# Observations of Red Wine Quality by Subham Biswas

# Load the Data  
setwd('C:/Users/abc/Downloads/Udacity\_R')  
red\_wine <-read.csv('wineQualityReds.csv',header= TRUE)

# removing the serial number column  
red\_wine$X <- NULL  
#summarising the data set  
summary(red\_wine)

## fixed.acidity volatile.acidity citric.acid residual.sugar   
## Min. : 4.60 Min. :0.1200 Min. :0.000 Min. : 0.900   
## 1st Qu.: 7.10 1st Qu.:0.3900 1st Qu.:0.090 1st Qu.: 1.900   
## Median : 7.90 Median :0.5200 Median :0.260 Median : 2.200   
## Mean : 8.32 Mean :0.5278 Mean :0.271 Mean : 2.539   
## 3rd Qu.: 9.20 3rd Qu.:0.6400 3rd Qu.:0.420 3rd Qu.: 2.600   
## Max. :15.90 Max. :1.5800 Max. :1.000 Max. :15.500   
## chlorides free.sulfur.dioxide total.sulfur.dioxide  
## Min. :0.01200 Min. : 1.00 Min. : 6.00   
## 1st Qu.:0.07000 1st Qu.: 7.00 1st Qu.: 22.00   
## Median :0.07900 Median :14.00 Median : 38.00   
## Mean :0.08747 Mean :15.87 Mean : 46.47   
## 3rd Qu.:0.09000 3rd Qu.:21.00 3rd Qu.: 62.00   
## Max. :0.61100 Max. :72.00 Max. :289.00   
## density pH sulphates alcohol   
## Min. :0.9901 Min. :2.740 Min. :0.3300 Min. : 8.40   
## 1st Qu.:0.9956 1st Qu.:3.210 1st Qu.:0.5500 1st Qu.: 9.50   
## Median :0.9968 Median :3.310 Median :0.6200 Median :10.20   
## Mean :0.9967 Mean :3.311 Mean :0.6581 Mean :10.42   
## 3rd Qu.:0.9978 3rd Qu.:3.400 3rd Qu.:0.7300 3rd Qu.:11.10   
## Max. :1.0037 Max. :4.010 Max. :2.0000 Max. :14.90   
## quality   
## Min. :3.000   
## 1st Qu.:5.000   
## Median :6.000   
## Mean :5.636   
## 3rd Qu.:6.000   
## Max. :8.000

# printing structure of the dataset  
str(red\_wine)

## 'data.frame': 1599 obs. of 12 variables:  
## $ 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 ...

#printing columns in the data set  
colnames(red\_wine)

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

#describing more  
describe(red\_wine)

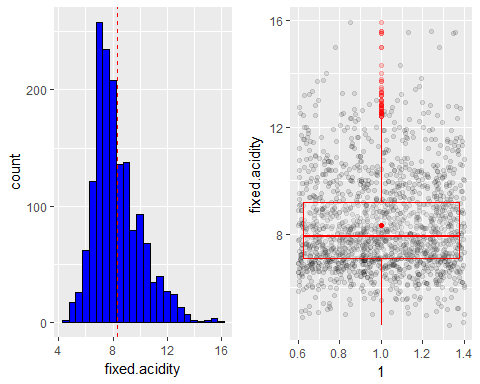
## vars n mean sd median trimmed mad min  
## fixed.acidity 1 1599 8.32 1.74 7.90 8.15 1.48 4.60  
## volatile.acidity 2 1599 0.53 0.18 0.52 0.52 0.18 0.12  
## citric.acid 3 1599 0.27 0.19 0.26 0.26 0.25 0.00  
## residual.sugar 4 1599 2.54 1.41 2.20 2.26 0.44 0.90  
## chlorides 5 1599 0.09 0.05 0.08 0.08 0.01 0.01  
## free.sulfur.dioxide 6 1599 15.87 10.46 14.00 14.58 10.38 1.00  
## total.sulfur.dioxide 7 1599 46.47 32.90 38.00 41.84 26.69 6.00  
## density 8 1599 1.00 0.00 1.00 1.00 0.00 0.99  
## pH 9 1599 3.31 0.15 3.31 3.31 0.15 2.74  
## sulphates 10 1599 0.66 0.17 0.62 0.64 0.12 0.33  
## alcohol 11 1599 10.42 1.07 10.20 10.31 1.04 8.40  
## quality 12 1599 5.64 0.81 6.00 5.59 1.48 3.00  
## max range skew kurtosis se  
## fixed.acidity 15.90 11.30 0.98 1.12 0.04  
## volatile.acidity 1.58 1.46 0.67 1.21 0.00  
## citric.acid 1.00 1.00 0.32 -0.79 0.00  
## residual.sugar 15.50 14.60 4.53 28.49 0.04  
## chlorides 0.61 0.60 5.67 41.53 0.00  
## free.sulfur.dioxide 72.00 71.00 1.25 2.01 0.26  
## total.sulfur.dioxide 289.00 283.00 1.51 3.79 0.82  
## density 1.00 0.01 0.07 0.92 0.00  
## pH 4.01 1.27 0.19 0.80 0.00  
## sulphates 2.00 1.67 2.42 11.66 0.00  
## alcohol 14.90 6.50 0.86 0.19 0.03  
## quality 8.00 5.00 0.22 0.29 0.02

Findings: 1.There are 1599 observations and 13 variables. 2. X is identification id for each observation. 3. Quality is the most important variable as it determines quality of individual sample or observation. 4.Mean Quality is 5.636 and Median is 6 , range is from 3 to 8.

# Univariate Plots Section

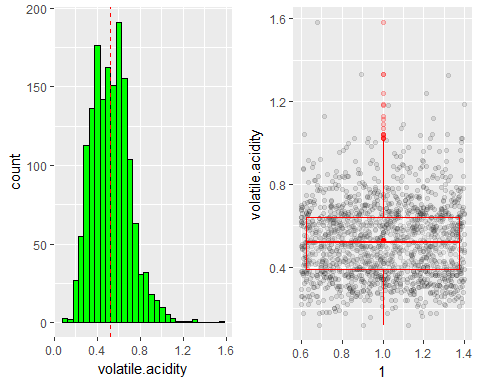
I have plotted the mean for all the variables

p1 <- ggplot(data = red\_wine, aes(x =fixed.acidity))+  
 geom\_histogram(binwidth = 1/2, color = 'black',fill = 'blue')+  
 geom\_vline(aes(xintercept=mean(fixed.acidity)),  
 color="red", linetype=2)  
p2 <- ggplot(red\_wine, aes( x = 1, y = fixed.acidity ) ) +   
 geom\_jitter(alpha = 0.1 ) +  
 geom\_boxplot(alpha = 0.2, color = 'red' )+  
 stat\_summary(fun.y = "mean",   
 geom = "point",   
 color = "red" )  
grid.arrange(p1 , p2, ncol = 2)



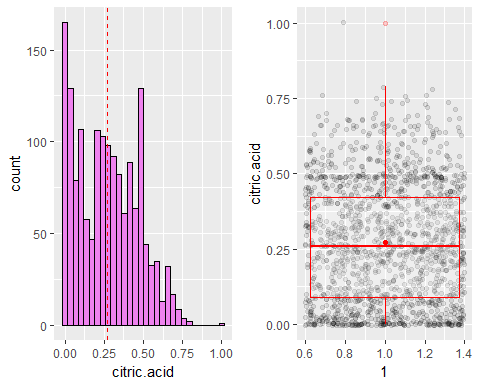
We can see from the plot ,it is almost normal distribution with majority falling between in range 6 to 10 , mean of around 8.5

p1 <- ggplot(data = red\_wine, aes(x =volatile.acidity))+  
 geom\_histogram( color = 'black',fill = 'green')+  
 geom\_vline(aes(xintercept=mean(volatile.acidity)),  
 color="red", linetype=2)  
p2 <- ggplot(red\_wine, aes( x = 1, y = volatile.acidity ) ) +   
 geom\_jitter(alpha = 0.1 ) +  
 geom\_boxplot(alpha = 0.2, color = 'red' )+  
 stat\_summary(fun.y = "mean",   
 geom = "point",   
 color = "red" )  
grid.arrange(p1 , p2, ncol = 2)

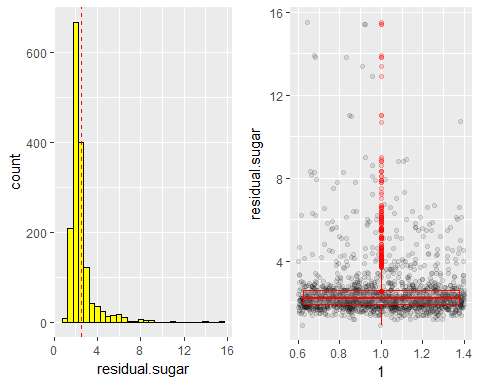


We can see from the plot ,it is almost normal distribution with majority falling between in range 0.2 to 0.8 , mean around 0.5

p1 <- ggplot(data = red\_wine, aes(x =citric.acid))+  
 geom\_histogram( color = 'black',fill = 'violet')+  
 geom\_vline(aes(xintercept=mean(citric.acid)),  
 color="red", linetype=2)  
p2 <- ggplot(red\_wine, aes( x = 1, y = citric.acid ) ) +   
 geom\_jitter(alpha = 0.1 ) +  
 geom\_boxplot(alpha = 0.2, color = 'red' )+  
 stat\_summary(fun.y = "mean",   
 geom = "point",   
 color = "red" )  
grid.arrange(p1 , p2, ncol = 2)

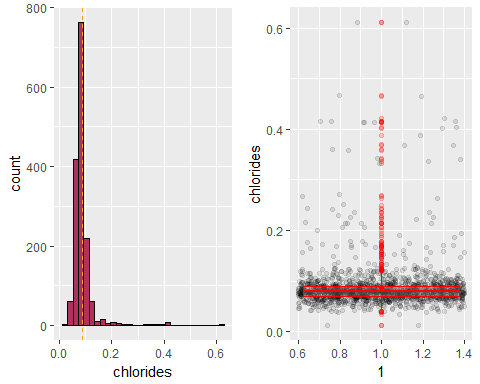


p1 <- ggplot(data = red\_wine, aes(x =residual.sugar))+  
 geom\_histogram(binwidth = 0.5, color = 'black', fill = 'yellow')+  
 geom\_vline(aes(xintercept=mean(residual.sugar)),  
 color="red", linetype=2)  
p2 <- ggplot(red\_wine, aes( x = 1, y = residual.sugar ) ) +   
 geom\_jitter(alpha = 0.1 ) +  
 geom\_boxplot(alpha = 0.2, color = 'red' )+  
 stat\_summary(fun.y = "mean",   
 geom = "point",   
 color = "red" )  
grid.arrange(p1 , p2, ncol = 2)

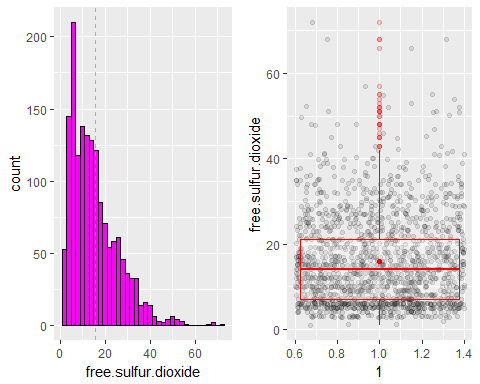


We can see from the plot ,it is almost normal distribution with majority falling between in range 1 to 3

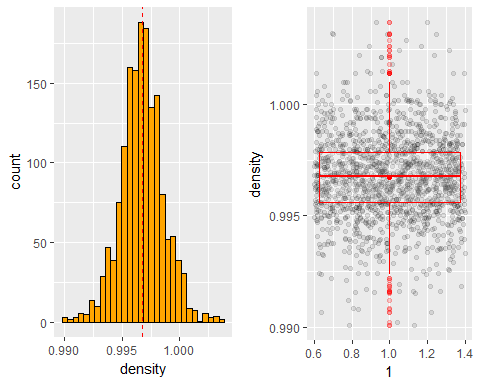
p1 <- ggplot(data = red\_wine, aes(x =chlorides))+  
 geom\_histogram( color = 'black',fill = 'maroon')+  
 geom\_vline(aes(xintercept=mean(chlorides)),  
 color="orange", linetype=2)  
p2 <- ggplot(red\_wine, aes( x = 1, y = chlorides ) ) +   
 geom\_jitter(alpha = 0.1 ) +  
 geom\_boxplot(alpha = 0.2, color = 'red' )+  
 stat\_summary(fun.y = "mean",   
 geom = "point",   
 color = "red" )  
grid.arrange(p1 , p2, ncol = 2)



p1 <- ggplot(data = red\_wine, aes(x =free.sulfur.dioxide))+  
 geom\_histogram(binwidth = 2, color = 'black',fill='magenta')+  
 geom\_vline(aes(xintercept=mean(free.sulfur.dioxide)),  
 color="orange", linetype=2)  
p2 <- ggplot(red\_wine, aes( x = 1, y = free.sulfur.dioxide ) ) +   
 geom\_jitter(alpha = 0.1 ) +  
 geom\_boxplot(alpha = 0.2, color = 'red' )+  
 stat\_summary(fun.y = "mean",   
 geom = "point",   
 color = "red" )  
grid.arrange(p1 , p2, ncol = 2)

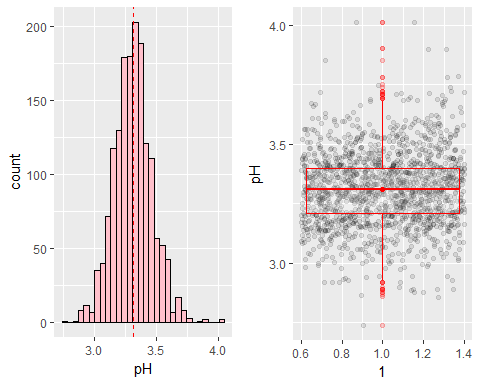


p1<- ggplot(data = red\_wine, aes(x =density))+  
 geom\_histogram(color = 'black',fill ='orange')+  
 geom\_vline(aes(xintercept=mean(density)),  
 color="red", linetype=2)  
p2 <- ggplot(red\_wine, aes( x = 1, y = density ) ) +   
 geom\_jitter(alpha = 0.1 ) +  
 geom\_boxplot(alpha = 0.2, color = 'red' )+  
 stat\_summary(fun.y = "mean",   
 geom = "point",   
 color = "red" )  
grid.arrange(p1 , p2, ncol = 2)

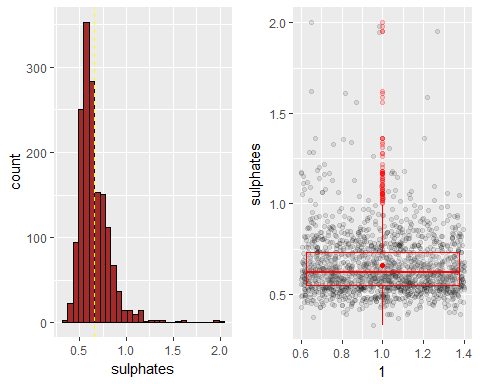


We can see from the plot ,it is almost normal distribution with majority falling between in range 0.995 to 1.0

p1 <- ggplot(data = red\_wine, aes(x =pH))+  
 geom\_histogram( color = 'black',fill = 'pink')+  
 geom\_vline(aes(xintercept=mean(pH)),  
 color="red", linetype=2)  
p2 <- ggplot(red\_wine, aes( x = 1, y = pH ) ) +   
 geom\_jitter(alpha = 0.1 ) +  
 geom\_boxplot(alpha = 0.2, color = 'red' )+  
 stat\_summary(fun.y = "mean",   
 geom = "point",   
 color = "red" )  
grid.arrange(p1 , p2, ncol = 2)

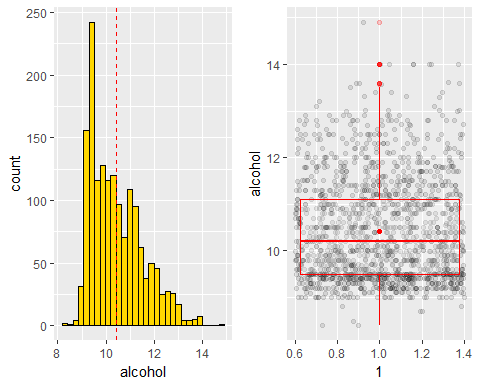
 We can see from the plot ,it is almost normal distribution with majority falling between in range 3.0 to 3.7

p1 <- ggplot(data = red\_wine, aes(x =sulphates))+  
 geom\_histogram( color = 'black',fill ='brown')+  
 geom\_vline(aes(xintercept=mean(sulphates)),  
 color="yellow", linetype=2)  
p2 <- ggplot(red\_wine, aes( x = 1, y = sulphates ) ) +   
 geom\_jitter(alpha = 0.1 ) +  
 geom\_boxplot(alpha = 0.2, color = 'red' )+  
 stat\_summary(fun.y = "mean",   
 geom = "point",   
 color = "red" )  
grid.arrange(p1 , p2, ncol = 2)

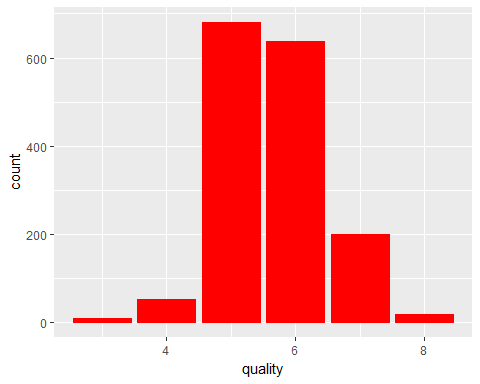


We can see from the plot ,it is almost normal distribution with majority falling between in range 0.4 to 0.9

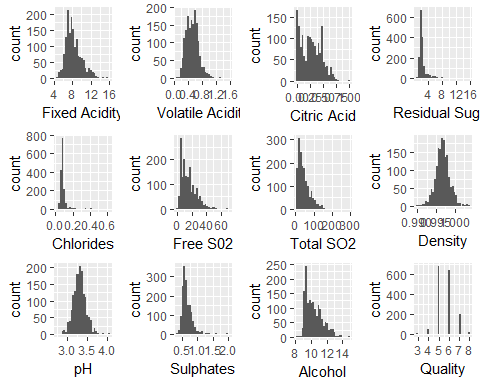
p1 <- ggplot(data = red\_wine, aes(x =alcohol))+  
 geom\_histogram( color = 'black',fill ='gold')+  
 geom\_vline(aes(xintercept=mean(alcohol)),  
 color="red", linetype=2)  
p2 <- ggplot(red\_wine, aes( x = 1, y = alcohol ) ) +   
 geom\_jitter(alpha = 0.1 ) +  
 geom\_boxplot(alpha = 0.2, color = 'red' )+  
 stat\_summary(fun.y = "mean",   
 geom = "point",   
 color = "red" )  
grid.arrange(p1 , p2, ncol = 2)



ggplot(data = red\_wine, aes(x =quality))+  
geom\_bar( color = 'red', fill = 'red')

 We can see from the plot ,the samples are mostly of quality 5 and 6 and least of 3 and 8.That means the bad and good quality wines are very less as compared to average quality wine

#Plotting all the graphs togetehr in a grid  
grid.arrange(qplot(red\_wine$fixed.acidity,xlab ='Fixed Acidity'),  
 qplot(red\_wine$volatile.acidity, xlab = 'Volatile Acidity'),  
 qplot(red\_wine$citric.acid, xlab = 'Citric Acid'),  
 qplot(red\_wine$residual.sugar, xlab = 'Residual Sugar'),  
 qplot(red\_wine$chlorides, xlab = 'Chlorides'),  
 qplot(red\_wine$free.sulfur.dioxide, xlab = 'Free S02'),  
 qplot(red\_wine$total.sulfur.dioxide, xlab = 'Total SO2'),  
 qplot(red\_wine$density, xlab = 'Density'),  
 qplot(red\_wine$pH, xlab = 'pH'),  
 qplot(red\_wine$sulphates, xlab = 'Sulphates'),  
 qplot(red\_wine$alcohol, xlab = 'Alcohol'),  
 qplot(red\_wine$quality, xlab = 'Quality'),  
 ncol = 4)

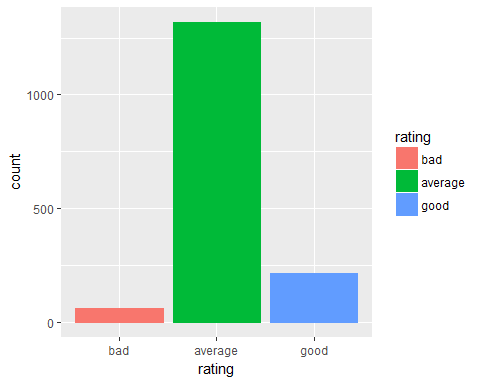


# Univariate Analysis

I first looked at wine quality. Although it has a discrete range of only 3-8, we can roughly see that there is some amount of normal distribution. A large majority of the wines examined received ratings of 5 or 6, and very few received 3, 4, or 8. There's not much more we can do with this histogram, as both decreasing or increasing bin sizes would distort the data.

Given the ratings and distribution of wine quality, I'll instantiate another categorical variable, classifying the wines as 'bad' (rating 0 to 4), 'average' (rating 5 or 6), and 'good' (rating 7 to 10).

## bad average good   
## 63 1319 217



### What is the structure of your dataset?

1.fixed.acidity, density, pH, sulphates,chlorides have normal distributions. 2.volatile.acidity,free.sulpfur.dioxide, total.sulfur.dioxide,alcohol are postively skewed. 3.citric.acid have lot of zero values, maybe some records are missing 4.residual.sugar and chlorides have extreme outliers.

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

quality , pH and density are proportional to each other

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

SO2, Acidity (both), density are likely to contribute to quality of wine.

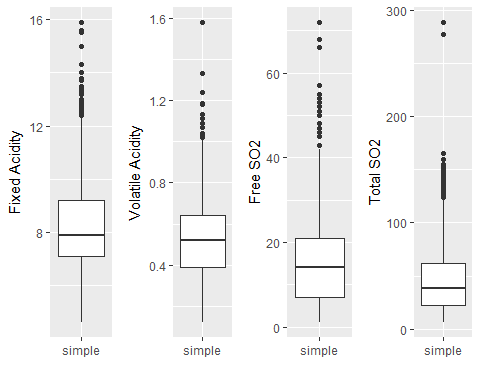
### Did you create any new variables from existing variables in the dataset?

I have created rating variable, which is ordered and devide samples in 'bad','average','good'

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

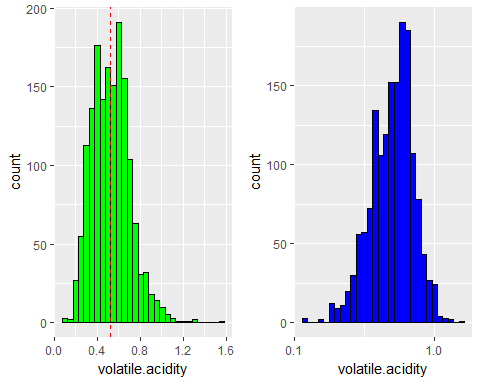
According to all the above plots, there are some outliers in some of the features like SO2(free and total), acidity (fixed and volatile).

do\_boxplot <- function(column, ylab) {  
 return(qplot(data = red\_wine, x = 'simple',  
 y = column, geom = 'boxplot',  
 xlab = '',  
 ylab = ylab))}  
grid.arrange(do\_boxplot(red\_wine$fixed.acidity,'Fixed Acidity'),  
 do\_boxplot(red\_wine$volatile.acidity,'Volatile Acidity'),  
 do\_boxplot(red\_wine$free.sulfur.dioxide, 'Free SO2'),  
 do\_boxplot(red\_wine$total.sulfur.dioxide,'Total SO2'),  
 ncol =4  
 )

 Box plots is very well suited for such plots which have outliers as it can be seen from above figure, we can also see the median for each of the plots.

Also the distribution for Volatile acidity apears to be bimodal normal distribution. But when taking log distribution, the plot becomes normal distributed.

p1 <- ggplot(data = red\_wine, aes(x =volatile.acidity))+  
 geom\_histogram(color = 'black',fill='green') +geom\_vline(aes(xintercept=mean(volatile.acidity)),  
 color="red", linetype=2)  
  
p2 <- ggplot(data = red\_wine, aes(x =volatile.acidity))+  
 geom\_histogram(color = 'black',fill='blue') +   
 scale\_x\_log10()  
  
grid.arrange(p1 ,p2 ,ncol =2)

 ###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? citric.acid have lot of zero values, maybe some records are missing or they are actually zero

summary(red\_wine$citric.acid)

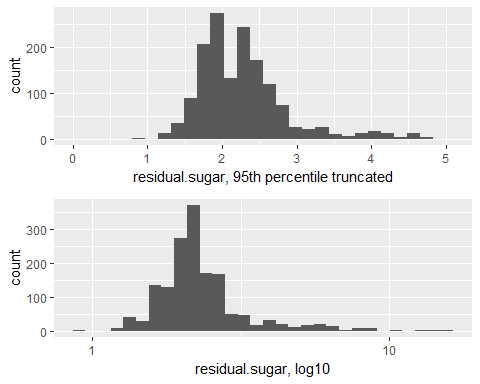
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.090 0.260 0.271 0.420 1.000

summary(log10(red\_wine$citric.acid))

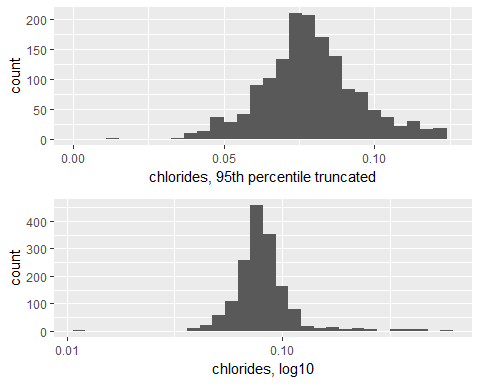
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -Inf -1.0458 -0.5850 -Inf -0.3768 0.0000

We can see that minimum count ,mean for citric.acid is negative infinity that means a lot of observations are zero.

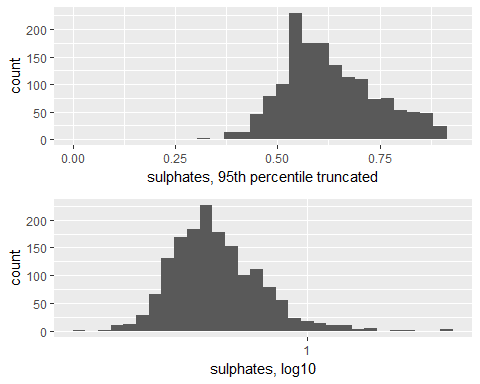
p1 <- ggplot(data = red\_wine, aes(x = residual.sugar)) +  
 geom\_histogram() +  
 scale\_x\_continuous(lim = c(0, quantile(red\_wine$residual.sugar, 0.95))) +  
 xlab('residual.sugar, 95th percentile truncated')  
  
p2 <- p1 + scale\_x\_log10() + xlab('residual.sugar, log10')  
grid.arrange(p1, p2, ncol=1)



p1 <- ggplot(data = red\_wine, aes(x = chlorides)) +  
 geom\_histogram() +  
 scale\_x\_continuous(lim = c(0, quantile(red\_wine$chlorides, 0.95))) +  
 xlab('chlorides, 95th percentile truncated')  
  
p2 <- p1 + scale\_x\_log10() + xlab('chlorides, log10')  
grid.arrange(p1, p2, ncol=1)

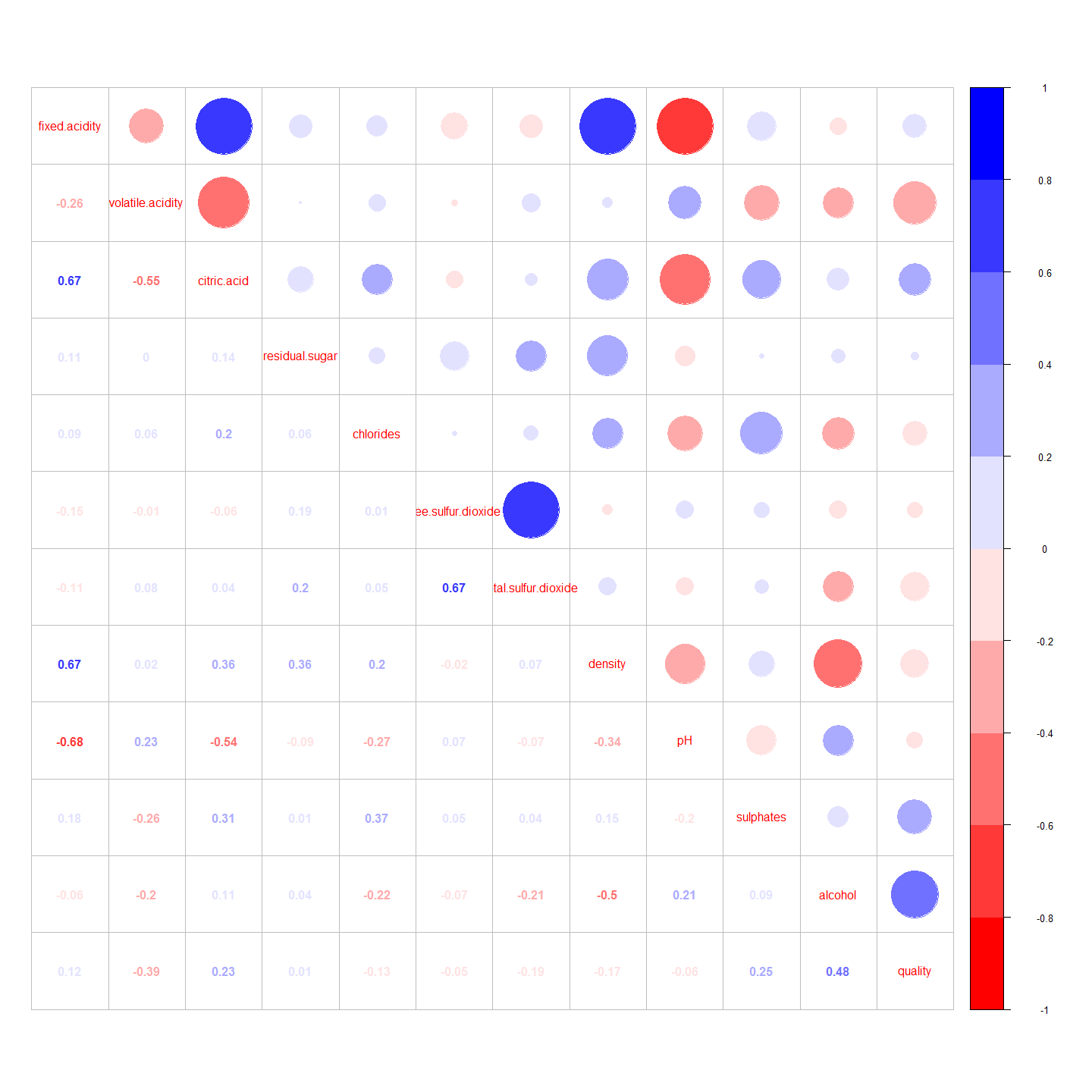


p1 <- ggplot(data = red\_wine, aes(x = sulphates)) +  
 geom\_histogram() +  
 scale\_x\_continuous(lim = c(0, quantile(red\_wine$sulphates, 0.95))) +  
 xlab('sulphates, 95th percentile truncated')  
  
p2 <- p1 + scale\_x\_log10() + xlab('sulphates, log10')  
grid.arrange(p1, p2, ncol=1)



rm(p1, p2)

# Bivariate Plots Section

Lets see the correlation behaviour between the features.  Correlation matrix output shows following behaviour:

## Fixed Acidity

It shows positive correlation with citric acid which is true since citric acid is one of the fixed acid. It also shows positive correlation with density. It also shows significant negative correlation with pH and volatile acidity.

## Volatile Acidity

It is highly negatively correlated with citric acid and quality.

## Free SO2

It shows significant positive correlation with total SO2, and very less correlation with sulphates.

## Density

Significant negative correlation is observed with alcohol, acidity (fixed and citric acid) and pH.

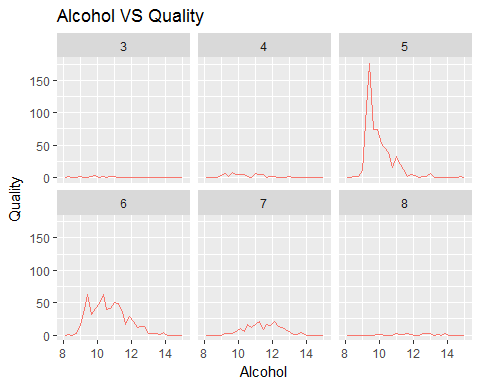
## Quality

Quality and alcohol is positively correlated along with negative correlation with volatile acidity.

First lets plot all the variables that have postive or negative correlation with quality

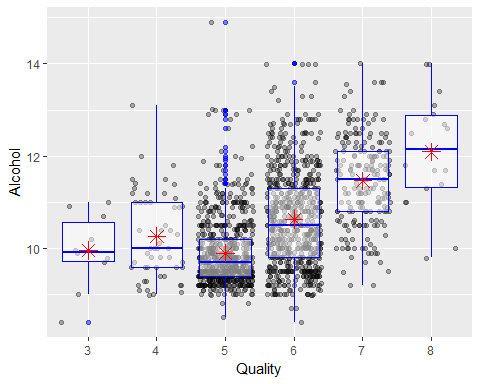
ggplot(red\_wine, aes(x =alcohol))+  
 geom\_freqpoly(aes(color ='red'))+ facet\_wrap(~quality)+  
 theme(legend.position = "none") +  
 ggtitle("Alcohol VS Quality") +  
 xlab("Alcohol") +  
 ylab("Quality")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 Only samples with quality 5 and 6 shows high quantity of alcohol ie average Rating samples. But samples with Rating good and bad doesnot show any significant change.

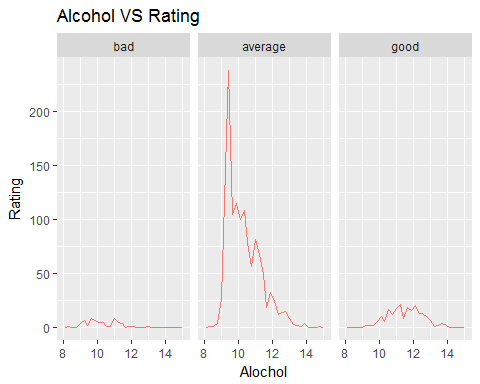
Let us plot and see the mean ,median for all the Quality ratings

ggplot(aes(factor(quality),   
 alcohol),   
 data = red\_wine) +  
 geom\_jitter( alpha = .3) +   
 geom\_boxplot( alpha = .5,color = 'blue')+   
 stat\_summary(fun.y = "mean",   
 geom = "point",   
 color = "red",   
 shape = 8,   
 size = 4)+  
 xlab('Quality')+  
 ylab('Alcohol')

 Higher Quality wine have high mean and median value of Alcohol as compared to Lowest Quality of Red Wine.

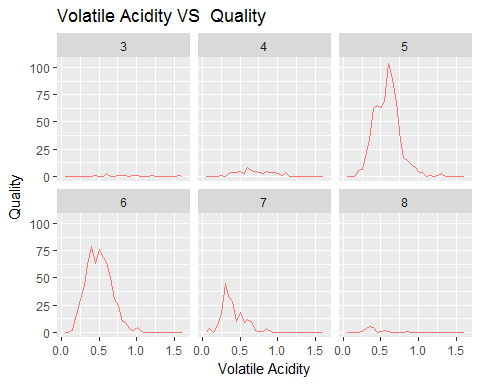
Let us plot the Alcohol vs Rating graph

ggplot(red\_wine, aes(x =alcohol))+  
 geom\_freqpoly(aes(color ='red'))+ facet\_wrap(~rating)+  
 theme(legend.position = "none") +  
 ggtitle("Alcohol VS Rating") +  
 xlab("Alochol") +  
 ylab("Rating")



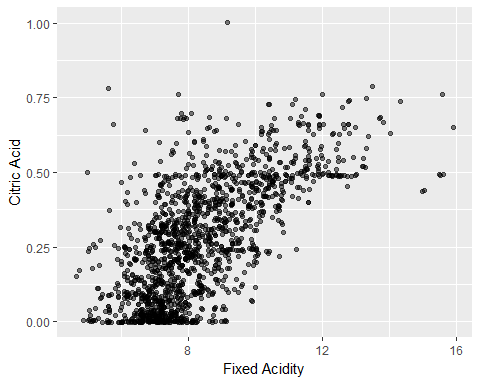
Now let us see quality and volatile acidity

ggplot(red\_wine, aes(x =volatile.acidity))+  
 geom\_freqpoly(aes(color ='red'))+ facet\_wrap(~quality)+  
 theme(legend.position = "none") +  
 ggtitle("Volatile Acidity VS Quality") +  
 xlab("Volatile Acidity") +  
 ylab("Quality")

 We can say the same for Volatile acidity as well that average Rating wine samples have high volume of volatile acid present in them.

# Fixed Acid VS Citric Acid

ggplot(red\_wine,aes(x = fixed.acidity, y = citric.acid))+  
 geom\_point(alpha = 0.5, position = "jitter")+  
 xlab('Fixed Acidity')+  
 ylab('Citric Acid')



Since citric acid is one of the components of fixed acidiy hence it is showing such strong correlation.

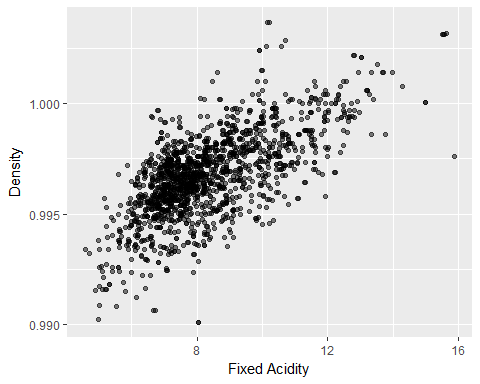
Lets check the correlation coefficient

cor.test(red\_wine$fixed.acidity, red\_wine$citric.acid, method ='pearson')

##   
## Pearson's product-moment correlation  
##   
## data: red\_wine$fixed.acidity and red\_wine$citric.acid  
## 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

I used pearson method to calculate pearson between fixed acid and citric acid and the coeficient cme to be 0.6717.A good rule of thumb says that coefficient above 0.5 is moderate and coefficient above 0.7 is good.So this shows almost good rleationship between the two. #Fixed Acid VS Density

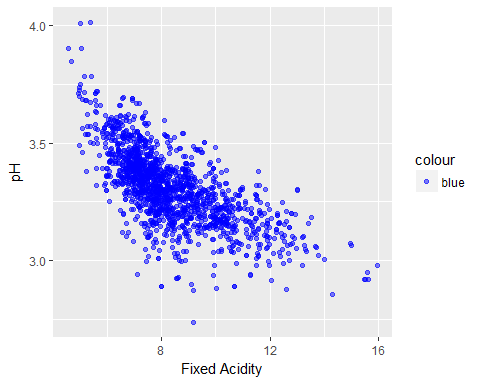
ggplot(red\_wine, aes(x = fixed.acidity, y =density))+  
 geom\_point(alpha = 0.5, position = "jitter")+  
 xlab('Fixed Acidity')+  
 ylab('Density')

 Most data is clustered around 0.995 to 1 g/dm3 density and 7 to 8.5 g/dm3. Here i used jitter to add small amount of random variation to the location of each point, and is a useful way of handling overplotting caused by discreteness in smaller datasets. Also alpha parameter is used to set transparency of each point.

# Fixed Acid vs PH

The above correlation matrix shows higly negative correlation ,lets check in the graph and by how much.

ggplot(red\_wine,aes(x = fixed.acidity, y = pH))+  
 geom\_point(alpha = 0.5, position = "jitter",aes(color = 'blue'))+  
 scale\_color\_manual(values = c("blue"))+  
 xlab('Fixed Acidity')+  
 ylab('pH')



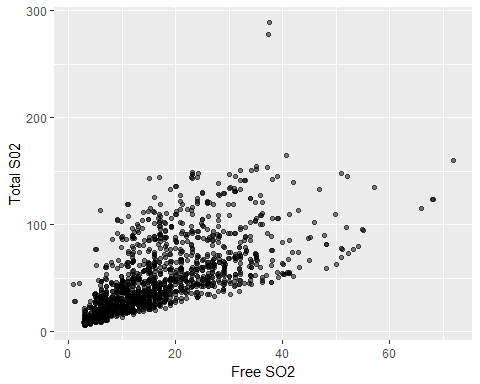
cor.test(red\_wine$fixed.acidity, red\_wine$pH, method ='spearman')

## Warning in cor.test.default(red\_wine$fixed.acidity, red\_wine$pH, method =  
## "spearman"): Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: red\_wine$fixed.acidity and red\_wine$pH  
## S = 1162900000, p-value < 2.2e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## -0.7066736

Solutions with a pH less than 7 are acidic and solutions with a pH greater than 7 are basic.So that means leeser the pH value higher the acidity .This is the reason we see negative correlation with coefficient of -0.7

ggplot(red\_wine,aes(x = free.sulfur.dioxide, y = total.sulfur.dioxide))+  
geom\_point(alpha = 0.5, position = "jitter")+  
xlab('Free SO2')+  
ylab('Total S02')



cor.test(red\_wine$fixed.acidity, red\_wine$citric.acid, method ='spearman')

## Warning in cor.test.default(red\_wine$fixed.acidity, red\_wine$citric.acid, :  
## Cannot compute exact p-value with ties

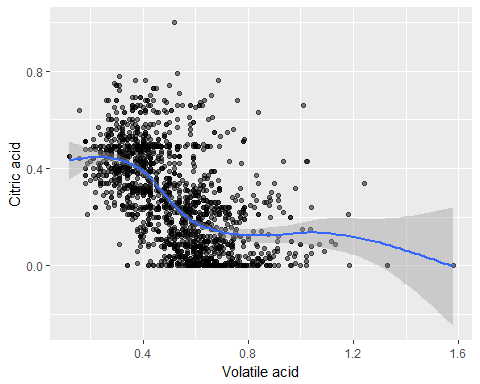
##   
## Spearman's rank correlation rho  
##   
## data: red\_wine$fixed.acidity and red\_wine$citric.acid  
## S = 230510000, p-value < 2.2e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.6617084

It has a correlation coefficient of 0.6617 which indicates a strong relationship between free and total sulphur dioxides and it is true because both are oxides.

# Volatile acidity vs citric acid

ggplot(red\_wine,aes(x = volatile.acidity, y = citric.acid))+  
 geom\_point(alpha = 0.5)+ geom\_smooth(method = 'auto')+  
 xlab('Volatile acid')+  
 ylab('Citric acid')

## `geom\_smooth()` using method = 'gam'

 My original hypothesis was since both are acids they will be positively correlated but according to scatter plot , these variables are negatively correlated.

# 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?

Volatile acid and alcohol vs quality showed similar behaviour , where rating with average show significant increase while rating with good and bad samples didnot show much which was totally different behaviour as compared to other relations of variables.

### Did you observe any interesting relationships between the other features Ctric acid and Volatile acid are negatively correlated with the coefficient of -0.61025

cor.test(red\_wine$volatile.acidity, red\_wine$citric.acid, method ='spearman')

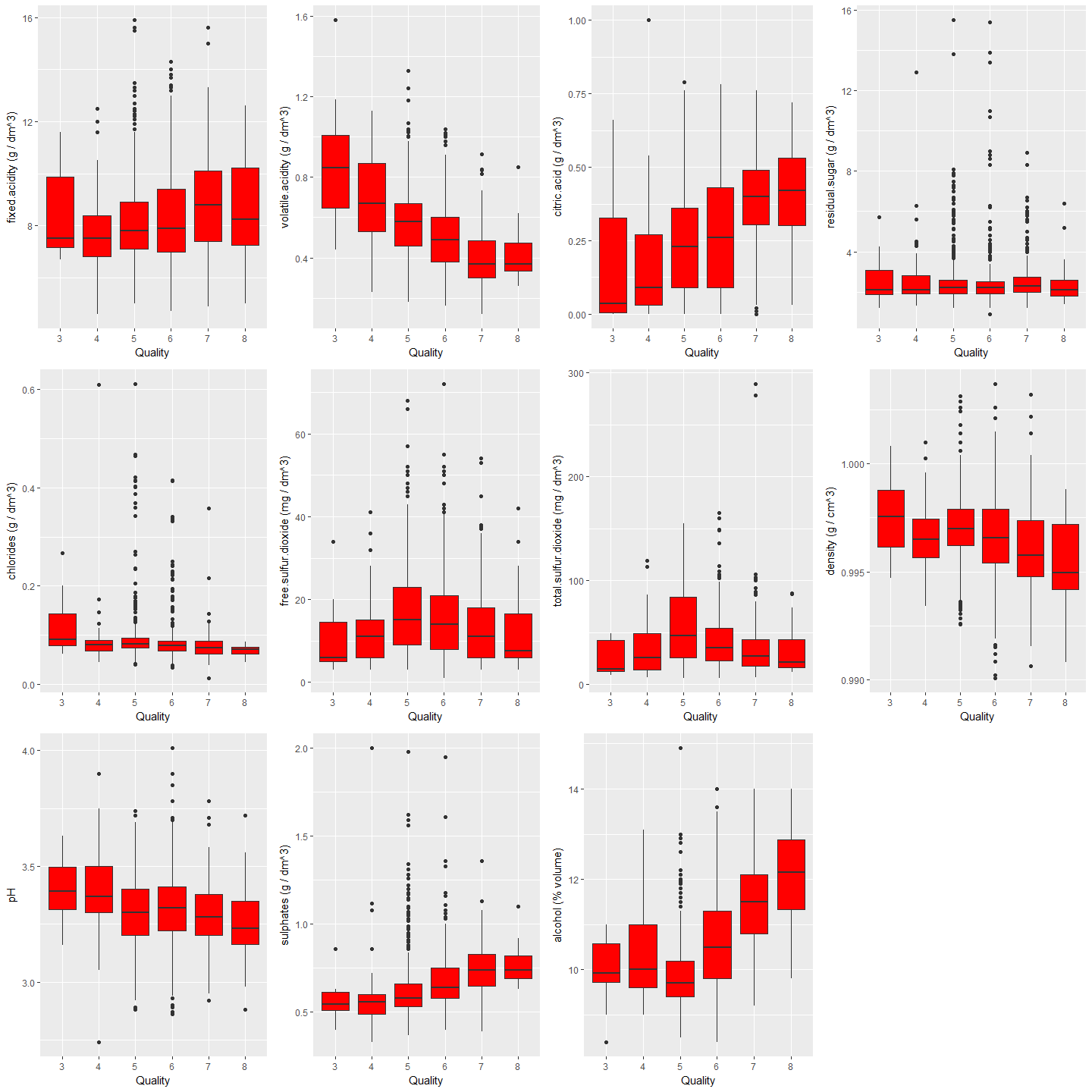
## Warning in cor.test.default(red\_wine$volatile.acidity, red\_wine  
## $citric.acid, : Cannot compute exact p-value with ties

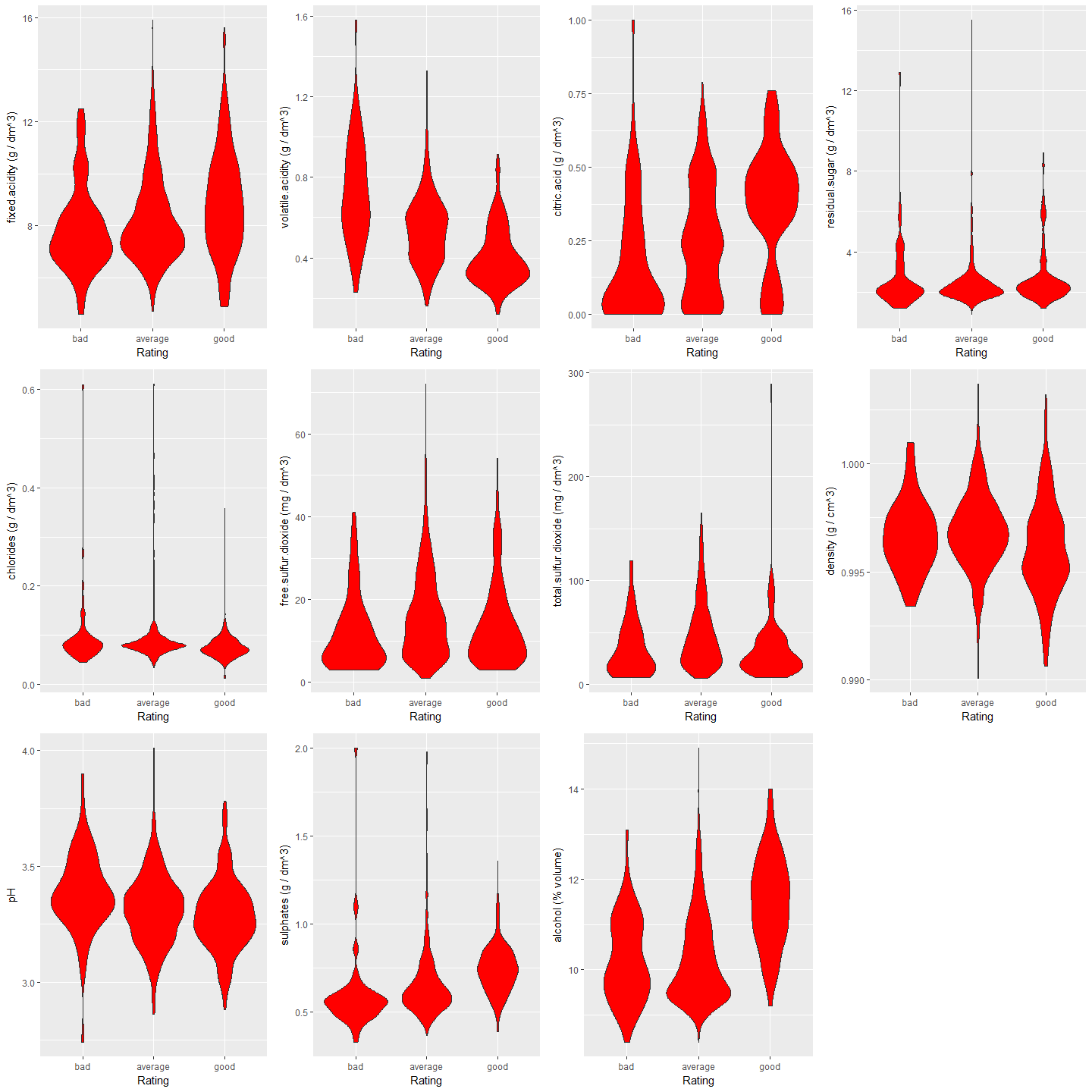
##   
## Spearman's rank correlation rho  
##   
## data: red\_wine$volatile.acidity and red\_wine$citric.acid  
## S = 1097200000, p-value < 2.2e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## -0.6102595

Because volatile acidity reduces the quality of wine (negative correlation) while citric acid increases wine quality (positive correlation).

### What was the strongest relationship you found?

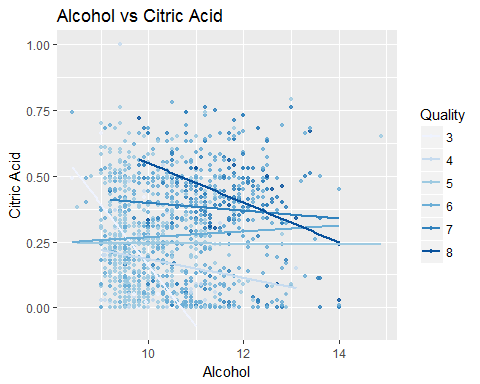
I found the relation between Fixed acid and Citric acid to be strongest ,with the correlation coefficiet around 0.7 and relation between fixed acid and pH to be highly neative with -0.7 coefficient .





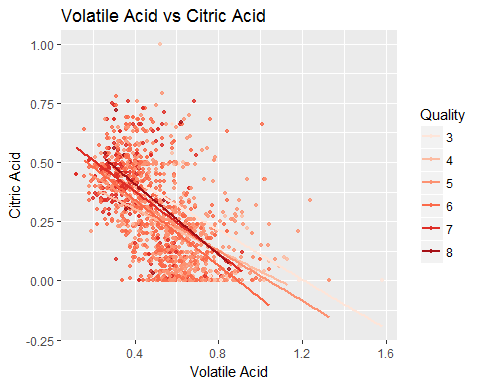
# Multivariate Plots Section

ggplot(aes(x = alcohol,   
 y = citric.acid , color = factor(quality)),   
 data = red\_wine) +  
 geom\_point(alpha = 0.8, size = 1) +  
 geom\_smooth(method = "lm", se = FALSE,size=1) +  
 scale\_color\_brewer(type='seq',  
 guide=guide\_legend(title='Quality'))+  
 ggtitle("Alcohol vs Citric Acid") +  
 xlab("Alcohol") +  
 ylab("Citric Acid")

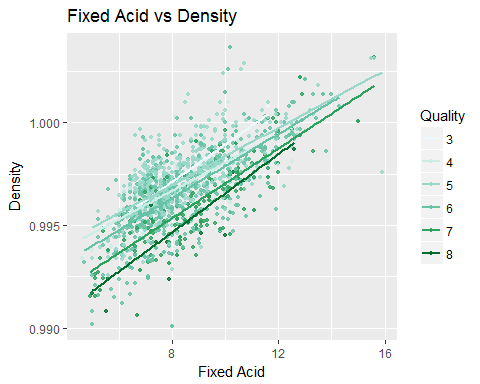


From this plot we can say that redwine with good rating is more clustered from 11 to 13/dm^3 alcohol and 0.25 to 0.75gm/dm^3 Citric acid

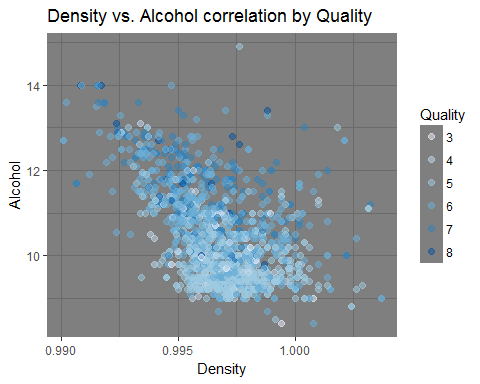
ggplot(aes(x = volatile.acidity,   
 y = citric.acid , color = factor(quality)),   
 data = red\_wine) +  
 geom\_point(alpha = 0.8, size = 1) +  
 geom\_smooth(method = "lm", se = FALSE,size=1) +  
 scale\_color\_brewer(type='seq',  
 guide=guide\_legend(title='Quality'), palette = "Reds")+  
 ggtitle("Volatile Acid vs Citric Acid") +  
 xlab("Volatile Acid") +  
 ylab("Citric Acid")

 We find out something interesting that most of the good wines are clustered around 0.2 to 0.6 gm/dm^3 Volatile acid and 0.25 to 0.75 gm/dm^3 Citric Acid

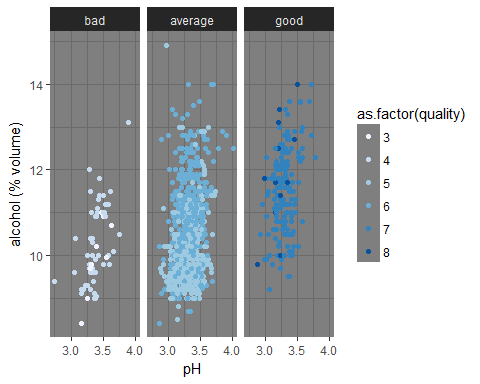
ggplot(aes(x = fixed.acidity,   
 y = density , color = factor(quality)),   
 data = red\_wine) +  
 geom\_point(alpha = 0.8, size = 1) +  
 geom\_smooth(method = "lm", se = FALSE,size=1) +  
 scale\_color\_brewer(type='seq',  
 guide=guide\_legend(title='Quality'), palette = 2)+  
 ggtitle("Fixed Acid vs Density") +  
 xlab("Fixed Acid") +  
 ylab("Density")

 Good wines seems to be distributed across the plot.

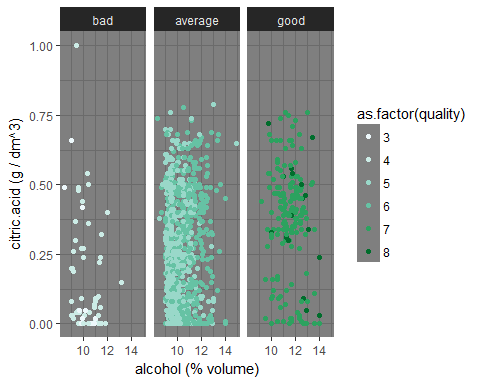
ggplot(red\_wine, aes(x = density, y = alcohol, color = as.factor(quality))) +  
 geom\_point(alpha = 1/2, size = 2) +  
 scale\_color\_brewer() +  
 ggtitle('Density vs. Alcohol correlation by Quality') +  
 ylab("Alcohol") +  
 xlab("Density") +  
 labs(color = "Quality")+theme\_dark()

 We can see that there is large conectration of average rating wines in range from 0.995 to 1 gm/cm^3 density and 9 to 10 (%volume) alchol. Whereas good rating wines have high concentraion of Alcohol

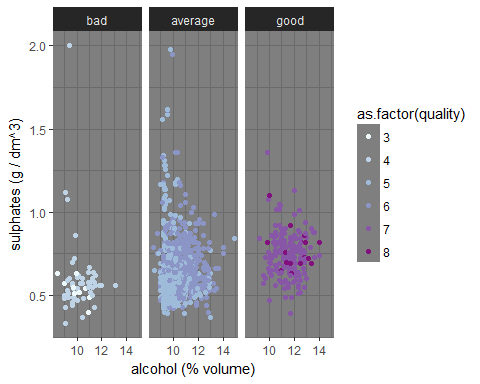
ggplot(data = red\_wine, aes(pH, alcohol, color = as.factor(quality))) +  
 geom\_point() +  
 facet\_wrap(~rating) +  
 ylab('alcohol (% volume)') +  
 theme\_dark() +  
 scale\_color\_brewer(type = 'seq', palette = 1)



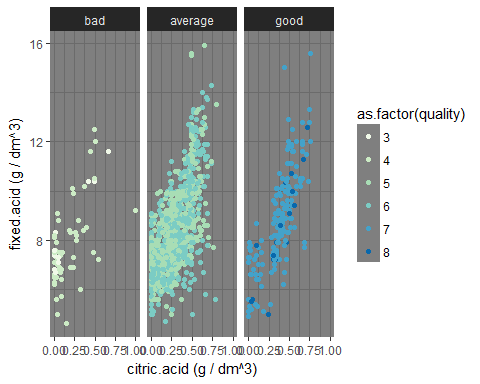
ggplot(data = red\_wine, aes(alcohol, citric.acid, color = as.factor(quality))) +  
 geom\_point() +  
 facet\_wrap(~rating) +  
 xlab('alcohol (% volume)') +  
 ylab('citric.acid (g / dm^3)') +  
 scale\_color\_brewer(type = 'seq', palette = 2) +  
 theme\_dark()



ggplot(data = red\_wine, aes(alcohol, sulphates, color = as.factor(quality))) +  
 geom\_point() +  
 facet\_wrap(~rating) +  
 xlab('alcohol (% volume)') +  
 ylab('sulphates (g / dm^3)') +  
 scale\_color\_brewer(type = 'seq', palette = 3) +  
 theme\_dark()



ggplot(data = red\_wine, aes(citric.acid, fixed.acidity, color = as.factor(quality))) +  
 geom\_point() +  
 facet\_wrap(~rating) +  
 xlab('citric.acid (g / dm^3)') +  
 ylab('fixed.acid (g / dm^3)') +  
 scale\_color\_brewer(type = 'seq', palette = 4) +  
 theme\_dark()



# 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.Average rating wine show high concentration of alcohol then comes good rating and then bad rating wine. 2.Average rating wine show high concentration of citric acid then comes good rating and then bad rating wine. 3.Average rating wine show high positive mutual relationship of Fixed acid and Citric Acid then comes good rating and then bad rating wine.

### Were there any interesting or surprising interactions between features?

Good rating wine have low concentration of Volatile acid and high concentration of Citric acid

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

Let us use decision tree as our model

set.seed(42)  
wine\_sample <- sample\_n(2, tbl = red\_wine, size = 150)  
t <- sample(2, nrow(wine\_sample),replace = TRUE, prob =c(0.7, 0.3))  
train <- wine\_sample[t ==1,]  
test <- wine\_sample[t ==2,]  
my\_formula <- rating ~ volatile.acidity+density+alcohol+pH+quality  
wine\_ctree <- ctree(my\_formula, data = train)  
table(predict(wine\_ctree),train$rating)

##   
## bad average good  
## bad 7 0 0  
## average 0 88 0  
## good 0 0 15

plot(wine\_ctree)

testpred <- predict(wine\_ctree, newdata = test)  
table(testpred, test$quality)

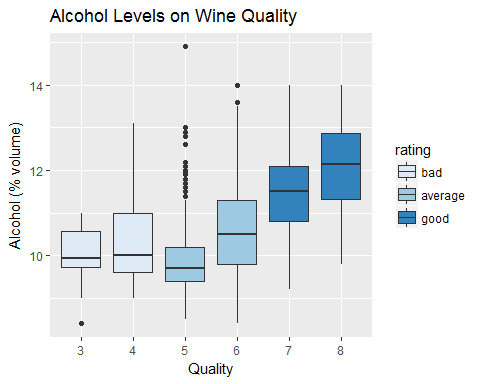
##   
## testpred 4 5 6 7  
## bad 1 0 0 0  
## average 0 21 13 0  
## good 0 0 0 5

No error in predicted values of test data , i think this is the strength of my model.Also by increasing the number of variables i can increase the efficiency of my model.

# Final Plots and Summary

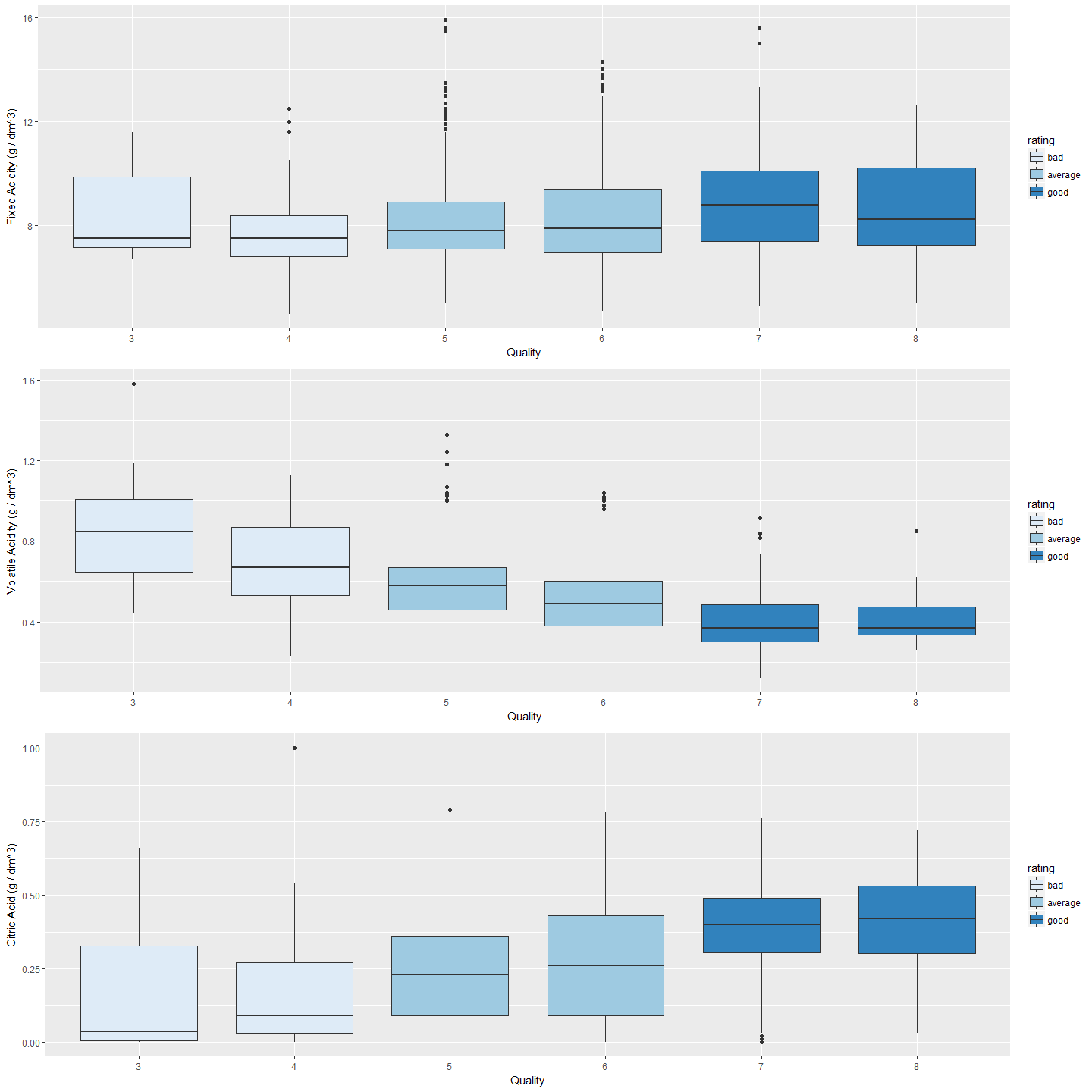
### Plot One : Effect of alcohol on wine quality

ggplot(data = red\_wine, aes(as.factor(quality), alcohol, fill = rating)) +  
 geom\_boxplot() +  
 ggtitle('Alcohol Levels on Wine Quality') +  
 xlab('Quality') +  
 ylab('Alcohol (% volume)') +  
 scale\_fill\_brewer(type = 'seq', palette = 1)

 ### Description One

As alcohol is highly correlated with the quality, it is better to see its pattern with varying rating. From the above plot, it clearly shows higher % of alcohol yields better wine.

### Plot Two : Effect of acids on wine quality

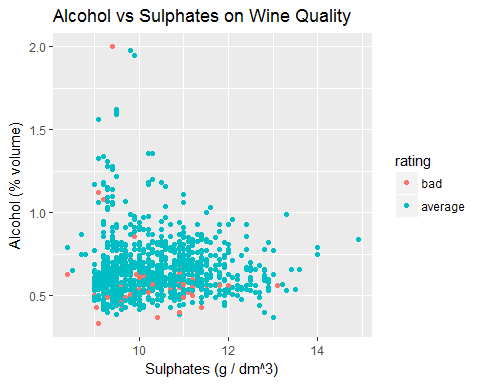


### Description Two

As more the acidic better is the wine. It would be better to see which acids have more impact on wine quality. Above plot shows, fixed.acidity and citric.acid have highly correlated with quality but volatile.acidity has negative impact on quality.

### Plot Three Ideal vs poor wine

ggplot(data = subset(red\_wine, rating != 'good'), aes(alcohol, sulphates, color = rating)) +  
 geom\_point() +  
 ggtitle('Alcohol vs Sulphates on Wine Quality') +  
 xlab('Sulphates (g / dm^3)') +  
 ylab('Alcohol (% volume)')



### Description Three

It would be great to see the real pattern between good and bad wines. Above plot differentiate between good and bad wines. It shows higher the % of alcohol and higher the sulphates give better wines.

# Reflection

So, there are lots of features on which the wine quality is depend. We do a lot to find the relationship between every variables in the dataset, try to get some linear relation using geom\_smooth(). And, after this EDA, I can conclude that the major factors for better wine quality is alcohol, acidity and sulphates. These features must also be in required content otherwise negative impact will effect the wine quality. Also, we can't be totally sure about quality index also it has been taken by some experts. We've also concluded that there is linear relationship between pH and quality with negative slope.

One thing that is still unclear is the amount of residual.sugar. It contains many outliers, also after doing some operation we get its common range from 1 to 4. But we can't find its amount for ideal wine quality. I think more future research need to be done to find its ideal quantity.

Also i have use Decision Tree as my model to train and test my dataset .