ANALYZING PHYSIOCHEMICAL PROPERTIES OF RED WINE AND PREDICTING ITS QUALITY

I. INTRODUCTION

In this project, we explore a data set on red wine quality and its physicochemical properties. First objective is to explore which chemical properties of red wines affect its quality. For that we'll start by exploring the data using the statistical program, R. As interesting relationships in the data are discovered, we'll produce and refine plots to illustrate them. We will then try to predict the quality of red wine using Support Vector Machines.

II. DATASET

This dataset is taken from UC Irvine Machine Learning Repository. The dataset contains two files containing red wine and white wine related data respectively. We choose only red wine dataset. The dataset has 1599 observations of 13 numeric variables. Following is the description of variables:

Input variables (based on physicochemical tests):

1. **ID** – Wine ID from 1 to 1599
2. **Fixed acidity**: most acids involved with wine or fixed or nonvolatile (do not evaporate readily) (g / dm^3)
3. **Volatile acidity**: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste (g / dm^3)
4. **Citric acid**: found in small quantities, citric acid can add ‘freshness’ and flavor to wines (g / dm^3)
5. **Residual sugar**: the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet (g / dm^3)
6. **Chlorides**: the amount of salt in the wine (g / dm^3)
7. **Free sulfur dioxide**: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine (mg / dm^3)
8. **Total sulfur dioxide**: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine (mg / dm^3)
9. **Density**: the density of water is close to that of water depending on the percent alcohol and sugar content (g / dm^3)

10. **pH**: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
11. **Sulphates**: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant (g / dm^3)
12. **Alcohol**: the percent alcohol content of the wine (% by volume)

Output variable (based on sensory data):

13. **Quality**: output variable (based on sensory data, score between 0 and 10)

III. EXPLORATORY DATA ANALYSIS

A. UNIVARIATE ANALYSIS

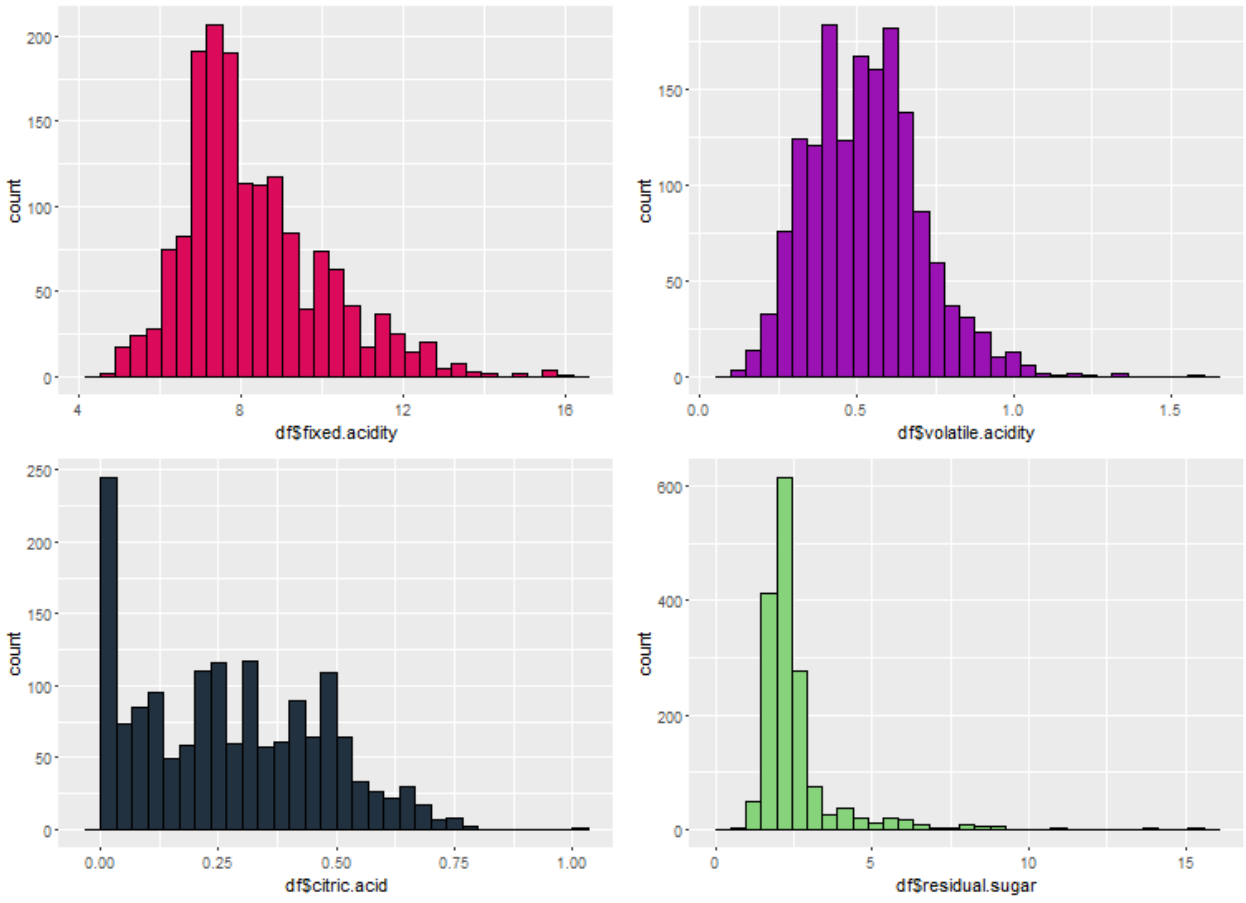


Figure 1: Histogram of 1st four variables fixed.acidity,volatile.acidity,citric.acid and residual.sugar

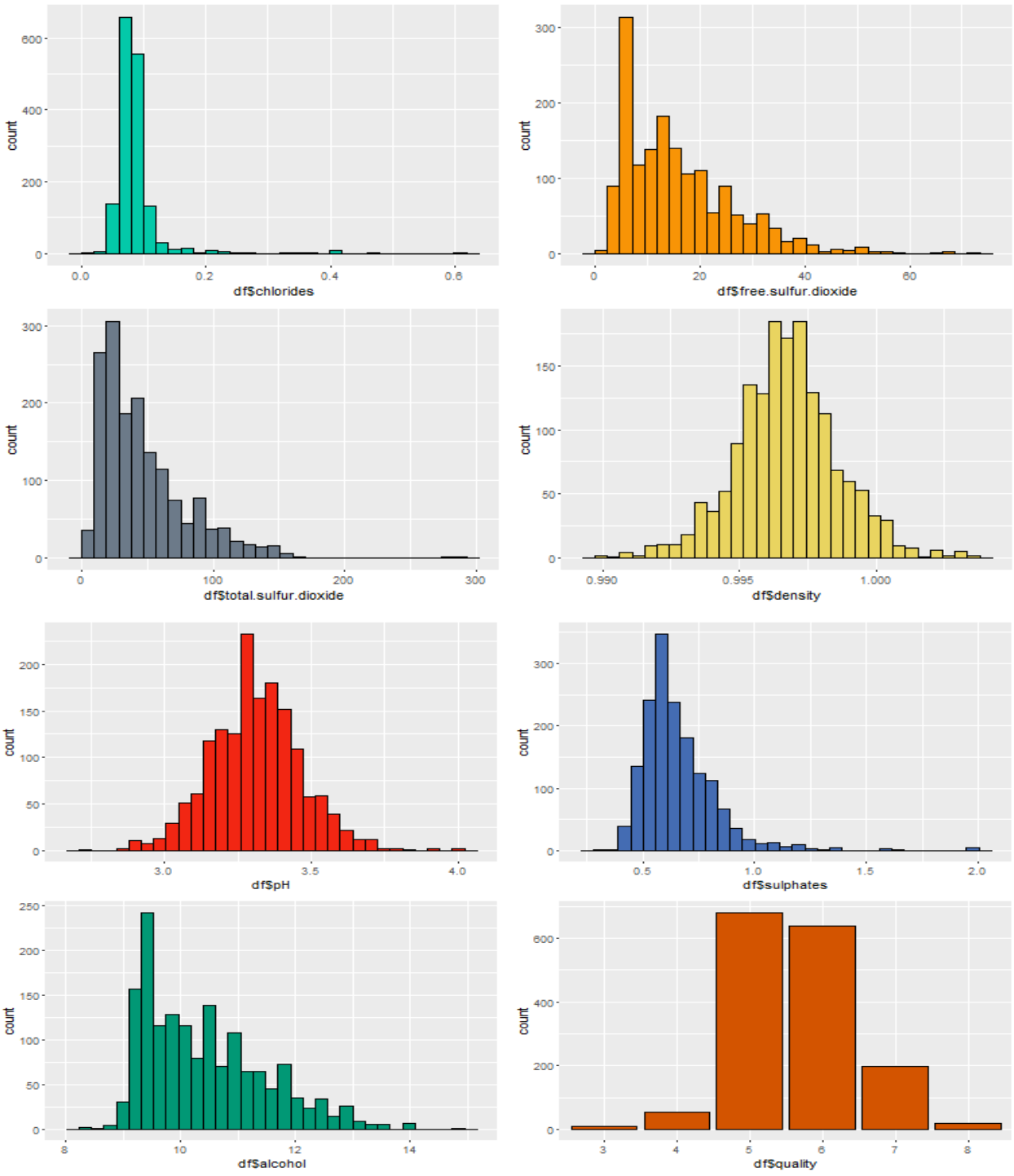


Figure 2: Histogram of next 8 variables chlorides, free.sulphur.dioxide, total.sulphur.dioxide, density, pH, sulphates, alcohol,quality

OBSERVATIONS:

1. Both distributions for fixed acidity and volatile acidity have long positive tails, this makes their mean higher than their medians, and make median better measure of central value. Moreover, volatile acidity distribution has a slight bimodal distribution.

   Summary fixed.acidty:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
4.60	7.10	7.90	8.32	9.20	15.90

   Summary volatile.acidty:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.12	0.39	0.52	0.5278	0.64	1.58

2. Citric acid distribution seems bimodal and there are few outliers too.
3. Residual sugar is highly positively skewed
4. The plot for chlorides shows the typical long positive tail, with the bulk of the values between 0.03 g/dm^3 and .20 g/dm^3

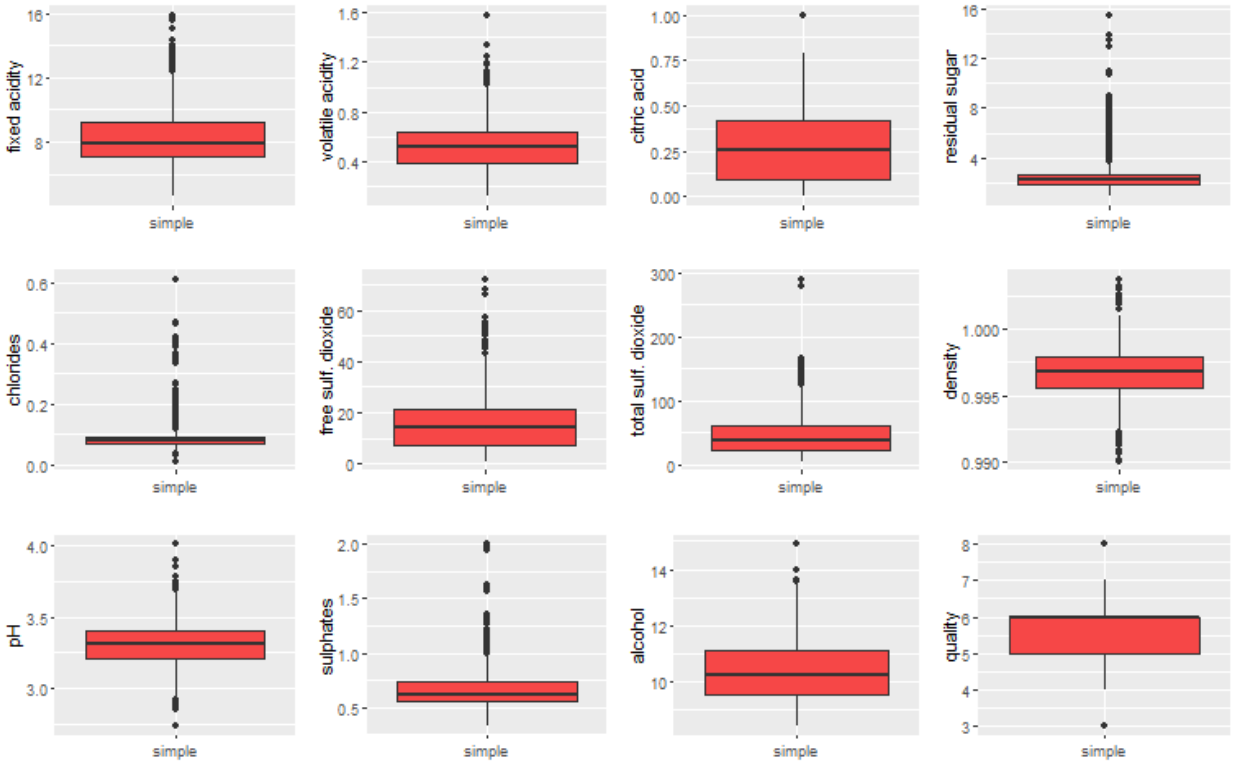


Figure 3: Boxplot of all variables

B. BIVARIATE ANALYSIS

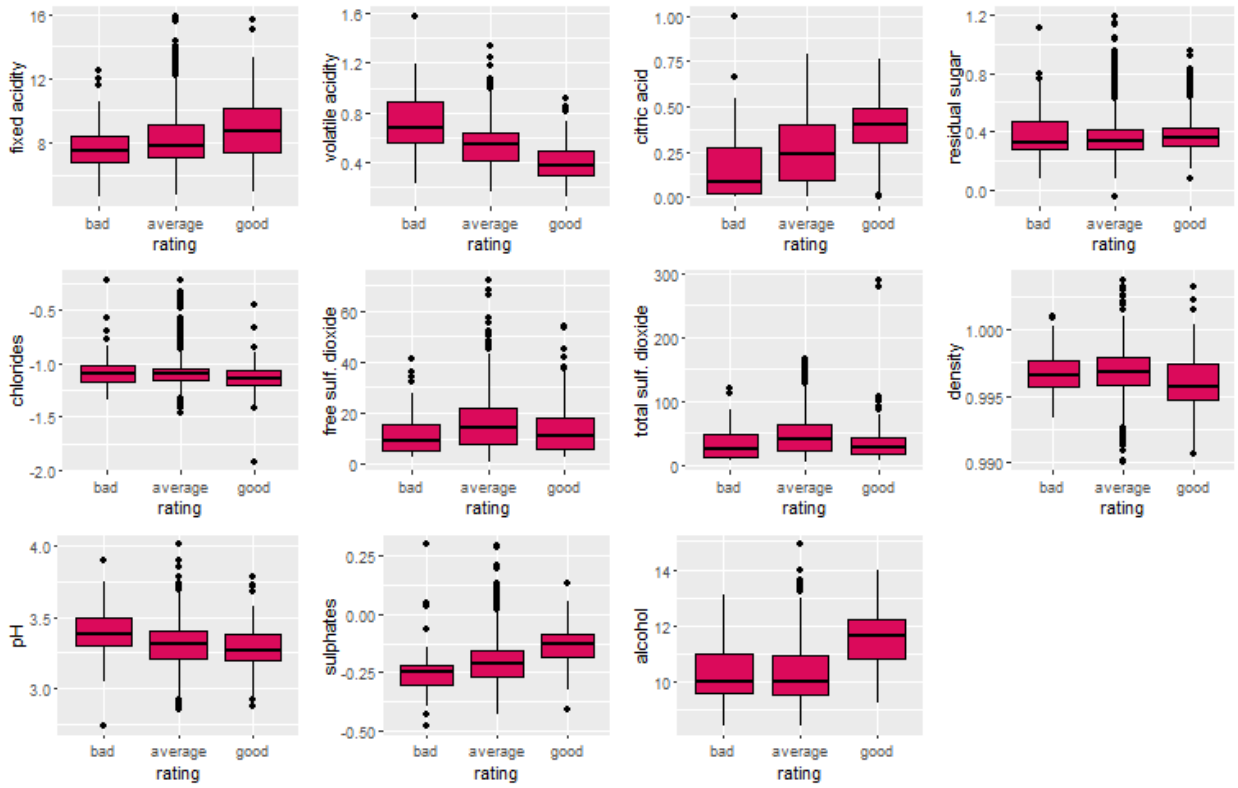


Figure 4: Boxplots of rating against all variables

OBSERVATIONS:

1. For Bivariate Analysis we created a new variable rating based on quality attribute. All the observations with quality 0-4 fall under bad rating, 5-6 under average rating and 7-10 under good rating.
2. From exploring these plots, it seems that a 'good' wine generally has these trends:

    - higher fixed acidity (tartaric acid)
    - high citric acid levels,
    - lower volatile acidity (acetic acid)
    - lower pH (i.e. more acidic)
    - higher sulphates
    - higher alcohol
    - to a lesser extend, lower chlorides and lower density

CORRELATIONS:

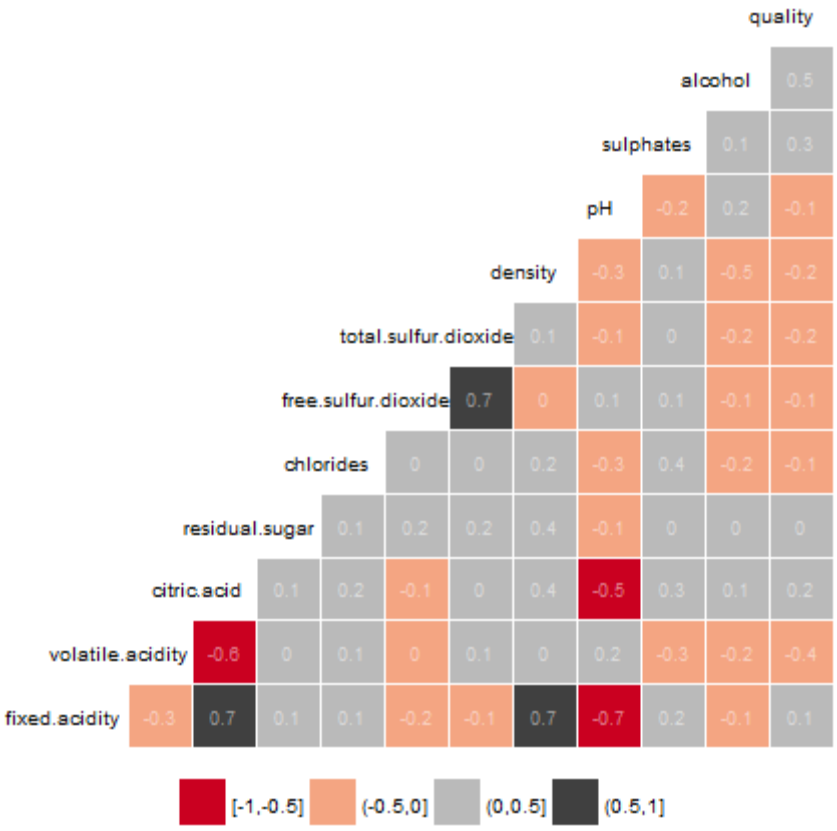


Figure 5: Correlation Matrix

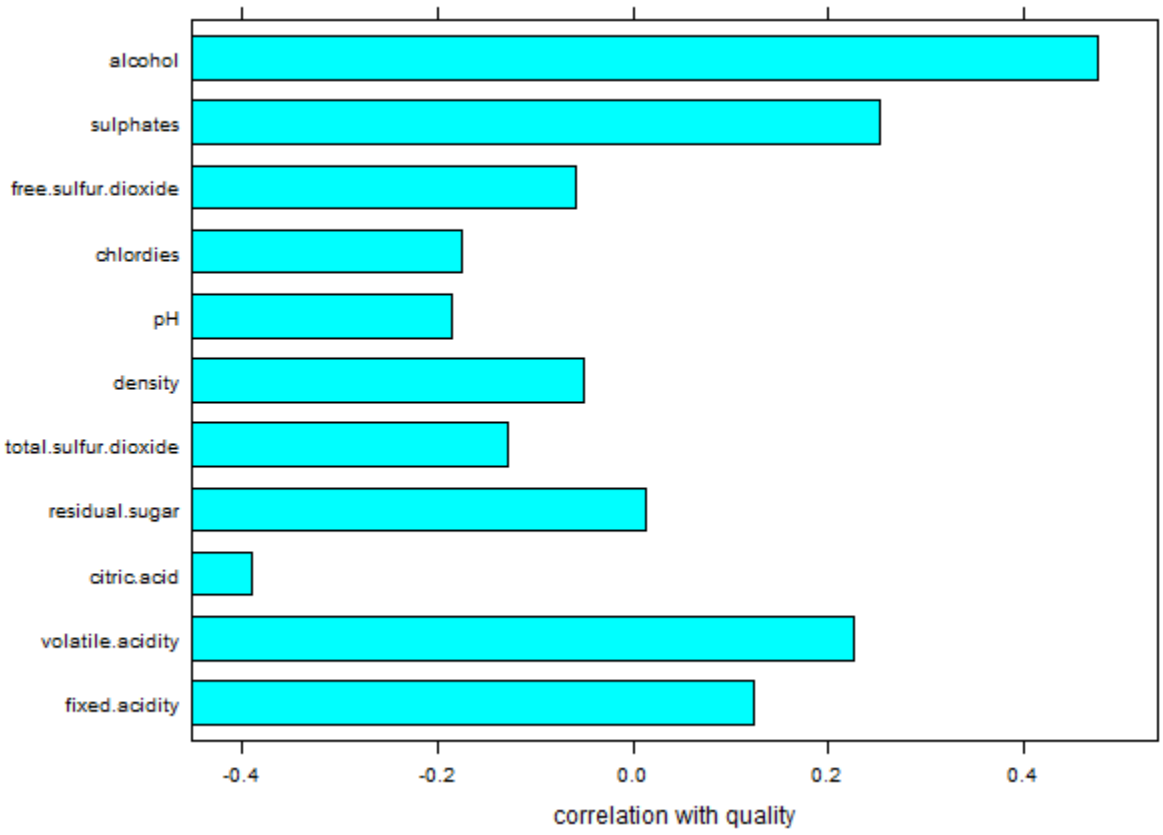


Figure 6: Bar chart showing correlation of all attributes with quality

From the above two correlation diagrams we have following observations:

1. From Figure 5 we can see that there is a strong positive correlation between fixed.acidity and volatile acidity.
2. There is strong positive correlation between free.sulphur.dioxide and total.sulphur.dioxide
3. There is strong positive correlation between fixed.acidity and density.
4. There is strong negative correlation between pH and fixed.acidity and pH and citric acid
5. There is strong negative correlation between volatile.acidity and citric acid
6. From Figure 6 we can find that following attributes are highly correlated with quality :
    a. alcohol
    b. sulphates
    c. volatile acidity
    d. citric acid

C. MULTIVARIATE ANALYSIS

We will examine 4 features which show high correlation with quality: alcohol, sulphates, volatile acidity and citric acid. These scatterplots are faceted by rating to illustrate the population differences between good wines, average wines, and bad wines.

I. Effect of acids on wine quality

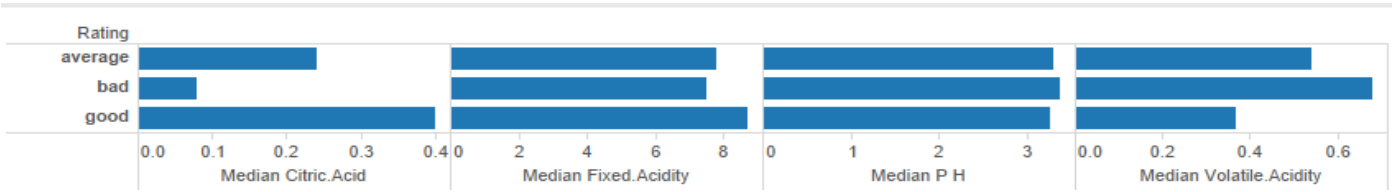


Figure 7: Effect of acids on wine quality

We plot median values citric acid, fixed.acidity and pH for average, bad and good wines. We see that bad wines have low pH and high acidity values and good wines have high acidity and low pH values. But volatile.acidity values are negatively correlated with wine quality.

II. Effect of alcohol on wine quality

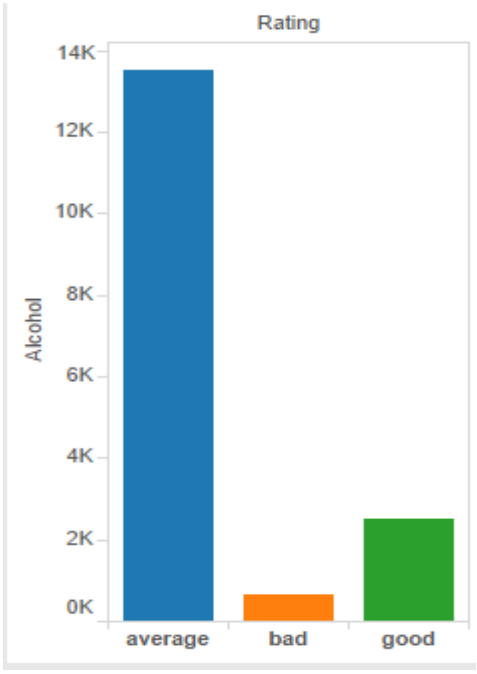
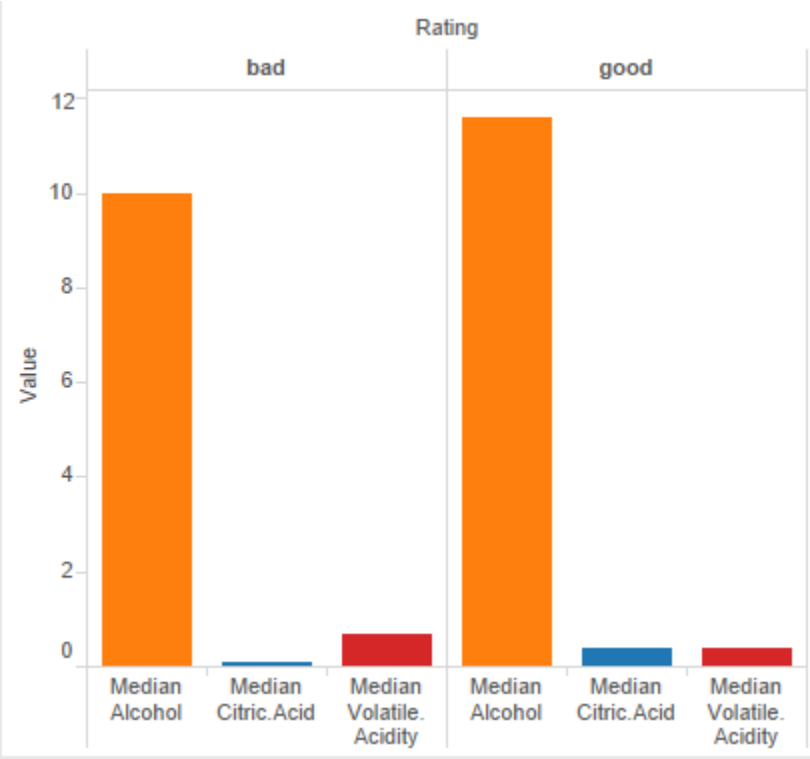


Figure 8: Effect of alcohol on wine quality

Good wines have more alcohol than bad wines.

III. Attributes influencing wine quality

We subset the data to remove the 'average' wines, or any wine with a rating of 5 or 6. As the correlation tests show, wine quality was affected most strongly by alcohol and volatile acidity and citric acid. While the boundaries are not as clear cut or modal, it's apparent that high volatile acidity--with few exceptions--kept wine quality down. A combination of high alcohol content, high citric acid content and low volatile acidity produced better wines.



IV. PREDICTION

The Second Objective of this project is to predict the quality of wine based on its physiochemical properties. For that we use SVM in R. We use 1300 observations for training and 299 for testing. We tuned the SVM model using 10 fold cross validation and fitted the best model to the data. We got an accuracy of 0.6355 if we use just the above 3 attributes i.e. alcohol, citric acid and volatile acidity. And we get and accuracy of 0.6421 if we use all the features. Also, We try to predict the wine quality as good, average or bad i.e. based on rating variable and obtain an accuracy of 0.8392.

V. EVALUATION

Model 1: We construct a confusion matrix for evaluation of the two models. The model with only 3 predictors had following statistics:

```
Overall Statistics

               Accuracy : 0.6355
                 95% CI : (0.5781, 0.6901)
    No Information Rate : 0.4515
    P-Value [Acc > NIR] : 1.23e-10

                  Kappa : 0.3648
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
Sensitivity           0.00000  0.00000   0.7037   0.7222  0.19048  0.00000
Specificity           1.00000  1.00000   0.7256   0.6532  0.98561  1.00000
Pos Pred Value            NaN      NaN   0.6786   0.6026  0.50000      NaN
Neg Pred Value        0.98662  0.96656   0.7484   0.7635  0.94158  0.98997
Prevalence            0.01338  0.03344   0.4515   0.4214  0.07023  0.01003
Detection Rate        0.00000  0.00000   0.3177   0.3043  0.01338  0.00000
Detection Prevalence  0.00000  0.00000   0.4682   0.5050  0.02676  0.00000
Balanced Accuracy     0.50000  0.50000   0.7147   0.6877  0.58804  0.50000
```

Model 2: The model with all 11 predictors (excluding ID) had following statistics:

```
Overall Statistics

               Accuracy : 0.6421
                 95% CI : (0.5849, 0.6965)
    No Information Rate : 0.4515
    P-Value [Acc > NIR] : 2.612e-11

                  Kappa : 0.3717
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
Sensitivity           0.00000  0.00000   0.7259   0.7143  0.19048  0.00000
Specificity           1.00000  1.00000   0.6951   0.6763  0.99640  1.00000
Pos Pred Value            NaN      NaN   0.6622   0.6164  0.80000      NaN
Neg Pred Value        0.98662  0.96656   0.7550   0.7647  0.94218  0.98997
Prevalence            0.01338  0.03344   0.4515   0.4214  0.07023  0.01003
Detection Rate        0.00000  0.00000   0.3278   0.3010  0.01338  0.00000
Detection Prevalence  0.00000  0.00000   0.4950   0.4883  0.01672  0.00000
Balanced Accuracy     0.50000  0.50000   0.7105   0.6953  0.59344  0.50000
```

Model 3 : If we try to predict wine quality based on rating variable i.e. good, bad and average we get following statistics:

```
Overall Statistics

               Accuracy : 0.8392
                 95% CI : (0.7806, 0.8873)
    No Information Rate : 0.8392
    P-Value [Acc > NIR] : 0.547

                  Kappa : 0
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: bad Class: average Class: good
Sensitivity             0.00000         1.0000      0.0000
Specificity             1.00000         0.0000      1.0000
Pos Pred Value              NaN         0.8392         NaN
Neg Pred Value          0.94975            NaN      0.8894
Prevalence              0.05025         0.8392      0.1106
Detection Rate          0.00000         0.8392      0.0000
Detection Prevalence    0.00000         1.0000      0.0000
Balanced Accuracy       0.50000         0.5000      0.5000
```

VI. CONCLUSION

We conducted Exploratory Data Analysis using univariate, bivariate and multivariate plots and obtained attributes which influence the quality of wine. We found that a combination of high alcohol content, high citric acid content and low volatile acidity produced better wines. Also, we used all features to predict the quality of wine as a score between 3-8 and we achieved best accuracy of 0.6421 and if we try to predict the accuracy of wine based on rating i.e. good, average and bad we get an accuracy of 0.8392 with support vector Machines using a linear kernel.