12/16/2019

Predicting Wine Quality

Laura Larregui, Angela Garcia, John Christman
IST 687 INTRODUCTION TO DATA SCIENCE
SYRACUSE UNIVERSITY

DATA SETS AND DISTRIBUTIONS

#Data Imports and Setup

```
readURL <- function(inputURL)  #Begin function named readURL that takes a URL


{
  csvFile <- read.csv(url(inputURL), sep = ';')  #assign the results of the URL call as a csv file to a dataframe named csvFile. Added sep = ';' to separate the data into columns
  return(csvFile)  # return the dataframe
}

redWine <- readURL("https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv")
whiteWine <- readURL("https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv")

str(redWine)

## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...

str(whiteWine)

## '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 ...
```

```r
summary(redWine)
```

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar  
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900  
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900  
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200  
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539  
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600  
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500  
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density      
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901  
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956  
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968  
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967  
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978  
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037  
##        pH          sulphates         alcohol         quality     
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000  
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000  
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000  
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636  
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000  
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000  
```

```r
summary(whiteWine)
```

```
##  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    density      
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871  
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917  
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937  
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940  
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961  
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390  
##        pH          sulphates         alcohol         quality     
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000  
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000  
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000  
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878  
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000  
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000  
```

```r
#The datasets only have one column of data.  The column names are separated by
periods the data by semi-colons
#1. Create columns
#2.  separate the data into the columns
#3.  Verify no NAs
redWine <- na.omit(redWine)
whiteWine<-na.omit (whiteWine)
```

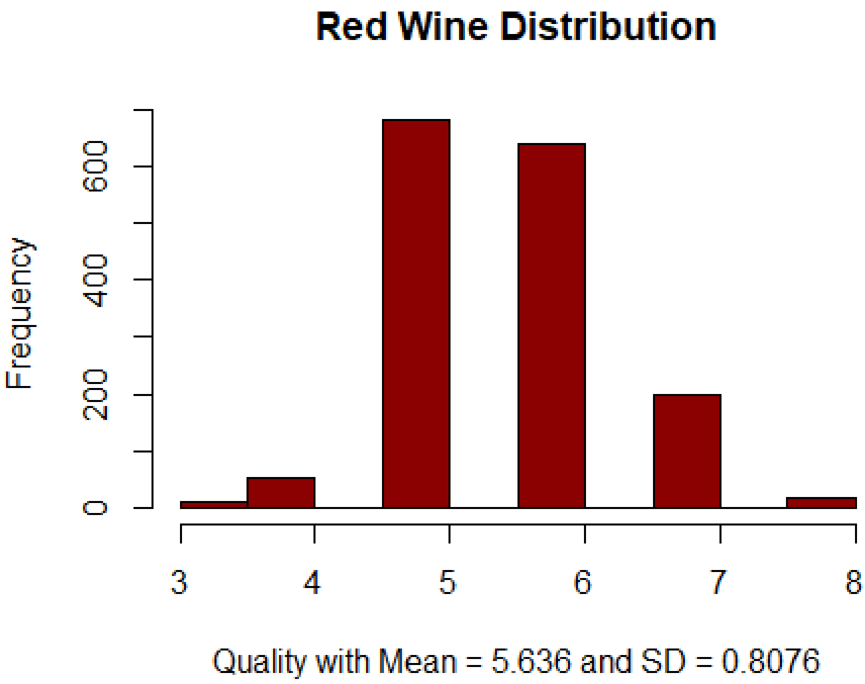
#Standard Deviations

```r
#Quality Standard Deviations
sd(redWine$quality)

## [1] 0.8075694

sd(whiteWine$quality)

## [1] 0.8856386

#Red wine Standard Deviations
sd(redWine$quality)

## [1] 0.8075694

sd(redWine$alcohol)

## [1] 1.065668

sd(redWine$residual.sugar)

## [1] 1.409928

sd(redWine$pH)

## [1] 0.1543865

#White wine standard deviations
sd(whiteWine$quality)

## [1] 0.8856386

sd(whiteWine$alcohol)

## [1] 1.230621

sd(whiteWine$residual.sugar)

## [1] 5.072058

sd(whiteWine$pH)

## [1] 0.1510006
```

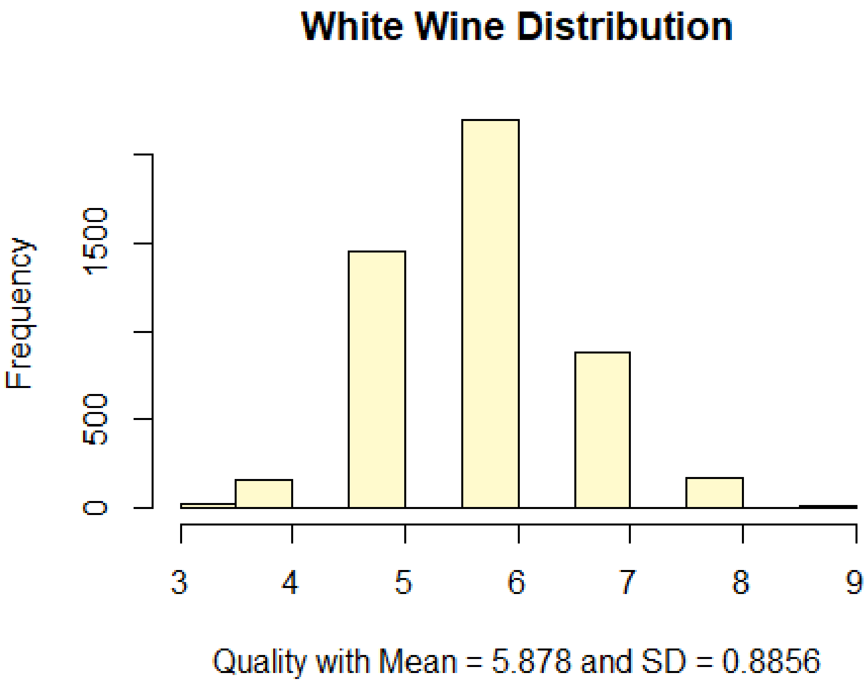
#1. Create visualizations for the data #heat maps, histograms and scatter plots

#Histograms

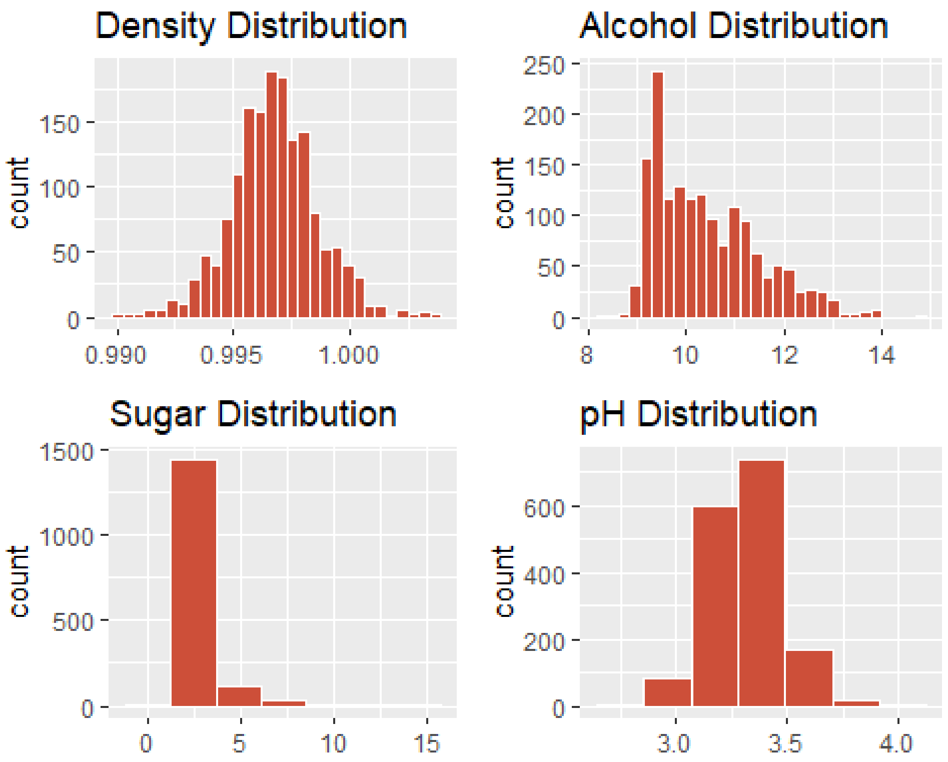
```r
hist(redWine$quality, main = "Red Wine Distribution", xlab = "Quality with Mean =
5.636 and SD = 0.8076", col ="red4")
```



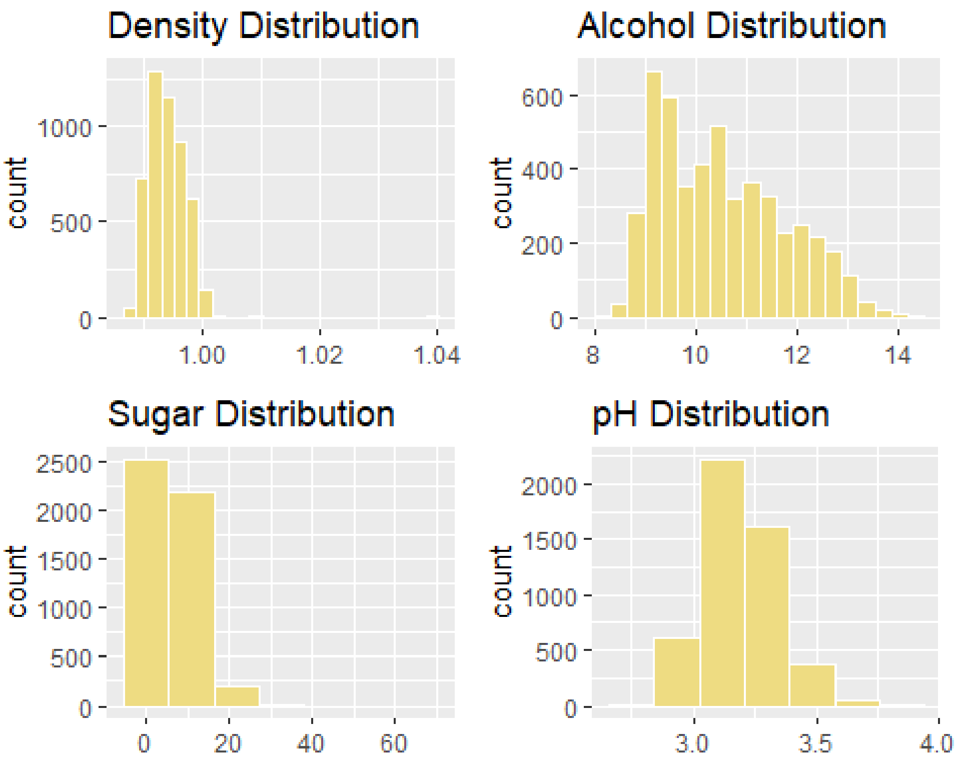
```
hist(whiteWine$quality, main = "White Wine Distribution", xlab = "Quality with Mean
= 5.878 and SD = 0.8856", col = "lemonchiffon")

library(grid, warn.conflicts = FALSE) # Eliminate warning when library is installed
```



```
library(gridExtra, warn.conflicts = FALSE)
library (ggplot2, warn.conflicts = FALSE)

h1 <- ggplot(aes(density), data = redWine) + geom_histogram(bins = 30,fill=
"tomato3",color="white")
h1 <- h1 + ggtitle("Density Distribution") +theme(axis.title.x = element_blank())

h2 <- ggplot(aes(alcohol), data = redWine) + geom_histogram(bins = 30, fill=
"tomato3",color="white")
h2 <- h2 + ggtitle("Alcohol Distribution") +theme(axis.title.x = element_blank())

h3 <- ggplot(aes(residual.sugar), data = redWine) + geom_histogram(bins = 7, fill=
"tomato3",color="white")
h3 <- h3 + ggtitle("Sugar Distribution") + theme(axis.title.x = element_blank())

h4 <- ggplot(aes(pH), data = redWine) + geom_histogram(bins = 7, fill=
"tomato3",color="white")
h4 <- h4 + ggtitle("pH Distribution")+theme(axis.title.x = element_blank())

grid.arrange(h1,h2,h3,h4,ncol=2)
```


Density Distribution
Alcohol Distribution
Sugar Distribution
pH Distribution

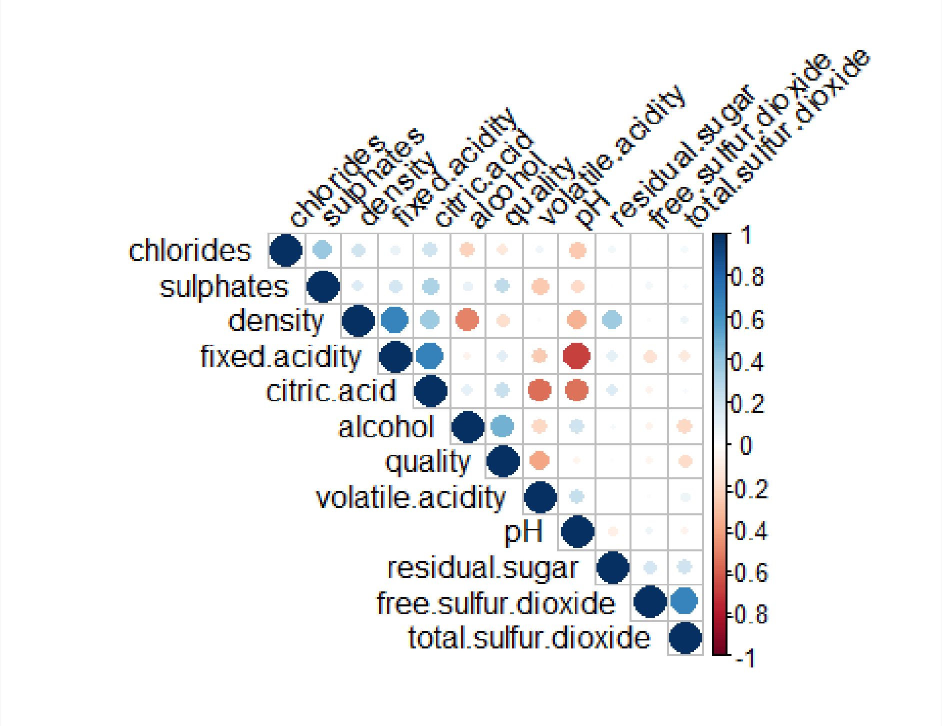
```
g1 <- ggplot(aes(density), data = whiteWine) + geom_histogram(bins = 25,fill=
"lightgoldenrod2",color="white")
g1 <- g1 +  ggtitle("Density Distribution") + theme(axis.title.x = element_blank())

g2 <- ggplot(aes(alcohol), data = whiteWine) + geom_histogram(bins = 20, fill=
"lightgoldenrod2",color="white")
g2 <- g2 + ggtitle("Alcohol Distribution") + theme(axis.title.x = element_blank())

g3 <- ggplot(aes(residual.sugar), data = whiteWine) + geom_histogram(bins = 7,
fill= "lightgoldenrod2",color="white")
g3 <- g3 + ggtitle("Sugar Distribution") + theme(axis.title.x = element_blank())

g4 <- ggplot(aes(pH), data = whiteWine) + geom_histogram(bins = 7, fill=
"lightgoldenrod2",color="white")
g4 <- g4 + ggtitle("pH Distribution") + theme(axis.title.x = element_blank())

grid.arrange(g1,g2,g3,g4,ncol=2)
```



#Correlation matrix

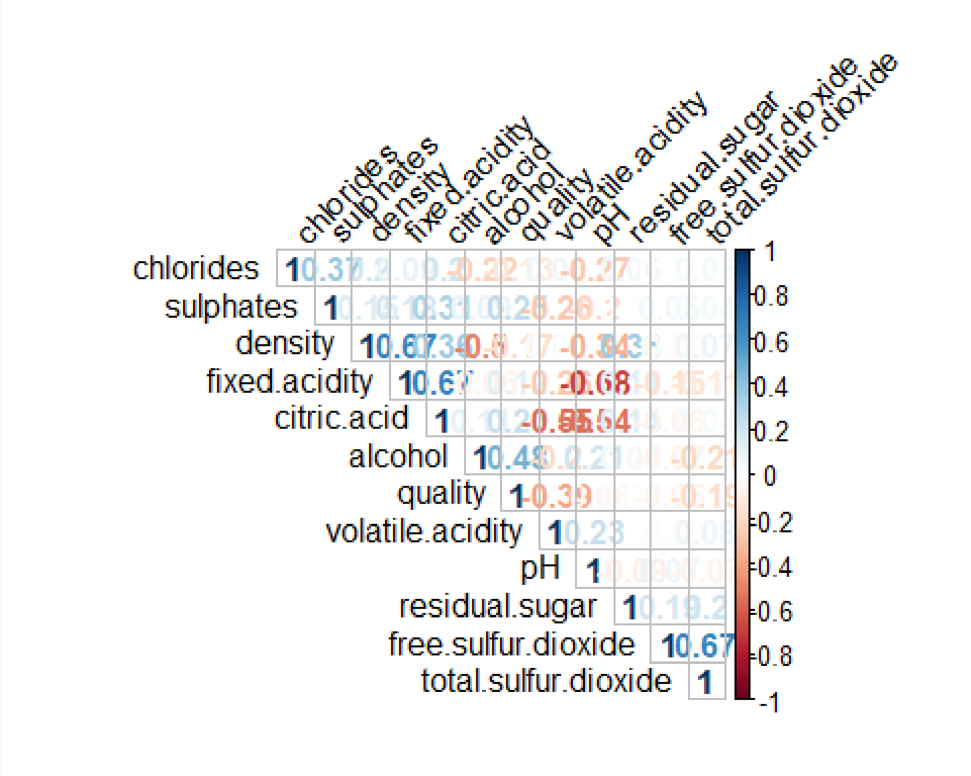
```
#1. Create the correlation matrix. ##Reference: http://www.sthda.com/english/wiki/correlation-matrix-a-quick-start-guide-to-analyze-format-and-visualize-a-correlation-matrix-using-r-software
#Red Wine Correlation Matrix
#install.packages("corrplot")
library(corrplot, warn.conflicts = FALSE)

## corrplot 0.84 loaded

red_cor <- cor(redWine)
round(red_cor, 2)

##                      fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity                 1.00            -0.26        0.67           0.11
## volatile.acidity             -0.26             1.00       -0.55           0.00
## citric.acid                   0.67            -0.55        1.00           0.14
## residual.sugar                0.11             0.00        0.14           1.00
## chlorides                     0.09             0.06        0.20           0.06
## free.sulfur.dioxide          -0.15            -0.01       -0.06           0.19
## total.sulfur.dioxide         -0.11             0.08        0.04           0.20
## density                       0.67             0.02        0.36           0.36
## pH                           -0.68             0.23       -0.54          -0.09
## sulphates                     0.18            -0.26        0.31           0.01
## alcohol                      -0.06            -0.20        0.11           0.04
## quality                       0.12            -0.39        0.23           0.01
##                      chlorides free.sulfur.dioxide total.sulfur.dioxide density
## fixed.acidity             0.09               -0.15                -0.11    0.67
## volatile.acidity          0.06               -0.01                 0.08    0.02
## citric.acid               0.20               -0.06                 0.04    0.36
## residual.sugar            0.06                0.19                 0.20    0.36
## chlorides                 1.00                0.01                 0.05    0.20
## free.sulfur.dioxide       0.01                1.00                 0.67   -0.02
## total.sulfur.dioxide      0.05                0.67                 1.00    0.07
## density                   0.20               -0.02                 0.07    1.00
## pH                       -0.27                0.07                -0.07   -0.34
## sulphates                 0.37                0.05                 0.04    0.15
## alcohol                  -0.22               -0.07                -0.21   -0.50
## quality                  -0.13               -0.05                -0.19   -0.17
##                         pH sulphates alcohol quality
## fixed.acidity        -0.68      0.18   -0.06    0.12
## volatile.acidity      0.23     -0.26   -0.20   -0.39
## citric.acid          -0.54      0.31    0.11    0.23
## residual.sugar       -0.09      0.01    0.04    0.01
## chlorides            -0.27      0.37   -0.22   -0.13
## free.sulfur.dioxide   0.07      0.05   -0.07   -0.05
## total.sulfur.dioxide -0.07      0.04   -0.21   -0.19
## density              -0.34      0.15   -0.50   -0.17
## pH                    1.00     -0.20    0.21   -0.06
## sulphates            -0.20      1.00    0.09    0.25
## alcohol               0.21      0.09    1.00    0.48
## quality              -0.06      0.25    0.48    1.00
```

```
corrplot(red_cor, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



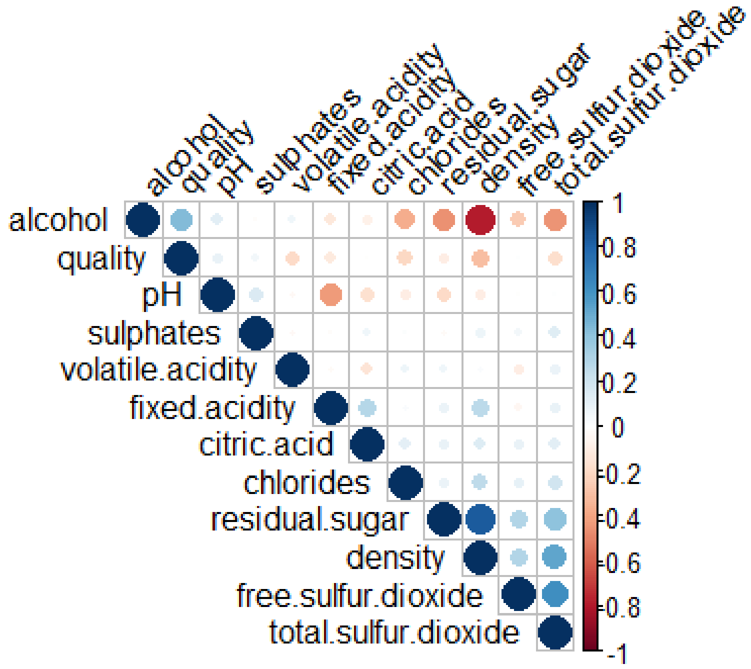
#Positive correlations are displayed in blue and negative correlations in red color. Color intensity and the size of the circle are proportional to the correlation coefficients

```
#Correlation matrix with numbers
corrplot(red_cor, method = 'number', type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```

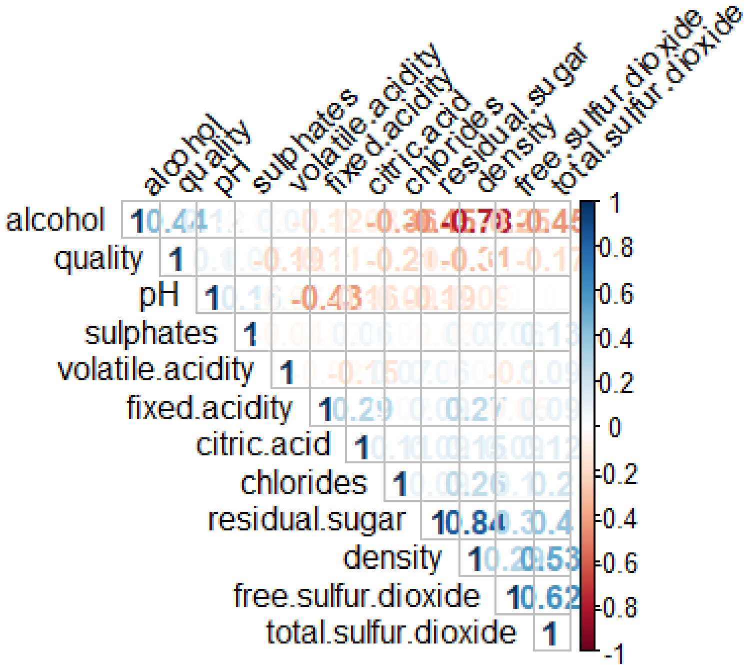


```
#White Wine Correlation Matrix

library(corrplot,warn.conflicts = FALSE)
white_cor <- cor(whiteWine)
round(white_cor, 2)

##                      fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity                 1.00            -0.02        0.29           0.09
## volatile.acidity             -0.02             1.00       -0.15           0.06
## citric.acid                   0.29            -0.15        1.00           0.09
## residual.sugar                0.09             0.06        0.09           1.00
## chlorides                     0.02             0.07        0.11           0.09
## free.sulfur.dioxide          -0.05            -0.10        0.09           0.30
## total.sulfur.dioxide          0.09             0.09        0.12           0.40
## density                       0.27             0.03        0.15           0.84
## pH                           -0.43            -0.03       -0.16          -0.19
## sulphates                    -0.02            -0.04        0.06          -0.03
## alcohol                      -0.12             0.07       -0.08          -0.45
## quality                      -0.11            -0.19       -0.01          -0.10
##                      chlorides free.sulfur.dioxide total.sulfur.dioxide density
## fixed.acidity             0.02               -0.05                 0.09    0.27
## volatile.acidity          0.07               -0.10                 0.09    0.03
## citric.acid               0.11                0.09                 0.12    0.15
## residual.sugar            0.09                0.30                 0.40    0.84
## chlorides                 1.00                0.10                 0.20    0.26
## free.sulfur.dioxide       0.10                1.00                 0.62    0.29
## total.sulfur.dioxide      0.20                0.62                 1.00    0.53
## density                   0.26                0.29                 0.53    1.00
## pH                       -0.09                0.00                 0.00   -0.09
## sulphates                 0.02                0.06                 0.13    0.07
## alcohol                  -0.36               -0.25                -0.45   -0.78
## quality                  -0.21                0.01                -0.17   -0.31
##                         pH sulphates alcohol quality
## fixed.acidity        -0.43     -0.02   -0.12   -0.11
## volatile.acidity     -0.03     -0.04    0.07   -0.19
## citric.acid          -0.16      0.06   -0.08   -0.01
## residual.sugar       -0.19     -0.03   -0.45   -0.10
## chlorides            -0.09      0.02   -0.36   -0.21
## free.sulfur.dioxide   0.00      0.06   -0.25    0.01
## total.sulfur.dioxide  0.00      0.13   -0.45   -0.17
## density              -0.09      0.07   -0.78   -0.31
## pH                    1.00      0.16    0.12    0.10
## sulphates             0.16      1.00   -0.02    0.05
## alcohol               0.12     -0.02    1.00    0.44
## quality               0.10      0.05    0.44    1.00
```

```
corrplot(white_cor, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



```
#Correlation matrix with numbers
corrplot(white_cor, method = 'number',type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



```
#Quality correlations
dfred_cor<-data.frame(red_cor)
dfwhite_cor<-data.frame(white_cor)
names<-row.names(dfred_cor)

QualityCor<-round(cbind(dfred_cor$quality,dfwhite_cor$quality),digits=4)
colnames(QualityCor)<-c("Red Quality", "White Quality")
row.names(QualityCor)<-names
QualityCor

##                        Red Quality White Quality
## fixed.acidity               0.1241       -0.1137
## volatile.acidity           -0.3906       -0.1947
## citric.acid                 0.2264       -0.0092
## residual.sugar              0.0137       -0.0976
## chlorides                  -0.1289       -0.2099
## free.sulfur.dioxide        -0.0507        0.0082
## total.sulfur.dioxide       -0.1851       -0.1747
## density                    -0.1749       -0.3071
## pH                         -0.0577        0.0994
## sulphates                   0.2514        0.0537
## alcohol                     0.4762        0.4356
## quality                     1.0000        1.0000

#table preview
knitr::kable(head(redWine))
```

```
        volat                           free.  total
 fixed  ile.a  citri  resid             sulfu  .sulf
 .acid  cidit  c.aci  ual.s  chlor      r.dio  ur.di  densi          sulph  alcoh  quali
   ity      y      d   ugar   ides       xide  oxide     ty     pH   ates     ol     ty
-----------------------------------------------------------------------------------------
   7.4   0.70   0.00    1.9  0.076         11     34 0.9978   3.51   0.56    9.4      5
   7.8   0.88   0.00    2.6  0.098         25     67 0.9968   3.20   0.68    9.8      5
   7.8   0.76   0.04    2.3  0.092         15     54 0.9970   3.26   0.65    9.8      5
  11.2   0.28   0.56    1.9  0.075         17     60 0.9980   3.16   0.58    9.8      6
   7.4   0.70   0.00    1.9  0.076         11     34 0.9978   3.51   0.56    9.4      5
   7.4   0.66   0.00    1.8  0.075         13     40 0.9978   3.51   0.56    9.4      5
```

```
#Converting the quality attribute from numeric to factor
redWine$bquality <- ifelse(redWine$quality < 5, "Mediocre", ifelse(redWine$quality
<7 , "Average", ifelse(redWine$quality >6, "Excellent", NA)))

whiteWine$bquality <- ifelse(whiteWine$quality < 5, "Mediocre",
ifelse(whiteWine$quality <7 , "Average", ifelse(whiteWine$quality >6, "Excellent",
NA)))
```

Exploratory Analysis

```r
readURL <- function(inputURL)  #Begin function named readURL that takes a URL
{
  csvFile <- read.csv(url(inputURL), sep = ';')  #assign the results of the URL call as a csv file to a dataframe named csvFile. Added sep = ';' to seperate the data into columns
  return(csvFile)  # return the dataframe
}

redWine <- readURL("https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv")
whiteWine <- readURL("https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-white.csv")
```

#Tree Models

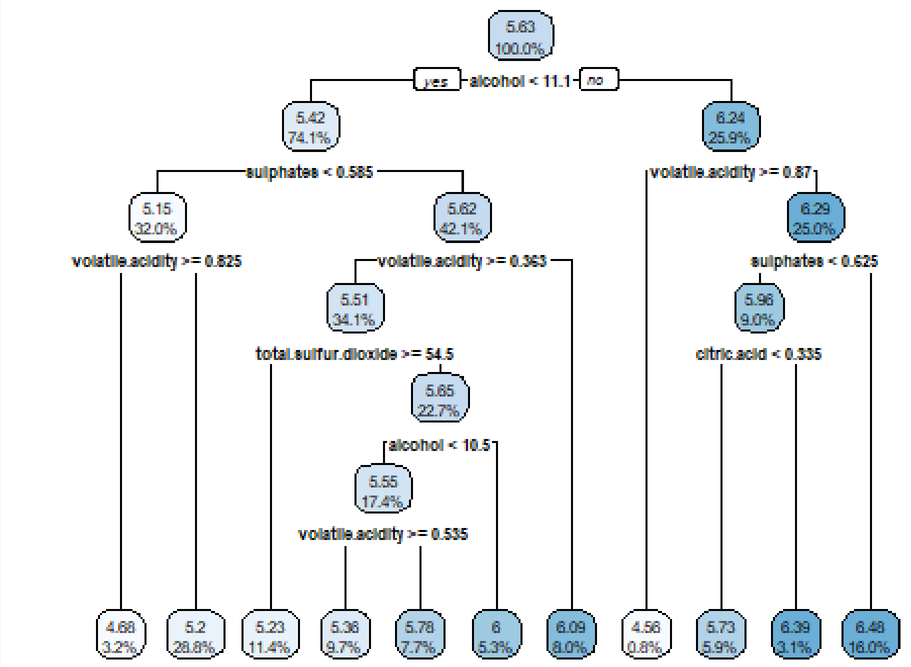
##Model Training (Regression Tree for Red Wine)
#Reference:https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart

```r
#install.packages("rpart")
#install.packages("rpart.plot")
#install.packages("rattle")

library(rpart,warn.conflicts = FALSE)
library(rpart.plot,warn.conflicts = FALSE)
library(rattle, warn.conflicts = FALSE)

## Rattle: A free graphical interface for data science with R.
## Version 5.2.0 Copyright (c) 2006-2018 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

#Recursive Partitioning and Regression Trees **Red Wine**

nrows<-nrow(redWine)
cutPoint<- floor(nrows/3*2)
cutPoint

## [1] 1066

rand<-sample(1:nrows)
#training set Red Wine
red_train <- redWine[rand[1:cutPoint],]

#test set Red Wine
red_test <- redWine[rand[(cutPoint+1:nrows)],]
red_test<-na.omit(red_test)

w.rpart <- rpart(quality ~. , data = red_train)
```

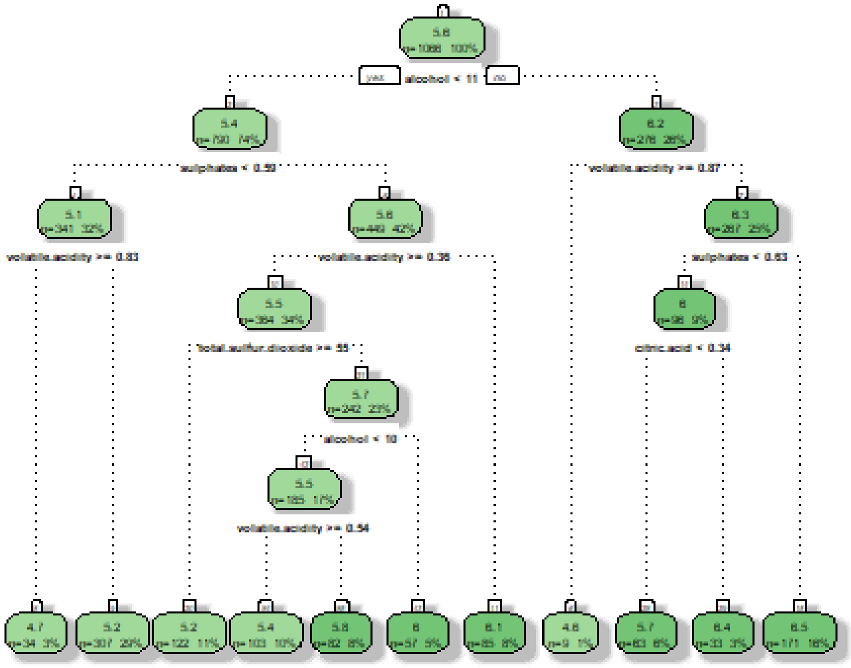
```
w.rpart
```

```
## n= 1066
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 1066 710.893100 5.628518
##    2) alcohol< 11.05 790 389.986100 5.416456
##      4) sulphates< 0.585 341 118.668600 5.146628
##        8) volatile.acidity>=0.825 34  23.441180 4.676471 *
##        9) volatile.acidity< 0.825 307  86.879480 5.198697 *
##      5) sulphates>=0.585 449 227.634700 5.621381
##       10) volatile.acidity>=0.3625 364 158.956000 5.510989
##         20) total.sulfur.dioxide>=54.5 122  29.573770 5.229508 *
##         21) total.sulfur.dioxide< 54.5 242 114.843000 5.652893
##           42) alcohol< 10.45 185  83.859460 5.545946
##             84) volatile.acidity>=0.535 103  41.708740 5.359223 *
##             85) volatile.acidity< 0.535 82  34.048780 5.780488 *
##           43) alcohol>=10.45 57  22.000000 6.000000 *
##       11) volatile.acidity< 0.3625 85  45.247060 6.094118 *
##    3) alcohol>=11.05 276 183.692000 6.235507
##      6) volatile.acidity>=0.87 9   4.222222 4.555556 *
##      7) volatile.acidity< 0.87 267 153.213500 6.292135
##       14) sulphates< 0.625 96  55.833330 5.958333
##         28) citric.acid< 0.335 63  34.412700 5.730159 *
##         29) citric.acid>=0.335 33  11.878790 6.393939 *
##       15) sulphates>=0.625 171  80.678360 6.479532 *
```

```r
rpart.plot(w.rpart, digits = 3)
```



```r
fancyRpartPlot(w.rpart)
```



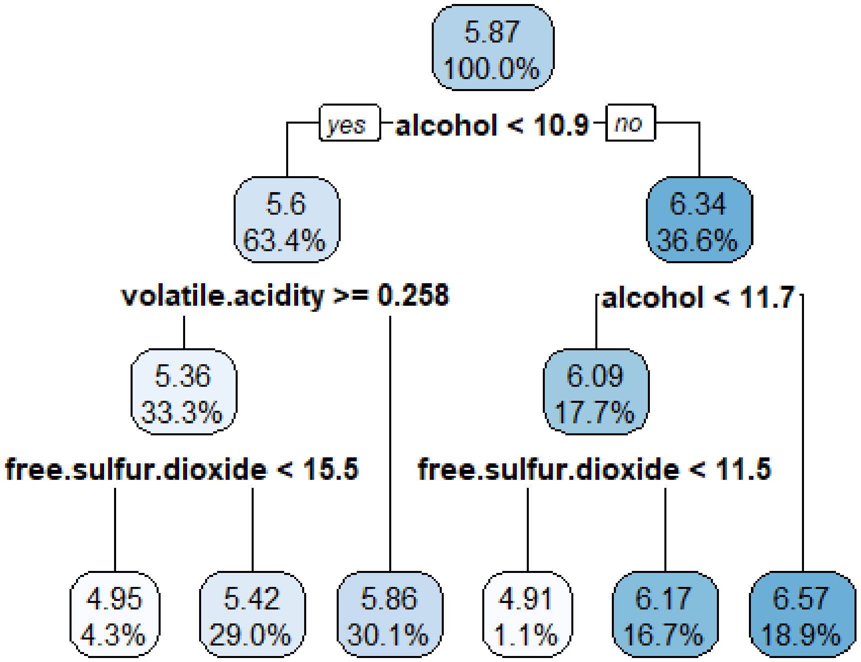
Rattle 2019-Dec-15 16:31:51 amrpo

```
prediction <- predict(w.rpart,red_test)

RWine.Pred<-as.matrix(summary(prediction)) # Summarizing results fro red Wine
colnames(RWine.Pred)<-"RWine.Pred" # Add column names
RWine.Test<-as.matrix(summary(red_test$quality))
colnames(RWine.Test)<-"RWine.Test"

RwineTree.Df<- data.frame(RWine.Pred,RWine.Test)



#Mean Absolute Error Function
MAE <- function(actual, predicted){
  MAE<-mean(abs(actual - predicted))
}

MAE.Red<-MAE(red_test$quality, prediction)
MAE.Red

## [1] 0.5479605
```

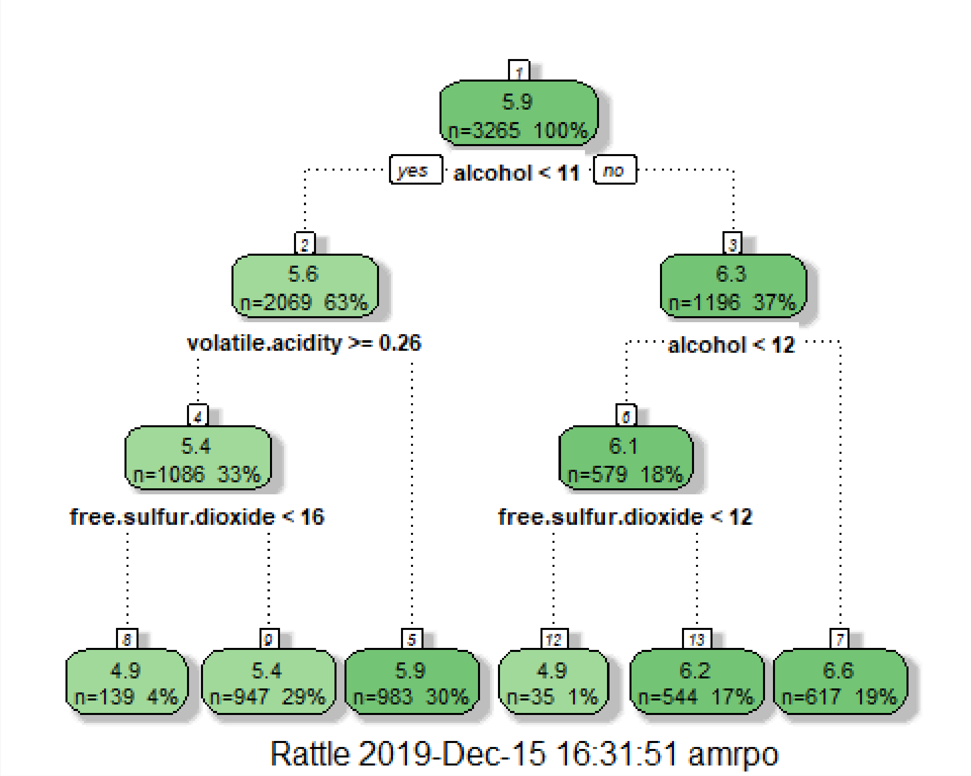
MAE=0.39

```
#Recursive Partitioning and Regression Trees **White Wine**
nrows.w<-nrow(whiteWine)
cutPoint.w<- floor(nrows.w/3*2)
cutPoint.w

## [1] 3265

rand.w<-sample(1:nrows.w)
#training set White Wine
white_train <- whiteWine[rand.w[1:cutPoint.w],]

#test set White Wine
white_test <- whiteWine[rand.w[(cutPoint.w+1:nrows.w)],]
white_test<-na.omit(white_test)

w.rpartw <- rpart(quality ~. , data = white_train)

w.rpartw

## n= 3265
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 3265 2531.15700 5.869219
##    2) alcohol< 10.85 2069 1225.03700 5.598840
##      4) volatile.acidity>=0.2575 1086  534.50460 5.360958
##        8) free.sulfur.dioxide< 15.5 139   72.64748 4.949640 *
##         9) free.sulfur.dioxide>=15.5 947  434.88910 5.421331 *
##       5) volatile.acidity< 0.2575 983  561.18410 5.861648 *
##     3) alcohol>=10.85 1196  893.20650 6.336957
##       6) alcohol< 11.725 579  428.14850 6.091537
##        12) free.sulfur.dioxide< 11.5 35   36.74286 4.914286 *
##        13) free.sulfur.dioxide>=11.5 544  339.77760 6.167279 *
##       7) alcohol>=11.725 617  397.45870 6.567261 *
```

```
rpart.plot(w.rpartw, digits = 3)
```



```
fancyRpartPlot(w.rpartw)
```



```
prediction.w <- predict(w.rpartw,white_test)

WWine.Pred<-as.matrix(summary(prediction.w)) #Summarize results for White Wine
colnames(WWine.Pred)<-"WWine.Pred" #Add column name
WWine.Test<-as.matrix(summary(white_test$quality))
colnames(WWine.Test)<-"WWine.Test"


#Mean Absolute Error
MAE.White<-MAE(white_test$quality, prediction.w)
MAE.White

## [1] 0.6117674
```

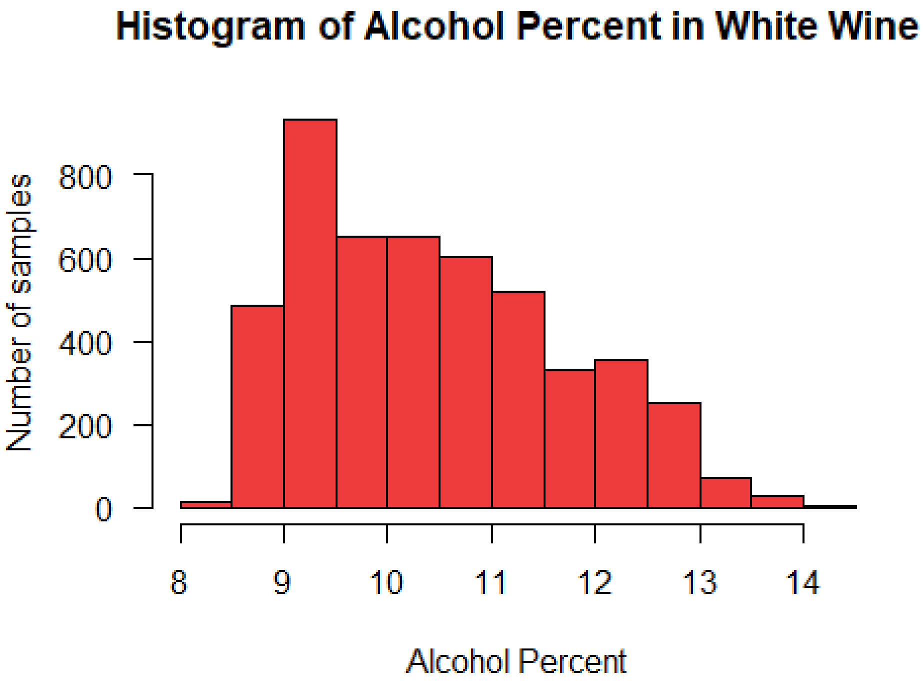
MAE=0.38

```
##Consolidated Results for both Wine types

wineTree.Df<- round(cbind(WWine.Pred,RWine.Pred,WWine.Test,RWine.Test),2)
wineTree.Df

##         WWine.Pred RWine.Pred WWine.Test RWine.Test
## Min.          4.91       4.56        3.0       3.00
## 1st Qu.       5.42       5.20        5.0       5.00
## Median        5.86       5.36        6.0       6.00
## Mean          5.88       5.65        5.9       5.65
## 3rd Qu.        6.17        6.09        6.0        6.00
## Max.           6.57        6.48        9.0        8.00
```

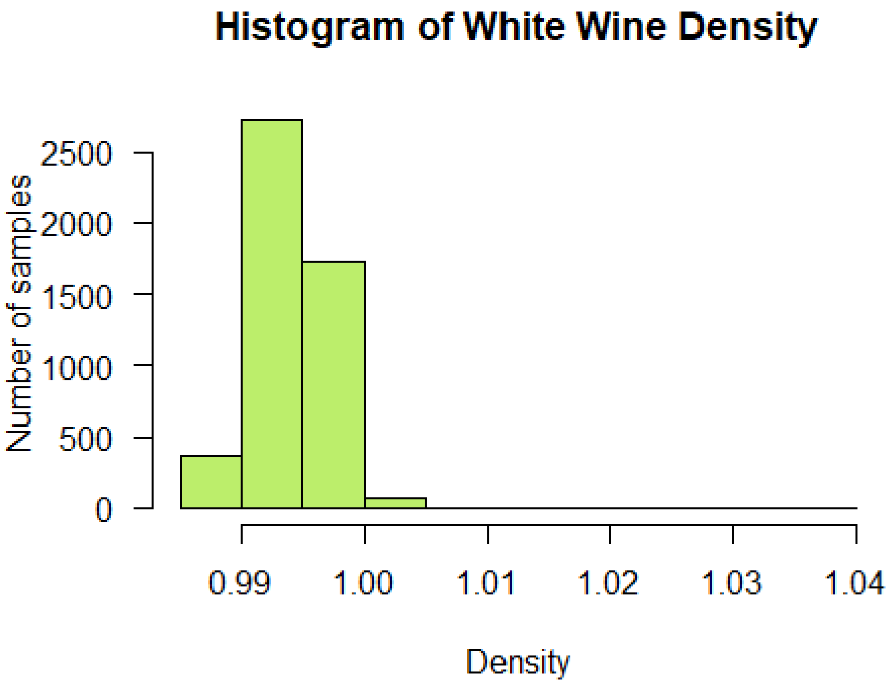
#White wine

#Exploratory Analysis

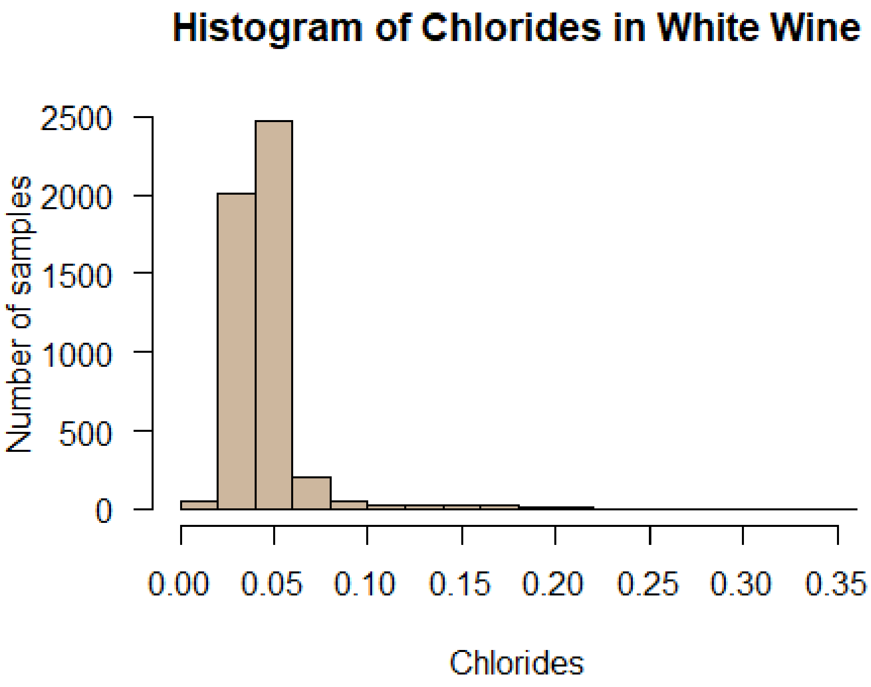
```
hist(whiteWine$alcohol, col="#EE3B3B", main="Histogram of Alcohol Percent in White
Wine", xlab="Alcohol Percent", ylab="Number of samples", las=1)
```



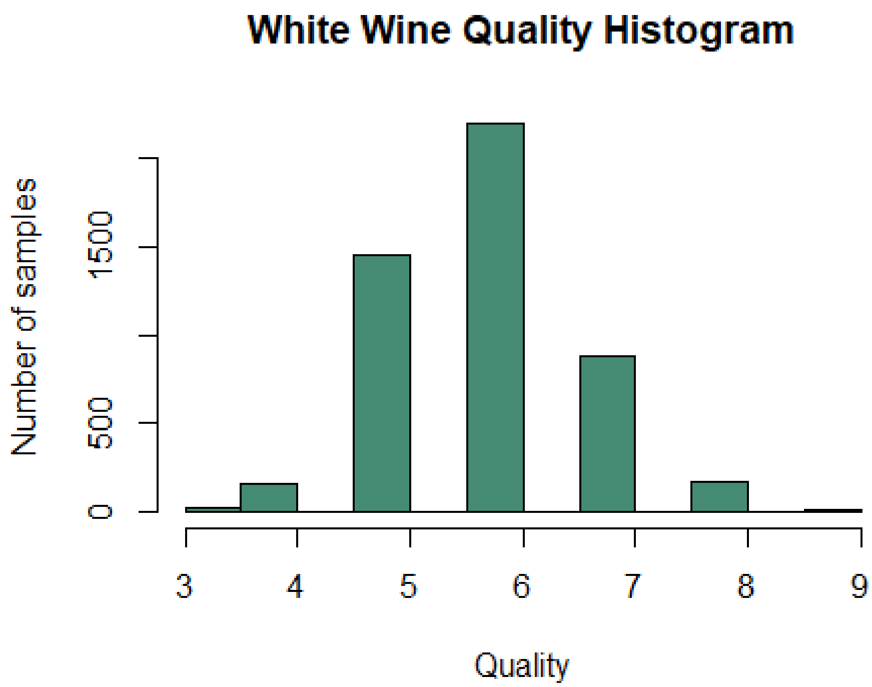
```
hist(whiteWine$density, col="#BCEE6B", main="Histogram of White Wine Density",
xlab="Density", ylab="Number of samples", las=1)
```



```r
hist(whiteWine$chlorides, col="#CDB79E", main="Histogram of Chlorides in White Wine", xlab="Chlorides", ylab="Number of samples", las=1)
```



```r
hist(whiteWine$quality, col="#458B74", main="White Wine Quality Histogram",
xlab="Quality", ylab="Number of samples")
```



# Factorizing a variable

```r
table(whiteWine$quality)

## 
##    3    4    5    6    7    8    9 
##   20  163 1457 2198  880  175    5
```

45 %of the scores are at score 6

The categorical variable we want is either: High or Low.

1 to 5 low and 6 to 9 high

```r
quality_fac <- ifelse(whiteWine$quality >= 6, "high", "low")
whitewine_data <- data.frame(whiteWine, quality_fac)
table(whitewine_data$quality_fac)

## 
## high  low 
## 3258 1640

#High  3258 Low 1640 We can now remove the old integer quality variable
whitewine_data <- whitewine_data[,-12]
```

#Splitting data into training and testing

```r
set.seed(71)
training_size <- round(0.8 * dim(whitewine_data)[1])
training_sample <- sample(dim(whitewine_data)[1], training_size, replace=FALSE)
training_data <- whitewine_data[training_sample,]
testing_data <- whitewine_data[-training_sample,]
testing_size <- round(0.2 * dim(whitewine_data)[1])
testing_sample <- sample(dim(whitewine_data)[1], testing_size, replace=FALSE)
```

#80 %of the data set is training data. 20%is testing data

#Using C50

```r
library(C50)
C50_model <- C5.0(quality_fac~., data=training_data)
predict_C50 <- predict(C50_model, testing_data[,-12])
testing_high <- quality_fac[testing_sample]
C50_model
```

```
##
## Call:
## C5.0.formula(formula = quality_fac ~ ., data = training_data)
##
## Classification Tree
## Number of samples: 3918
## Number of predictors: 11
##
## Tree size: 133
##
## Non-standard options: attempt to group attributes
```

```r
summary(C50_model)
```

```
##
## Call:
## C5.0.formula(formula = quality_fac ~ ., data = training_data)
##
##
## C5.0 [Release 2.07 GPL Edition]      Sun Dec 15 16:31:53 2019
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 3918 cases (12 attributes) from undefined.data
##
## Decision tree:
##
## alcohol > 10.7:
## :...free.sulfur.dioxide <= 11.5:
## :   :...citric.acid <= 0.2: low (11)
## :   :   citric.acid > 0.2:
## :   :   :...alcohol <= 11.8:
## :   :       :...alcohol <= 11.1:
## :   :       :   :...citric.acid <= 0.31: high (8)
## :   :       :   :   citric.acid > 0.31: low (18/7)
## :   :       :   alcohol > 11.1:
```

```
## :   :       :   :...sulphates <= 0.62: low (24/4)
## :   :       :       sulphates > 0.62: high (4/1)
## :   :       alcohol > 11.8:
## :   :       :...citric.acid > 0.39:
## :   :           :...total.sulfur.dioxide <= 70: low (5)
## :   :           :   total.sulfur.dioxide > 70: high (5)
## :   :           citric.acid <= 0.39:
## :   :           :...density <= 0.99138: high (20)
## :   :               density > 0.99138:
## :   :               :...residual.sugar <= 7.35: low (3)
## :   :                   residual.sugar > 7.35: high (6/1)
## :   free.sulfur.dioxide > 11.5:
## :   :...alcohol > 11.6:
## :       :...volatile.acidity <= 0.48: high (689/25)
## :       :   volatile.acidity > 0.48:
## :       :   :...chlorides <= 0.026: low (6/1)
## :       :       chlorides > 0.026: high (33/3)
## :       alcohol <= 11.6:
## :       :...fixed.acidity <= 6.8: high (386/39)
## :           fixed.acidity > 6.8:
## :           :...pH <= 2.97:
## :               :...total.sulfur.dioxide > 119: low (11)
## :               :   total.sulfur.dioxide <= 119:
## :               :   :...total.sulfur.dioxide <= 72: low (2)
## :               :       total.sulfur.dioxide > 72: high (14/2)
## :               pH > 2.97:
## :               :...volatile.acidity <= 0.275: high (201/34)
## :                   volatile.acidity > 0.275:
## :                   :...total.sulfur.dioxide <= 101:
## :                       :...residual.sugar > 16.1: high (2)
## :                       :   residual.sugar <= 16.1:
## :                       :   :...pH <= 3.19: low (14)
## :                       :       pH > 3.19: high (4/1)
## :                       total.sulfur.dioxide > 101:
## :                       :...residual.sugar > 2.25: high (53/6)
## :                           residual.sugar <= 2.25:
## :                           :...volatile.acidity > 0.43: low (5)
## :                               volatile.acidity <= 0.43:
## :                               :...free.sulfur.dioxide <= 34: low (15/6)
## :                                   free.sulfur.dioxide > 34: high (8)
## alcohol <= 10.7:
## :...volatile.acidity <= 0.25:
##     :...residual.sugar > 17.6:
##     :   :...fixed.acidity > 6.7: low (19/2)
##     :   :   fixed.acidity <= 6.7:
##     :   :   :...chlorides <= 0.068: high (6)
##     :   :       chlorides > 0.068: low (2)
##     :   residual.sugar <= 17.6:
##     :   :...free.sulfur.dioxide <= 14:
##     :       :...residual.sugar > 4.2: high (19/3)
##     :       :   residual.sugar <= 4.2:
##     :       :   :...pH > 3.4: high (3)
##     :       :       pH <= 3.4:
##     :       :       :...citric.acid <= 0.31: low (16)
##     :       :           citric.acid > 0.31:
```

```
##     :       :           :...free.sulfur.dioxide <= 7: low (4)
##     :       :               free.sulfur.dioxide > 7:
##     :       :               :...volatile.acidity <= 0.235: high (10/3)
##     :       :                   volatile.acidity > 0.235: low (4)
##     :       free.sulfur.dioxide > 14:
##     :       :...volatile.acidity > 0.205:
##     :           :...residual.sugar <= 13.8:
##     :           :   :...alcohol > 9.733334:
##     :           :   :   :...sulphates <= 0.39:
##     :           :   :   :   :...sulphates > 0.38: low (10/2)
##     :           :   :   :   :   sulphates <= 0.38:
##     :           :   :   :   :   :...sulphates <= 0.31: low (2)
##     :           :   :   :   :       sulphates > 0.31: high (17/3)
##     :           :   :   :   sulphates > 0.39:
##     :           :   :   :   :...fixed.acidity <= 7:
##     :           :   :   :       :...citric.acid <= 0.53: high (116/12)
##     :           :   :   :       :   citric.acid > 0.53: low (4/1)
##     :           :   :   :       fixed.acidity > 7:
##     :           :   :   :       :...residual.sugar > 9.65: high (7)
##     :           :   :   :           residual.sugar <= 9.65:
##     :           :   :   :           :...pH > 3.12: high (32/6)
##     :           :   :   :               pH <= 3.12:
##     :           :   :   :               :...volatile.acidity <= 0.215: high (2)
##     :           :   :   :                   volatile.acidity > 0.215: low (12/1)
##     :           :   :   alcohol <= 9.733334:
##     :           :   :   :...pH <= 2.94: high (10)
##     :           :   :       pH > 2.94:
##     :           :   :       :...sulphates > 0.44:
##     :           :   :           :...citric.acid > 0.48: low (13/1)
##     :           :   :           :   citric.acid <= 0.48:
##     :           :   :           :   :...chlorides <= 0.05: high (71/20)
##     :           :   :           :       chlorides > 0.05: [S1]
##     :           :   :           sulphates <= 0.44:
##     :           :   :           :...citric.acid > 0.7: high (5)
##     :           :   :               citric.acid <= 0.7:
##     :           :   :               :...residual.sugar <= 4.6:
##     :           :   :                   :...alcohol <= 9.2: low (3)
##     :           :   :                   :   alcohol > 9.2: high (9/1)
##     :           :   :                   residual.sugar > 4.6:
##     :           :   :                   :...volatile.acidity > 0.225: low (28)
##     :           :   :                       volatile.acidity <= 0.225: [S2]
##     :           :   residual.sugar > 13.8:
##     :           :   :...alcohol <= 9.1:
##     :           :       :...sulphates > 0.41: high (46/1)
##     :           :       :   sulphates <= 0.41:
##     :           :       :   :...fixed.acidity <= 6.6: low (5)
##     :           :       :       fixed.acidity > 6.6: high (10)
##     :           :       alcohol > 9.1:
##     :           :       :...residual.sugar > 17.2: high (8)
##     :           :           residual.sugar <= 17.2:
##     :           :           :...pH > 3.26: high (9/1)
##     :           :               pH <= 3.26:
##     :           :               :...pH > 3.18: low (14/1)
##     :           :                   pH <= 3.18:
##     :           :                   :...pH <= 3.02: low (6/1)
##     :           :                       pH > 3.02:
##     :           :                       :...chlorides <= 0.052: high (9)
##     :           :                           chlorides > 0.052: [S3]
##     :           volatile.acidity <= 0.205:
##     :           :...density > 0.997: high (107/4)
##     :               density <= 0.997:
##     :               :...sulphates > 0.48: high (180/25)
##     :                   sulphates <= 0.48:
##     :                   :...alcohol <= 9.8:
##     :                       :...sulphates <= 0.43:
##     :                       :   :...free.sulfur.dioxide <= 61.5: high (69/19)
##     :                       :   :   free.sulfur.dioxide > 61.5: low (7/1)
##     :                       :   sulphates > 0.43:
##     :                       :   :...volatile.acidity <= 0.19: low (24/6)
##     :                       :       volatile.acidity > 0.19: high (3)
##     :                       alcohol > 9.8:
##     :                       :...volatile.acidity <= 0.13: high (14)
##     :                           volatile.acidity > 0.13:
##     :                           :...residual.sugar > 2.9: high (50/7)
##     :                               residual.sugar <= 2.9:
##     :                               :...alcohol > 10.55: high (8)
##     :                                   alcohol <= 10.55:
##     :                                   :...residual.sugar > 1.75: low (11/2)
##     :                                       residual.sugar <= 1.75:
##     :                                       :...residual.sugar <= 1.3: [S4]
##     :                                           residual.sugar > 1.3:
##     :                                           :...sulphates <= 0.46: high (17/1)
##     :                                               sulphates > 0.46: low (2)
##     volatile.acidity > 0.25:
##     :...free.sulfur.dioxide <= 17: low (175/34)
##         free.sulfur.dioxide > 17:
##         :...alcohol > 10:
##             :...total.sulfur.dioxide > 159: low (87/35)
##             :   total.sulfur.dioxide <= 159:
##             :   :...pH > 3.33: high (33/1)
##             :       pH <= 3.33:
##             :       :...free.sulfur.dioxide <= 22:
##             :           :...pH > 3.21: low (7)
##             :           :   pH <= 3.21:
##             :           :   :...alcohol <= 10.15: high (3)
##             :           :       alcohol > 10.15:
##             :           :       :...pH <= 3.16: low (8/1)
##             :           :           pH > 3.16: high (2)
##             :           free.sulfur.dioxide > 22:
##             :           :...alcohol <= 10.3:
##             :               :...residual.sugar <= 1.3: low (7)
##             :               :   residual.sugar > 1.3:
##             :               :   :...total.sulfur.dioxide <= 144: high (34/9)
##             :               :       total.sulfur.dioxide > 144: low (7/1)
##             :               alcohol > 10.3:
##             :               :...fixed.acidity <= 7.1: high (33/1)
##             :                   fixed.acidity > 7.1:
##             :                   :...pH > 3.18: high (13/1)
##             :                       pH <= 3.18:
##             :                       :...sulphates <= 0.39: high (3)
##             :                           sulphates > 0.39: low (12/2)
##             alcohol <= 10:
##             :...volatile.acidity > 0.425: low (108/18)
##                 volatile.acidity <= 0.425:
##                 :...citric.acid <= 0.23:
##                     :...sulphates > 0.55:
##                     :   :...residual.sugar <= 6.1: low (14/1)
##                     :   :   residual.sugar > 6.1:
##                     :   :   :...free.sulfur.dioxide <= 45: high (14/1)
##                     :   :       free.sulfur.dioxide > 45: low (3)
##                     :   sulphates <= 0.55:
##                     :   :...sulphates > 0.45: low (76/9)
##                     :       sulphates <= 0.45:
##                     :       :...sulphates <= 0.36: low (9)
##                     :           sulphates > 0.36:
##                     :           :...fixed.acidity <= 5.9: high (5)
##                     :               fixed.acidity > 5.9:
##                     :               :...citric.acid <= 0.15: low (15/1)
##                     :                   citric.acid > 0.15:
##                     :                   :...sulphates <= 0.39: high (4)
##                     :                       sulphates > 0.39:
##                     :                       :...alcohol <= 9.1: low (5)
##                     :                           alcohol > 9.1: [S5]
##                     citric.acid > 0.23:
##                     :...alcohol <= 8.7: low (39/7)
##                         alcohol > 8.7:
##                         :...chlorides <= 0.04:
##                             :...fixed.acidity > 7.6: high (18/1)
##                             :   fixed.acidity <= 7.6:
##                             :   :...free.sulfur.dioxide <= 61: high (63/22)
##                             :       free.sulfur.dioxide > 61: low (6)
##                             chlorides > 0.04:
##                             :...free.sulfur.dioxide > 67:
##                                 :...pH <= 3.14: low (5)
##                                 :   pH > 3.14: high (19/1)
##                                 free.sulfur.dioxide <= 67:
##                                 :...total.sulfur.dioxide <= 135:
##                                     :...fixed.acidity <= 7.3:
##                                     :   :...citric.acid <= 0.78: high (37/6)
##                                     :   :   citric.acid > 0.78: low (2)
##                                     :   fixed.acidity > 7.3:
##                                     :   :...volatile.acidity <= 0.285: low (12)
##                                     :       volatile.acidity > 0.285: high (6/1)
##                                     total.sulfur.dioxide > 135:
##                                     :...alcohol > 9.4:
##                                         :...volatile.acidity > 0.355:
##                                         :   :...chlorides <= 0.094: high (23/3)
##                                         :   :   chlorides > 0.094: low (2)
##                                         :   volatile.acidity <= 0.355: [S6]
##                                         alcohol <= 9.4:
##                                         :...density <= 0.9944: [S7]
##                                             density > 0.9944:
##                                             :...pH <= 2.99: low (17)
##                                                 pH > 2.99: [S8]
##
```

```
## SubTree [S1]
##
## free.sulfur.dioxide <= 50.5: low (33/10)
## free.sulfur.dioxide > 50.5: high (8)
##
## SubTree [S2]
##
## free.sulfur.dioxide <= 56: low (6)
## free.sulfur.dioxide > 56: high (5/1)
##
## SubTree [S3]
##
## free.sulfur.dioxide <= 55: low (3)
## free.sulfur.dioxide > 55: high (2)
##
## SubTree [S4]
##
## residual.sugar <= 1.05: high (2)
## residual.sugar > 1.05: low (16/7)
##
## SubTree [S5]
##
## free.sulfur.dioxide <= 34: low (5)
## free.sulfur.dioxide > 34:
## :...alcohol <= 9.9: high (11/2)
##     alcohol > 9.9: low (3)
##
## SubTree [S6]
##
## volatile.acidity > 0.295: low (63/12)
## volatile.acidity <= 0.295:
## :...chlorides <= 0.05: high (43/13)
##     chlorides > 0.05:
##     :...volatile.acidity <= 0.275: low (12)
##         volatile.acidity > 0.275: high (3)
##
## SubTree [S7]
##
## total.sulfur.dioxide <= 172: high (8)
## total.sulfur.dioxide > 172: low (4/1)
##
## SubTree [S8]
##
## fixed.acidity <= 6.9: low (117/24)
## fixed.acidity > 6.9:
## :...citric.acid > 0.57: low (10)
##     citric.acid <= 0.57:
##     :...citric.acid > 0.52: high (7)
##         citric.acid <= 0.52:
##         :...citric.acid > 0.36: low (18/2)
##             citric.acid <= 0.36:
##             :...total.sulfur.dioxide <= 154: low (6)
##                 total.sulfur.dioxide > 154:
##                 :...volatile.acidity <= 0.28: high (13)
##                     volatile.acidity > 0.28:
```

```
##                      :...fixed.acidity > 7.6: high (8)
##                          fixed.acidity <= 7.6:
##                          :...free.sulfur.dioxide <= 31: high (4)
##                              free.sulfur.dioxide > 31: low (12/1)
##
##
## Evaluation on training data (3918 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##     133  483(12.3%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##    2413   202    (a): class high
##     281  1022    (b): class low
##
##
##  Attribute usage:
##
##  100.00% alcohol
##   99.31% free.sulfur.dioxide
##   86.80% volatile.acidity
##   35.17% fixed.acidity
##   32.85% residual.sugar
##   30.12% citric.acid
##   26.90% sulphates
##   26.65% pH
##   20.78% total.sulfur.dioxide
##   19.47% density
##   18.15% chlorides
##
##
## Time: 0.1 secs
```

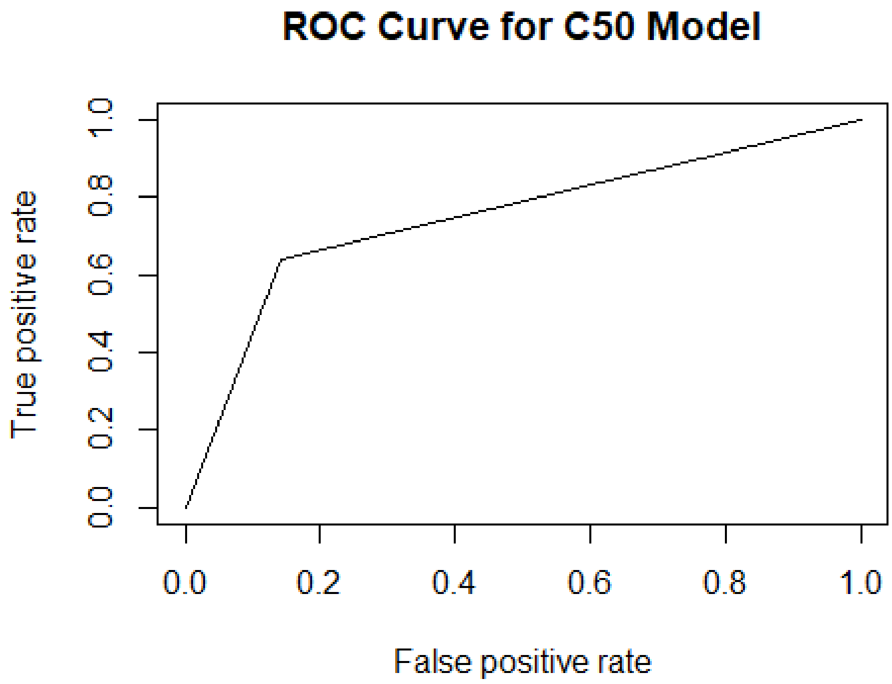
```r
# missclassification error
mean(predict_C50 != testing_high)
```

```
## [1] 0.4316327
```

```r
#0.4316327  So the misclassification error for this model is 43%
```

The misclassification error for this model is 43%

```r
library(ROCR)

predict_C50_num <- as.numeric(predict_C50)
actual_num <- as.numeric(testing_data$quality_fac)
pr <- prediction(predict_C50_num, actual_num)
auc_data1 <- performance(pr, "tpr", "fpr")
plot(auc_data1, main="ROC Curve for C50 Model")
```

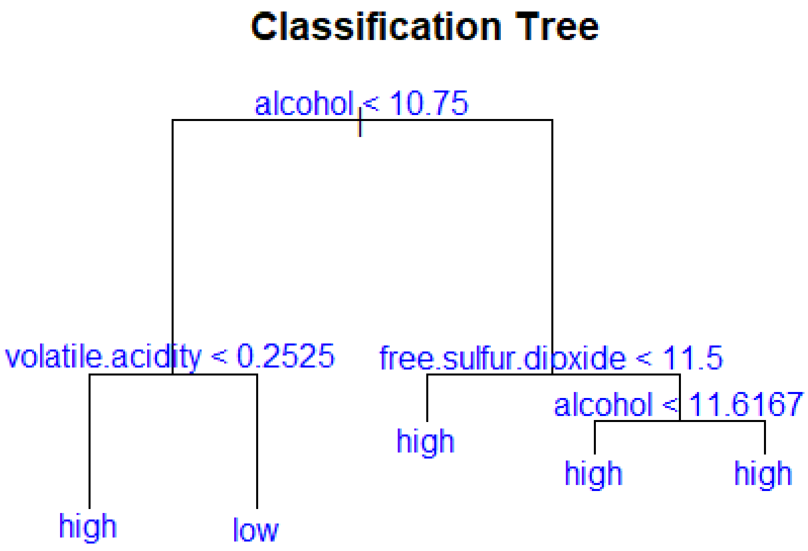


```
aucval1 <- performance(pr, measure="auc")
aucval1@y.values[[1]]

## [1] 0.7497127

# area under the curve value = 0.7497127.
```

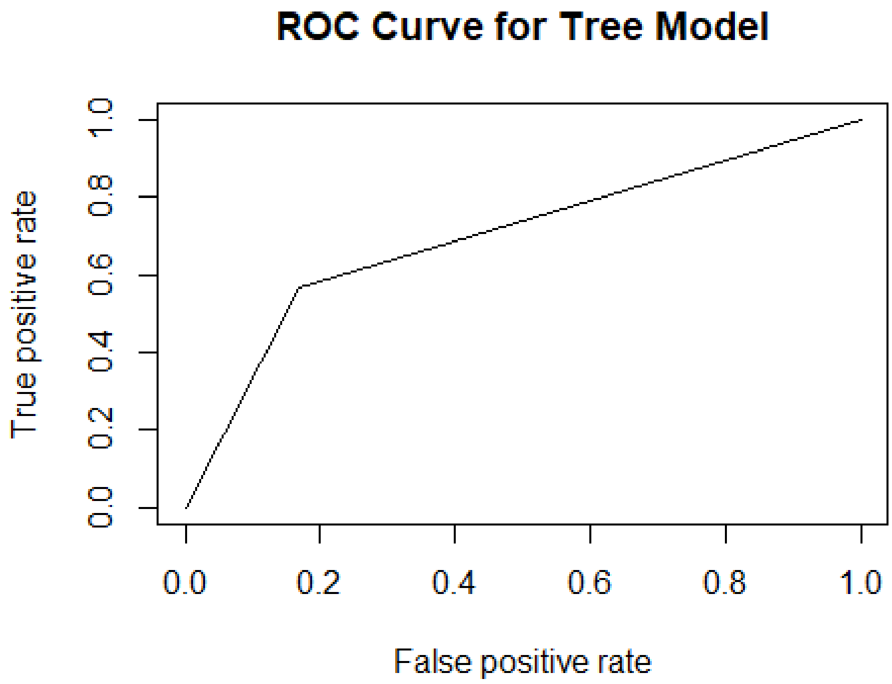
The area under the curve value for the C50 model =0.7497127

#Using The Tree Model

```
library(tree)
tree_model <- tree(quality_fac~., data=training_data)
predict_tree <- predict(tree_model, testing_data[,-12], type="class")
mean(predict_tree != testing_high)

## [1] 0.4346939

#So the misclassification error for the tree model is almost 43%
tree_model

## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
##  1) root 3918 4984.0 high ( 0.66743 0.33257 )
##    2) alcohol < 10.75 2371 3273.0 high ( 0.53817 0.46183 )
##      4) volatile.acidity < 0.2525 1098 1318.0 high ( 0.71220 0.28780 ) *
##      5) volatile.acidity > 0.2525 1273 1700.0 low ( 0.38806 0.61194 ) *
##    3) alcohol > 10.75 1547 1221.0 high ( 0.86555 0.13445 )
##      6) free.sulfur.dioxide < 11.5 104  144.2 high ( 0.50000 0.50000 ) *
##      7) free.sulfur.dioxide > 11.5 1443  988.6 high ( 0.89189 0.10811 )
##       14) alcohol < 11.6167 715  656.5 high ( 0.82797 0.17203 ) *
##       15) alcohol > 11.6167 728  268.7 high ( 0.95467 0.04533 ) *

summary(tree_model)

##
## Classification tree:
## tree(formula = quality_fac ~ ., data = training_data)
## Variables actually used in tree construction:
## [1] "alcohol"             "volatile.acidity"    "free.sulfur.dioxide"
## Number of terminal nodes:  5
## Residual mean deviance:  1.045 = 4088 / 3913
## Misclassification error rate: 0.2598 = 1018 / 3918

plot(tree_model)
text(tree_model, pretty = 0, cex = 1, col = "blue")
title("Classification Tree")
```



```
predict_tree_num <- as.numeric(predict_tree)
pr2 <- prediction(predict_tree_num, actual_num)
auc_data2 <- performance(pr2, "tpr", "fpr")
plot(auc_data2, main="ROC Curve for Tree Model")
```



```
aucval2 <- performance(pr2, measure="auc")
aucval2@y.values[[1]]

## [1] 0.6986238
```

The area under the curve value for the tree model =0.6948789

#Using rpart

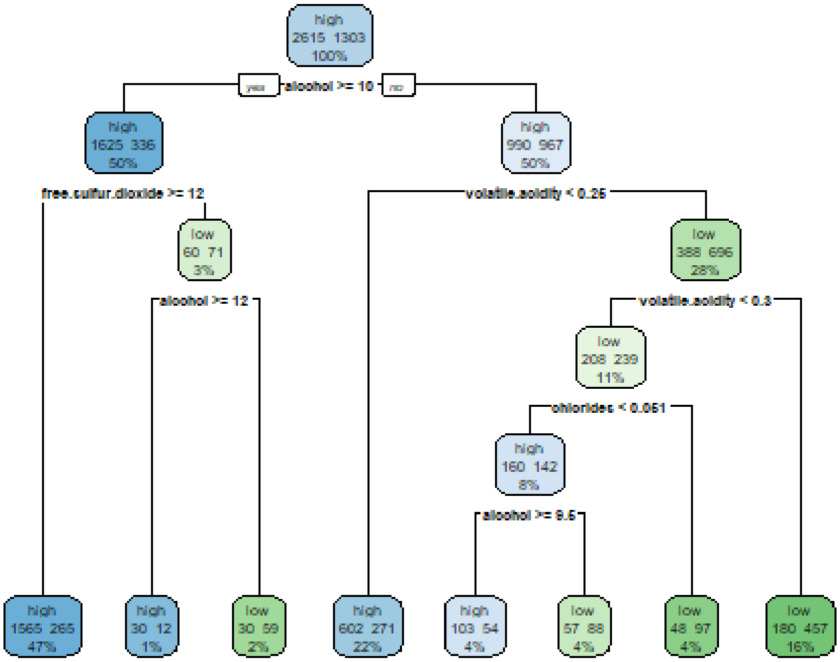
```
library (rpart)
library(rpart.plot)
rpart_model <- rpart(quality_fac~., data=training_data, method="class")
rpart_model

## n= 3918
## 
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
## 
##  1) root 3918 1303 high (0.6674324 0.3325676)  
##    2) alcohol>=10.35 1961  336 high (0.8286588 0.1713412)  
##      4) free.sulfur.dioxide>=11.5 1830  265 high (0.8551913 0.1448087) *
##      5) free.sulfur.dioxide< 11.5 131   60 low (0.4580153 0.5419847)  
##       10) alcohol>=11.85 42   12 high (0.7142857 0.2857143) *
##       11) alcohol< 11.85 89   30 low (0.3370787 0.6629213) *
##    3) alcohol< 10.35 1957  967 high (0.5058763 0.4941237)  
##      6) volatile.acidity< 0.2525 873  271 high (0.6895762 0.3104238) *
##      7) volatile.acidity>=0.2525 1084  388 low (0.3579336 0.6420664)  
##       14) volatile.acidity< 0.3025 447  208 low (0.4653244 0.5346756)
##         28) chlorides< 0.0505 302  142 high (0.5298013 0.4701987)  
##           56) alcohol>=9.45 157   54 high (0.6560510 0.3439490) *
##           57) alcohol< 9.45 145   57 low (0.3931034 0.6068966) *
##         29) chlorides>=0.0505 145   48 low (0.3310345 0.6689655) *
##       15) volatile.acidity>=0.3025 637  180 low (0.2825746 0.7174254) *
```

```
predict_rpart <- predict(rpart_model, testing_data[,-12], type="class")
mean(predict_rpart != testing_high)
```

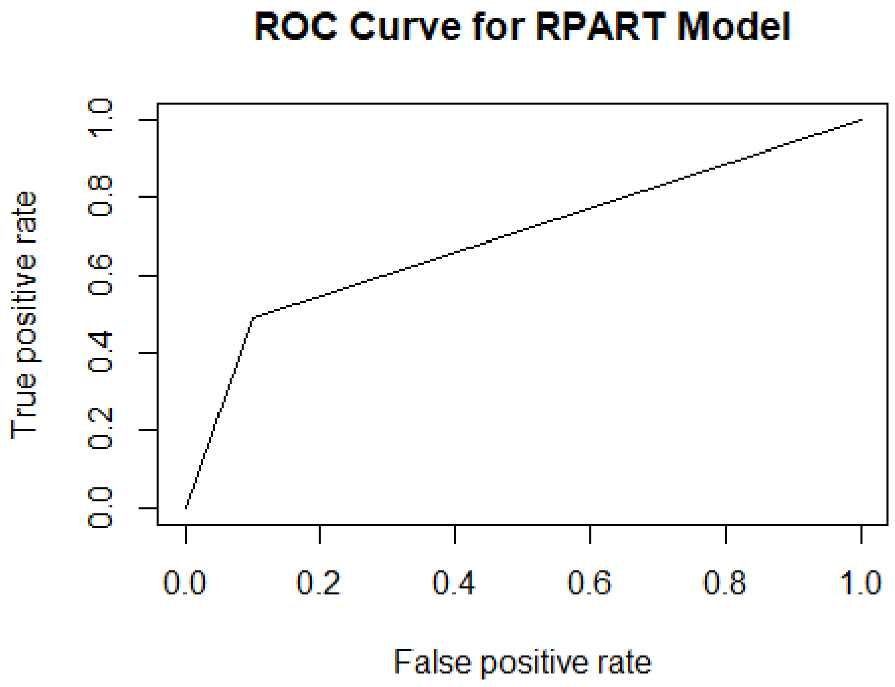
```
## [1] 0.4061224
```

So the misclassification error for the tree model is 40%

```
rpart.plot(rpart_model, extra=101)
```



```
#We can plot the tree and show the correctly and incorrectly classified instances
predict_rpart_num <- as.numeric(predict_rpart)
pr3 <- prediction(predict_rpart_num, actual_num)
auc_data3 <- performance(pr3, "tpr", "fpr")
plot(auc_data3, main="ROC Curve for RPART Model")
```



```
aucval3 <- performance(pr3, measure="auc")
aucval3@y.values[[1]]

## [1] 0.6935567
```

The area under the curve value for the tree model =0.6935567

#Results Comparison

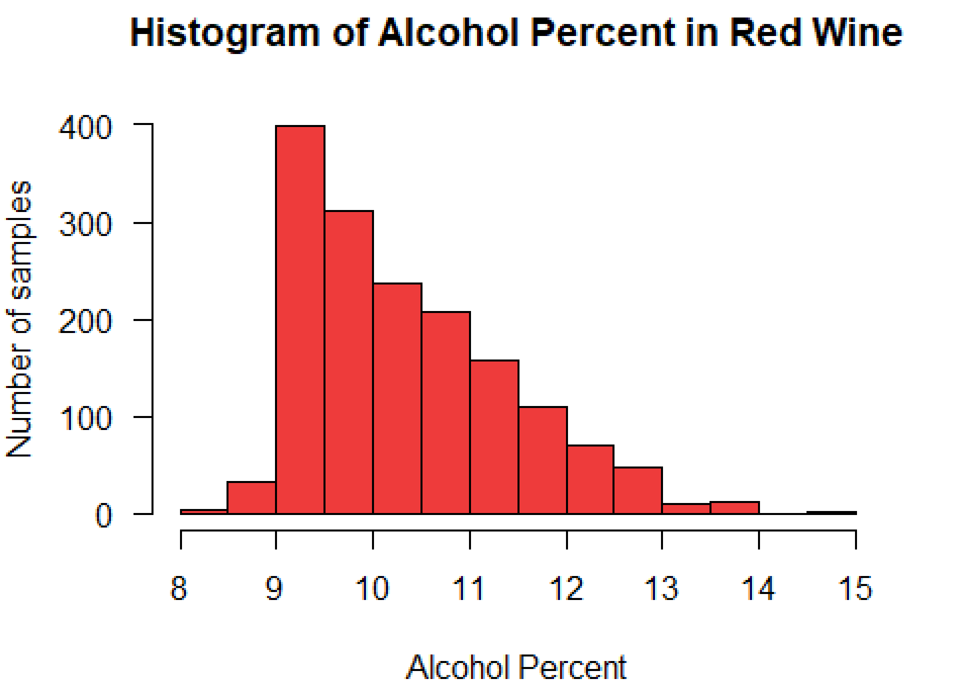
```
#nrowsw<-nrow(whitewine_data)
#cutPoint<- floor(nrowsw/3*2)
#cutPoint
#rand<-sample(1:nrowsw)
#training set
#white_train <- whitewine_data[rand[1:cutPoint],]

#test set
#white_test <- whitewine_data[rand[(cutPoint+1:nrows)],]
#white_test<-na.omit(white_test)

testing<- quality_fac[testing_sample]

#C50 Model
table(testing,predicted=predict_C50)

##         predicted
## testing high low
##    high  458 208
##    low   215  99

#557 correctly classified (57%)
#423 incorrectly classified (43%)

# Tree Model
table(testing,predicted=predict_tree)

##         predicted
## testing high low
##    high  460 206
##    low   220  94

#557 correctly classified (57%)
#423 incorrectly classified (43%)


# RPart Model
table(testing,predicted=predict_rpart)

##         predicted
## testing high low
##    high  510 156
##    low   242  72

#557 correctly classified (57%)
#423 incorrectly classified (43%)
```

#Red Wine

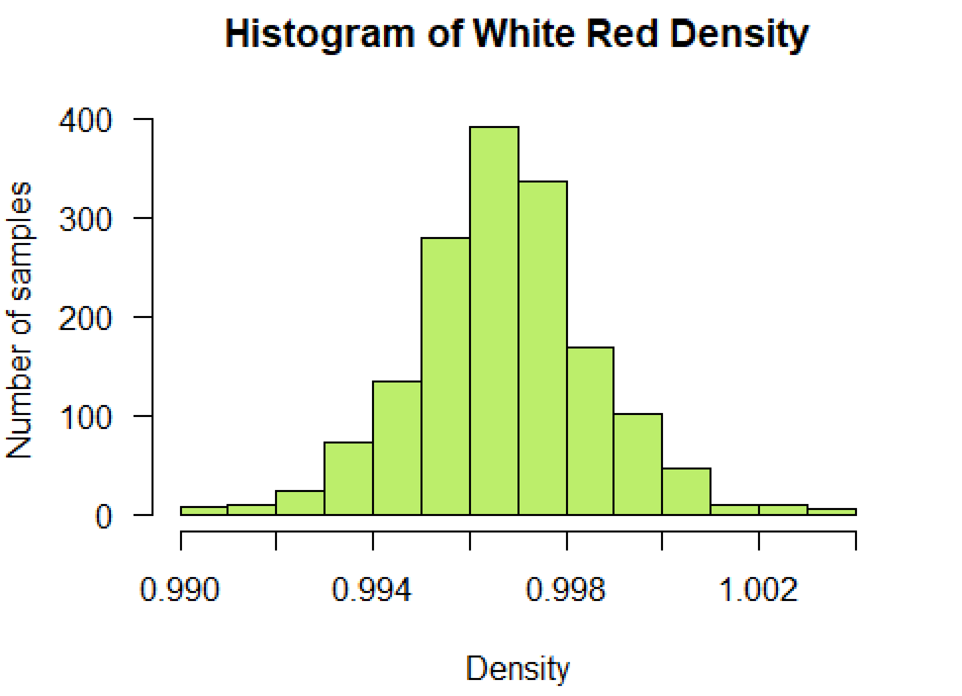
#Exploratory Analysis

```
cor(redWine)

##                      fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity           1.00000000     -0.256130895  0.67170343    0.114776724
## volatile.acidity       -0.25613089      1.000000000 -0.55249568    0.001917882
## citric.acid             0.67170343     -0.552495685  1.00000000    0.143577162
## residual.sugar          0.11477672      0.001917882  0.14357716    1.000000000
## chlorides               0.09370519      0.061297772  0.20382291    0.055609535
## free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813    0.187048995
## total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302    0.203027882
## density                 0.66804729      0.022026232  0.36494718    0.355283371
## pH                     -0.68297819      0.234937294 -0.54190414   -0.085652422
## sulphates               0.18300566     -0.260986685  0.31277004    0.005527121
## alcohol                -0.06166827     -0.202288027  0.10990325    0.042075437
## quality                 0.12405165     -0.390557780  0.22637251    0.013731637
##                        chlorides free.sulfur.dioxide total.sulfur.dioxide
## fixed.acidity        0.093705186        -0.153794193          -0.11318144
## volatile.acidity     0.061297772        -0.010503827           0.07647000
## citric.acid          0.203822914        -0.060978129           0.03553302
## residual.sugar       0.055609535         0.187048995           0.20302788
## chlorides            1.000000000         0.005562147           0.04740047
## free.sulfur.dioxide   0.005562147         1.000000000           0.66766645
## total.sulfur.dioxide  0.047400468         0.667666450           1.00000000
## density               0.200632327        -0.021945831           0.07126948
## pH                   -0.265026131         0.070377499          -0.06649456
## sulphates             0.371260481         0.051657572           0.04294684
## alcohol              -0.221140545        -0.069408354          -0.20565394
## quality              -0.128906560        -0.050656057          -0.18510029
##                          density          pH    sulphates     alcohol
## fixed.acidity         0.66804729 -0.68297819  0.183005664 -0.06166827
## volatile.acidity      0.02202623  0.23493729 -0.260986685 -0.20228803
## citric.acid           0.36494718 -0.54190414  0.312770044  0.10990325
## residual.sugar        0.35528337 -0.08565242  0.005527121  0.04207544
## chlorides             0.20063233 -0.26502613  0.371260481 -0.22114054
## free.sulfur.dioxide  -0.02194583  0.07037750  0.051657572 -0.06940835
## total.sulfur.dioxide  0.07126948 -0.06649456  0.042946836 -0.20565394
## density               1.00000000 -0.34169933  0.148506412 -0.49617977
## pH                   -0.34169933  1.00000000 -0.196647602  0.20563251
## sulphates             0.14850641 -0.19664760  1.000000000  0.09359475
## alcohol              -0.49617977  0.20563251  0.093594750  1.00000000
## quality              -0.17491923 -0.05773139  0.251397079  0.47616632
##                          quality
## fixed.acidity         0.12405165
## volatile.acidity     -0.39055778
## citric.acid           0.22637251
## residual.sugar        0.01373164
## chlorides            -0.12890656
## free.sulfur.dioxide  -0.05065606
## total.sulfur.dioxide -0.18510029
## density              -0.17491923
## pH                   -0.05773139
## sulphates             0.25139708
## alcohol               0.47616632
## quality               1.00000000
```

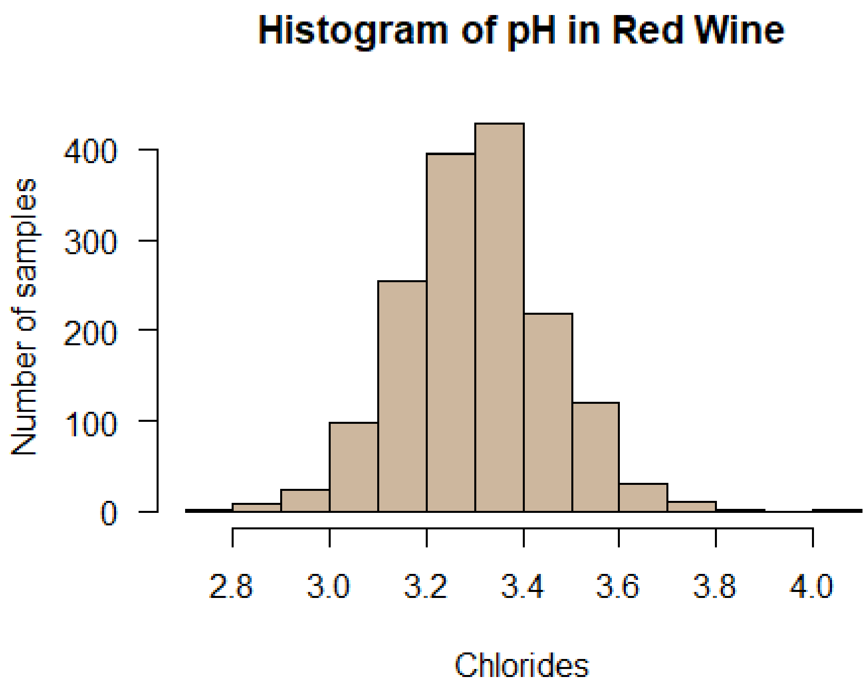
```r
hist(redWine$alcohol, col="#EE3B3B", main="Histogram of Alcohol Percent in Red
Wine", xlab="Alcohol Percent", ylab="Number of samples", las=1)
```



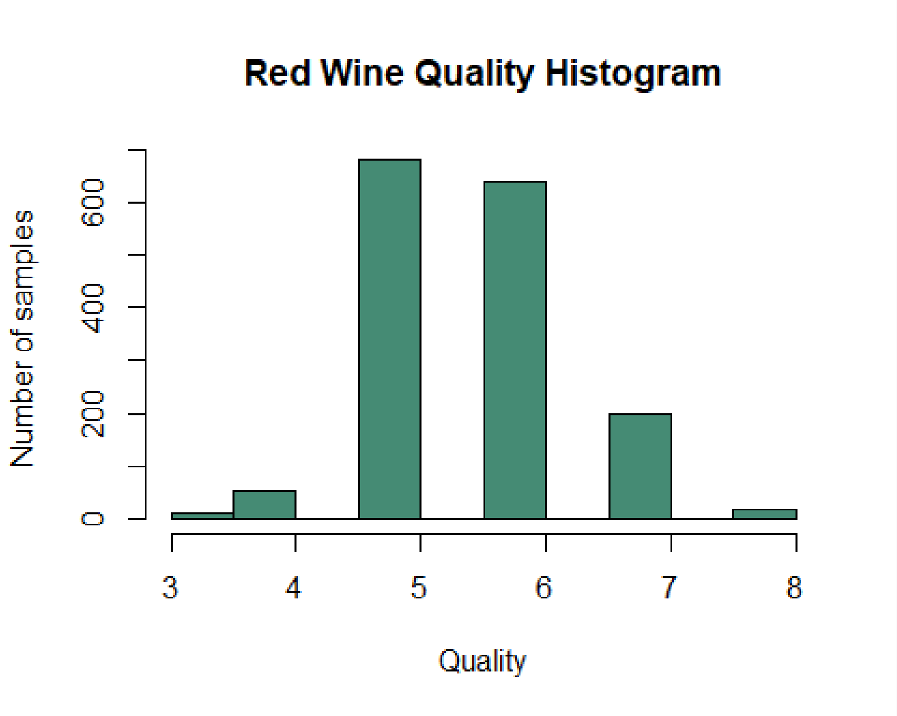
```r
hist(redWine$density, col="#BCEE6B", main="Histogram of White Red Density",
xlab="Density", ylab="Number of samples", las=1)
```



```r
hist(redWine$pH, col="#CDB79E", main="Histogram of pH in Red Wine",
xlab="Chlorides", ylab="Number of samples", las=1)
```



```r
hist(redWine$quality, col="#458B74", main="Red Wine Quality Histogram",
xlab="Quality", ylab="Number of samples")
```



# Factorizing a variable

```
table(redWine$quality)

## 
##   3   4   5   6   7   8 
##  10  53 681 638 199  18
```

43 %of the scores are at score 5

The categorical variable we want is either: High or Low.

3 to 5 low and 6 to 8 high

```
rquality_fac <- ifelse(redWine$quality >= 6, "high", "low")
redwine_data <- data.frame(redWine, rquality_fac)
table(redwine_data$rquality_fac)

## 
## high  low 
##  855  744

#High  855 Low 744 We can now remove the old integer quality variable
redwine_data <- redwine_data[,-12]
```

#Splitting data into training and testing

```
set.seed(71)
rtraining_size <- round(0.8 * dim(redwine_data)[1])
rtraining_sample <- sample(dim(redwine_data)[1], rtraining_size, replace=FALSE)
rtraining_data <- redwine_data[rtraining_sample,]
rtesting_data <- redwine_data[-rtraining_sample,]
rtesting_size <- round(0.2 * dim(redwine_data)[1])
rtesting_sample <- sample(dim(redwine_data)[1], rtesting_size, replace=FALSE)
```

#80 %of the data set is training data. 20%is testing data

#Using C50

```
library(C50)
rC50_model <- C5.0(rquality_fac~., data=rtraining_data)
rpredict_C50 <- predict(rC50_model, rtesting_data[,-12])
rtesting_high <- rquality_fac[rtesting_sample]
rC50_model

## 
## Call:
## C5.0.formula(formula = rquality_fac ~ ., data = rtraining_data)
## 
## Classification Tree
## Number of samples: 1279 
## Number of predictors: 11 
## 
## Tree size: 91 
## 
## Non-standard options: attempt to group attributes

summary(rC50_model)
```

```
##
## Call:
## C5.0.formula(formula = rquality_fac ~ ., data = rtraining_data)
##
##
## C5.0 [Release 2.07 GPL Edition]      Sun Dec 15 16:31:53 2019
## ------------------------------
##
## Class specified by attribute `outcome'
##
## Read 1279 cases (12 attributes) from undefined.data
##
## Decision tree:
##
## total.sulfur.dioxide > 109:
## :...density <= 0.99323: high (4)
## :   density > 0.99323: low (72/2)
## total.sulfur.dioxide <= 109:
## :...alcohol > 10.2:
##     :...sulphates <= 0.63:
##     :   :...alcohol > 11.4:
##     :   :   :...alcohol > 12.8:
##     :   :   :   :...density <= 0.99252: high (3)
##     :   :   :   :   density > 0.99252: low (7/1)
##     :   :   :   alcohol <= 12.8:
##     :   :   :   :...volatile.acidity <= 0.565: high (53/2)
##     :   :   :       volatile.acidity > 0.565:
##     :   :   :       :...citric.acid <= 0.05:
##     :   :   :           :...density <= 0.99553: high (20/1)
##     :   :   :           :   density > 0.99553: low (3/1)
##     :   :   :           citric.acid > 0.05:
##     :   :   :           :...sulphates <= 0.55: low (6)
##     :   :   :               sulphates > 0.55:
##     :   :   :               :...volatile.acidity <= 0.665: low (2)
##     :   :   :                   volatile.acidity > 0.665: high (4)
##     :   :   alcohol <= 11.4:
##     :   :   :...free.sulfur.dioxide <= 7: low (55/16)
##     :   :       free.sulfur.dioxide > 7:
##     :   :       :...sulphates <= 0.53:
##     :   :           :...volatile.acidity <= 0.41: high (4)
##     :   :           :   volatile.acidity > 0.41: low (22/6)
##     :   :           sulphates > 0.53:
##     :   :           :...citric.acid <= 0.07:
##     :   :               :...density <= 0.99596: high (23)
##     :   :               :   density > 0.99596:
##     :   :               :   :...sulphates <= 0.61: high (11/1)
##     :   :               :       sulphates > 0.61: low (2)
##     :   :               citric.acid > 0.07:
##     :   :               :...chlorides <= 0.072: low (12/2)
##     :   :                   chlorides > 0.072:
##     :   :                   :...chlorides <= 0.089: high (24/3)
##     :   :                       chlorides > 0.089:
##     :   :                       :...sulphates <= 0.62: low (8/1)
##     :   :                           sulphates > 0.62: high (2)
##     :   sulphates > 0.63:
```

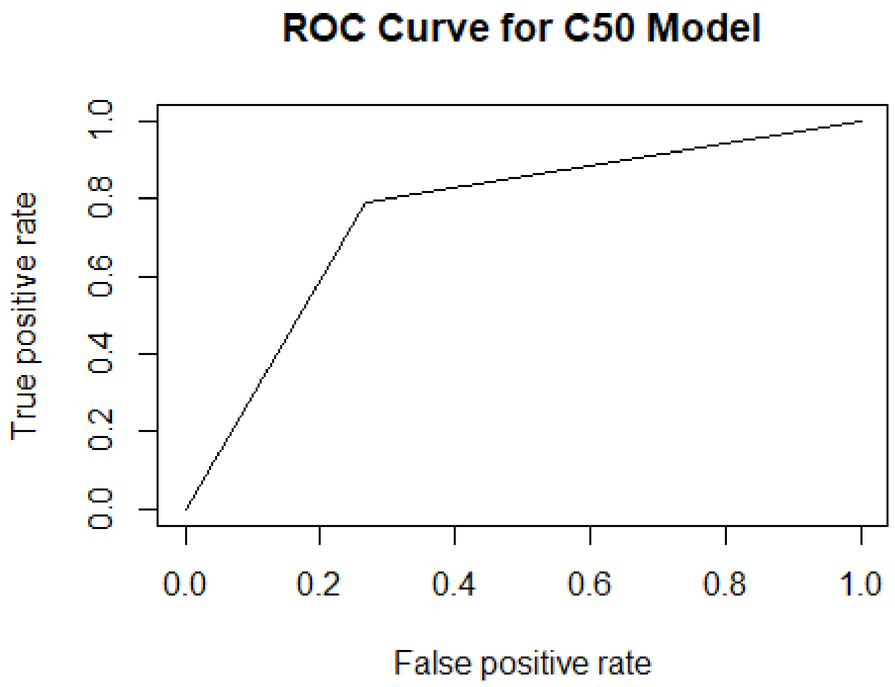
```
##     :   :...alcohol > 11.5: high (104/2)
##     :       alcohol <= 11.5:
##     :       :...residual.sugar > 3.8:
##     :           :...chlorides > 0.092: low (5)
##     :           :   chlorides <= 0.092:
##     :           :   :...total.sulfur.dioxide <= 78: high (10/1)
##     :           :       total.sulfur.dioxide > 78: low (3)
##     :           residual.sugar <= 3.8:
##     :           :...alcohol <= 10.6:
##     :               :...pH <= 3.41: high (42/6)
##     :               :   pH > 3.41:
##     :               :   :...chlorides <= 0.065: high (3)
##     :               :       chlorides > 0.065: low (9/1)
##     :               alcohol > 10.6:
##     :               :...total.sulfur.dioxide > 66:
##     :                   :...pH <= 3.32: high (9)
##     :                   :   pH > 3.32: low (8/2)
##     :                   total.sulfur.dioxide <= 66:
##     :                   :...alcohol > 11.2: high (27)
##     :                       alcohol <= 11.2:
##     :                       :...chlorides > 0.097:
##     :                           :...citric.acid <= 0.36: low (3)
##     :                           :   citric.acid > 0.36: high (5/1)
##     :                           chlorides <= 0.097:
##     :                           :...chlorides > 0.064: high (63/1)
##     :                               chlorides <= 0.064:
##     :                               :...free.sulfur.dioxide > 14: high (14)
##     :                                   free.sulfur.dioxide <= 14:
##     :                                   :...volatile.acidity <= 0.42: high (7/1)
##     :                                       volatile.acidity > 0.42: low (4)
##     alcohol <= 10.2:
##     :...sulphates > 0.57:
##         :...volatile.acidity <= 0.315:
##         :   :...volatile.acidity <= 0.27: high (19)
##         :   :   volatile.acidity > 0.27:
##         :   :   :...chlorides > 0.095: low (2)
##         :   :       chlorides <= 0.095:
##         :   :       :...citric.acid <= 0.41: low (3/1)
##         :   :           citric.acid > 0.41: high (13)
##         :   volatile.acidity > 0.315:
##         :   :...volatile.acidity > 0.655:
##         :       :...density > 0.99765:
##         :       :   :...free.sulfur.dioxide > 18: low (6)
##         :       :   :   free.sulfur.dioxide <= 18:
##         :       :   :   :...fixed.acidity <= 9.5: high (11/1)
##         :       :   :       fixed.acidity > 9.5: low (4/1)
##         :       :   density <= 0.99765:
##         :       :   :...chlorides <= 0.091: low (26)
##         :       :       chlorides > 0.091:
##         :       :       :...citric.acid > 0.18: low (8)
##         :       :           citric.acid <= 0.18:
##         :       :           :...citric.acid > 0.09: high (3)
##         :       :               citric.acid <= 0.09:
##         :       :               :...chlorides <= 0.094: high (3/1)
##         :       :                   chlorides > 0.094: low (4)
##         :       volatile.acidity <= 0.655:
##         :       :...chlorides > 0.098:
##         :           :...chlorides <= 0.107: low (13)
##         :           :   chlorides > 0.107:
##         :           :   :...alcohol <= 9.4: low (23/4)
##         :           :       alcohol > 9.4:
##         :           :       :...volatile.acidity <= 0.43: high (6)
##         :           :           volatile.acidity > 0.43:
##         :           :           :...volatile.acidity <= 0.475: low (4)
##         :           :               volatile.acidity > 0.475: high (7/2)
##         :           chlorides <= 0.098:
##         :           :...pH > 3.53:
##         :               :...volatile.acidity > 0.56: low (10)
##         :               :   volatile.acidity <= 0.56:
##         :               :   :...density <= 0.99735: low (2)
##         :               :       density > 0.99735: high (2)
##         :               pH <= 3.53:
##         :               :...alcohol <= 9.8:
##         :                   :...free.sulfur.dioxide <= 4: low (7)
##         :                   :   free.sulfur.dioxide > 4:
##         :                   :   :...chlorides <= 0.089:
##         :                   :       :...total.sulfur.dioxide <= 38: high (46/14)
##         :                   :       :   total.sulfur.dioxide > 38: low (61/23)
##         :                   :       chlorides > 0.089:
##         :                   :       :...free.sulfur.dioxide <= 23: high (17/1)
##         :                   :           free.sulfur.dioxide > 23: low (3/1)
##         :                   alcohol > 9.8:
##         :                   :...sulphates <= 0.61:
##         :                       :...volatile.acidity > 0.6: high (4)
##         :                       :   volatile.acidity <= 0.6:
##         :                       :   :...volatile.acidity <= 0.545: high (3/1)
##         :                       :       volatile.acidity > 0.545: low (6)
##         :                       sulphates > 0.61:
##         :                       :...density <= 0.99836: high (28)
##         :                           density > 0.99836:
##         :                           :...volatile.acidity > 0.52: high (4)
##         :                               volatile.acidity <= 0.52:
##         :                               :...sulphates <= 0.69: low (3)
##         :                                   sulphates > 0.69:
##         :                                   :...alcohol <= 10.03333: high (6)
##         :                                       alcohol > 10.03333: low (3/1)
##         sulphates <= 0.57:
##         :...alcohol > 9.7:
##             :...volatile.acidity > 0.585: low (47/7)
##             :   volatile.acidity <= 0.585:
##             :   :...sulphates <= 0.47: low (4)
##             :       sulphates > 0.47:
##             :       :...alcohol > 10.03333: high (9/1)
##             :           alcohol <= 10.03333:
##             :           :...alcohol > 9.95: low (7/1)
##             :               alcohol <= 9.95:
##             :               :...free.sulfur.dioxide <= 9: high (5)
##             :                   free.sulfur.dioxide > 9:
##             :                   :...volatile.acidity <= 0.45: high (3)
##             :                       volatile.acidity > 0.45: low (5/1)
##             alcohol <= 9.7:
##             :...chlorides > 0.082:
##                 :...residual.sugar <= 4.8: low (67/2)
##                 :   residual.sugar > 4.8:
##                 :   :...free.sulfur.dioxide <= 24: high (3)
##                 :       free.sulfur.dioxide > 24: low (3)
##                 chlorides <= 0.082:
##                 :...alcohol <= 9:
##                     :...residual.sugar <= 2.05: low (3)
##                     :   residual.sugar > 2.05: high (6/1)
##                     alcohol > 9:
##                     :...density > 0.99744: low (27/1)
##                         density <= 0.99744:
##                         :...alcohol <= 9.3: low (24/3)
##                             alcohol > 9.3:
##                             :...residual.sugar <= 1.65: low (8)
##                                 residual.sugar > 1.65:
##                                 :...free.sulfur.dioxide <= 8: low (16/2)
##                                     free.sulfur.dioxide > 8:
##                                     :...pH <= 3.16: low (4)
##                                         pH > 3.16:
##                                         :...fixed.acidity > 7.8: high (6)
##                                             fixed.acidity <= 7.8: [S1]
##
## SubTree [S1]
##
## volatile.acidity > 0.645: low (7)
## volatile.acidity <= 0.645:
## :...citric.acid <= 0.14: high (6)
##     citric.acid > 0.14:
##     :...citric.acid <= 0.29: low (3)
##         citric.acid > 0.29: high (3)
##
##
## Evaluation on training data (1279 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##      91  121( 9.5%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     604    80    (a): class high
##      41   554    (b): class low
##
##
##  Attribute usage:
##
##  100.00% total.sulfur.dioxide
##   94.06% sulphates
##   94.06% alcohol
##   53.01% chlorides
##   45.66% volatile.acidity
##   31.82% free.sulfur.dioxide
##   28.30% density
##   27.13% residual.sugar
##   23.85% pH
##   13.37% citric.acid
##    3.13% fixed.acidity
##
##
## Time: 0.0 secs
```

```
# missclassification error
mean(rpredict_C50 != rtesting_high)
```

```
## [1] 0.54375
```

```
#The misclassification error for this model is 54%
```

```
library(ROCR)
rpredict_C50_num <- as.numeric(rpredict_C50)
ractual_num <- as.numeric(rtesting_data$rquality_fac)
rpr <- prediction(rpredict_C50_num, ractual_num)
rauc_data1 <- performance(rpr, "tpr", "fpr")
plot(rauc_data1, main="ROC Curve for C50 Model")
```



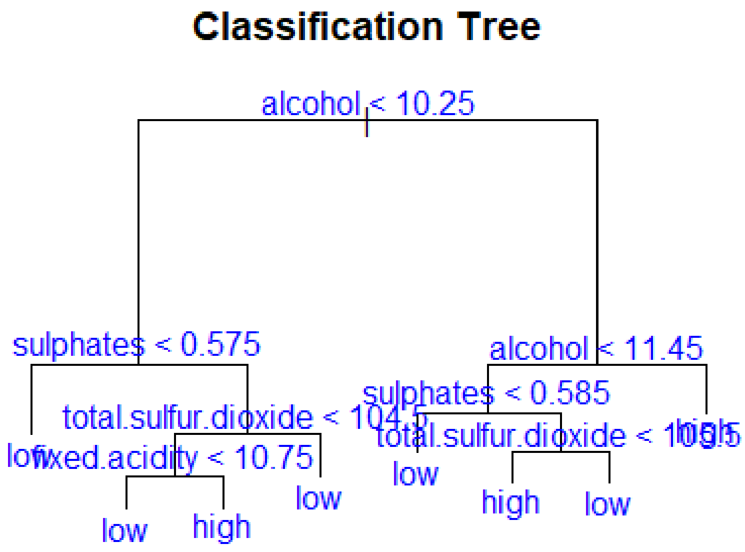
```
raucval1 <- performance(rpr, measure="auc")
raucval1@y.values[[1]]
```

```
## [1] 0.7614702
```

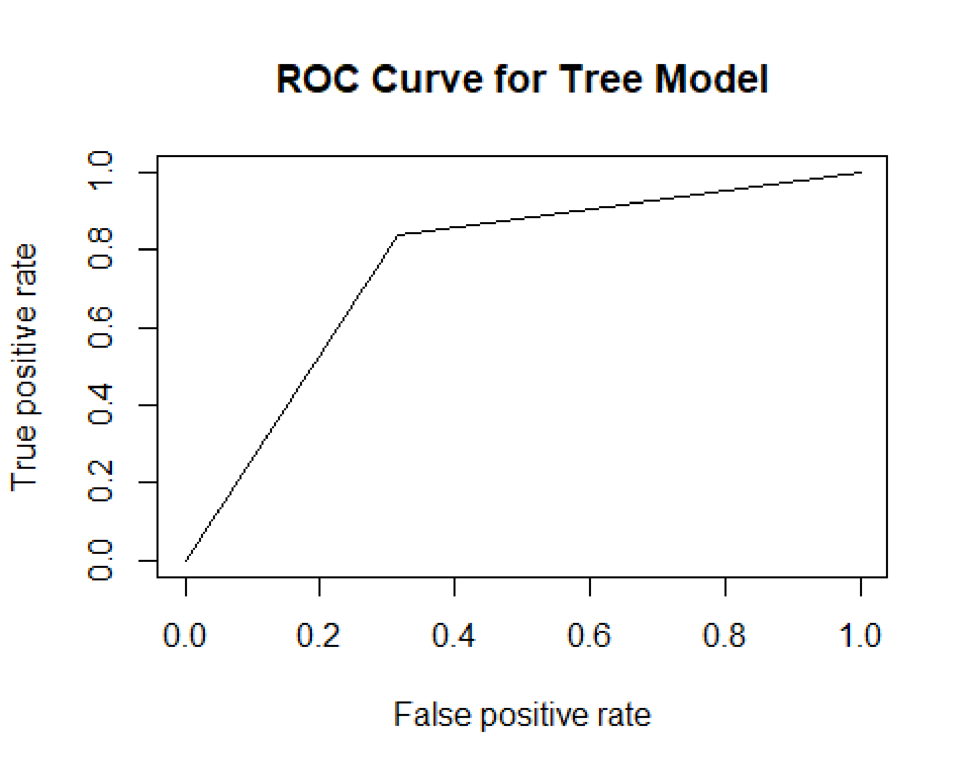
```
# area under the curve value = 0.7614702.
```

#Using the Tree Model

```
library(tree)
rtree_model <- tree(rquality_fac~., data=rtraining_data)
rpredict_tree <- predict(rtree_model, rtesting_data[,-12], type="class")
mean(rpredict_tree != rtesting_high)

## [1] 0.515625

#Misclassification error for the tree model is almost 52%
rtree_model

## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
## 
##  1) root 1279 1767.000 high ( 0.53479 0.46521 )  
##    2) alcohol < 10.25 688  891.100 low ( 0.35029 0.64971 )  
##      4) sulphates < 0.575 296  290.000 low ( 0.19257 0.80743 ) *
##      5) sulphates > 0.575 392  542.000 low ( 0.46939 0.53061 )  
##       10) total.sulfur.dioxide < 104.5 358  496.100 high ( 0.51117 0.48883 )  
##         20) fixed.acidity < 10.75 311  429.100 low ( 0.45981 0.54019 ) *
##         21) fixed.acidity > 10.75 47   39.560 high ( 0.85106 0.14894 ) *
##       11) total.sulfur.dioxide > 104.5 34    9.023 low ( 0.02941 0.97059 ) *
##    3) alcohol > 10.25 591  665.200 high ( 0.74958 0.25042 )  
##      6) alcohol < 11.45 375  477.400 high ( 0.66667 0.33333 )  
##       12) sulphates < 0.585 105  143.900 low ( 0.43810 0.56190 ) *
##       13) sulphates > 0.585 270  300.300 high ( 0.75556 0.24444 )  
##         26) total.sulfur.dioxide < 105.5 261  274.000 high ( 0.78161 0.21839 ) *
##         27) total.sulfur.dioxide > 105.5 9    0.000 low ( 0.00000 1.00000 ) *
##      7) alcohol > 11.45 216  146.500 high ( 0.89352 0.10648 ) *

summary(rtree_model)

## 
## Classification tree:
## tree(formula = rquality_fac ~ ., data = rtraining_data)
## Variables actually used in tree construction:
## [1] "alcohol"              "sulphates"            "total.sulfur.dioxide"
## [4] "fixed.acidity"       
## Number of terminal nodes:  8 
## Residual mean deviance:  1.048 = 1332 / 1271 
## Misclassification error rate: 0.2611 = 334 / 1279

plot(rtree_model)
text(rtree_model, pretty = 0, cex = 1, col = "blue")
title("Classification Tree")
```



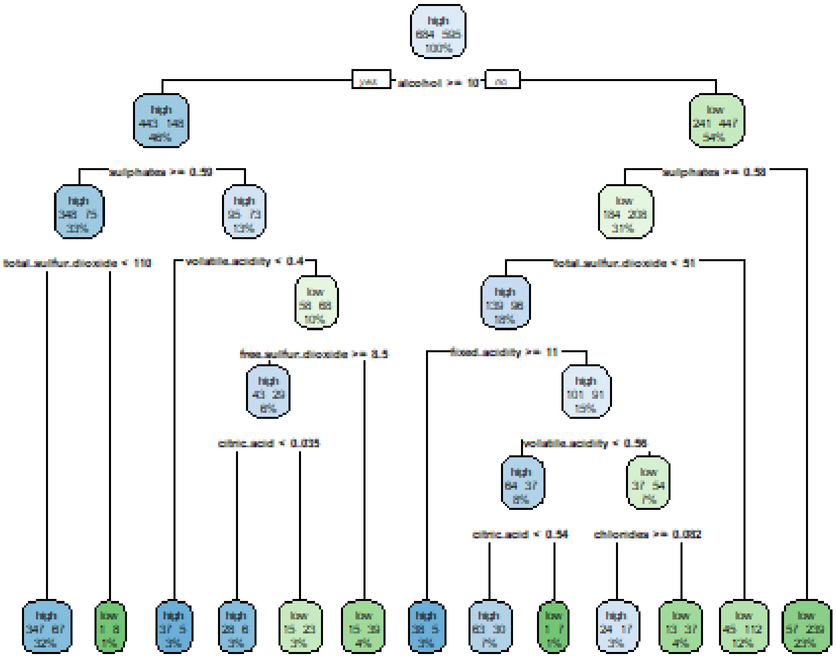
```
rpredict_tree_num <- as.numeric(rpredict_tree)
rpr2 <- prediction(rpredict_tree_num, ractual_num)
rauc_data2 <- performance(rpr2, "tpr", "fpr")
plot(rauc_data2, main="ROC Curve for Tree Model")
```



```
raucval2 <- performance(rpr2, measure="auc")
raucval2@y.values[[1]]

## [1] 0.7615684

#The area under the curve value for the tree model = 0.7615684
```

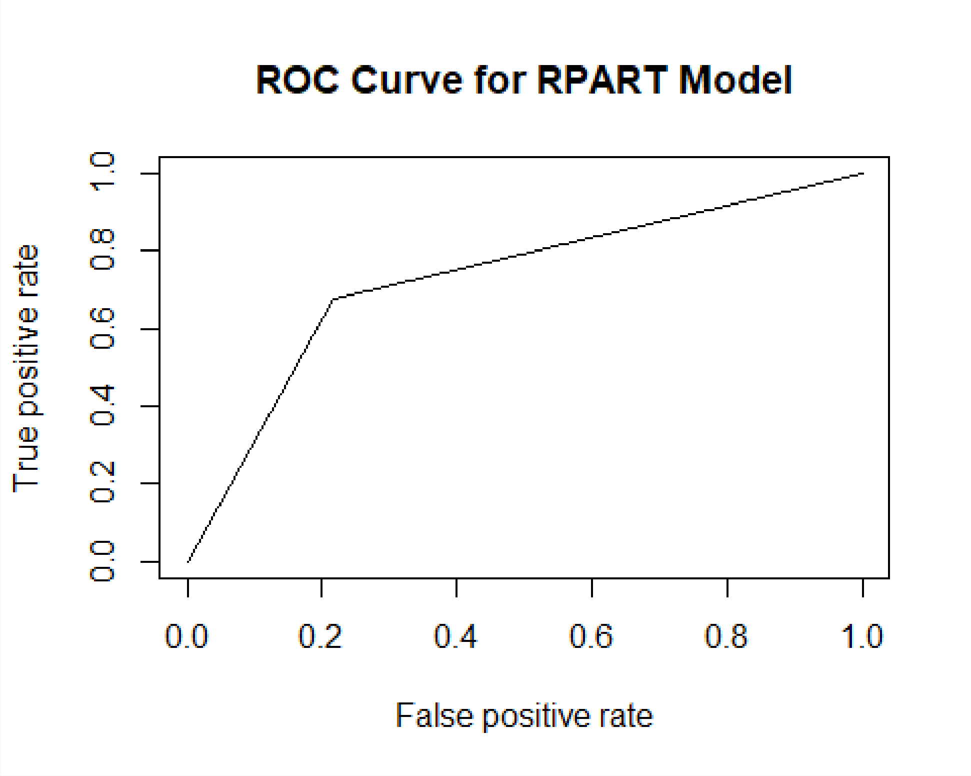
#Using rpart

```
library (rpart)
library(rpart.plot)
rrpart_model <- rpart(rquality_fac~., data=rtraining_data, method="class")
rrpart_model

## n= 1279
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##   1) root 1279 595 high (0.5347928 0.4652072)
##     2) alcohol>=10.25 591 148 high (0.7495770 0.2504230)
##       4) sulphates>=0.585 423  75 high (0.8226950 0.1773050)
##         8) total.sulfur.dioxide< 109.5 414  67 high (0.8381643 0.1618357) *
##         9) total.sulfur.dioxide>=109.5 9   1 low (0.1111111 0.8888889) *
##       5) sulphates< 0.585 168  73 high (0.5654762 0.4345238)
##        10) volatile.acidity< 0.395 42   5 high (0.8809524 0.1190476) *
##        11) volatile.acidity>=0.395 126  58 low (0.4603175 0.5396825)
##          22) free.sulfur.dioxide>=8.5 72  29 high (0.5972222 0.4027778)
##            44) citric.acid< 0.035 34   6 high (0.8235294 0.1764706) *
##            45) citric.acid>=0.035 38  15 low (0.3947368 0.6052632) *
##          23) free.sulfur.dioxide< 8.5 54  15 low (0.2777778 0.7222222) *
##     3) alcohol< 10.25 688 241 low (0.3502907 0.6497093)
##       6) sulphates>=0.575 392 184 low (0.4693878 0.5306122)
##        12) total.sulfur.dioxide< 50.5 235  96 high (0.5914894 0.4085106)
##          24) fixed.acidity>=10.75 43   5 high (0.8837209 0.1162791) *
##          25) fixed.acidity< 10.75 192  91 high (0.5260417 0.4739583)
##            50) volatile.acidity< 0.555 101  37 high (0.6336634 0.3663366)
##             100) citric.acid< 0.535 93  30 high (0.6774194 0.3225806) *
##             101) citric.acid>=0.535 8   1 low (0.1250000 0.8750000) *
##            51) volatile.acidity>=0.555 91  37 low (0.4065934 0.5934066)
##             102) chlorides>=0.082 41  17 high (0.5853659 0.4146341) *
##             103) chlorides< 0.082 50  13 low (0.2600000 0.7400000) *
##        13) total.sulfur.dioxide>=50.5 157  45 low (0.2866242 0.7133758) *
##       7) sulphates< 0.575 296  57 low (0.1925676 0.8074324) *

rpredict_rpart <- predict(rrpart_model, rtesting_data[,-12], type="class")
mean(rpredict_rpart != rtesting_high)

## [1] 0.5125

#So the misclassification error for the tree model is 51.25%
rpart.plot(rrpart_model, extra=101)
```



```r
#We can plot the tree and show the correctly and incorrectly classified instances
rpredict_rpart_num <- as.numeric(rpredict_rpart)
rpr3 <- prediction(rpredict_rpart_num, ractual_num)
rauc_data3 <- performance(rpr3, "tpr", "fpr")
```

```
plot(rauc_data3, main="ROC Curve for RPART Model")
```



```
raucval3 <- performance(rpr3, measure="auc")
raucval3@y.values[[1]]

## [1] 0.730739

#So, the area under the curve value for the tree model = 0.730739
```

#Results Comparison

```
rtesting<- rquality_fac[rtesting_sample]

#C50 Model
table(rtesting,predicted=rpredict_C50)

##          predicted
## rtesting high low
##     high   77  95
##     low    79  69

#146 correctly classified (46%)
#174 incorrectly classified (54%)


# Tree Model
table(rtesting,predicted=rpredict_tree)

##          predicted
## rtesting high low
##     high   74  98
##     low    67  81

#155 correctly classified (48%)
#165 incorrectly classified (52%)


# RPart Model
table(rtesting,predicted=rpredict_rpart)

##         predicted
## rtesting high low
##     high   95  77
##     low    87  61

#156 correctly classified (49%)
#164 incorrectly classified (51%)
```

Red Wine Analytics

##Analytics RED WINE

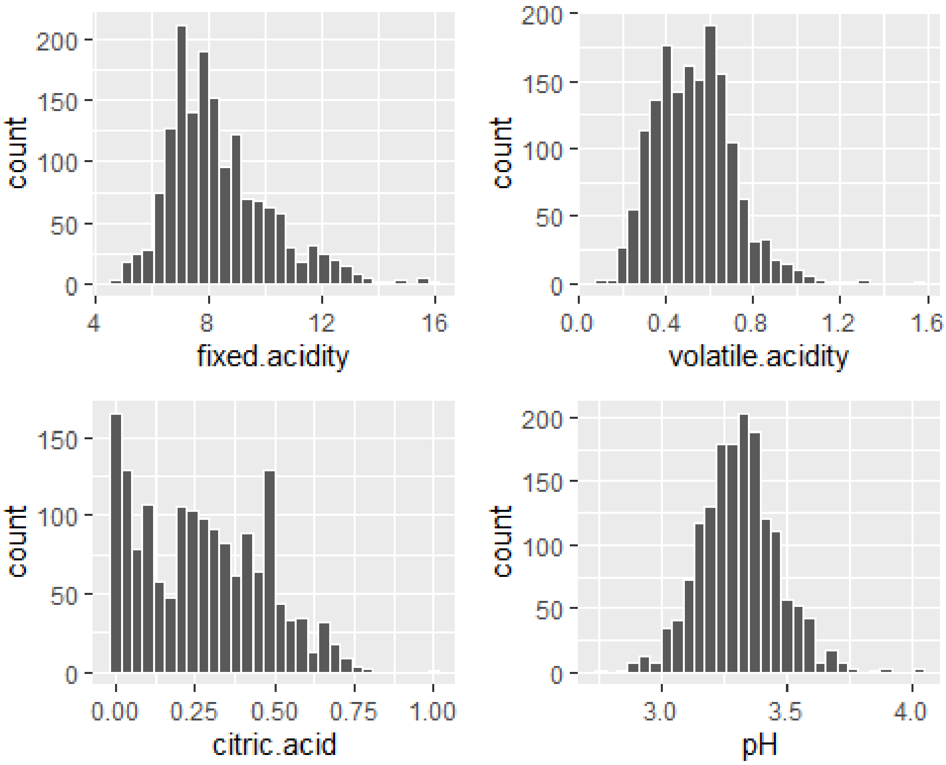
```
# Load and view the variables in data.
readURL <- function(inputURL)  #Begin function named readURL that takes a URL
{
  csvFile <- read.csv(url(inputURL), sep = ';')  #assign the results of the URL
call as a csv file to a dataframe named csvFile. Added sep = ';' to seperate the
data into columns
  return(csvFile)  # return the dataframe
}
#Using URL Functions on Red Wine URL
redWine <- readURL("https://archive.ics.uci.edu/ml/machine-learning-databases/wine-
quality/winequality-red.csv")

# Determine whether there are any 'NA' values in the dataset
redWine <- na.omit(redWine)
# The resulting dataframe is same size, so there are no NA values
data <- redWine
```

#Univariate Plots and Analysis Section

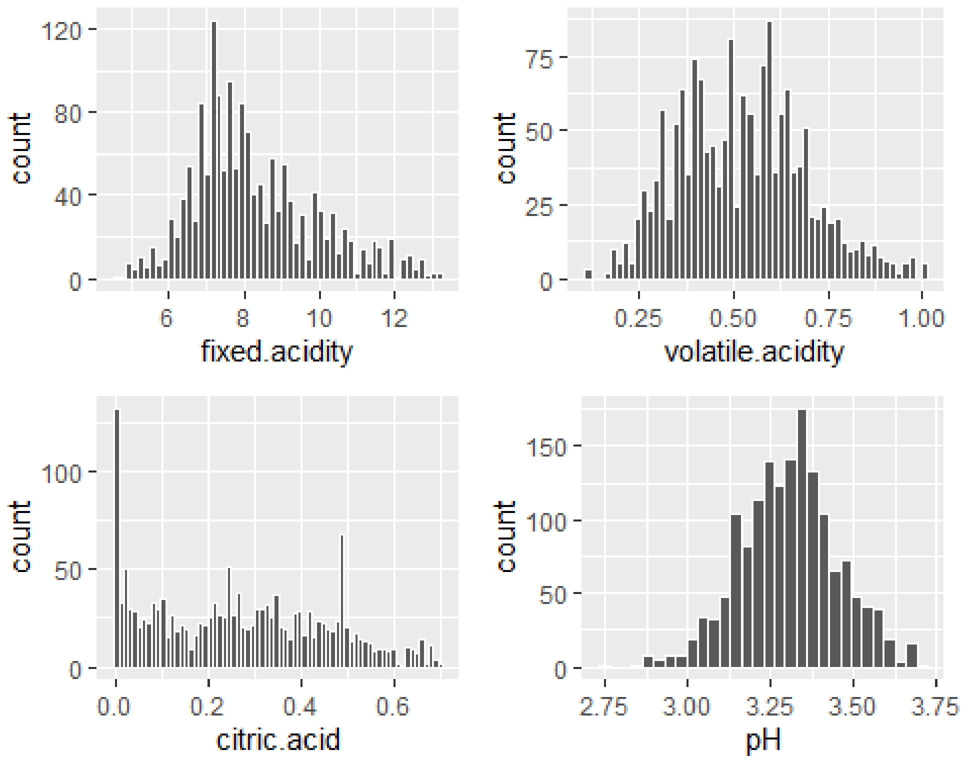
#Distribution and Histograms

##Rather than simply output 12 histograms, we will group the 12 properties into 3 different categories and look at each category in turn. Since pH is a measure of acidity, we will group pH together with the graphs showing the 3 acid levels (fixed.acidity, volatile.acidity, and citric.acid). Next,we will group together the 5 remaining concentration measurements (residual.sugar, chlorides,free.sulfur.dioxide, total.sulfur.dioxide, and sulphates). Finally, we will group together alcohol, density and quality.

```
### "Acidity" Related Histograms:
library(pdp, warn.conflicts = FALSE)
library(ggplot2, warn.conflicts = FALSE)
p1 <- ggplot(aes(fixed.acidity), data = data) + geom_histogram(bins = 30,
color="white")
p2 <- ggplot(aes(volatile.acidity), data = data) + geom_histogram(bins = 30,
color="white")
p3 <- ggplot(aes(citric.acid), data = data) + geom_histogram(bins = 30,
color="white")
p4 <- ggplot(aes(pH), data = data) + geom_histogram(bins = 30, color="white")

grid.arrange(p1,p2,p3,p4,ncol=2)
```

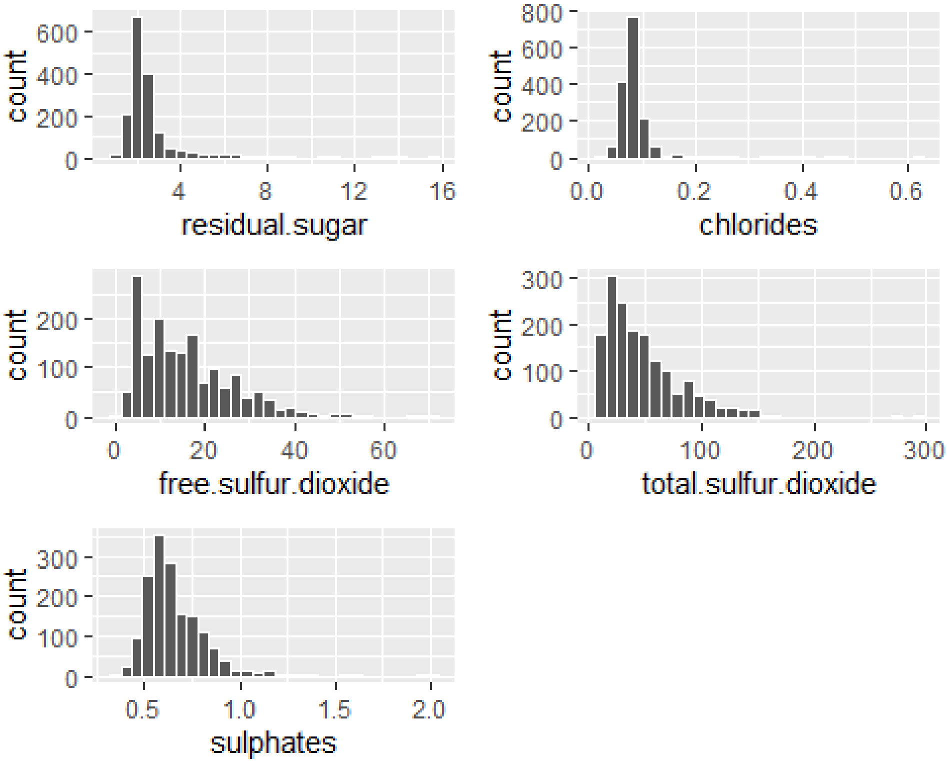


## These four parameters all look reasonably normally distributed. In all four cases, there is some positive skewing, as can be judged by the long extension on the right-hand side of the graph, with very low 'count' values for the higher x-axis values. As we get deeper into the analysis, it might make sense to exclude the upper most quantile (e.g. 1%) of each of these parameters, to remove this skewing, which appears to impact only a small number of wines (as judged by the very small count values).

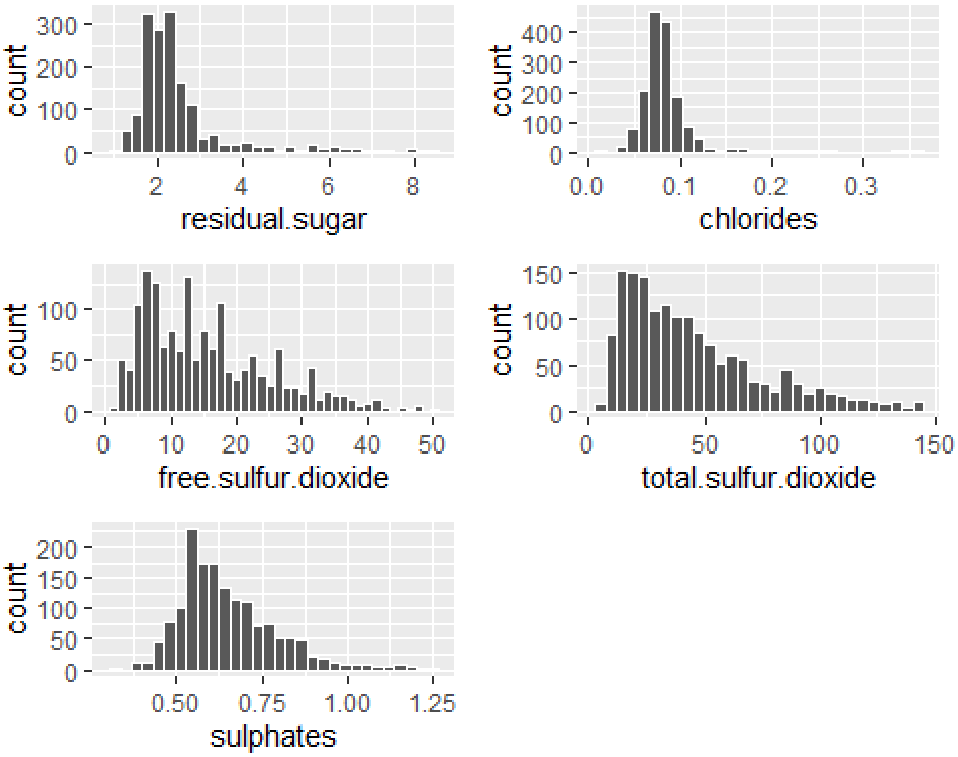


##Once the top 1%of each parameter is excluded, it is easier to see the shape of the bulk of the data. All four parameters appear to be approximately normally distributed. There are two interesting 'spikes' in the citric acid profile, one near the median and a second smaller one near a value of 0.5.

#Histograms:

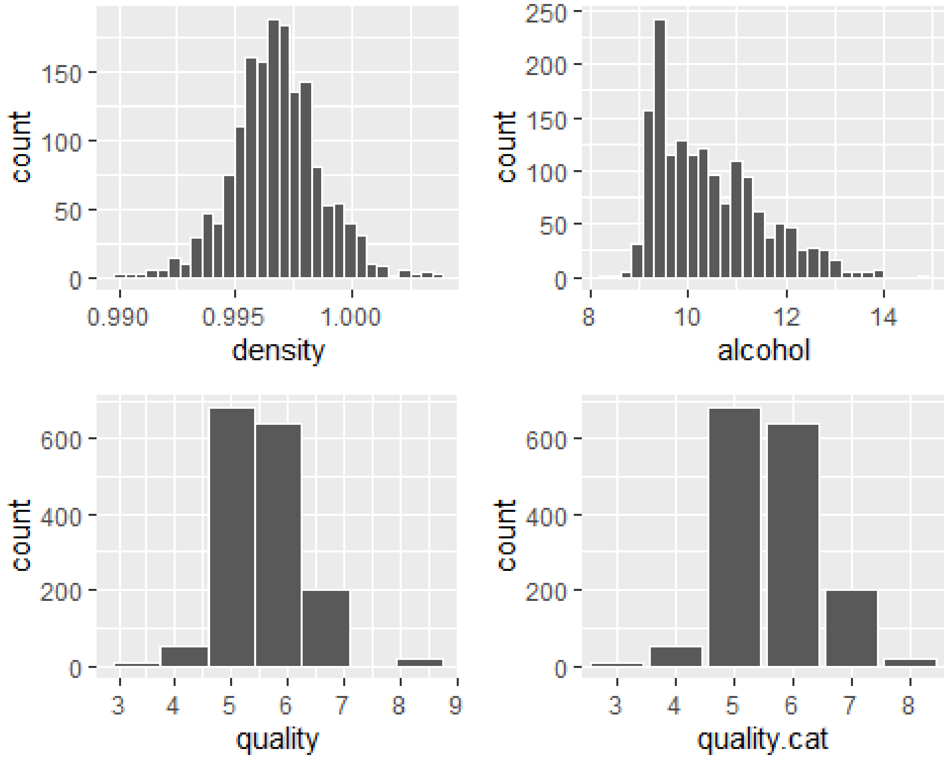


##As was seen with the four “acid” related parameters, the five graphs above also exhibit positive skew.

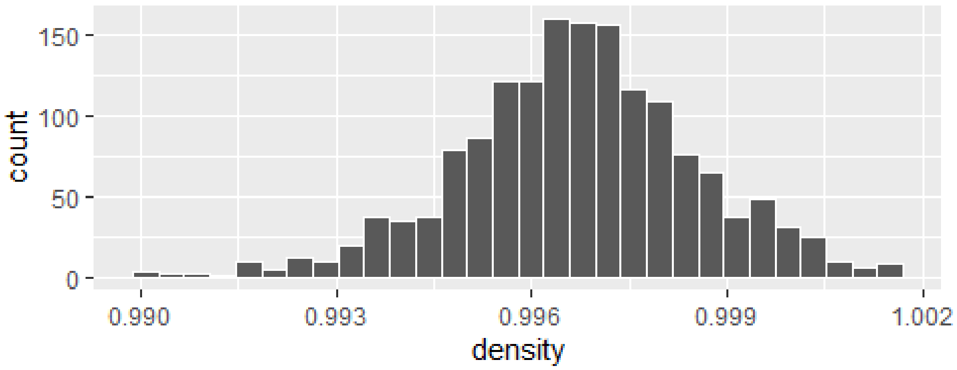


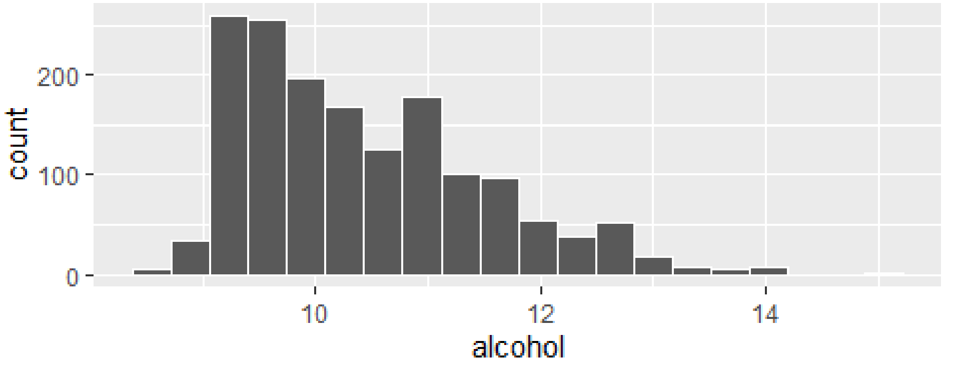
##Once the top 1%of each parameter is excluded, it is easier to see the shape of the bulk of the data. Most parameters appear to be approximately normally distributed here, with the exception of residual sugar. ###(Note: a bar chart is used in the case of 'quality.cat', since it is categorical):

```
##  Ord.factor w/ 6 levels "3"<"4"<"5"<"6"<..: 3 3 3 4 3 3 3 5 5 3 ...
```



## The quality rating appears to be normally distributed, with the bulk of assessments in the middle bins. Density appears normal too, but with some positive skew. The alcohol content looks interesting. ##Density looks fairly normally distributed, whereas alcohol content does not.





#Create New Variables:

Reference: http://beerandwinejournal.com/chloride-and-sulfate/

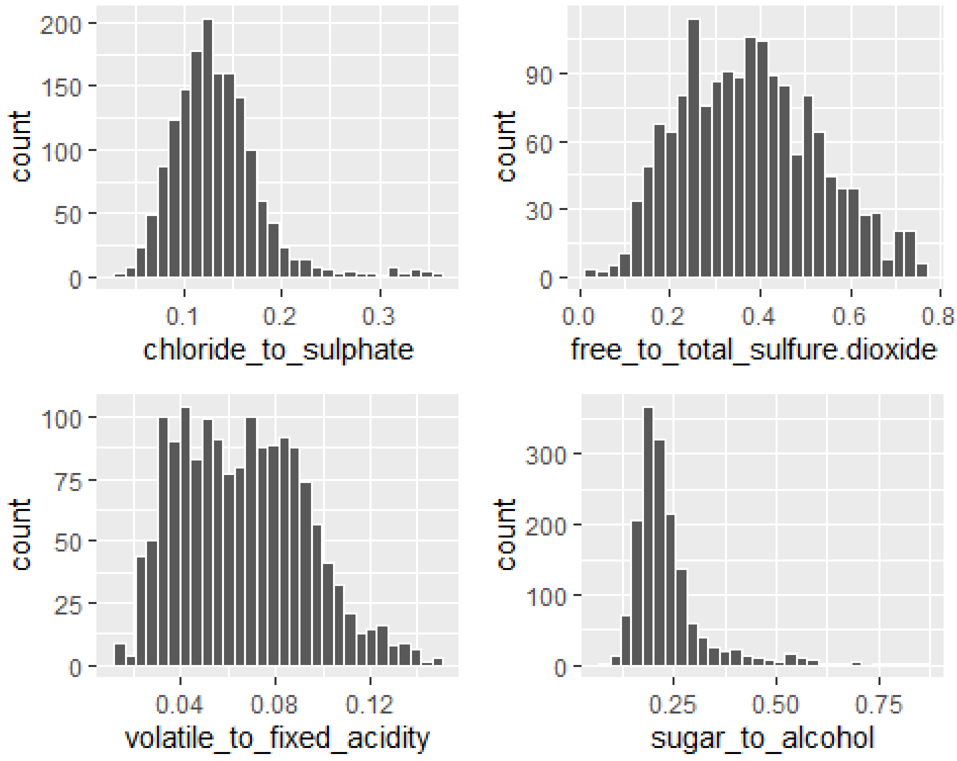
The chlorides to sulphates ratio might be a far more important measure of quality than the individual levels of either ion. Thus, we will create a chlorides-to-sulphate ratio variable.

```
# Create and add four new variables to the dataframe:
data$chloride_to_sulphate <-with(data,chlorides / sulphates)
data$free_to_total_sulfure.dioxide <-with(data,free.sulfur.dioxide /
total.sulfur.dioxide)
data$volatile_to_fixed_acidity <-with(data,volatile.acidity / fixed.acidity)
data$sugar_to_alcohol <-with(data,residual.sugar / alcohol)


# Output summary data on the new variables:
str(subset(data,select =
c(chloride_to_sulphate,free_to_total_sulfure.dioxide,volatile_to_fixed_acidity,suga
r_to_alcohol)))

## 'data.frame':    1599 obs. of  4 variables:
##  $ chloride_to_sulphate         : num  0.136 0.144 0.142 0.129 0.136 ...
##  $ free_to_total_sulfure.dioxide: num  0.324 0.373 0.278 0.283 0.324 ...
##  $ volatile_to_fixed_acidity    : num  0.0946 0.1128 0.0974 0.025 0.0946 ...
##  $ sugar_to_alcohol             : num  0.202 0.265 0.235 0.194 0.202 ...
```

```
summary(subset(data,select =
c(chloride_to_sulphate,free_to_total_sulfure.dioxide,volatile_to_fixed_acidity,suga
r_to_alcohol)))
```

```
##  chloride_to_sulphate free_to_total_sulfure.dioxide volatile_to_fixed_acidity
##  Min.   :0.03077      Min.   :0.02273               Min.   :0.01348
##  1st Qu.:0.10455      1st Qu.:0.25926               1st Qu.:0.04405
##  Median :0.12833      Median :0.37500               Median :0.06569
##  Mean   :0.13572      Mean   :0.38231               Mean   :0.06706
##  3rd Qu.:0.15581      3rd Qu.:0.48485               3rd Qu.:0.08581
##  Max.   :0.57761      Max.   :0.85714               Max.   :0.20800
##  sugar_to_alcohol
##  Min.   :0.07087
##  1st Qu.:0.18306
##  Median :0.21111
##  Mean   :0.24550
##  3rd Qu.:0.25481
##  Max.   :1.71111
```

#Plot the new parameters as a group:



## The free:total sulfur dioxide graph looks normally distributed. The chloride:sulphate, volatile:fixed acidity and sugar:alcohol graphs look positively skewed.

#Bivariate Plots and Analysis Section

#Linear Model Red Wine

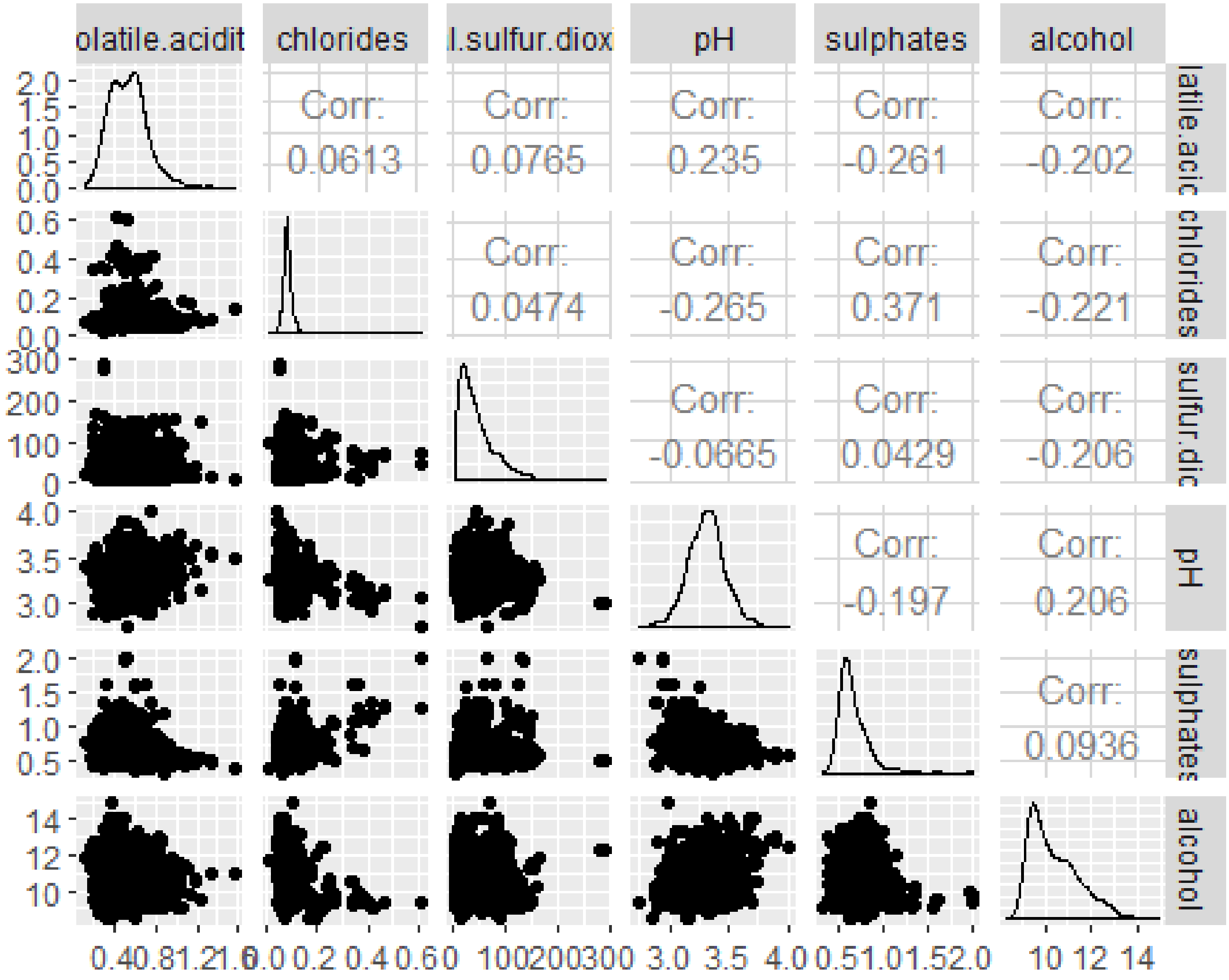
```
##
## Call:
## lm(formula = quality ~ ., data = subset(data, select = -c(quality.cat,
##     chloride_to_sulphate, free_to_total_sulfure.dioxide, volatile_to_fixed_acidity,
##     sugar_to_alcohol)))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.68911 -0.36652 -0.04699  0.45202  2.02498
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.197e+01  2.119e+01   1.036   0.3002
## fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357
## volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
## citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150
## residual.sugar        1.633e-02  1.500e-02   1.089   0.2765
## chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
## free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *
## total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
## density              -1.788e+01  2.163e+01  -0.827   0.4086
## pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *
## sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
## alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.648 on 1587 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3561
## F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

##An initial look at the Linear Regression Model shows multiple variables are statistically significant (p-value<0.05). Running the regression model fora subset of data based statistical significance

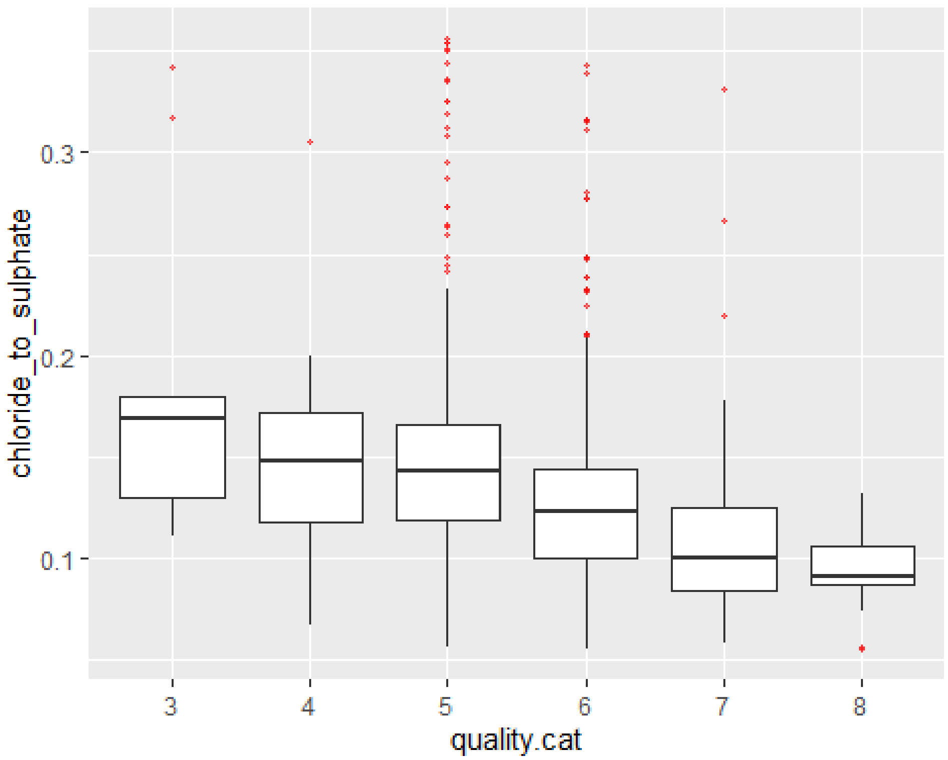
```
linRegressionWine2<-lm(formula =
quality~volatile.acidity+chlorides+total.sulfur.dioxide+pH+sulphates+alcohol,data =
data)
summary(linRegressionWine2)

##
## Call:
## lm(formula = quality ~ volatile.acidity + chlorides + total.sulfur.dioxide +
##     pH + sulphates + alcohol, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60575 -0.35883 -0.04806  0.46079  1.95643
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.2957316  0.3995603  10.751  < 2e-16 ***
## volatile.acidity     -1.0381945  0.1004270 -10.338  < 2e-16 ***
## chlorides            -2.0022839  0.3980757  -5.030 5.46e-07 ***
## total.sulfur.dioxide -0.0023721  0.0005064  -4.684 3.05e-06 ***
## pH                   -0.4351830  0.1160368  -3.750 0.000183 ***
## sulphates             0.8886802  0.1100419   8.076 1.31e-15 ***
## alcohol               0.2906738  0.0168108  17.291  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6487 on 1592 degrees of freedom
## Multiple R-squared:  0.3572, Adjusted R-squared:  0.3548
## F-statistic: 147.4 on 6 and 1592 DF,  p-value: < 2.2e-16
```

#Determination Coefficient: 35.48%of Quality can be explained by these attributes

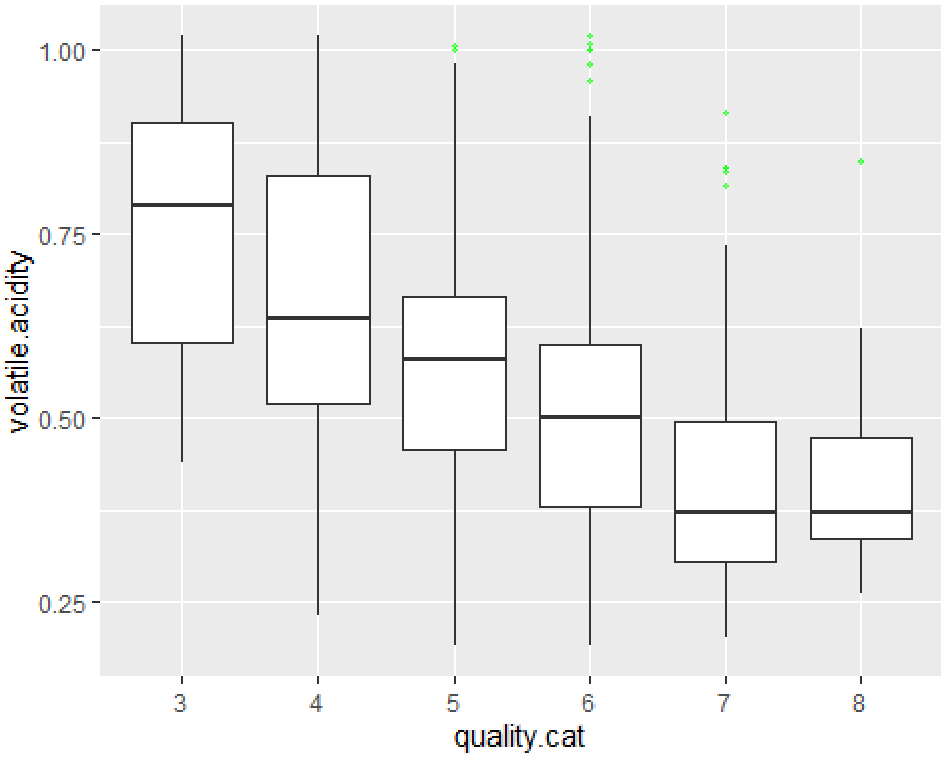
# Scatterplot matrix

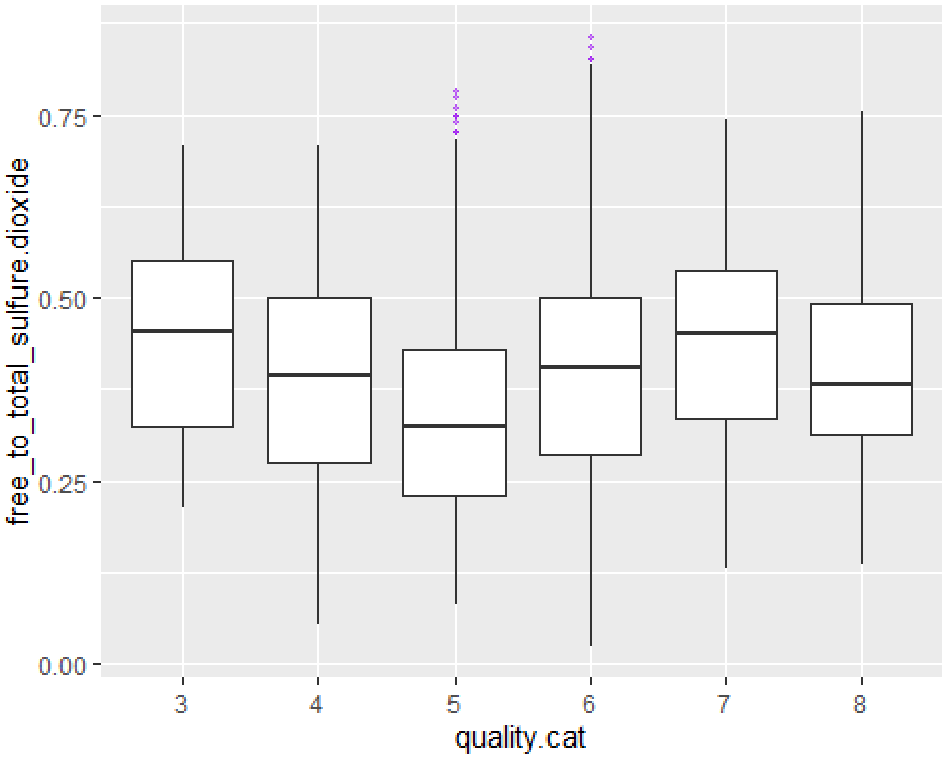


#Expanding on the highest correlation coefficients,as this graph is too dense to draw conclusions # Bivariate pairs
##a. Quality and the Chloride:Sulphate Ratio

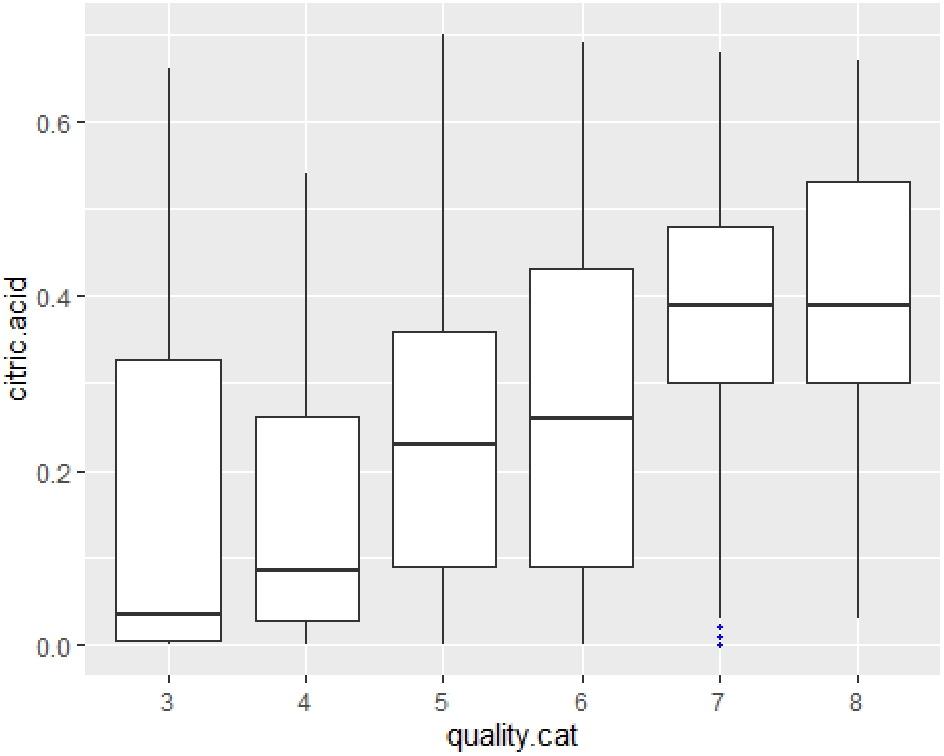


##It appears that higher quality wines have lower chloride:sulphate ratios ##b. Quality and Volatile Acidity

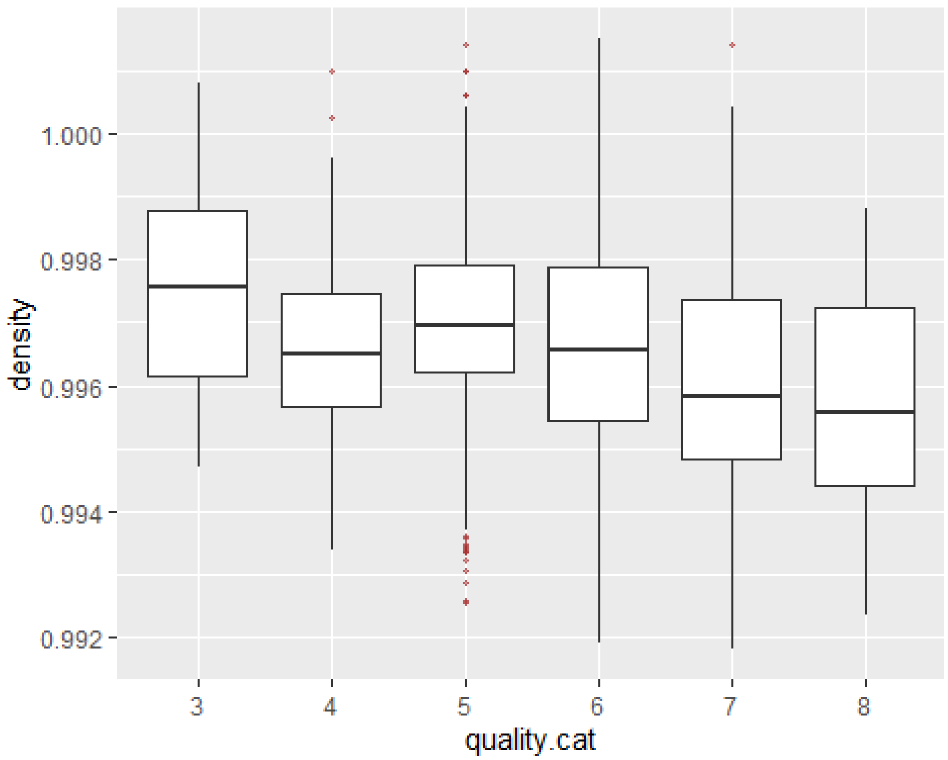




##It appears that higher quality wines have lower volatile acidity #c. Quality and the Free:Total Sulfur Dioxide Ratio ##d. Quality and citric.acid

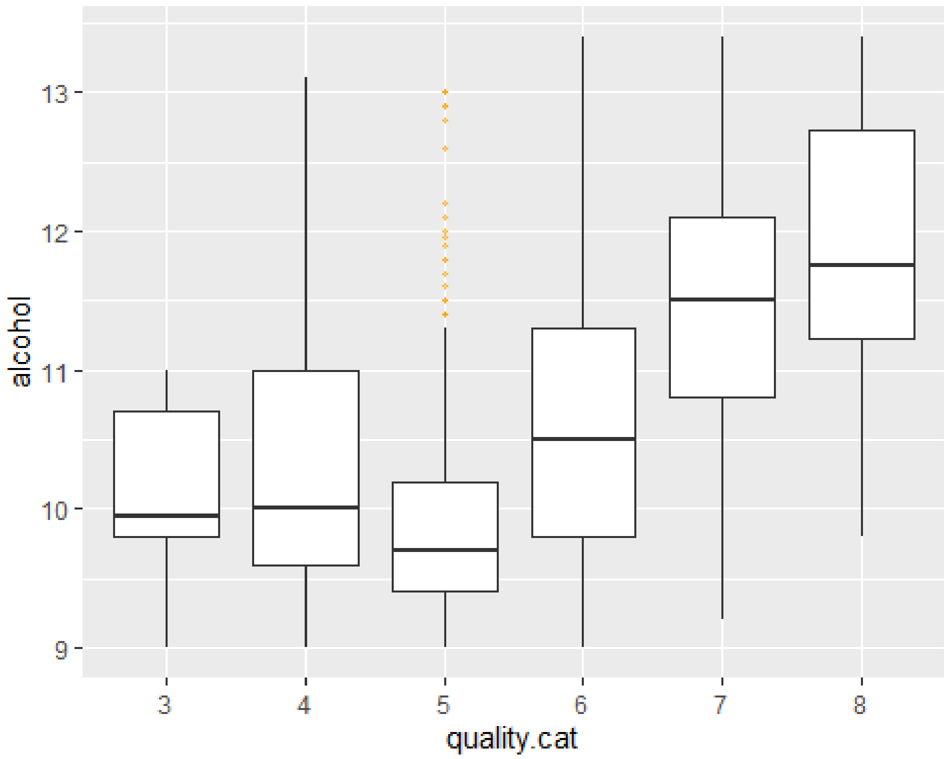


##It appears that in general, higher quality wines have higher citric acid levels. ##e. Quality and Density

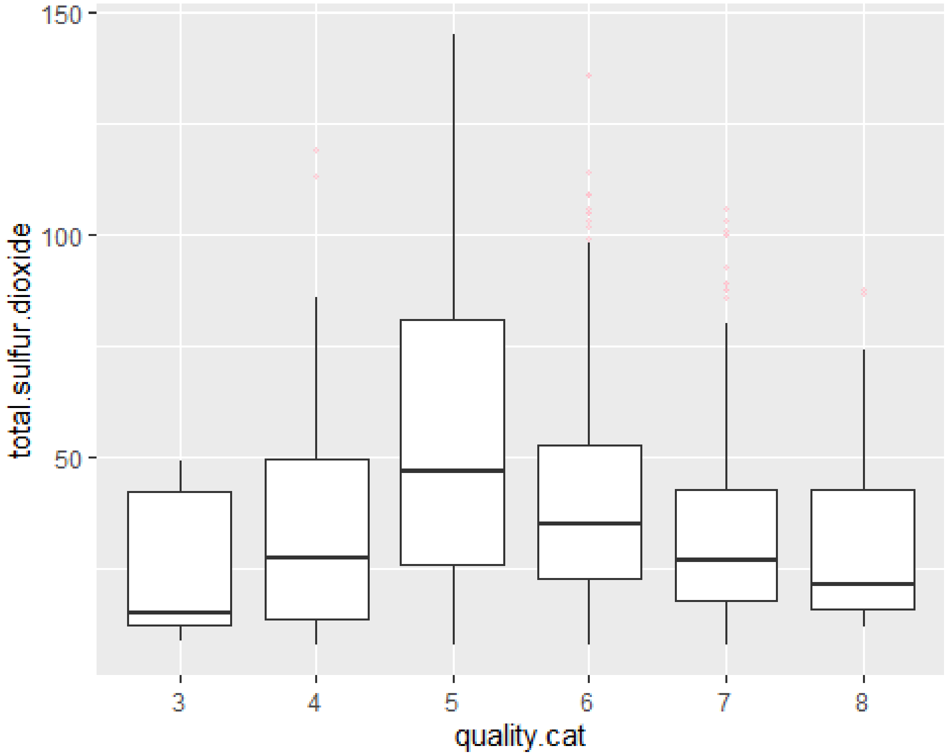


# It is hard to discern any clear trend between the density and a wine’s quality, given that the median values move up and down as the quality improves.

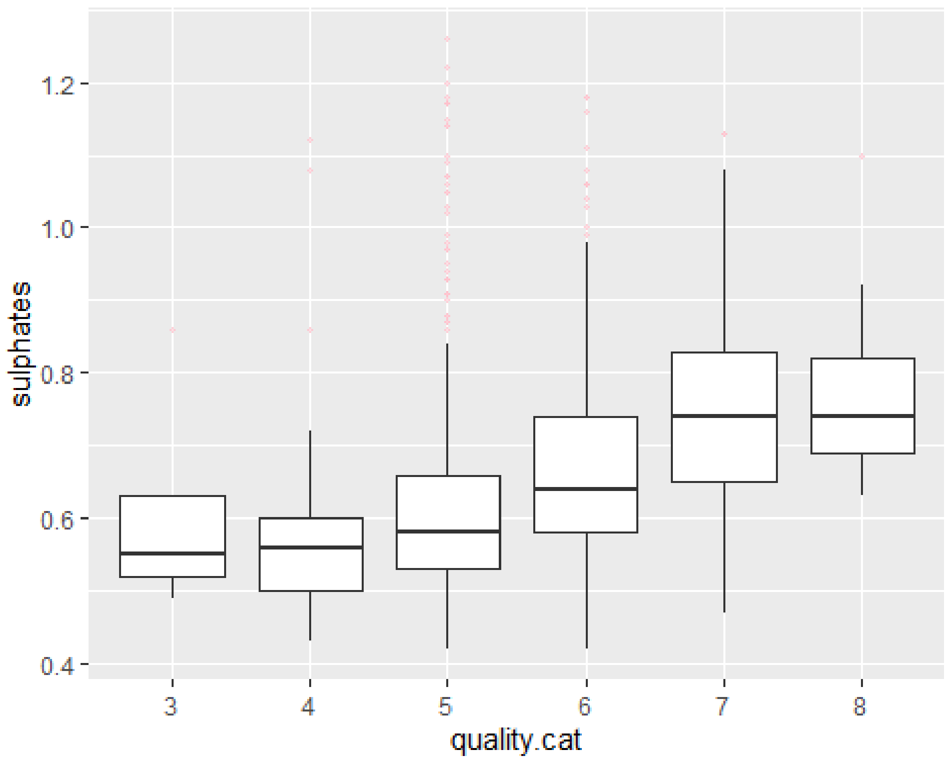
##f. Quality and Alcohol Content



##The relationship between alcohol content and quality appears potentially promising, particularly at the higher end of the quality scale, where there is a clear upwards trend in quality (from levels 6 through 8). ##g. Quality and Total Sulphur Dioxide



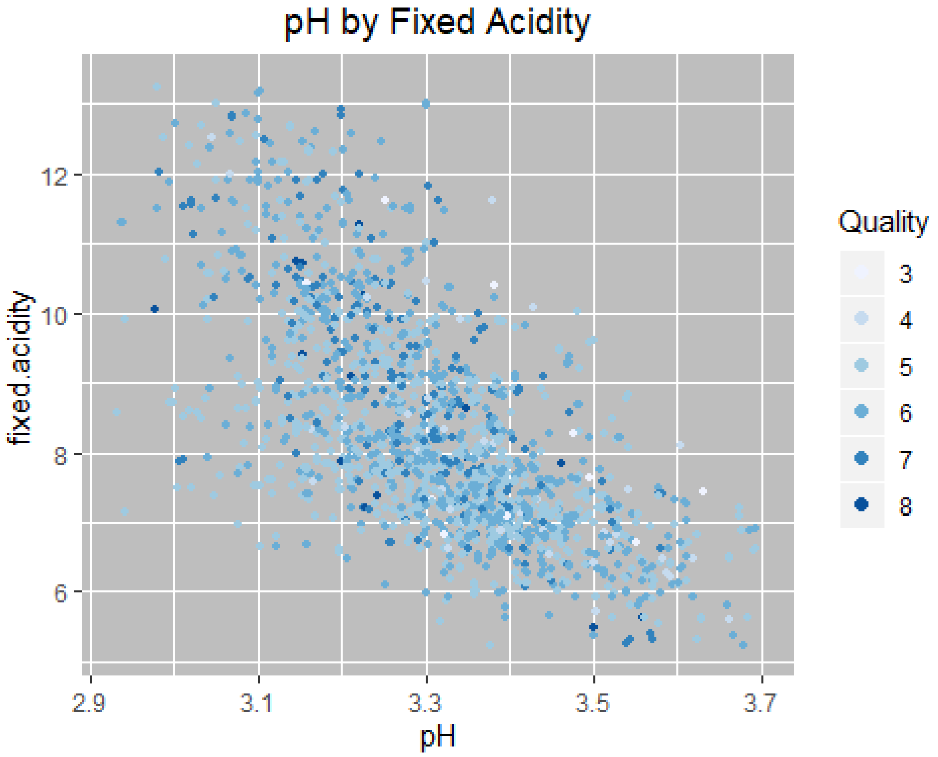
##It is hard to discern any clear trend between the total sulfur dioxide and a wine's quality.
##h. Quality and Sulphates

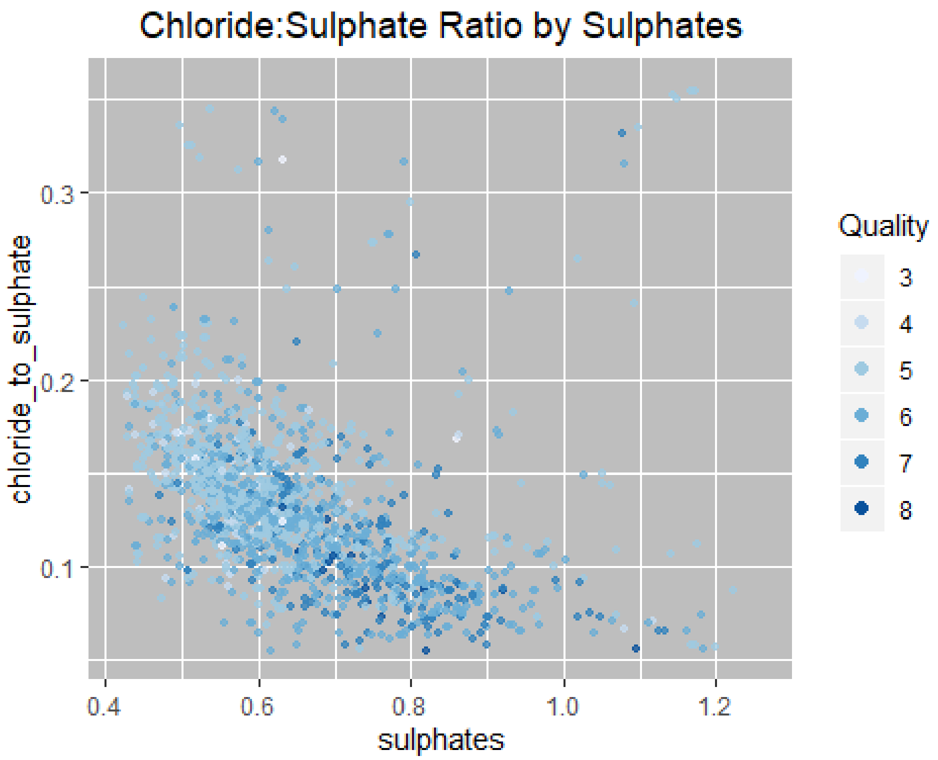


##It appears that in general, higher quality wines have higher sulphate levels.

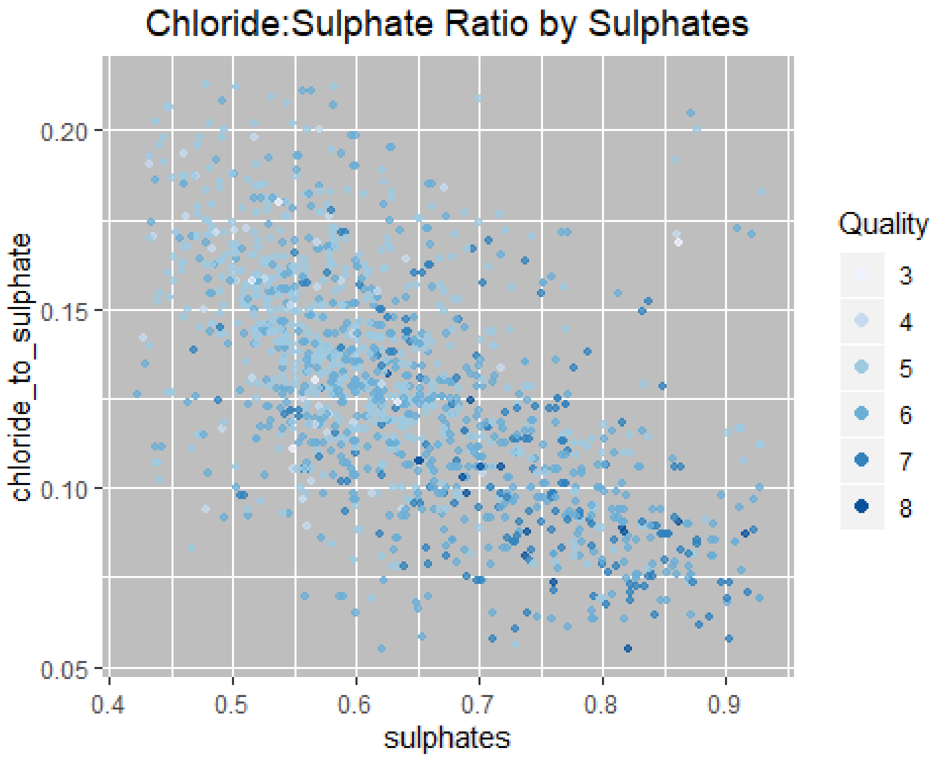
# Multivariate Plots and Analysis Section

##We will now consider the interaction of multiple variables. First, it was observed in the bivariate analysis that there is a relatively strong inverse relationship between fixed acidity and pH(correlation coefficient of -0.68).

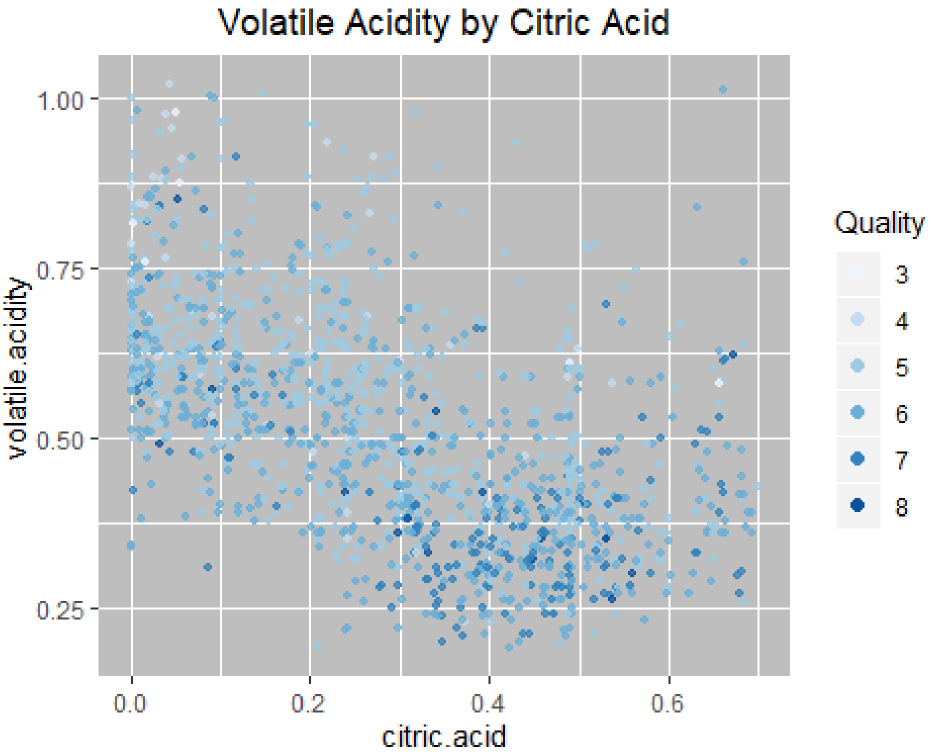




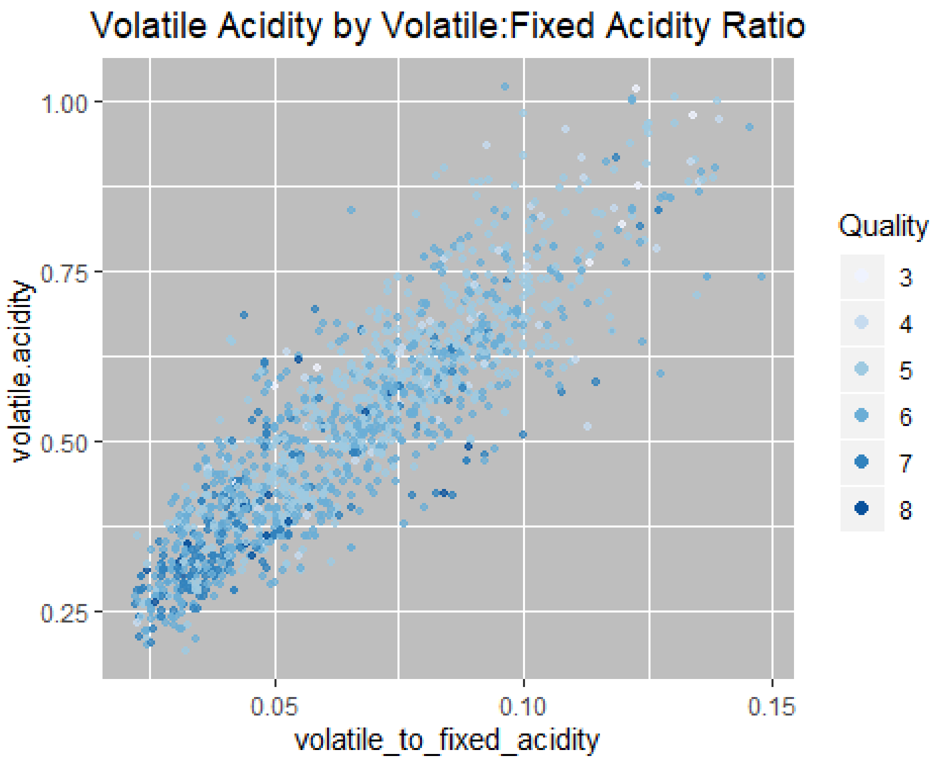
# It appears there might be a tendency for high quality wines to be high sulphate levels and low chloride:sulphate ratio. Let’s zoom in on the lower left portion of the graph, which contains most of the data points, by truncating out the top 5%quantile for each variable:



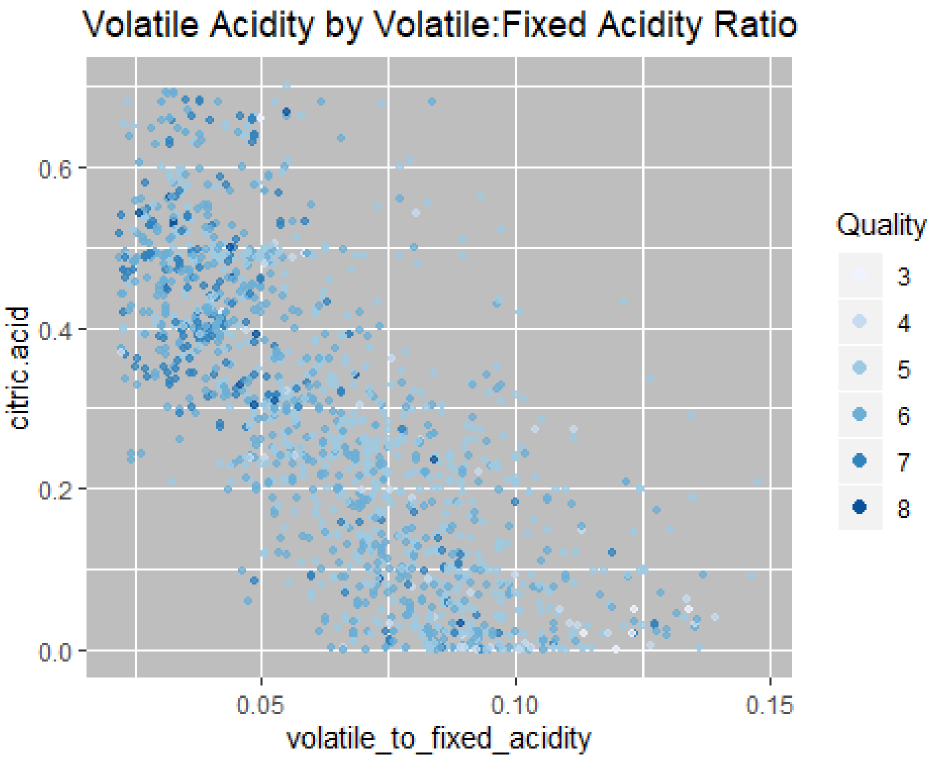
## There does indeed appear to be a tendency for the higher quality wines to be higher in sulphates and lower chloride:sulphate ratio, given that the quality 7-8 wines have tended to cluster in the lower right portion of the graph, whereas the quality 3-5 wines are more in the upper left portion.



# There is no strong pattern regarding where the higher versus lower quality wines fall on the graph. The quality points are dispersed throughout, even though there might be some weak relationships in terms of where they tend to fall.



## There does indeed appear to be a tendency for the higher quality wines to be lower in volatile acidity and volatile:fixed acidity ratio, given that the quality 7-8 wines have tended to cluster in the lower left portion of the graph, whereas the quality 3-5 wines are more in the upper right portion.

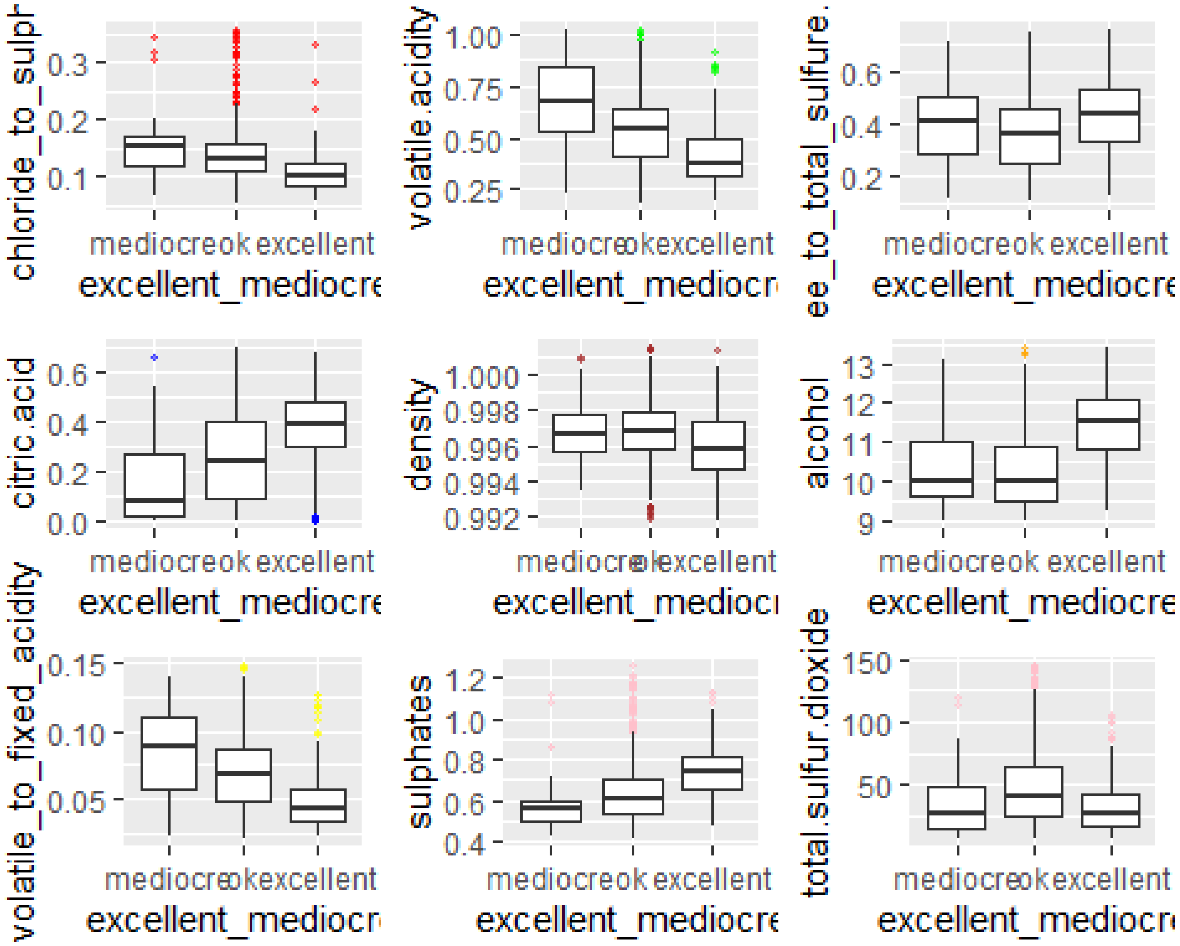


## There does indeed appear to be a tendency for the higher quality wines to be higher in citric acid levels and lower volatile:fixed acidity ratio, given that the quality 7-8 wines have tended to cluster in the upper left portion of the graph, whereas the quality 3-5 wines are more in the lower right portion.

#Additional Data Transformation

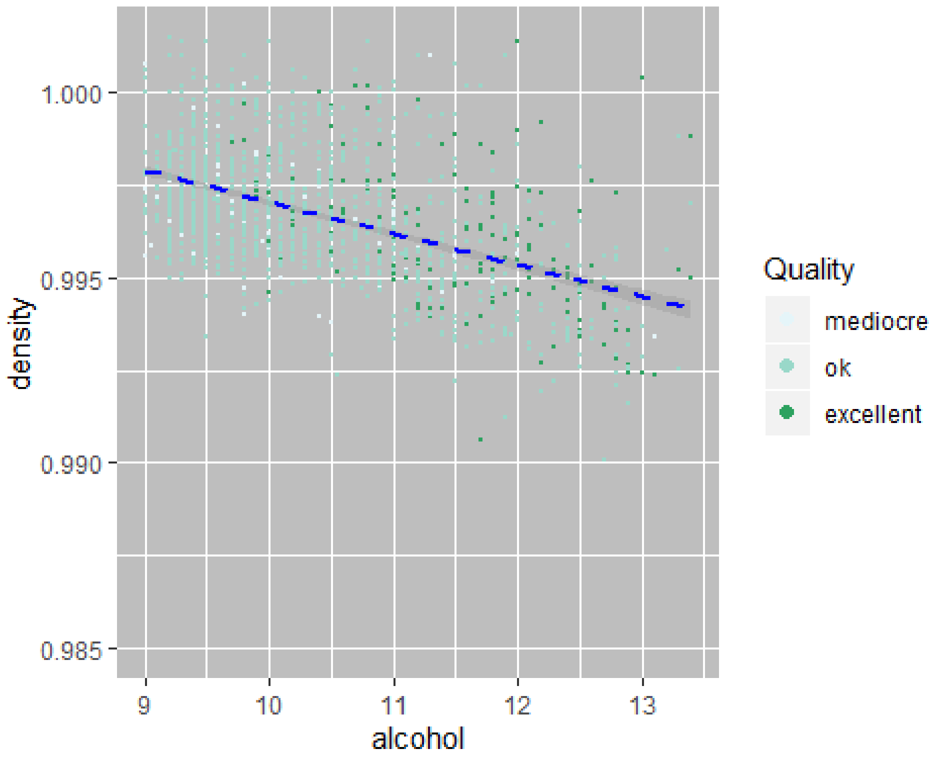
##Let’s consider any wine with a 3-4 rating as ‘mediocre’, a wine with a 5-6 rating as ‘ok’ and a wine with a 7-8 rating as ‘excellent’.

```
## mediocre        ok excellent 
##       63      1319       217
```



#The following variables correlate inversely with quality (i.e. quality decreases as these variables increase in value):
##* chloride:sulphate ratio
##* volatile acidity
##* Volatile:fixed acidity
##* density
#The following variables correlate with quality (i.e. quality increases as these variables increase in value):

##* alcohol content
##* citric acid ##* sulphates # the new quality categories on the density vs. alcohol content graph:



#The categories split quite well: good wines tend to have higher alcohol content and lower density levels.

#Predictive Models

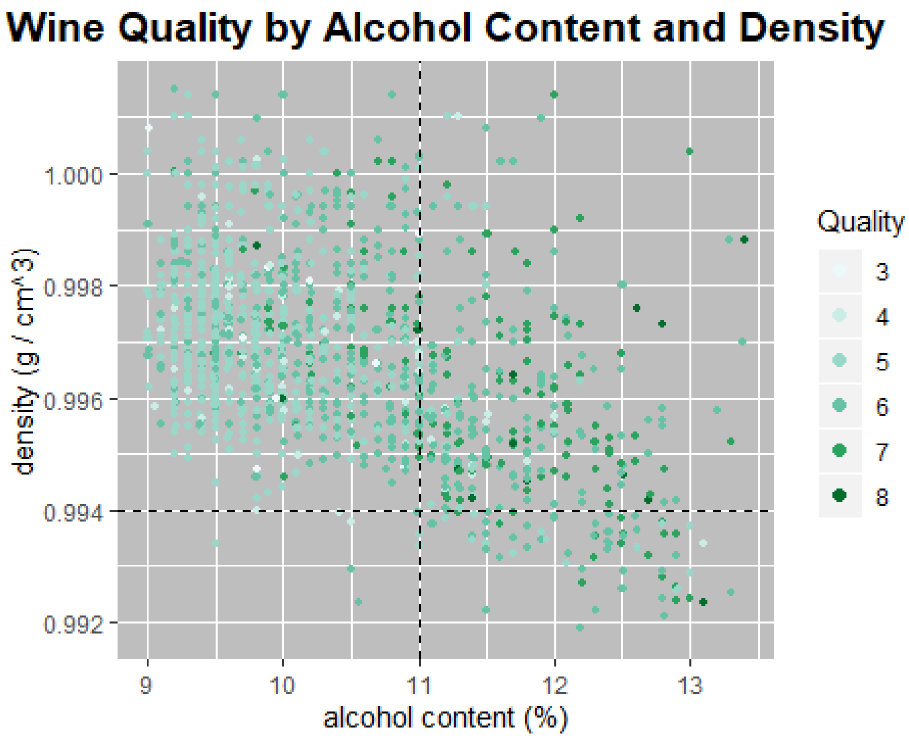
```
## Call:
## polr(formula = quality.cat ~ alcohol + density + sulphates +
##     citric.acid + volatile.acidity + total.sulfur.dioxide + chloride_to_sulphate
##     volatile_to_fixed_acidity + free_to_total_sulfure.dioxide,
##     data = data, Hess = TRUE)
##
## Coefficients:
##                                    Value Std. Error t value
## alcohol                         0.869203   0.059471  14.616
## density                        -9.426778   0.427938 -22.028
## sulphates                       2.070837   0.331911   6.239
## citric.acid                    -0.628833   0.423761  -1.484
## volatile.acidity               -0.841046   0.720169  -1.168
## total.sulfur.dioxide           -0.005829   0.001742  -3.346
## chloride_to_sulphate           -4.259009   1.093675  -3.894
## volatile_to_fixed_acidity     -20.428885   5.339168  -3.826
## free_to_total_sulfure.dioxide   1.041803   0.365607   2.850
##
## Intercepts:
##     Value    Std. Error t value
## 3|4  -7.3751   0.4681   -15.7550
## 4|5  -5.4386   0.4695   -11.5847
## 5|6  -1.7003   0.4804    -3.5396
## 6|7   1.1474   0.5160     2.2238
## 7|8   4.1439   0.5785     7.1634
##
## Residual Deviance: 3073.782
## AIC: 3101.782
## [1] "Confidence Levels:"
##                                      2.5 %       97.5 %
## alcohol                          0.752641665  0.985764951
## density                        -10.265521765 -8.588034596
## sulphates                        1.420303554  2.721370147
## citric.acid                     -1.459389427  0.201723332
## volatile.acidity                -2.252550543  0.570458515
## total.sulfur.dioxide            -0.009242937 -0.002414235
## chloride_to_sulphate            -6.402572413 -2.115445487
## volatile_to_fixed_acidity      -30.893461066 -9.964308536
## free_to_total_sulfure.dioxide    0.325227053  1.758379079
```

#The model can also be built for the scenario where the 'transformed' quality categories of 'mediocre', 'ok', and 'excellent' are the desired prediction outcome, and those modeling results are as follows:

```
##
## Re-fitting to get Hessian
## Call:
## polr(formula = data$excellent_mediocre ~ alcohol + density +
##     sulphates + citric.acid + volatile.acidity + total.sulfur.dioxide +
##     chloride_to_sulphate + volatile_to_fixed_acidity +
free_to_total_sulfure.dioxide,
##     data = data)
##
## Coefficients:
##                                   Value Std. Error   t value
## alcohol                        6.968e-01   0.078474    8.8790
## density                       -1.515e+02   0.552851 -274.0595
## sulphates                      2.026e+00   0.450147    4.5015
## citric.acid                    6.508e-01   0.608719    1.0691
## volatile.acidity               1.583e-01   1.043732    0.1516
## total.sulfur.dioxide          -1.173e-03   0.002376   -0.4937
## chloride_to_sulphate          -5.147e+00   1.521937   -3.3819
## volatile_to_fixed_acidity     -2.938e+01   7.774516   -3.7789
## free_to_total_sulfure.dioxide  6.140e-01   0.520367    1.1800
##
## Intercepts:
##              Value     Std. Error t value
## mediocre|ok  -148.7238    0.6197  -240.0107
## ok|excellent -142.0578    0.6863  -206.9824
##
## Residual Deviance: 1378.793
## AIC: 1400.793
## [1] "Confidence Levels:"
## Re-fitting to get Hessian
##                                     2.5 %        97.5 %
## alcohol                        0.54295899  8.505704e-01
## density                       -152.59758633 -1.504305e+02
## sulphates                        1.14405899  2.908601e+00
## citric.acid                     -0.54225803  1.843875e+00
## volatile.acidity                -1.88740658  2.203948e+00
## total.sulfur.dioxide            -0.00582904  3.483237e-03
## chloride_to_sulphate            -8.13002194 -2.164140e+00
## volatile_to_fixed_acidity      -44.61673771 -1.414119e+01
## free_to_total_sulfure.dioxide   -0.40589034  1.633911e+00
```

#Both models appear to fit the data well, with the estimated value to standard error ratio (i.e. the t-value) exceeding 2.9 for all parameters. Both models have limitations, however. First, they are only valid for the quality range exhibited in the dataset. Since the dataset only contained wines in the 3-9 quality range, these models would be unreliable at identifying wines outside of this range. Second, the models are only valid for the wine under consideration here (i.e. Portuguese "Vinho Verde" wines). A new model would likely be needed for each wine variety, or at the very least, this model would need to be validated against a new set of data before one could make any claims about its applicability beyond this particular dataset and wine variety.

#Final Plot and Summary



#This plot demonstrates that in general, the high-quality wines (quality 7-8) tend to have high alcohol content and low density, as shown by the preponderance of green shaded points in the lower right quadrant of the graph. Conversely, the poor-quality wines (quality 3-4) tend to have low alcohol content and high density, dominating the two left side quadrants.

WHITE WINE ANALYTICS

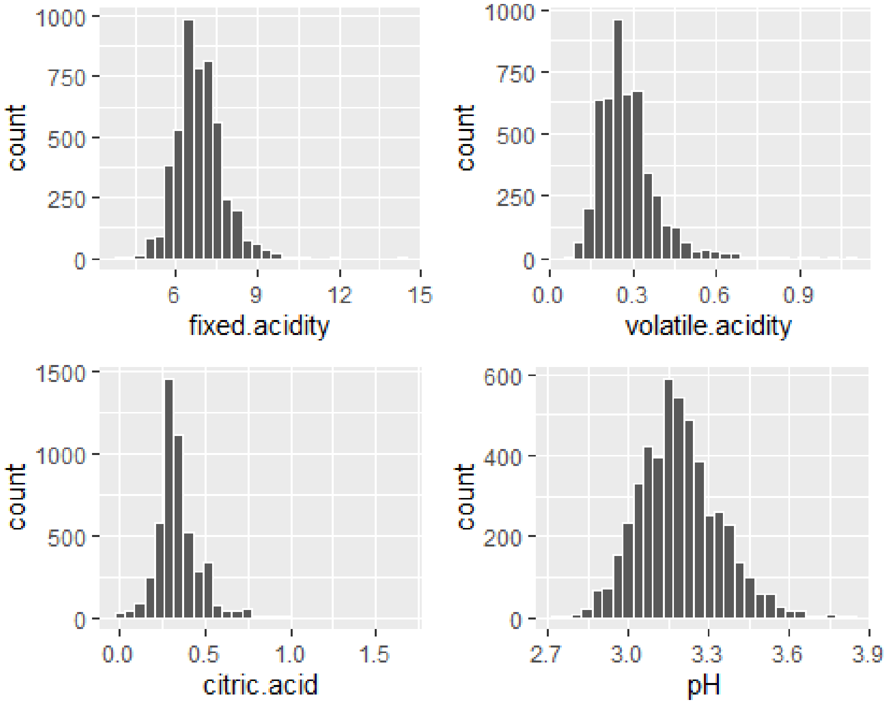
```
# Load and view the variables in data.

readURL <- function(inputURL)  #Begin function named readURL that takes a URL
{
  csvFile <- read.csv(url(inputURL), sep = ';')  #assign the results of the URL
call as a csv file to a dataframe named csvFile. Added sep = ';' to seperate the
data into columns
  return(csvFile)  # return the dataframe
}
#Using URL Functions on Red Wine URL
WhiteWine <- readURL("https://archive.ics.uci.edu/ml/machine-learning-
databases/wine-quality/winequality-white.csv")

# Determine whether there are any 'NA' values in the dataset
WhiteWine <- na.omit(WhiteWine)
# The resulting dataframe is same size, so there are no NA values
data <- WhiteWine
```

#Univariate Plots and Analysis Section

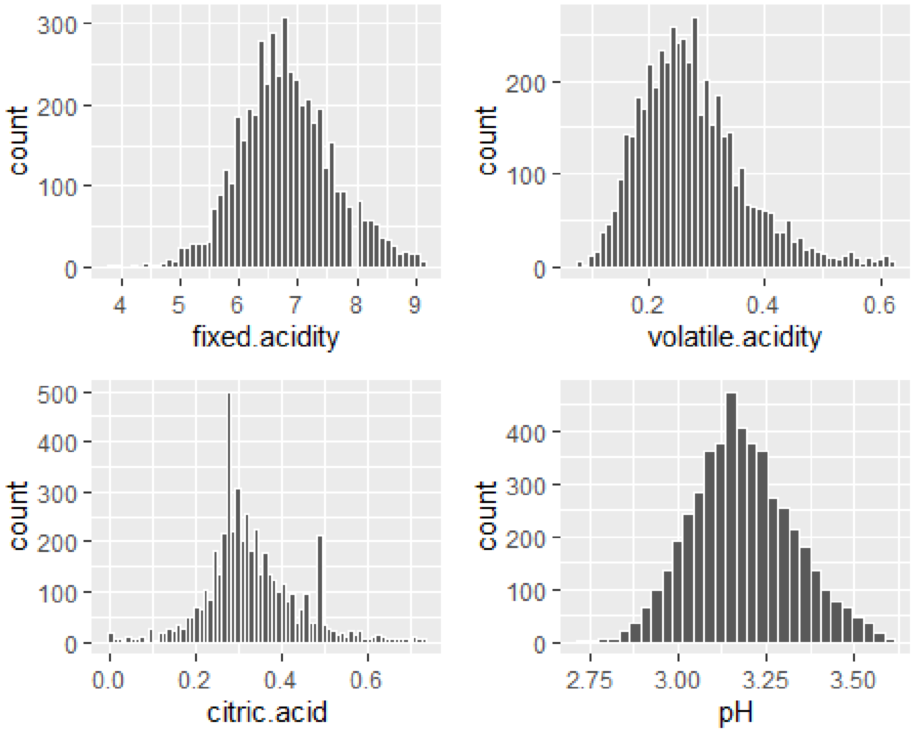
##Rather than simply output 12 histograms, we will group the 12 properties into 3 different categories and look at each category in turn. Since pH is a measure of acidity, we will group pH together with the graphs showing the 3 acid levels (fixed.acidity, volatile.acidity, and citric.acid). Next, we will group together the 5 remaining concentration measurements (residual.sugar, chlorides,free.sulfur.dioxide, total.sulfur.dioxide, and sulphates). Finally, we will group together alcohol, density and quality.

#Distribution and Histograms

```
### "Acidity" Related Histograms:
library(pdp, warn.conflicts = FALSE)
library(ggplot2, warn.conflicts = FALSE)
p1 <- ggplot(aes(fixed.acidity), data = data) + geom_histogram(bins = 30,
color="white")
p2 <- ggplot(aes(volatile.acidity), data = data) + geom_histogram(bins = 30,
color="white")
p3 <- ggplot(aes(citric.acid), data = data) + geom_histogram(bins = 30,
color="white")
p4 <- ggplot(aes(pH), data = data) + geom_histogram(bins = 30, color="white")
grid.arrange(p1,p2,p3,p4,ncol=2)
```

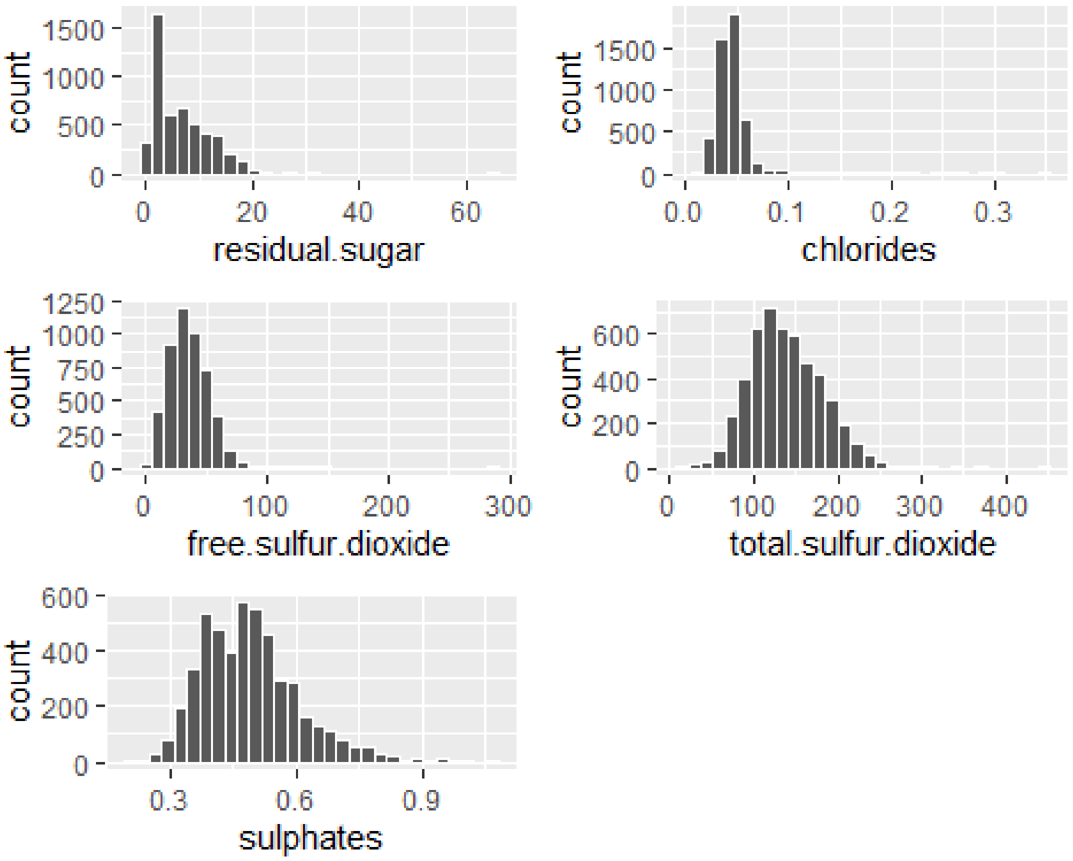


## These four parameters look normally distributed with a positive skew (long extension on the right-hand side of the graph), with very low 'count' values for the higher x-axis values. It might make sense to exclude the upper most quantile (e.g. 1%) of each of these parameters, to remove this skewing. Note the 'spikes' in the citric acid profile, could be potential outliers.

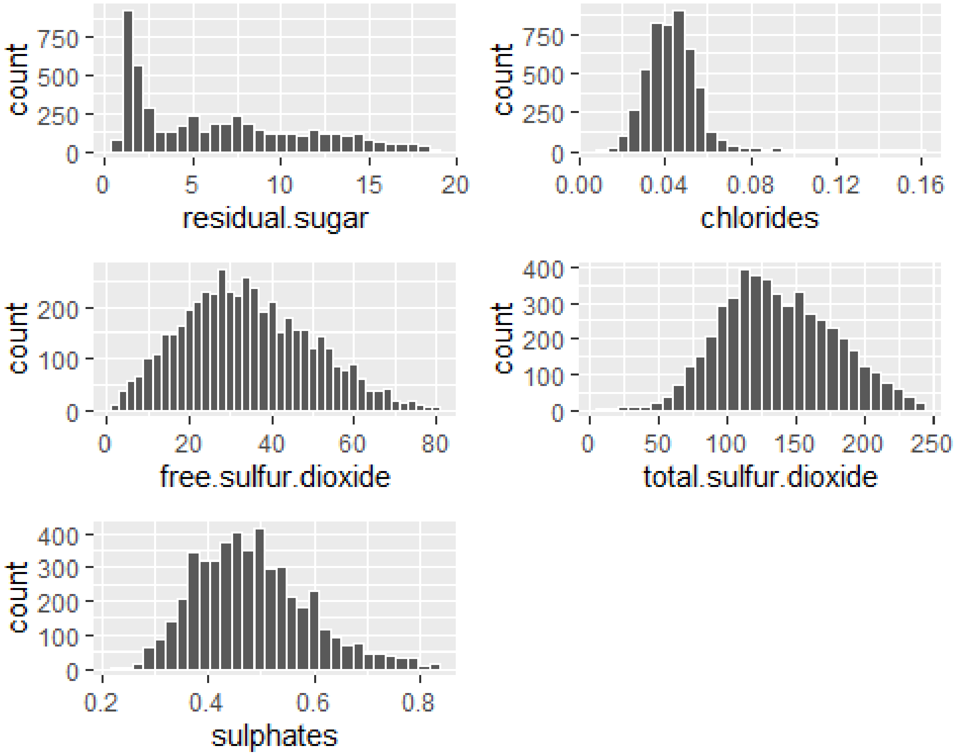


##Once the top 1%of each parameter is excluded, it is easier to see the shape of the bulk of the data. All parameters seem normally distributed. Note the 'spikes' in the citric acid profile.

##As was seen with the four “acid” related parameters, the five graphs above also exhibit positive skew.

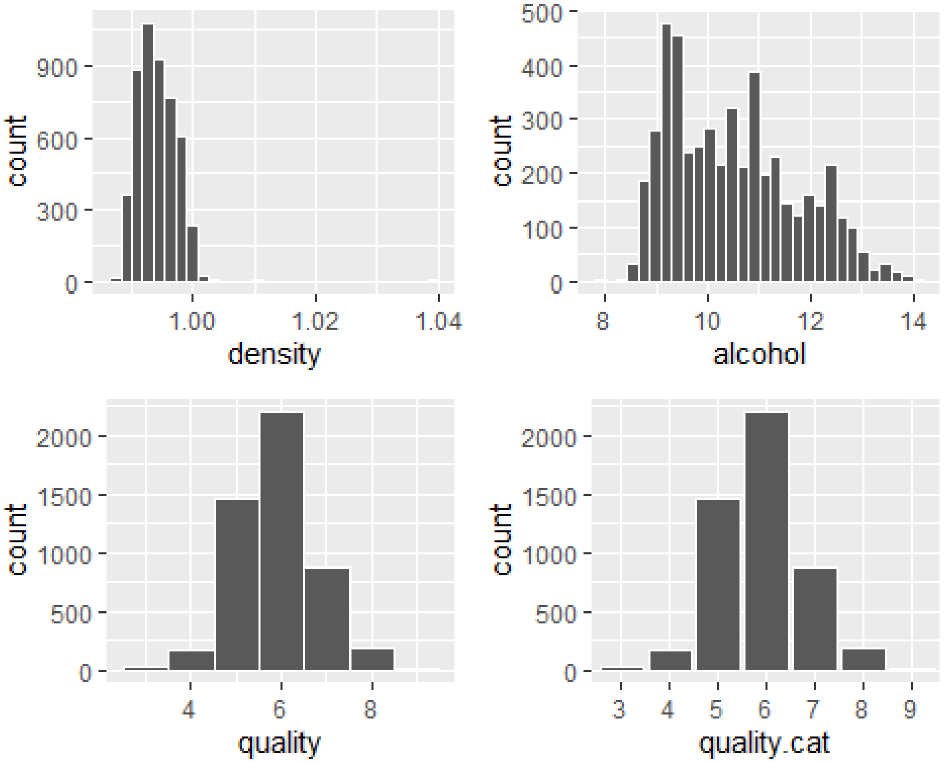


#Excluding the upper most quantile (e.g. 1%) of each of these parameters

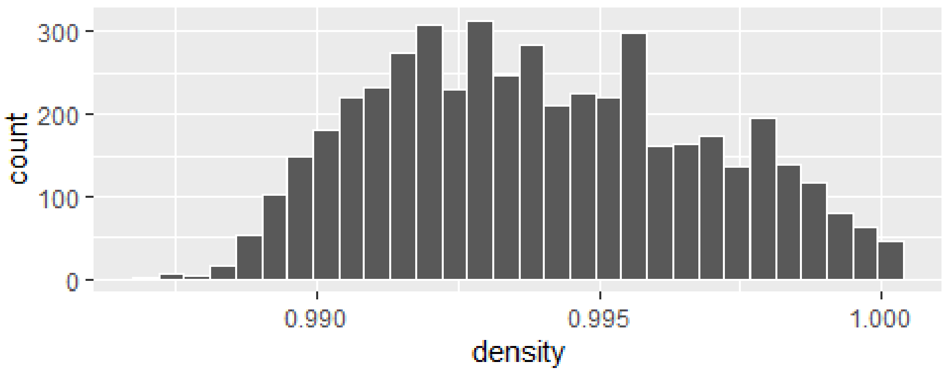


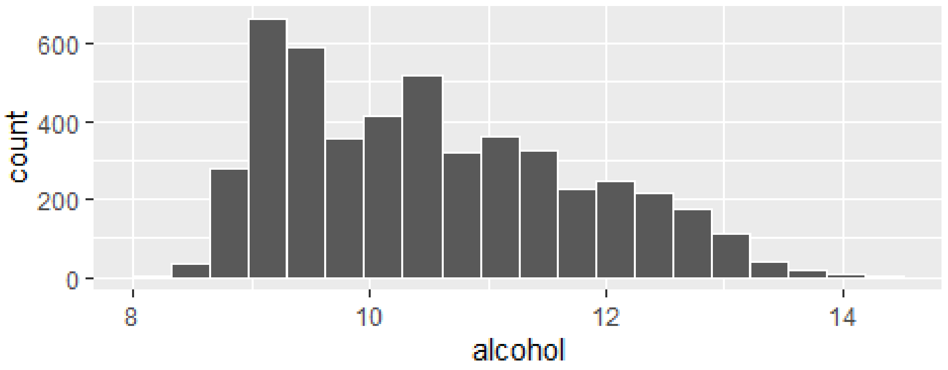
##Once the top 1%of each parameter is excluded, it is easier to see the shape of the bulk of the data. Most parameters appear to be approximately normally distributed with the exception of Residual Sugar

```
##(Note: a bar chart is used in the case of ‘quality.cat’, since it is
categorical):
##  Ord.factor w/ 7 levels "3"<"4"<"5"<"6"<..: 4 4 4 4 4 4 4 4 4 4 ...
```



## The quality rating appears to be normally distributed, with the bulk of assessments in the middle bins. Density appears normal too, but with some positive skew. The alcohol content looks interesting.





##Density looks normally distributed, whereas alcohol content does not.

#Create New Variables:

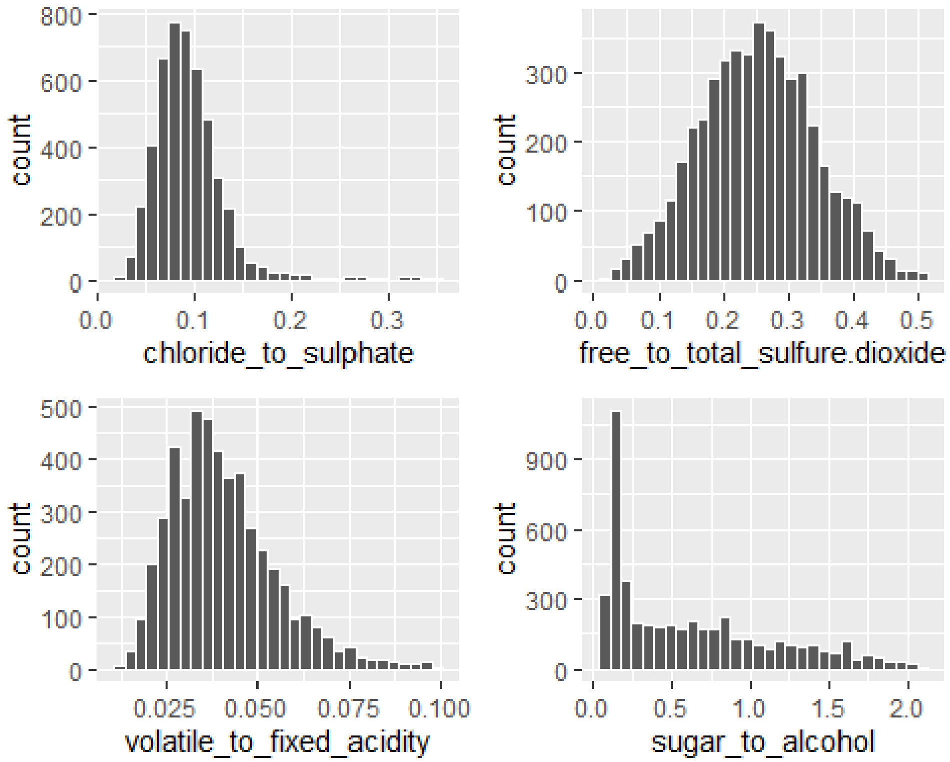
Reference: http://beerandwinejournal.com/chloride-and-sulfate/

The chlorides to sulphates ratio might be a far more important measure of quality than the individual levels of either ion. Thus, we will create a chlorides-to-sulphate ratio variable.

```
# Create and add four new variables to the dataframe:
data$chloride_to_sulphate <-with(data,chlorides / sulphates)
data$free_to_total_sulfure.dioxide <-with(data,free.sulfur.dioxide /
total.sulfur.dioxide)
data$volatile_to_fixed_acidity <-with(data,volatile.acidity / fixed.acidity)
data$sugar_to_alcohol <-with(data,residual.sugar / alcohol)
# Output summary data on the new variables:
str(subset(data,select =
c(chloride_to_sulphate,free_to_total_sulfure.dioxide,volatile_to_fixed_acidity,suga
r_to_alcohol)))

## 'data.frame':    4898 obs. of  4 variables:
##  $ chloride_to_sulphate         : num  0.1 0.1 0.114 0.145 0.145 ...
##  $ free_to_total_sulfure.dioxide: num  0.265 0.106 0.309 0.253 0.253 ...
##  $ volatile_to_fixed_acidity    : num  0.0386 0.0476 0.0346 0.0319 0.0319 ...
##  $ sugar_to_alcohol             : num  2.352 0.168 0.683 0.859 0.859 ...

summary(subset(data,select =
c(chloride_to_sulphate,free_to_total_sulfure.dioxide,volatile_to_fixed_acidity,suga
r_to_alcohol)))
```

```
##  chloride_to_sulphate free_to_total_sulfure.dioxide volatile_to_fixed_acidity
##  Min.   :0.02121      Min.   :0.02362               Min.   :0.01111          
##  1st Qu.:0.07143      1st Qu.:0.19093               1st Qu.:0.03030          
##  Median :0.08980      Median :0.25368               Median :0.03836          
##  Mean   :0.09774      Mean   :0.25558               Mean   :0.04126          
##  3rd Qu.:0.11053      3rd Qu.:0.31579               3rd Qu.:0.04848          
##  Max.   :0.62708      Max.   :0.71053               Max.   :0.18033          
##  sugar_to_alcohol
##  Min.   :0.0566  
##  1st Qu.:0.1575  
##  Median :0.4906  
##  Mean   :0.6423  
##  3rd Qu.:0.9773  
##  Max.   :5.6239
```

#Plot the new parameters as a group:



## The free:total sulfur dioxide graph looks normally distributed. The chloride:sulphate, volatile:fixed acidity and sugar:alcohol graphs look positively skewed.

Bivariate Plots and Analysis Section

#Linear Model Red Wine

```
## 
## Call:
## lm(formula = quality ~ ., data = subset(data, select = -c(quality.cat, 
##     chloride_to_sulphate, free_to_total_sulfure.dioxide, 
volatile_to_fixed_acidity,
##     sugar_to_alcohol)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8348 -0.4934 -0.0379  0.4637  3.1143
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.502e+02  1.880e+01   7.987 1.71e-15 ***
## fixed.acidity         6.552e-02  2.087e-02   3.139  0.00171 **
## volatile.acidity     -1.863e+00  1.138e-01 -16.373  < 2e-16 ***
## citric.acid           2.209e-02  9.577e-02   0.231  0.81759
## residual.sugar        8.148e-02  7.527e-03  10.825  < 2e-16 ***
## chlorides            -2.473e-01  5.465e-01  -0.452  0.65097
## free.sulfur.dioxide   3.733e-03  8.441e-04   4.422 9.99e-06 ***
## total.sulfur.dioxide -2.857e-04  3.781e-04  -0.756  0.44979
## density              -1.503e+02  1.907e+01  -7.879 4.04e-15 ***
## pH                    6.863e-01  1.054e-01   6.513 8.10e-11 ***
## sulphates             6.315e-01  1.004e-01   6.291 3.44e-10 ***
## alcohol               1.935e-01  2.422e-02   7.988 1.70e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7514 on 4886 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2803
## F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

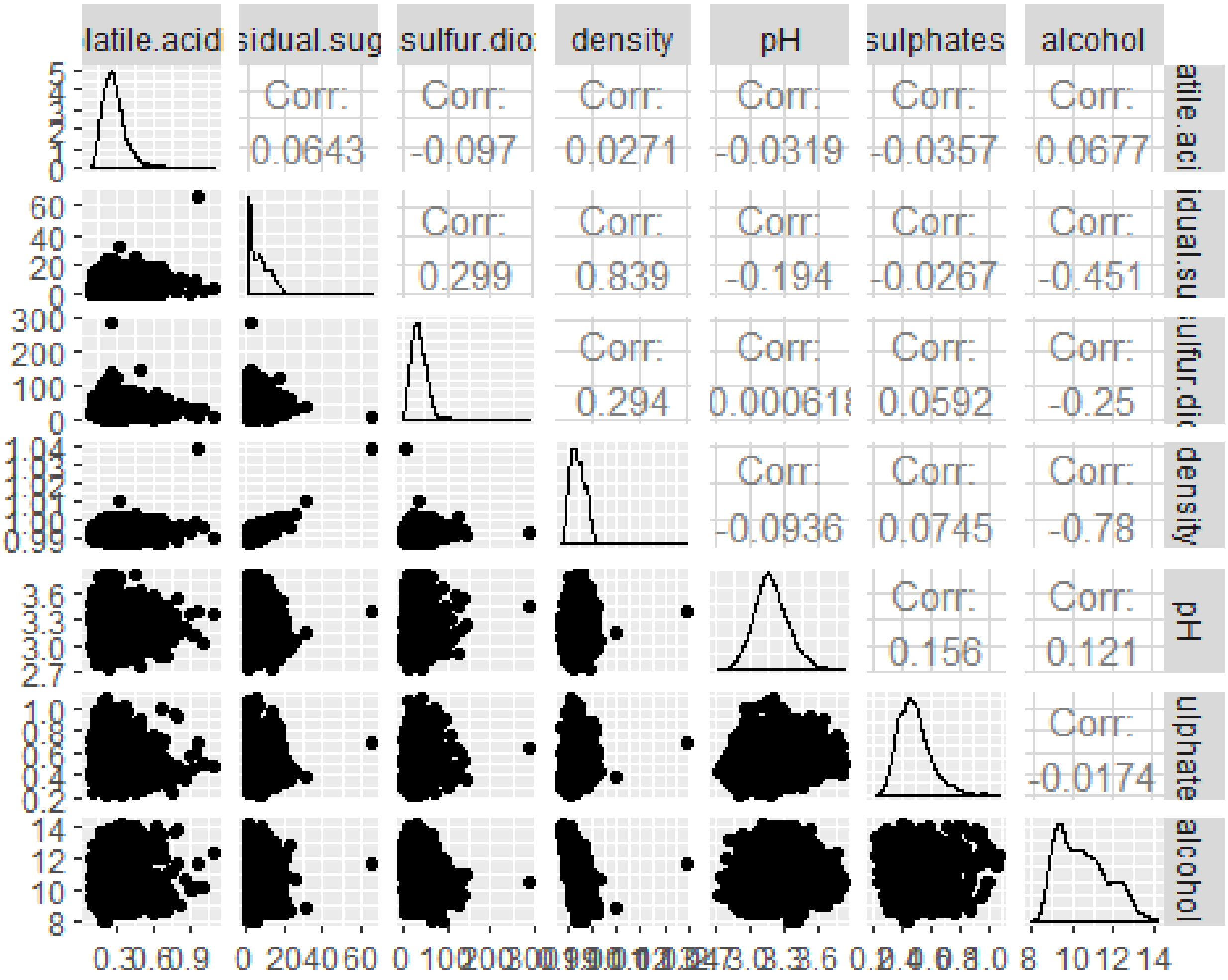
##An initial look at the Linear Regression Model shows the majority of variables are statistically significant (p-value< 0.05). Running the regression model fora subset of data based on statistical significance

```
linRegressionWine3<-lm(formula = quality ~ ., data = subset(data, select = -
c(quality.cat, chloride_to_sulphate, free_to_total_sulfure.dioxide,
volatile_to_fixed_acidity, sugar_to_alcohol, fixed.acidity,citric.acid,
chlorides,total.sulfur.dioxide)))
summary(linRegressionWine3)
```

```
##
## Call:
## lm(formula = quality ~ ., data = subset(data, select = -c(quality.cat,
##     chloride_to_sulphate, free_to_total_sulfure.dioxide,
volatile_to_fixed_acidity, sugar_to_alcohol, fixed.acidity, citric.acid, chlorides,
##     total.sulfur.dioxide)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8107 -0.4999 -0.0375  0.4636  3.2180
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.112e+02  1.273e+01   8.734  < 2e-16 ***
## volatile.acidity    -1.940e+00  1.085e-01 -17.872  < 2e-16 ***
## residual.sugar       6.637e-02  5.358e-03  12.386  < 2e-16 ***
## free.sulfur.dioxide  3.283e-03  6.770e-04   4.849 1.28e-06 ***
## density             -1.103e+02  1.274e+01  -8.653  < 2e-16 ***
## pH                   4.619e-01  7.638e-02   6.046 1.59e-09 ***
## sulphates            5.708e-01  9.856e-02   5.791 7.42e-09 ***
## alcohol              2.438e-01  1.870e-02  13.035  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.752 on 4890 degrees of freedom
## Multiple R-squared:  0.2801, Adjusted R-squared:  0.2791
## F-statistic: 271.8 on 7 and 4890 DF,  p-value: < 2.2e-16
```

#Determination Coefficient: 27.91%of Quality can be explained by these attributes

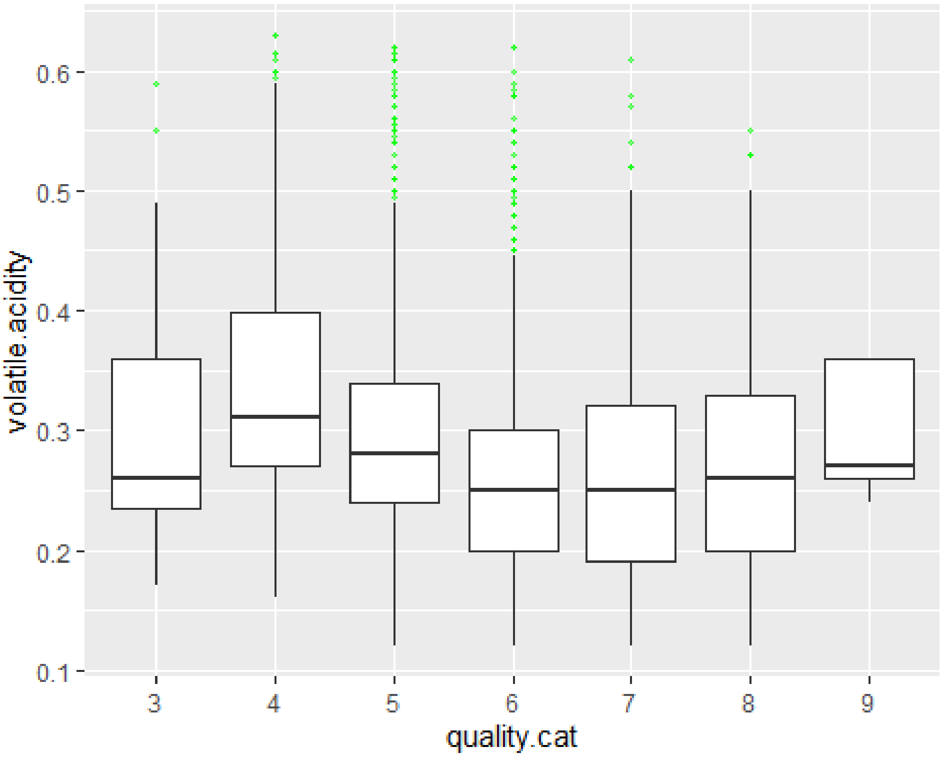
# Scatterplot matrix



#Expanding on the highest correlation coefficients, as this graph is too dense to draw conclusions

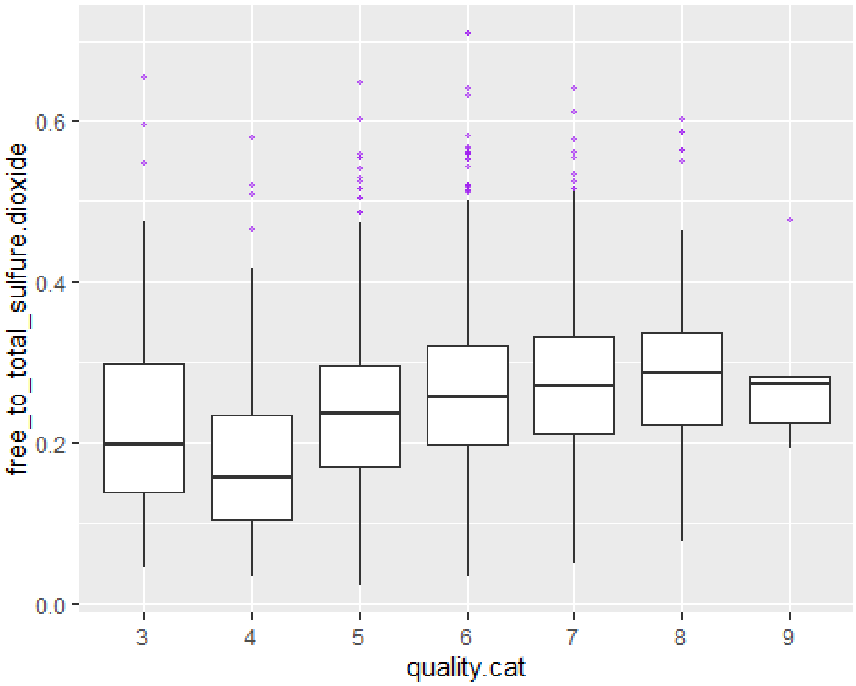
# Bivariate Plots and Analysis

##. Quality and Volatile Acidity

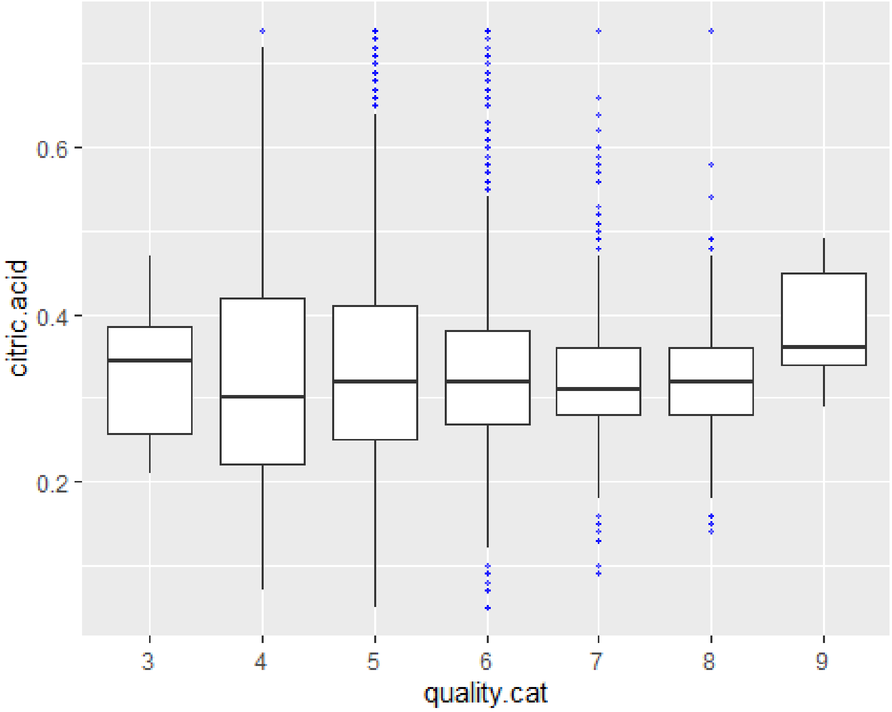


## Volatile Acidity doesn't seem to change much from low to higher quality wines

#Quality and the Free:Total Sulfur Dioxide Ratio

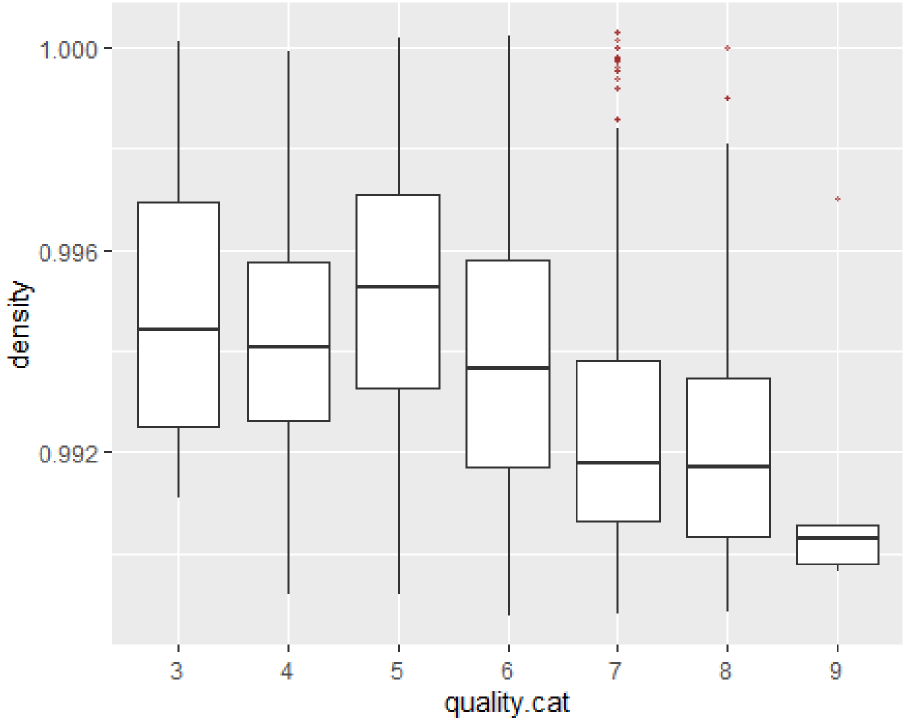


## Quality and citric.acid



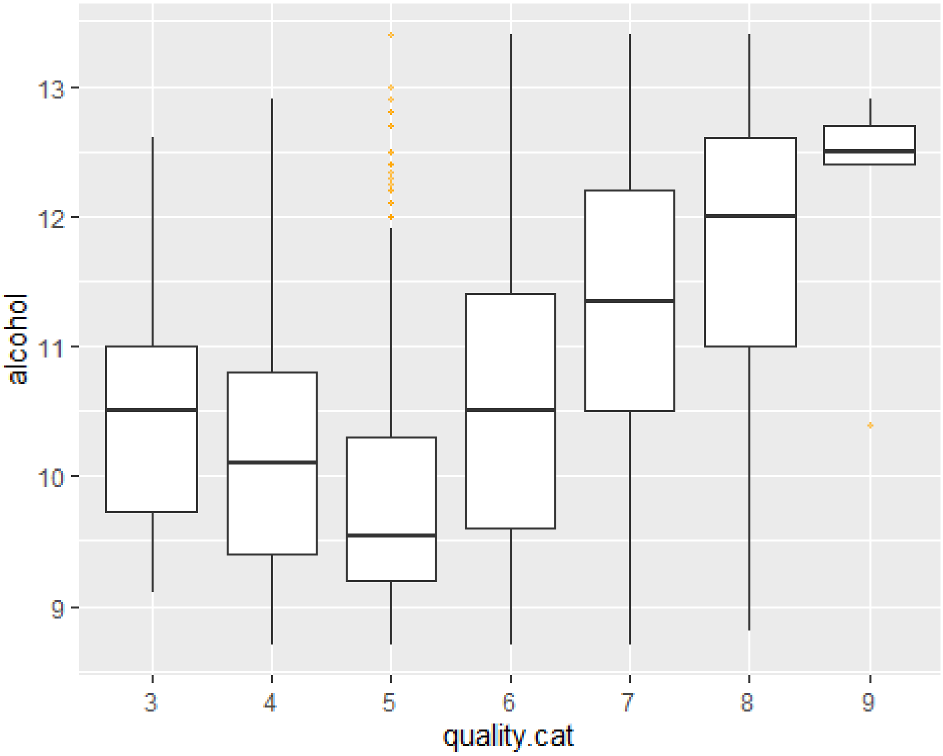
##It appears that in general, higher quality wines have slightly higher citric acid levels.

## Quality and Density



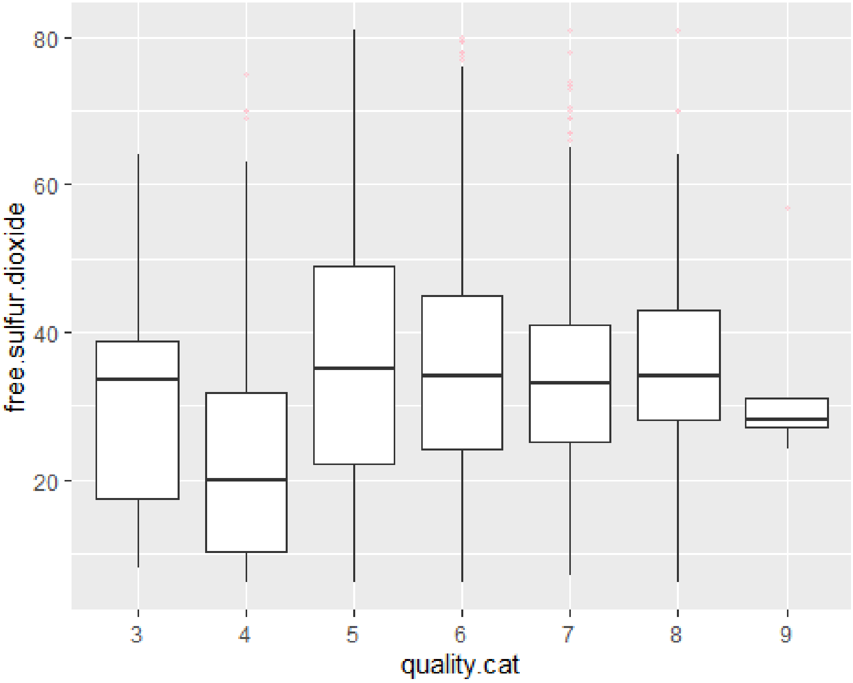
# High quality wines have lower density

##f. Quality and Alcohol Content



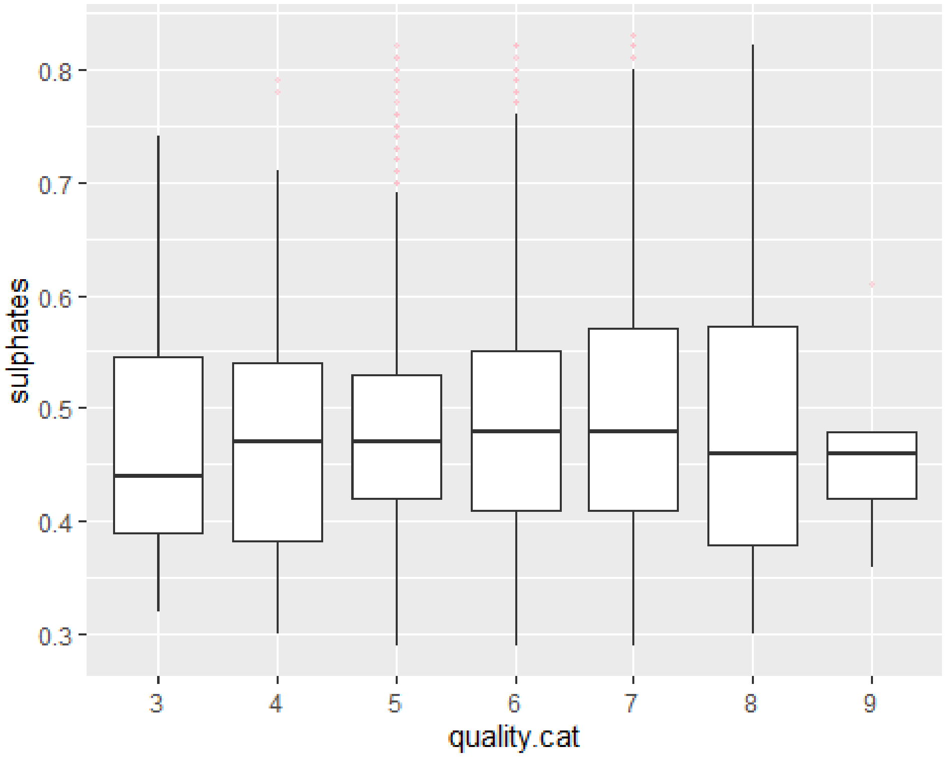
##The relationship between alcohol content and quality appears potentially promising, particularly at the higher end of the quality scale, where there is a clear upwards trend in quality (from levels 6 through 8).

##g. Quality and free Sulphur Dioxide



##It is hard to discern any clear trend between the total sulfur dioxide and a wine's quality.

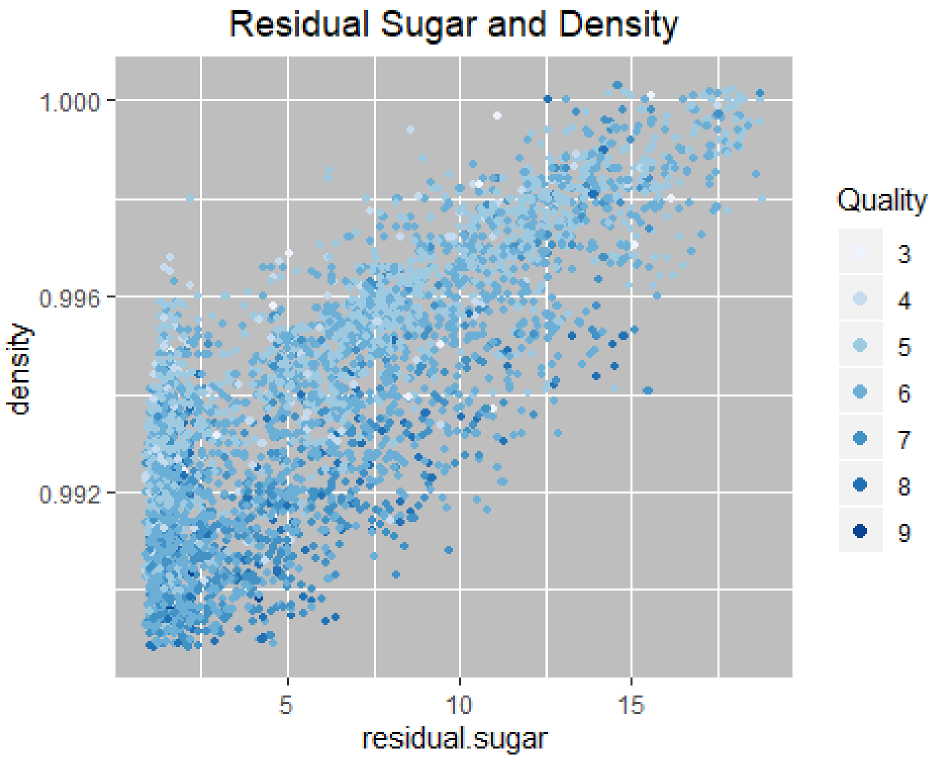
##. Quality and Sulphates

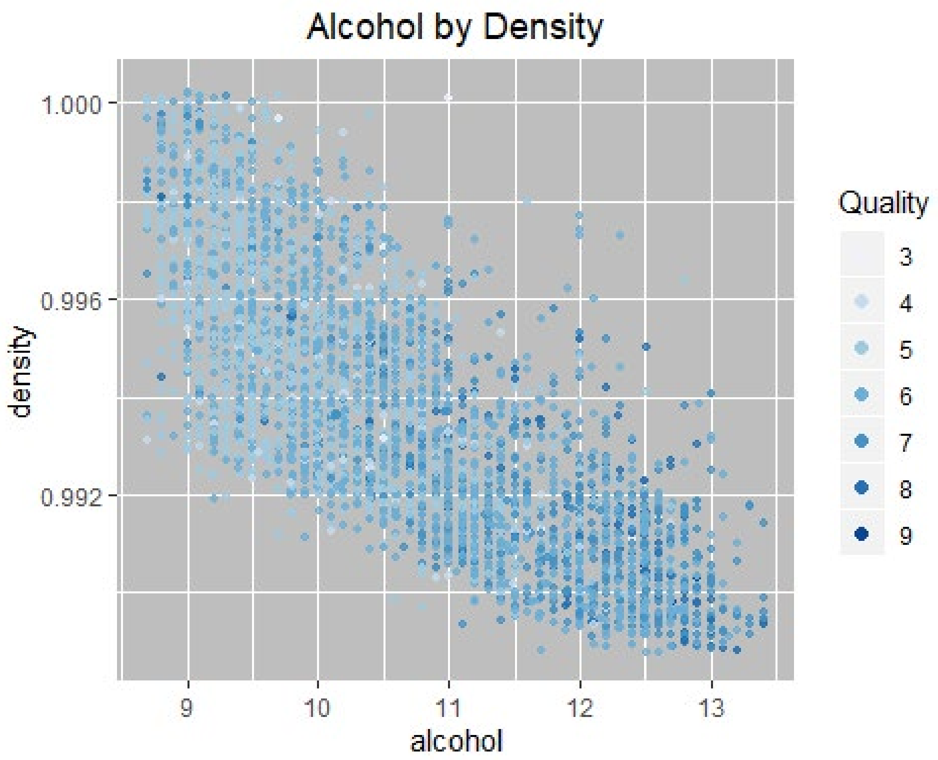


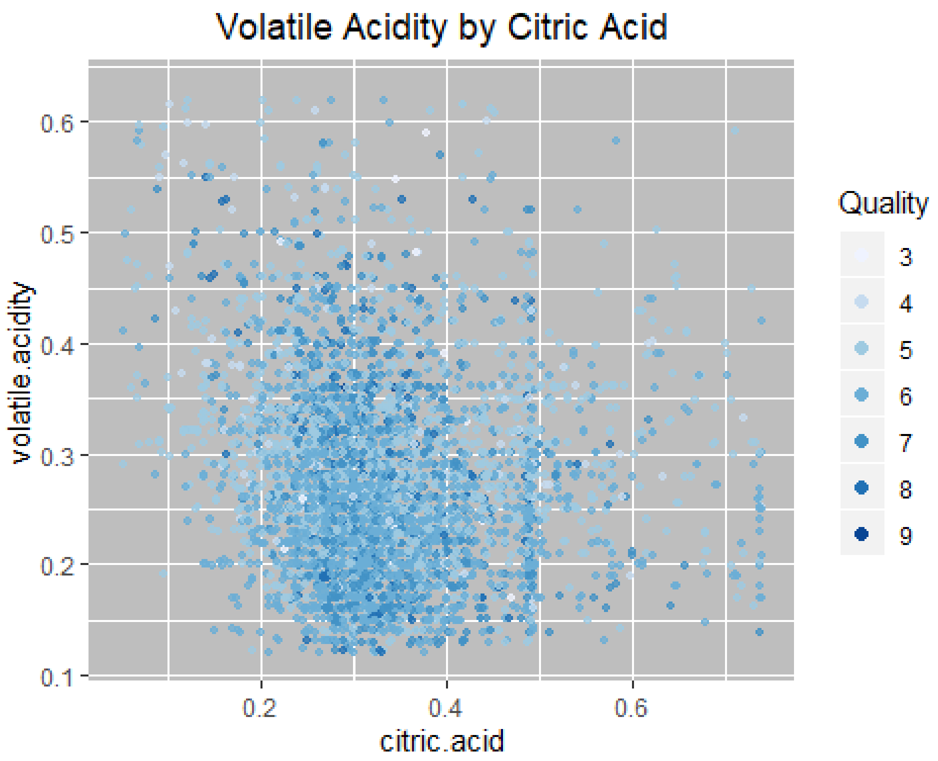
##t is hard to discern any clear trend between sulphate levels and quality.

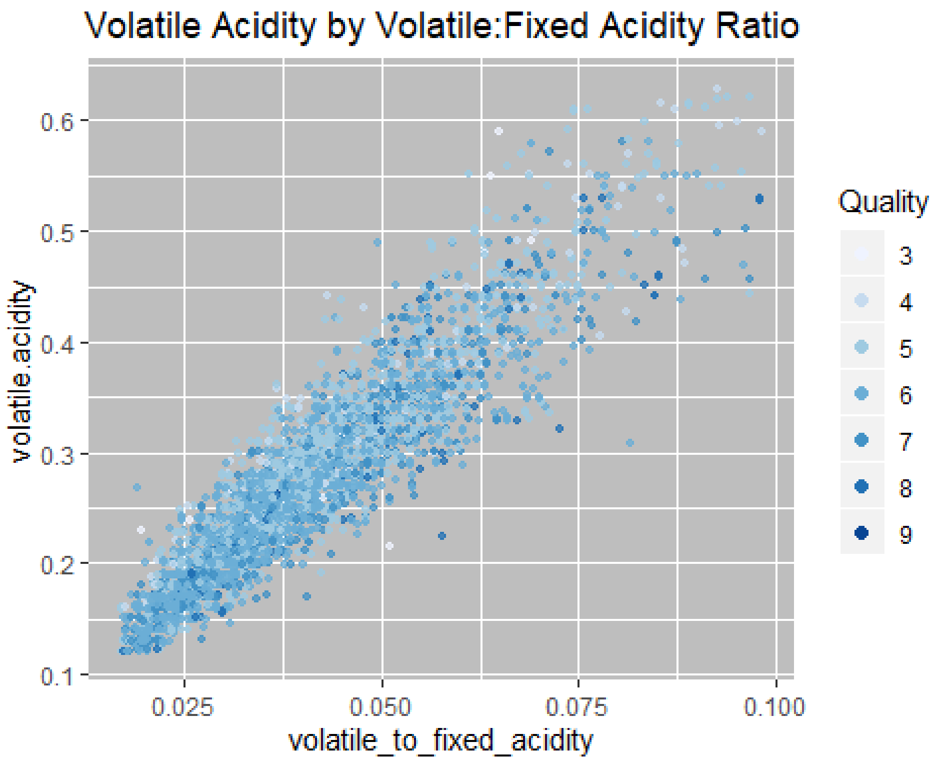
# Multivariate Plots and Analysis Section

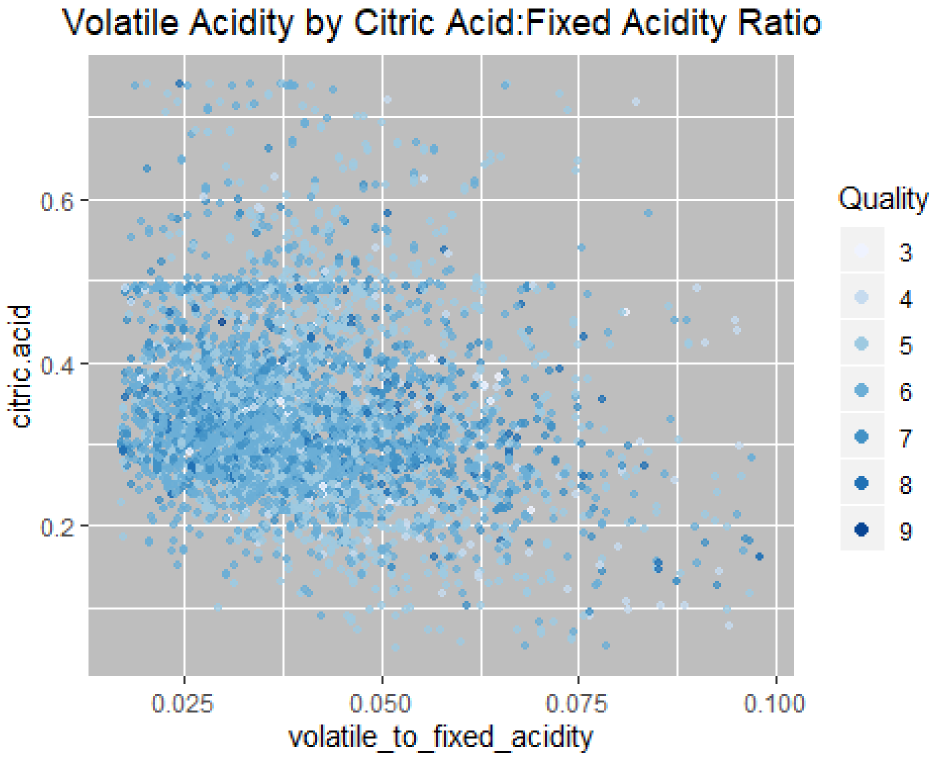
#We will now consider the interaction of multiple variables.




Alcohol by Density
density
1.000
0.996
0.992
alcohol
9
10
11
12
13
Quality
3
4
5
6
7
8
9


Volatile Acidity by Citric Acid
volatile.acidity
0.6
0.5
0.4
0.3
0.2
0.1
citric.acid
0.2
0.4
0.6
Quality
3
4
5
6
7
8
9

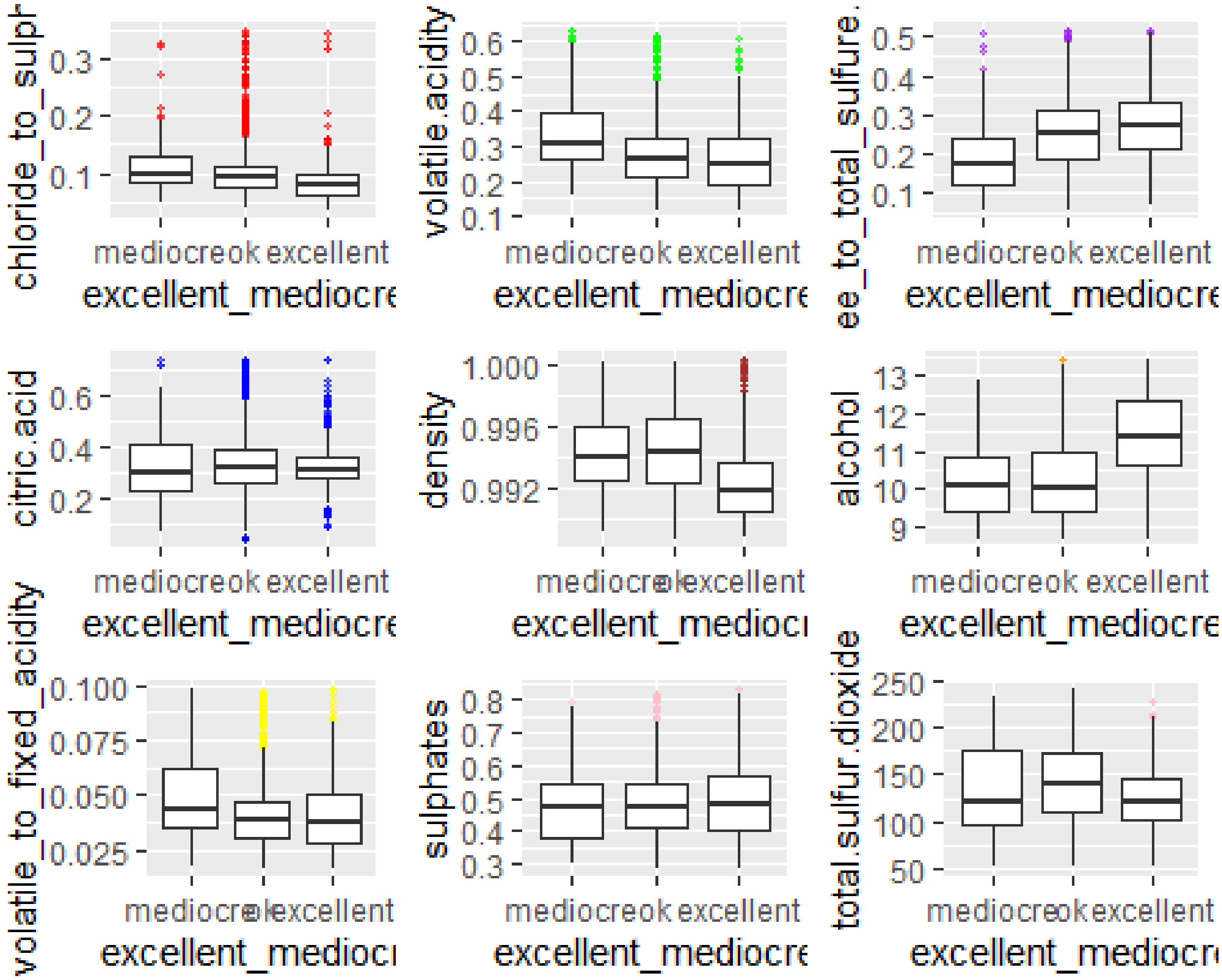


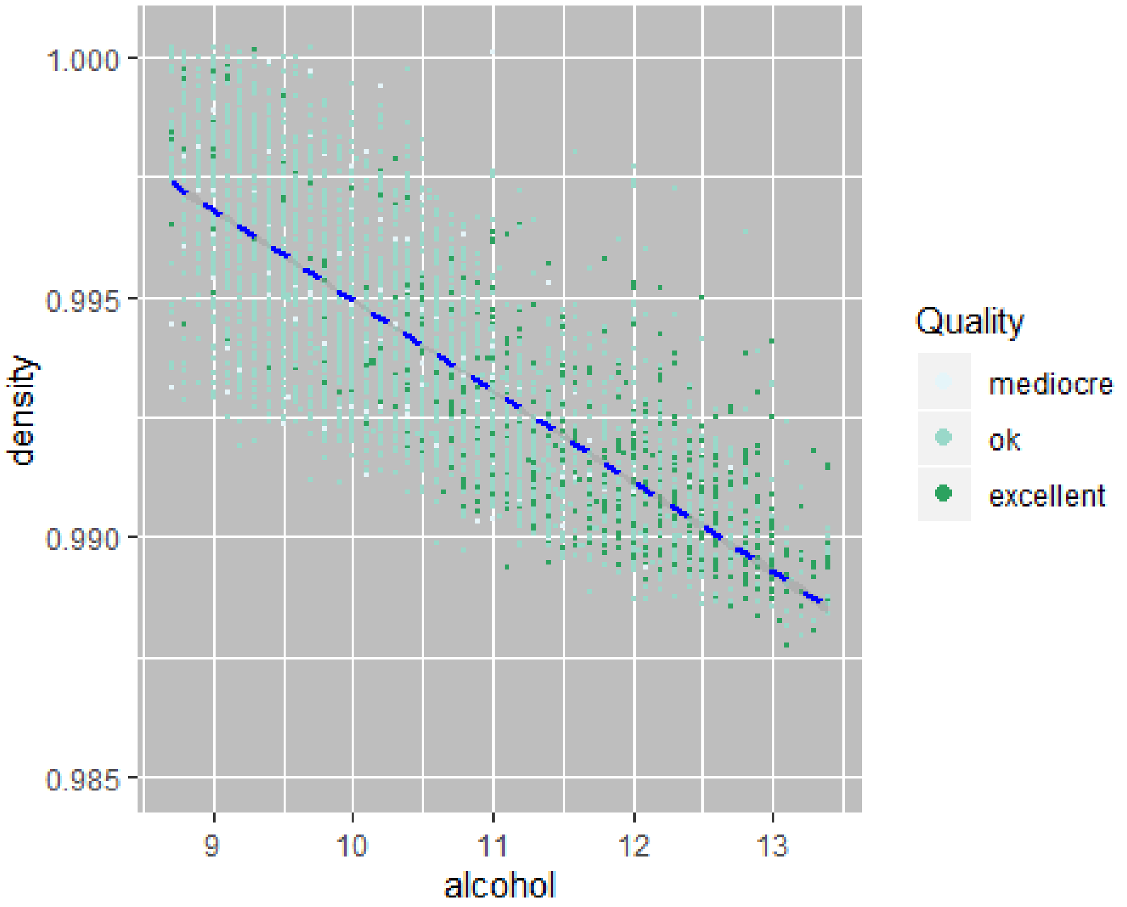


# One Final Data Transformation

##Lets consider any wine with a 3-4 rating as ‘mediocre’, a wine with a 5-6 rating as ‘ok’ and a wine with a 7-8 rating as ‘excellent’.

```
##   mediocre        ok  excellent
##        183      3655       1060
```





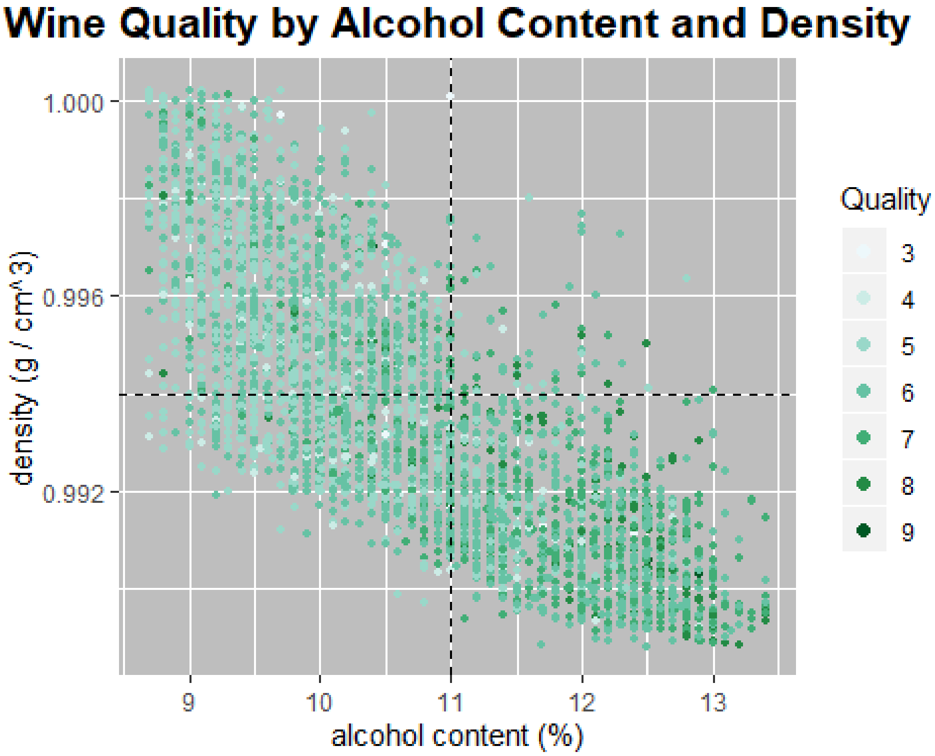
#The categories split quite well: good wines tend to have higher alcohol content and lower density levels.

#Predictive Models

```
## Call:
## polr(formula = quality.cat ~ alcohol + density + sulphates + 
##     citric.acid + volatile.acidity + total.sulfur.dioxide + chloride_to_sulphate 
+ 
##     volatile_to_fixed_acidity + free_to_total_sulfure.dioxide, 
##     data = data, Hess = TRUE)
## 
## Coefficients:
##                                    Value Std. Error  t value
## alcohol                         1.027663  0.0286945  35.8139
## density                       106.662507  0.2306949 462.3531
## sulphates                       0.790275  0.2629990   3.0049
## citric.acid                    -0.039837  0.2308559  -0.1726
## volatile.acidity               -9.066165  0.2954961 -30.6812
## total.sulfur.dioxide            0.001183  0.0007562   1.5646
## chloride_to_sulphate           -1.250571  0.6100165  -2.0501
## volatile_to_fixed_acidity      27.933222  0.0440556 634.0446
## free_to_total_sulfure.dioxide   3.070066  0.3076082   9.9804
## 
## Intercepts:
```

```
##     Value    Std. Error t value
## 3|4 110.4168   0.2278   484.6871
## 4|5 112.7554   0.2302   489.8279
## 5|6 115.7658   0.2362   490.1709
## 6|7 118.3238   0.2511   471.1495
## 7|8 120.5611   0.2682   449.4542
## 8|9 124.2369   0.5172   240.1983
##
## Residual Deviance: 10983.75
## AIC: 11013.75

## [1] "Confidence Levels:"

##                                        2.5 %        97.5 %
## alcohol                         9.714230e-01   1.083903486
## density                         1.062104e+02 107.114660813
## sulphates                       2.748068e-01   1.305743728
## citric.acid                    -4.923067e-01   0.412631972
## volatile.acidity               -9.645326e+00  -8.487002819
## total.sulfur.dioxide           -2.989697e-04   0.002665218
## chloride_to_sulphate           -2.446182e+00  -0.054961091
## volatile_to_fixed_acidity       2.784687e+01  28.019568926
## free_to_total_sulfure.dioxide  2.467165e+00   3.672966567

##
## Re-fitting to get Hessian

## Call:
## polr(formula = data$excellent_mediocre ~ alcohol + density +
##     sulphates + citric.acid + volatile.acidity + total.sulfur.dioxide +
##     chloride_to_sulphate + volatile_to_fixed_acidity +
free_to_total_sulfure.dioxide,
##     data = data)
##
## Coefficients:
##                                    Value Std. Error  t value
## alcohol                         0.913252  0.0354709  25.7465
## density                        74.432466  0.2555931 291.2147
## sulphates                       0.603296  0.3233984   1.8655
## citric.acid                    -0.238817  0.3067408  -0.7786
## volatile.acidity               -7.388277  0.3688747 -20.0292
## total.sulfur.dioxide            0.002828  0.0009876   2.8632
## chloride_to_sulphate           -2.612945  0.8328124  -3.1375
## volatile_to_fixed_acidity      20.829167  0.0547422 380.4956
## free_to_total_sulfure.dioxide  3.002316  0.3812168   7.8756
##
## Intercepts:
##              Value    Std. Error t value
## mediocre|ok   79.7021   0.2527   315.4449
## ok|excellent  85.1066   0.2803   303.6106
##
## Residual Deviance: 5624.808
## AIC: 5646.808
## [1] "Confidence Levels:"
## Re-fitting to get Hessian
##                                      2.5 %       97.5 %
## alcohol                        0.843730023  0.982773558
## density                       73.931512330 74.933418983
## sulphates                     -0.030553379  1.237145141
## citric.acid                   -0.840018272  0.362383505
## volatile.acidity              -8.111257783 -6.665295408
## total.sulfur.dioxide           0.000891956  0.004763108
## chloride_to_sulphate          -4.245227719 -0.980663068
## volatile_to_fixed_acidity     20.721874177 20.936459669
## free_to_total_sulfure.dioxide  2.255144810  3.749487148
```

#Final Plot and Summary



#This plot demonstrates that in general, the high-quality wines (quality 7-8) tend to have high alcohol content and low density, as shown by the preponderance of green shaded points in the lower right quadrant of the graph. Conversely, the poor-quality wines (quality 3-4) tend to have low alcohol content and high density, dominating the two left side quadrants.

Random Forest

```
#Resetting the data set
readURL <- function(inputURL)  #Begin function named readURL that takes a URL
{
csvFile <- read.csv(url(inputURL), sep = ';')  #assign the results of the URL call
as a csv file to a dataframe named csvFile. Added sep = ';' to seperate the data
into columns
  return(csvFile)  # return the dataframe
}

redWine <- readURL("https://archive.ics.uci.edu/ml/machine-learning-databases/wine-
quality/winequality-red.csv")

whiteWine <- readURL("https://archive.ics.uci.edu/ml/machine-learning-
databases/wine-quality/winequality-white.csv")

# Verify no NAs

redWine <- na.omit(redWine)

whiteWine<-na.omit (whiteWine)

#table preview

#Red Wine Random Forest

library(randomForest)  #call random forest library

## randomForest 4.6-14

set.seed(100)  #set seed value

red_rftrain <- na.omit(sample(nrow(redWine), 0.7*nrow(redWine), replace=FALSE))
#create a sample of values for training

rwTrainSet <- na.omit(redWine[red_rftrain,])  #red wine training data
rwTestSet <- na.omit(redWine[-red_rftrain,])  #red wine testing data
rwTrainSet <- rwTrainSet[,-13]  #remove the last column (text of quality score)
rwTestSet <- rwTestSet[,-13] #remove the last column (text of quality score)

str(rwTrainSet)  #show training set details

## 'data.frame':    1119 obs. of  12 variables:
##  $ fixed.acidity       : num  10.4 6.8 12.2 6.8 8.4 8.7 7.2 6.7 6.7 6.6 ...
##  $ volatile.acidity    : num  0.44 0.83 0.45 0.36 0.36 0.82 0.39 0.54 0.28 0.5 ...
##  $ citric.acid         : num  0.73 0.09 0.49 0.32 0.32 0.02 0.32 0.13 0.28 0 ...
##  $ residual.sugar      : num  6.55 1.8 1.4 1.8 2.2 1.2 1.8 2 2.4 1.8 ...
##  $ chlorides           : num  0.074 0.074 0.075 0.067 0.081 0.07 0.065 0.076 0.012 0.062
...
```

```
##  $ free.sulfur.dioxide : num  38 4 3 4 32 36 34 15 36 21 ...
##  $ total.sulfur.dioxide: num  76 25 6 8 79 48 60 36 100 28 ...
##  $ density             : num  0.999 0.995 0.997 0.993 0.996 ...
##  $ pH                  : num  3.17 3.38 3.13 3.36 3.3 3.2 3.46 3.61 3.26 3.44 ...
##  $ sulphates           : num  0.85 0.45 0.63 0.55 0.72 0.58 0.78 0.64 0.39 0.55 ...
##  $ alcohol             : num  12 9.6 10.4 12.8 11 9.8 9.9 9.8 11.7 12.3 ...
##  $ quality             : int  7 5 5 7 6 5 5 5 7 6 ...
```

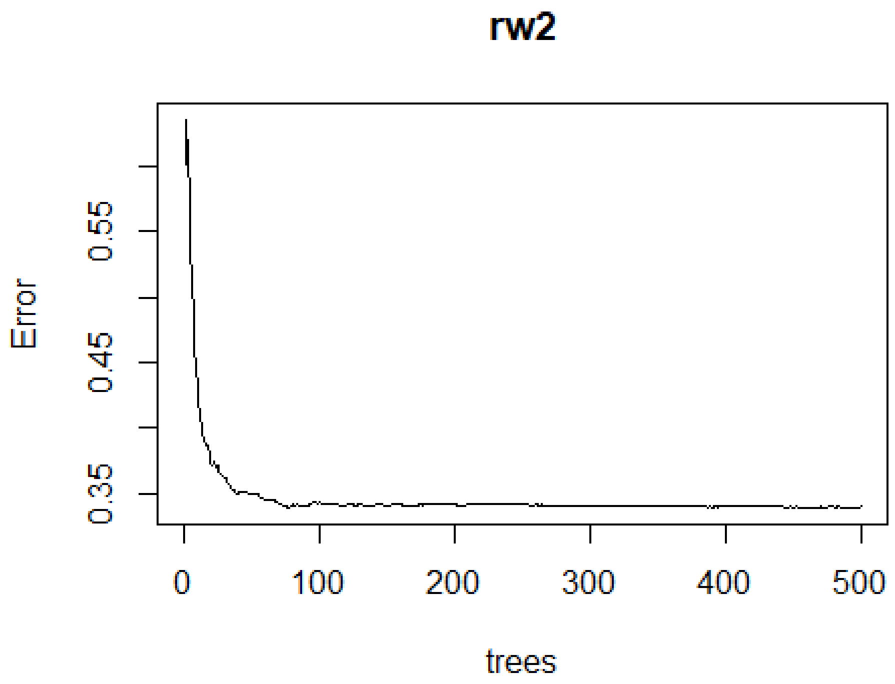
```
str(rwTestSet) #show testing set details
```

```
## 'data.frame':    480 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.9 7.3 7.8 7.5 7.8 8.9 7.6 6.9 8.3 ...
##  $ volatile.acidity    : num  0.7 0.6 0.65 0.58 0.5 0.61 0.62 0.39 0.4 0.655 ...
##  $ citric.acid         : num  0 0.06 0 0.02 0.36 0.29 0.18 0.31 0.14 0.12 ...
##  $ residual.sugar      : num  1.9 1.6 1.2 2 6.1 1.6 3.8 2.3 2.4 2.3 ...
##  $ chlorides           : num  0.076 0.069 0.065 0.073 0.071 0.114 0.176 0.082 0.085
0.083 ...
##  $ free.sulfur.dioxide : num  11 15 15 9 17 9 52 23 21 15 ...
##  $ total.sulfur.dioxide: num  34 59 21 18 102 29 145 71 40 113 ...
##  $ density             : num  0.998 0.996 0.995 0.997 0.998 ...
##  $ pH                  : num  3.51 3.3 3.39 3.36 3.35 3.26 3.16 3.52 3.43 3.17 ...
##  $ sulphates           : num  0.56 0.46 0.47 0.57 0.8 1.56 0.88 0.65 0.63 0.66 ...
##  $ alcohol             : num  9.4 9.4 10 9.5 10.5 9.1 9.2 9.7 9.7 9.8 ...
##  $ quality             : int  5 5 7 7 5 5 5 5 6 5 ...
```

```
rw2 <-  randomForest(quality ~
volatile.acidity+chlorides+total.sulfur.dioxide+pH+sulphates+alcohol,data =
rwTrainSet, ntree = 500, mtry = 6, importance = TRUE)

#create a random forest for red wine quality based on volatile acidity, chlorides,
total.sulfur.dioxide, pH, sulphates, alcohol from the training data

rw2  #display results of the random forest
```

```
##
## Call:
##  randomForest(formula = quality ~ volatile.acidity + chlorides +
total.sulfur.dioxide + pH + sulphates + alcohol, data = rwTrainSet,      ntree =
500, mtry = 6, importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 6
##
##           Mean of squared residuals: 0.3395346
##                     % Var explained: 45.98
```

```
plot(rw2)  #plot results
```

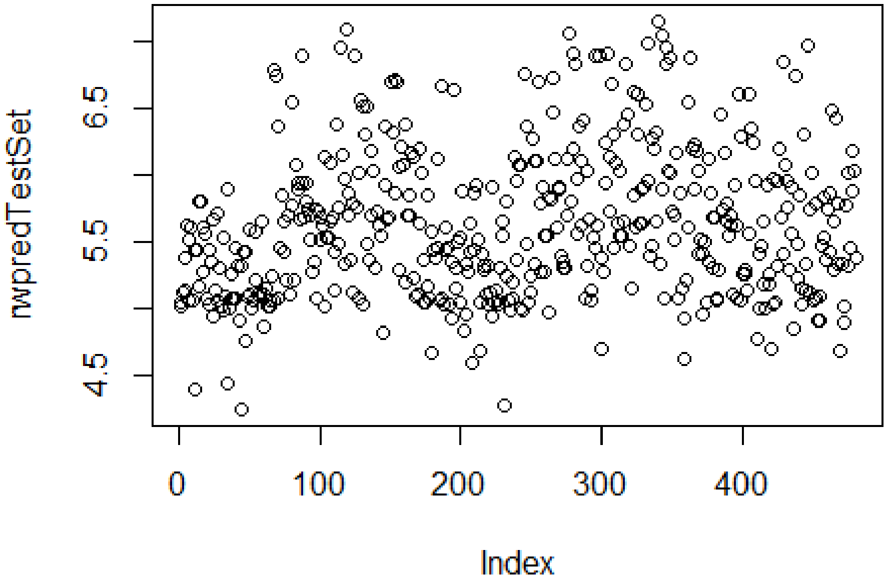


```
importance(rw2)  #display importance statistics

##                       %IncMSE IncNodePurity
## volatile.acidity     36.57271     108.42529
## chlorides            32.10525      75.04634
## total.sulfur.dioxide 35.96481      83.63795
## pH                   23.76536      64.95998
## sulphates            52.47342     107.79968
## alcohol              79.27348     233.20962

rwpredTrainSet <- predict(rw2, rwTrainSet, type = "class")  #predict the red wine
quality score against the training data

head(table(format(round(rwpredTrainSet,4),nsmall=4), rwTrainSet$quality), n=25)
#show the top 25 results of the confusion matrix.  Rounded the training set data
prediction to 4 decimal places

##
##           3 4 5 6 7 8
##   3.4406 1 0 0 0 0 0
##   3.5946 1 0 0 0 0 0
##   3.6080 1 0 0 0 0 0
##   3.8441 1 0 0 0 0 0
##   3.9100 1 0 0 0 0 0
##   3.9438 1 0 0 0 0 0
##   3.9444 1 0 0 0 0 0
##   4.0353 1 0 0 0 0 0
##   4.0514 0 1 0 0 0 0
##   4.1179 1 0 0 0 0 0
##   4.1914 0 1 0 0 0 0
##   4.2481 0 1 0 0 0 0
##   4.2522 0 1 0 0 0 0
##   4.2558 1 0 0 0 0 0
##   4.3237 0 1 0 0 0 0
##   4.3751 0 0 1 0 0 0
##   4.3829 0 1 0 0 0 0
##   4.3961 0 1 0 0 0 0
##   4.4075 0 1 0 0 0 0
##   4.4262 0 1 0 0 0 0
##   4.4848 0 1 0 0 0 0
##   4.4884 0 1 0 0 0 0
##   4.5011 0 1 0 0 0 0
##   4.5100 0 1 0 0 0 0
##   4.5162 0 1 0 0 0 0
```

```
tail(table(format(round(rwpredTrainSet,4),nsmall=4), rwTrainSet$quality), n=25) #show the last 25 results of the confusion matrix.  Rounded the training set data prediction to 4 decimal places
```

```
## 
##          3 4 5 6 7 8
##   6.9361 0 0 0 0 1 0
##   6.9435 0 0 0 0 1 0
##   6.9441 0 0 0 0 1 0
##   6.9509 0 0 0 0 1 0
##   6.9571 0 0 0 0 1 0
##   6.9678 0 0 0 0 1 0
##   6.9699 0 0 0 0 1 0
##   6.9792 0 0 0 0 1 0
##   6.9817 0 0 0 0 1 0
##   6.9945 0 0 0 0 1 0
##   6.9948 0 0 0 0 1 0
##   7.0067 0 0 0 0 2 0
##   7.0157 0 0 0 0 1 0
##   7.0194 0 0 0 0 1 0
##   7.0310 0 0 0 0 1 0
##   7.2649 0 0 0 0 0 1
##   7.3963 0 0 0 0 0 1
##   7.4092 0 0 0 0 0 1
##   7.4221 0 0 0 0 0 1
##   7.4279 0 0 0 0 0 1
##   7.4473 0 0 0 0 0 1
##   7.4628 0 0 0 0 0 1
##   7.4744 0 0 0 0 0 1
##   7.5889 0 0 0 0 0 2
##   7.5996 0 0 0 0 0 1
```

```
rwpredTestSet <- predict(rw2, rwTestSet, type = "class")  #predict the  red wine
quality score against the test data

plot(rwpredTestSet) #plot results
```



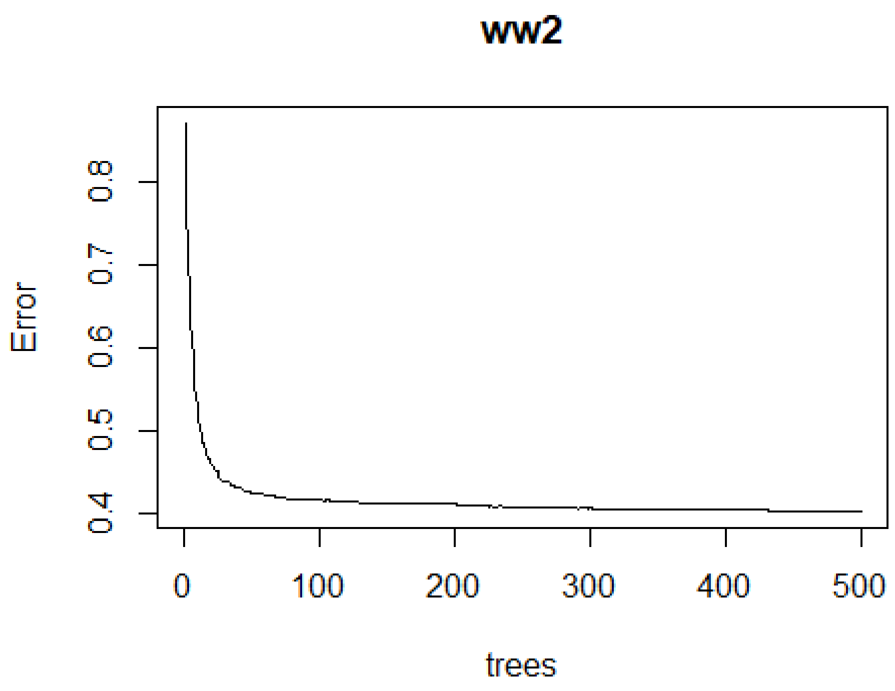
```
paste0("The mean of values where the predicted value equals actual is ",
mean((round(rwpredTestSet)) == rwTestSet$quality)) #Display the mean of matched
values

## [1] "The mean of values where the predicted value equals actual is 0.625"

head(table(format(round(rwpredTestSet,4),nsmall=4), rwTestSet$quality), n=25) #show
the top 25 results of the confusion matrix.  Rounded the test set data prediction
to 4 decimal places

## 
##           4 5 6 7 8
##   4.2505 0 1 0 0 0
##   4.2853 0 1 0 0 0
##   4.4062 1 0 0 0 0
##   4.4398 0 1 0 0 0
##   4.5976 0 1 0 0 0
##   4.6293 1 0 0 0 0
##   4.6679 0 1 0 0 0
##   4.6892 1 0 0 0 0
##   4.6945 0 1 0 0 0
##   4.6977 0 1 0 0 0
##   4.7077 0 1 0 0 0
##   4.7637 0 0 1 0 0
##   4.7760 0 1 0 0 0
##   4.8233 0 1 0 0 0
##   4.8398 0 1 0 0 0
##   4.8542 0 1 0 0 0
##   4.8713 0 1 0 0 0
##   4.9064 0 1 0 0 0
##   4.9125 0 0 2 0 0
##   4.9127 0 1 0 0 0
##   4.9355 0 1 0 0 0
##   4.9382 0 1 0 0 0
##   4.9447 0 1 0 0 0
##   4.9476 0 0 1 0 0
##   4.9607 1 0 0 0 0
```

```
tail(table(format(round(rwpredTestSet,4),nsmall=4), rwTestSet$quality), n=25) #show the last 25 results of the confusion matrix.  Rounded the test set data prediction to 4 decimal places
```

```
## 
##           4 5 6 7 8
##   6.7084 0 0 1 0 0
##   6.7219 0 0 0 1 0
##   6.7285 0 0 0 1 0
##   6.7424 0 0 0 1 0
##   6.7463 0 0 0 1 0
##   6.7619 0 0 0 1 0
##   6.7892 0 0 0 1 0
##   6.8389 0 0 0 1 0
##   6.8398 0 0 0 1 0
##   6.8399 0 0 0 1 0
##   6.8531 0 0 0 1 0
##   6.8816 0 0 1 0 0
##   6.8918 0 0 0 1 0
##   6.8929 0 0 0 2 1
##   6.8938 0 0 0 0 1
##   6.9116 0 0 0 1 0
##   6.9121 0 0 0 1 0
##   6.9594 0 0 0 1 0
##   6.9678 0 0 0 1 0
##   6.9726 0 0 0 1 0
##   6.9923 0 0 1 0 0
##   7.0467 0 0 0 1 0
##   7.0648 0 0 0 1 0
##   7.1032 0 0 1 0 0
##   7.1590 0 0 0 1 0
```

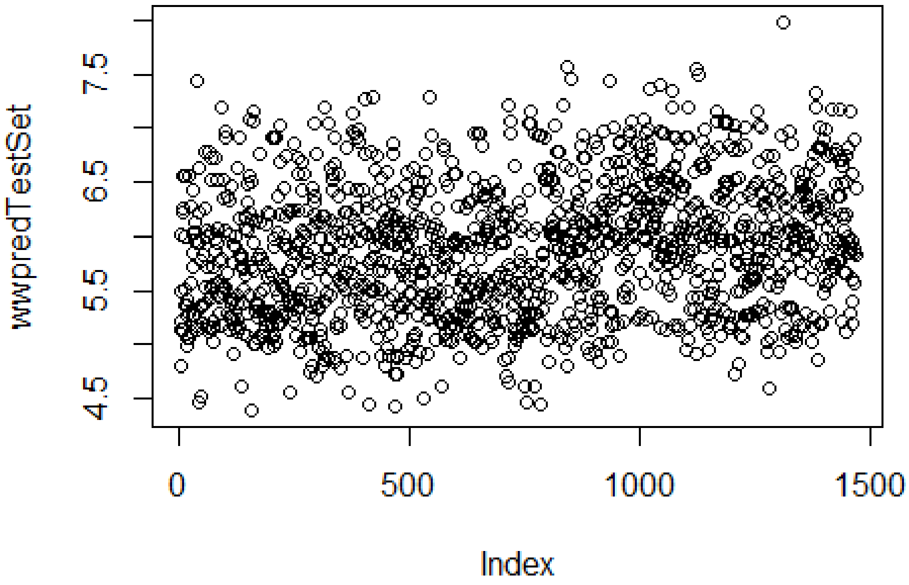
```
#White Wine Random Forest

white_rftrain <- na.omit(sample(nrow(whiteWine), 0.7*nrow(whiteWine),
replace=FALSE)) #create a sample of values for training

wwTrainSet <- na.omit(whiteWine[white_rftrain,]) #white wine training data

wwTestSet <- na.omit(whiteWine[-white_rftrain,]) #white wine testing data

wwTrainSet <- wwTrainSet[,-13]  #remove the last column (text of quality score)

wwTestSet <- wwTestSet[,-13]  #remove the last column (text of quality score)

str(wwTrainSet) #show training set details

## 'data.frame':    3428 obs. of  12 variables:
##  $ fixed.acidity       : num  7 5.8 7.4 7.2 6.9 7.1 6 6.9 6.8 6.3 ...
##  $ volatile.acidity    : num  0.15 0.12 0.18 0.23 0.24 0.49 0.28 0.22 0.22 0.27 ...
##  $ citric.acid         : num  0.38 0.21 0.3 0.25 0.37 0.22 0.35 0.28 0.32 0.37 ...
##  $ residual.sugar      : num  15.3 1.3 8.8 18.8 6.1 2 1.9 7.8 5.9 7.9 ...
##  $ chlorides           : num  0.045 0.056 0.064 0.085 0.027 0.047 0.037 0.05 0.054 0.047
...
##  $ free.sulfur.dioxide : num  54 35 26 19 38 ...
##  $ total.sulfur.dioxide: num  120 121 103 111 112 ...
##  $ density             : num  0.998 0.991 0.996 1 0.991 ...
##  $ pH                  : num  3.18 3.32 2.94 3.1 3.19 3.24 3.16 3.22 3.2 3.19 ...
##  $ sulphates           : num  0.42 0.33 0.56 0.51 0.34 0.37 0.69 0.6 0.57 0.48 ...
##  $ alcohol             : num  9.8 11.4 9.3 8.7 12.4 11 10.6 11.5 10.8 9.5 ...
##  $ quality             : int  6 6 5 5 6 3 5 8 6 6 ...

str(wwTestSet) #show testing set details

## 'data.frame':    1470 obs. of  12 variables:
##  $ fixed.acidity       : num  6.3 6.2 8.1 6.3 6.6 7 7.2 7.3 7.2 6.6 ...
##  $ volatile.acidity    : num  0.3 0.32 0.22 0.48 0.27 0.28 0.32 0.24 0.19 0.25 ...
##  $ citric.acid         : num  0.34 0.16 0.43 0.04 0.41 0.39 0.36 0.39 0.31 0.29 ...
##  $ residual.sugar      : num  1.6 7 1.5 1.1 1.3 ...
##  $ chlorides           : num  0.049 0.045 0.044 0.046 0.052 0.051 0.033 0.057 0.062
0.068 ...
##  $ free.sulfur.dioxide : num  14 30 28 30 16 32 37 45 31 39 ...
##  $ total.sulfur.dioxide: num  132 136 129 99 142 141 114 149 173 124 ...
##  $ density             : num  0.994 0.995 0.994 0.993 0.995 ...
##  $ pH                  : num  3.3 3.18 3.22 3.24 3.42 3.38 3.1 3.21 3.35 3.34 ...
##  $ sulphates           : num  0.49 0.47 0.45 0.36 0.47 0.53 0.71 0.36 0.44 0.58 ...
##  $ alcohol             : num  9.5 9.6 11 9.6 10 10.5 12.3 8.6 11.7 11 ...
##  $ quality             : int  6 6 6 6 6 6 7 5 6 7 ...

ww2 <-  randomForest(quality ~
volatile.acidity+residual.sugar+free.sulfur.dioxide+density+pH+sulphates+alcohol,da
ta = wwTrainSet, ntree = 500, mtry = 6, importance = TRUE)

#create a random forest for white wine quality based on volatile acidity, residual
sugar, free sulfur dioxide, density, pH, sulphates and alcohol from the training
```

```
data

ww2 #display results of the random forest

##
## Call:
##  randomForest(formula = quality ~ volatile.acidity + residual.sugar +
free.sulfur.dioxide + density + pH + sulphates + alcohol,      data = wwTrainSet,
ntree = 500, mtry = 6, importance = TRUE)
##                 Type of random forest: regression
##                       Number of trees: 500
## No. of variables tried at each split: 6
##
##           Mean of squared residuals: 0.4033124
##                     % Var explained: 49.55

plot(ww2) #plot results
```



```
importance(ww2) #display importance statistics

##                       %IncMSE IncNodePurity
## volatile.acidity    117.94407      419.5121
## residual.sugar       67.10755      282.9295
## free.sulfur.dioxide  86.90598      394.8253
## density              49.20727      310.1492
## pH                   60.01782      284.4473
## sulphates             55.47379      241.3552
## alcohol              124.94191      692.0755

wwpredTrainSet <- predict(ww2, wwTrainSet, type = "class") #predict the white wine
quality score against the training data

head(table(format(round(wwpredTrainSet,4),nsmall=4), wwTrainSet$quality), n=25)
#show the top 25 results of the confusion matrix.  Rounded the training set data
prediction to 4 decimal places

##
##          3 4 5 6 7 8 9
##   3.7061 1 0 0 0 0 0 0
##   3.8181 1 0 0 0 0 0 0
##   3.9086 1 0 0 0 0 0 0
##   3.9318 1 0 0 0 0 0 0
##   3.9320 1 0 0 0 0 0 0
##   3.9443 1 0 0 0 0 0 0
##   3.9658 1 0 0 0 0 0 0
##   4.0714 0 1 0 0 0 0 0
##   4.0837 0 1 0 0 0 0 0
##   4.1090 0 1 0 0 0 0 0
##   4.1253 0 2 0 0 0 0 0
##   4.1604 0 2 0 0 0 0 0
##   4.1854 0 2 0 0 0 0 0
##   4.1904 0 1 0 0 0 0 0
##   4.2019 0 1 0 0 0 0 0
##   4.2038 0 1 0 0 0 0 0
##   4.2059 0 1 0 0 0 0 0
##   4.2100 0 1 0 0 0 0 0
##   4.2154 0 1 0 0 0 0 0
##   4.2174 0 1 0 0 0 0 0
##   4.2175 0 1 0 0 0 0 0
##   4.2259 0 1 0 0 0 0 0
##   4.2451 1 0 0 0 0 0 0
##   4.2475 0 1 0 0 0 0 0
##   4.2542 1 0 0 0 0 0 0

tail(table(format(round(wwpredTrainSet,4),nsmall=4), wwTrainSet$quality), n=25)
#show the bottom 25 results of the confusion matrix.  Rounded the training set data
prediction to 4 decimal places

##
##          3 4 5 6 7 8 9
##   7.5906 0 0 0 0 0 1 0
##   7.5986 0 0 0 0 0 2 0
##   7.6014 0 0 0 0 0 1 0
##   7.6060 0 0 0 0 0 1 0
##   7.6063 0 0 0 0 0 1 0
##   7.6075 0 0 0 0 0 2 0
##   7.6133 0 0 0 0 0 1 0
##   7.6431 0 0 0 0 0 1 0
##   7.6497 0 0 0 0 0 1 0
##   7.6511 0 0 0 0 0 2 0
##   7.6526 0 0 0 0 0 2 0
##   7.6680 0 0 0 0 0 2 0
##   7.6859 0 0 0 0 0 2 0
##   7.6916 0 0 0 0 0 2 0
##   7.7253 0 0 0 0 0 2 0
##   7.7285 0 0 0 0 0 3 0
##   7.7522 0 0 0 0 0 2 0
##   7.7596 0 0 0 0 0 2 0
##   7.7654 0 0 0 0 0 2 0
##   7.8607 0 0 0 0 0 0 1
##   7.8772 0 0 0 0 0 2 0
##   7.9000 0 0 0 0 0 3 0
##   7.9424 0 0 0 0 0 0 1
##   7.9787 0 0 0 0 0 5 0
##   7.9996 0 0 0 0 0 6 0
```

```
wwpredTestSet <- predict(ww2, wwTestSet, type = "class")

plot(wwpredTestSet) #plot results
```



```
paste0("The mean of values where the predicted value equals actual is ",
mean((round(wwpredTestSet)) == wwTestSet$quality))  #Display the mean of matched
values
```

```
## [1] "The mean of values where the predicted value equals actual is
0.660544217687075"
```

```
head(table(format(round(wwpredTestSet,4),nsmall=4), wwTestSet$quality), n=25)
#show the top 25 results of the confusion matrix.  Rounded the testing set data
prediction to 4 decimal places
```

```
##
##          3 4 5 6 7 8 9
##   4.3947 0 0 1 0 0 0 0
##   4.4302 0 1 0 0 0 0 0
##   4.4418 0 1 0 0 0 0 0
##   4.4558 0 0 1 0 0 0 0
##   4.4665 0 1 0 0 0 0 0
##   4.4670 0 0 1 0 0 0 0
##   4.4961 0 1 0 0 0 0 0
##   4.5308 0 1 0 0 0 0 0
##   4.5568 0 1 0 0 0 0 0
##   4.5591 0 0 1 0 0 0 0
##   4.5934 0 1 0 0 0 0 0
##   4.6081 0 0 1 0 0 0 0
##   4.6097 0 0 1 0 0 0 0
##   4.6147 0 1 0 0 0 0 0
##   4.6235 0 0 1 0 0 0 0
##   4.6530 0 1 0 0 0 0 0
##   4.7143 0 1 0 0 0 0 0
##   4.7159 0 1 0 0 0 0 0
##   4.7198 0 1 0 0 0 0 0
##   4.7227 0 0 2 0 0 0 0
##   4.7434 0 0 1 0 0 0 0
##   4.7756 0 0 1 0 0 0 0
##   4.7954 0 1 0 0 0 0 0
##   4.7957 0 1 0 0 0 0 0
##   4.7982 0 0 0 1 0 0 0
```

```
tail(table(format(round(wwpredTestSet,4),nsmall=4), wwTestSet$quality), n=25)
#show the bottom 25 results of the confusion matrix.  Rounded the testing set data
prediction to 4 decimal places
```

```
##
##          3 4 5 6 7 8 9
##   7.1628 0 0 0 0 1 0 0
##   7.1643 0 0 0 0 1 0 0
##   7.1700 0 0 0 1 0 0 0
##   7.1787 0 0 0 0 0 1 0
##   7.1838 0 0 0 1 0 0 0
##   7.1896 0 0 0 0 0 1 0
##   7.1935 0 0 0 0 0 1 0
##   7.1948 0 0 0 0 1 0 0
##   7.1990 0 0 0 0 1 0 0
##   7.2199 0 0 0 0 1 0 0
##   7.2204 0 0 1 0 0 0 0
##   7.2714 0 0 0 0 0 1 0
##   7.2848 0 0 0 0 0 1 0
##   7.2849 0 0 0 0 0 1 0
##   7.3300 0 0 0 0 1 0 0
##   7.3495 0 0 0 0 0 1 0
##   7.3614 0 0 0 0 0 1 0
##   7.3937 0 0 0 0 0 1 0
##   7.4348 0 0 0 0 0 1 0
##   7.4427 0 0 0 0 0 1 0
##   7.4490 0 0 0 0 0 1 0
##   7.4979 0 0 0 0 0 1 0
##   7.5487 0 0 0 0 0 1 0
##   7.5728 0 0 0 0 0 1 0
##   7.9787 0 0 0 0 0 3 0
```