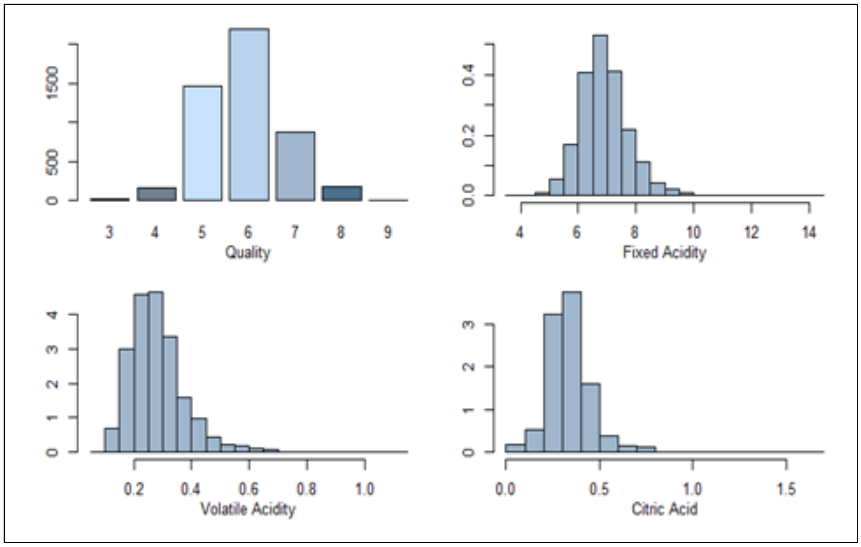
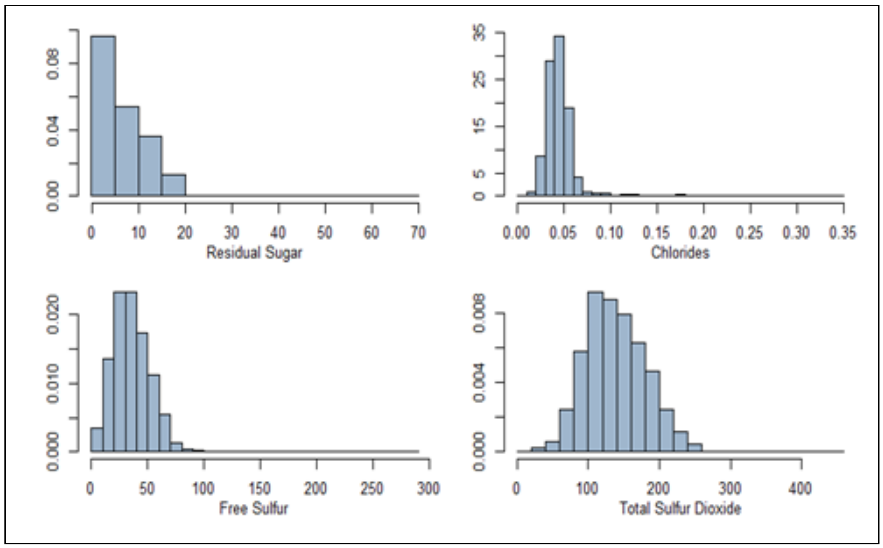
WQD.1 - Exploratory Data Analysis (EDA) and Data Pre-processing

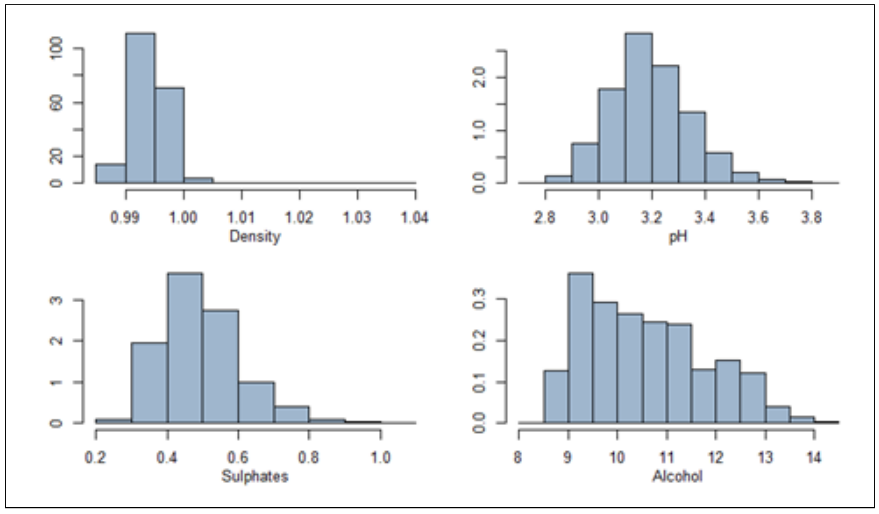
All variables are summarized and univariate analysis with plots are shown below.

Sample R code for EDA

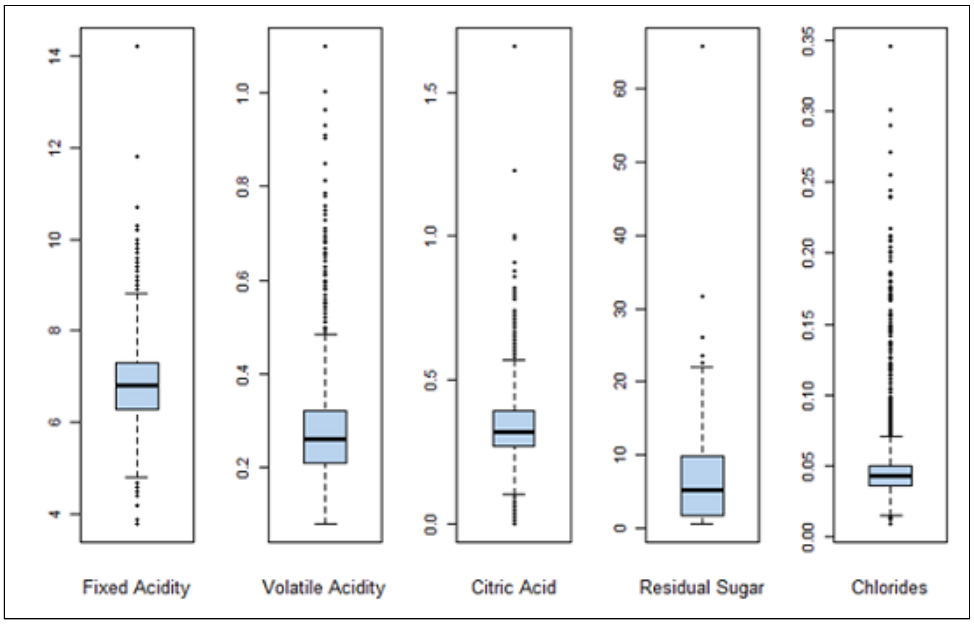
Histograms to show the distribution of the variable values:

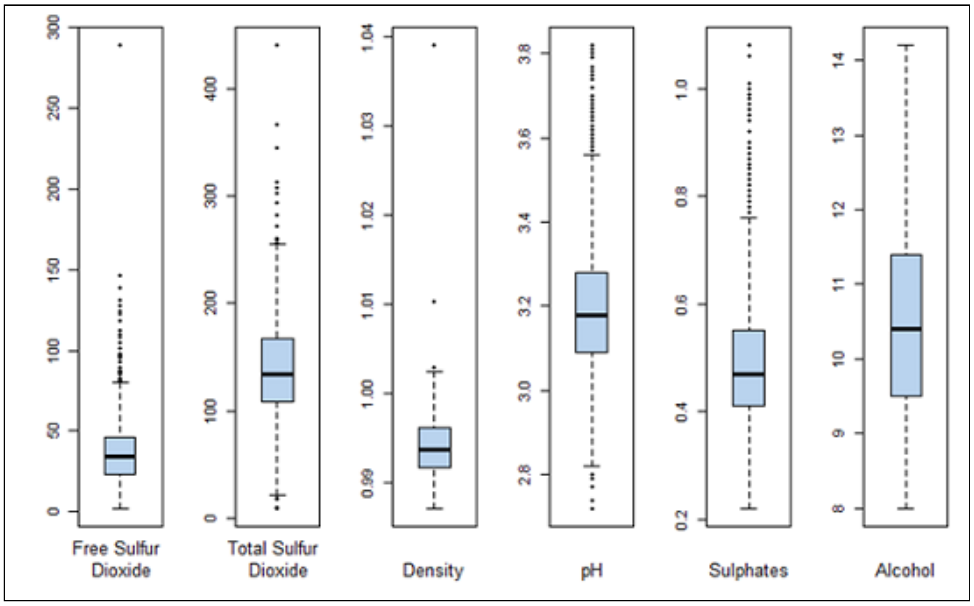






Boxplots for each of the variables as another indicator of spread.





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.

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

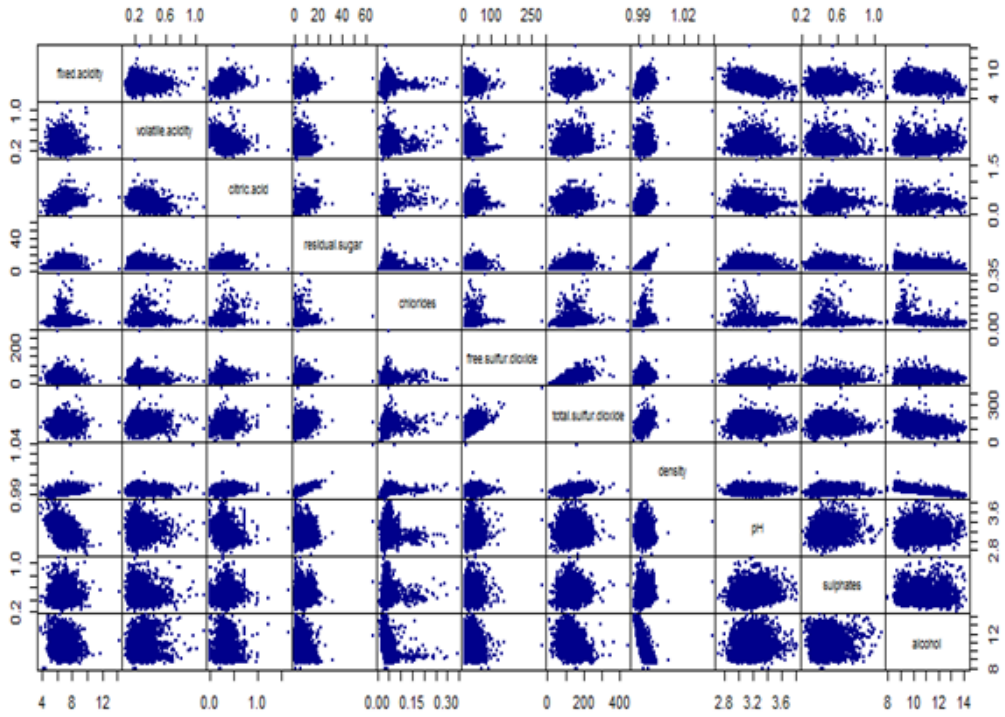
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
fixed acidity	1.00	-0.02	0.29	0.09	0.02	-0.05	0.09	0.27	-0.43	-0.02	-0.12
volatile acidity	-0.02	1.00	-0.15	0.06	0.07	-0.10	0.09	0.03	-0.03	-0.04	0.07
citric acid	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).
