Red Wine Quality Analysis

Krishna P Koirala

1/23/2018

Table of Contents

[My Words 2](#_Toc511079909)

[Reading data 2](#_Toc511079910)

[Getting overview of data 2](#_Toc511079911)

[Removing X column(this is unnecessary) 4](#_Toc511079912)

[Histogram of quality 5](#_Toc511079913)

[Conclusion 5](#_Toc511079914)

[Acidity Study 5](#_Toc511079915)

[Residual sugar Study 11](#_Toc511079916)

[Conclusion (sugar vs quality) 16](#_Toc511079917)

[Alcohol Study 16](#_Toc511079918)

[Free Sulpher dioxide Study 20](#_Toc511079919)

[pH Study 25](#_Toc511079920)

[Chlorides Study 27](#_Toc511079921)

[Conclusion (Chlorides) 31](#_Toc511079922)

[Sulphates Study 31](#_Toc511079923)

[Conclusion(sulphates vs quality) 34](#_Toc511079924)

[Density Study 34](#_Toc511079925)

[Conclusion(Density vs quality) 35](#_Toc511079926)

[Correlation coefficient 35](#_Toc511079927)

[Bar plots 36](#_Toc511079928)

[Question. 40](#_Toc511079929)

[Predictive modeling (Linear modeling for quality) 41](#_Toc511079930)

[Data partition 41](#_Toc511079931)

[Fit Linear model on the traingin dataset 42](#_Toc511079932)

[Stepwise variable selection 42](#_Toc511079933)

[Evaluate Predictive performance of the final model 49](#_Toc511079934)

[Three final plots 49](#_Toc511079935)

[First Plot 50](#_Toc511079936)

[Second Plot 51](#_Toc511079937)

[Third Plots 52](#_Toc511079938)

[Reflection 53](#_Toc511079939)

[Alternative way of analysis 53](#_Toc511079940)

# My Words

I am very excited to do analysis of Red wine data which contains 1,599 red wines with 11 variables on the chemical properties of the wine. Although I am not a greate fan of Wine, my focus would be to see how each chemical component influences the quality of wine (0 'very bad' to 10 'very excellent'). This dataset is public available for research. The details are described in [Cortez et al., 2009].

A dm^3 unit is mentioned in the data set. Where dm stands for Decimeter, where 1 decimeter = 10 centimeters. Other units are familier.

# Reading data

rm(list = ls())  
RedWineQuality <- read.csv("~/Desktop/Nanodegree/wineQualityReds.csv")

# Getting overview of data

head(RedWineQuality)

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

tail(RedWineQuality)

## X fixed.acidity volatile.acidity citric.acid residual.sugar  
## 1594 1594 6.8 0.620 0.08 1.9  
## 1595 1595 6.2 0.600 0.08 2.0  
## 1596 1596 5.9 0.550 0.10 2.2  
## 1597 1597 6.3 0.510 0.13 2.3  
## 1598 1598 5.9 0.645 0.12 2.0  
## 1599 1599 6.0 0.310 0.47 3.6  
## chlorides free.sulfur.dioxide total.sulfur.dioxide density pH  
## 1594 0.068 28 38 0.99651 3.42  
## 1595 0.090 32 44 0.99490 3.45  
## 1596 0.062 39 51 0.99512 3.52  
## 1597 0.076 29 40 0.99574 3.42  
## 1598 0.075 32 44 0.99547 3.57  
## 1599 0.067 18 42 0.99549 3.39  
## sulphates alcohol quality  
## 1594 0.82 9.5 6  
## 1595 0.58 10.5 5  
## 1596 0.76 11.2 6  
## 1597 0.75 11.0 6  
## 1598 0.71 10.2 5  
## 1599 0.66 11.0 6

summary(RedWineQuality)

## 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

str(RedWineQuality)

## '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 ...

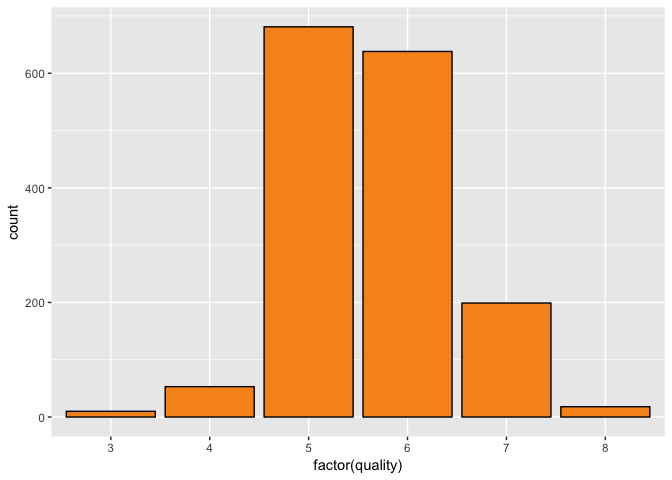
The data does not contains NA values which is cool.

# Removing X column(this is unnecessary)

RedWineQuality$X <- NULL  
colnames(RedWineQuality)

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

# Histogram of quality



table(RedWineQuality$quality)

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

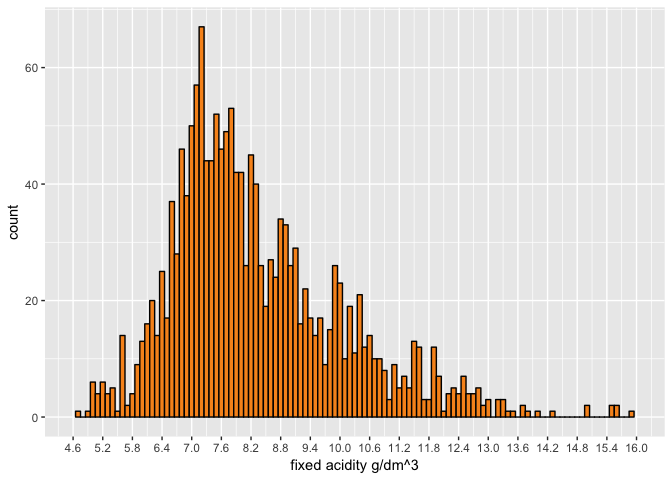
# Conclusion

Most of the people gave rating 5 and 6. Few of the people really did not like the quality of wine.

Nobody gave rating 0, 1, 2, 9, 10. This might be because most of the people randomly choose the rating 5 and 6. And surprisingly no body rated 9 and 10 means the wine quality looks not so good.

# Acidity Study

Let's see how fixed acidity is distributed



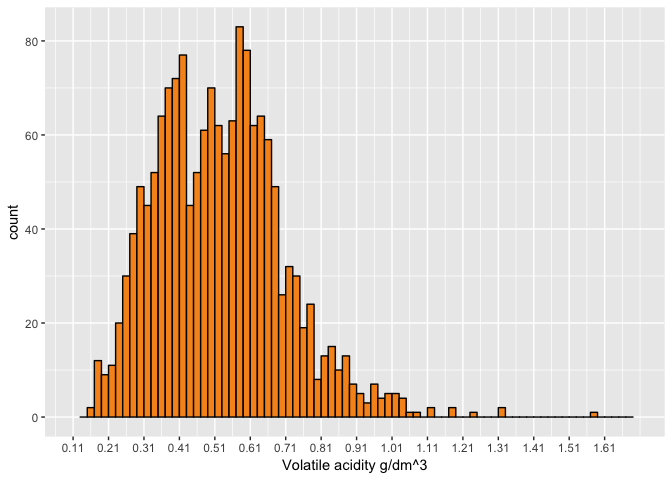
Fixed acidity almost have normal distribution

Lets see the summary of fixed acidity vs quality. I can find summary of all variables but I do not see that is so important here.

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

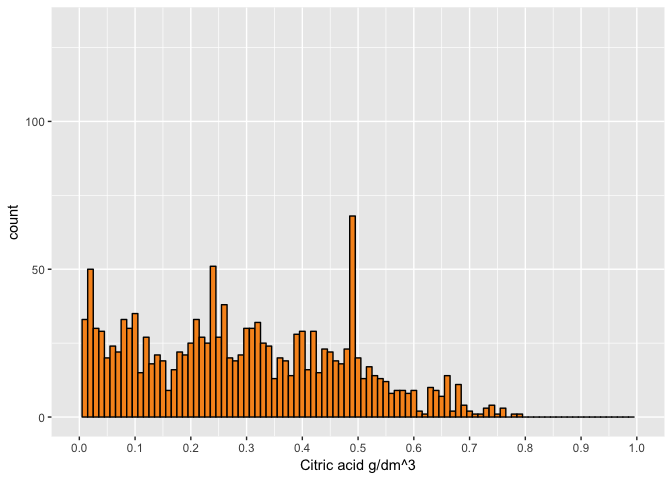
Range of fixed acid is low for rating 3 and 8. This is because very few people gave those rating.

Lets see how volatile acidity distributed



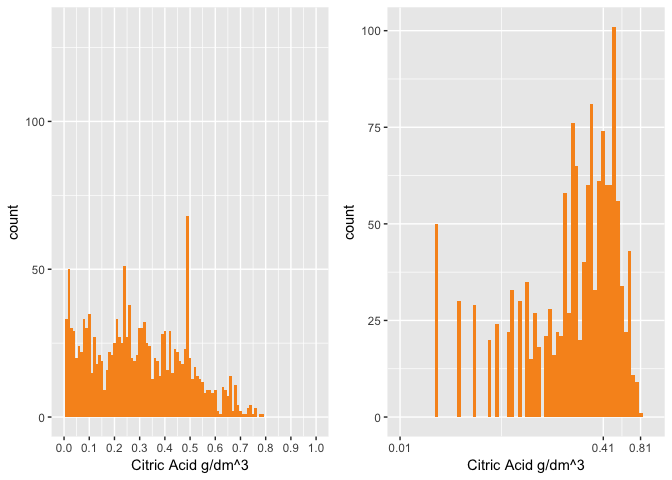
Volatile acidity almost have normal distribution.

Lets see how citric acid is distributed



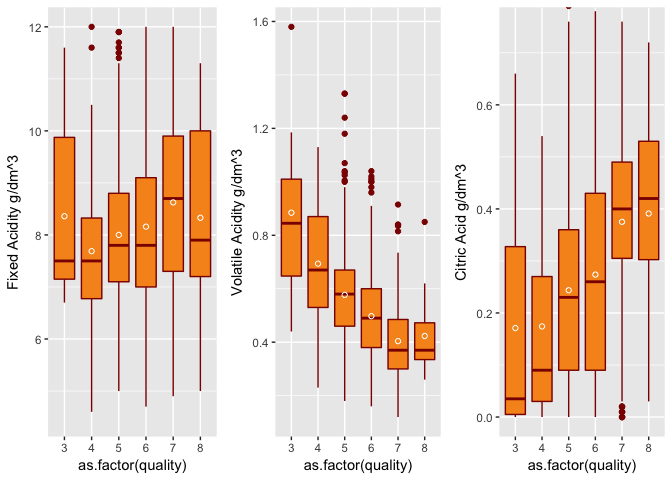
Citric acid is not normally distributed.

What happens if I transfer the citric acid plot using log10 First I need to make subset of data which contains no zero values of citric acid otherwise log gives infinity.



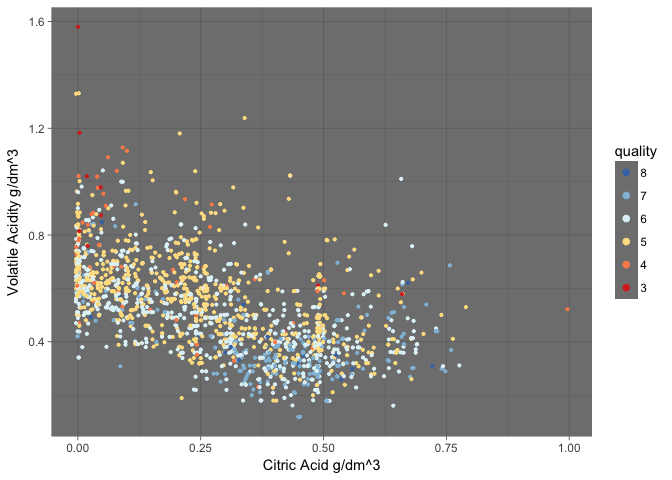
The transformation is not depecting normality there are lots of peaks which is bad.

Plot Acidity vs quality



It looks like acidity has predictive capability. As rating increases with increase value of citric acid. The rating decreases with increased value of volatile acidity.

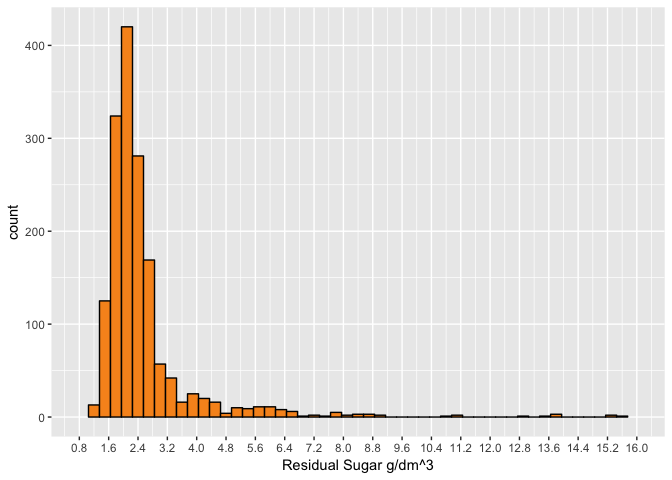
Lets see how quality realtes to those acids value



From above we can see that most of the high rating is below volatile acidity 0.4 and citric acid between 0.25 to 0.62. There are wines having citric acid values 0 having volatile acidity greater than 0. Rating 6 is randomly distributed throughout the values of citric and volatile acid Most of the rating 5 is above the volatile acidity value 0.4 and below 1 except some outliers.

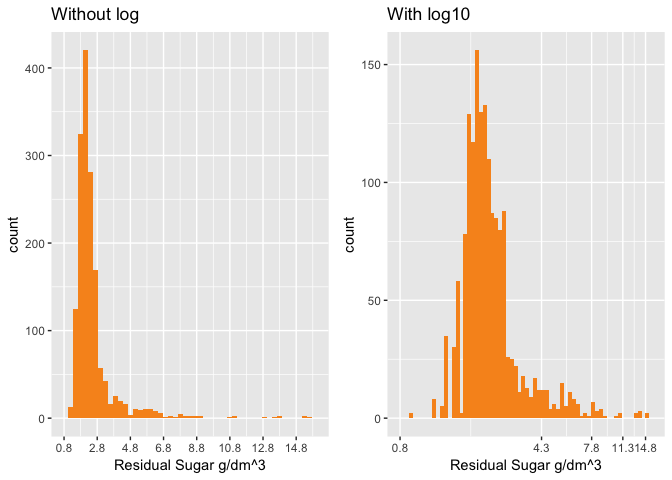
# Residual sugar Study

Residual sugar plot



Count of wines having sugars more than 6.8 are very low. Wines having sugars value lower than 6.8 are high in number. Although the range is about 14.

Lets do the log10 transformation of residual sugar and see what can be seen

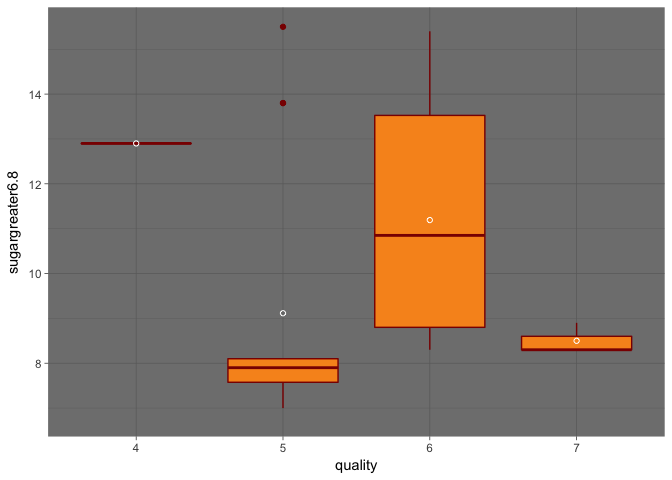


The log transformation gave slightly better normality than without transformation. Still it looks skewed to the right.

Lets break the sugars value and see what we can find.

sugargreater6.8 <- subset(RedWineQuality, residual.sugar >= 6.8)  
sugarlessthn6.8 <- subset(RedWineQuality, residual.sugar < 6.8)

Let's make plot of sugar greater than 6.8

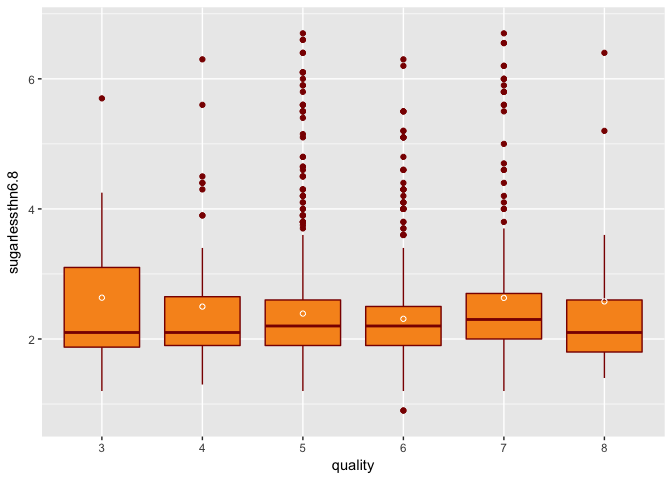


by(sugargreater6.8$residual.sugar, sugargreater6.8$quality, table)

## sugargreater6.8$quality: 4  
##   
## 12.9   
## 1   
## --------------------------------------------------------   
## sugargreater6.8$quality: 5  
##   
## 7 7.2 7.3 7.5 7.8 7.9 8.1 13.8 15.5   
## 1 1 1 1 2 3 2 2 1   
## --------------------------------------------------------   
## sugargreater6.8$quality: 6  
##   
## 8.3 8.6 8.8 9 10.7 11 13.4 13.9 15.4   
## 1 1 2 1 1 2 1 1 2   
## --------------------------------------------------------   
## sugargreater6.8$quality: 7  
##   
## 8.3 8.9   
## 2 1

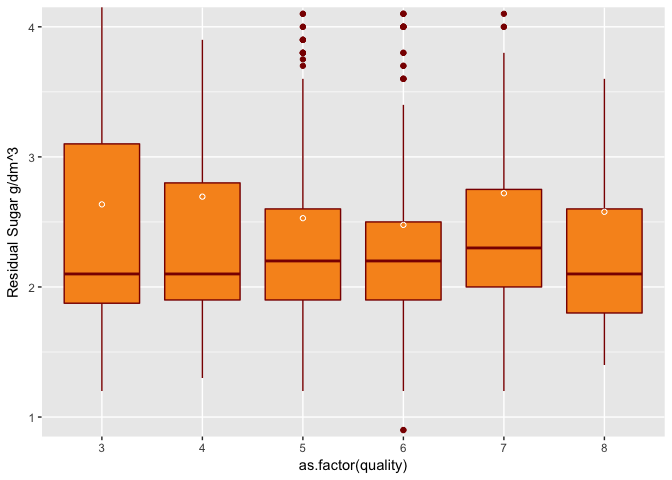
Nobody gave rating 3 and 8 for residual sugar value greater than 6.8. Only one people gave rating 4 for residual sugar value greater than 6.8. Three people gave raing rating 7. Range of the sugar value for rating 6 would be higher than all others if we remove two outliers of rating 5. The mean of residual sugar for rating 5 is out of the box because of presence of two outliers.

Lets make a plot of sugar less then 6.8



Quality 3, 4 and 8 had few outliers thane quality 5, 6 and 7.

Lets make a box plot of residual sugar vs quality

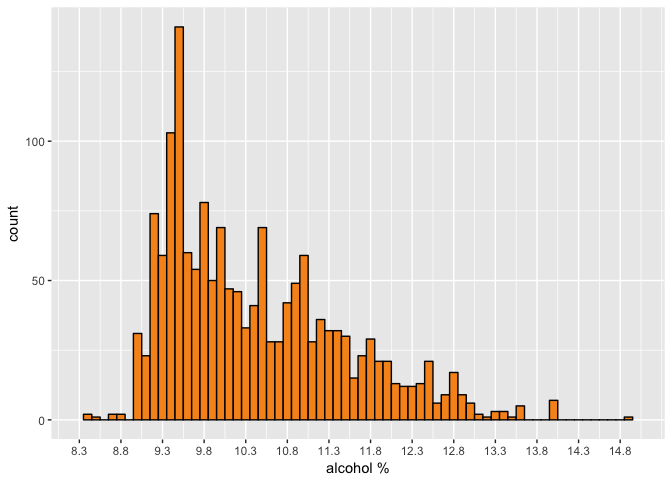


# Conclusion (sugar vs quality)

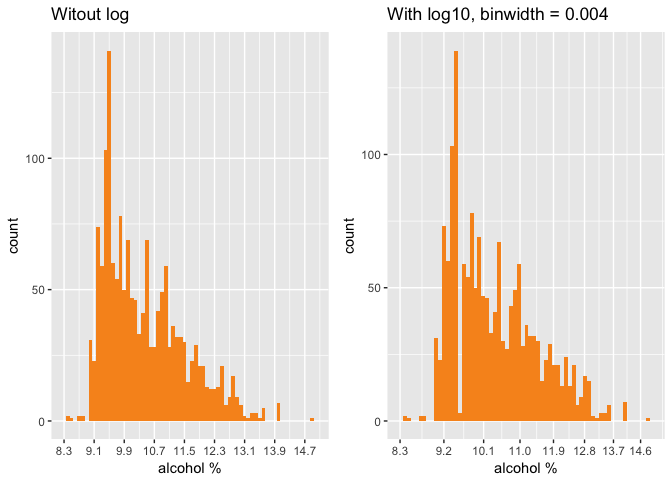
No people gave rating 3, 4 and 8 for high sugar. This could be because number of people who gave rating 3, 4, 8 are very low all together. For sugar less than 6.8 relation between quality vs sugar has similar as the original relation. It looks like residual sugar is not the best to predict the quality.

# Alcohol Study

Lets see the distribution of alcohol

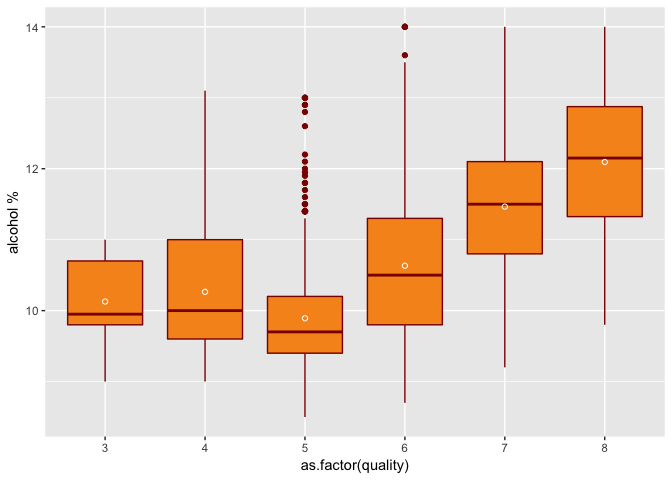


Alcohal distribution is not normal. It has long tail on the right side. Lets see how the log transformation of alcohol looks like

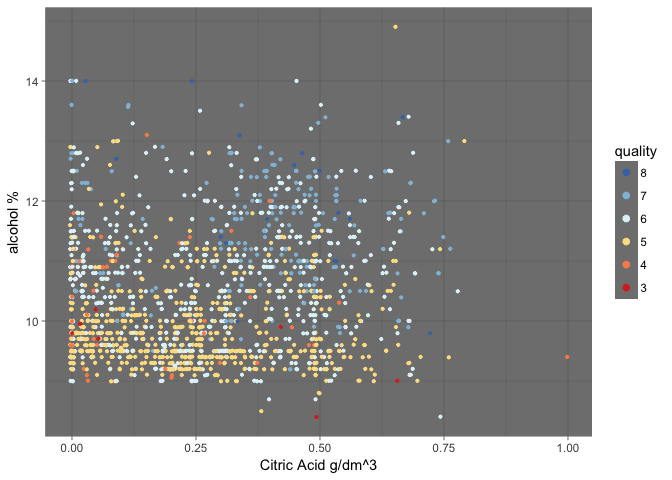


The log10 distribution almost looks similar to non log. I tried various binwidth, I found with the binwidth = 0.004 something different in log10 transformation plot. This might be something we need to consider.

Lets see the box plot of alcohol vs quality

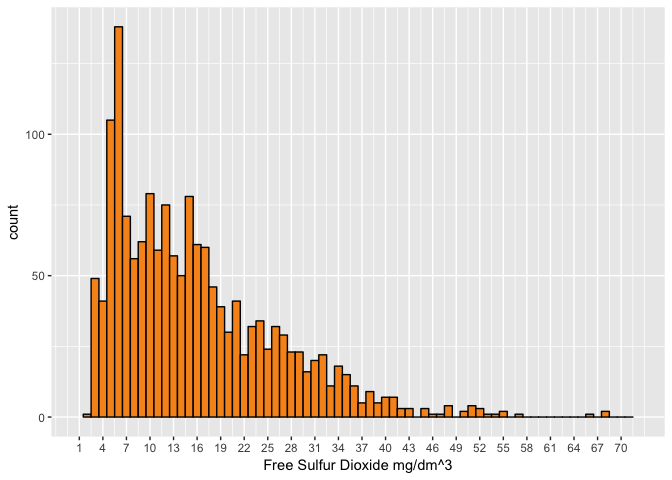


Since I found citric acid and alcohol have same type of trend over quality. I want to see how citric acid and alcohol are related vs quality



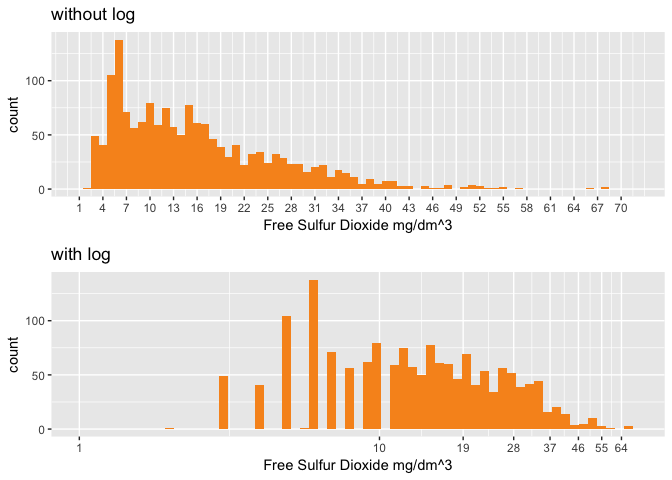
Most of rating 7, 8 are above alcohol value 10 and citric acid between 0.25 to 0.75 Rating 5 is mostly below the alcohol value 11, and bulk of rating 5 falls below alcohol value 10. Rating 6 is randomly distributed throughout the values of alcohol and citric acid. Rating 3 falls below alcohol value 10.

# Free Sulpher dioxide Study



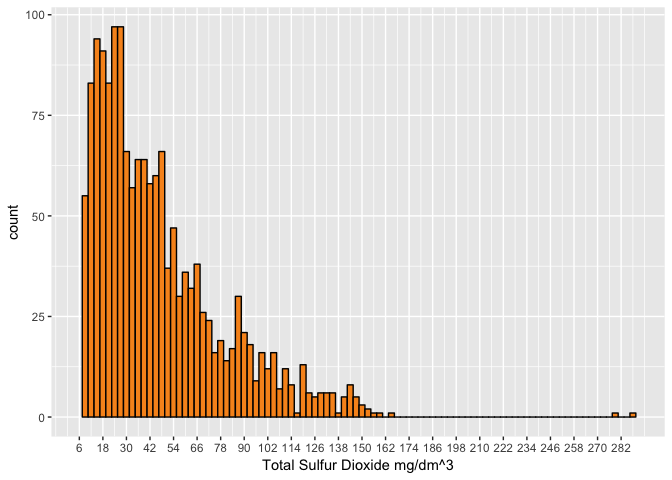
Free sulfur dioxide is not normally distributed. It has long tail on right.

Lets do the log transformation of free sulfur dioxide and see what we can find

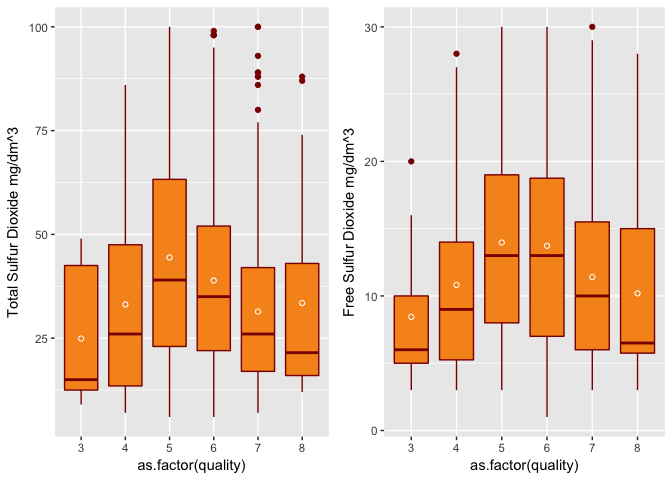


Not showing normality, not so informative.

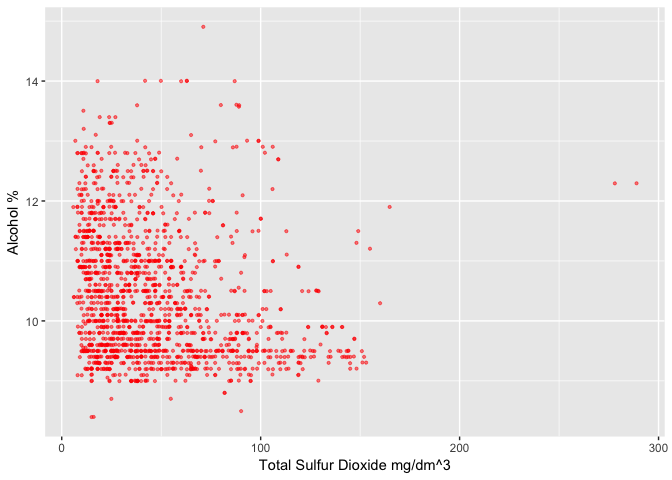
Lets plot the total sulfur dioxide



Plots of sulfur dioxied vs quality



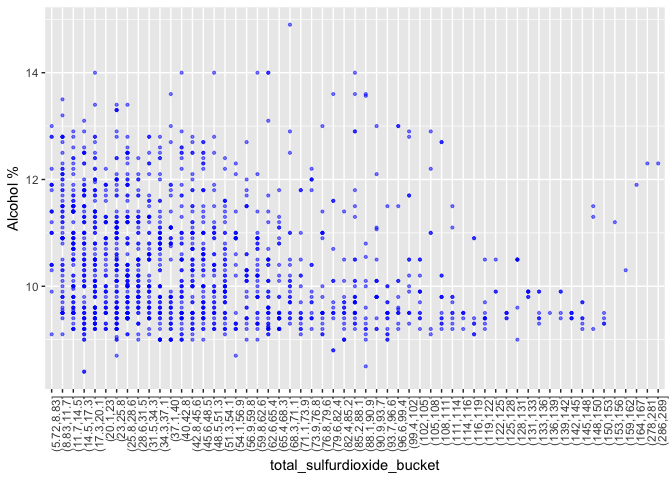
I want to see how total sulfur dioxide relates with alcohol. As I saw kind of opposite relation in box plot



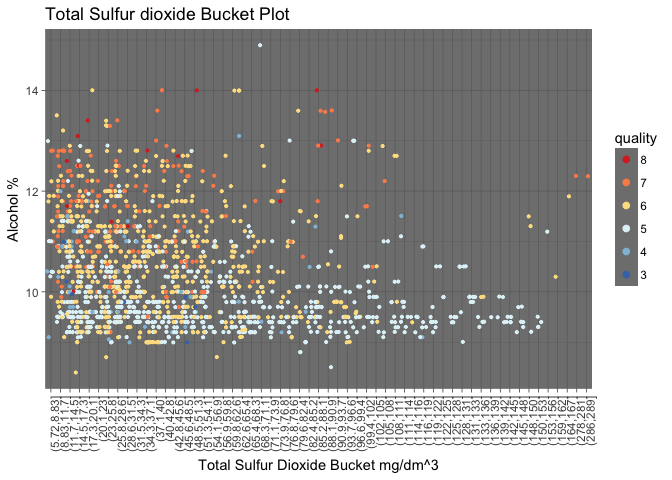
Not getting much information. Lets create the bucket of total sulfur dioxide

RedWineQuality$total\_sulfurdioxide\_bucket <- cut(RedWineQuality$total.sulfur.dioxide,   
 100)

Lets plot the total\_sulfurdioxide\_bucket



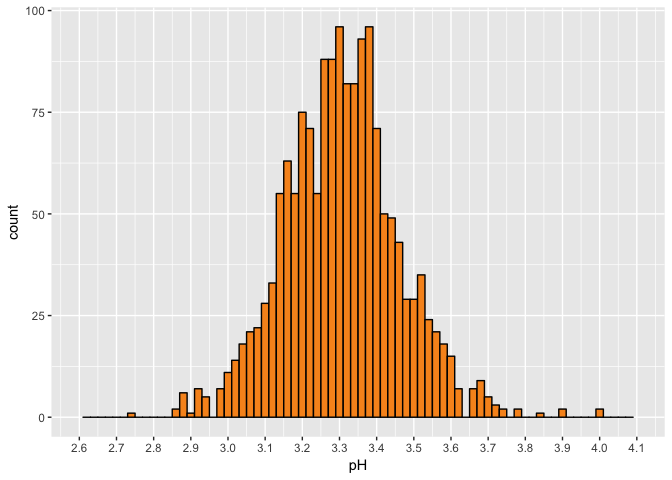
Lets make above plot more nicer and clear using colorbrewer



The interesting fact is for the total.sulfur.dioxide value from 99 to 153 people gave rating 5 except of some outliers. From above plot we see that people rated 5 for alcohol value lower than 11, high for high value of alcohol. Most of the rating 5 falls below the alcohol value 11. Most of the rating 7 lies above the alcohol value 11. Rating 4, 6 are randomly distributed.

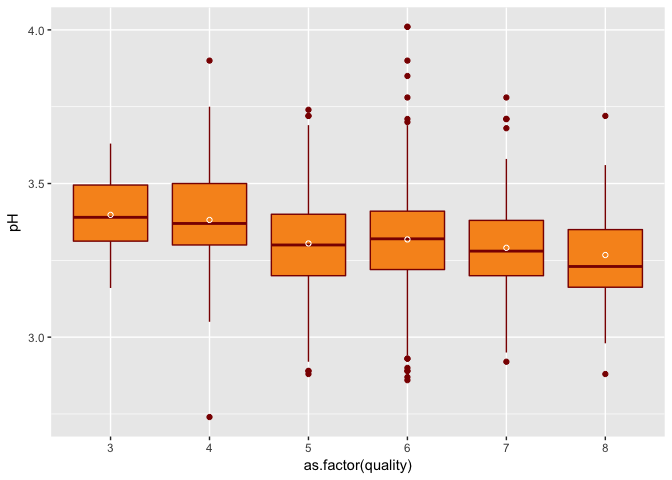
# pH Study

Plotting pH to see its distribution



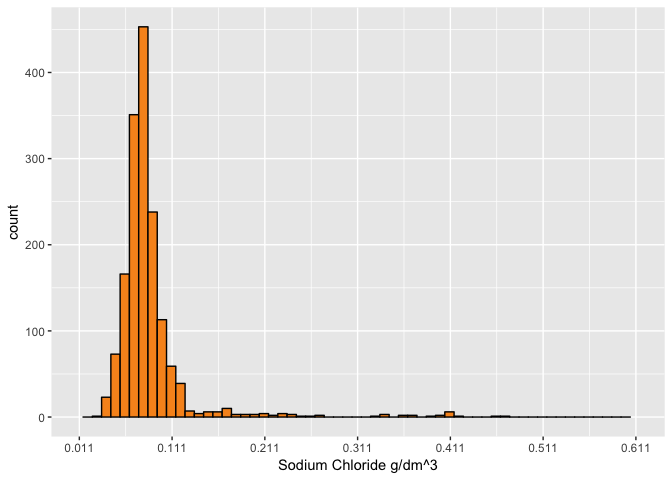
pH is normally distributed.

pH vs quality plot



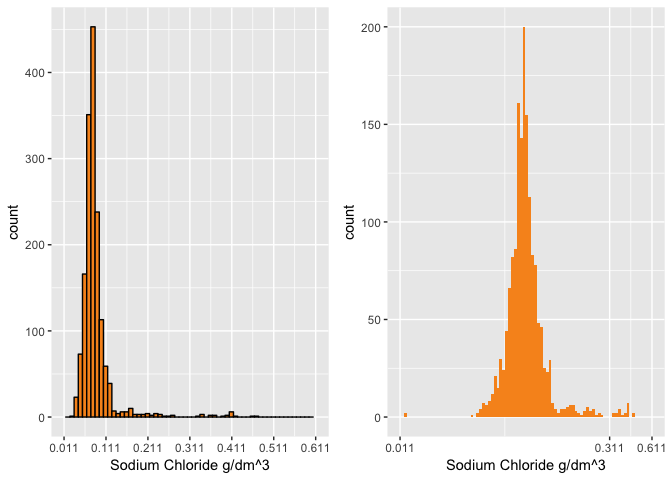
People gave high rating for low value of pH. pH could be helpful for prediction.

# Chlorides Study



Chloride is not normally distributed, it has long tail on the right.

Lets create log10 plot for chloride and compare with non log chloride.

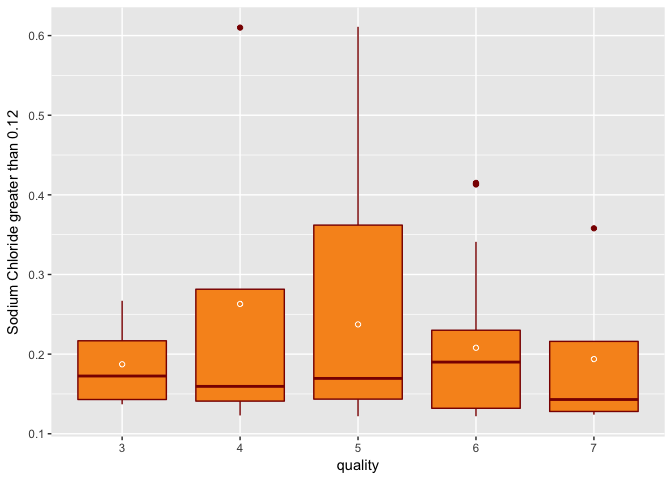


The log transformation of chlorides gave almost normal distribution except of 3 small picks on the right and one negligible peak around chloride value 0.011.

Lets study the distribution of chlorides more. It has long tail on the right. I want to subset chlorides value after 0.121

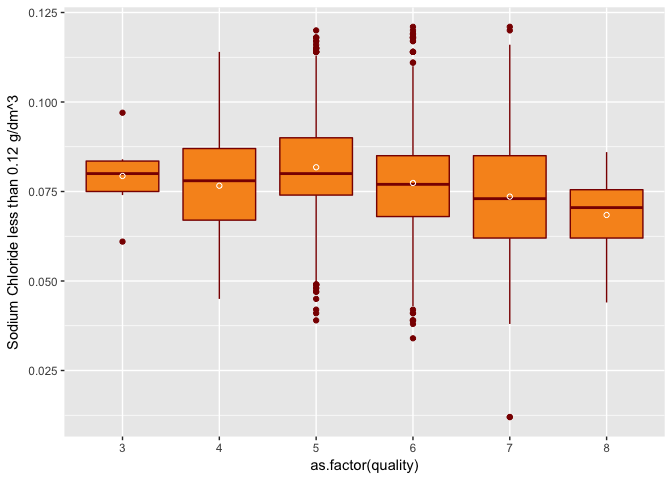
chloridegreater0.12 <- subset(RedWineQuality, chlorides > 0.121)  
chloridelessthn0.12 <- subset(RedWineQuality, chlorides <= 0.121)

Making plot of chlorides greater than 0.12

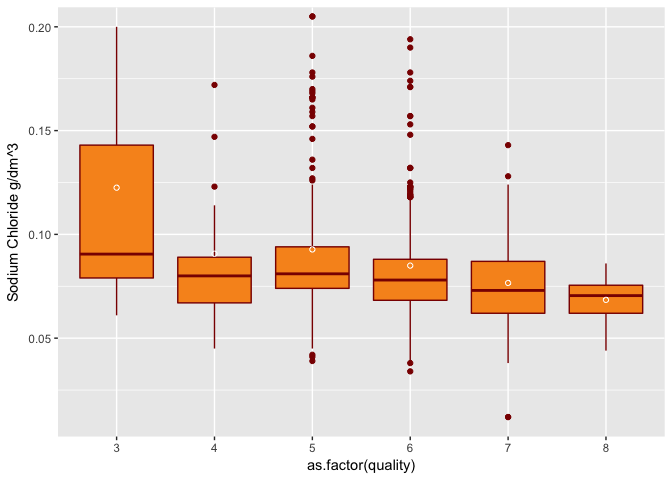


People did not gave rating 8 for chlorides having value greater than 0.12. Large number of people gave rating 5. Few people gave rating 3 and 7.

Making plot of chlorides less then 0.12



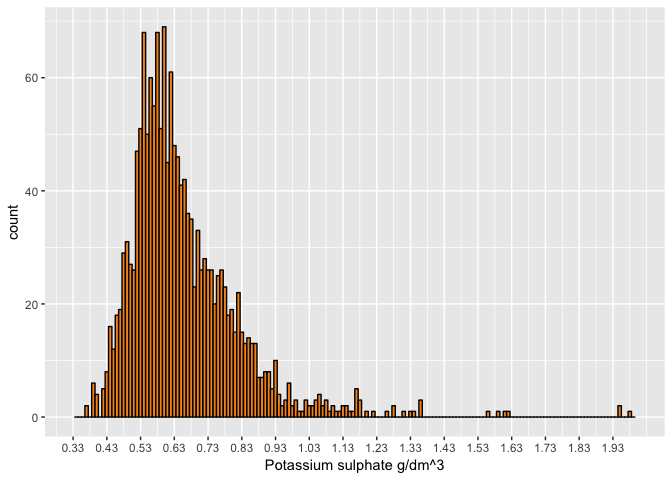
plot of quality vs chloride



# Conclusion (Chlorides)

No people rated 8 for having chloride value greater than 0.121. People rated 8 for lowest range of chloride value than the other rating. Chloride could be a variable for prediction. Because people really not liking the higher value of chloride.

# Sulphates Study

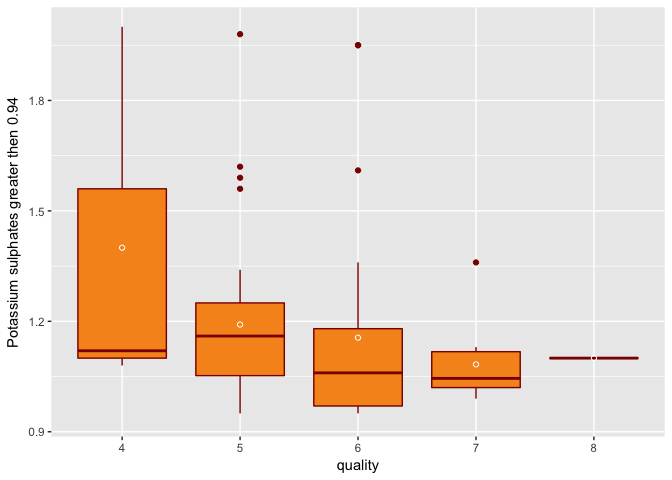


Not normally distributed. Long tail on the right.

Lets explore more about sulphates. I want to see how quality relates with higher values of sulphate. I want to see the subset of sulphates value.

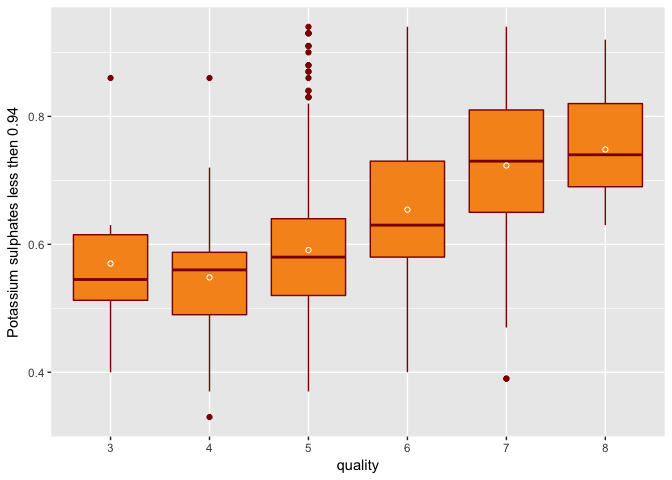
sulphatesgreaterthn\_0.94 <- subset(RedWineQuality, sulphates >   
 0.94)  
sulphateslessthn\_0.94 <- subset(RedWineQuality, sulphates <=   
 0.94)

Plotting sulphates greater than 0.94

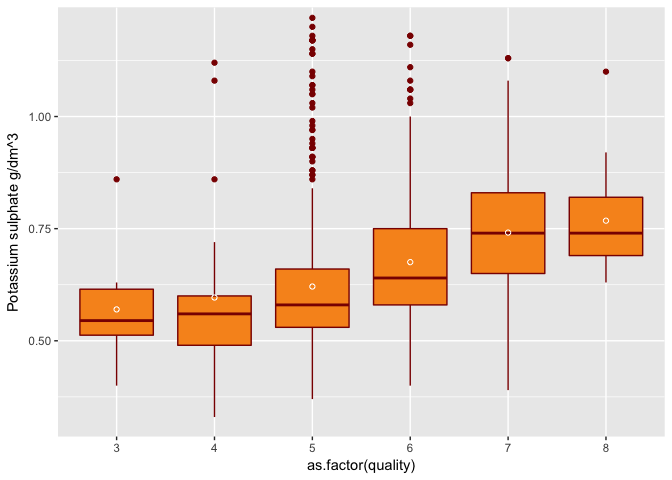


For sulphate value greater than 0.94 people did not give rating 3. May be only one people gave rating 8. Most of the people gave rating 4.

Plotting sulphates less than 0.94



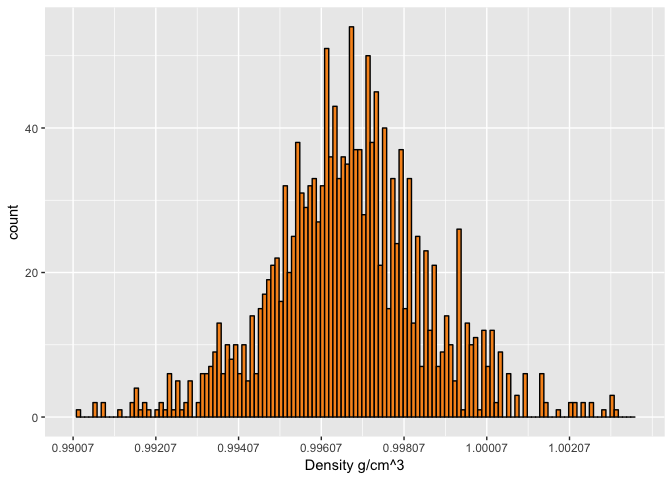
Plotting quality vs sulphates



# Conclusion(sulphates vs quality)

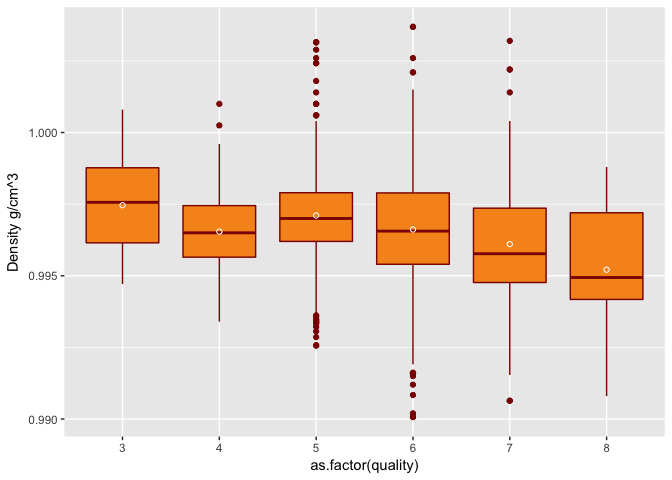
Sulphates greater than 0.94 has no significant contribution quality of wine. For sulphates value lower than 0.94, quality of wine increases with increase in sulphates value.

# Density Study



Density plot looks normally distributed.

Plotting quality vs density



# Conclusion(Density vs quality)

Density could be predictor for quality as it has trend. For higher value of density quality is low and for lower value of density, quality is high.

Let's find out correlation coefficient between quality and other variables. Remove sulfurdioxide\_bucket as it does not necessary for prediction.

# Correlation coefficient

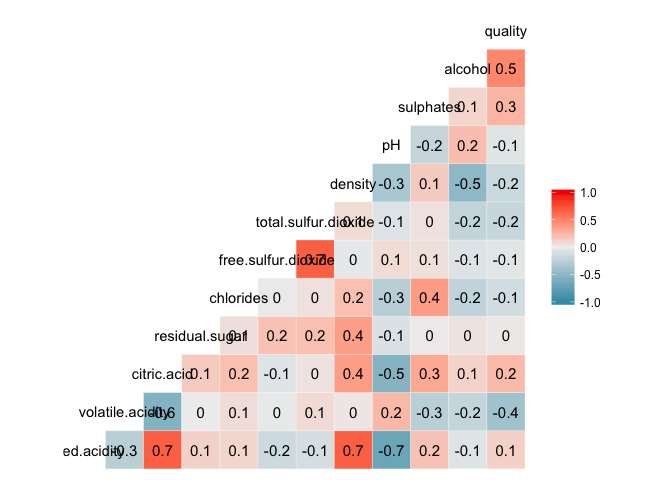
# Removing unneeded columns  
  
x <- subset(RedWineQuality, select = -c(total\_sulfurdioxide\_bucket,   
 quality))  
y <- RedWineQuality$quality  
cor(x, y)

## [,1]  
## fixed.acidity 0.12405165  
## volatile.acidity -0.39055778  
## citric.acid 0.22637251  
## residual.sugar 0.01373164  
## chlorides -0.12890656  
## free.sulfur.dioxide -0.05065606  
## total.sulfur.dioxide -0.18510029  
## density -0.17491923  
## pH -0.05773139  
## sulphates 0.25139708  
## alcohol 0.47616632

Only alcohol seems slighty better than other variables for prediction.

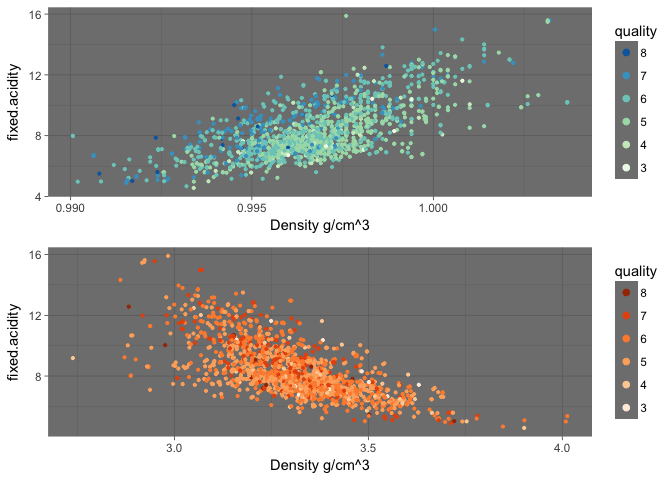
Using ggcorr function from GGally library to see the correlation matrix. Which giving same realtion as we saw above.

x1 <- subset(RedWineQuality, select = -c(total\_sulfurdioxide\_bucket))  
library(GGally)  
ggcorr(x1, label = TRUE)



# Bar plots

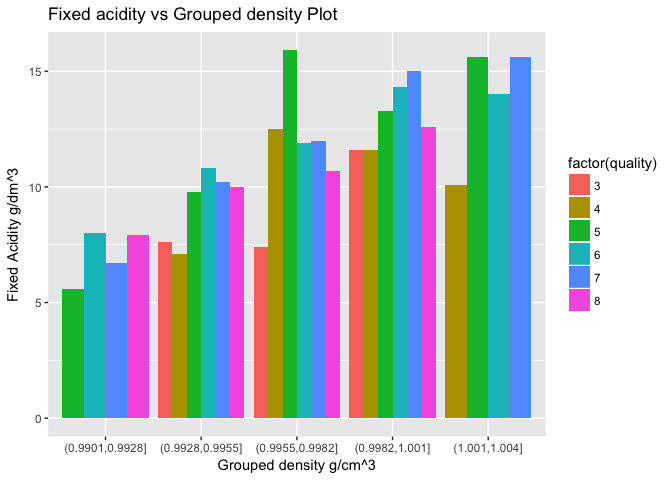
From above correlation matrix it is seen that fixed acidity and density are +vely strongly related. pH and fixed acidity are -Vely strongly correlated. So I want to see their line graph and how they relates with quality. I am inserting quality in everywhere because I am looking how quality of wine influenced by other variables.



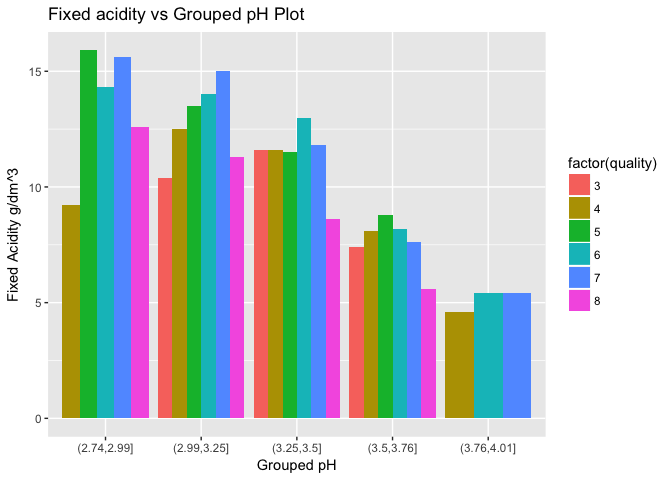
What is +ve and -ve correlation mean here? In +ve correlation : Value of fixed acidity increasing with increasing value of density In -ve correlation : Value of fixed acidity increasing with decreasing value of pH

Lets make bins of density and pH and plot them

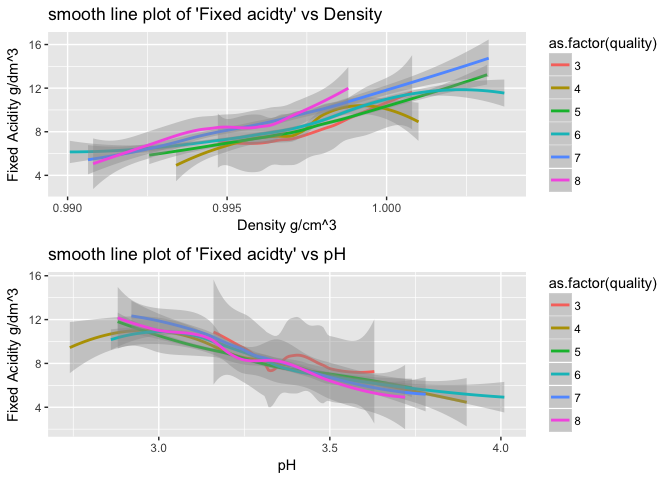
RedWineQuality$density\_bins <- cut(RedWineQuality$density, 5)  
RedWineQuality$pH\_bins <- cut(RedWineQuality$pH, 5)



No 3,4 rating for low value of acidtity and fixed acidity. No 3, 8 rating for high value of density and Fixed acidity. More people gave rating 5 for high value of Fixed acidity and middle value of density.



No 3, 5, 8 rating for high value of pH. If we consider rating 4 is outlier and keeping in mind that count of people for rating 8 is low, we could say that people liked wine having low value of fixed acidity and high value of pH.



It is seen from above plots that quality 3 requires highest value of confidence interval for prediction.

# Question.

Can I increase linearity if I transfer the variables? Lets see.

log10\_residual\_sugar <- log10(RedWineQuality$residual.sugar)  
a <- (RedWineQuality$quality)  
cor(a, log10\_residual\_sugar)

## [1] 0.02353331

cor(RedWineQuality$quality, RedWineQuality$residual.sugar)

## [1] 0.01373164

log10\_chlorides <- log10(RedWineQuality$chlorides)  
a <- (RedWineQuality$quality)  
cor(a, log10\_chlorides)

## [1] -0.17614

cor(RedWineQuality$quality, RedWineQuality$chlorides)

## [1] -0.1289066

log10\_sulphates <- log10(RedWineQuality$sulphates)  
a <- (RedWineQuality$quality)  
cor(a, log10\_sulphates)

## [1] 0.3086419

cor(RedWineQuality$quality, RedWineQuality$sulphates)

## [1] 0.2513971

What happened when I did the log10 transformation of other variables.

1. I found decrease in R for doing log10 of quality and fixed.acidity.
2. Same for volatile aciditly.
3. I found R low for removing citric acid equals to zero and doing log10 transformation.
4. R for log10\_residual\_sugar is slightly greater.
5. I found log\_10\_chlorides more predictive than chlorides.
6. R for free sulfur dioxide remainded same.
7. Remained almost same for total sulfur dioxide.
8. R remained same for density after transformation.
9. Remained same for pH.
10. R increases with log10 in sulphates.
11. Remained same for alcohol.

# Predictive modeling (Linear modeling for quality)

# Data partition

library(caret)

## Loading required package: lattice

library(lattice)  
set.seed(1234)  
trainIndex <- createDataPartition(RedWineQuality$quality, p = 0.8,   
 list = FALSE)  
head(trainIndex)

## Resample1  
## [1,] 1  
## [2,] 2  
## [3,] 4  
## [4,] 8  
## [5,] 9  
## [6,] 10

train\_data <- RedWineQuality[trainIndex, ]  
test\_data <- RedWineQuality[-trainIndex, ]

# Fit Linear model on the traingin dataset

Since log10 transformation of sugar, chloride and sulphates have slightly higher correlation coeff with quality. So make the model accordingly

fit <- lm(quality ~ fixed.acidity + volatile.acidity + citric.acid +   
 I(log10(residual.sugar)) + I(log10(chlorides)) + free.sulfur.dioxide +   
 total.sulfur.dioxide + density + pH + I(log10(sulphates)) +   
 alcohol, data = train\_data)  
summary(fit)

##   
## Call:  
## lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +   
## I(log10(residual.sugar)) + I(log10(chlorides)) + free.sulfur.dioxide +   
## total.sulfur.dioxide + density + pH + I(log10(sulphates)) +   
## alcohol, data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.41068 -0.37726 -0.04284 0.44787 1.87633   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.037e+01 2.588e+01 1.174 0.240797   
## fixed.acidity 1.492e-02 2.932e-02 0.509 0.610875   
## volatile.acidity -9.419e-01 1.349e-01 -6.984 4.60e-12 \*\*\*  
## citric.acid -1.704e-01 1.600e-01 -1.065 0.286933   
## I(log10(residual.sugar)) 2.250e-01 1.665e-01 1.351 0.176910   
## I(log10(chlorides)) -6.002e-01 1.505e-01 -3.988 7.04e-05 \*\*\*  
## free.sulfur.dioxide 3.291e-03 2.410e-03 1.366 0.172277   
## total.sulfur.dioxide -3.018e-03 8.027e-04 -3.760 0.000178 \*\*\*  
## density -2.583e+01 2.632e+01 -0.981 0.326578   
## pH -4.896e-01 2.187e-01 -2.239 0.025334 \*   
## I(log10(sulphates)) 1.855e+00 2.245e-01 8.265 3.49e-16 \*\*\*  
## alcohol 2.654e-01 3.166e-02 8.382 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6511 on 1269 degrees of freedom  
## Multiple R-squared: 0.3674, Adjusted R-squared: 0.3619   
## F-statistic: 66.99 on 11 and 1269 DF, p-value: < 2.2e-16

# Stepwise variable selection

I am using the stepAIC() method in MASS package to do the variable selection. The stepAIC() method performs stepwise model selection by AIC( Akaike information criterion, a measure used to evalluate the relative quality of statistical models)

library(MASS)  
# Perform stepwise model selection  
step <- stepAIC(fit, direction = "both")

## Start: AIC=-1087.21  
## quality ~ fixed.acidity + volatile.acidity + citric.acid + I(log10(residual.sugar)) +   
## I(log10(chlorides)) + free.sulfur.dioxide + total.sulfur.dioxide +   
## density + pH + I(log10(sulphates)) + alcohol  
##   
## Df Sum of Sq RSS AIC  
## - fixed.acidity 1 0.1098 538.15 -1089.0  
## - density 1 0.4084 538.45 -1088.2  
## - citric.acid 1 0.4812 538.53 -1088.1  
## - I(log10(residual.sugar)) 1 0.7740 538.82 -1087.4  
## - free.sulfur.dioxide 1 0.7908 538.84 -1087.3  
## <none> 538.04 -1087.2  
## - pH 1 2.1254 540.17 -1084.2  
## - total.sulfur.dioxide 1 5.9947 544.04 -1075.0  
## - I(log10(chlorides)) 1 6.7432 544.79 -1073.2  
## - volatile.acidity 1 20.6835 558.73 -1040.9  
## - I(log10(sulphates)) 1 28.9611 567.01 -1022.0  
## - alcohol 1 29.7866 567.83 -1020.2  
##   
## Step: AIC=-1088.95  
## quality ~ volatile.acidity + citric.acid + I(log10(residual.sugar)) +   
## I(log10(chlorides)) + free.sulfur.dioxide + total.sulfur.dioxide +   
## density + pH + I(log10(sulphates)) + alcohol  
##   
## Df Sum of Sq RSS AIC  
## - density 1 0.385 538.54 -1090.03  
## - citric.acid 1 0.387 538.54 -1090.03  
## - I(log10(residual.sugar)) 1 0.669 538.82 -1089.36  
## - free.sulfur.dioxide 1 0.835 538.99 -1088.96  
## <none> 538.15 -1088.95  
## + fixed.acidity 1 0.110 538.04 -1087.21  
## - pH 1 5.970 544.12 -1076.82  
## - total.sulfur.dioxide 1 6.520 544.67 -1075.52  
## - I(log10(chlorides)) 1 7.598 545.75 -1072.99  
## - volatile.acidity 1 20.583 558.74 -1042.87  
## - I(log10(sulphates)) 1 29.471 567.63 -1022.65  
## - alcohol 1 48.644 586.80 -980.09  
##   
## Step: AIC=-1090.03  
## quality ~ volatile.acidity + citric.acid + I(log10(residual.sugar)) +   
## I(log10(chlorides)) + free.sulfur.dioxide + total.sulfur.dioxide +   
## pH + I(log10(sulphates)) + alcohol  
##   
## Df Sum of Sq RSS AIC  
## - I(log10(residual.sugar)) 1 0.327 538.87 -1091.2  
## - free.sulfur.dioxide 1 0.835 539.37 -1090.0  
## <none> 538.54 -1090.0  
## - citric.acid 1 0.945 539.49 -1089.8  
## + density 1 0.385 538.15 -1089.0  
## + fixed.acidity 1 0.087 538.45 -1088.2  
## - pH 1 6.071 544.61 -1077.7  
## - total.sulfur.dioxide 1 6.160 544.70 -1077.5  
## - I(log10(chlorides)) 1 7.617 546.16 -1074.0  
## - volatile.acidity 1 22.895 561.43 -1038.7  
## - I(log10(sulphates)) 1 29.472 568.01 -1023.8  
## - alcohol 1 92.937 631.48 -888.1  
##   
## Step: AIC=-1091.25  
## quality ~ volatile.acidity + citric.acid + I(log10(chlorides)) +   
## free.sulfur.dioxide + total.sulfur.dioxide + pH + I(log10(sulphates)) +   
## alcohol  
##   
## Df Sum of Sq RSS AIC  
## - citric.acid 1 0.801 539.67 -1091.35  
## <none> 538.87 -1091.25  
## - free.sulfur.dioxide 1 0.959 539.83 -1090.97  
## + I(log10(residual.sugar)) 1 0.327 538.54 -1090.03  
## + density 1 0.043 538.82 -1089.36  
## + fixed.acidity 1 0.040 538.83 -1089.35  
## - total.sulfur.dioxide 1 5.942 544.81 -1079.21  
## - pH 1 6.129 545.00 -1078.77  
## - I(log10(chlorides)) 1 7.363 546.23 -1075.87  
## - volatile.acidity 1 22.571 561.44 -1040.69  
## - I(log10(sulphates)) 1 29.192 568.06 -1025.67  
## - alcohol 1 96.502 635.37 -882.23  
##   
## Step: AIC=-1091.35  
## quality ~ volatile.acidity + I(log10(chlorides)) + free.sulfur.dioxide +   
## total.sulfur.dioxide + pH + I(log10(sulphates)) + alcohol  
##   
## Df Sum of Sq RSS AIC  
## <none> 539.67 -1091.35  
## + citric.acid 1 0.801 538.87 -1091.25  
## - free.sulfur.dioxide 1 1.268 540.94 -1090.35  
## + fixed.acidity 1 0.403 539.26 -1090.31  
## + density 1 0.328 539.34 -1090.13  
## + I(log10(residual.sugar)) 1 0.182 539.49 -1089.78  
## - pH 1 5.401 545.07 -1080.59  
## - total.sulfur.dioxide 1 6.879 546.55 -1077.13  
## - I(log10(chlorides)) 1 8.506 548.17 -1073.32  
## - volatile.acidity 1 24.480 564.15 -1036.52  
## - I(log10(sulphates)) 1 28.463 568.13 -1027.51  
## - alcohol 1 96.746 636.41 -882.12

# Show results  
step$anova

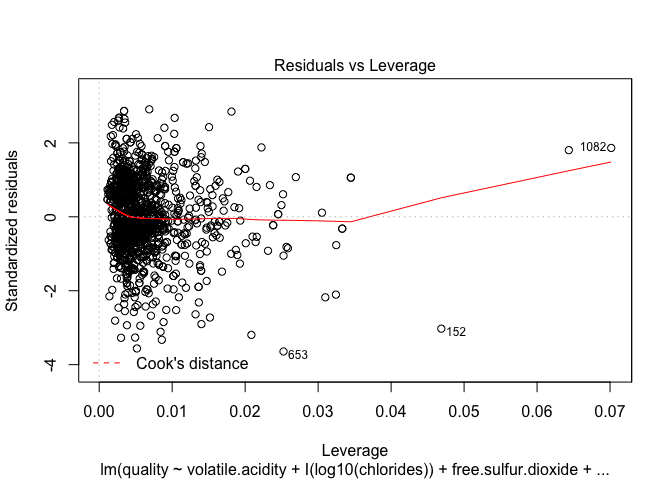
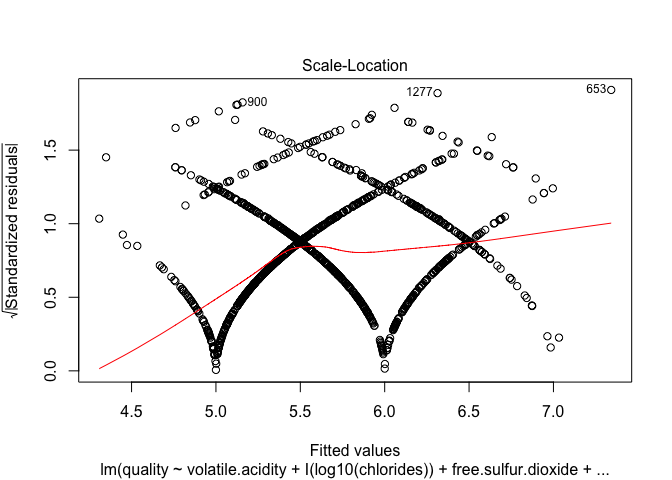
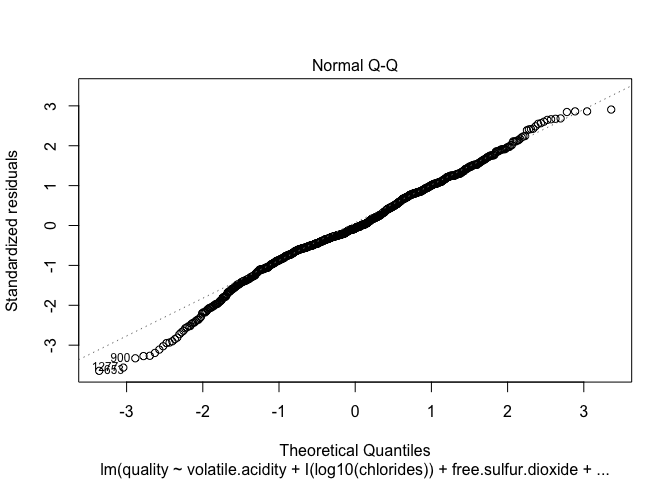
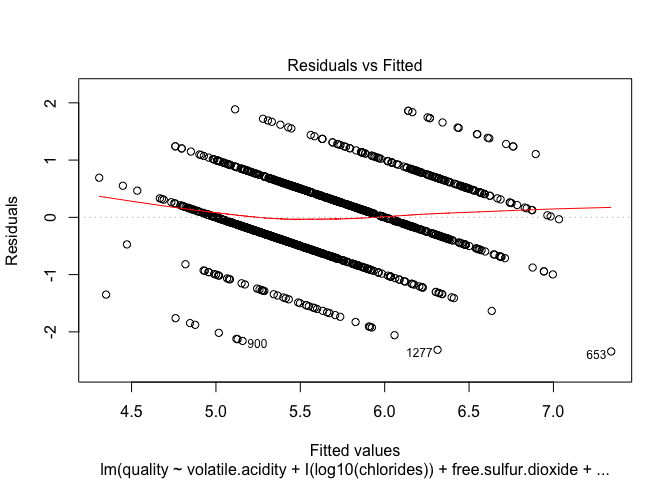
## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## quality ~ fixed.acidity + volatile.acidity + citric.acid + I(log10(residual.sugar)) +   
## I(log10(chlorides)) + free.sulfur.dioxide + total.sulfur.dioxide +   
## density + pH + I(log10(sulphates)) + alcohol  
##   
## Final Model:  
## quality ~ volatile.acidity + I(log10(chlorides)) + free.sulfur.dioxide +   
## total.sulfur.dioxide + pH + I(log10(sulphates)) + alcohol  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 1269 538.0448 -1087.209  
## 2 - fixed.acidity 1 0.1098272 1270 538.1547 -1088.948  
## 3 - density 1 0.3851033 1271 538.5398 -1090.031  
## 4 - I(log10(residual.sugar)) 1 0.3269728 1272 538.8667 -1091.254  
## 5 - citric.acid 1 0.8006678 1273 539.6674 -1091.352

Now I am going to use the model selected by the stepwise selection procedure as the final model

fit\_final <- lm(quality ~ volatile.acidity + I(log10(chlorides)) +   
 free.sulfur.dioxide + total.sulfur.dioxide + pH + I(log10(sulphates)) +   
 alcohol, data = train\_data)  
  
summary(fit\_final)

##   
## Call:  
## lm(formula = quality ~ volatile.acidity + I(log10(chlorides)) +   
## free.sulfur.dioxide + total.sulfur.dioxide + pH + I(log10(sulphates)) +   
## alcohol, data = train\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.34228 -0.36872 -0.04523 0.46151 1.88643   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.4058975 0.4182231 10.535 < 2e-16 \*\*\*  
## volatile.acidity -0.8632808 0.1136034 -7.599 5.76e-14 \*\*\*  
## I(log10(chlorides)) -0.6403236 0.1429486 -4.479 8.16e-06 \*\*\*  
## free.sulfur.dioxide 0.0040877 0.0023637 1.729 0.083979 .   
## total.sulfur.dioxide -0.0030665 0.0007612 -4.028 5.95e-05 \*\*\*  
## pH -0.4813264 0.1348439 -3.570 0.000371 \*\*\*  
## I(log10(sulphates)) 1.7382025 0.2121316 8.194 6.10e-16 \*\*\*  
## alcohol 0.2878693 0.0190559 15.107 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6511 on 1273 degrees of freedom  
## Multiple R-squared: 0.3655, Adjusted R-squared: 0.362   
## F-statistic: 104.7 on 7 and 1273 DF, p-value: < 2.2e-16

# Plot fit\_final  
plot(fit\_final)



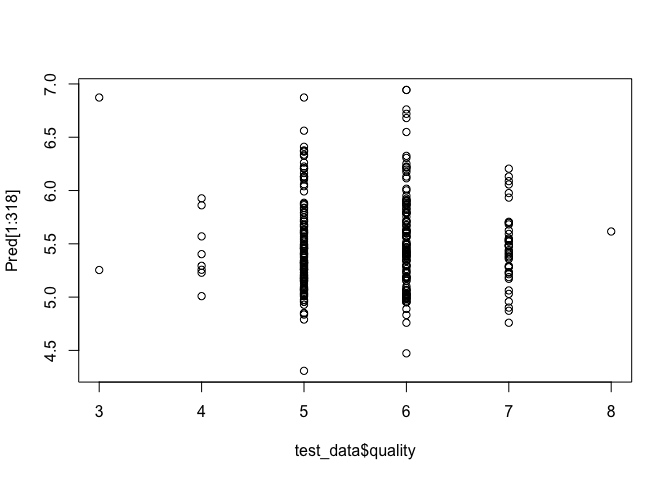
# Evaluate Predictive performance of the final model

# Predict the rating on testing set  
nrow(test\_data)

## [1] 318

Pred <- predict(fit\_final, data = test\_data)

# Plot the Rating vs Predicted rating  
plot(test\_data$quality, Pred[1:318])

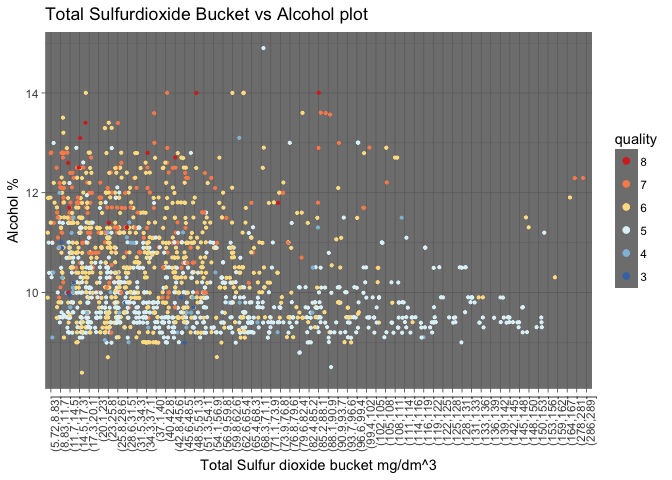


The test data predicted all quality well.

# Three final plots

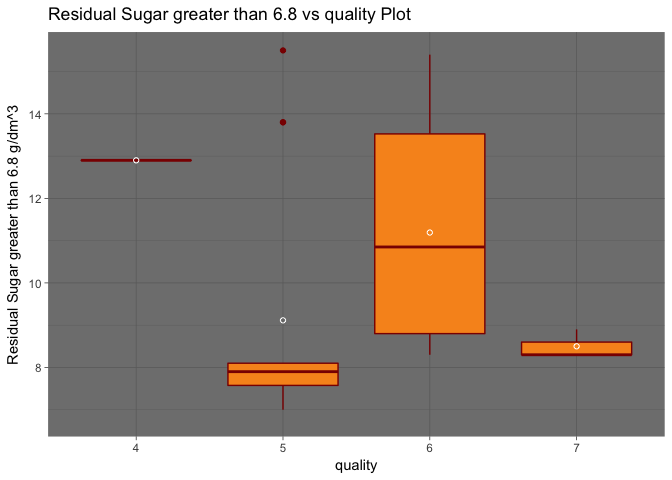
These are the three plots which caught my eyes with some interesting informations.

# First Plot



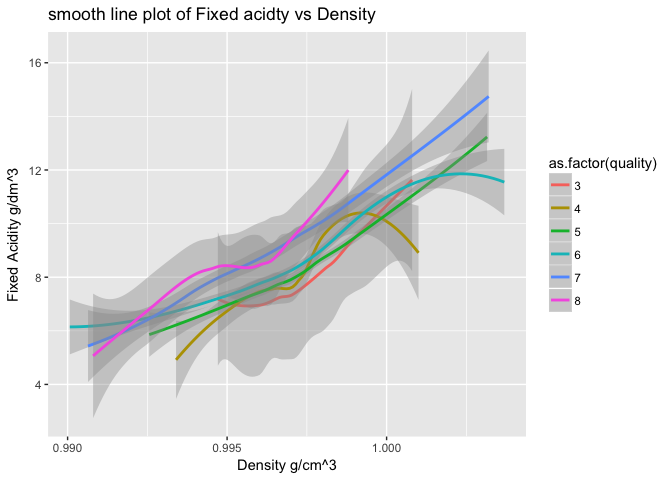
The above plot giving me the information that for the total sulfur dioxide value greater than 105, people gave rating only 5 except of some outliers of 6 and 7 rating. Which is quite interesting.

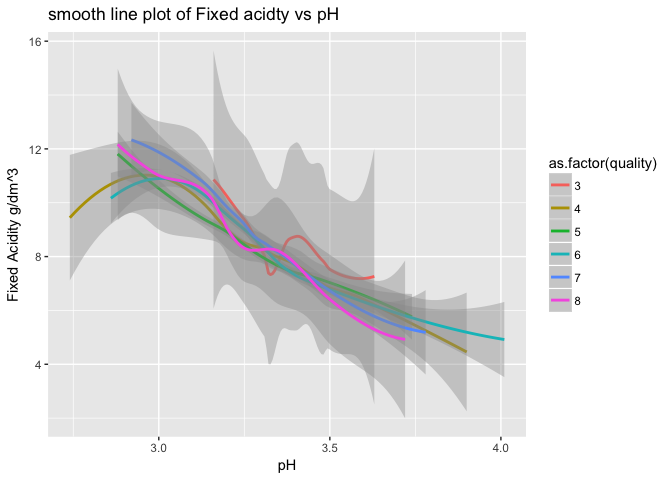
# Second Plot



Nobody gave rating 3 and 8 for residual sugar value greater than 6.8. Only one people gave rating 4 for residual sugar value greater than 6.8. Three people gave raing rating 7. Range of the sugar value for rating 6 would be higher than all others if we remove two outliers of rating 5. The mean of residual sugar for rating 5 is out of the box because of presence of two outliers.

# Third Plots





The above 2 smoothed plots were interesting to me because those were providing tentetive idea how smoothed data of fixed acidity vs density and fixed acidity vs pH are helpful to predict the quality of wine. But interesting thing is that my modedl only selected pH for quality prediction. While exploring density I was thinking that density could be the good predictor but my model disproved that.

# Reflection

For the whole data set most of the people gave rating 5 and 6. Nobody gave rating 0, 1, 2, 9, 10. This might be because most of the people randomly choose the rating 5 and 6. And surprisingly no body rated 9 and 10 means the wine quality might not be good in reality.

I first thought that acidity has predictive capability. As quality increases with increase value of citric acid and decreases with increased value of volatile acidity.

For residual sugar nobody gave rating 3 and 8 for the value greater than 6.8. May be only one people gave rating 4 for residual sugar value greater than 6.8. Most of the rating 5 falls below the alcohol value 11. Most of the rating 7 lies above the alcohol value 11. Rating 4, 6 are randomly distributed.

The interesting fact is for the total.sulfur.dioxide value from 99 to 153 people gave rating 5 except of some outliers.

People gave high rating for low value of pH.

No people rated 8 for having chloride value greater than 0.121.

For sulphate value greater than 0.94 people did not give rating 3. May be only one people gave rating 8. Most of the people gave rating 4.

Density showed predictor for quality as it has trend. For higher value of density quality is low and for lower value of density, quality is high.

The linear model gave me seven final variabes (volatile.acidity, log10(chlorides), free.sulfur.dioxide, total.sulfur.dioxide, pH, log10(sulphates), alcohol) for prediction of quality of wine.

But it is not the final conclusion. There might be other variables(which are not present in our data) we need to consider for wine quality prediction.

# Alternative way of analysis

We could go with the other way of analysis by dividing the quality into three groups eg: low, medium and high