Wine Quality ML Regression Models

Josh Bicer

2023-06-01

Part 1: Install libraries and Read Data

```
# Install required packages and libraries
library(ggplot2)
library(MASS)
library(pscl)
library(dplyr)
library(readr)
library(corrplot)
library(zoo)
library(Amelia)
library(psych)
library(car)
library(glm2)
library(leaps)

# Read in data from CSV file
wine = read.csv("wine_training.csv")
wine_test = read.csv("wine_testing.csv")
```

Part 2: Data Exploration

This section explores the wine quality dataset to gain a better understanding of the data used for the model. This includes histogram and boxplot visualizations for the target variable and all applicable input variables. Correlation plots and summaries of the dataset are included below.

```
# Review data frame summary
summary(wine)
```

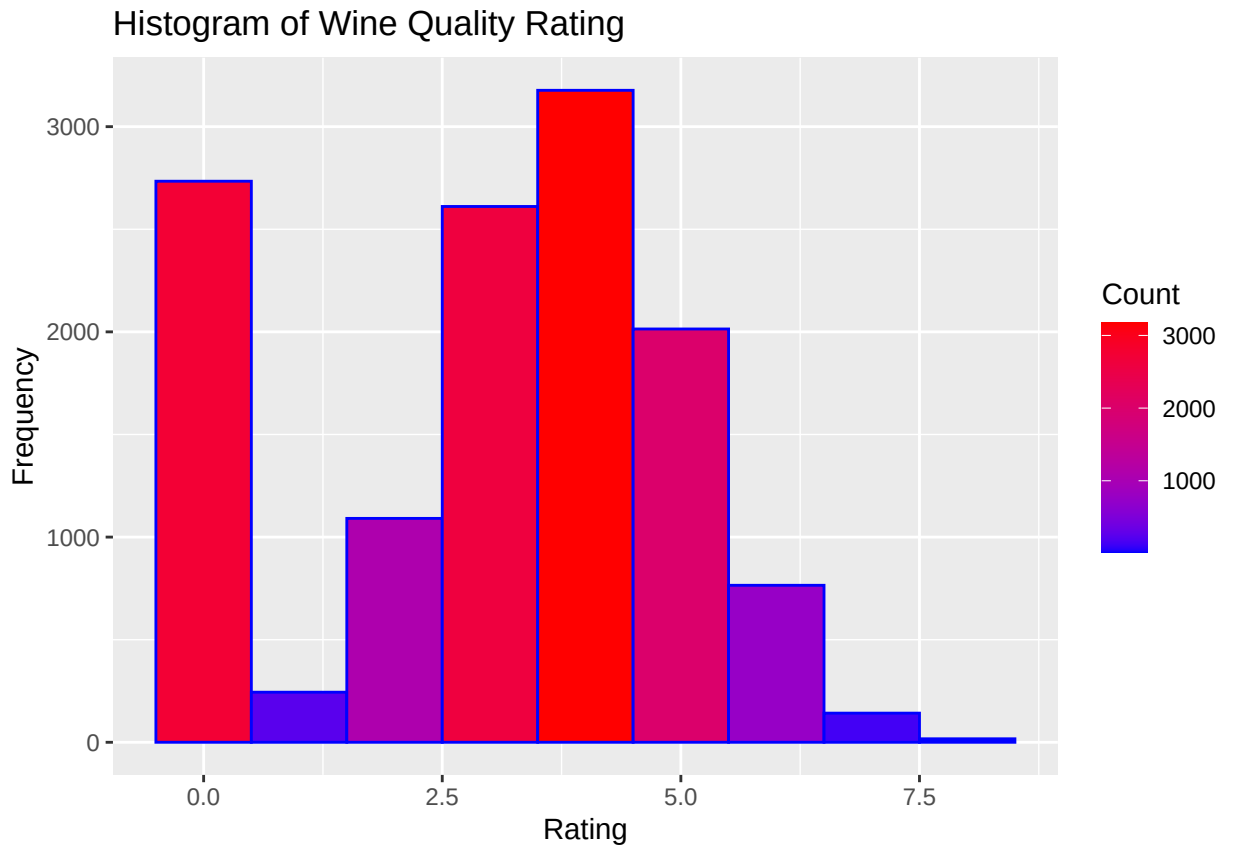
```
##      INDEX           TARGET       FixedAcidity     VolatileAcidity
##  Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900
##  1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300
##  Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800
##  Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241
##  3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400
##  Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800
##
##    CitricAcid      ResidualSugar        Chlorides       FreeSulfurDioxide
##  Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00
##  1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00
##  Median : 0.3100   Median :   3.900   Median : 0.0460   Median :  30.00
##  Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85
##  3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00
```

```
##  Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00
##                    NA's   :616       NA's   :638       NA's   :647
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
##  1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800
##  Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000
##  Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
##  3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600
##  Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
##  NA's   :682                         NA's   :395     NA's   :1210
##     Alcohol       LabelAppeal          AcidIndex          STARS
##  Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median :10.40   Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##  NA's   :653                                          NA's   :3359
```

```
# Review structure of the data frame
str(wine)
```

```
## 'data.frame':    12795 obs. of  16 variables:
##  $ INDEX             : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET            : int  3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity      : num  3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity   : num  1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
##  $ CitricAcid        : num  -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
##  $ ResidualSugar     : num  54.2 26.1 14.8 18.8 9.4 ...
##  $ Chlorides         : num  -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...
##  $ FreeSulfurDioxide : num  NA 15 214 22 -167 -37 287 523 -213 62 ...
##  $ TotalSulfurDioxide: num  268 -327 142 115 108 15 156 551 NA 180 ...
##  $ Density           : num  0.993 1.028 0.995 0.996 0.995 ...
##  $ pH                : num  3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
##  $ Sulphates         : num  -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...
##  $ Alcohol           : num  9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
##  $ LabelAppeal       : int  0 -1 -1 -1 0 0 0 1 0 0 ...
##  $ AcidIndex         : int  8 7 8 6 9 11 8 7 6 8 ...
##  $ STARS             : int  2 3 3 1 2 NA NA 3 NA 4 ...
```

```
# Histogram of target variable for Star Rating
# Rating is on a scale between 0 - 10
ggplot(data=wine, aes(TARGET)) +
  geom_histogram(binwidth =1,
                 col="BLUE",
                 aes(fill=..count..))+
  scale_fill_gradient("Count", low = "blue", high = "red") +
  ggtitle("Histogram of Wine Quality Rating") +
  xlab("Rating") +
  ylab("Frequency")
```




```
zero_count = length(wine$TARGET[wine$TARGET == 0])
print(paste('The number of zeros in Target Variable is: ', zero_count))
```

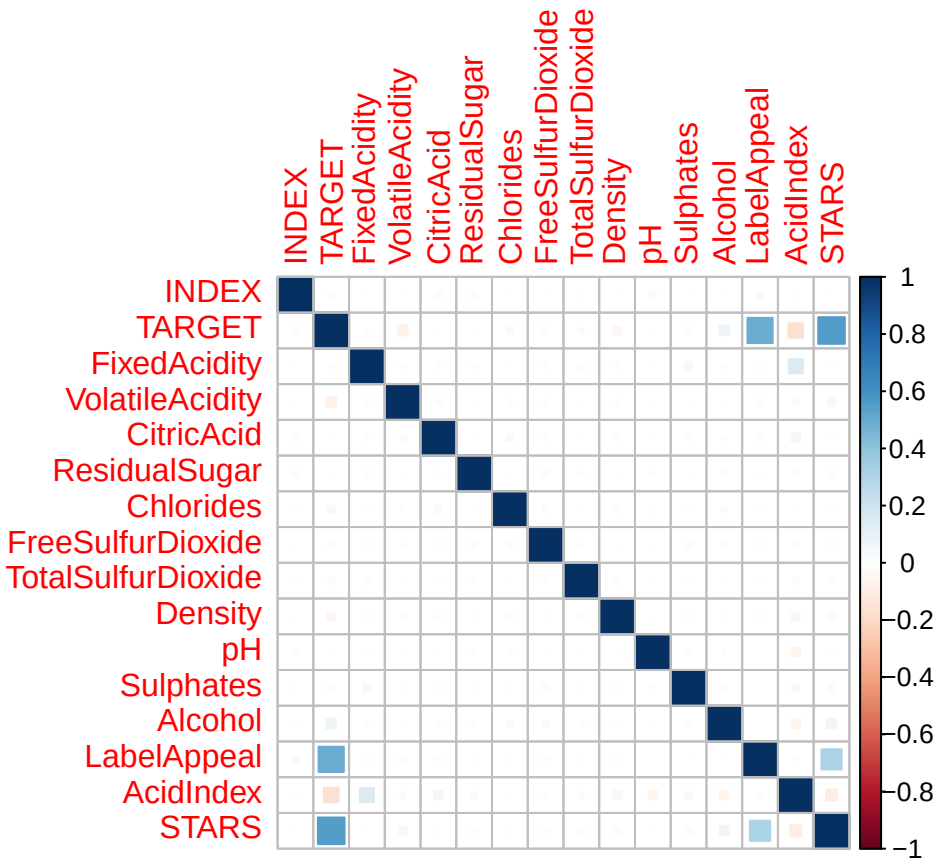
```
## [1] "The number of zeros in Target Variable is:  2734"
```

```
# Examine correlation among variables
wine_clean = na.omit(wine)
cor(wine_clean[sapply(wine_clean, is.numeric)])
```

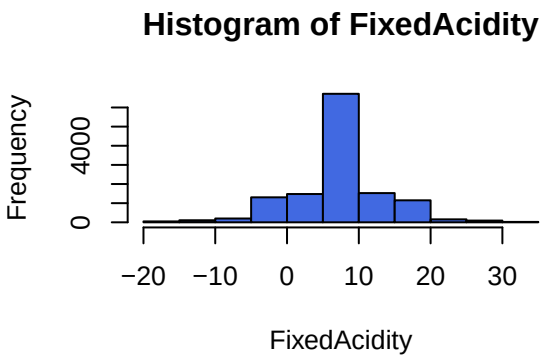
```
##                          INDEX        TARGET FixedAcidity VolatileAcidity
## INDEX              1.0000000000  0.0236764338 -0.002831415   -0.0008743296
## TARGET             0.0236764338  1.0000000000 -0.012538100   -0.0759978765
## FixedAcidity      -0.0028314152 -0.0125380998  1.000000000    0.0190109733
## VolatileAcidity   -0.0008743296 -0.0759978765  0.019010973    1.0000000000
## CitricAcid         0.0278869710  0.0023450490  0.014000376   -0.0234315631
## ResidualSugar      0.0208952098  0.0035195999 -0.015429391    0.0015279517
## Chlorides          0.0026827829 -0.0304301331 -0.006104447    0.0148489225
## FreeSulfurDioxide  0.0046416504  0.0226398054  0.015438463   -0.0114408079
## TotalSulfurDioxide 0.0064949038  0.0216020726 -0.023323485   -0.0007434083
## Density           -0.0034840089 -0.0475989086  0.011574241    0.0130977690
## pH                -0.0274556333  0.0002198557 -0.004553886    0.0072030364
## Sulphates         -0.0053946247 -0.0212203783  0.042229181    0.0015161001
## Alcohol           -0.0024453460  0.0737771084 -0.013085026    0.0002603082
## LabelAppeal        0.0314911460  0.4979464796  0.011375965   -0.0202419713
## AcidIndex          0.0055244862 -0.1676430648  0.154167846    0.0250529742
## STARS             -0.0057807296  0.5546857223 -0.004937345   -0.0402432388
##                      CitricAcid ResidualSugar     Chlorides FreeSulfurDioxide
```

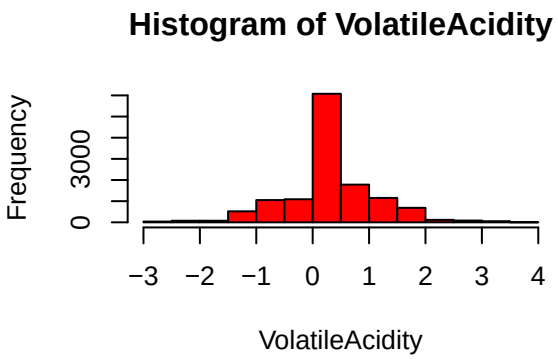
```
## INDEX               0.0278869710   0.020895210  0.0026827829       0.004641650
## TARGET              0.0023450490   0.003519600 -0.0304301331       0.022639805
## FixedAcidity        0.0140003760  -0.015429391 -0.0061044471       0.015438463
## VolatileAcidity    -0.0234315631   0.001527952  0.0148489225      -0.011440808
## CitricAcid          1.0000000000  -0.009843146 -0.0335608661       0.012113248
## ResidualSugar      -0.0098431456   1.000000000  0.0041215692       0.021959113
## Chlorides          -0.0335608661   0.004121569  1.0000000000      -0.020492488
## FreeSulfurDioxide   0.0121132485   0.021959113 -0.0204924876       1.000000000
## TotalSulfurDioxide -0.0099174506   0.017030939  0.0004188605       0.013461673
## Density            -0.0169919691  -0.007120841  0.0206724860      -0.008663509
## pH                 -0.0007581304   0.017563769 -0.0179702278      -0.002008516
## Sulphates          -0.0144237270  -0.002705775  0.0026187777       0.026829029
## Alcohol             0.0169864284  -0.018943324 -0.0228849573      -0.023867458
## LabelAppeal         0.0153315666  -0.004579308 -0.0063870237       0.014960087
## AcidIndex           0.0545838104  -0.020301890 -0.0017134096      -0.014733717
## STARS               0.0071401699   0.019665541 -0.0063242568      -0.015390398
##                    TotalSulfurDioxide      Density            pH    Sulphates
## INDEX                   0.0064949038 -0.003484009 -0.0274556333 -0.005394625
## TARGET                  0.0216020726 -0.047598909  0.0002198557 -0.021220378
## FixedAcidity           -0.0233234848  0.011574241 -0.0045538857  0.042229181
## VolatileAcidity        -0.0007434083  0.013097769  0.0072030364  0.001516100
## CitricAcid             -0.0099174506 -0.016991969 -0.0007581304 -0.014423727
## ResidualSugar           0.0170309394 -0.007120841  0.0175637691 -0.002705775
## Chlorides               0.0004188605  0.020672486 -0.0179702278  0.002618778
## FreeSulfurDioxide       0.0134616726 -0.008663509 -0.0020085157  0.026829029
## TotalSulfurDioxide      1.0000000000  0.023167955 -0.0034227601  0.002504051
## Density                 0.0231679548  1.000000000 -0.0020192285 -0.010609294
## pH                     -0.0034227601 -0.002019229  1.0000000000  0.010449255
## Sulphates               0.0025040509 -0.010609294  0.0104492547  1.000000000
## Alcohol                -0.0168515467 -0.006128355 -0.0122034469  0.010844330
## LabelAppeal            -0.0027237419 -0.018094403  0.0002181758  0.003768700
## AcidIndex              -0.0221292631  0.047778830 -0.0537128921  0.031071782
## STARS                   0.0220949002 -0.028492455 -0.0044002985 -0.023135130
##                          Alcohol   LabelAppeal    AcidIndex        STARS
## INDEX               -0.0024453460  0.0314911460  0.005524486 -0.005780730
## TARGET               0.0737771084  0.4979464796 -0.167643065  0.554685722
## FixedAcidity        -0.0130850260  0.0113759650  0.154167846 -0.004937345
## VolatileAcidity      0.0002603082 -0.0202419713  0.025052974 -0.040243239
## CitricAcid           0.0169864284  0.0153315666  0.054583810  0.007140170
## ResidualSugar       -0.0189433242 -0.0045793083 -0.020301890  0.019665541
## Chlorides           -0.0228849573 -0.0063870237 -0.001713410 -0.006324257
## FreeSulfurDioxide   -0.0238674577  0.0149600871 -0.014733717 -0.015390398
## TotalSulfurDioxide  -0.0168515467 -0.0027237419 -0.022129263  0.022094900
## Density             -0.0061283546 -0.0180944026  0.047778830 -0.028492455
## pH                  -0.0122034469  0.0002181758 -0.053712892 -0.004400299
## Sulphates            0.0108443299  0.0037686996  0.031071782 -0.023135130
## Alcohol              1.0000000000 -0.0006449123 -0.055891906  0.064854486
## LabelAppeal         -0.0006449123  1.0000000000  0.010300984  0.318897022
## AcidIndex           -0.0558919056  0.0103009840  1.000000000 -0.095482582
## STARS                0.0648544864  0.3188970216 -0.095482582  1.000000000
```

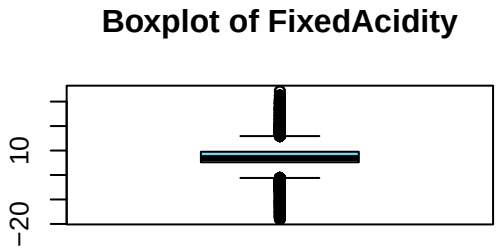
```
# Plot correlation
corrplot(cor(wine_clean), method = "square")
```

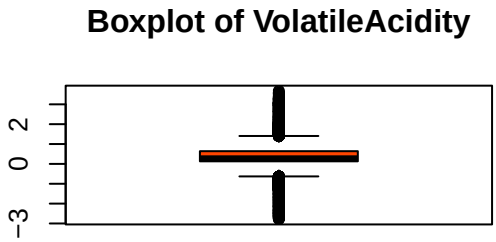



```
### Explore histograms and boxplots of independent variables
# Fixed Acidity and Volatile Acidity
par(mfrow=c(2,2))
hist(wine$FixedAcidity, col = "royalblue", xlab = "FixedAcidity", main = "Histogram of FixedAcidity")
hist(wine$VolatileAcidity, col = "red", xlab = "VolatileAcidity", main = "Histogram of VolatileAcidity")
boxplot(wine$FixedAcidity, col = "skyblue", main = "Boxplot of FixedAcidity")
boxplot(wine$VolatileAcidity, col = "orangered", main = "Boxplot of VolatileAcidity")
```

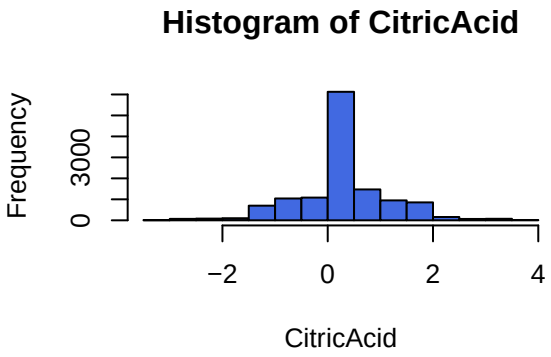


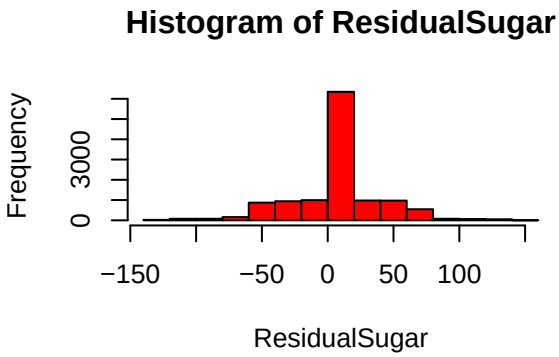


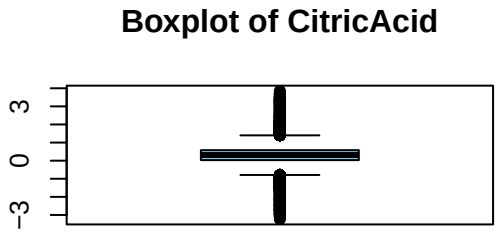


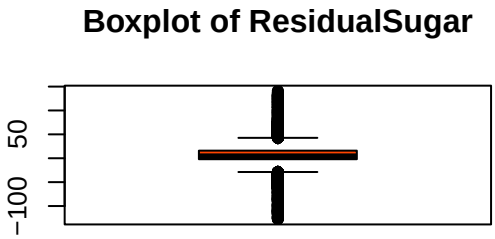



```r
par(mfrow=c(1,1))

# Citric Acid and Residual Sugar
par(mfrow=c(2,2))
hist(wine$CitricAcid, col = "royalblue", xlab = "CitricAcid", main = "Histogram of CitricAcid")
hist(wine$ResidualSugar, col = "red", xlab = "ResidualSugar", main = "Histogram of ResidualSugar")
boxplot(wine$CitricAcid, col = "skyblue", main = "Boxplot of CitricAcid")
boxplot(wine$ResidualSugar, col = "orangered", main = "Boxplot of ResidualSugar")
```

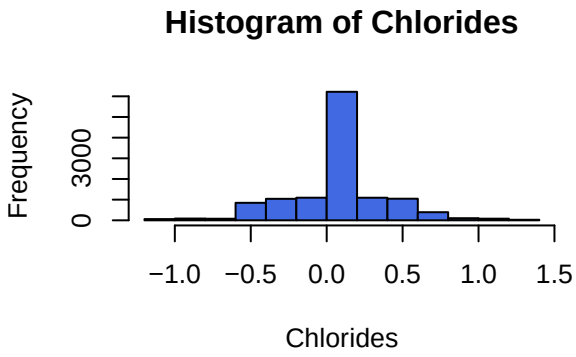


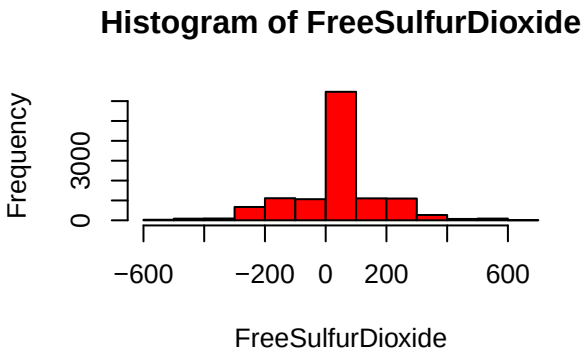


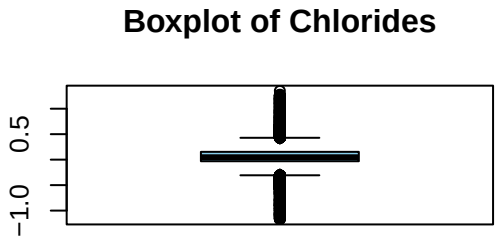


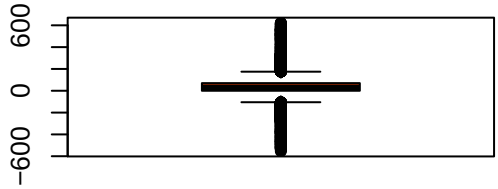



```
par(mfrow=c(1,1))

# Chlorides and Free Sulfur Dioxide
par(mfrow=c(2,2))
hist(wine$Chlorides, col = "royalblue", xlab = "Chlorides", main = "Histogram of Chlorides")
hist(wine$FreeSulfurDioxide, col = "red", xlab = "FreeSulfurDioxide", main = "Histogram of FreeSulfurDi
boxplot(wine$Chlorides, col = "skyblue", main = "Boxplot of Chlorides")
boxplot(wine$FreeSulfurDioxide, col = "orangered", main = "Boxplot of FreeSulfurDioxide")
```

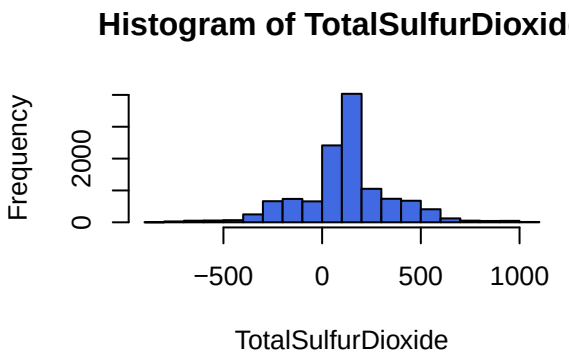


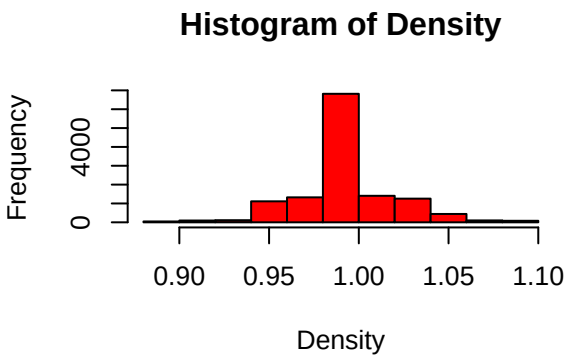


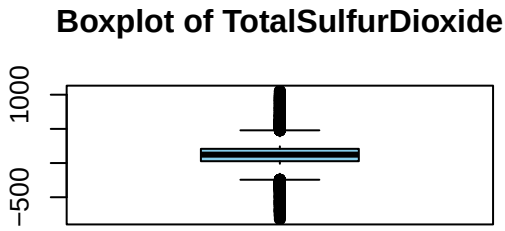


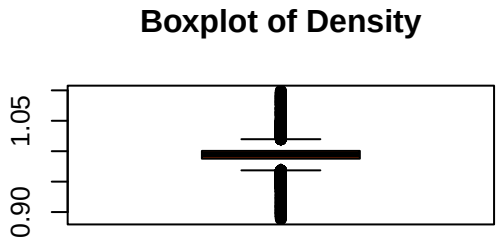



```r
par(mfrow=c(1,1))

# Total Sulfur Dioxide and Density
par(mfrow=c(2,2))
hist(wine$TotalSulfurDioxide, col = "royalblue", xlab = "TotalSulfurDioxide", main = "Histogram of Total
hist(wine$Density, col = "red", xlab = "Density", main = "Histogram of Density")
boxplot(wine$TotalSulfurDioxide, col = "skyblue", main = "Boxplot of TotalSulfurDioxide")
boxplot(wine$Density, col = "orangered", main = "Boxplot of Density")
```

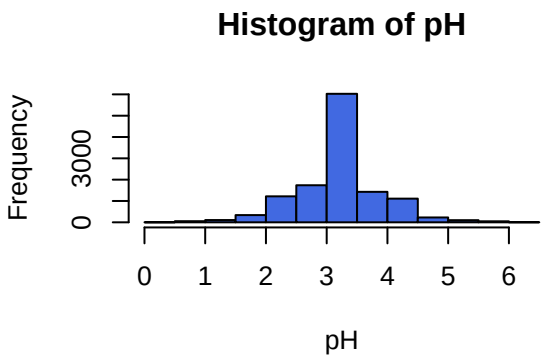


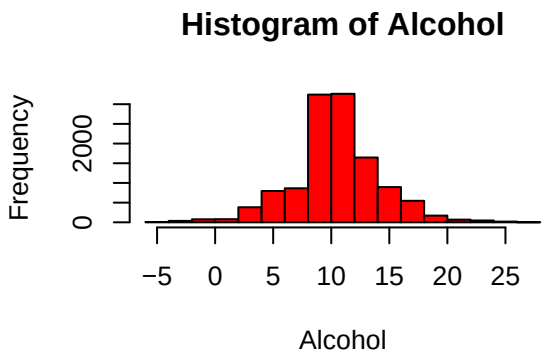


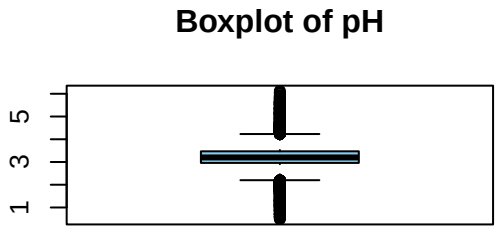


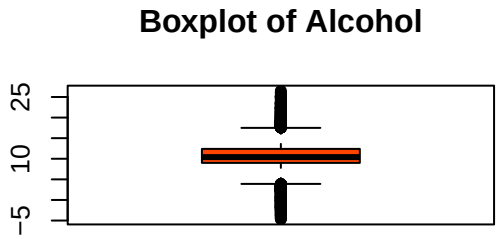



```
par(mfrow=c(1,1))

# pH Level and Alocohol Content
par(mfrow=c(2,2))
hist(wine$pH, col = "royalblue", xlab = "pH", main = "Histogram of pH")
hist(wine$Alcohol, col = "red", xlab = "Alcohol", main = "Histogram of Alcohol")
boxplot(wine$pH, col = "skyblue", main = "Boxplot of pH")
boxplot(wine$Alcohol, col = "orangered", main = "Boxplot of Alcohol")
```

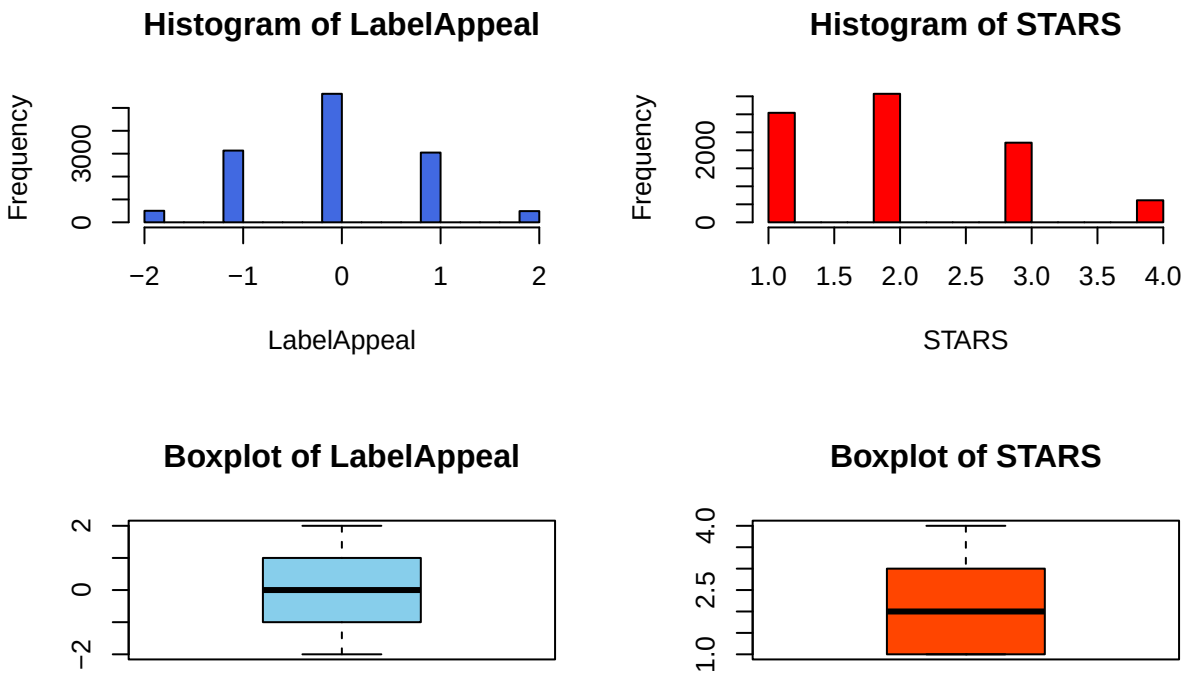









```
par(mfrow=c(1,1))

# Label appeal and STARS
par(mfrow=c(2,2))
hist(wine$LabelAppeal, col = "royalblue", xlab = "LabelAppeal", main = "Histogram of LabelAppeal")
hist(wine$STARS, col = "red", xlab = "STARS", main = "Histogram of STARS")
boxplot(wine$LabelAppeal, col = "skyblue", main = "Boxplot of LabelAppeal")
boxplot(wine$STARS, col = "orangered", main = "Boxplot of STARS")
```




```
par(mfrow=c(1,1))
```

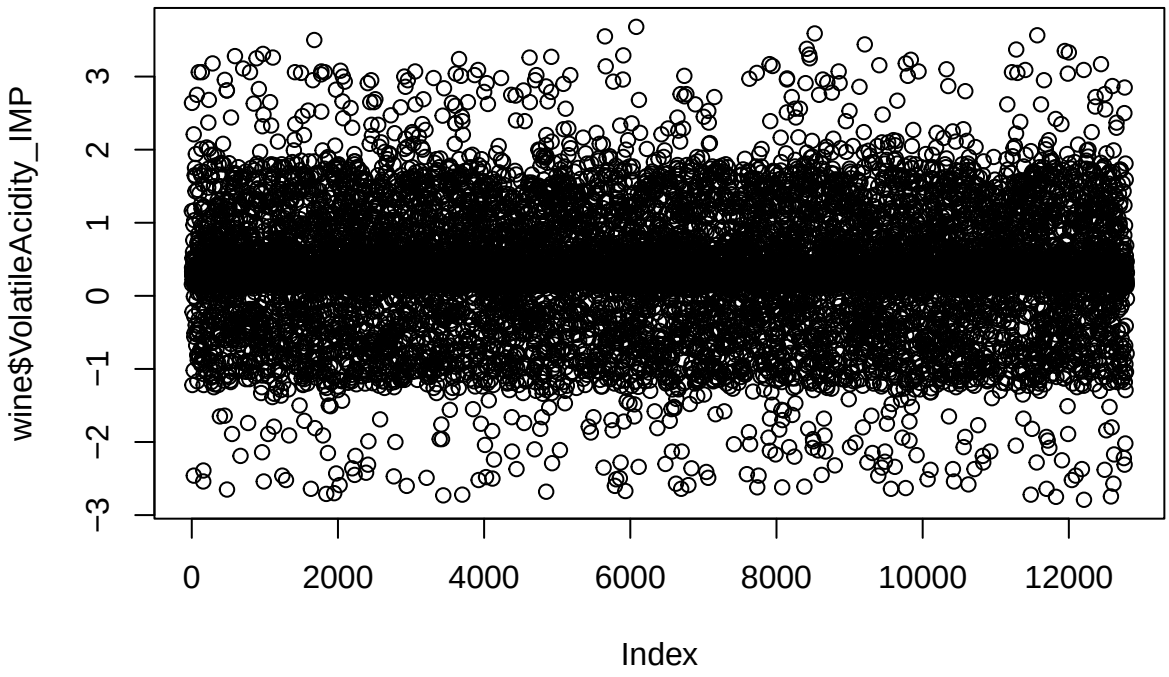
Part 3: Data Preparation

This section is used to clean the data to prepare for use in the machine learning models. All data cleaning steps are performed on both the training and testing datasets for consistency. Any missing values in the dataset are replaced with the median value and any significant outliers are replaced with the 1% or 99% value. Additionally, all original data is preserved and any imputed values in the model use the _IMP label to differentiate.

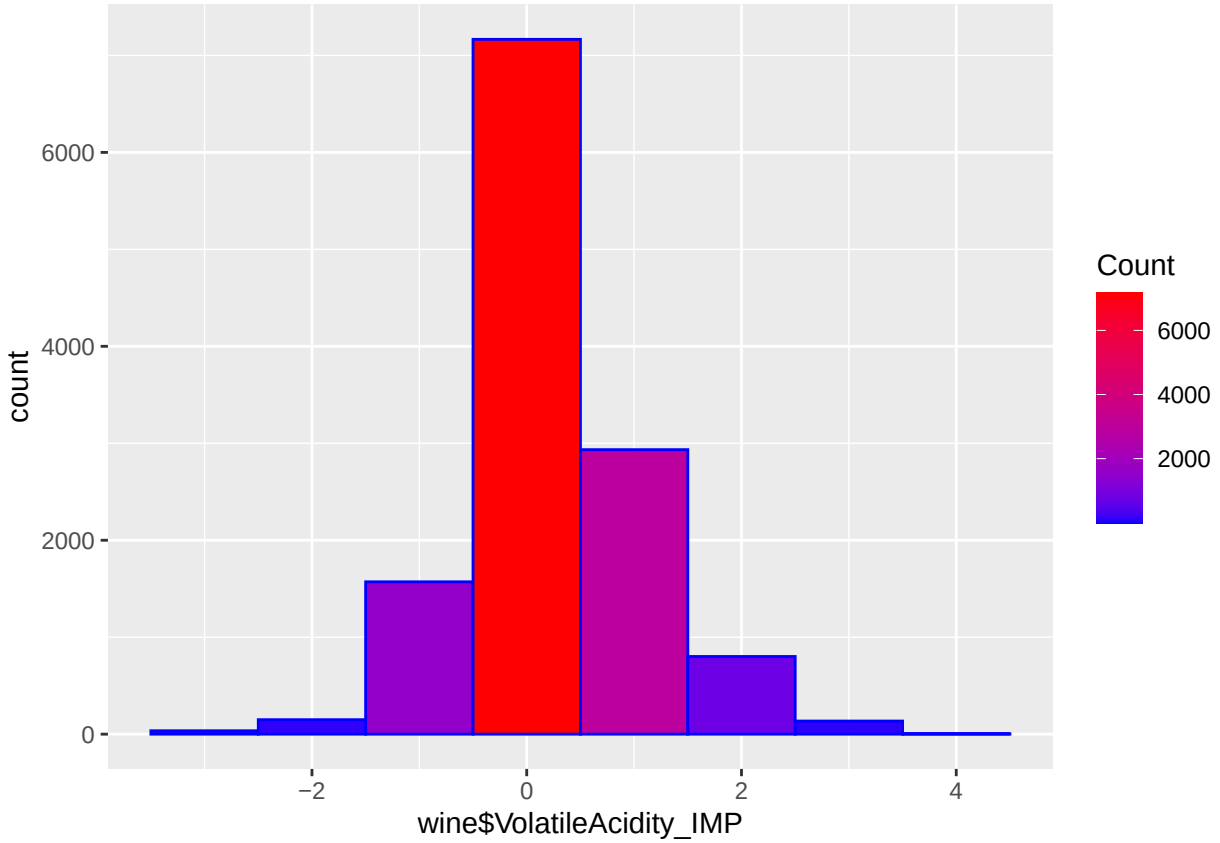
```
### Training Data Set
# Create imputed variables for each column
wine$FixedAcidity_IMP = wine$FixedAcidity
wine$VolatileAcidity_IMP = wine$VolatileAcidity
wine$CitricAcid_IMP = wine$CitricAcid
wine$ResidualSugar_IMP = wine$ResidualSugar
wine$Chlorides_IMP = wine$Chlorides
wine$FreeSulfurDioxide_IMP = wine$FreeSulfurDioxide
wine$TotalSulfurDioxide_IMP = wine$TotalSulfurDioxide
wine$Density_IMP = wine$Density
wine$pH_IMP = wine$pH
wine$Sulphates_IMP = wine$Sulphates
wine$Alcohol_IMP = wine$Alcohol
wine$LabelAppeal_IMP = wine$LabelAppeal
wine$AcidIndex_IMP = wine$AcidIndex
wine$STARS_IMP = wine$STARS
```

```
### Replace missing NA values with median and outliers with 1st or 99th percentile
### Create FLAG variables where replaced.
#ResidualSugar
wine$ResidualSugar_FLAG = as.factor(ifelse(wine$ResidualSugar_IMP < quantile(wine$ResidualSugar_IMP, c(
                                           | wine$ResidualSugar_IMP > quantile(wine$ResidualSugar_IMP,c
                                           | is.na(wine$ResidualSugar_IMP),1,0))
wine$ResidualSugar_IMP[is.na(wine$ResidualSugar_IMP)] = median(wine$ResidualSugar_IMP, na.rm=TRUE)
wine$ResidualSugar_IMP = as.numeric(ifelse(wine$ResidualSugar_IMP < quantile(wine$ResidualSugar_IMP, c(
                                           quantile(wine$ResidualSugar_IMP,c(.01)), wine$ResidualSugar_
wine$ResidualSugar_IMP = as.numeric(ifelse(wine$ResidualSugar_IMP > quantile(wine$ResidualSugar_IMP, c(
                                           quantile(wine$ResidualSugar_IMP,c(.99)), wine$ResidualSugar_

#Chlorides
wine$Chlorides_FLAG = as.factor(ifelse(is.na(wine$Chlorides),1,0))
wine$Chlorides_IMP[is.na(wine$Chlorides_IMP)] = median(wine$Chlorides, na.rm = TRUE)

#FreeSulfurDioxide
wine$FreeSulfurDioxide_FLAG = as.factor(ifelse(wine$FreeSulfurDioxide_IMP < quantile(wine$FreeSulfurDio
                                           | wine$FreeSulfurDioxide_IMP > quantile(wine$FreeSulfurDioxi
                                           | is.na(wine$FreeSulfurDioxide_IMP),1,0))
wine$FreeSulfurDioxide_IMP[is.na(wine$FreeSulfurDioxide_IMP)] = median(wine$FreeSulfurDioxide_IMP, na.r
wine$FreeSulfurDioxide_IMP = as.numeric(ifelse(wine$FreeSulfurDioxide_IMP < quantile(wine$FreeSulfurDio
                                           quantile(wine$FreeSulfurDioxide_IMP,c(.01)), wine$FreeSulfur
wine$FreeSulfurDioxide_IMP = as.numeric(ifelse(wine$FreeSulfurDioxide_IMP > quantile(wine$FreeSulfurDio
                                           quantile(wine$FreeSulfurDioxide_IMP,c(.99)), wine$FreeSulfur

#TotalSulfurDioxide
wine$TotalSulfurDioxide_FLAG = as.factor(ifelse(wine$TotalSulfurDioxide_IMP < quantile(wine$TotalSulfur
                                                | wine$TotalSulfurDioxide_IMP > quantile(wine$TotalSulfu
                                                | is.na(wine$TotalSulfurDioxide_IMP),1,0))
wine$TotalSulfurDioxide_IMP[is.na(wine$TotalSulfurDioxide_IMP)] = median(wine$TotalSulfurDioxide_IMP, na
wine$TotalSulfurDioxide_IMP = as.numeric(ifelse(wine$TotalSulfurDioxide_IMP < quantile(wine$TotalSulfur
                                                quantile(wine$TotalSulfurDioxide_IMP,c(.01)), wine$Total
wine$TotalSulfurDioxide_IMP = as.numeric(ifelse(wine$TotalSulfurDioxide_IMP > quantile(wine$TotalSulfur
                                                quantile(wine$TotalSulfurDioxide_IMP,c(.99)), wine$Total

#pH
wine$pH_FLAG = as.factor(ifelse(is.na(wine$pH),1,0))
wine$pH_IMP[is.na(wine$pH_IMP)] = median(wine$pH, na.rm = TRUE)

#Sulfates
wine$Sulphates_FLAG = as.factor(ifelse(is.na(wine$Sulphates),1,0))
wine$Sulphates_IMP[is.na(wine$Sulphates_IMP)] = median(wine$Sulphates, na.rm = TRUE)

#Alcohol
wine$Alcohol_FLAG = as.factor(ifelse(is.na(wine$Alcohol),1,0))
wine$Alcohol_IMP[is.na(wine$Alcohol_IMP)] = median(wine$Alcohol, na.rm = TRUE)

#Stars
wine$STARS_FLAG = as.factor(ifelse(is.na(wine$STARS),1,0))
wine$STARS_IMP = na.aggregate(wine$STARS_IMP, wine$LabelAppeal)

# Create Red/Wine Flag Variable
```

```
plot(wine$VolatileAcidity_IMP)
```



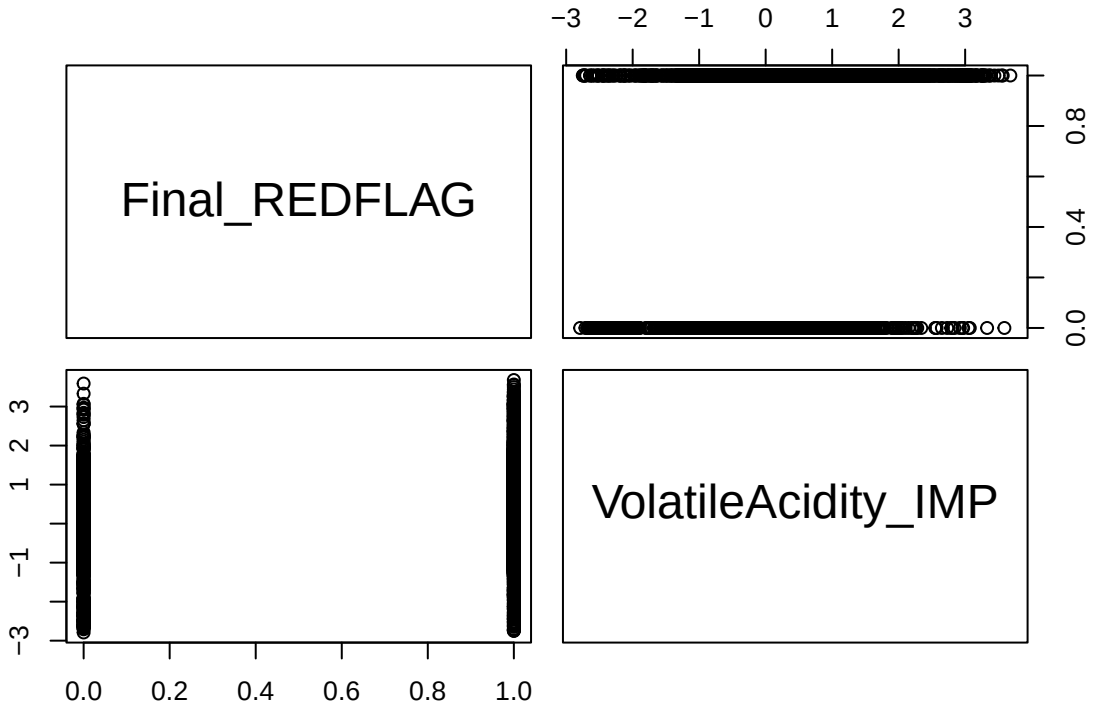

```
# Visualize volatile acidity
ggplot(data=wine, aes(wine$VolatileAcidity_IMP)) +
  geom_histogram(binwidth =1,
                 col="BLUE",
                 aes(fill=..count..))+
  scale_fill_gradient("Count", low = "blue", high = "red")
```



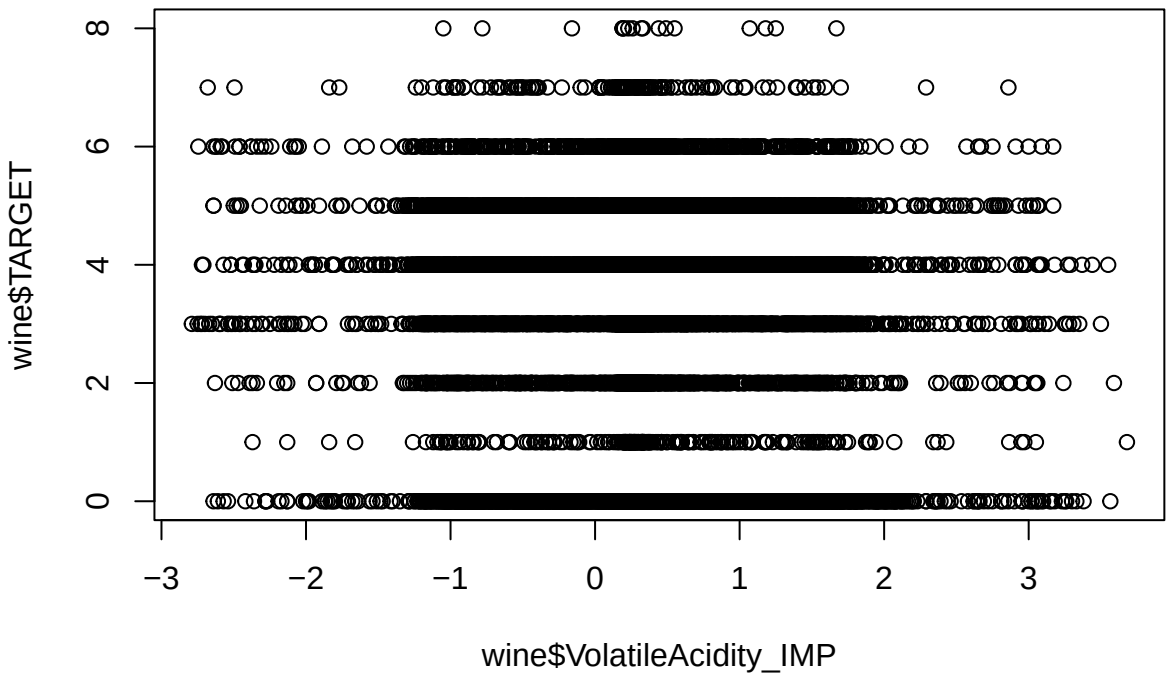

```
summary(wine$VolatileAcidity_IMP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.7900  0.1300  0.2800  0.3241  0.6400  3.6800
```

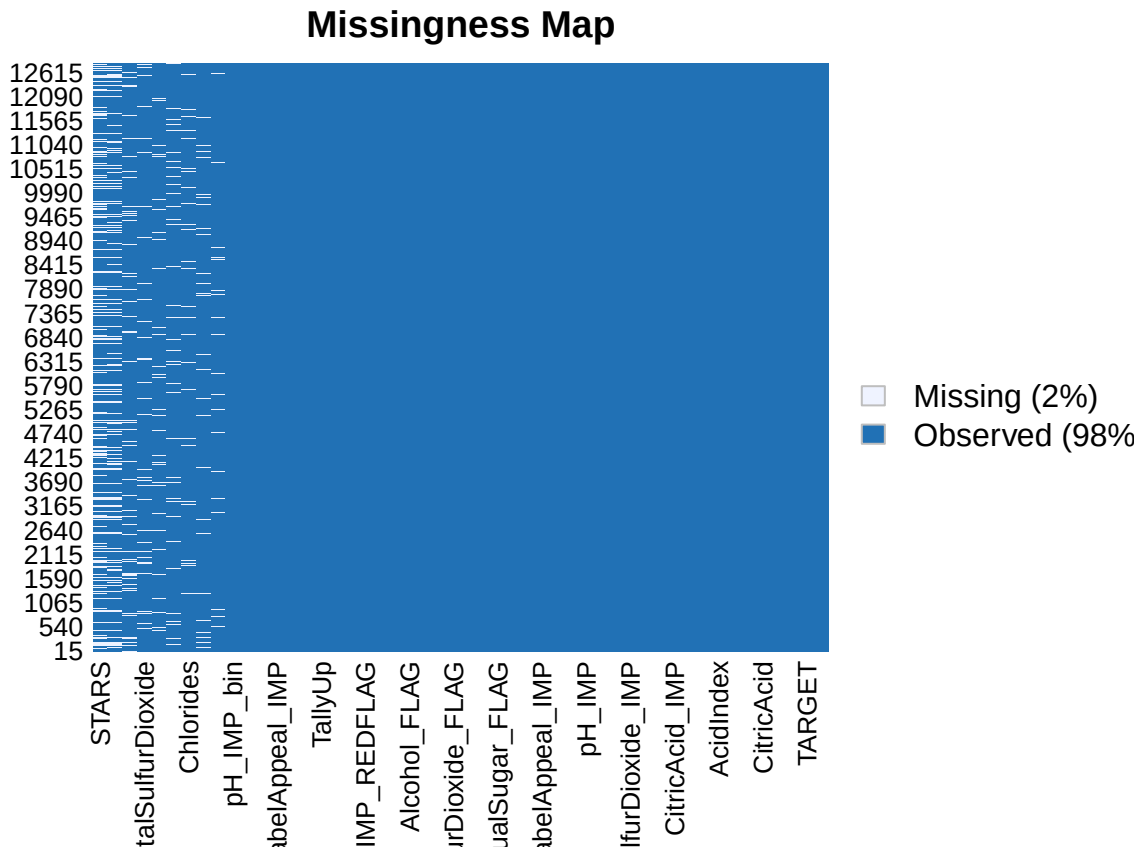
```
# Make new indicator that indicates red vs white based on volatile acidity
wine$VolatileAcidity_IMP_REDFLAG = ifelse(wine$VolatileAcidity_IMP > mean(wine$VolatileAcidity_IMP),1,0
wine$ResidualSugar_IMP_REDFLAG = ifelse(wine$ResidualSugar_IMP < mean(wine$ResidualSugar_IMP),1,0)
wine$TotalSulfurDioxide_IMP_REDFLAG = ifelse(wine$TotalSulfurDioxide_IMP < mean(wine$TotalSulfurDioxide_
wine$Density_IMP_REDFLAG = ifelse(wine$Density_IMP > mean(wine$Density_IMP),1,0)
wine$TallyUp = wine$VolatileAcidity_IMP_REDFLAG + wine$ResidualSugar_IMP_REDFLAG + wine$TotalSulfurDioxi
wine$Final_REDFLAG = ifelse(wine$TallyUp > mean(wine$TallyUp),1,0)

pairs(wine[,c("Final_REDFLAG","VolatileAcidity_IMP")])
```




```
plot(wine$VolatileAcidity_IMP,wine$TARGET)
```




```
# Add Target Flag for 0 sale scenarios
wine$TARGET_Flag = ifelse(wine$TARGET > 0,1,0)
wine$TARGET_AMT = wine$TARGET - 1
wine$TARGET_AMT = ifelse(wine$TARGET_Flag == 0,NA,wine$TARGET-1)

# Create interaction terms and imputed variables from input variables
wine$STARSxLabelAppeal_IMP = wine$STARS_IMP * wine$LabelAppeal_IMP
wine$STARSxAlcohol_IMP = wine$STARS_IMP * wine$Alcohol_IMP

# Create bins for 3 levels of Residual Sugar
wine$ResidualSugar_bin[wine$ResidualSugar_IMP < quantile(wine$ResidualSugar_IMP, c(.25))] = "Low Sweetn
wine$ResidualSugar_bin[wine$ResidualSugar_IMP >= quantile(wine$ResidualSugar_IMP, c(.25)) &
                       wine$ResidualSugar_IMP < quantile(wine$ResidualSugar_IMP, c(.75))] = "Medium Sw
wine$ResidualSugar_bin[wine$ResidualSugar_IMP >= quantile(wine$ResidualSugar_IMP, c(.75))] = "High Swee
wine$ResidualSugar_bin = factor(wine$ResidualSugar_bin, levels = c("Low Sweetness", "Medium Sweetness",

# Create bins for 3 levels of acidity
wine$pH_IMP_bin[wine$pH_IMP <= 3] = "Low Acidity"
wine$pH_IMP_bin[wine$pH_IMP > 3 & wine$pH_IMP < 3.5] = "Medium Acidity"
wine$pH_IMP_bin[wine$pH_IMP >= 3.5] = "High Acidity"
wine$pH_IMP_bin = factor(wine$pH_IMP_bin, levels = c("Low Acidity", "Medium Acidity", "High Acidity"))

# Confirm no NAs for imputed variables
missmap(wine)
```




```
summary(wine)
```

```
##      INDEX           TARGET       FixedAcidity     VolatileAcidity
##  Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900
##  1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300
##  Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800
##  Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241
##  3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400
##  Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800
##
##    CitricAcid      ResidualSugar        Chlorides       FreeSulfurDioxide
##  Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00
##  1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00
##  Median : 0.3100   Median :   3.900   Median : 0.0460   Median :  30.00
##  Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85
##  3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00
##  Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00
##                    NA's   :616        NA's   :638       NA's   :647
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
##  1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800
##  Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000
##  Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
##  3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600
##  Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
##  NA's   :682                         NA's   :395     NA's   :1210
```

```
##     Alcohol       LabelAppeal          AcidIndex          STARS
##  Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median :10.40   Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##  NA's   :653                                          NA's   :3359
##  FixedAcidity_IMP  VolatileAcidity_IMP CitricAcid_IMP    ResidualSugar_IMP
##  Min.   :-18.100   Min.   :-2.7900     Min.   :-3.2400   Min.   :-89.036
##  1st Qu.:  5.200   1st Qu.: 0.1300     1st Qu.: 0.0300   1st Qu.:  0.900
##  Median :  6.900   Median : 0.2800     Median : 0.3100   Median :  3.900
##  Mean   :  7.076   Mean   : 0.3241     Mean   : 0.3084   Mean   :  5.339
##  3rd Qu.:  9.500   3rd Qu.: 0.6400     3rd Qu.: 0.5800   3rd Qu.: 14.900
##  Max.   : 34.400   Max.   : 3.6800     Max.   : 3.8600   Max.   : 96.912
##
##  Chlorides_IMP      FreeSulfurDioxide_IMP TotalSulfurDioxide_IMP
##  Min.   :-1.17100   Min.   :-381.06       Min.   :-515.1
##  1st Qu.: 0.00000   1st Qu.:   5.00       1st Qu.:  34.0
##  Median : 0.04600   Median :  30.00       Median : 123.0
##  Mean   : 0.05438   Mean   :  30.85       Mean   : 120.7
##  3rd Qu.: 0.12800   3rd Qu.:  64.00       3rd Qu.: 198.0
##  Max.   : 1.35100   Max.   : 460.24       Max.   : 744.1
##
##   Density_IMP         pH_IMP      Sulphates_IMP      Alcohol_IMP
##  Min.   :0.8881   Min.   :0.480   Min.   :-3.1300   Min.   :-4.70
##  1st Qu.:0.9877   1st Qu.:2.970   1st Qu.: 0.3400   1st Qu.: 9.10
##  Median :0.9945   Median :3.200   Median : 0.5000   Median :10.40
##  Mean   :0.9942   Mean   :3.207   Mean   : 0.5245   Mean   :10.48
##  3rd Qu.:1.0005   3rd Qu.:3.450   3rd Qu.: 0.7700   3rd Qu.:12.20
##  Max.   :1.0992   Max.   :6.130   Max.   : 4.2400   Max.   :26.50
##
##  LabelAppeal_IMP     AcidIndex_IMP      STARS_IMP     ResidualSugar_FLAG
##  Min.   :-2.000000   Min.   : 4.000   Min.   :1.000   0:11935
##  1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.381   1:  860
##  Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :-0.009066   Mean   : 7.773   Mean   :2.023
##  3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:2.391
##  Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##
##  Chlorides_FLAG FreeSulfurDioxide_FLAG TotalSulfurDioxide_FLAG pH_FLAG
##  0:12157        0:11908                0:11869                 0:12400
##  1:  638        1:  887                1:  926                 1:  395
##
##
##
##
##
##  Sulphates_FLAG Alcohol_FLAG STARS_FLAG VolatileAcidity_IMP_REDFLAG
##  0:11585        0:12142      0:9436     Min.   :0.0000
##  1: 1210        1:  653      1:3359     1st Qu.:0.0000
##                                         Median :0.0000
##                                         Mean   :0.4331
##                                         3rd Qu.:1.0000
```

```
##                                       Max.   :1.0000
##
##  ResidualSugar_IMP_REDFLAG TotalSulfurDioxide_IMP_REDFLAG Density_IMP_REDFLAG
##  Min.   :0.00              Min.   :0.0000                 Min.   :0.0000
##  1st Qu.:0.00              1st Qu.:0.0000                 1st Qu.:0.0000
##  Median :1.00              Median :0.0000                 Median :1.0000
##  Mean   :0.56              Mean   :0.4624                 Mean   :0.5126
##  3rd Qu.:1.00              3rd Qu.:1.0000                 3rd Qu.:1.0000
##  Max.   :1.00              Max.   :1.0000                 Max.   :1.0000
##
##     TallyUp      Final_REDFLAG     TARGET_Flag       TARGET_AMT
##  Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:2.000
##  Median :2.000   Median :1.0000   Median :1.0000   Median :3.000
##  Mean   :1.968   Mean   :0.6642   Mean   :0.7863   Mean   :2.852
##  3rd Qu.:3.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000
##  Max.   :4.000   Max.   :1.0000   Max.   :1.0000   Max.   :7.000
##                                                    NA's   :2734
##  STARSxLabelAppeal_IMP STARSxAlcohol_IMP        ResidualSugar_bin
##  Min.   :-6.0000       Min.   :-12.00    Low Sweetness   :3194
##  1st Qu.:-1.0000       1st Qu.: 12.10    Medium Sweetness:6391
##  Median : 0.0000       Median : 19.80    High Sweetness  :3210
##  Mean   : 0.2543       Mean   : 21.36
##  3rd Qu.: 1.0000       3rd Qu.: 27.72
##  Max.   : 8.0000       Max.   :102.40
##
##           pH_IMP_bin
##  Low Acidity   :3455
##  Medium Acidity:6376
##  High Acidity  :2964
##
##
##
##
```

The same data preparation steps are performed on the testing dataset below

```
### Testing Data Set
# Create imputed variables for each column
wine_test$FixedAcidity_IMP = wine_test$FixedAcidity
wine_test$VolatileAcidity_IMP = wine_test$VolatileAcidity
wine_test$CitricAcid_IMP = wine_test$CitricAcid
wine_test$ResidualSugar_IMP = wine_test$ResidualSugar
wine_test$Chlorides_IMP = wine_test$Chlorides
wine_test$FreeSulfurDioxide_IMP = wine_test$FreeSulfurDioxide
wine_test$TotalSulfurDioxide_IMP = wine_test$TotalSulfurDioxide
wine_test$Density_IMP = wine_test$Density
wine_test$pH_IMP = wine_test$pH
wine_test$Sulphates_IMP = wine_test$Sulphates
wine_test$Alcohol_IMP = wine_test$Alcohol
wine_test$LabelAppeal_IMP = wine_test$LabelAppeal
wine_test$AcidIndex_IMP = wine_test$AcidIndex
wine_test$STARS_IMP = wine_test$STARS
```

```
### Replace missing NA values with median and outliers with 1st or 99th percentile of training data
### Create FLAG variables where replaced.

#ResidualSugar
wine_test$ResidualSugar_FLAG = as.factor(ifelse(wine_test$ResidualSugar_IMP < quantile(wine$ResidualSuga
                                        | wine_test$ResidualSugar_IMP > quantile(wine_test$ResidualSu
                                        | is.na(wine_test$ResidualSugar_IMP),1,0))
wine_test$ResidualSugar_IMP[is.na(wine_test$ResidualSugar_IMP)] = median(wine$ResidualSugar_IMP, na.rm=
wine_test$ResidualSugar_IMP = as.numeric(ifelse(wine_test$ResidualSugar_IMP < quantile(wine_test$Residua
                                        quantile(wine$ResidualSugar_IMP,c(.01)), wine_test$ResidualSu
wine_test$ResidualSugar_IMP = as.numeric(ifelse(wine_test$ResidualSugar_IMP > quantile(wine_test$Residua
                                        quantile(wine$ResidualSugar_IMP,c(.99)), wine_test$ResidualSu

#Chlorides
wine_test$Chlorides_FLAG = as.factor(ifelse(is.na(wine_test$Chlorides),1,0))
wine_test$Chlorides_IMP[is.na(wine_test$Chlorides_IMP)] = median(wine$Chlorides, na.rm = TRUE)

#FreeSulfurDioxide
wine_test$FreeSulfurDioxide_FLAG = as.factor(ifelse(wine_test$FreeSulfurDioxide_IMP < quantile(wine$Free
                                            | wine_test$FreeSulfurDioxide_IMP > quantile(wine_test$Fr
                                            | is.na(wine_test$FreeSulfurDioxide_IMP),1,0))
wine_test$FreeSulfurDioxide_IMP[is.na(wine_test$FreeSulfurDioxide_IMP)] = median(wine$FreeSulfurDioxide_
wine_test$FreeSulfurDioxide_IMP = as.numeric(ifelse(wine_test$FreeSulfurDioxide_IMP < quantile(wine_test
                                            quantile(wine$FreeSulfurDioxide_IMP,c(.01)), wine_test$Fr
wine_test$FreeSulfurDioxide_IMP = as.numeric(ifelse(wine_test$FreeSulfurDioxide_IMP > quantile(wine_test
                                            quantile(wine$FreeSulfurDioxide_IMP,c(.99)), wine_test$Fr

#TotalSulfurDioxide
wine_test$TotalSulfurDioxide_FLAG = as.factor(ifelse(wine_test$TotalSulfurDioxide_IMP < quantile(wine$To
                                             | wine_test$TotalSulfurDioxide_IMP > quantile(wine_testS
                                             | is.na(wine_test$TotalSulfurDioxide_IMP),1,0))
wine_test$TotalSulfurDioxide_IMP[is.na(wine_test$TotalSulfurDioxide_IMP)] = median(wine$TotalSulfurDioxi
wine_test$TotalSulfurDioxide_IMP = as.numeric(ifelse(wine_test$TotalSulfurDioxide_IMP < quantile(wine_te
                                             quantile(wine$TotalSulfurDioxide_IMP,c(.01)), wine_testS
wine_test$TotalSulfurDioxide_IMP = as.numeric(ifelse(wine_test$TotalSulfurDioxide_IMP > quantile(wine_te
                                             quantile(wine$TotalSulfurDioxide_IMP,c(.99)), wine_testS

#pH
wine_test$pH_FLAG = as.factor(ifelse(is.na(wine_test$pH),1,0))
wine_test$pH_IMP[is.na(wine_test$pH_IMP)] = median(wine$pH, na.rm = TRUE)

#Sulfates
wine_test$Sulphates_FLAG = as.factor(ifelse(is.na(wine_test$Sulphates),1,0))
wine_test$Sulphates_IMP[is.na(wine_test$Sulphates_IMP)] = median(wine$Sulphates, na.rm = TRUE)

#Alcohol
wine_test$Alcohol_FLAG = as.factor(ifelse(is.na(wine_test$Alcohol),1,0))
wine_test$Alcohol_IMP[is.na(wine_test$Alcohol_IMP)] = median(wine$Alcohol, na.rm = TRUE)

#Stars
wine_test$STARS_FLAG = as.factor(ifelse(is.na(wine_test$STARS),1,0))
wine_test$STARS_IMP = na.aggregate(wine_test$STARS_IMP, wine_test$LabelAppeal)
```

```
# Make new indicator that indicates red vs white based on volatile acidity
wine_test$VolatileAcidity_IMP_REDFLAG = ifelse(wine_test$VolatileAcidity_IMP > mean(wine_test$VolatileA
wine_test$ResidualSugar_IMP_REDFLAG = ifelse(wine_test$ResidualSugar_IMP < mean(wine_test$ResidualSugar
wine_test$TotalSulfurDioxide_IMP_REDFLAG = ifelse(wine_test$TotalSulfurDioxide_IMP < mean(wine_test$Tota
wine_test$Density_IMP_REDFLAG = ifelse(wine_test$Density_IMP > mean(wine_test$Density_IMP),1,0)
wine_test$TallyUp = wine_test$VolatileAcidity_IMP_REDFLAG + wine_test$ResidualSugar_IMP_REDFLAG + wine_
wine_test$Final_REDFLAG = ifelse(wine_test$TallyUp > mean(wine_test$TallyUp),1,0)

# Add Target Flag for 0 sale scenarios
wine_test$TARGET_Flag = ifelse(wine_test$TARGET > 0,1,0)
wine_test$TARGET_AMT = wine_test$TARGET - 1
wine_test$TARGET_AMT = ifelse(wine_test$TARGET_Flag == 0,NA,wine_test$TARGET-1)

# Create interaction terms and imputed variables from input variables
wine_test$STARSxLabelAppeal_IMP = wine_test$STARS_IMP * wine_test$LabelAppeal_IMP
wine_test$STARSxAlcohol_IMP = wine_test$STARS_IMP * wine_test$Alcohol_IMP

# Create bins for 3 levels of sweetness from the ResidualSugar
wine_test$ResidualSugar_bin[wine_test$ResidualSugar_IMP < quantile(wine$ResidualSugar_IMP, c(.25))] = "
wine_test$ResidualSugar_bin[wine_test$ResidualSugar_IMP >= quantile(wine$ResidualSugar_IMP, c(.25)) &
                            wine_test$ResidualSugar_IMP < quantile(wine$ResidualSugar_IMP, c(.75))] = "Medi
wine_test$ResidualSugar_bin[wine_test$ResidualSugar_IMP >= quantile(wine$ResidualSugar_IMP, c(.75))] =
wine_test$ResidualSugar_bin = factor(wine_test$ResidualSugar_bin, levels = c("Low Sweetness", "Medium Su

# Create bins for acidity based on pH level
wine_test$pH_IMP_bin[wine_test$pH_IMP <= 3] = "Low Acidity"
wine_test$pH_IMP_bin[wine_test$pH_IMP > 3 & wine_test$pH_IMP < 3.5] = "Medium Acidity"
wine_test$pH_IMP_bin[wine_test$pH_IMP >= 3.5] = "High Acidity"
wine_test$pH_IMP_bin = factor(wine_test$pH_IMP_bin, levels = c("Low Acidity", "Medium Acidity", "High A

# Confirm there are no missing or empty values remaining in the dataset.
#summary(wine_test)
#missmap(wine_test)
```

Part 4: Model Development

This section of the model is used to define and create 7 machine learning models. The models developed in this section are included below. This section begins with 2 linear regression models that are simple in nature to establish a baseline understanding of the data and relationships. The next five models compare to common Counting regression models, the Poisson and Negative Binomial regressions. These two models are then developed using the zero-inflated approach to control for the high volume of zeros in the dataset. Finally, a hurdle model is used to understand the probabilities of zero and non-zero target values. 1. Linear Regression 2. Stepwise Linear Regression 3. Poisson Regression 4. Negative Binomial Regression 5. Zero-Inflated Poisson (ZIP) Regression 6. Zero-Inflated Negative-Binomial (ZINB) Regression 7. Hurdle Model Regression

```
### Model 1: Linear Regression
lm_fit = lm(TARGET~ ResidualSugar_bin + VolatileAcidity_IMP + CitricAcid_IMP +
              Chlorides_IMP + FreeSulfurDioxide_IMP + TotalSulfurDioxide_IMP + STARSxAlcohol_IMP +
              pH_IMP_bin + Sulphates_IMP + Alcohol_IMP + LabelAppeal_IMP + AcidIndex_IMP + STARS_IMP +
              Final_REDFLAG + STARSxLabelAppeal_IMP + ResidualSugar_FLAG + Chlorides_FLAG + FreeSulfurDi
              TotalSulfurDioxide_FLAG + pH_FLAG + Sulphates_FLAG + Alcohol_FLAG + STARS_FLAG
             ,data = wine)

summary(lm_fit)
```

```
##
## Call:
## lm(formula = TARGET ~ ResidualSugar_bin + VolatileAcidity_IMP +
##     CitricAcid_IMP + Chlorides_IMP + FreeSulfurDioxide_IMP +
##     TotalSulfurDioxide_IMP + STARSxAlcohol_IMP + pH_IMP_bin +
##     Sulphates_IMP + Alcohol_IMP + LabelAppeal_IMP + AcidIndex_IMP +
##     STARS_IMP + Final_REDFLAG + STARSxLabelAppeal_IMP + ResidualSugar_FLAG +
##     Chlorides_FLAG + FreeSulfurDioxide_FLAG + TotalSulfurDioxide_FLAG +
##     pH_FLAG + Sulphates_FLAG + Alcohol_FLAG + STARS_FLAG, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4460 -0.8155  0.0451  0.8631  5.7238
##
## Coefficients:
##                                    Estimate  Std. Error t value Pr(>|t|)
## (Intercept)                         3.98218019  0.12873952  30.932  < 2e-16 ***
## ResidualSugar_binMedium Sweetness -0.02726432  0.02943432  -0.926 0.354320
## ResidualSugar_binHigh Sweetness   -0.07521594  0.03526108  -2.133 0.032934 *
## VolatileAcidity_IMP               -0.06105497  0.01581099  -3.862 0.000113 ***
## CitricAcid_IMP                     0.01903113  0.01367130   1.392 0.163932
## Chlorides_IMP                     -0.11989697  0.03789980  -3.164 0.001562 **
## FreeSulfurDioxide_IMP              0.00028816  0.00008410   3.426 0.000613 ***
## TotalSulfurDioxide_IMP             0.00009757  0.00005723   1.705 0.088256 .
## STARSxAlcohol_IMP                  0.00673806  0.00408905   1.648 0.099413 .
## pH_IMP_binMedium Acidity           0.00184094  0.02842110   0.065 0.948355
## pH_IMP_binHigh Acidity            -0.10630087  0.03337698  -3.185 0.001452 **
## Sulphates_IMP                     -0.03034096  0.01326037  -2.288 0.022148 *
## Alcohol_IMP                       -0.00041211  0.00879428  -0.047 0.962625
## LabelAppeal_IMP                    0.22487357  0.03590737   6.263 3.91e-10 ***
## AcidIndex_IMP                     -0.20234578  0.00915313 -22.107  < 2e-16 ***
## STARS_IMP                          0.61675795  0.04680790  13.176  < 2e-16 ***
## Final_REDFLAG                     -0.21432265  0.02939369  -7.291 3.25e-13 ***
## STARSxLabelAppeal_IMP              0.09859538  0.01677445   5.878 4.26e-09 ***
## ResidualSugar_FLAG1                0.06989816  0.04742235   1.474 0.140520
## Chlorides_FLAG1                   -0.00228989  0.05404092  -0.042 0.966202
## FreeSulfurDioxide_FLAG1            0.03926428  0.04628011   0.848 0.396228
## TotalSulfurDioxide_FLAG1           0.05701726  0.04547789   1.254 0.209961
## pH_FLAG1                          -0.12281605  0.06906836  -1.778 0.075398 .
## Sulphates_FLAG1                   -0.03607084  0.04017539  -0.898 0.369291
## Alcohol_FLAG1                      0.06940938  0.05340002   1.300 0.193693
## STARS_FLAG1                       -2.22459901  0.02737777 -81.256  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.329 on 12769 degrees of freedom
## Multiple R-squared:  0.5253, Adjusted R-squared:  0.5243
## F-statistic: 565.1 on 25 and 12769 DF,  p-value: < 2.2e-16
```

```
wine$lm_fit = fitted(lm_fit)

### Model 2: Stepwise Linear Regression
stepwise_lm = stepAIC(lm_fit, direction="both", trace = 0)
stepwise_lm$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## TARGET ~ ResidualSugar_bin + VolatileAcidity_IMP + CitricAcid_IMP +
##     Chlorides_IMP + FreeSulfurDioxide_IMP + TotalSulfurDioxide_IMP +
##     STARSxAlcohol_IMP + pH_IMP_bin + Sulphates_IMP + Alcohol_IMP +
##     LabelAppeal_IMP + AcidIndex_IMP + STARS_IMP + Final_REDFLAG +
##     STARSxLabelAppeal_IMP + ResidualSugar_FLAG + Chlorides_FLAG +
##     FreeSulfurDioxide_FLAG + TotalSulfurDioxide_FLAG + pH_FLAG +
##     Sulphates_FLAG + Alcohol_FLAG + STARS_FLAG
##
## Final Model:
## TARGET ~ ResidualSugar_bin + VolatileAcidity_IMP + Chlorides_IMP +
##     FreeSulfurDioxide_IMP + TotalSulfurDioxide_IMP + STARSxAlcohol_IMP +
##     pH_IMP_bin + Sulphates_IMP + LabelAppeal_IMP + AcidIndex_IMP +
##     STARS_IMP + Final_REDFLAG + STARSxLabelAppeal_IMP + ResidualSugar_FLAG +
##     pH_FLAG + STARS_FLAG
##
##
##                         Step Df    Deviance Resid. Df Resid. Dev      AIC
## 1                                             12769   22539.45 7296.689
## 2          - Chlorides_FLAG  1 0.003169361     12770   22539.45 7294.691
## 3             - Alcohol_IMP  1 0.003900928     12771   22539.45 7292.693
## 4  - FreeSulfurDioxide_FLAG  1 1.270147221     12772   22540.72 7291.414
## 5          - Sulphates_FLAG  1 1.437308028     12773   22542.16 7290.230
## 6 - TotalSulfurDioxide_FLAG  1 2.833033820     12774   22544.99 7289.838
## 7            - Alcohol_FLAG  1 2.939386842     12775   22547.93 7289.506
## 8          - CitricAcid_IMP  1 3.369018364     12776   22551.30 7289.418
```

```
lm_fit_stepwise = lm(TARGET~ ResidualSugar_bin + VolatileAcidity_IMP +
                       Chlorides_IMP + FreeSulfurDioxide_IMP + TotalSulfurDioxide_IMP +
                       pH_IMP_bin + Sulphates_IMP + STARSxAlcohol_IMP + LabelAppeal_IMP + AcidIndex_IMP
                       Final_REDFLAG + STARSxLabelAppeal_IMP + ResidualSugar_FLAG
                       + pH_FLAG + STARS_FLAG
                       , data=wine)

summary(lm_fit_stepwise)
```

```
##
## Call:
## lm(formula = TARGET ~ ResidualSugar_bin + VolatileAcidity_IMP +
##     Chlorides_IMP + FreeSulfurDioxide_IMP + TotalSulfurDioxide_IMP +
##     pH_IMP_bin + Sulphates_IMP + STARSxAlcohol_IMP + LabelAppeal_IMP +
##     AcidIndex_IMP + STARS_IMP + Final_REDFLAG + STARSxLabelAppeal_IMP +
##     ResidualSugar_FLAG + pH_FLAG + STARS_FLAG, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4571 -0.8143  0.0454  0.8657  5.7136
##
## Coefficients:
##                                    Estimate  Std. Error t value Pr(>|t|)
## (Intercept)                      3.98577873  0.08855392  45.010  < 2e-16 ***
## ResidualSugar_binMedium Sweetness -0.02880802  0.02942092  -0.979 0.327516
```

```
## ResidualSugar_binHigh Sweetness   -0.07727849  0.03522758  -2.194 0.028275 *
## VolatileAcidity_IMP               -0.06141665  0.01580268  -3.886 0.000102 ***
## Chlorides_IMP                     -0.12068437  0.03789021  -3.185 0.001450 **
## FreeSulfurDioxide_IMP              0.00028991  0.00008409   3.448 0.000568 ***
## TotalSulfurDioxide_IMP             0.00009640  0.00005720   1.685 0.091984 .
## pH_IMP_binMedium Acidity           0.00171453  0.02841968   0.060 0.951895
## pH_IMP_binHigh Acidity            -0.10646204  0.03336980  -3.190 0.001424 **
## Sulphates_IMP                     -0.03019439  0.01325631  -2.278 0.022759 *
## STARSxAlcohol_IMP                  0.00657543  0.00150894   4.358 1.32e-05 ***
## LabelAppeal_IMP                    0.22475081  0.03588921   6.262 3.91e-10 ***
## AcidIndex_IMP                     -0.20135683  0.00912930 -22.056  < 2e-16 ***
## STARS_IMP                          0.61866992  0.02366659  26.141  < 2e-16 ***
## Final_REDFLAG                     -0.21697817  0.02931883  -7.401 1.44e-13 ***
## STARSxLabelAppeal_IMP              0.09861045  0.01676547   5.882 4.16e-09 ***
## ResidualSugar_FLAG1                0.07115846  0.04740279   1.501 0.133343
## pH_FLAG1                          -0.12169808  0.06904870  -1.762 0.078009 .
## STARS_FLAG1                       -2.22484847  0.02736292 -81.309  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.329 on 12776 degrees of freedom
## Multiple R-squared:  0.525,  Adjusted R-squared:  0.5243
## F-statistic: 784.5 on 18 and 12776 DF,  p-value: < 2.2e-16
```

```
wine$fittedLMStepwise = fitted(lm_fit_stepwise)

### Model 3: Poisson Regression

poisson_model = glm(TARGET ~ ResidualSugar_bin + VolatileAcidity_IMP +
                      Chlorides_IMP + FreeSulfurDioxide_IMP + TotalSulfurDioxide_IMP +
                      pH_IMP_bin + Sulphates_IMP + Alcohol_IMP + LabelAppeal_IMP + AcidIndex_IMP + STARS
                      Final_REDFLAG + STARSxLabelAppeal_IMP + STARS_FLAG,
                      family="poisson"(link="log"), data=wine)

summary(poisson_model)
```

```
##
## Call:
## glm(formula = TARGET ~ ResidualSugar_bin + VolatileAcidity_IMP +
##     Chlorides_IMP + FreeSulfurDioxide_IMP + TotalSulfurDioxide_IMP +
##     pH_IMP_bin + Sulphates_IMP + Alcohol_IMP + LabelAppeal_IMP +
##     AcidIndex_IMP + STARS_IMP + Final_REDFLAG + STARSxLabelAppeal_IMP +
##     STARS_FLAG, family = poisson(link = "log"), data = wine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1178  -0.6496   0.0116   0.4476   3.5418
##
## Coefficients:
##                                    Estimate  Std. Error z value     Pr(>|z|)
## (Intercept)                       1.53072369  0.04425504  34.589      < 2e-16
## ResidualSugar_binMedium Sweetness -0.01071785  0.01273220  -0.842      0.39991
## ResidualSugar_binHigh Sweetness   -0.02604795  0.01531433  -1.701      0.08896
## VolatileAcidity_IMP               -0.02012329  0.00686547  -2.931      0.00338
## Chlorides_IMP                     -0.03774491  0.01647989  -2.290      0.02200
```

```
## FreeSulfurDioxide_IMP              0.00009767  0.00003637   2.685      0.00724
## TotalSulfurDioxide_IMP             0.00003764  0.00002493   1.509      0.13117
## pH_IMP_binMedium Acidity          -0.00220061  0.01211741  -0.182      0.85589
## pH_IMP_binHigh Acidity            -0.03722314  0.01455075  -2.558      0.01052
## Sulphates_IMP                     -0.01146160  0.00575172  -1.993      0.04629
## Alcohol_IMP                        0.00376295  0.00140739   2.674      0.00750
## LabelAppeal_IMP                    0.19873321  0.01544639  12.866      < 2e-16
## AcidIndex_IMP                     -0.08033430  0.00452594 -17.750      < 2e-16
## STARS_IMP                          0.18551534  0.00645722  28.730      < 2e-16
## Final_REDFLAG                     -0.06853349  0.01235085  -5.549 0.0000000287
## STARSxLabelAppeal_IMP             -0.02026557  0.00655782  -3.090      0.00200
## STARS_FLAG1                       -1.02517806  0.01696271 -60.437      < 2e-16
##
## (Intercept)                       ***
## ResidualSugar_binMedium Sweetness
## ResidualSugar_binHigh Sweetness   .
## VolatileAcidity_IMP               **
## Chlorides_IMP                     *
## FreeSulfurDioxide_IMP             **
## TotalSulfurDioxide_IMP
## pH_IMP_binMedium Acidity
## pH_IMP_binHigh Acidity            *
## Sulphates_IMP                     *
## Alcohol_IMP                       **
## LabelAppeal_IMP                   ***
## AcidIndex_IMP                     ***
## STARS_IMP                         ***
## Final_REDFLAG                     ***
## STARSxLabelAppeal_IMP             **
## STARS_FLAG1                       ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 13811  on 12778  degrees of freedom
## AIC: 45787
##
## Number of Fisher Scoring iterations: 6
```

```r
wine$poisson_fit = predict(poisson_model, newdata = wine, type = "response")

### Model 4: Negative Binomial Regression

NBR_Model = glm.nb(TARGET ~ ResidualSugar_bin + VolatileAcidity_IMP +
                     Chlorides_IMP + FreeSulfurDioxide_IMP + TotalSulfurDioxide_IMP +
                     pH_IMP_bin + Sulphates_IMP + STARSxAlcohol_IMP + LabelAppeal_IMP + AcidIndex_IMP +
                     Final_REDFLAG + STARSxLabelAppeal_IMP + ResidualSugar_FLAG +
                     pH_FLAG + STARS_FLAG, data=wine)

summary(NBR_Model)
```

```
##
## Call:
```

```
## glm.nb(formula = TARGET ~ ResidualSugar_bin + VolatileAcidity_IMP +
##     Chlorides_IMP + FreeSulfurDioxide_IMP + TotalSulfurDioxide_IMP +
##     pH_IMP_bin + Sulphates_IMP + STARSxAlcohol_IMP + LabelAppeal_IMP +
##     AcidIndex_IMP + STARS_IMP + Final_REDFLAG + STARSxLabelAppeal_IMP +
##     ResidualSugar_FLAG + pH_FLAG + STARS_FLAG, data = wine, init.theta = 40188.8343,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1112  -0.6520   0.0118   0.4486   3.5405
##
## Coefficients:
##                                    Estimate  Std. Error z value     Pr(>|z|)
## (Intercept)                      1.57140310  0.04137663  37.978      < 2e-16
## ResidualSugar_binMedium Sweetness -0.01249020  0.01281378  -0.975      0.32969
## ResidualSugar_binHigh Sweetness   -0.02637148  0.01532495  -1.721      0.08528
## VolatileAcidity_IMP               -0.02000211  0.00686779  -2.912      0.00359
## Chlorides_IMP                     -0.03761472  0.01648384  -2.282      0.02249
## FreeSulfurDioxide_IMP              0.00009671  0.00003637   2.659      0.00784
## TotalSulfurDioxide_IMP             0.00003620  0.00002495   1.451      0.14672
## pH_IMP_binMedium Acidity           0.00057046  0.01224874   0.047      0.96285
## pH_IMP_binHigh Acidity            -0.03749361  0.01455217  -2.576      0.00998
## Sulphates_IMP                     -0.01154786  0.00575228  -2.008      0.04469
## STARSxAlcohol_IMP                  0.00141439  0.00059404   2.381      0.01727
## LabelAppeal_IMP                    0.19863497  0.01544565  12.860      < 2e-16
## AcidIndex_IMP                     -0.08034457  0.00452741 -17.746      < 2e-16
## STARS_IMP                          0.17029570  0.00935361  18.206      < 2e-16
## Final_REDFLAG                     -0.07017179  0.01241212  -5.653 0.0000000157
## STARSxLabelAppeal_IMP             -0.02018184  0.00655790  -3.077      0.00209
## ResidualSugar_FLAG1                0.02598634  0.02035244   1.277      0.20167
## pH_FLAG1                          -0.04870917  0.03036035  -1.604      0.10863
## STARS_FLAG1                       -1.02491926  0.01696444 -60.416      < 2e-16
##
## (Intercept)                       ***
## ResidualSugar_binMedium Sweetness
## ResidualSugar_binHigh Sweetness   .
## VolatileAcidity_IMP               **
## Chlorides_IMP                     *
## FreeSulfurDioxide_IMP             **
## TotalSulfurDioxide_IMP
## pH_IMP_binMedium Acidity
## pH_IMP_binHigh Acidity            **
## Sulphates_IMP                     *
## STARSxAlcohol_IMP                 *
## LabelAppeal_IMP                   ***
## AcidIndex_IMP                     ***
## STARS_IMP                         ***
## Final_REDFLAG                     ***
## STARSxLabelAppeal_IMP             **
## ResidualSugar_FLAG1
## pH_FLAG1
## STARS_FLAG1                       ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for Negative Binomial(40188.83) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 13808  on 12776  degrees of freedom
## AIC: 45791
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  40189
##           Std. Err.:  34067
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -45750.73
```

```
wine$NBR_fit = predict(NBR_Model, newdata = wine, type = "response")

### Model 5: Zero-Inflated Poisson (ZIP) Regression

ZIP_Model = zeroinfl(TARGET ~ VolatileAcidity_IMP +
                     FreeSulfurDioxide_IMP + TotalSulfurDioxide_IMP +
                     pH_IMP_bin + Sulphates_IMP + STARSxAlcohol_IMP + LabelAppeal_IMP + AcidIndex_IMP +
                     Final_REDFLAG + STARSxLabelAppeal_IMP + STARS_IMP +
                     pH_FLAG + STARS_FLAG, data=wine)

summary(ZIP_Model)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ VolatileAcidity_IMP + FreeSulfurDioxide_IMP +
##     TotalSulfurDioxide_IMP + pH_IMP_bin + Sulphates_IMP + STARSxAlcohol_IMP +
##     LabelAppeal_IMP + AcidIndex_IMP + Final_REDFLAG + STARSxLabelAppeal_IMP +
##     STARS_IMP + pH_FLAG + STARS_FLAG, data = wine)
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.148109 -0.410854  0.008162  0.393418  5.914926
##
## Count model coefficients (poisson with log link):
##                             Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)               1.21908307  0.04230044  28.820   < 2e-16 ***
## VolatileAcidity_IMP      -0.01126922  0.00701048  -1.607     0.108
## FreeSulfurDioxide_IMP     0.00002284  0.00003678   0.621     0.535
## TotalSulfurDioxide_IMP   -0.00003028  0.00002480  -1.221     0.222
## pH_IMP_binMedium Acidity  0.00389382  0.01251560   0.311     0.756
## pH_IMP_binHigh Acidity    0.00209152  0.01486810   0.141     0.888
## Sulphates_IMP             0.00055485  0.00590163   0.094     0.925
## STARSxAlcohol_IMP         0.00253658  0.00059996   4.228 0.0000236 ***
## LabelAppeal_IMP           0.35760390  0.01629121  21.951   < 2e-16 ***
## AcidIndex_IMP            -0.02072806  0.00484586  -4.277 0.0000189 ***
## Final_REDFLAG            -0.01847337  0.01177551  -1.569     0.117
## STARSxLabelAppeal_IMP    -0.05927835  0.00685795  -8.644   < 2e-16 ***
## STARS_IMP                 0.10678586  0.00969506  11.014   < 2e-16 ***
## pH_FLAG1                 -0.00435409  0.03103851  -0.140     0.888
```

```
## STARS_FLAG1              -0.15766130  0.01847795  -8.532   < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -2.3377632  0.3449746  -6.777 1.23e-11 ***
## VolatileAcidity_IMP        0.0973003  0.0456575   2.131  0.03308 *
## FreeSulfurDioxide_IMP     -0.0007780  0.0002510  -3.100  0.00194 **
## TotalSulfurDioxide_IMP    -0.0006633  0.0001670  -3.973 7.10e-05 ***
## pH_IMP_binMedium Acidity   0.0169091  0.0842844   0.201  0.84100
## pH_IMP_binHigh Acidity     0.4149675  0.0967765   4.288 1.80e-05 ***
## Sulphates_IMP              0.1246948  0.0388785   3.207  0.00134 **
## STARSxAlcohol_IMP          0.0124130  0.0053535   2.319  0.02041 *
## LabelAppeal_IMP            0.7550921  0.1684452   4.483 7.37e-06 ***
## AcidIndex_IMP              0.4116854  0.0259826  15.845  < 2e-16 ***
## Final_REDFLAG              0.5205022  0.0859240   6.058 1.38e-09 ***
## STARSxLabelAppeal_IMP      0.4986693  0.1134071   4.397 1.10e-05 ***
## STARS_IMP                 -3.5311720  0.2403284 -14.693  < 2e-16 ***
## pH_FLAG1                   0.4650572  0.1983101   2.345  0.01902 *
## STARS_FLAG1                5.7406009  0.2268701  25.303  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 37
## Log-likelihood: -2.037e+04 on 30 Df
```

```
wine$ZIP_fit = predict(ZIP_Model, newdata = wine, type = "response")

### Model 6: Zero-Inflated Negative Binomial (ZINB) Regression

ZINB_Model = zeroinfl(TARGET ~ VolatileAcidity_IMP + Alcohol_IMP +
                        FreeSulfurDioxide_IMP + TotalSulfurDioxide_IMP +
                        pH_IMP_bin + Sulphates_IMP + STARSxAlcohol_IMP + LabelAppeal_IMP + AcidIndex_IM
                        Final_REDFLAG + STARSxLabelAppeal_IMP + STARS_IMP +
                        pH_FLAG + STARS_FLAG, data=wine, dist = "negbin")

summary(ZINB_Model)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ VolatileAcidity_IMP + Alcohol_IMP + FreeSulfurDioxide_IMP +
##     TotalSulfurDioxide_IMP + pH_IMP_bin + Sulphates_IMP + STARSxAlcohol_IMP +
##     LabelAppeal_IMP + AcidIndex_IMP + Final_REDFLAG + STARSxLabelAppeal_IMP +
##     STARS_IMP + pH_FLAG + STARS_FLAG, data = wine, dist = "negbin")
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.147608 -0.413902  0.001158  0.390821  5.969478
##
## Count model coefficients (negbin with log link):
##                             Estimate  Std. Error z value Pr(>|z|)
## (Intercept)               1.11001293  0.05951305  18.652  < 2e-16 ***
## VolatileAcidity_IMP      -0.01153524  0.00701060  -1.645  0.09989 .
## Alcohol_IMP               0.01030699  0.00392804   2.624  0.00869 **
## FreeSulfurDioxide_IMP     0.00002239  0.00003679   0.609  0.54270
## TotalSulfurDioxide_IMP   -0.00002930  0.00002480  -1.181  0.23748
```

```
## pH_IMP_binMedium Acidity  0.00417612  0.01251523   0.334  0.73862
## pH_IMP_binHigh Acidity    0.00227196  0.01486768   0.153  0.87855
## Sulphates_IMP             0.00049107  0.00590198   0.083  0.93369
## STARSxAlcohol_IMP        -0.00144628  0.00164018  -0.882  0.37789
## LabelAppeal_IMP           0.35792892  0.01630073  21.958  < 2e-16 ***
## AcidIndex_IMP            -0.02071985  0.00484553  -4.276 1.90e-05 ***
## Final_REDFLAG            -0.01847686  0.01177549  -1.569  0.11663
## STARSxLabelAppeal_IMP    -0.05929674  0.00686224  -8.641  < 2e-16 ***
## STARS_IMP                 0.14922500  0.01894618   7.876 3.37e-15 ***
## pH_FLAG1                 -0.00626041  0.03105603  -0.202  0.84024
## STARS_FLAG1              -0.15757018  0.01847490  -8.529  < 2e-16 ***
## Log(theta)               17.22969313  1.35531387  12.713  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                            Estimate Std. Error z value      Pr(>|z|)
## (Intercept)              -2.9618674  0.5130921  -5.773 0.00000000781 ***
## VolatileAcidity_IMP       0.0965650  0.0456335   2.116       0.03434 *
## Alcohol_IMP               0.0591999  0.0351980   1.682       0.09259 .
## FreeSulfurDioxide_IMP    -0.0007914  0.0002513  -3.149       0.00164 **
## TotalSulfurDioxide_IMP   -0.0006585  0.0001670  -3.943 0.00008053560 ***
## pH_IMP_binMedium Acidity  0.0194917  0.0843258   0.231       0.81720
## pH_IMP_binHigh Acidity    0.4191711  0.0968765   4.327 0.00001512485 ***
## Sulphates_IMP             0.1232320  0.0389208   3.166       0.00154 **
## STARSxAlcohol_IMP        -0.0183014  0.0195435  -0.936       0.34904
## LabelAppeal_IMP           0.7634888  0.1689841   4.518 0.00000623937 ***
## AcidIndex_IMP             0.4114553  0.0259799  15.837       < 2e-16 ***
## Final_REDFLAG             0.5198706  0.0859458   6.049 0.00000000146 ***
## STARSxLabelAppeal_IMP     0.4929433  0.1136050   4.339 0.00001430682 ***
## STARS_IMP                -3.2048560  0.3129119 -10.242       < 2e-16 ***
## pH_FLAG1                  0.4522132  0.1985006   2.278       0.02272 *
## STARS_FLAG1               5.7367875  0.2271547  25.255       < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 30392096.0417
## Number of iterations in BFGS optimization: 39
## Log-likelihood: -2.036e+04 on 33 Df
```

```
wine$ZINB_fit = predict(ZINB_Model, newdata = wine, type = "response")

### Model 7: Hurdle Model Regression
hurdle_model = hurdle(TARGET ~ VolatileAcidity_IMP +
                        FreeSulfurDioxide_IMP + TotalSulfurDioxide_IMP +
                        pH_IMP_bin + Sulphates_IMP + STARSxAlcohol_IMP + LabelAppeal_IMP + AcidIndex_IM
                        Final_REDFLAG + STARSxLabelAppeal_IMP + STARS_IMP +
                        pH_FLAG + STARS_FLAG, data = wine)
summary(hurdle_model)
```

```
##
## Call:
## hurdle(formula = TARGET ~ VolatileAcidity_IMP + FreeSulfurDioxide_IMP +
##     TotalSulfurDioxide_IMP + pH_IMP_bin + Sulphates_IMP + STARSxAlcohol_IMP +
##     LabelAppeal_IMP + AcidIndex_IMP + Final_REDFLAG + STARSxLabelAppeal_IMP +
##     STARS_IMP + pH_FLAG + STARS_FLAG, data = wine)
##
```

```
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.089872 -0.423370 -0.003424  0.398538  5.098284
##
## Count model coefficients (truncated poisson with log link):
##                             Estimate  Std. Error z value Pr(>|z|)
## (Intercept)               1.19268733  0.04348042  27.430  < 2e-16 ***
## VolatileAcidity_IMP      -0.00939733  0.00723421  -1.299 0.193940
## FreeSulfurDioxide_IMP     0.00002279  0.00003788   0.602 0.547352
## TotalSulfurDioxide_IMP   -0.00004222  0.00002557  -1.651 0.098660 .
## pH_IMP_binMedium Acidity  0.00489687  0.01291374   0.379 0.704541
## pH_IMP_binHigh Acidity    0.00719517  0.01532599   0.469 0.638730
## Sulphates_IMP             0.00081559  0.00607896   0.134 0.893271
## STARSxAlcohol_IMP         0.00271554  0.00061630   4.406 1.05e-05 ***
## LabelAppeal_IMP           0.36471717  0.01717865  21.231  < 2e-16 ***
## AcidIndex_IMP            -0.01669935  0.00494715  -3.376 0.000737 ***
## Final_REDFLAG            -0.01646618  0.01215201  -1.355 0.175412
## STARSxLabelAppeal_IMP    -0.05821111  0.00717252  -8.116 4.82e-16 ***
## STARS_IMP                 0.09775213  0.00998011   9.795  < 2e-16 ***
## pH_FLAG1                 -0.00172514  0.03189746  -0.054 0.956868
## STARS_FLAG1              -0.15681422  0.01873496  -8.370  < 2e-16 ***
## Zero hurdle model coefficients (binomial with logit link):
##                             Estimate Std. Error z value       Pr(>|z|)
## (Intercept)               2.5291440  0.2284704  11.070        < 2e-16 ***
## VolatileAcidity_IMP      -0.1112128  0.0378899  -2.935        0.00333 **
## FreeSulfurDioxide_IMP     0.0006083  0.0002046   2.973        0.00295 **
## TotalSulfurDioxide_IMP    0.0006022  0.0001375   4.381 0.000011822377 ***
## pH_IMP_binMedium Acidity -0.0184632  0.0693324  -0.266        0.79001
## pH_IMP_binHigh Acidity   -0.3663723  0.0798080  -4.591 0.000004418225 ***
## Sulphates_IMP            -0.0998439  0.0320144  -3.119        0.00182 **
## STARSxAlcohol_IMP        -0.0095752  0.0045537  -2.103        0.03549 *
## LabelAppeal_IMP          -0.2085871  0.1066407  -1.956        0.05047 .
## AcidIndex_IMP            -0.3749952  0.0211447 -17.735        < 2e-16 ***
## Final_REDFLAG            -0.4495899  0.0702683  -6.398 0.000000000157 ***
## STARSxLabelAppeal_IMP    -0.4276271  0.0688038  -6.215 0.000000000513 ***
## STARS_IMP                 2.4132104  0.1192782  20.232        < 2e-16 ***
## pH_FLAG1                 -0.3883279  0.1640528  -2.367        0.01793 *
## STARS_FLAG1              -4.1465848  0.1010847 -41.021        < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 22
## Log-likelihood: -2.038e+04 on 30 Df
```

```
wine$hurdle_fit = predict(hurdle_model, newdata = wine, type = "response")
```

Part 5: Model Evaluation

This section of the model is used to evaluate and compare the results of the ML regression models above. The models are compared using the following metrics:

1. Root Mean Squared Error (RMSE)
2. Mean Squared Error (MSE)
3. Akaike's Information Criteria (AIC)
4. Bayesian Information Criteria (BIC)

Additionally, the Vuong test is applied to compare the Poisson and Negative Binomial against their zero-inflated counterparts in order to establish statistical significance.

```
# Define function for MSE calculation
mse = function(sm)
  mean(sm$residuals^2)

# Define function for RMSE calculation
rmse = function(pred, target)
  sqrt(mean((pred-target)**2))

# AIC
lm_fit_AIC = AIC(lm_fit)
lm_fit_stepwise_AIC = AIC(lm_fit_stepwise)
poisson_model_AIC = AIC(poisson_model)
NBR_Model_AIC = AIC(NBR_Model)
ZIP_Model_AIC = AIC(ZIP_Model)
ZINB_Model_AIC = AIC(ZINB_Model)
hurdle_model_AIC = AIC(hurdle_model)

# BIC
lm_fit_BIC = BIC(lm_fit)
lm_fit_stepwise_BIC = BIC(lm_fit_stepwise)
poisson_model_BIC = BIC(poisson_model)
NBR_Model_BIC = BIC(NBR_Model)
ZIP_Model_BIC = BIC(ZIP_Model)
ZINB_Model_BIC = BIC(ZINB_Model)
hurdle_model_BIC = BIC(hurdle_model)

# MSE
lm_fit_mse = mse(lm_fit)
lm_fit_stepwise_mse = mse(lm_fit_stepwise)
poisson_model_mse = mse(poisson_model)
NBR_Model_mse = mse(NBR_Model)
ZIP_Model_mse = mse(ZIP_Model)
ZINB_Model_mse = mse(ZINB_Model)
hurdle_model_mse = mse(hurdle_model)

# RMSE
lm_fit_rmse = rmse(wine$lm_fit, wine$TARGET)
lm_fit_stepwise_rmse = rmse(wine$fittedLMStepwise, wine$TARGET)
poisson_model_rmse = rmse(wine$poisson_fit, wine$TARGET)
NBR_Model_rmse = rmse(wine$NBR_fit, wine$TARGET)
ZIP_Model_rmse = rmse(wine$ZIP_fit, wine$TARGET)
ZINB_Model_rmse = rmse(wine$ZINB_fit, wine$TARGET)
hurdle_model_rmse = rmse(wine$hurdle_fit, wine$TARGET)
```

```
# Create table to display model evaluation results
data = matrix(c(lm_fit_AIC, lm_fit_stepwise_AIC, poisson_model_AIC, NBR_Model_AIC, ZIP_Model_AIC, ZINB_

# Add row and column names for results output
colnames(data) = c('LRM', 'LRM Stepwise','Poisson','Neg Binom', 'ZIP', 'ZINB', 'Hurdle')
rownames(data) = c('AIC', 'BIC', 'MSE', 'RMSE')
output = as.table(data)
```

```
print('The results of all models are compared in the table below: ')
```

```
## [1] "The results of all models are compared in the table below: "
```

```
print(output)
```

```
##                LRM  LRM Stepwise       Poisson     Neg Binom           ZIP
## AIC  43609.3262721 43602.0547948 45786.9884318 45790.7260926 40793.2032168
## BIC  43810.6601353 43751.1909898 45913.7541975 45939.8622875 41016.9075093
## MSE      1.7615824     1.7625091     0.5807014     0.5806571     1.5863207
## RMSE     1.3272462     1.3275952     1.3166069     1.3167305     1.2594922
##               ZINB        Hurdle
## AIC  40790.3957387 40823.8037585
## BIC  41036.4704604 41047.5080509
## MSE      1.5843981     1.5919465
## RMSE     1.2587288     1.2617236
```

```
# Perform Vuong test to compare 2 poisson models
vuong(poisson_model, ZIP_Model)
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A    p-value
## Raw                   -42.75778 model2 > model1 < 2.22e-16
## AIC-corrected         -42.53632 model2 > model1 < 2.22e-16
## BIC-corrected         -41.71061 model2 > model1 < 2.22e-16
```

```
# Perform Vuong test to compare 2 negative binomial models
vuong(NBR_Model, ZINB_Model)
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##                Vuong z-statistic             H_A    p-value
## Raw                   -42.86864 model2 > model1 < 2.22e-16
## AIC-corrected         -42.64689 model2 > model1 < 2.22e-16
## BIC-corrected         -41.82012 model2 > model1 < 2.22e-16
```

```
# A summary of the 3 best-performing models that capture overdispered and zero-adjusted results
summary(wine$ZIP_fit)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.006222 1.816144 2.984052 3.026069 4.179592 7.284493
```

```
summary(wine$ZINB_fit)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.006989 1.824661 2.984462 3.026105 4.190134 6.872129
```

```
summary(wine$hurdle_fit)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01872 1.90573 3.03602 3.02506 4.15582 7.24648
```

Part 6: Model Selection and Testing Prediction

The final portion of the script selects the highest-performing model and runs the prediction on the testing data set.

```
### Highest-Performing Model: Zero-Inflated Negative Binomial Regression Model
# The results of the Vuong test demonstrated that the zero-inflated models scored much better compared
summary(ZINB_Model)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ VolatileAcidity_IMP + Alcohol_IMP + FreeSulfurDioxide_IMP +
##     TotalSulfurDioxide_IMP + pH_IMP_bin + Sulphates_IMP + STARSxAlcohol_IMP +
##     LabelAppeal_IMP + AcidIndex_IMP + Final_REDFLAG + STARSxLabelAppeal_IMP +
##     STARS_IMP + pH_FLAG + STARS_FLAG, data = wine, dist = "negbin")
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.147608 -0.413902  0.001158  0.390821  5.969478
##
## Count model coefficients (negbin with log link):
##                             Estimate  Std. Error z value Pr(>|z|)
## (Intercept)               1.11001293  0.05951305  18.652  < 2e-16 ***
## VolatileAcidity_IMP      -0.01153524  0.00701060  -1.645  0.09989 .
## Alcohol_IMP               0.01030699  0.00392804   2.624  0.00869 **
## FreeSulfurDioxide_IMP     0.00002239  0.00003679   0.609  0.54270
## TotalSulfurDioxide_IMP   -0.00002930  0.00002480  -1.181  0.23748
## pH_IMP_binMedium Acidity  0.00417612  0.01251523   0.334  0.73862
## pH_IMP_binHigh Acidity    0.00227196  0.01486768   0.153  0.87855
## Sulphates_IMP             0.00049107  0.00590198   0.083  0.93369
## STARSxAlcohol_IMP        -0.00144628  0.00164018  -0.882  0.37789
## LabelAppeal_IMP           0.35792892  0.01630073  21.958  < 2e-16 ***
## AcidIndex_IMP            -0.02071985  0.00484553  -4.276 1.90e-05 ***
## Final_REDFLAG            -0.01847686  0.01177549  -1.569  0.11663
## STARSxLabelAppeal_IMP    -0.05929674  0.00686224  -8.641  < 2e-16 ***
## STARS_IMP                 0.14922500  0.01894618   7.876 3.37e-15 ***
## pH_FLAG1                 -0.00626041  0.03105603  -0.202  0.84024
## STARS_FLAG1              -0.15757018  0.01847490  -8.529  < 2e-16 ***
## Log(theta)               17.22969313  1.35531387  12.713  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                             Estimate Std. Error z value      Pr(>|z|)
## (Intercept)              -2.9618674  0.5130921  -5.773 0.00000000781 ***
## VolatileAcidity_IMP       0.0965650  0.0456335   2.116       0.03434 *
## Alcohol_IMP               0.0591999  0.0351980   1.682       0.09259 .
## FreeSulfurDioxide_IMP    -0.0007914  0.0002513  -3.149       0.00164 **
## TotalSulfurDioxide_IMP   -0.0006585  0.0001670  -3.943 0.00008053560 ***
## pH_IMP_binMedium Acidity  0.0194917  0.0843258   0.231       0.81720
## pH_IMP_binHigh Acidity    0.4191711  0.0968765   4.327 0.00001512485 ***
## Sulphates_IMP             0.1232320  0.0389208   3.166       0.00154 **
## STARSxAlcohol_IMP        -0.0183014  0.0195435  -0.936       0.34904
## LabelAppeal_IMP           0.7634888  0.1689841   4.518 0.00000623937 ***
## AcidIndex_IMP             0.4114553  0.0259799  15.837       < 2e-16 ***
## Final_REDFLAG             0.5198706  0.0859458   6.049 0.00000000146 ***
## STARSxLabelAppeal_IMP     0.4929433  0.1136050   4.339 0.00001430682 ***
```

```
## STARS_IMP                -3.2048560  0.3129119 -10.242       < 2e-16 ***
## pH_FLAG1                  0.4522132  0.1985006   2.278       0.02272 *
## STARS_FLAG1               5.7367875  0.2271547  25.255       < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 30392096.0417
## Number of iterations in BFGS optimization: 39
## Log-likelihood: -2.036e+04 on 33 Df
```

```
# Run prediction on testing dataset
wine_test$TARGET = predict(ZIP_Model, newdata = wine_test, type = "response")

summary(wine_test)
```

```
##      INDEX           TARGET         FixedAcidity     VolatileAcidity
##  Min.   :    3   Min.   :0.02706   Min.   :-18.200   Min.   :-2.8300
##  1st Qu.: 4018   1st Qu.:2.03466   1st Qu.:  5.200   1st Qu.: 0.0800
##  Median : 7906   Median :3.07335   Median :  6.900   Median : 0.2800
##  Mean   : 8048   Mean   :3.12215   Mean   :  6.864   Mean   : 0.3103
##  3rd Qu.:12061   3rd Qu.:4.16337   3rd Qu.:  9.000   3rd Qu.: 0.6300
##  Max.   :16130   Max.   :6.97634   Max.   : 33.500   Max.   : 3.6100
##
##    CitricAcid      ResidualSugar        Chlorides        FreeSulfurDioxide
##  Min.   :-3.1200   Min.   :-128.300   Min.   :-1.15000   Min.   :-563.00
##  1st Qu.: 0.0000   1st Qu.:  -2.600   1st Qu.: 0.01600   1st Qu.:   3.00
##  Median : 0.3100   Median :   3.600   Median : 0.04700   Median :  30.00
##  Mean   : 0.3124   Mean   :   5.319   Mean   : 0.06143   Mean   :  34.95
##  3rd Qu.: 0.6050   3rd Qu.:  17.200   3rd Qu.: 0.17100   3rd Qu.:  79.25
##  Max.   : 3.7600   Max.   : 145.400   Max.   : 1.26300   Max.   : 617.00
##                    NA's   :168        NA's   :138        NA's   :152
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-769.00    Min.   :0.8898   Min.   :0.600   Min.   :-3.0700
##  1st Qu.:  27.25    1st Qu.:0.9883   1st Qu.:2.980   1st Qu.: 0.3300
##  Median : 124.00    Median :0.9946   Median :3.210   Median : 0.5000
##  Mean   : 123.41    Mean   :0.9947   Mean   :3.237   Mean   : 0.5346
##  3rd Qu.: 210.00    3rd Qu.:1.0005   3rd Qu.:3.490   3rd Qu.: 0.8200
##  Max.   :1004.00    Max.   :1.0998   Max.   :6.210   Max.   : 4.1800
##  NA's   :157                         NA's   :104     NA's   :310
##     Alcohol       LabelAppeal         AcidIndex          STARS
##  Min.   :-4.20   Min.   :-2.00000   Min.   : 5.000   Min.   :1.00
##  1st Qu.: 9.00   1st Qu.:-1.00000   1st Qu.: 7.000   1st Qu.:1.00
##  Median :10.40   Median : 0.00000   Median : 8.000   Median :2.00
##  Mean   :10.58   Mean   : 0.01349   Mean   : 7.748   Mean   :2.04
##  3rd Qu.:12.50   3rd Qu.: 1.00000   3rd Qu.: 8.000   3rd Qu.:3.00
##  Max.   :25.60   Max.   : 2.00000   Max.   :17.000   Max.   :4.00
##  NA's   :185                                         NA's   :841
##  FixedAcidity_IMP  VolatileAcidity_IMP CitricAcid_IMP    ResidualSugar_IMP
##  Min.   :-18.200   Min.   :-2.8300     Min.   :-3.1200   Min.   :-97.900
##  1st Qu.:  5.200   1st Qu.: 0.0800     1st Qu.: 0.0000   1st Qu.:  0.500
##  Median :  6.900   Median : 0.2800     Median : 0.3100   Median :  3.900
##  Mean   :  6.864   Mean   : 0.3103     Mean   : 0.3124   Mean   :  5.238
##  3rd Qu.:  9.000   3rd Qu.: 0.6300     3rd Qu.: 0.6050   3rd Qu.: 15.525
##  Max.   : 33.500   Max.   : 3.6100     Max.   : 3.7600   Max.   : 98.500
##
```

```
##  Chlorides_IMP      FreeSulfurDioxide_IMP TotalSulfurDioxide_IMP
##  Min.   :-1.15000   Min.   :-381.00       Min.   :-515.0
##  1st Qu.: 0.02400   1st Qu.:   5.00       1st Qu.:  32.0
##  Median : 0.04600   Median :  30.00       Median : 123.0
##  Mean   : 0.06079   Mean   :  34.74       Mean   : 123.5
##  3rd Qu.: 0.14350   3rd Qu.:  70.00       3rd Qu.: 201.0
##  Max.   : 1.26300   Max.   : 470.00       Max.   : 744.0
##
##   Density_IMP         pH_IMP      Sulphates_IMP      Alcohol_IMP
##  Min.   :0.8898   Min.   :0.600   Min.   :-3.0700   Min.   :-4.20
##  1st Qu.:0.9883   1st Qu.:2.990   1st Qu.: 0.3600   1st Qu.: 9.10
##  Median :0.9946   Median :3.200   Median : 0.5000   Median :10.40
##  Mean   :0.9947   Mean   :3.235   Mean   : 0.5314   Mean   :10.57
##  3rd Qu.:1.0005   3rd Qu.:3.460   3rd Qu.: 0.7550   3rd Qu.:12.40
##  Max.   :1.0998   Max.   :6.210   Max.   : 4.1800   Max.   :25.60
##
##  LabelAppeal_IMP    AcidIndex_IMP      STARS_IMP     ResidualSugar_FLAG
##  Min.   :-2.00000   Min.   : 5.000   Min.   :1.000   0:3093
##  1st Qu.:-1.00000   1st Qu.: 7.000   1st Qu.:1.586   1: 242
##  Median : 0.00000   Median : 8.000   Median :2.000
##  Mean   : 0.01349   Mean   : 7.748   Mean   :2.023
##  3rd Qu.: 1.00000   3rd Qu.: 8.000   3rd Qu.:2.364
##  Max.   : 2.00000   Max.   :17.000   Max.   :4.000
##
##  Chlorides_FLAG FreeSulfurDioxide_FLAG TotalSulfurDioxide_FLAG pH_FLAG
##  0:3197         0:3123                 0:3116                  0:3231
##  1: 138         1: 212                 1: 219                  1: 104
##
##
##
##
##
##  Sulphates_FLAG Alcohol_FLAG STARS_FLAG VolatileAcidity_IMP_REDFLAG
##  0:3025         0:3150       0:2494     Min.   :0.0000
##  1: 310         1: 185       1: 841     1st Qu.:0.0000
##                                         Median :0.0000
##                                         Mean   :0.4483
##                                         3rd Qu.:1.0000
##                                         Max.   :1.0000
##
##  ResidualSugar_IMP_REDFLAG TotalSulfurDioxide_IMP_REDFLAG Density_IMP_REDFLAG
##  Min.   :0.0000            Min.   :0.0000                 Min.   :0.0000
##  1st Qu.:0.0000            1st Qu.:0.0000                 1st Qu.:0.0000
##  Median :1.0000            Median :1.0000                 Median :0.0000
##  Mean   :0.5667            Mean   :0.5205                 Mean   :0.4972
##  3rd Qu.:1.0000            3rd Qu.:1.0000                 3rd Qu.:1.0000
##  Max.   :1.0000            Max.   :1.0000                 Max.   :1.0000
##
##     TallyUp      Final_REDFLAG    TARGET_Flag    TARGET_AMT
##  Min.   :0.000   Min.   :0.0000   Mode:logical   Mode:logical
##  1st Qu.:1.000   1st Qu.:0.0000   NA's:3335      NA's:3335
##  Median :2.000   Median :0.0000
##  Mean   :2.033   Mean   :0.3241
##  3rd Qu.:3.000   3rd Qu.:1.0000
```

```
##  Max.   :4.000   Max.   :1.0000
##
##  STARSxLabelAppeal_IMP STARSxAlcohol_IMP        ResidualSugar_bin
##  Min.   :-6.0000       Min.   :-8.40     Low Sweetness   : 854
##  1st Qu.:-1.0000       1st Qu.:12.10     Medium Sweetness:1621
##  Median : 0.0000       Median :19.67     High Sweetness  : 860
##  Mean   : 0.2823       Mean   :21.59
##  3rd Qu.: 1.0000       3rd Qu.:27.95
##  Max.   : 8.0000       Max.   :94.40
##
##           pH_IMP_bin
##  Low Acidity   : 861
##  Medium Acidity:1678
##  High Acidity  : 796
##
##
##
##
```

```
summary(wine_test$TARGET)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02706 2.03466 3.07335 3.12215 4.16337 6.97634
```

```
length(wine_test$TARGET[wine_test$TARGET<1]) / length(wine_test$TARGET)
```

```
## [1] 0.07436282
```

```
summary(wine$TARGET)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   3.000   3.029   4.000   8.000
```