EVALUATING WINE QUALITY VIA PHYSICOCHEMICAL TESTS

MATH 2319 Machine Learning Applied Project Phase I

Huynh Ai Loan (s3655461)

1. Introduction

The objective of this project is to build classifiers to predict whether physicochemical tests make the quality of wine larger than 5 grade in range of score between 0 (very bad) and 10 (very excellent) which are made by wine experts. The data sets were collected from the UCI Machine Learning Repository. There are two phases in this project. Phase 1 focuses on data preprocessing and exploration as described in this report. Phase 2 will be implemented later

2. Dataset

The dataset is used to evaluate Wine Quality via physicochemical tests. There are two read and white wine samples provided in this dataset. However, in this report, we only explore data for white wine sample.

The inputs consist of objective tests and the output is the median of at least 3 evaluations made by wine experts. The output of quality score is between 0 (very bad) and 10 (very excellent). Based on the range of score, grade 5 is considered to threshold of average wine quality to explore whether a wine gains 5 score in quality evaluation.

Descriptive Features:

The input attributes description is described in winequality.names.txt:

- Fixed acidity
- Volatile acidity
- Citric Acid
- Residual Sugar
- Chlorides
- Free Sulfur Dioxide
- Total sulfur dioxide
- Density
- pH: integer
- Sulphates
- Alcohol

Output variable is Quality

3. Data Pre-processing:

We use str(wine_white) to see the structure of dataset. Look at the result in figure 1, we see that data structure is defined with wrong data types for attributes. The 11 input variables should be numeric datatype, therefore we need to transform this dataset before jumping next steps.

```r
# Load dataset
wine_while <- read.csv2(file = "Dataset/winequality-white.csv", header = TRUE, sep = ";")
str(wine_while)
```

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : Factor w/ 68 levels "10","10.2","10.3",..: 38 30 50 41 41 50 29 38 30 50 ...
##  $ volatile.acidity    : Factor w/ 125 levels "0.08","0.085",..: 37 43 39 29 29 39 47 37 43 27 ...
##  $ citric.acid         : Factor w/ 87 levels "0","0.01","0.02",..: 37 35 41 33 33 41 17 37 35 44 ...
##  $ residual.sugar      : Factor w/ 310 levels "0.6","0.7","0.8",..: 190 18 258 286 286 258 261 190 18 16 ...
##  $ chlorides           : Factor w/ 160 levels "0.009","0.012",..: 35 39 40 48 48 40 35 35 39 34 ...
##  $ free.sulfur.dioxide : Factor w/ 132 levels "10","101","105",..: 66 17 41 68 68 41 41 66 17 36 ...
##  $ total.sulfur.dioxide: Factor w/ 251 levels "10","100","101",..: 76 36 249 94 94 249 40 76 36 32 ...
##  $ density             : Factor w/ 890 levels "0.98711","0.98713",..: 879 472 561 602 602 561 545 879 472 45
4 ...
##  $ pH                  : Factor w/ 103 levels "2.72","2.74",..: 24 54 50 43 43 50 42 24 54 46 ...
##  $ sulphates           : Factor w/ 79 levels "0.22","0.23",..: 23 27 22 18 18 22 25 23 27 23 ...
##  $ alcohol             : Factor w/ 104 levels "10","10.0333333333333",..: 88 95 3 104 104 3 98 88 95 22 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

Figure 1

Transform factor data type into numeric data

```r
# Transform Data: Convert datatype from Factor into Numeric:
indx <- sapply(wine_white, is.factor) # get Index of factor columns
wine_white[indx] <- lapply(wine_white[indx], function(x) as.numeric(as.character(x)))
str(wine_white)
```

```
## 'data.frame':    4898 obs. of  12 variables:
##  $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
##  $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
##  $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
##  $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
##  $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
##  $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
##  $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
##  $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
##  $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
##  $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
##  $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
##  $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

Figure 2

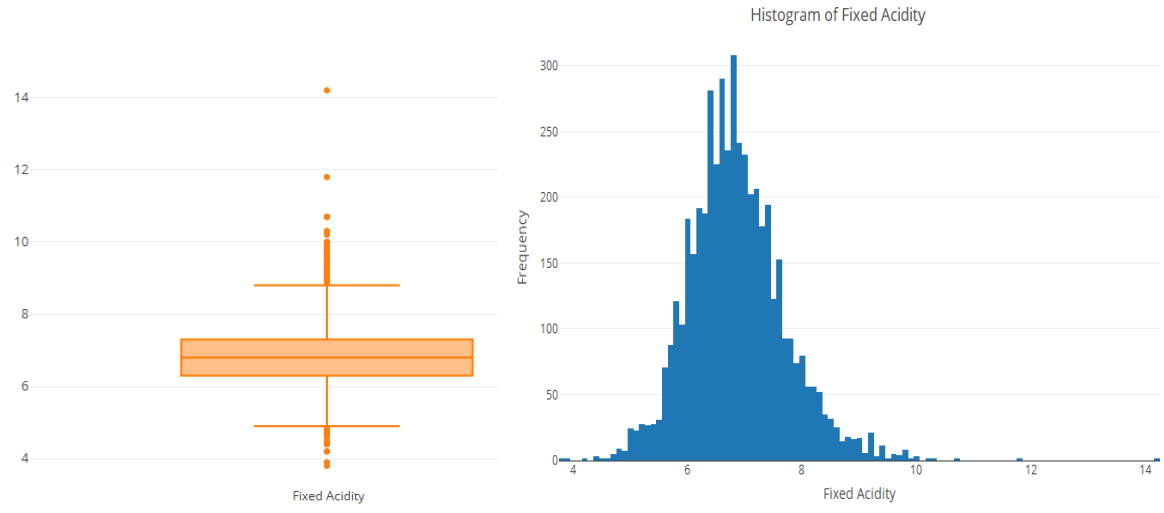
In Data Summary (Figure 3), there are no missing values in this dataset. However, we need to deal with outliers for each variable. In order to identify outliers, the boxplot and histogram are used.

```
# Data Summary
summary(wine_white)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0
##  Median :0.04300   Median : 34.00      Median :134.0
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0
##     density             pH          sulphates         alcohol
##  Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00
##  1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50
##  Median :0.9937   Median :3.180   Median :0.4700   Median :10.40
##  Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51
##  3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40
##  Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.878
##  3rd Qu.:6.000
##  Max.   :9.000
```

Figure 3

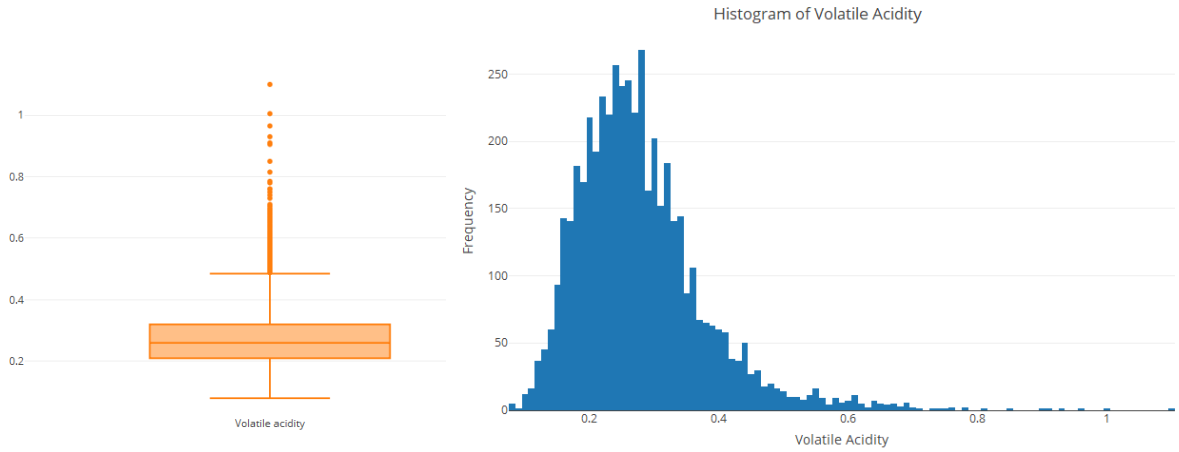
Fixed Acidity:



Observing boxplot and histogram before processing as above, we will remove outliers which have fixed acidity >= 0 and <= 4.

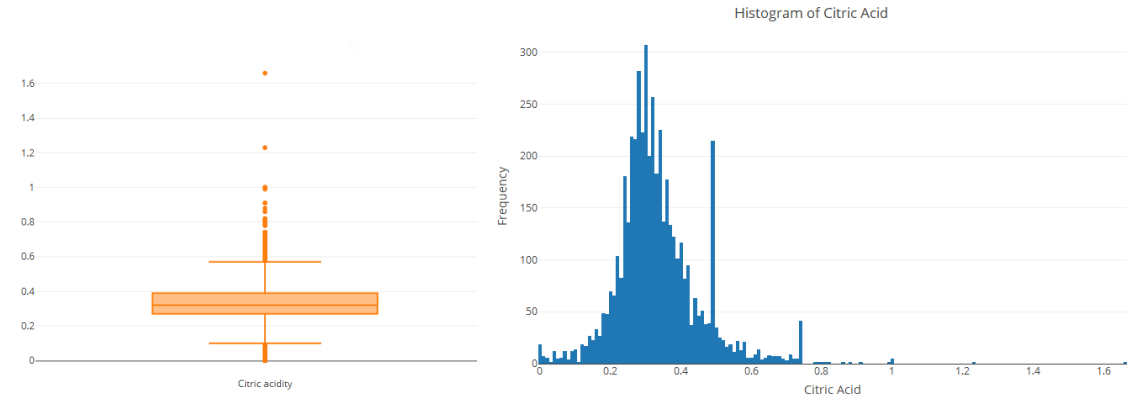
Volatile Acidity

Looking at boxplot and histogram before removing outlier in picture as below, we will remove outliers which have volatile acidity >= 0.6



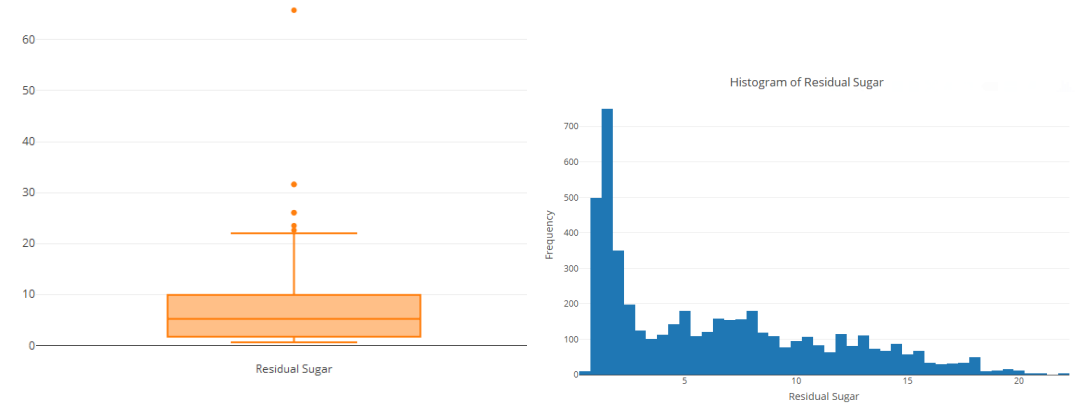
Citric Acidity:

Remove outliers which have citric acid is 0 or larger than 0.75



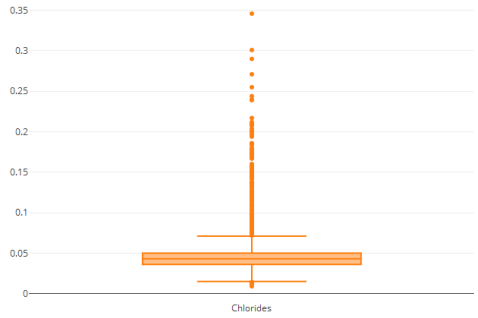
Residual Sugar:

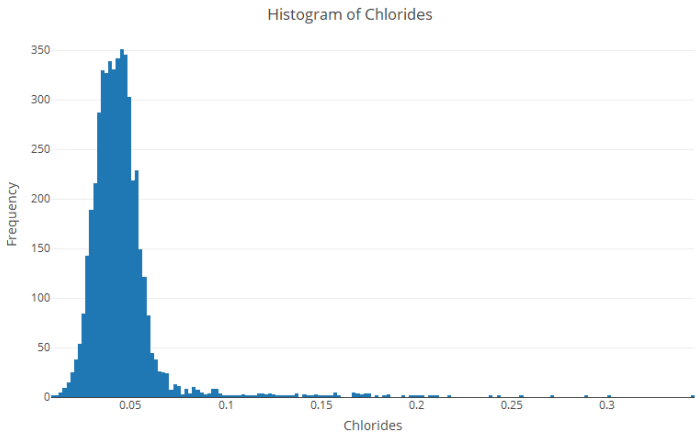
Remove outliers which have residual sugar larger than 22



Chlorides

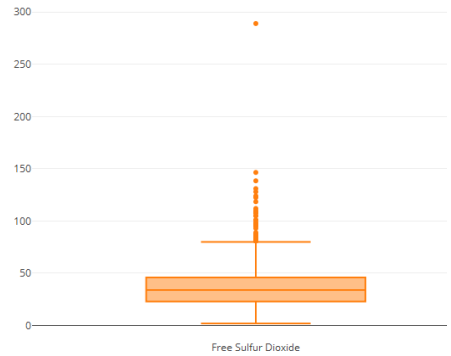
Remove outliers which have chlorides larger than 0.1

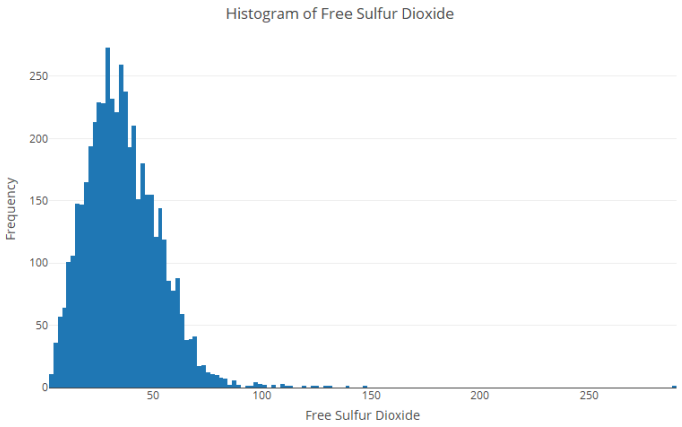




Free Sulfur Dioxide

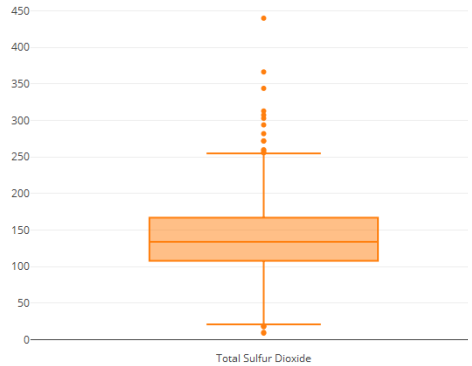
Remove outliers which have fee sulfur dioxide >= 75

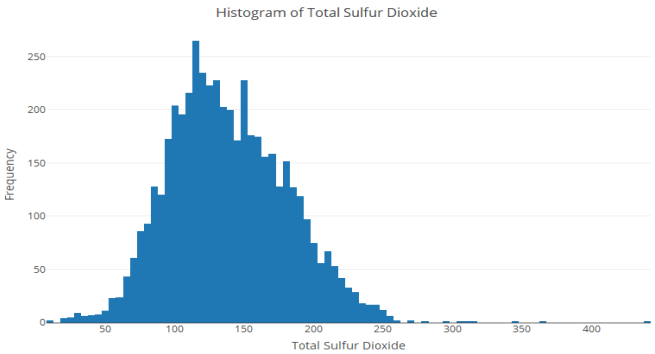




Total Sulfur Dioxide

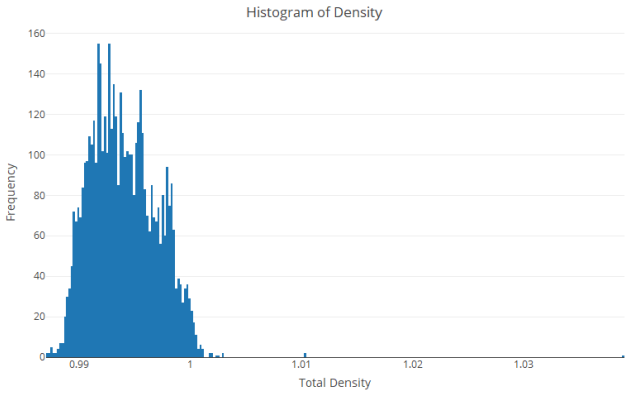
Remove outliers which have total sulfur dioxide >= 270 or <=25





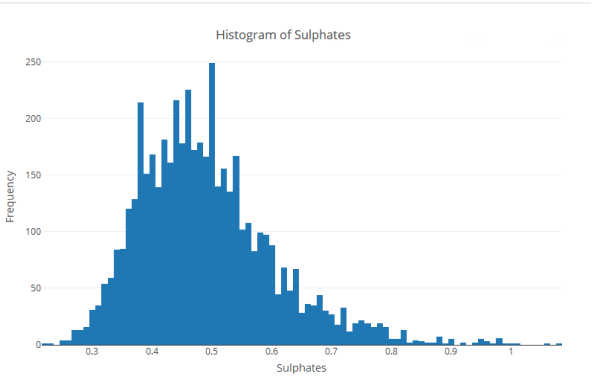
Density

Remove outliers which have density >= 1.001



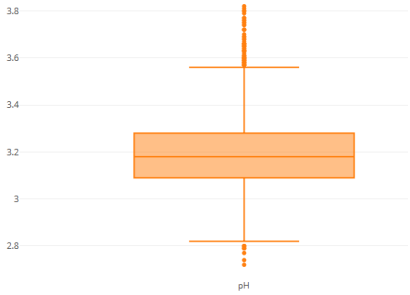
Sulphates

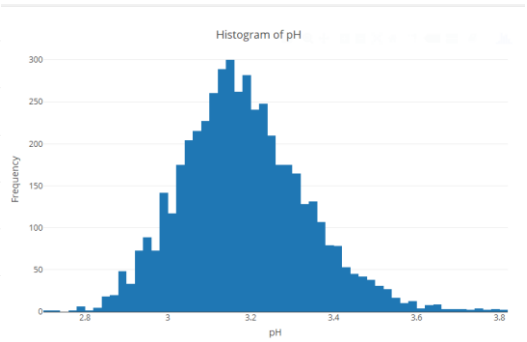
Remove outliers which have sulphates >= 0.8



pH

No need to remove outliers for pH





For Alcohol, we don’t need to do removing outlier for this variable

The table 1 is shown the number of quality classes before Data Processing

Table 1: Number of quality classes before Data Processing

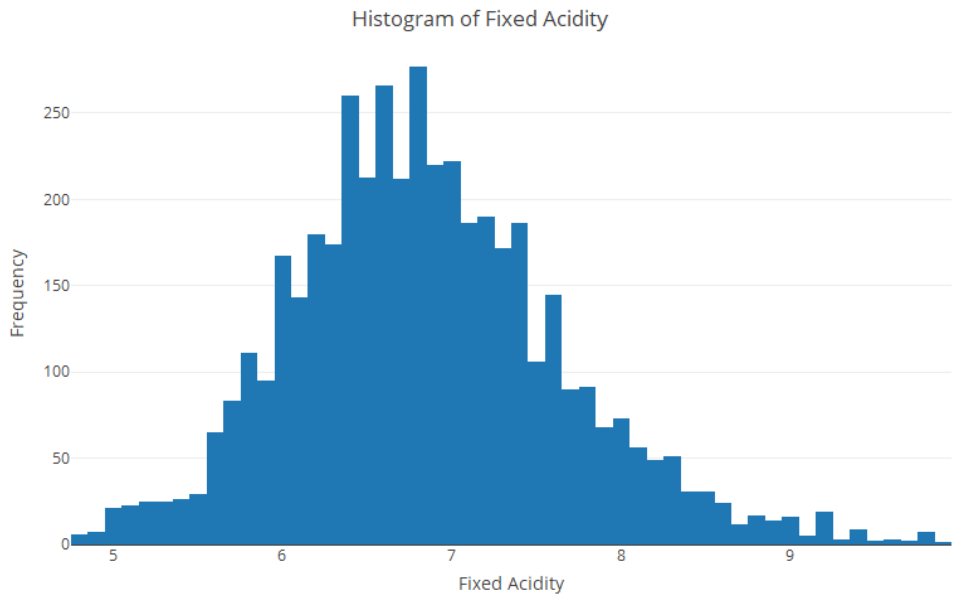
Var1	Freq
3	20
4	163
5	1457
6	2198
7	880
8	175
9	5

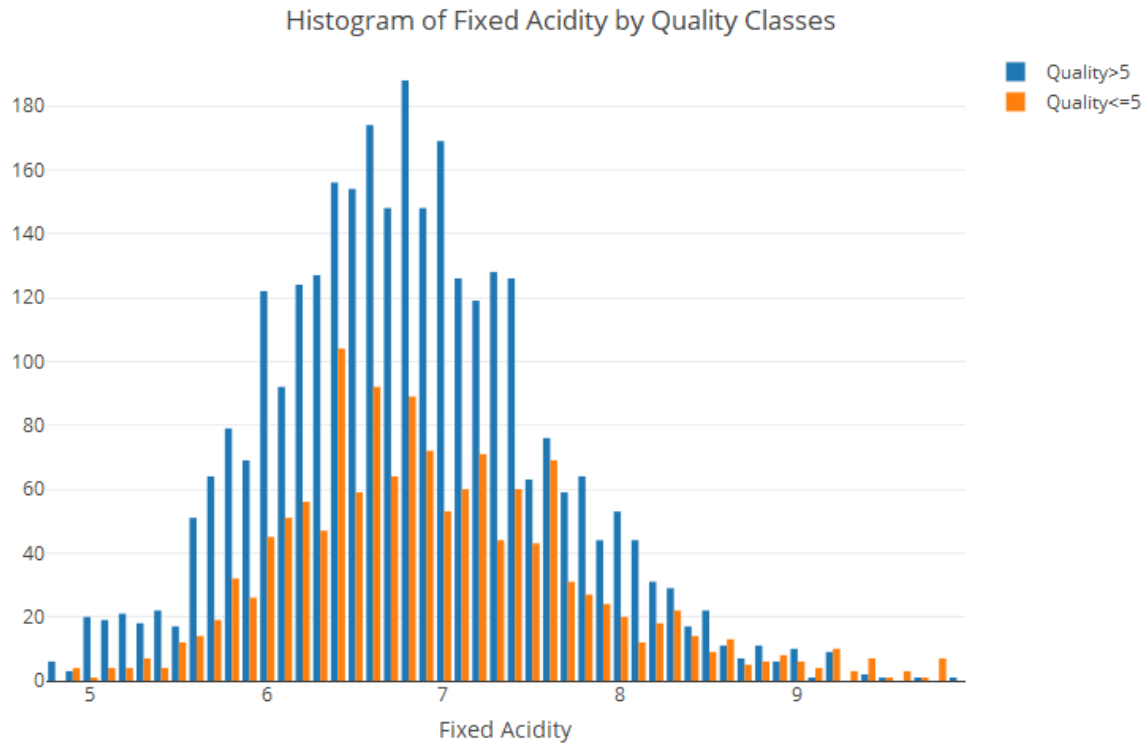
Table 2: Number of quality classes after processing

Var1	Freq
3	10
4	125
5	1322
6	2041
7	846
8	160
9	5

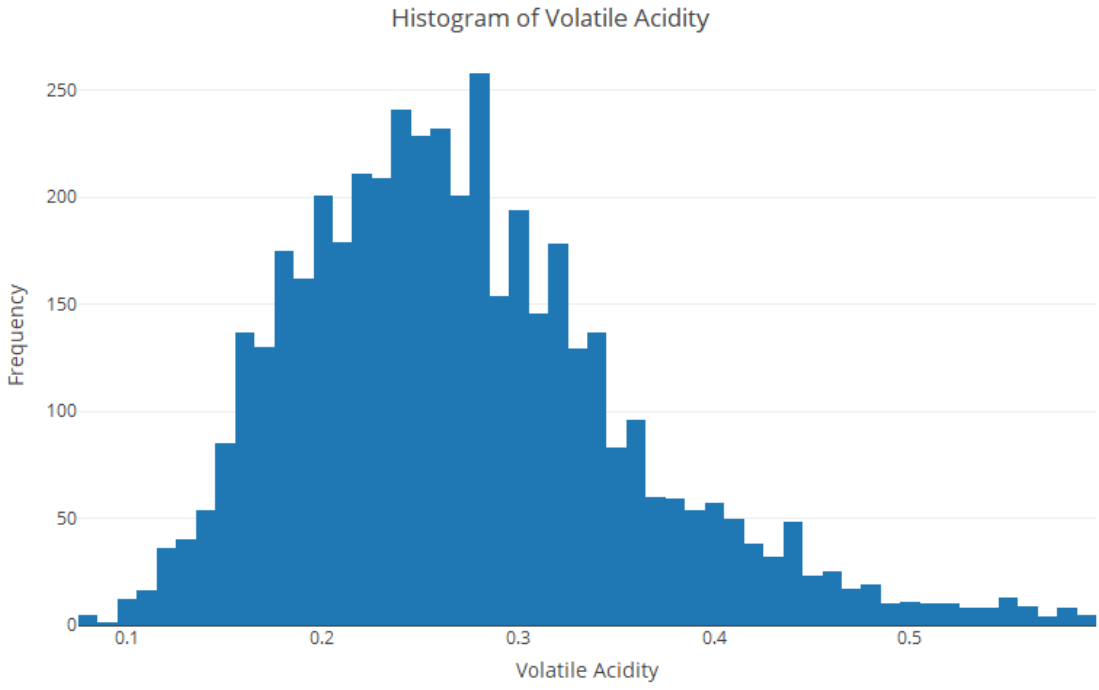
4. Data Explorer:

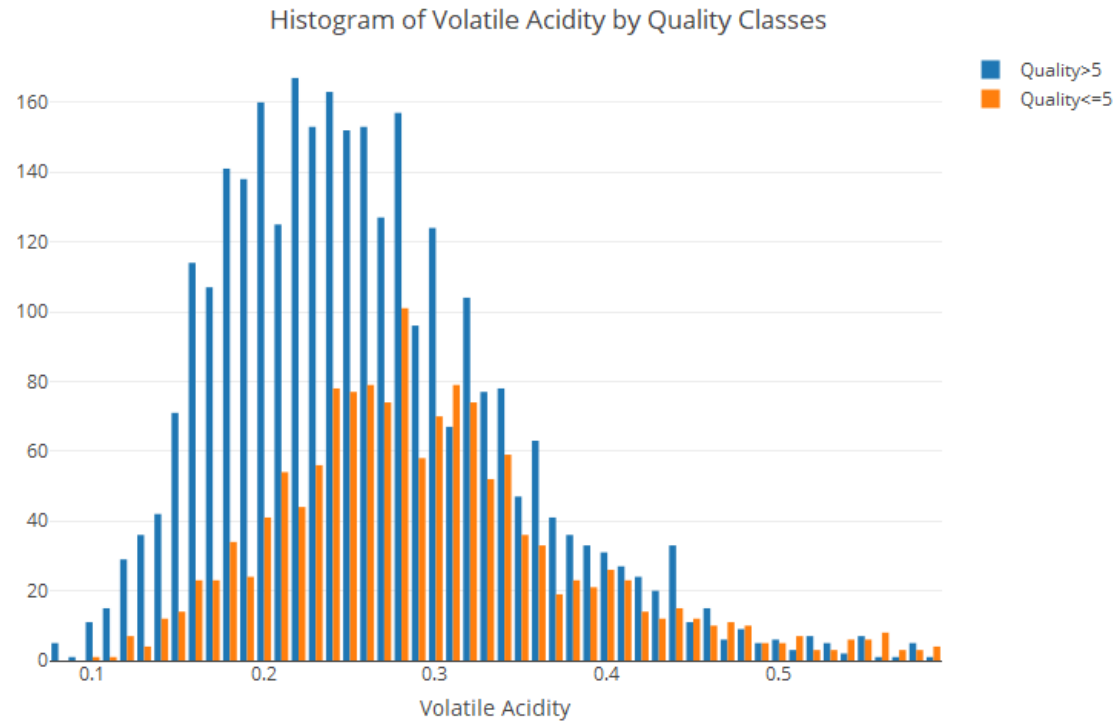
Fix Acidity



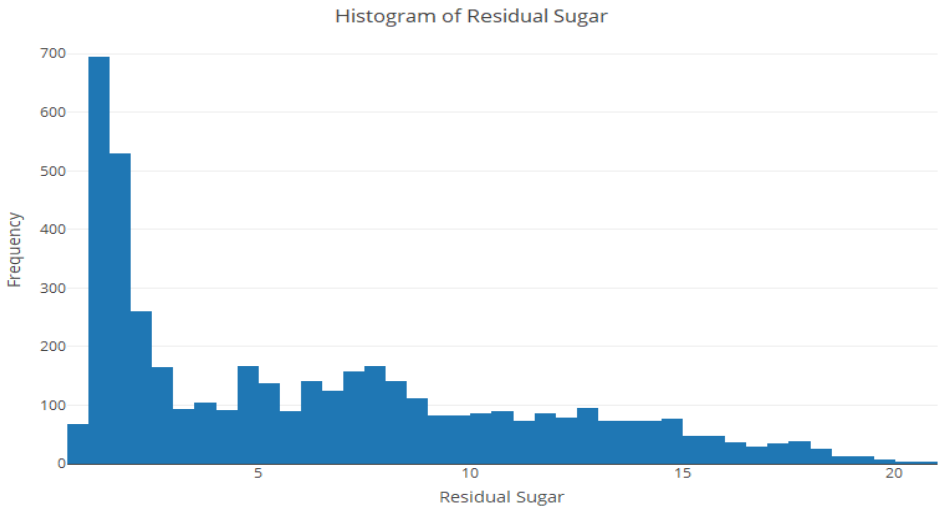


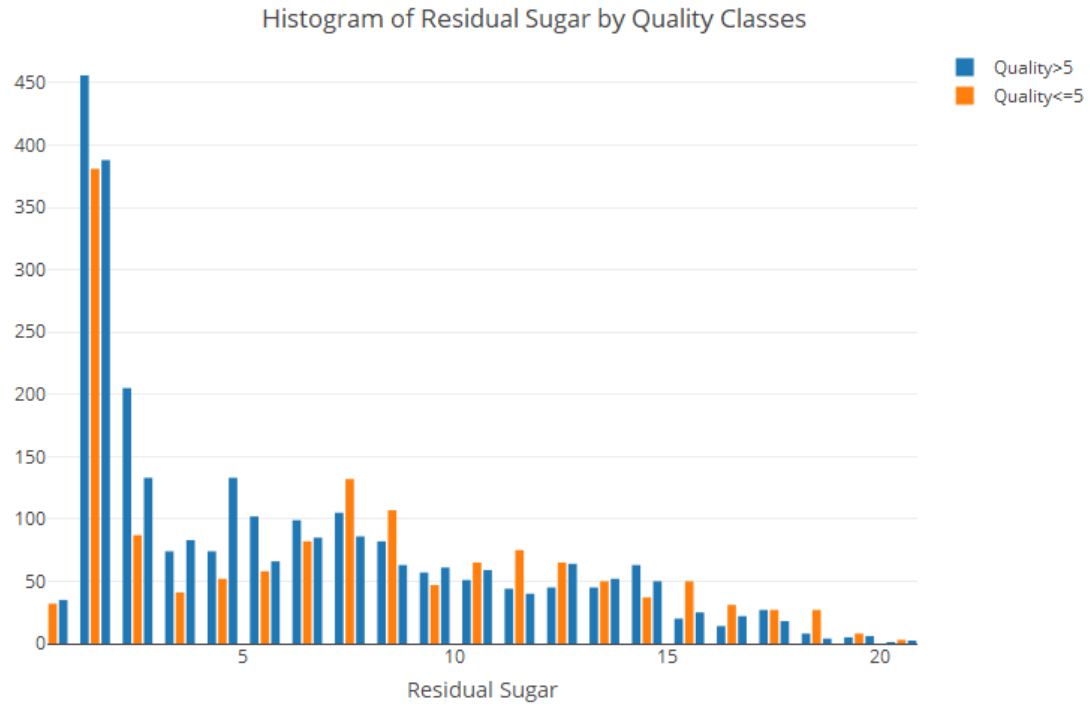
Volatile Acidity



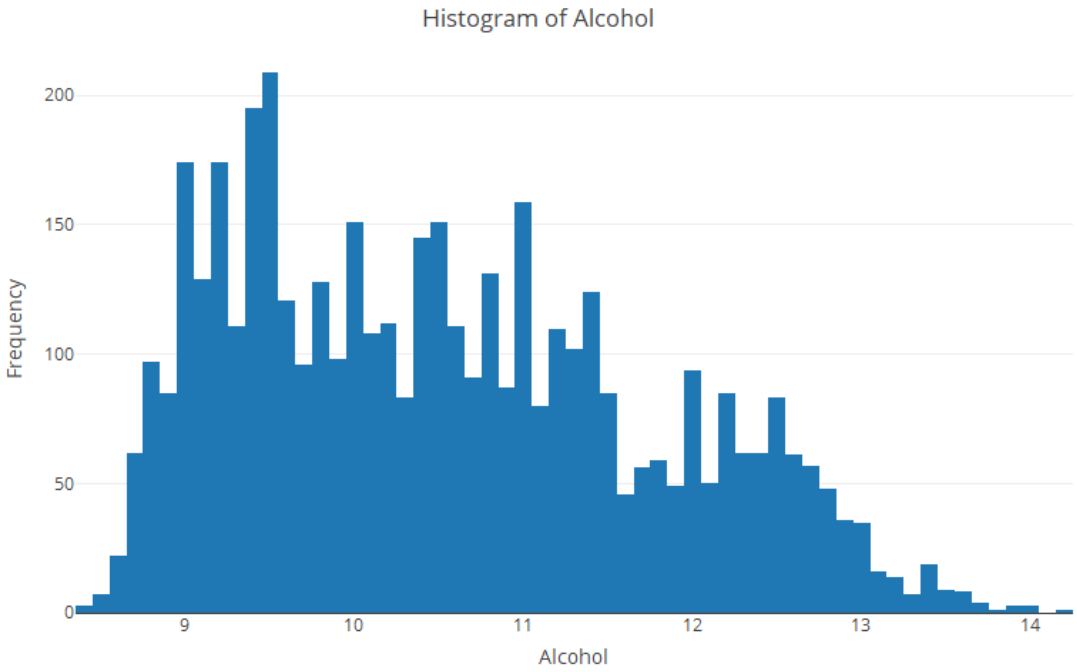


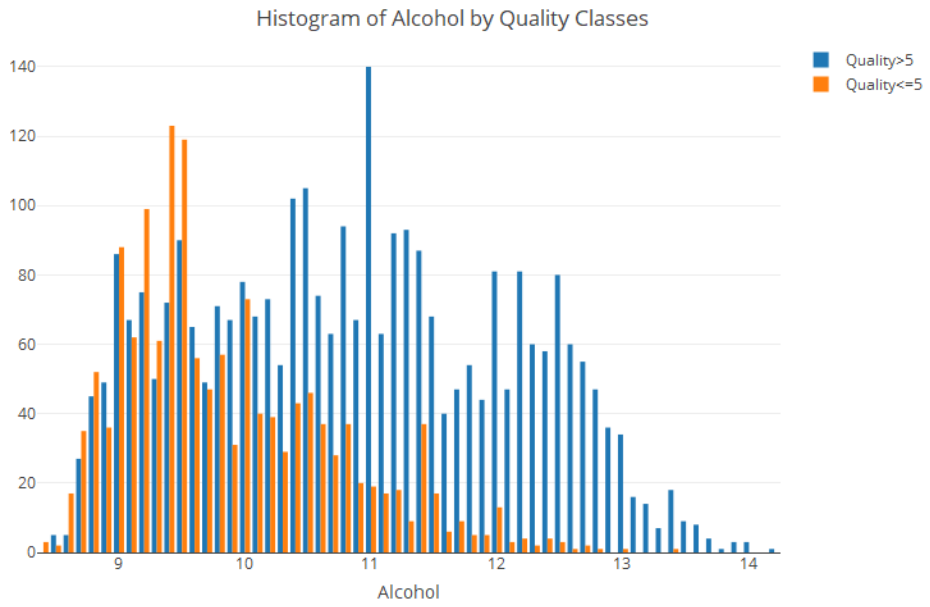
Residual Sugar




Histogram of Residual Sugar by Quality Classes
Quality>5
Quality<=5
450
400
350
300
250
200
150
100
50
0
5
10
15
20
Residual Sugar

Alcohol


Histogram of Alcohol
200
150
100
50
0
Frequency
9
10
11
12
13
14
Alcohol



We do the same thing with other features, it implies that each feature impact on the quality of wine. For Volatile Acidity, the lower result may lead the higher quality score. However, for fixed acidity, it does not clearly show whether the higher or lower result will impact strongly on the quality of score. This is also the same with Residual Sugar. However, with Alcohol, it identifies that the higher alcohol result will lead to higher quality larger than 5 score.

IV. SUMMARY:

After removing outliers in numeric features, in explore part, we see that all of features affects on the quality of wine. It will be helpful to predict the quality of wine