Predict 411

WINE SALES PROJECT

Zeeshan Latifi

Northwestern University Winter 2018

Introduction:

The purpose of this analysis is to determine whether we can predict the number of wine sample cases that were purchased by a wine distribution company after sampling the wine by using the wine characteristics as the predictor variables. The data set contains information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The target variable is what we will be predicting for as mentioned above. These cases are used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. If we can accurately predict the number of cases, then that manufacturer will be able to adjust their wine offerings to maximize sales.

Data Exploration:

The data contains 16 variables with 12,795 records of wine data. Figure 1 below highlights the summary of the data for each variable. As you can see there are several variables that contain missing values. For these variables with "NA's" we'll impute values using the simple average method.

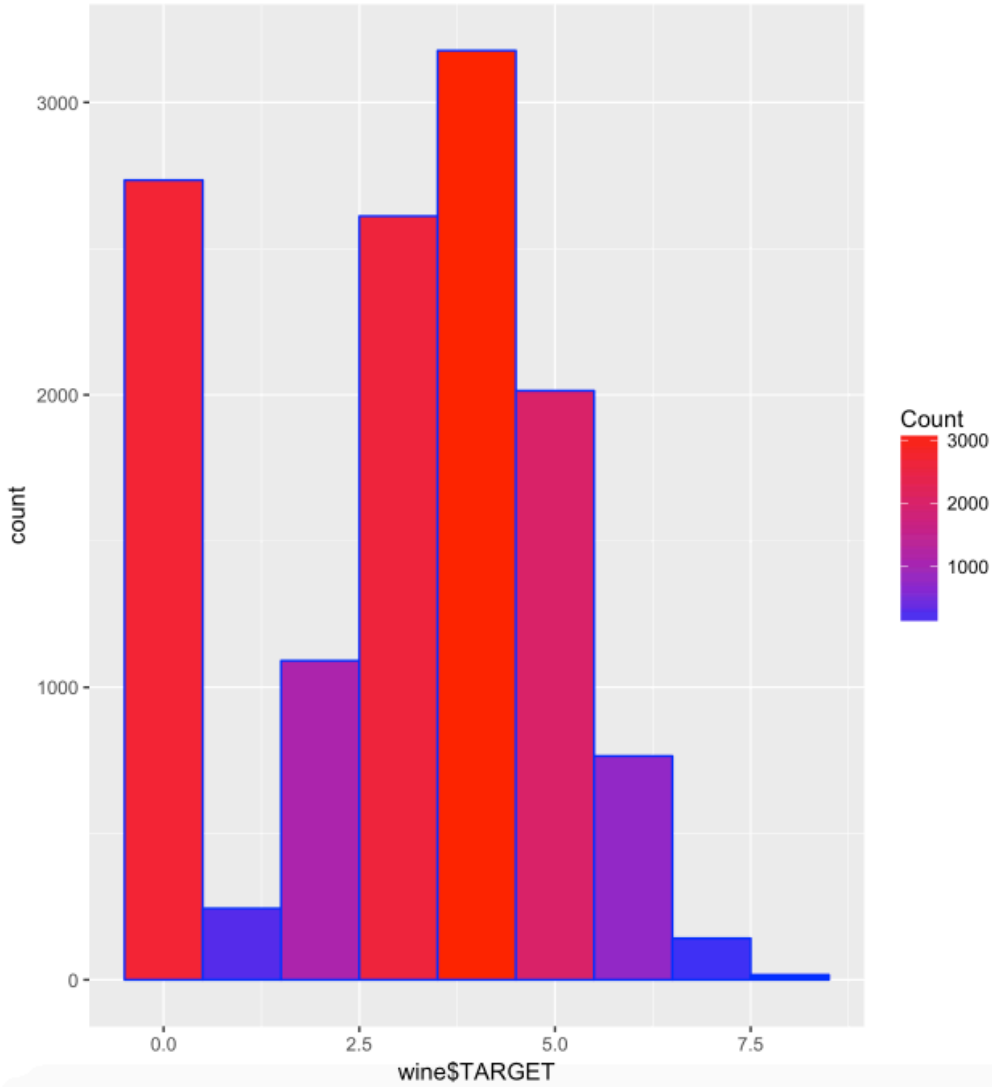
Figure 1

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

Notice how the Target variable minimum and maximum are 0 and 8 respectively with the mean being just over 3. We would expect these results when evaluating our model with our test data set. About 25% of the data for STARS is missing, we'll need to impute these values before using that variable in our model.

First let's begin by looking at our Target variable. Figure 2 below outlines the Target variable count. You can see that there is a large proportion of wine orders that are 0. Other than this, the rest of the data seems to be fairly distributed with a slight right skew.

Figure 2



Let's now look at some of the predictor variables. Figure 3 below shows the distribution for both Acid Index and well as the Alcohol content in the wine. Acid Index seems to be skewed right, with many upper outliers. Alcohol content levels between 7 and 12 seem to most prevalent; however, we can say the distribution is fairly normal.

Figure 3

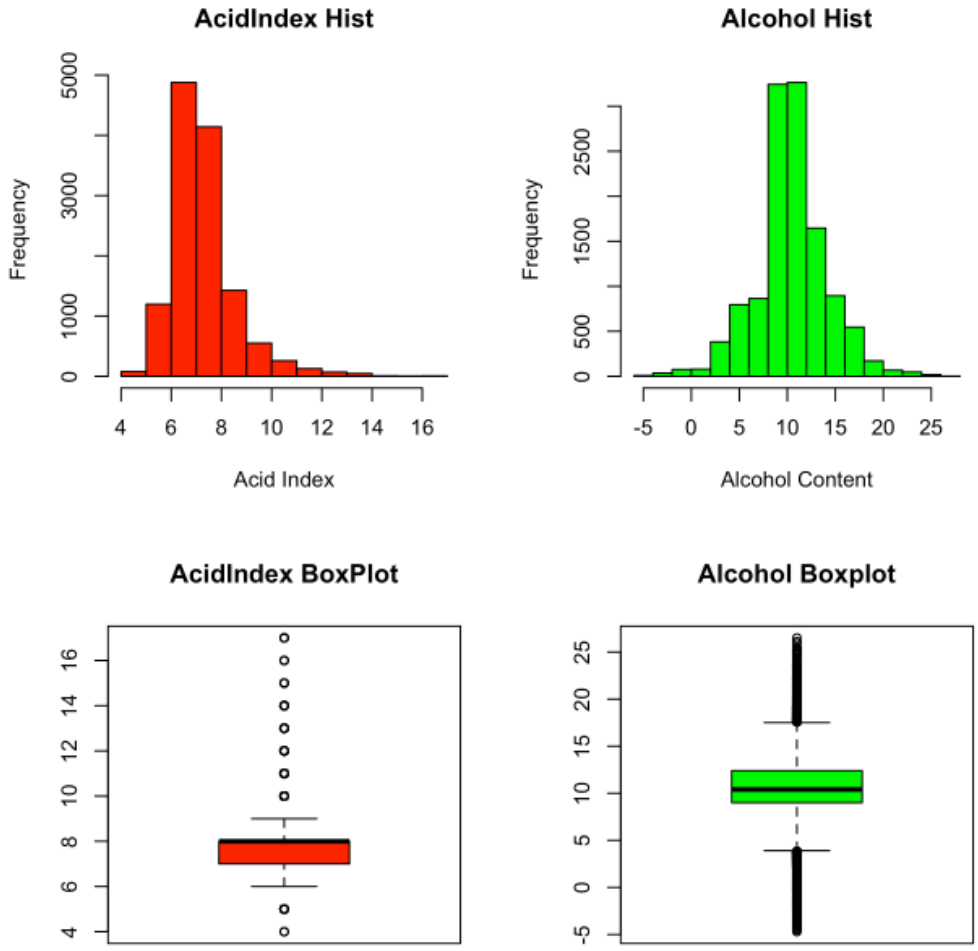
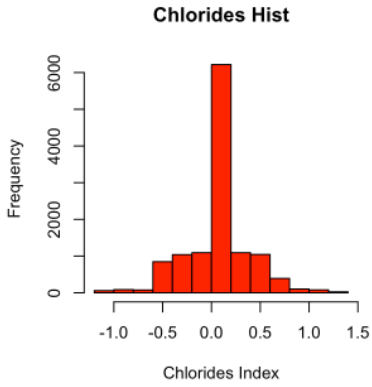
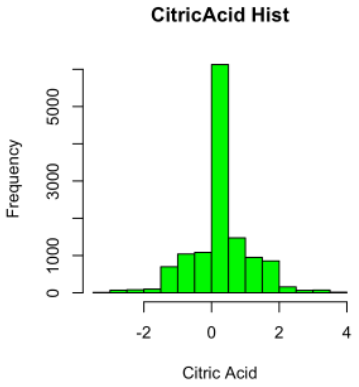
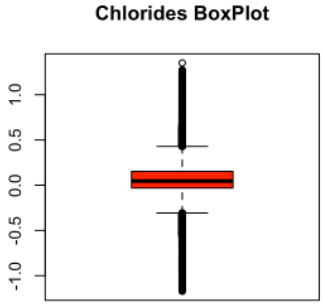


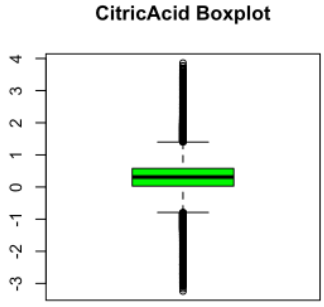
Figure 4 below outlines the histogram and boxplot for Chlorides and Citric Acid. Here there is a little different story. We can see that there is a large number of wines that fall within the a very low amount of Chloride as well as Citric Acid. Because of this, the boxplots for each are showing many outliers. The same can be said for Density and Fixed Acidity.

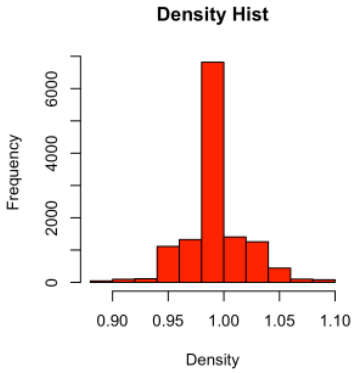
Figure 4

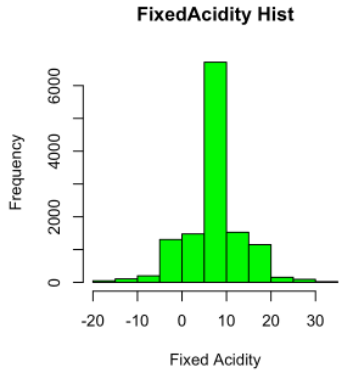

Chlorides Hist
Frequency
0
2000
4000
6000
-1.0
-0.5
0.0
0.5
1.0
1.5
Chlorides Index

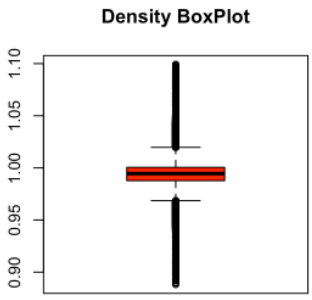

CitricAcid Hist
Frequency
0
1000
3000
5000
-2
0
2
4
Citric Acid


Chlorides BoxPlot
-1.0
-0.5
0.0
0.5
1.0


CitricAcid Boxplot
-3
-2
-1
0
1
2
3
4


Density Hist
Frequency
0
2000
4000
6000
0.90
0.95
1.00
1.05
1.10
Density


FixedAcidity Hist
Frequency
0
2000
4000
6000
-20
-10
0
10
20
30
Fixed Acidity


Density BoxPlot
0.90
0.95
1.00
1.05
1.10

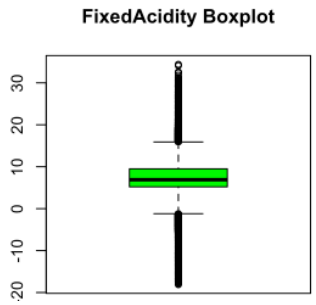

FixedAcidity Boxplot
-20
-10
0
10
20
30

Figure 5 below showcases a little different story. We wanted to see the distribution of the Label Appeal and number of Stars for each wine (if the data exists). We can see that most of the data that we have is a Star rating of 2 for the wines. The majority of wines showcase a Label Appeal of 0.

Figure 5

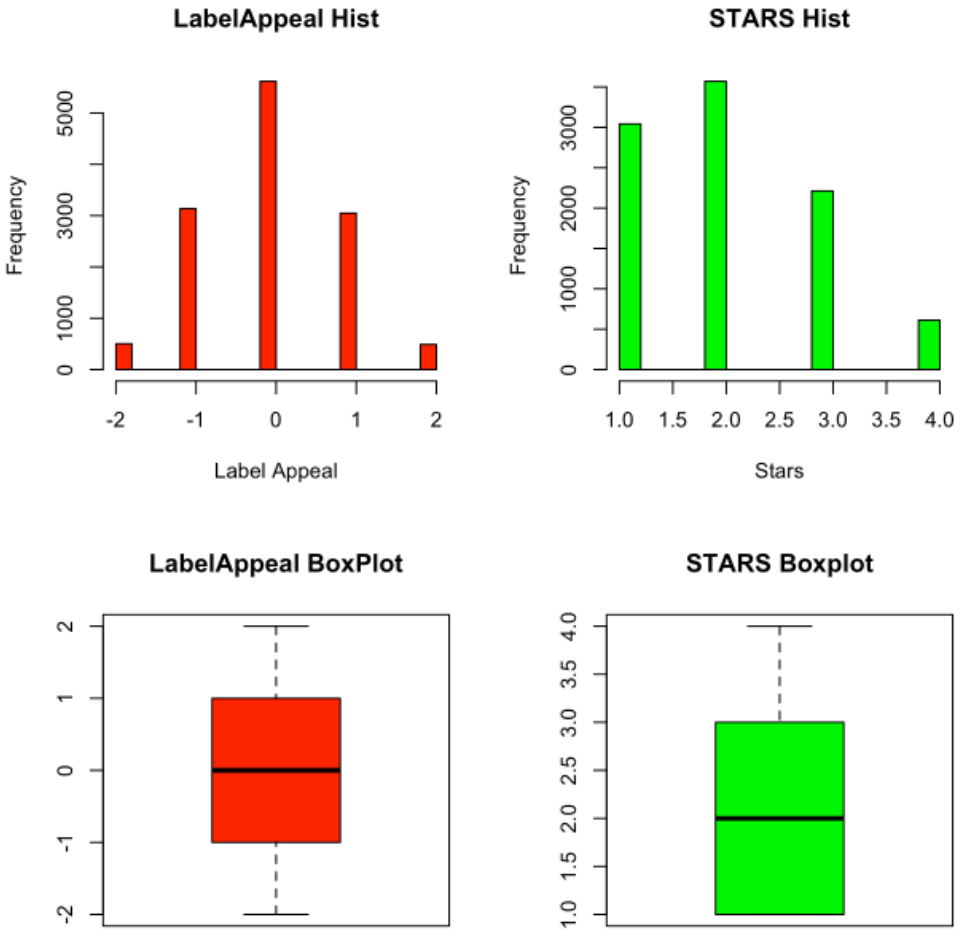
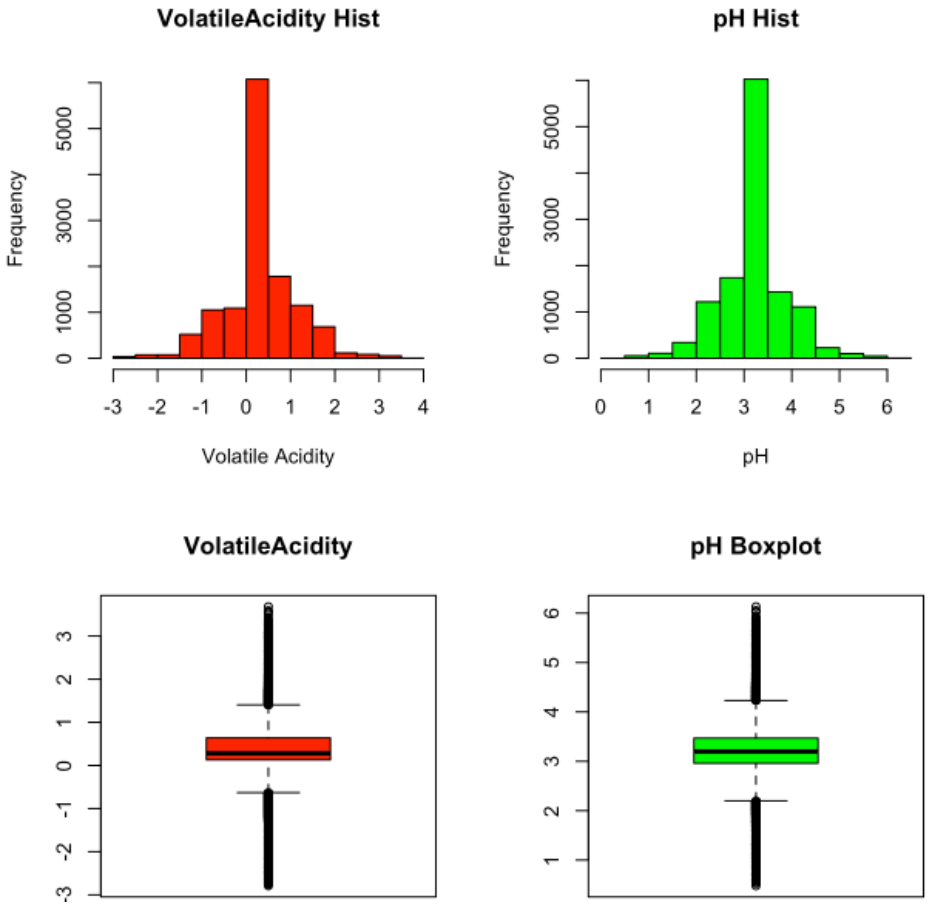


Figure 6 below looks very similar to the variables above in Figures 3 and 4. Volatile Acidity and pH levels seems to be predominately in the middle of the pack. Therefore, the boxplots show many outliers on either tail end. There may be a correlation to these variables. We'll dive deeper in the next section.

Figure 6



Data Preparation:

To being our data preparation, we'll impute the values needed for our missing variables. To do this, we'll use the simple average method and just replace our missing values with NA's.

To start, we need to impute the missing values in our data. In order to do this, we've replaced our missing values with the mean of the actual values. The newly imputed values are given a new name with the suffix: IMP. It should be mentioned that every transformation, categorization, or imputation were also replicated on our test data set. This is to ensure proper predictability when deploying our model.

Figure 7

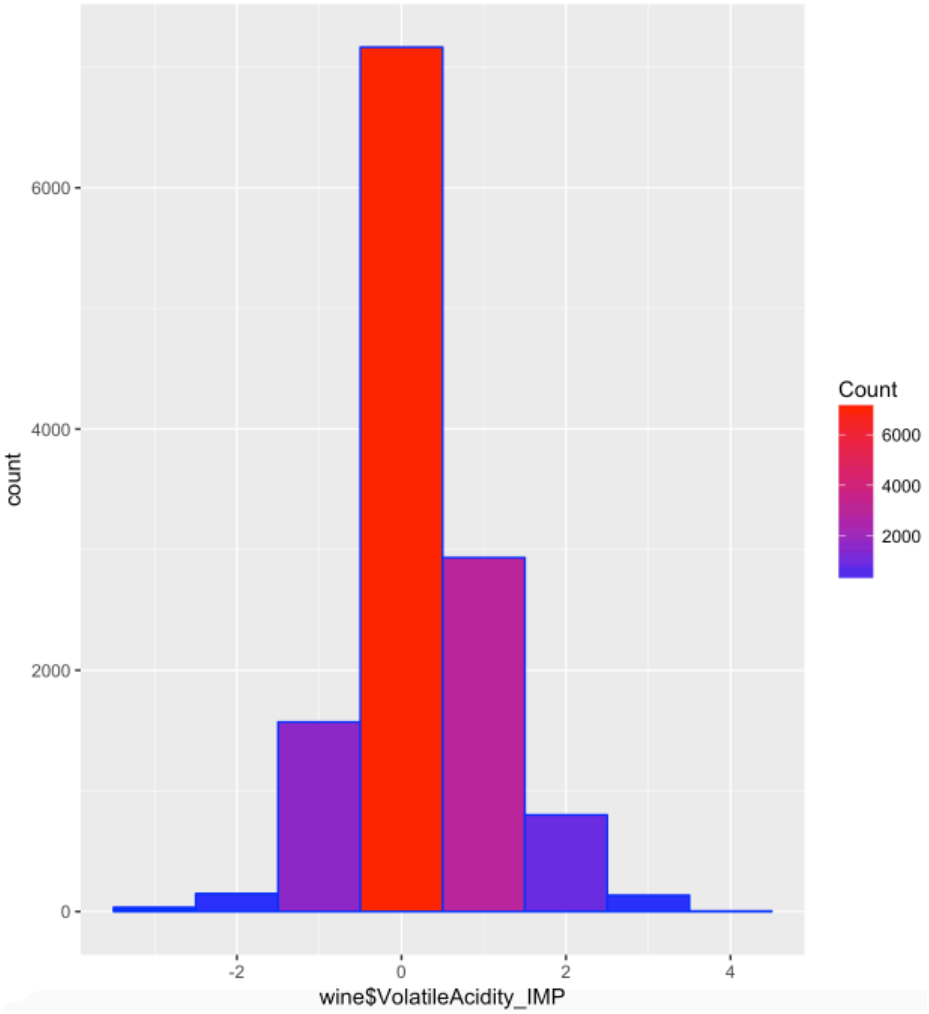
```
     INDEX           TARGET       FixedAcidity     VolatileAcidity     CitricAcid      ResidualSugar
 Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400   Min.   :-127.800
 1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300   1st Qu.:  -2.000
 Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100   Median :   3.900
 Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084   Mean   :   5.419
 3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800   3rd Qu.:  15.900
 Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600   Max.   : 141.150
                                                                                       NA's   :616
   Chlorides       FreeSulfurDioxide TotalSulfurDioxide    Density             pH          Sulphates
 Min.   :-1.1710   Min.   :-555.00   Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
 1st Qu.:-0.0310   1st Qu.:   0.00   1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800
 Median : 0.0460   Median :  30.00   Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000
 Mean   : 0.0548   Mean   :  30.85   Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
 3rd Qu.: 0.1530   3rd Qu.:  70.00   3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600
 Max.   : 1.3510   Max.   : 623.00   Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
 NA's   :638       NA's   :647       NA's   :682                         NA's   :395     NA's   :1210
    Alcohol       LabelAppeal          AcidIndex          STARS       FixedAcidity_IMP  VolatileAcidity_IMP
 Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000   Min.   :-18.100   Min.   :-2.7900
 1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000   1st Qu.:  5.200   1st Qu.: 0.1300
 Median :10.40   Median : 0.000000   Median : 8.000   Median :2.000   Median :  6.900   Median : 0.2800
 Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042   Mean   :  7.076   Mean   : 0.3241
 3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000   3rd Qu.:  9.500   3rd Qu.: 0.6400
 Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000   Max.   : 34.400   Max.   : 3.6800
 NA's   :653                                          NA's   :3359
 CitricAcid_IMP    ResidualSugar_IMP  Chlorides_IMP      FreeSulfurDioxide_IMP TotalSulfurDioxide_IMP  Density_IMP
 Min.   :-3.2400   Min.   :-127.800   Min.   :-1.17100   Min.   :-555.00       Min.   :-823.0         Min.   :0.8881
 1st Qu.: 0.0300   1st Qu.:   0.900   1st Qu.: 0.00000   1st Qu.:   5.00       1st Qu.:  34.0         1st Qu.:0.9877
 Median : 0.3100   Median :   4.900   Median : 0.04800   Median :  30.85       Median : 120.7         Median :0.9945
 Mean   : 0.3084   Mean   :   5.419   Mean   : 0.05482   Mean   :  30.85       Mean   : 120.7         Mean   :0.9942
 3rd Qu.: 0.5800   3rd Qu.:  14.900   3rd Qu.: 0.12800   3rd Qu.:  64.00       3rd Qu.: 198.0         3rd Qu.:1.0005
 Max.   : 3.8600   Max.   : 141.150   Max.   : 1.35100   Max.   : 623.00       Max.   :1057.0         Max.   :1.0992
    pH_IMP      Sulphates_IMP      Alcohol_IMP     LabelAppeal_IMP     AcidIndex_IMP      STARS_IMP
 Min.   :0.480   Min.   :-3.1300   Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
 1st Qu.:2.970   1st Qu.: 0.3400   1st Qu.: 9.10   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:2.000
 Median :3.208   Median : 0.5271   Median :10.49   Median : 0.000000   Median : 8.000   Median :2.000
 Mean   :3.208   Mean   : 0.5271   Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
 3rd Qu.:3.450   3rd Qu.: 0.7700   3rd Qu.:12.20   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:2.042
 Max.   :6.130   Max.   : 4.2400   Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000

 ResidualSugar_IMP_Flag Chlorides_IMP_Flag FreeSulfurDioxide_IMP_Flag TotalSulfurDioxide_IMP_Flag  pH_IMP_Flag
 Min.   :0.00000        Min.   :0.00000    Min.   :0.00000            Min.   :0.0000              Min.   :0.00000
 1st Qu.:0.00000        1st Qu.:0.00000    1st Qu.:0.00000            1st Qu.:0.0000              1st Qu.:0.00000
 Median :0.00000        Median :0.00000    Median :0.00000            Median :0.0000              Median :0.00000
 Mean   :0.04814        Mean   :0.04986    Mean   :0.05057            Mean   :0.0533              Mean   :0.03087
 3rd Qu.:0.00000        3rd Qu.:0.00000    3rd Qu.:0.00000            3rd Qu.:0.0000              3rd Qu.:0.00000
 Max.   :1.00000        Max.   :1.00000    Max.   :1.00000            Max.   :1.0000              Max.   :1.00000

 Sulphates_IMP_Flag Alcohol_IMP_Flag  STARS_IMP_Flag
 Min.   :0.00000    Min.   :0.00000   Min.   :0.0000
 1st Qu.:0.00000    1st Qu.:0.00000   1st Qu.:0.0000
 Median :0.00000    Median :0.00000   Median :0.0000
 Mean   :0.09457    Mean   :0.05104   Mean   :0.2625
 3rd Qu.:0.00000    3rd Qu.:0.00000   3rd Qu.:1.0000
 Max.   :1.00000    Max.   :1.00000   Max.   :1.0000
```

Next, we also created some flags within our data set to ensure accuracy and detectability. We flagged our variables that were imputed with a 1, 0 if actual value. We also broke out the type of wine, red or white, by looking at the volatile acidity. Figure 8 below has the breakdown of wine of red versus white. White wine is represented by the blue color.

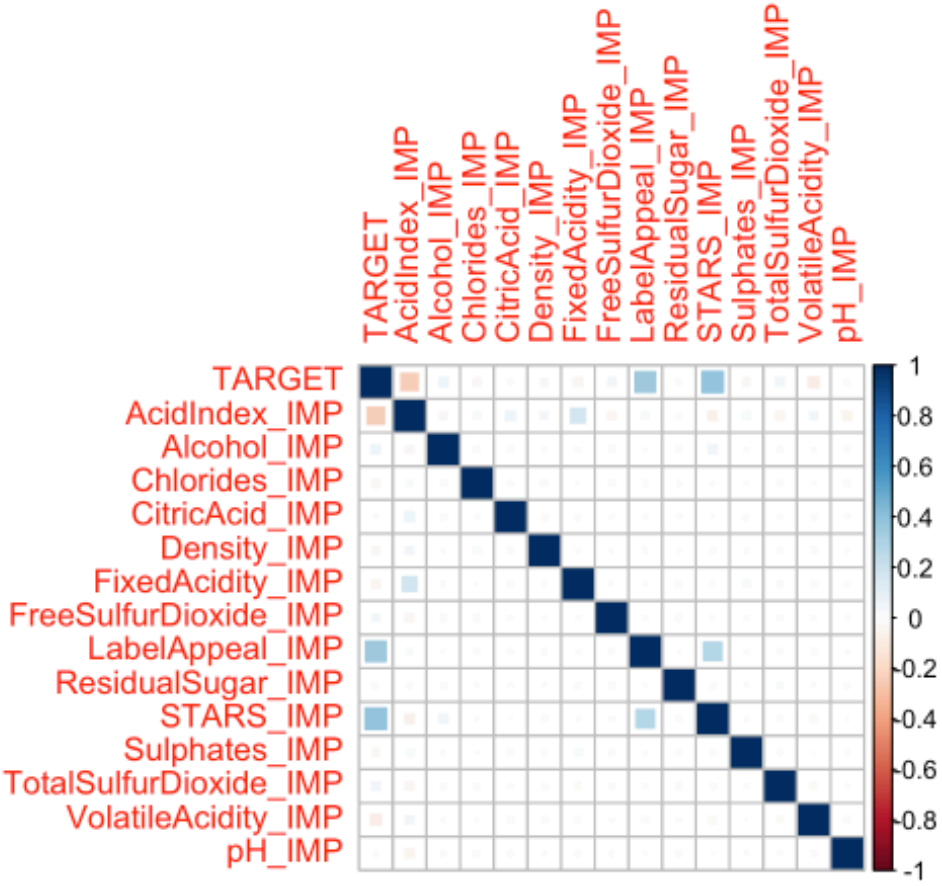
Figure 8



This indicator was created so that when our model is complete, we can see the breakdown of what cases of wine were ordered by the type of wine. As shown the majority of the wine that falls within a 0 range for Volatile Acidity happens to be red wine.

Additionally, we also built a correlation plot in Figure 9. You can see that there is a positive correlation between our Target and the number of Stars a wine gets as well as the Label Appeal. This makes sense logically, the better the rating, the more you'll sell. There is also somewhat of a negative correlation between the Acid Index and the Target variable as shown.

Figure 9



Build Linear Regression Models:

From our cleansed data, we now begin building an optimal model for predicting our response variable. In order to optimize our search for the most accurate model, we will use an R function (regsubsets) to score the variables that will have the most impact on generating the best model. Based on the automated variable selection, the best variables are as follows:

- AcidIndex_IMP
- Alcohol_IMP
- Density_IMP
- LabelAppeal_IMP
- STARS_IMP

- Density_IMP_REDFLAG
- ResidualSugar_IMP_REDFLAG
- STARS_IMP_Flag
- TotalSulfurDioxide_IMP_REDFLAG
- VolatileAcidity_IMP_REDFLAG

Figures 10-15 below indicate the summary output for each of our models. The following models were used for this analysis (in order of occurrence):

1. Regular Linear Regression
2. Regular Linear Regression using Stepwise Variable Selection
3. Poisson
4. Negative Binomial Distribution
5. Zero Inflated Poisson
6. Zero Inflated Negative Binomial Regression

Figure 10 (1)

```
Call:
lm(formula = TARGET ~ AcidIndex_IMP + Alcohol_IMP + Density_IMP +
    LabelAppeal_IMP + STARS_IMP + Density_IMP_REDFLAG + ResidualSugar_IMP_REDFLAG +
    STARS_IMP_Flag + TotalSulfurDioxide_IMP_REDFLAG + VolatileAcidity_IMP_REDFLAG,
    data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6602 -0.8438  0.0316  0.8376  6.0613

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   1.175431   0.574950   2.044 0.040934 *
AcidIndex_IMP                -0.180784   0.009030 -20.020  < 2e-16 ***
Alcohol_IMP                   0.011579   0.003198   3.621 0.000295 ***
Density_IMP                   2.537781   0.581203   4.366 1.27e-05 ***
LabelAppeal_IMP               0.469723   0.013598  34.543  < 2e-16 ***
STARS_IMP                     0.766373   0.015629  49.034  < 2e-16 ***
Density_IMP_REDFLAG          -0.280109   0.031485  -8.897  < 2e-16 ***
ResidualSugar_IMP_REDFLAG    -0.095602   0.023372  -4.090 4.33e-05 ***
STARS_IMP_Flag               -2.256707   0.026893 -83.913  < 2e-16 ***
TotalSulfurDioxide_IMP_REDFLAG -0.141326   0.023364  -6.049 1.50e-09 ***
VolatileAcidity_IMP_REDFLAG  -0.221031   0.023531  -9.393  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.303 on 12784 degrees of freedom
Multiple R-squared:  0.5428,    Adjusted R-squared:  0.5424
F-statistic:  1517 on 10 and 12784 DF,  p-value: < 2.2e-16
```

As expected, the regular linear regression is not going to be optimal here as we have zero inflated variables. However, for a good threshold, it's always good to see what the model output is. Here we can see that Density and STARS have a significant impact on the model. STARS seems to be counterintuitive. Having more stars should be a positive impact on the cases of wine being sold.

Figure 11 (2)

```
Call:
lm(formula = TARGET ~ AcidIndex_IMP + Alcohol_IMP + Density_IMP +
    LabelAppeal_IMP + STARS_IMP + Density_IMP_REDFLAG + ResidualSugar_IMP_REDFLAG +
    STARS_IMP_Flag + TotalSulfurDioxide_IMP_REDFLAG + VolatileAcidity_IMP_REDFLAG,
    data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6602 -0.8438  0.0316  0.8376  6.0613

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.175431   0.574950   2.044 0.040934 *
AcidIndex_IMP                 -0.180784   0.009030 -20.020  < 2e-16 ***
Alcohol_IMP                    0.011579   0.003198   3.621 0.000295 ***
Density_IMP                    2.537781   0.581203   4.366 1.27e-05 ***
LabelAppeal_IMP                0.469723   0.013598  34.543  < 2e-16 ***
STARS_IMP                      0.766373   0.015629  49.034  < 2e-16 ***
Density_IMP_REDFLAG           -0.280109   0.031485  -8.897  < 2e-16 ***
ResidualSugar_IMP_REDFLAG     -0.095602   0.023372  -4.090 4.33e-05 ***
STARS_IMP_Flag                -2.256707   0.026893 -83.913  < 2e-16 ***
TotalSulfurDioxide_IMP_REDFLAG -0.141326   0.023364  -6.049 1.50e-09 ***
VolatileAcidity_IMP_REDFLAG   -0.221031   0.023531  -9.393  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.303 on 12784 degrees of freedom
Multiple R-squared:  0.5428,    Adjusted R-squared:  0.5424
F-statistic:  1517 on 10 and 12784 DF,  p-value: < 2.2e-16
```

For our regular linear regression with a stepwise variable selection (Figure 11), we have very similar results to the initial regression model as shown in Figure 10. The adjusted R^2 is the same for both models at 0.5424.

Figure 12 (3)

```
Call:
glm(formula = TARGET ~ AcidIndex_IMP + Alcohol_IMP + Density_IMP +
    LabelAppeal_IMP + STARS_IMP + Density_IMP_REDFLAG + ResidualSugar_IMP_REDFLAG +
    STARS_IMP_Flag + TotalSulfurDioxide_IMP_REDFLAG + VolatileAcidity_IMP_REDFLAG,
    family = poisson(link = "log"), data = wine)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1770  -0.6551   0.0090   0.4551   3.7050

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                     0.685022   0.254837   2.688 0.007186 **
AcidIndex_IMP                  -0.074255   0.004555 -16.301  < 2e-16 ***
Alcohol_IMP                     0.003258   0.001416   2.302 0.021359 *
Density_IMP                     0.852063   0.257285   3.312 0.000927 ***
LabelAppeal_IMP                 0.159715   0.006127  26.066  < 2e-16 ***
STARS_IMP                       0.183538   0.006111  30.036  < 2e-16 ***
Density_IMP_REDFLAG            -0.093952   0.013919  -6.750 1.48e-11 ***
ResidualSugar_IMP_REDFLAG      -0.034682   0.010286  -3.372 0.000747 ***
STARS_IMP_Flag                 -1.024305   0.016988 -60.296  < 2e-16 ***
TotalSulfurDioxide_IMP_REDFLAG -0.049784   0.010317  -4.825 1.40e-06 ***
VolatileAcidity_IMP_REDFLAG    -0.076514   0.010496  -7.289 3.11e-13 ***
---
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: 13691  on 12784  degrees of freedom
AIC: 45655

Number of Fisher Scoring iterations: 6
```

For our Poisson regression model, we can see that the STARS variable is also having a positive impact, although not at the magnitude at our previous models. Of the remaining variables, no other variable stands out as a significant coefficient of magnitude.

Figure 13 (4)

```
Call:
glm.nb(formula = TARGET ~ +AcidIndex_IMP + Alcohol_IMP + Density_IMP +
    LabelAppeal_IMP + STARS_IMP + Density_IMP_REDFLAG + ResidualSugar_IMP_REDFLAG +
    STARS_IMP_Flag + TotalSulfurDioxide_IMP_REDFLAG + VolatileAcidity_IMP_REDFLAG,
    data = wine, init.theta = 40944.55286, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1769  -0.6551   0.0090   0.4551   3.7048

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   0.685015   0.254848   2.688 0.007190 **
AcidIndex_IMP                -0.074257   0.004555 -16.301  < 2e-16 ***
Alcohol_IMP                   0.003258   0.001416   2.301 0.021367 *
Density_IMP                   0.852086   0.257296   3.312 0.000927 ***
LabelAppeal_IMP               0.159714   0.006128  26.065  < 2e-16 ***
STARS_IMP                     0.183540   0.006111  30.035  < 2e-16 ***
Density_IMP_REDFLAG          -0.093954   0.013920  -6.750 1.48e-11 ***
ResidualSugar_IMP_REDFLAG    -0.034683   0.010287  -3.372 0.000747 ***
STARS_IMP_Flag               -1.024305   0.016988 -60.295  < 2e-16 ***
TotalSulfurDioxide_IMP_REDFLAG -0.049786   0.010318  -4.825 1.40e-06 ***
VolatileAcidity_IMP_REDFLAG  -0.076516   0.010497  -7.289 3.11e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(40944.55) family taken to be 1)

    Null deviance: 22860  on 12794  degrees of freedom
Residual deviance: 13690  on 12784  degrees of freedom
AIC: 45657
Number of Fisher Scoring iterations: 1


              Theta:  40945
          Std. Err.:  34907
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -45633.26
```

For our negative binomial distribution, we can see that there is again a positive coefficient for STARS. This makes logical sense, however, the STARS_IMP_FLAG is negative here.

Figure 14 (5)

```
Call:
zeroinfl(formula = TARGET ~ AcidIndex_IMP + Alcohol_IMP + Density_IMP + LabelAppeal_IMP + STARS_IMP +
    Density_IMP_REDFLAG + ResidualSugar_IMP_REDFLAG + STARS_IMP_Flag + TotalSulfurDioxide_IMP_REDFLAG +
    VolatileAcidity_IMP_REDFLAG, data = wine)

Pearson residuals:
     Min       1Q   Median       3Q      Max
-2.32984 -0.41535 -0.00237  0.37691  6.53993

Count model coefficients (poisson with log link):
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 0.7722549  0.2625470   2.941 0.003267 **
AcidIndex_IMP              -0.0162503  0.0048651  -3.340 0.000837 ***
Alcohol_IMP                 0.0064377  0.0014455   4.454 8.44e-06 ***
Density_IMP                 0.4233243  0.2653593   1.595 0.110648
LabelAppeal_IMP             0.2325997  0.0063183  36.813  < 2e-16 ***
STARS_IMP                   0.1022796  0.0064164  15.940  < 2e-16 ***
Density_IMP_REDFLAG        -0.0570562  0.0142523  -4.003 6.25e-05 ***
ResidualSugar_IMP_REDFLAG   0.0009837  0.0105322   0.093 0.925588
STARS_IMP_Flag             -0.1866301  0.0185596 -10.056  < 2e-16 ***
TotalSulfurDioxide_IMP_REDFLAG  0.0093632  0.0105432   0.888 0.374498
VolatileAcidity_IMP_REDFLAG    -0.0308224  0.0107159  -2.876 0.004024 **

Zero-inflation model coefficients (binomial with logit link):
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 1.595006   1.735823   0.919  0.35816
AcidIndex_IMP               0.382491   0.025627  14.925  < 2e-16 ***
Alcohol_IMP                 0.027553   0.009497   2.901  0.00372 **
Density_IMP                -3.640226   1.731307  -2.103  0.03550 *
LabelAppeal_IMP             0.713761   0.042474  16.805  < 2e-16 ***
STARS_IMP                  -3.784662   0.332988 -11.366  < 2e-16 ***
Density_IMP_REDFLAG         0.373102   0.093149   4.005 6.19e-05 ***
ResidualSugar_IMP_REDFLAG   0.311561   0.069735   4.468 7.90e-06 ***
STARS_IMP_Flag              6.009677   0.346825  17.328  < 2e-16 ***
TotalSulfurDioxide_IMP_REDFLAG  0.546234   0.069114   7.903 2.71e-15 ***
VolatileAcidity_IMP_REDFLAG     0.433461   0.068711   6.308 2.82e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 32
Log-likelihood: -2.034e+04 on 22 Df
```

Based on our data, we're predicting this model to work best. The Zero Inflated Poisson model, accounts for the zero inflated variables. We can also see the count model above to have a positive coefficient for STARS. The data for many of our variables are zero inflated. Therefore, this model would be best for using that as a reliable predictor.

Figure 15 (6)

```
Call:
zeroinfl(formula = TARGET ~ AcidIndex_IMP + Alcohol_IMP + TotalSulfurDioxide_IMP + VolatileAcidity_IMP +
    STARS_IMP, data = wine, dist = "negbin", EM = TRUE)

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.8283 -0.5215  0.1158  0.5185  5.2282

Count model coefficients (negbin with log link):
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            9.359e-01  4.392e-02  21.308  < 2e-16 ***
AcidIndex_IMP         -1.193e-02  4.924e-03  -2.422  0.01542 *
Alcohol_IMP            6.927e-03  1.466e-03   4.725  2.3e-06 ***
TotalSulfurDioxide_IMP -4.256e-05  2.294e-05  -1.855  0.06356 .
VolatileAcidity_IMP   -1.676e-02  6.871e-03  -2.439  0.01472 *
STARS_IMP              1.901e-01  6.092e-03  31.207  < 2e-16 ***
Log(theta)             1.197e+01  3.655e+00   3.276  0.00105 **

Zero-inflation model coefficients (binomial with logit link):
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           -4.3709657  0.1860691 -23.491  < 2e-16 ***
AcidIndex_IMP          0.4804932  0.0187401  25.640  < 2e-16 ***
Alcohol_IMP            0.0072418  0.0071568   1.012    0.312
TotalSulfurDioxide_IMP -0.0008546  0.0001152  -7.418 1.19e-13 ***
VolatileAcidity_IMP    0.2395260  0.0333908   7.173 7.31e-13 ***
STARS_IMP             -0.5099835  0.0353555 -14.424  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 158460.9464
Number of iterations in BFGS optimization: 1
Log-likelihood: -2.337e+04 on 13 Df
```

Figure 15 above outlines the last of our models. Here we've accounted for the zero inflated variables, but now we have negative binomial regression. For the count models, no variable stands out as significant for our regression, indicating a more conservative mode. However, when accounting for zero inflation, the intercept is of high magnitude.

Select Best Model and Stand Alone Scoring:

To asses our model, we will use the AIC and prediction values. The table below outlines the model metrics.

Figure 16

Model	AIC
Model 1	43098.72
Model 2	43098.72
Model 3	45654.86
Model 4	45657.26
Model 5	40730.06
Model 6	46761.49

Based on the values above in Figure 16, we can see the best model fit is when using Zero Inflated Poisson (Model 5). Just as we predicted, this was the best model. This is because when looking at the data, we had several zero inflated variables. In addition, when looking at the target variable, we noticed the variance is greater than the mean. The variance is 3.71 and the mean 3.03. This means that using a Poisson regression model would work best. When looking back at the values from the original wine data set, the min was 0, the max was 8, and the mean was just a little over 3. When building our target prediction using our test data set, the values came very close to the actual values from the training data set. We can then reasonably infer that the Zero Inflation Poisson Regression model is the best fit model to predict the cases of wine that will be sold.