Analysis of Wine Quality Data

In the second example of data mining for knowledge discovery we consider a set of observations on a number of red and white wine varieties involving their chemical properties and ranking by tasters. Wine industry shows a recent growth spurt as social drinking is on the rise. The price of wine depends on a rather abstract concept of wine appreciation by wine tasters, opinion among whom may have a high degree of variability. Pricing of wine depends on such a volatile factor to some extent. Another key factor in wine certification and quality assessment is physicochemical tests which are laboratory-based and takes into account factors like acidity, pH level, presence of sugar and other chemical properties. For the wine market, it would be of interest if human quality of tasting can be related to the chemical properties of wine so that certification and quality assessment and assurance process is more controlled.

Two datasets are available of which one dataset is on red wine and have 1599 different varieties and the other is on white wine and have 4898 varieties. Only white wine data is analysed. All wines are produced in a particular area of Portugal. Data are collected on 12 different properties of the wines one of which is Quality, based on sensory data, and the rest are on chemical properties of the wines including density, acidity, alcohol content etc. All chemical properties of wines are continuous variables. Quality is an ordinal variable with possible ranking from 1 (worst) to 10 (best). Each variety of wine is tasted by three independent tasters and the final rank assigned is the median rank given by the tasters.

Objective of the Analysis

Prediction of Quality ranking from the chemical properties of the wines

A predictive model developed on this data is expected to provide guidance to vineyards regarding quality and price expected on their produce without heavy reliance on volatility of wine tasters.

Data Files for this case (*right-click and "save as"*) :

- Wine data - Wine_data.csv [1]
- Training dataset - Training50_winedata.csv [2]
- Test dataset - Test50_winedata.csv [3]

The following analytical approaches are taken:

- Multiple regression: The response Quality is assumed to be a continuous variable and is

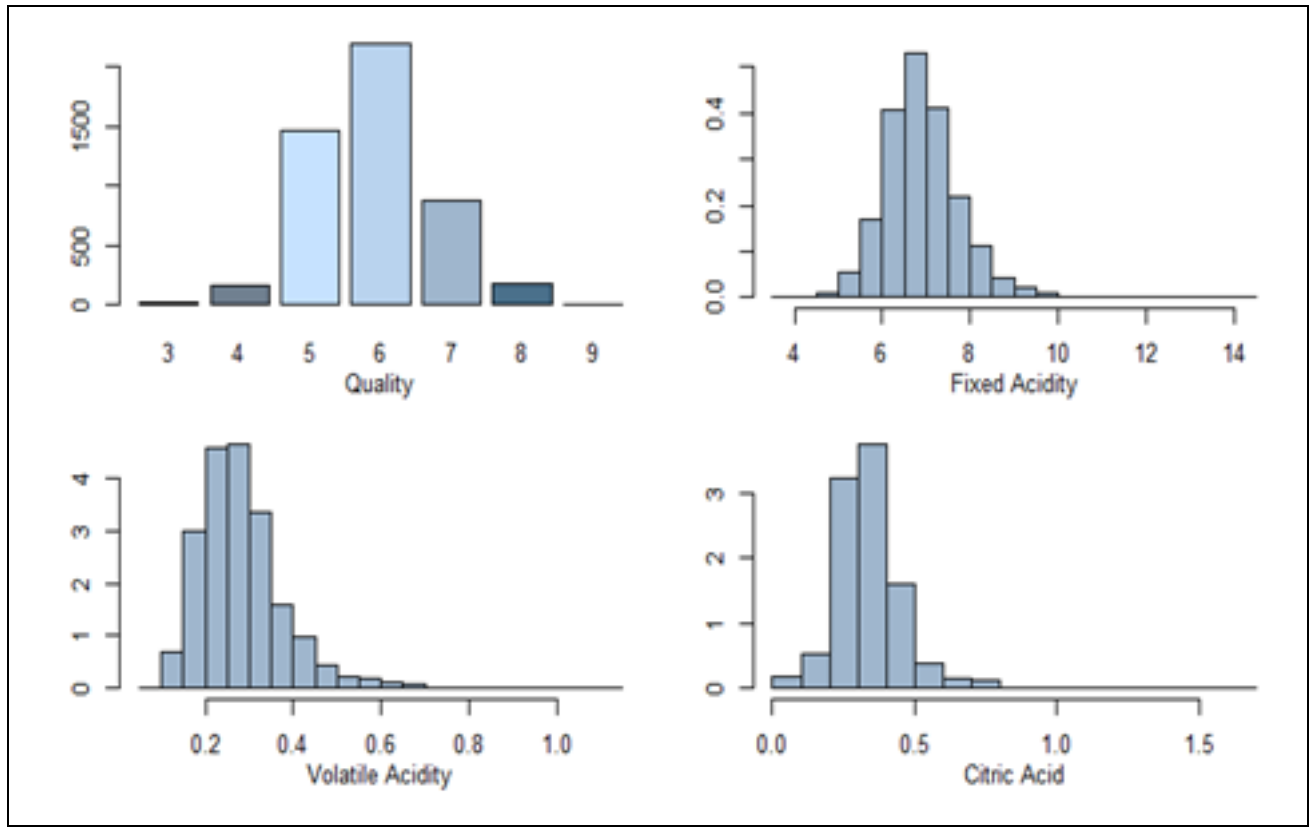
predicted by the independent predictors, all of which are continuous
- Regression Tree
- Classification of wines based on the chemical properties: Unsupervised analysis

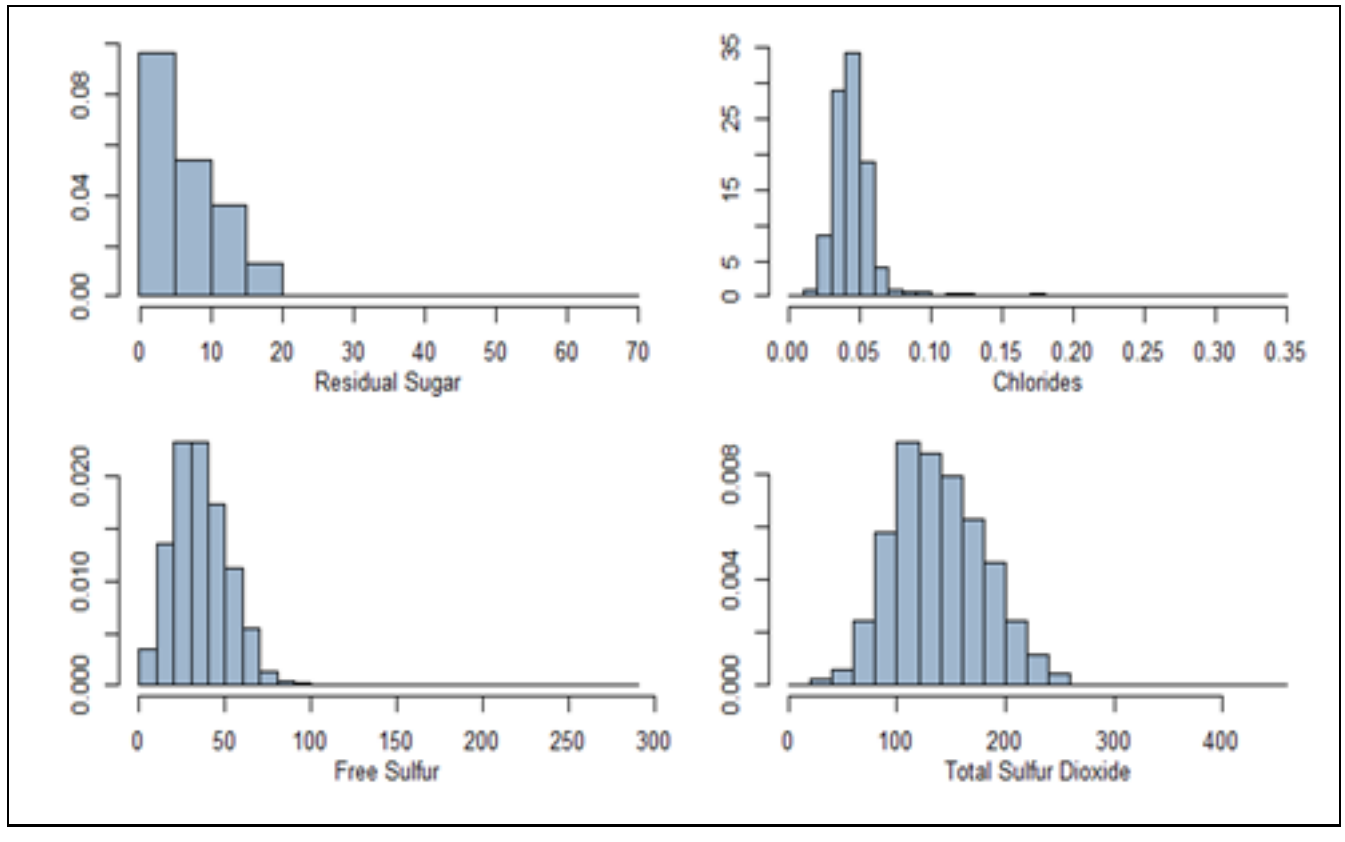
WQD.1 - Exploratory Data Analysis (EDA) and Data Pre-processing

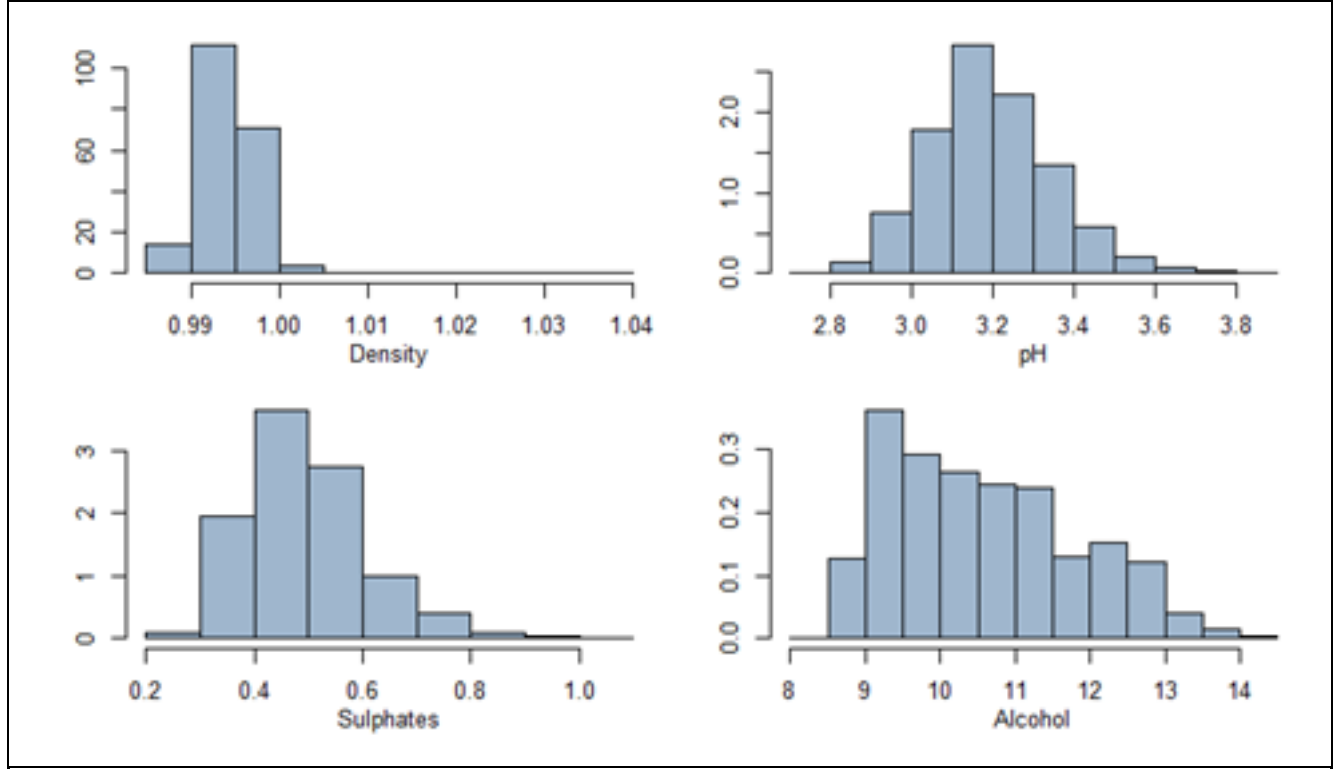
All variables are summarized and univariate analysis with plots are shown below.

Sample R code for EDA

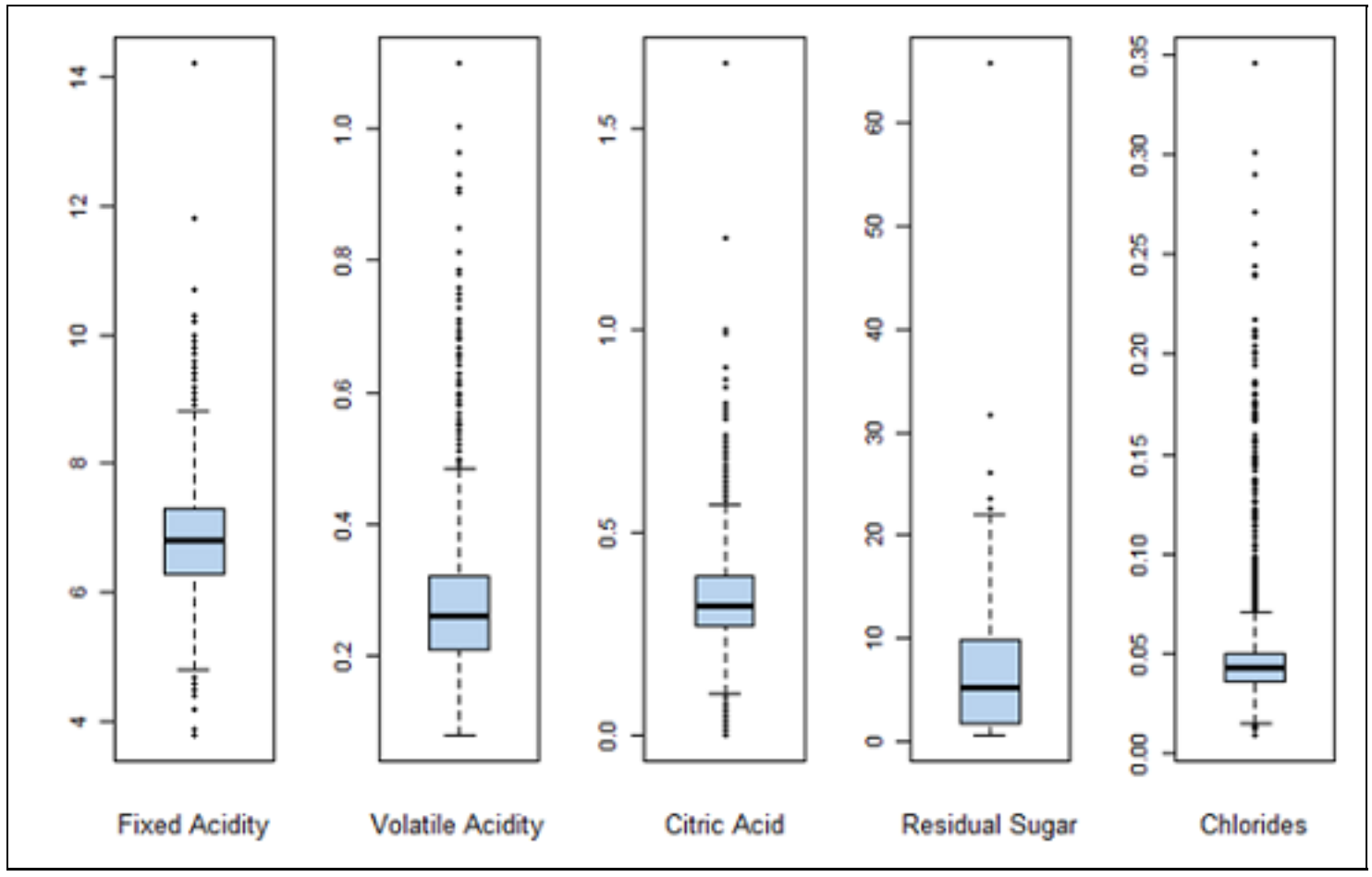
Histograms to show the distribution of the variable values:

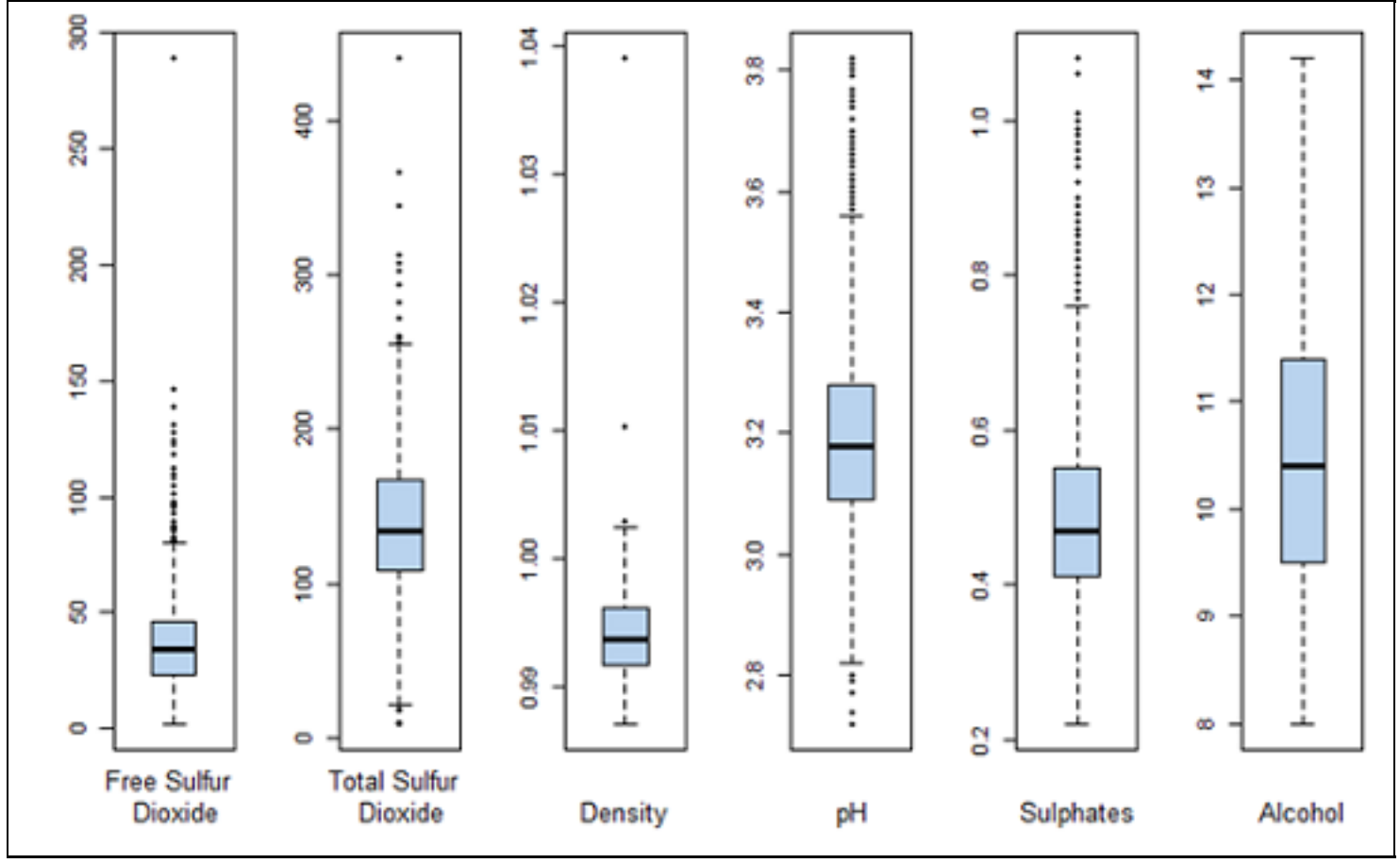






Boxplots for each of the variables as another indicator of spread.





Observations regarding variables: All variables have outliers

- Quality has most values concentrated in the categories 5, 6 and 7. Only a small proportion is in the categories [3, 4] and [8, 9] and none in the categories [1, 2] and 10.
- 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; even after eliminating the outliers distribution will remain skewed.
- Some of the variables, e.g . free sulphur dioxide, density, have a few outliers but these are very different from the rest.
- Mostly outliers are on the larger side.
- Alcohol has an irregular shaped distribution but it does not have pronounced outliers.

Sample R code for Summary Statistics & Correlations

These observations are supported by the summary statistics also, as shown in the following table:

	Minimum	Q1	Median	Q3	Maximum	Range	IQR	MAD	Mean	StDev	StErr
fixed.acidity	3.80	6.30	6.80	7.30	14.20	10.40	1.00	0.74	6.85	0.84	0.012
volatile.acidity	0.08	0.21	0.26	0.32	1.10	1.02	0.11	0.09	0.28	0.10	0.001
citric.acid	0.00	0.27	0.32	0.39	1.66	1.66	0.12	0.09	0.33	0.12	0.002
residual.sugar	0.60	1.70	5.20	9.90	65.80	65.20	8.20	5.34	6.39	5.07	0.072
chlorides	0.01	0.04	0.04	0.05	0.35	0.34	0.01	0.01	0.05	0.02	0.000
free.sulfur.dioxide	2.00	23.00	34.00	46.00	289.00	287.00	23.00	16.31	35.31	17.01	0.243
total.sulfur.dioxide	9.00	108.00	134.00	167.00	440.00	431.00	59.00	43.00	138.36	42.50	0.607
density	0.99	0.99	0.99	1.00	1.04	0.05	0.00	0.00	0.99	0.00	0.000
pH	2.72	3.09	3.18	3.28	3.82	1.10	0.19	0.15	3.19	0.15	0.002
sulphates	0.22	0.41	0.47	0.55	1.08	0.86	0.14	0.10	0.49	0.11	0.002
alcohol	8.00	9.50	10.40	11.40	14.20	6.20	1.90	1.48	10.51	1.23	0.018

Range is much larger compared to the IQR. Mean is usually greater than the median. These observations indicate that there are outliers in the data set and before any analysis is performed outliers must be taken care of.

Next we look at the bivariate analysis, including all pairwise scatterplot and correlation coefficients. Since the variables have non-normal distribution, we have considered both person and spearman rank correlations.

Table: Pearson's Correlation

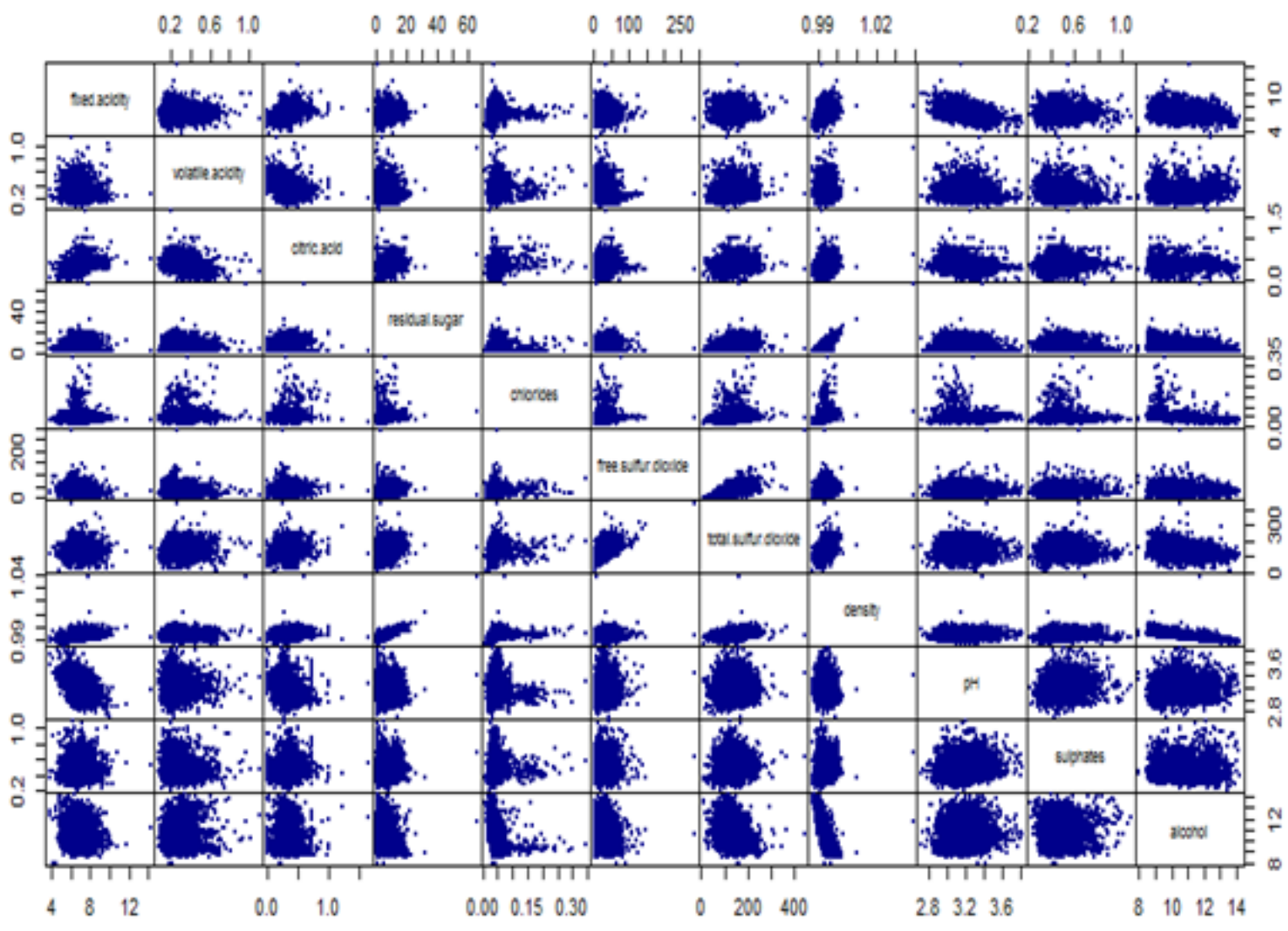
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
fixed.acidity	1.00	-0.02	0.29	0.09	0.02	-0.05	0.09	0.27	-0.43	-0.02	-0.12
volatile.acidity	-0.02	1.00	-0.15	0.06	0.07	-0.10	0.09	0.03	-0.03	-0.04	0.07
citric.acid	0.29	-0.15	1.00	0.09	0.11	0.09	0.12	0.15	-0.16	0.06	-0.08
residual.sugar	0.09	0.06	0.09	1.00	0.09	0.30	0.40	0.84	-0.19	-0.03	-0.45
chlorides	0.02	0.07	0.11	0.09	1.00	0.10	0.20	0.26	-0.09	0.02	-0.36
free.sulfur.dioxide	-0.05	-0.10	0.09	0.30	0.10	1.00	0.62	0.29	0.00	0.06	-0.25
total.sulfur.dioxide	0.09	0.09	0.12	0.40	0.20	0.62	1.00	0.53	0.00	0.13	-0.45
density	0.27	0.03	0.15	0.84	0.26	0.29	0.53	1.00	-0.09	0.07	-0.78
pH	-0.43	-0.03	-0.16	-0.19	-0.09	0.00	0.00	-0.09	1.00	0.16	0.12
sulphates	-0.02	-0.04	0.06	-0.03	0.02	0.06	0.13	0.07	0.16	1.00	-0.02
alcohol	-0.12	0.07	-0.08	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.02	1.00

Table: Spearman Rank Correlation

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
fixed.acidity	1.00	-0.04	0.30	0.11	0.09	-0.02	0.11	0.27	-0.42	-0.01	-0.11
volatile.acidity	-0.04	1.00	-0.15	0.11	0.00	-0.08	0.12	0.01	-0.05	-0.02	0.03
citric.acid	0.30	-0.15	1.00	0.02	0.03	0.09	0.09	0.09	-0.15	0.08	-0.03
residual.sugar	0.11	0.11	0.02	1.00	0.23	0.35	0.43	0.78	-0.18	0.00	-0.45
chlorides	0.09	0.00	0.03	0.23	1.00	0.17	0.38	0.51	-0.05	0.09	-0.57
free.sulfur.dioxide	-0.02	-0.08	0.09	0.35	0.17	1.00	0.62	0.33	-0.01	0.05	-0.27
total.sulfur.dioxide	0.11	0.12	0.09	0.43	0.38	0.62	1.00	0.56	-0.01	0.16	-0.48
density	0.27	0.01	0.09	0.78	0.51	0.33	0.56	1.00	-0.11	0.10	-0.82
pH	-0.42	-0.05	-0.15	-0.18	-0.05	-0.01	-0.01	-0.11	1.00	0.14	0.15
sulphates	-0.01	-0.02	0.08	0.00	0.09	0.05	0.16	0.10	0.14	1.00	-0.04
alcohol	-0.11	0.03	-0.03	-0.45	-0.57	-0.27	-0.48	-0.82	0.15	-0.04	1.00

Pearson's correlation and rank correlations are very close, hence only the former is considered. High correlations (≥ 40% in absolute value) are identified and marked in red. Pairwise scatterplots are also shown below.

Scatterplot of Predictors



Sample R code for Preparing Data

Data Preparation

Possibly the most important step in data preparation is to identify outliers. Since this is a multivariate data, we consider only those points which do not have any predictor variable value to be outside of limits constructed by boxplots. The following rule is applied:

- A predictor value is considered to be an outlier only if it is greater than $Q_3 + 1.5IQR$

The rationale behind this rule is that the extreme outliers are all on the higher end of the values and the distributions are all positively skewed. Application of this rule reduces the data size from 4899 to 4074.

Data is randomly divided into Training data and Test Data of equal sizes (50% each).

WQD.2 - Multiple Regression

Sample R code for Multiple Regression

Linear regression is fitted to the Training data.

Model I: All predictors in the model

Regression Coefficients	Estimate	Std. Error	t	Pr(>|t|)	VIF
(Intercept)	166.40	38.60	4.31	0.00	
fixed.acidity	0.13	0.04	3.38	0.00	3.15
volatile.acidity	-1.85	0.22	-8.27	0.00	1.12
citric.acid	0.05	0.18	0.30	0.76	1.12
residual.sugar	0.08	0.01	5.59	0.00	19.06
chlorides	-3.63	2.05	-1.77	0.08	1.59
free.sulfur.dioxide	0.00	0.00	2.88	0.00	1.87
total.sulfur.dioxide	0.00	0.00	0.40	0.69	2.52
density	-167.20	39.12	-4.27	0.00	47.87
pH	0.85	0.18	4.68	0.00	2.41
sulphates	0.81	0.18	4.64	0.00	1.14
alcohol	0.16	0.05	3.24	0.00	13.03

For extremely high VIF density was removed from the model. There are other predictors with high VIF, but they were not removed at this step.

Model II: After removal of density VIFs improved

Regression Coefficients	Estimate	Std. Error	t	Pr(>|t|)	VIF
(Intercept)	1.47	0.56	2.61	0.01	
fixed.acidity	0.00	0.02	0.14	0.89	1.28
volatile.acidity	-1.91	0.22	-8.54	0.00	1.12
citric.acid	-0.01	0.18	-0.03	0.97	1.11
residual.sugar	0.02	0.00	5.29	0.00	1.49
chlorides	-5.47	2.01	-2.72	0.01	1.52
free.sulfur.dioxide	0.01	0.00	3.63	0.00	1.82
total.sulfur.dioxide	0.00	0.00	-0.79	0.43	2.34
pH	0.31	0.13	2.38	0.02	1.27
sulphates	0.60	0.17	3.58	0.00	1.05
alcohol	0.34	0.02	18.32	0.00	1.98

Not all predictors are significant. A forward selection method is employed to build a working model. The sample R output follows:

```
Start:  AIC=-680.65
quality ~ 1

                      Df Sum of Sq    RSS     AIC
+ alcohol              1   283.176 1173.8 -1118.89
+ chlorides            1   130.970 1326.0  -870.52
+ total.sulfur.dioxide 1    48.673 1408.3  -747.86
+ residual.sugar       1    25.581 1431.4  -714.73
+ volatile.acidity     1    21.608 1435.3  -709.09
+ pH                   1    13.037 1443.9  -696.96
+ fixed.acidity        1     4.089 1452.9  -684.38
<none>                             1457.0  -680.65
+ citric.acid          1     1.238 1455.7  -680.38
+ sulphates            1     1.002 1456.0  -680.05
+ free.sulfur.dioxide  1     0.240 1456.7  -678.99
```

```
Step:  AIC=-1118.89
quality ~ alcohol
```

	Df	Sum of Sq	RSS	AIC
+ volatile.acidity	1	40.585	1133.2	-1188.6
+ free.sulfur.dioxide	1	17.564	1156.2	-1147.6
+ residual.sugar	1	11.564	1162.2	-1137.0
+ sulphates	1	5.907	1167.9	-1127.2
+ chlorides	1	5.772	1168.0	-1126.9
+ pH	1	2.837	1171.0	-1121.8
+ total.sulfur.dioxide	1	1.706	1172.1	-1119.8
+ citric.acid	1	1.421	1172.4	-1119.3
<none>			1173.8	-1118.9
+ fixed.acidity	1	0.244	1173.5	-1117.3

```
Step:  AIC=-1188.56
quality ~ alcohol + volatile.acidity
```

	Df	Sum of Sq	RSS	AIC
+ residual.sugar	1	18.8659	1114.3	-1220.8
+ free.sulfur.dioxide	1	16.5614	1116.6	-1216.5
+ sulphates	1	5.7863	1127.4	-1197.0
+ total.sulfur.dioxide	1	5.6644	1127.5	-1196.8
+ chlorides	1	4.1606	1129.0	-1194.1
+ pH	1	2.3397	1130.9	-1190.8
<none>			1133.2	-1188.6
+ fixed.acidity	1	0.8165	1132.4	-1188.0
+ citric.acid	1	0.0908	1133.1	-1186.7

```
Step:  AIC=-1220.76
quality ~ alcohol + volatile.acidity + residual.sugar
```

	Df	Sum of Sq	RSS	AIC
+ free.sulfur.dioxide	1	9.2408	1105.1	-1235.7
+ sulphates	1	7.9372	1106.4	-1233.3
+ pH	1	4.9199	1109.4	-1227.8
+ chlorides	1	3.8495	1110.5	-1225.8
+ total.sulfur.dioxide	1	2.0808	1112.2	-1222.6
<none>			1114.3	-1220.8
+ fixed.acidity	1	1.0437	1113.3	-1220.7
+ citric.acid	1	0.0002	1114.3	-1218.8

```
Step:  AIC=-1235.72
quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide
```

	Df	Sum of Sq	RSS	AIC
+ sulphates	1	7.4161	1097.7	-1247.4
+ pH	1	4.2500	1100.8	-1241.6
+ chlorides	1	4.0537	1101.0	-1241.2
<none>			1105.1	-1235.7
+ fixed.acidity	1	0.6325	1104.5	-1234.9
+ total.sulfur.dioxide	1	0.1885	1104.9	-1234.1
+ citric.acid	1	0.0467	1105.0	-1233.8

```
Step:  AIC=-1247.44
quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
    sulphates
```

	Df	Sum of Sq	RSS	AIC
+ chlorides	1	4.3656	1093.3	-1253.6
+ pH	1	3.1849	1094.5	-1251.4
<none>			1097.7	-1247.4
+ total.sulfur.dioxide	1	0.6983	1097.0	-1246.7
+ fixed.acidity	1	0.6129	1097.1	-1246.6
+ citric.acid	1	0.1387	1097.5	-1245.7

```
Step:  AIC=-1253.56
quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
    sulphates + chlorides

                      Df Sum of Sq    RSS     AIC
+ pH                   1    3.3650 1090.0 -1257.8
<none>                             1093.3 -1253.6
+ fixed.acidity        1    0.4827 1092.8 -1252.5
+ total.sulfur.dioxide 1    0.2152 1093.1 -1252.0
+ citric.acid          1    0.0848 1093.2 -1251.7

Step:  AIC=-1257.84
quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide + sulphates
+ chlorides + pH

                      Df Sum of Sq    RSS     AIC
<none>                             1090.0 -1257.8
+ total.sulfur.dioxide 1   0.33456 1089.6 -1256.5
+ citric.acid          1   0.00436 1089.9 -1255.8
+ fixed.acidity        1   0.00167 1089.9 -1255.8
```

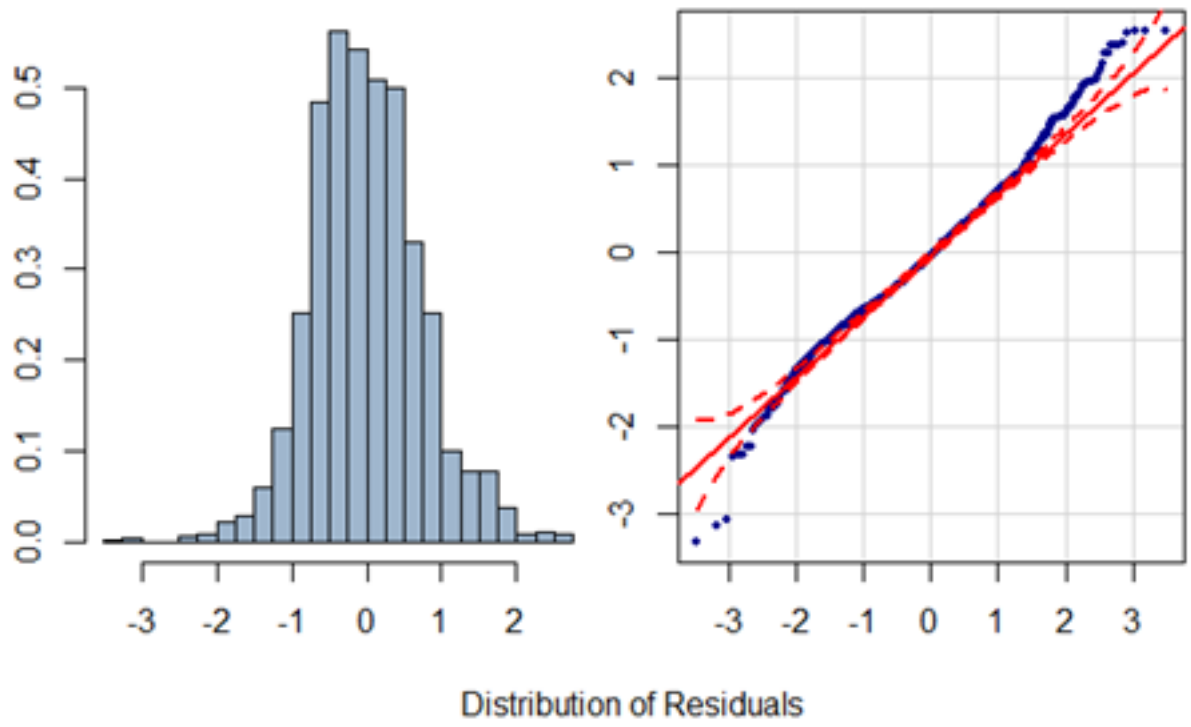
Model III: Working model

Regression Coefficients	Estimate	Std. Error	t	Pr(>|t|)
(Intercept)	1.49	0.46	3.27	0.00
alcohol	0.35	0.02	19.20	0.00
volatile.acidity	-1.95	0.22	-9.03	0.00
residual.sugar	0.02	0.00	5.24	0.00
free.sulfur.dioxide	0.005	0.001	3.95	0.00
sulphates	0.59	0.17	3.51	0.00
chlorides	-5.74	1.97	-2.91	0.00
pH	0.30	0.12	2.50	0.01

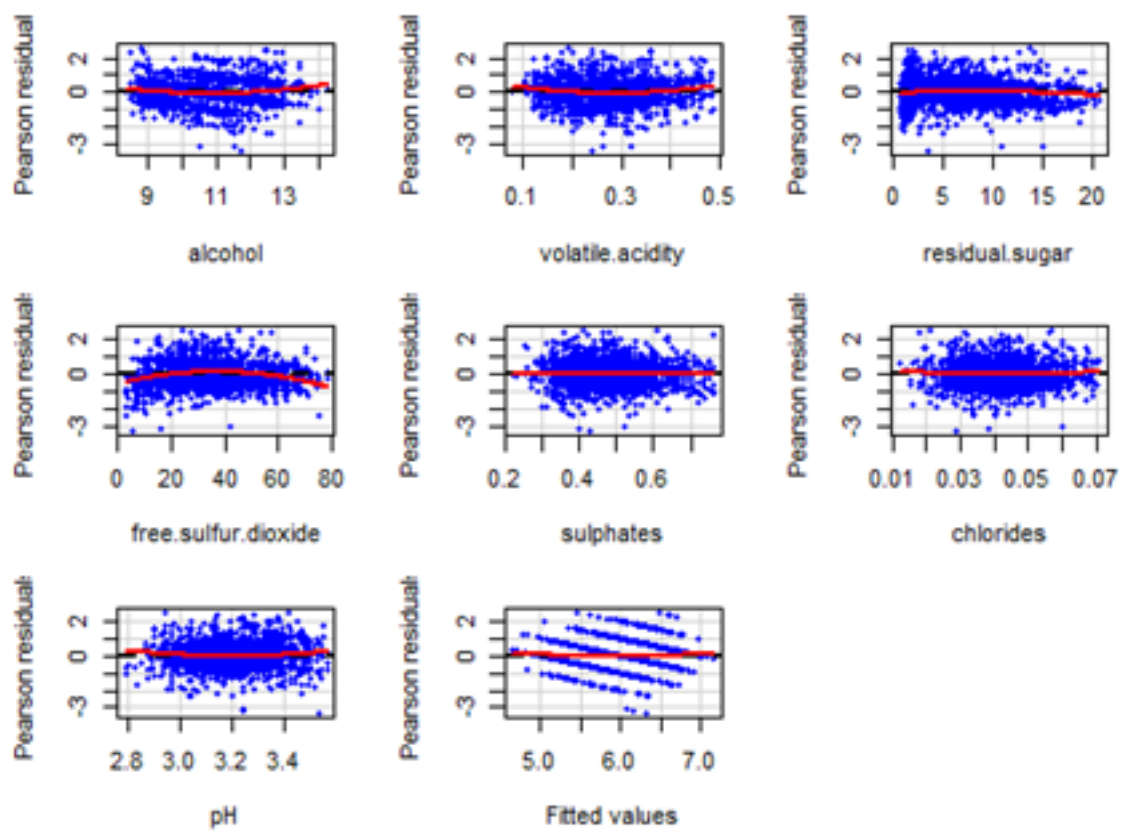
Sample R output:

```
Residual standard error: 0.7329 on 2029 degrees of freedom
Multiple R-squared:  0.2519,   Adjusted R-squared:  0.2493
F-statistic:  97.6 on 7 and 2029 DF,  p-value: < 2.2e-16
```

Note that multiple R^2 is 25%. Regression diagnostics are examined for possible improvement of the model.



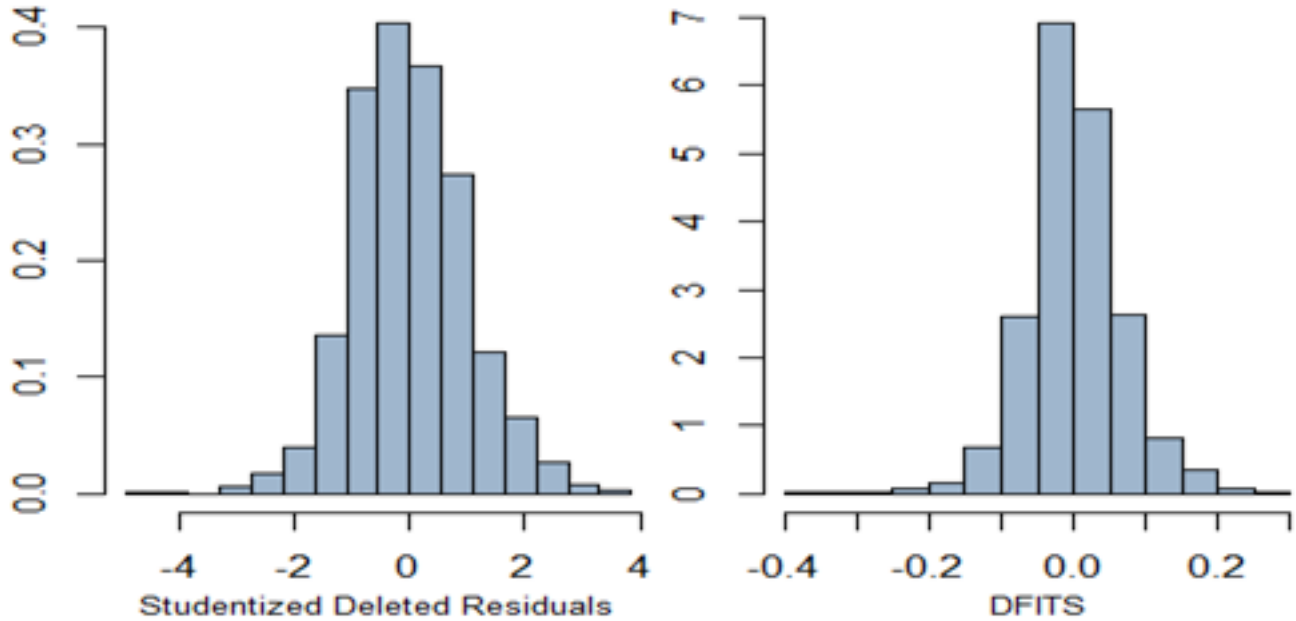
Residuals have an approximately symmetric distribution but there seems to be outliers at both ends. Partial residual plots are given below. Note the pattern in the fitted value plot. Since the response actually takes only integer values but has been assumed to be continuous, such pattern arises.

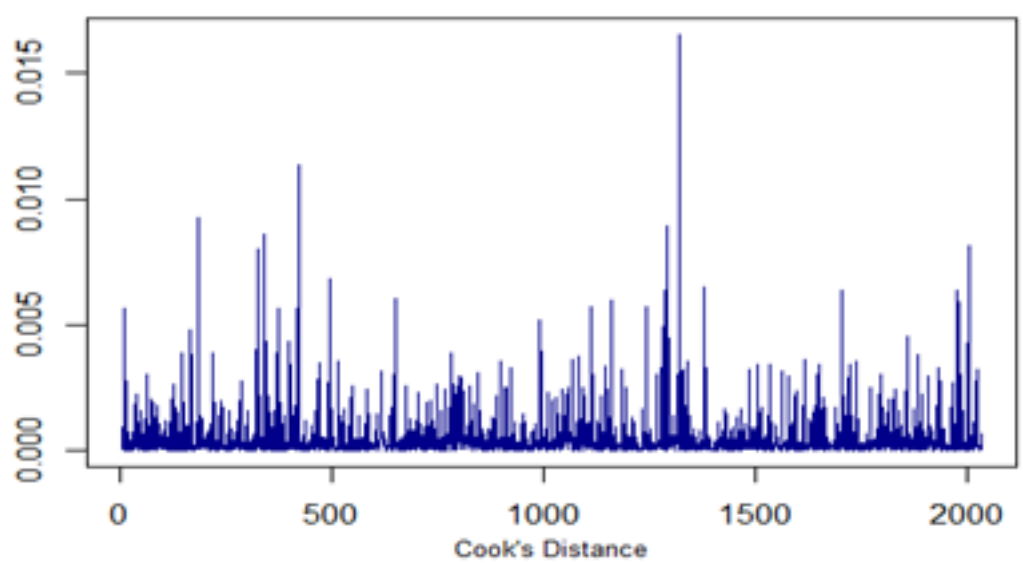


Outliers and leverage points are identified through the following:

- Studentized deleted residuals (a point is outlier if residual is outside of [-3, 3] limits
- DFITS (a point is outlier if residual is outside of [-1, 1] limits
- Cook's distance

All three plots are given below. Note that no point is identified as outlier with DFITS value.





Only 26 points are identified as outliers according to the above criteria. A final model is fit after eliminating these points and a slight improvement in the R^2 value is noted.

Model IV: Final model

Regression Coefficients	Estimate	Std. Error	t	Pr(>|t|)
(Intercept)	1.41	0.43	3.25	0.00
alcohol	0.35	0.02	20.42	0.00
volatile.acidity	-1.99	0.20	-9.72	0.00
residual.sugar	0.02	0.00	5.58	0.00
free.sulfur.dioxide	0.004	0.001	3.21	0.00
sulphates	0.56	0.16	3.57	0.00
chlorides	-5.79	1.87	-3.10	0.00
pH	0.34	0.11	2.94	0.00

Sample R output:

```
Residual standard error: 0.6884 on 2003 degrees of freedom
Multiple R-squared:  0.2809,   Adjusted R-squared:  0.2784
F-statistic: 111.8 on 7 and 2003 DF,  p-value: < 2.2e-16
```

Application of this model on test data gives sum of square of differences between the actual response

and predicted response to be 1196.205 whereas sum of square of deviations of actual response is 1554.754. Ratio of these two may be taken as the ratio of Error sum of squares and total sum of squares. Hence a measure similar to that of R^2 may be computed as 1 – 1196.205/1554.754 = 0.2306.

Sample R code for Final Model

WQD.3 - Application of Polynomial Regression

In order to investigate whether a polynomial relationship fits the model better, an alternative model with squared terms of the significant variables is tried, which improves R^2 value to 31%.

Regression Coefficients	Estimate	Std. Error	t	Pr(>|t|)
(Intercept)	5.78	0.11	51.91	0.00
poly(alcohol, 2)1	18.03	0.92	19.52	0.00
poly(alcohol, 2)2	2.01	0.72	2.78	0.00
poly(volatile.acidity, 2)1	-6.52	0.70	-9.31	0.00
poly(volatile.acidity, 2)2	2.70	0.68	3.96	0.00
residual.sugar	0.02	0.00	4.57	0.00
poly(free.sulfur.dioxide, 2)1	2.57	0.73	3.52	0.00
poly(free.sulfur.dioxide, 2)2	-0.24	0.68	-7.70	0.00
chlorides	-0.64	1.83	-3.63	0.00
sulphates	0.65	0.16	4.21	0.00
poly(pH, 2)1	1.75	0.70	2.52	0.00
poly(pH, 2)2	2.39	0.68	3.50	0.00

Sample R output:

```
Residual standard error: 0.6727 on 1999 degrees of freedom
Multiple R-squared:  0.3146,   Adjusted R-squared:  0.3108
F-statistic: 83.42 on 11 and 1999 DF,  p-value: < 2.2e-16
```

Application of this model on test data gives sum of square of differences between the actual response and predicted response to be 1139.41 whereas sum of square of deviations of actual response is 1554.754. Ratio of these two may be taken as the ratio of Error sum of squares and total sum of squares. Hence a measure similar to that of R^2 may be computed as 1 – 1139.41/1554.754 = 0.2671.

WQD.4 - Applying Tree-Based Methods

Sample R code for Tree-based Models and Random Forest

The response variable quality is assumed to be an ordinal variable, not a continuous variable. It has been noted before that proportions in too low (4 or less) or too high (8 or above) categories are small.

Quality Category	3	4	5	6	7	8	9
Proportion	0.4%	3.3%	29.7%	44.9%	18.0%	3.6%	0.1%

Hence wines are classified into three categories by combining 3, 4, and 5 into one category (Low), 6 (Medium) and 7, 8 and 9 into another (High).

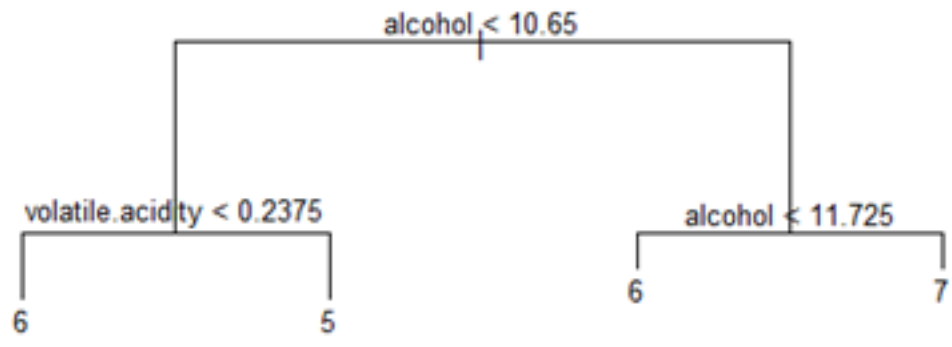
The following regression tree is obtained:

```
n = 2037

Classification tree:
tree(formula = FactQ ~ ., method = "class")
Variables actually used in tree construction:
[1] "alcohol"          "volatile.acidity"
Number of terminal nodes:  4
Residual mean deviance:  1.811 = 3681 / 2033
Misclassification error rate: 0.4502 = 917 / 2037

node), split, n, deviance, yval, (yprob)
      * denotes terminal node

1) root 2037 4301.0 6 ( 0.31026 0.46343 0.22631 )
  2) alcohol < 10.65 1107 2064.0 5 ( 0.46161 0.44896 0.08943 )
    4) volatile.acidity < 0.2375 396  764.0 6 ( 0.24495 0.58081 0.17424 ) *
    5) volatile.acidity > 0.2375 711 1161.0 5 ( 0.58228 0.37553 0.04219 ) *
  3) alcohol > 10.65 930 1832.0 6 ( 0.13011 0.48065 0.38925 )
    6) alcohol < 11.725 511 1044.0 6 ( 0.19765 0.51272 0.28963 ) *
    7) alcohol > 11.725 419  711.7 7 ( 0.04773 0.44153 0.51074 ) *
```

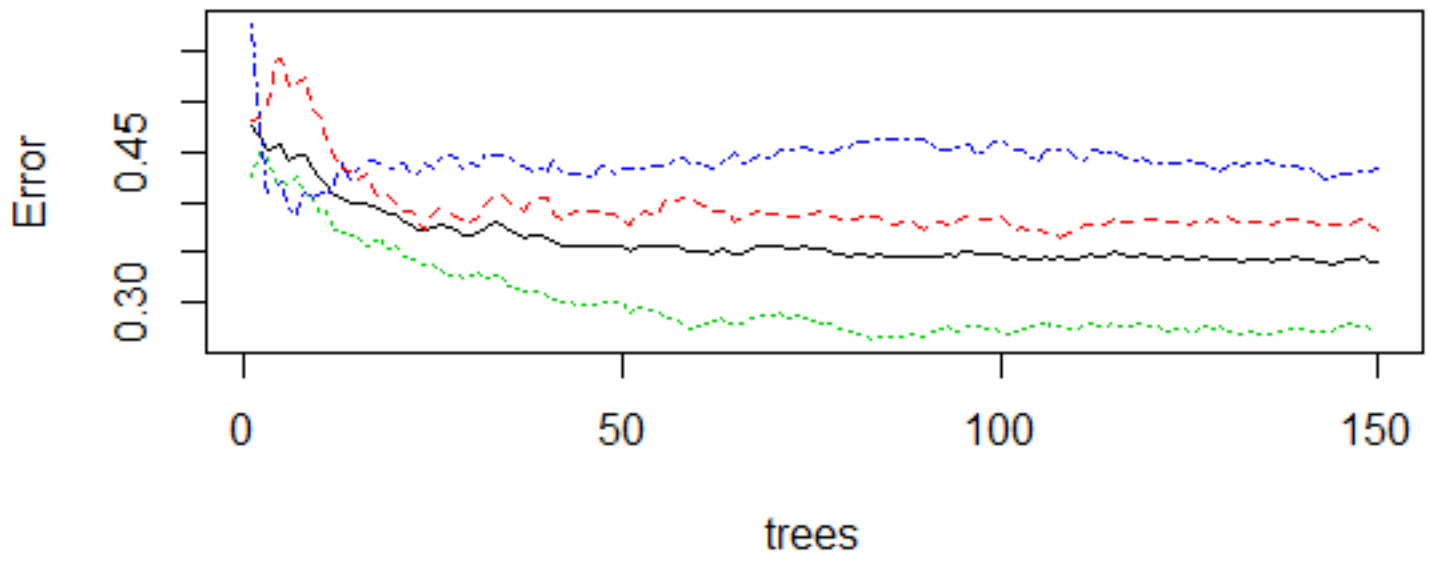


Applying the procedure on Test data, the following mis-classification table is obtained:

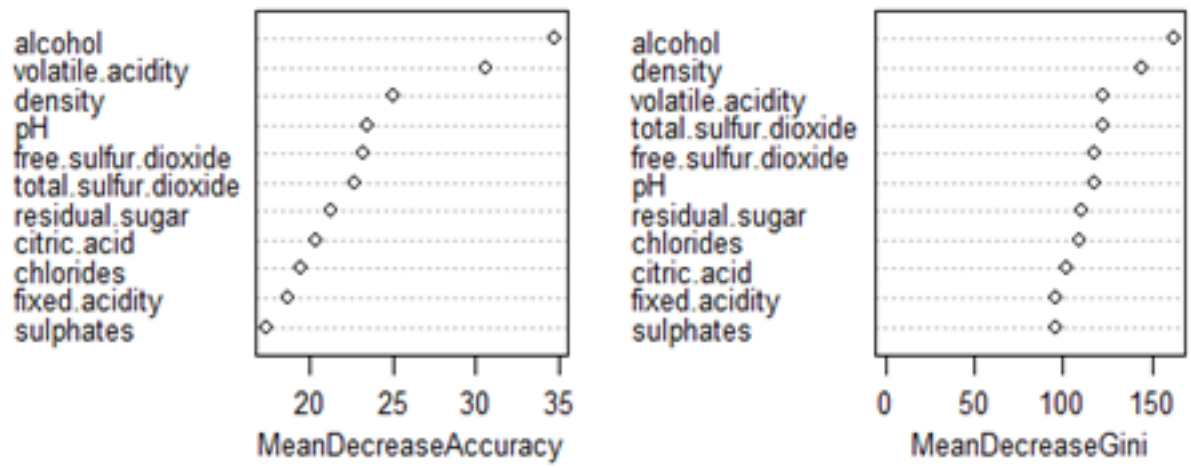
	Quality Classification		
Test Data	**Low**	**Medium**	**High**
Low	371	277	38
Medium	214	495	251
High	19	167	205
Accuracy	(371 + 495 + 205) / 2037 = 50%		

WQD.5 - Random Forest

Completely unsupervised random forest method on Training data with *ntree* = 150 leads to the following error plot:



Importance of predictors are given in the following dotplot:



Accuracy improves from 50% to 67.7%.

	Quality Classification		
Test Data	**Low**	**Medium**	**High**
Low	378	277	17
Medium	219	726	201
High	7	83	276
Accuracy	(378 + 726 + 276) / 2037 = 67.7%		

WQD.6 - Classification

Sample R code for Classification

Nearest neighbor classifier is used with three levels (Low, Medium, High) of quality. It turned out that for $k = 5$, test data misclassification rate is lowest, when all predictors are being used.

	Quality Classification		
Test Data	**Low**	**Medium**	**High**
Low	293	185	46
Medium	267	595	215
High	44	159	233
Accuracy	(293 + 595 + 233) / 2037 = 55%		

WQD.7 - Conclusion

It does not look like wine quality is well supported by its chemical properties. At each quality level variability of the predictors is high and the groups are not well separated.

Food for thought!

> Think of some visualization techniques that may help in bringing out these features in the data.

Source URL: https://onlinecourses.science.psu.edu/stat857/node/223

Links:
[1] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/Wine_data.xlsx
[2] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/Training50_winedata.csv
[3] https://onlinecourses.science.psu.edu/stat857/sites/onlinecourses.science.psu.edu.stat857/files/Test50_winedata.csv