Red Wine Quality Exploration by Samuel Duan

Univariate Plots Section

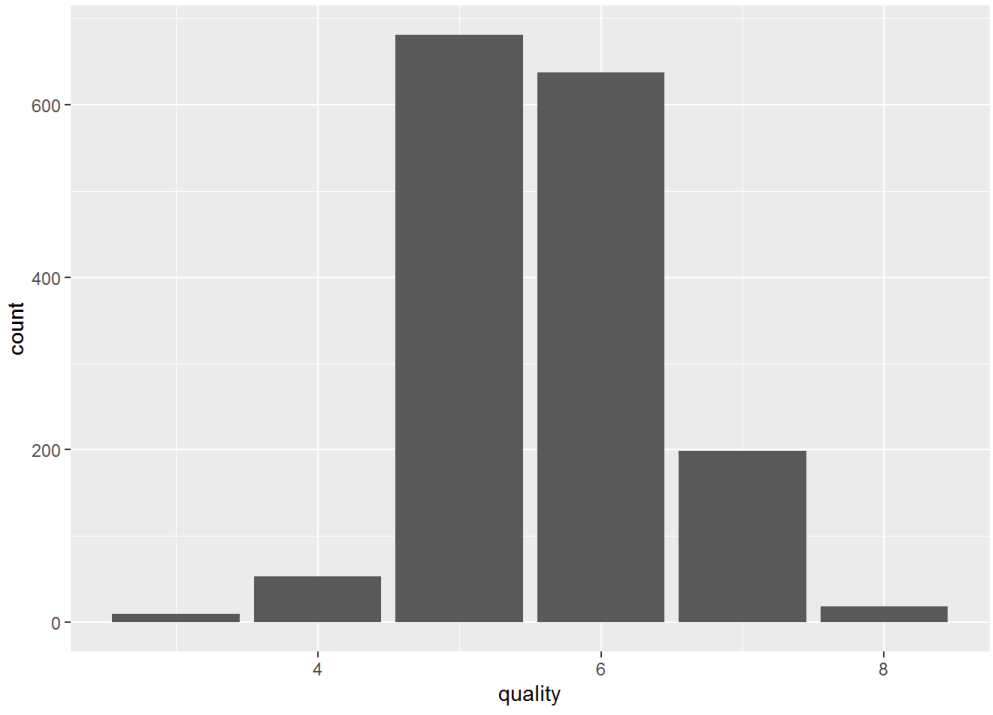
In this section, some preliminary exploration of the dataset is performed, including data set summary and histogram is plotted to show the distribution of each particular variable.

0. Summary of the Data Set

```
## 'data.frame':    1599 obs. of  13 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

This data set contains 1599 observations and 11 chemical attributes that could affect the wine quality.

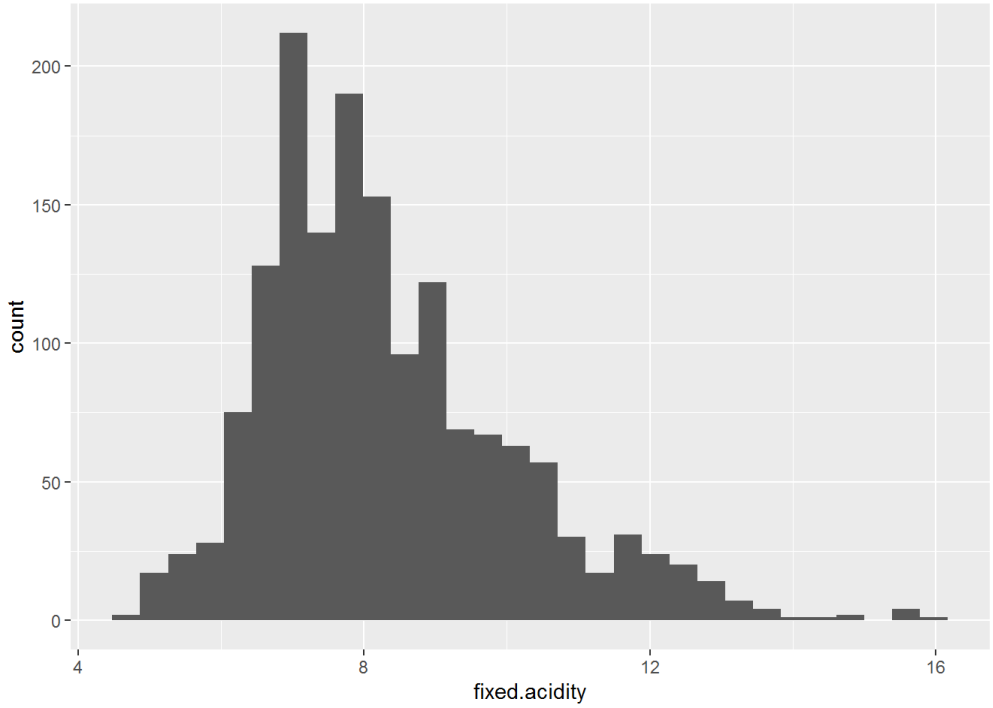
1. Histogram of wine quality




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

As shown in the plot above that the mode of the wine quality is 5.

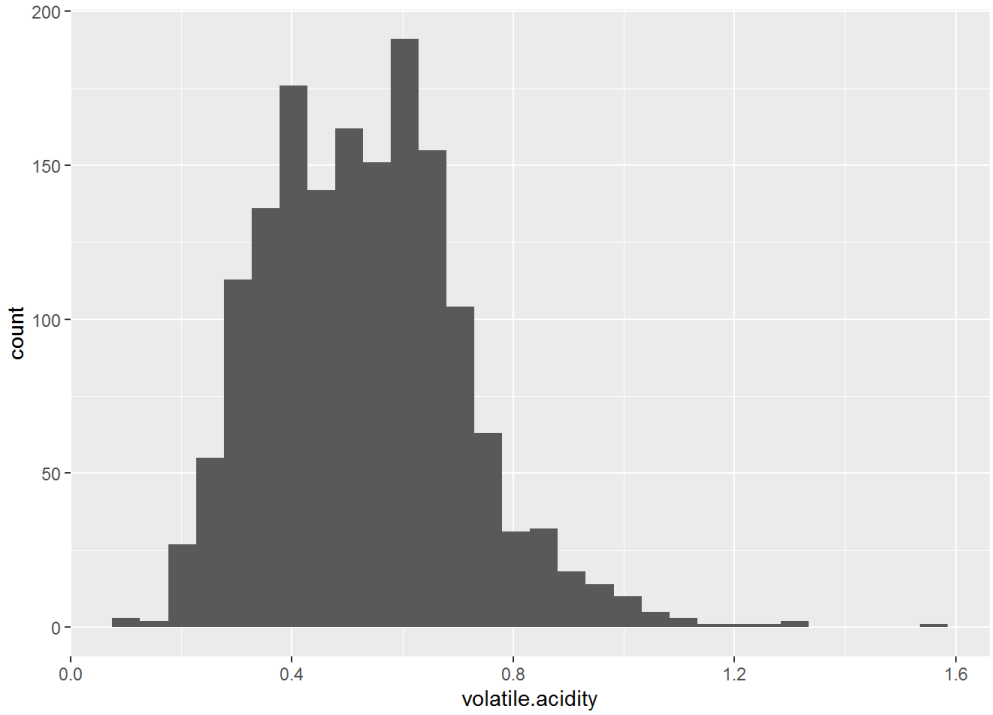
2. Histogram of fixed acidity




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.60    7.10    7.90    8.32    9.20   15.90
```

The distribution of fixed acidity is skewed to the right.

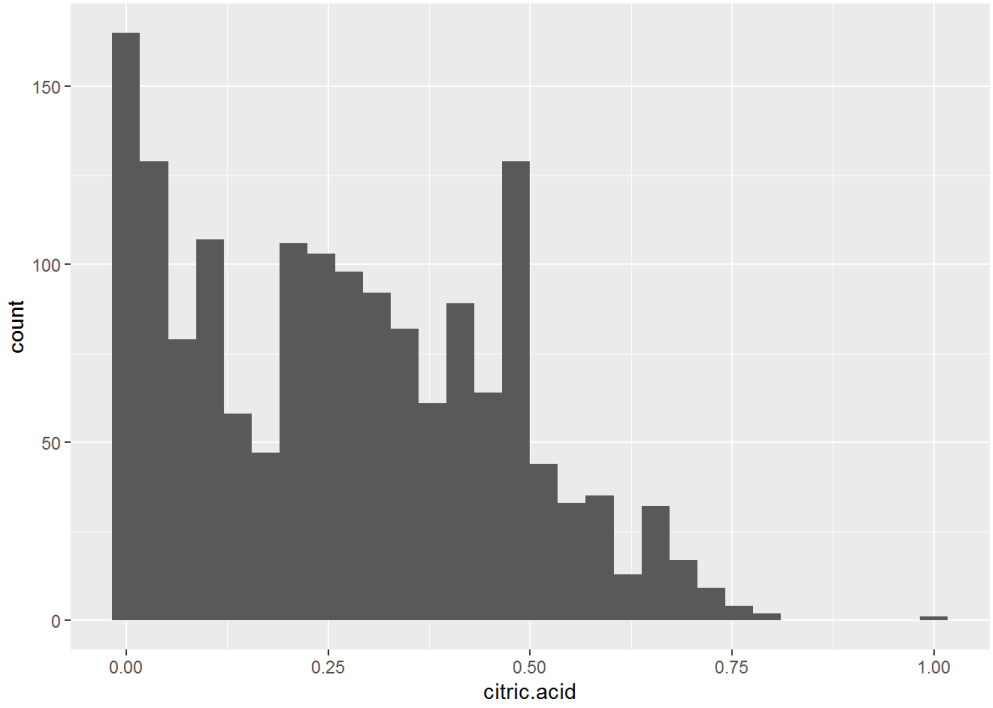
3. Histogram of volatile acidity

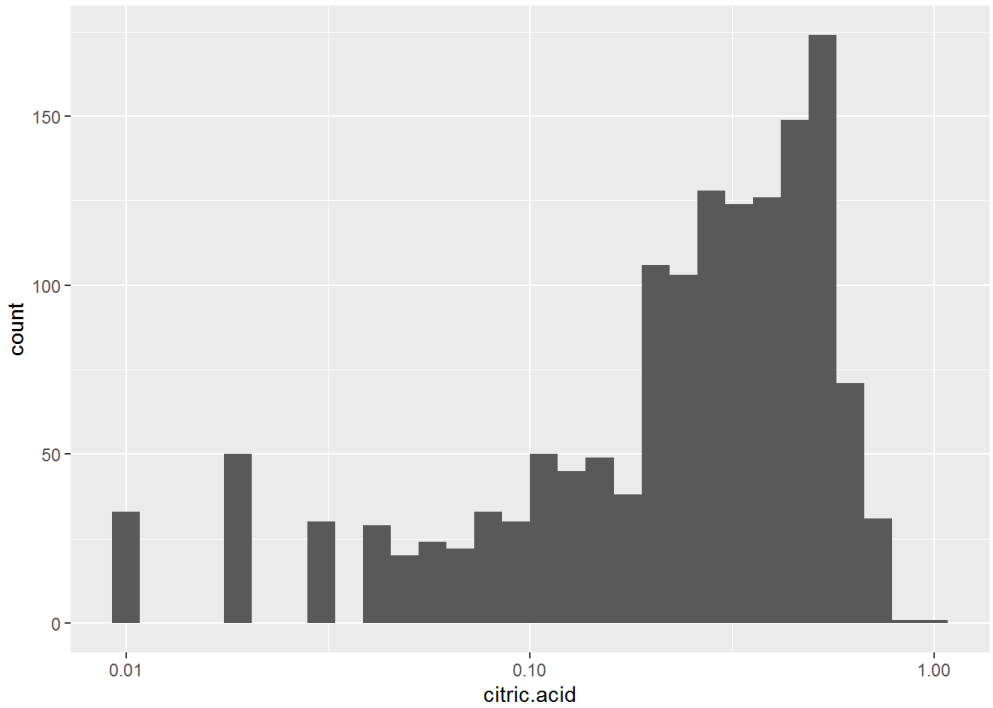



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3900  0.5200  0.5278  0.6400  1.5800
```

The distribution of volatile acidity is skewed to the right.

4. Histogram of citric acid

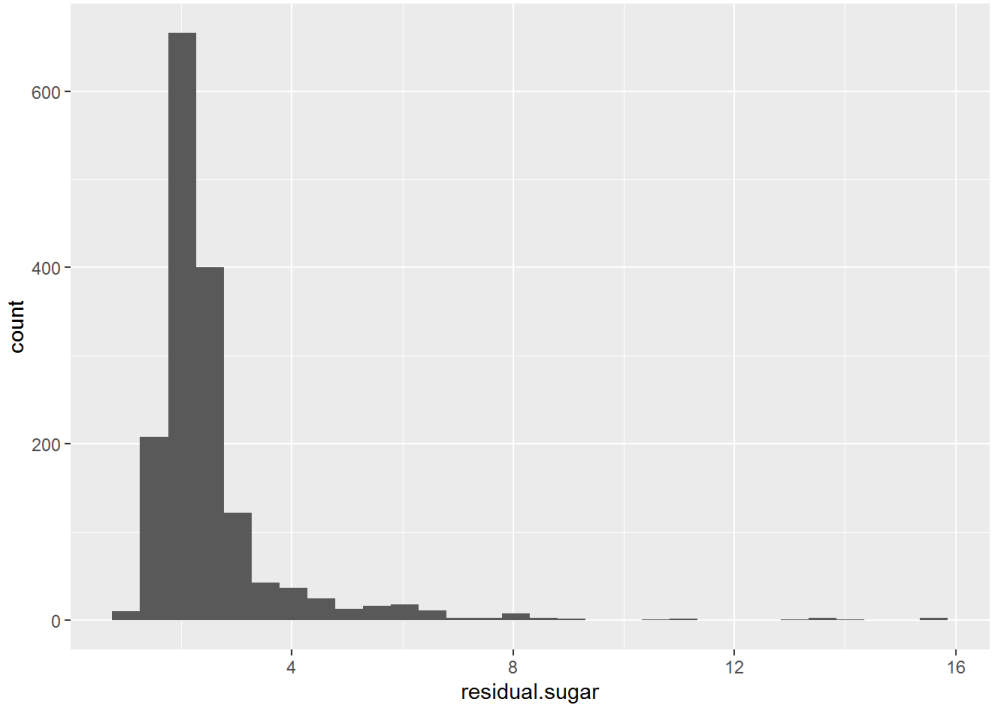


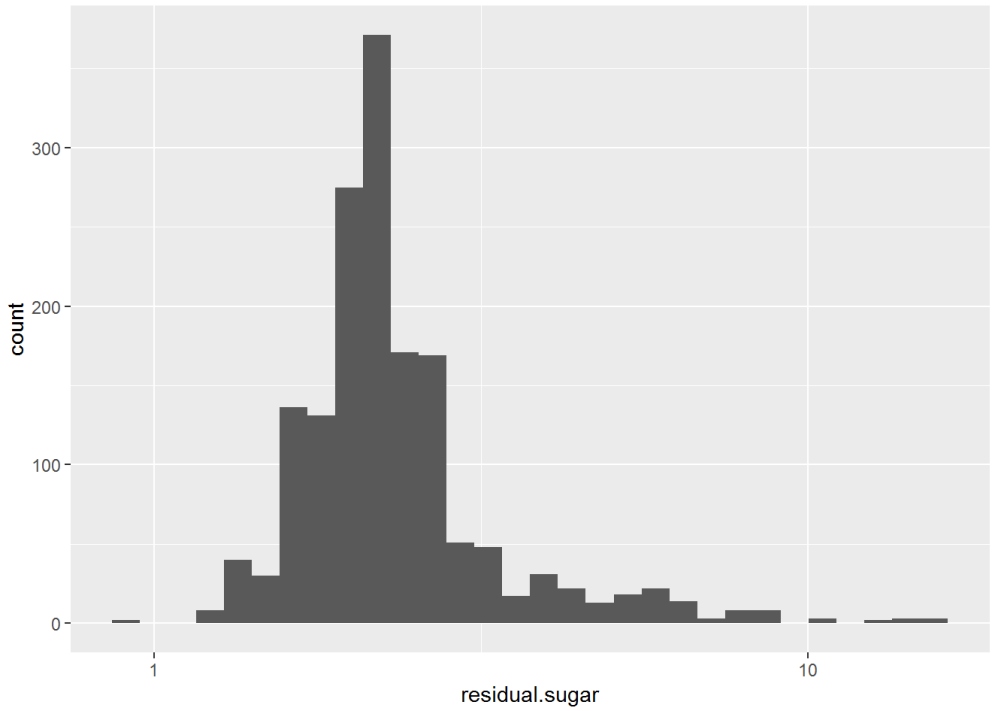



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.090   0.260   0.271   0.420   1.000
```

The distribution of citric acid is skewed to the right, and it seems to be multimodal. Thus, it is plotted on a log scale to explore any unseen features.

5. Histogram of residual sugar

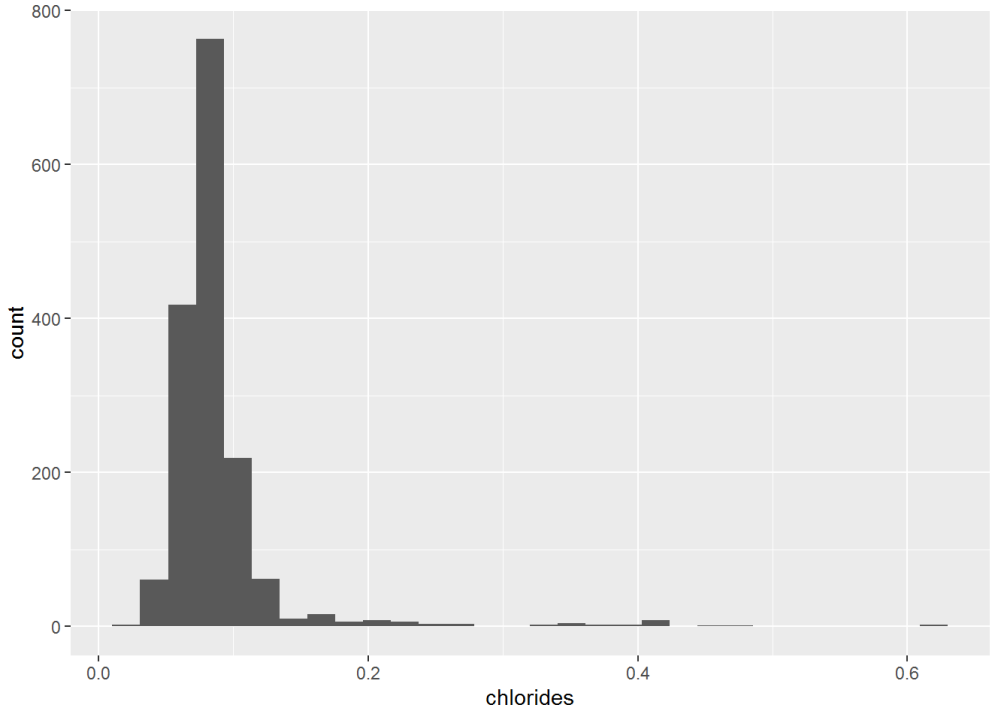


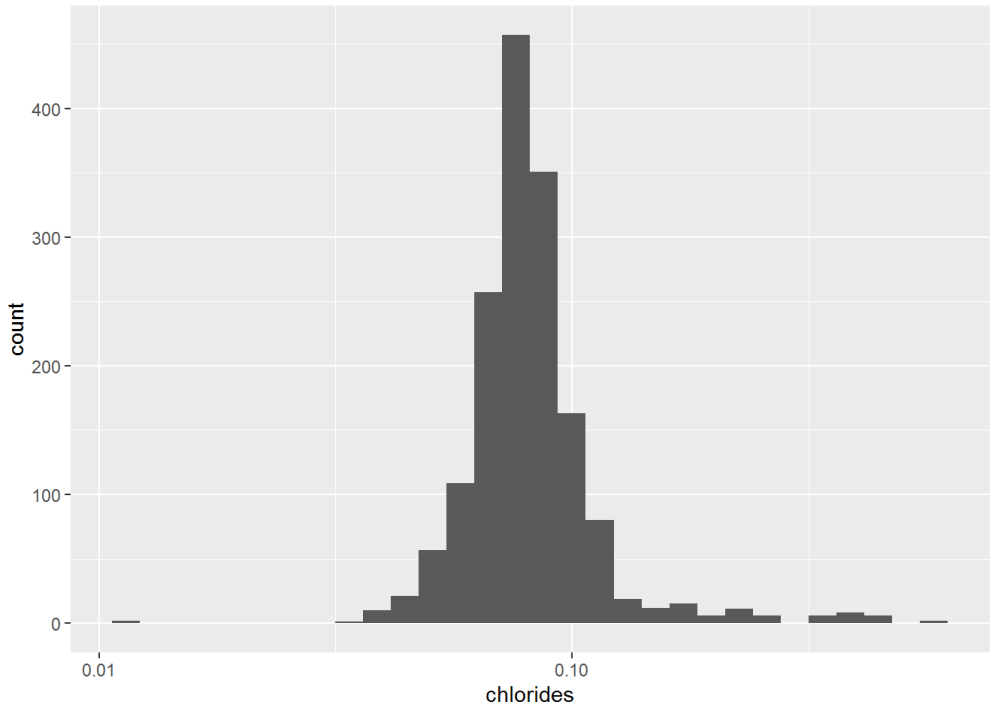



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.539   2.600  15.500
```

Since the distriution of residual sugar is significantly skewed to the right and concentrated at the left, it is plotted on a log scale to explore explore any unseen features.

6. Histogram of chlorides

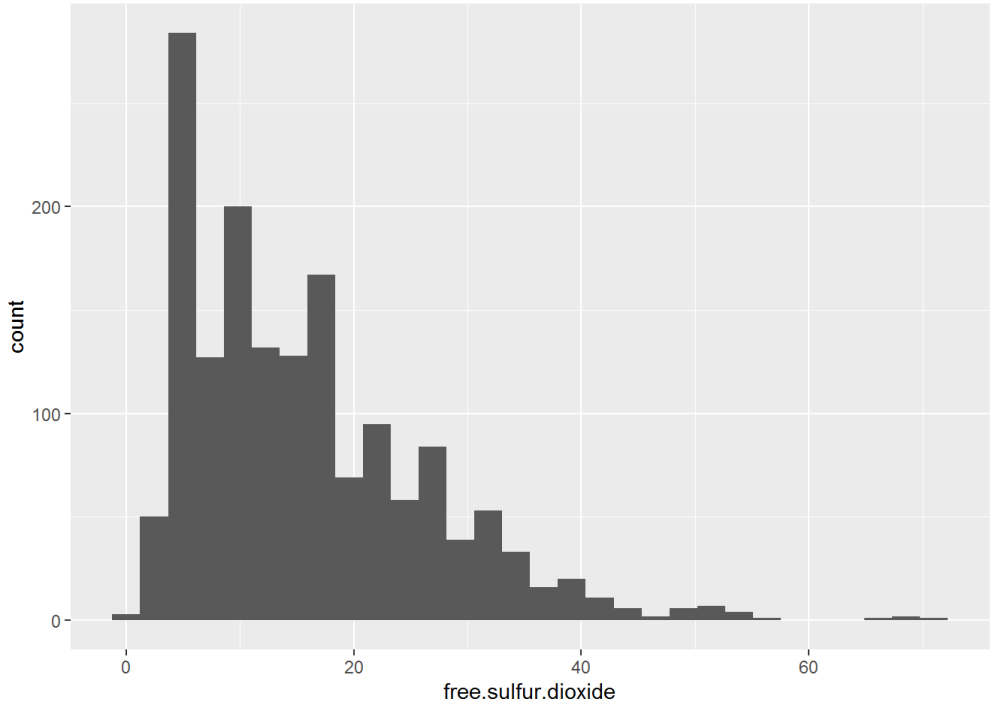





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.07000 0.07900 0.08747 0.09000 0.61100
```

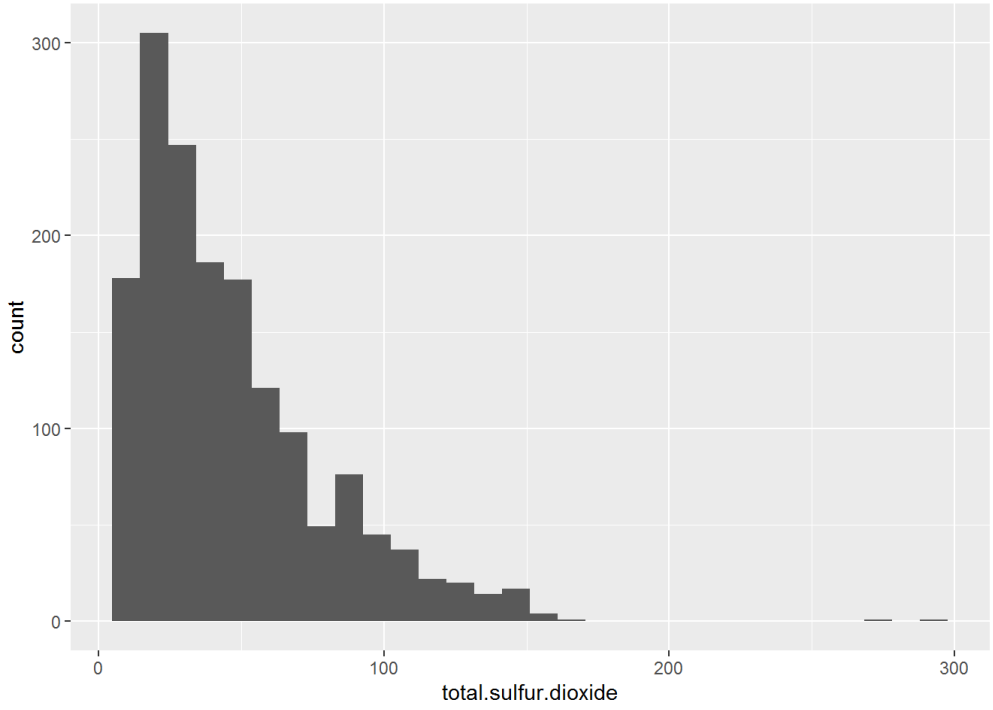
Again, since the distriution of chlorides is significantly skewed to the right, it is plotted on a log scale to explore explore any unseen features.

7. Histogram of free sulfur dioxide




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    7.00   14.00   15.87   21.00   72.00
```

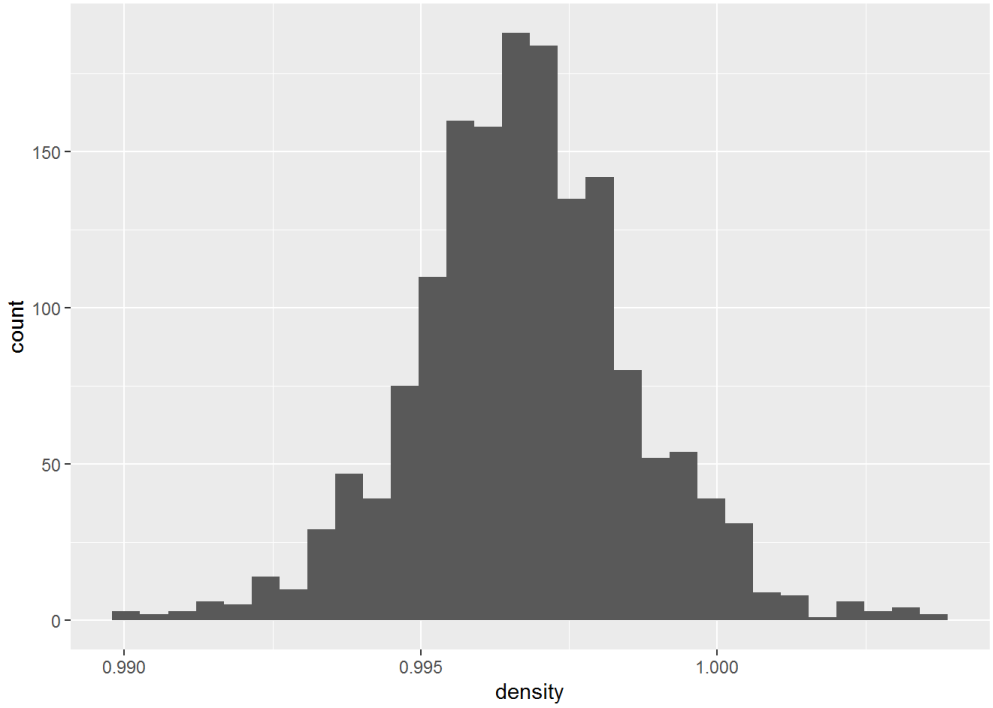
8. Histogram of total sulfur dioxide




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

Both the distributions of free and total sulfur dioxide are skewed to the right.

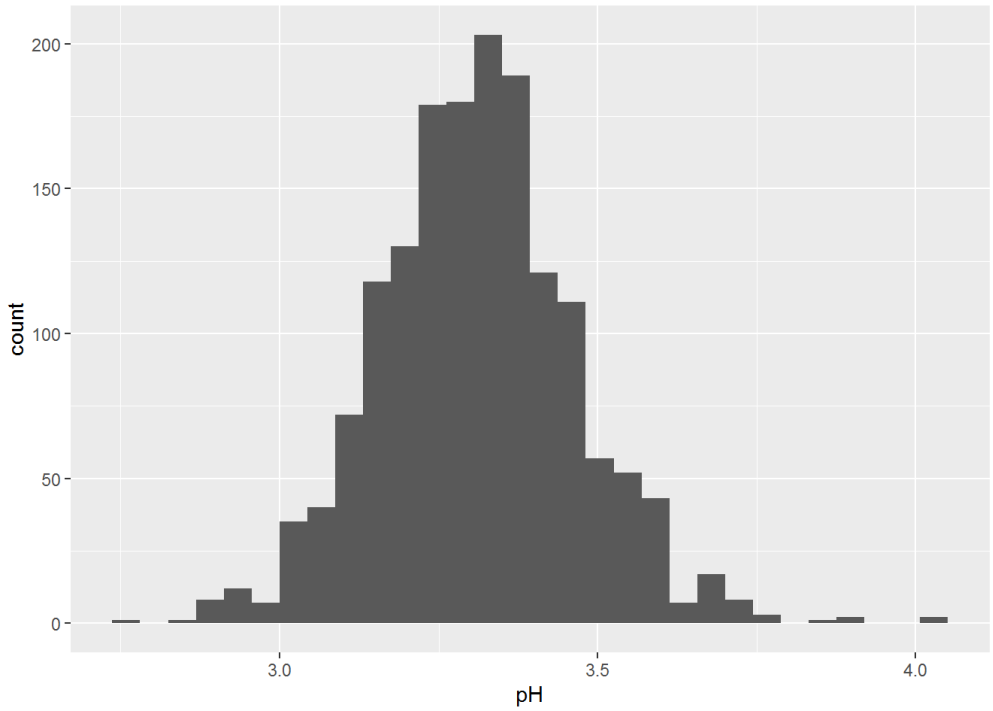
9. Histogram of density




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9956  0.9968  0.9967  0.9978  1.0040
```

While most of the other parameters are positively skewed, the distribution of density is quite symmetric.

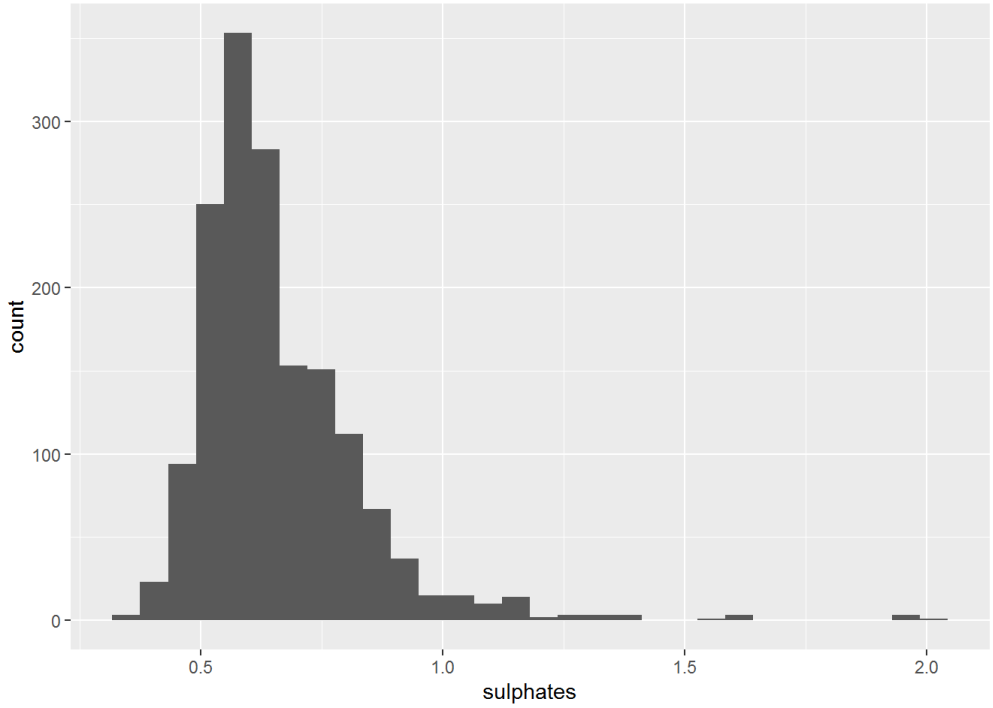
10. Histogram of pH value




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.210   3.310   3.311   3.400   4.010
```

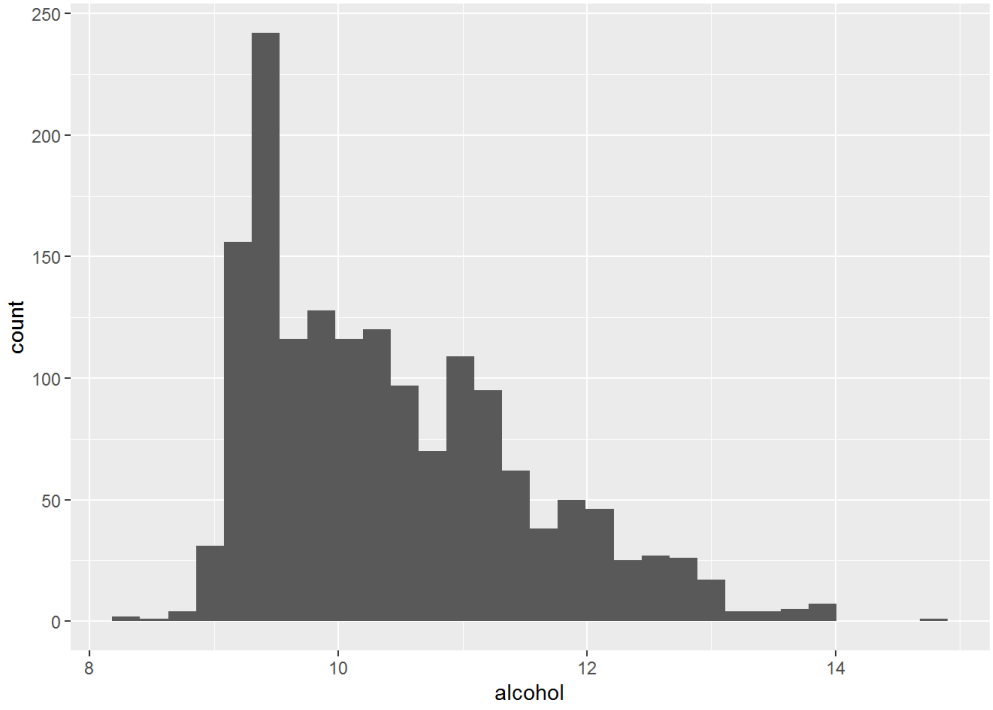
The distribution of pH value also is symmetric.

11. Histogram of sulphates




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.5500  0.6200  0.6581  0.7300  2.0000
```

12. Histogram of alcohol




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.50   10.20   10.42   11.10   14.90
```

Both the distribution of sulphates and alcohol are skewed to the right.

Univariate Analysis

What is the structure of your dataset?

This data set contains 1599 observations and 11 chemical attributes that could affect the wine quality. All of 11 chemical attributes are numerical variables.

Other observations:
1. The distributions of density and pH value are symmetic while most of the attributes are skewed to the right.
2. Most of the attributes are unimodal distributions, but Citric acid's has prominent peaks.

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

The main features of interest after univariate analysis are density and pH value, since they resemble the distribution of the quality.

What other features in the dataset do you think will help support your

investigation into your feature(s) of interest? Both the distributions of density and pH value are unimodal.

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

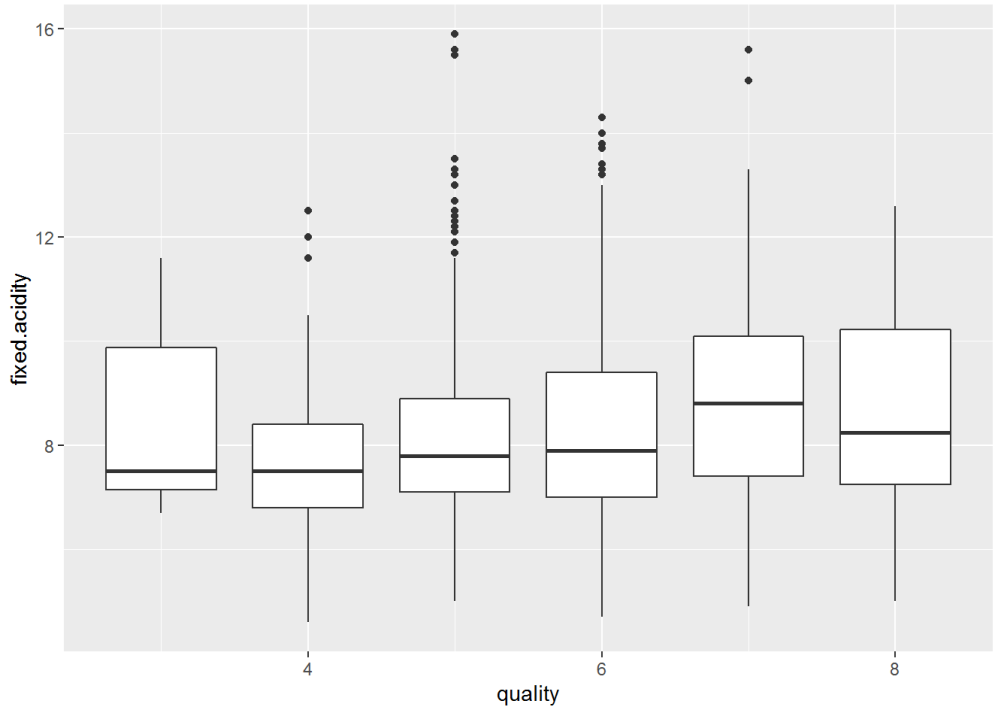
No.

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?
Some of the attributes (e.g. residual sugar, chlorides) have narrow distributions and are positively skewed. Thus, these qualities are plotted on log scale to explore any hidden features.

Bivariate Plots Section

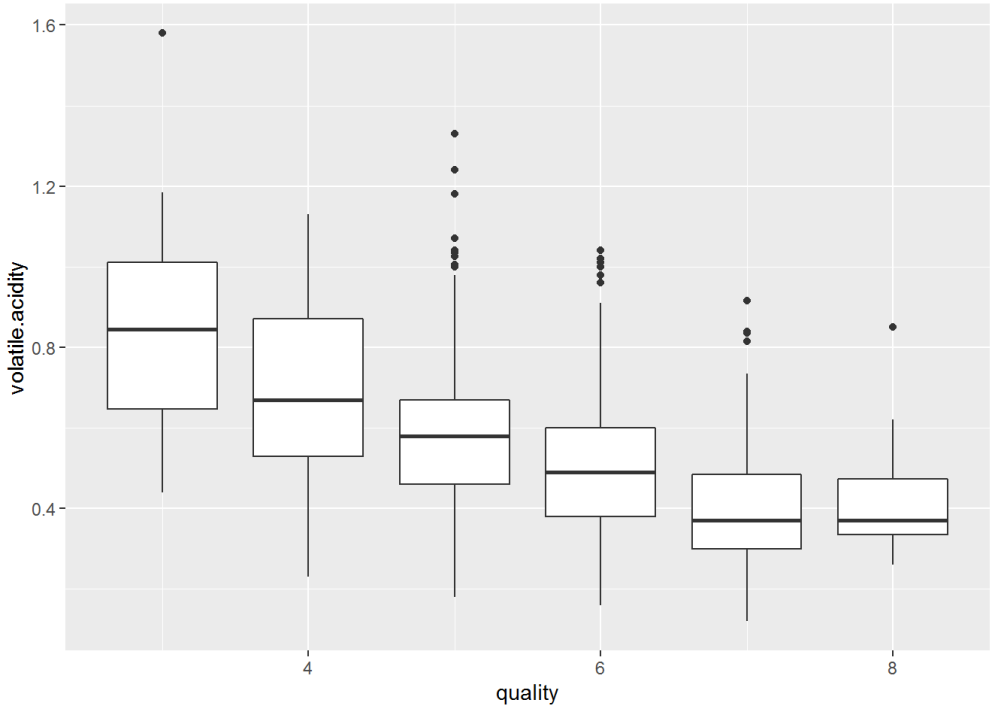
1. Boxplot: fixed.acidity vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.700   7.150   7.500   8.360   9.875  11.600
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.600   6.800   7.500   7.779   8.400  12.500
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   7.100   7.800   8.167   8.900  15.900
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.700   7.000   7.900   8.347   9.400  14.300
##
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.900   7.400   8.800   8.872  10.100  15.600
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.000   7.250   8.250   8.567  10.230  12.600
```

No strong correlation is found between fixed acidity and quality.

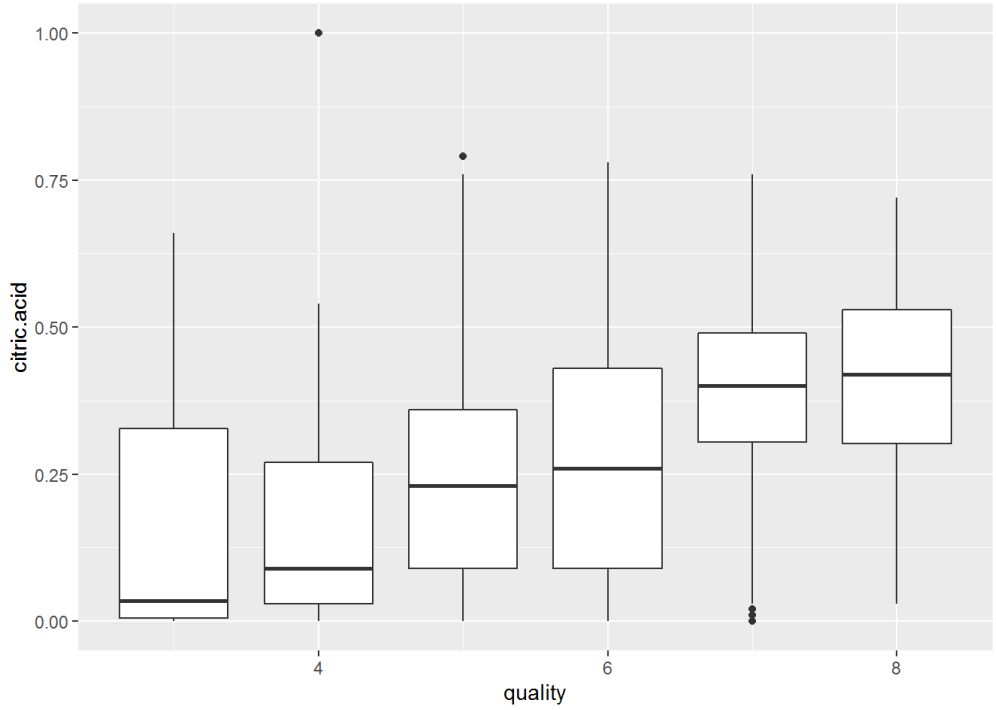
2. Boxplot: volatile.acidity vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4400  0.6475  0.8450  0.8845  1.0100  1.5800
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.230   0.530   0.670   0.694   0.870   1.130
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.180   0.460   0.580   0.577   0.670   1.330
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1600  0.3800  0.4900  0.4975  0.6000  1.0400
##
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1200  0.3000  0.3700  0.4039  0.4850  0.9150
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2600  0.3350  0.3700  0.4233  0.4725  0.8500
```

Volatile acidity and quality are negatively correlated.

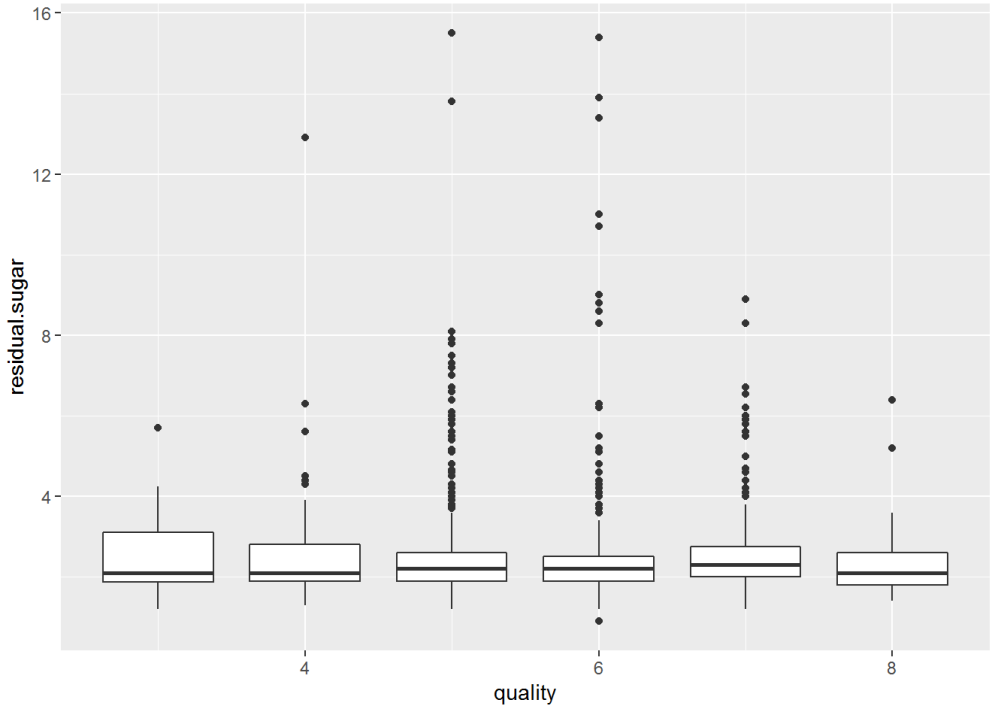
3. Boxplot: citric.acid vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0050  0.0350  0.1710  0.3275  0.6600
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0300  0.0900  0.1742  0.2700  1.0000
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2300  0.2437  0.3600  0.7900
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0900  0.2600  0.2738  0.4300  0.7800
##
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3050  0.4000  0.3752  0.4900  0.7600
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0300  0.3025  0.4200  0.3911  0.5300  0.7200
```

Citric acid and quality are positively correlated.

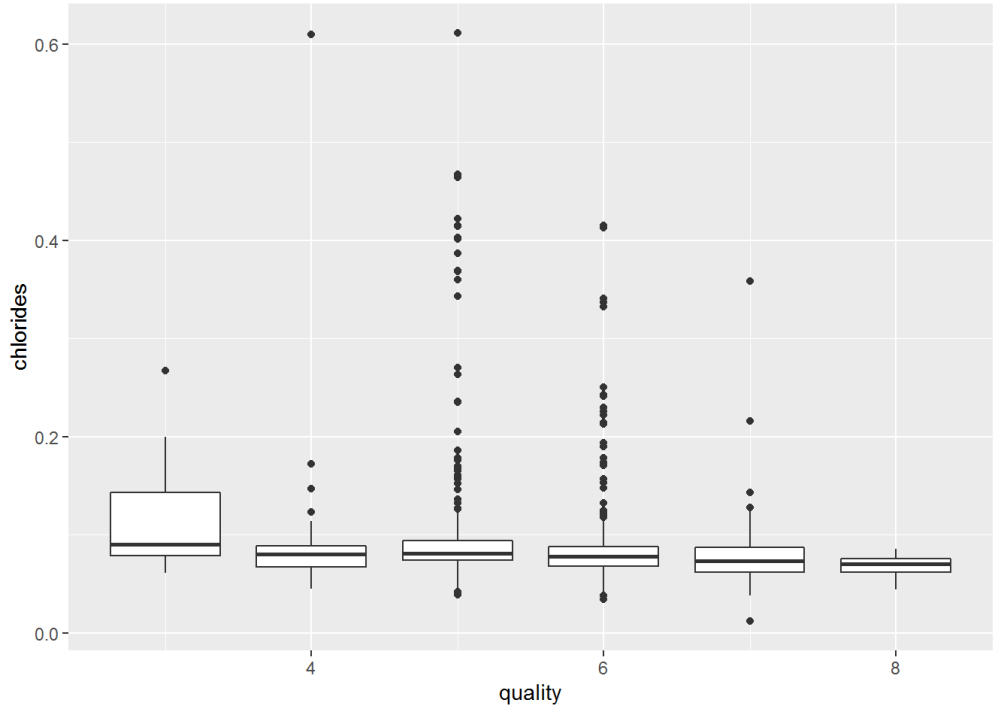
4. Boxplot: residual.sugar vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   1.875   2.100   2.635   3.100   5.700
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.300   1.900   2.100   2.694   2.800  12.900
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   1.900   2.200   2.529   2.600  15.500
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   1.900   2.200   2.477   2.500  15.400
##
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.200   2.000   2.300   2.721   2.750   8.900
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.400   1.800   2.100   2.578   2.600   6.400
```

No strong correlation is found between residual sugar and quality.

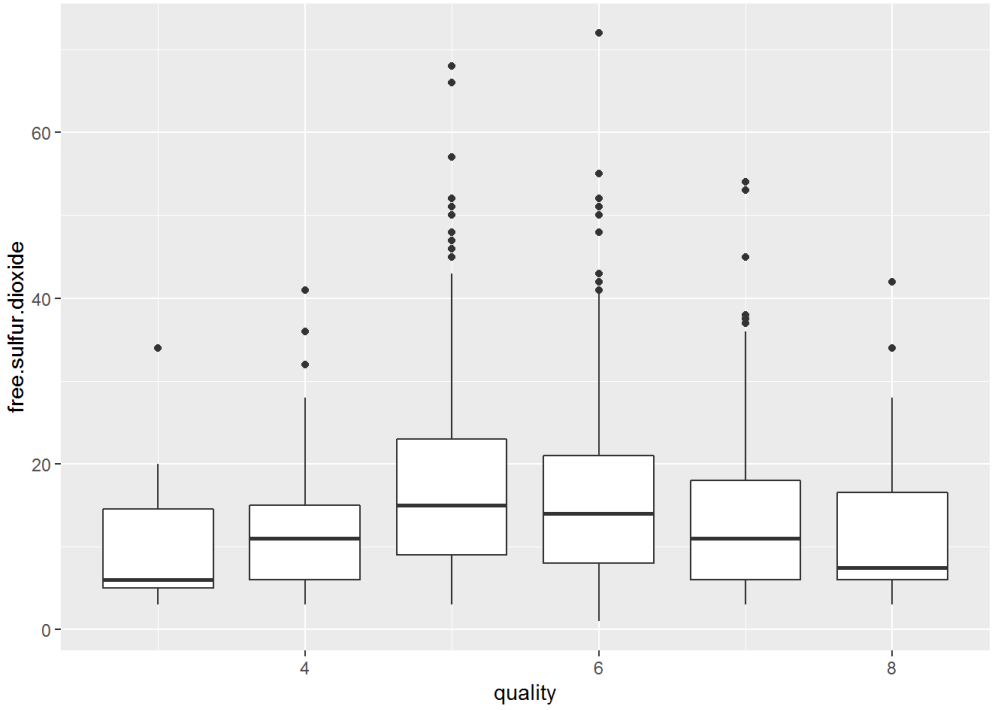
5. Boxplot: chlorides vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0610  0.0790  0.0905  0.1225  0.1430  0.2670
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04500 0.06700 0.08000 0.09068 0.08900 0.61000
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03900 0.07400 0.08100 0.09274 0.09400 0.61100
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03400 0.06825 0.07800 0.08496 0.08800 0.41500
##
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01200 0.06200 0.07300 0.07659 0.08700 0.35800
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04400 0.06200 0.07050 0.06844 0.07550 0.08600
```

No strong correlation is found between residual sugar and quality.

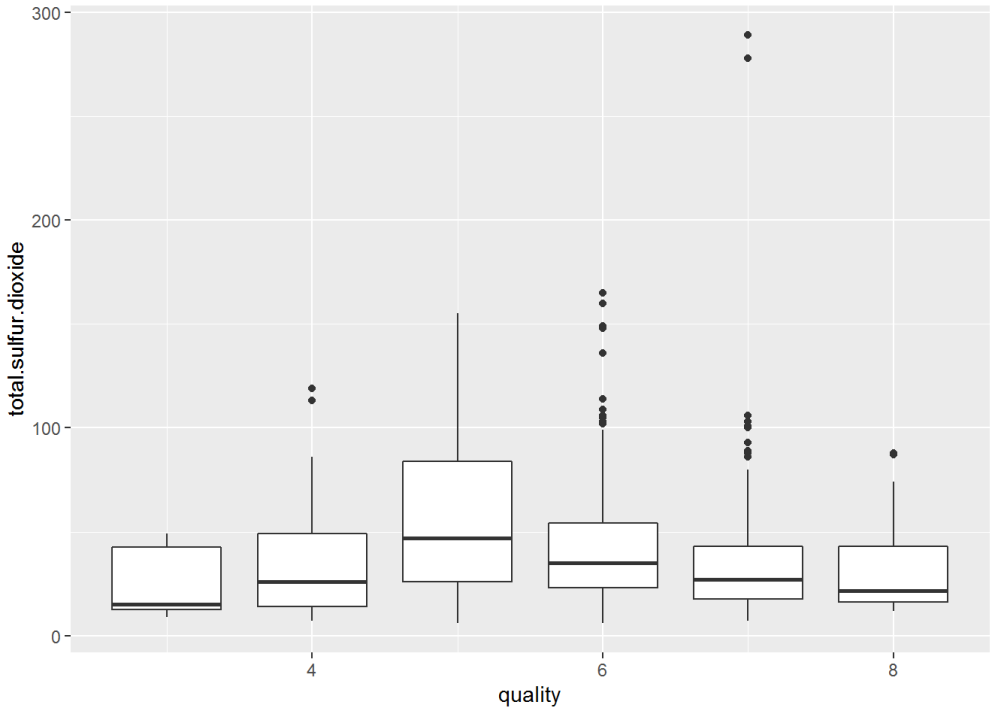
6. Boxplot: free.sulfur.dioxide vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0     5.0     6.0    11.0    14.5    34.0
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    6.00   11.00   12.26   15.00   41.00
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    9.00   15.00   16.98   23.00   68.00
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    8.00   14.00   15.71   21.00   72.00
##
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    6.00   11.00   14.05   18.00   54.00
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    6.00    7.50   13.28   16.50   42.00
```

No strong correlation is found between free sulfur dioxide and quality.

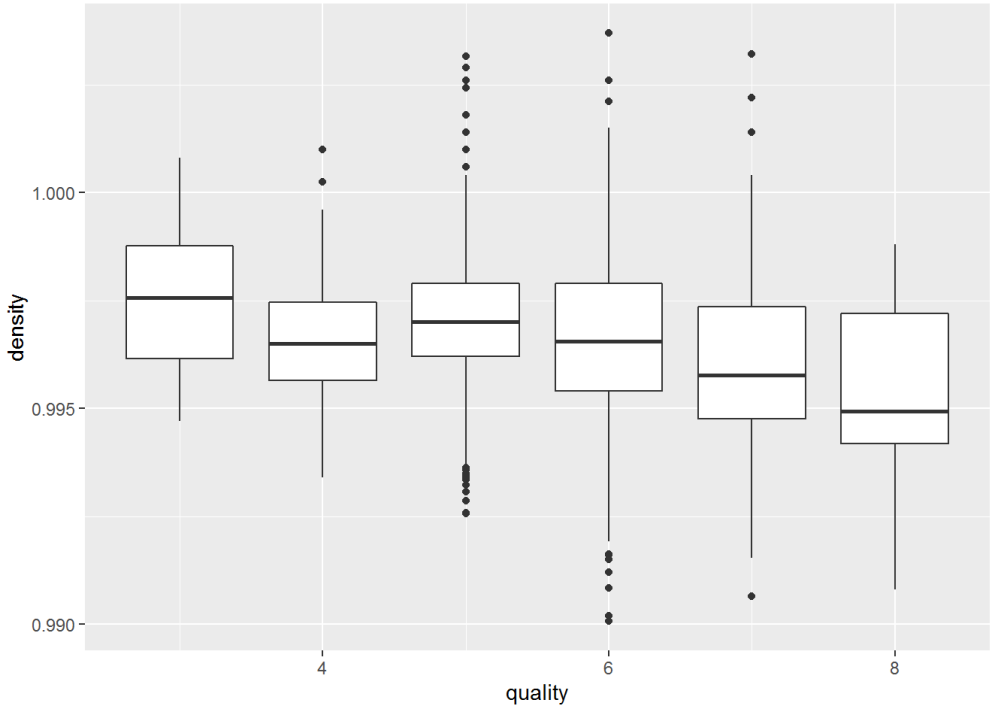
7. Boxplot: total.sulfur.dioxide vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0    12.5    15.0    24.9    42.5    49.0
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   14.00   26.00   36.25   49.00  119.00
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   26.00   47.00   56.51   84.00  155.00
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   23.00   35.00   40.87   54.00  165.00
##
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   17.50   27.00   35.02   43.00  289.00
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   16.00   21.50   33.44   43.00   88.00
```

No strong correlation is found between total sulfur dioxide and quality.

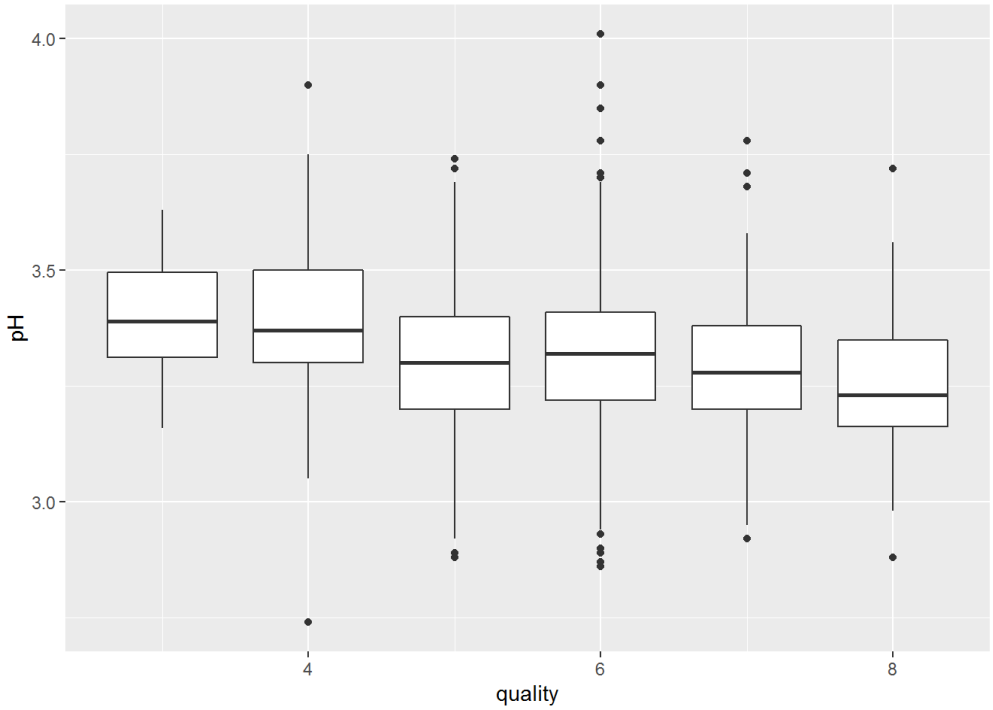
8. Boxplot: density vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9947  0.9962  0.9976  0.9975  0.9988  1.0010
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9934  0.9956  0.9965  0.9965  0.9974  1.0010
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9926  0.9962  0.9970  0.9971  0.9979  1.0030
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9901  0.9954  0.9966  0.9966  0.9979  1.0040
##
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9906  0.9948  0.9958  0.9961  0.9974  1.0030
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9908  0.9942  0.9949  0.9952  0.9972  0.9988
```

A week and negative correlation is found between density and quality. Despite both of the distributions of density and quality are symmetric, it does not necessarily indicate they are highly correlated.

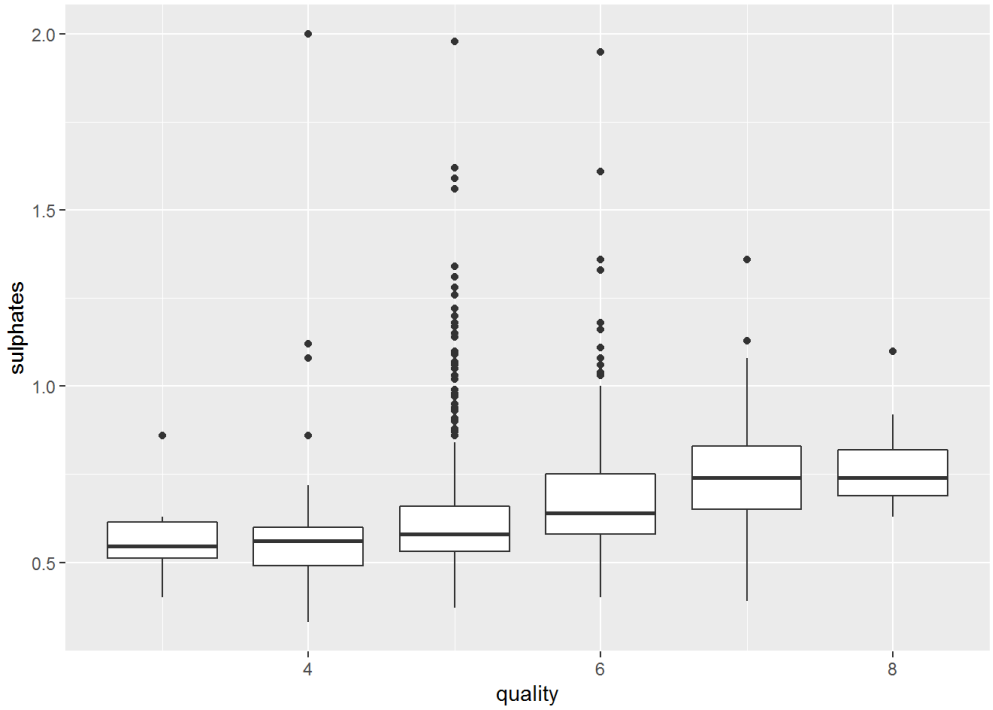
9. Boxplot: pH vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.160   3.312   3.390   3.398   3.495   3.630
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.740   3.300   3.370   3.382   3.500   3.900
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.880   3.200   3.300   3.305   3.400   3.740
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.860   3.220   3.320   3.318   3.410   4.010
##
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.920   3.200   3.280   3.291   3.380   3.780
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.880   3.162   3.230   3.267   3.350   3.720
```

Similiar to the previous plot, a week and negative correlation is found between density and quality. Their resemblance does not lead to a strong correlation.

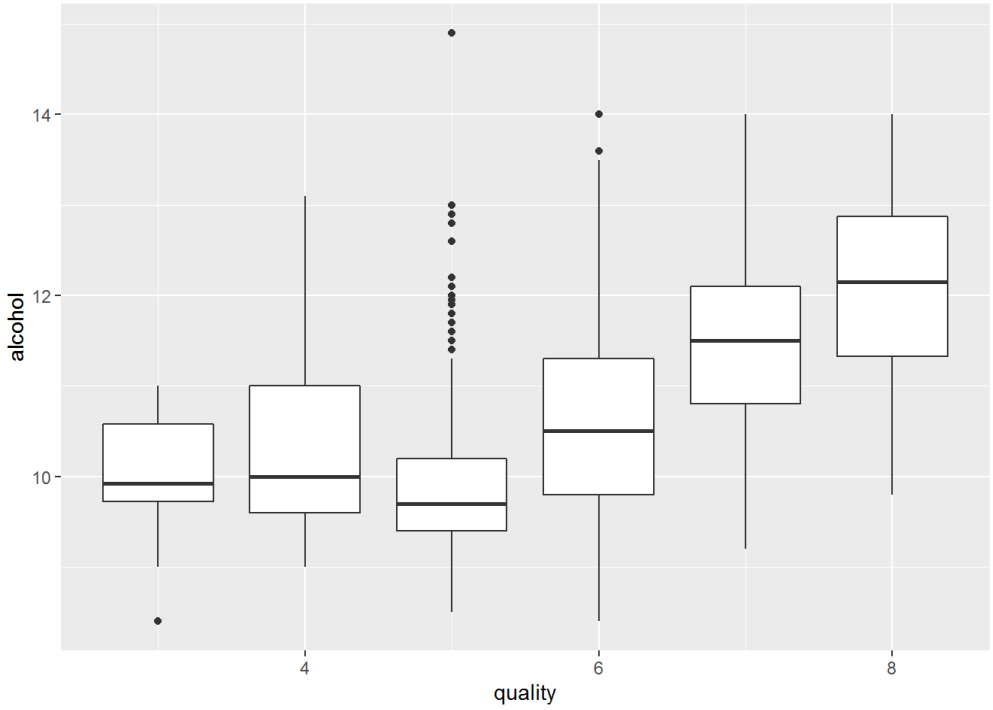
10. Boxplot: sulphates vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5125  0.5450  0.5700  0.6150  0.8600
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3300  0.4900  0.5600  0.5964  0.6000  2.0000
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.370   0.530   0.580   0.621   0.660   1.980
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4000  0.5800  0.6400  0.6753  0.7500  1.9500
##
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3900  0.6500  0.7400  0.7413  0.8300  1.3600
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.6300  0.6900  0.7400  0.7678  0.8200  1.1000
```

The correlation between Sulphates and quality are positive, yet weak.

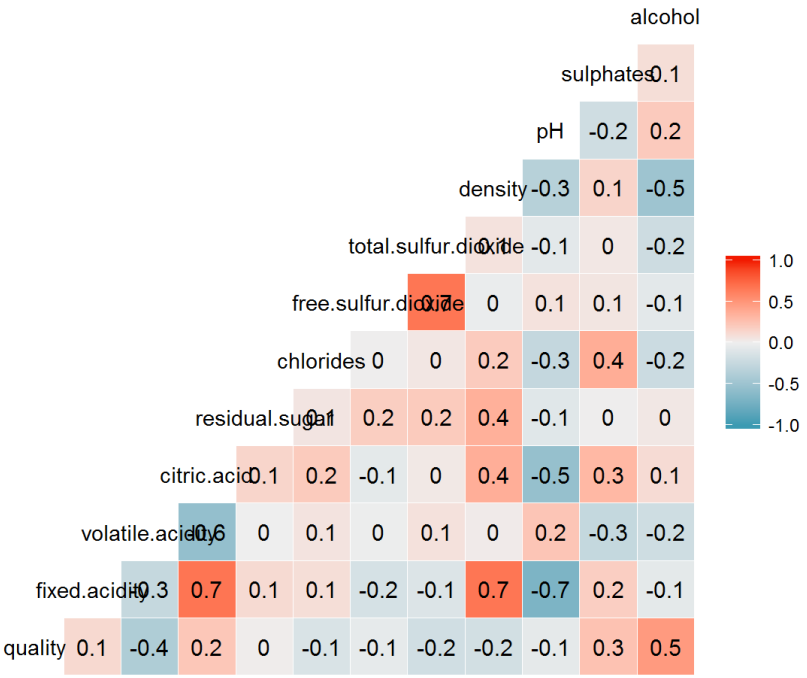
11. Boxplot: alcohol vs quality



```
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.400   9.725   9.925   9.955  10.580  11.000
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00    9.60   10.00   10.27   11.00   13.10
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     8.5     9.4     9.7     9.9    10.2    14.9
##
## $`6`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.40    9.80   10.50   10.63   11.30   14.00
##
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.20   10.80   11.50   11.47   12.10   14.00
##
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.80   11.32   12.15   12.09   12.88   14.00
```

The correlation between alcohol and quality are positive and strong.

12. The correlation plot between chemical attributes



The correlation between chemical attributes are calculated to provide a quantitative insight on their relationship. Both this chart and the above boxplots draw the same conclusion: volatile acidity and alcohol show stronger correlation with the wine quality. While the correlation between sulphate and wine quality is 0.3, many outliers were found in the boxplot.

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?

1. When we explored the dataset by univariate plots, density and pH value are the only two attributes show symmetric distributions. At that time, it was suspected that they may have a stronger correlation with wine quality.

2. However, as discussed briefly in the section of bavariate plots, volatile acidity and alcohol, rather than density and pH value show stronger correlation with the wine quality.
3. Thus, the features of interest now are the votalie acidity and alcohol.

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

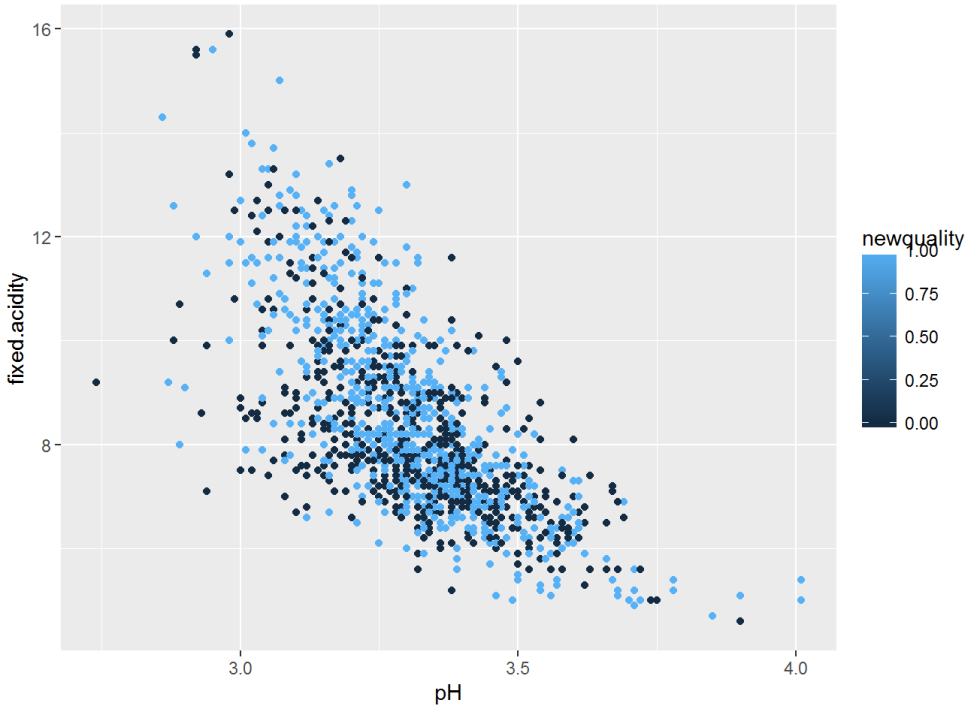
3 chemical attributes show strong correlation with fixed acidity (which is stronger than those with wine quality) and they are pH value, density and citric acid.

What was the strongest relationship you found?

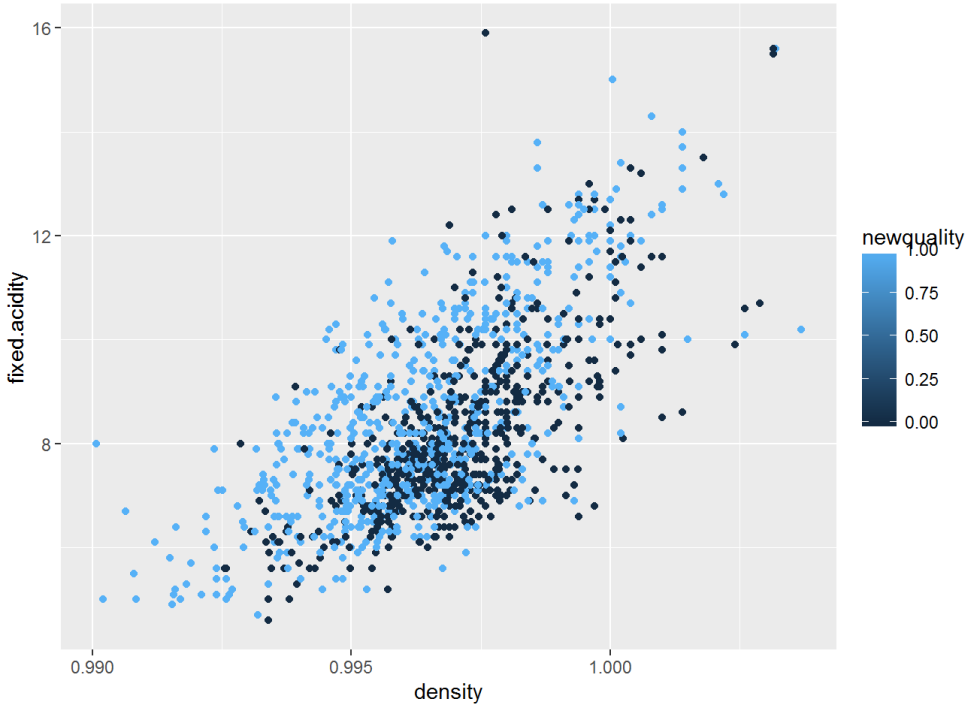
Among the chemical attribute: (1) pH value and fixed acidity, (2) density and fixed acidity, (3) citric acid and fixed acidity
Between the wine quality and a sole chemical attribute: alcohol and wine quality

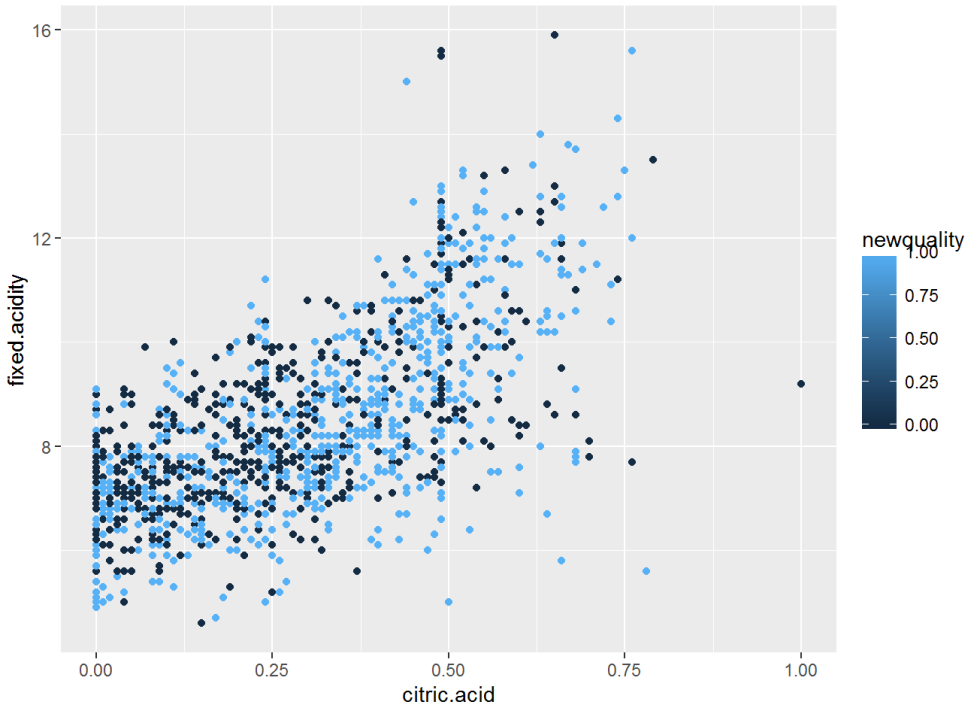
Multivariate Plots Section

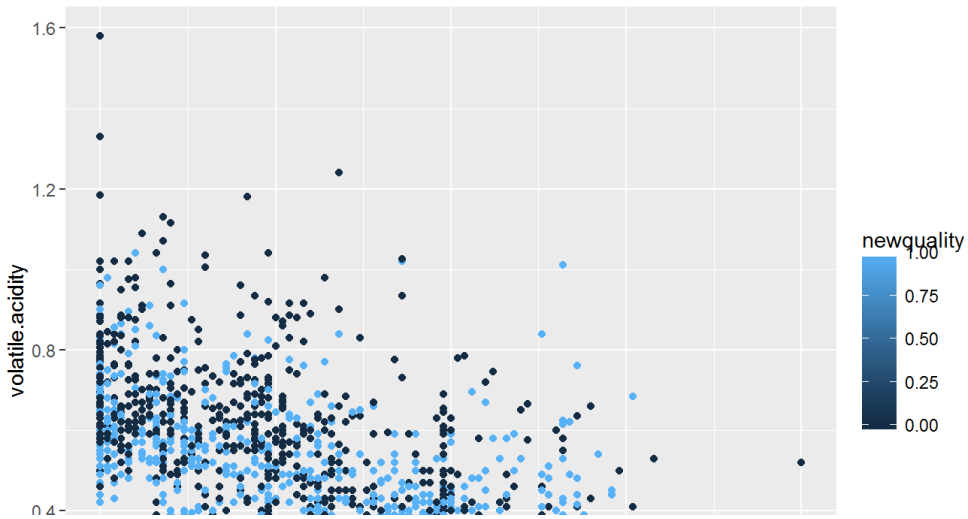
> **Tip**: Now it's time to put everything together. Based on what you found in the bivariate plots section, create a few multivariate plots to investigate more complex interactions between variables. Make sure that the plots that you create here are justified by the plots you explored in the previous section. If you plan on creating any mathematical models, this is the section where you will do that.

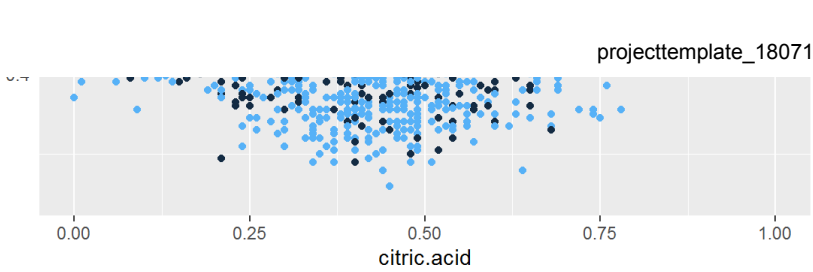


[Not sure which variables to choose to do the multivariate plots] In this plot, I create a new variable to indicate the wine quality is high or low. If wine quality is larger than 5, 1 is assigned, indicating high quliaty. If it is smaller than 5, 0 is assigned, indicating low quality. Then fixed acidity is plotted against pH value for both high and low quliaty. As shown in the plot, wine quality is in effect independent of the correlation between fixed acidity and pH value.


fixed.acidity
16
12
8
0.990
0.995
1.000
density
newquality
1.00
0.75
0.50
0.25
0.00


fixed.acidity
16
12
8
0.00
0.25
0.50
0.75
1.00
citric.acid
newquality
1.00
0.75
0.50
0.25
0.00


volatile.acidity
1.6
1.2
0.8
0.4
newquality
1.00
0.75
0.50
0.25
0.00



Now I employ stepwise model selection to determine which chemical attributes have an important role on the wine quality.

```
## Start:  AIC=-1375.49
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                       Df Sum of Sq    RSS     AIC
## - density              1     0.287 666.70 -1376.8
## - fixed.acidity        1     0.389 666.80 -1376.5
## - residual.sugar       1     0.498 666.91 -1376.3
## - citric.acid          1     0.646 667.06 -1375.9
## <none>                             666.41 -1375.5
## - free.sulfur.dioxide  1     1.694 668.10 -1373.4
## - pH                   1     1.957 668.37 -1372.8
## - chlorides            1     8.391 674.80 -1357.5
## - total.sulfur.dioxide 1     8.427 674.84 -1357.4
## - sulphates            1    26.971 693.38 -1314.0
## - volatile.acidity     1    33.620 700.03 -1298.8
## - alcohol              1    45.672 712.08 -1271.5
##
## Step:  AIC=-1376.8
## quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol
##
##                       Df Sum of Sq    RSS     AIC
## - fixed.acidity        1     0.108 666.81 -1378.5
## - residual.sugar       1     0.231 666.93 -1378.2
## - citric.acid          1     0.654 667.35 -1377.2
## <none>                             666.70 -1376.8
## + density              1     0.287 666.41 -1375.5
## - free.sulfur.dioxide  1     1.829 668.53 -1374.4
## - pH                   1     4.325 671.02 -1368.5
## - total.sulfur.dioxide 1     8.728 675.43 -1358.0
## - chlorides            1     8.761 675.46 -1357.9
## - sulphates            1    27.287 693.98 -1314.7
## - volatile.acidity     1    35.000 701.70 -1297.0
## - alcohol              1   119.669 786.37 -1114.8
##
## Step:  AIC=-1378.54
## quality ~ volatile.acidity + citric.acid + residual.sugar + chlorides +
##     free.sulfur.dioxide + total.sulfur.dioxide + pH + sulphates +
##     alcohol
##
##                       Df Sum of Sq    RSS     AIC
## - residual.sugar       1     0.257 667.06 -1379.9
## - citric.acid          1     0.565 667.37 -1379.2
## <none>                             666.81 -1378.5
## + fixed.acidity        1     0.108 666.70 -1376.8
## + density              1     0.005 666.80 -1376.5
## - free.sulfur.dioxide  1     1.901 668.71 -1376.0
## - pH                   1     7.065 673.87 -1363.7
## - chlorides            1     9.940 676.75 -1356.9
## - total.sulfur.dioxide 1    10.031 676.84 -1356.7
## - sulphates            1    27.673 694.48 -1315.5
## - volatile.acidity     1    36.234 703.04 -1295.9
## - alcohol              1   120.633 787.44 -1114.7
##
## Step:  AIC=-1379.93
## quality ~ volatile.acidity + citric.acid + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol
##
##                       Df Sum of Sq    RSS     AIC
## - citric.acid          1     0.475 667.54 -1380.8
## <none>                             667.06 -1379.9
## + residual.sugar       1     0.257 666.81 -1378.5
## + fixed.acidity        1     0.133 666.93 -1378.2
## + density              1     0.028 667.03 -1378.0
## - free.sulfur.dioxide  1     2.064 669.13 -1377.0
## - pH                   1     7.138 674.20 -1364.9
## - total.sulfur.dioxide 1     9.828 676.89 -1358.5
## - chlorides            1     9.832 676.89 -1358.5
```

```
## - sulphates             1    27.446 694.51 -1317.5
## - volatile.acidity      1    35.977 703.04 -1297.9
## - alcohol               1   122.667 789.73 -1112.0
##
## Step:  AIC=-1380.79
## quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + pH + sulphates + alcohol
##
##                        Df Sum of Sq    RSS     AIC
## <none>                              667.54 -1380.8
## + citric.acid           1     0.475 667.06 -1379.9
## + residual.sugar        1     0.167 667.37 -1379.2
## + density               1     0.031 667.51 -1378.9
## + fixed.acidity         1     0.007 667.53 -1378.8
## - free.sulfur.dioxide   1     2.394 669.93 -1377.1
## - pH                    1     7.073 674.61 -1365.9
## - total.sulfur.dioxide  1    10.787 678.32 -1357.2
## - chlorides             1    10.809 678.35 -1357.1
## - sulphates             1    27.060 694.60 -1319.2
## - volatile.acidity      1    42.318 709.85 -1284.5
## - alcohol               1   124.483 792.02 -1109.4
```

```
##               Step Df  Deviance Resid. Df Resid. Dev       AIC
## 1                   NA        NA      1587   666.4107 -1375.489
## 2        - density  1 0.2868924      1588   666.6976 -1376.801
## 3  - fixed.acidity  1 0.1079824      1589   666.8056 -1378.542
## 4 - residual.sugar  1 0.2566805      1590   667.0623 -1379.926
## 5    - citric.acid  1 0.4748034      1591   667.5371 -1380.789
```

According the result from the model selection, the amount of (1) volatile acidity, (2) chlorides, (3) free.sulfur.dioxide, (4) total.sulfur.dioxide, (5) pH, (6) sulphates, and (7) alcohol would have an affect on the wine quality.

In the prior bivariate analysis, alcohol is the only attribute that shows strong correlation with the wine quality. At first, it seems to be contray to the model selection result. As I look into the model selection as shown below:
Df Sum of Sq RSS AIC - citric.acid 1 0.475 667.54 -1380.8 667.06 -1379.9 + residual.sugar 1 0.257 666.81 -1378.5 + fixed.acidity 1 0.133 666.93 -1378.2 + density 1 0.028 667.03 -1378.0 - free.sulfur.dioxide 1 2.064 669.13 -1377.0 - pH 1 7.138 674.20 -1364.9 - total.sulfur.dioxide 1 9.828 676.89 -1358.5 - chlorides 1 9.832 676.89 -1358.5 - sulphates 1 27.446 694.51 -1317.5 - volatile.acidity 1 35.977 703.04 -1297.9 - alcohol 1 122.667 789.73 -1112.0 the removal of alcohol would cause huge decrease in AIC value, while the removal of others affect AIC less significantly. In fact, this list shows roughly the importance of that attribute in determing the wine quality in an ascending order (that is from less important to more important)

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?

Were there any interesting or surprising interactions between features?

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

Final Plots and Summary

> **Tip**: You've done a lot of exploration and have built up an understanding of the structure of and relationships between the variables in your dataset. Here, you will select three plots from all of your previous exploration to present here as a summary of some of your most interesting findings. Make sure that you have refined your selected plots for good titling, axis labels (with units), and good aesthetic choices (e.g. color, transparency). After each plot, make sure you justify why you chose each plot by describing what it shows.

Plot One

Description One

Plot Two

Description Two

Plot Three

Description Three

Reflection

> **Tip**: Here's the final step! Reflect on the exploration you performed and the insights you found. What were some of the struggles that you went through? What went well? What was surprising? Make sure you include an insight into future work that could be done with the dataset.

> **Tip**: Don't forget to remove this, and the other **Tip** sections before saving your final work and knitting the final report!