1. **Description of the Data Set**

**Problem Statement**

Can we predict the quality of wine based on its chemical properties?

**Overview**

Data on 12 different properties of the wines are collected. One of the properties is Quality, and the rest are chemical properties of the wines. The dataset contains a total of **4898 observations** and **12 variables**.

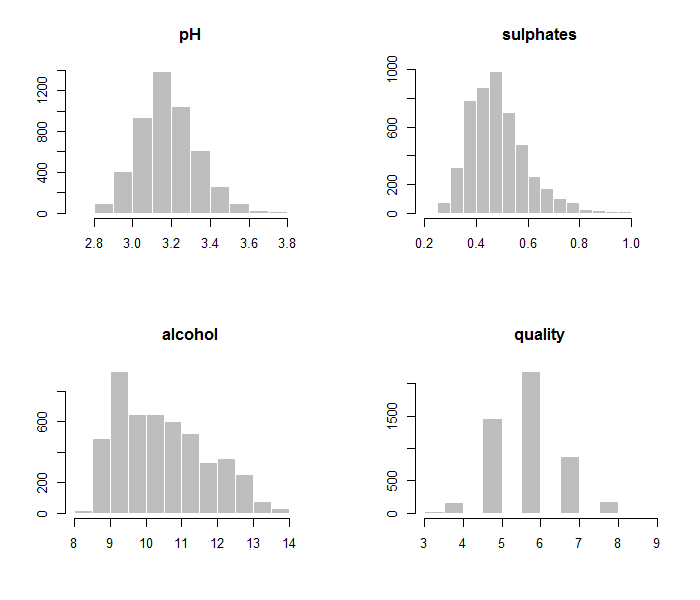
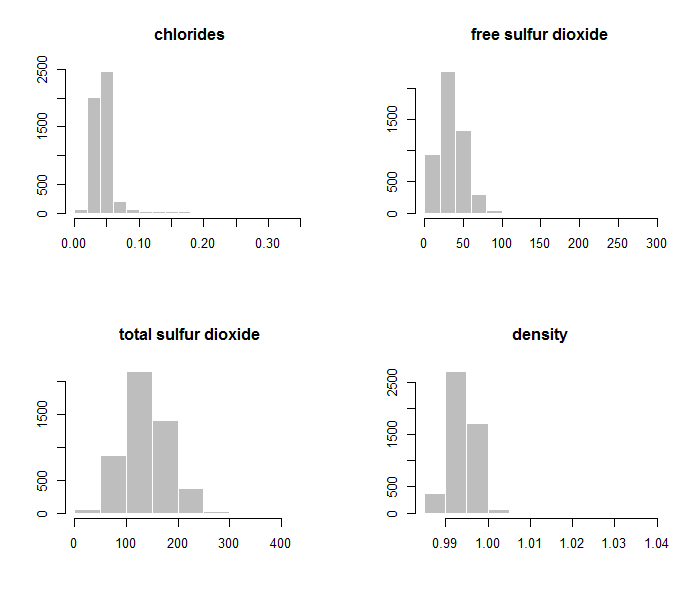
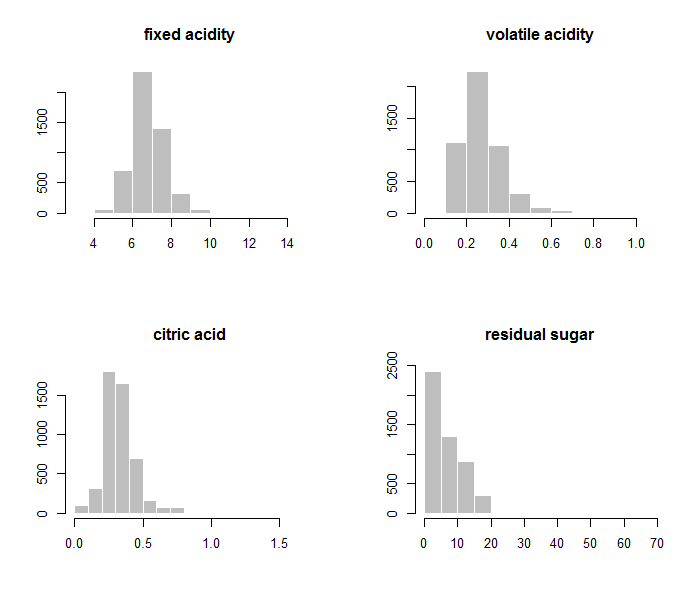
Of the 12 variables:

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulphur dioxide
7. Total sulphur dioxide
8. Density
9. pH
10. Sulphates
11. Alcohol
12. Quality

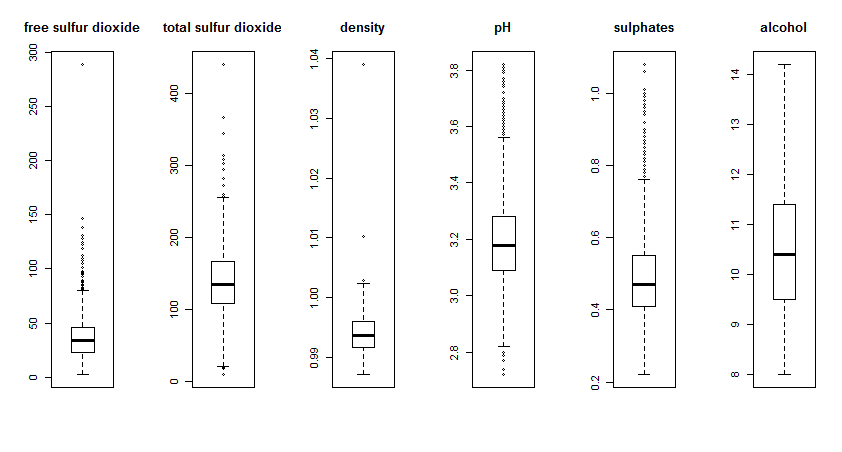
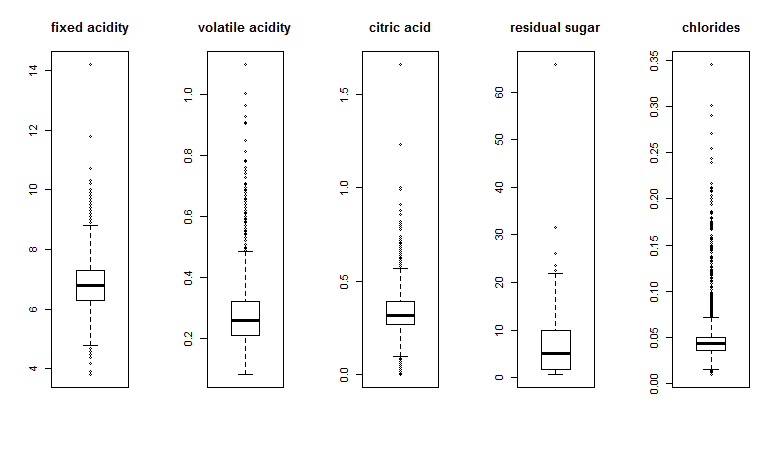
**B. What possible insights can be obtained from mining the chosen dataset?**

Pattern of which chemical properties of wine will affect its quality might be mined.

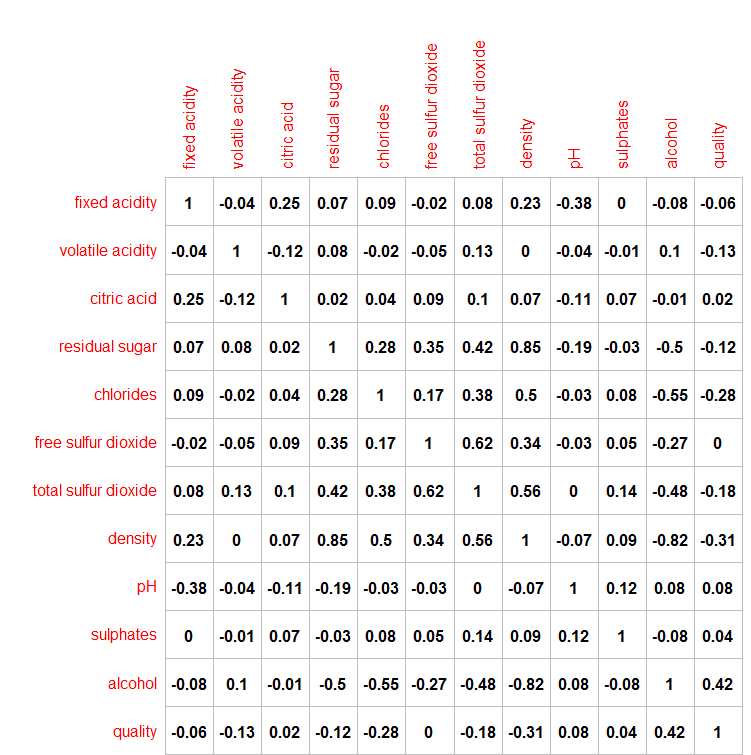
Below is the distribution of each properties:



Boxplots for each of the variables are shown below:



Correlations between variables:



Based on the above observations regarding each variables, we can observe that:

1. Quality are mostly concentrated in categories of 5 to 7. Small amount is observed in categories 3 to 4 and 8 to 9 and none is observed in categories 0 to 3 and 10.
2. All variables have outliers.
3. Residual sugar’s distribution is skewed left, it remains skewed left after removing outliers.
4. Alcohol percentage and quality of wine has a correlation of 0.42 but it is not high enough to be used as quality prediction for quality of wine.

**C. What type of data mining technique (association rule mining, classification or clustering) would be relevant? Give an example, for example, if you think classification is suitable, describe what will be classified and what the possible classes are.**

Classification can be used. Quality of the wine will be classified as Low (3, 4 and 5), Medium (6) and High (7, 8 and 9). Based on the classification, we might find out if there is any correlation between quality and chemical properties of the wine.

**D. Describe data quality issues, and be specific. Identify which attribute (column) has issues, or if the structure of the data has problems.**

The data quality issue in this dataset is outliers. All variables other than quality has outliers.

**E. Perform a pre-processing task on the dataset chosen.**#create new data frame without outliers

wineNoOut <- as.data.frame(sapply(wine,remove\_outliers))

winenNoOut <- na.omit(wineNoOut)

#create new data frame without outliers and with new category Low, Medium and High

winenNoOutCategory["category"]<-NA

winenNoOutCategory$category <- ifelse(winenNoOutCategory$quality== 3, "Low",

ifelse(winenNoOutCategory$quality== 4, "Low",

ifelse(winenNoOutCategory$quality== 5, "Low",

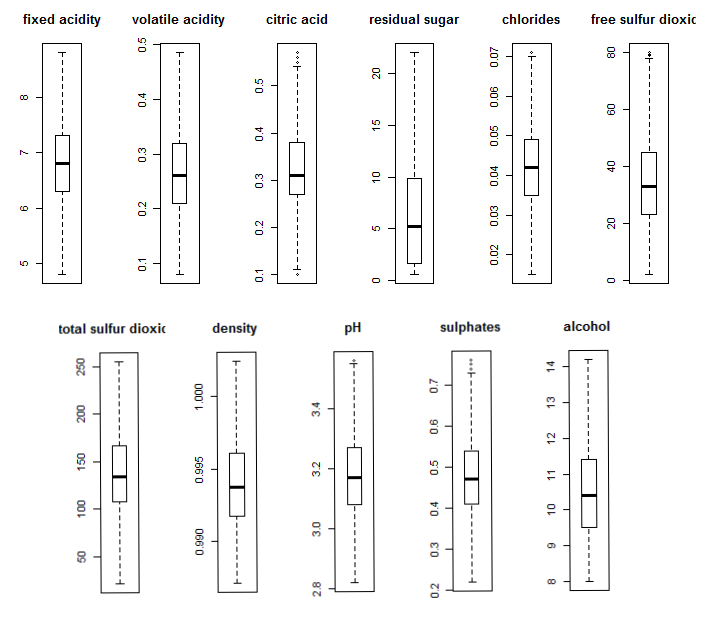
ifelse(winenNoOutCategory$quality==6, "Medium",

ifelse(winenNoOutCategory$quality==7, "High",

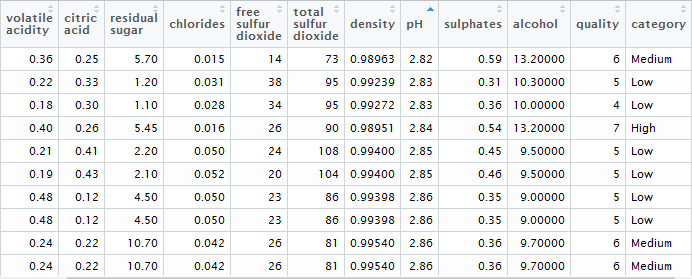
ifelse(winenNoOutCategory$quality== 8, "High",

ifelse(winenNoOutCategory$quality== 9, "High",NA

)))))))



Boxplot shows nearly no outliers.



New category added