Project3.1

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Wine Quality Reds:-

In these project we are discussing about Red wine contains 13 variables of 1599 observations, first changing the directory to our file location as follows.

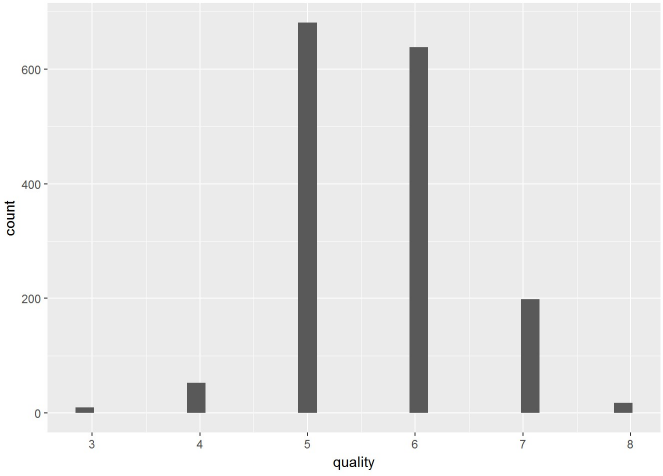
Installing and loading the files required to complete the project.

UNIVARIATE PLOTS:-

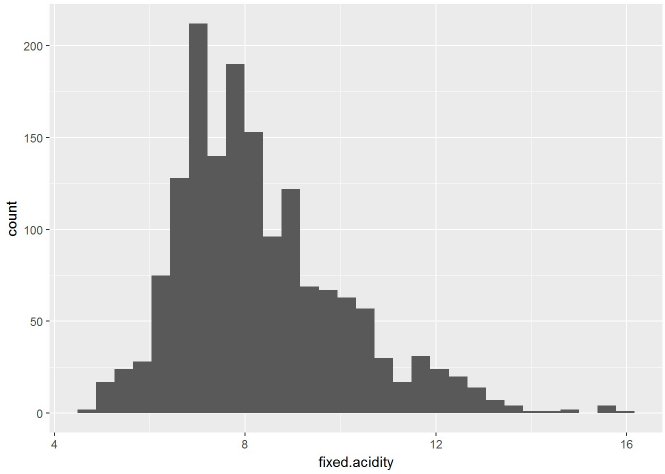
we are going to explore the data using each variable to get clear view about the variables and applying the statistical methods to find the relations between them. The summary is the overview of the each variable.

```
##        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
```

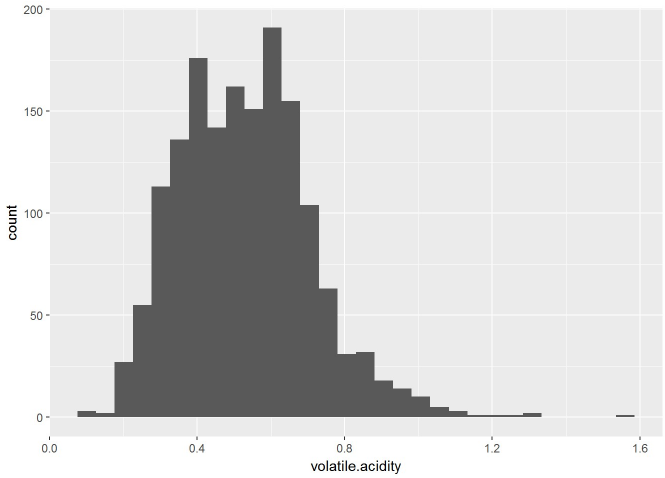
```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num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



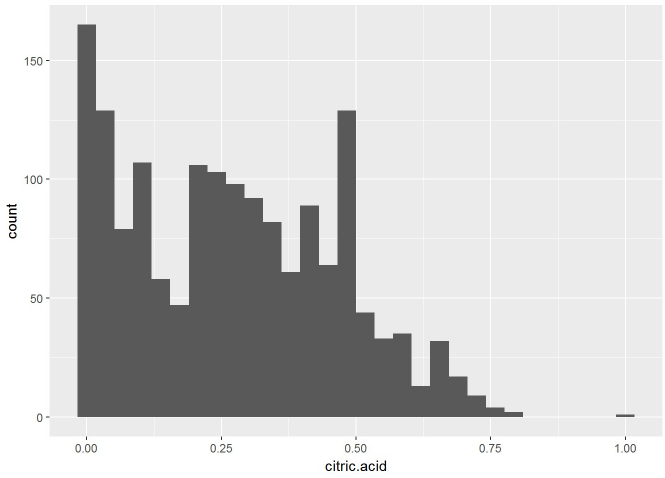
In the above plot it is showing that in given observation there are very less amount of low and high quality wines are given.Now lets see the another plot



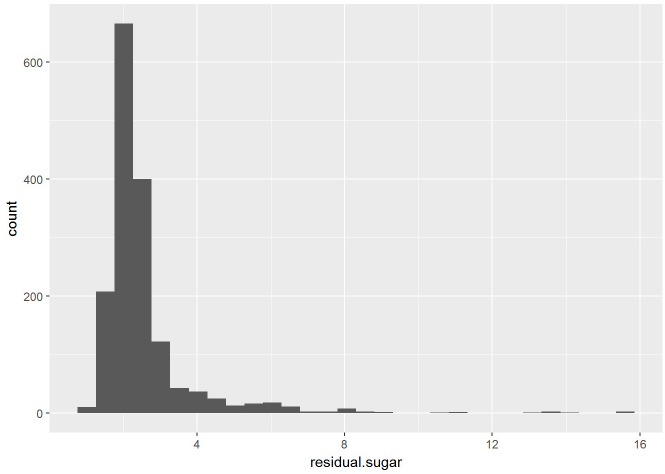
Approximately by seeing we can say that the mean is to be near by 10 because of the outliers and the median is slightly shifts to the 8.



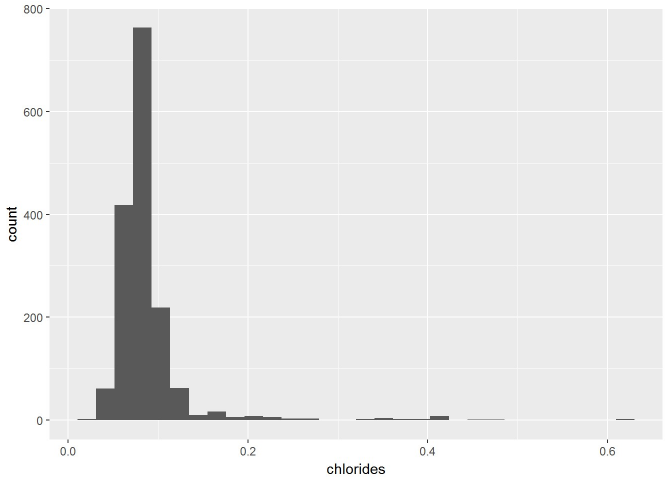
comparing to the fixed acidity the volatile acidity concentration is very less and appers to be bimodal with high range of outliers



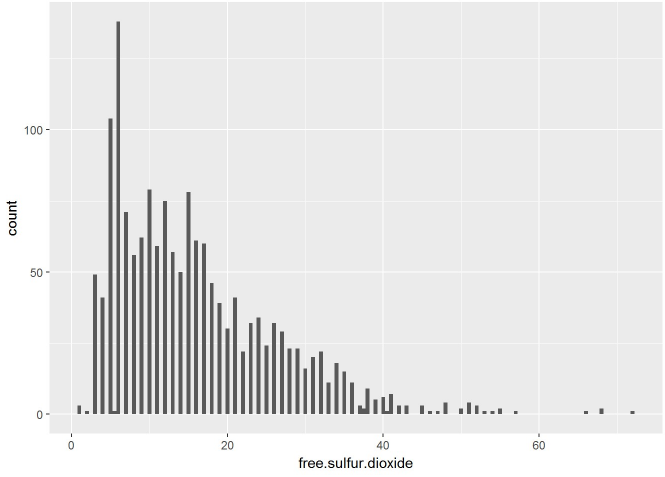
It is seem to be very intreasting plot and the concentration level is very less compare to the other acidic variables.



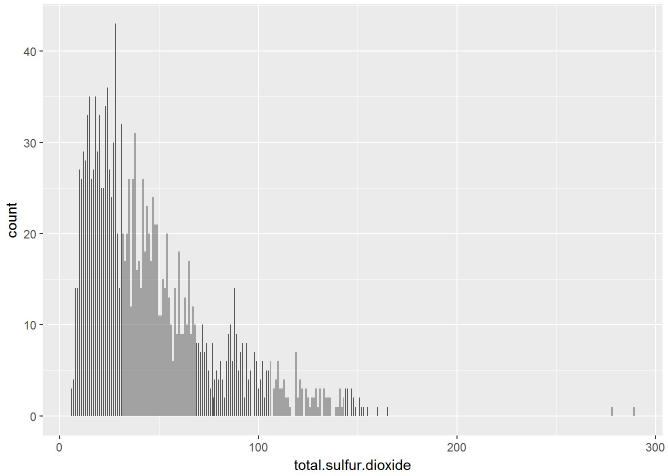
The median is nearby 2 which mostly used for the high quality wines.



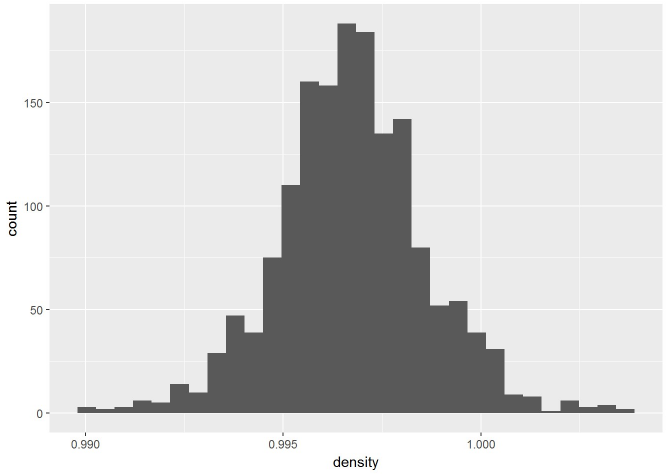
It is very interesting that the chlorides have similar distribuation to residual sugar level in in the wine but the meadian and mean changes by values.



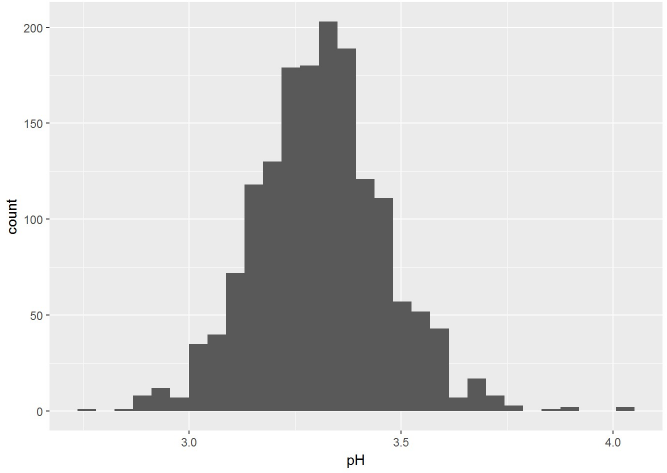
Here the meadian may less then 10 but the mean will shifts to near to 30 because of the outliers which are seen to be after 60 too.



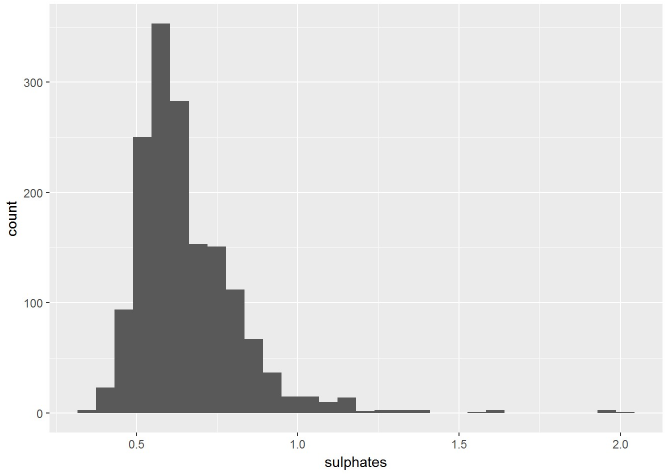
The mean is 46 and median is 38 and which is similar to the free sulfur dioxide.



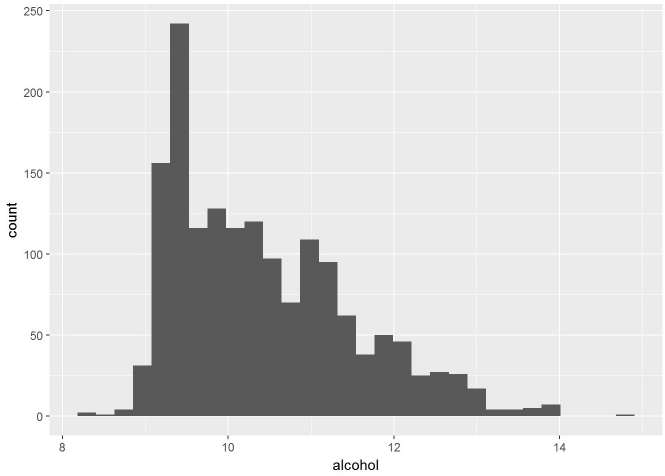
Density is seem to be noramlly distributed.



The Density and the PH value are having similar normal distributions.



The above distribaution is similar to the sugar varibale in the data.



It is having rapid growth in the distribution and here some the variables are having similar distribution.

UNIVARIATE ANALYSIS: -

What is the th structure of yor dataset?

The data consists of 1599 observations with 12 different variables, Here the quality of a wine is denoted by the scale 0-10 were 0 is bad and 10 is good with there alcohol % by volume and the pH is a logarithemic value, density is calculated with (g/cm^3) and all other variables are with (mg/dm^3).

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

The volatile acidity,sulphates,citric acidity,alcohol,chlorides are the interesting variables and the distributions between the variables are different with the quality and quality is one of the important variable in the data set.

Did you create any new variable fro existing variablesin the datset?

The given group of wines are divided in t three groups bad,better and good according to there quality.

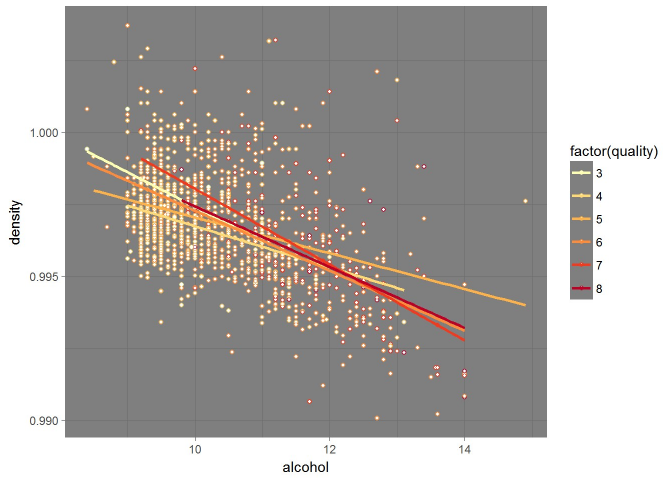
Of the features you investigated, were there any unusauldistributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data?is so, why did you do this?

Among all those variables citric acid has a unusual distribution.Which has the values of wines are 1 mg/dm^3.

BIVARIATE PLOTS SECTION: -

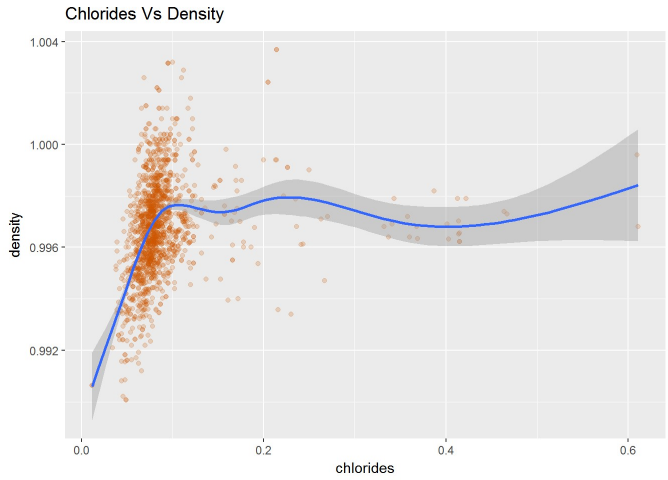
CORRELATION BETWEEN VARIABLES SHOWN IN THE PLOTS

Exploring the factors with density of wine




```
## $title
## [1] "Density vs Alcohol"
##
## $subtitle
## NULL
##
## attr(,"class")
## [1] "labels"
```

It is showing that as alcohol % increases the density level decreases and the good wine has less density with high alcohol % for some wine.

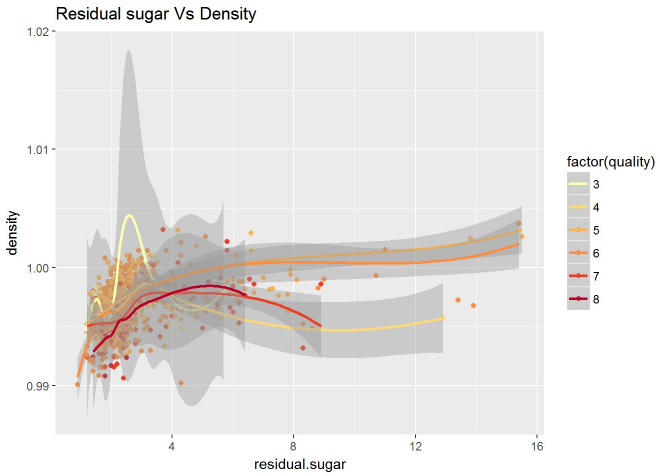



```
##
##  Pearson's product-moment correlation
##
## data:  chlorides and density
## t = 8.1842, df = 1597, p-value = 5.541e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1531171 0.2472220
## sample estimates:
##       cor
## 0.2006323
```

```
## [1] "summary"
```

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

The chlorides for the wine has very less with a density below 1 and above 0.992

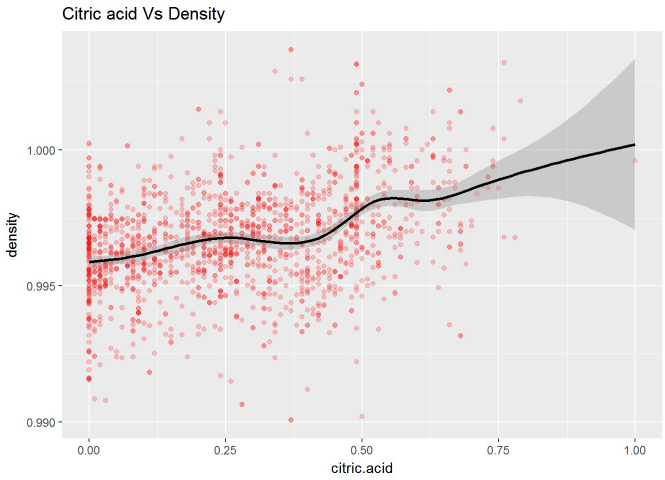



```
##
##  Pearson's product-moment correlation
##
## data:  residual.sugar and density
## t = 15.189, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3116908 0.3973835
## sample estimates:
##       cor
## 0.3552834
```

```
## [1] "summarty of residual sugar"
```

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

It showing that the highest residual suagr is 15.5 but there large varience between 3rd quardrant and maximum.




```
##
##  Pearson's product-moment correlation
##
## data:  citric.acid and density
## t = 15.665, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3216809 0.4066925
## sample estimates:
##       cor
## 0.3649472
```

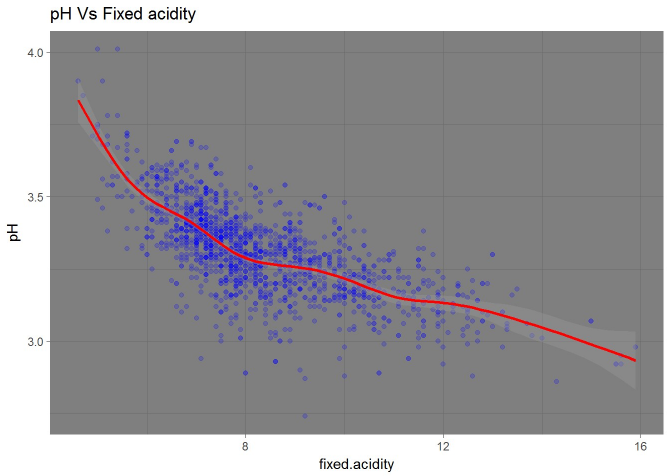
```
## [1] "summary of citric acid comparing with density"
```

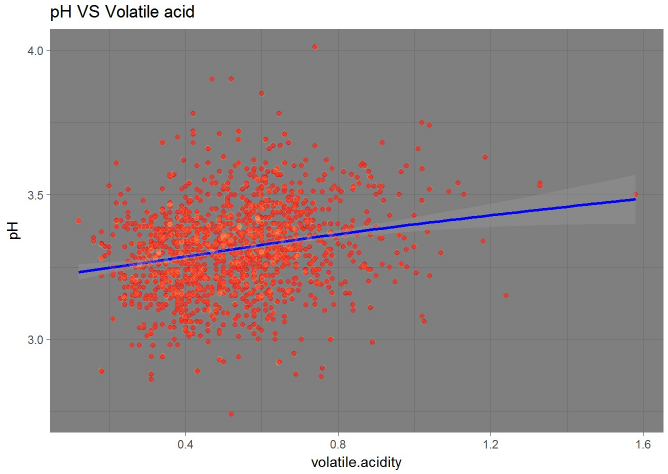
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

we can see a very intresting plot here that the citric acid is citric acid is slightly increased with the density.

Exploration of acid variables

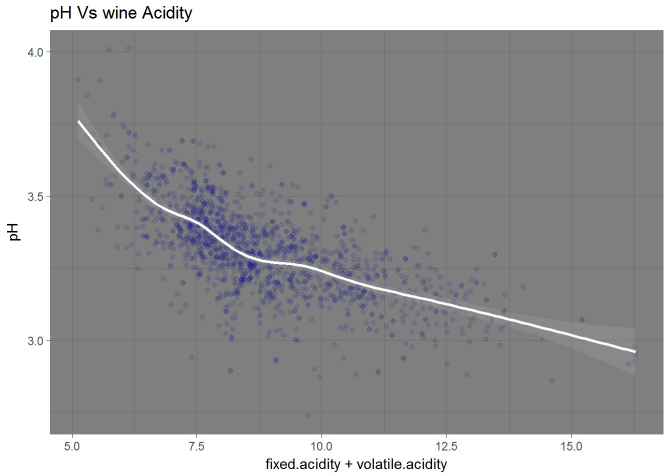
The pH is a value shows the acidity level of wine.





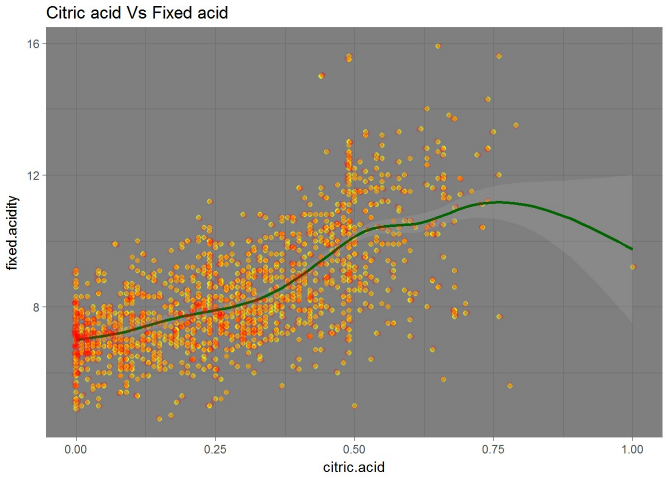

```
##
##  Pearson's product-moment correlation
##
## data:  volatile.acidity and pH
## t = 9.659, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1880823 0.2807254
## sample estimates:
##       cor
## 0.2349373
```

volatile acidity is positively recating with the pH.



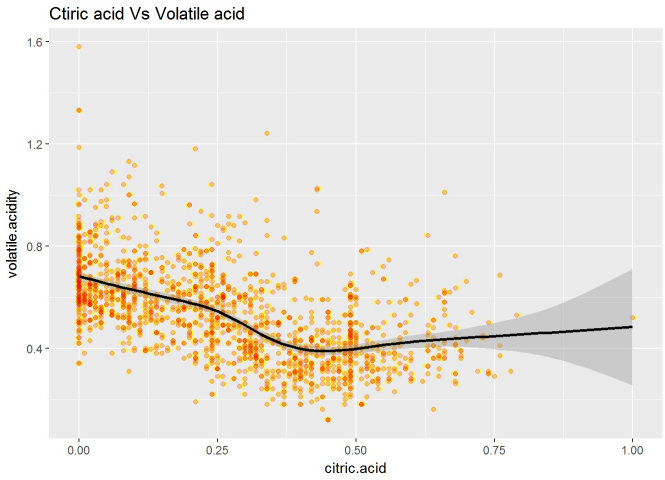
```
##
##  Pearson's product-moment correlation
##
## data:  fixed.acidity + volatile.acidity and pH
## t = -36.376, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6990929 -0.6454169
## sample estimates:
##        cor
## -0.6731405
```

Comparing to volatile acidity the sum of both acids showing us negative corelation with pH value.



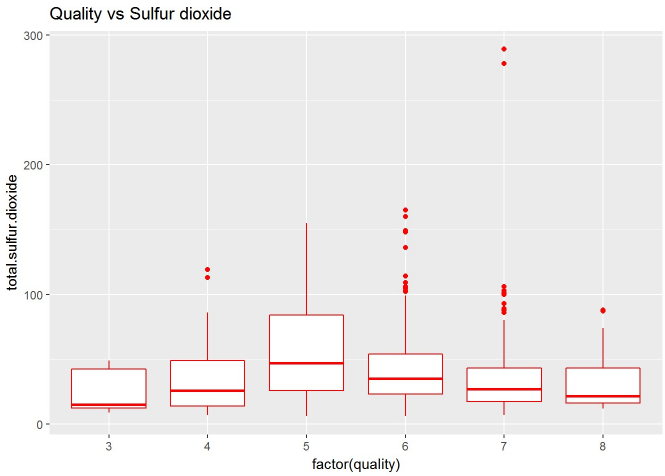

```
##
##  Pearson's product-moment correlation
##
## data:  citric.acid and fixed.acidity
## t = 36.234, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6438839 0.6977493
## sample estimates:
##       cor
## 0.6717034
```

we can see a rapid growth in a plot until to some end but a outlier on citric acid made it to drop a bit down.

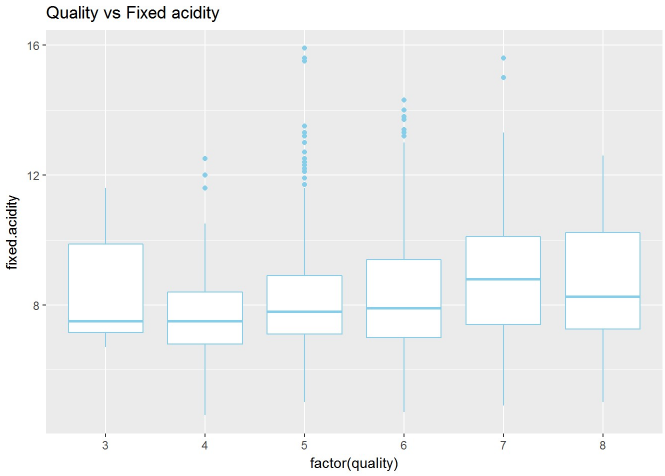


By comparing the citic acid with fixed and volatile acidities it can be konwn that volatile acid is having rapid decrease but the fixed acid has growth. so, it says that the taoal acidic nature of wine is trying to keep in equilbrium state.

Exploring the the quality of wine with other variables

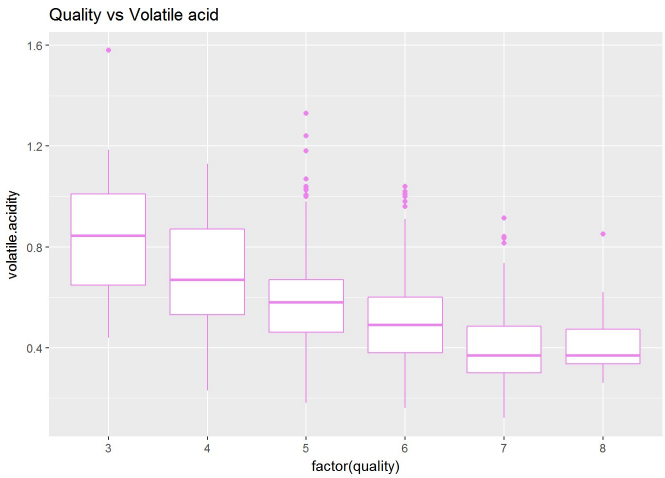


The sulfur dioxide is get known by the smell of wine here in the box plot it is clearly showing that the quality of wines with 5 rating has more smell and taste then others.



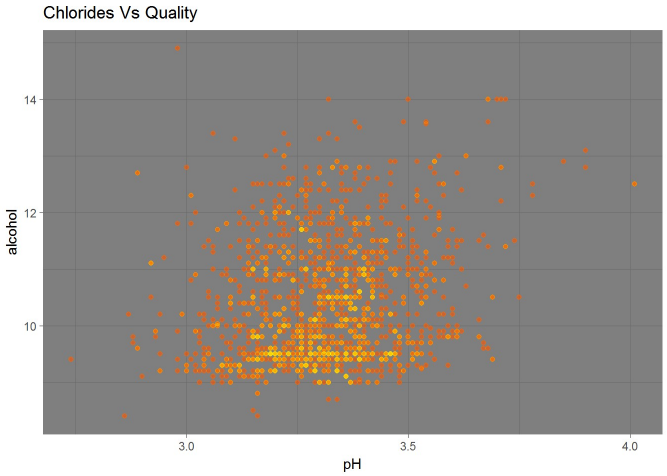

```
##
##  Pearson's product-moment correlation
##
## data:  fixed.acidity and quality
## t = 4.996, df = 1597, p-value = 6.496e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07548957 0.17202667
## sample estimates:
##       cor
## 0.1240516
```

The fixed acidity in the wine is used to do not ecvaporate the wine easily, so, the countries were the high temperatures they try to keep the fixed acidity more and it is almost same for all the above quality wines given above there is a little varience between them.

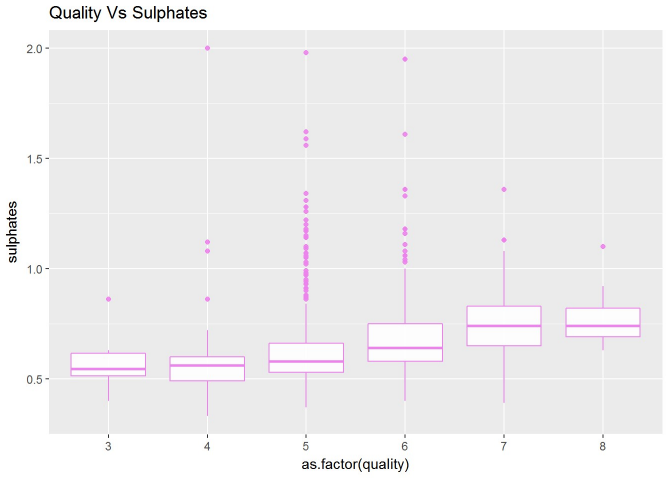


```
##
##  Pearson's product-moment correlation
##
## data:  volatile.acidity and quality
## t = -16.954, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4313210 -0.3482032
## sample estimates:
##        cor
## -0.3905578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

As the volatile acid nature increase the taste of wine changes as venigar so, for as the quality increases the volatile acid nature decreased.

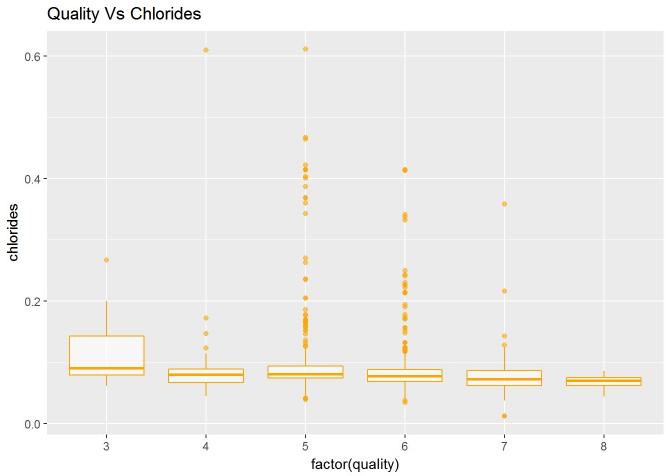



```
##
##  Pearson's product-moment correlation
##
## data:  chlorides and quality
## t = -5.1948, df = 1597, p-value = 2.313e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.17681041 -0.08039344
## sample estimates:
##        cor
## -0.1289066
```

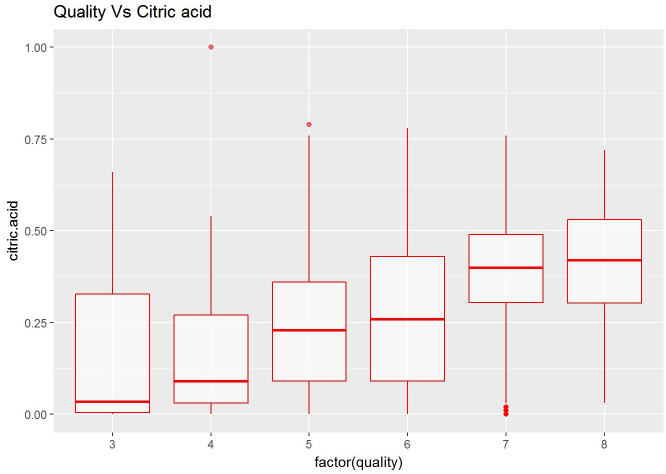



```
##
##  Pearson's product-moment correlation
##
## data:  sulphates and quality
## t = 10.38, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2049011 0.2967610
## sample estimates:
##       cor
## 0.2513971
```

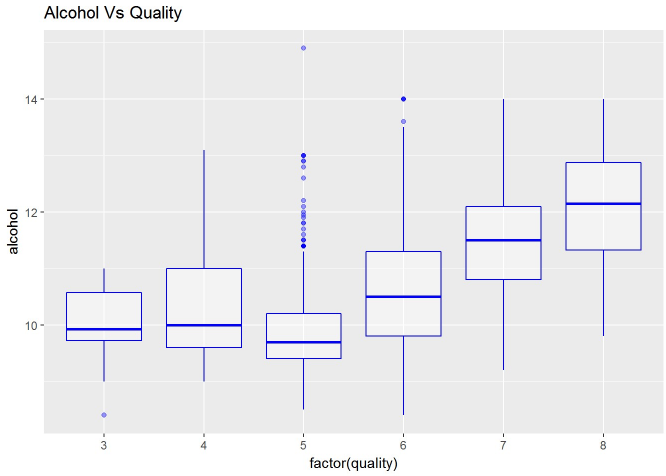
The sulpahates are used to maintain a s02 gas level in the wine and its act like a antioxidant.



The chlorides add salt nature to the wines so, for the good quality wines have less quantity of chlorides.

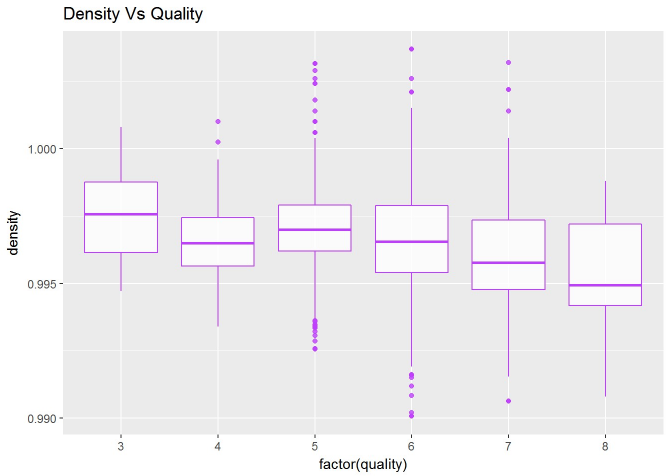


In the real life the citric acid is used to keep the wine fresh so, for the good quality of wines have high rate of citric acid but in small quantites.




```
##
##  Pearson's product-moment correlation
##
## data:  alcohol and 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
```

Here in the given data the alcohol is goven in % and as usually the good quality of wine having higher level of alcohol.




```
##
##  Pearson's product-moment correlation
##
## data:  density and quality
## t = -7.0997, df = 1597, p-value = 1.875e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2220365 -0.1269870
## sample estimates:
##        cor
## -0.1749192
```

Density is the water quantity that is maintained depending on the alcohol and the sugar level of the wine most of the time they try to balance the things here.

BIVARIATE ANALYSIS:-

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

By the above investigation there are some relations which can be said easily by above plots are sulphates,citric acid,alcohol which are directly proportional to the quality factor and inversly proportional to the chlorides and volatile acid.It had good scatter plots with different variables and there correlation coefficients.

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

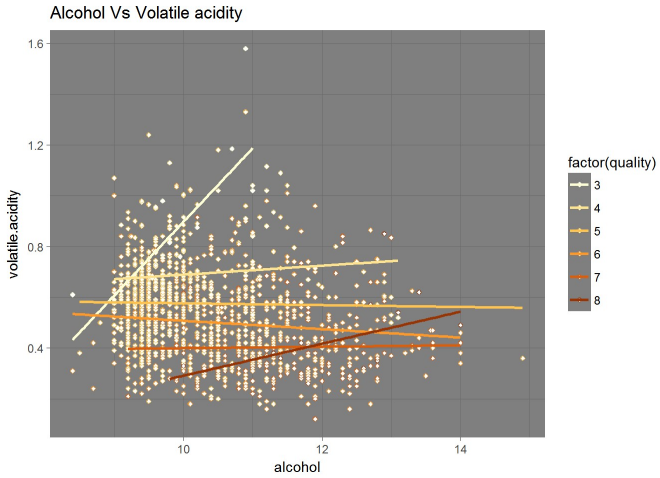
Yes, There is a strong correalation between pH and the acidity of wine which come from fixed and volatile acidity, other one is the negative correlation with the alcohol level in the wine.

What was the strongest relationship you found?

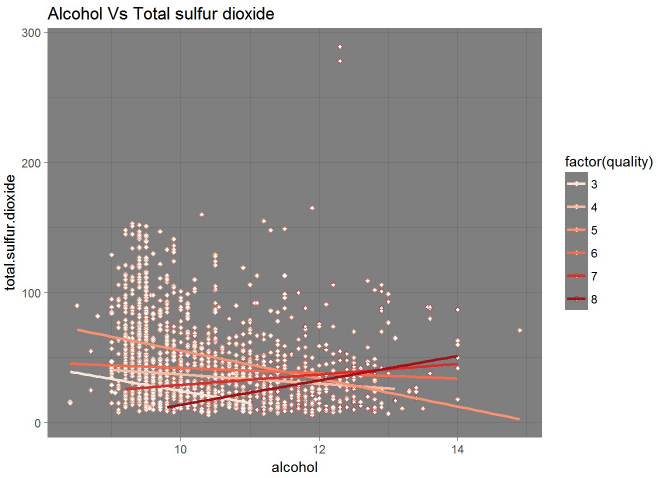
Yes there is a strongest relation between alcohol and the quality of wine which is directly proportional and also density and alcohol are negativley correlated.

MUTIVARIATE ANALYSIS :-

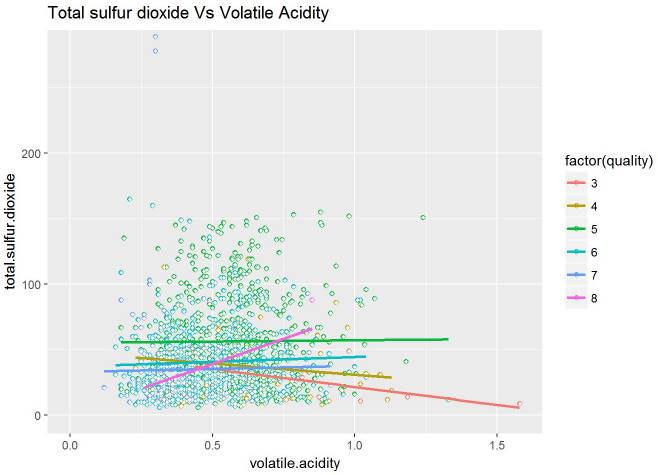
By seeing above analysis the variables alcohol,citric acid,volatile acid,chlorides and total sufur dioxide has great impact on deciding the good quality wine.



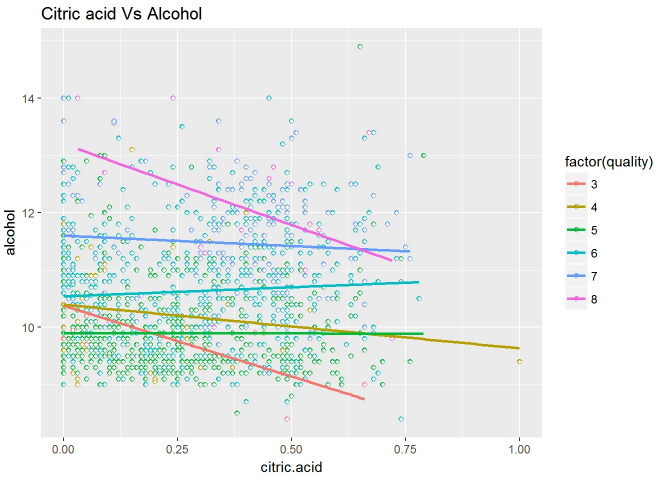
Here there is clear view about the volatile acidity and alcohol showing that the high qualtiy wines having little bit volatile acid with high alcohol level the brown line indicates positive ralation between them.



There is a simialr observation between alcohol and total sulfur dioxide which means with high alcohol and little bit of total sufu dioxide with possitive coefficient the red line showing the relation.



The above plot shows that the positive relation between total sulfur dioxide and vlatile acid for a good quality of a wine.



It is very clear that the alcohol and citric acid has negative relation here.Let's train the line model to find a good quality wine.

```
##
## Call:
## lm(formula = quality ~ alcohol + volatile.acidity + total.sulfur.dioxide +
##     citric.acid, data = wineQualityReds)
##
## Coefficients:
##          (Intercept)               alcohol      volatile.acidity
##              3.247836             0.301524            -1.306618
## total.sulfur.dioxide           citric.acid
##             -0.002014             0.105683
```

It is clear that the alcohol and the citric acid has the positive relation and the total sulfur dioxide and volatile acid has negative ralation.

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?

Here in the mutivariate plots there are few variables mostly concentrated making the quality as a factor because to find the good quality of wine.

The features made clear view that the volatile acid,total sufur dioxides are positively reacted they are little bit high for the good quality of wine comapre to others and the alcohol level for good wine is also a positive relation but in the last there is negative relation with the citric acid were there is small quatities are used for the good quality of wine. **Were threre any interesting or surprising interactions between features?**

Yes, there is a surprising thing here that the positive realtion between the quality and citric acid in bivariate analysis but here there is a negative relation found between other variables for a good quality of wine, so, we can predict that they use small quantities of citric acid for good quality wine.

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

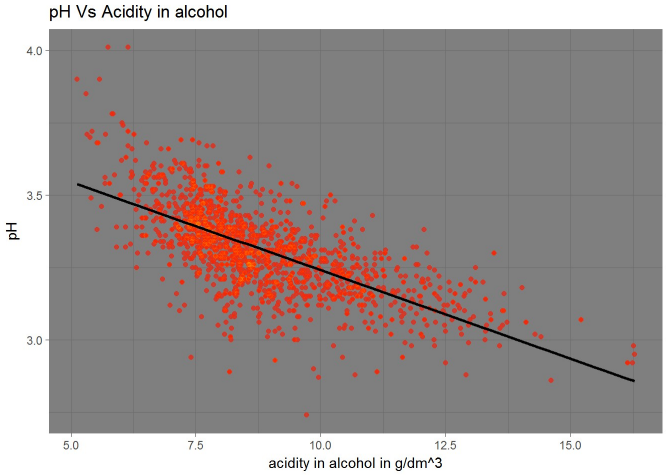
Yes, created a linear model between between different variables which shown that the aicidity levels like volatile acid and sulfur dioxide are negatively related it means they are used in less quantites and alcohol & citric acid are in positive relations for a good quality wine, alcohol level is medium-high and the citric acid in low-high quantities.

FINAL PLOTS :-

Here few main variables are taken from the analysis to examine the quality of wine.

Final plot1:-

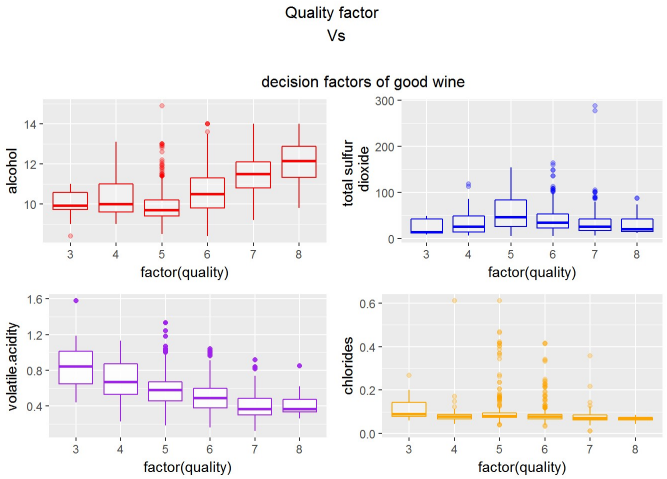
It is drawn between the acid level(sum of fixed and volatile acids) and pH level which used to calculate the acid quantity of wine.



Description:

In the above plot as the volume of alcohol increases the pH level is decresed it predict that the high volumed alcoholic wine has less pH level.so, it predicts that less pH and high alcohol levels wich makes wine good.

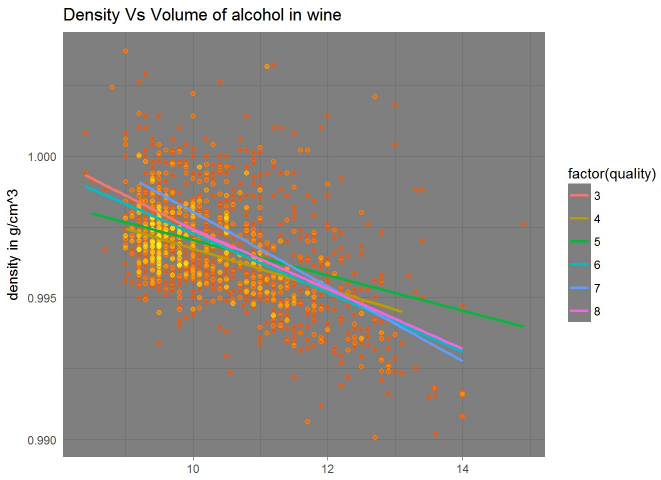
Final Plot2:-



Description:

The above shown boxplots are the directly and inversly poportional to the wine quality. The alcohol level in the wine is directly proportional to the wine which means the good quality wine has the high volume alcohol. The total sulfur dioxidewhich should be used in low contration levels which is used for the good taste and the smell puposes high concentration levels which spoils the taste of wine so, it is inversly proportional to the quality and also shown in above plot. The high amount of volatile acid leads to the unpleasent and vinegar taste so, it also used in low concentration levels which makes the wine good quality. chlorides are in the salty nature so, the high amount of chlorides makes the wine salty but for the better and good quality wines contain very less concentrations.

Final Plot3:-



Description:

Here it is showing a relation between density and the volume alcohol in the wine.The density is high for the low volume alcohol. In the last the good quality wine has less density with high alcoholic nature.

REFLECTION :-

The Quality Wine Reds contains the data of 13 different chemical variables wich are used in the 1599 observations which differnt properties. Basically when I started these project I surprised to see all the chemicals which are used to prepare wine and I only no that the good quality of wine has high concentraions of alcohol. And In my intial stages of project I felt difficulties to decide the variables which are used to find the good wine, after making a research of project using different anlysis like univariate,bi and multi varite plots I got a knowledge about the chemical properties and there reactions with the variables. I started plotting a scatter,box plots using different varibles in the data got conclusion about the main features of wine which discussed in the final plot section. And also learnt diffrent methods to find the correlation between the varibles and there purpose of use in the wine. In the last the good quality of wine consits high alcohol with low density and pH level which is noticed by the acidic nature of wine and also citric acid is used to keep the wine freshness in low levels.