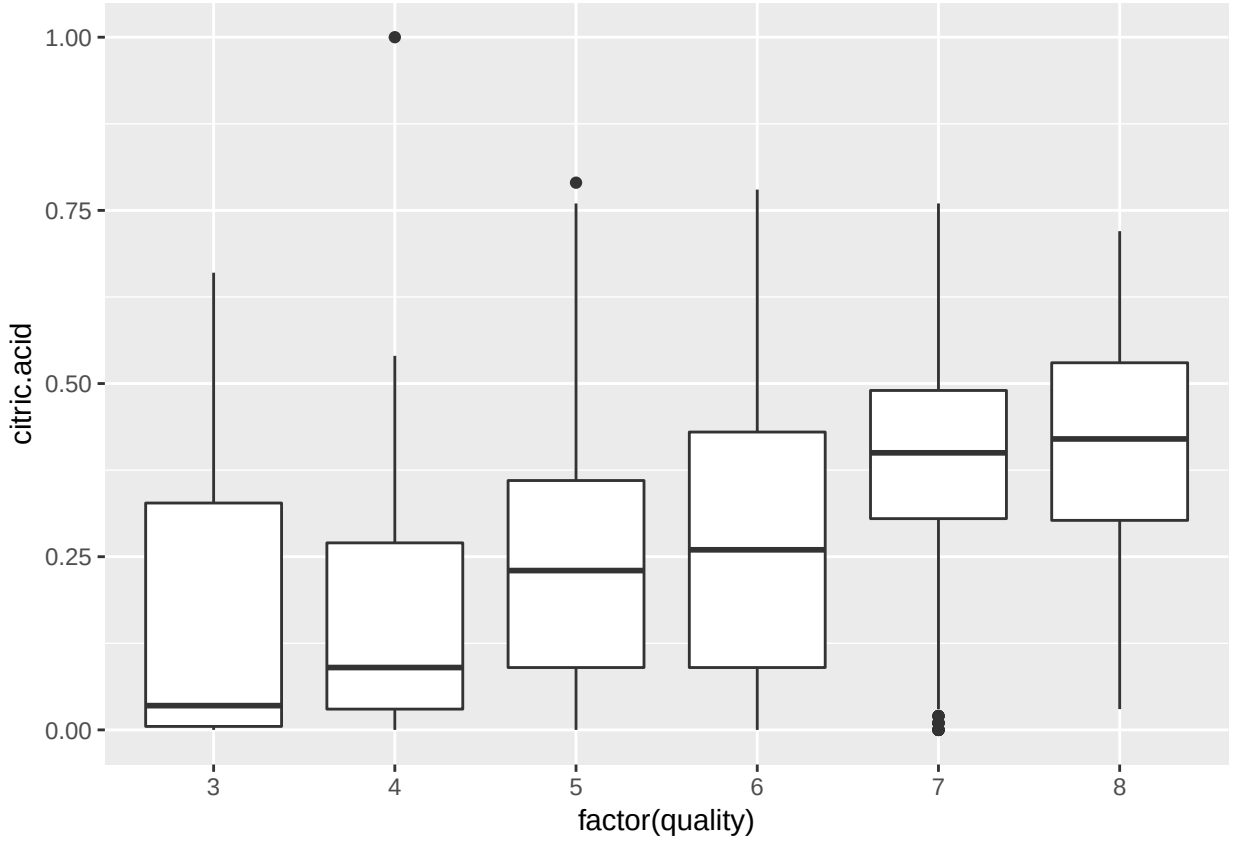
It interesting to see that the correlation between alcohol and quality diminishes and even becomes negative(reciprocal effect) when we approach above the 13% alcohol by volume mark

###Quality vs citric acid

```
##
##  Pearson's product-moment correlation
##
## data:  rwine$citric.acid and rwine$quality
## t = 9.2875, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1793415 0.2723711
```

```
## sample estimates:
##       cor
## 0.2263725
```

The correlation between citric acid and quality is 0.226 though being a weak correlation it do effect the quality of

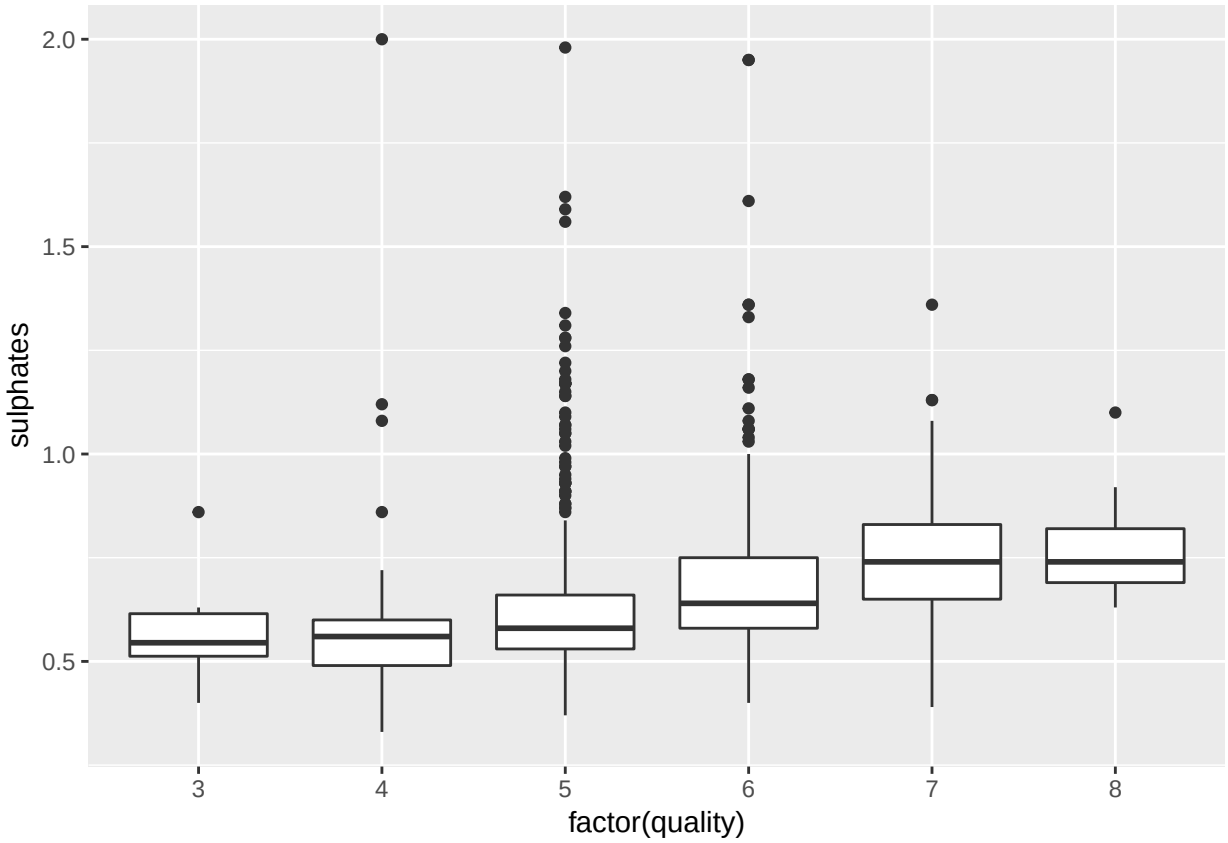


wine

The boxplots shows how the citric acid median values shows a steady increase as we move to better quality wines

Quality vs Suplhates

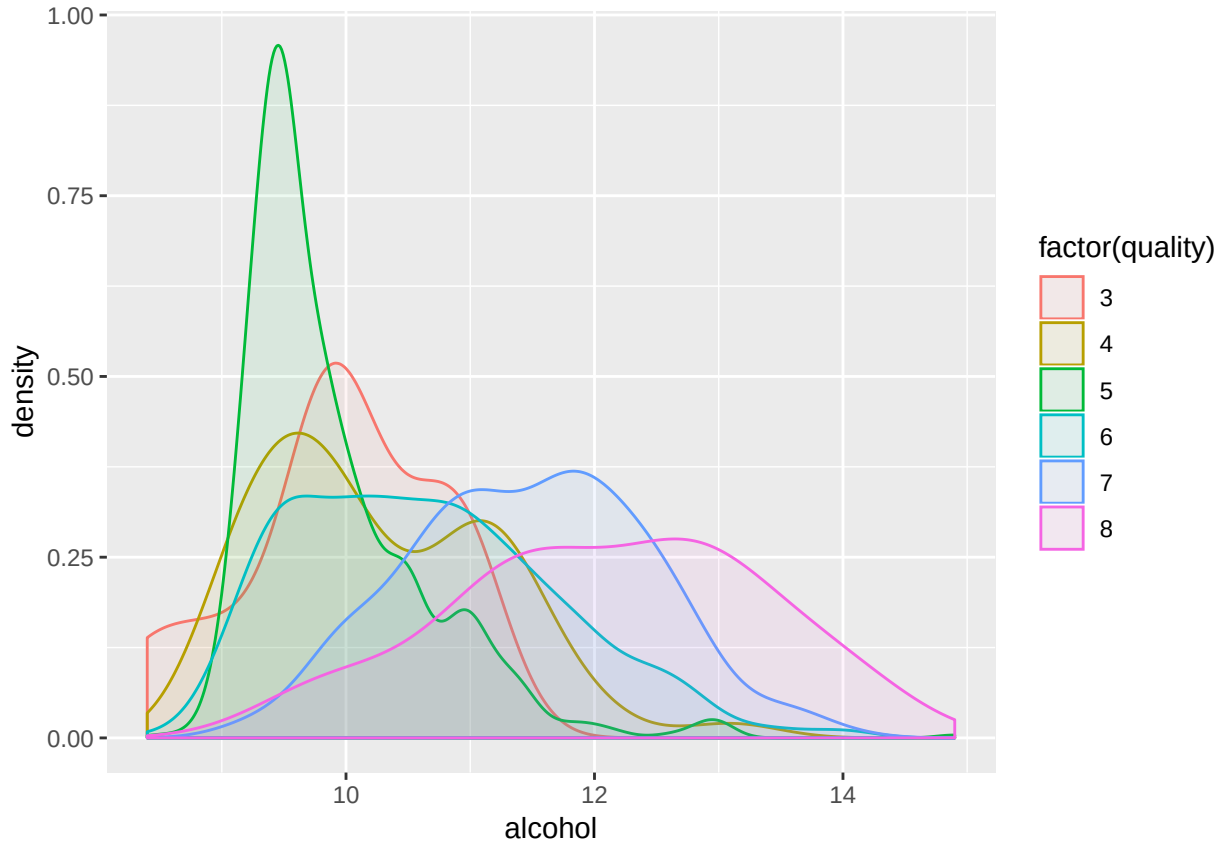
```
##
##  Pearson's product-moment correlation
##
## data:  rwine$sulphates and rwine$quality
## t = 10.38, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2049011 0.2967610
## sample estimates:
##       cor
## 0.2513971
```



Good wines have higher sulphates values than bad wines, though the difference is not that wide. The correlation between these two variables is 0.251

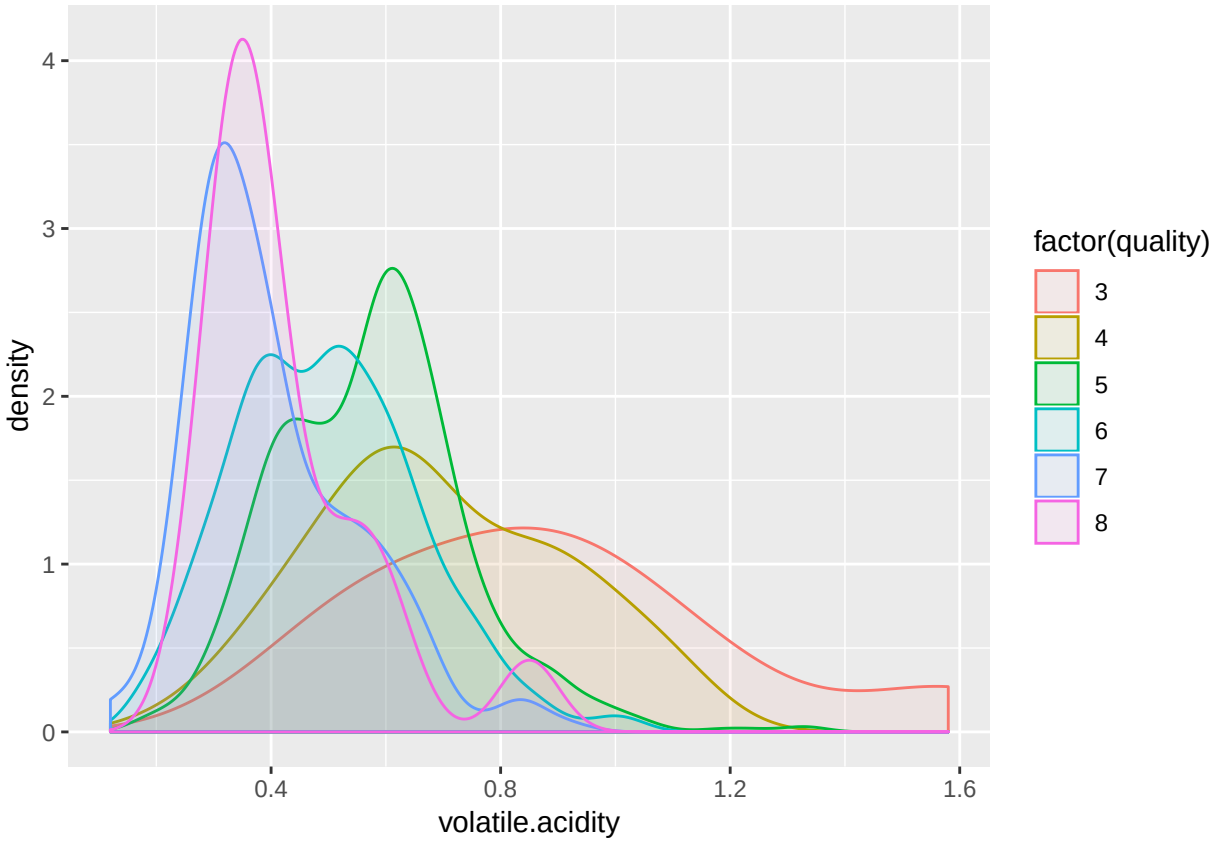
####Veiw of Quality through density plots Density plots are an interesting way of visulisation ,I have provided them for some variables in relation with quality

For alcohol



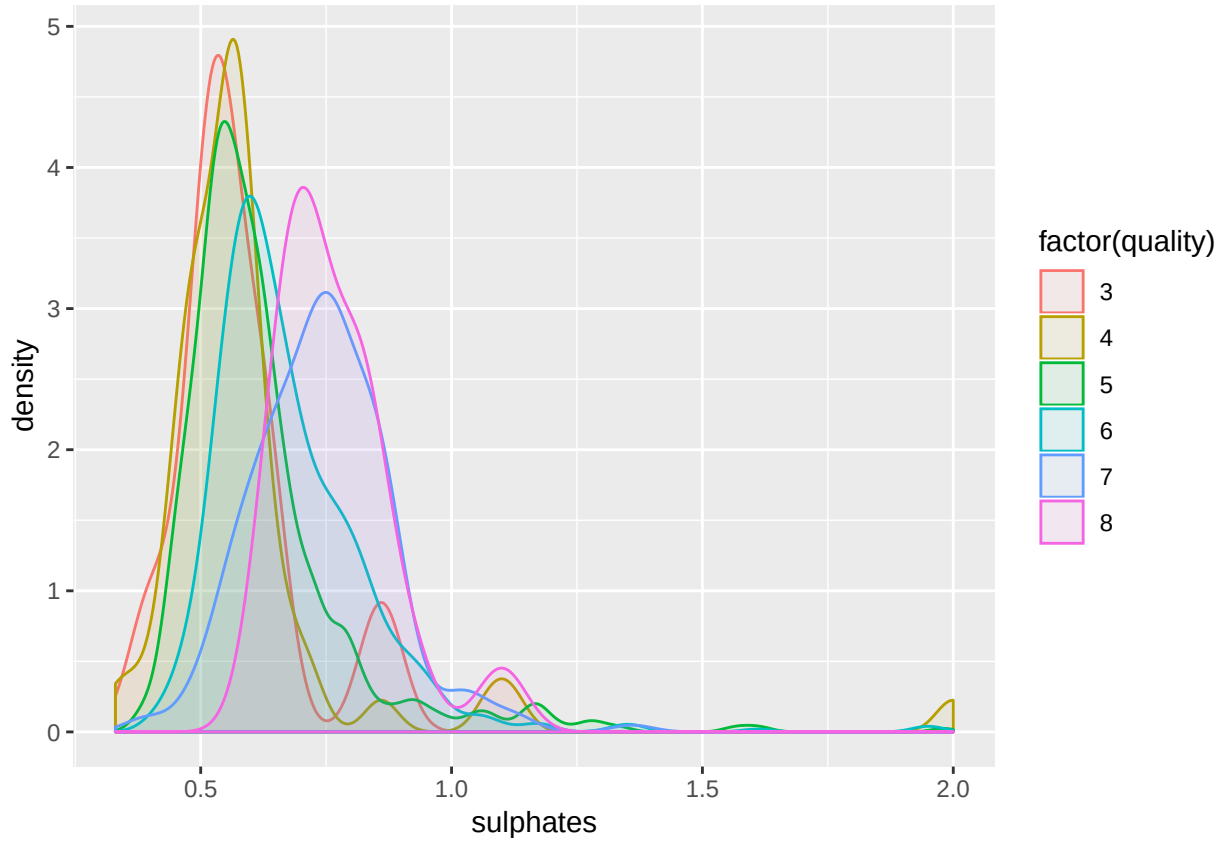
For 5 quality wine we can see a distinct peak at about 8% alcohol. For increase in quality the maximum value moves towards right hand side(more alcohol)

For volatile acidity



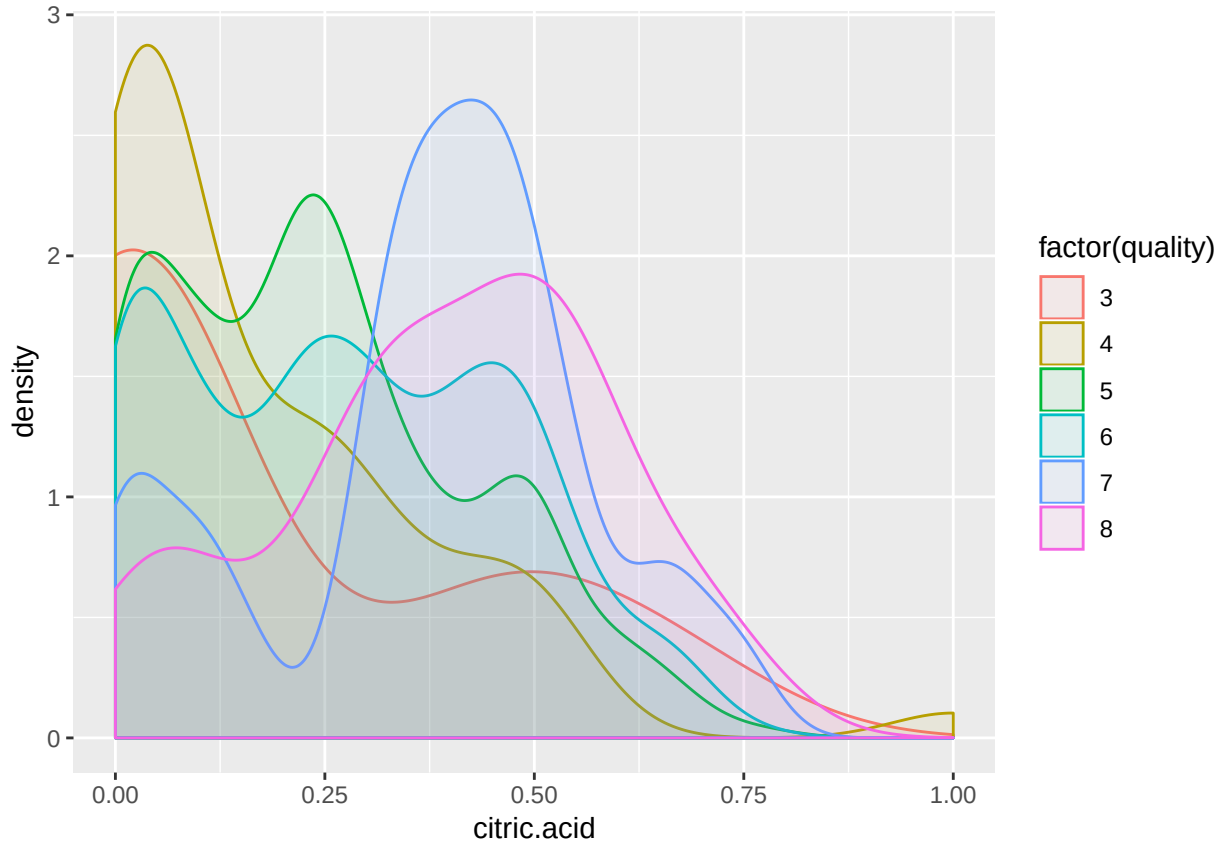
Redwine of quality 7 and 8 have their peaks for volatile acidity well below the 0.4 mark.Wine with quality 3 has the pick at the most right handside (towards more volatile acidity)

For sulphates



Nearly all wines are below 1.0 marks.We see a sideward movement of peaks towards more sulphate side as we increase the quality .This was also observered with alcohol

For citric acid

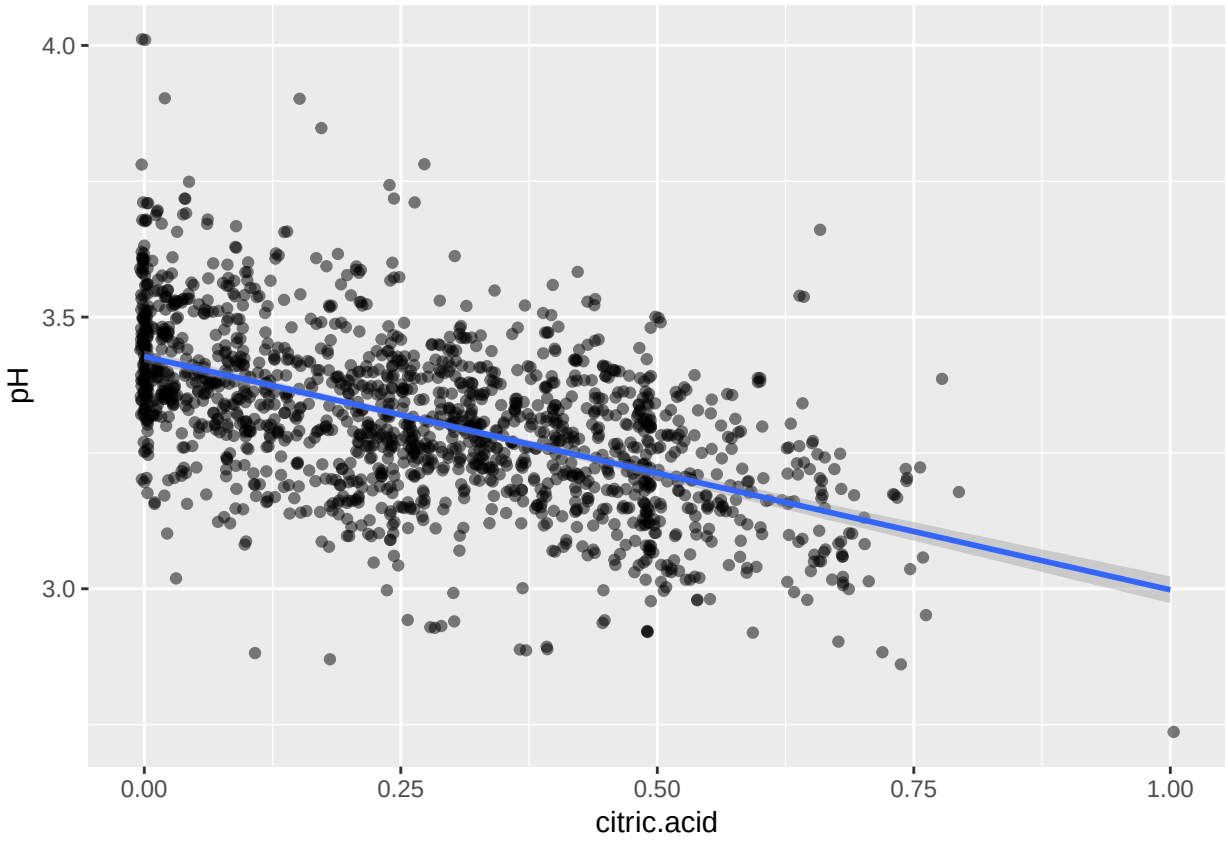


7 and 8 quality wine appear to be clear winners here .Citric acid had a positive correlation with quality which is depicted well by density plots

Fixed acidity and citric acid relation with pH

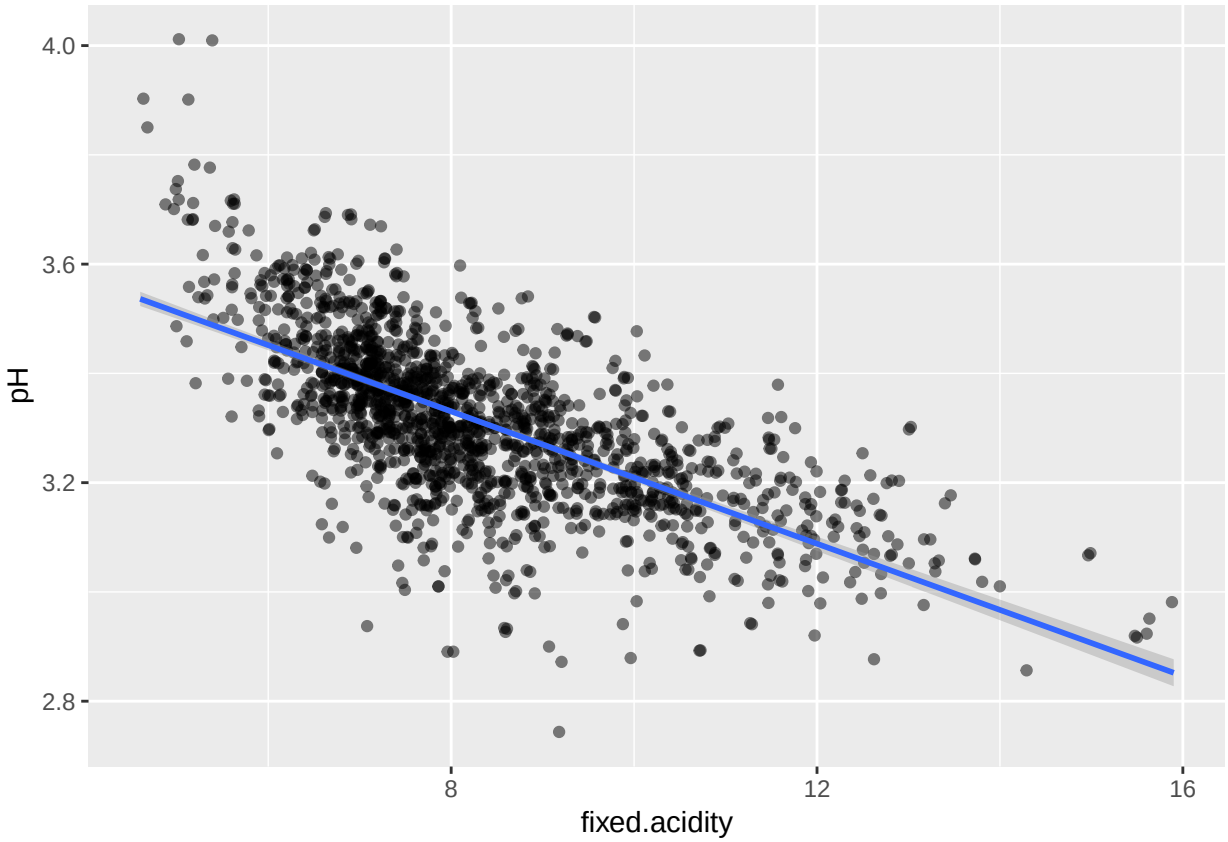
Both acidity should have a correlation with pH as pH scales is made for acids and bases

```
##
##  Pearson's product-moment correlation
##
## data:  rwine$pH and rwine$citric.acid
## t = -25.767, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5756337 -0.5063336
## sample estimates:
##        cor
## -0.5419041
```

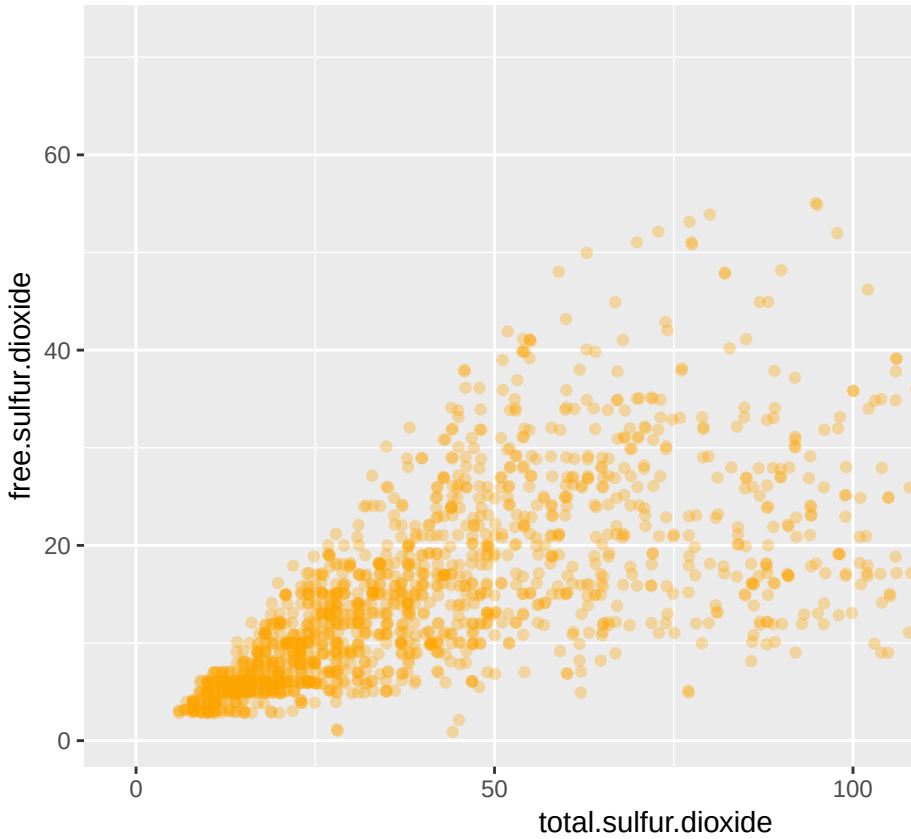


Well negative correlation and the scatter plot shows how the citric acid increment lowers pH

```
##
##  Pearson's product-moment correlation
##
## data:  rwine$fixed.acidity and rwine$pH
## t = -37.366, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7082857 -0.6559174
## sample estimates:
##       cor
## -0.6829782
```



Same is for fixed acidity the correlation is at wooping -0.68



###Free sulfur dioxide vs total sulfur dioxide

A scatter plot between total sulfur dioxide and free sulfur dioxide.shows that the both entities are related. The following scatter plot doesnot shows a linear trend but shows a conical arrangement

```
##
##  Pearson's product-moment correlation
##
## data:  rwine$free.sulfur.dioxide and rwine$total.sulfur.dioxide
## t = 35.84, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6395786 0.6939740
## sample estimates:
##       cor
## 0.6676665
```

The correlation between the two explains the scatterplot further.In most cases correlation does not implies causation but in this case the increase in total sulphur dioxide is causing increase in free sulphur dioxide due to its decay

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?

My main analysis was regarding quality these are some of the relationsips i observed

Alcohol

Below 13 % - Alcohol has a positive correlation of 0.47 the heighest correlation on positive side with quality of red wine.The boxplots also show an increase in median alcohol % with increase in quality

above 13% - Trends seem to be different as we epproach above 13% mark.The correlation diminish rather become negative

#####Volatile acidity Volatile acidity had a pearson's R of -0.39 and kendall's tau of -0.30 which shows the inverse relation between quality and volatile acidity.Volatile acidity Even the boxplots showed the median for each increase in quality has a lower value of volatile acidity

#####Citric acid Citric acid showed an positive correlation of 0.22 and the boxplots depicted a rise in its concentrations from low quality wines to high quality onces.It incresting to note that the values of citric acid for a quality lying in same rating group had nearly same concetrations .i.e the concentration of 3-4 were nearly same.A exponetial rise in citric acid came then we moved from C rating wine to B rating one or B rating one to A rating one

#####Sulphates Sulphates are suspected to have a positive inpact on quality ,i.e increase in there value will increase in wine quality.We will see in our further analysis

Fixed acidity and citric acid relation with pH was also observed .Though results were not that surprising

Did you observe any interesting relationships between the other features

(not the main feature(s) of interest)?

Yes, I found the relationship between free sulfur dioxide and total sulfurdioxide which were not main part of my research also high correlation between density and fixed acidity was observed

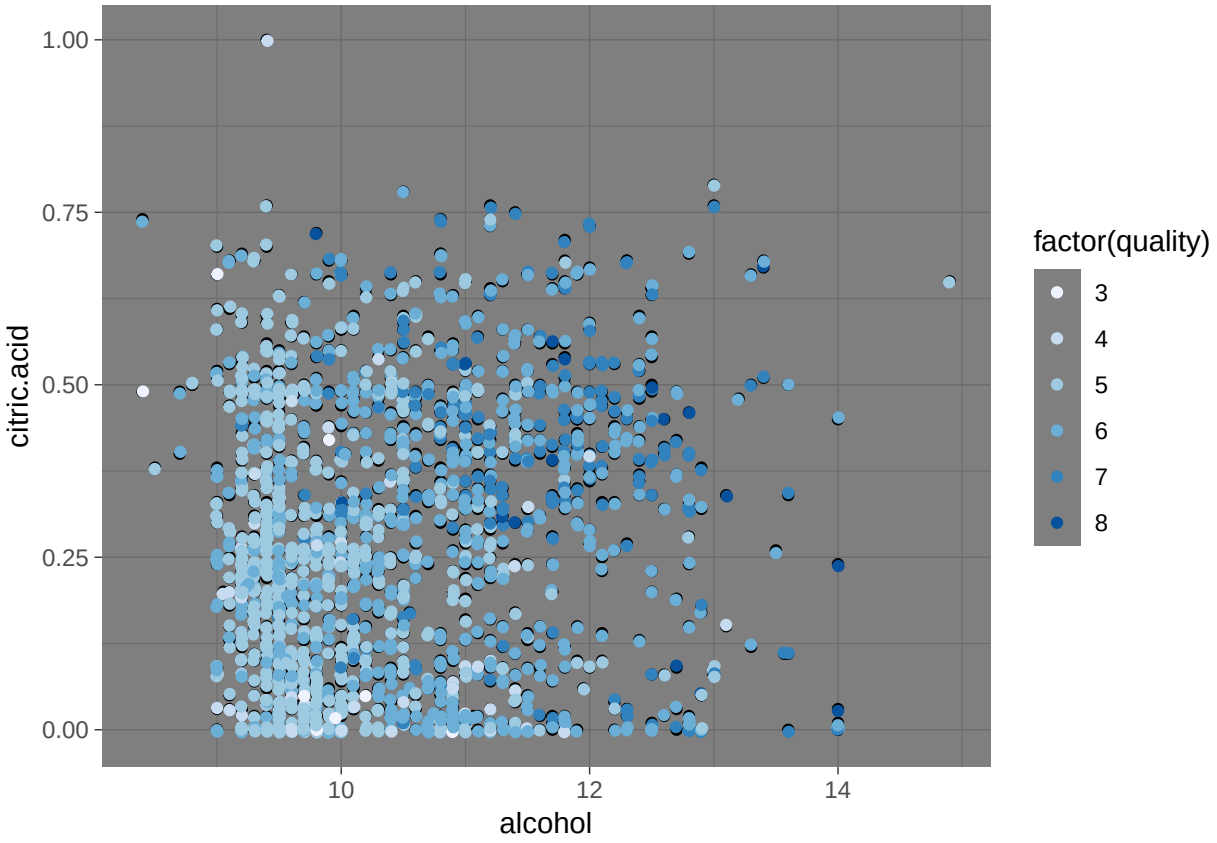
What was the strongest relationship you found?

I found the strongest relation between pH and fixed acidity at wooping -0.68.I was expecting it,The second strongest trend was between totalsulfurdioxide and free sulphur dioxide.

For quality analysis of red wine i found that alcohol and volatile acidity had the strongest relationship

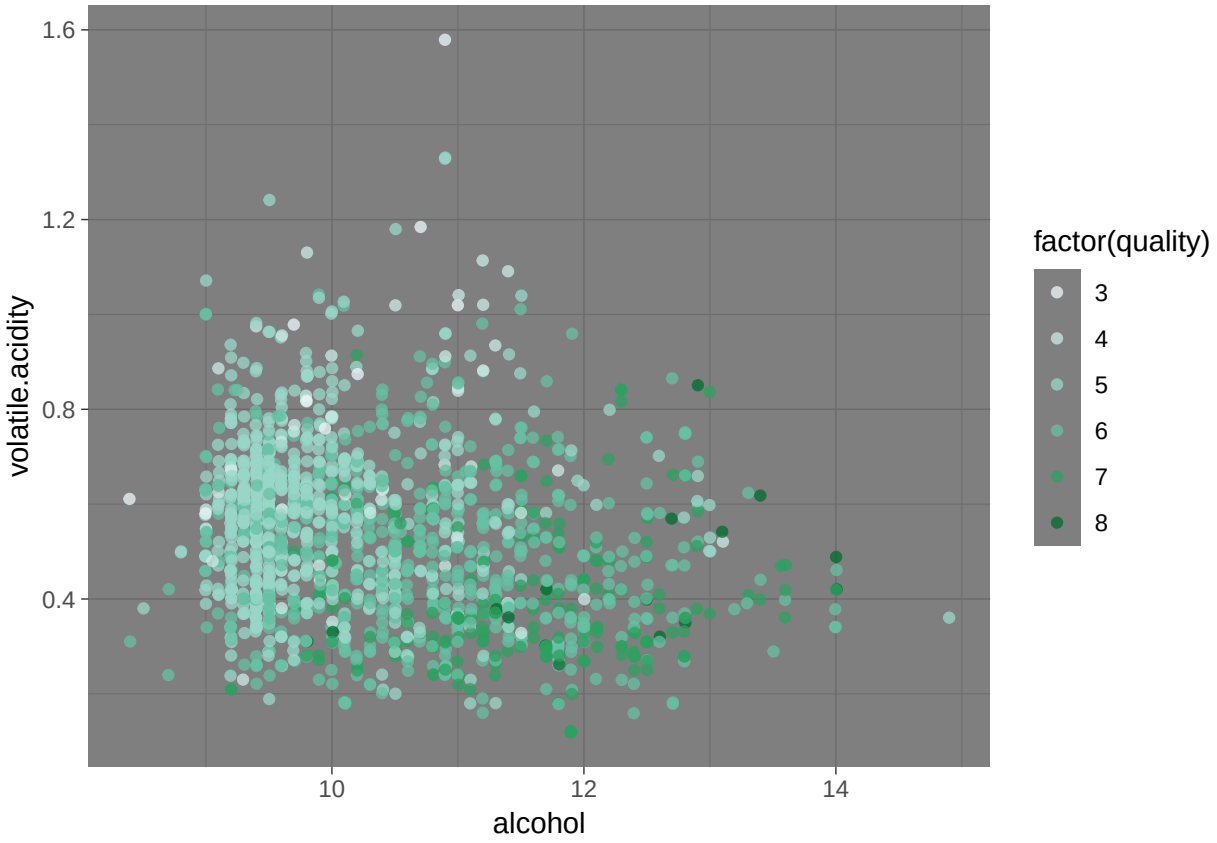
Multivariate Plots Section

Alcohol and citric Acid



Values above 11% by volume of alcohol and 0.25 of citric acid result in rating A wines which are quality 7 and above . Mosr rating B wines lave a lower alcohol concentrations while the rating C wine also have low citric acid

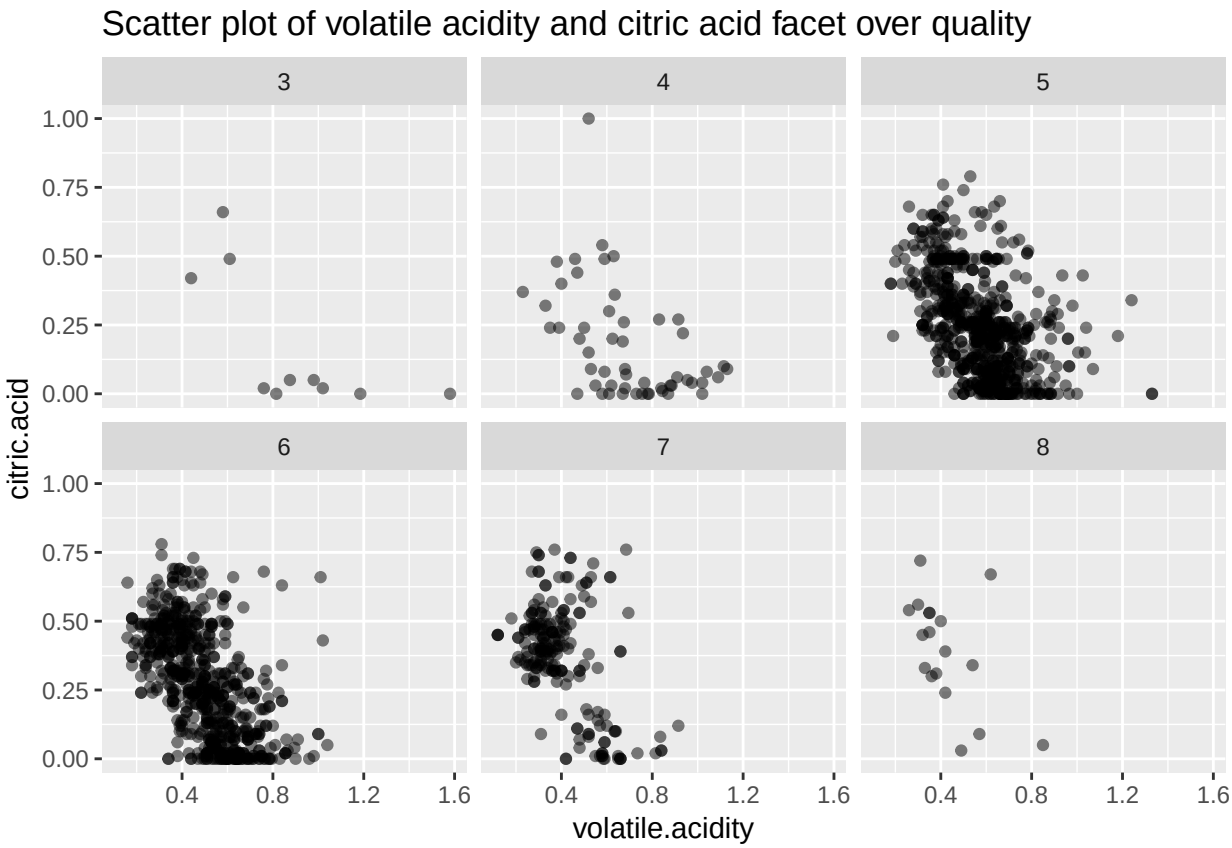
Alcohol and Volatile acidity

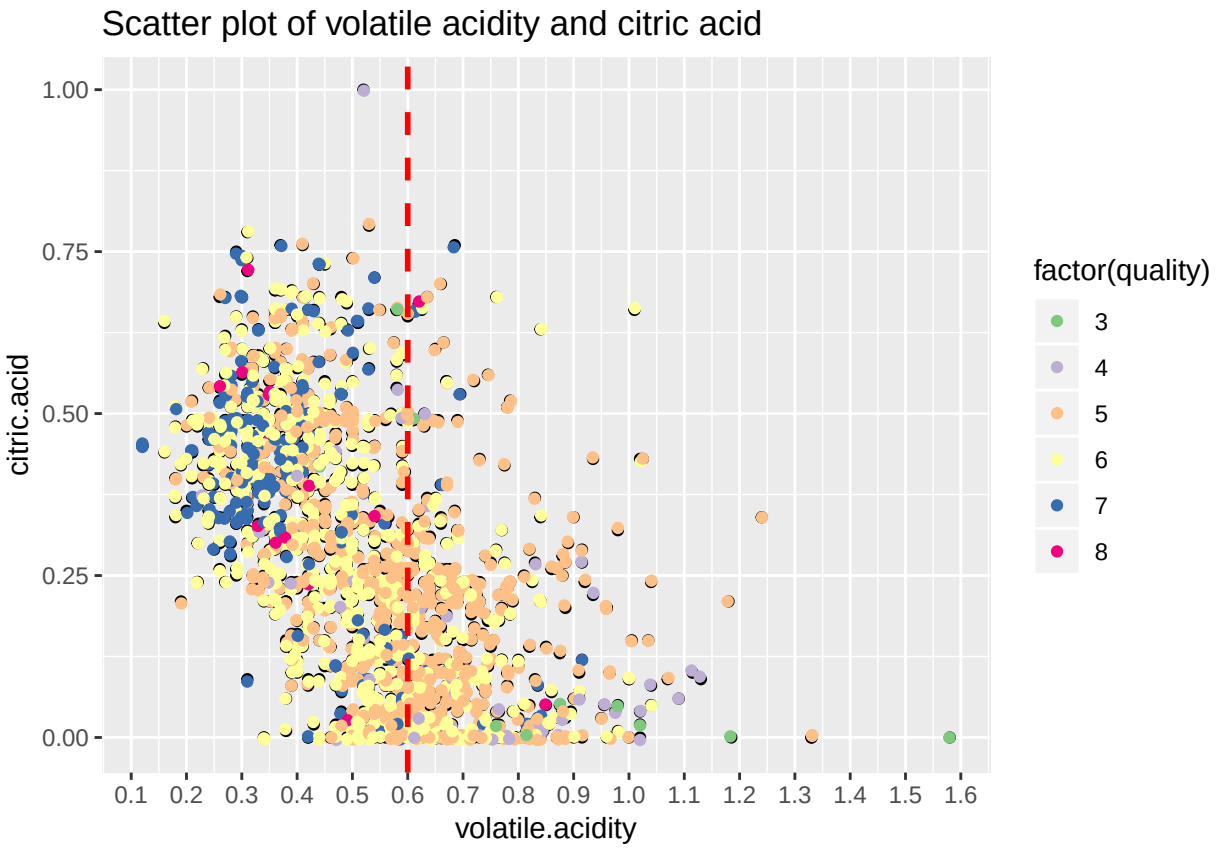


Even though the volatile acidity and alcohol had high correlations values of negative and positive.Alcohol seems to vary more than volatile acidity when we talk about quality,nearly every Rating A wine has less than 0.6 volatile acidity.Well I thinks its acceptable as volatile acid is acetic acid, no one wants their redwine to taste like vinegar

Lets talk acid!

volatile acidity has negative corelation while citric acidity has a positive one I was interested in seeing how they related with each other

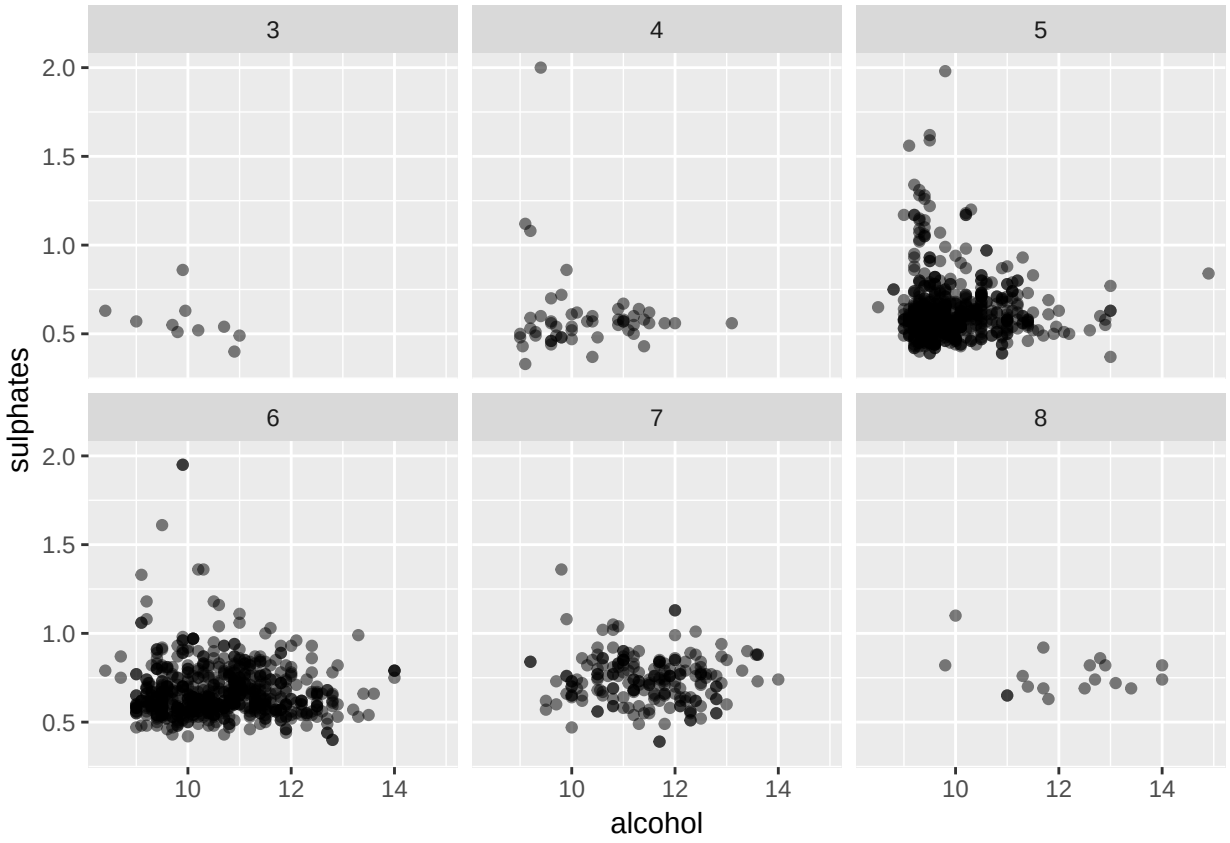

Scatter plot of volatile acidity and citric acid facet over quality
3
4
5
6
7
8
citric.acid
1.00
0.75
0.50
0.25
0.00
0.4
0.8
1.2
1.6
volatile.acidity

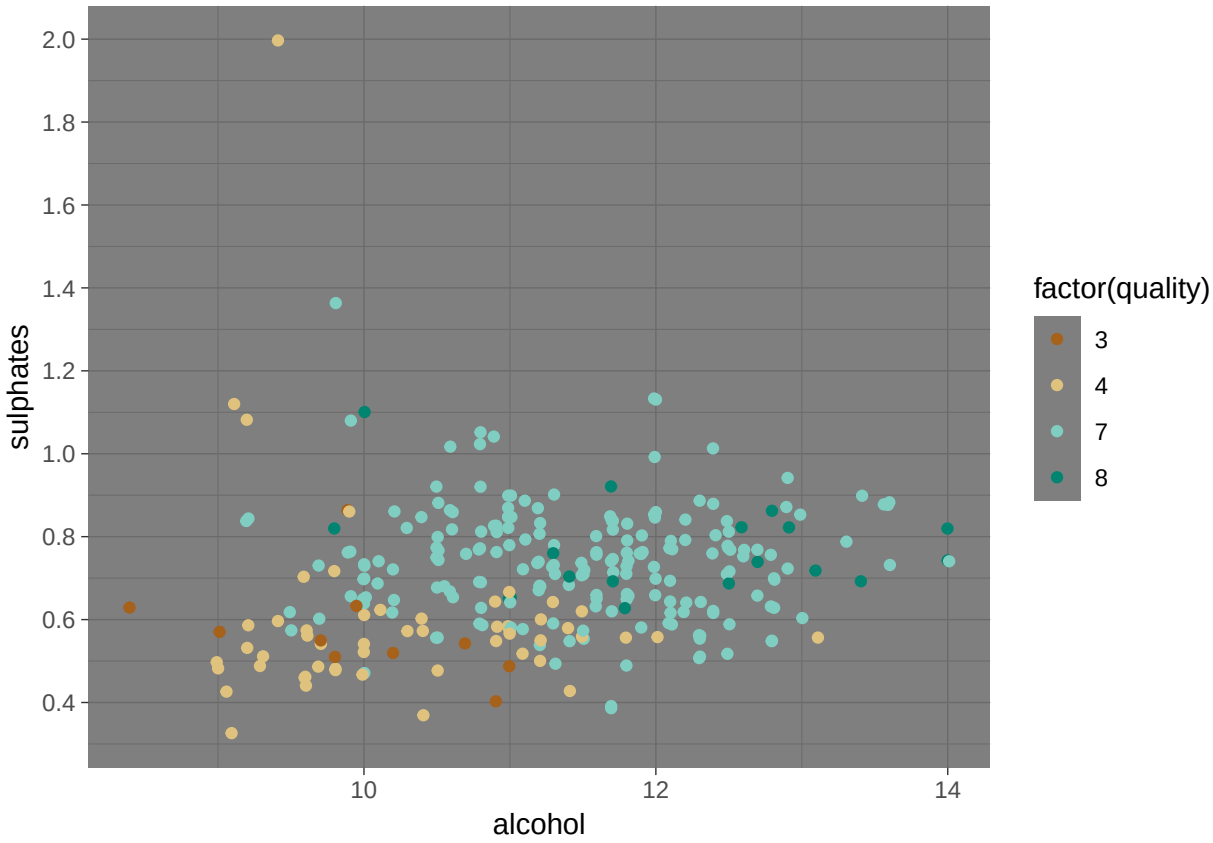


Nearly every wine has volatile acidity less than 0.8.And like we disscussed earlier the 7 and 8 quality wines all have volatile acidity of less than 0.6 .For wines with quality of 6 and 5(rating =B) the volatile acidity is between 0.4 and 0.8. Some C rating wine have a volatile acidity value of more than 0.8

Most A rating wines (quality 7-8) have citric acid value of 0.25 to 0.75 ,While the B rating wines(quality 6-5) have citric acid value below 0.50

Alcohol and Sulphates





I was expecting this plot to be like the citric acid and alcohol one,But i was amazed to see that nearly all wine lie below 1.0 sulphates level except quality 5.

Because of overploting i have removed the the rating B from next plot to visualize the trends more deeply.We can see rating A wines mostly have sulphates values between 0.5 and 1 and the best wine quality 8 have sulphate values between 0.6 and 1.Alcohol has the same values as seen before

Most C rating wines have sulphate value below 0.6

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?

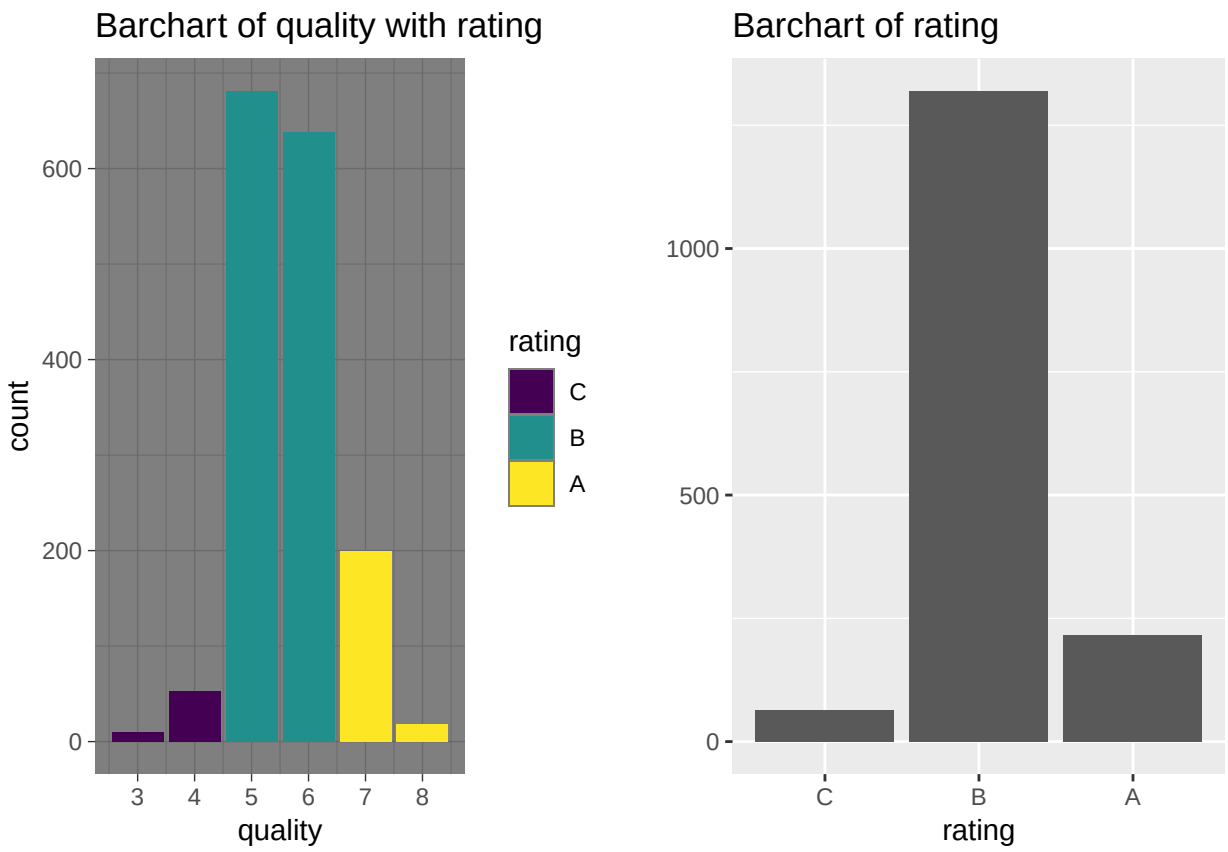
I am quite optimistic that higher citric.acid and lower volatile.acidity contribute towards better wines. Also, better wines tend to have higher alcohol content

Were there any interesting or surprising interactions between features?

Yes, I was little surpised with suplhates and alcohol graphs . Sulphates had a better correlation with quality than citric acid still the distribution was not that distinct between the different quality wines.Further nearly all wines had a sulphuate contant of less than One . Irrespective of alcohol content.Even though suplhate is a biproduct of fermantation just like alcohol . So longer fermantation should have produced more sulphates and alcohol which is clearly not the case.

Final Plots and Summary

Plot One



Description One

The plot is from univarent section.The first plot which introduced the idea of this analysis.I believe it is an important plot as it was the one which was one of the main driving idea behind his whole analysis into quality of red wine

Lets break this distribution for one final time an look at the percentage

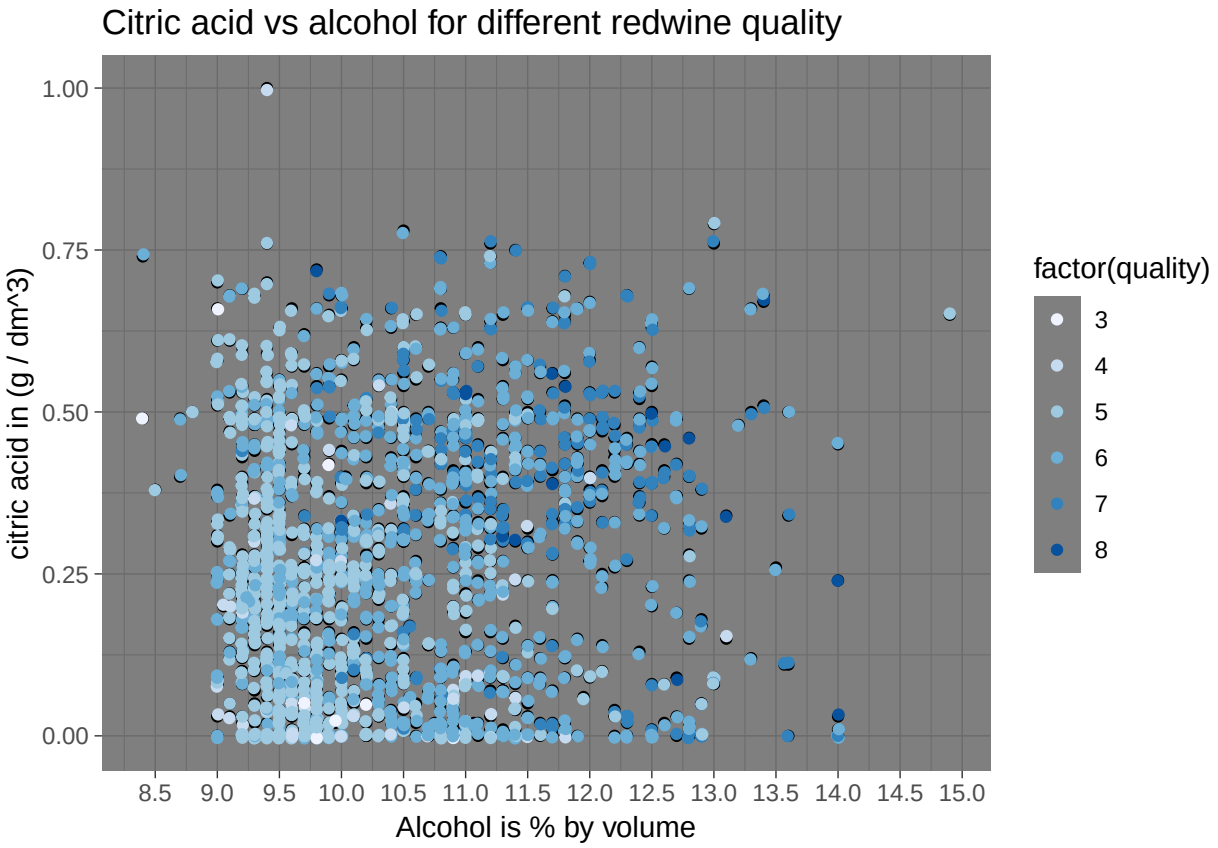
By rating

A rated=13.5% wines

B rated=82.48% wines

C rated=3.9% wines

Plot Two

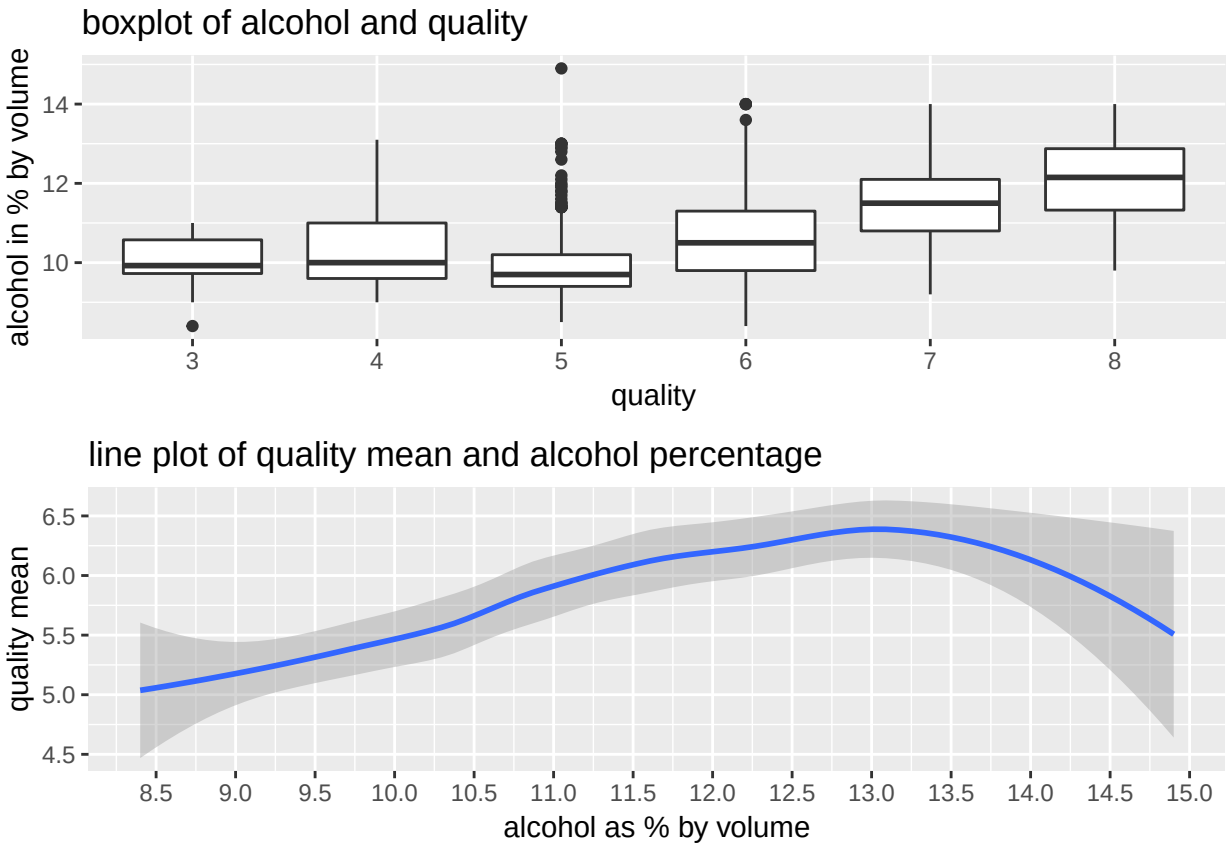


Description Two

The plot is from multivarent section discussing the relation of alcohol and citric acid concentration on quality

Values above 11% by volume of alcohol and 0.25 of citric acid result in rating A wines which are quality 7 and above . Most rating B wines lave a lower alcohol concentrations below 11% while the rating C specially quality 4 wines also have low citric acid with some exceptions

Plot Three



Description Three

Plots taken from bivarate analysis section discussing about effect of alcohol percentage on quality

The above boxplots show a peroidic rise in level of alcohol.One of the interesting findings i came across arised from this quality mean and alcohol% graph. An interesting trend of a exponetial decrement above 13% alcohol gave way to futher analysis which show that the alcohol which had a positive correlation of 0.47 becomes negatively correlated above 13%

This behaviour led to two seperate findings about alcohol in relation with quality

Below 13 % - Alcohol has a positive correlation of 0.47 the heighest correlation on positive side with quality of red wine.The boxplots also show an increase in median alcohol % with increase in quality

above 13% - Trends seem to be different as we epproach above 13% mark.The correlation diminish rather become negative

---

Reflection

The red wine dataset contains 1,599 observation with 13 variables of which 11 were on the chemical properties and 1 was for numbering and one for quality. I was interested in the correlation between the features and wine quality.I also created a new variable rating which had values from A for 8-7 ,B for 6-5 and C for 4-3

Unlike the diamond case study,the wine quality is more complex. Most of the data visualization in this project was done on the 4 features that have the highest correlation coefficient: alcohol(0.476), volatile acidity(-0.391),

sulphates(0.251),citric acid(0.226).

Most helping was the correlation matrix which i was able to make with ggcorr function.I helped to figure out the correlations between other variables too.Although I was unable to do as much intervariable analysis because so many of the variables were numeric instead of factoral.In the latter anaylis is grouped quality as a factor for bringing out important observations

I noticed that the dataset is highly unbalanced. It has many data points for medium quality wine (5, 6). However, for low quality (3,4) and high quality (7, 8) wine, it has fewer data points.The data set for medium quality wine covered above 80% of the distribution

If the data set has more records on both the low end and high end, the quality of analysis can be improved. We can be more certain about whether there is a significant correlation between a chemical component and the wine quality.

I would also like to see the data about temperaturem region of wine making, Quality of grapes used for making wine. These thinks will help research to go further.