Project 2 Data Set Exploration

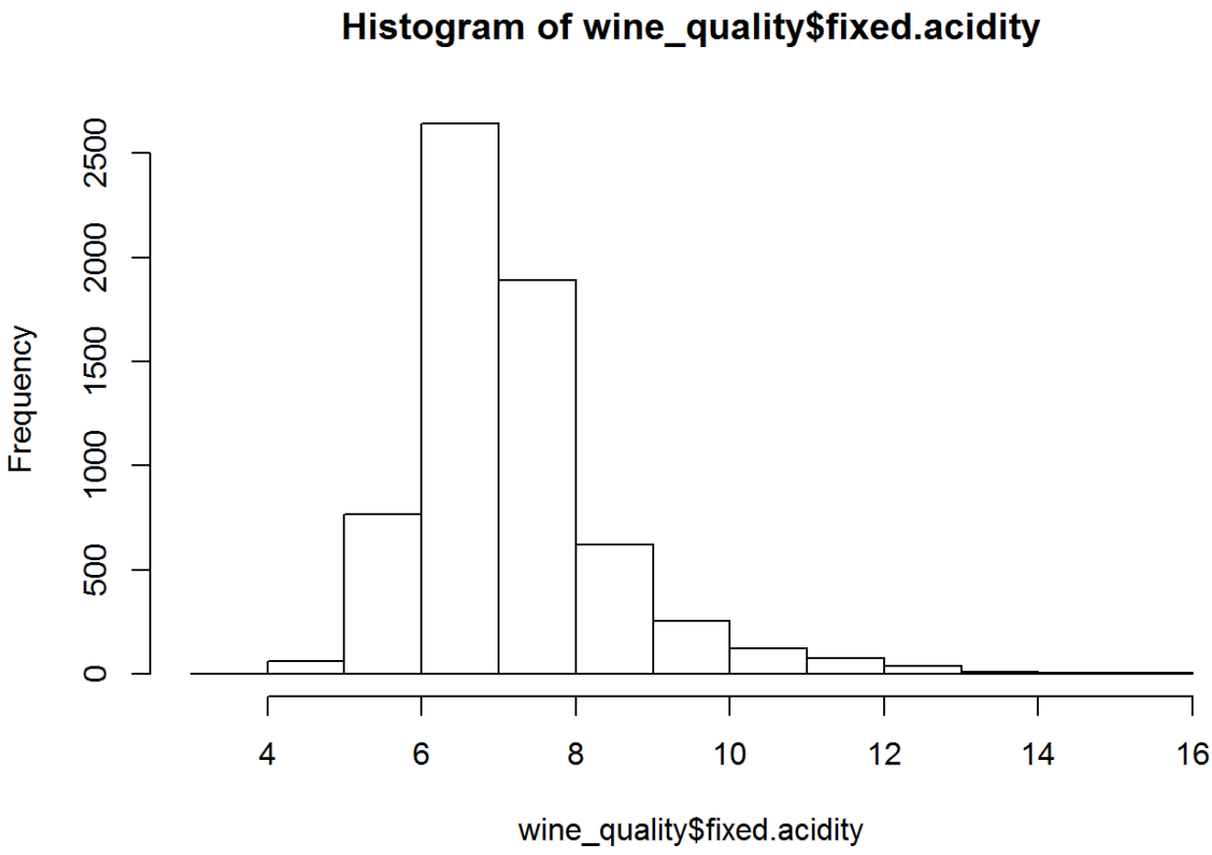
Adam Stopek

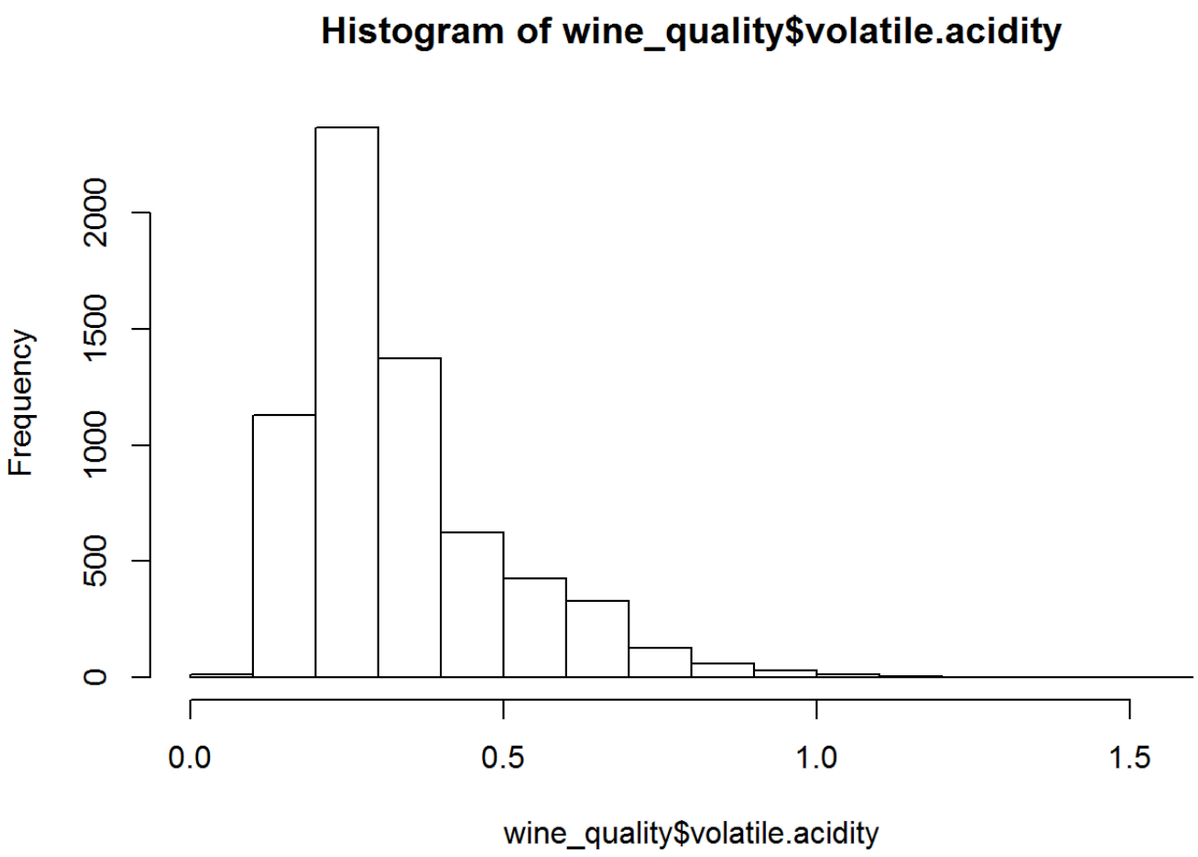
Sunday, October 05, 2014

The data set I chose to work on was a data set about wine quality related to red and white variants of the Portuguese “Vinho Verde” wine. There are 6497 data observations with 12 different variables. Here is a discription of the data set. There are no missing variables in any of the observations.

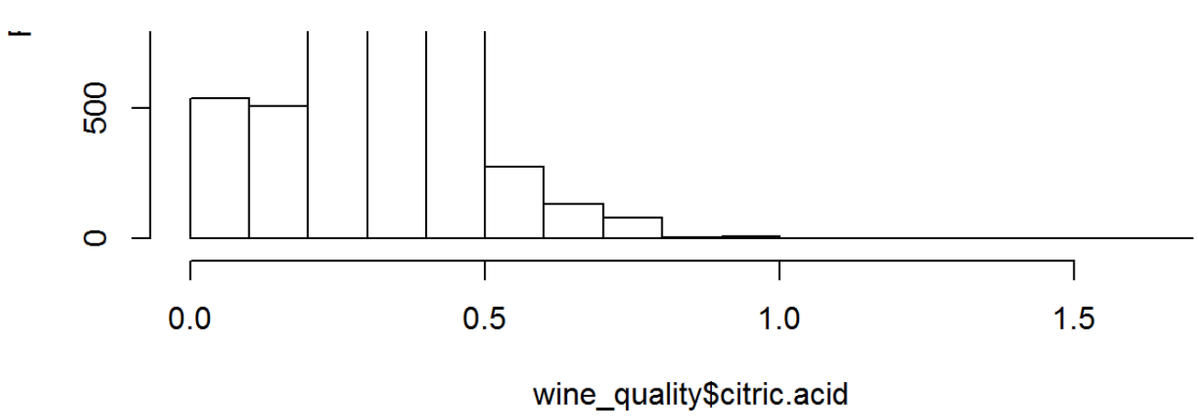
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 3.80   Min.   :0.08     Min.   :0.000   Min.   : 0.60
##  1st Qu.: 6.40   1st Qu.:0.23     1st Qu.:0.250   1st Qu.: 1.80
##  Median : 7.00   Median :0.29     Median :0.310   Median : 3.00
##  Mean   : 7.21   Mean   :0.34     Mean   :0.319   Mean   : 5.44
##  3rd Qu.: 7.70   3rd Qu.:0.40     3rd Qu.:0.390   3rd Qu.: 8.10
##  Max.   :15.90   Max.   :1.58     Max.   :1.660   Max.   :65.80
##    chlorides     free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.009   Min.   :  1.0       Min.   :  6          Min.   :0.987
##  1st Qu.:0.038   1st Qu.: 17.0       1st Qu.: 77          1st Qu.:0.992
##  Median :0.047   Median : 29.0       Median :118          Median :0.995
##  Mean   :0.056   Mean   : 30.5       Mean   :116          Mean   :0.995
##  3rd Qu.:0.065   3rd Qu.: 41.0       3rd Qu.:156          3rd Qu.:0.997
##  Max.   :0.611   Max.   :289.0       Max.   :440          Max.   :1.039
##        pH         sulphates        alcohol        quality       color
##  Min.   :2.72   Min.   :0.220   Min.   : 8.0   Min.   :3.00   red  :1599
##  1st Qu.:3.11   1st Qu.:0.430   1st Qu.: 9.5   1st Qu.:5.00   white:4898
##  Median :3.21   Median :0.510   Median :10.3   Median :6.00
##  Mean   :3.22   Mean   :0.531   Mean   :10.5   Mean   :5.82
##  3rd Qu.:3.32   3rd Qu.:0.600   3rd Qu.:11.3   3rd Qu.:6.00
##  Max.   :4.01   Max.   :2.000   Max.   :14.9   Max.   :9.00
```

Here are distributions of all the data


Histogram of wine_quality$fixed.acidity
Frequency
0
500
1000
1500
2000
2500
4
6
8
10
12
14
16
wine_quality$fixed.acidity

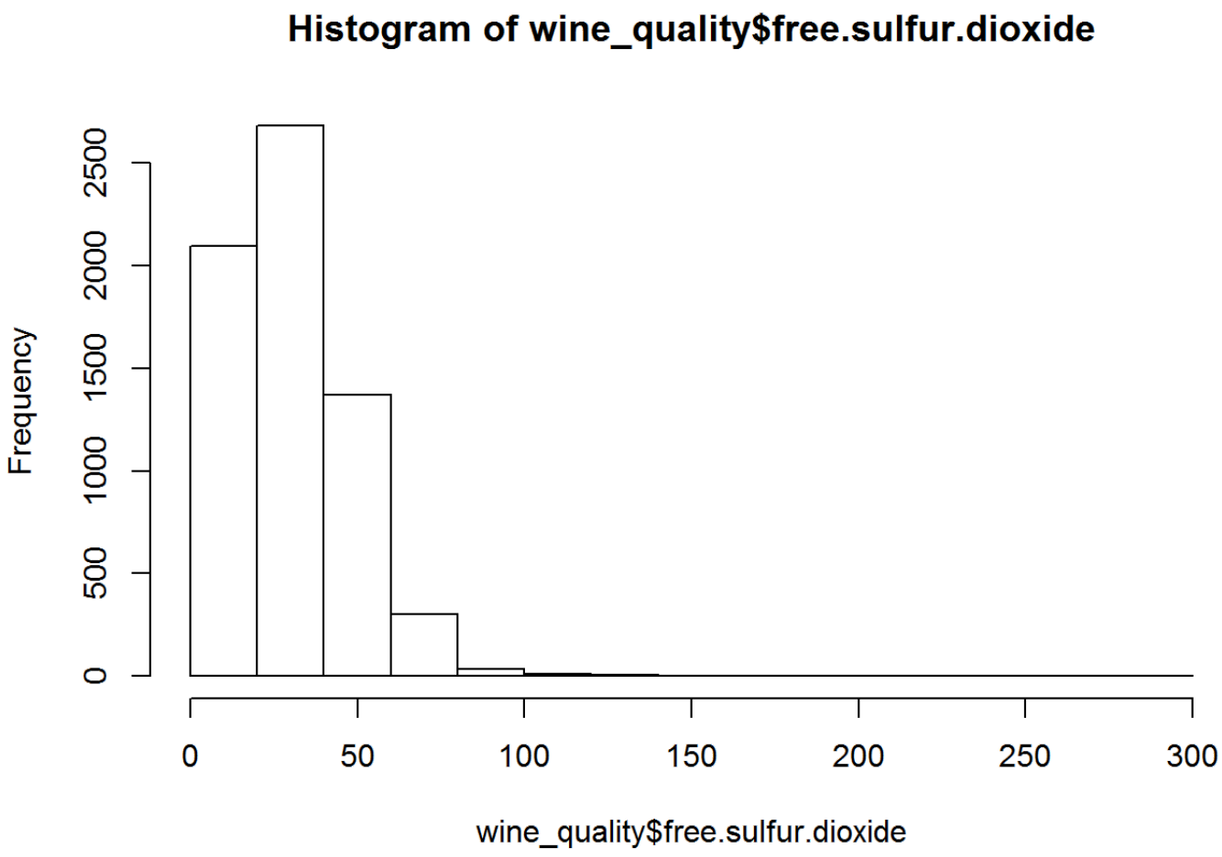

Histogram of wine_quality$volatile.acidity
Frequency
0
500
1000
1500
2000
0.0
0.5
1.0
1.5
wine_quality$volatile.acidity

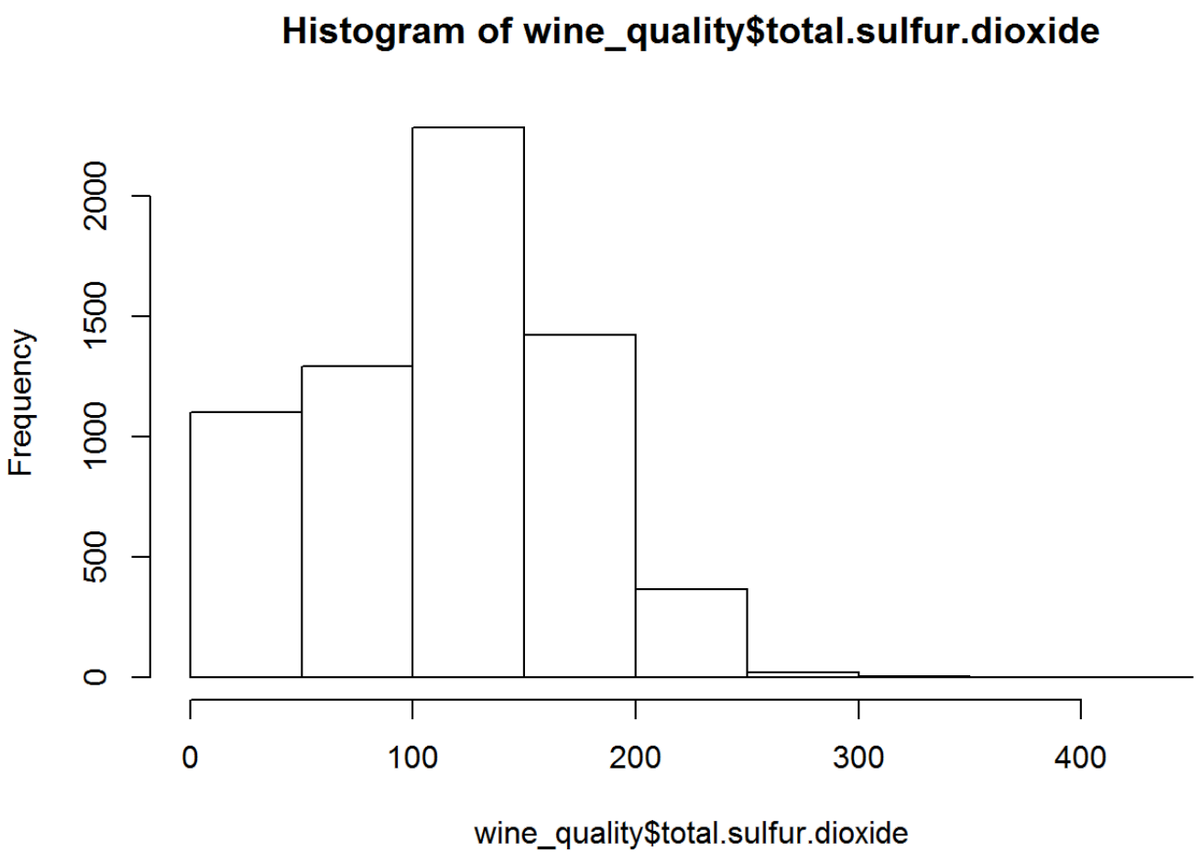

Histogram of wine_quality$citric.acid
Frequency
1000
1500
2000

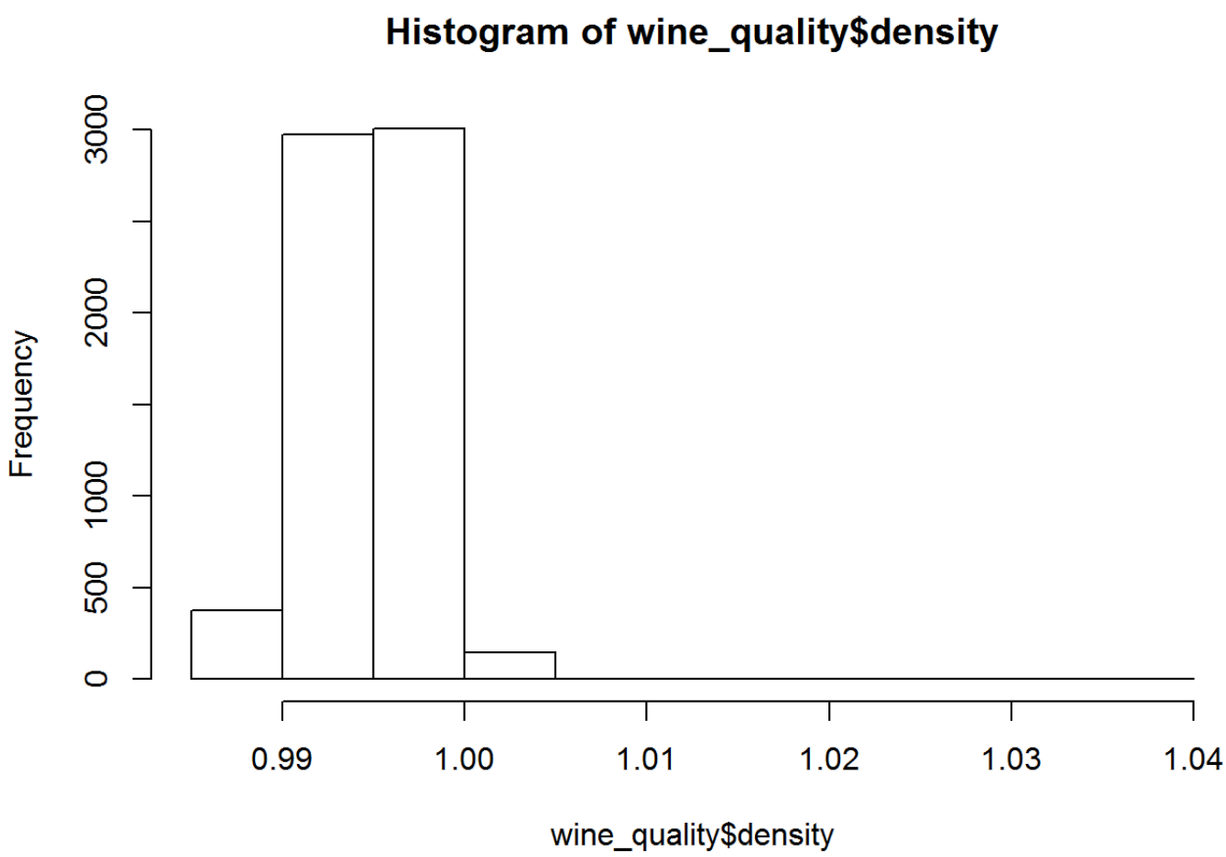

0
500
0.0
0.5
1.0
1.5
wine_quality$citric.acid

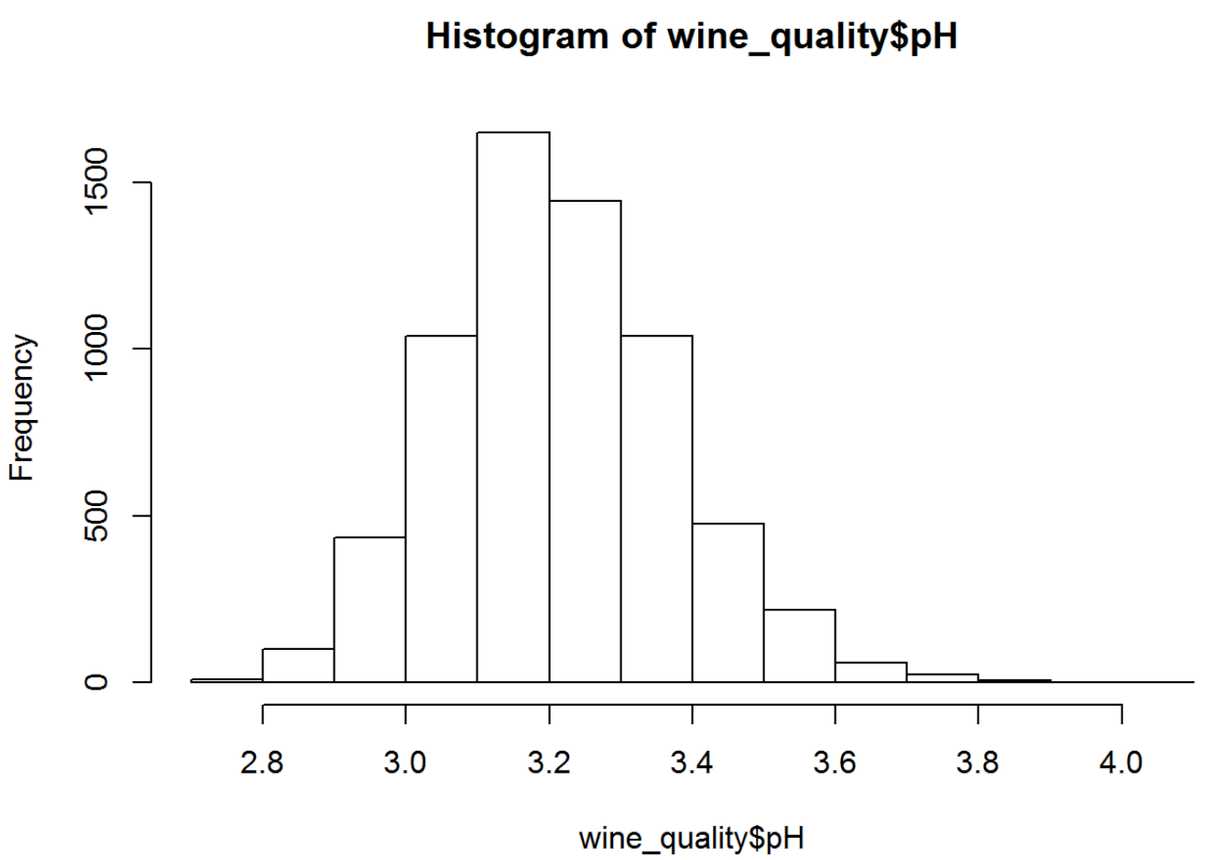

Histogram of wine_quality$residual.sugar
Frequency
0
1000
2000
3000
4000
0
10
20
30
40
50
60
70
wine_quality$residual.sugar

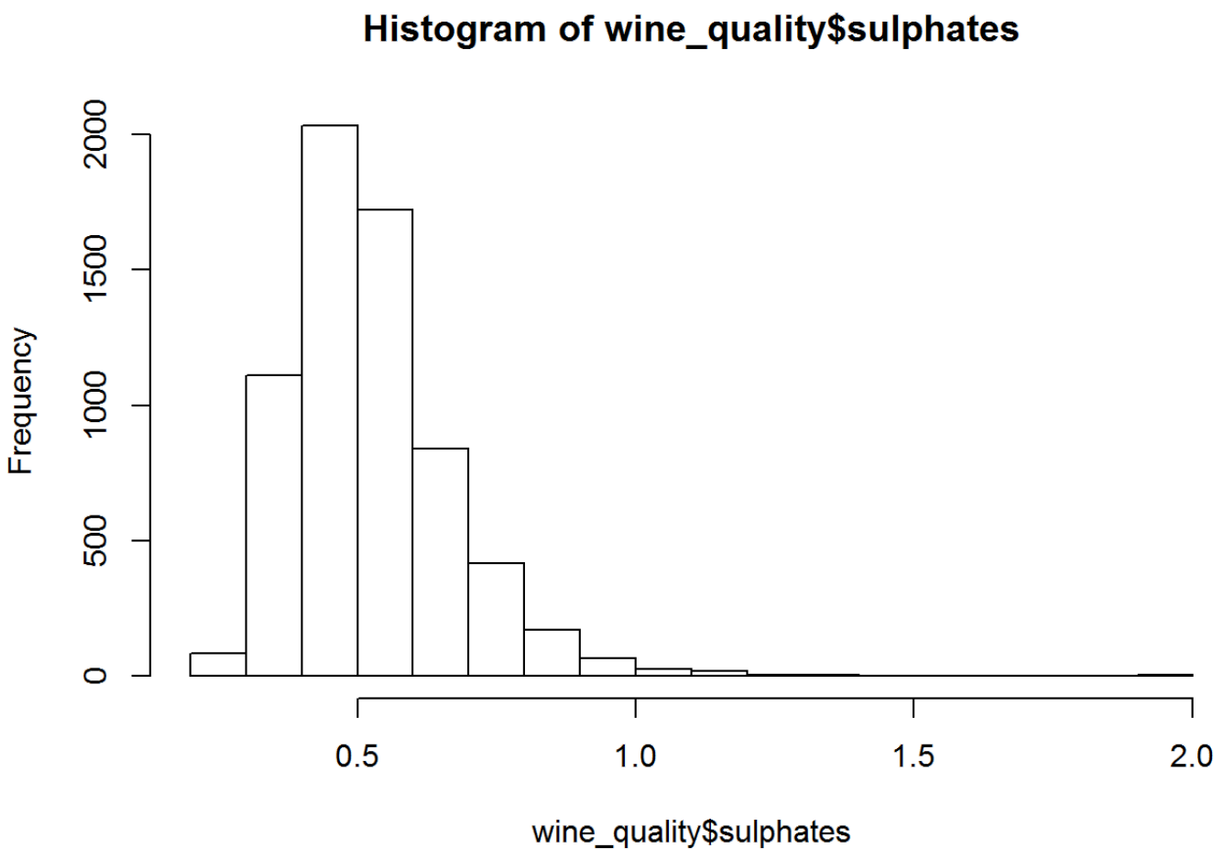

Histogram of wine_quality$chlorides
Frequency
0
1000
2000
3000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
wine_quality$chlorides


Histogram of wine_quality$free.sulfur.dioxide
Frequency
0
500
1000
1500
2000
2500
0
50
100
150
200
250
300
wine_quality$free.sulfur.dioxide

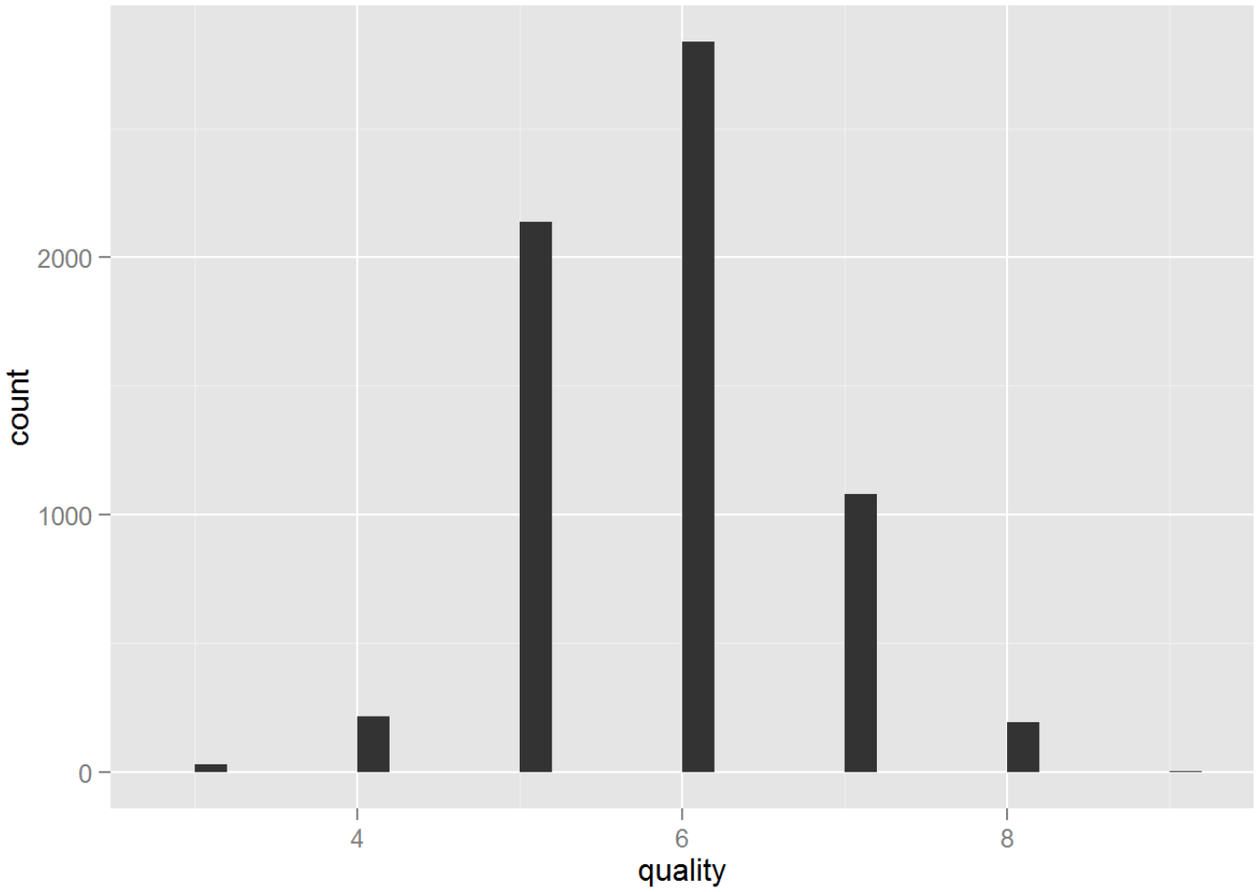

Histogram of wine_quality$total.sulfur.dioxide
Frequency
0
500
1000
1500
2000
0
100
200
300
400
wine_quality$total.sulfur.dioxide

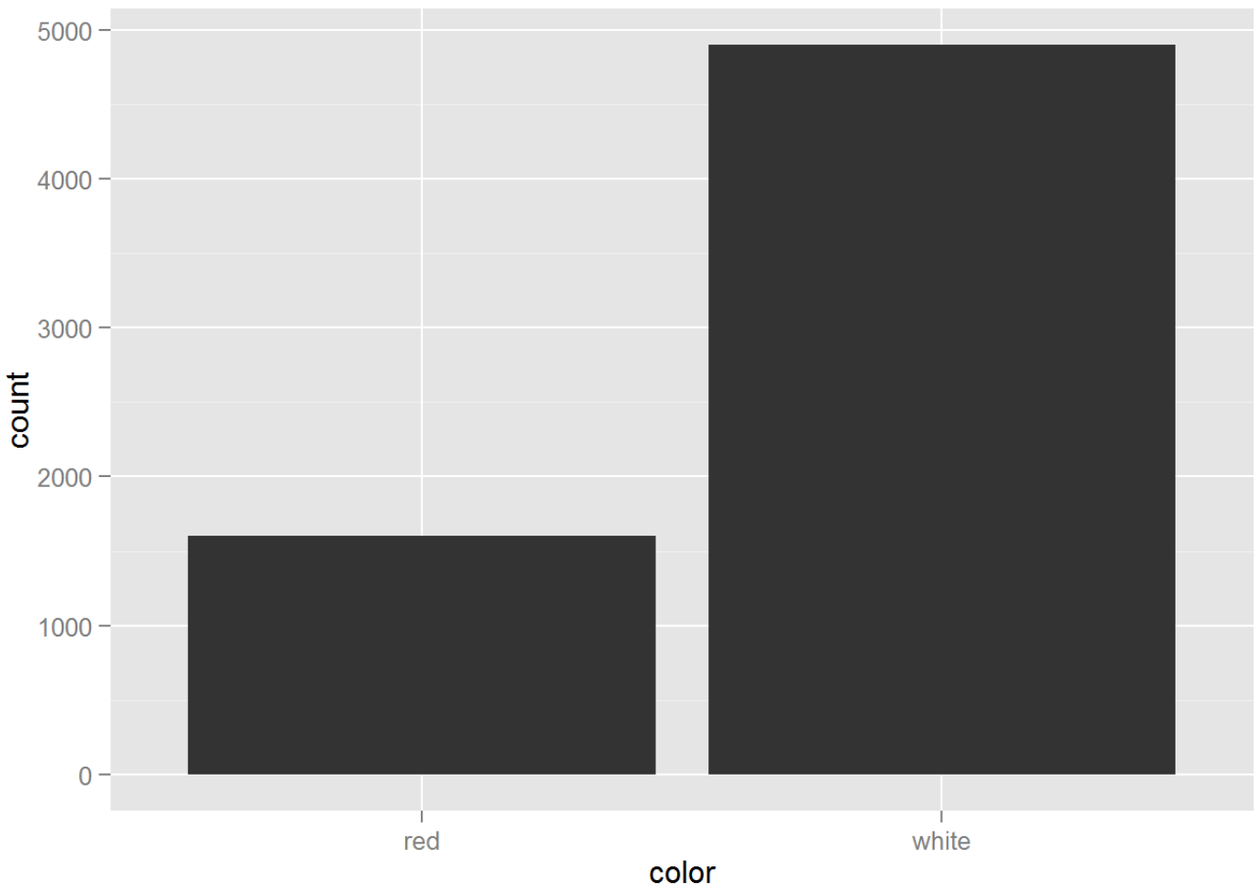

Histogram of wine_quality$density
Frequency
0
500
1000
2000
3000
0.99
1.00
1.01
1.02
1.03
1.04
wine_quality$density


Histogram of wine_quality$pH
Frequency
0
500
1000
1500
2.8
3.0
3.2
3.4
3.6
3.8
4.0
wine_quality$pH

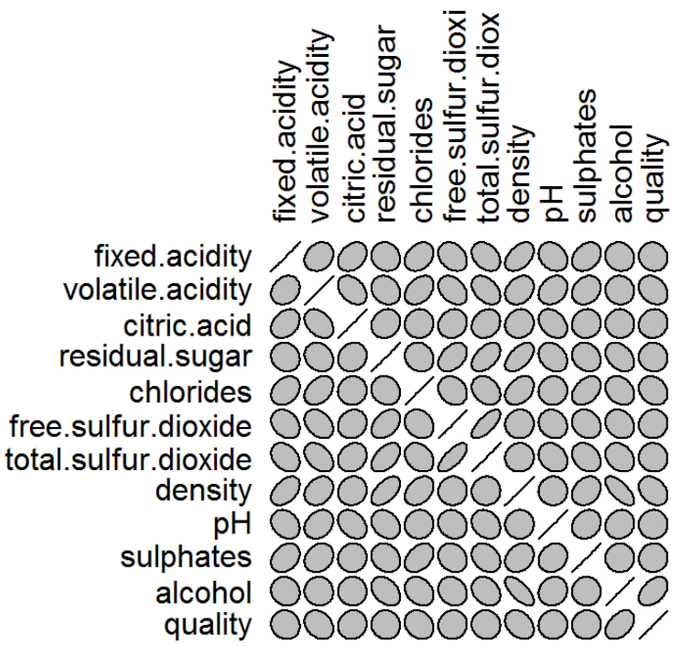

Histogram of wine_quality$sulphates
Frequency
0
500
1000
1500
2000
0.5
1.0
1.5
2.0
wine_quality$sulphates


Histogram of wine_quality$alcohol
Frequency
0
200
400
600
800
1200
8
9
10
11
12
13
14
15
wine_quality$alcohol





Lets look at whether there are any correlated variables. Below is the cross correlation matrix and a visuallization of it



Here are the list of Strong Correlations: Alchohol with Density, Total Sulfur Dioxide with Free sulfar dioxide, Density with Residual Sugar, Alchohol with Quality

What makes a good wine? This question can be answered using Logistic Regression on Quality. The mean value given to quality was 5.8, so I will create a new variable called above average quality and run a classification model on it. I will run a logistic regression and then a step wise procedure on the regression to get only the best variables

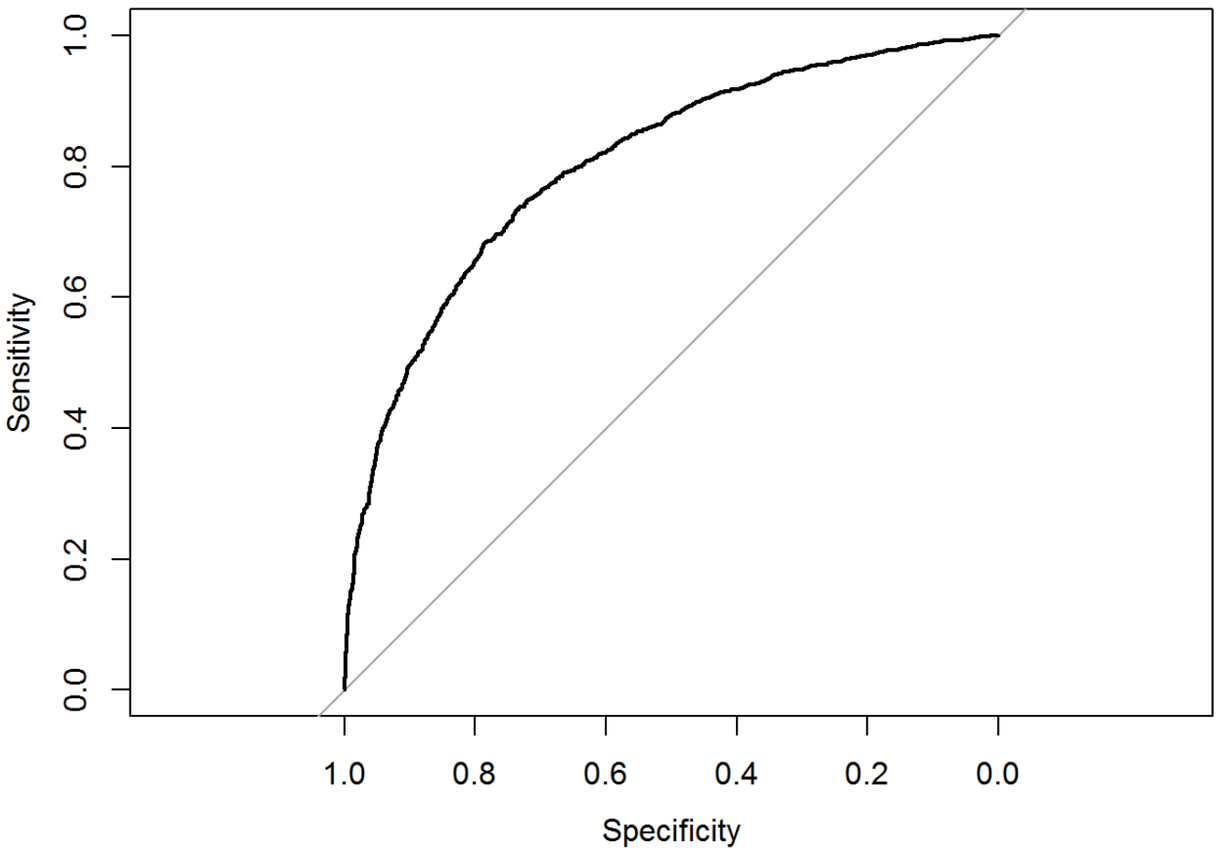
```
##
## Call:
## glm(formula = above_avg_quality ~ fixed.acidity + volatile.acidity +
##     color + citric.acid + residual.sugar + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + density + pH + sulphates + alcohol,
##     family = "binomial", data = wine_quality)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.408  -0.898   0.437   0.817   2.606
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.29e+02   4.54e+01    2.84  0.00449 **
## fixed.acidity         1.06e-01   5.09e-02    2.08  0.03735 *
## volatile.acidity     -4.78e+00   2.95e-01  -16.18  < 2e-16 ***
## colorwhite           -6.61e-01   1.90e-01   -3.47  0.00051 ***
## citric.acid          -4.93e-01   2.56e-01   -1.92  0.05438 .
## residual.sugar        1.20e-01   1.92e-02    6.26  3.8e-10 ***
## chlorides            -1.35e+00   1.05e+00   -1.29  0.19639
## free.sulfur.dioxide   1.47e-02   2.57e-03    5.72  1.1e-08 ***
## total.sulfur.dioxide -5.72e-03   1.05e-03   -5.42  6.0e-08 ***
## density              -1.40e+02   4.60e+01   -3.04  0.00239 **
## pH                    7.88e-01   2.98e-01    2.65  0.00813 **
## sulphates             2.01e+00   2.67e-01    7.52  5.7e-14 ***
## alcohol               8.05e-01   6.10e-02   13.19  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8541.0  on 6496  degrees of freedom
## Residual deviance: 6693.9  on 6484  degrees of freedom
## AIC: 6720
##
## Number of Fisher Scoring iterations: 4
```

```
## Start:  AIC=6720
## above_avg_quality ~ fixed.acidity + volatile.acidity + color +
##     citric.acid + residual.sugar + chlorides + free.sulfur.dioxide +
##     total.sulfur.dioxide + density + pH + sulphates + alcohol
##
##                        Df Deviance  AIC
## - chlorides             1     6696 6720
## <none>                        6694 6720
## - citric.acid           1     6698 6722
## - fixed.acidity         1     6698 6722
## - pH                    1     6701 6725
## - density               1     6703 6727
## - color                 1     6706 6730
## - total.sulfur.dioxide  1     6723 6747
## - free.sulfur.dioxide   1     6728 6752
## - residual.sugar        1     6733 6757
## - sulphates             1     6753 6777
## - alcohol               1     6849 6873
## - volatile.acidity      1     6990 7014
##
## Step:  AIC=6720
## above_avg_quality ~ fixed.acidity + volatile.acidity + color +
##     citric.acid + residual.sugar + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + alcohol
##
##                        Df Deviance  AIC
## <none>                        6696 6720
## + chlorides             1     6694 6720
## - citric.acid           1     6700 6722
## - fixed.acidity         1     6701 6723
## - pH                    1     6704 6726
## - density               1     6706 6728
## - color                 1     6707 6729
## - total.sulfur.dioxide  1     6724 6746
## - free.sulfur.dioxide   1     6729 6751
## - residual.sugar        1     6737 6759
## - sulphates             1     6753 6775
## - alcohol               1     6853 6875
## - volatile.acidity      1     7002 7024
```

This is the model I used to predict quality. It is a stepwise logistic regression model.

```
##
## Call:
## glm(formula = above_avg_quality ~ fixed.acidity + volatile.acidity +
##     color + citric.acid + residual.sugar + free.sulfur.dioxide +
##     total.sulfur.dioxide + density + pH + sulphates + alcohol,
##     family = "binomial", data = wine_quality)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.428  -0.897   0.436   0.818   2.621
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.34e+02   4.52e+01    2.97  0.00297 **
## fixed.acidity         1.16e-01   5.04e-02    2.31  0.02117 *
## volatile.acidity     -4.82e+00   2.94e-01  -16.43  < 2e-16 ***
## colorwhite           -6.32e-01   1.89e-01   -3.34  0.00082 ***
## citric.acid          -5.48e-01   2.53e-01   -2.17  0.03007 *
## residual.sugar        1.23e-01   1.90e-02    6.49  8.6e-11 ***
## free.sulfur.dioxide   1.46e-02   2.57e-03    5.67  1.4e-08 ***
## total.sulfur.dioxide -5.65e-03   1.05e-03   -5.37  8.0e-08 ***
## density              -1.46e+02   4.59e+01   -3.18  0.00149 **
## pH                    8.57e-01   2.93e-01    2.93  0.00343 **
## sulphates             1.93e+00   2.59e-01    7.44  1.0e-13 ***
## alcohol               8.10e-01   6.10e-02   13.28  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8541.0  on 6496  degrees of freedom
## Residual deviance: 6695.6  on 6485  degrees of freedom
## AIC: 6720
##
## Number of Fisher Scoring iterations: 4
```

Here is the ROC curve which describes the quality of this model




```
##
## Call:
## roc.formula(formula = above_avg_quality ~ prob, data = wine_quality)
##
## Data: prob in 2384 controls (above_avg_quality 0) < 4113 cases (above_avg_quality 1).
## Area under the curve: 0.804
```

Citation:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.