WINE QUALITY ANALYSIS

Presented By:
Alekya Kumar
Monalisa Mishra
Phaneendra Ramachandraiah
Trupti Jadhav

University at Buffalo The State University of New York

Motivation

- 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 has shown a recent growth spurt as social drinking is on the rise.
- The price of the wine depends on rather abstract concept of wine appreciation by wine tasters whose opinions have a high degree of variability.
- Another important 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.
- If human quality of tasting can be related to the chemical properties of wine, the certification and quality assessment can be more controlled
- The main objective of our analysis is to predict the quality rankings from the chemical properties of the wines.
- A predictive model developed on this data can be used to provide guidance to Wine Makers regarding quality and price expected on their produce without heavy reliance on volatility of wine tasters.

Motivation

- 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 has shown a recent growth spurt as social drinking is on the rise.
- The price of the wine depends on rather abstract concept of wine appreciation by wine tasters whose opinions have a high degree of variability.
- Another important 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.
- If human quality of tasting can be related to the chemical properties of wine, the certification and quality assessment can be more controlled
- The main objective of our analysis is to predict the quality rankings from the chemical properties of the wines.
- A predictive model developed on this data can be used to provide guidance to Wine Makers regarding quality and price expected on their produce without heavy reliance on volatility of wine tasters.

Motivation

- 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 has shown a recent growth spurt as social drinking is on the rise.
- The price of the wine depends on rather abstract concept of wine appreciation by wine tasters whose opinions have a high degree of variability.
- Another important 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.
- If human quality of tasting can be related to the chemical properties of wine, the certification and quality assessment can be more controlled
- The main objective of our analysis is to predict the quality rankings from the chemical properties of the wines.
- A predictive model developed on this data can be used to provide guidance to Wine Makers regarding quality and price expected on their produce without heavy reliance on volatility of wine tasters.

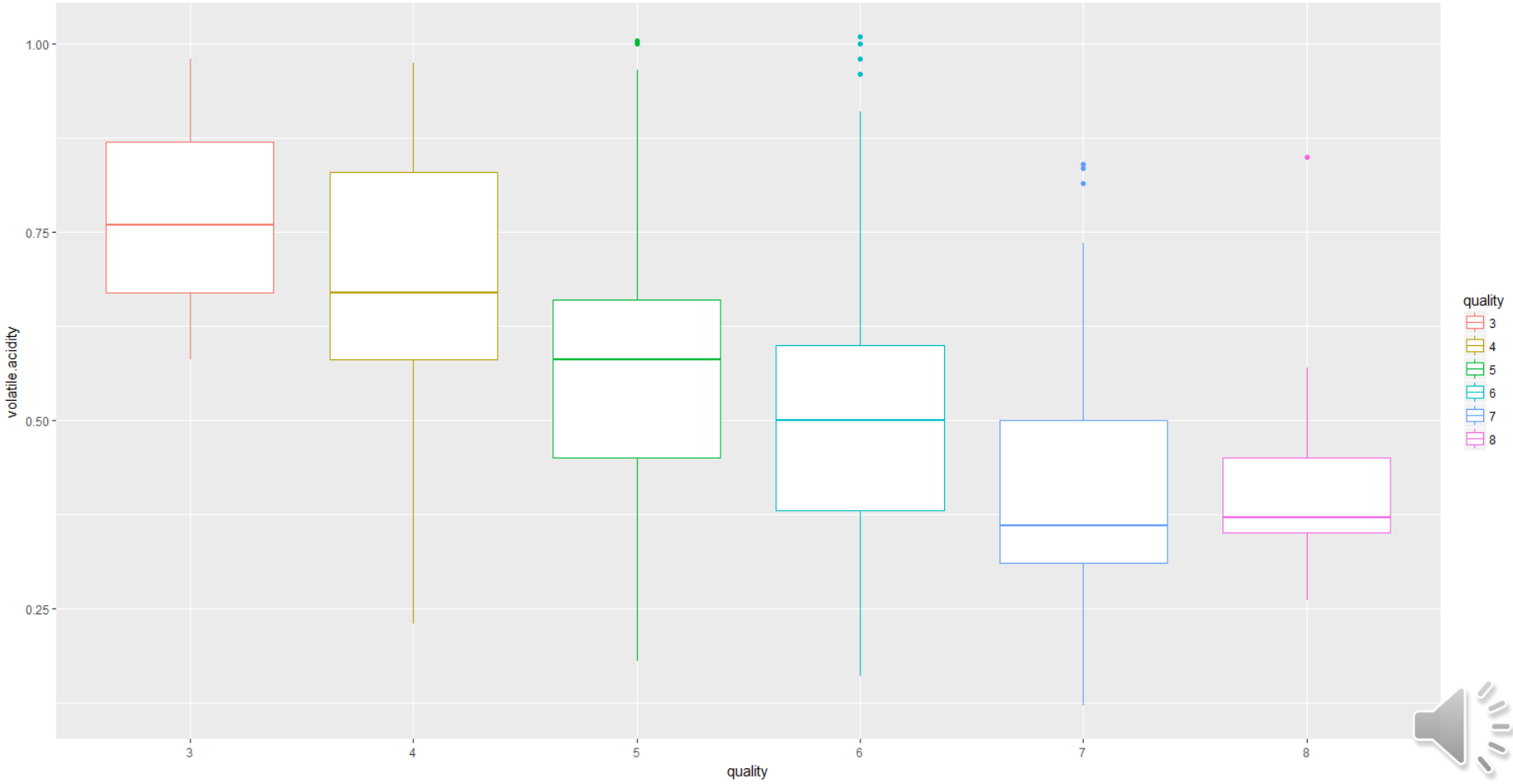
Dataset Introduction

- The Dataset was Downloaded from the University of California-Irvine archive.
- Two Different Datasets was used for analysis.
- Red Wine has around 1599 Observations .
- White Wine data set has around 4898 Observations
- The Dataset contains 12 variables out of which Quality is selected as the Response Variable.
- The analysis was done separately and the results are interpreted accordingly.

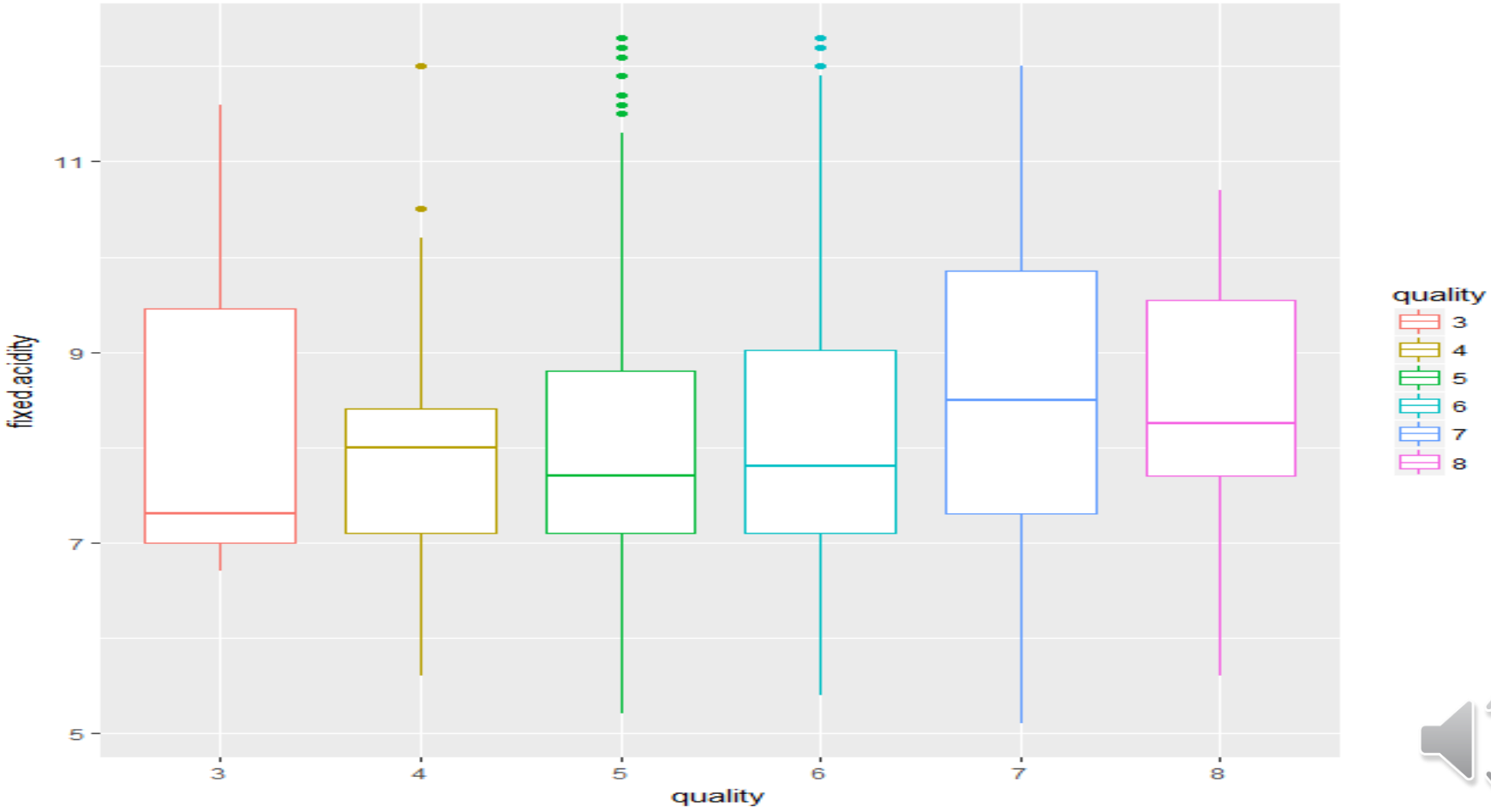
Dataset Description

- Fixed Acidity: measure of the total concentration of titratable acids and free hydrogen ions present in wine
- Volatile Acidity: measure of steam distillable acids present in a wine
- Citric Acid: used to boost the wine's total acidity
- Residual Sugar: measure of any natural grape sugars that are leftover after fermentation ceases
- Chlorides: amount of salt in the wine
- Free SO2: prevents microbial growth and the oxidation of wine
- Total SO2: amount of free and bound forms of SO2
- Density: measure of density of wine
- pH: Low pH wines will taste tart and crisp, while higher pH wines are more susceptible to bacterial growth. Most wine pH's fall around 3 or 4; about 3.0 to 3.4 is desirable for white wines, while about 3.3 to 3.6 is best for red wine
- Sulfates: a wine additive which can contribute to sulfur dioxide gas (SO2) levels, which acts as an antimicrobial and antioxidant
- Alcohol: the percentage of alcohol present in the wine

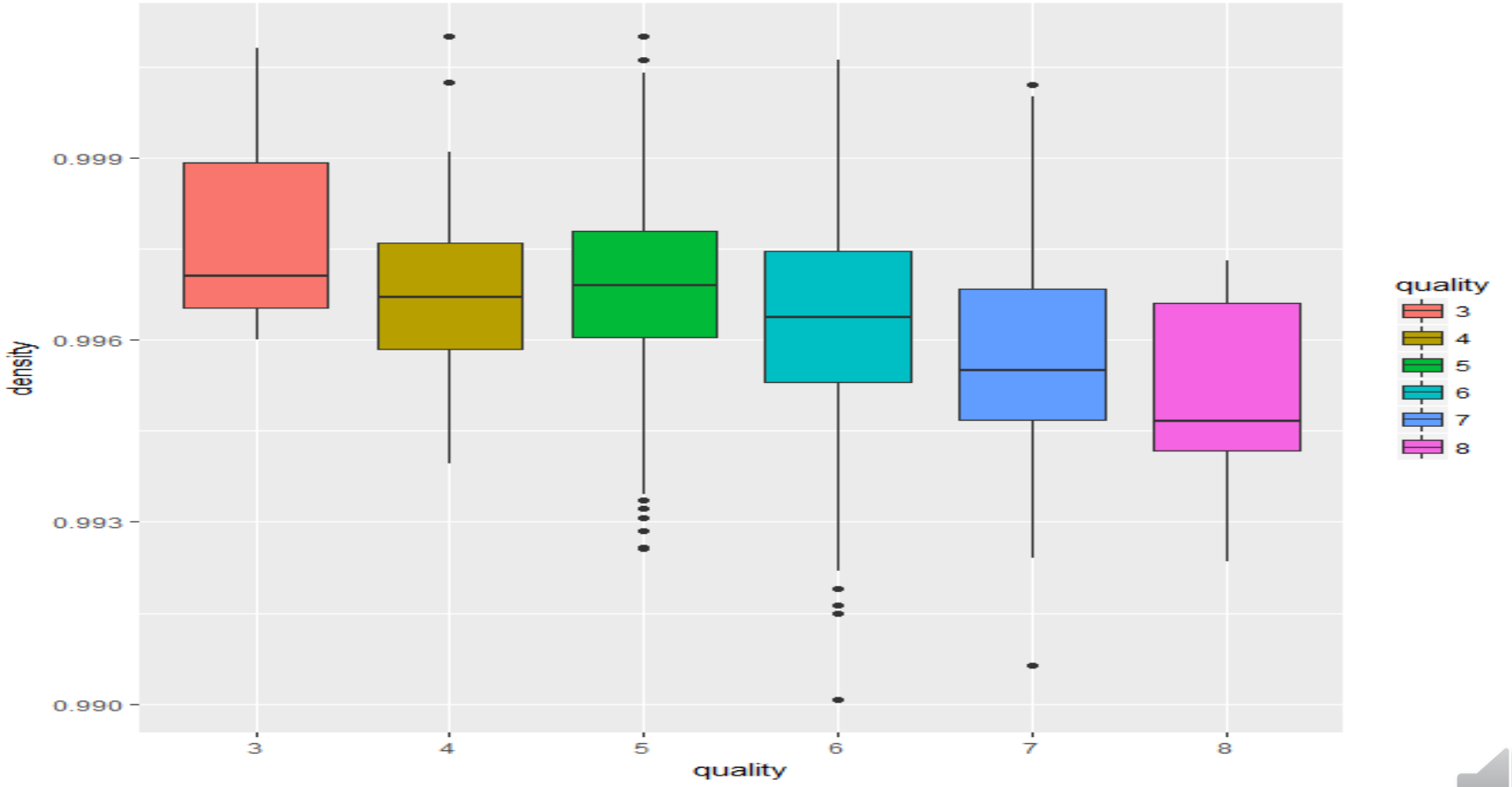
Exploratory Data Analysis – Volatile Acidity Vs Quality



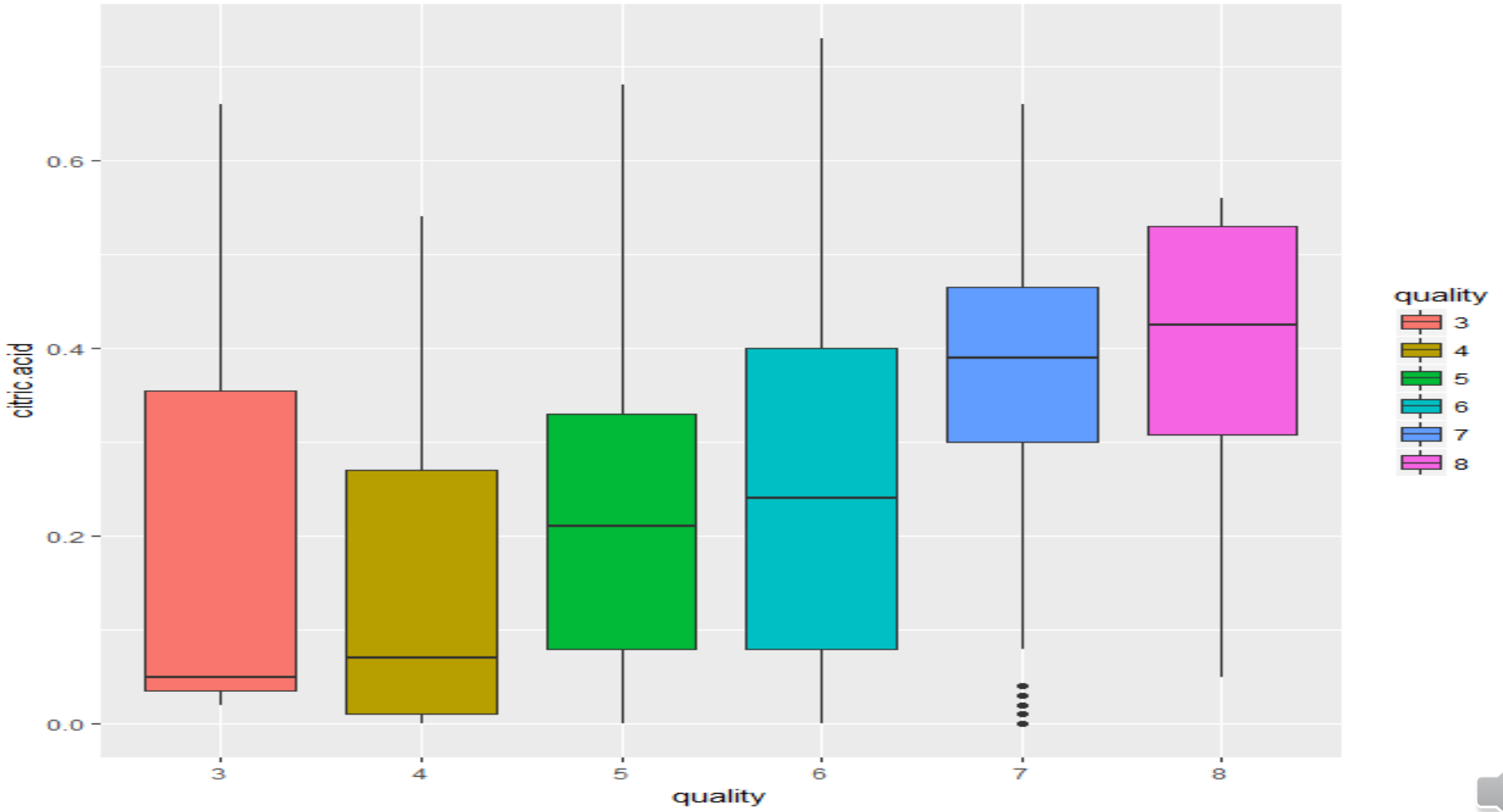
Fixed Acidity Vs Quality



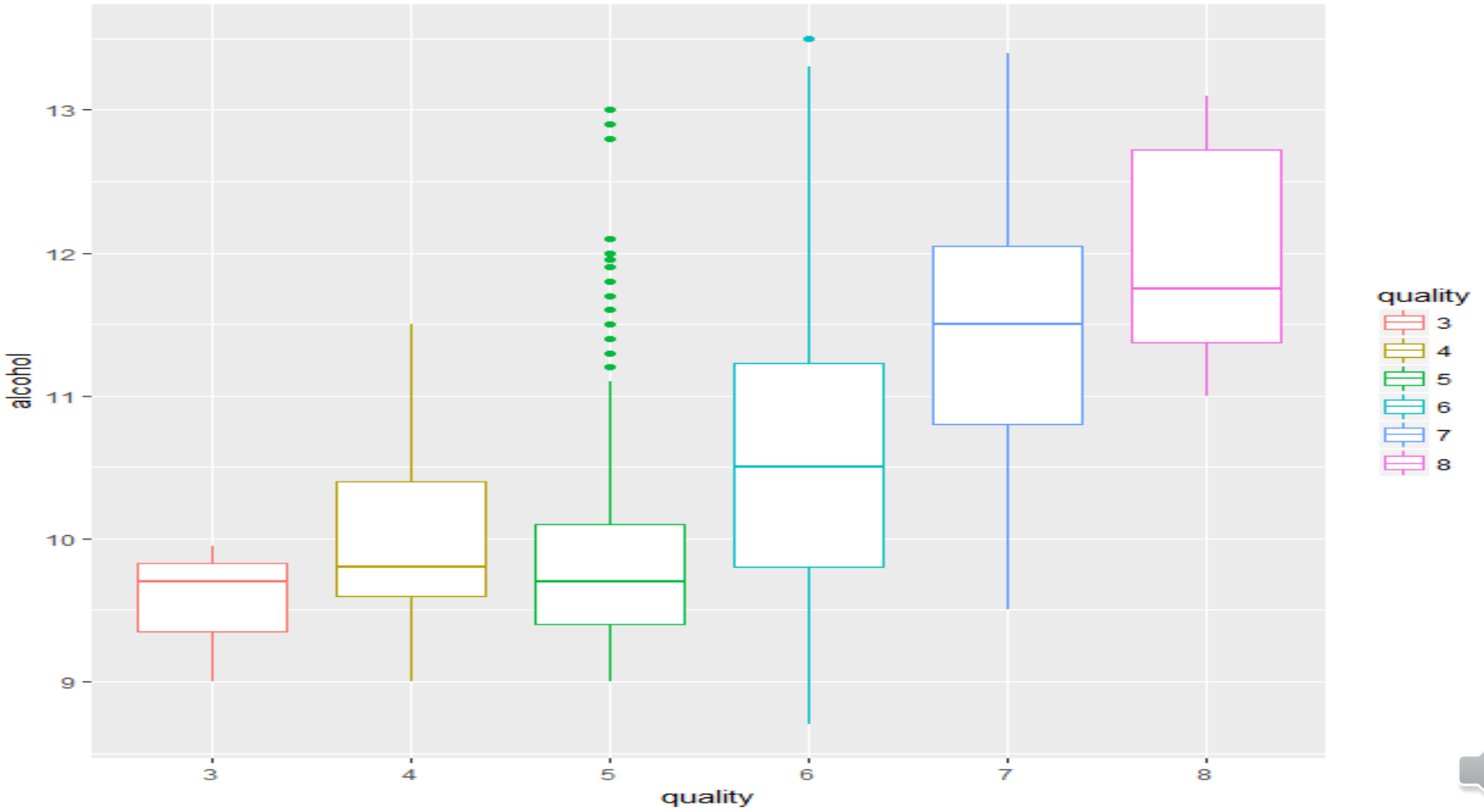
Density Vs Quality



Citric Acid Vs Quality

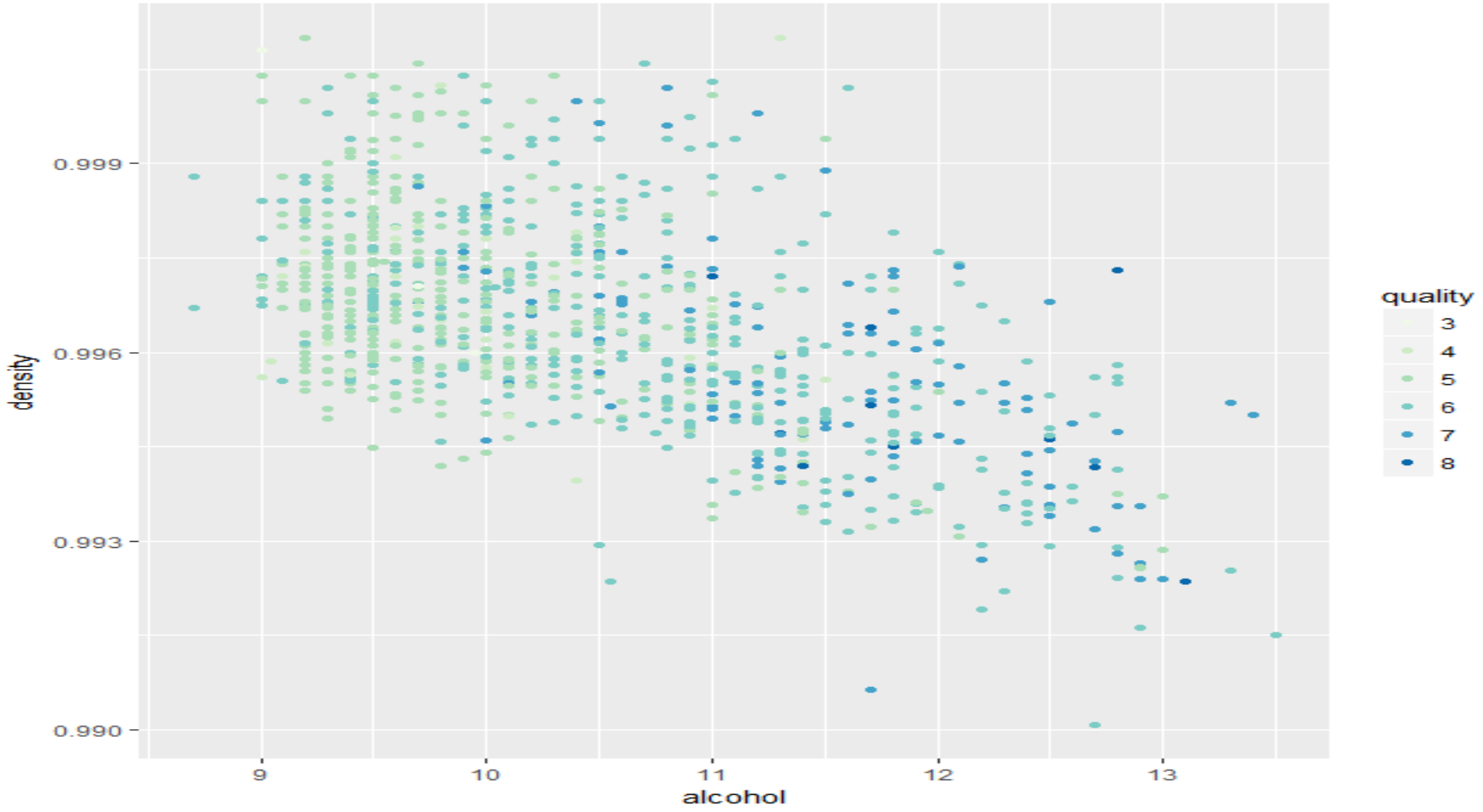


Alcohol Vs Quality

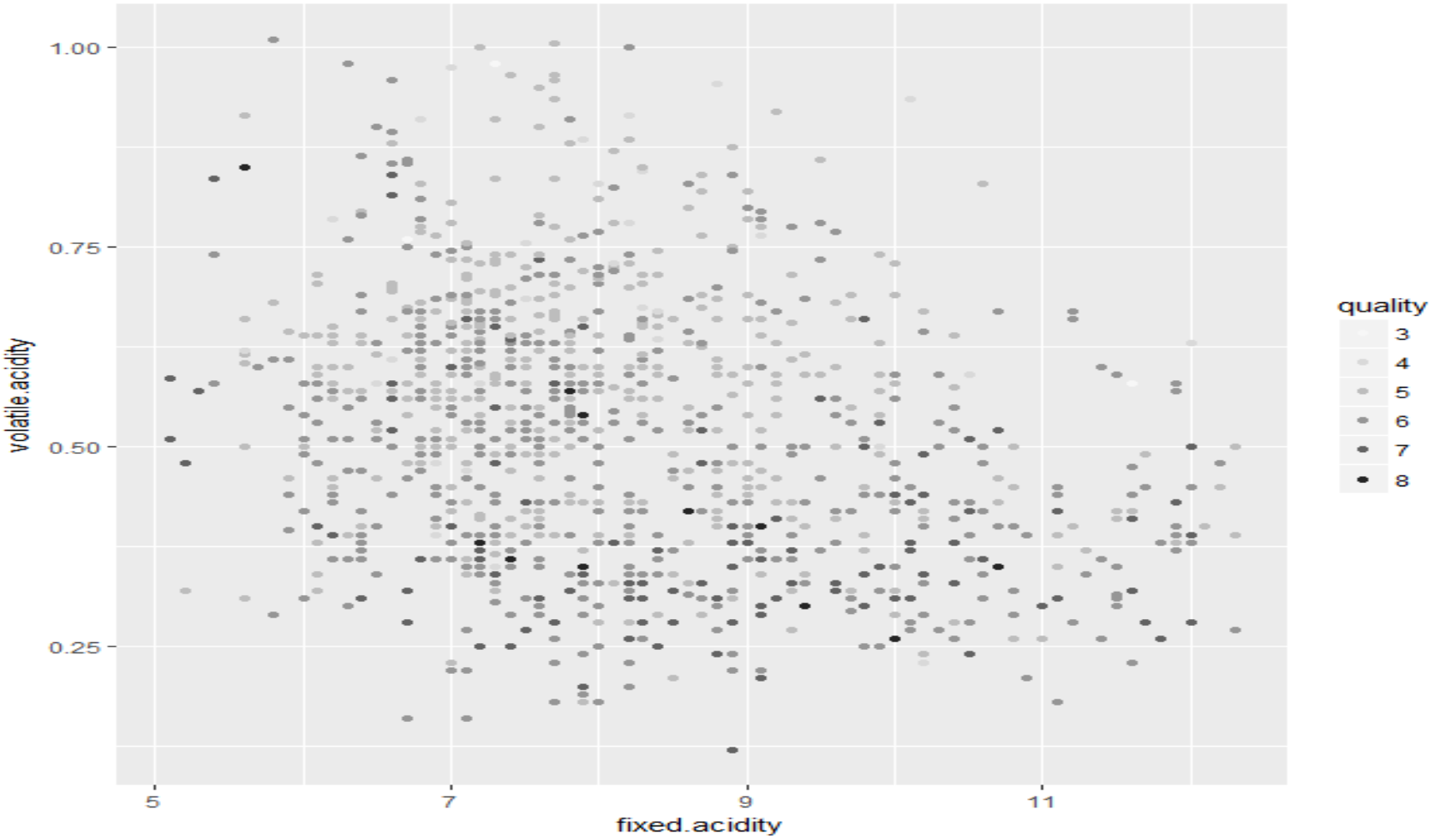


Volatile Acidity Vs Fixed Acidity

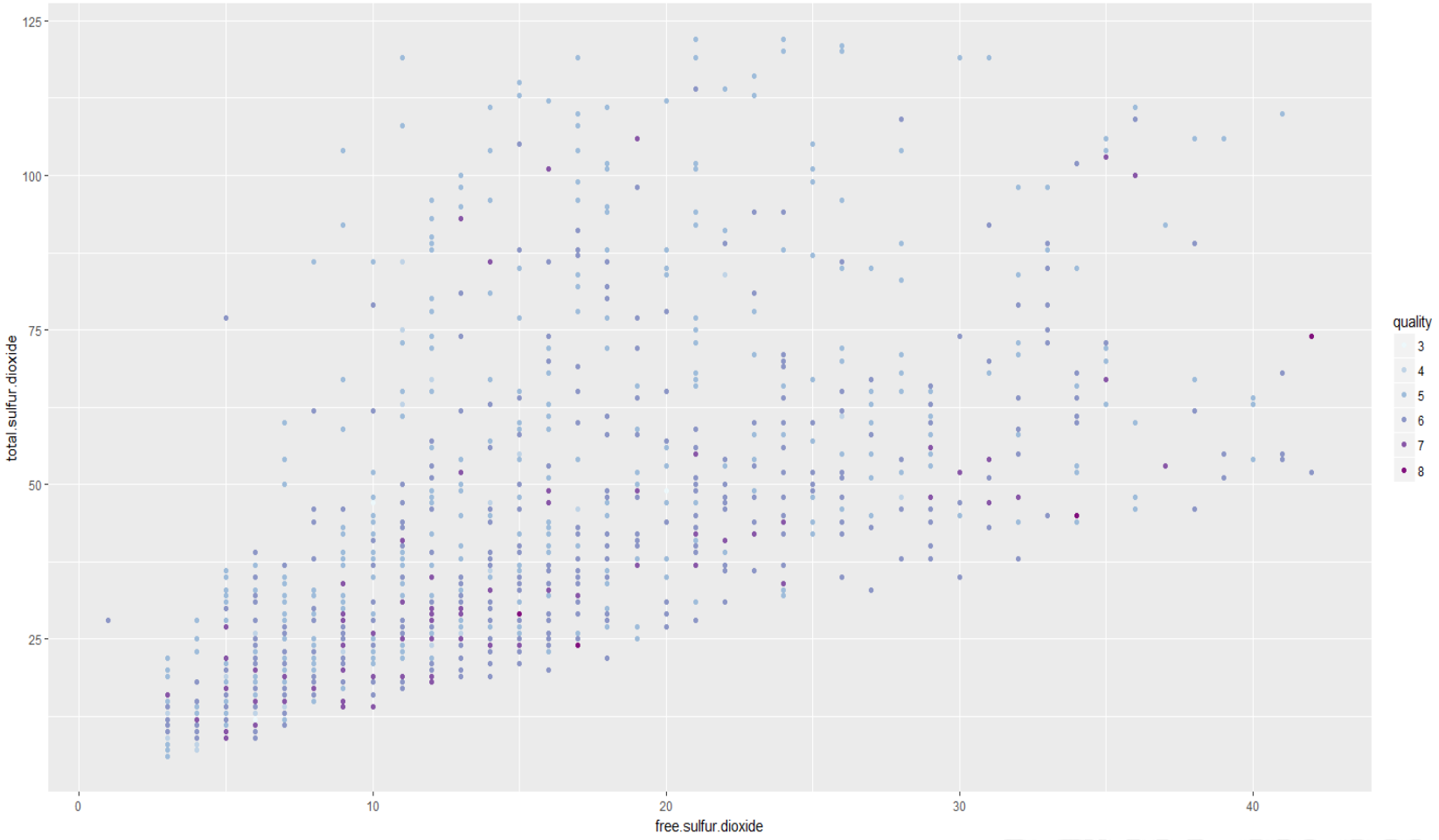
Density Vs Alcohol



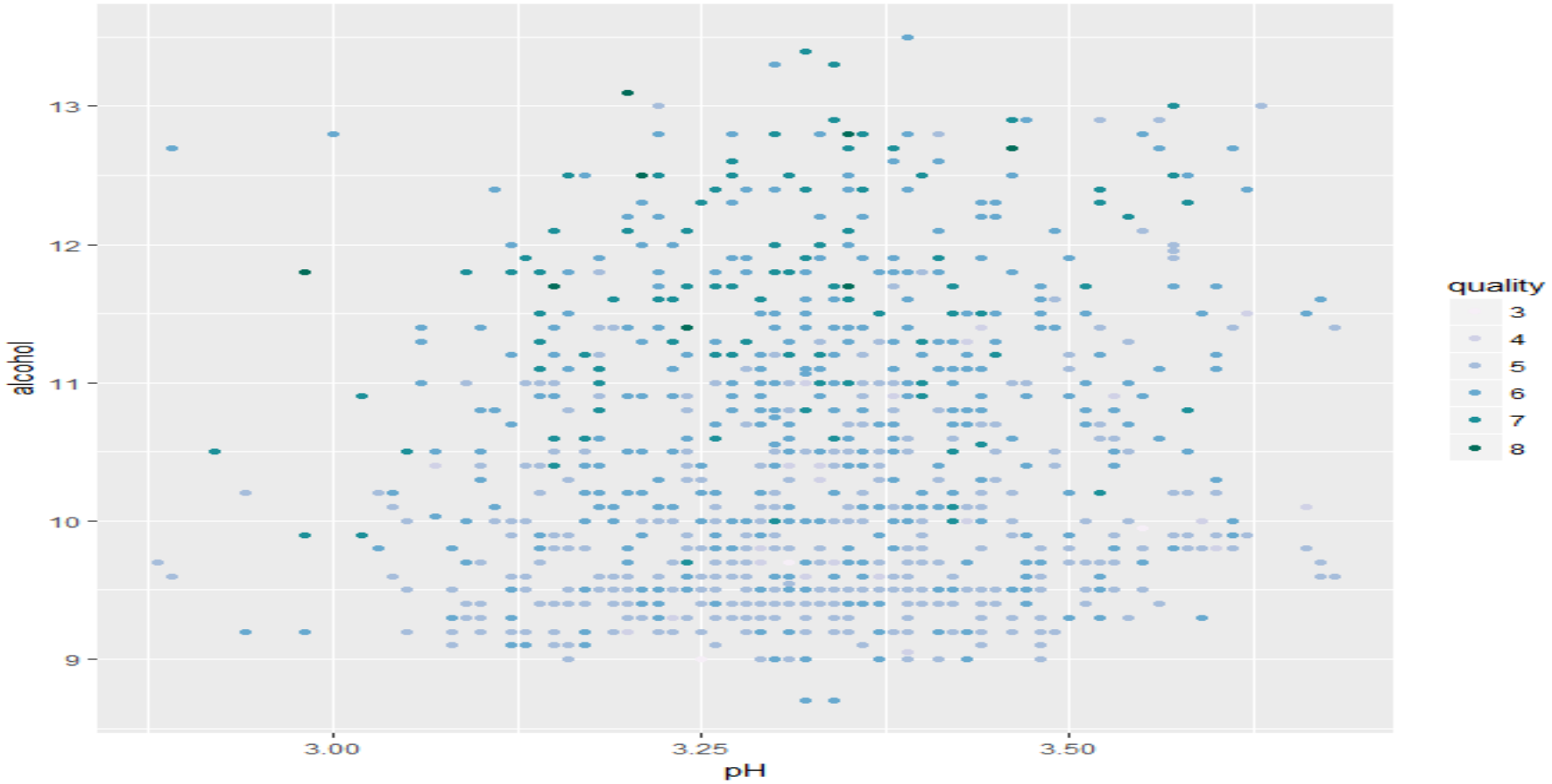
Volatile Acidity Vs Fixed Acidity



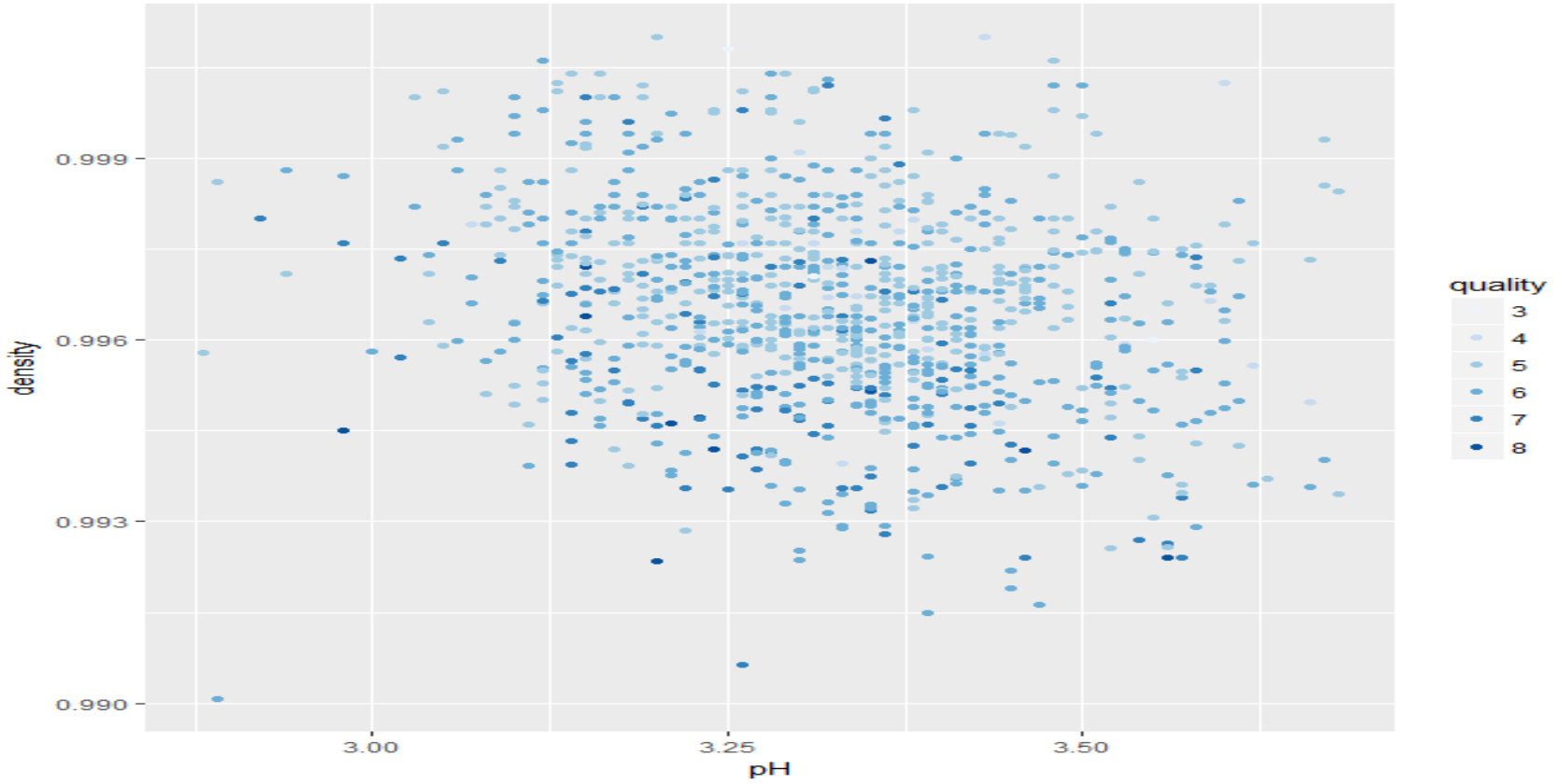
Total Sulfur Dioxide Vs Free Sulfur Dioxide



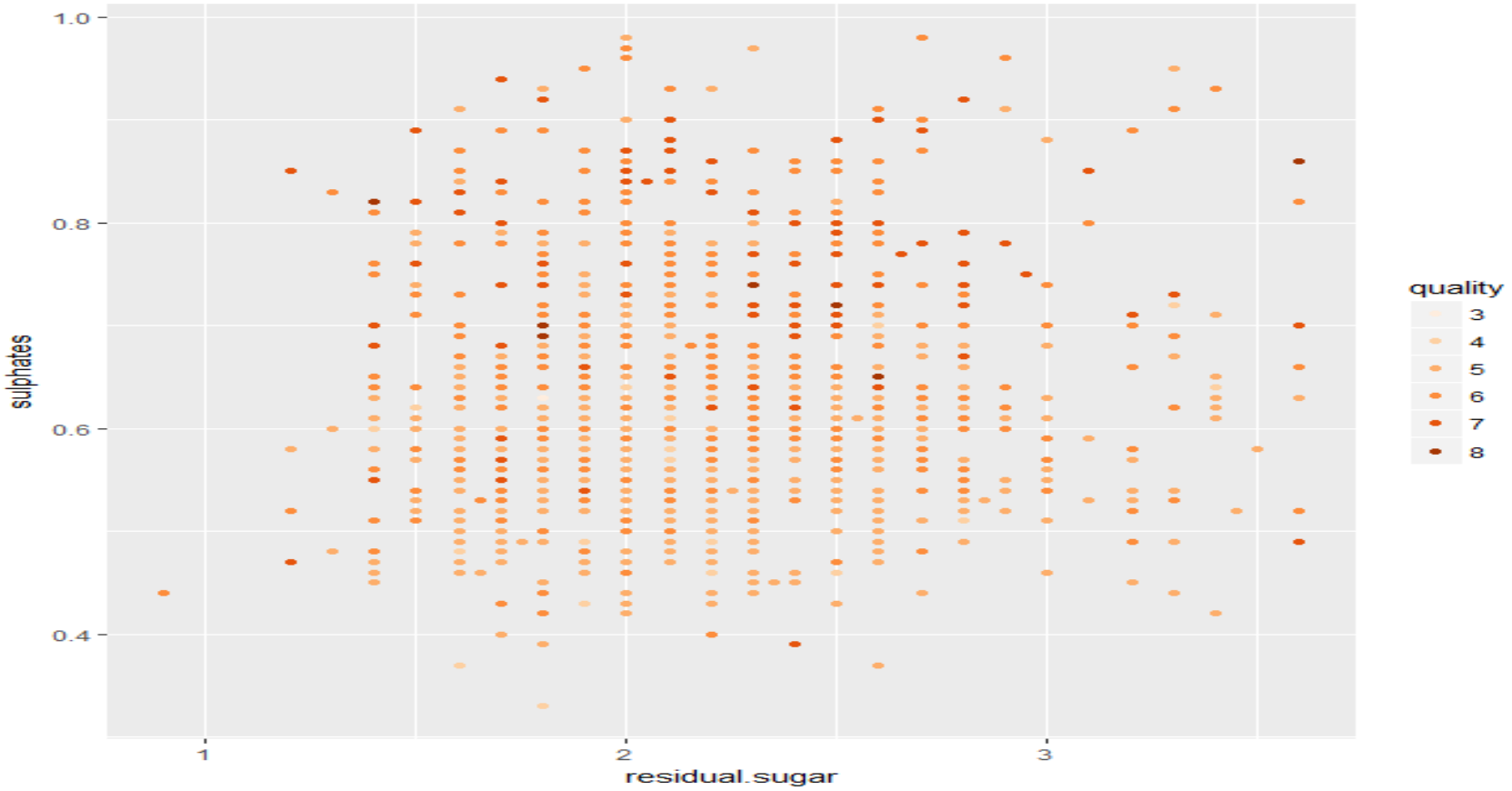
Alcohol Vs pH



Density Vs pH

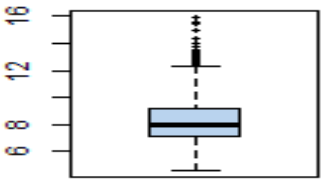


Sulphates Vs Residual Sugar

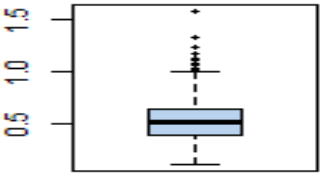


Data Cleaning

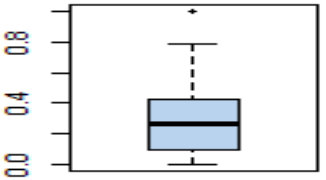
There were significant outliers in most of the predictors and we employed histograms to identify them. The outliers were then removed to proceed with Modelling.



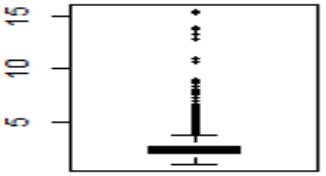
Fixed Acidity



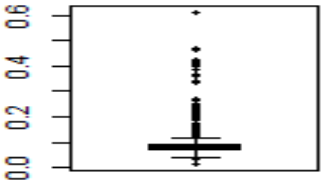
volatile.acidity



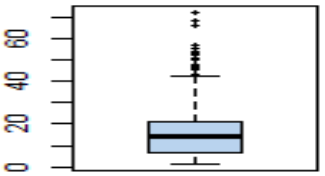
citric.acid



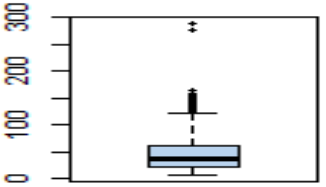
residual.sugar



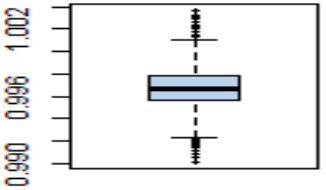
chlorides



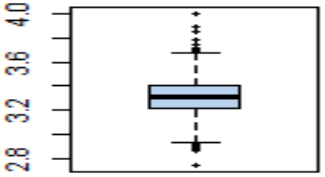
free.sulfur.dioxide



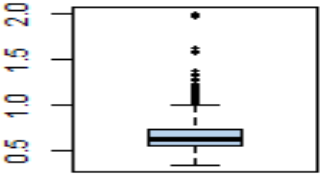
total.sulfur.dioxide



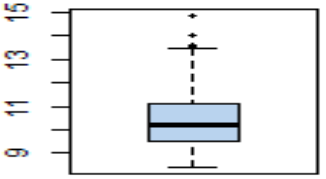
density



pH



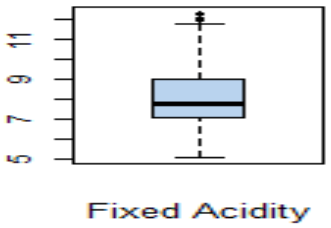
sulphates

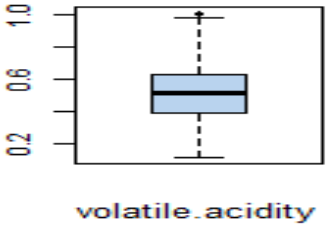


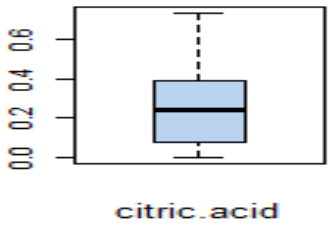
alcohol

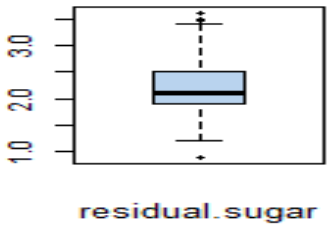
Data Cleaning

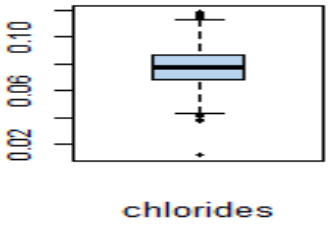
After the Outlier removal, the number of observations in Red wine was reduced to 1212 observations.

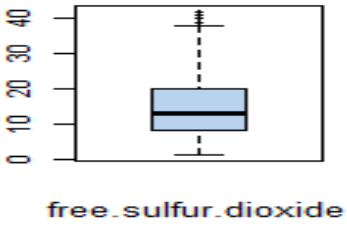


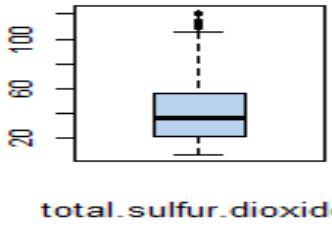


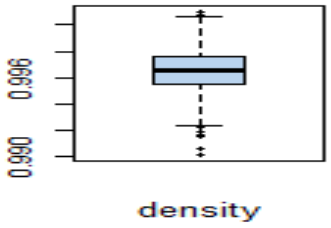


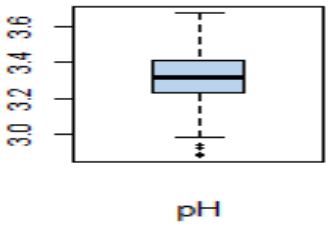


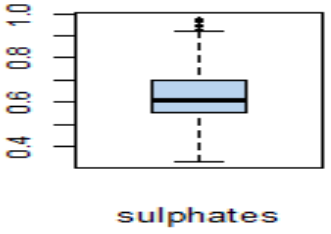


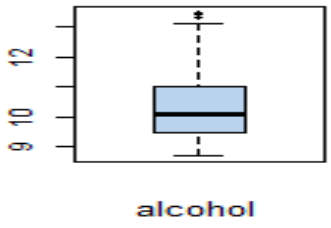




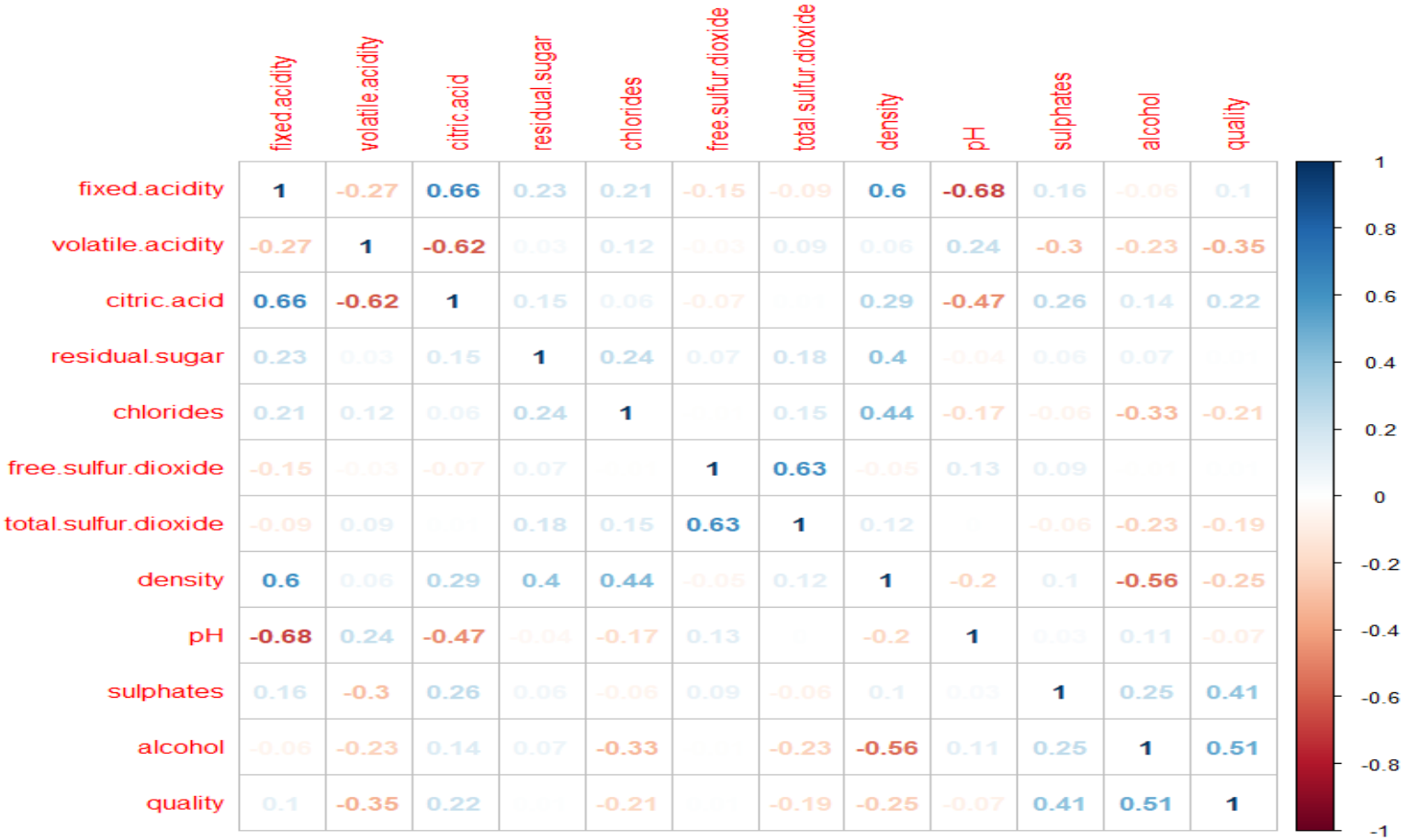








Correlation Plot for Red Wine Dataset



	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
fixed.acidity	1	-0.27	0.66	0.23	0.21	-0.15	-0.09	0.6	-0.68	0.16	-0.06	0.1
volatile.acidity	-0.27	1	-0.62	0.03	0.12	-0.03	0.09	0.06	0.24	-0.3	-0.23	-0.35
citric.acid	0.66	-0.62	1	0.15	0.06	-0.07	0.04	0.29	-0.47	0.26	0.14	0.22
residual.sugar	0.23	0.03	0.15	1	0.24	0.07	0.18	0.4	-0.04	0.06	0.07	0.01
chlorides	0.21	0.12	0.06	0.24	1	0.05	0.15	0.44	-0.17	-0.06	-0.33	-0.21
free.sulfur.dioxide	-0.15	-0.03	-0.07	0.07	0.05	1	0.63	-0.05	0.13	0.09	0.05	0.05
total.sulfur.dioxide	-0.09	0.09	0.04	0.18	0.15	0.63	1	0.12	0	-0.06	-0.23	-0.19
density	0.6	0.06	0.29	0.4	0.44	-0.05	0.12	1	-0.2	0.1	-0.56	-0.25
pH	-0.68	0.24	-0.47	-0.04	-0.17	0.13	0	-0.2	1	0.03	0.11	-0.07
sulphates	0.16	-0.3	0.26	0.06	-0.06	0.09	-0.06	0.1	0.03	1	0.25	0.41
alcohol	-0.06	-0.23	0.14	0.07	-0.33	0.05	-0.23	-0.56	0.11	0.25	1	0.51
quality	0.1	-0.35	0.22	0.01	-0.21	0.05	-0.19	-0.25	-0.07	0.41	0.51	1

Manual Selection of Variables

- The variables were selected manually by looking at the correlation plot.
- These variables were fitted into a linear Regression Model to check the Variance Inflation Factor (VIF)

1. Volatile Acidity
2. Citric acid
3. Density
4. pH
5. Sulphates

6 – Variable Model
Volatile Acidity
Citric Acid
Density
pH
Sulphates
Alcohol

- VIF was not above 10 for any of the variables . There is no proper correlation.

volatile.acidity	citric.acid	density	pH	sulphates	alcohol
1.896891	2.600279	2.149449	1.362884	1.265113	1.977165

Automatic Selection of Variables

8 Variable Model
Volatile Acidity
Citric Acid
Chlorides
Free Sulfur Dioxide
Total Sulfur Dioxide
pH
Sulphates
Alcohol

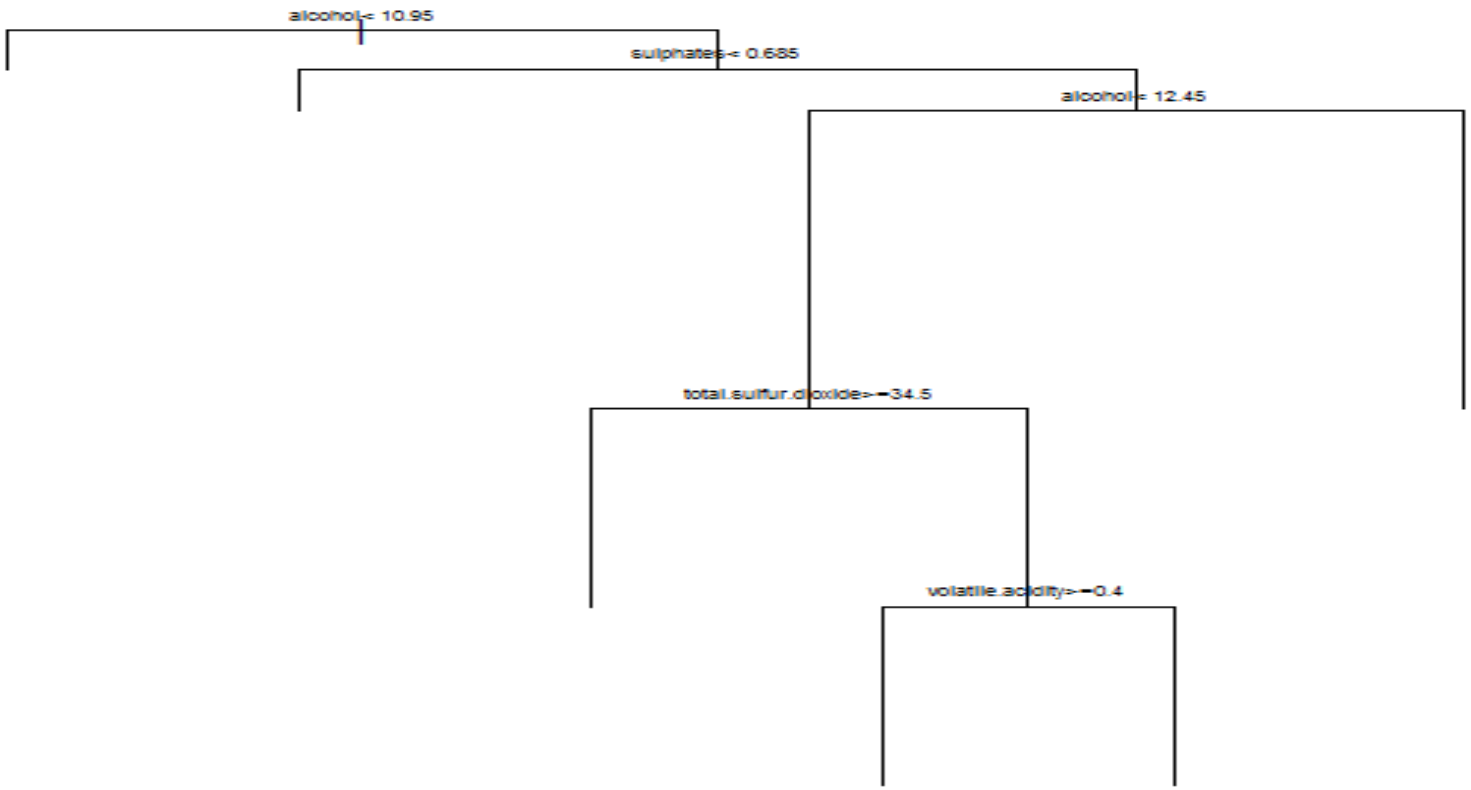
- The variables were selected based on Subset selection method.
- The nth variable model that subset selection method returned was in par with the model returned by Bootstrap method(8th variable model)
- Linear Regression was performed to check the VIF .

```
volatile.acidity        citric.acid           chlorides free.sulfur.dioxide total.sulfur.dioxide
        1.764914            2.081330            1.158510            1.769128            1.762358
              pH           sulphates             alcohol
        1.414034            1.204807            1.276195
```

- Again we see VIF does not give us much information about the correlation.

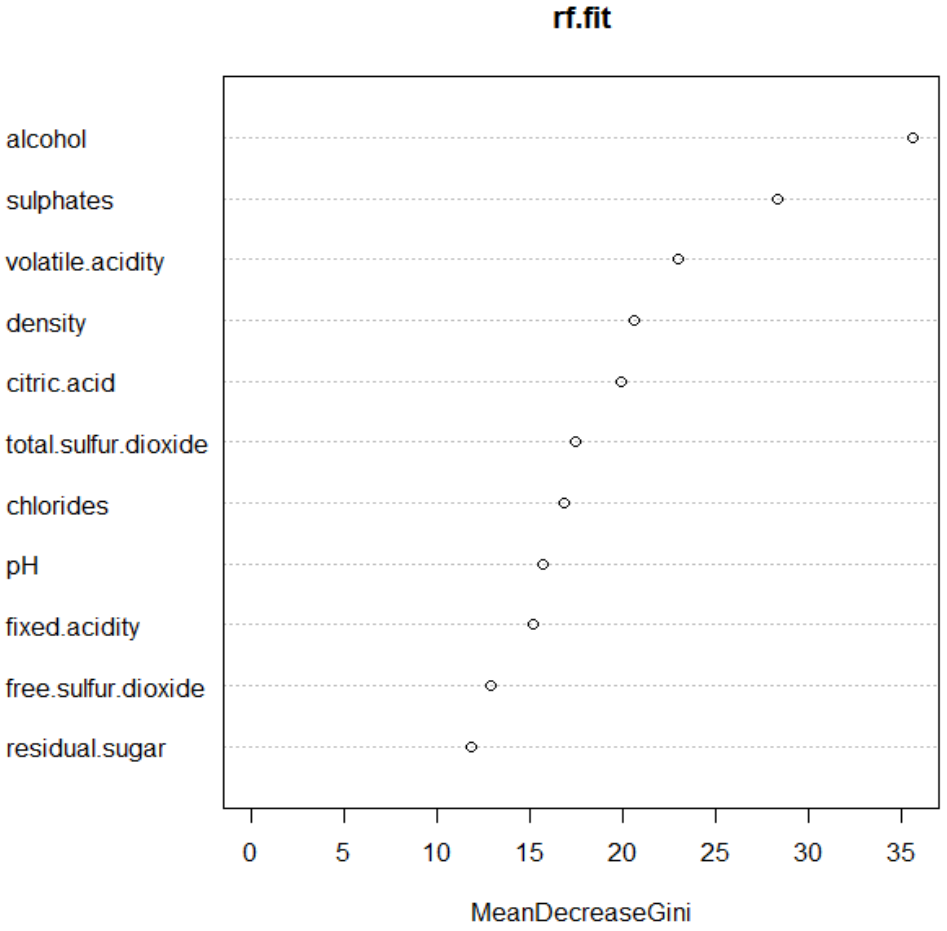
Decision Tree

From the Decision Tree, we can see that the Quality is highly dependent on Alchohol,Sulphates,volatile acidity and Total Sulphur Dioxide. The model achieved an accuracy rate of around 93.%



Random Forest

The Random Forest method obtained an accuracy of around 94.703%



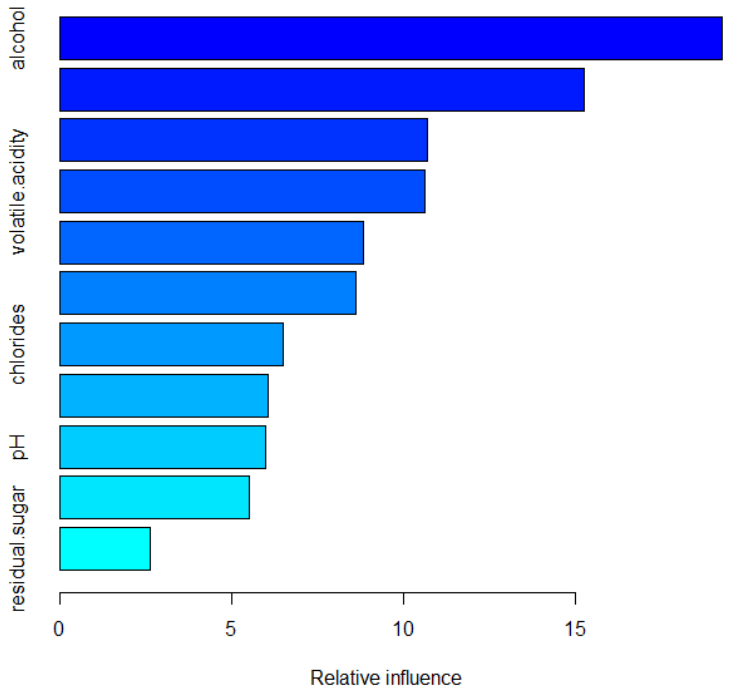
Importance(rf.fit)

fixed.acidity	15.22754
volatile.acidity	22.97286
citric.acid	19.93858
residual.sugar	11.89042
chlorides	16.87289
free.sulfur.dioxide	12.94350
total.sulfur.dioxide	17.49340
density	20.65072
ph	15.71268
sulphates	28.38139

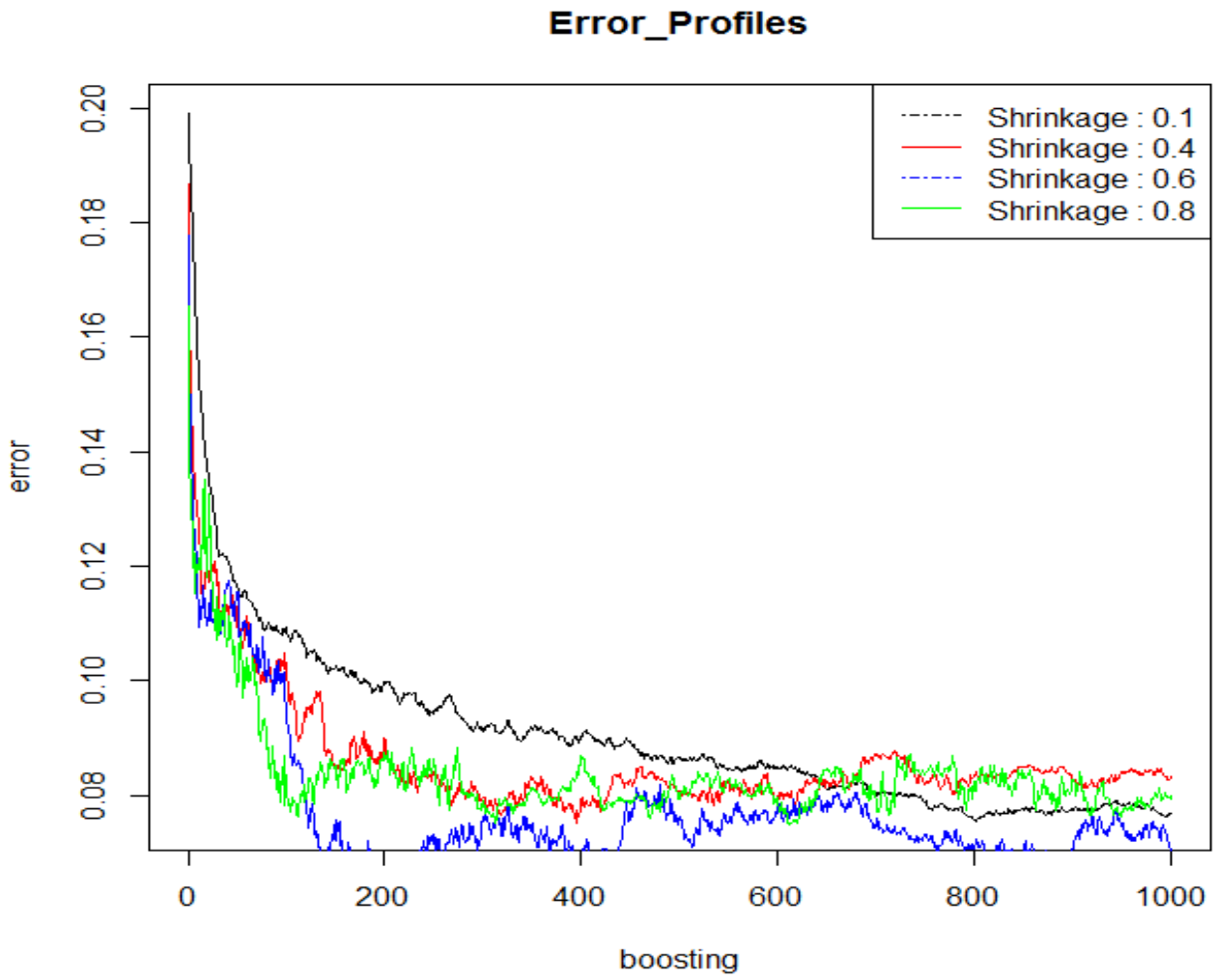
Boosting

Boosting shows relative importance of variables towards response

var	rel.inf
alcohol	19.833222
sulphates	15.999266
volatile.acidity	11.003070
density	9.777465
total.sulfur.dioxide	8.949794
pH	7.280583
free.sulfur.dioxide	6.505134
citric.acid	6.292692
chlorides	6.044518
fixed.acidity	5.858226
residual.sugar	2.456030



Boosting



Shrinkage	Error
Shrinkage:0.1	0.0769452
Shrinkage:0.4	0.0833345
Shrinkage:0.6	0.0706016
Shrinkage:0.8	0.0793255

Confusion Matrix for SVM – Radial Kernel

Original Set of Features

	NO	YES
NO	339	23
YES	11	26

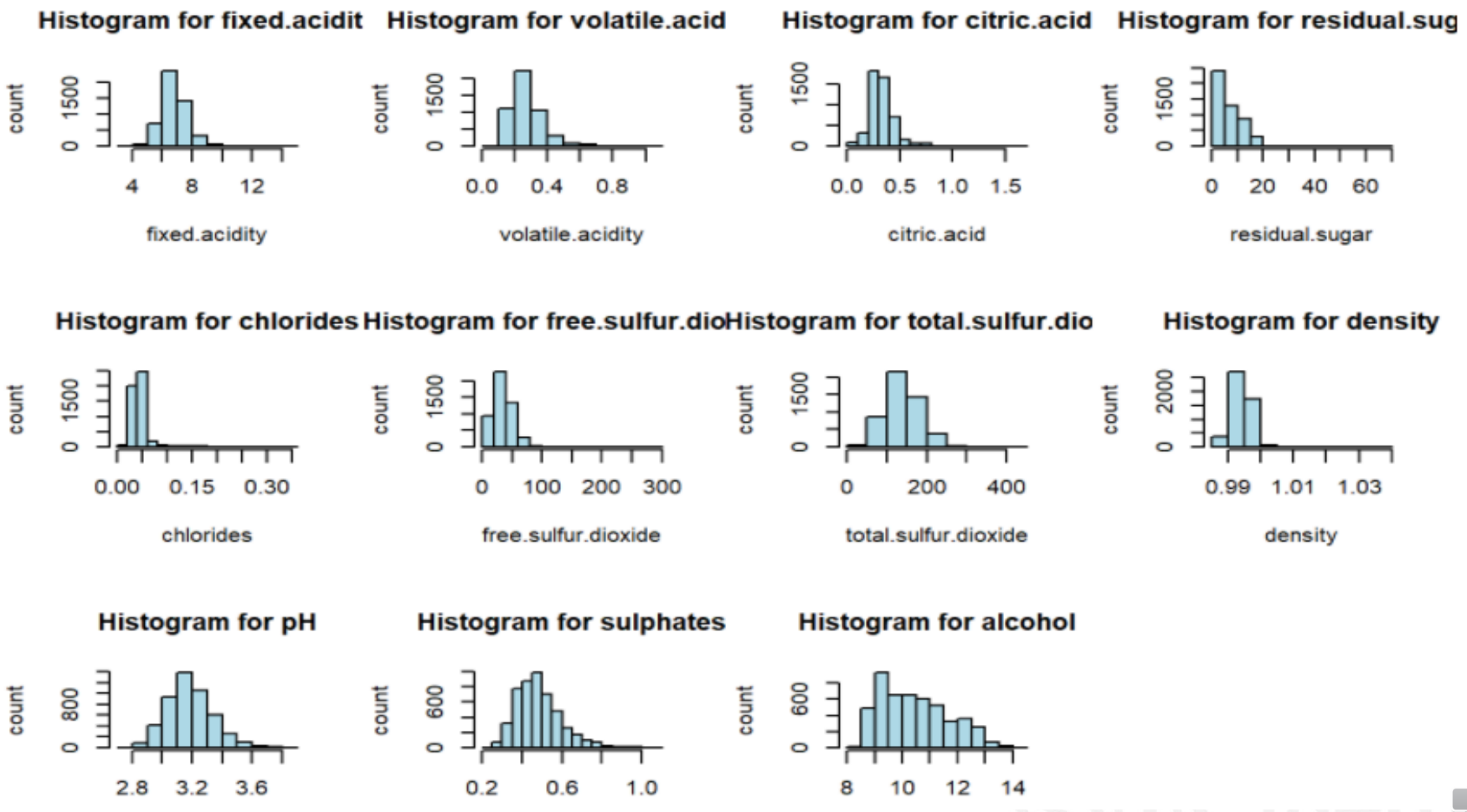
Manually Selected Features

	NO	YES
NO	329	22
YES	21	27

Features selected using Automatic selection methods

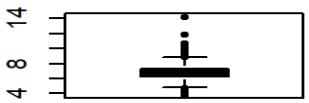
	NO	YES
NO	333	19
YES	17	30

Data Cleaning(White Wine)

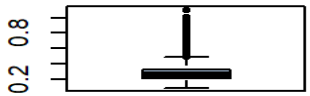


Data Cleaning(White Wine)

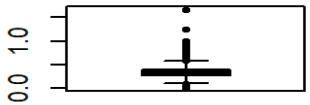
- In EDA, outlier identification and removal is most significant as it indicates bad data.



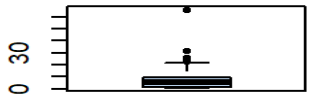
Fixed Acidity



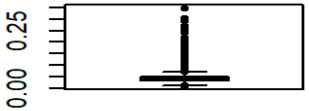
volatile Acidity



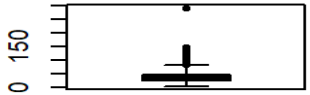
citric Acidity



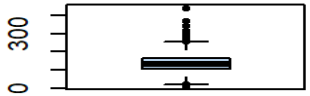
Residual sugar



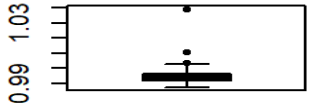
Chlorides



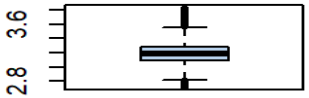
free sulfur dioxide



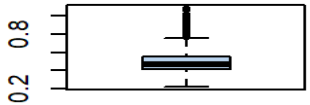
total sulfur dioxide



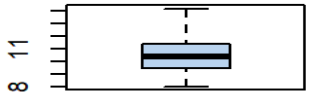
density



pH

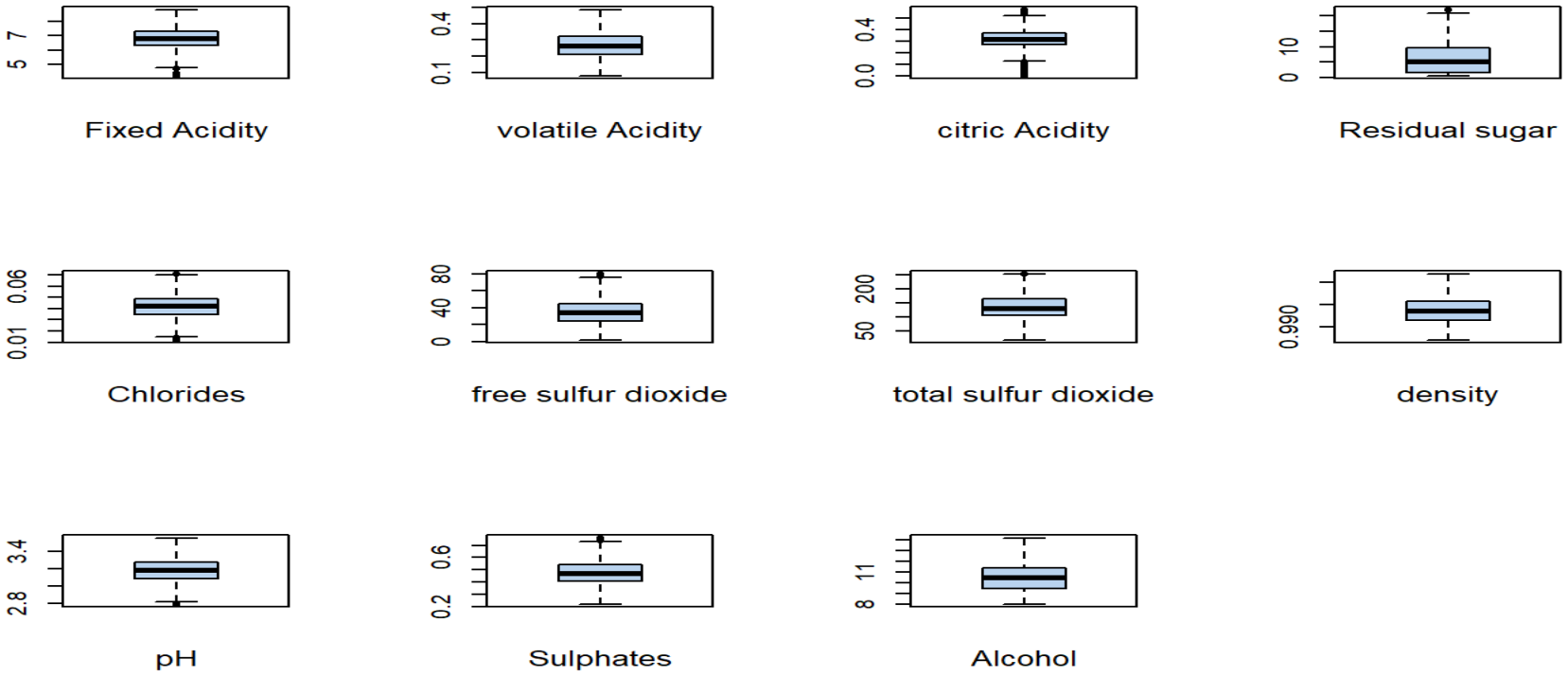


Sulphates



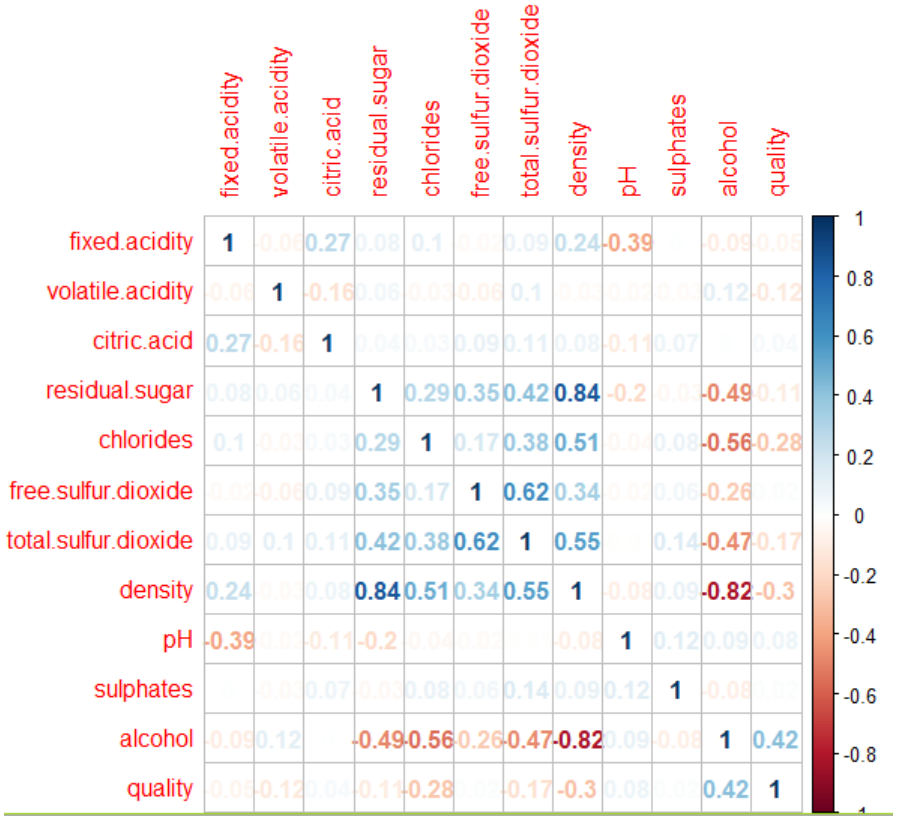
Alcohol

Data Cleaning(White Wine)



- As mostly outliers are on the larger side, we considered removal of outlier if it is greater than Q_3 + 1.5IQR
- After removing outliers, number of observations: white wine – 4074

Manual Feature Selection using Correlation



5-variable model
Volatile Acidity
Residual Sugar
Chlorides
pH
Alcohol

- Variance inflation factor (VIF) from the result of lm is checked to test density and it resulted in maximum value around 16.7
- If VIF is more than 10, multicollinearity is strongly suggested.

So, we move ahead with above 5-variable model

Manual Feature Selection using Correlation

- With 5-variable model selected, performed

  - Multiple linear regression

  - SVM

  to check the accuracy of the model
- Error rate in Multiple Regression – 56.6%
- Error in SVM with radial kernel – 19%

Automatic Feature Selection

- Performed best subset selection in which we get the models with significant variables
- AdjR^2 and Mallow's Cp: 9-variable model
- 10-fold cross validation - 9-variable model
- Linear regression is fitted to the Training data and checked for VIF
- VIF of density – 43.68
- Removed density and move ahead with 8-variable model
- These variables are significant variables

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.983e+02  2.637e+01   7.520 6.68e-14 ***
## fixed.acidity        1.578e-01  2.700e-02   5.844 5.50e-09 ***
## volatile.acidity    -1.868e+00  1.550e-01 -12.049  < 2e-16 ***
## residual.sugar       9.731e-02  9.970e-03   9.760  < 2e-16 ***
## chlorides           -3.297e+00  1.447e+00  -2.279 0.022726 *
## free.sulfur.dioxide  5.345e-03  8.386e-04   6.374 2.05e-10 ***
## density             -1.997e+02  2.672e+01  -7.472 9.61e-14 ***
## pH                   9.809e-01  1.281e-01   7.657 2.37e-14 ***
## sulphates            7.504e-01  1.257e-01   5.971 2.56e-09 ***
## alcohol              1.241e-01  3.377e-02   3.675 0.000241 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9-variable model
Fixed Acidity
Volatile Acidity
Residual Sugar
Chlorides
Free Sulphur Dioxide
Density
pH
Sulphates
Alcohol

Methods for White Wine

- Regression

i. Multiple Regression

ii. Regression Tree

iii. Boosting for Regression Tree

- Classification

i. Classification Tree

ii. Boosting for Classification Tree

iii. Support Vector Machine (radial kernel)

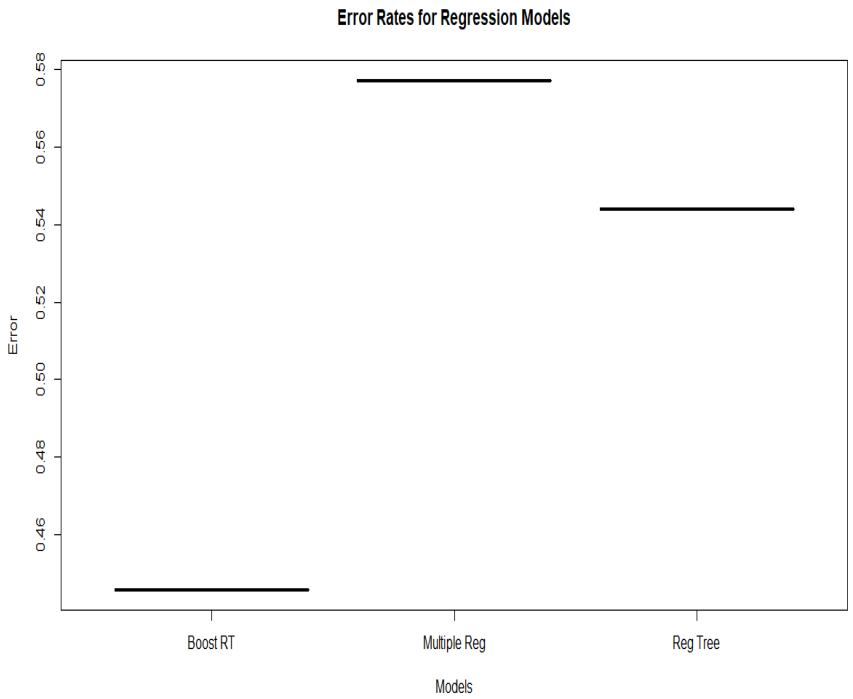
In classification, categorizing the wine quality as Good and Bad:

- When the quality <= 6, Bad
- When the quality >= 7, Good

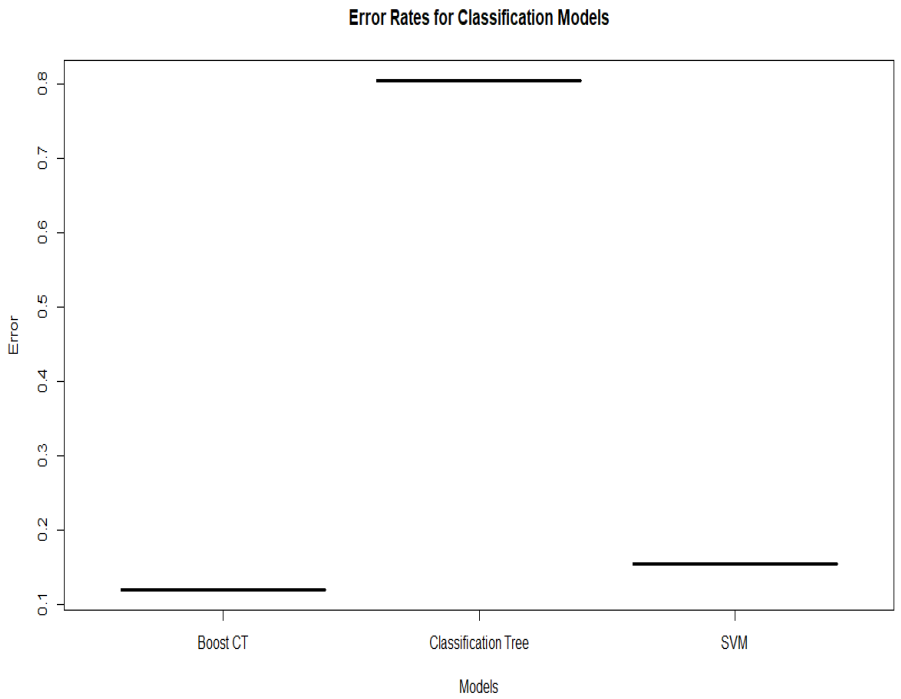
8-variable model
Fixed Acidity
Volatile Acidity
Residual Sugar
Chlorides
Free Sulphur Dioxide
pH
Sulphates
Alcohol

Accuracy of the model

Regression



Classification

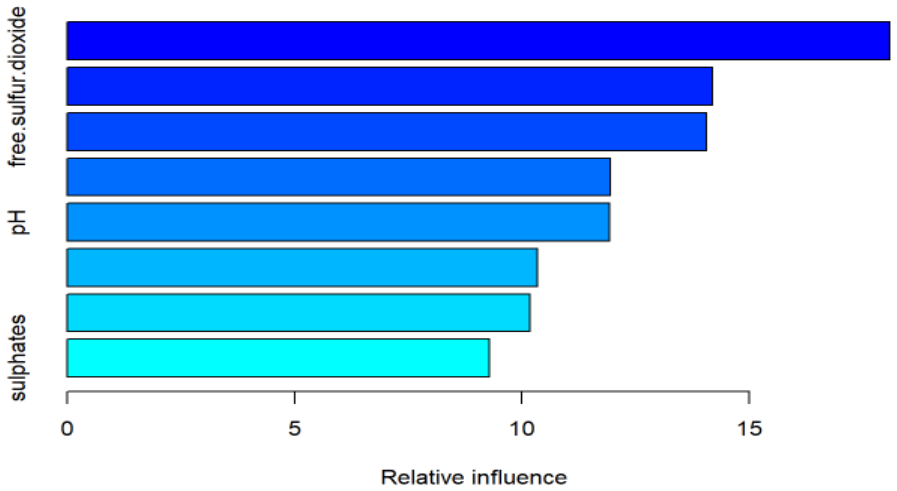


- Regression – Boosting gives less error; Classification – Boosting gives less error

Boosting for Regression Tree

- Good thing about boosting - shows relative importance of the variables towards the response
- Sequentially applies the weak classification algorithm to repeatedly modified versions of the data, thereby producing a powerful model

```
##                                 var   rel.inf
## alcohol                     alcohol 18.081948
## free.sulfur.dioxide free.sulfur.dioxide 14.186929
## residual.sugar           residual.sugar 14.058282
## volatile.acidity     volatile.acidity 11.937945
## pH                                pH 11.923468
## chlorides                   chlorides 10.343779
## fixed.acidity           fixed.acidity 10.180910
## sulphates                   sulphates  9.286738
```

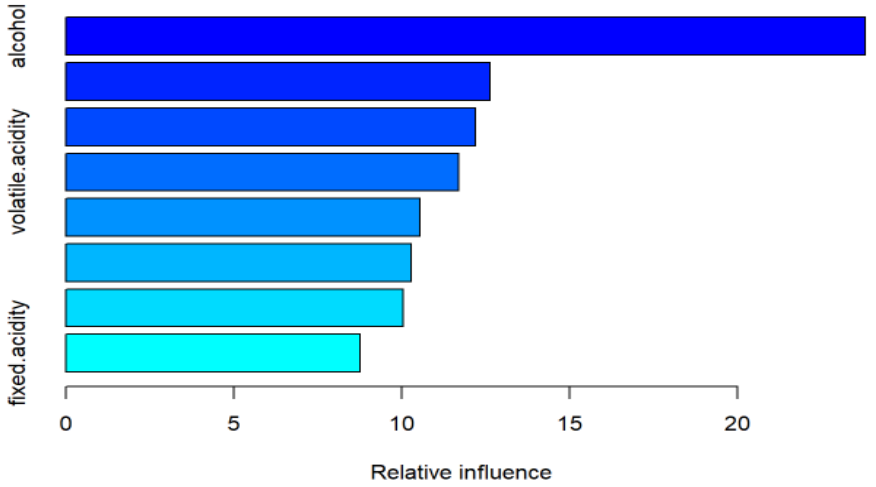


- In regression, while predicting the quality alcohol, free SO2 and residual sugar are most important.

Boosting for Classification Tree

- In classification, while predicting the class of quality alcohol, residual sugar and pH are most important.
- For best wine, the percentage of alcohol and the taste (pH) really matters

```
##                                  var   rel.inf
## alcohol                      alcohol 23.812213
## residual.sugar        residual.sugar 12.629944
## pH                                pH 12.210719
## volatile.acidity     volatile.acidity 11.694584
## free.sulfur.dioxide free.sulfur.dioxide 10.558240
## chlorides                  chlorides 10.281417
## sulphates                  sulphates 10.052990
## fixed.acidity          fixed.acidity  8.759892
```



Conclusion

White Wine Error Rates – Manual Feature Selection

Method Used	Error Rate(%)
Multiple Regression	56
SVM – Linear	24
SVM – Radial	19

White Wine Error Rates – Automatic Feature Selection - Regression

Method Used	Error Rate(%)
Multiple Regression	57.7
Regression Tree	54.3
Boosting with Regression Tree	44.0

Conclusion

White Wine Error Rates – Automatic Feature Selection - Classification

Method Used	Error Rate(%)
Classification Tree	80
Boosting	12
SVM – Linear	21
SVM - Radial	15.6

Conclusion

Red Wine Error Rates – Tree Based Methods

Method Used	Error Rate(%)
Decision Tree	7.69
Random Forest	5.37
Boosting	7.06

Red Wine Error Rates – Support Vector Machine – Radial Kernel

Feature Selection Method Used	Error Rate(%)
Manual	10.78
Automatic	9.2
All	8.53

Critical Questions

1) How can the histogram plots of variables of white wine by interpreted w.r.t outlier ?
2) We have used multiple regression and classification models for the predictions. According to you, which other predictive models we could have used for the analysis?