White Wine Dataset Exploratory Data Analysis

Introduction

The White Wine dataset is a public available dataset. The dataset is related to white variaants of Portuguese "Vinho Verde" wine.

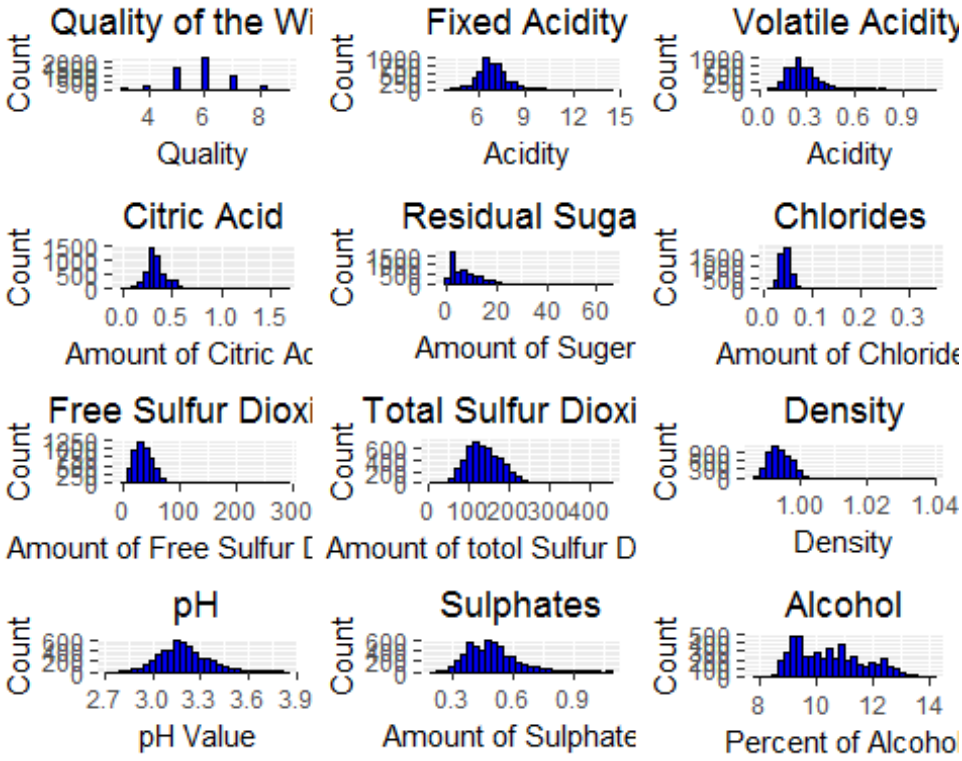
```
## [1] 4898   12

## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.04
4 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...

##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median :0.04300   Median : 34.00      Median :134.0
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

The dataset contains 4898 white wine instances. There are 12 variables. 11 of them are input variables based on physicochemical tests and their data type are numeric.

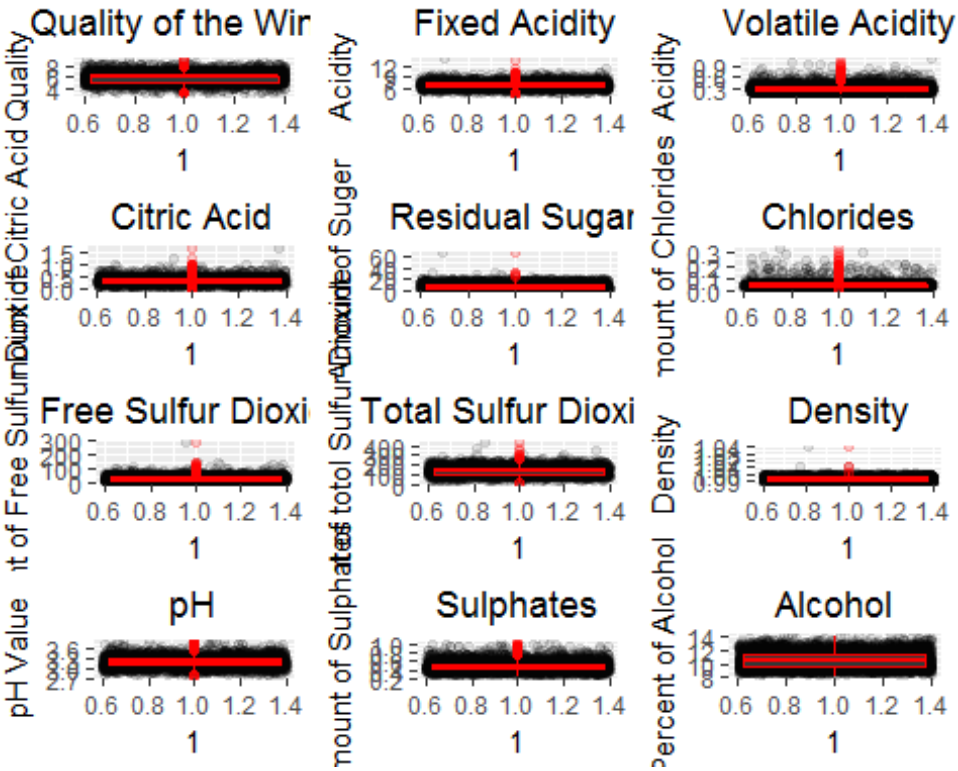
Input variables (based on physicochemical tests):
Fixed Acidity: Amount of tartaric acid in the wine
Volatile Acidity: Amount of acetic acid in wine
Citric Acid: Amount of citric acid in the wine, it is usually used to add 'freshness' and flavor to wines
Residual Sugar: Amount of sugar remaining after fermentation stops
Chlorides: Amount of salt in the wine
Free Sulfur Dioxide: Amount of free form of sulfur dioxide gas, it can prevent microbial growth and the oxidation of wine
Total Sulfur Dioxide: Amount of free and bound forms of sulfur dioxide gas
Density: Density of water
pH: Acidicity or basicity of the wine
Sulphates: Amount of sulphates in wine, it is a wine additive used as an antimicrobial and antioxidant
Alcohol: Percent of alcohol content of the wine

Output variable (based on sensory data):
Quality: Qualiy of the wine

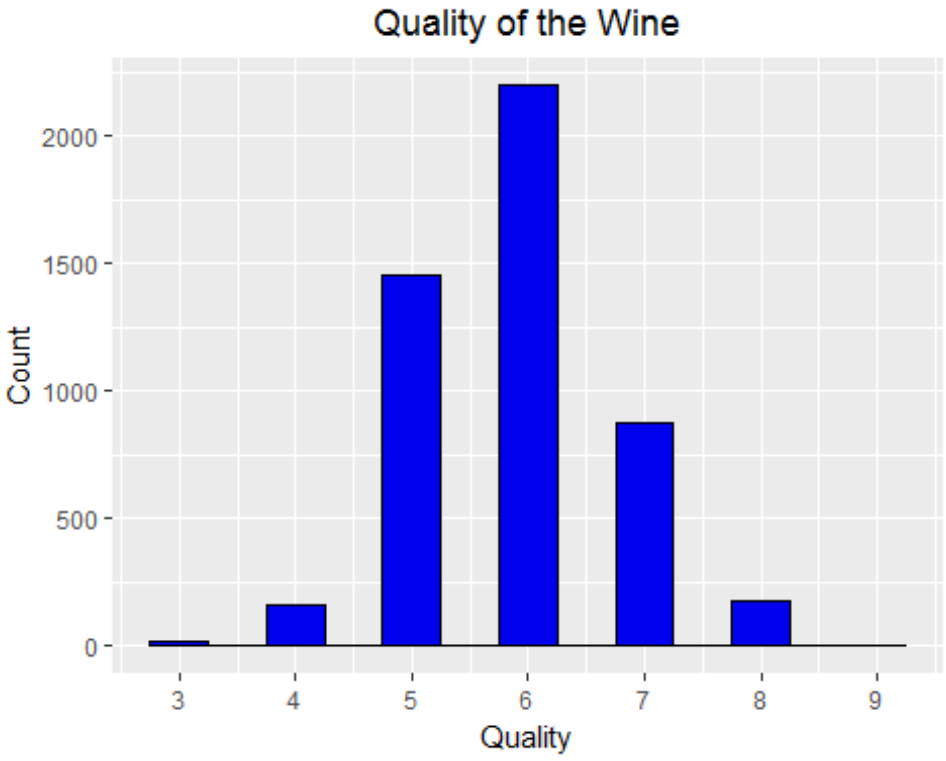
Univariate Plots Section



Looking at the histograms, I find that some variables, such as residual sugar, chlorides and alcohol, are right-skewed. I also find that some variables contain outliners. In order to take a closer look, I created some boxplots.



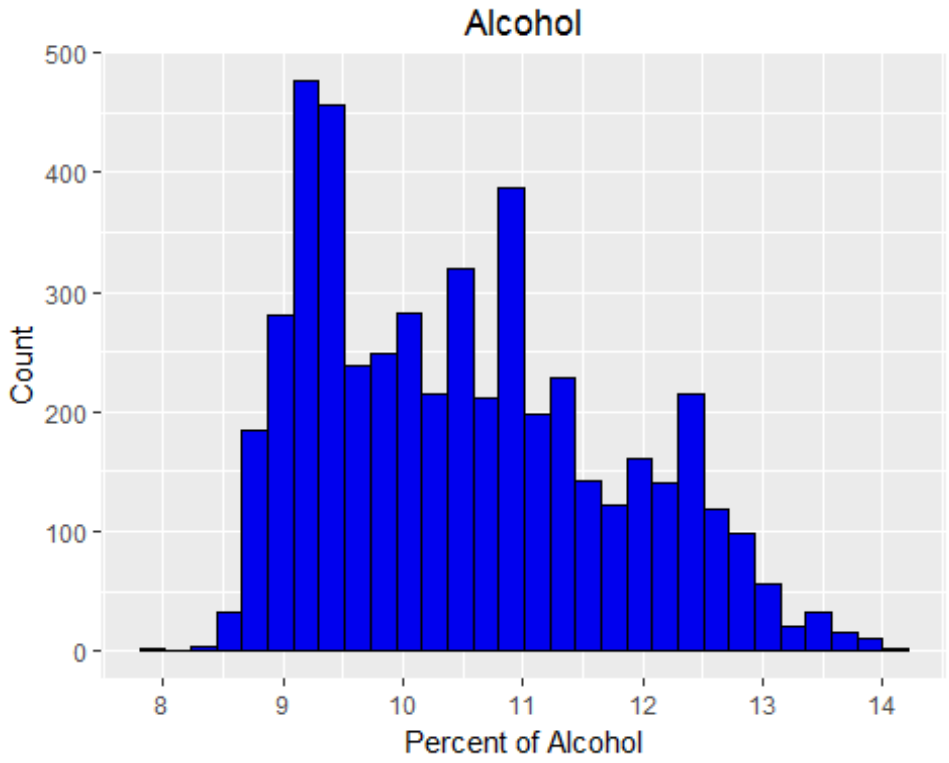
According to the boxplot, it is clear that a number of variables contain outliners. And among these variables, variables such as volatile acidity, chlorides and sulphates contains relatively more outliers than the others.



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.878   6.000   9.000

##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5

##
##           3           4           5           6           7           8
## 0.004083299 0.033278889 0.297468354 0.448754594 0.179665169 0.035728869
##           9
## 0.001020825
```

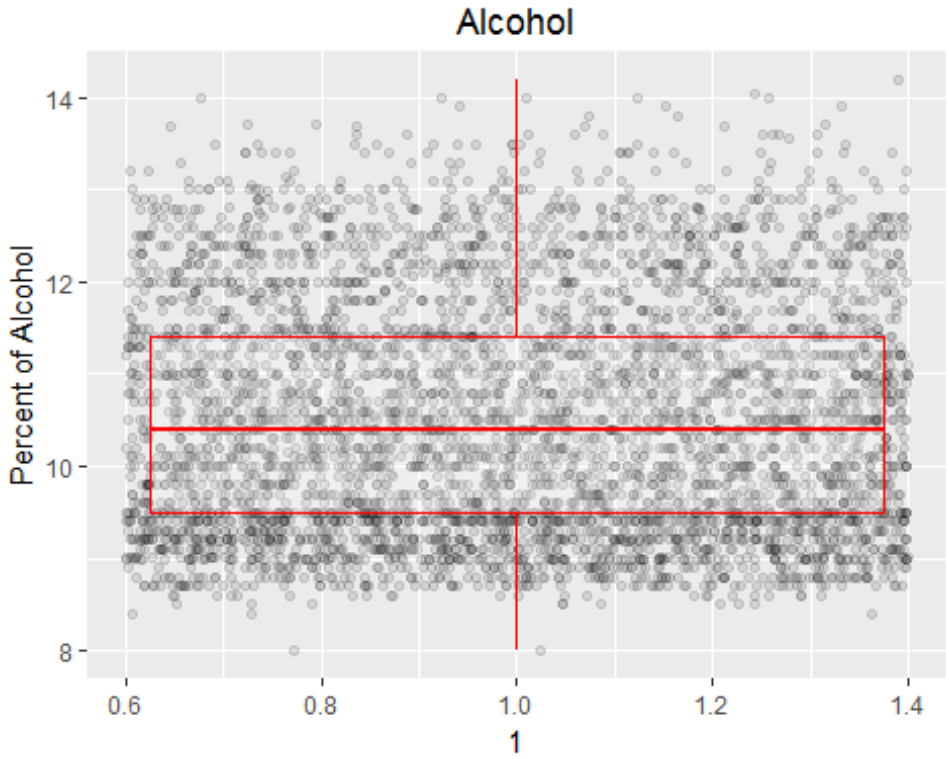
From the histogram and the table, I observe that the quality of wines scored from 3 to 9, while the mean is 5.88 and the median is 6.00. I also find that most of the wine are scored 6(44.88%), and only a very small amount of wine are scored 9(0.1%).

I created a new quality variable by converting the integer variable to factor variable. Since the quality of wines scored from 3 to 9, I created 7 levels and named them based on scores, such as "Score 3", "Score 5", "Score7".

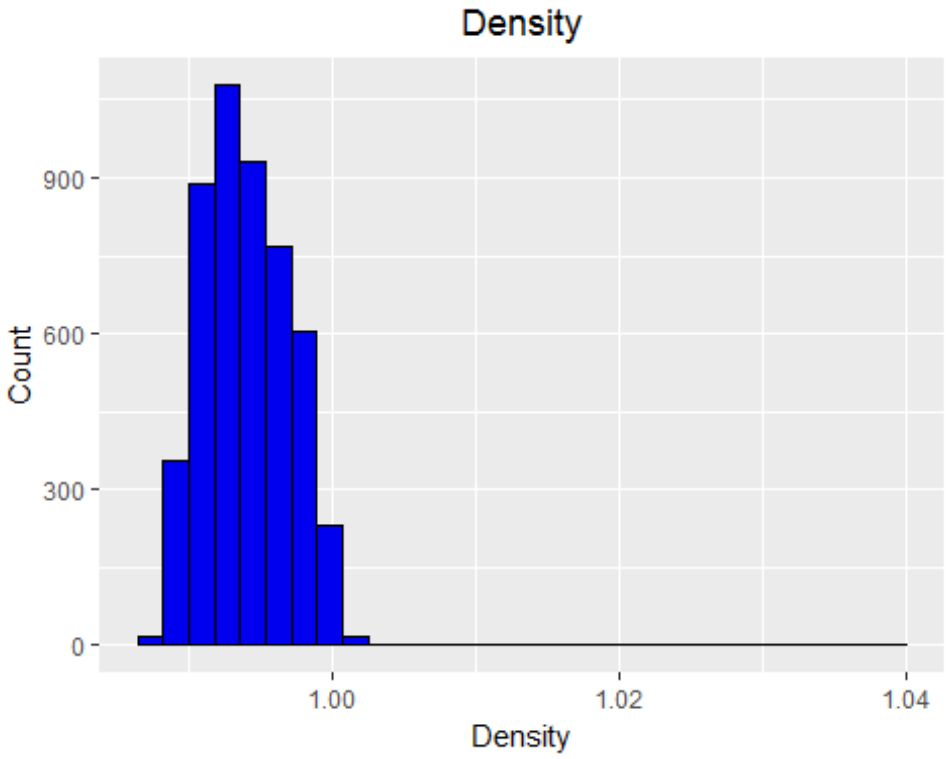



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.50   10.40   10.51   11.40   14.20
```

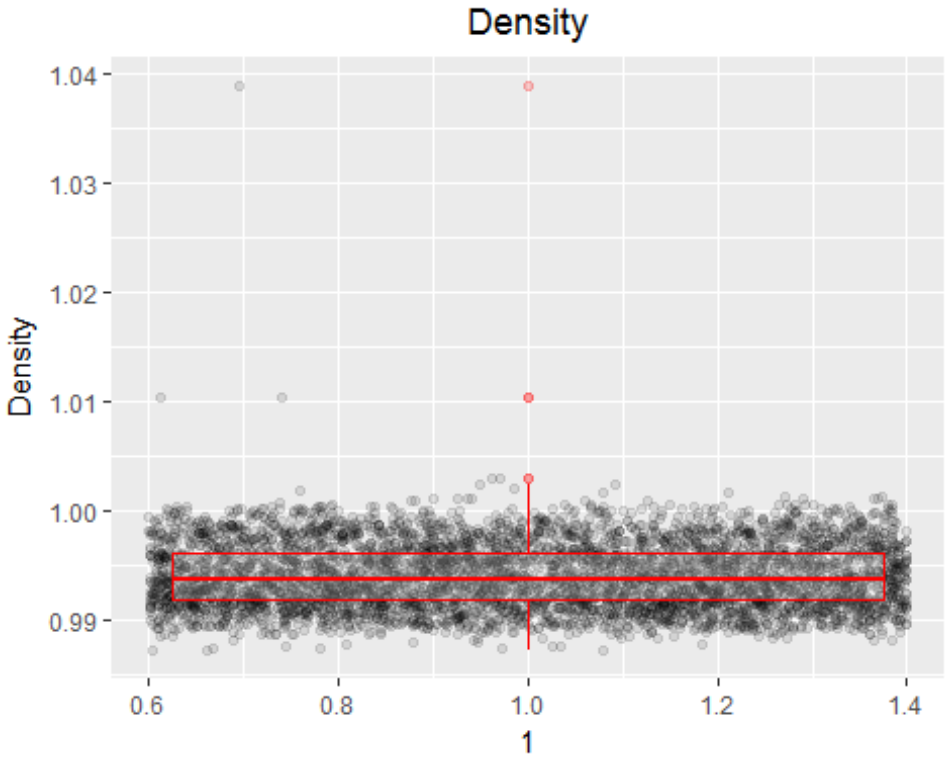
According to the histogram and the table, I find that the percent of alcohol ranged from 8.00%/volume to 14.20%/volume, the mean is 10.51%/volume and the median is 10.40%/volume.



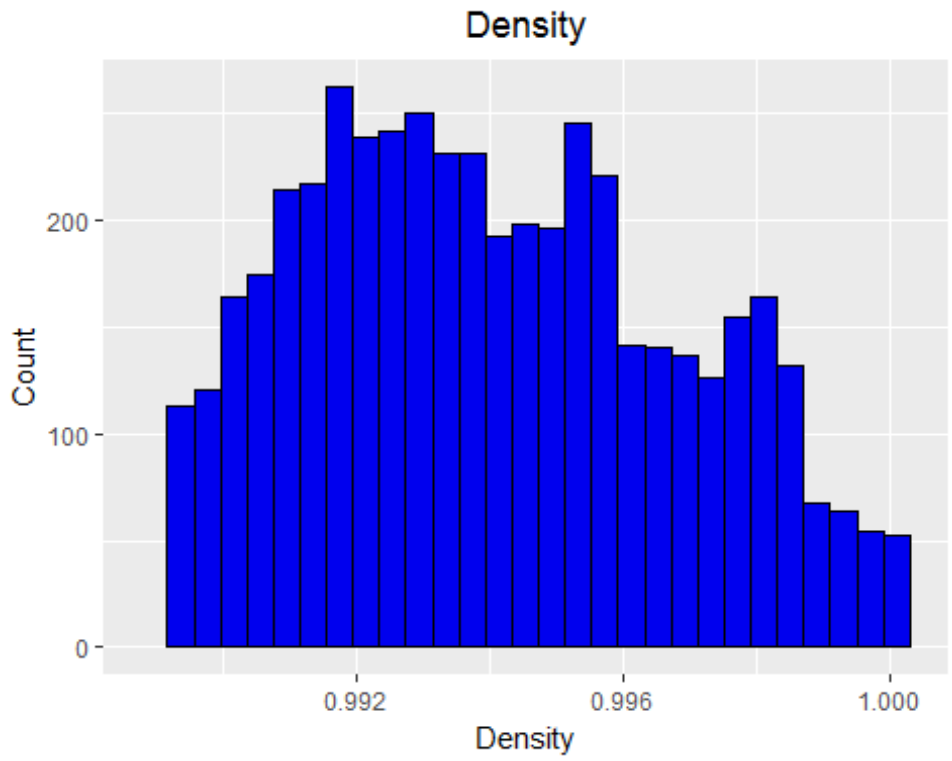
Although 75% of wines have 11.40% of alcohol in wine and the wine that contains that highest percent of alcohol has 14.20% of alcohol, according to the boxplot there does not have any outliners.



The histogram is shifted to the left, which means density variable has at least one outliners. Let's take a closer look by creating a boxplot.

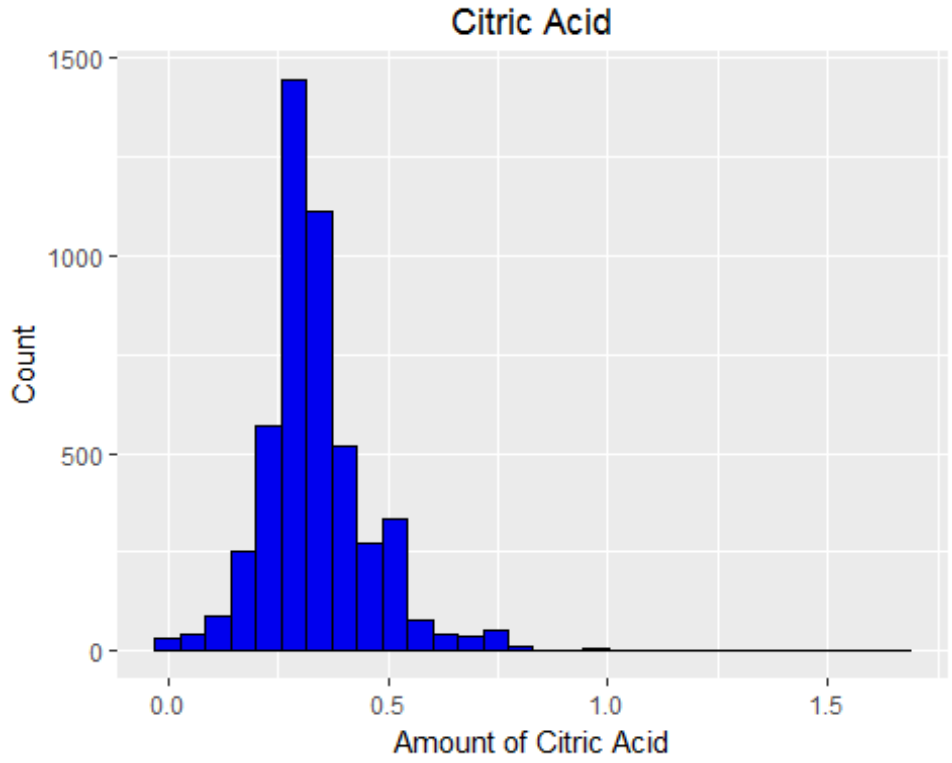


The boxplot concurs my observaton. It showed that there are three outliers. Thus, I created another histogram that contains only the 1th to 99th percentile of the density level.

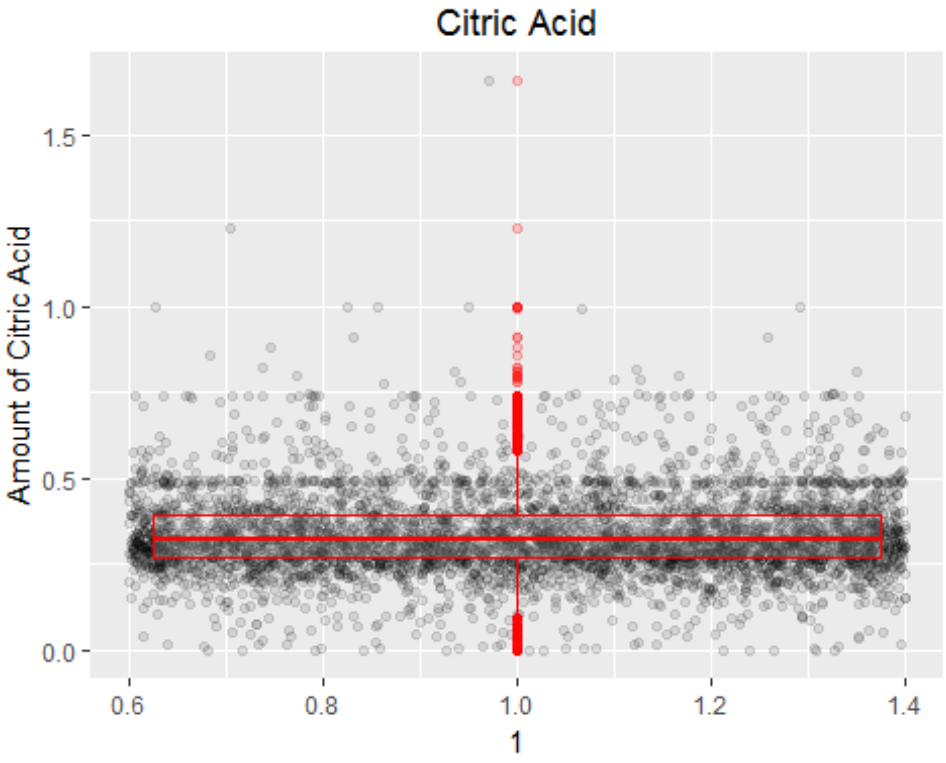



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9917  0.9937  0.9940  0.9961  1.0390
```

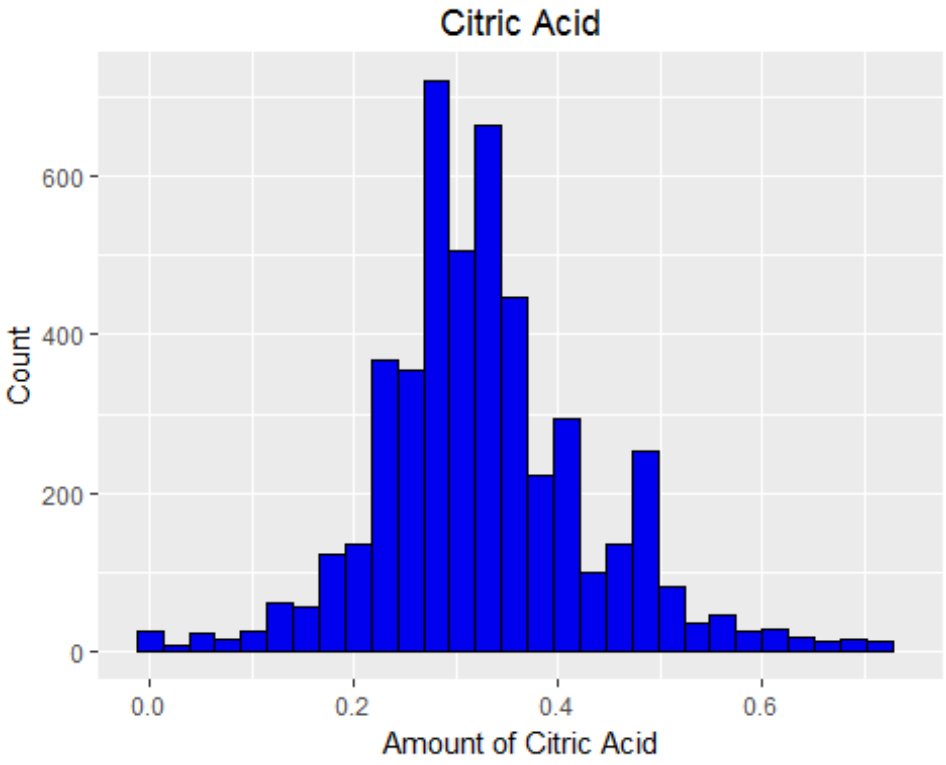
The histogram is a little right-skewed. The density of wines ranged from 0.98g/dm^3 to 1.03g/dm^3, the mean and median are 0.99g/dm^3.



Similar with the density of wines, the histogram of the amount of citirc acid in wine is shifted to the left. So I created a boxplot to check if there are any outliers.

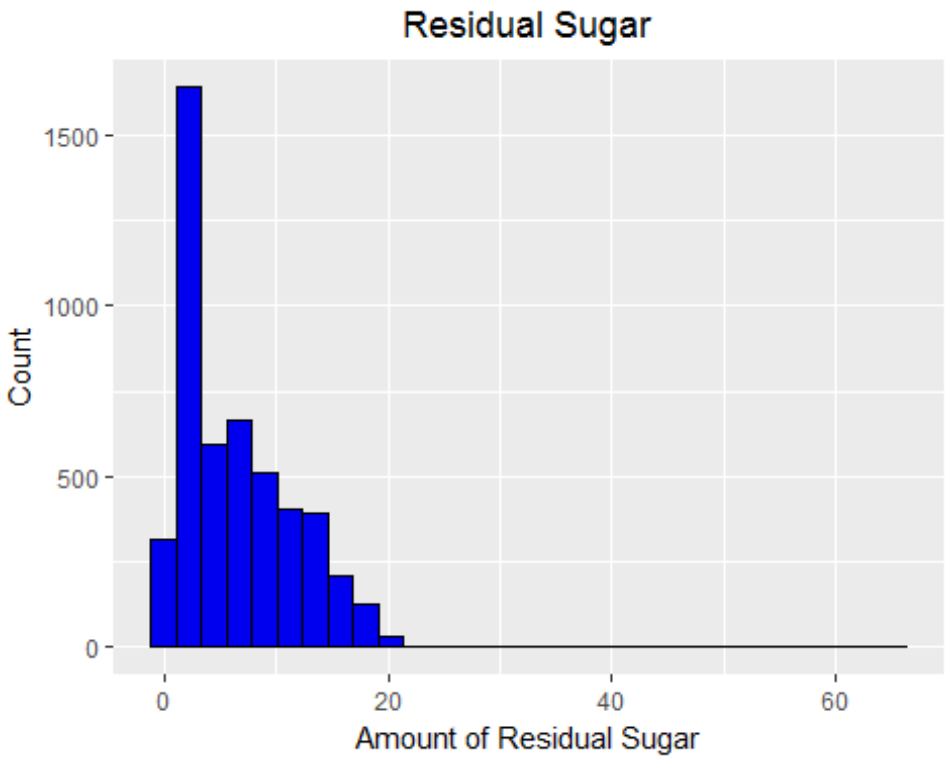


The boxplot showed that there are a number of outliers in both tails. Thus, I created another histogram that contains only the 1th to 99th percentile of the amount of critic acid in wine to take a close look of this variable.

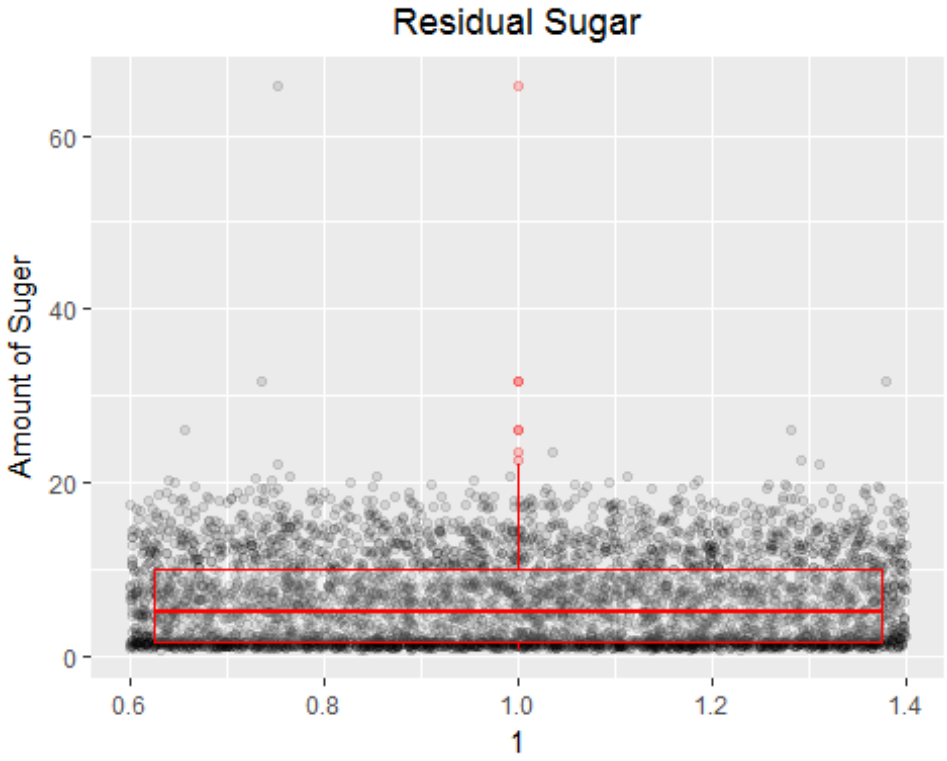



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

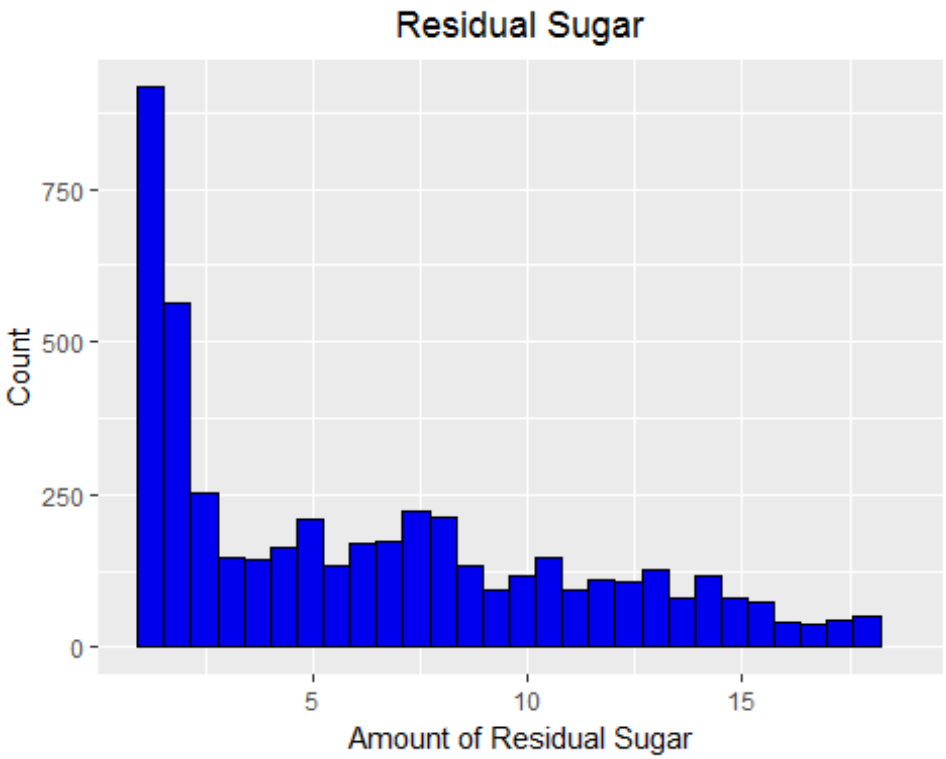
According to the histogram and the table, I find that the mean of the amount of citric acid in wine is 0.33g/dm^3 and the median is 0.32g/dm^3.



Similar with the density of wine and the amount of citric acid in wine, the histogram of the amount of residual suger in wine is shifted to the left. So I guess it contains at least one outlier.



The boxplot confirmed my guess and there are a few outliers.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

The above histogram contains only the 1th to 99th percentile of the amount of residual suger in wine and it is heavy right skewed, The histogram and the table show that the mean is 6.39g/dm^3 and the median is 5.20g/dm^3. It consist a very large outliners as the max is 65.8g/dm^3, while 75% of the wine contains only 9.90g/dm^3 of residual suger.

Univariate Analysis

What is the structure of your dataset?

The data set consist 4898 white wine instances and consists 12 variables. The 11 input vairable are numeric variables, while the output variable(quality of wine) is intger variable.

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

The quality of wine is the main feature as it is an output variable. Moreover, the quality(score) of the wine plays an important role in influencing people's decision in picking wines.

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

The density, ph value and the amount of citric acid and residual sugar are also interesting features. Becauses the thickness and flavor of the wine will also influence people's preference. Moreover, citric acid is usually used to add "freshness" and flavor to wines, thus a right amount of critic acid could increase the quality of the wine.

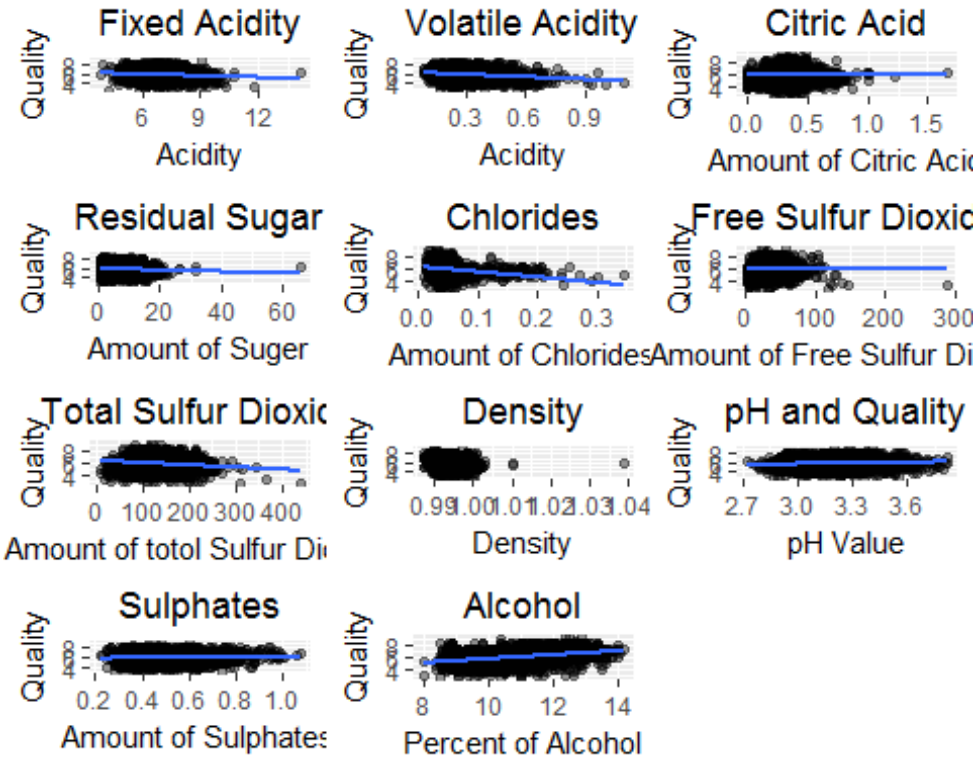
Did you create any new variables from existing variables in the dataset?

I created a factor variable for the quality variable. I created 7 levels and named them based on the score. I think it will be a way

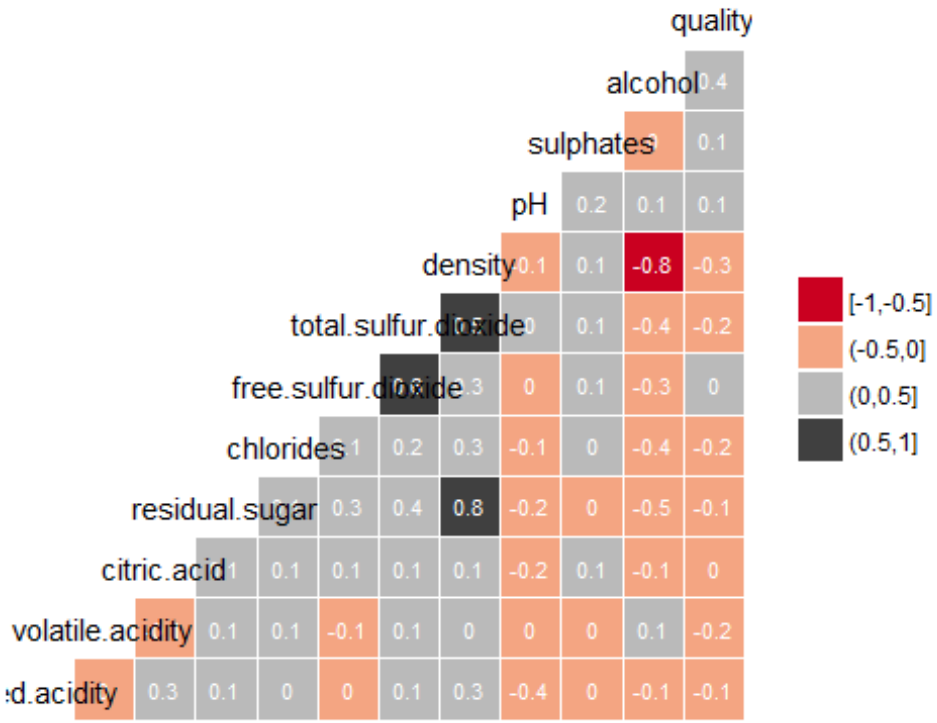
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

The dataset is very tidy and has no missing attribute value, but some of the variables are right-skeded and consist outliers.

Bivariate Plots Section

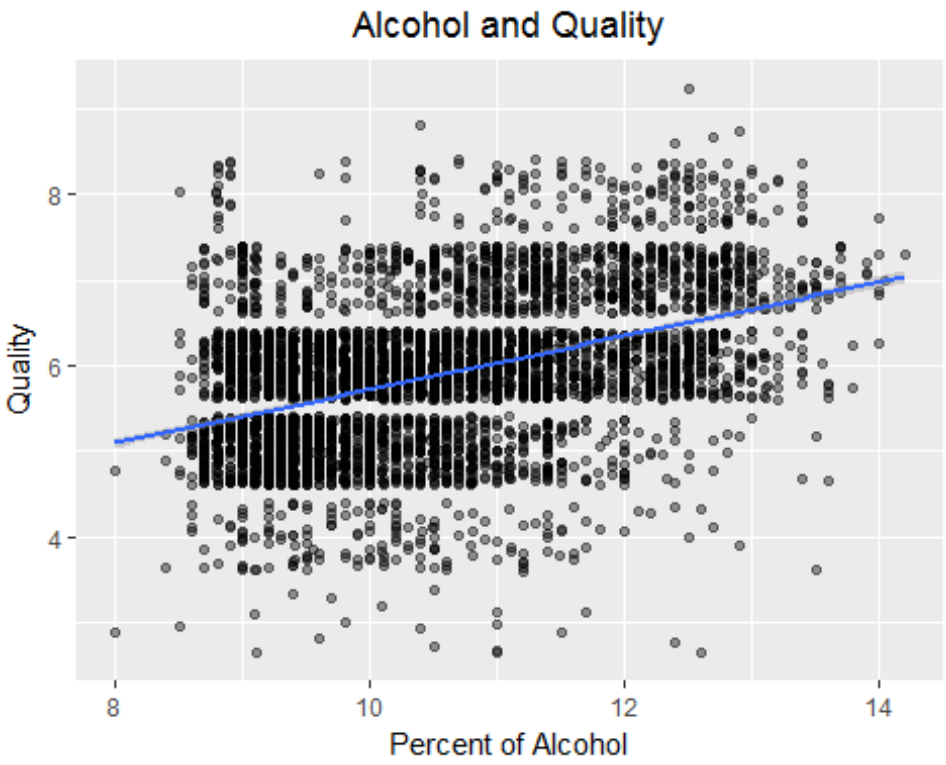


I ploted some scatter plots to explore the input varibles' relationship with quality of wine. The plot showed that the percent of alcohol in wine and pH value seems to have a positive relationship with the quality of wine, the amount of chlorides and the amount of volatile acidity seems to have a relationship with the quality of wine.

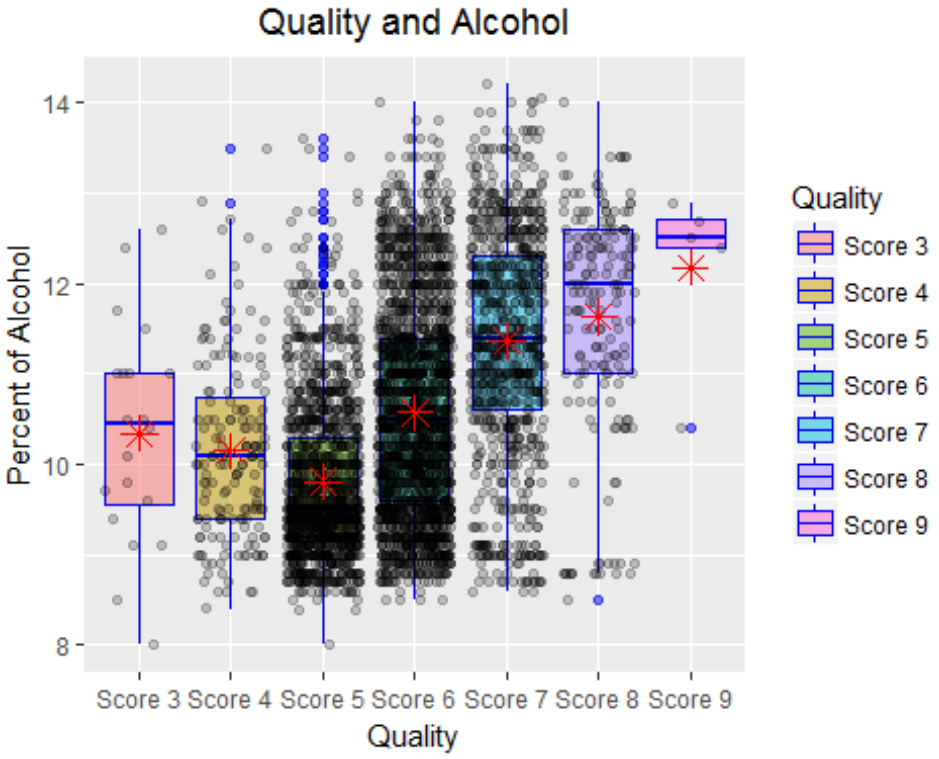


The correlation matrix confirmed my guesses that the percent of alcohol in wine and the pH value have a positive relationship with the quality of wine, however the relationship between the pH value and the quality of wine is not as strong as I thought. And althought the correlation matrix concurs with my guess that the amount of chlorides and volatile acidity have a negative relationship with the quality of wine, their relationship are weak. The matrix also showed that the amount of free silfur dioxide and critic acid do not have any relationship with the quality of wine.

Besides, the density of wine has a strong positive relationship with the amount of total sulfur dioxide and residual sugar in wine. The amount of total sulfur dioxide in wine also have a strong relationship with the amount of free sulfur dioxide in wine.



The scatter plot demonstrates that there is the percent of alcohol has a positive relationship with the quality of wine.

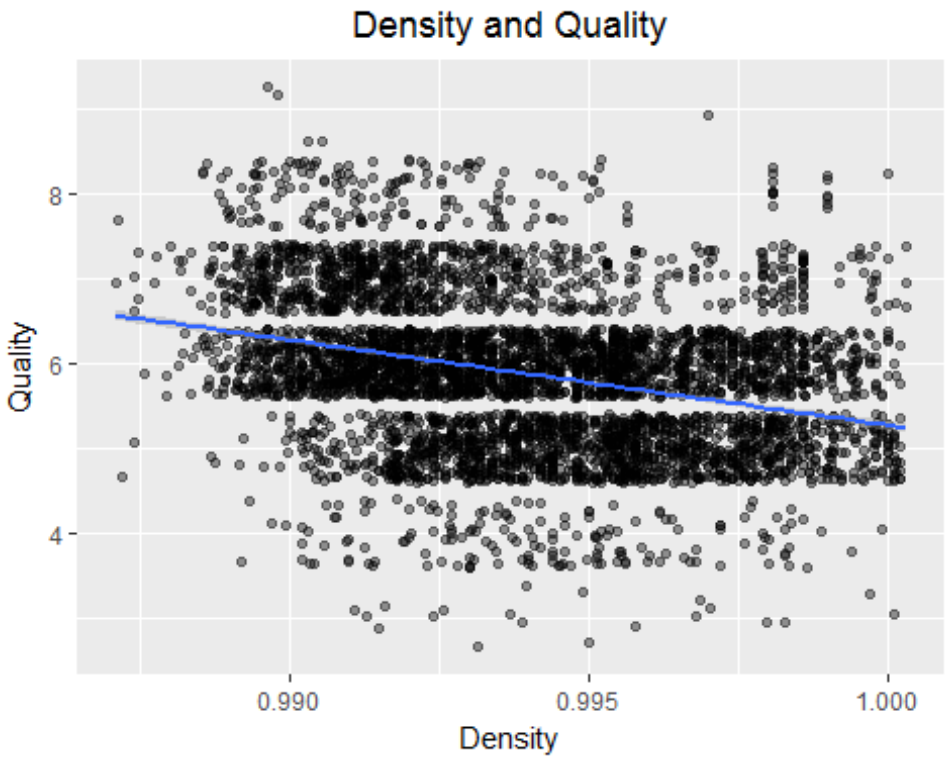


```
## data$quality_c: Score 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00    9.55   10.45   10.34   11.00   12.60
## ------------------------------------------------------------
## data$quality_c: Score 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.40   10.10   10.15   10.75   13.50
## ------------------------------------------------------------
## data$quality_c: Score 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.000   9.200   9.500   9.809  10.300  13.600
## ------------------------------------------------------------
## data$quality_c: Score 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50    9.60   10.50   10.58   11.40   14.00
## ------------------------------------------------------------
## data$quality_c: Score 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.60   10.60   11.40   11.37   12.30   14.20
## ------------------------------------------------------------
## data$quality_c: Score 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.50   11.00   12.00   11.64   12.60   14.00
## ------------------------------------------------------------
## data$quality_c: Score 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   12.40   12.50   12.18   12.70   12.90
```

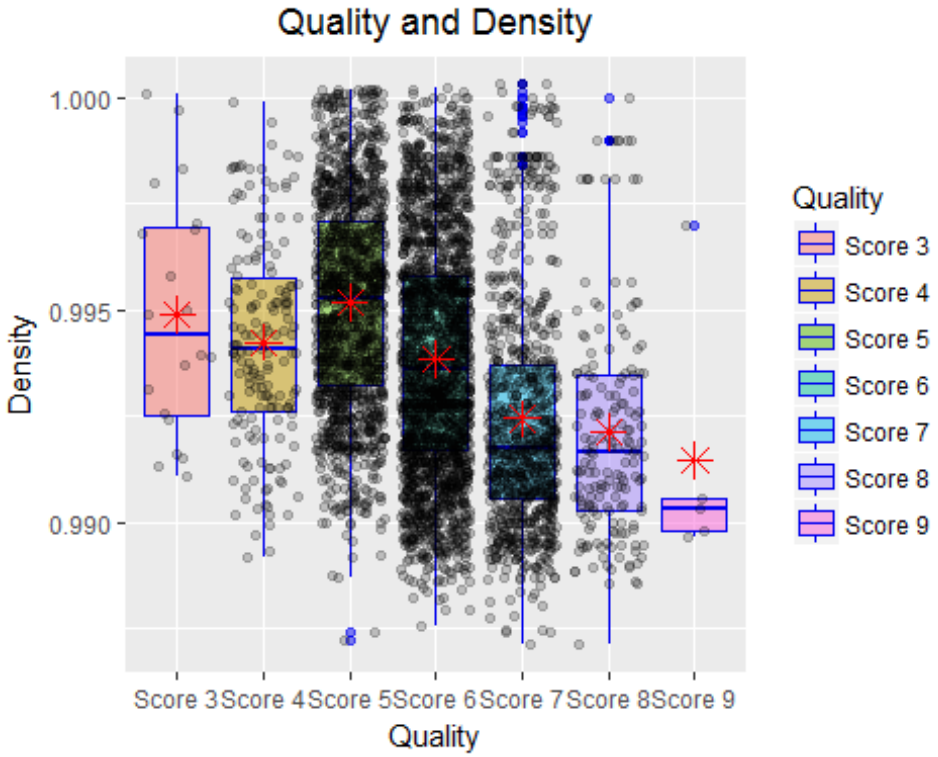
The boxplot and the table confirmed my observation, wines with a higher score tends to have a higher percent of alcohol. However, some wines with a score 5 have a very high percent of alcohol too.

```
##
##  Pearson's product-moment correlation
##
## data:  data$alcohol and data$quality
## t = 33.858, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4126015 0.4579941
## sample estimates:
##       cor
## 0.4355747
```

The correlation test I computed consist with my observation. The result shows that there is a moderate positive correlation between the two variables ($r = 0.44$, $p < 0.05$).



The scatter plot shows that the density of wine has a negative relationship with the qaulity of wine.

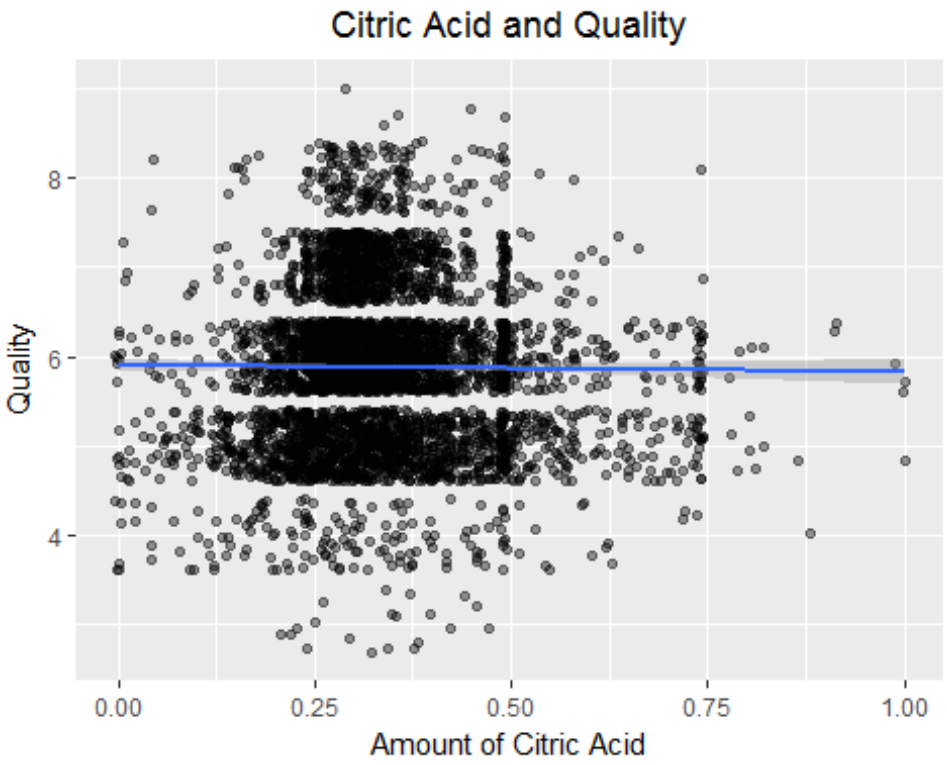


```
## data$quality_c: Score 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9911  0.9925  0.9944  0.9949  0.9969  1.0000
## ------------------------------------------------------------
## data$quality_c: Score 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9892  0.9926  0.9941  0.9943  0.9958  1.0000
## ------------------------------------------------------------
## data$quality_c: Score 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9872  0.9933  0.9953  0.9953  0.9972  1.0020
## ------------------------------------------------------------
## data$quality_c: Score 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9876  0.9917  0.9937  0.9940  0.9959  1.0390
## ------------------------------------------------------------
## data$quality_c: Score 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9906  0.9918  0.9925  0.9937  1.0000
## ------------------------------------------------------------
## data$quality_c: Score 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9871  0.9903  0.9916  0.9922  0.9935  1.0010
## ------------------------------------------------------------
## data$quality_c: Score 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9896  0.9898  0.9903  0.9915  0.9906  0.9970
```

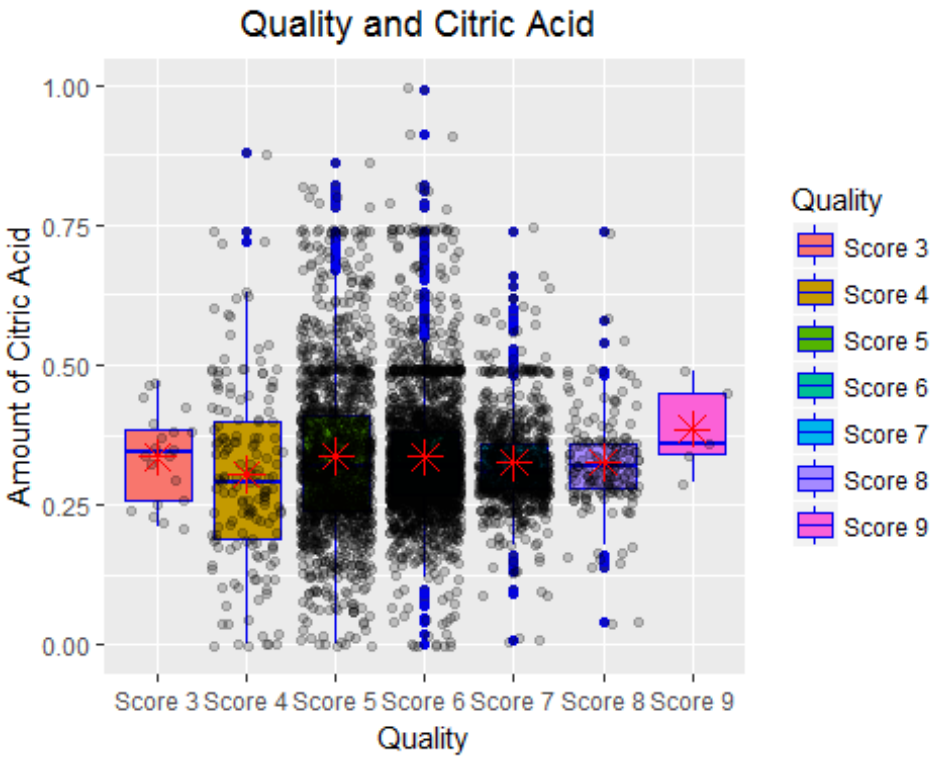
The boxplot and the table confirmed my observation, wines with a higher score tends to have a lower density.

```
##
##  Pearson's product-moment correlation
##
## data:  data$density and data$quality
## t = -22.581, df = 4896, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3322718 -0.2815385
## sample estimates:
##       cor
## -0.3071233
```

And the result of the pearsone correlation test shows that there is a weak negative correlation between these variables ($r = -0.31$, $p < 0.05$).



The scatter demonstrates no relationship bewteen the amout of critic acid and the quality of wine.

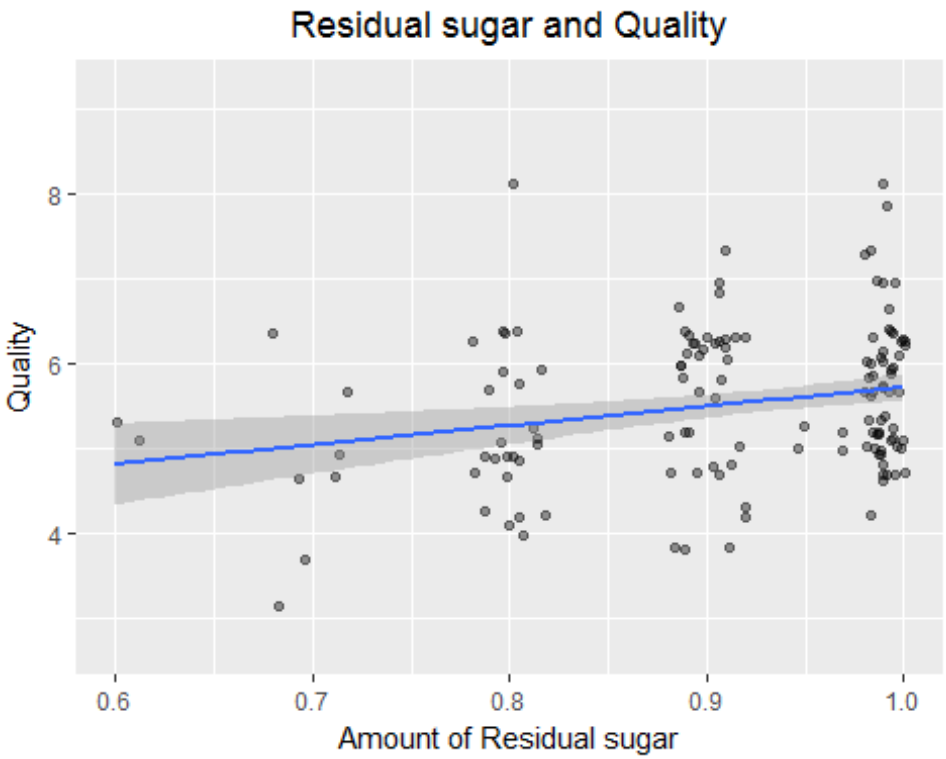


```
## data$quality_c: Score 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2100  0.2575  0.3450  0.3360  0.3850  0.4700
## ------------------------------------------------------------
## data$quality_c: Score 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1900  0.2900  0.3042  0.4000  0.8800
## ------------------------------------------------------------
## data$quality_c: Score 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2400  0.3200  0.3377  0.4100  1.0000
## ------------------------------------------------------------
## data$quality_c: Score 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.270   0.320   0.338   0.380   1.660
## ------------------------------------------------------------
## data$quality_c: Score 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0100  0.2800  0.3100  0.3256  0.3600  0.7400
## ------------------------------------------------------------
## data$quality_c: Score 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0400  0.2800  0.3200  0.3265  0.3600  0.7400
## ------------------------------------------------------------
## data$quality_c: Score 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.290   0.340   0.360   0.386   0.450   0.490
```

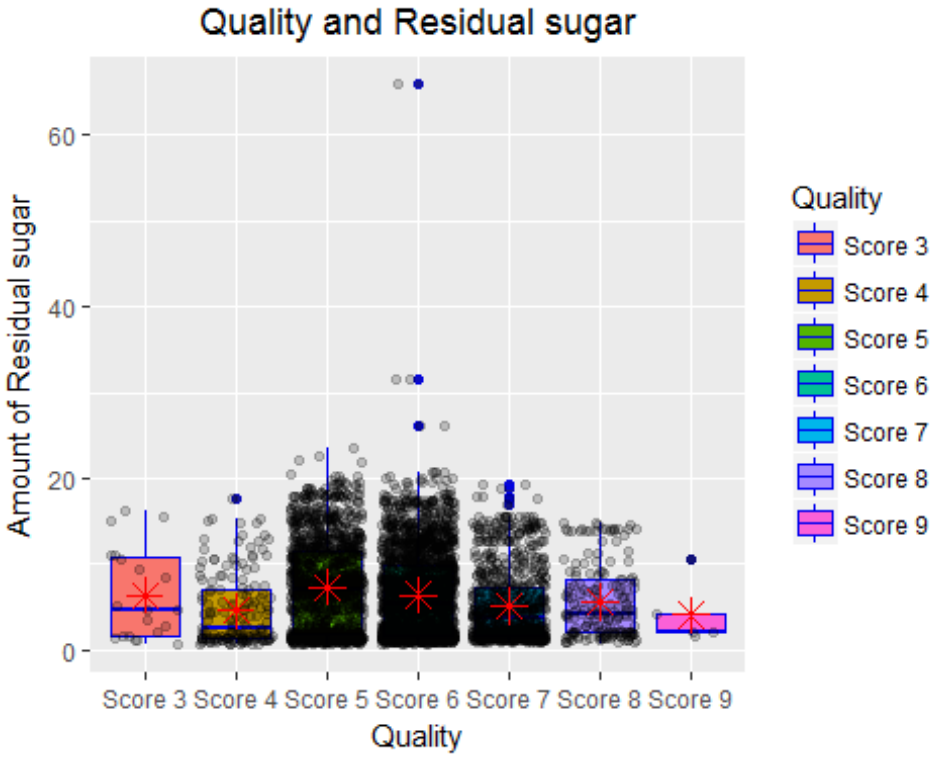
The boxplot and the table confirmed my observation, the quality of wine does not seems to have any relationship with the amount of citric acid in wine.

```
##
##  Pearson's product-moment correlation
##
## data:  data$citric.acid and data$quality
## t = -0.6444, df = 4896, p-value = 0.5193
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03720595  0.01880221
## sample estimates:
##          cor
## -0.009209091
```

The result of the pearson correlation test confirmed my guess. The result shows that there is no correlation between the two variables ($r = -0.01$, $p < 0.05$).



The scatter plot demonstratess that the amount of residual sugar has a weak relationship wih the quality of wine.



```
## data$quality_c: Score 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.700   1.588   4.600   6.392  10.700  16.200
## ------------------------------------------------------------
## data$quality_c: Score 4
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.700   1.300   2.500   4.628   7.100  17.550
## ------------------------------------------------------------
## data$quality_c: Score 5
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.800   7.000   7.335  11.500  23.500
## ------------------------------------------------------------
## data$quality_c: Score 6
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.700   1.700   5.300   6.442   9.900  65.800
## ------------------------------------------------------------
## data$quality_c: Score 7
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.700   3.650   5.186   7.325  19.250
## ------------------------------------------------------------
## data$quality_c: Score 8
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.800   2.100   4.300   5.671   8.200  14.800
## ------------------------------------------------------------
## data$quality_c: Score 9
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.60    2.00    2.20    4.12    4.20   10.60
```

The boxplot and the table show that wine with a higher score tend to have a lower amount of residual sugar.

```
##
##  Pearson's product-moment correlation
##
## data:  data$residual.sugar and data$quality
## t = -6.8603, df = 4896, p-value = 7.724e-12
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.12524103 -0.06976101
## sample estimates:
##         cor
## -0.09757683
```

I computed a pearson correlation test to access the relationship between these variable. The result demonstrates that there is a very weak negative correlation between thequality of wine and the amount of residual suger in wine($r = -0.10$, $p < 0.05$).

Bivariate Analysis

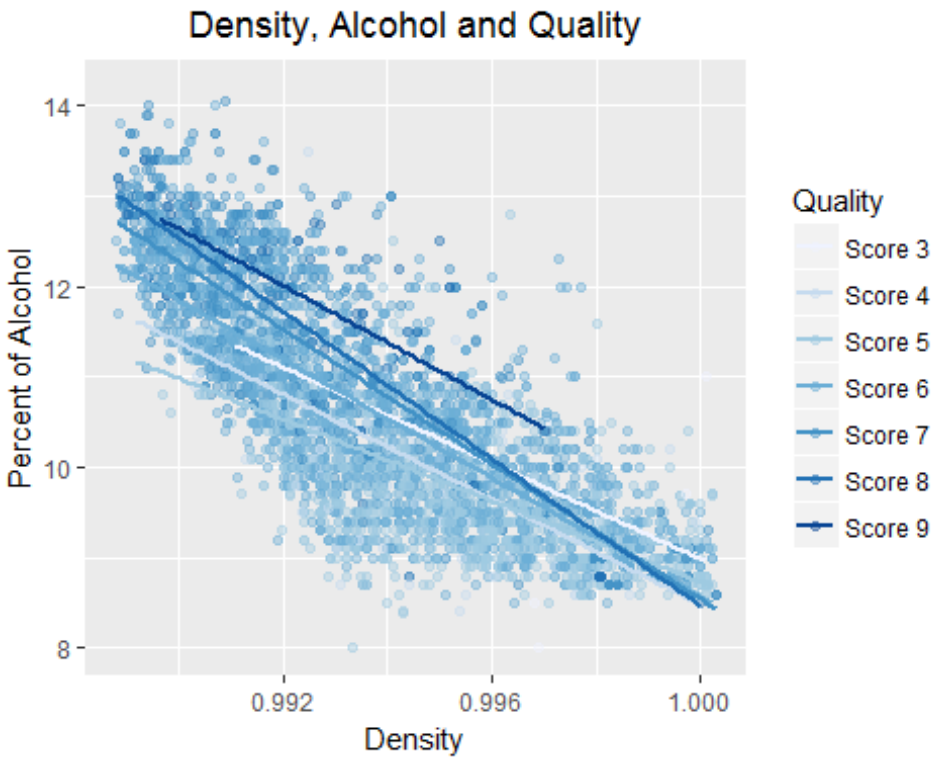
Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Based on the scatter plots and boxplots, I observated the quality of wine has a positive moderate relationship with the percent of alcohol in wine, a weak negative relationship with the density of wine, a very weak negative relationship with the amount of residual sugar in wine, and no relationship with the amount of citric acid in wine. All of the observations have been confirmed by the result of pearson correlation tests.

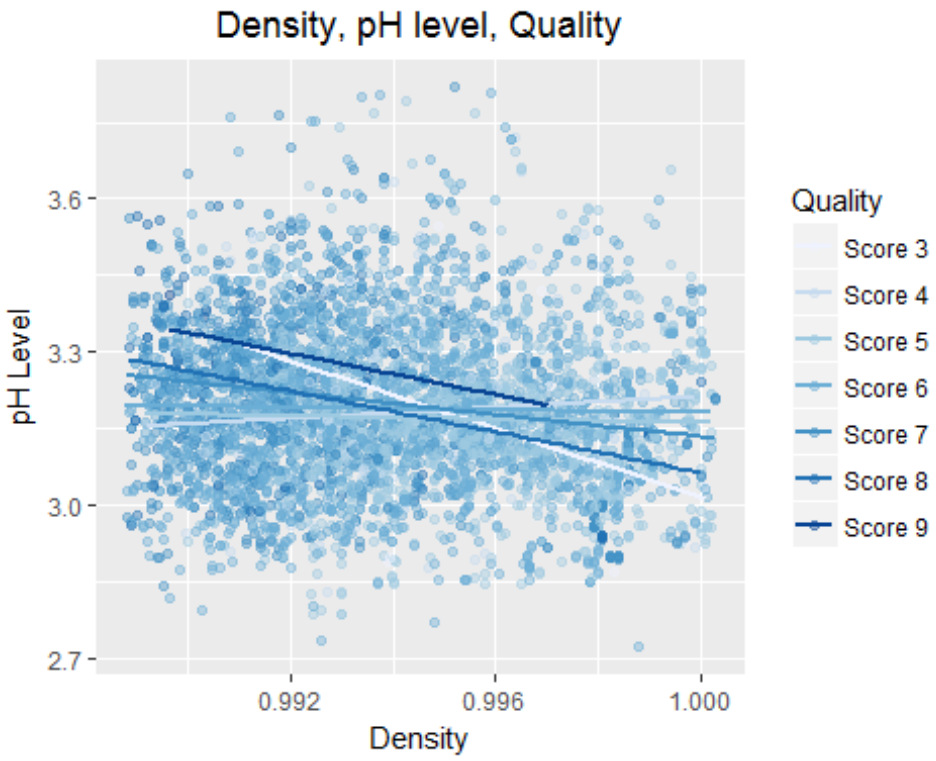
What was the strongest relationship you found?

The strongest relationship I found is the relationship between the quality of wine and the percent of alcohol in wine. The result of the pearson correlation test shows that there is a moderate positive correlation between the two variables ($r = 0.44$, $p < 0.05$). Thus, wines that content a higher percent of alcohol tend to have a better quality.

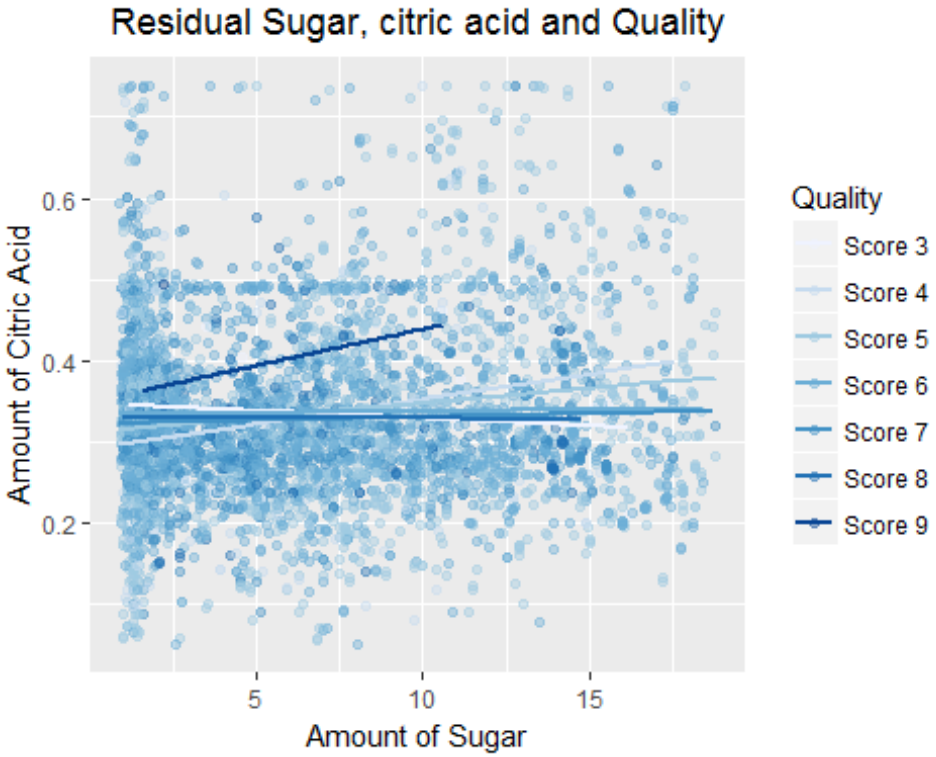
Multivariate Plots Section



The scatterplot demonstrated there is a relationship between the density of wine, the percent of alcohol in wine and the quality of wine. Wines with a higher quality tend to have a higher percent of acohol in wine and density of wine ratio.



The scatterplot fails to show any relationship between the density, the pH value and the quality of wine.



Although the scatterplot shows wines with a score 9 tends to have a higher amount of citric acid in wine and amount of sugar in wine ratio, it fails to show any relationship between the amount of residual sugar, citric acid in wine and the quality of wine.

```
##
## Calls:
## m1: lm(formula = quality ~ fixed.acidity, data = data)
## m2: lm(formula = quality ~ fixed.acidity + volatile.acidity, data = data)
## m3: lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid,
##     data = data)
## m4: lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar, data = data)
## m5: lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides, data = data)
## m6: lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide, data = data)
## m7: lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide,
##     data = data)
## m8: lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density, data = data)
## m9: lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH, data = data)
## m10: lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates, data = data)
## m11: lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol, data = data)
##
## ================================================================================
================================================================
##                              m1         m2         m3         m4         m5         m6
m7          m8           m9           m10          m11
## --------------------------------------------------------------------------------
----------------------------------------------------------------
##   (Intercept)            6.696***   7.210***   7.214***   7.232***   7.499***   7.447*
**   7.519***   237.792***   275.802***   283.857***   150.193***
##                         (0.103)    (0.107)    (0.108)    (0.108)    (0.107)    (0.112)
(0.111)      (8.375)      (8.643)      (8.633)     (18.804)
##   fixed.acidity         -0.119***  -0.124***  -0.122***  -0.117***  -0.121***  -0.118*
**  -0.100***     0.043**      0.161***     0.166***     0.066**
##                         (0.015)    (0.015)    (0.015)    (0.015)    (0.015)    (0.015)
(0.015)      (0.015)      (0.017)      (0.017)      (0.021)
##   volatile.acidity                 -1.735***  -1.741***  -1.689***  -1.547***  -1.523*
**  -1.308***    -1.737***    -1.720***    -1.699***    -1.863***
##                                    (0.122)    (0.124)    (0.124)    (0.122)    (0.123)
(0.123)      (0.116)      (0.113)      (0.113)      (0.114)
##   citric.acid                                 -0.037      0.011      0.187      0.175
0.216*       0.108        0.148        0.094        0.022
##                                               (0.108)    (0.108)    (0.106)    (0.107)
(0.106)      (0.098)      (0.096)      (0.096)      (0.096)
##   residual.sugar                                          -0.013***  -0.011***  -0.012*
**  -0.005*       0.097***     0.124***     0.128***     0.081***
##                                                          (0.002)    (0.002)    (0.003)
(0.003)      (0.004)      (0.005)      (0.005)      (0.008)
##   chlorides                                                         -7.797***  -7.871*
```

```
**  -7.029***    -2.118***    -0.477       -0.286       -0.247
##                                                                  (0.559)    (0.561)
(0.561)      (0.552)      (0.554)      (0.550)      (0.547)
##   free.sulfur.dioxide                                             0.001
0.007***     0.003***     0.003***     0.003***     0.004***
##                                                                  (0.001)
(0.001)      (0.001)      (0.001)      (0.001)      (0.001)
##   total.sulfur.dioxide
-0.004***     0.000        0.000       -0.000       -0.000
##
(0.000)      (0.000)      (0.000)      (0.000)      (0.000)
##   density
-233.806***  -277.081***  -285.389***  -150.284***
##
(8.503)      (8.884)      (8.875)     (19.075)
##   pH
1.236***     1.162***     0.686***
##
(0.088)      (0.087)      (0.105)
##   sulphates
0.829***     0.631***
##
(0.098)      (0.100)
##   alcohol
0.193***
##
(0.024)
## ------------------------------------------------------------------------------------
-------------------------------------------------------------------
##   R-squared                  0.0        0.1        0.1        0.1        0.1        0.
1        0.1         0.2          0.3          0.3          0.3
##   adj. R-squared             0.0        0.1        0.1        0.1        0.1        0.
1        0.1         0.2          0.3          0.3          0.3
##   sigma                      0.9        0.9        0.9        0.9        0.8        0.
8        0.8         0.8          0.8          0.8          0.8
##   F                         64.1      133.9       89.3       74.6      100.9       84.
6       89.0       184.4        192.6        183.0        174.3
##   p                          0.0        0.0        0.0        0.0        0.0        0.
0        0.0         0.0          0.0          0.0          0.0
##   Log-likelihood         -6322.8    -6224.2    -6224.1    -6209.7    -6114.2    -6112.
9    -6060.9     -5708.8      -5611.2      -5575.5      -5543.7
##   Deviance                3791.4     3641.8     3641.7     3620.3     3481.9     3480.
0     3406.9      2950.6       2835.4       2794.3       2758.3
##   AIC                    12651.5    12456.4    12458.2    12431.3    12242.5    12241.
8    12139.8     11437.5      11244.4      11175.0      11113.5
##   BIC                    12671.0    12482.3    12490.7    12470.3    12287.9    12293.
8    12198.3     11502.5      11315.9      11253.0      11197.9
##   N                       4898       4898       4898       4898       4898       4898
4898        4898         4898         4898         4898
## ====================================================================================
===================================================================
```

I used the multiple linear regression analysis to develop a model for predicting the quality of wine from the 11 input variables(i.e. density, amount of critical aicd, chlorides). It was found that density(beta = -150.28, p < .001), volatile acidity(beta = -1.86, p < .001), amount of alcohol(beta = 0.19, p < .001), pH

level(beta = 0.69, p < .001), amount of sulphates(beta = 0.63, p < .001), amount of residual sugar(beta = 0.08, p < .001), fixed acidity (beta = 0.66, p < .01) and amout of free sulfur dioxide(beta = 0.004, p < .001) were significant predictors. The amount of citric acid, chlorides and total sulfur dioxide are not a significant predictor. The overal model fit was R^2 = 0.3.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

The scatterplots I created shows that wines that the quality of wine has a postive relationship with the density of wine and a negaitve relatioinship with the percent of alcohol in wine. Wines with a higher quality tend to have a higher percent of acohol in wine and density of wine ratio. However, the scatters fails to show any relationships between the quality, density and pH level of wine, and between the quality of wine with the amount of residual sugar and citric acid in wine.

Were there any interesting or surprising interactions between features?

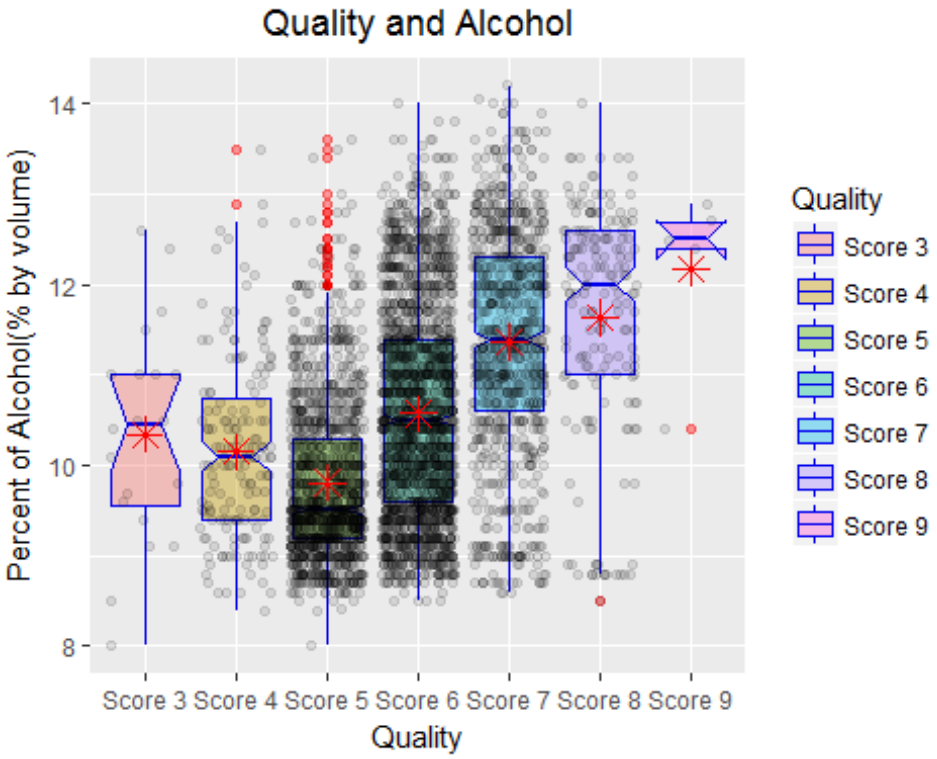
It is pretty suprise to find that the amount of residual sugar and citric acid do not influence the quality of wine. Because citric acid is usually used to add 'freshness' to wines. So it was interesting to find that they have no relationship with the quality of wine as both of them were related with the favor of wines.

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

I created a linear model to predict the quality of white wine from the 11 input variables. The result shows that density is the best predictor of the quality of wine, and volatile acidity, pH level, fixed acidity and amount of alcohol, sulphates, residual sugar and free sulfur dioxide are also predictors of the quality of wine.

Final Plots and Summary

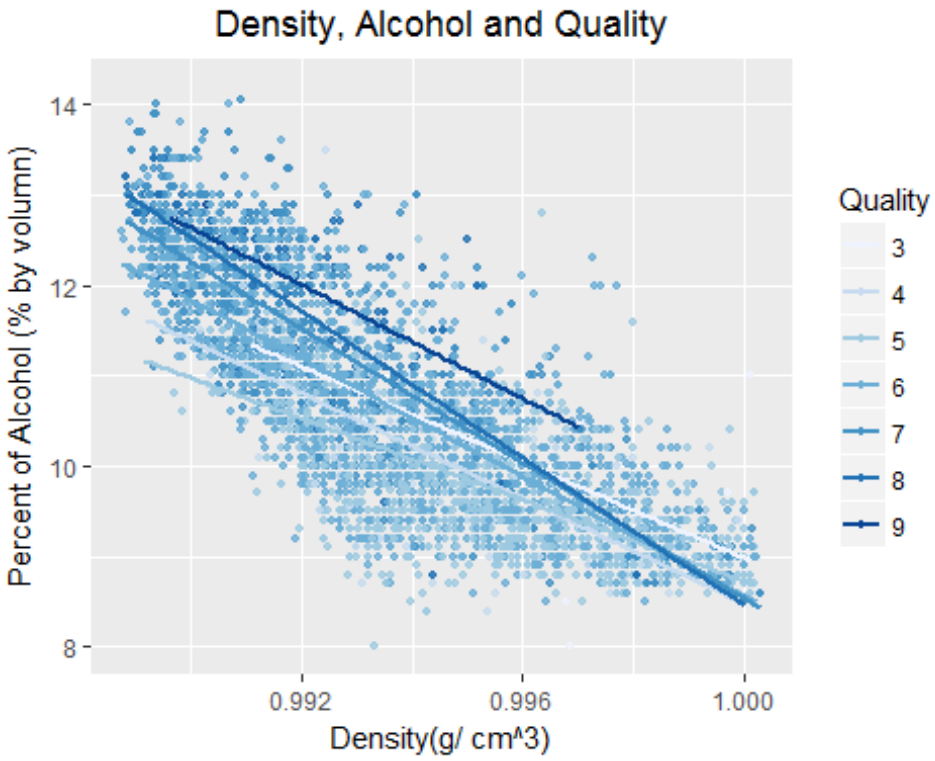
Plot One



Description One

This boxplot demostrated the relationship between the quality of wine and the percentage of alcohol in wine. Although there are a few outliners, such as a number of wines that are scored 5 have a relatively high percentage of alcohol in wine, it shows a pretty clear picture that the wines with a higher quality tends to have a higher percentage of alcohol. The plot also shows that wines that are scored 6 has a wider range of percent of alcohol in wine that wines that have a different score.

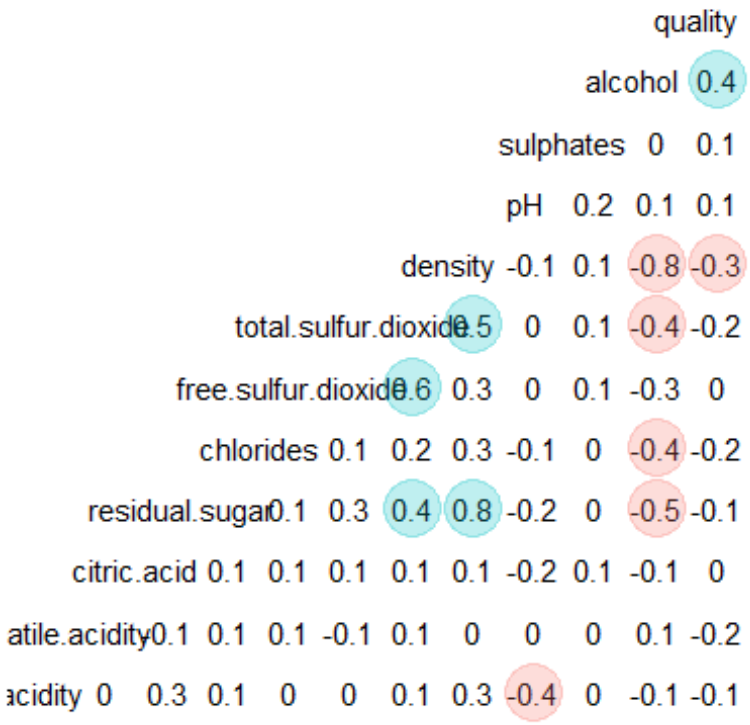
Plot Two



Description Two

This scatter plot shows that relationship between the density and the quality of wine and the percent of alcohol in wine. It domostrates that the density of wine has a negative correlation with the percentage of alcohol in wine. It also shows wines that have a higher score (darker color) tend to have a lower density and percent of alcohol in wine ratio, and wine that have a lower score (lighter color) tend to have a higher density and percent of alcohol in wine ratio.

Plot Three



Description Three

I choose the correlation matrix because it is able to provide a very clear picture about the relationships between variables. The plot highlighted the correlations that is bigger than +/-0.3. It shows that the quality of wine has a moderate positive relationship wtth alcohol and weak negative relationship with density. The matrix also demostrates that the percent of alcohol in wine is the variable that has the most strong and moderate correlation with others variables. It has a strong negative relationship with density and moderate relationship with the amount of total sulfur dioxide, chlorides and residual sugar in wine.

Reflection

I became more proficient in exploring and visualizing data with R after working on this exercise. The exercise advance my skills in using histograms, boxplots and scatter plots to present the data and computing pearson correlation test and multiple linear regression analysis to find the relationship between variables. It also taught me how to use correlation matrix, which is a efficient way to explore the data as it is able to provide a very clear picture about the correlation between variables.