White Wine Data Exploration by Arda Tasci

Arda Tasci

September 22, 2017

Univariate Plots Section

```
## [1] 4898   13
```

There are 4898 observations with 13 variables in the dataset which are :

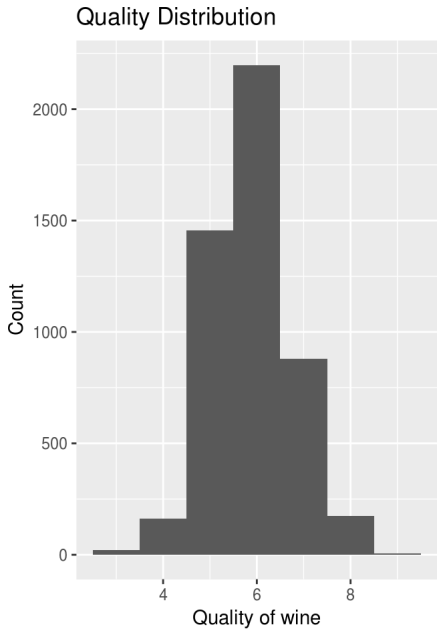
```
##  [1] "X"                    "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"          "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "quality"
```

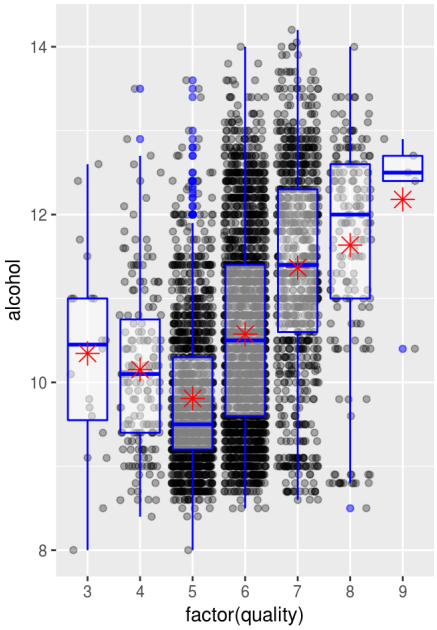
```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  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 0.044 ...
##  $ 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 ...
```

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##     alcohol         quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.40   Median :6.000
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

The target variable is the quality which depends on the other variables. Although the scale of quality shall be between 1 and 10, the quality metric varies between 3 and 9. Median is 6 and mean is 5.878. Most of the observations probably have quality of 6 based on this distribution.

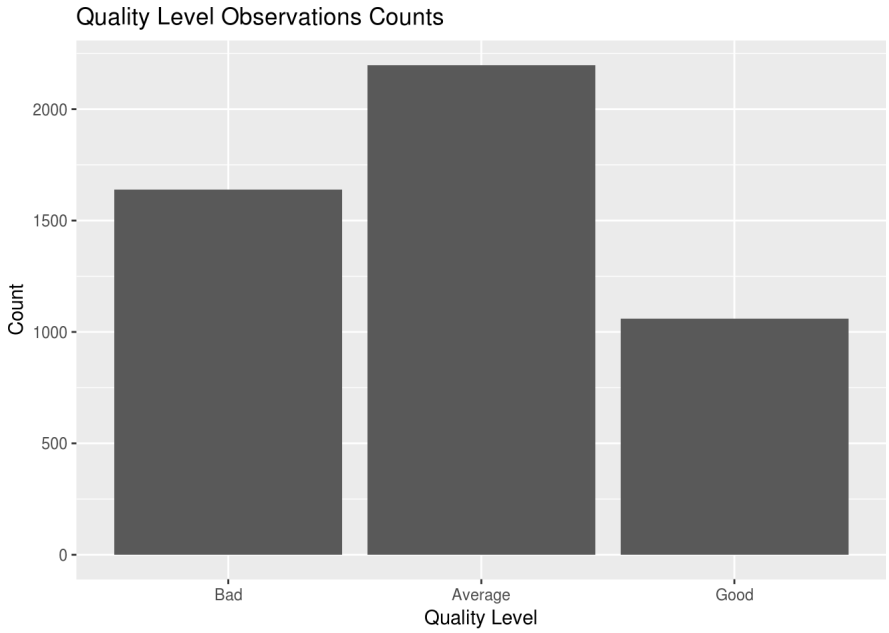
Quality variable is a discrete variable which shall be cast as an ordered factor.

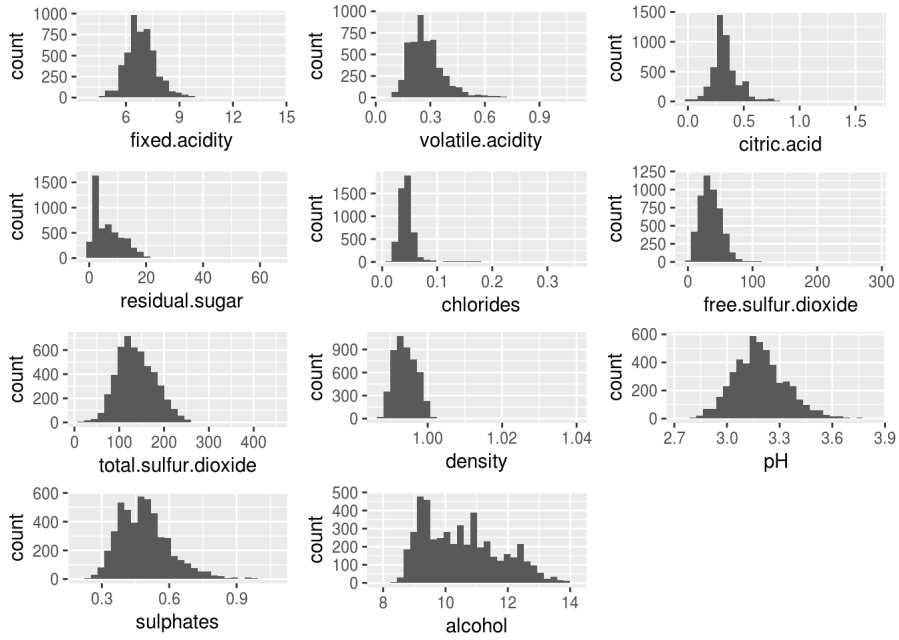




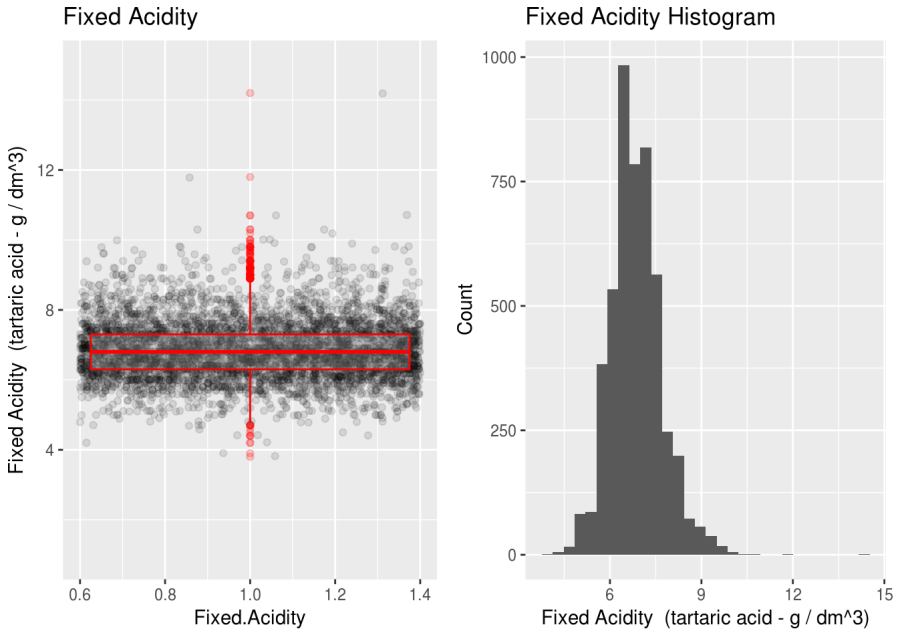

```
##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

As I stated, highest count of observations have quality of 6. Most of the wines are graded 5, 6, and 7. Quality is an ordered factor variable so that I created a new variable using quality which is quality.level having values “Bad”, “Average” and “Good”. The quality values of this variable is (0,5], (5, 6], (6,9] respectively.



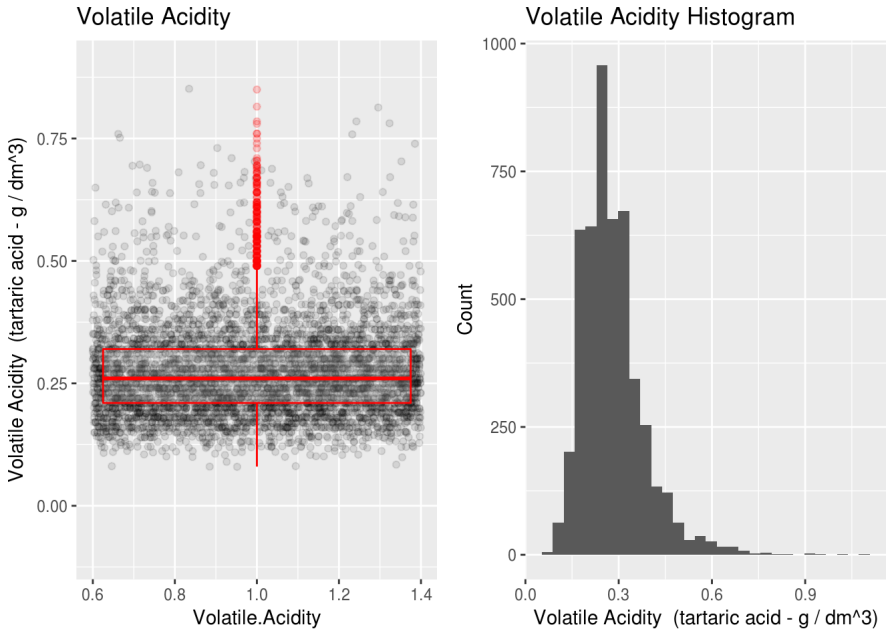


Let's look deeper into the variables.



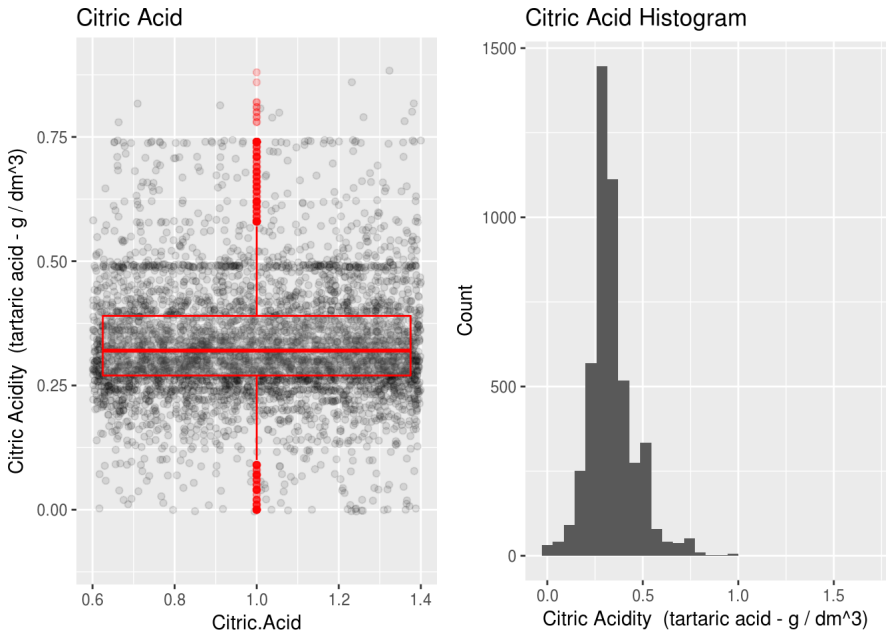

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

As seen Fixed Acidity values are mostly fall between 6 and 7.3. Observations with fixed.acidity values less than 5 and more than 10 can be accepted as outliers.



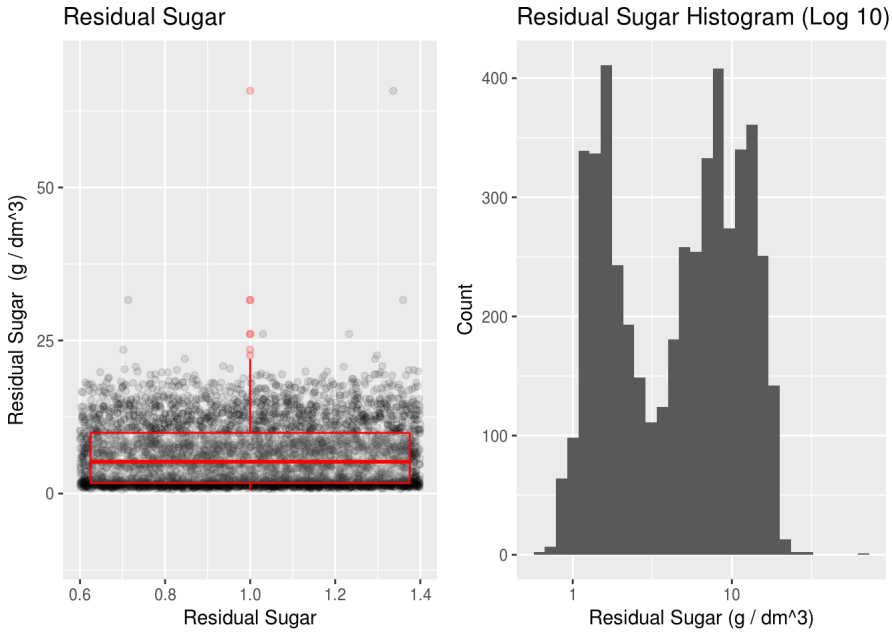

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

Volatile acidity is positively skewed. There are outliers in observations. As stated in dataset description, both fixed and volatile acidity are related with acidity of the wine. So that there might be a relation between those two variables which will be analyzed further later in this report.



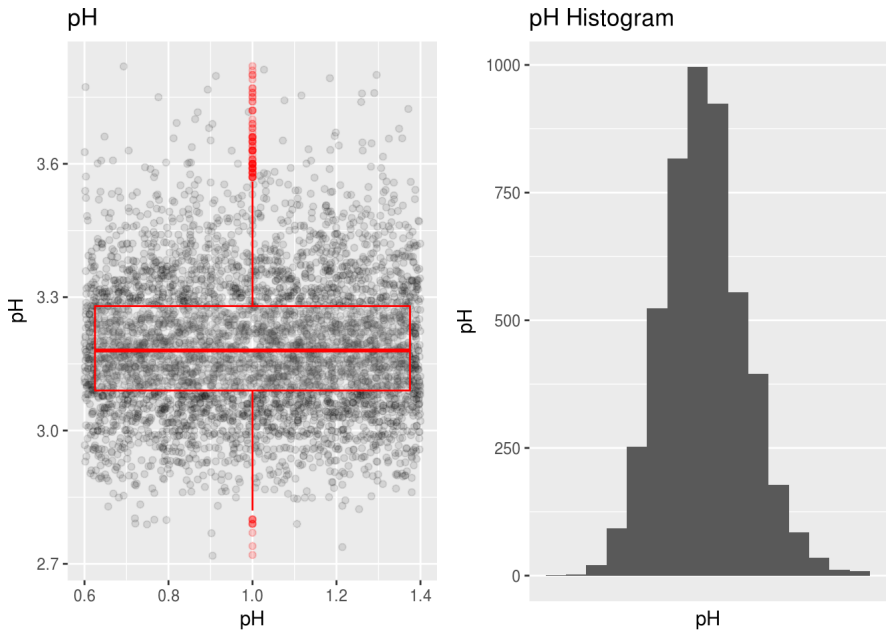

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

Citric acid values of wines are mostly fall between 0.49 and 0.74. Values are long tailed so that there are outliers in the dataset.

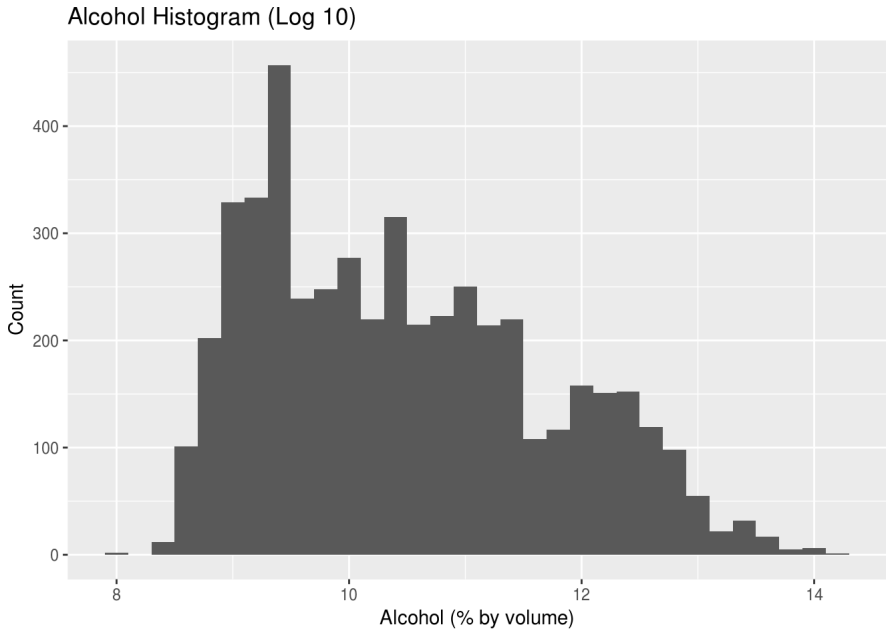


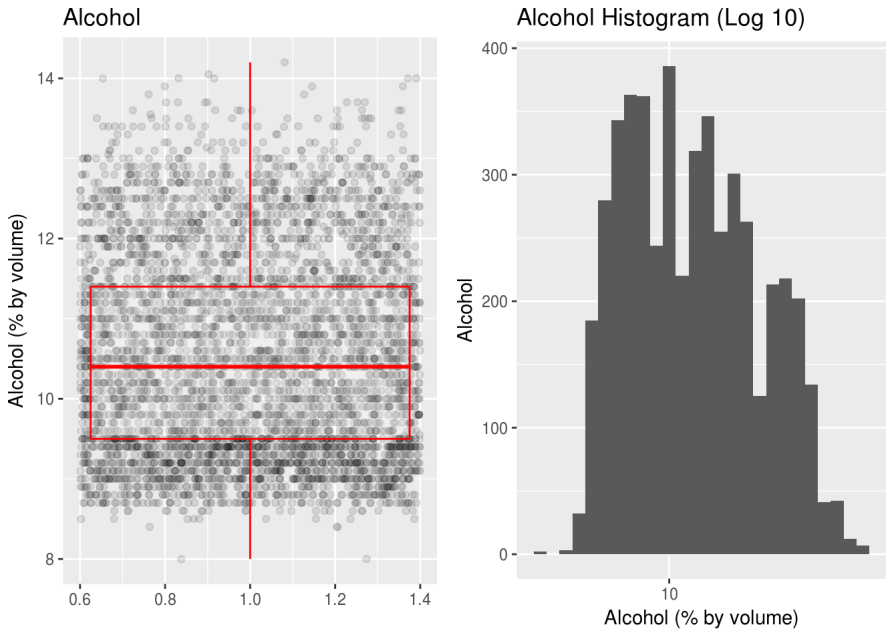

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

Residual sugar is highly skewed. There are extreme outliers as 65.8. So that in the plot of residual sugar above the x axis is log transformed to see the dsitribution much easier.



PH values of observations are seem to have a normal distirbution.






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

Alcohol values are skewed positively. Most of the values fall between 7 and 14.

Univariate Analysis

What is the structure of your dataset?

The dataset consists of 4898 observations which includes 12 variables and the quality is the target variable (total 13) which meant to be vary between 0 and 10. The variable X is the observation number (identifier) and the rest of the variables are numerical.

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

The main feature of interest is quality, obviously. However, alcohol, density, acidity variables and sugar are also features of interest.

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

The acid-related features such as fixed acidity, citric acid, pH are of interest since thay might be related with the quality. Moreover, residual suger and the alcohol level may also be related with the level of quality.

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

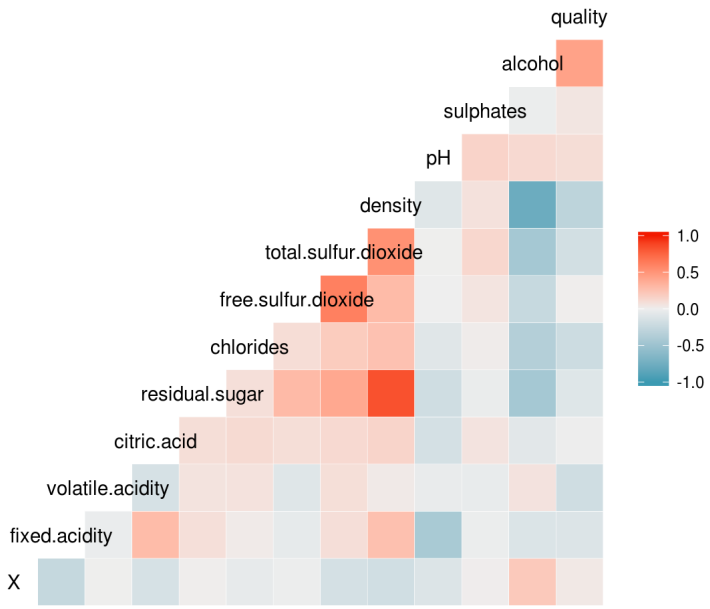
Yes. Although quality values are numeric, those values represents an oredered factor. So I have created a categorical varaible “quality.level” which represent the quality level of wine which I have described in previous section.

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?

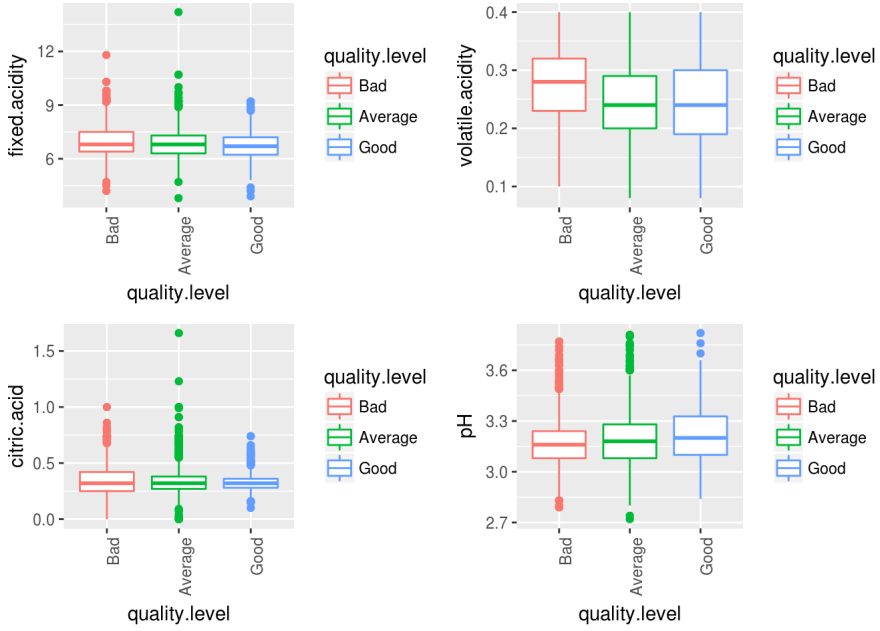
Resiual sugar has a log tail distribution and the histogram was skewed. Therefore in order to investigate this variable I applied log transformation to residual sugar values.

Bivariate Plots Section

In order to see the correlations between variables, a correlation matrix is visualized.



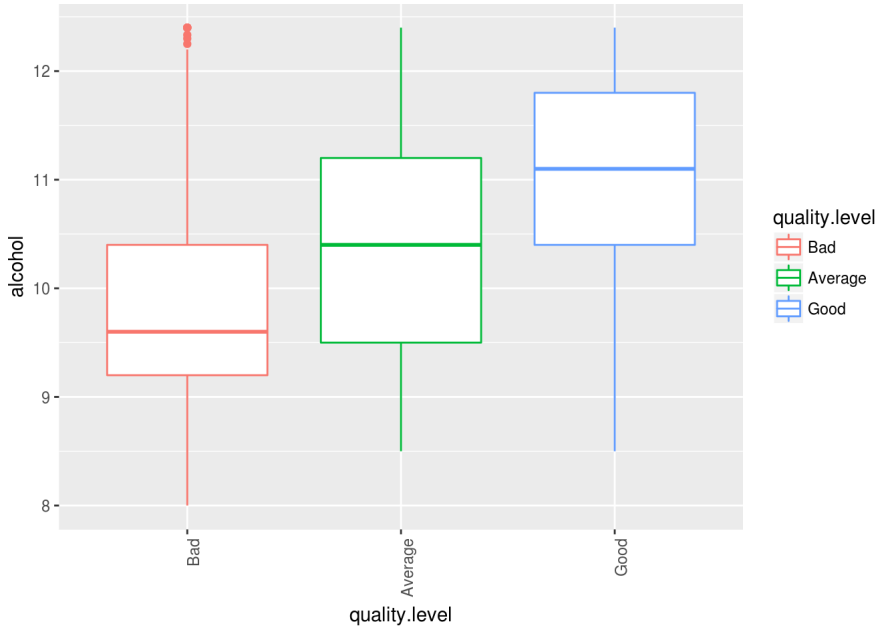
First of all, I have investigate the relation between quality variable and acidity variables. As seen in the correlation matrix, volatile.acidity seem to be negatively correlated with quality.




```
##
##  Pearson's product-moment correlation
##
## data:  quality and volatile.acidity
## t = -12.286, df = 4848, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2009321 -0.1463411
## sample estimates:
##       cor
## -0.1737701
```

Excluding the outliers, wine quality may be negatively correalted with volatile.acidity, however we cannot say this is exact since the correlation coefficient is very low. The description of the dataset indicates that the more the volatile acidity, the taste of the wine becomes more like a vinegar. Other acidity variables do not seem to be correlated with quality.

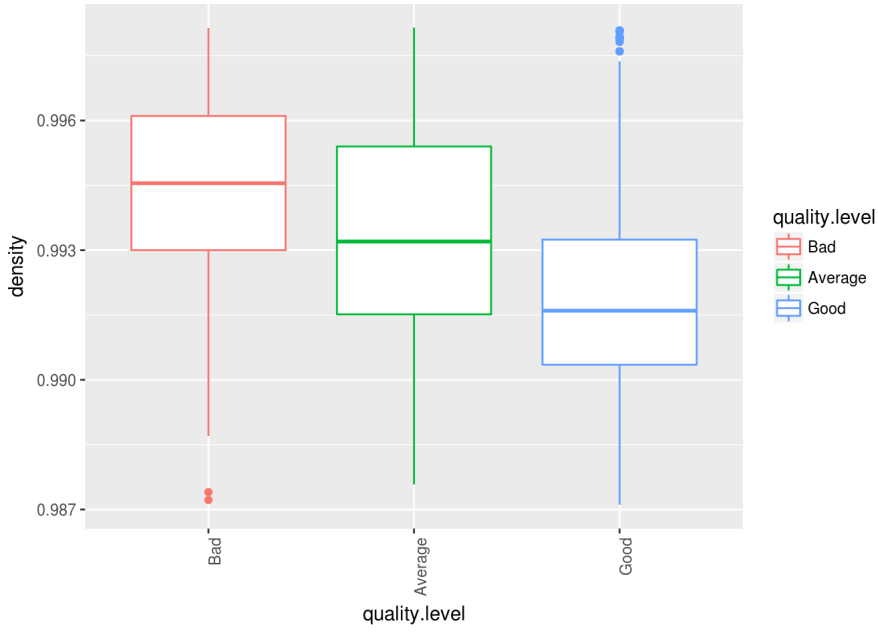
Let's see how alcohol level is related with quality. As seen the correlation matrix, alcohol is the highest correlated variable with quality.




```
##
##  Pearson's product-moment correlation
##
## data:  quality and alcohol
## t = 33.279, df = 4855, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4077970 0.4536048
## sample estimates:
##       cor
## 0.4309785
```

When outliers removed from dataset, alcholol seems to be positively correlated with quality (again Pearson Coeficcient very low to indicate a strong correlation).

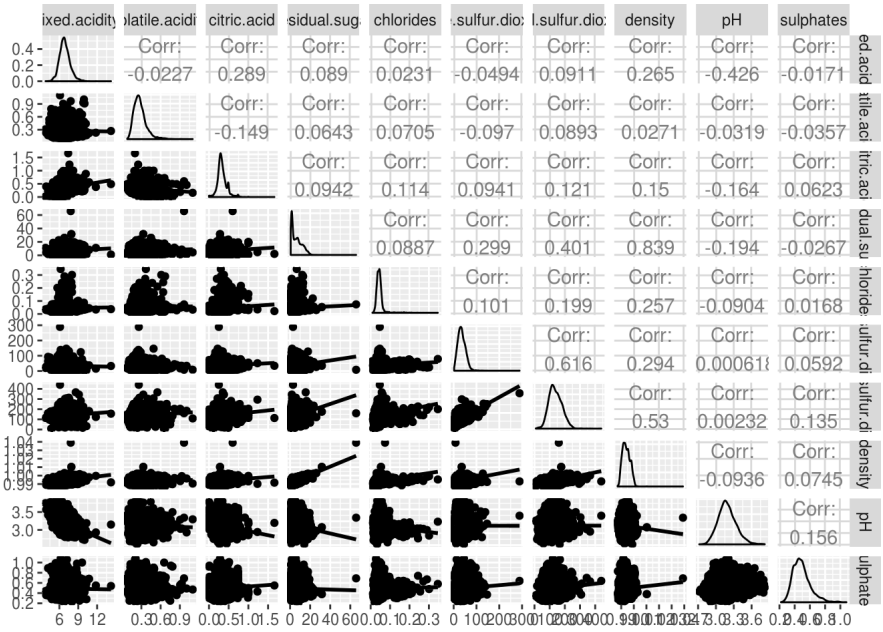
Now I want to investigate the relationship between quality and density. The correlation matrix plot shows that those two variables are corelated.




```
##
##  Pearson's product-moment correlation
##
## data:  quality and density
## t = -23.578, df = 4847, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3457932 -0.2952862
## sample estimates:
##       cor
## -0.3207677
```

Density and the quality of wine seems to be negatively correlated.

Now let's see how the variables other than quality correlated with each other. In order to clearly see the relations between variables, the plot below is created.



The most corelated variables are "denisty" and "residual.sugar". The correlation co-efficient for these two variable is 0.839. "total.sulfur.dioxide" and "free.sulfur.dioxide" are also positively correlated (0.616). All the variables related with acidicty are negatively correlated with pH which is a well-known fact in chemistry.

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?

Feature of interest is selected as quality. The quality variable has highest poistive correlation with alcohol (0.436) and highest negative correlation with density(-0.307).

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

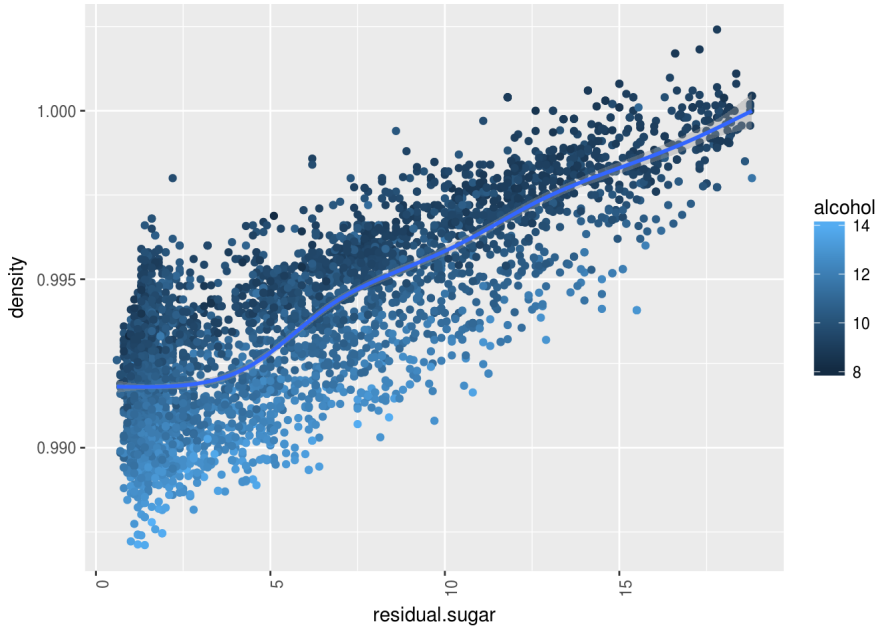
Yes. When all variables other than quality is investigated, density and residual sugar have strong positive correlation with the correlation coefficient of 0.839. The sulfur dioxide variables are also positively correlated with each other. pH and acidicity variables have negative correlation.

What was the strongest relationship you found?

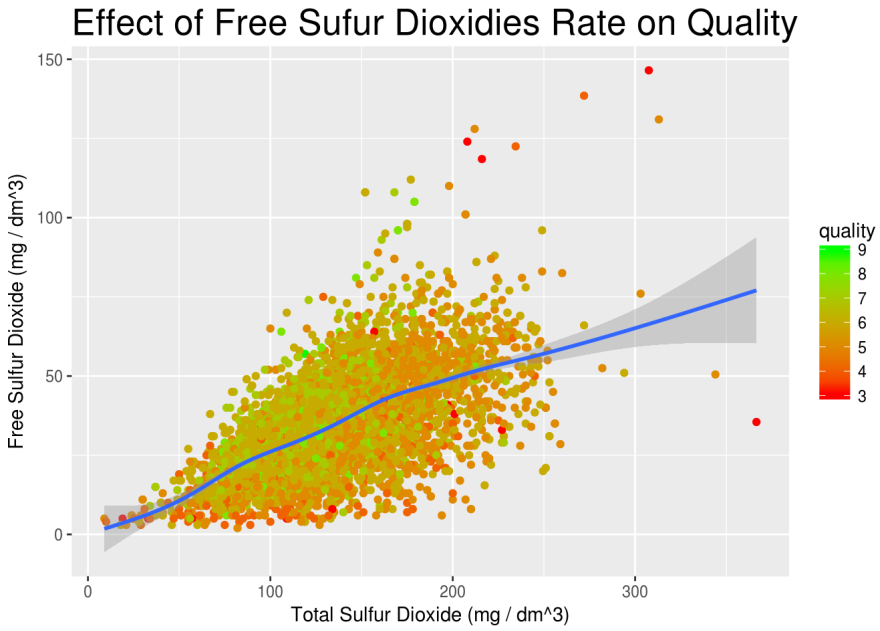
Residual sugar and density.

Multivariate Plots Section

In the bivariete analysis, it is observed that the highest correlated variable with quality was alcohol. Now, in the multivariate analysis section, I want to investigate the effect of multiple variables on quality.



For the same density, when the residual suger increases, the alcohol content is also increases. Since we have shown that alcohol and quality are positively correlated, residual sugar and the density have also effect on the quality of wine.



I assume that every white wine contains sulfur dioxide. There are different types of sulfur dioxides in wine. In the dataset, free sulfur dioxidies and the total sulfur dioxidies are present. If we want to investigate the impact of free sulfur dioxides on quality of wine among other dioxides, we can analyse three variables “total.sulfur.dioxide”, “free.sulfur.dioxide” and “quality” variables together. As seen in the plot above, for a certain amount of sulfur dioxides, high quality wines contains more free sulfur dioxides with respect to bad ones.

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?

I observed the relationship between the residual sugar, density and quality. For the same density, “Good” wines have more residual sugar and “Bad” ones have less. Probably, density and the residual sugar affects the level of alcohol which also affects the quality. Moreover the type of sulfur diodixes used in wine affects the quality. Free sulfur dioxides makes the quality of wine higher.

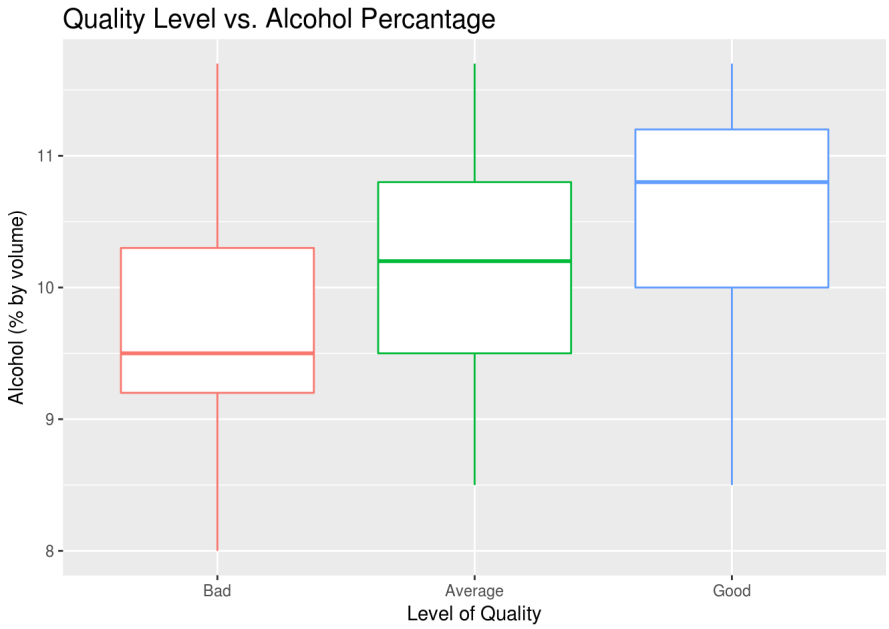
Were there any interesting or surprising interactions between features?

Sulfur dioxides was not seem to be related much with the quality of wine. However, in multivariate analysis, it is observed that the rate of free sulfur dioxidies is related with the quality of wine.

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

Final Plots and Summary

Plot One



Description One

This plot shows that alcohol content percantage is higher in good wines with respect to average and bad wines. In the box plot, for each level of quality mean values of alcohol percantage are clearly seen.

Plot Two

White Wine Data Exploration by Arda Tasci

Arda Tasci

September 22, 2017

Univariate Plots Section

```
## [1] 4898   13
```

There are 4898 observations with 13 variables in the dataset which are :

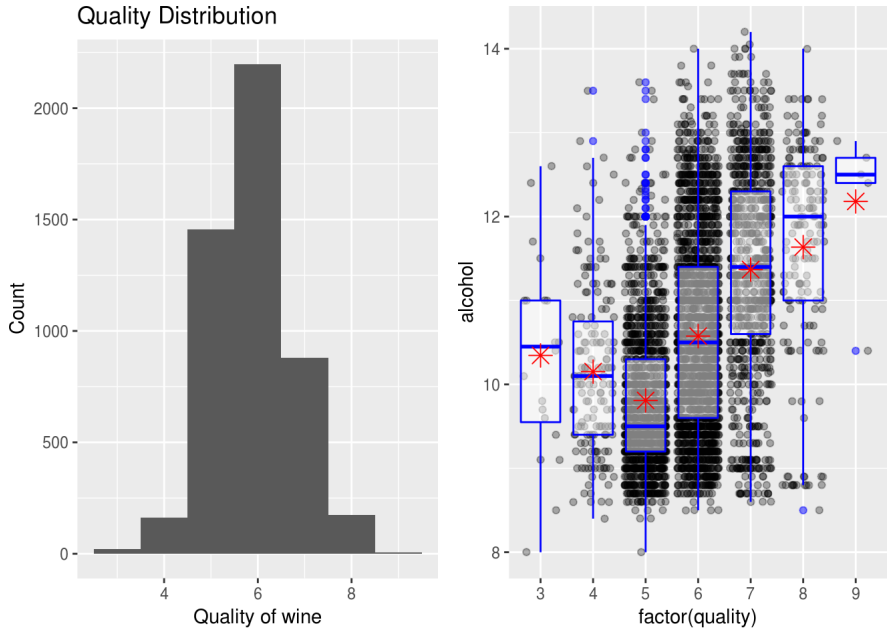
```
##  [1] "X"                    "fixed.acidity"        "volatile.acidity"
##  [4] "citric.acid"          "residual.sugar"       "chlorides"
##  [7] "free.sulfur.dioxide"  "total.sulfur.dioxide" "density"
## [10] "pH"                   "sulphates"            "alcohol"
## [13] "quality"
```

```
## 'data.frame':    4898 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ 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 0.044 ...
##  $ 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 ...
```

```
##        X        fixed.acidity    volatile.acidity  citric.acid
##  Min.   :   1   Min.   : 3.800   Min.   :0.0800   Min.   :0.0000
##  1st Qu.:1225   1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700
##  Median :2450   Median : 6.800   Median :0.2600   Median :0.3200
##  Mean   :2450   Mean   : 6.855   Mean   :0.2782   Mean   :0.3342
##  3rd Qu.:3674   3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900
##  Max.   :4898   Max.   :14.200   Max.   :1.1000   Max.   :1.6600
##  residual.sugar     chlorides       free.sulfur.dioxide
##  Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
##  1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
##  Median : 5.200   Median :0.04300   Median : 34.00
##  Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
##  3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
##  Max.   :65.800   Max.   :0.34600   Max.   :289.00
##  total.sulfur.dioxide    density             pH          sulphates
##  Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200
##  1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100
##  Median :134.0        Median :0.9937   Median :3.180   Median :0.4700
##  Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898
##  3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500
##  Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800
##     alcohol         quality
##  Min.   : 8.00   Min.   :3.000
##  1st Qu.: 9.50   1st Qu.:5.000
##  Median :10.40   Median :6.000
##  Mean   :10.51   Mean   :5.878
##  3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :14.20   Max.   :9.000
```

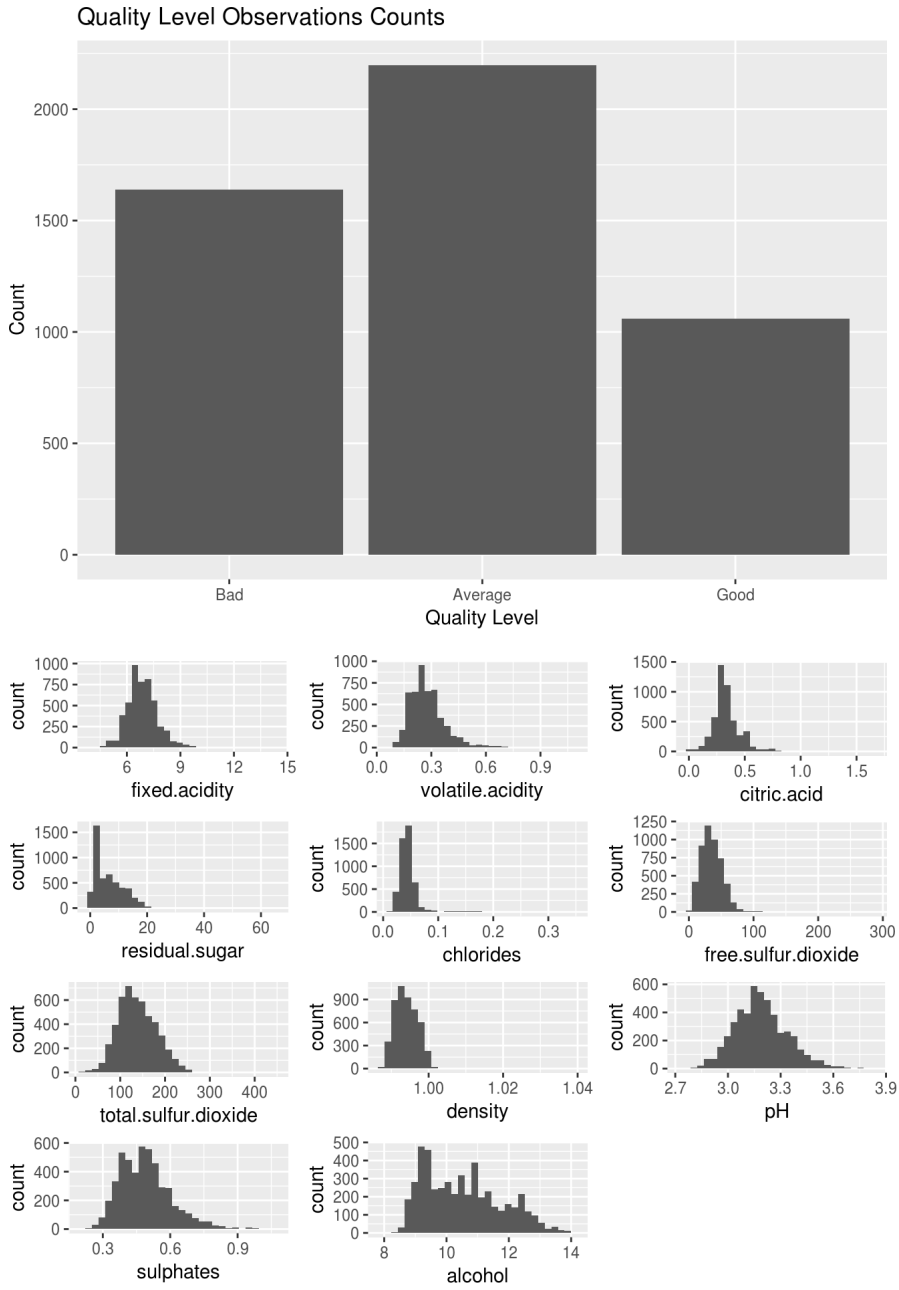
The target variable is the quality which depends on the other variables. Although the scale of quality shall be between 1 and 10, the quality metric varies between 3 and 9. Median is 6 and mean is 5.878. Most of the observations probably have quality of 6 based on this distribution.

Quality variable is a discrete variable which shall be cast as an ordered factor.

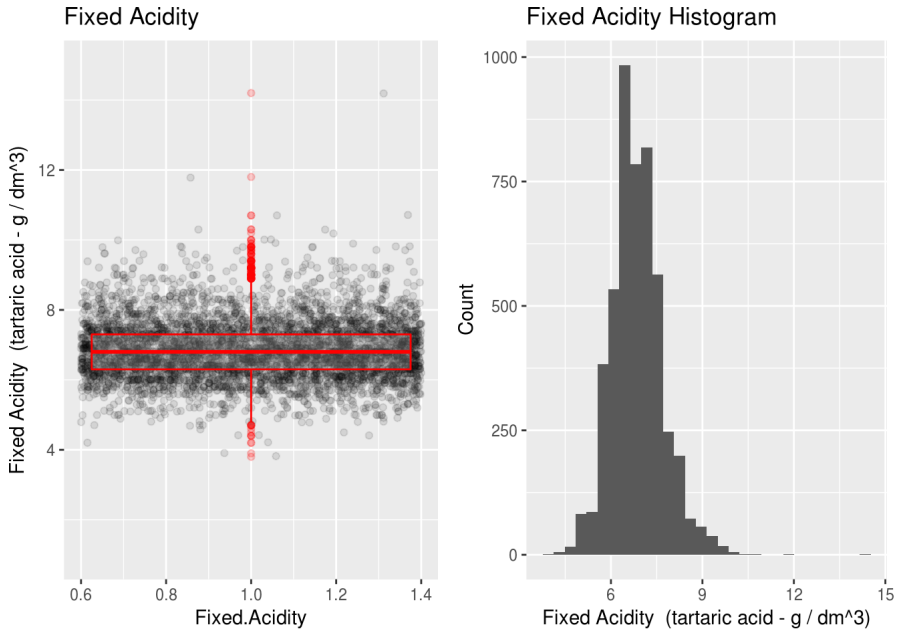



```
##
##    3    4    5    6    7    8    9
##   20  163 1457 2198  880  175    5
```

As I stated, highest count of observations have quality of 6. Most of the wines are graded 5, 6, and 7. Quality is an ordered factor variable so that I created a new variable using quality which is quality.level having values “Bad”, “Average” and “Good”. The quality values of this variable is (0,5], (5, 6], (6,9] respectively.

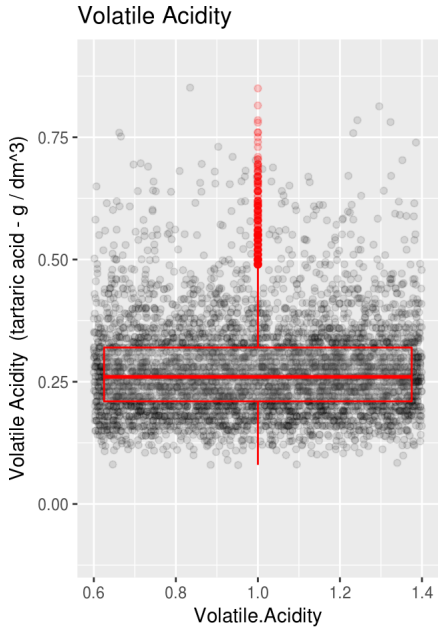


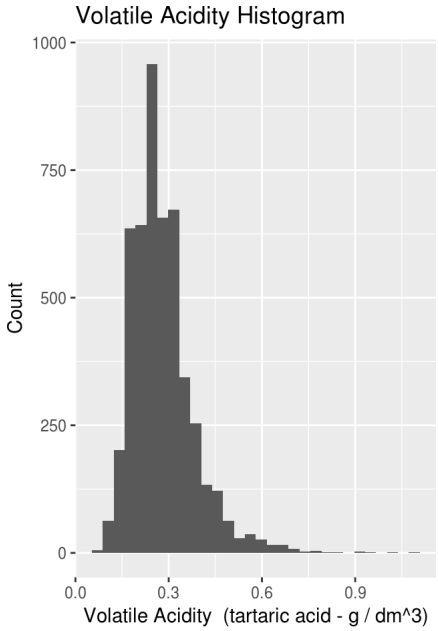
Let's look deeper into the variables.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.800   6.300   6.800   6.855   7.300  14.200
```

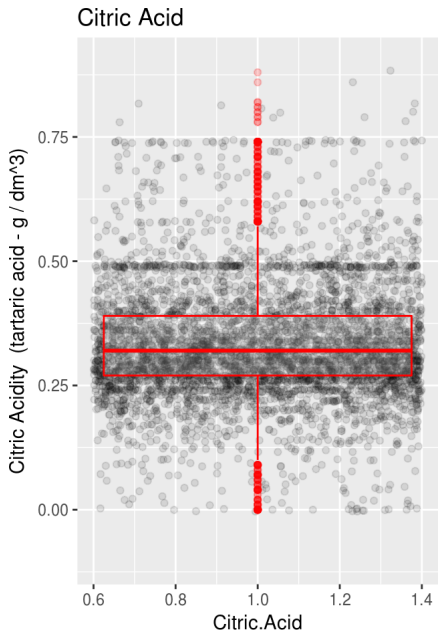
As seen Fixed Acidity values are mostly fall between 6 and 7.3. Observations with fixed.acidity values less than 5 and more than 10 can be accepted as outliers.

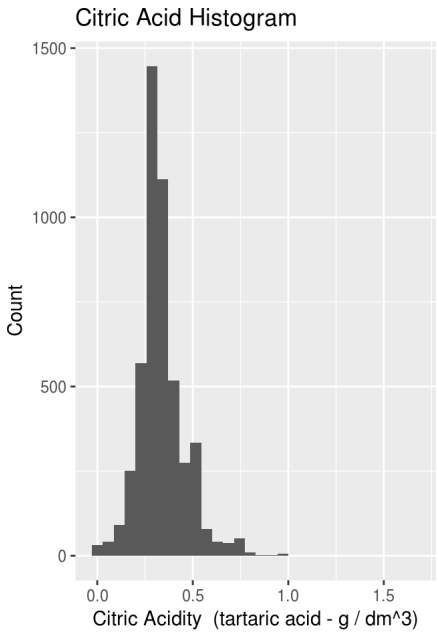





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2100  0.2600  0.2782  0.3200  1.1000
```

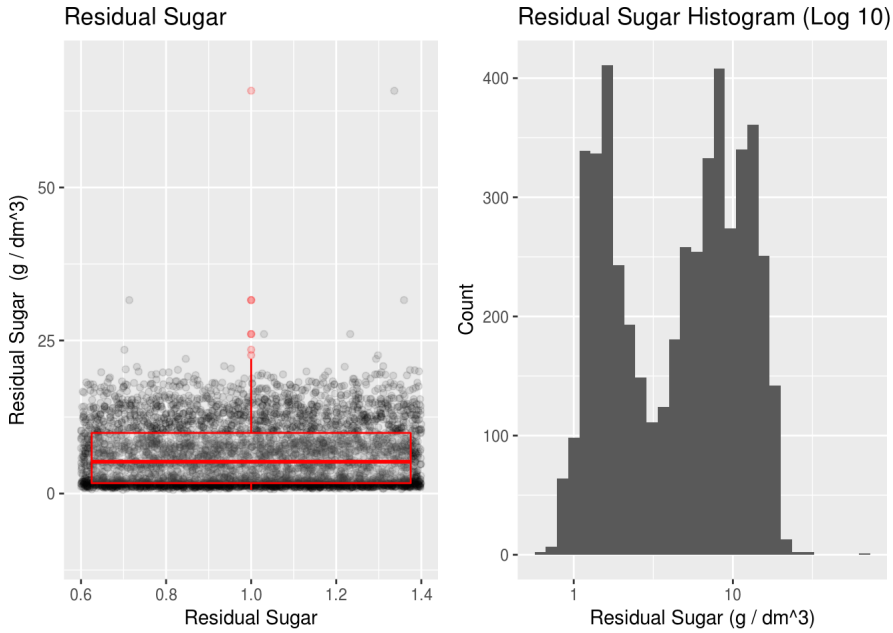
Volatile acidity is positively skewed. There are outliers in observations. As stated in dataset description, both fixed and volatile acidity are related with acidity of the wine. So that there might be a relation between those two variables which will be analyzed further later in this report.





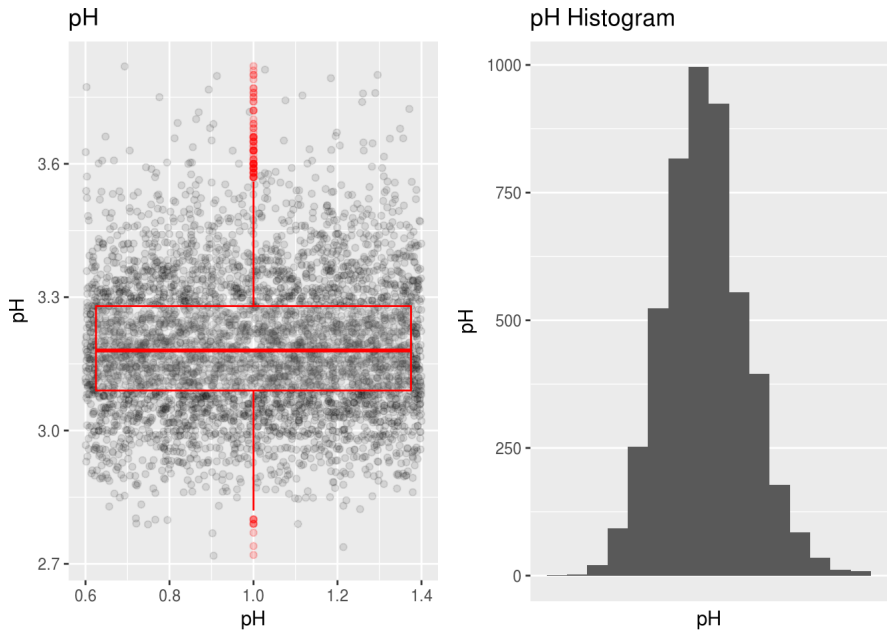

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2700  0.3200  0.3342  0.3900  1.6600
```

Citric acid values of wines are mostly fall between 0.49 and 0.74. Values are long tailed so that there are outliers in the dataset.

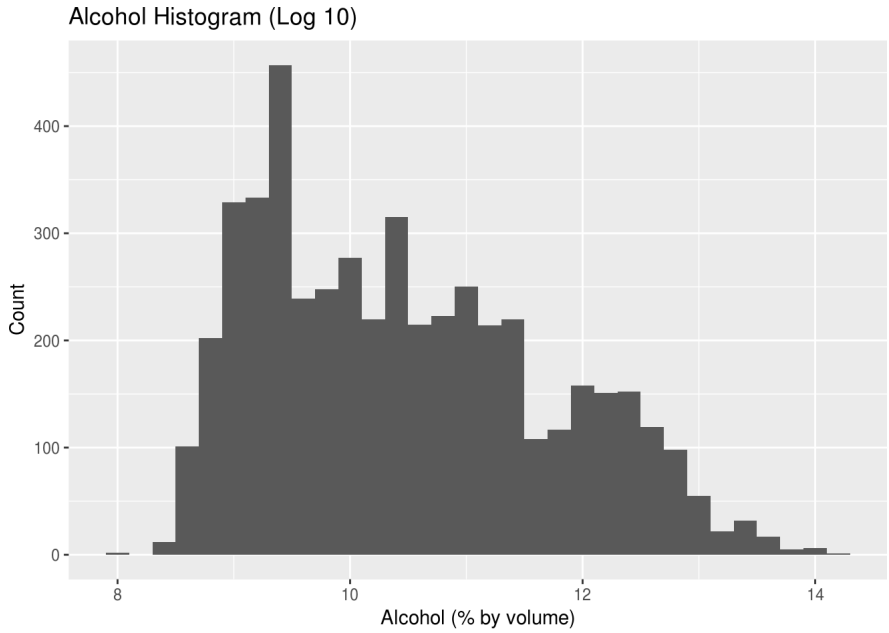


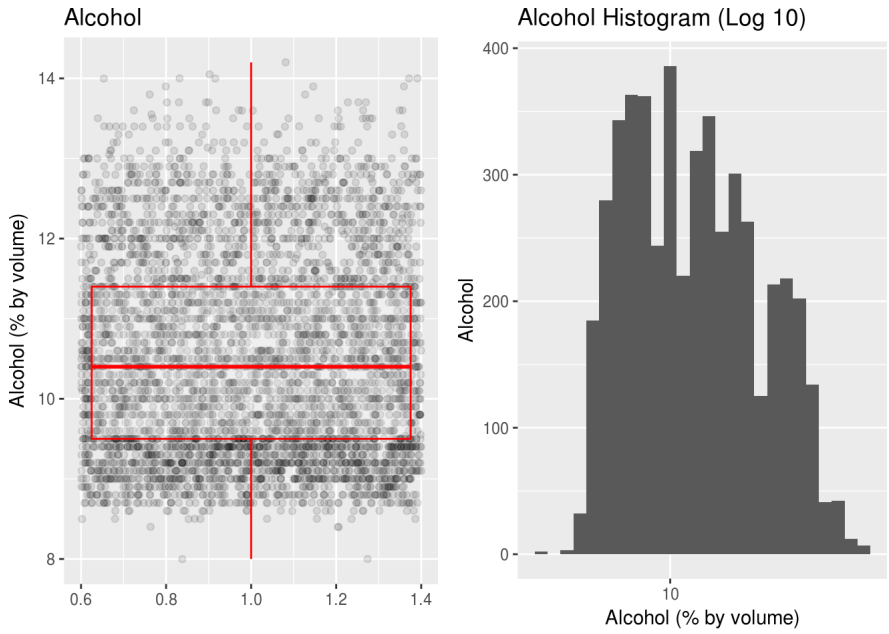

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.600   1.700   5.200   6.391   9.900  65.800
```

Residual sugar is highly skewed. There are extreme outliers as 65.8. So that in the plot of residual sugar above the x axis is log transformed to see the dsitribution much easier.



PH values of observations are seem to have a normal distirbution.






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

Alcohol values are skewed positively. Most of the values fall between 7 and 14.

Univariate Analysis

What is the structure of your dataset?

The dataset consists of 4898 observations which includes 12 variables and the quality is the target variable (total 13) which meant to be vary between 0 and 10. The variable X is the observation number (identifier) and the rest of the variables are numerical.

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

The main feature of interest is quality, obviously. However, alcohol, density, acidity variables and sugar are also features of interest.

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

The acid-related features such as fixed acidity, citric acid, pH are of interest since thay might be related with the quality. Moreover, residual suger and the alcohol level may also be related with the level of quality.

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

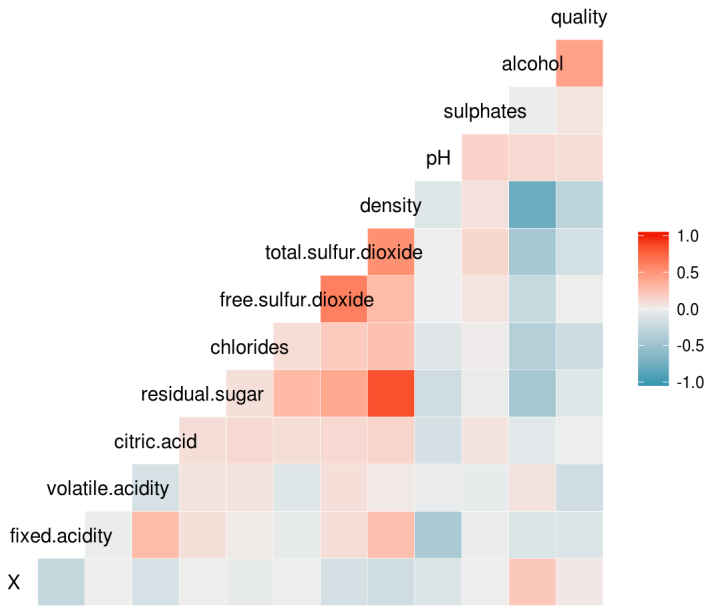
Yes. Although quality values are numeric, those values represents an oredered factor. So I have created a categorical varaible “quality.level” which represent the quality level of wine which I have described in previous section.

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?

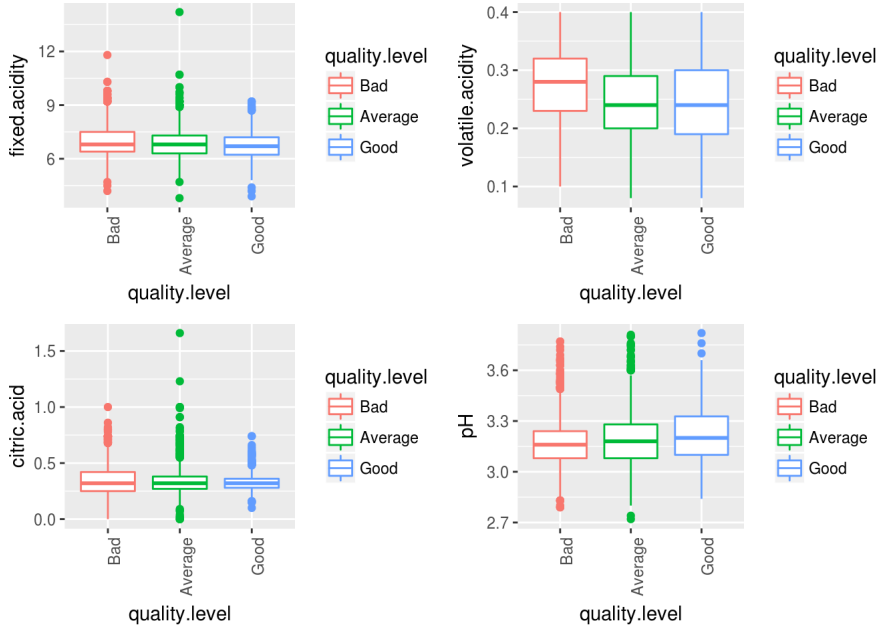
Resiual sugar has a log tail distribution and the histogram was skewed. Therefore in order to investigate this variable I applied log transformation to residual sugar values.

Bivariate Plots Section

In order to see the correlations between variables, a correlation matrix is visualized.



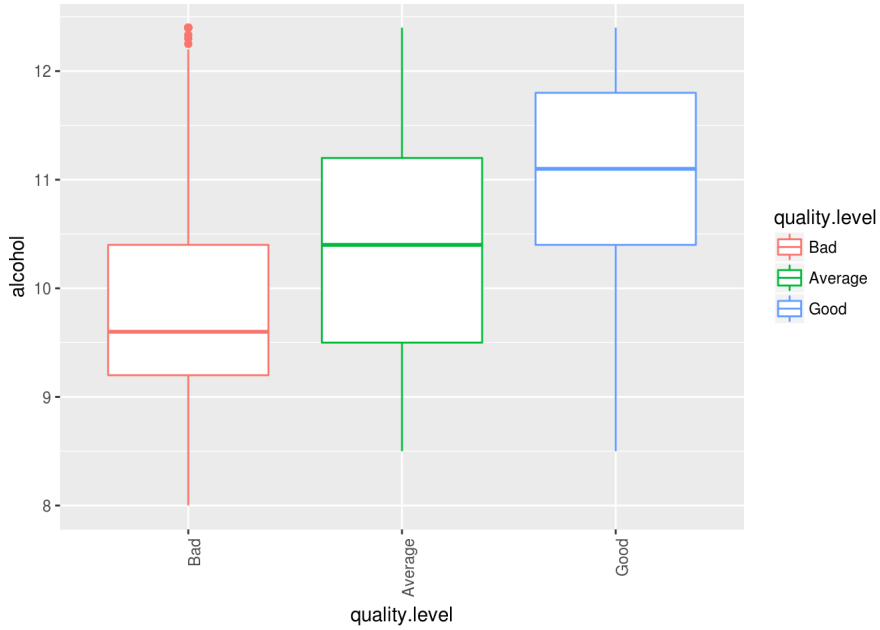
First of all, I have investigate the relation between quality variable and acidity variables. As seen in the correlation matrix, volatile.acidity seem to be negatively correlated with quality.




```
##
##  Pearson's product-moment correlation
##
## data:  quality and volatile.acidity
## t = -12.286, df = 4848, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2009321 -0.1463411
## sample estimates:
##       cor
## -0.1737701
```

Excluding the outliers, wine quality may be negatively correalted with volatile.acidity, however we cannot say this is exact since the correlation coefficient is very low. The description of the dataset indicates that the more the volatile acidity, the taste of the wine becomes more like a vinegar. Other acidity variables do not seem to be correlated with quality.

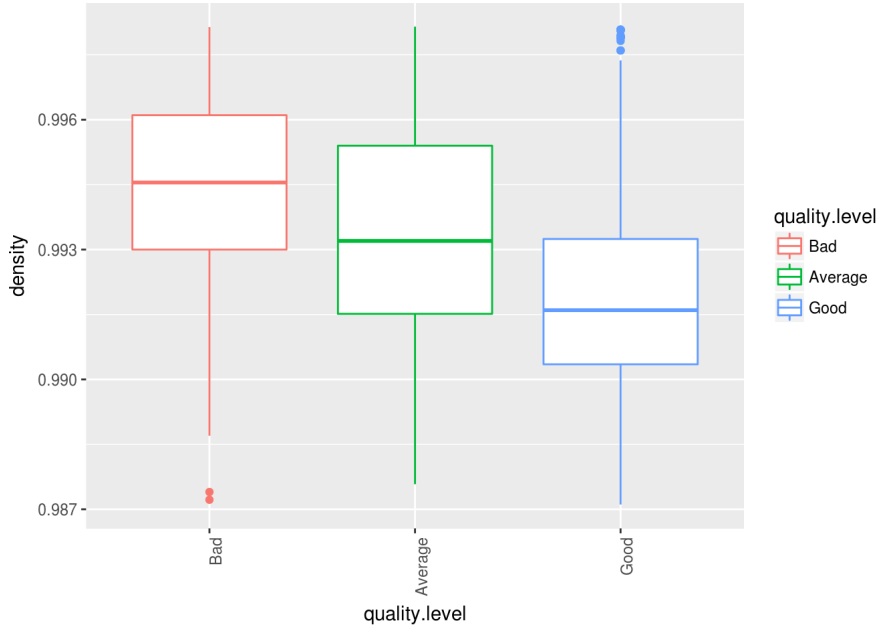
Let's see how alcohol level is related with quality. As seen the correlation matrix, alcohol is the highest correlated variable with quality.




```
##
##  Pearson's product-moment correlation
##
## data:  quality and alcohol
## t = 33.279, df = 4855, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4077970 0.4536048
## sample estimates:
##       cor
## 0.4309785
```

When outliers removed from dataset, alcholol seems to be positively correlated with quality (again Pearson Coeficcient very low to indicate a strong correlation).

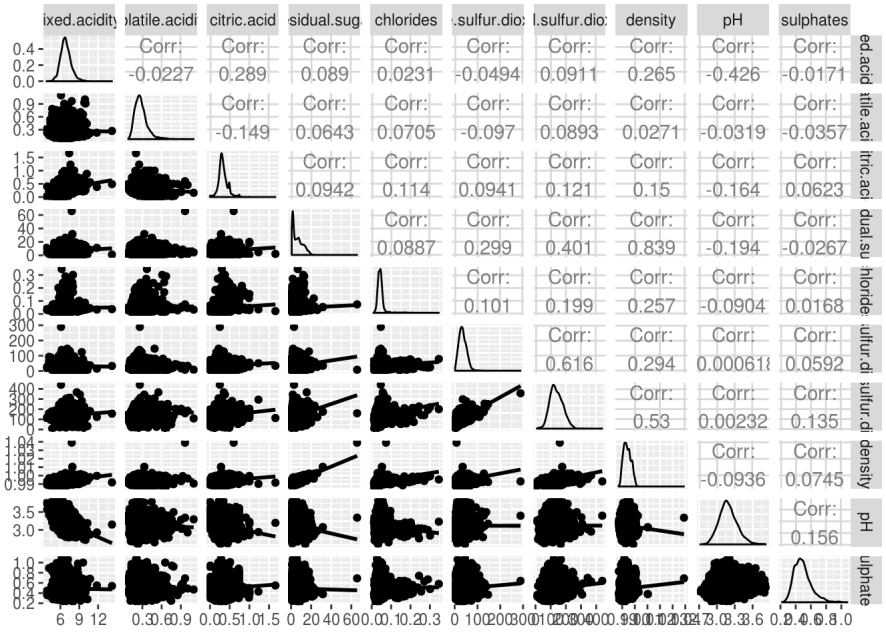
Now I want to investigate the relationship between quality and density. The correlation matrix plot shows that those two variables are corelated.




```
##
##  Pearson's product-moment correlation
##
## data:  quality and density
## t = -23.578, df = 4847, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3457932 -0.2952862
## sample estimates:
##       cor
## -0.3207677
```

Density and the quality of wine seems to be negatively correlated.

Now let's see how the variables other than quality correlated with each other. In order to clearly see the relations between variables, the plot below is created.



The most corelated variables are "denisty" and "residual.sugar". The correlation co-efficient for these two variable is 0.839. "total.sulfur.dioxide" and "free.sulfur.dioxide" are also positively correlated (0.616). All the variables related with acidicty are negatively correlated with pH which is a well-known fact in chemistry.

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?

Feature of interest is selected as quality. The quality variable has highest poistive correlation with alcohol (0.436) and highest negative correlation with density(-0.307).

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

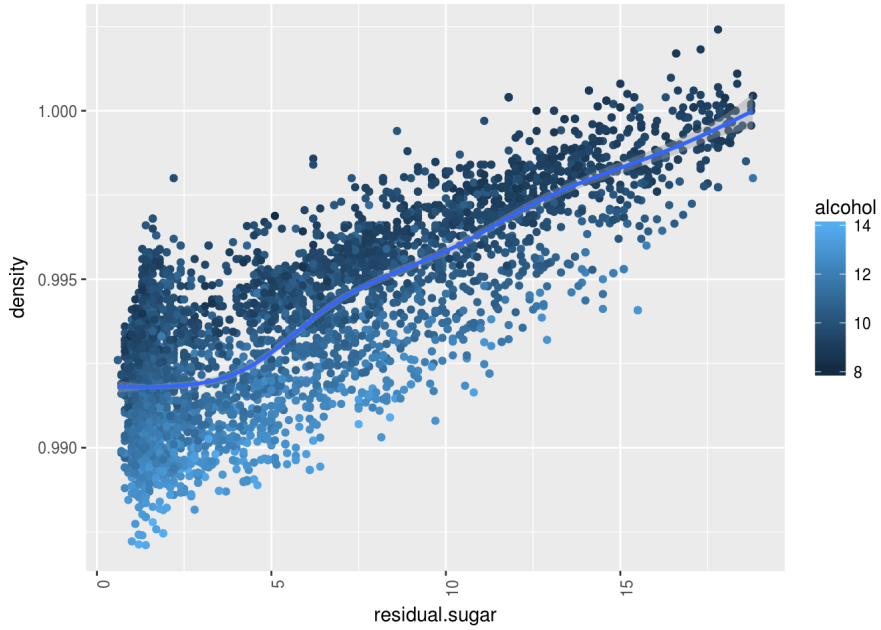
Yes. When all variables other than quality is investigated, density and residual sugar have strong positive correlation with the correlation coefficient of 0.839. The sulfur dioxide variables are also positively correlated with each other. pH and acidicity variables have negative correlation.

What was the strongest relationship you found?

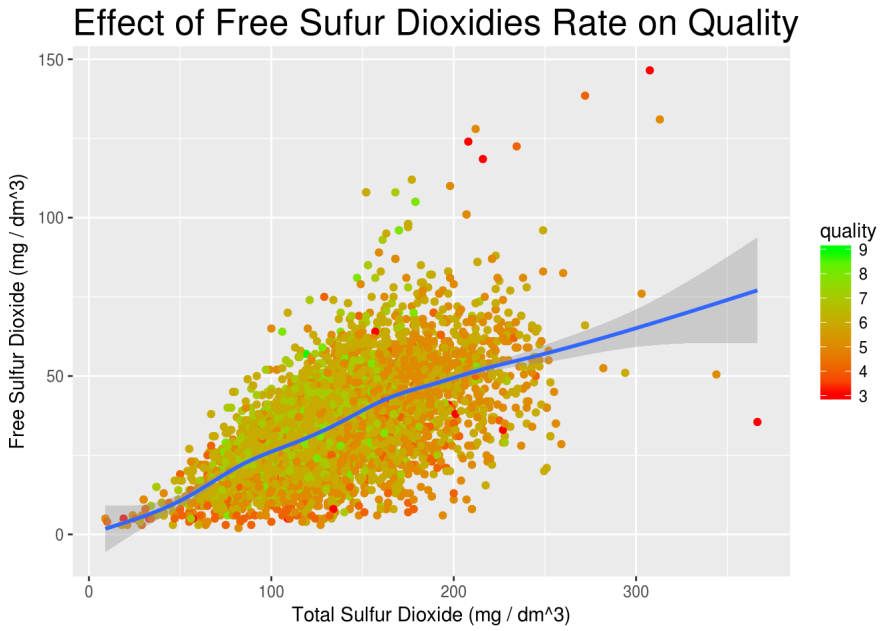
Residual sugar and density.

Multivariate Plots Section

In the bivariete analysis, it is observed that the highest correlated variable with quality was alcohol. Now, in the multivariate analysis section, I want to investigate the effect of multiple variables on quality.



For the same density, when the residual suger increases, the alcohol content is also increases. Since we have shown that alcohol and quality are positively correlated, residual sugar and the density have also effect on the quality of wine.



I assume that every white wine contains sulfur dioxide. There are different types of sulfur dioxides in wine. In the dataset, free sulfur dioxidies and the total sulfur dioxidies are present. If we want to investigate the impact of free sulfur dioxides on quality of wine among other dioxides, we can analyse three variables “total.sulfur.dioxide”, “free.sulfur.dioxide” and “quality” variables together. As seen in the plot above, for a certain amount of sulfur dioxides, high quality wines contains more free sulfur dioxides with respect to bad ones.

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?

I observed the relationship between the residual sugar, density and quality. For the same density, “Good” wines have more residual sugar and “Bad” ones have less. Probably, density and the residual sugar affects the level of alcohol which also affects the quality. Moreover the type of sulfur diodixes used in wine affects the quality. Free sulfur dioxides makes the quality of wine higher.

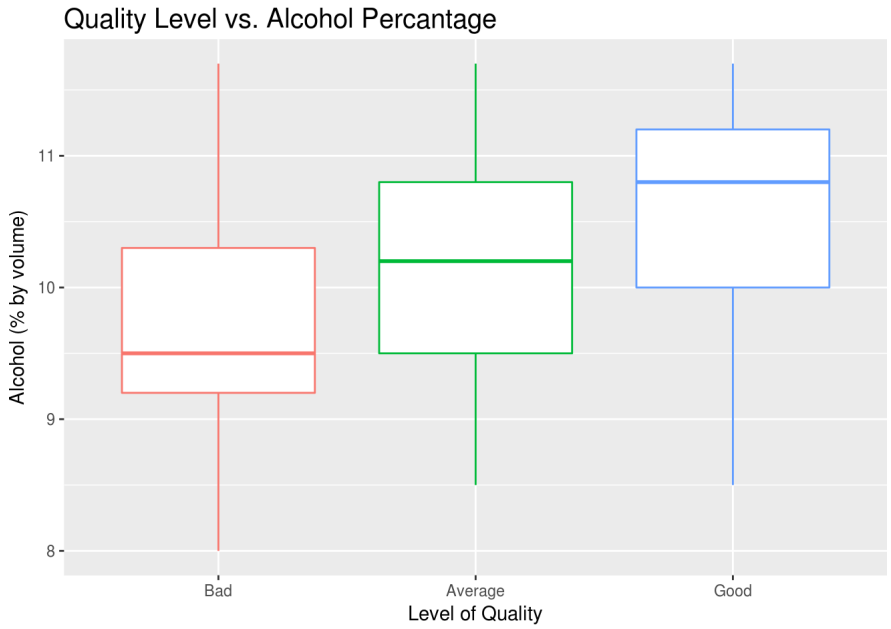
Were there any interesting or surprising interactions between features?

Sulfur dioxides was not seem to be related much with the quality of wine. However, in multivariate analysis, it is observed that the rate of free sulfur dioxidies is related with the quality of wine.

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

Final Plots and Summary

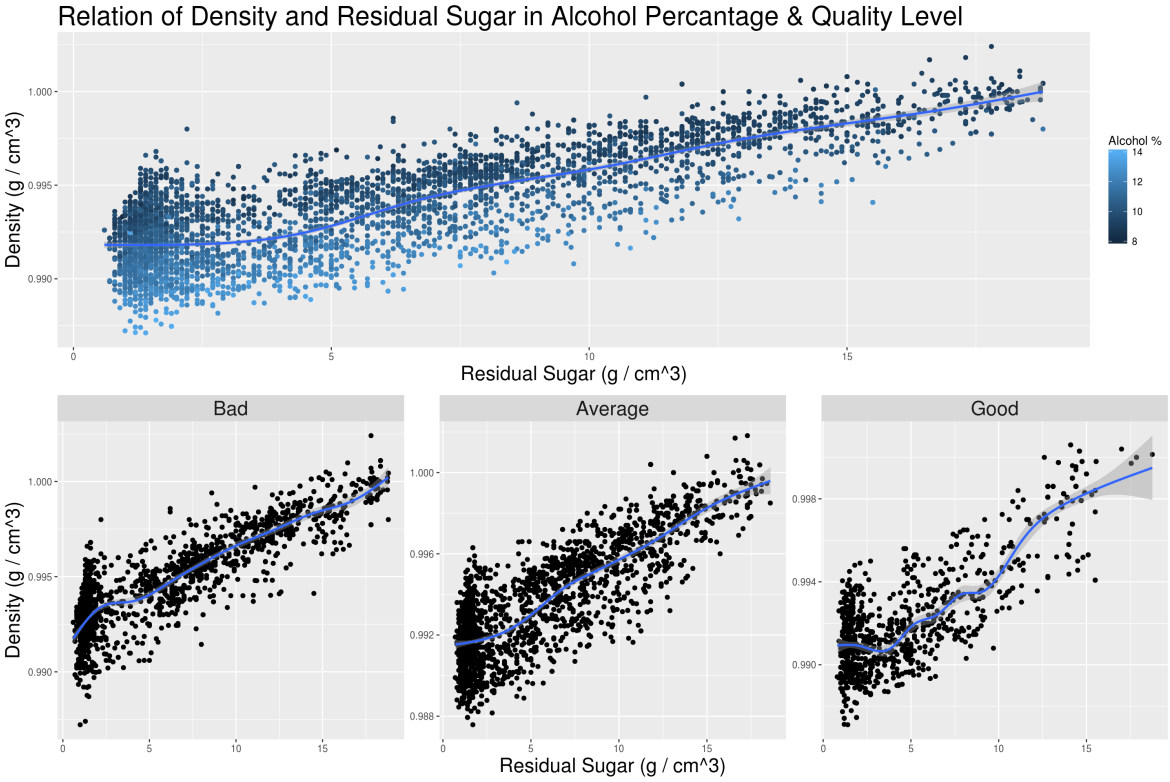
Plot One



Description One

This plot shows that alcohol content percantage is higher in good wines with respect to average and bad wines. In the box plot, for each level of quality mean values of alcohol percantage are clearly seen.

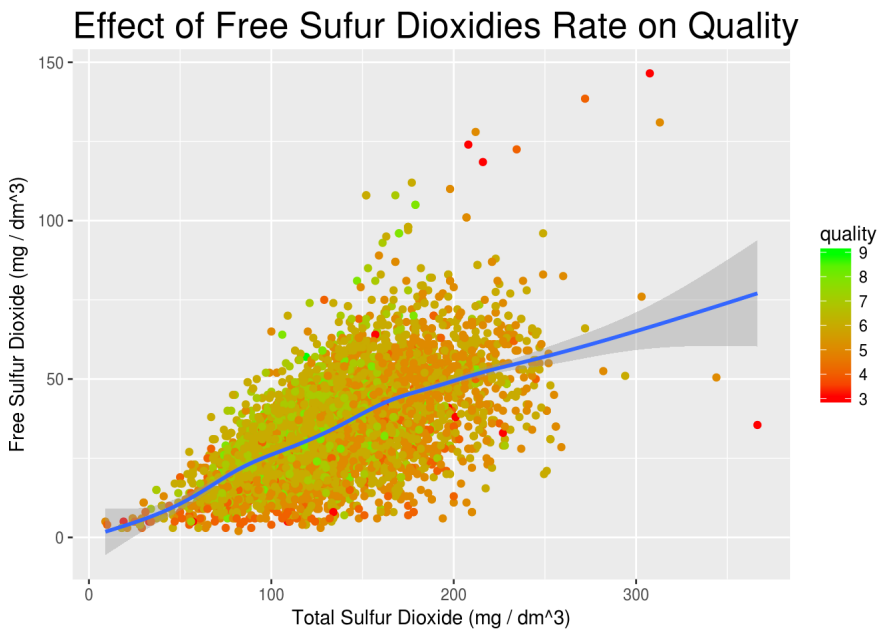
Plot Two



Description Two

This plot clearly shows that density and residual sugar have affect on alcohol level of wine. When the residual suger increases, the alcohol content is also increases. It is clearly seen that the good quality of wines have lower density for the same amount of residual sugar. Since we have shown that alcohol and quality are positively correlated, residual sugar and the density have also effect on the quality of wine.

Plot Three



Description Three

When the amount of free sulfur dioxidies in total sulfur dioxidies increases, the wine quality also increases.

Reflection

First of all, I investigate the distributon of the variables in the dataset. Based on the dataset description, the quality variable is the dependent variable in the dataset. When investigated, it varies between 3 and 9 based on the ratings given by people. Although those values are numerical, actually they are ordered factors. Therefore, I create categories as “Bad”, “Average” and “Good” and cut the quality values into those categories. This helped me a lot in investigating the data easier. Moreover, histograms showed that some of the variables are skewed and have outliers. In the further analysis, I take the subset of the original data by excluding the outliers.

In bivariate analysis section, I have explore the correlation between quality and the other variables in order to see the effect of the variables on quality of wine. Although there is no strong correlation between quality and other variables, the highest correlation was the alcohol level of the wine. Other than the quality, the most correlated variables were residual sugar and density, which I have investigated deeper in multivariate analysis section.

In multivariate analysis section, the relation between residual sugar, density and alcohol was very interesting. Those variables are related with each other (which was explained in detail). Residual sugar and density has affect on the alcohol directly and quality indirectly.

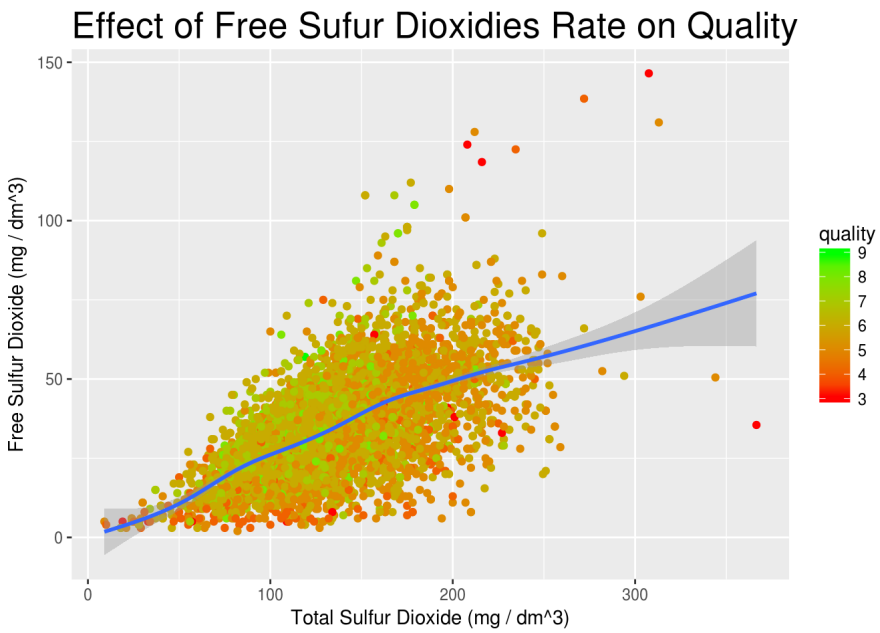
The toughest struggle I have faced was creating plots to reflect the relationsships between a ordered factor and continious variables. Moreover, the correlations were not clear between variables. However, in each section I have analyzed the variables their relations and come up with an idea about the effect of those variables on quality.

N0Qdfbo7vGhLg6WkWdrAQFeTDZfCernZOfN2CXP1dBOyKjXbhJJMyIKaRuxF5GMHdJNqgkQcNTuZPr+6e+WEhwAQoKaihoFaAaiho/K9v14BLt4sBAERA67X\
width="1440" />

Description Two

This plot clearly shows that density and residual sugar have affect on alcohol level of wine. When the residual suger increases, the alcohol content is also increases. It is clearly seen that the good quality of wines have lower density for the same amount of residual sugar. Since we have shown that alcohol and quality are positively correlated, residual sugar and the density have also effect on the quality of wine.

Plot Three



Description Three

When the amount of free sulfur dioxidies in total sulfur dioxidies increases, the wine quality also increases.

Reflection

First of all, I investigate the distributon of the variables in the dataset. Based on the dataset description, the quality variable is the dependent variable in the dataset. When investigated, it varies between 3 and 9 based on the ratings given by people. Although those values are numerical, actually they are ordered factors. Therefore, I create categories as “Bad”, “Average” and “Good” and cut the quality values into those categories. This helped me a lot in investigating the data easier. Moreover, histograms showed that some of the variables are skewed and have outliers. In the further analysis, I take the subset of the original data by excluding the outliers.

In bivariate analysis section, I have explore the correlation between quality and the other variables in order to see the effect of the variables on quality of wine. Although there is no strong correlation between quality and other variables, the highest correlation was the alcohol level of the wine. Other than the quality, the most correlated variables were residual sugar and density, which I have investigated deeper in multivariate analysis section.

In multivariate analysis section, the relation between residual sugar, density and alcohol was very interesting. Those variables are related with each other (which was explained in detail). Residual sugar and density has affect on the alcohol directly and quality indirectly.

The toughest struggle I have faced was creating plots to reflect the relationsships between a ordered factor and continious variables. Moreover, the correlations were not clear between variables. However, in each section I have analyzed the variables their relations and come up with an idea about the effect of those variables on quality.