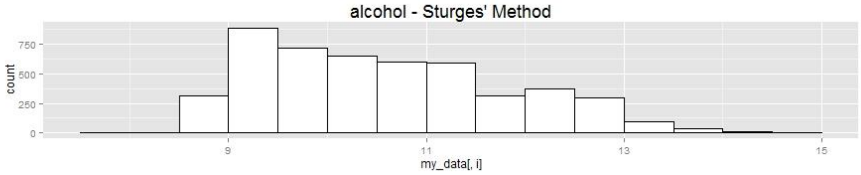
Data understanding and visualization

Janne Flinck

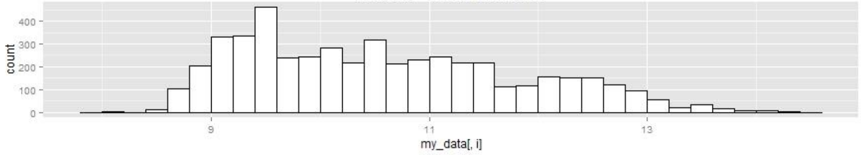
79428

Histograms of the attributes

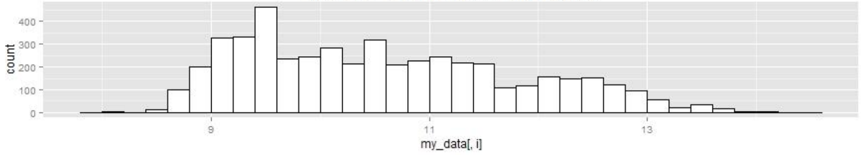
alcohol - Sturges' Method



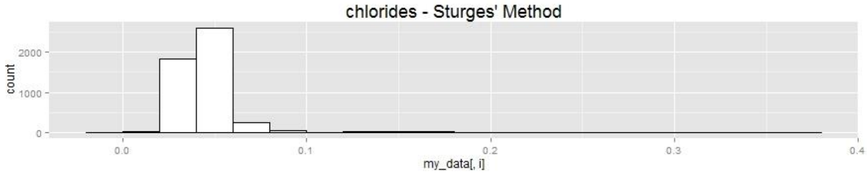
alcohol - Scott's Method



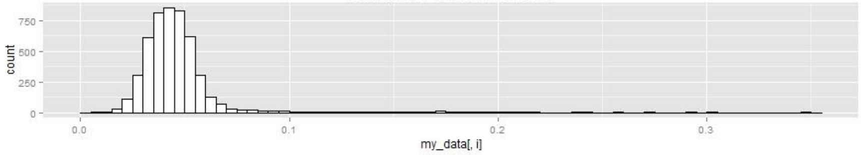
alcohol - Freedman-Diaconis' Method



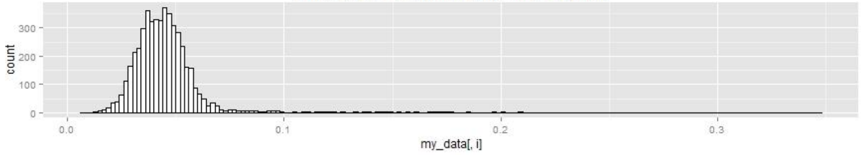
chlorides - Sturges' Method



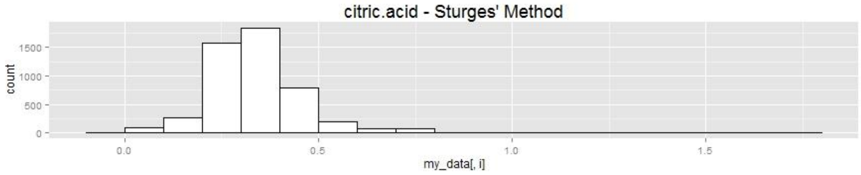
chlorides - Scott's Method



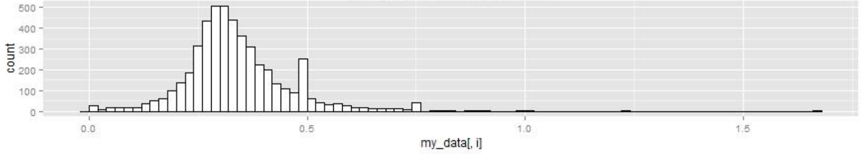
chlorides - Freedman-Diaconis' Method



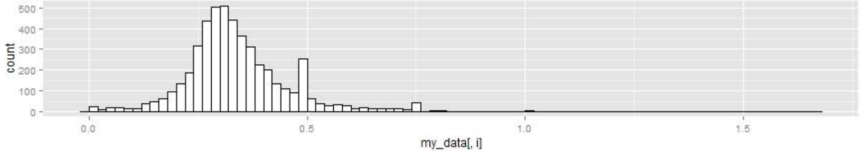
citric.acid - Sturges' Method



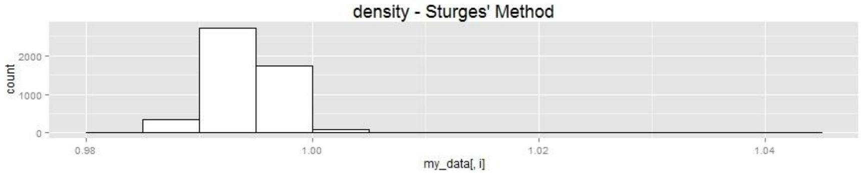
citric.acid - Scott's Method



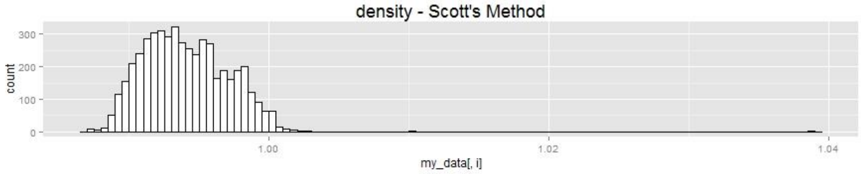
citric.acid - Freedman-Diaconis' Method



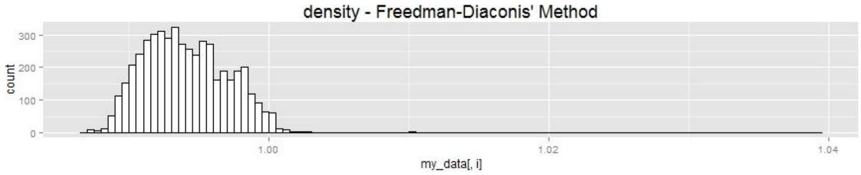
density - Sturges' Method

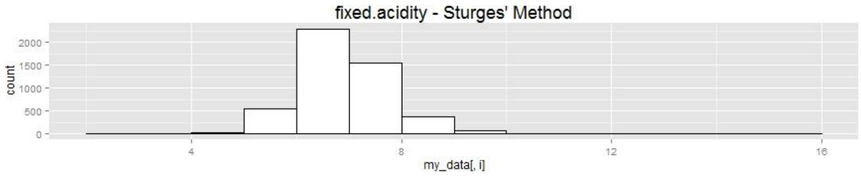


density - Scott's Method

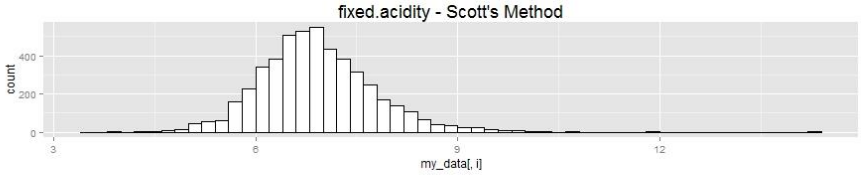


density - Freedman-Diaconis' Method

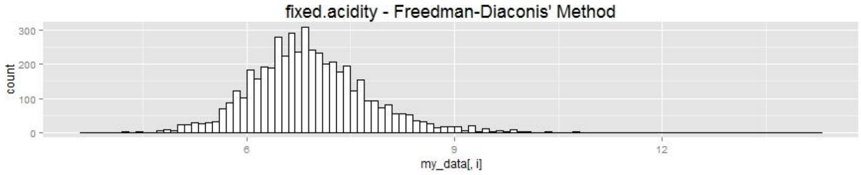




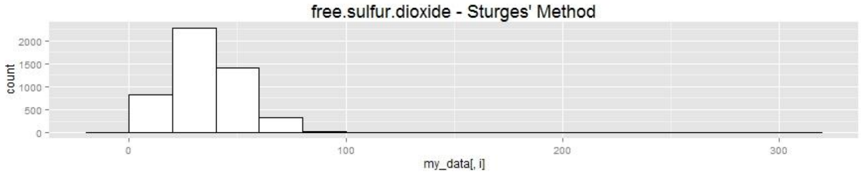
fixed.acidity - Sturges' Method



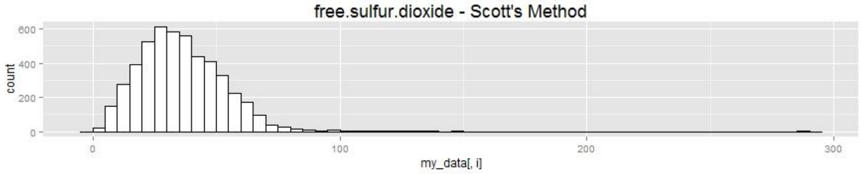
fixed.acidity - Scott's Method



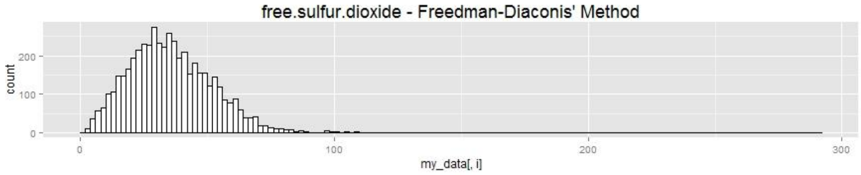
fixed.acidity - Freedman-Diaconis' Method



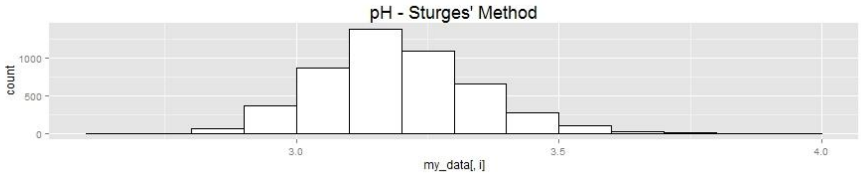
free.sulfur.dioxide - Sturges' Method

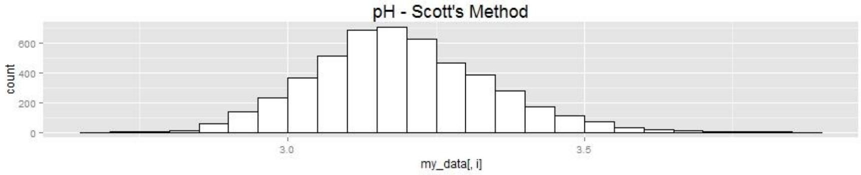


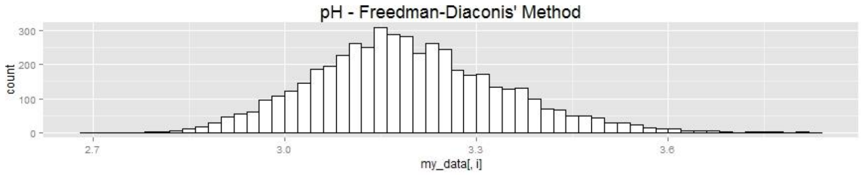
free.sulfur.dioxide - Scott's Method

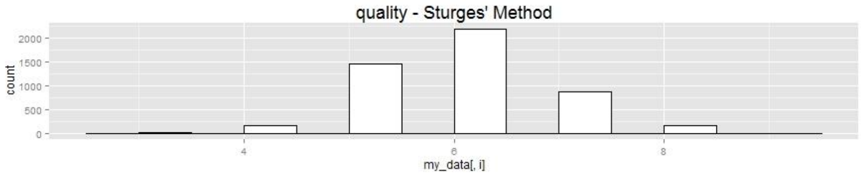


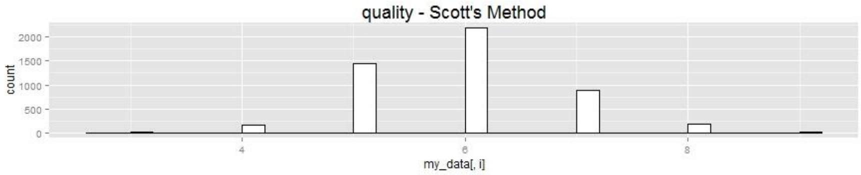
free.sulfur.dioxide - Freedman-Diaconis' Method

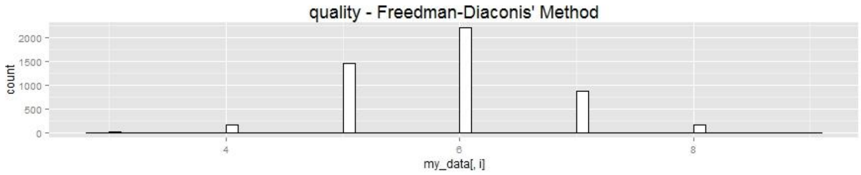

pH - Sturges' Method
count
1000
500
0
3.0
3.5
4.0
my_data[, i]

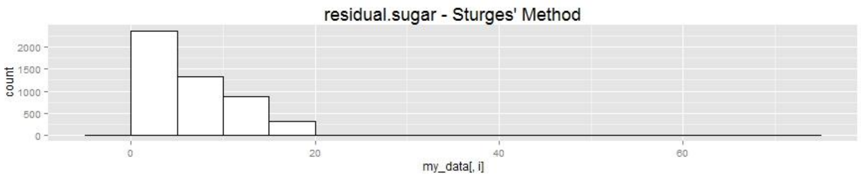

pH - Scott's Method
count
600
400
200
0
3.0
3.5
my_data[, i]

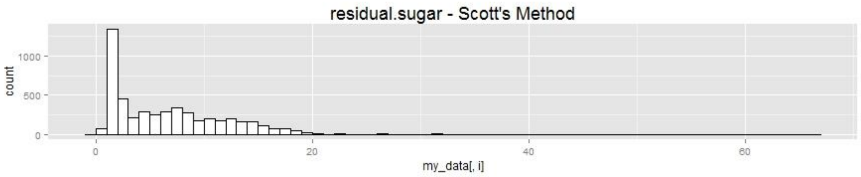

pH - Freedman-Diaconis' Method
count
300
200
100
0
2.7
3.0
3.3
3.6
my_data[, i]

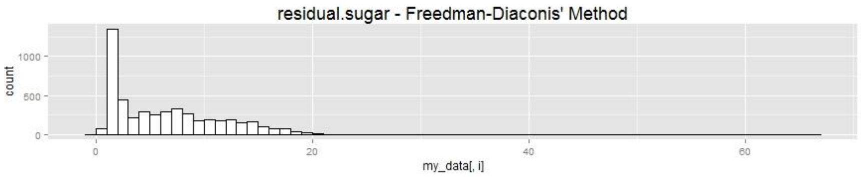

quality - Sturges' Method
count
2000
1500
1000
500
0
4
6
8
my_data[, i]

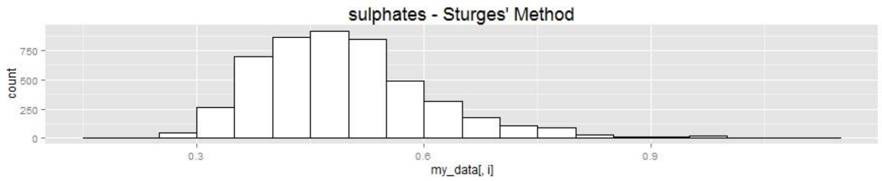

quality - Scott's Method
count
2000
1500
1000
500
0
4
6
8
my_data[, i]

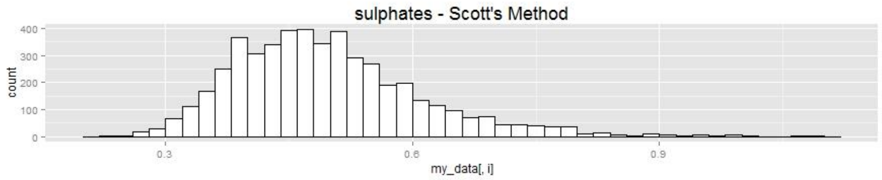

quality - Freedman-Diaconis' Method
count
2000
1500
1000
500
0
4
6
8
my_data[, i]

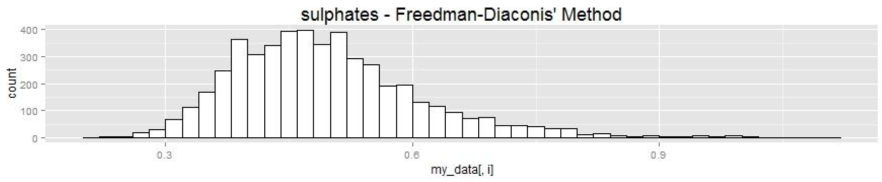

residual.sugar - Sturges' Method
count
2000
1500
1000
500
0
0
20
40
60
my_data[, i]

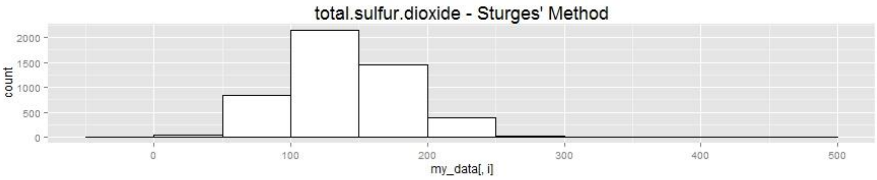

residual.sugar - Scott's Method
count
1000
500
0
0
20
40
60
my_data[, i]

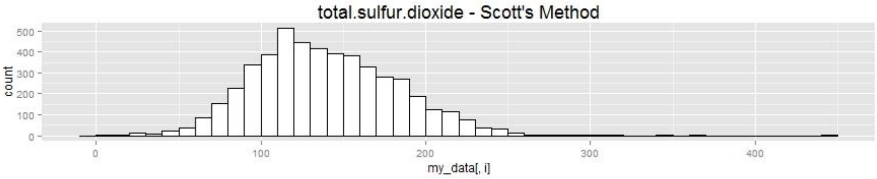

residual.sugar - Freedman-Diaconis' Method
count
1000
500
0
0
20
40
60
my_data[, i]

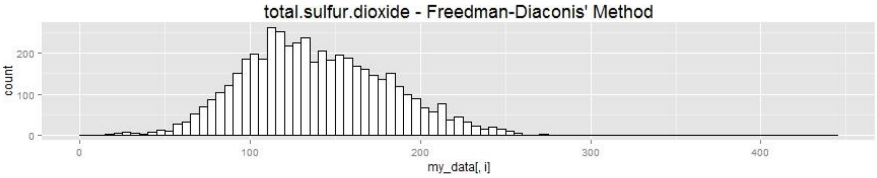

sulphates - Sturges' Method
count
750
500
250
0
0.3
0.6
0.9
my_data[, i]

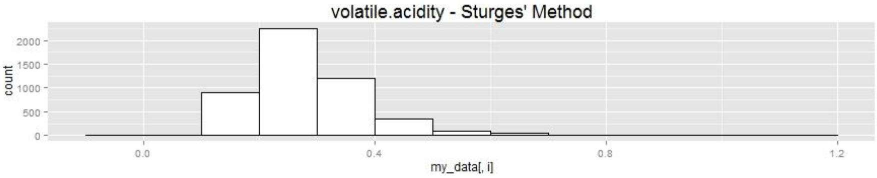

sulphates - Scott's Method
count
400
300
200
100
0
0.3
0.6
0.9
my_data[, i]

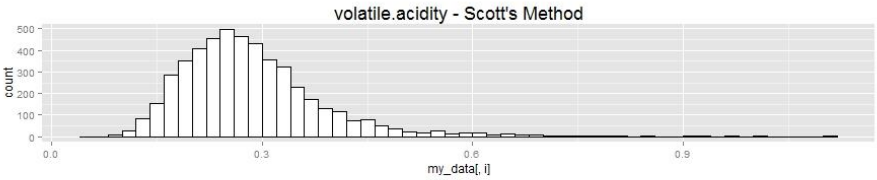

sulphates - Freedman-Diaconis' Method
count
400
300
200
100
0
0.3
0.6
0.9
my_data[, i]

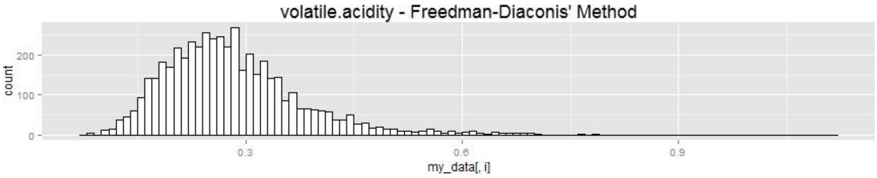










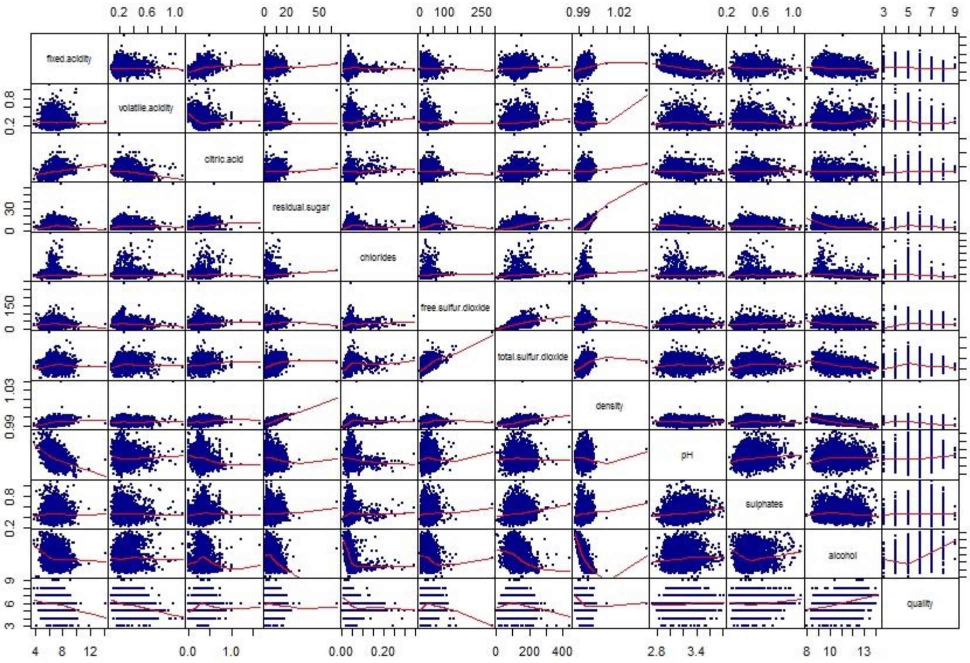


We can observer the following things:

- Fixed acidity, volatile acidity and citric acid have outliers. If those outliers are eliminated distribution of the variables may be taken to be symmetric.
- Residual sugar has a positively skewed distribution. With a spike at the left hand side, which can not been seen in Sturges' binning.
- Citric acid also has a spike at about 0.5, which can not been seen with Sturges' binning.
- Some of the variables, like free sulfur dioxide and density, have a few outliers but these are very different from the rest.
- Chlorides have a very high positive kurtosis and so do free sulfur dioxides.
- Total sulfur dioxide looks to have the largest range between min and max value.
- Alcohol has an irregular shaped distribution but it does not have pronounced outliers.

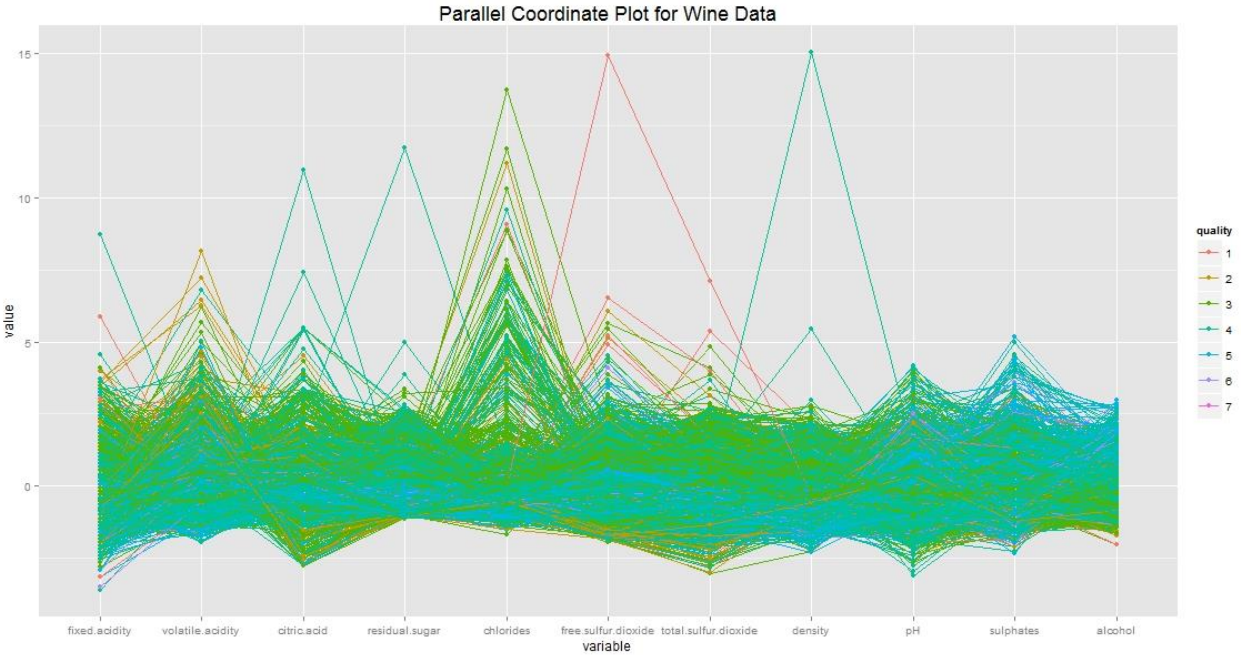
- The spread for the quality of White wine seems to exhibit a normal distribution, with a peak quality around quality rating 6.
- We can see that the variables have vastly different means and variances.

Scatter plot of the data and the parallel coordinates representation

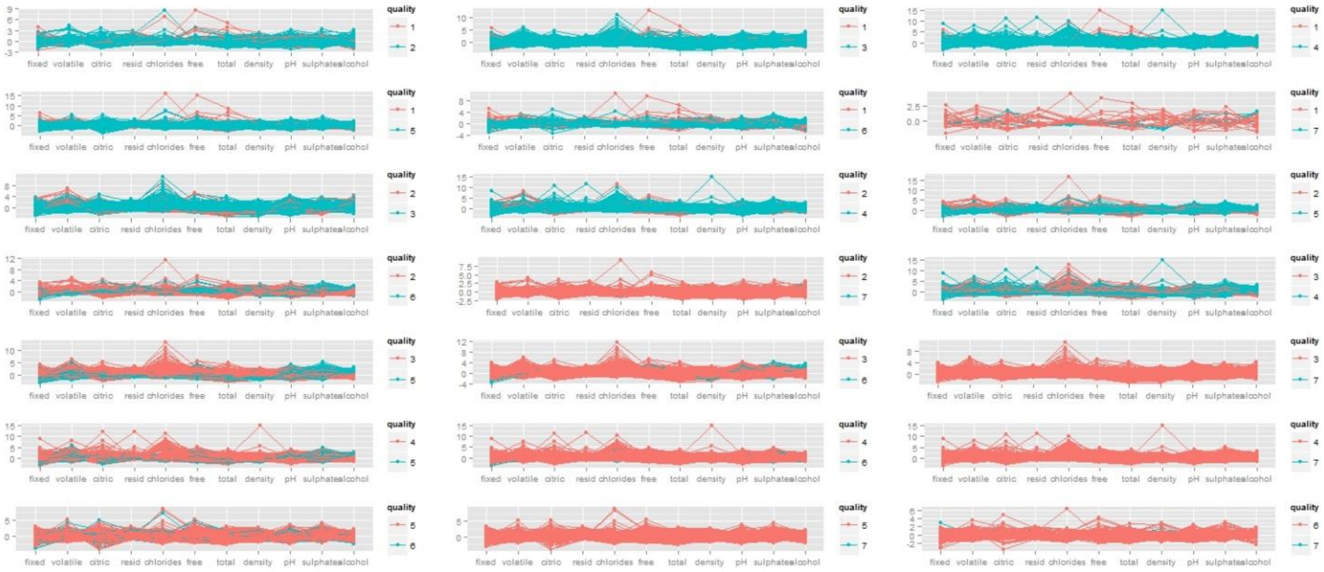


By plotting the variables against each other it becomes obvious that some are strongly correlated: in other words, there is an overlap in the power of some variables at explaining/accounting for the data variability. For example density seems to be correlated with residual sugar, total sulfur dioxide seems to be correlated with free sulfur dioxide and alcohol and density have a negative correlation but there are some influential points in this plot.

Here are all the quality categories in the same parallel coordinate plot:



Here I compare all quality categories to each other:

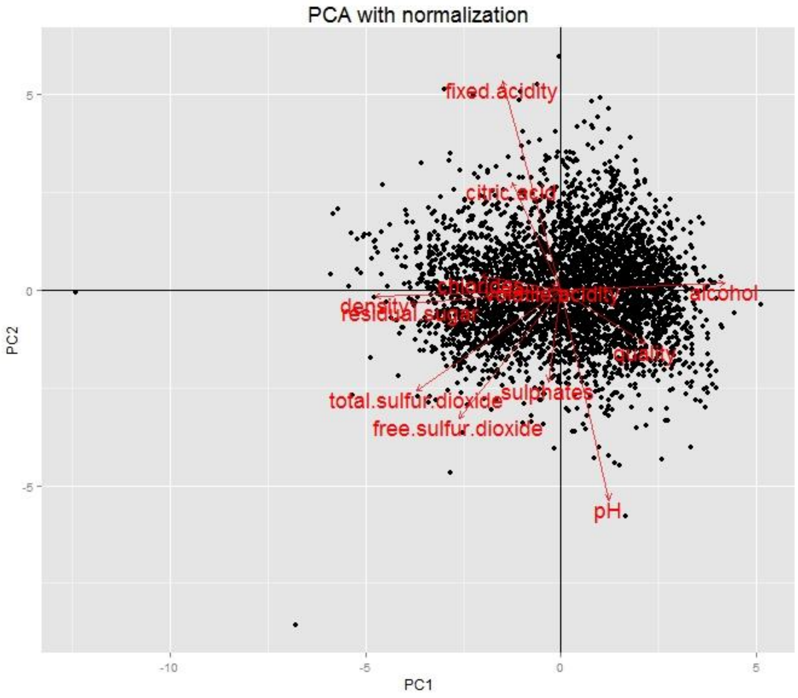


In the parallel coordinate plots we see some outliers in the mess. For example free sulfur dioxide has a large spike for one observation of quality 1 and density for an observation for quality 4. There are no clear separating variables.

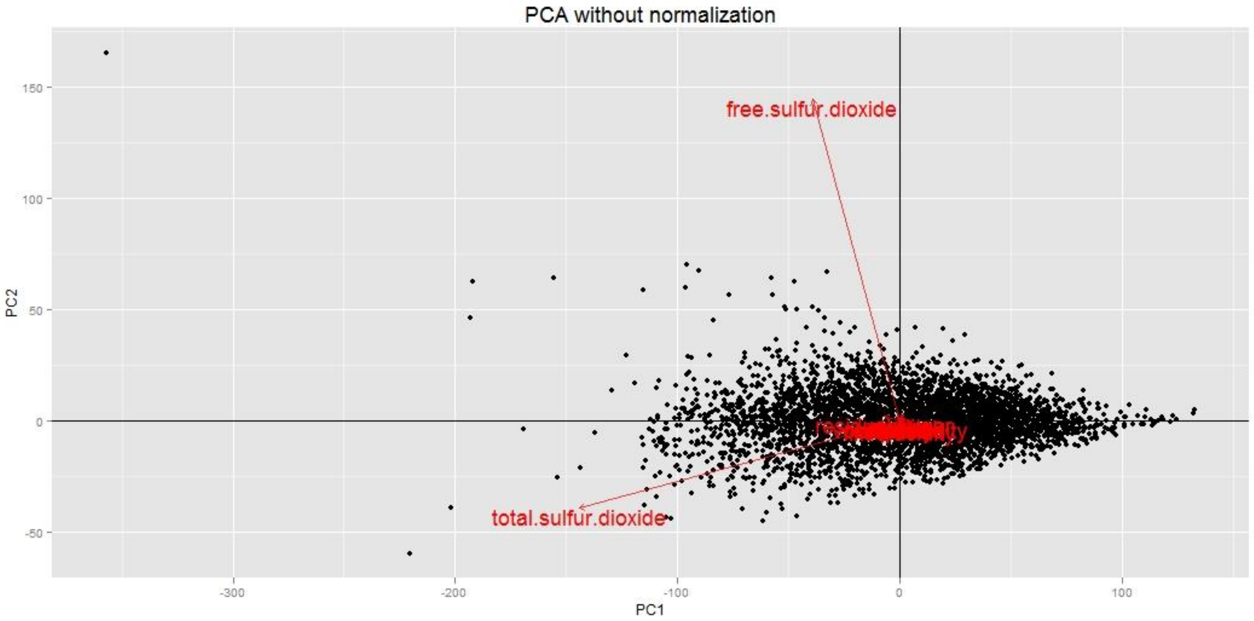
Principal components analysis

Here I perform principal component analysis (PCA) and show a resulting plot for white wine. What's interesting about this plot is that judging by the first two principal components; a quality is very much correlated with alcohol content and pH.

Here's the plot with PC1 on the horizontal axis and PC2 on the vertical axis:

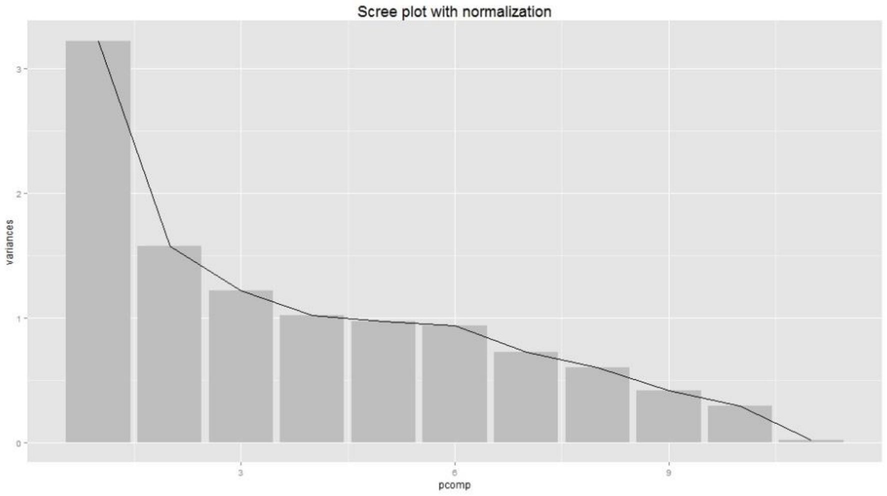


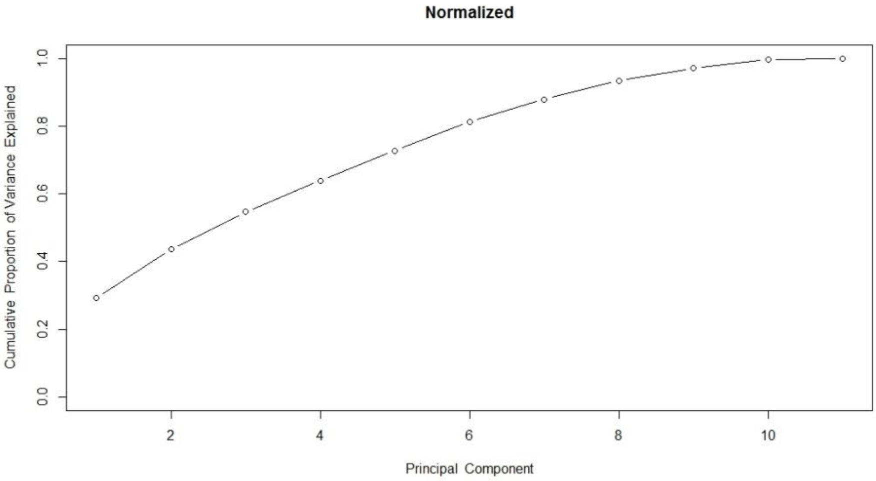
In the above plot we can see two points clearly being away from the center of the point cloud. Both are in the bottom left quadrant.



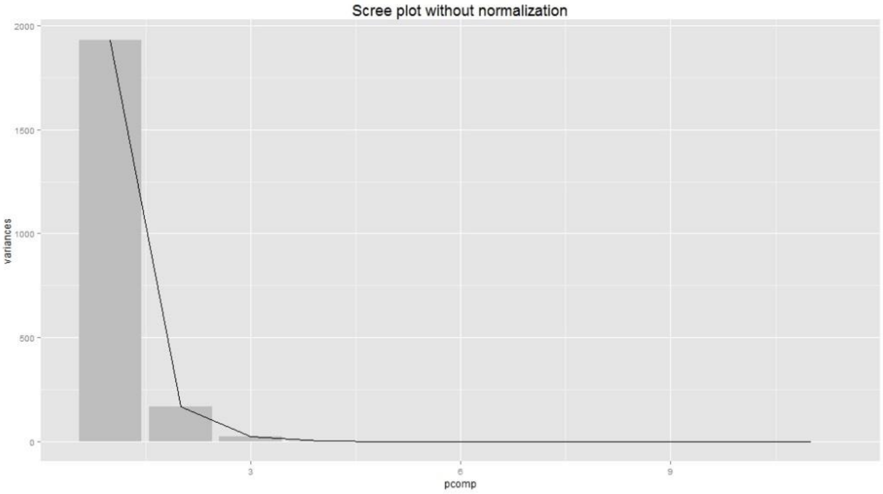
Here is PCA without normalization. We can see that free sulfur dioxide and total sulfur dioxide dominate because of their large measurement units.

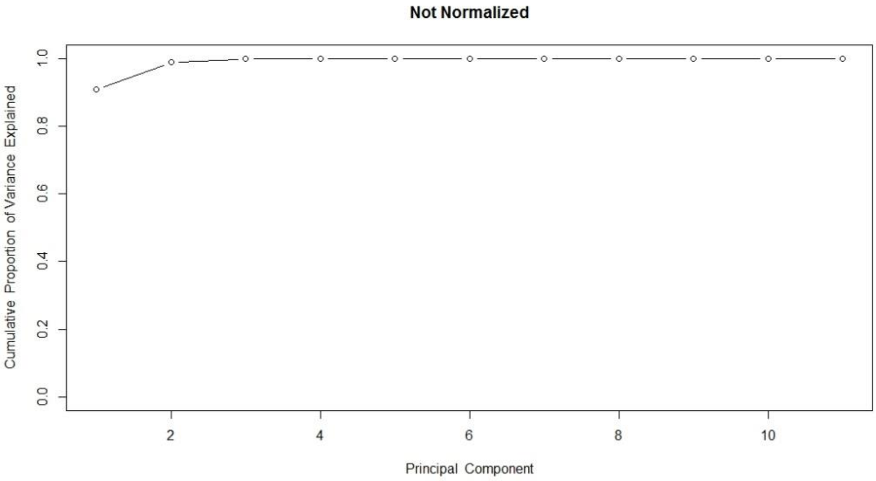
These plots only show the first two principal components. So what we see here is not the whole story, because there are more components, which we can see in the scree plots:





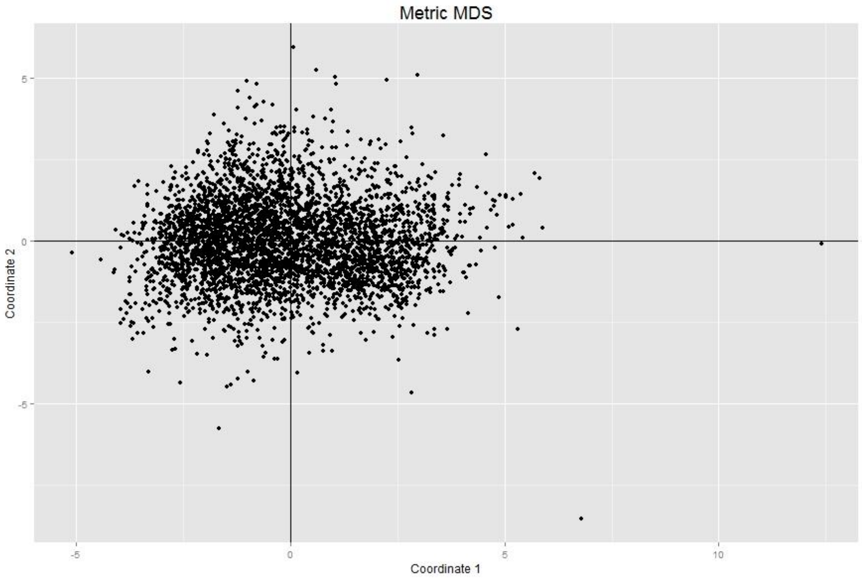
In the normalized case PC1 and PC2 explain respectively ~30% and ~14% of the data's total variability, summing up to a 44% of the total variability. In order to get, for example, 80% variance explained we would need to use the first 6 principle components.





A scree plot allows a graphical assessment of the relative contribution of the PCs in explaining the variability of the data. As we can see from the PCA plots and scree plots, if we do not normalize the data, free sulfur dioxide and total sulfur dioxide become dominant just because of their large measurement units.

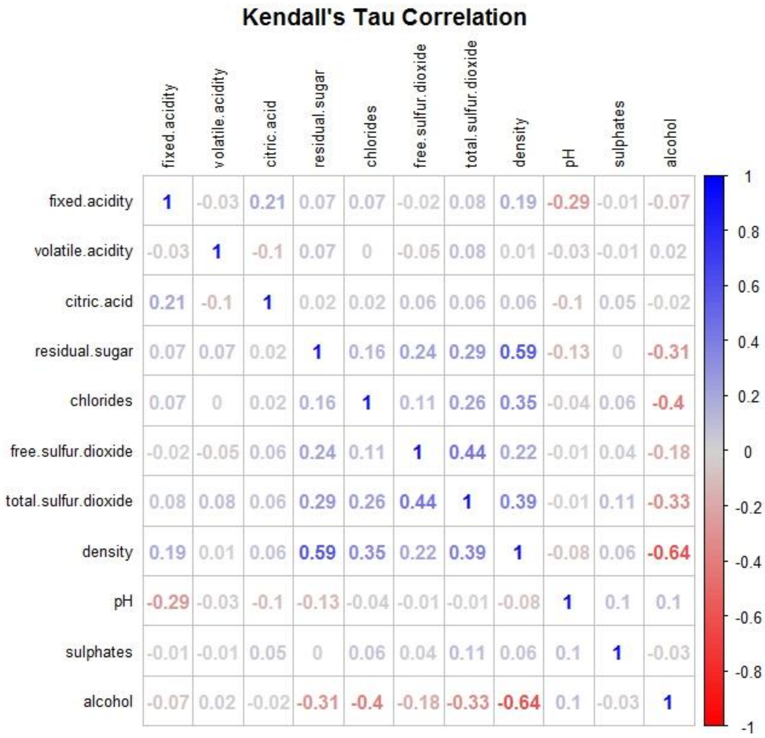
2D MDS representation



Seems to be the same as the normalized PCA plot, where 2 points are clearly far from the center.

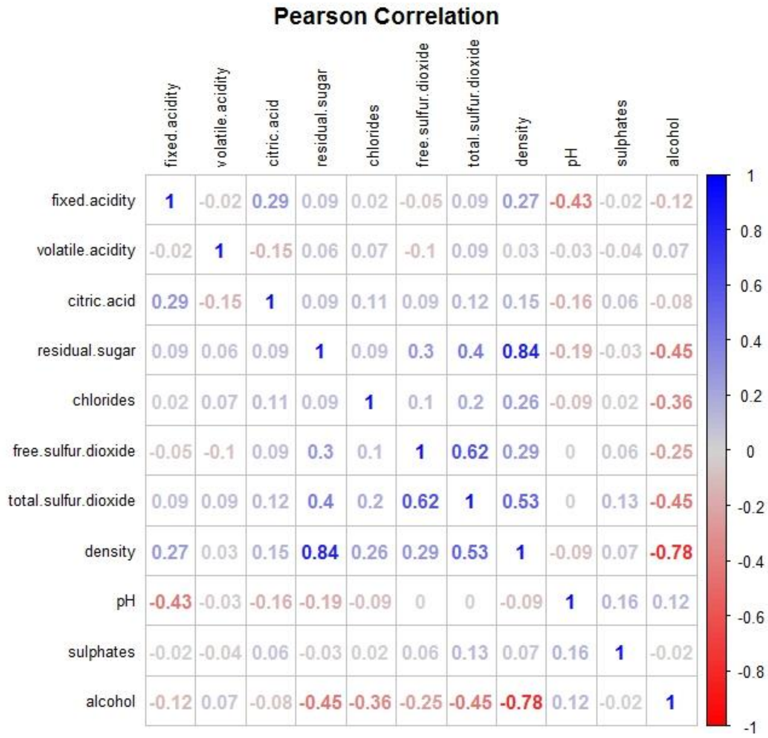
Pearson and Kendall's tau correlation tables

Next are correlation tables for the attributes. I understand this as looking only at the possible predictors for quality.



Kendall's Tau Correlation

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol
fixed.acidity	1	-0.03	0.21	0.07	0.07	-0.02	0.08	0.19	-0.29	-0.01	-0.07
volatile.acidity	-0.03	1	-0.1	0.07	0	-0.05	0.08	0.01	-0.03	-0.01	0.02
citric.acid	0.21	-0.1	1	0.02	0.02	0.06	0.06	0.06	-0.1	0.05	-0.02
residual.sugar	0.07	0.07	0.02	1	0.16	0.24	0.29	0.59	-0.13	0	-0.31
chlorides	0.07	0	0.02	0.16	1	0.11	0.26	0.35	-0.04	0.06	-0.4
free.sulfur.dioxide	-0.02	-0.05	0.06	0.24	0.11	1	0.44	0.22	-0.01	0.04	-0.18
total.sulfur.dioxide	0.08	0.08	0.06	0.29	0.26	0.44	1	0.39	-0.01	0.11	-0.33
density	0.19	0.01	0.06	0.59	0.35	0.22	0.39	1	-0.08	0.06	-0.64
pH	-0.29	-0.03	-0.1	-0.13	-0.04	-0.01	-0.01	-0.08	1	0.1	0.1
sulphates	-0.01	-0.01	0.05	0	0.06	0.04	0.11	0.06	0.1	1	-0.03
alcohol	-0.07	0.02	-0.02	-0.31	-0.4	-0.18	-0.33	-0.64	0.1	-0.03	1



	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol
fixed.acidity	1	-0.02	0.29	0.09	0.02	-0.05	0.09	0.27	-0.43	-0.02	-0.12
volatile.acidity	-0.02	1	-0.15	0.06	0.07	-0.1	0.09	0.03	-0.03	-0.04	0.07
citric.acid	0.29	-0.15	1	0.09	0.11	0.09	0.12	0.15	-0.16	0.06	-0.08
residual.sugar	0.09	0.06	0.09	1	0.09	0.3	0.4	0.84	-0.19	-0.03	-0.45
chlorides	0.02	0.07	0.11	0.09	1	0.1	0.2	0.26	-0.09	0.02	-0.36
free.sulfur.dioxide	-0.05	-0.1	0.09	0.3	0.1	1	0.62	0.29	0	0.06	-0.25
total.sulfur.dioxide	0.09	0.09	0.12	0.4	0.2	0.62	1	0.53	0	0.13	-0.45
density	0.27	0.03	0.15	0.84	0.26	0.29	0.53	1	-0.09	0.07	-0.78
pH	-0.43	-0.03	-0.16	-0.19	-0.09	0	0	-0.09	1	0.16	0.12
sulphates	-0.02	-0.04	0.06	-0.03	0.02	0.06	0.13	0.07	0.16	1	-0.02
alcohol	-0.12	0.07	-0.08	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.02	1

In both of these correlation tables we can see that density correlates with residual sugar (0.59 and 0.84), total sulfur dioxide correlations with free sulfur dioxide (0.44 and 0.62), alcohol and density have a negative correlation (-0.64, -0.78)

Remember to include a short explanation of what you did to get the visualizations.

For the histograms I made a loop that prints the plots with the right histogram bin specifications.

For the scatter plot and parallel coordinates I used the built in R command: pairs(data) and for the parallel coordinates I used a library specific to this plot.

2D MDS representation was done by calculating the Euclidean distance between the rows and then calculating the multidimensional scaling of my data matrix with the maximum dimension of the space which the data are to be represented in specified as 2. Then I made a loop to plot the qualities 2 per plot for individual assessment of the qualities and variables.

PCA was done by a built in function with some additional plotting commands.

Correlations were done with built in command specifying what correlation to use with some plot commands.