Exploratory Data Analysis on Quality of Red Wine

Sai Saketh Boyanapalli

October 15, 2017

RedWineQualityDataSet: This data set contains data on 1599 samples of red wines, there chemical Properties and quality. The inputs include objective tests and the output is based on sensory data from a wine expert(median of at least 3 evaluations from expert are taken)

I want to perform a exploratory data analysis on this data set to find interesting relationships between the different chemical properties and how they all relate to quality of the Red wine.

Univariate Analysis

```
str(redWine) # a look at structure of the data
```

```
## '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 ...
```

```
names(redWine) # different variables in our data set.
```

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

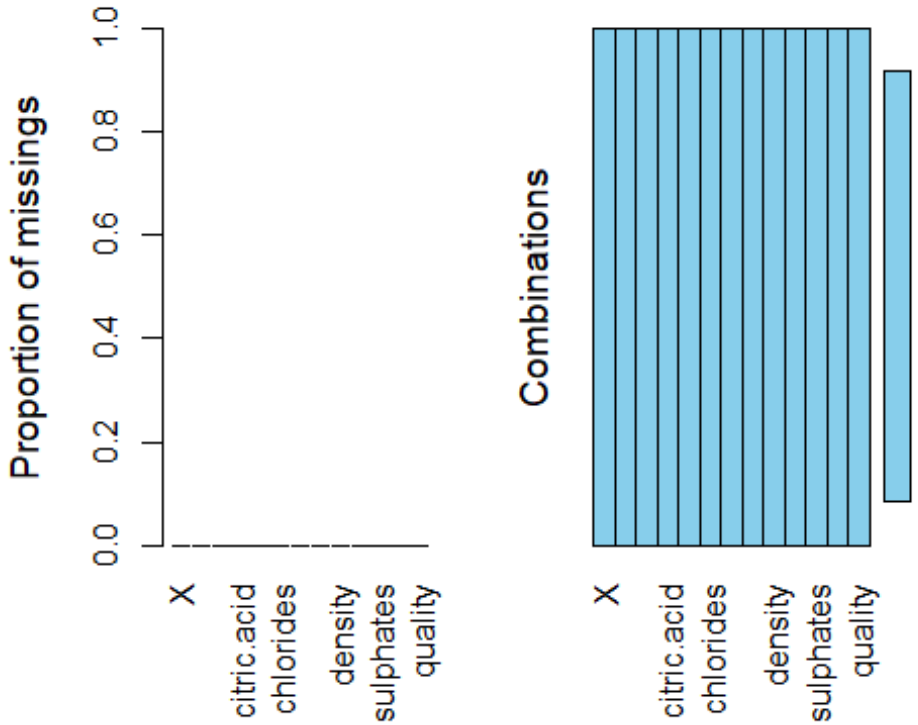
```
dim(redWine) # dimensions of our data.
```

```
## [1] 1599   13
```

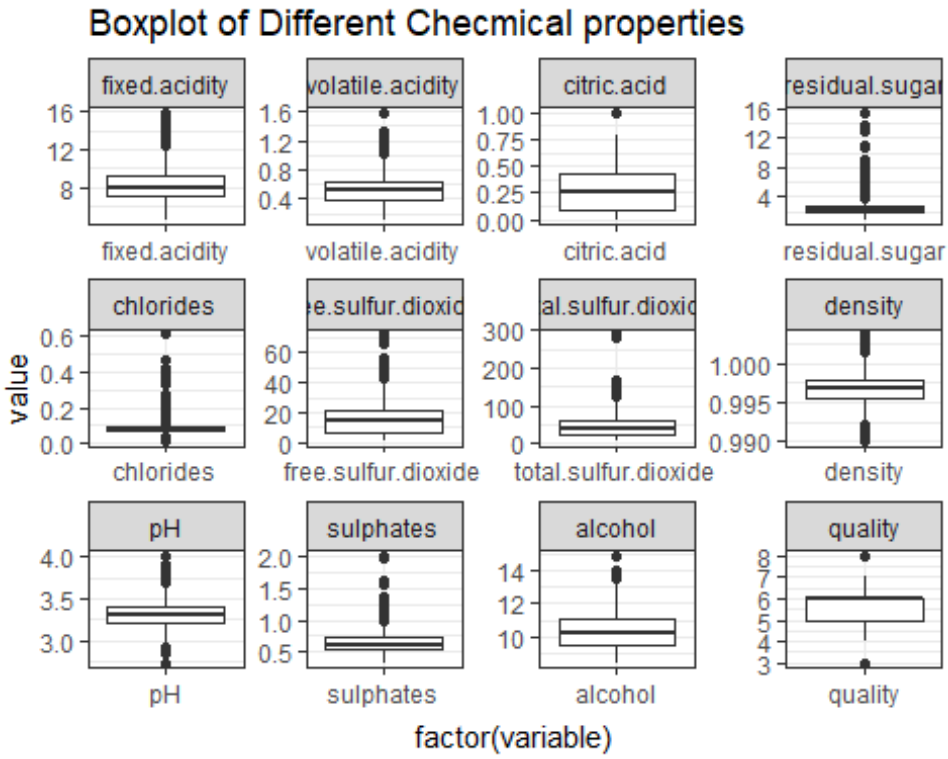
```
summary(redWine) # a brief summary of all variables in the data set.
```

```
##        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
```

```
aggr(redWine) # A look at missigness in our data (function taken from VIM
library)
```

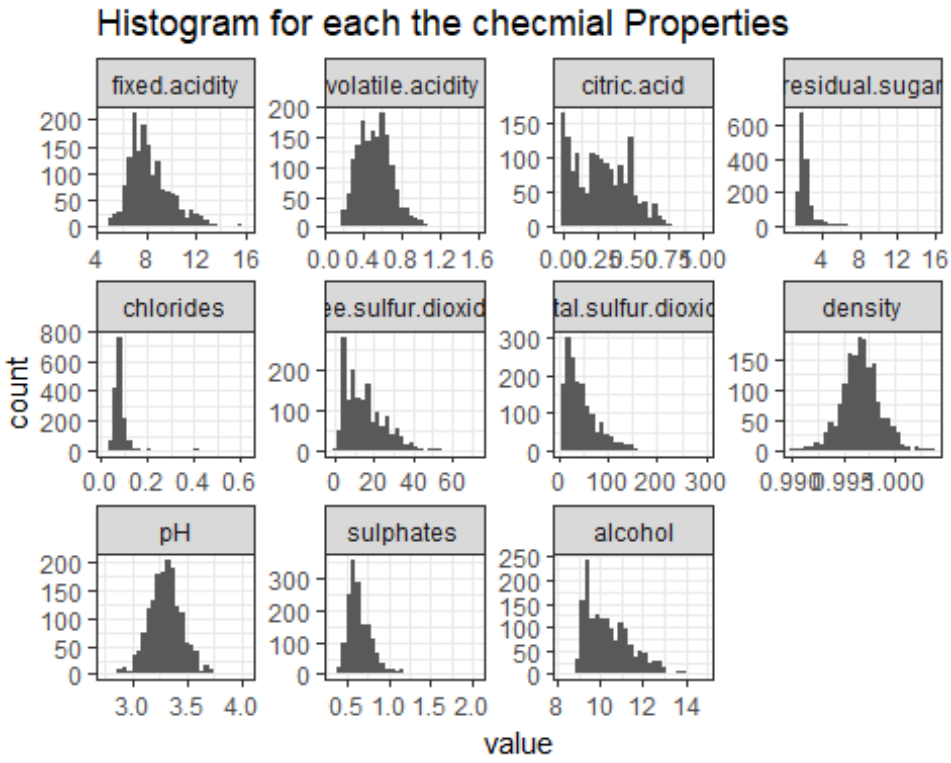


We can see that our data set is complete.



This boxplot is a good start to outlier identification we can see that few variables have lot of ouliers and some of them have only few outliers. we can get a better understanding by taking a look at each variable.

A look at how the data is distributed for each of the chemical properties

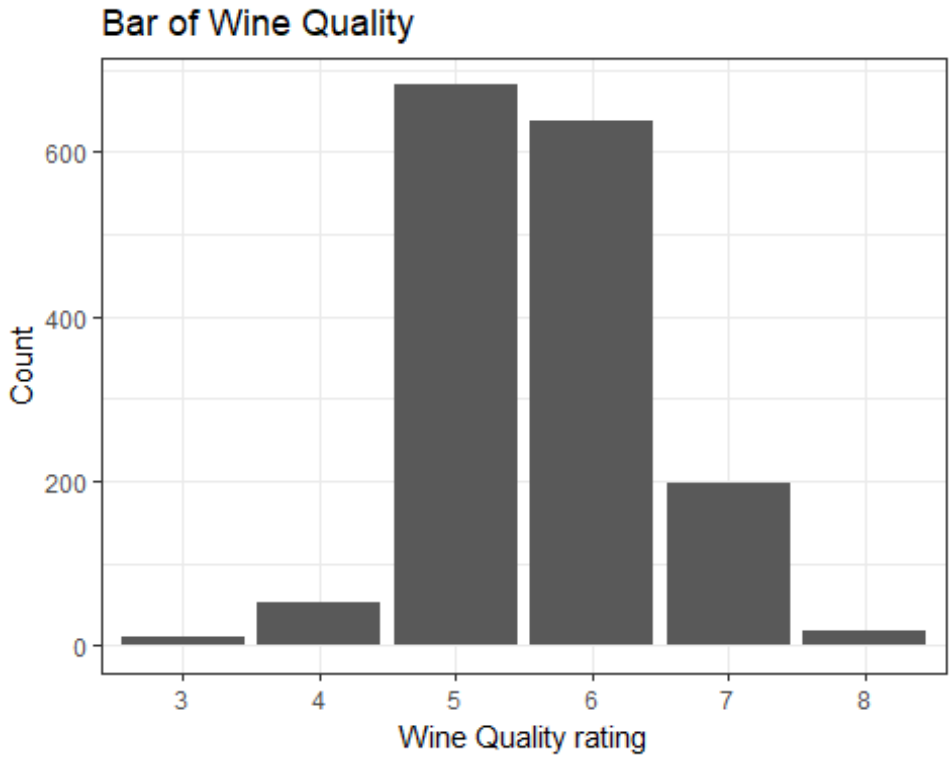


As indicated by the boxplot of chemical properties the plots with most outliers have skewed distributions in case of histogram.

if we look at histograms of **fixed Acidity, Density, PH, sulphates, volatile Acidity** they all follow distributions close to normal. and have very few outliers so, we can ignore those outliers for now.

Now we are left with few chemical properties which might need a closer look.

Exploring Quality




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

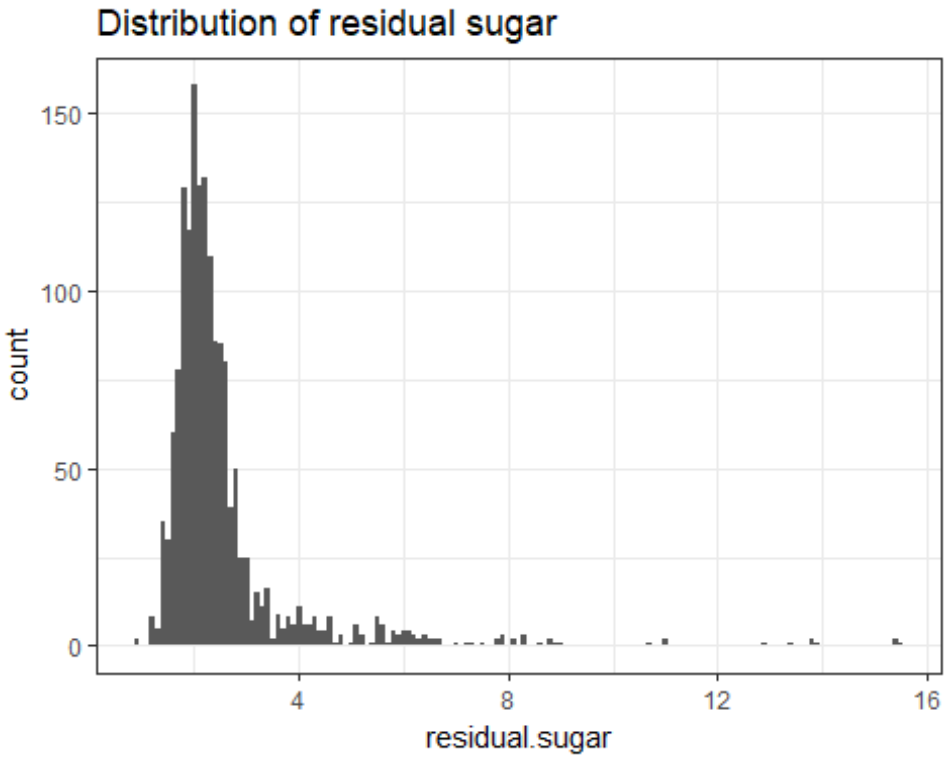
We can see that minimum rated quality is 3 in our data set and the best quality is 8, these were indicated as outliers in the above boxplot which means only few of the red wines received low quality and high quality rating and most of the ratings are in the middle. And another interesting finding is none of them received Highest quality rating - 10 or lowest quality rating - 0.

lets create factors for quality this might be useful later.

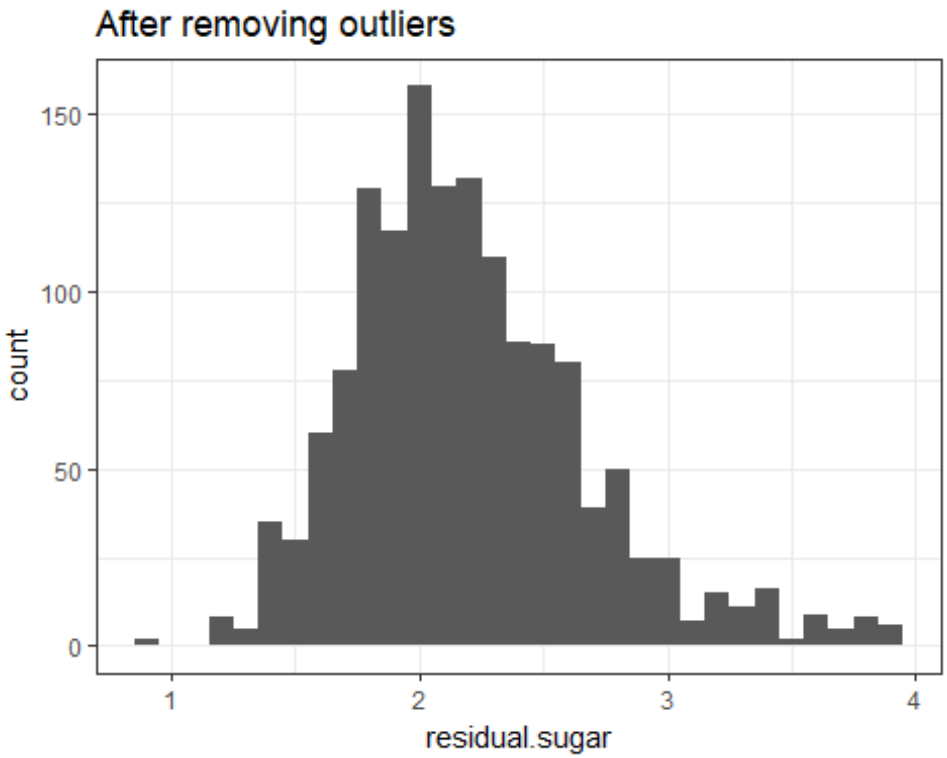
```
redWine$qualityFactor <- factor(redWine$quality) # created new variable in
data frame for quality factors.
```

Let's look at other variables that are not normal.

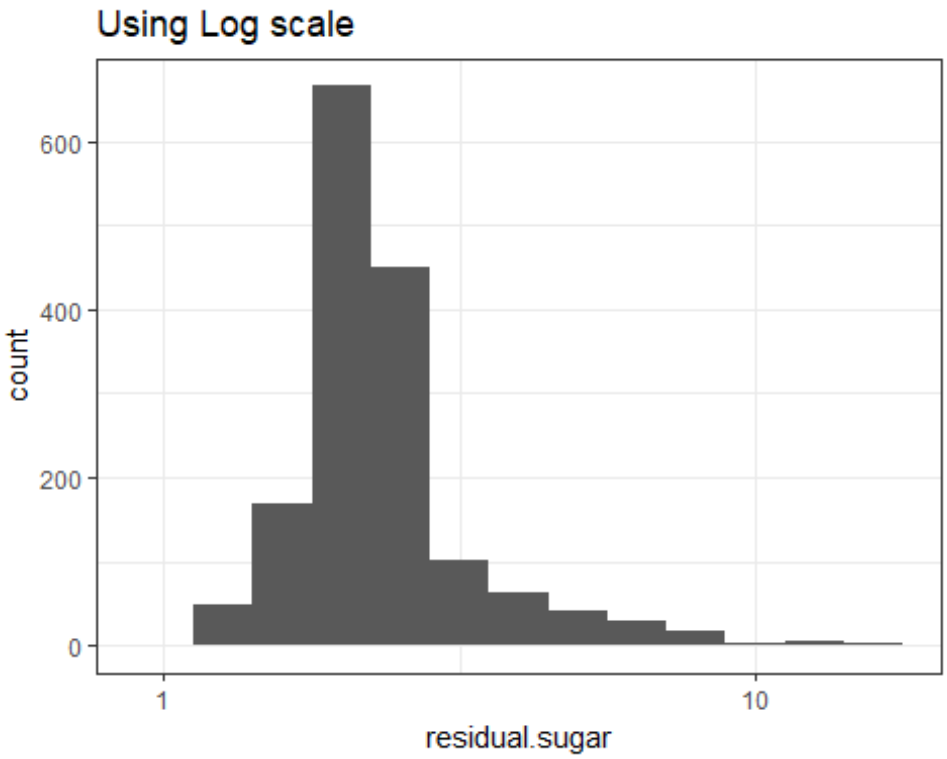
Exploring residual.sugar



The distribution of **Residual.sugar**, the distribution is skewed to left and there are many outliers to the right.

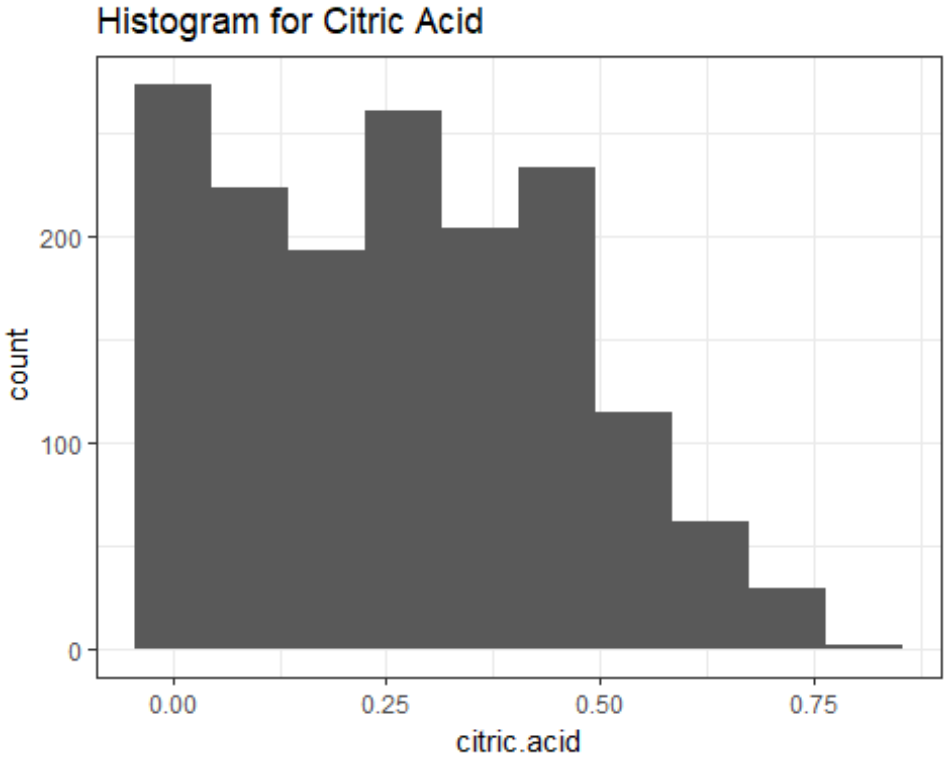


removing all the residual.sugar > 4 from the data the histogram changes from being skewed distibution to a normal distribution. We only have few samples with residual sugar content greater than 4 this might be due to less preference of sweety alcohol. This will be useful when modelling.



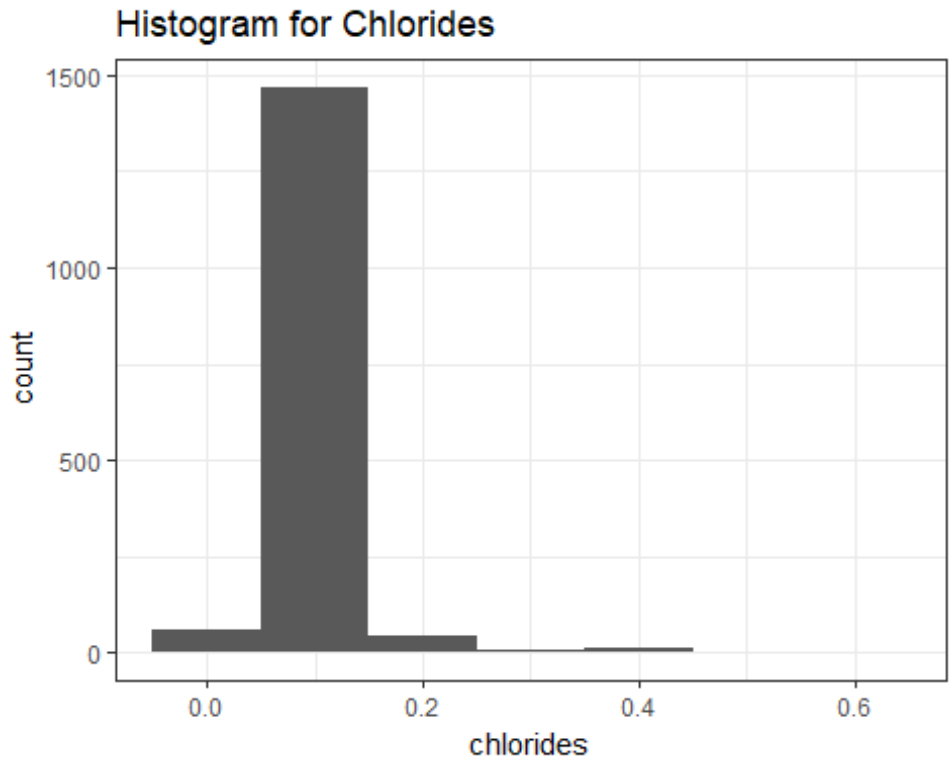
Applying log scale also yeilds close to normal distribution of data.

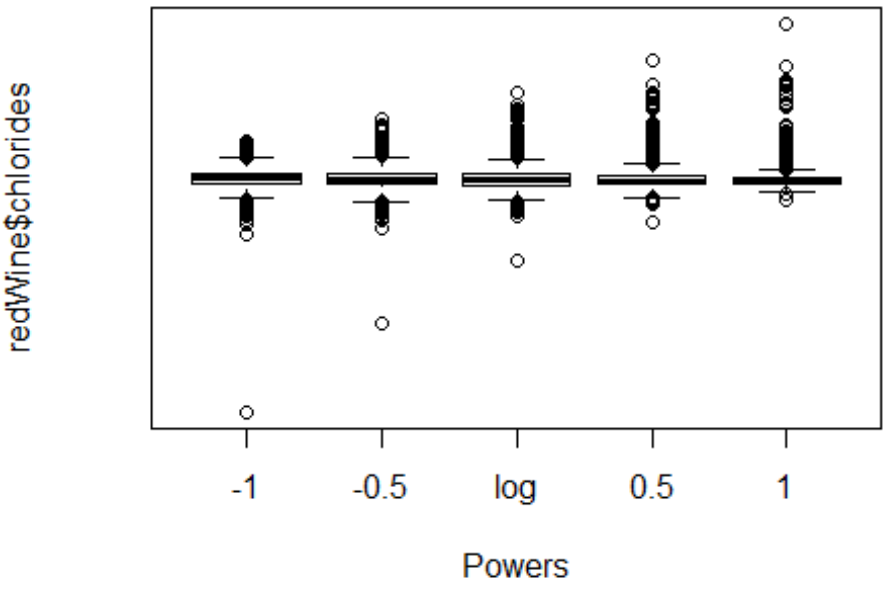
Exploring citric acid



After removing outliers the distribution of citric acid is not normal. but instead close to uniform.

Exploring Chlorides

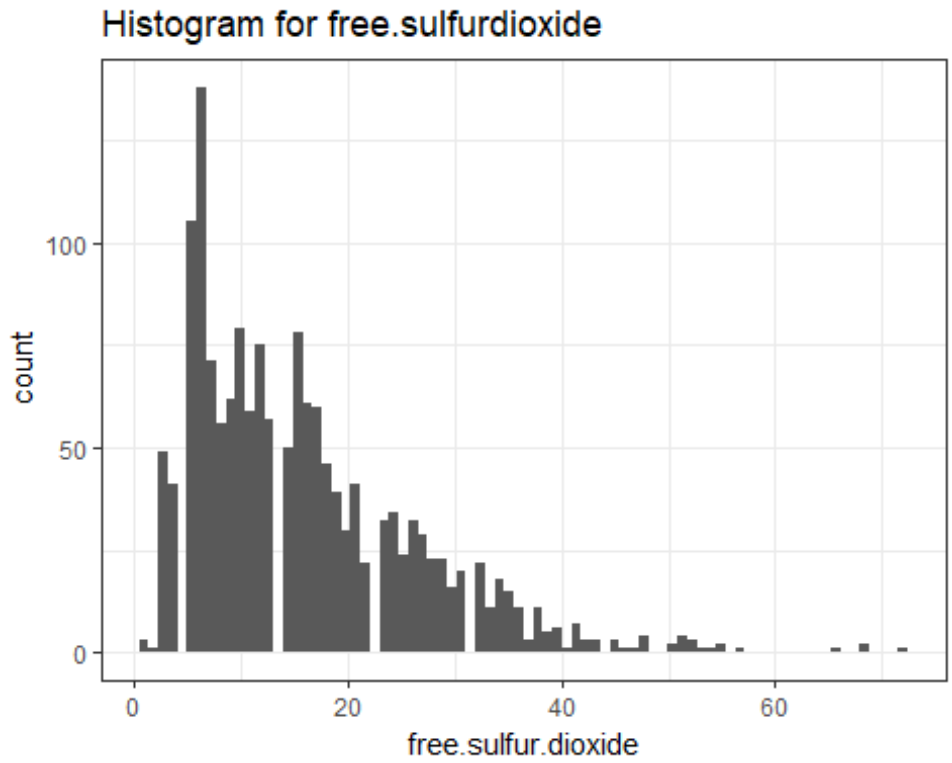




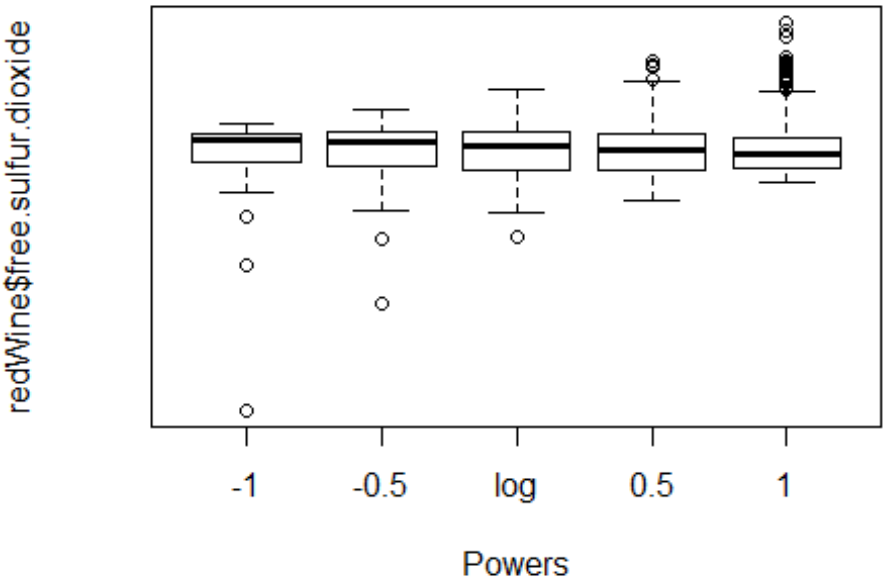
This symbox fuction will help us decide which powers will help us get closer to normal distribution for our data. if we look at the output of symbox fuction, the boxplot with few outliers will give us a power transformation which will distribute our data close to normal.

for chlorodes log transformation might yeild best result.

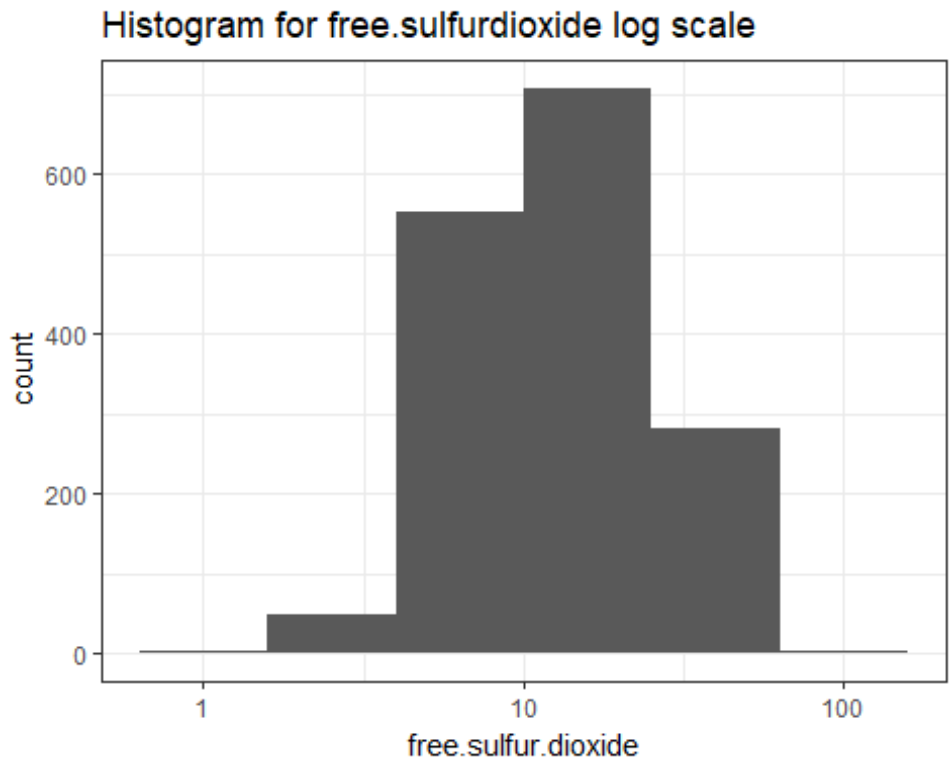
Exploring free Sulfur.dioxide



The distribution is skewed to left.

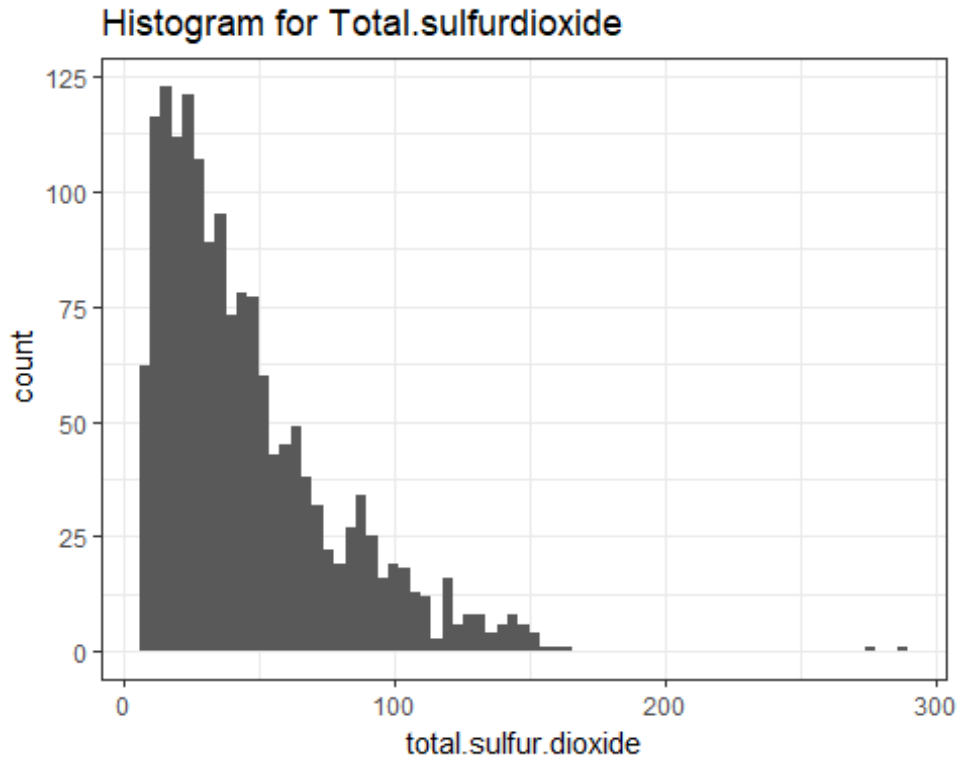


Looks like log transformation will yeild a distribution close to normal

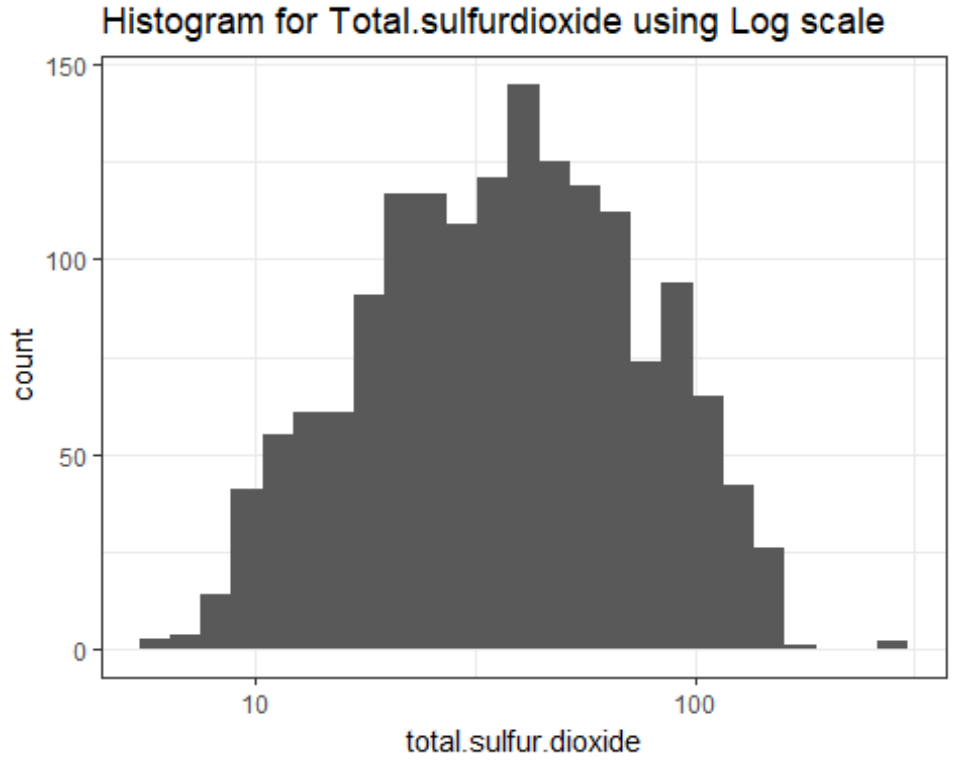


Our above assumptions seems to hold true this yeilds a close to normal distribution.

Exploring Total sulfur.dioxide

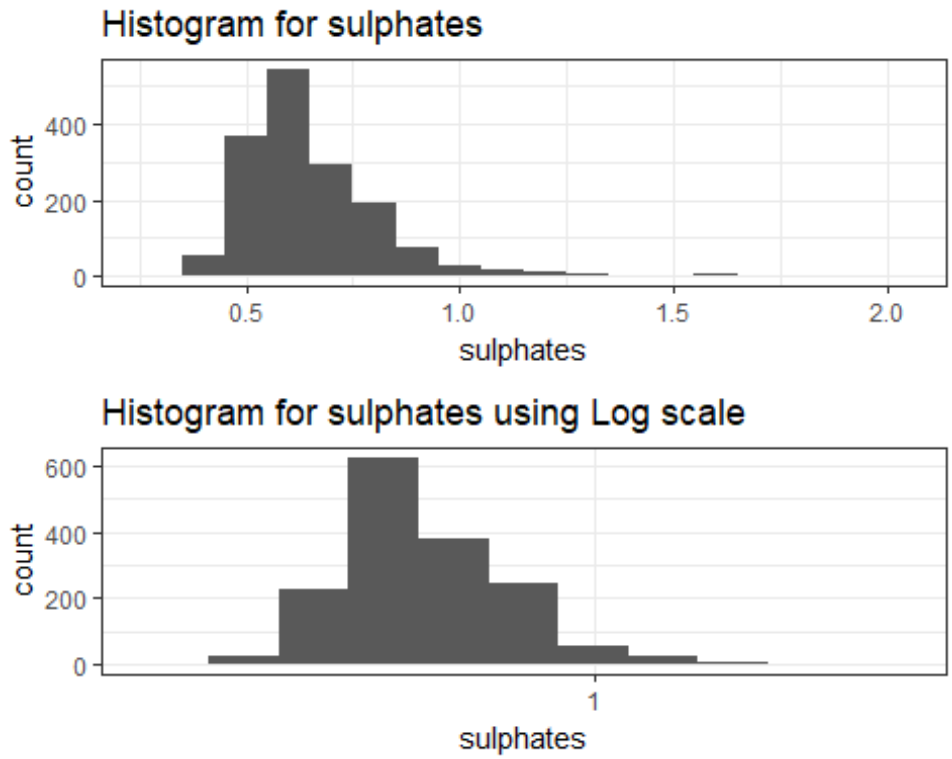


The distribution of total sulfurdioxide looks skewed to left.



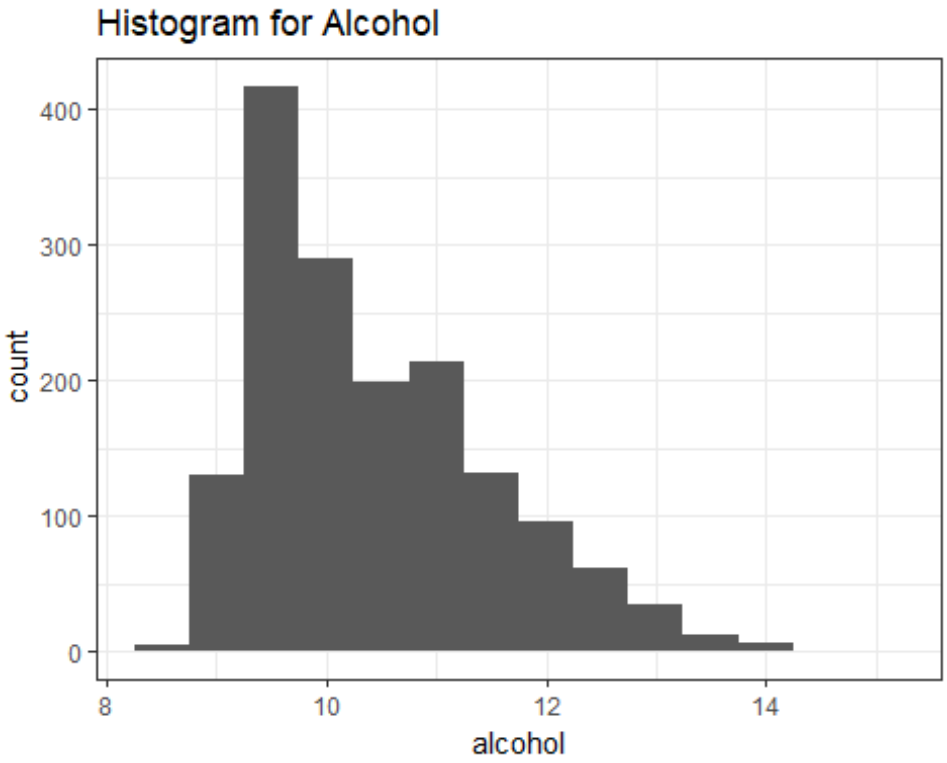
The similar trend follows here, the log transformation yeilds a close to normal distribution

Exploring Sulphates



Similar trend for sulphates.

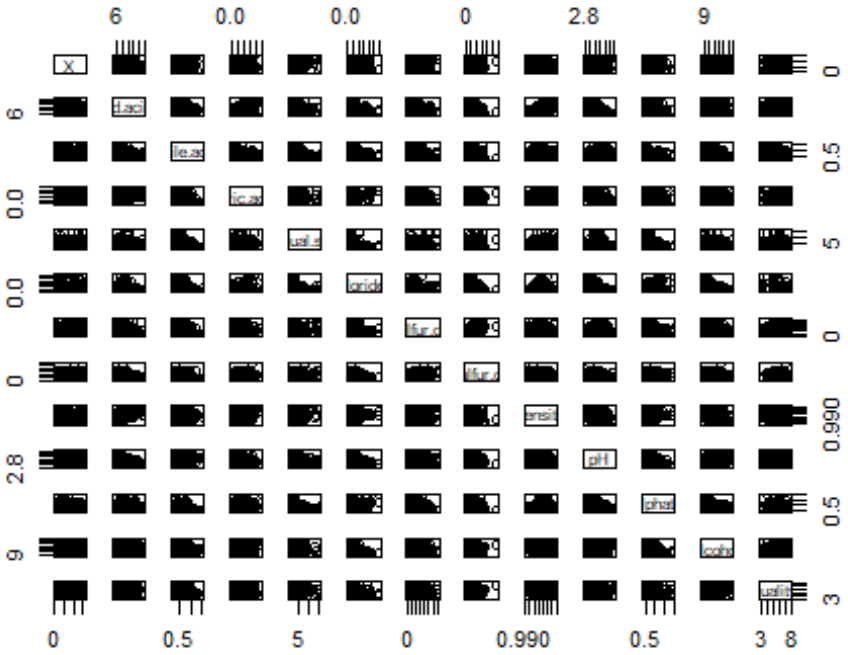
Exploring Alcohol



If we look at distributions for chemical properties they all tend to follow normal distributions when we change the scale to log10 for most of the distributions except for citric acid which seems to be closer to uniform distribution all the other follow close to normal distribution.

In our data set the chemical properties of **free sulfur dioxide and total sulphur dioxide** are in milligrams per 1dm^3 where as the other chemical properties are in grams per dm^3 to maintain consistency thoroughout I will be converting those two to g/dm^3.

Bivariate Plots Section

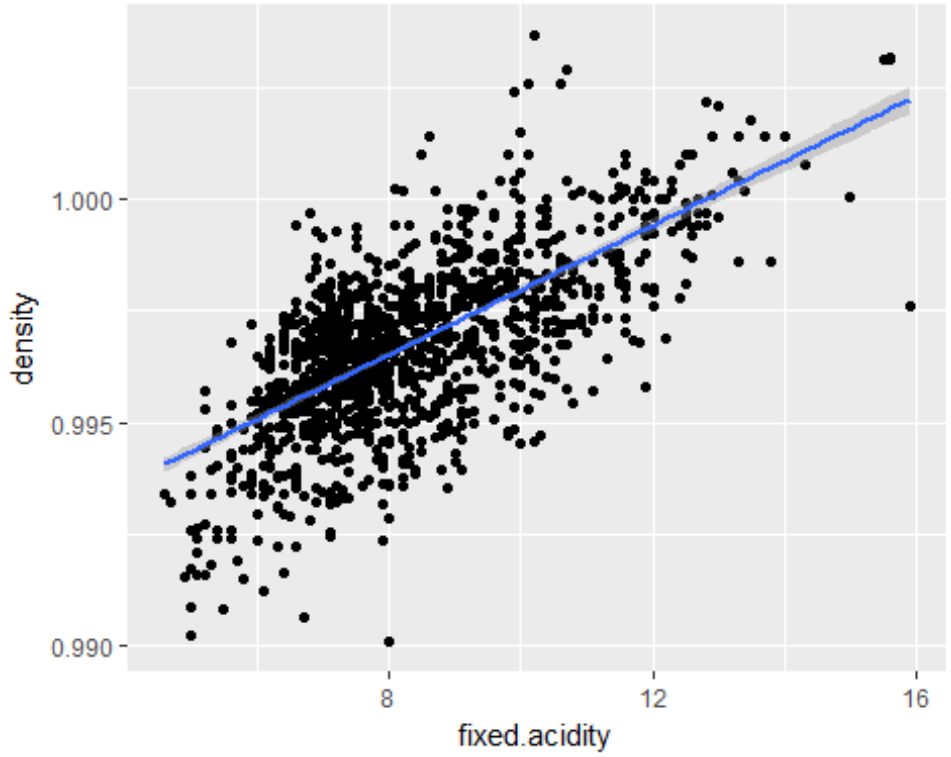


To look at relationship between all variables in data set

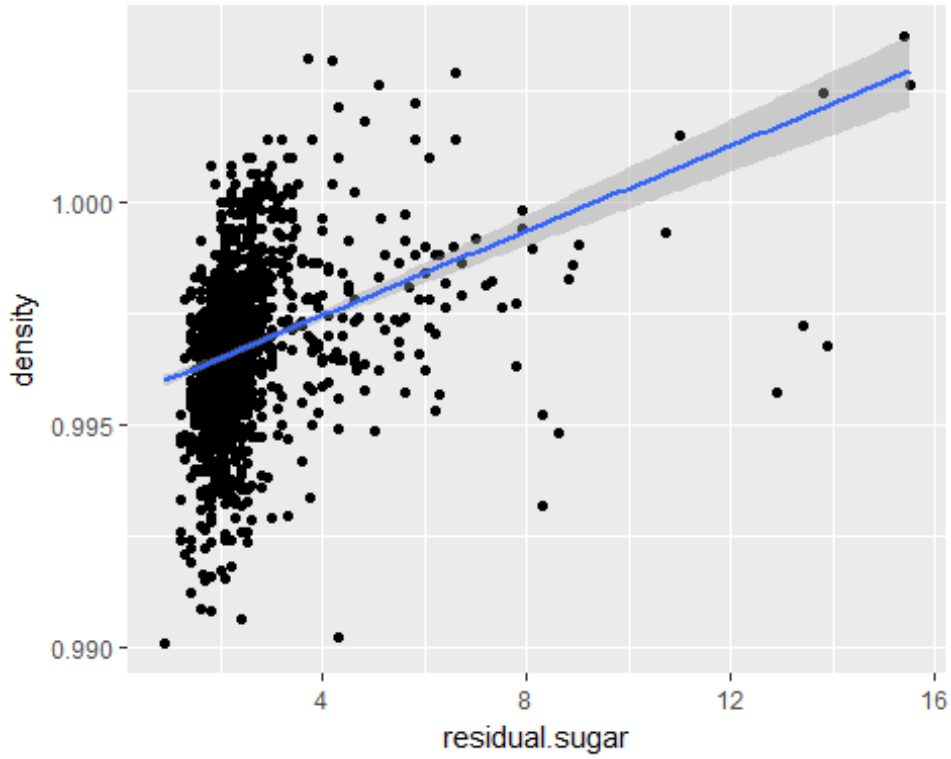
```
##                              X fixed.acidity volatile.acidity
## X                  1.000000000   -0.26848392     -0.008815099
## fixed.acidity     -0.268483920    1.00000000     -0.256130895
## volatile.acidity  -0.008815099   -0.25613089      1.000000000
## citric.acid       -0.153551355    0.67170343     -0.552495685
## residual.sugar    -0.031260835    0.11477672      0.001917882
## chlorides         -0.119868519    0.09370519      0.061297772
## free.sulfur.dioxide  0.090479643   -0.15379419     -0.010503827
## total.sulfur.dioxide -0.117849669   -0.11318144      0.076470005
## density           -0.368372087    0.66804729      0.022026232
## pH                 0.136005328   -0.68297819      0.234937294
## sulphates         -0.125306999    0.18300566     -0.260986685
## alcohol            0.245122841   -0.06166827     -0.202288027
## quality            0.066452608    0.12405165     -0.390557780
##                   citric.acid residual.sugar    chlorides
## X                 -0.15355136   -0.031260835 -0.119868519
## fixed.acidity      0.67170343    0.114776724  0.093705186
## volatile.acidity  -0.55249568    0.001917882  0.061297772
## citric.acid        1.00000000    0.143577162  0.203822914
## residual.sugar     0.14357716    1.000000000  0.055609535
## chlorides          0.20382291    0.055609535  1.000000000
## free.sulfur.dioxide  -0.06097813    0.187048995  0.005562147
## total.sulfur.dioxide  0.03553302    0.203027882  0.047400468
```

```
## density               0.36494718    0.355283371  0.200632327
## pH                   -0.54190414   -0.085652422 -0.265026131
## sulphates             0.31277004    0.005527121  0.371260481
## alcohol               0.10990325    0.042075437 -0.221140545
## quality               0.22637251    0.013731637 -0.128906560
##                      free.sulfur.dioxide total.sulfur.dioxide     density
## X                        0.090479643          -0.11784967 -0.36837209
## fixed.acidity           -0.153794193          -0.11318144  0.66804729
## volatile.acidity        -0.010503827           0.07647000  0.02202623
## citric.acid             -0.060978129           0.03553302  0.36494718
## residual.sugar           0.187048995           0.20302788  0.35528337
## chlorides                0.005562147           0.04740047  0.20063233
## free.sulfur.dioxide      1.000000000           0.66766645 -0.02194583
## total.sulfur.dioxide     0.667666450           1.00000000  0.07126948
## density                 -0.021945831           0.07126948  1.00000000
## pH                       0.070377499          -0.06649456 -0.34169933
## sulphates                0.051657572           0.04294684  0.14850641
## alcohol                 -0.069408354          -0.20565394 -0.49617977
## quality                 -0.050656057          -0.18510029 -0.17491923
##                               pH    sulphates     alcohol     quality
## X                     0.13600533 -0.125306999  0.24512284  0.06645261
## fixed.acidity        -0.68297819  0.183005664 -0.06166827  0.12405165
## volatile.acidity      0.23493729 -0.260986685 -0.20228803 -0.39055778
## citric.acid          -0.54190414  0.312770044  0.10990325  0.22637251
## residual.sugar       -0.08565242  0.005527121  0.04207544  0.01373164
## chlorides            -0.26502613  0.371260481 -0.22114054 -0.12890656
## free.sulfur.dioxide   0.07037750  0.051657572 -0.06940835 -0.05065606
## total.sulfur.dioxide -0.06649456  0.042946836 -0.20565394 -0.18510029
## density              -0.34169933  0.148506412 -0.49617977 -0.17491923
## pH                    1.00000000 -0.196647602  0.20563251 -0.05773139
## sulphates            -0.19664760  1.000000000  0.09359475  0.25139708
## alcohol               0.20563251  0.093594750  1.00000000  0.47616632
## quality              -0.05773139  0.251397079  0.47616632  1.00000000
```

Corelation between all variables in data set

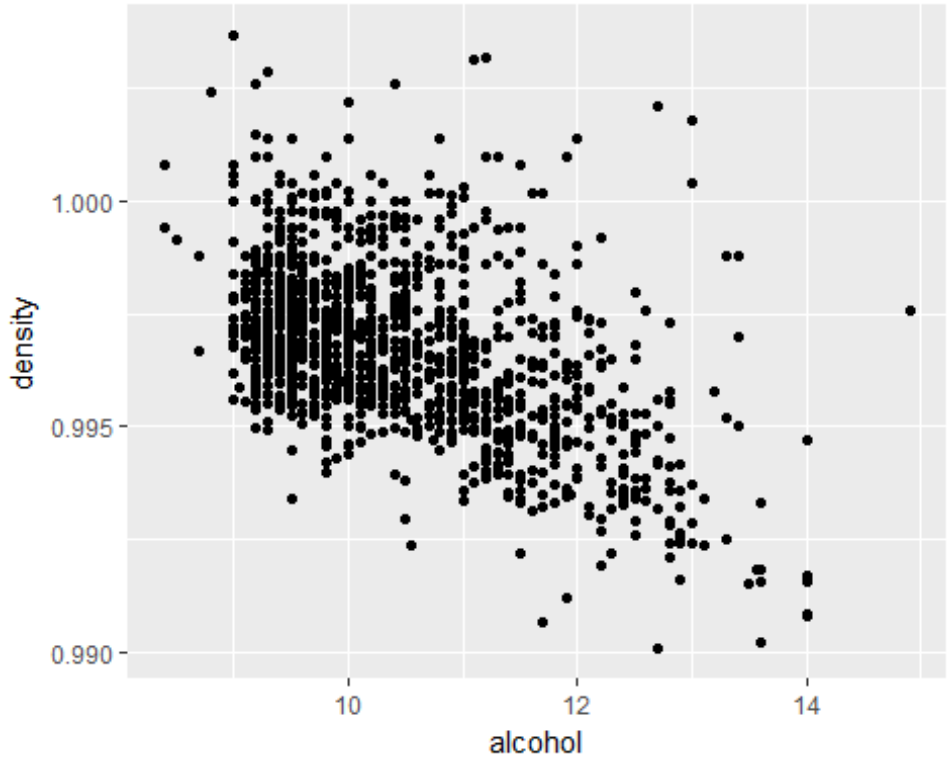


scatterplot for fixed acidity over density, this shows they are closely related.



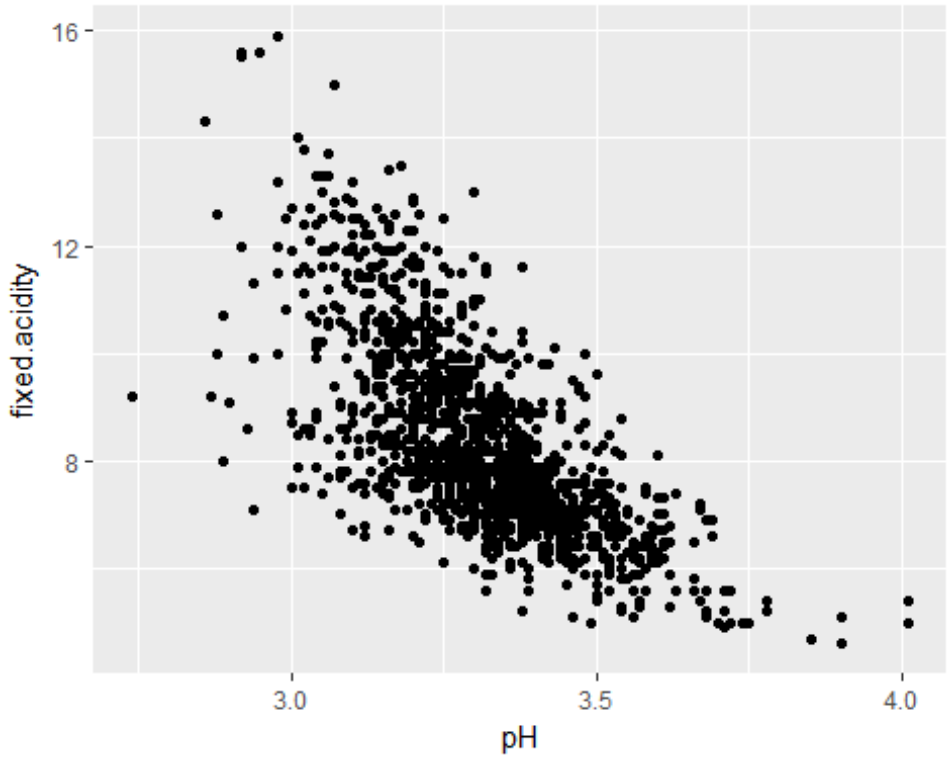
```
##
##  Pearson's product-moment correlation
##
## data:  redWine$residual.sugar and redWine$density
## t = 15.189, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3116908 0.3973835
## sample estimates:
##       cor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

scatterplot for residual.sugar over density, this shows they are many outliers in the data.



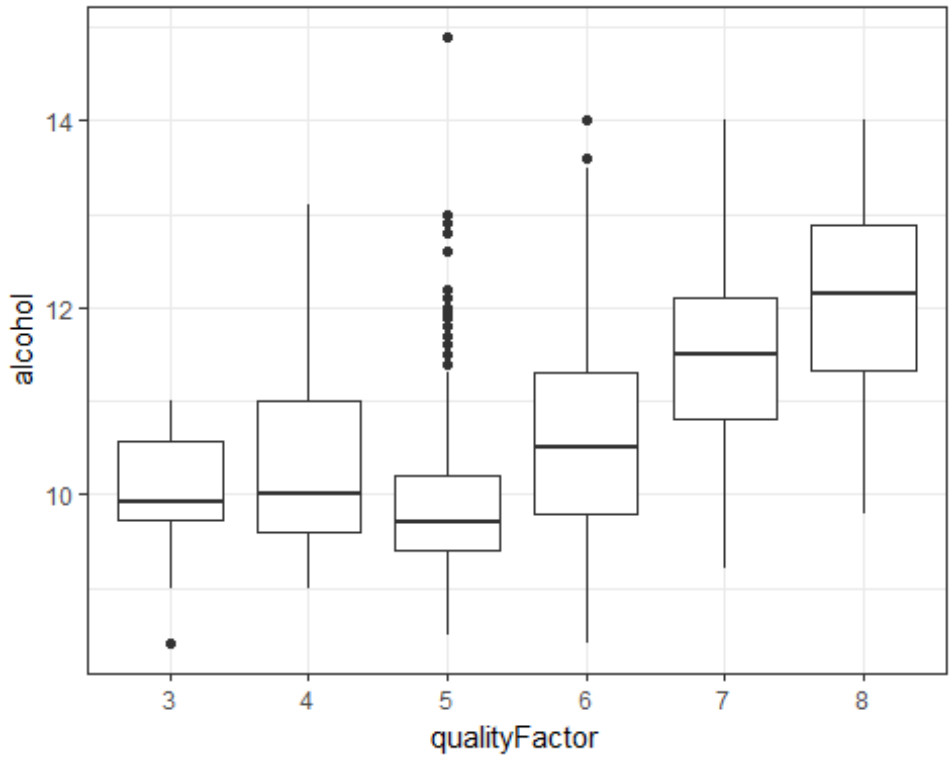

```
##
##  Pearson's product-moment correlation
##
## data:  redWine$alcohol and redWine$density
## t = -22.838, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5322547 -0.4583061
## sample estimates:
##       cor
## -0.4961798
```

We can see a negative relationship between density over alcohol with correlation coefficeint -0.49

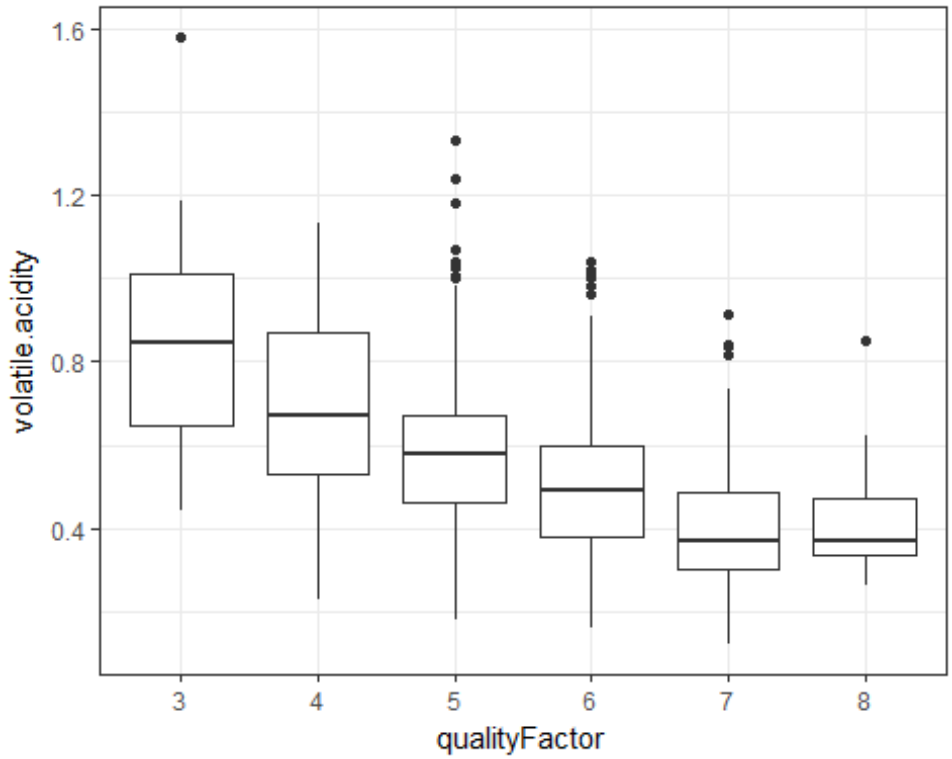



```
##
##  Pearson's product-moment correlation
##
## data:  redWine$fixed.acidity and redWine$pH
## t = -37.366, df = 1597, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7082857 -0.6559174
## sample estimates:
##       cor
##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

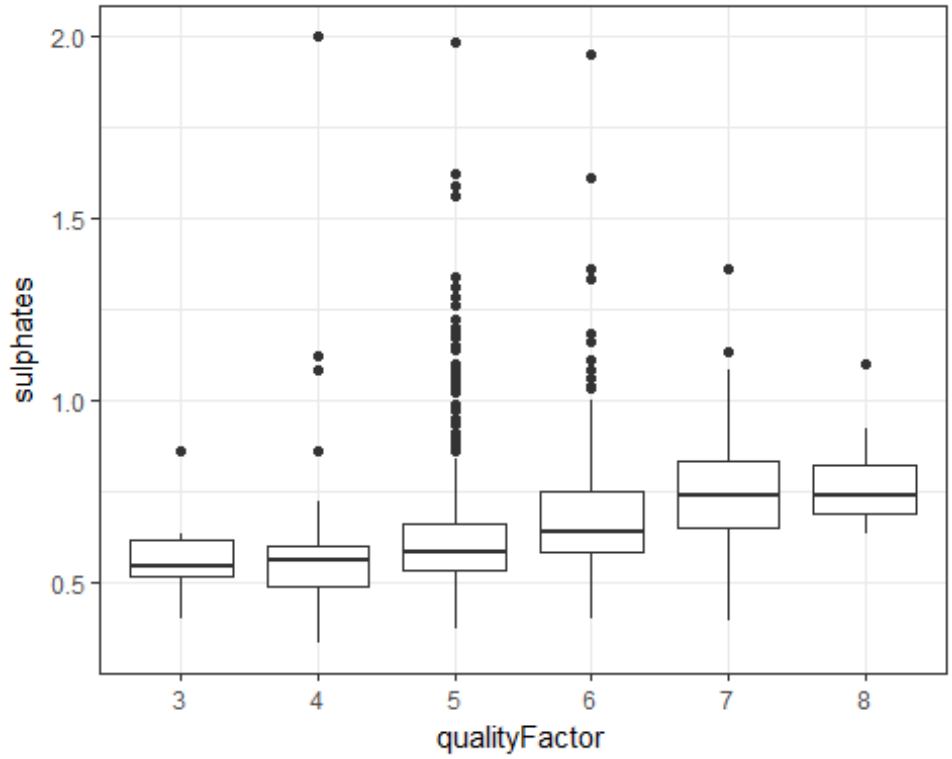
we can see that fixed acidity and ph are negatively correlated.



How does relationship of alcohol and quality change for different qualities of red wine. we can see mean alcohol % by volume increases with quality.



How does relationship of volatile acidity and quality change for different qualities of red wine. we can see mean volatile Acidity decreases with increase in qulaity of redWine.



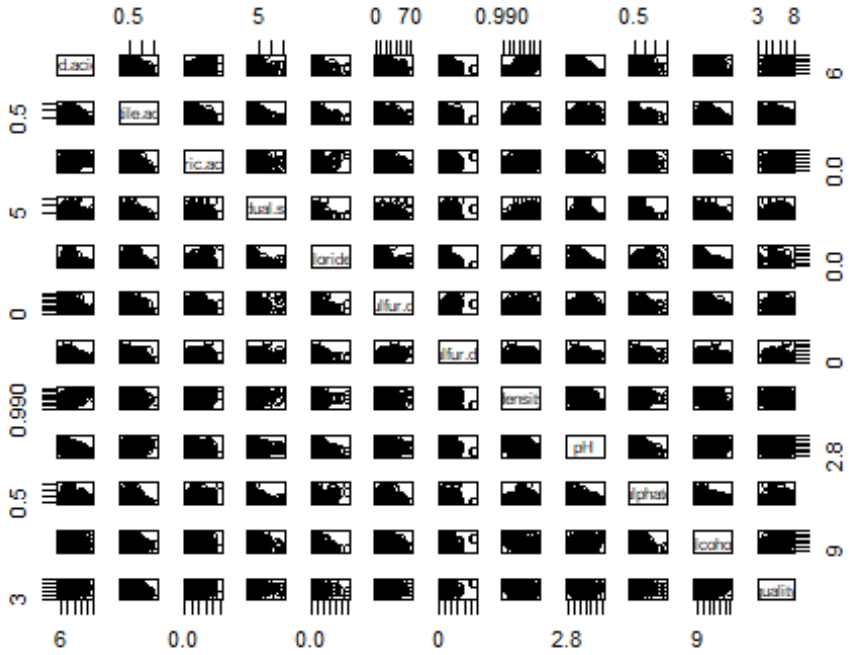
How does the mean of sulphates changes with quality.

In the above section I have created different plots to see if there is a relationship between quality and other chemical properties & realtionship between other chemical properties in general.

The plots that require further analysis to gain more insight were scatterplot matrix along with correlation coefficeint table, relationship between quality over alcohol, sulphates and volatile acidity.

Few relations between other chemical properties pH over fixed.acidity and density over fixed.acidity.

Bivariate Analysis:



Here I have created a scatter plot matrix to take a look at correlation between variables in our data set visually.

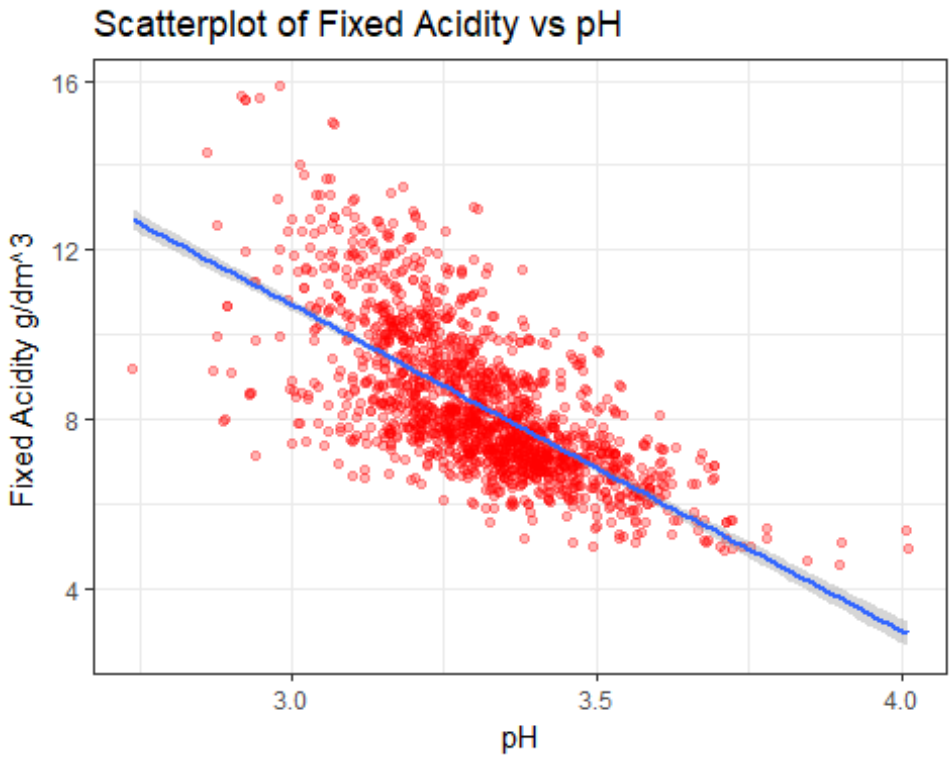
```
##                       fixed.acidity volatile.acidity citric.acid
## fixed.acidity           1.00000000     -0.256130895  0.67170343
## volatile.acidity       -0.25613089      1.000000000 -0.55249568
## citric.acid             0.67170343     -0.552495685  1.00000000
## residual.sugar          0.11477672      0.001917882  0.14357716
## chlorides               0.09370519      0.061297772  0.20382291
## free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813
## total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302
## density                 0.66804729      0.022026232  0.36494718
## pH                     -0.68297819      0.234937294 -0.54190414
## sulphates               0.18300566     -0.260986685  0.31277004
## alcohol                -0.06166827     -0.202288027  0.10990325
## quality                 0.12405165     -0.390557780  0.22637251
##                       residual.sugar    chlorides free.sulfur.dioxide
## fixed.acidity            0.114776724  0.093705186        -0.153794193
## volatile.acidity         0.001917882  0.061297772        -0.010503827
## citric.acid              0.143577162  0.203822914        -0.060978129
## residual.sugar           1.000000000  0.055609535         0.187048995
## chlorides                0.055609535  1.000000000         0.005562147
## free.sulfur.dioxide      0.187048995  0.005562147         1.000000000
## total.sulfur.dioxide     0.203027882  0.047400468         0.667666450
```

```
## density                 0.355283371  0.200632327        -0.021945831
## pH                     -0.085652422 -0.265026131         0.070377499
## sulphates               0.005527121  0.371260481         0.051657572
## alcohol                 0.042075437 -0.221140545        -0.069408354
## quality                 0.013731637 -0.128906560        -0.050656057
##                      total.sulfur.dioxide     density          pH
## fixed.acidity                -0.11318144  0.66804729 -0.68297819
## volatile.acidity              0.07647000  0.02202623  0.23493729
## citric.acid                   0.03553302  0.36494718 -0.54190414
## residual.sugar                0.20302788  0.35528337 -0.08565242
## chlorides                     0.04740047  0.20063233 -0.26502613
## free.sulfur.dioxide           0.66766645 -0.02194583  0.07037750
## total.sulfur.dioxide          1.00000000  0.07126948 -0.06649456
## density                       0.07126948  1.00000000 -0.34169933
## pH                           -0.06649456 -0.34169933  1.00000000
## sulphates                     0.04294684  0.14850641 -0.19664760
## alcohol                      -0.20565394 -0.49617977  0.20563251
## quality                      -0.18510029 -0.17491923 -0.05773139
##                         sulphates     alcohol     quality
## fixed.acidity         0.183005664 -0.06166827  0.12405165
## volatile.acidity     -0.260986685 -0.20228803 -0.39055778
## citric.acid           0.312770044  0.10990325  0.22637251
## residual.sugar        0.005527121  0.04207544  0.01373164
## chlorides             0.371260481 -0.22114054 -0.12890656
## free.sulfur.dioxide   0.051657572 -0.06940835 -0.05065606
## total.sulfur.dioxide  0.042946836 -0.20565394 -0.18510029
## density               0.148506412 -0.49617977 -0.17491923
## pH                   -0.196647602  0.20563251 -0.05773139
## sulphates             1.000000000  0.09359475  0.25139708
## alcohol               0.093594750  1.00000000  0.47616632
## quality               0.251397079  0.47616632  1.00000000
```

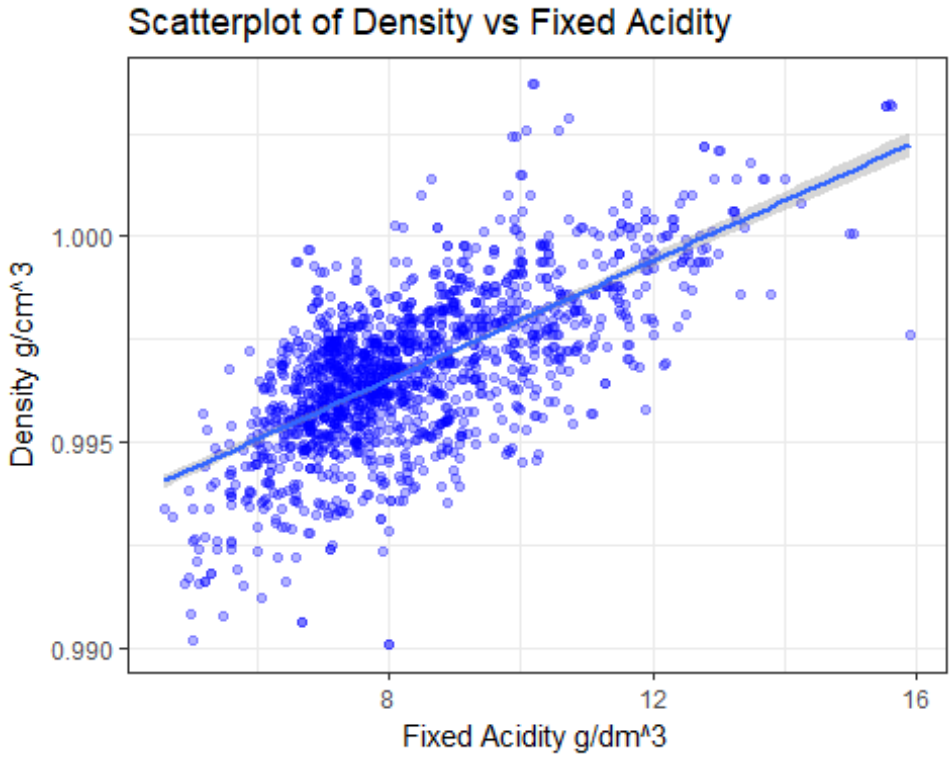
Then performed a correlation test to get corelation coefficients between different variables.

From the two plots we can see that **quality is kind of related to Alcohol > Volatile Acidity > sulphates** of all available chemical properties.

If we look at relationship between chemical properties. pH is highly correlated to fixed acidity and fixed acidity and desity are correlated.

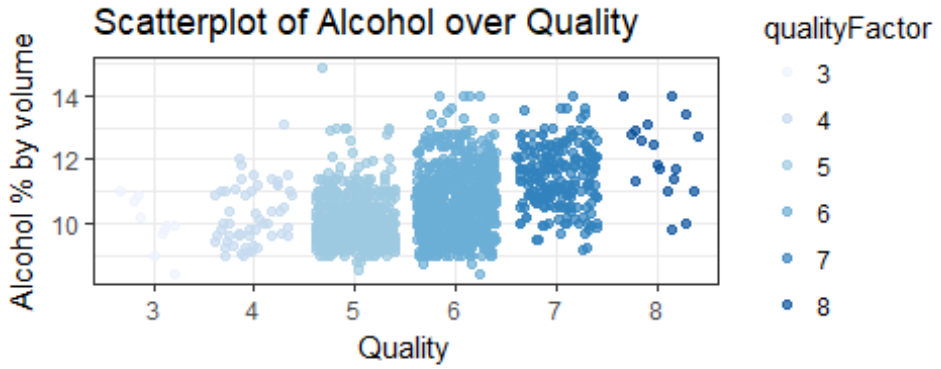


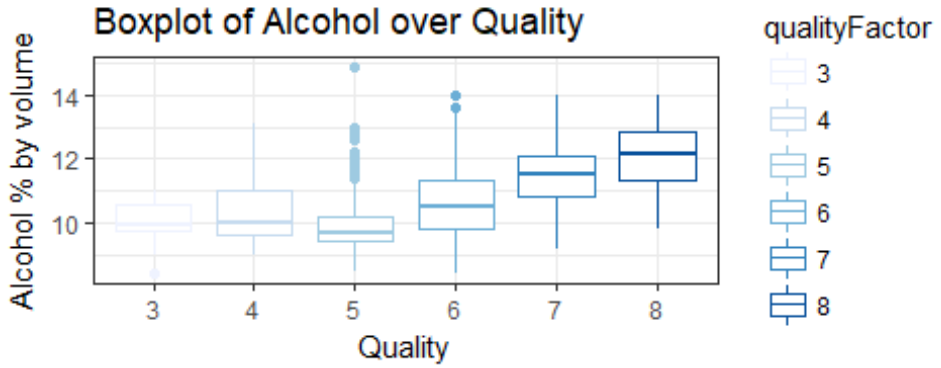
PH is negatively related to acidity which is a well know fact that the pH value decreases with increase in acidity



Density is positevly correlated to fixed acidity with a correlation coefficient of 0.66.

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

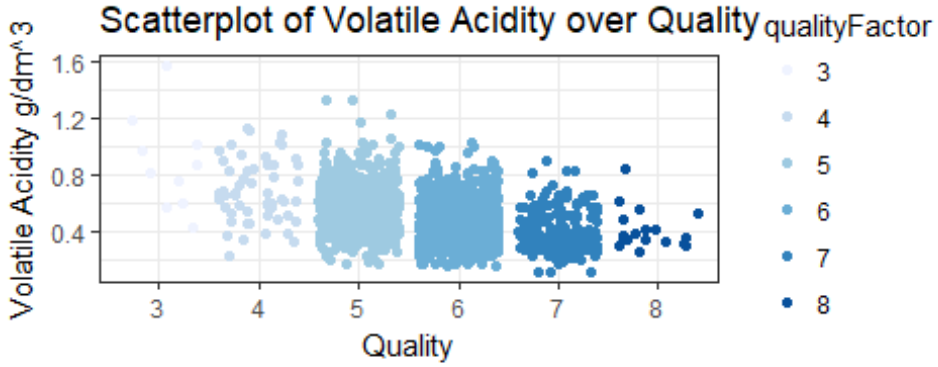


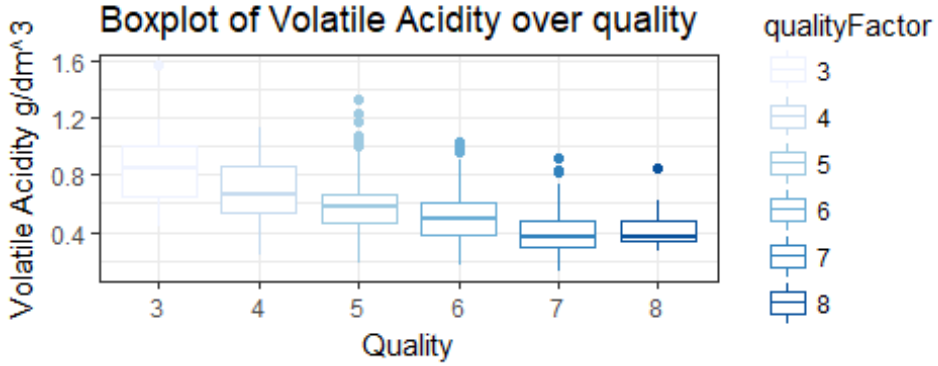


Looking at the above scatter plot we can see a positive realtionship between **quality and alcohol** This is more evident from the boxplot when we can see an increasing trend for mean of alcohol over quality.

```
##
##  Pearson's product-moment correlation
##
## data:  redWine$volatile.acidity and redWine$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
```

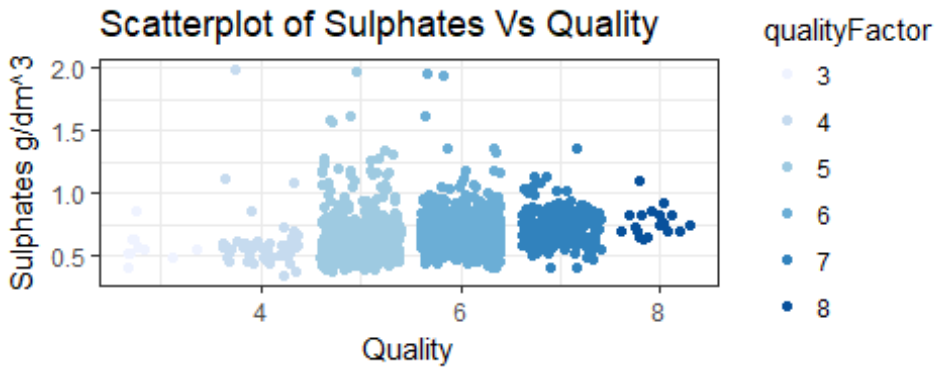


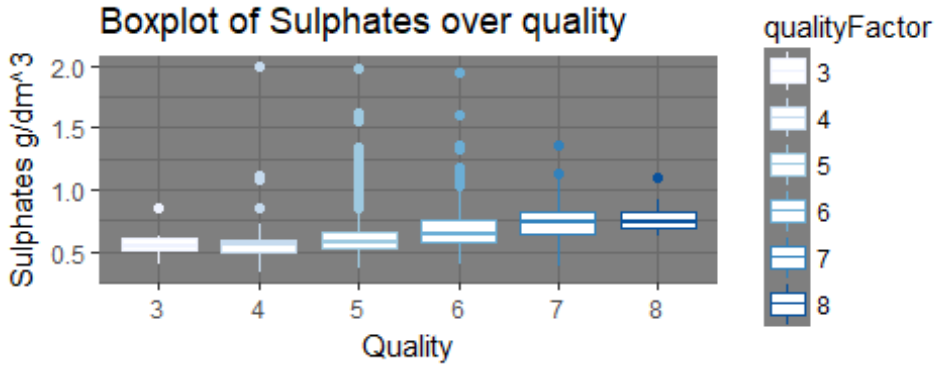


Here we see a inverse linear relationship between quality and Volatile Acidity. the trend is very sutle. similarly, it is more evident in the boxplot of volatile acidity over quality.

As the amount of volatile Acidity (acetic acid) in wine increase it leads to unpleasant, vulger taste of wine. which is evident with this inverse relationship with quaity of wine.

```
##
##  Pearson's product-moment correlation
##
## data:  redWine$sulphates and redWine$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
```





there is a slight linear relationship between quality and sulphates but as we look at box plot we can see as the the quality is increasing there are few outliers in the data which might be pulling the mean towards a higher value.

Correlation test reveals that there is slight to no relationship between quality and sulphates. As we consider correlation coefficient > 0.3 to have any kind of relationship.

Multivariate plots section

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = redWine)
## m2: lm(formula = quality ~ alcohol + log10(sulphates), data = redWine)
## m3: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity,
##     data = redWine)
## m4: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity +
##     log10(residual.sugar), data = redWine)
## m5: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity +
##     log10(residual.sugar) + log10(chlorides), data = redWine)
## m6: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity +
##     log10(residual.sugar) + log10(chlorides) +
log10(free.sulfur.dioxide1),
##     data = redWine)
## m7: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity +
##     log10(residual.sugar) + log10(chlorides) + log10(free.sulfur.dioxide1)
+
```

```
##     log10(total.sulfur.dioxide1), data = redWine)
## m8: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity +
##     log10(residual.sugar) + log10(chlorides) + log10(free.sulfur.dioxide1)
+
##     log10(total.sulfur.dioxide1) + density, data = redWine)
## m9: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity +
##     log10(residual.sugar) + log10(chlorides) + log10(free.sulfur.dioxide1)
+
##     log10(total.sulfur.dioxide1) + density + pH, data = redWine)
## m10: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity
+
##     log10(residual.sugar) + log10(chlorides) + log10(free.sulfur.dioxide1)
+
##     log10(total.sulfur.dioxide1) + density + pH, data = redWine)
## m11: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity
+
##     log10(residual.sugar) + log10(chlorides) + log10(free.sulfur.dioxide1)
+
##     log10(total.sulfur.dioxide1) + density + pH + citric.acid,
##     data = redWine)
##
##
================================================================================
================================================================================
================================
##                                        m1            m2            m3
m4            m5            m6            m7            m8            m9
m10           m11
## --------------------------------------------------------------------------
--------------------------------------------------------------------------
----------------------------------
##   (Intercept)                          1.875***      2.541***      3.369***
3.377***      3.062***      3.314***      3.923***     -1.636         8.747
8.747        -1.210
##                                       (0.175)       (0.177)       (0.184)
(0.185)       (0.202)       (0.270)       (0.304)      (12.737)      (12.929)
(12.929)      (14.041)
##   alcohol                             0.361***      0.335***      0.303***
0.304***      0.282***      0.279***      0.260***      0.266***
0.271***      0.271***      0.287***
##                                       (0.017)       (0.016)       (0.016)
(0.016)       (0.017)       (0.017)       (0.017)       (0.021)       (0.021)
(0.021)       (0.023)
##   log10(sulphates)                                   2.070***      1.477***
1.478***      1.714***      1.738***      1.788***      1.766***
1.770***      1.770***      1.778***
##                                                     (0.177)       (0.177)
(0.177)       (0.187)       (0.188)       (0.187)       (0.194)       (0.193)
(0.193)       (0.193)
##   volatile.acidity                                                 -1.156***
```

```
-1.154***     -1.100***     -1.097***     -1.074***     -1.071***     -
0.952***     -0.952***     -1.064***
##                                                                 (0.097)
(0.097)       (0.098)       (0.098)       (0.098)       (0.098)       (0.102)
(0.102)       (0.119)
##   log10(residual.sugar)
-0.036         0.024         0.041         0.111         0.080         0.088
0.088         0.066
##
(0.106)       (0.107)       (0.107)       (0.108)       (0.129)       (0.128)
(0.128)       (0.129)
##   log10(chlorides)
-0.495***     -0.510***     -0.505***     -0.511***     -0.619***     -
0.619***     -0.590***
##
(0.129)       (0.129)       (0.129)       (0.130)       (0.132)       (0.132)
(0.132)
##   log10(free.sulfur.dioxide1)
-0.078         0.221*        0.226*        0.272**       0.272**
0.252**
##
(0.056)       (0.089)       (0.090)       (0.090)       (0.090)       (0.091)
##   log10(total.sulfur.dioxide1)
-0.383***     -0.383***     -0.414***     -0.414***     -0.388***
##
(0.090)       (0.090)       (0.090)       (0.090)       (0.091)
##   density
5.505        -3.566        -3.566         6.747
##
(12.610)      (12.745)      (12.745)      (13.950)
##   pH
-0.486***     -0.486***     -0.591***
##
(0.119)       (0.119)       (0.133)
##   citric.acid
-0.245
##
(0.135)
## ------------------------------------------------------------------------
------------------------------------------------------------------------
-----------------------------------
##   R-squared                          0.227         0.288         0.345
0.346         0.352         0.352         0.360         0.360         0.366
0.366         0.368
##   adj. R-squared                     0.226         0.287         0.344
0.344         0.350         0.350         0.357         0.356         0.363
0.363         0.364
##   sigma                              0.710         0.682         0.654
0.654         0.651         0.651         0.648         0.648         0.645
0.645         0.644
```

```
##   F                                  468.267       322.031       280.646
210.397       172.715       144.349       127.638       111.651       102.059
102.059        92.313
##   p                                    0.000         0.000         0.000
0.000         0.000         0.000         0.000         0.000         0.000
0.000         0.000
##   Log-likelihood                -1721.057     -1655.601     -1587.752
-1587.693     -1580.332     -1579.336     -1570.300     -1570.204     -
1561.907     -1561.907     -1560.257
##   Deviance                       805.870       742.522       682.108
682.058       675.807       674.965       667.380       667.300       660.410
660.410       659.049
##   AIC                          3448.114      3319.202      3185.503
3187.386      3174.664      3174.671      3158.600      3160.409
3145.814      3145.814      3144.513
##   BIC                          3464.245      3340.711      3212.389
3219.649      3212.304      3217.688      3206.994      3214.180
3204.962      3204.962      3209.039
##   N                            1599          1599          1599
1599          1599          1599          1599          1599          1599
1599          1599
##
================================================================================
================================================================================
================================
```

here I have used multiple linear regression to predict the quality using chemical properties. I have converted the chemical properties to log10 based on previous findings from univariate analysis.

Alcohol, Sulphates & volatile acidity account for most variation in the quality.

```
##
## Calls:
## n1: lm(formula = quality ~ alcohol, data = newdataredwine)
## n2: lm(formula = quality ~ alcohol + sulphates, data = newdataredwine)
## n3: lm(formula = quality ~ alcohol + sulphates + volatile.acidity,
##     data = newdataredwine)
## n4: lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
##     residual.sugar, data = newdataredwine)
## n5: lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
##     residual.sugar + chlorides, data = newdataredwine)
## n6: lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
##     residual.sugar + chlorides + free.sulfur.dioxide1, data =
newdataredwine)
## n7: lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
##     residual.sugar + chlorides + free.sulfur.dioxide1 +
total.sulfur.dioxide1,
##     data = newdataredwine)
## n8: lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
```

```
##     residual.sugar + chlorides + free.sulfur.dioxide1 +
total.sulfur.dioxide1 +
##     density, data = newdataredwine)
## n9: lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
##     residual.sugar + chlorides + free.sulfur.dioxide1 +
total.sulfur.dioxide1 +
##     density + pH, data = newdataredwine)
## n10: lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
##     residual.sugar + chlorides + free.sulfur.dioxide1 +
total.sulfur.dioxide1 +
##     density + pH, data = newdataredwine)
## n11: lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
##     residual.sugar + chlorides + free.sulfur.dioxide1 +
total.sulfur.dioxide1 +
##     density + pH + citric.acid, data = newdataredwine)
##
##
================================================================================
================================================================================
=========================
##                                n1            n2            n3            n4
n5            n6            n7            n8            n9           n10
n11
## --------------------------------------------------------------------------------
--------------------------------------------------------------------------------
---------------------------
##   (Intercept)               1.800***      1.191***      2.280***
2.354***      2.611***      2.666***      2.798***      9.017        23.933
23.933        13.062
##                            (0.184)       (0.181)       (0.201)
(0.210)       (0.227)       (0.229)       (0.231)      (14.598)      (14.810)
(14.810)      (16.587)
##   alcohol                   0.368***      0.323***      0.299***
0.300***      0.286***      0.284***      0.271***      0.264***
0.270***      0.270***      0.285***
##                            (0.018)       (0.017)       (0.017)
(0.017)       (0.017)       (0.017)       (0.018)       (0.024)       (0.024)
(0.024)       (0.026)
##   sulphates                                1.690***      1.254***
1.263***      1.312***      1.330***      1.303***      1.321***
1.359***      1.359***      1.356***
##                                           (0.133)       (0.134)
(0.134)       (0.135)       (0.135)       (0.135)       (0.142)       (0.141)
(0.141)       (0.141)
##   volatile.acidity                                      -1.075***     -
1.068***     -1.038***     -1.040***     -1.017***     -1.020***     -
0.866***     -0.866***     -0.964***
##                                                         (0.099)
(0.100)       (0.100)       (0.100)       (0.100)       (0.100)       (0.104)
(0.104)       (0.124)
```

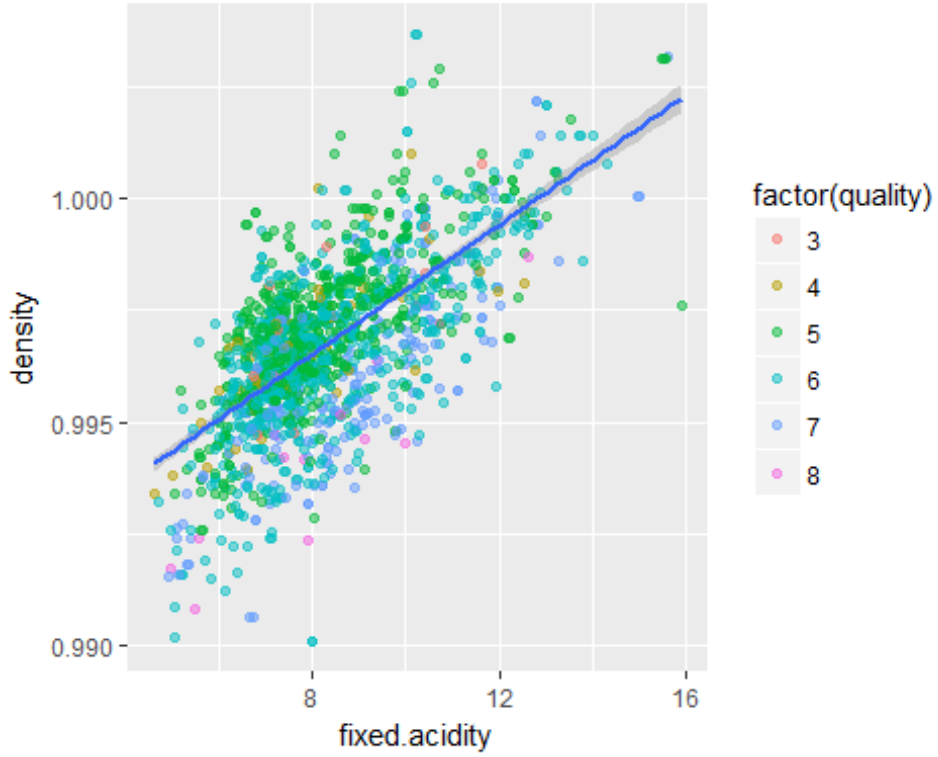
```
##   residual.sugar                                                          -
0.042        -0.018        -0.012         0.004         0.013         0.017
0.017         0.008
##
(0.035)       (0.036)       (0.036)       (0.036)       (0.042)       (0.042)
(0.042)       (0.042)
##   chlorides
-2.604**      -2.701**      -2.553**      -2.500**      -2.968**      -
2.968**      -2.932**
##
(0.898)       (0.899)       (0.897)       (0.906)       (0.904)       (0.904)
(0.904)
##   free.sulfur.dioxide1
-0.000         0.000         0.000         0.000         0.000         0.000
##
(0.000)       (0.000)       (0.000)       (0.000)       (0.000)       (0.000)
##   total.sulfur.dioxide1
-0.000***     -0.000***     -0.000***     -0.000***     -0.000***
##
(0.000)       (0.000)       (0.000)       (0.000)       (0.000)
##   density
-6.203       -19.328       -19.328        -8.174
##
(14.558)      (14.697)      (14.697)      (16.577)
##   pH
-0.596***     -0.596***     -0.682***
##
(0.123)       (0.123)       (0.137)
##   citric.acid
-0.210
##
(0.145)
## ------------------------------------------------------------------------
------------------------------------------------------------------------
--------------------------
##   R-squared                   0.237         0.315         0.367
0.368         0.372         0.373         0.379         0.379         0.389
0.389         0.390
##   adj. R-squared              0.237         0.314         0.366
0.366         0.370         0.371         0.376         0.375         0.385
0.385         0.386
##   sigma                       0.690         0.654         0.629
0.629         0.627         0.627         0.624         0.625         0.620
0.620         0.619
##   F                         439.833       325.191       273.411
205.498       166.940       139.946       122.641       107.272        99.457
99.457        89.793
##   p                           0.000         0.000         0.000
0.000         0.000         0.000         0.000         0.000         0.000
0.000         0.000
```

```
##   Log-likelihood          -1483.659     -1407.338     -1351.148     -
1350.406     -1346.202     -1344.446     -1338.357     -1338.266     -
1326.616     -1326.616     -1325.553
##   Deviance                  674.033       605.152       558.981
558.396       555.089       553.715       548.973       548.902       539.944
539.944       539.134
##   AIC                      2973.319      2822.677      2712.296
2712.813      2706.404      2704.892      2694.714      2696.531
2675.232      2675.232      2675.106
##   BIC                      2989.086      2843.699      2738.574
2744.346      2743.193      2746.937      2742.014      2749.087
2733.043      2733.043      2738.173
##   N                         1416          1416          1416          1416
1416          1416          1416          1416          1416          1416
1416
##
================================================================================
================================================================================
========================
```

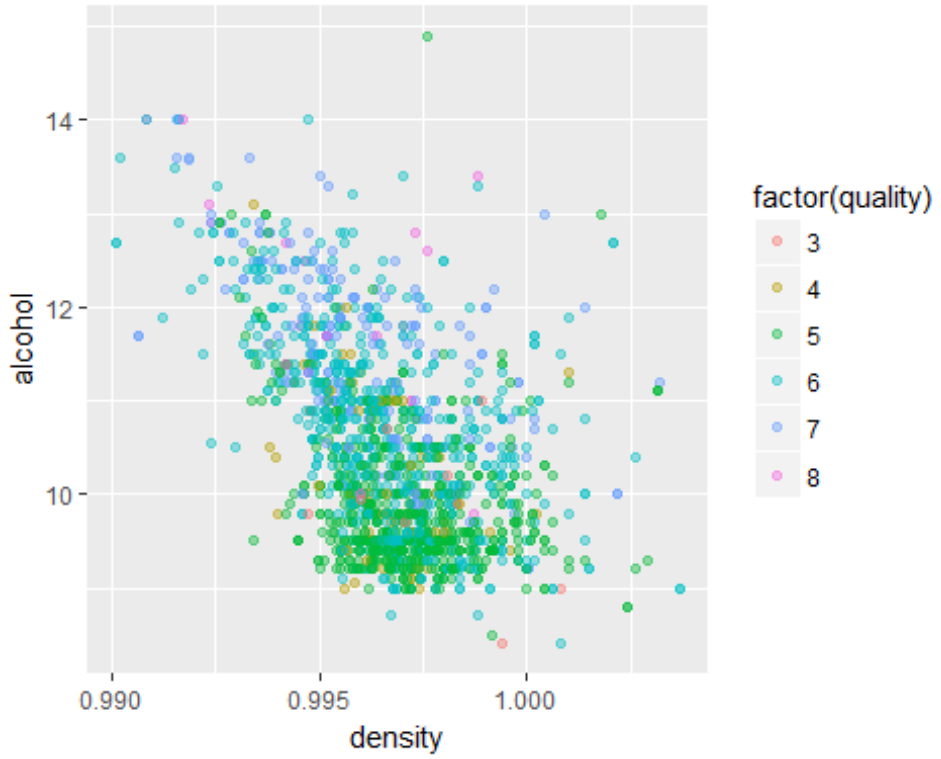
similarly performed multiple linear regression but here I have removed all the outlier's from the data as sepecified by boxplot from univariate analysis section.

We can see an improvement in the overall value of R - Squared compared to the above one.

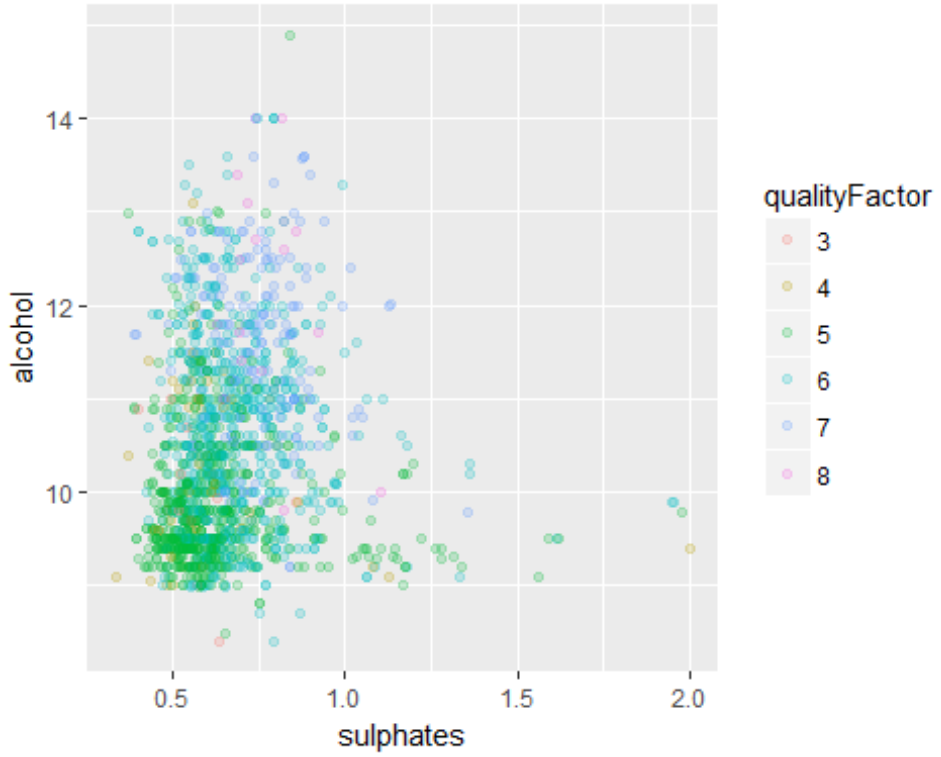
Alcohol, Sulphates & volatile acidity account for most variation in the quality.



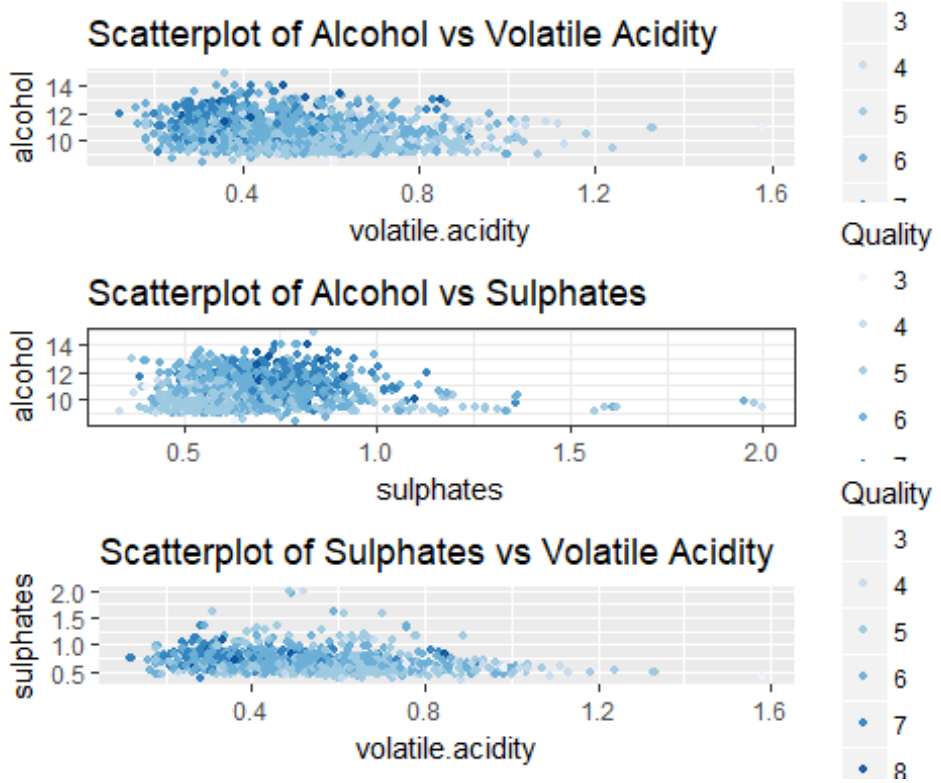
we can see the relationship between fixed acidity and density with points colored with respect to quality.



scatterplot of alcohol over density with points colored with respect to quality.

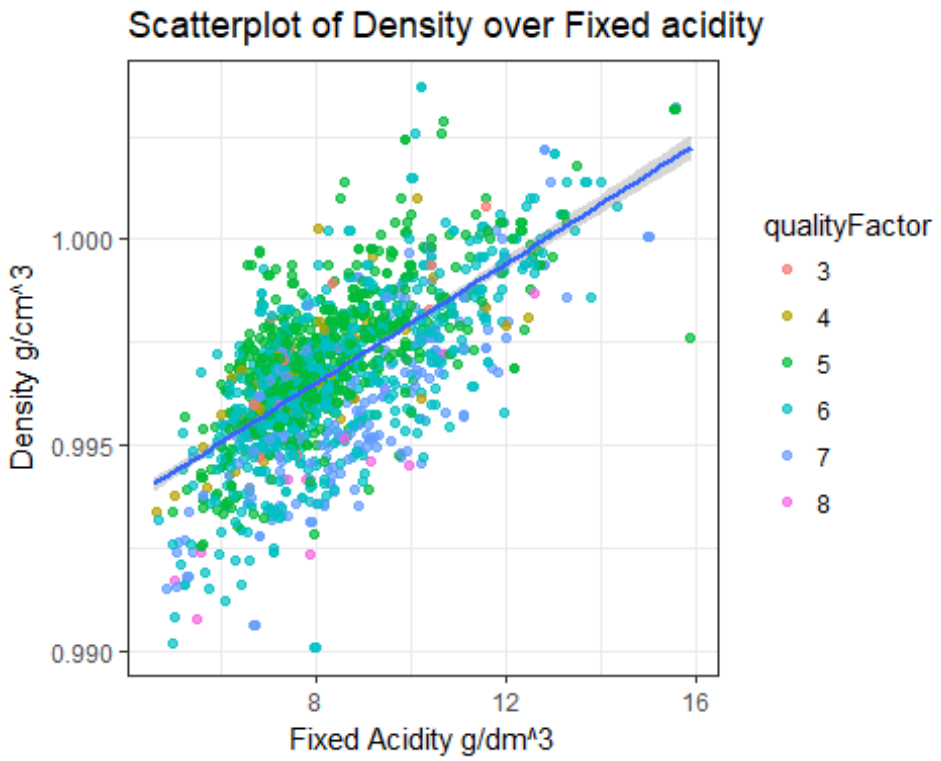


scatterplot of alcohol over sulphates with points colored with respect to quality.

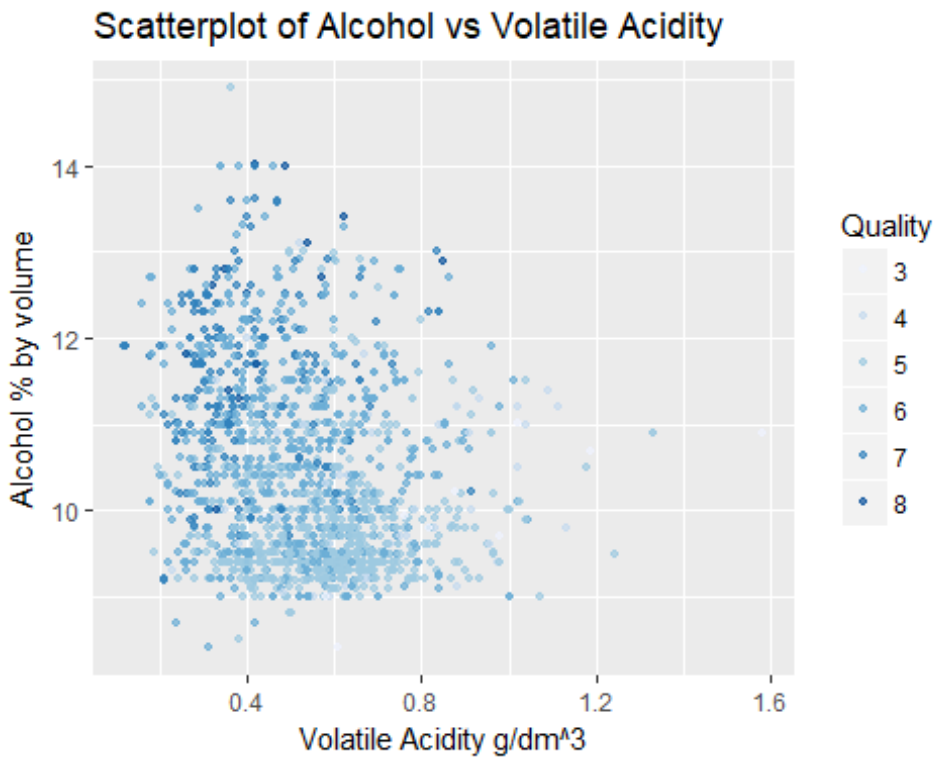


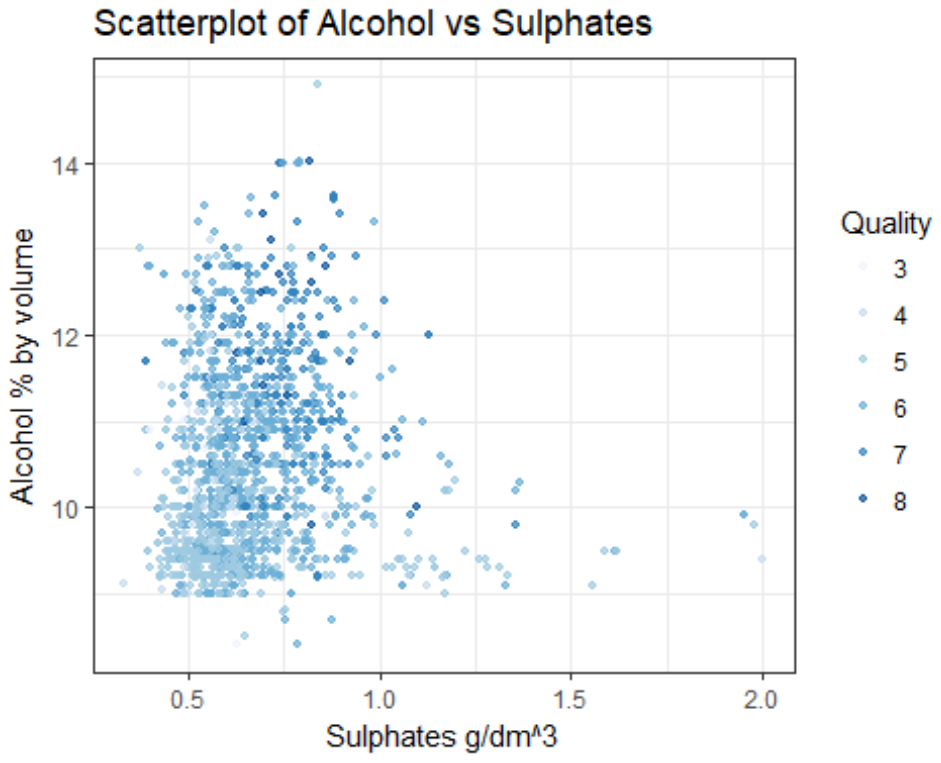
Looking at the variation from the correlation test and multiple linear regression I have created the above plots to look for the relationship between quality and other chemical properties. These plots are discussed in detail below.

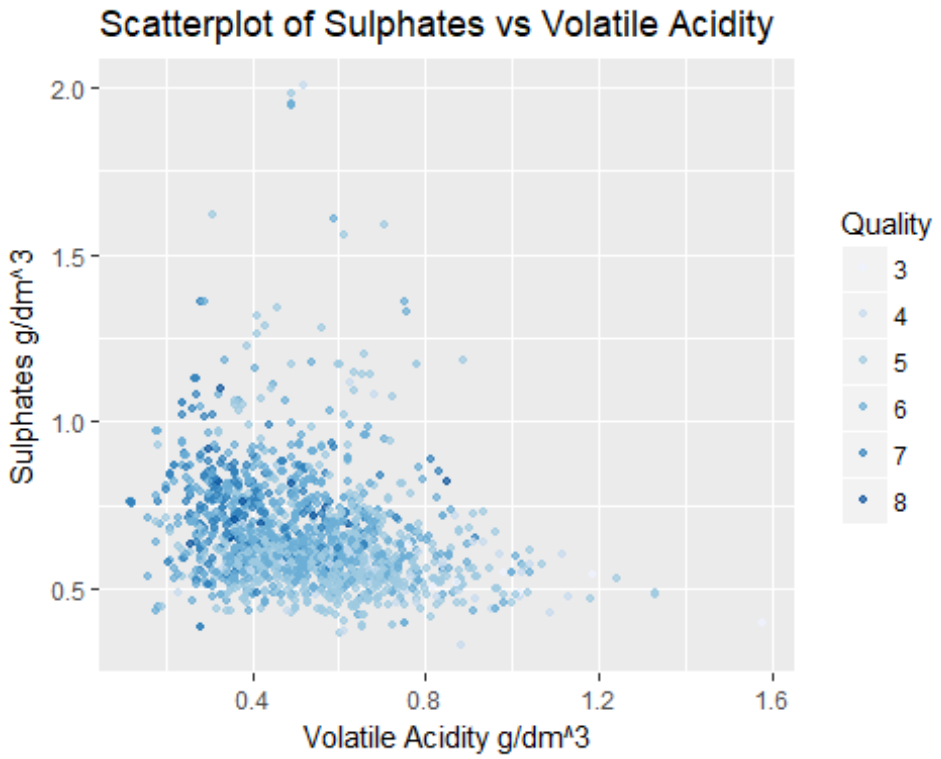
Multivariate Analysis:



Here we can see that for a given value of Fixed Acidity the lower density generally tends to yeild a better quality of wine.


Scatterplot of Alcohol vs Volatile Acidity
Alcohol % by volume
14
12
10
0.4
0.8
1.2
1.6
Volatile Acidity g/dm^3
Quality
3
4
5
6
7
8


Scatterplot of Alcohol vs Sulphates
Alcohol % by volume
14
12
10
0.5
1.0
1.5
2.0
Sulphates g/dm^3
Quality
3
4
5
6
7
8



This graph revalidates our claim that quality is somewhat linearly related to **Alcohol & Sulphates** and inversely related to **Volatile Acidity**

We can see that for first plot the quality increases with Alcohol and decreases with Volatile Acidity.

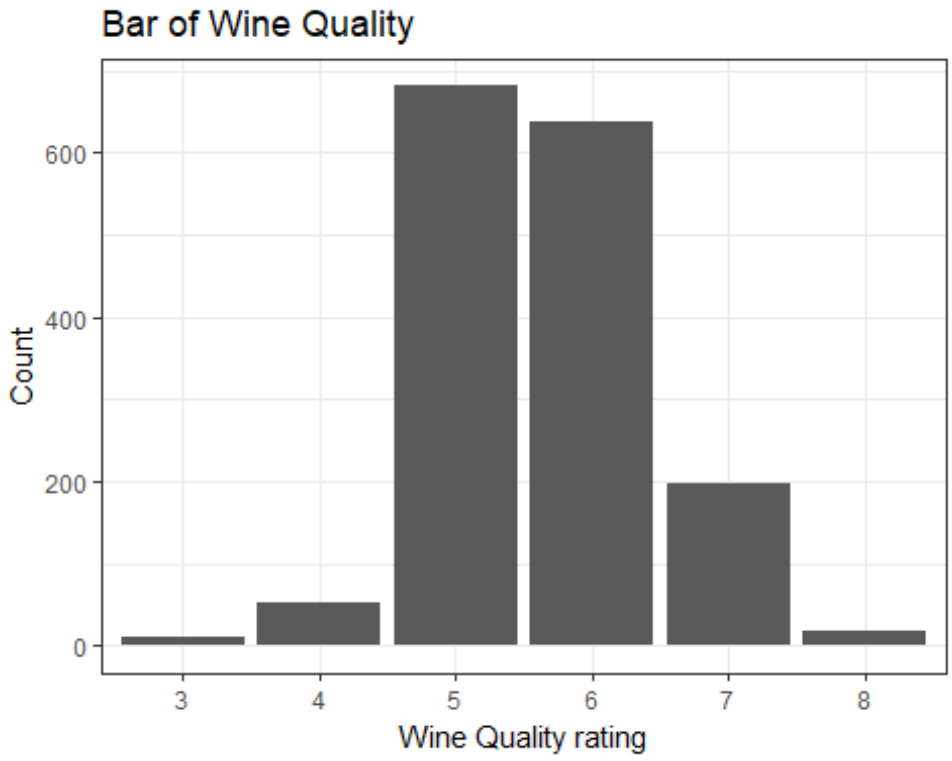
In Second plot Quality increases with Alcohol and sulphates.

Thrid plot it Increases with sulphates and decreases with Volatile Acidity.

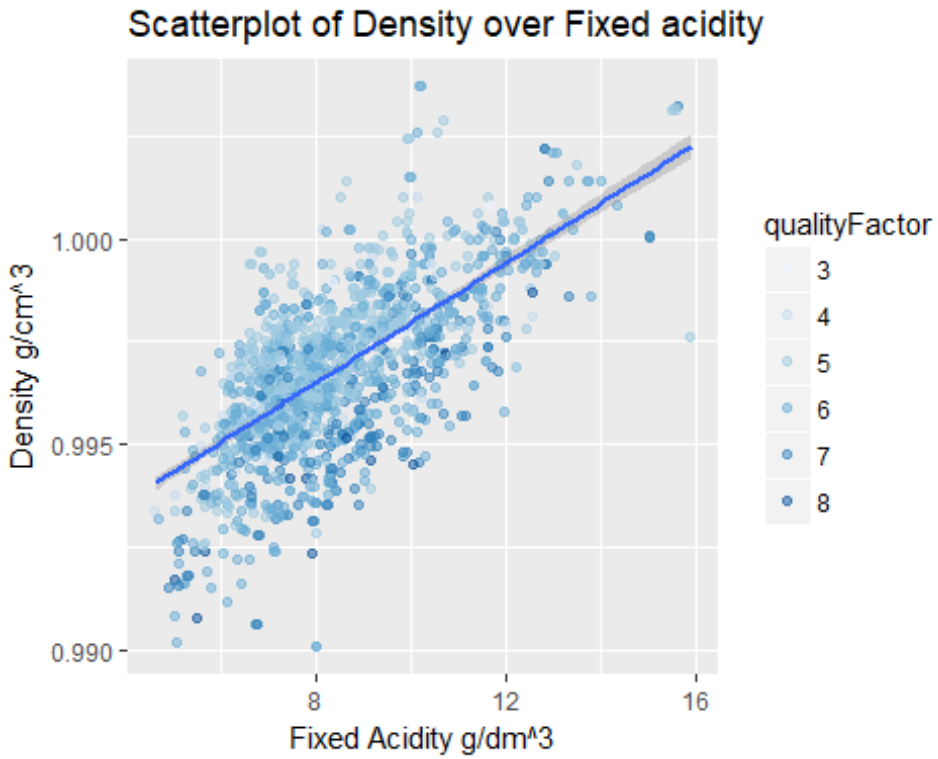
Final Plots & Summary

There are many interesting plots that we have come across during the analysis of redwine sample dataset.

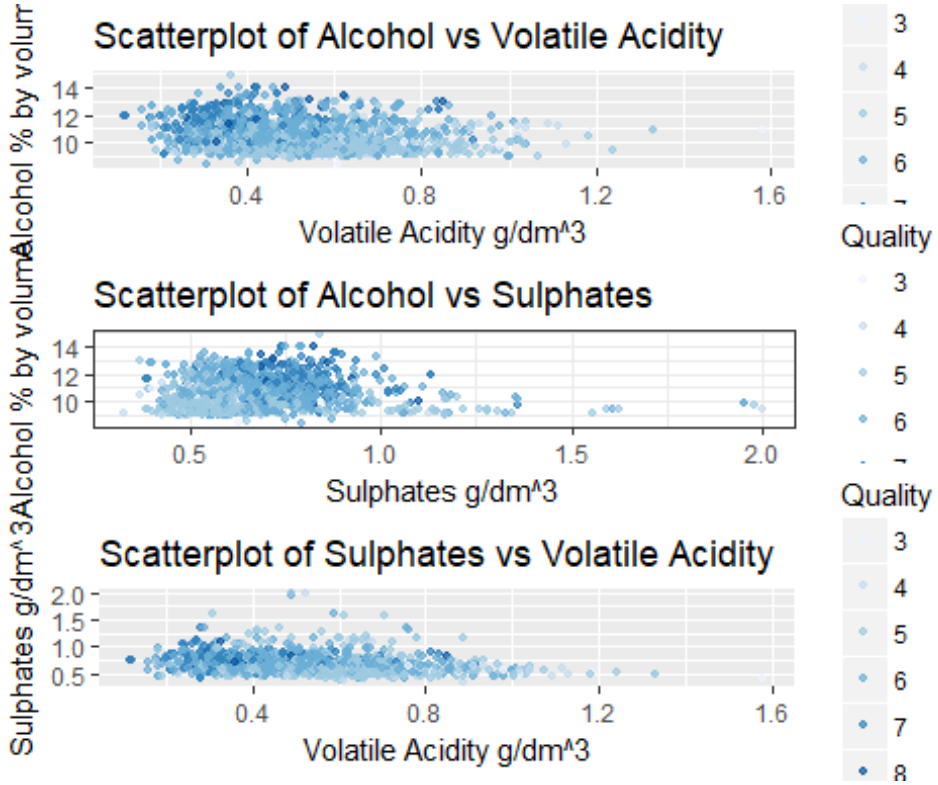
But these plots have helped us answer the question that we have raised at the beginning regarding the quality of red wine.



This plot helped us see that there are many medium quality red wines in our sample data set and the lowest rated quality sample is 3 and the highest rated quality sample is 8. There are no redwines with highest quality rating 10 or lowest quality rating 0.



The density and fixed acidity have a linear relationshio among them and these are most related in our data set. interesting enough we can see that the quality for a value of Fixed acidity decreases with an increase in density.



This plot helps to potray the chemical properties with strong/weak relationship to quality in our sample data set of red wines.

As we have found before __ % of Alcohol and amount of sulphates__ present in redwine tend to have positve relationship with quality of redwine. whereas **Amt of Volatile Acidity** tends to have inverse relationship with quality of red wine.

```
##
## Calls:
## m1: lm(formula = quality ~ alcohol, data = redWine)
## m2: lm(formula = quality ~ alcohol + log10(sulphates), data = redWine)
## m3: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity,
##     data = redWine)
## m4: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity +
##     log10(residual.sugar), data = redWine)
## m5: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity +
##     log10(residual.sugar) + log10(chlorides), data = redWine)
## m6: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity +
##     log10(residual.sugar) + log10(chlorides) +
log10(free.sulfur.dioxide1),
##     data = redWine)
## m7: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity +
##     log10(residual.sugar) + log10(chlorides) + log10(free.sulfur.dioxide1)
```

```
+
##     log10(total.sulfur.dioxide1), data = redWine)
## m8: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity +
##     log10(residual.sugar) + log10(chlorides) + log10(free.sulfur.dioxide1)
+
##     log10(total.sulfur.dioxide1) + density, data = redWine)
## m9: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity +
##     log10(residual.sugar) + log10(chlorides) + log10(free.sulfur.dioxide1)
+
##     log10(total.sulfur.dioxide1) + density + pH, data = redWine)
## m10: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity
+
##     log10(residual.sugar) + log10(chlorides) + log10(free.sulfur.dioxide1)
+
##     log10(total.sulfur.dioxide1) + density + pH, data = redWine)
## m11: lm(formula = quality ~ alcohol + log10(sulphates) + volatile.acidity
+
##     log10(residual.sugar) + log10(chlorides) + log10(free.sulfur.dioxide1)
+
##     log10(total.sulfur.dioxide1) + density + pH + citric.acid,
##     data = redWine)
##
##
================================================================================
================================================================================
================================
##                                        m1            m2            m3
m4            m5            m6            m7            m8            m9
m10           m11
## --------------------------------------------------------------------------------
--------------------------------------------------------------------------------
----------------------------------
##   (Intercept)                        1.875***      2.541***      3.369***
3.377***      3.062***      3.314***      3.923***     -1.636         8.747
8.747        -1.210
##                                     (0.175)       (0.177)       (0.184)
(0.185)       (0.202)       (0.270)       (0.304)      (12.737)      (12.929)
(12.929)      (14.041)
##   alcohol                           0.361***      0.335***      0.303***
0.304***      0.282***      0.279***      0.260***      0.266***
0.271***      0.271***      0.287***
##                                     (0.017)       (0.016)       (0.016)
(0.016)       (0.017)       (0.017)       (0.017)       (0.021)       (0.021)
(0.021)       (0.023)
##   log10(sulphates)                                   2.070***      1.477***
1.478***      1.714***      1.738***      1.788***      1.766***
1.770***      1.770***      1.778***
##                                                     (0.177)       (0.177)
(0.177)       (0.187)       (0.188)       (0.187)       (0.194)       (0.193)
(0.193)       (0.193)
```

```
##   volatile.acidity                                                -1.156***
-1.154***     -1.100***     -1.097***     -1.074***     -1.071***     -
0.952***     -0.952***     -1.064***
##                                                                   (0.097)
(0.097)       (0.098)       (0.098)       (0.098)       (0.098)       (0.102)
(0.102)       (0.119)
##   log10(residual.sugar)
-0.036         0.024         0.041         0.111         0.080         0.088
0.088         0.066
##
(0.106)       (0.107)       (0.107)       (0.108)       (0.129)       (0.128)
(0.128)       (0.129)
##   log10(chlorides)
-0.495***     -0.510***     -0.505***     -0.511***     -0.619***     -
0.619***     -0.590***
##
(0.129)       (0.129)       (0.129)       (0.130)       (0.132)       (0.132)
(0.132)
##   log10(free.sulfur.dioxide1)
-0.078         0.221*        0.226*        0.272**       0.272**
0.252**
##
(0.056)       (0.089)       (0.090)       (0.090)       (0.090)       (0.091)
##   log10(total.sulfur.dioxide1)
-0.383***     -0.383***     -0.414***     -0.414***     -0.388***
##
(0.090)       (0.090)       (0.090)       (0.090)       (0.091)
##   density
5.505        -3.566        -3.566         6.747
##
(12.610)      (12.745)      (12.745)      (13.950)
##   pH
-0.486***     -0.486***     -0.591***
##
(0.119)       (0.119)       (0.133)
##   citric.acid
-0.245
##
(0.135)
## ------------------------------------------------------------------------
------------------------------------------------------------------------
-----------------------------------
##   R-squared                          0.227         0.288         0.345
0.346         0.352         0.352         0.360         0.360         0.366
0.366         0.368
##   adj. R-squared                     0.226         0.287         0.344
0.344         0.350         0.350         0.357         0.356         0.363
0.363         0.364
##   sigma                              0.710         0.682         0.654
0.654         0.651         0.651         0.648         0.648         0.645
```

```
0.645         0.644
##   F                                  468.267       322.031       280.646
210.397       172.715       144.349       127.638       111.651       102.059
102.059        92.313
##   p                                    0.000         0.000         0.000
0.000         0.000         0.000         0.000         0.000         0.000
0.000         0.000
##   Log-likelihood                -1721.057     -1655.601     -1587.752
-1587.693     -1580.332     -1579.336     -1570.300     -1570.204     -
1561.907     -1561.907     -1560.257
##   Deviance                           805.870       742.522       682.108
682.058       675.807       674.965       667.380       667.300       660.410
660.410       659.049
##   AIC                           3448.114      3319.202      3185.503
3187.386      3174.664      3174.671      3158.600      3160.409
3145.814      3145.814      3144.513
##   BIC                           3464.245      3340.711      3212.389
3219.649      3212.304      3217.688      3206.994      3214.180
3204.962      3204.962      3209.039
##   N                             1599          1599          1599
1599          1599          1599          1599          1599          1599
1599          1599
##
================================================================================
================================================================================
================================
```

I have perfromed a multiple regression on all the available chemical properties againts quality to find which chemical properties offer maximum variance i.e., r^2 value interesting enough i found that Alcohol, volatile acidity and sulphates offer upto 34.5% variance and the overall value of r^2 is 36.8% which indicates that the other chemical properties govern little variance of quality.

This answers over question earlier that properties Alcohol, volatile acidity and sulphates account for quality of redwine but these have a very small influence over the rating of redwine none the less offer something when compared to other chemical properties.

Reflection:

As the first data set available in the projects page I found myself tending towrds this one. And it was taking about redwine of which I have heard plenty of healthy benifits about and I myself started drinking in right does to get the benifits. so, i was interested about what chemical properties are in redwine and how they all govern quality.

EDA module in udacity and problem sets for analysis of univariate, bivariate and multivariate variables availabe in our data set made me familiar with the process of exploratory data analysis.

when I initially started looking at the data and performed univariate analysis nothing excited me than I was not understanding anything. I later tracked back to basics that I have to get familiar with documentation then can figure out whats happening with the variables.

This helped me and those graphs started making sense. Another tricky part was that I thought there will be a clear relationship with chemical properties and quality expect few but that was not the case. then realized whats said in theory only reflects a bit practically.

When I have drawn scatterplot for chemical properties with high correlation coefficient the graphs were not showing a trend I was expecting than I figured out that boxplot will help showing the relationship.

I have initially missed that quality is ordinal and i have to use sequential color pallete and bar plot is used to dipict an ordinal variable.

I feel like everything was straight forward except finding the relationship. programming with r is simple and easy thanks to practice problem sets.