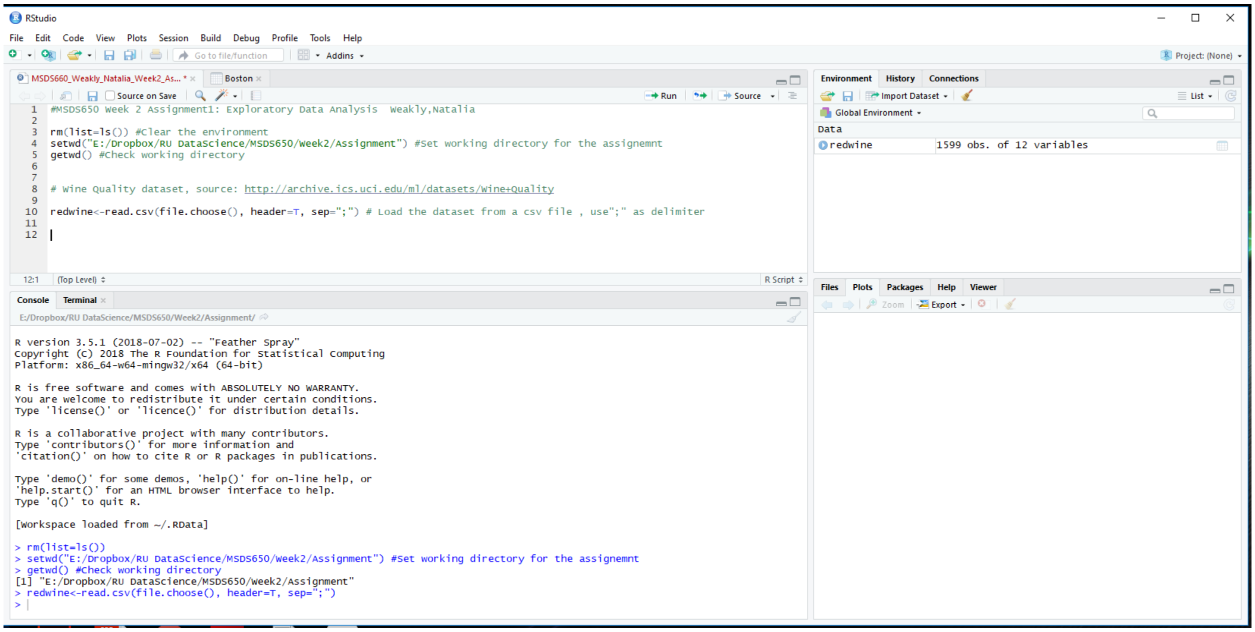
Data Set Description

For this assignment I chose to use the ***Wine Quality*** dataset that I downloaded from http://archive.ics.uci.edu/ml/datasets/Wine+Quality .as a csv file.

I loaded the dataset into R using the following command:

redwine<-read.csv(file.choose(), header=T, sep=";") # Load the dataset from a csv file , use";" as delimiter

Which loaded a dataset containing 1599 observations with 12 variables, as shown below:



```r
#MSDS650 Week 2 Assignment1: Exploratory Data Analysis  Weakly,Natalia

rm(list=ls()) #Clear the environment
setwd("E:/Dropbox/RU DataScience/MSDS650/Week2/Assignment") #Set working directory for the assignemnt
getwd() #Check working directory


# Wine Quality dataset, source: http://archive.ics.uci.edu/ml/datasets/Wine+Quality

redwine<-read.csv(file.choose(), header=T, sep=";") # Load the dataset from a csv file , use";" as delimiter

```

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> rm(list=ls())
> setwd("E:/Dropbox/RU DataScience/MSDS650/Week2/Assignment") #Set working directory for the assignemnt
> getwd() #Check working directory
[1] "E:/Dropbox/RU DataScience/MSDS650/Week2/Assignment"
> redwine<-read.csv(file.choose(), header=T, sep=";")
> 
```

```
Data
redwine          1599 obs. of 12 variables
```

According to the meta data file, this data set was created by Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) in 2009 to study wine preferences. The inputs include objective tests (e.g. PH values) and the output variable is based on sensory data (median of at least three evaluations made by wine experts). Each wine expert graded the wine quality between 0 (very bad) and 10 (very excellent).

Overall, the following 12 variables are included in the dataset:

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density

9. pH
10. sulphates
11. alcohol
12. quality

Objective

Using this data set I would like to explore wine quality ratings and factors that might influence them. In order to narrow down the initial data analysis, I will mostly focus on three variables – level of sulphates, alcohol content, and quality rating.

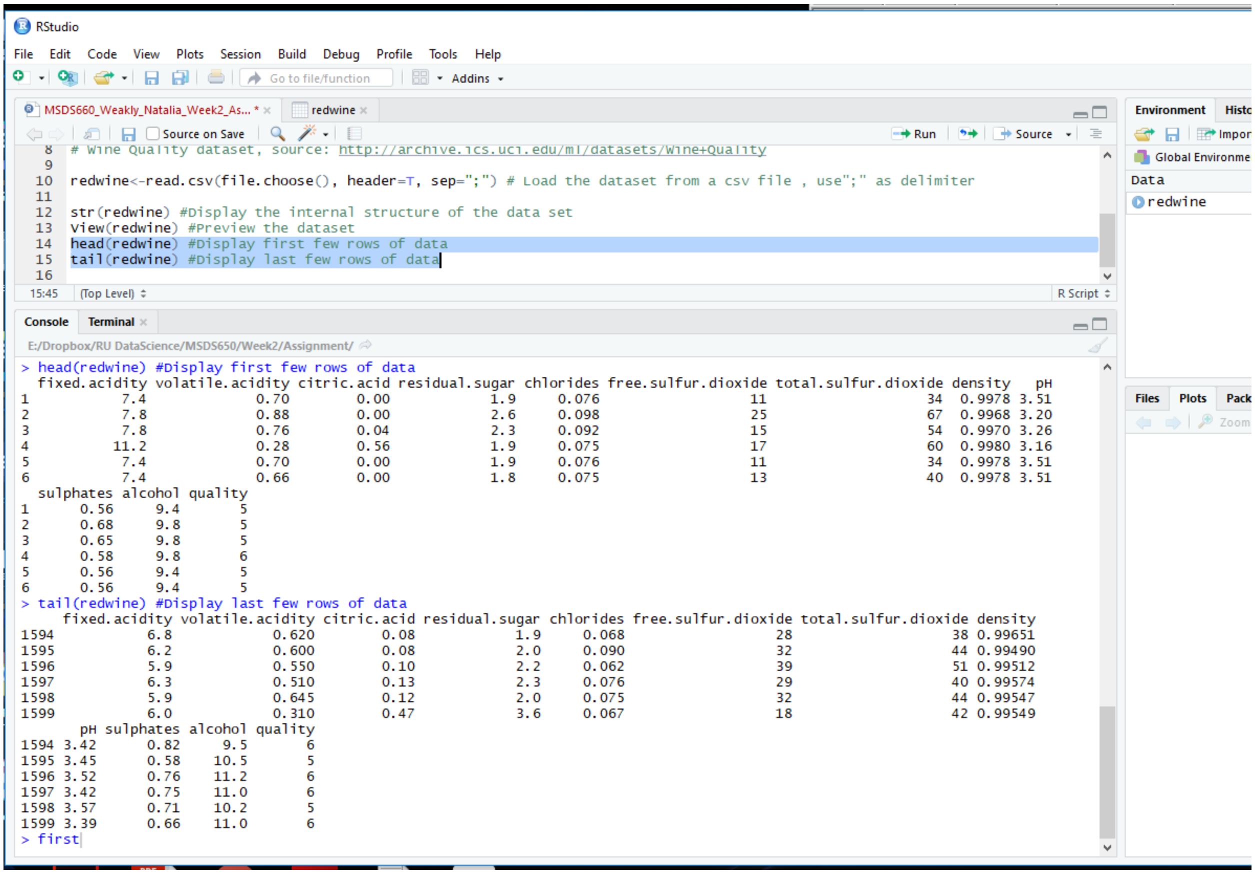
Exploratory Data Analysis in R

I used the ***str(redwine)*** command again to display the internal structure, which provided the following output:

```
> str(redwine)
'data.frame':	1599 obs. of  12 variables:
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

This command returns data types and a few first values for each variable. All but one of the variables use num type, quality variable is int data type.

In order to look at the sample data, I used ***head(redwine)*** and ***tail(redwine)*** commands with the sample output as follows:



```
# Wine Quality dataset, source: http://archive.ics.uci.edu/ml/datasets/Wine+Quality

redwine<-read.csv(file.choose(), header=T, sep=";") # Load the dataset from a csv file , use";" as delimiter

str(redwine) #Display the internal structure of the data set
View(redwine) #Preview the dataset
head(redwine) #Display first few rows of data
tail(redwine) #Display last few rows of data

```

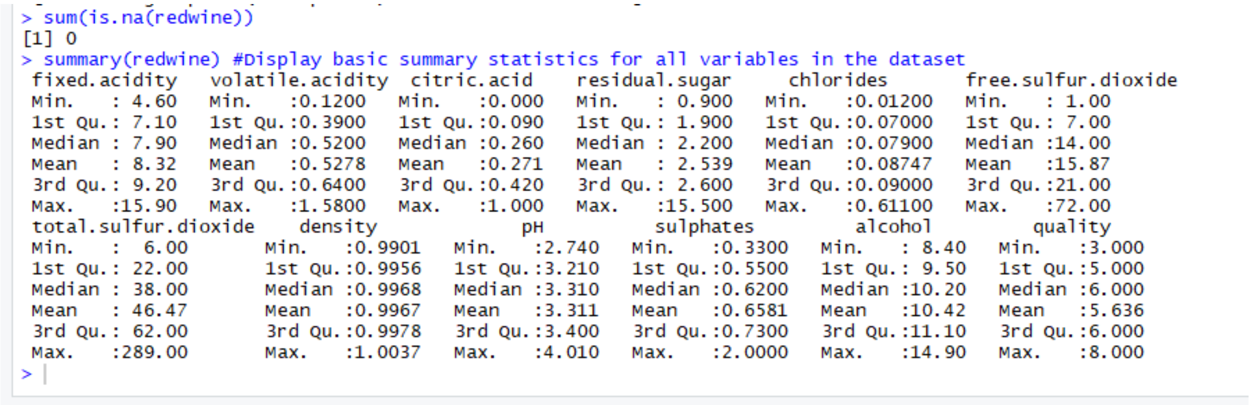
```
> head(redwine) #Display first few rows of data
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH
1           7.4             0.70        0.00            1.9     0.076                  11                   34  0.9978 3.51
2           7.8             0.88        0.00            2.6     0.098                  25                   67  0.9968 3.20
3           7.8             0.76        0.04            2.3     0.092                  15                   54  0.9970 3.26
4          11.2             0.28        0.56            1.9     0.075                  17                   60  0.9980 3.16
5           7.4             0.70        0.00            1.9     0.076                  11                   34  0.9978 3.51
6           7.4             0.66        0.00            1.8     0.075                  13                   40  0.9978 3.51
  sulphates alcohol quality
1      0.56     9.4       5
2      0.68     9.8       5
3      0.65     9.8       5
4      0.58     9.8       6
5      0.56     9.4       5
6      0.56     9.4       5
> tail(redwine) #Display last few rows of data
     fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density
1594           6.8            0.620        0.08            1.9     0.068                  28                   38 0.99651
1595           6.2            0.600        0.08            2.0     0.090                  32                   44 0.99490
1596           5.9            0.550        0.10            2.2     0.062                  39                   51 0.99512
1597           6.3            0.510        0.13            2.3     0.076                  29                   40 0.99574
1598           5.9            0.645        0.12            2.0     0.075                  32                   44 0.99547
1599           6.0            0.310        0.47            3.6     0.067                  18                   42 0.99549
       pH sulphates alcohol quality
1594 3.42      0.82     9.5       6
1595 3.45      0.58    10.5       5
1596 3.52      0.76    11.2       6
1597 3.42      0.75    11.0       6
1598 3.57      0.71    10.2       5
1599 3.39      0.66    11.0       6
> first
```

Visual inspection using View(redwine) did not show any missing values, but in order to confirm that I used ***sum(is.na(redwine))***, with the following output:

```
> sum(is.na(redwine))
[1] 0
```

So, I have not detected any irregularities in the data requiring clean-up, so I proceeded with the exploratory analysis. I used ***summary(redwine)*** to display basic summary statistics for all variables in the dataset. It returned the following:

Exploratory Data Analysis



```
> sum(is.na(redwine))
[1] 0
> summary(redwine) #Display basic summary statistics for all variables in the dataset
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides       free.sulfur.dioxide
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200   Min.   : 1.00     
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00     
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00     
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747   Mean   :15.87     
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00     
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100   Max.   :72.00     
 total.sulfur.dioxide    density             pH          sulphates         alcohol          quality     
 Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000  
 1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000  
 Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200   Median :10.20   Median :6.000  
 Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636  
 3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000  
 Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000  
> 
```

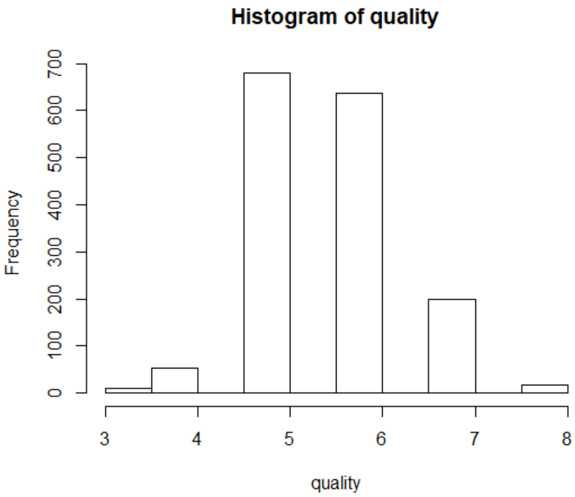
The quality variable, representing expert opinions on a scale from 1 to 10, for this data set actually varies from 3.0 (min value) to a maximum of 8.0, with the median value of 6.00. The middle 50% of ratings (interquartile range) are between 5.0 and 6.0.

Sulphates measurements range from 0.33 to 2.00 with the mean value of 0.6581 and slightly lower median of 0.6581, this along with the IQR information (first quartile 0.55, and third quartile 0.73) suggest that the sulphates distribution might be skewed to the right.

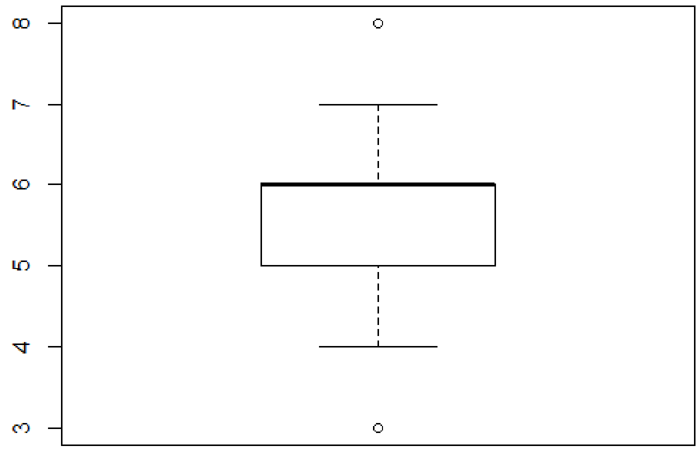
The wine sample included in the dataset contained from 8.40% to 14.90% percent of alcohol, with the mean of 10.42% and median of 10.20%.

Then, I used visual tools to look at the distributions of those variables a little closer.

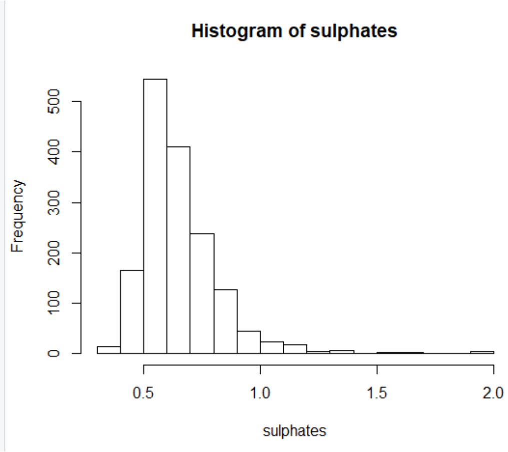
The ***hist(quality)*** command showed frequencies distributions for all ratings, making it apparent that the vast majority of the samples received ratings of 5 and 6. Experts apparently were avoiding both extremes .



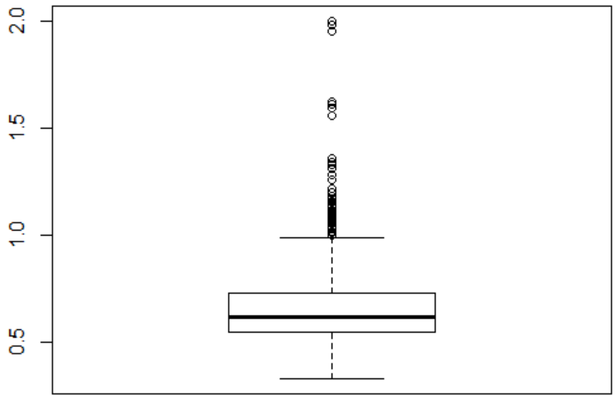
Displaying simple boxplot, using ***boxplot(quality)*** command, confirmed that ratings of 3 and 8 were treated as outliers.



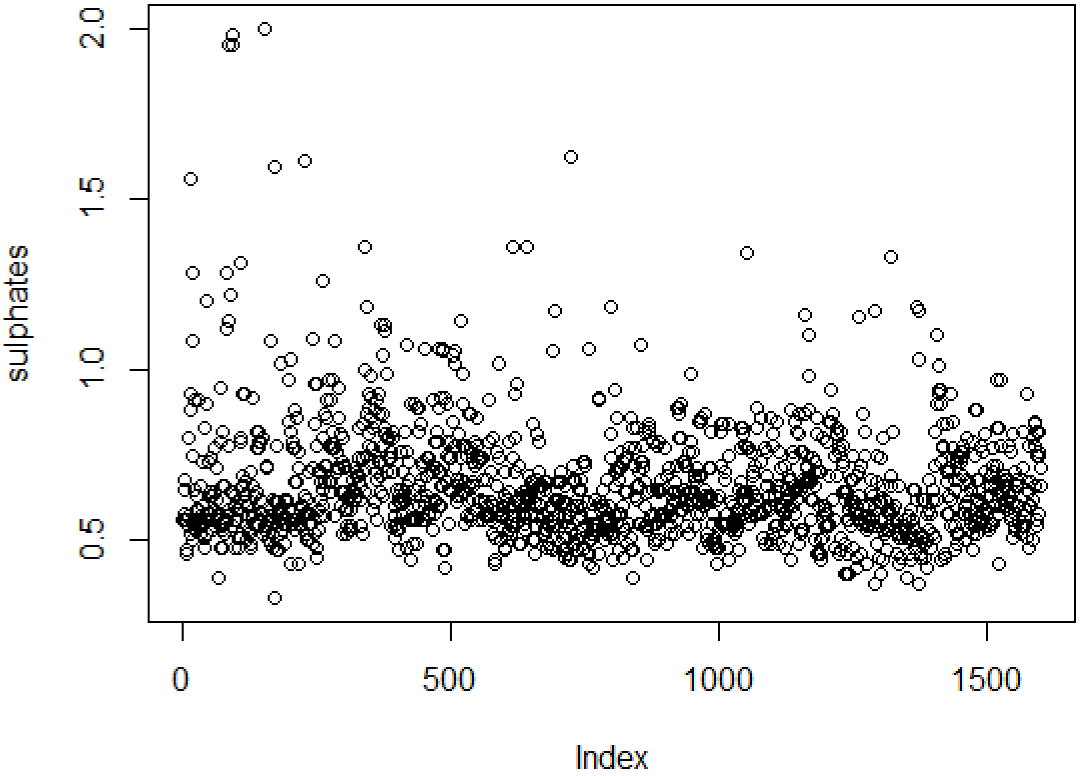
Displaying a histogram for sulphates variable (***hist(sulphates))*** confirmed right-skewed distribution suggested earlier by the comparison of basic statistics.



Using ***boxplot(sulphates)*** helped visualizing sulphated distribution and the large number of outliers, samples with the sulphate content between 1.00 and 2.00.



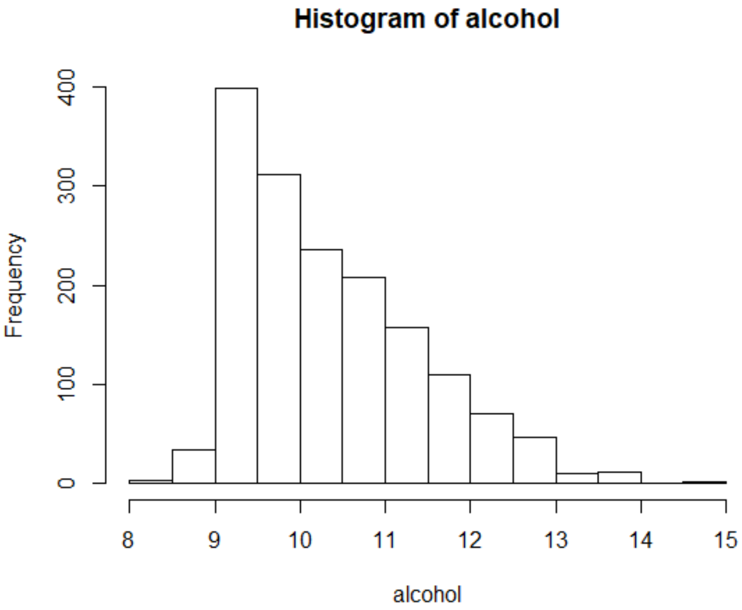
Scatter plot (***plot(sulphates)*** command) showed a relatively narrow band of varying sulphates values.



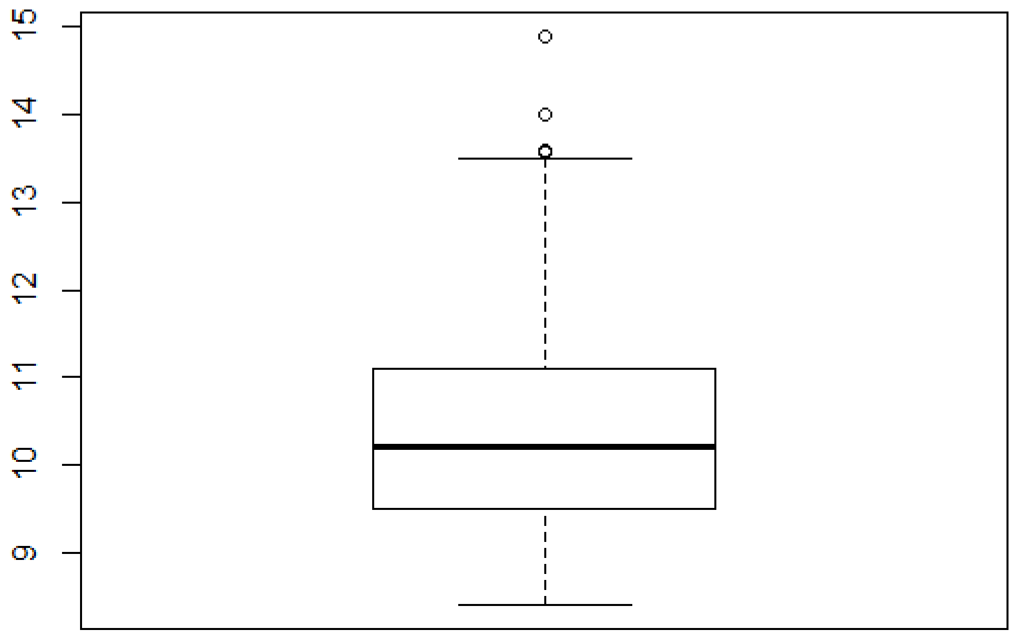
The mean value for sulphates, returned earlier by the summary function, is 0.6581 and the standard deviation is 0.169507. The distribution has a very long right tail – the maximum values for sulphates (2.0) lay more than 7.9 standard deviations away from the mean ((2.0-0.6581)/0.169507).

```
> sd(sulphates) #Standard deviation for sulphates
[1] 0.169507
```

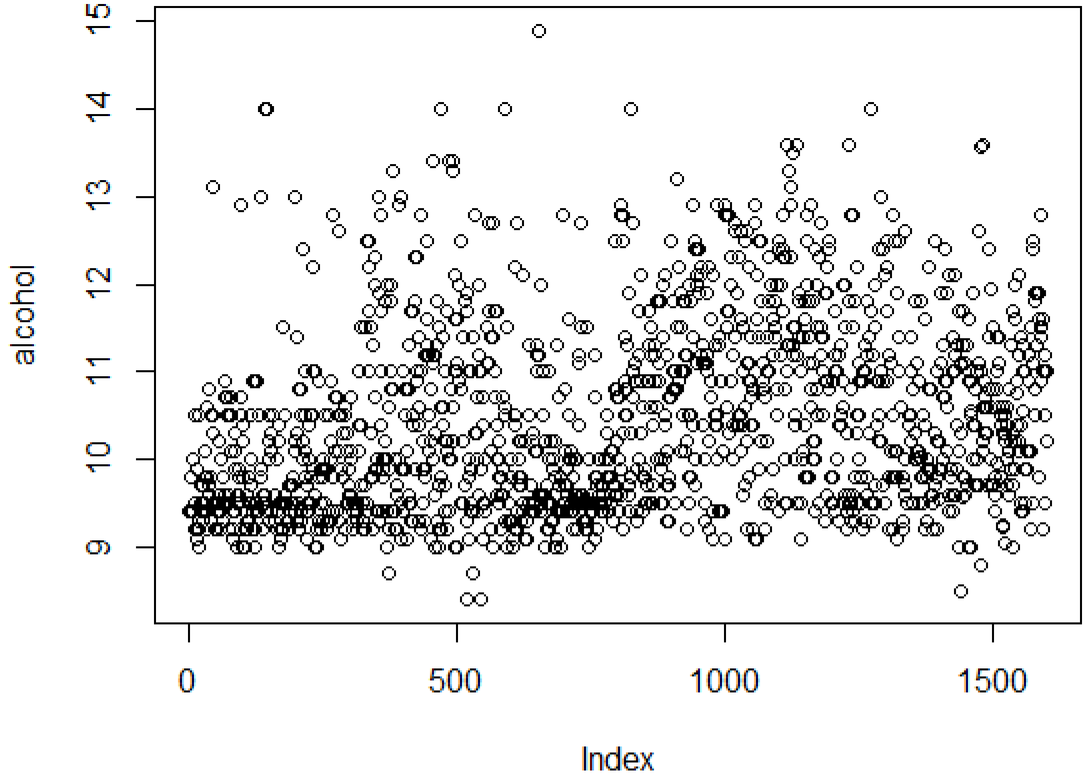
Histogram for the alcohol content (***hist(alcohol)***) showed that the distribution is not normal, it is substantially right-skewed.



Displaying a boxplot helped visualize that the middle 50% of the samples had from 9.50% to a little over 11% alcohol content with a large number of samples with higher alcohol content.



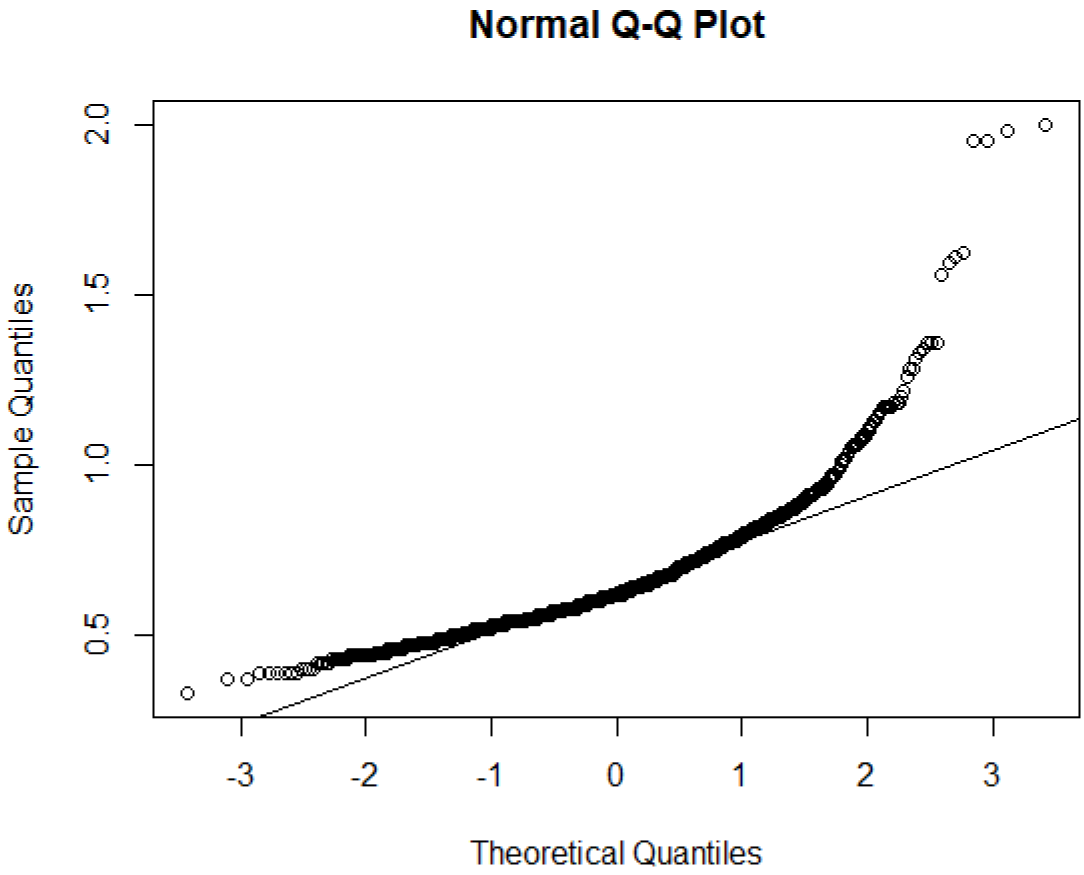
The command ***plot(alcohol)*** displayed scatter plot for the alcohol variable, confirming that relatively few samples had alcohol content below 9.00%.



Previous steps suggested that sulphates and alcohol distributions significantly deviate from normal, but I used normal QQ plots to confirm it.

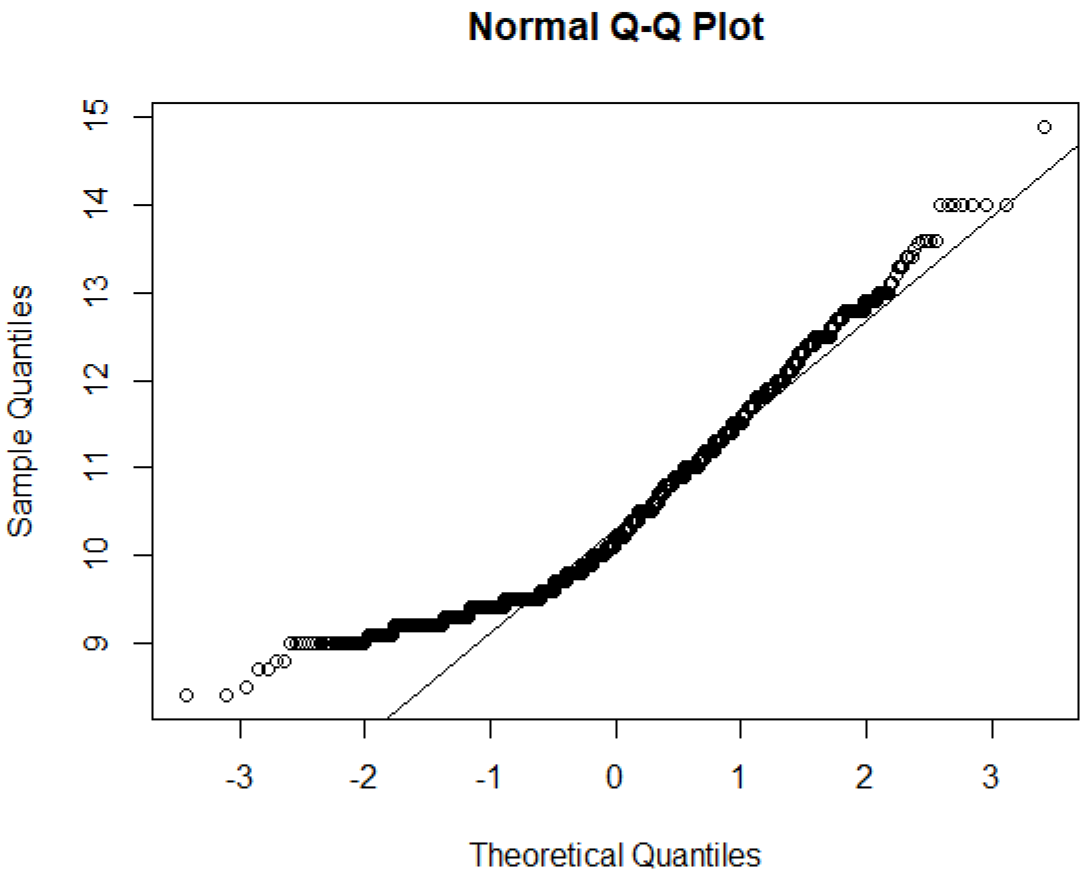
```
> qqnorm(sulphates)#Norm Q-Q plot for sulphates values
> qqline(sulphates)
```

Output for sulphates:



```
> qqnorm(alcohol)#Norm Q-Q plot for alcohol values
> qqline(alcohol)
```

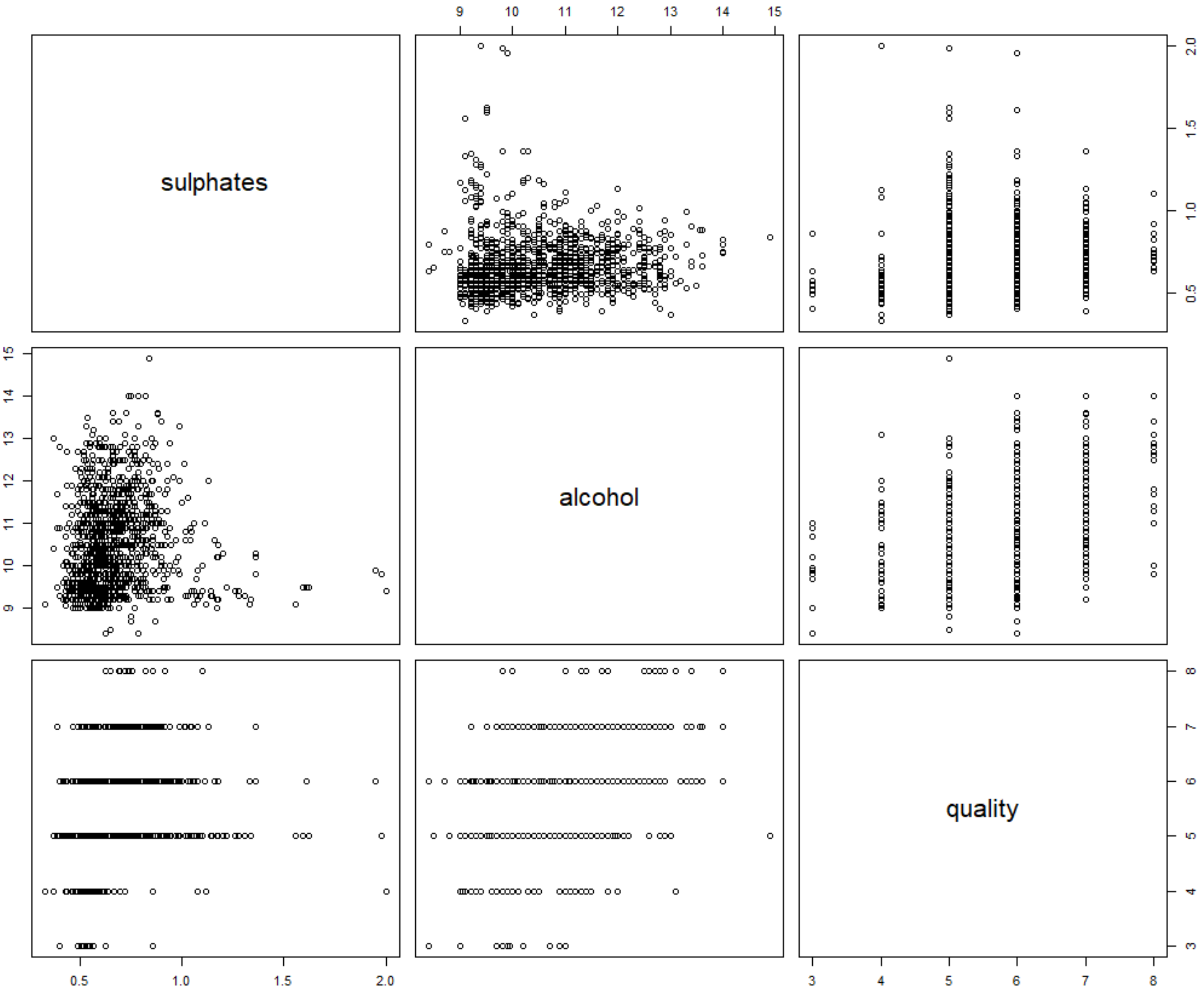
Output for alcohol:



In order to try to detect any possible trends in the relationships between these three variables I created a scatter plot matrix using ***pairs()*** command.

```
> pairs(~sulphates+alcohol+quality, data=redwine) #Display scatterplot matrix for three variables
```

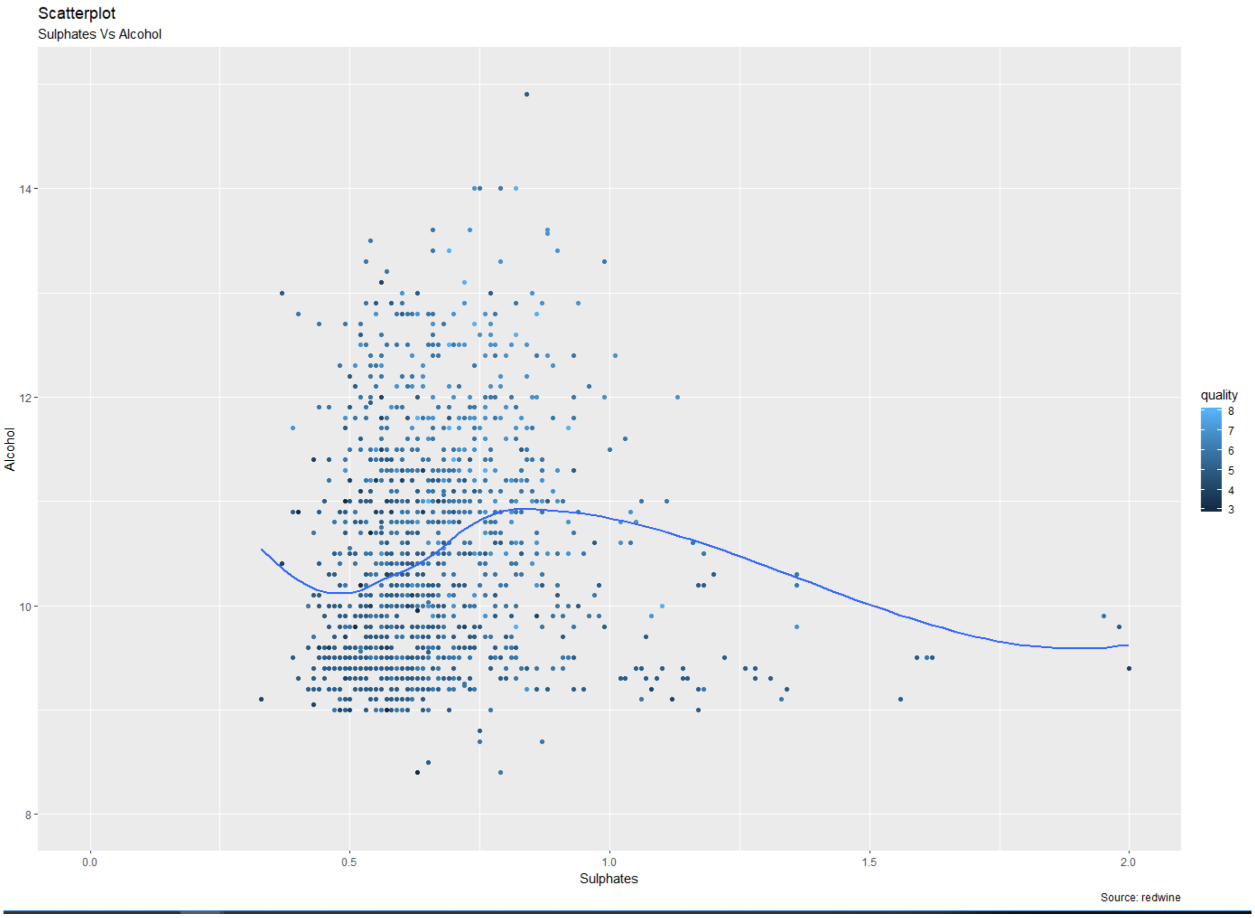
Exploratory Data Analysis



Unfortunately, visual analysis of this matrix did not yield any significant trends.

Since neither pair of the variables shoed significant rends, I tried to plot all three variables on the same graph using ggplot2 capabilities looking for more insights.

```
> gg <- ggplot(redwine, aes(x=sulphates, y=alcohol)) +
+   geom_point(aes(col=quality)) +
+   geom_smooth(method="loess", se=F) +
+   xlim(c(0, 2.0)) +
+   ylim(c(8, 15)) +
+   labs(subtitle="Sulphates Vs Alcohol",
+        y="Alcohol",
+        x="Sulphates",
+        title="Scatterplot",
+        caption = "Source: redwine")
>
> plot(gg)
```

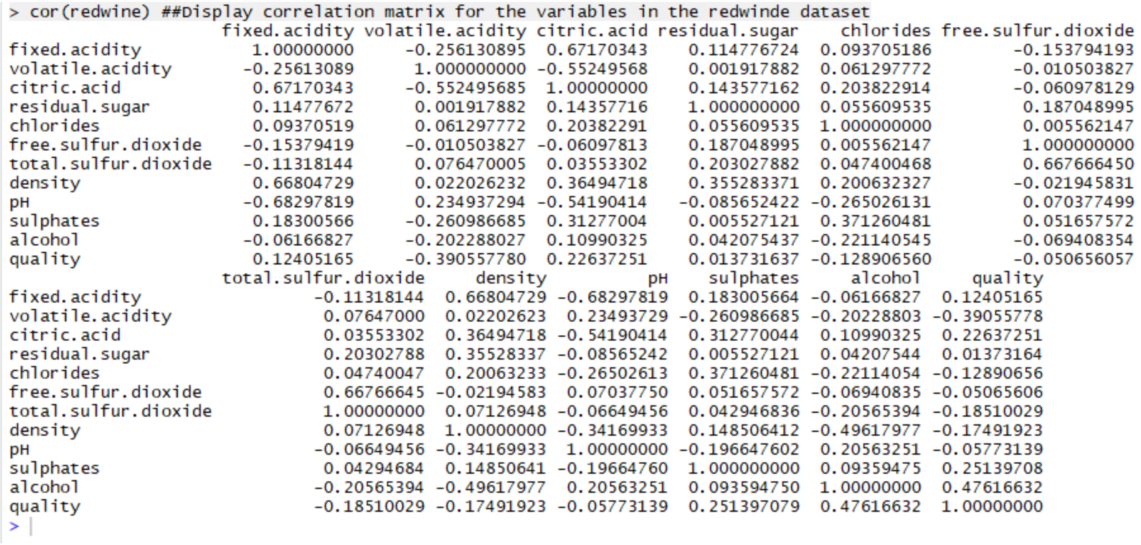


Unfortunately, again this graph did not reveal any significant relationship between sulphates and alcohol content and wine rating.

I looked at the correlation matrix for the variables using ***cor(redwine)*** command:

```
> cor(redwine) ##Display correlation matrix for the variables in the redwinde dataset
```

It provided the following output:



```
> cor(redwine) ##Display correlation matrix for the variables in the redwinde dataset
                     fixed.acidity volatile.acidity citric.acid residual.sugar    chlorides free.sulfur.dioxide
fixed.acidity           1.00000000     -0.256130895  0.67170343    0.114776724  0.093705186        -0.153794193
volatile.acidity       -0.25613089      1.000000000 -0.55249568    0.001917882  0.061297772        -0.010503827
citric.acid             0.67170343     -0.552495685  1.00000000    0.143577162  0.203822914        -0.060978129
residual.sugar          0.11477672      0.001917882  0.14357716    1.000000000  0.055609535         0.187048995
chlorides               0.09370519      0.061297772  0.20382291    0.055609535  1.000000000         0.005562147
free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813    0.187048995  0.005562147         1.000000000
total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302    0.203027882  0.047400468         0.667666450
density                 0.66804729      0.022026232  0.36494718    0.355283371  0.200632327        -0.021945831
pH                     -0.68297819      0.234937294 -0.54190414   -0.085652422 -0.265026131         0.070377499
sulphates               0.18300566     -0.260986685  0.31277004    0.005527121  0.371260481         0.051657572
alcohol                -0.06166827     -0.202288027  0.10990325    0.042075437 -0.221140545        -0.069408354
quality                 0.12405165     -0.390557780  0.22637251    0.013731637 -0.128906560        -0.050656057
                     total.sulfur.dioxide     density          pH    sulphates     alcohol     quality
fixed.acidity                 -0.11318144  0.66804729 -0.68297819  0.183005664 -0.06166827  0.12405165
volatile.acidity               0.07647000  0.02202623  0.23493729 -0.260986685 -0.20228803 -0.39055778
citric.acid                    0.03553302  0.36494718 -0.54190414  0.312770044  0.10990325  0.22637251
residual.sugar                 0.20302788  0.35528337 -0.08565242  0.005527121  0.04207544  0.01373164
chlorides                      0.04740047  0.20063233 -0.26502613  0.371260481 -0.22114054 -0.12890656
free.sulfur.dioxide            0.66766645 -0.02194583  0.07037750  0.051657572 -0.06940835 -0.05065606
total.sulfur.dioxide           1.00000000  0.07126948 -0.06649456  0.042946836 -0.20565394 -0.18510029
density                        0.07126948  1.00000000 -0.34169933  0.148506412 -0.49617977 -0.17491923
pH                            -0.06649456 -0.34169933  1.00000000 -0.196647602  0.20563251 -0.05773139
sulphates                      0.04294684  0.14850641 -0.19664760  1.000000000  0.09359475  0.25139708
alcohol                       -0.20565394 -0.49617977  0.20563251  0.093594750  1.00000000  0.47616632
quality                       -0.18510029 -0.17491923 -0.05773139  0.251397079  0.47616632  1.00000000
> 
```

This correlation coefficients confirmed the absence of strong correlation between the three variables included in the analysis (alcohol vs. quality 0.47616632, sulphates vs. quality 0.25139708, alcohol vs. sulphates 0.09359475).

<u>Conclusions:</u>

Exploratory analysis of the three variables (sulphates, alcohol and quality) did not reveal any significant trends in the data. Neither sulphates level nor alcohol separately, or in combination could be used as a significant predictor of subjective rating of red wine quality in this dataset. Other variables need to be included in future analysis.

Personally, I was surprised with a lack of significant relationship between the level of sulphates in the samples and wine quality ratings. Before the analysis, I expected a negative correlation, with higher quality wines having lower sulphates levels (no added sulphates).

The distribution of quality ratings with majority of the samples receiving 5 and 6 was also unexpected. It might be due to the data collection mechanism (average of at least three expert ratings), or it might potentially show the difficulty of blind wine ratings even for experts.