# Unit 03 Assignment

## Wine Sales Poisson Regression Project

**Name:** Young, Brent

**Predict 411 Section #:** 56

**Quarter:** Fall 2017

***Bingo Bonus:***

- Used MICE package for missing value imputation (20 Points)

## Introduction

### *Context*

The dataset that we will be working with is called wine (includes approximately 12,000 records). Each record represents commercially available wines. Additionally, each record has variables that are mostly related to chemical properties of the wine being sold. The target variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases are used to provide tasting samples to restaurants and wine stores across the United States. The more sample cases that are purchased, the more likely a wine is to be sold at a high end restaurant.

### *Objectives/Purpose*

The purpose of unit 3 assignment is to analyze wine data using Poisson Regression, Zero-Inflated Poisson Regression, Negative Binomial Regression, Zero-Inflated Negative Binomial Regression, and Multiple Linear Regression in order to predict the number of wine cases that will be sold based on certain properties/characteristics of the wine. If the wine manufacturer can predict the number of cases, then the manufacturer will be able to adjust their wine offerings to maximize their sales. This will be accomplished by generating at least five models using the following procedures: Poisson Regression, Zero-Inflated Poisson Regression, Negative Binomial Regression, Zero-Inflated Negative Binomial Regression, and Multiple Linear Regression. From these models, the best model will be selected. First, an initial exploratory data analysis will be conducted using scatterplots, boxplots, summary statistics, etc. to help understand important characteristics and properties of the data that may be disguised by numerical summaries. Second, data preparation/transformations of the data will begin. This includes, but not limited to fixing missing values, conducting data transformations, and creating new/flag variables. Third, we will begin building at least five different models using the procedures: Poisson Regression, Zero-Inflated Poisson Regression, Negative Binomial Regression, Zero-Inflated Negative Binomial Regression, and Multiple Linear Regression. This will be conducted by manually selecting the variables or using variable selection techniques (e.g., stepwise). We will then discuss the coefficients in the model to ensure that it makes intuitive wine sense (e.g., number of stars, wine label appeal, and similarities/difference in the coefficients and magnitude of variables model to model). Fourth, we will then decide on the "best model" using metrics such as AIC, or Average Squared Error. Fourth, a Stand Alone scoring program will be conducted that will score the new data and predict the number of wine cases that will be sold based upon the qualities of the wine. The data step within the Poisson family will include all the variable transformations such as fixing missing values and the poisson regression formula. Lastly, a scored data file will be produced that will contain two variables for each record: INDEX, P_TARGET.

## Section 1: Data Exploration

### Figure 1: Structure and Size of the Data

```
'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 ...
```

**Observations:** Figure 1 shows the structure of the data, which comes out to 12795 rows and 16 variables (integers/numbers). INDEX is not considered a true variable, while TARGET is considered our response variable, and the rest of the variables are considered our predictors.

### Figure 2: Definitions of the Variables (Data Dictionary)

| VARIABLE NAME | DEFINITION | THEORETICAL EFFECT |
|---|---|---|
| **INDEX** | Identification Variable (do not use) | None |
| TARGET | Number of Cases Purchased | None |
| | | |
| | | |
| AcidIndex | Proprietary method of testing total acidity of wine by using a weighted average | |
| Alcohol | Alcohol Content | |
| Chlorides | Chloride content of wine | |
| CitricAcid | Citric Acid Content | |
| Density | Density of Wine | |
| FixedAcidity | Fixed Acidity of Wine | |
| FreeSulfurDioxide | Sulfur Dioxide content of wine | |
| LabelAppeal | Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customes don't like the design. | Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales. |
| ResidualSugar | Residual Sugar of wine | |
| STARS | Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor | A high number of stars suggests high sales |
| Sulphates | Sulfate conten of wine | |
| TotalSulfurDioxide | Total Sulfur Dioxide of Wine | |
| VolatileAcidity | Volatile Acid content of wine | |
| pH | pH of wine | |

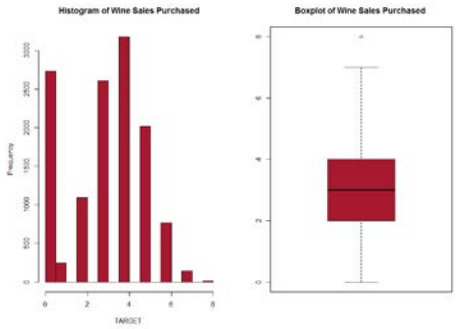
**Observations:** Figure 2 shows the definitions of the variables that are included in the dataset and the theoretical effect in the third column.

**Figure 3: Data Quality Check** *(see appendix)*

```
> summary(wine)
     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
```

**Observations:** Figure 3 *(also see appendix for additional data quality checks)* shows summary statistics so that we can check for missing values, outliers, etc. The data shows that the mean number of wine cases purchased is 3.029, the median number of wine cases purchased is 3, and that 21.4% or 2734 records have values of 0 (e.g., did not purchase any sample cases of wine). The data quality check also revealed that there are missing values for 8 variables: ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol, and Stars. Furthermore, the data quality check also revealed outliers and a “wide range” of values for majority of the variables (e.g., very low or very high totals). For example, ResidualSugar, FreeSulfurDioxide, and TotalSulfurDioxide. It’s also important to note that some of the variables have negative numbers when they technically shouldn’t (based on laws of chemistry, etc.). We will ignore these issues and use the data as we normally would. Additionally, only 6.5% of the wine case purchases received a 4 star, with majority receiving 1 to 3 Stars. Moreover, 43.9% of the wine case purchases received a LabelAppeal of 0 (Neutral). As we go on, we will have to investigate these outliers, missing values, and decide what to do with them (e.g., conducting imputation, etc.).

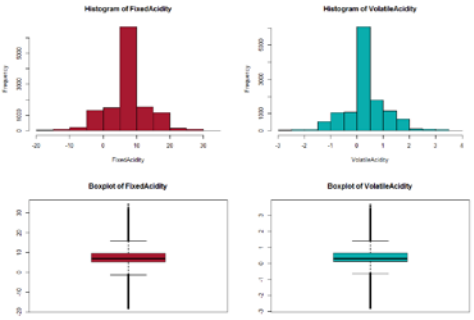
**Figure 4: Histogram and Boxplot of Wine Sales Purchased**




**Observations:** Figure 4 shows a histogram and boxplot of sample cases of wine that were purchased by wine distribution companies after sampling a wine (response variable). The histogram shows that there are an excess of zero counts, which suggests that a lot of wine distribution companies opted not to purchase any sample cases. There is also a few outliers around 8 sample cases of wine that were purchased. Additionally, most of the values hover around the mean of 3. The box plot also shows that the median number of sample cases of wine that were purchased is 3 and shows the outliers that are seen in the histogram. Majority of the sample cases of wine that were purchased fall in-between 2 to 4.

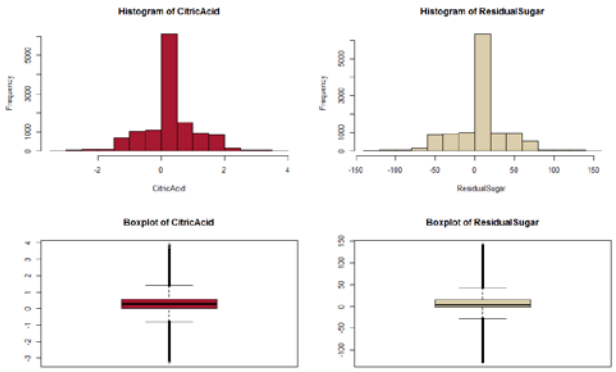
## Chemistry

**Figure 5: Histogram and Boxplot of FixedAcidity & VolatileAcidity**



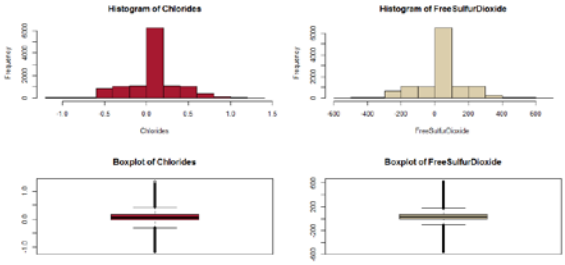

**Observations:** Figure 5 shows a histogram and boxplot of FixedAcidity & VolatileAcidity. The histogram of FixedAcidity shows a symmetric bell shape with noticeable outliers around less than -5 and greater than 20. Additionally, majority of the values hover around the mean of 7.076. The box plot of FixedAcidity also shows that the median number is 6.9 and shows the outliers that are seen in the histogram. Majority of FixedAcidity fall in-between 5.200 to 9.500. The histogram of VolatileAcidity shows a symmetric bell shape with noticeable outliers around less than -1.5 and greater than 2. Additionally, majority of the values hover around the mean of 0.3241. The box plot of VolatileAcidity also shows that the median number is 0.2800 and shows the outliers that are seen in the histogram. Majority of VolatileAcidity fall in-between 0.1300 to 0.6400.

**Figure 6: Histogram and Boxplot of CitricAcid & ResidualSugar**



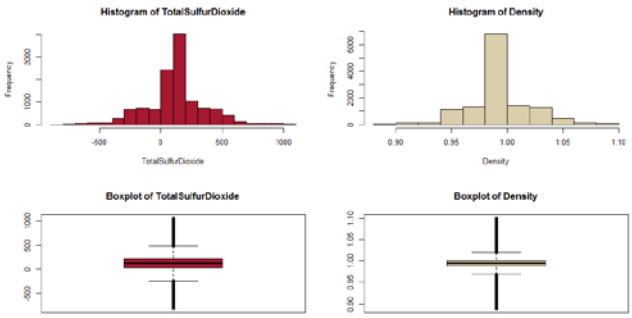

**Observations:** Figure 6 shows a histogram and boxplot of CitricAcid & ResidualSugar. The histogram of CitricAcid shows a symmetric bell shape with noticeable outliers around less than -1.5 and greater than 2. Additionally, majority of the values hover around the mean of 0.3084. The box plot of CitricAcid also shows that the median number is 0.3100 and shows the outliers that are seen in the histogram. Majority of CitricAcid fall in-between 0.0300 to 0.5800. The histogram of ResidualSugar shows a symmetric bell shape with noticeable outliers around less than -65 and greater than 65. Additionally, majority of the values hover around the mean of 5.419. The box plot of ResidualSugar also shows that the median number is 3.900 and shows the outliers that are seen in the histogram. Majority of ResidualSugar fall in-between -2.000 to 15.900.

**Figure 7: Histogram and Boxplot of Chlorides & FreeSulfurDioxide**



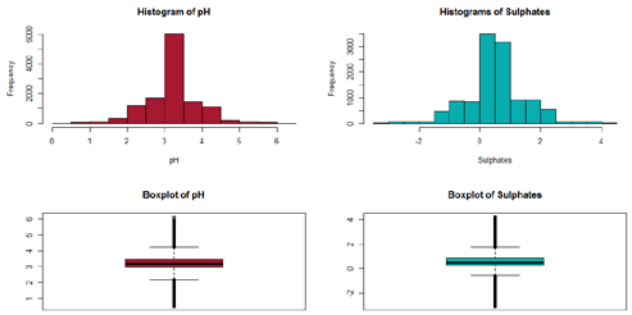

**Observations:** Figure 7 shows a histogram and boxplot of Chlorides & FreeSulfurDioxide. The histogram of Chlorides shows a symmetric bell shape with noticeable outliers around less than -0.6 and greater than 0.7. Additionally, majority of the values hover around the mean of 0.05482. The box plot of Chlorides also shows that the median number is 0.0460 and shows the outliers that are seen in the histogram. Majority of Chlorides fall in-between -0.0310 to 0.1530. The histogram of FreeSulfurDioxide shows a symmetric bell shape with noticeable outliers around less than -275 and greater than 350. Additionally, majority of the values hover around the mean of 30.85. The box plot of FreeSulfurDioxide also shows that the median number is 30.00 and shows the outliers that are seen in the histogram. Majority of FreeSulfurDioxide fall in-between 0.00 to 70.00.

**Figure 8: Histogram and Boxplot of TotalSulfurDioxide and Density**



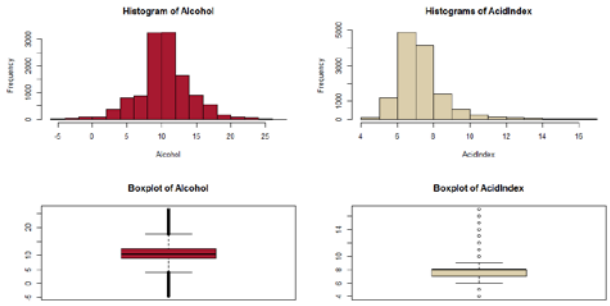

**Observations:** Figure 8 shows a histogram and boxplot of TotalSulfurDioxide and Density. The histogram of TotalSulfurDioxide shows a symmetric bell shape with noticeable outliers around less than -400 and greater than 725. Additionally, majority of the values hover around the mean of 120.7. The box plot of TotalSulfurDioxide also shows that the median number is 123.0 and shows the outliers that are seen in the histogram. Majority of TotalSulfurDioxide fall in-between 27.0 to 208.0. The histogram of Density shows a symmetric bell shape with noticeable outliers around less than 0.93 and greater than 1.06. Additionally, majority of the values hover around the mean of 0.9942. The box plot of Density also shows that the median number is 0.9945 and shows the outliers that are seen in the histogram. Majority of Density fall in-between 0.9877 to 1.0005.

**Figure 9: Histogram and Boxplot of pH & Sulphates**




**Observations:** Figure 9 shows a histogram and boxplot of pH & Sulphates. The histogram of pH shows a symmetric bell shape with noticeable outliers around less than 1.5 and greater than 4.75. Additionally, majority of the values hover around the mean of 3.208. The box plot of pH also shows that the median number is 3.200 and shows the outliers that are seen in the histogram. Majority of pH fall in-between 2.960 to 3.470. The histogram of Sulphates shows a symmetric bell shape with noticeable outliers around less than -1.5 and greater than 2.5. Additionally, majority of the values hover around the mean of 0.5271. The box plot of Sulphates also shows that the median number is 0.5000 and shows the outliers that are seen in the histogram. Majority of Sulphates fall in-between 0.2800 to 0.8600. It's also important to note that Suphates has the second most missing values in the data set.

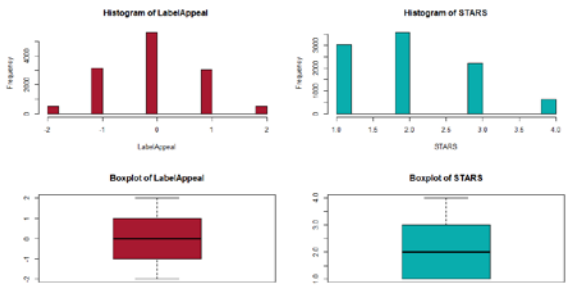
**Figure 10: Histogram and Boxplot of Alcohol & AcidIndex**




**Observations:** Figure 10 shows a histogram and boxplot of Alcohol & AcidIndex. The histogram of Alcohol shows a symmetric bell shape with noticeable outliers around less than 3 and greater than 20. Additionally, majority of the values hover around the mean of 10.49. The box plot of Alcohol also shows that the median number is 10.40 and shows the outliers that are seen in the histogram. Majority of Alcohol fall in-between 9.00 to 12.40. The histogram of AcidIndex shows a slight right skew with some outliers around less than 5 and greater than 11. Additionally, majority of the values hover around the mean of 7.773. The box plot of AcidIndex also shows that the median number is 8.000 and shows the outliers that are seen in the histogram. Majority of AcidIndex fall in-between 7.000 to 8.000. It's also important to note that AcidIndex has the second most missing values in the data set.

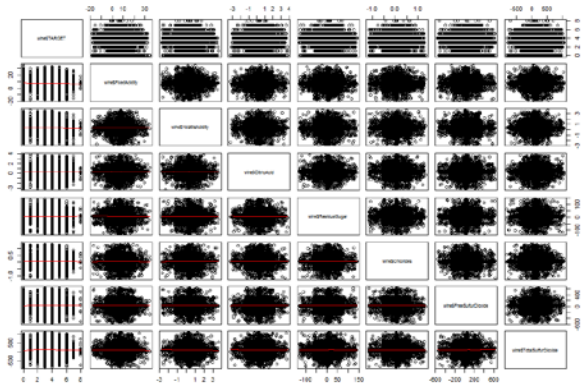
## Marketing

**Figure 11: Histogram and Boxplot of LabelAppeal & STARS**

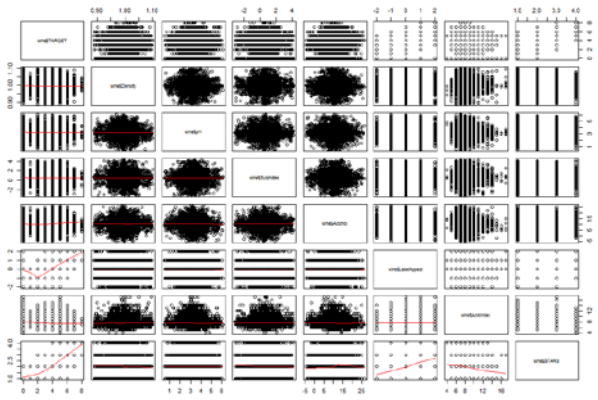



**Observations:** Figure 11 shows a histogram and boxplot of LabelAppeal & STARS. The histogram of LabelAppeal shows a symmetric bell shape. Majority of the values hover around the mean of -0.009066. The box plot of LabelAppeal also shows that the median number is 0.0. Majority of LabelAppeal fall in-between -1.0 to 1.0. The histogram of STARS shows a slight right skew. Majority of the values hover around the mean of 2.042. The box plot of STARS also shows that the median number is 2.000. Majority of STARS fall in-between 1 to 3. It's also important to note that STARS has the most missing values in the data set.

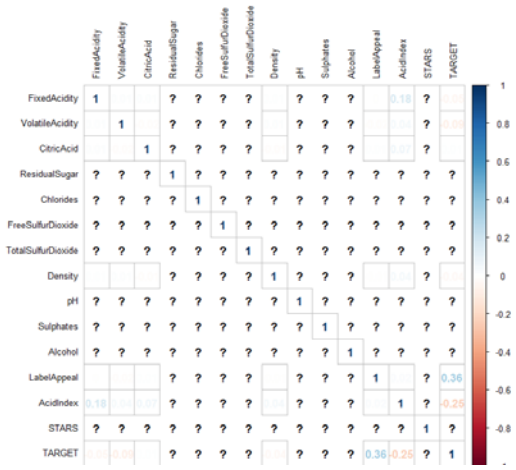
**Figure 12: Scatterplot Matrices and Correlation Matrix**



*Scatterplot Matrix*



*Scatterplot Matrix 2*




*Correlation Matrix*

**Observations:** Figure 12 shows scatterplot matrices and a correlation matrix of the variables that were included in the dataset (excluding INDEX). This gives us an idea of the most promising predictor variables based on the predictors that are most correlated with TARGET. This also allows us to see which variables may be correlated with each other (e.g., potential multicollinearity concerns). Also, note that the correlation matrix is incomplete due to the missing values for the following 8 variables: ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol, and Stars. The scatterplot matrix for most of these variables also show N/A or no correlations. The scatterplot matrix and correlation matrix shows that LabelAppeal and STARS have the strongest positive correlations with TARGET, while AcidIndex was moderate negatively correlated with TARGET. This suggests that as LabelAppeal and STARS increase, the number of sample cases of wine that are purchased also increases, which makes intuitive wine sales sense. Additionally, as AcidIndex increases, the number of sample cases of wine that were purchased decreases.

## Section 2: Data Preparation

### Figure 13: Missing Values for Variables

| INDEX | TARGET | FixedAcidity | VolatileAcidity | CitricAcid | ResidualSugar |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 616 |
| **Chlorides** | **FreeSulfurDioxide** | **TotalSulfurDioxide** | **Density** | **pH** | **Sulphates** |
| 638 | 647 | 682 | 0 | 395 | 1210 |
| **Alcohol** | **LabelAppeal** | **AcidIndex** | **STARS** | | |
| 653 | 0 | 0 | 3359 | | |

**Observations:** Figure 13 shows variables in the wine data set that have missing data. We will use the MICE package (pmm = predictive mean matching) to impute the missing data. The MICE package uses an algorithm in such a way that uses information from other variables in dataset to predict and impute the missing values. We need to address the missing values because Poisson, Binomial, and OLS regression cannot handle missing values and must be addressed prior to utilizing these modeling techniques.

### Figure 14: Percentage of Missing Values for Variables

| INDEX | TARGET | FixedAcidity | VolatileAcidity | CitricAcid | ResidualSugar |
|---|---|---|---|---|---|
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 4.814381 |
| **Chlorides** | **FreeSulfurDioxide** | **TotalSulfurDioxide** | **Density** | **pH** | **Sulphates** |
| 4.986323 | 5.056663 | 5.330207 | 0.000000 | 3.087143 | 9.456819 |
| **Alcohol** | **LabelAppeal** | **AcidIndex** | **STARS** | | |
| 5.103556 | 0.000000 | 0.000000 | **26.252442** | | |

**Observations:** Figure 14 shows the percentage of missing variables in the wine data set. STARS had the highest percentage of missing data at 26.25%.

**Figure 15: Summary of Imputation using Predictive Mean Matching**

```
Multiply imputed data set
Call:
mice(data = wine, m = 5, method = "pmm", maxit = 50, seed = 500)
Number of multiple imputations:  5
Missing cells per column:
              INDEX             TARGET        FixedAcidity     VolatileAcidity          CitricAcid
                  0                  0                   0                   0                   0
    ResidualSugar          Chlorides    FreeSulfurDioxide   TotalSulfurDioxide             Density
                616                638                 647                 682                   0
                 pH          Sulphates            Alcohol        LabelAppeal           AcidIndex
                395               1210                 653                   0                   0
              STARS    NoResidualSugar       NoChlorides  NoFreeSulfurDioxide NoTotalSulfurDioxide
               3359                  0                   0                   0                   0
              NopH       NoSulphates         NoAlcohol             NoSTARS
                 0                  0                   0                   0
Imputation methods:
              INDEX             TARGET        FixedAcidity     VolatileAcidity          CitricAcid
             "pmm"              "pmm"               "pmm"               "pmm"               "pmm"
    ResidualSugar          Chlorides    FreeSulfurDioxide   TotalSulfurDioxide             Density
             "pmm"              "pmm"               "pmm"               "pmm"               "pmm"
                pH          Sulphates            Alcohol        LabelAppeal           AcidIndex
             "pmm"              "pmm"               "pmm"               "pmm"               "pmm"
             STARS    NoResidualSugar       NoChlorides  NoFreeSulfurDioxide NoTotalSulfurDioxide
             "pmm"              "pmm"               "pmm"               "pmm"               "pmm"
              NopH       NoSulphates         NoAlcohol             NoSTARS
             "pmm"              "pmm"               "pmm"               "pmm"
```

```
Number of N/A's:
```

```
          INDEX             TARGET        FixedAcidity     VolatileAcidity          CitricAcid
              0                  0                   0                   0                   0
 ResidualSugar          Chlorides    FreeSulfurDioxide   TotalSulfurDioxide             Density
              0                  0                   0                   0                   0
             pH          Sulphates            Alcohol        LabelAppeal           AcidIndex
              0                  0                   0                   0                   0
         STARS    NoResidualSugar       NoChlorides  NoFreeSulfurDioxide NoTotalSulfurDioxide
              0                  0                   0                   0                   0
          NopH       NoSulphates         NoAlcohol             NoSTARS
              0                  0                   0                   0
```

**Observations:** Figure 15 shows imputation being applied to the missing values using predictive mean matching. The result shows that all the missing values have been replaced.

**Figure 16: Transformation of Variables**

**Discussion:** I created 8 flag variables called: NoResidualSugar, NoChlorides, NoFreeSulfurDioxide, NopH, NoSulphates, NoResidualSugar, NoAlcohol, and NoSTARS. These variables were created by assigning 0 to values that had data and assigning 1 to data that was missing. Here is the code that I used:

- wine$NoResidualSugar <- 0
- wine$NoResidualSugar [is.na(wine$ResidualSugar)] <- 1
- wine$NoChlorides  <- 0
- wine$NoChlorides [is.na(wine$Chlorides)] <- 1
- wine$NoFreeSulfurDioxide <- 0
- wine$NoFreeSulfurDioxide[is.na(wine$FreeSulfurDioxide)] <- 1
- wine$NoTotalSulfurDioxide <- 0
- wine$NoTotalSulfurDioxide[is.na(wine$TotalSulfurDioxide)] <- 1
- wine$NopH <- 0
- wine$NopH[is.na(wine$pH)] <- 1

- wine$NoSulphates <- 0
- wine$NoSulphates [is.na(wine$Sulphates)] <- 1
- wine$NoResidualSugar <- 0
- wine$NoResidualSugar [is.na(wine$ResidualSugar)] <- 1
- wine$NoAlcohol <- 0
- wine$NoAlcohol [is.na(wine$Alcohol)] <- 1
- wine$NoSTARS<- 0
- wine$NoSTARS [is.na(wine$STARS)] <- 1

I created these variables because the data set contained a lot of variables that had missing data. Additionally, there is a good chance that a variable that is missing is actually predictive of the target, which would increase the accuracy of my model. Furthermore, I also conducted log transformations on all 14 original predictor variables and SQRT transformations on AcidIndex and STARS. I conducted log transformations on these variables since these variables either contained missing data from the original data set, contained outliers, or were somewhat correlated to TARGET. I will experiment with these transformations later on in the analysis.

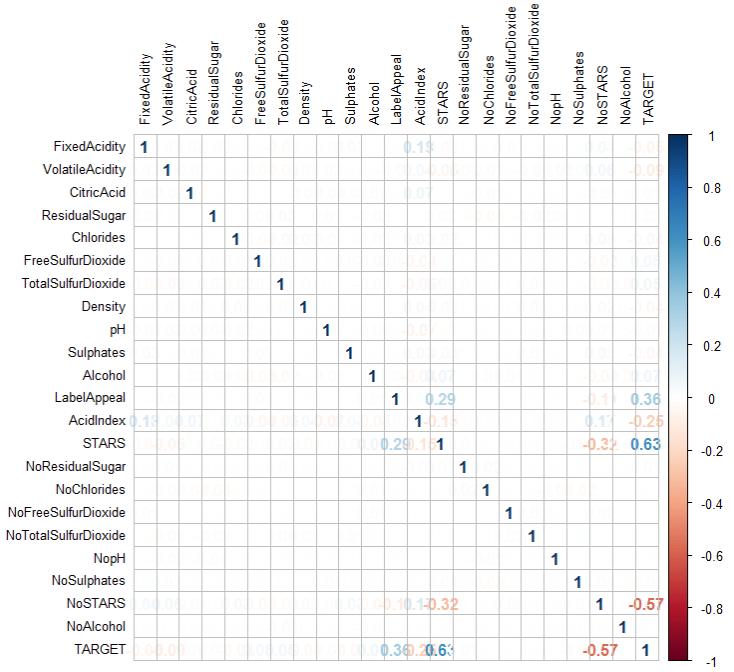
**Figure 17: Handling Outliers**

- wine2$FixedAcidity [(wine2$FixedAcidity >= 20)] = 20
- wine2$FixedAcidity [(wine2$FixedAcidity <= -5)] = -5
- wine2$VolatileAcidity [(wine2$VolatileAcidity >= 2)] = 2
- wine2$VolatileAcidity [(wine2$VolatileAcidity <= -1.5)] = -1.5
- wine2$CitricAcid [(wine2$CitricAcid >= 2)] = 2
- wine2$CitricAcid [(wine2$CitricAcid <= -1.5)] = -1.5
- wine2$ResidualSugar [(wine2$ResidualSugar >= 65)] = 65
- wine2$ResidualSugar [(wine2$ResidualSugar <= -65)] = -65
- wine2$Chlorides [(wine2$Chlorides >= 0.7)] = 0.7
- wine2$Chlorides [(wine2$Chlorides <= -0.6)] = -0.6
- wine2$FreeSulfurDioxide [(wine2$FreeSulfurDioxide >= 350)] = 350
- wine2$FreeSulfurDioxide [(wine2$FreeSulfurDioxide <= -275)] = -275
- wine2$TotalSulfurDioxide [(wine2$TotalSulfurDioxide >= 725)] = 725
- wine2$TotalSulfurDioxide [(wine2$TotalSulfurDioxide <= -400)] = -400
- wine2$Density [(wine2$Density >= 1.06)] = 1.06
- wine2$Density [(wine2$Density <= 0.93)] = 0.93
- wine2$pH [(wine2$pH >= 4.75)] = 4.75
- wine2$pH [(wine2$pH <= 1.5)] = 1.5
- wine2$Sulphates [(wine2$Sulphates >= 2.5)] = 2.5
- wine2$Sulphates [(wine2$Sulphates <= -1.5)] = -1.5
- wine2$Alcohol [(wine2$Alcohol >= 20)] = 20
- wine2$Alcohol [(wine2$Alcohol <= 3)] = 3
- wine2$AcidIndex [(wine2$AcidIndex >= 11)] = 11

- wine2$AcidIndex [(wine2$AcidIndex <= 5)] = 5

**Discussion:** Figure 17 shows how I handled the outliers from the wine data set based on the EDA in section 1 (e.g., box plots, bar graphs, and summary statistics). Addressing outliers is important because outliers can exert significant influence on model parameters. For instance, the model may be less accurate and the model may give a different interpretation or understanding that actually exists. Additionally, outliers can significantly impact a predictive model. For example, an outlier can cause a large difference in the coefficient or "Beta" value in a regression model. As a result, the primary technique that I used to handle the outliers was trimming the data (e.g., when a variable exceeds a certain limit, it is simply truncated so that it cannot exceed the limit).

**Figure 18: Correlation Matrix After Missing Values, Outliers, etc. Have Been Addressed**




**Observations:** Figure 18 shows a correlation matrix of all the numeric variables that were included in the dataset (excluding INDEX) in addition to the 8 flag variables that I created after the missing values, outliers, etc. have been addressed. This gives us an idea of the most promising predictor variables based on the predictors that are most correlated with TARGET. This also allows us to see which variables may be correlated with each other (e.g., potential multicollinearity concerns) and uncover new insights that we did not see before due to missing values. The correlation matrix shows that STARS has the strongest positive correlation with TARGET, while LabelAppeal has a moderate positively correlation with TARGET. Additionally, NoSTARS has the strongest negative correlation with TARGET, while AcidIndex was moderate negatively correlated with TARGET. This suggests that as LabelAppeal and STARS increase, the number of sample cases of wine that are purchased also increases, which makes intuitive wine sales sense. Additionally, the more wines that did not have any STARS (e.g., N/A), the number of sample cases of wine that were purchased decreases. Lastly, as

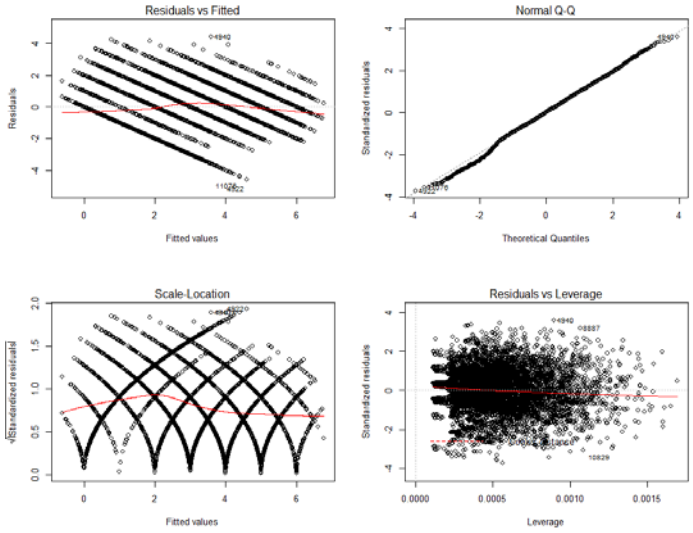
AcidIndex increases, the number of sample cases of wine that were purchased decreases. It's also interesting to note that STARS vs. LabelAppeal and STARS vs. NoSTARS were also somewhat correlated. For example, as the number of STARS increases, LabelAppeal also increases.

## Section 3: Build Models

### Model 1A, 1B, 1C: Multiple Linear Regression Models

**Figure 19: Model MLRResult1**

```
Analysis of Variance Table

Response: TARGET
                Df  Sum Sq Mean Sq  F value    Pr(>F)
STARS            1 18590.5 18590.5 12236.129 < 2.2e-16 ***
NoSTARS          1  7260.8  7260.8  4778.974 < 2.2e-16 ***
LabelAppeal      1  1464.5  1464.5   963.922 < 2.2e-16 ***
AcidIndex        1   667.8   667.8   439.524 < 2.2e-16 ***
VolatileAcidity  1    63.1    63.1    41.504 1.219e-10 ***
Residuals    12789 19430.5     1.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(MLRResult1)

Call:
lm(formula = TARGET ~ STARS + NoSTARS + LabelAppeal + AcidIndex +
    VolatileAcidity, data = wine2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5910 -0.7601  0.0300  0.8329  4.4144

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.32634    0.08090  41.115  < 2e-16 ***
STARS            0.91218    0.01350  67.567  < 2e-16 ***
NoSTARS         -1.72000    0.02640 -65.147  < 2e-16 ***
LabelAppeal      0.41561    0.01283  32.398  < 2e-16 ***
AcidIndex       -0.19613    0.00945 -20.754  < 2e-16 ***
VolatileAcidity -0.09727    0.01510  -6.442 1.22e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.233 on 12789 degrees of freedom
Multiple R-squared:  0.5907,    Adjusted R-squared:  0.5906
F-statistic:  3692 on 5 and 12789 DF,  p-value: < 2.2e-16
```




```
> coefficients(MLRResult1)
    (Intercept)           STARS         NoSTARS      LabelAppeal       AcidIndex VolatileAcidity
     3.32633715      0.91217710     -1.72000212      0.41561440     -0.19612882     -0.09726844

> vif(MLRResult1)
         STARS         NoSTARS      LabelAppeal       AcidIndex VolatileAcidity
      1.225376        1.136535        1.100393        1.045533        1.005709
```

**Observations:** Figure 19 shows a summary of the multiple regression model TARGET ~ STARS + NoSTARS + LabelAppeal + AcidIndex + VolatileAcidity. These variables were chosen based on the correlation matrix (variables that had a high correlation with TARGET). The results show that there are no multicollinearity issues in this model. A statistically significant result was obtained overall as indicated by the F-statistic which is 3692 with a p-value = < 2.2e-16. This indicates the model has produced statistically significant results to be investigated. Additionally, the t-test of all the predictor variables are statistically significant. The residual standard error of 1.233, shows us that when predicting TARGET, one standard error = 1.233. The adjusted R-squared value of 0.5906, indicates that 59.06% of the variation in TARGET is explained by the predictor variables. Furthermore, the scatterplots with residuals and qq-plots of residuals is shown so that we can check to make sure the model is meeting all the assumptions. The QQ plot reveals that the density distribution is slightly non-

normal. This is present in the plot where some of the data points are departing from the line. The scatterplot of residuals vs. fitted also shows some heteroscedascity. The plot is relatively non-linear and has a non-random scatter of data over the range of values for the independent variable.

In regards to the coefficients, the coefficients in the model makes intuitive wine sense. For instance, STARS and LabelAppeal are positive. This suggests that as LabelAppeal and STARS increase, the number of sample cases of wine that are purchased also increases, which makes intuitive wine sales sense. Additionally, the fact that NoSTARS is negative, suggests that the more wines that did not have any STARS (e.g., N/A), the number of sample cases of wine that were purchased decreases, which also make intuitive wine sales sense. Lastly, the coefficients of AcidIndex and VolatileAcidity are negative. AcidIndex seems to negatively impact the model more, while VolatileAcidity seems to negatively impact the model less.

**Figure 20: Stepwise Regression Model**

```
Analysis of Variance Table

Response: TARGET
                    Df  Sum Sq Mean Sq   F value    Pr(>F)
VolatileAcidity      1   388.9   388.9  257.2188 < 2.2e-16 ***
CitricAcid           1     4.2     4.2    2.8085   0.09379 .
Chlorides            1    75.3    75.3   49.8067 1.784e-12 ***
FreeSulfurDioxide    1   101.5   101.5   67.1327 2.779e-16 ***
TotalSulfurDioxide   1   103.2   103.2   68.2793 < 2.2e-16 ***
Density              1    54.2    54.2   35.8609 2.176e-09 ***
pH                   1     4.7     4.7    3.1348   0.07666 .
Sulphates            1    75.8    75.8   50.1099 1.530e-12 ***
Alcohol              1   218.7   218.7  144.6716 < 2.2e-16 ***
LabelAppeal          1  5952.4  5952.4 3936.6837 < 2.2e-16 ***
AcidIndex            1  2790.5  2790.5 1845.5564 < 2.2e-16 ***
STARS                1 12026.7 12026.7 7954.0077 < 2.2e-16 ***
NoFreeSulfurDioxide  1     0.3     0.3    0.1882   0.66443
NoAlcohol            1     3.8     3.8    2.5242   0.11214
NoSTARS              1  6354.4  6354.4 4202.5467 < 2.2e-16 ***
Residuals        12779 19322.3     1.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

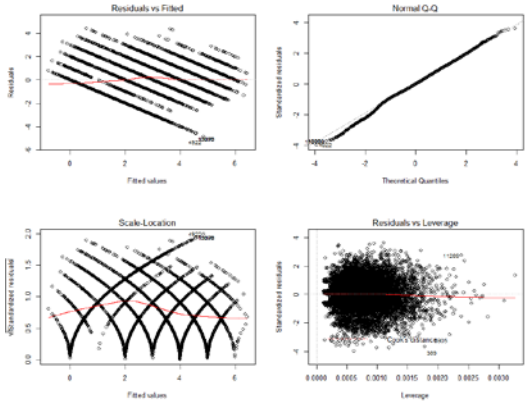
```
Call:
lm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
    FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
    Alcohol + LabelAppeal + AcidIndex + STARS + NoFreeSulfurDioxide +
    NoAlcohol + NoSTARS, data = wine2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5713 -0.7725  0.0252  0.8321  4.3620

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.009e+00  4.449e-01   9.011  < 2e-16 ***
VolatileAcidity     -9.470e-02  1.507e-02  -6.282 3.44e-10 ***
CitricAcid           2.234e-02  1.396e-02   1.600 0.109528
Chlorides           -1.073e-01  3.769e-02  -2.846 0.004433 **
FreeSulfurDioxide    3.378e-04  8.026e-05   4.209 2.59e-05 ***
TotalSulfurDioxide   1.923e-04  4.979e-05   3.861 0.000114 ***
Density             -7.259e-01  4.372e-01  -1.660 0.096903 .
pH                  -3.670e-02  1.710e-02  -2.147 0.031835 *
Sulphates           -2.787e-02  1.267e-02  -2.199 0.027866 *
Alcohol              1.018e-02  3.131e-03   3.252 0.001150 **
LabelAppeal          4.169e-01  1.280e-02  32.565  < 2e-16 ***
AcidIndex           -1.929e-01  9.506e-03 -20.291  < 2e-16 ***
STARS                9.079e-01  1.350e-02  67.272  < 2e-16 ***
NoFreeSulfurDioxide  7.574e-02  4.963e-02   1.526 0.127004
NoAlcohol            8.871e-02  4.942e-02   1.795 0.072643 .
NoSTARS             -1.710e+00  2.638e-02 -64.827  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.23 on 12779 degrees of freedom
Multiple R-squared:  0.593,     Adjusted R-squared:  0.5925
F-statistic:  1241 on 15 and 12779 DF,  p-value: < 2.2e-16
```

```
> coefficients(stepwise)
        (Intercept)     VolatileAcidity          CitricAcid           Chlorides   FreeSulfurDioxide
       4.0092222788       -0.0946981011        0.0223424360       -0.1072681705        0.0003378100
 TotalSulfurDioxide             Density                  pH           Sulphates             Alcohol
       0.0001922511       -0.7259117105       -0.0367001148       -0.0278653689        0.0101821196
        LabelAppeal           AcidIndex                STARS NoFreeSulfurDioxide           NoAlcohol
       0.4169459883       -0.1928867784        0.9078705637        0.0757387133        0.0887123910
            NoSTARS
      -1.7101455250

> vif(stepwise)
    VolatileAcidity          CitricAcid           Chlorides   FreeSulfurDioxide  TotalSulfurDioxide
           1.007264            1.006408            1.003706            1.004370            1.005326
            Density                  pH           Sulphates             Alcohol         LabelAppeal
           1.003319            1.006126            1.002870            1.008306            1.101391
          AcidIndex                STARS NoFreeSulfurDioxide           NoAlcohol             NoSTARS
           1.063014            1.230372            1.000608            1.000772            1.140117
```

**Observations:** Figure 20 shows a summary of the stepwise model: TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS + NoFreeSulfurDioxide + NoAlcohol + NoSTARS. These variables were chosen using stepwise regression. The results show that there are no multicollinearity issues in this model. A statistically significant result was obtained overall as indicated by the F-statistic which is 1241 with a p-value = < 2.2e-16. This indicates the model has produced statistically significant results to be investigated. The residual standard error of 1.23, shows us that when predicting TARGET, one standard error = 1.23. The adjusted R-squared value of 0.5925, indicates that 59.25% of the variation in TARGET is explained by the predictor variables. The ANOVA table and summary table shows that the following variables were not highly significant: CitricAcid, Density, pH, Sulphates, NoFreeSulfurDioxide, and NoAlcohol. These variables will be removed from the next model to create a more parsimonious model. I will also experiment with log transformations on some of the variables as well.

### Figure 21: Reduced Stepwise Model + Transformations

```
Analysis of Variance Table

Response: TARGET
                     Df  Sum Sq Mean Sq  F value    Pr(>F)
VolatileAcidity       1   388.9   388.9  259.782 < 2.2e-16 ***
FreeSulfurDioxide     1   103.9   103.9   69.423 < 2.2e-16 ***
TotalSulfurDioxide    1   106.9   106.9   71.375 < 2.2e-16 ***
Chlorides             1    69.8    69.8   46.597 9.111e-12 ***
Alcohol               1   221.7   221.7  148.087 < 2.2e-16 ***
LabelAppeal           1  5962.5  5962.5 3982.653 < 2.2e-16 ***
logAcidIndex          1  2631.9  2631.9 1757.974 < 2.2e-16 ***
logSTARS              1 12847.8 12847.8 8581.728 < 2.2e-16 ***
NoSTARS               1  6003.1  6003.1 4009.796 < 2.2e-16 ***
Residuals         12785 19140.6     1.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(Model3)

Call:
lm(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide + TotalSulfurDioxide +
    Chlorides + Alcohol + LabelAppeal + logAcidIndex + logSTARS +
    NoSTARS, data = wine2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8476 -0.7552  0.0148  0.8212  4.4136

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.256e+00  1.505e-01  34.923  < 2e-16 ***
VolatileAcidity    -9.489e-02  1.499e-02  -6.329 2.55e-10 ***
FreeSulfurDioxide   3.134e-04  7.984e-05   3.925 8.70e-05 ***
TotalSulfurDioxide  1.946e-04  4.953e-05   3.929 8.56e-05 ***
Chlorides          -1.100e-01  3.748e-02  -2.934 0.003357 **
Alcohol             1.134e-02  3.114e-03   3.642 0.000272 ***
LabelAppeal         4.267e-01  1.268e-02  33.653  < 2e-16 ***
logAcidIndex       -1.372e+00  7.039e-02 -19.497  < 2e-16 ***
logSTARS            1.746e+00  2.528e-02  69.035  < 2e-16 ***
NoSTARS            -1.670e+00  2.637e-02 -63.323  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.224 on 12785 degrees of freedom
Multiple R-squared:  0.5968,    Adjusted R-squared:  0.5966
F-statistic:  2103 on 9 and 12785 DF,  p-value: < 2.2e-16
```




```
> coefficients(Model3)
       (Intercept)    VolatileAcidity  FreeSulfurDioxide TotalSulfurDioxide          Chlorides            Alcohol
      5.2555134376      -0.0948892009       0.0003134193       0.0001946357      -0.1099559032       0.0113400481
       LabelAppeal       logAcidIndex            logSTARS            NoSTARS
      0.4266972658      -1.3724036164       1.7455407518      -1.6701325978

> vif(Model3)
   VolatileAcidity  FreeSulfurDioxide TotalSulfurDioxide          Chlorides            Alcohol        LabelAppeal
          1.006484           1.003777           1.004735           1.002598           1.006919           1.090875
      logAcidIndex            logSTARS            NoSTARS
          1.048137           1.234412           1.15101
```

**Observations:** Figure 21 shows a summary of the reduced stepwise model with transformations: TARGET ~ VolatileAcidity + FreeSulfurDioxide + TotalSulfurDioxide + Chlorides + Alcohol + LabelAppeal + logAcidIndex + logSTARS + NoSTARS. The results show that there are no multicollinearity issues in this model. A statistically significant result was obtained overall as indicated by the F-statistic which is 2103 with a p-value = < 2.2e-16. This indicates the model has produced statistically significant results to be investigated. Additionally, the t-test of all the predictor variables are statistically significant. The residual standard error of 1.224, shows us that when predicting TARGET, one standard error = 1.224. The adjusted R-squared value of 0.5966, indicates that 59.66% of the variation in TARGET is explained by the predictor variables. Furthermore, the scatterplots with residuals and qq-plots of residuals is shown so that we can check to make sure the model is meeting all the assumptions. The QQ plot reveals that the density distribution is slightly non-normal. This is present in the plot where some of the data points are departing from the line. The scatterplot of residuals vs. fitted also shows some heteroscedascity. The plot is relatively non-linear and has a non-random scatter of data over the range of values for the independent variable. This is similar to what we saw in the previous multiple linear regression models.

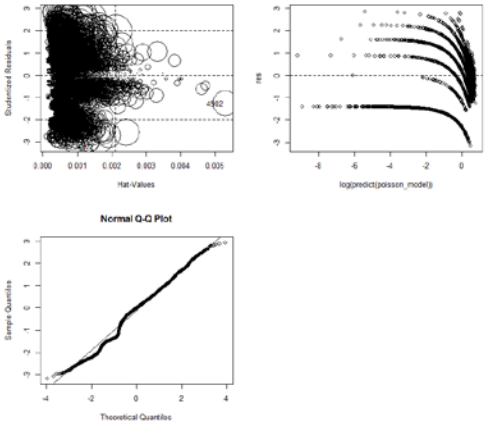
In regards to the coefficients, the coefficients in the model makes intuitive wine sense. For instance, logSTARS and LabelAppeal are positive, similar to Model 1. This suggests that as LabelAppeal and STARS increase, the number of sample cases of wine that are purchased also increases, which makes intuitive wine sales sense. Additionally, the fact that NoSTARS is negative, suggests that the more wines that did not have any STARS (e.g., N/A), the number of sample cases of wine that were purchased decreases, which also make intuitive wine sales sense. Furthermore, the coefficients of FreeSulfurDioxide, TotalSulfurDioxide, and Alcohol are positive, while VolatileAcidity, Chlorides, and logAcidIndex are negative. Out of these variables, logAcidIndex seems to have the highest negative impact on the model.

## Model 2: Poisson Regression (Figure 22)

```
Analysis of Deviance Table

Model: poisson, link: log

Response: TARGET

Terms added sequentially (first to last)


                   Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                12794      22861
VolatileAcidity     1    128.3     12793      22733 < 2.2e-16 ***
FreeSulfurDioxide   1     34.3     12792      22698 4.766e-09 ***
TotalSulfurDioxide  1     35.4     12791      22663 2.718e-09 ***
Chlorides           1     23.0     12790      22640 1.617e-06 ***
LabelAppeal         1   1979.0     12789      20661 < 2.2e-16 ***
logAcidIndex        1    922.8     12788      19738 < 2.2e-16 ***
logSTARS            1   4211.0     12787      15527 < 2.2e-16 ***
NoSTARS             1   2907.2     12786      12620 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Call:
glm(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide +
    TotalSulfurDioxide + Chlorides + LabelAppeal + logAcidIndex +
    logSTARS + NoSTARS, family = poisson(link = "log"), data = wine2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1798  -0.7183  -0.0203   0.4812   2.9194

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.983e+00  7.339e-02  27.018  < 2e-16 ***
VolatileAcidity    -3.048e-02  7.078e-03  -4.306 1.66e-05 ***
FreeSulfurDioxide   1.139e-04  3.747e-05   3.040  0.00236 **
TotalSulfurDioxide  6.887e-05  2.335e-05   2.949  0.00318 **
Chlorides          -3.875e-02  1.766e-02  -2.195  0.02819 *
LabelAppeal         1.319e-01  6.103e-03  21.615  < 2e-16 ***
logAcidIndex       -5.333e-01  3.570e-02 -14.939  < 2e-16 ***
logSTARS            5.541e-01  1.206e-02  45.943  < 2e-16 ***
NoSTARS            -8.554e-01  1.741e-02 -49.121  < 2e-16 ***
---
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: 12620  on 12786  degrees of freedom
AIC: 44580

Number of Fisher Scoring iterations: 6
```

```
> with(poisson_model, cbind(res.deviance = deviance, df = df.residual,
+                           p = pchisq(deviance, df.residual, lower.tail=FALSE)))
     res.deviance    df         p
[1,]     12619.97 12786 0.8505122
>
> library(AER)
> deviance(poisson_model)/poisson_model$df.residual
[1] 0.9870144
> dispersiontest(poisson_model)

        Overdispersion test

data:  poisson_model
z = -17.924, p-value = 1
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
 0.8100002

>
> #what type of dispersion does sample have?
> mean(wine2$TARGET)
[1] 3.029074
> var(wine2$TARGET)
[1] 3.710895
```




**Observations:** Figure 22 shows a summary of the poisson model VolatileAcidity + FreeSulfurDioxide + TotalSulfurDioxide + Chlorides + LabelAppeal + logAcidIndex + logSTARS + NoSTARS, which is used to model count variables (mean = variance). These variables were chosen based on the reduced stepwise regression seen in the multiple regression models, but with Alcohol removed. The model also has log transformations on AcidIndex and STARS. The deviance of residuals, which is a measure of model fit of a generalized linear model, shows that the null deviance is 22861 and the residual deviance is 12620. We can use the residual deviance to perform a goodness of fit test for the overall model. As a result, since the goodness-of-fit chi-squared test is not statistically significant we conclude that the model fits reasonably well. The Analysis of Deviance table shows the difference between the null deviance and the residual deviance (wider the gap, the better). The table shows that logSTARS, LabelAppeal, NoSTARS, and logAcidIndex significantly reduces the residual deviance, whereas variables such as Chlorides, FreeSulfurDioxide, and TotalSulfurDioxide seem to improve the model less as indicated by the low deviance and large p-values (without the variable explains more or less the same amount of variation). Furthermore, the results show that the data is not overdispersed

(as indicated by the dispersion test). The QQ plot also shows that the data is relatively normal, with most of the residuals falling on the diagonal line. The results also show an AIC of 44580.

In regards to the coefficients, the coefficients in the model makes intuitive wine sense. For instance, logSTARS and LabelAppeal are positive. This suggests that as LabelAppeal and STARS increase, the number of sample cases of wine that are purchased also increases, which makes intuitive wine sales sense. Additionally, the fact that NoSTARS is negative, suggests that the more wines that did not have any STARS (e.g., N/A), the number of sample cases of wine that were purchased decreases, which also make intuitive wine sales sense. Furthermore, the coefficients of FreeSulfurDioxide and TotalSulfurDioxide are positive, while VolatileAcidity, Chlorides, and logAcidIndex are negative. Overall, this is similar to what we saw in the multiple regression models.

## Model 3: Negative Binomial Regression (Figure 23)

```
Call:
glm.nb(formula = TARGET ~ VolatileAcidity + LabelAppeal + SQRT_AcidIndex +
    logSTARS + NoSTARS, data = wine2, init.theta = 45669.57633,
    link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.15046  -0.71493  -0.02147   0.47621   2.86393

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.013120   0.071304   28.23  < 2e-16 ***
VolatileAcidity -0.030782   0.007076   -4.35 1.36e-05 ***
LabelAppeal      0.132246   0.006103   21.67  < 2e-16 ***
SQRT_AcidIndex  -0.398184   0.025490  -15.62  < 2e-16 ***
logSTARS         0.553391   0.012059   45.89  < 2e-16 ***
NoSTARS         -0.855640   0.017415  -49.13  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(45669.58) family taken to be 1)

    Null deviance: 22860  on 12794  degrees of freedom
Residual deviance: 12625  on 12789  degrees of freedom
AIC: 44582

Number of Fisher Scoring iterations: 1

              Theta:  45670
          Std. Err.:  39444
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -44568.13

> odTest(NBR_Model)
Likelihood ratio test of H0: Poisson, as restricted NB model:
n.b., the distribution of the test-statistic under H0 is non-standard
e.g., see help(odTest) for details/references

Critical value of test statistic at the alpha= 0.05 level: 2.7055
Chi-Square Test Statistic =  -0.3973 p-value = 0.5
```

**Observations:** Figure 23 shows a summary of the negative binomial model TARGET ~ VolatileAcidity + LabelAppeal + SQRT_AcidIndex + logSTARS + NoSTARS, used for modeling count variables, usually for over-dispersed count outcome variables (variance > mean). These variables were chosen based on the correlation matrix (variables that had a high correlation with TARGET). The model also has SQRT transformations on AcidIndex and log transformations on STARS. Please note that I ran both the Poisson and Negative Binomial models prior to running this model and I received the same results. As a result, I decided to remove FreeSulfurDioxide, TotalSulfurDioxide, Chlorides, and Alcohol and transformed AcidIndex using SQRT so that I could get a different model. The odTest compares the log-likelihood ratios of a Negative Binomial regression to the restriction of a Poisson regression

mean=variance. The results show that we should accept the Poisson regression model because the test statistic of -0.3973 is less than 2.7055 with a p-value of 0.5. The deviance of residuals, which is a measure of model fit of a generalized linear model, shows that the null deviance is 22860 and the residual deviance is 12625. The results also show an AIC of 44582, 2*log likelihood of -44568.13, and Theta of 45670. Additionally, it's important to highlight that one common cause of over-dispersion is excess zeros. As a result, zero-inflated models will be generated and will be validated to see if it's better than the standard models.

In regards to the coefficients, the coefficients in the model makes intuitive wine sense. For instance, logSTARS and LabelAppeal are positive. This suggests that as LabelAppeal and STARS increase, the number of sample cases of wine that are purchased also increases, which makes intuitive wine sales sense. Additionally, the fact that NoSTARS is negative, suggests that the more wines that did not have any STARS (e.g., N/A), the number of sample cases of wine that were purchased decreases, which also make intuitive wine sales sense. This is similar to the other models. Lastly, the coefficients of SQRT_AcidIndex and VolatileAcidity are negative.  AcidIndex seems to negatively impact the model more, while VolatileAcidity seems to negatively impact the model less. Interestingly, the coefficients in this model are a lot smaller than the previous models.

## Model 4: Zero Inflated Poisson Regression (Figure 24)

```
> summary(ZIP_Model)

Call:
zeroinfl(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide + TotalSulfurDioxide + Chlorides + Alco
hol + LabelAppeal + logAcidIndex + logSTARS + NoSTARS, data = wine2)

Pearson residuals:
     Min       1Q   Median       3Q      Max
-2.20110 -0.34015 -0.02626  0.35421  6.52093

Count model coefficients (poisson with log link):
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.404e+00  8.054e-02  17.431  < 2e-16 ***
VolatileAcidity    -1.374e-02  7.278e-03  -1.889   0.0589 .
FreeSulfurDioxide   3.427e-05  3.817e-05   0.898   0.3693
TotalSulfurDioxide -1.725e-05  2.345e-05  -0.736   0.4619
Chlorides          -2.440e-02  1.808e-02  -1.349   0.1772
Alcohol             7.024e-03  1.499e-03   4.684 2.81e-06 ***
LabelAppeal         2.269e-01  6.342e-03  35.768  < 2e-16 ***
logAcidIndex       -1.542e-01  3.828e-02  -4.028 5.63e-05 ***
logSTARS            2.374e-01  1.276e-02  18.604  < 2e-16 ***
NoSTARS            -1.741e-01  1.839e-02  -9.468  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -9.6131557  0.5502345 -17.471  < 2e-16 ***
VolatileAcidity     0.2224465  0.0547646   4.062 4.87e-05 ***
FreeSulfurDioxide  -0.0009165  0.0003002  -3.053 0.002267 **
TotalSulfurDioxide -0.0012227  0.0001864  -6.560 5.36e-11 ***
Chlorides           0.0214782  0.1380551   0.156 0.876367
Alcohol             0.0436926  0.0114584   3.813 0.000137 ***
LabelAppeal         1.0474991  0.0543345  19.279  < 2e-16 ***
logAcidIndex        3.5871868  0.2503873  14.327  < 2e-16 ***
logSTARS           -4.3227736  0.1466999 -29.467  < 2e-16 ***
NoSTARS             3.3623327  0.0950601  35.371  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 27
Log-likelihood: -1.989e+04 on 20 Df

> 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                   -46.01624 model2 > model1 < 2.22e-16
AIC-corrected         -45.82332 model2 > model1 < 2.22e-16
BIC-corrected         -45.10404 model2 > model1 < 2.22e-16
```

**Observations:** Figure 24 shows a summary of the zero-inflated poisson regression model TARGET ~ VolatileAcidity + FreeSulfurDioxide + TotalSulfurDioxide + Chlorides + Alcohol + LabelAppeal + logAcidIndex + logSTARS + NoSTARS, which is used to model count data that has an excess of zero

counts and when the data is not overdispersed (e.g., when the variance is not much larger than the mean count). These variables were chosen based on the reduced stepwise regression seen in the multiple regression models. The model also has log transformations on AcidIndex and STARS. The log-likehood of the model is -1.989e+04 on 20 Df. The vuong test compares the zero-inflated model with a standard Poisson regression model. The vuong test shows that our test statistic is significant, indicating that the zero-inflated model is an improvement over the standard Poisson model.

In regards to the coefficients, the coefficients in the model makes intuitive wine sense. For instance, logSTARS and LabelAppeal are positive. This suggests that as LabelAppeal and STARS increase, the number of sample cases of wine that are purchased also increases, which makes intuitive wine sales sense. Additionally, the fact that NoSTARS is negative, suggests that the more wines that did not have any STARS (e.g., N/A), the number of sample cases of wine that were purchased decreases, which also make intuitive wine sales sense. This is similar to the other models. Furthermore, the coefficients of FreeSulfurDioxide and Alcohol are positive, while VolatileAcidity, Chlorides, and logAcidIndex are negative. Overall, this is similar to what we saw in the other models. However, one slight difference is that the coefficient of TotalSulfurDioxide is negative, whereas TotalSulfurDioxide was positive in the other models.

## Model 5: Zero Inflated Negative Binomial Regression (Figure 25)

```
Call:
zeroinfl(formula = TARGET ~ Alcohol + LabelAppeal + logAcidIndex + logSTARS + NoSTARS, data = wine2,
    dist = "negbin", EM = TRUE)

Pearson residuals:
     Min       1Q   Median       3Q      Max
-2.14445 -0.34500 -0.02292  0.35787  6.80174

Count model coefficients (negbin with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.406393   0.080408  17.491  < 2e-16 ***
Alcohol       0.006989   0.001498   4.664 3.10e-06 ***
LabelAppeal   0.227486   0.006340  35.884  < 2e-16 ***
logAcidIndex -0.158333   0.038283  -4.136 3.54e-05 ***
logSTARS      0.237580   0.012766  18.610  < 2e-16 ***
NoSTARS      -0.175839   0.018400  -9.556  < 2e-16 ***
Log(theta)   12.271800   3.793535   3.235  0.00122 **

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.95510    0.54814 -18.162  < 2e-16 ***
Alcohol       0.04621    0.01135   4.072 4.66e-05 ***
LabelAppeal   1.04820    0.05380  19.482  < 2e-16 ***
logAcidIndex  3.70673    0.25002  14.826  < 2e-16 ***
logSTARS     -4.30864    0.14542 -29.628  < 2e-16 ***
NoSTARS       3.34552    0.09371  35.700  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 213587.0809
Number of iterations in BFGS optimization: 1
Log-likelihood: -1.994e+04 on 13 Df

> 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                  -45.73077 model2 > model1 < 2.22e-16
AIC-corrected        -45.61439 model2 > model1 < 2.22e-16
BIC-corrected        -45.18049 model2 > model1 < 2.22e-16

> #what type of dispersion does sample have?
> mean(wine2$TARGET)
[1] 3.029074
> var(wine2$TARGET)
[1] 3.710895
```

**Observations:** Figure 25 shows a summary of the zero-inflated negative binomial regression model TARGET ~ Alcohol + LabelAppeal + logAcidIndex + logSTARS + NoSTARS, which is used to model count data that has an excess of zero counts and is usually for overdispersed outcome variables (e.g., when the variance is much larger than the mean). These variables were chosen based on the correlation

matrix (variables that had a high correlation with TARGET), but with Alcohol substituted for VolatileAcidity. The model also has log transformations on AcidIndex and STARS. Please note that I ran both the Poisson and Negative Binomial models prior to running this model and I received the same results. As a result, I decided to remove FreeSulfurDioxide, TotalSulfurDioxide and Chlorides so that I could get a different model. The log-likelihood of the model is -1.994e+04 on 13 Df, while the theta is 213587.0809. The vuong test compares the zero-inflated model with a standard negative binomial regression model. The vuong test shows that our test statistic is significant, indicating that the zero-inflated negative binomial regression model is an improvement over the standard negative binomial regression model.

In regards to the coefficients, the coefficients in the model makes intuitive wine sense. For instance, logSTARS and LabelAppeal are positive. This suggests that as LabelAppeal and STARS increase, the number of sample cases of wine that are purchased also increases, which makes intuitive wine sales sense. Additionally, the fact that NoSTARS is negative, suggests that the more wines that did not have any STARS (e.g., N/A), the number of sample cases of wine that were purchased decreases, which also make intuitive wine sales sense. This is similar to the other models. Lastly, the coefficients of logAcidIndex is negative, while Alcohol is positive. AcidIndex seems to negatively impact the model more, while Alcohol seems to positively impact the model less. This is also similar to what we saw in the other models.

## Section 4: Selection Models

### Figure 26: Model Comparison and Criteria for Selecting the "Best Model"

| Model Name | AIC | Rank | MSE | Rank | Handles Excess Zeros? | Total Points & Best Model* |
|---|---|---|---|---|---|---|
| MLRResult1 | 41670.28 | 5 | 1.518604 | 7 | No | 4 |
| stepwisemodel | 41628.22 | 4 | 1.509604 | 6 | No | 6 |
| Model3 | 41485.94 | 3 | 1.495947 | 5 | No | 8 |
| poisson_model | 44579.99 | 6 | 0.5045696 | 1 | No | 9 |
| NBR_Model | 44582.13 | 7 | 0.5055159 | 2 | No | 7 |
| ZIP_Model | 39828.88 | 1 | 1.338761 | 3 | Yes | 13 |
| ZINB_Model | 39896.1 | 2 | 1.350946 | 4 | Yes | 11 |

**Points: Rank 1 = 7 points, Rank 2 = 6 points, Rank 3 = 5 points, Rank 4 = 4 point, Rank 5 = 3 points, Rank 6 = 2 points, Rank 7 = 1 point; Yes =2, No = 1*

**Observations:** Figure 26 shows the model comparisons so that we can compare the in-sample fit and predictive accuracy of our models so that we can select the best model. The results above show the computations for AIC and mean squared error for each of these models. Each of these metrics represent some concept of 'fit' (e.g., rewarding for accuracy). Additionally, each model was ranked on each metric. Points were then allotted to each model based on how they ranked on each metric. As a result, given the criteria above. ZIP_Model is the best model because it ranked in the upper echelon on all the metrics and as a result received the most total points (e.g., was the most accurate and most parsimonious model). Additionally, the analysis in the modeling building section shows that the data is not overdispersed (e.g., when the variance is not much larger than the mean count) as indicated by the dispersion test. Furthermore, the vuong test shows that our test statistic was significant, which indicates that the zero-inflated model (ZIP_Model) is an improvement over the standard Poisson model.

The formula/coefficients for the ZIP_Model to predict the number of wine cases that will be sold based on certain properties/characteristics of the wine is:

```
> coefficients(ZIP_Model)
      count_(Intercept)    count_VolatileAcidity  count_FreeSulfurDioxide count_TotalSulfurDioxide          count_Chlorides
           1.403925e+00            -1.374487e-02             3.427232e-05            -1.725135e-05            -2.439722e-02
            count_Alcohol        count_LabelAppeal       count_logAcidIndex          count_logSTARS            count_NoSTARS
            7.023510e-03             2.268554e-01            -1.541670e-01             2.374461e-01            -1.741359e-01
```

In regards to the coefficients, the coefficients in the model makes intuitive wine sense. For instance, logSTARS and LabelAppeal are positive. This suggests that as LabelAppeal and STARS increase, the number of sample cases of wine that are purchased also increases, which makes intuitive wine sales sense. Additionally, the fact that NoSTARS is negative, suggests that the more wines that did not have any STARS (e.g., N/A), the number of sample cases of wine that were purchased decreases, which also make intuitive wine sales sense. This is similar to the other models. Furthermore, the coefficients of FreeSulfurDioxide and Alcohol are positive, while VolatileAcidity, Chlorides, and logAcidIndex are negative. Overall, this is similar to what we saw in the other models. However, one slight difference is that the coefficient of TotalSulfurDioxide is negative, whereas TotalSulfurDioxide was positive in the other models.

## Section 5: Stand Alone Scoring Program

```
#Part 5: Test Data

wine_test=read.csv("wine_test.csv",header=T)

#Part 2: Data Preparation
library(mice)

#Check for missing values
sapply(wine_test, function(x) sum(is.na(x)))

#Check missing data percentage
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(wine_test,2,pMiss)

#Create Flag Variables (Missing Data)

wine_test$NoResidualSugar <- 0
wine_test$NoResidualSugar [is.na(wine_test$ResidualSugar)] <- 1

wine_test$NoChlorides  <- 0
wine_test$NoChlorides [is.na(wine_test$Chlorides)] <- 1

wine_test$NoFreeSulfurDioxide <- 0
wine_test$NoFreeSulfurDioxide[is.na(wine_test$FreeSulfurDioxide)] <- 1

wine_test$NoTotalSulfurDioxide <- 0
wine_test$NoTotalSulfurDioxide[is.na(wine_test$TotalSulfurDioxide)] <- 1

wine_test$NopH <- 0
wine_test$NopH[is.na(wine_test$pH)] <- 1

wine_test$NoSulphates <- 0
wine_test$NoSulphates [is.na(wine_test$Sulphates)] <- 1

wine_test$NoResidualSugar <- 0
wine_test$NoResidualSugar [is.na(wine_test$ResidualSugar)] <- 1

wine_test$NoAlcohol <- 0
wine_test$NoAlcohol [is.na(wine_test$Alcohol)] <- 1

wine_test$NoSTARS<- 0
wine_test$NoSTARS [is.na(wine_test$STARS)] <- 1

str(wine_test)
```

**#Run imputation**

```
tempData <- mice(wine_test,m=5,maxit=50,meth='pmm',seed=500)
summary(tempData)
```

**#Check N/A values have been removed**

```
wine3 <- complete(tempData,1)
apply(wine3,2,pMiss)
summary(wine3)

densityplot(tempData)
```

**#Straighten Relationships – Create transformed variables that we can look at later**

```
wine3$logFixedAcidity <- log(wine3$FixedAcidity)
wine3$logVolatileAcidity <- log(wine3$VolatileAcidity)
wine3$logCitricAcid <- log(wine3$CitricAcid)
wine3$logResidualSugar <- log(wine3$ResidualSugar)
wine3$logChlorides <- log(wine3$Chlorides)
wine3$logFreeSulfurDioxide <- log(wine3$FreeSulfurDioxide)
wine3$logTotalSulfurDioxide <- log(wine3$TotalSulfurDioxide)
wine3$logDensity <- log(wine3$Density)
wine3$logpH <- log(wine3$pH)
wine3$logSulphates <- log(wine3$Sulphates)
wine3$logAlcohol <- log(wine3$Alcohol)
wine3$logLabelAppeal <- log(wine3$LabelAppeal)
wine3$logAcidIndex <- log(wine3$AcidIndex)
wine3$logSTARS <- log(wine3$STARS)
```

**#Create SQRT Transformations of Some of the Variables**

```
wine3$SQRT_STARS <- sqrt(wine3$ STARS)
wine3$SQRT_AcidIndex <- sqrt(wine3$AcidIndex)
```

**#Trim Data**

```
wine3$FixedAcidity [(wine3$FixedAcidity >= 20)] = 20
wine3$FixedAcidity [(wine3$FixedAcidity <= -5)] = -5

wine3$VolatileAcidity [(wine3$VolatileAcidity >= 2)] = 2
wine3$VolatileAcidity [(wine3$VolatileAcidity  <= -1.5)] = -1.5

wine3$CitricAcid [(wine3$CitricAcid >= 2)] = 2
wine3$CitricAcid [(wine3$CitricAcid <= -1.5)] = -1.5

wine3$ResidualSugar [(wine3$ResidualSugar >= 65)] = 65
wine3$ResidualSugar [(wine3$ResidualSugar <= -65)] = -65

wine3$Chlorides [(wine3$Chlorides >= 0.7)] = 0.7
wine3$Chlorides [(wine3$Chlorides <= -0.6)] = -0.6
```

```
wine3$FreeSulfurDioxide [(wine3$FreeSulfurDioxide >= 350)] = 350
wine3$FreeSulfurDioxide [(wine3$FreeSulfurDioxide <= -275)] = -275

wine3$TotalSulfurDioxide [(wine3$TotalSulfurDioxide >= 725)] = 725
wine3$TotalSulfurDioxide [(wine3$TotalSulfurDioxide <= -400)] = -400

wine3$Density [(wine3$Density >= 1.06)] = 1.06
wine3$Density [(wine3$Density <= 0.93)] = 0.93

wine3$pH [(wine3$pH >= 4.75)] = 4.75
wine3$pH [(wine3$pH <= 1.5)] = 1.5

wine3$Sulphates [(wine3$Sulphates >= 2.5)] = 2.5
wine3$Sulphates [(wine3$Sulphates <= -1.5)] = -1.5

wine3$Alcohol [(wine3$Alcohol >= 20)] = 20
wine3$Alcohol [(wine3$Alcohol <= 3)] = 3

wine3$AcidIndex [(wine3$AcidIndex >= 11)] = 11
wine3$AcidIndex [(wine3$AcidIndex <= 5)] = 5

summary(wine3)

# Stand Alone Scoring

ZIP_Model<-zeroinfl(TARGET ~ VolatileAcidity + FreeSulfurDioxide + TotalSulfurDioxide + Chlorides + Alcohol +
LabelAppeal + logAcidIndex + logSTARS + NoSTARS, data=wine2)

wine3$P_TARGET <- predict(ZIP_Model, newdata = wine3, type = "response")

summary(wine3)

select <- dplyr::select


# Scored Data File
scores <- wine3[c("INDEX","P_TARGET")]
write.csv(scores, file = "U3_Scored2.csv", row.names = FALSE)
write.csv(as.data.frame(scores), file = "WINE_TEST.csv",
      sheetName = "Scored Data File", row.names = FALSE)
```

## Section 6: Scored Data File

| **Summary Statistics** *Predicted Number of Wins for Quality Control Purposes* | |
|---|---|
| MEAN | 3.27 |
| MEDIAN | 3.13 |
| MAX | 7.61 |
| MIN | 0.02 |

## Conclusion

In section 1, we conducted an initial exploratory data analysis using scatterplots, boxplots, summary statistics, etc. to help understand important characteristics and properties of the data that may be disguised by numerical summaries. The EDA revealed outliers and missing values for 8 variables: ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol, and Stars.

In section 2, we conducted data preparation/transformations of the data by fixing the missing values using predictive mean matching, conducting data transformations, creating flag variables, and handling outliers.

In section 3, we built seven models using the following procedures: Poisson Regression, Zero-Inflated Poisson Regression, Negative Binomial Regression, Zero-Inflated Negative Binomial Regression, and Multiple Linear Regression. This was conducted by manually selecting the variables or using variable selection techniques. We then ran model diagnostics and discussed the coefficients in the model to ensure that it makes intuitive wine sales sense.

In section 4, we selected ZIP_Model as our “best model” based on ‘fit’ (AIC, MSE) metrics and the fact that the model could handle count variables with excess number of zeros.

Lastly, in section 5, a Stand Alone scoring program was conducted that scored the new data and predicted the number of wine cases that will be sold based on certain properties/characteristics of the wine. The summary statistics showed the following: mean (3.27), median (3.13), max (7.61), and min (0.02). The data step also included all the variable transformations such as fixing missing values and the Zero-Inflated Poisson Regression formula.

## Full Code

```
#Part 0: Load & Prepare Data

library(readr)
library(dplyr)
library(zoo)
library(psych)
library(ROCR)
library(corrplot)
library(car)
library(InformationValue)
library(rJava)
library(pbkrtest)
library(car)
library(leaps)
library(MASS)
library(corrplot)
library(glm2)
library(aod)
library(mice)
library(Hmisc)
library(xlsxjars)
library(xlsx)
library(VIM)
library(pROC)
library(pscl) # For "counting" models (e.g., Poisson and Negative Binomial)
library(ggplot2) # For graphical tools
library(readr)
library(corrplot)


setwd("~/R/Wine")

wine=read.csv("Wine_Training.csv",header=T)

#Part 1: Data Exploration

#Data Quality Check
str(wine)
summary(wine)

library(Hmisc)
describe(wine)

#TARGET
par(mfrow=c(1,2))
```

```
hist(wine$TARGET, col = "#A71930", xlab = "TARGET ", main = "Histogram of Wine Sales Purchased")

boxplot(wine$TARGET, col = "#A71930", main = "Boxplot of Wine Sales Purchased ")

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

**#Chemistry**
**# FixedAcidity and VolatileAcidity**

```
par(mfrow=c(2,2))
hist(wine$FixedAcidity, col = "#A71930", xlab = "FixedAcidity", main = "Histogram of FixedAcidity")
hist(wine$VolatileAcidity, col = "#09ADAD", xlab = "VolatileAcidity", main = "Histogram of VolatileAcidity")
boxplot(wine$FixedAcidity, col = "#A71930", main = "Boxplot of FixedAcidity")
boxplot(wine$VolatileAcidity, col = "#09ADAD", main = "Boxplot of VolatileAcidity")
par(mfrow=c(1,1))
```

**# CitricAcid and ResidualSugar**

```
par(mfrow=c(2,2))
hist(wine$CitricAcid, col = "#A71930", xlab = "CitricAcid", main = "Histogram of CitricAcid")
hist(wine$ResidualSugar, col = "#DBCEAC", xlab = "ResidualSugar ", main = "Histogram of ResidualSugar")
boxplot(wine$CitricAcid, col = "#A71930", main = "Boxplot of CitricAcid")
boxplot(wine$ResidualSugar, col = "#DBCEAC", main = "Boxplot of ResidualSugar")
par(mfrow=c(1,1))
```

**#Chlorides and FreeSulfur Dioxide**

```
par(mfrow=c(2,2))
hist(wine$Chlorides, col = "#A71930", xlab = "Chlorides", main = "Histogram of Chlorides")
hist(wine$FreeSulfurDioxide, col = "#DBCEAC", xlab = "FreeSulfurDioxide ", main = "Histogram of
FreeSulfurDioxide")
boxplot(wine$Chlorides, col = "#A71930", main = "Boxplot of Chlorides")
boxplot(wine$FreeSulfurDioxide, col = "#DBCEAC", main = "Boxplot of FreeSulfurDioxide")
par(mfrow=c(1,1))
```

**#TotalSulfurDioxide and Density**

```
par(mfrow=c(2,2))
hist(wine$TotalSulfurDioxide, col = "#A71930", xlab = "TotalSulfurDioxide", main = "Histogram of
TotalSulfurDioxide")
hist(wine$Density, col = "#DBCEAC", xlab = "Density", main = "Histogram of Density")
boxplot(wine$TotalSulfurDioxide, col = "#A71930", main = "Boxplot of TotalSulfurDioxide")
boxplot(wine$Density, col = "#DBCEAC", main = "Boxplot of Density")
par(mfrow=c(1,1))
```

**#pH and Sulphates**

```
par(mfrow=c(2,2))
hist(wine$pH, col = "#A71930", xlab = "pH", main = "Histogram of pH")
hist(wine$Sulphates, col = "#09ADAD", xlab = "Sulphates", main = "Histograms of Sulphates")
boxplot(wine$pH, col = "#A71930", main = "Boxplot of pH")
boxplot(wine$Sulphates, col = "#09ADAD", main = "Boxplot of Sulphates")
```

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

#Alcohol and Acid Index
par(mfrow=c(2,2))
hist(wine$Alcohol, col = "#A71930", xlab = "Alcohol", main = "Histogram of Alcohol")
hist(wine$AcidIndex, col = "#DBCEAC", xlab = "AcidIndex", main = "Histograms of AcidIndex")
boxplot(wine$Alcohol, col = "#A71930", main = "Boxplot of Alcohol")
boxplot(wine$AcidIndex, col = "#DBCEAC", main = "Boxplot of AcidIndex")
par(mfrow=c(1,1))

#Label Appeal and STARS
par(mfrow=c(2,2))
hist(wine$LabelAppeal, col = "#A71930", xlab = "LabelAppeal", main = "Histogram of LabelAppeal ")
hist(wine$STARS, col = "#09ADAD", xlab = "STARS", main = "Histogram of STARS")
boxplot(wine$LabelAppeal, col = "#A71930", main = "Boxplot of LabelAppeal")
boxplot(wine$STARS, col = "#09ADAD", main = "Boxplot of STARS")
par(mfrow=c(1,1))

# Scatterplot Matrix

panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
  if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
  text(0.5, 0.5, txt, cex = cex.cor * r)
}

pairs(~ wine$TARGET + wine$FixedAcidity+ wine$VolatileAcidity+ wine$CitricAcid+ wine$ResidualSugar+
wine$Chlorides+ wine$FreeSulfurDioxide+ wine$TotalSulfurDioxide, lower.panel = panel.smooth)
par(mfrow=c(1,1))

pairs(~ wine$TARGET + wine$Density+ wine$pH+ wine$Sulphates+ wine$Alcohol + wine$LabelAppeal+
wine$AcidIndex + wine$STARS, lower.panel = panel.smooth)
par(mfrow=c(1,1))

#Correlation Matrix
subdatnum <- subset(wine, select=c(
"FixedAcidity",
"VolatileAcidity",
"CitricAcid",
"ResidualSugar",
"Chlorides",
"FreeSulfurDioxide",
"TotalSulfurDioxide",
```

```
"Density",
"pH",
"Sulphates",
"Alcohol",
"LabelAppeal",
"AcidIndex",
"STARS",
"TARGET"))

require(corrplot)
mcor <- cor(subdatnum)
corrplot(mcor, method="number", shade.col=NA, tl.col="black",tl.cex=0.8)

#Part 2: Data Preparation
library(mice)

#Check for missing values
sapply(wine, function(x) sum(is.na(x)))

#Check missing data percentage
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(wine,2,pMiss)

#Create Flag Variables (Missing Data)

wine$NoResidualSugar <- 0
wine$NoResidualSugar [is.na(wine$ResidualSugar)] <- 1

wine$NoChlorides  <- 0
wine$NoChlorides [is.na(wine$Chlorides)] <- 1

wine$NoFreeSulfurDioxide <- 0
wine$NoFreeSulfurDioxide[is.na(wine$FreeSulfurDioxide)] <- 1

wine$NoTotalSulfurDioxide <- 0
wine$NoTotalSulfurDioxide[is.na(wine$TotalSulfurDioxide)] <- 1

wine$NopH <- 0
wine$NopH[is.na(wine$pH)] <- 1

wine$NoSulphates <- 0
wine$NoSulphates [is.na(wine$Sulphates)] <- 1

wine$NoResidualSugar <- 0
wine$NoResidualSugar [is.na(wine$ResidualSugar)] <- 1

wine$NoAlcohol <- 0
wine$NoAlcohol [is.na(wine$Alcohol)] <- 1
```

```
wine$NoSTARS<- 0
wine$NoSTARS [is.na(wine$STARS)] <- 1

str(wine)

#Run imputation
tempData <- mice(wine,m=5,maxit=50,meth='pmm',seed=500)
summary(tempData)

#Check N/A values have been removed
wine2 <- complete(tempData,1)
apply(wine2,2,pMiss)
summary(wine2)

densityplot(tempData)


#Straighten Relationships – Create transformed variables that we can look at later

wine2$logFixedAcidity <- log(wine2$FixedAcidity)
wine2$logVolatileAcidity <- log(wine2$VolatileAcidity)
wine2$logCitricAcid <- log(wine2$CitricAcid)
wine2$logResidualSugar <- log(wine2$ResidualSugar)
wine2$logChlorides <- log(wine2$Chlorides)
wine2$logFreeSulfurDioxide <- log(wine2$FreeSulfurDioxide)
wine2$logTotalSulfurDioxide <- log(wine2$TotalSulfurDioxide)
wine2$logDensity <- log(wine2$Density)
wine2$logpH <- log(wine2$pH)
wine2$logSulphates <- log(wine2$Sulphates)
wine2$logAlcohol <- log(wine2$Alcohol)
wine2$logLabelAppeal <- log(wine2$LabelAppeal)
wine2$logAcidIndex <- log(wine2$AcidIndex)
wine2$logSTARS <- log(wine2$STARS)

#Create SQRT Transformations of Some of the Variables
wine2$SQRT_STARS <- sqrt(wine2$STARS)
wine2$SQRT_AcidIndex <- sqrt(wine2$AcidIndex)


#Trim Data

wine2$FixedAcidity [(wine2$FixedAcidity >= 20)] = 20
wine2$FixedAcidity [(wine2$FixedAcidity <= -5)] = -5

wine2$VolatileAcidity [(wine2$VolatileAcidity >= 2)] = 2
wine2$VolatileAcidity [(wine2$VolatileAcidity  <= -1.5)] = -1.5

wine2$CitricAcid [(wine2$CitricAcid >= 2)] = 2
```

```
wine2$CitricAcid [(wine2$CitricAcid <= -1.5)] = -1.5

wine2$ResidualSugar [(wine2$ResidualSugar >= 65)] = 65
wine2$ResidualSugar [(wine2$ResidualSugar <= -65)] = -65

wine2$Chlorides [(wine2$Chlorides >= 0.7)] = 0.7
wine2$Chlorides [(wine2$Chlorides <= -0.6)] = -0.6

wine2$FreeSulfurDioxide [(wine2$FreeSulfurDioxide >= 350)] = 350
wine2$FreeSulfurDioxide [(wine2$FreeSulfurDioxide <= -275)] = -275

wine2$TotalSulfurDioxide [(wine2$TotalSulfurDioxide >= 725)] = 725
wine2$TotalSulfurDioxide [(wine2$TotalSulfurDioxide <= -400)] = -400

wine2$Density [(wine2$Density >= 1.06)] = 1.06
wine2$Density [(wine2$Density <= 0.93)] = 0.93

wine2$pH [(wine2$pH >= 4.75)] = 4.75
wine2$pH [(wine2$pH <= 1.5)] = 1.5

wine2$Sulphates [(wine2$Sulphates >= 2.5)] = 2.5
wine2$Sulphates [(wine2$Sulphates <= -1.5)] = -1.5

wine2$Alcohol [(wine2$Alcohol >= 20)] = 20
wine2$Alcohol [(wine2$Alcohol <= 3)] = 3

wine2$AcidIndex [(wine2$AcidIndex >= 11)] = 11
wine2$AcidIndex [(wine2$AcidIndex <= 5)] = 5

summary(wine2)

#Correlation Matrix
subdatnum2 <- subset(wine2, select=c(
"FixedAcidity",
"VolatileAcidity",
"CitricAcid",
"ResidualSugar",
"Chlorides",
"FreeSulfurDioxide",
"TotalSulfurDioxide",
"Density",
"pH",
"Sulphates",
"Alcohol",
"LabelAppeal",
"AcidIndex",
"STARS",
"NoResidualSugar",
```

```
"NoChlorides",
"NoFreeSulfurDioxide",
"NoTotalSulfurDioxide",
"NopH",
"NoSulphates",
"NoSTARS",
"NoAlcohol",
"TARGET"))

require(corrplot)
mcor <- cor(subdatnum2)
corrplot(mcor, method="number", shade.col=NA, tl.col="black",tl.cex=0.8)
par(mfrow=c(1,1))
```

**#Part 3: Model Creation**

```
MLRResult1<- lm(formula = TARGET ~ STARS + NoSTARS + LabelAppeal + AcidIndex + VolatileAcidity, data =
wine2)

anova(MLRResult1)
summary(MLRResult1)
par(mfrow=c(2,2))  # visualize four graphs at once
plot(MLRResult1)
vif(MLRResult1)
coefficients(MLRResult1)
```

**# Stepwise Approach**

```
stepwisemodel <- lm(formula = TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar + Chlorides
+ FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex +
STARS + NoResidualSugar + NoChlorides + NoFreeSulfurDioxide + NoTotalSulfurDioxide + NopH + NoSulphates
+ NoAlcohol + NoSTARS, data = wine2)

stepwise <- stepAIC(stepwisemodel, direction = "both")

summary(stepwise)

anova(stepwise)
summary(stepwise)
par(mfrow=c(2,2))  # visualize four graphs at once
plot(stepwise)
vif(stepwise)
coefficients(stepwise)
```

**#Model 3 (reduced Stepwise)**

```
Model3 <- lm(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide + TotalSulfurDioxide + Chlorides +
Alcohol + LabelAppeal + logAcidIndex + logSTARS + NoSTARS, data = wine2)

anova(Model3)
summary(Model3)
par(mfrow=c(2,2))  # visualize four graphs at once
plot(Model3)
vif(Model3)
coefficients(Model3)
```

**#Poisson Model**

```
poisson_model <- glm(TARGET ~ VolatileAcidity + FreeSulfurDioxide + TotalSulfurDioxide + Chlorides +
LabelAppeal + logAcidIndex + logSTARS + NoSTARS, family="poisson"(link="log"), data=wine2)

anova(poisson_model, test="Chisq")
summary(poisson_model)
coef(poisson_model)

wine2$poisson_yhat <- predict(poisson_model, newdata = wine2, type = "response")

with(poisson_model, cbind(res.deviance = deviance, df = df.residual,
  p = pchisq(deviance, df.residual, lower.tail=FALSE)))

library(AER)
deviance(poisson_model)/poisson_model$df.residual
dispersiontest(poisson_model)

#what type of dispersion does sample have?
mean(wine2$TARGET)
var(wine2$TARGET)

library(car)
influencePlot(poisson_model)
res <- residuals(poisson_model, type="deviance")
plot(log(predict(poisson_model)), res)
abline(h=0, lty=2)
qqnorm(res)
qqline(res)
```

**#Negative Binomial Distribution**

```
NBR_Model<-glm.nb(TARGET ~ VolatileAcidity + LabelAppeal + SQRT_AcidIndex + logSTARS + NoSTARS,
data=wine2)

summary(NBR_Model)
```

```
wine2$NBRphat <- predict(NBR_Model, newdata = wine2, type = "response")

odTest(NBR_Model)
```

**#ZERO INFLATED POISSON (ZIP)**

```
ZIP_Model<-zeroinfl(TARGET ~ VolatileAcidity + FreeSulfurDioxide + TotalSulfurDioxide + Chlorides + Alcohol + LabelAppeal + logAcidIndex + logSTARS + NoSTARS, data=wine2)

summary(ZIP_Model)

vuong(poisson_model, ZIP_Model)

wine2$ZIPphat <- predict(ZIP_Model, newdata = wine2, type = "response")
```

**#ZERO INFLATED NEGATIVE BINOMIAL REGRESSION (ZINB)**

```
ZINB_Model<-zeroinfl(TARGET ~ Alcohol + LabelAppeal + logAcidIndex + logSTARS + NoSTARS, data=wine2, dist = "negbin", EM=TRUE)

summary(ZINB_Model)

vuong(NBR_Model, ZINB_Model)

wine2$ZINBphat <- predict(ZINB_Model, newdata = wine2, type = "response")

#what type of dispersion does sample have?
mean(wine2$TARGET)
var(wine2$TARGET)
```

**#Part 4: Performance**

```
#Function for Mean Square Error Calculation
mse <- function(sm)
  mean(sm$residuals^2)

AIC(MLRResult1)
AIC(stepwisemodel)
AIC(Model3)
AIC(poisson_model)
AIC(NBR_Model)
AIC (ZIP_Model)
AIC(ZINB_Model)


BIC(MLRResult1)
BIC(stepwisemodel)
BIC(Model3)
```

```
BIC(poisson_model)
BIC(NBR_Model)
BIC(ZIP_Model)
BIC(ZINB_Model)


mse(MLRResult1)
mse(stepwisemodel)
mse(Model3)
mse(poisson_model)
mse(NBR_Model)
mse(ZIP_Model)
mse(ZINB_Model)

#####
#Designated proper working environment on my computer. You will want to make sure it is in proper place for your computer.
#####
```

**#Part 5: Test Data**

```
wine_test=read.csv("wine_test.csv",header=T)
```

**#Part 2: Data Preparation**

```
library(mice)
```

**#Check for missing values**

```
sapply(wine_test, function(x) sum(is.na(x)))
```

**#Check missing data percentage**

```
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(wine_test,2,pMiss)
```

**#Create Flag Variables (Missing Data)**

```
wine_test$NoResidualSugar <- 0
wine_test$NoResidualSugar [is.na(wine_test$ResidualSugar)] <- 1

wine_test$NoChlorides  <- 0
wine_test$NoChlorides [is.na(wine_test$Chlorides)] <- 1

wine_test$NoFreeSulfurDioxide <- 0
wine_test$NoFreeSulfurDioxide[is.na(wine_test$FreeSulfurDioxide)] <- 1

wine_test$NoTotalSulfurDioxide <- 0
wine_test$NoTotalSulfurDioxide[is.na(wine_test$TotalSulfurDioxide)] <- 1

wine_test$NopH <- 0
```

```
wine_test$NopH[is.na(wine_test$pH)] <- 1

wine_test$NoSulphates <- 0
wine_test$NoSulphates [is.na(wine_test$Sulphates)] <- 1

wine_test$NoResidualSugar <- 0
wine_test$NoResidualSugar [is.na(wine_test$ResidualSugar)] <- 1

wine_test$NoAlcohol <- 0
wine_test$NoAlcohol [is.na(wine_test$Alcohol)] <- 1

wine_test$NoSTARS<- 0
wine_test$NoSTARS [is.na(wine_test$STARS)] <- 1

str(wine_test)
```

**#Run imputation**

```
tempData <- mice(wine_test,m=5,maxit=50,meth='pmm',seed=500)
summary(tempData)
```

**#Check N/A values have been removed**

```
wine3 <- complete(tempData,1)
apply(wine3,2,pMiss)
summary(wine3)

densityplot(tempData)
```

**#Straighten Relationships – Create transformed variables that we can look at later**

```
wine3$logFixedAcidity <- log(wine3$FixedAcidity)
wine3$logVolatileAcidity <- log(wine3$VolatileAcidity)
wine3$logCitricAcid <- log(wine3$CitricAcid)
wine3$logResidualSugar <- log(wine3$ResidualSugar)
wine3$logChlorides <- log(wine3$Chlorides)
wine3$logFreeSulfurDioxide <- log(wine3$FreeSulfurDioxide)
wine3$logTotalSulfurDioxide <- log(wine3$TotalSulfurDioxide)
wine3$logDensity <- log(wine3$Density)
wine3$logpH <- log(wine3$pH)
wine3$logSulphates <- log(wine3$Sulphates)
wine3$logAlcohol <- log(wine3$Alcohol)
wine3$logLabelAppeal <- log(wine3$LabelAppeal)
wine3$logAcidIndex <- log(wine3$AcidIndex)
wine3$logSTARS <- log(wine3$STARS)
```

**#Create SQRT Transformations of Some of the Variables**

```
wine3$SQRT_STARS <- sqrt(wine3$ STARS)
wine3$SQRT_AcidIndex <- sqrt(wine3$AcidIndex)
```

**#Trim Data**

```
wine3$FixedAcidity [(wine3$FixedAcidity >= 20)] = 20
wine3$FixedAcidity [(wine3$FixedAcidity <= -5)] = -5

wine3$VolatileAcidity [(wine3$VolatileAcidity >= 2)] = 2
wine3$VolatileAcidity [(wine3$VolatileAcidity  <= -1.5)] = -1.5

wine3$CitricAcid [(wine3$CitricAcid >= 2)] = 2
wine3$CitricAcid [(wine3$CitricAcid <= -1.5)] = -1.5

wine3$ResidualSugar [(wine3$ResidualSugar >= 65)] = 65
wine3$ResidualSugar [(wine3$ResidualSugar <= -65)] = -65

wine3$Chlorides [(wine3$Chlorides >= 0.7)] = 0.7
wine3$Chlorides [(wine3$Chlorides <= -0.6)] = -0.6

wine3$FreeSulfurDioxide [(wine3$FreeSulfurDioxide >= 350)] = 350
wine3$FreeSulfurDioxide [(wine3$FreeSulfurDioxide <= -275)] = -275

wine3$TotalSulfurDioxide [(wine3$TotalSulfurDioxide >= 725)] = 725
wine3$TotalSulfurDioxide [(wine3$TotalSulfurDioxide <= -400)] = -400

wine3$Density [(wine3$Density >= 1.06)] = 1.06
wine3$Density [(wine3$Density <= 0.93)] = 0.93

wine3$pH [(wine3$pH >= 4.75)] = 4.75
wine3$pH [(wine3$pH <= 1.5)] = 1.5

wine3$Sulphates [(wine3$Sulphates >= 2.5)] = 2.5
wine3$Sulphates [(wine3$Sulphates <= -1.5)] = -1.5

wine3$Alcohol [(wine3$Alcohol >= 20)] = 20
wine3$Alcohol [(wine3$Alcohol <= 3)] = 3

wine3$AcidIndex [(wine3$AcidIndex >= 11)] = 11
wine3$AcidIndex [(wine3$AcidIndex <= 5)] = 5

summary(wine3)
```

**# Stand Alone Scoring**

```
ZIP_Model<-zeroinfl(TARGET ~ VolatileAcidity + FreeSulfurDioxide + TotalSulfurDioxide + Chlorides + Alcohol +
LabelAppeal + logAcidIndex + logSTARS + NoSTARS, data=wine2)

wine3$P_TARGET <- predict(ZIP_Model, newdata = wine3, type = "response")
```

```
summary(wine3)

select <- dplyr::select


# Scored Data File
scores <- wine3[c("INDEX","P_TARGET")]
write.csv(scores, file = "U3_Scored2.csv", row.names = FALSE)
write.csv(as.data.frame(scores), file = "WINE_TEST.csv",
        sheetName = "Scored Data File", row.names = FALSE)
```

# Appendix

## Data Quality Check (Figure 3)

```
> describe(wine)
wine

 16  Variables      12795  Observations
--------------------------------------------------------------------------------------------------------------------------
-----
INDEX
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   12795        0    12795        1     8070     5378    804.7   1610.4   4037.5   8110.0  12106.5  14515.6  15309.3

lowest :     1     2     4     5     6, highest: 16120 16123 16127 16128 16129
--------------------------------------------------------------------------------------------------------------------------
-----
TARGET
       n  missing distinct     Info     Mean      Gmd
   12795        0        9    0.962    3.029    2.141

Value          0     1     2     3     4     5     6     7     8
Frequency   2734   244  1091  2611  3177  2014   765   142    17
Proportion 0.214 0.019 0.085 0.204 0.248 0.157 0.060 0.011 0.001
--------------------------------------------------------------------------------------------------------------------------
-----
FixedAcidity
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   12795        0      470        1    7.076    6.688     -3.6     -1.2      5.2      6.9      9.5     15.6     17.8

lowest : -18.1 -18.0 -17.7 -17.5 -17.4, highest:  32.4  32.5  32.6  34.1  34.4
--------------------------------------------------------------------------------------------------------------------------
-----
VolatileAcidity
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   12795        0      815        1   0.3241   0.8262   -1.023   -0.720    0.130    0.280    0.640    1.350    1.640

lowest : -2.790 -2.750 -2.745 -2.730 -2.720, highest:  3.500  3.550  3.565  3.590  3.680
--------------------------------------------------------------------------------------------------------------------------
-----
CitricAcid
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   12795        0      602        1   0.3084   0.9057    -1.16    -0.84     0.03     0.31     0.58     1.43     1.79

lowest : -3.24 -3.16 -3.10 -3.08 -3.06, highest:  3.63  3.68  3.70  3.77  3.86
--------------------------------------------------------------------------------------------------------------------------
-----
ResidualSugar
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   12179      616     2077        1    5.419    35.31   -52.70   -39.66    -2.00     3.90    15.90    49.72    62.70

lowest : -127.80 -127.10 -126.20 -126.10 -125.70, highest:  136.50  137.60  138.00  140.65  141.15
--------------------------------------------------------------------------------------------------------------------------
-----
Chlorides
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   12157      638     1663        1  0.05482   0.3311   -0.489   -0.372   -0.031    0.046    0.153    0.481    0.598
```

```
lowest : -1.171 -1.170 -1.158 -1.156 -1.155, highest:  1.260  1.261  1.270  1.275  1.351
-------------------------------------------------------------------------------------------------------------------
-----
FreeSulfurDioxide
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   12148      647      999        1    30.85    155.2     -224     -171        0       30       70      230      284

lowest : -555 -546 -536 -535 -532, highest:  613  617  618  622  623
-------------------------------------------------------------------------------------------------------------------
-----
TotalSulfurDioxide
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   12113      682     1370        1    120.7    246.9   -273.0   -185.0     27.0    123.0    208.0    421.8    513.4

lowest : -823 -816 -793 -781 -779, highest: 1032 1041 1048 1054 1057
-------------------------------------------------------------------------------------------------------------------
-----
Density
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   12795        0     5933        1   0.9942  0.02769   0.9488   0.9587   0.9877   0.9945   1.0005   1.0295   1.0398

lowest : 0.88809 0.88949 0.88978 0.88983 0.89167, highest: 1.09658 1.09679 1.09695 1.09791 1.09924
-------------------------------------------------------------------------------------------------------------------
-----
pH
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   12400      395      497        1    3.208   0.7242     2.06     2.31     2.96     3.20     3.47     4.10     4.37

lowest : 0.48 0.53 0.54 0.58 0.59, highest: 5.91 5.94 6.02 6.05 6.13
-------------------------------------------------------------------------------------------------------------------
-----
Sulphates
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   11585     1210      630        1   0.5271   0.9827    -1.05    -0.70     0.28     0.50     0.86     1.77     2.09

lowest : -3.13 -3.12 -3.10 -3.07 -3.03, highest:  4.11  4.16  4.19  4.21  4.24
-------------------------------------------------------------------------------------------------------------------
-----
Alcohol
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
   12142      653      401        1    10.49    4.015      4.1      5.7      9.0     10.4     12.4     15.2     16.7

lowest : -4.7 -4.5 -4.4 -4.3 -4.1, highest: 25.4 25.6 26.0 26.1 26.5
-------------------------------------------------------------------------------------------------------------------
-----
LabelAppeal
        n   missing  distinct      Info      Mean       Gmd
    12795         0         5     0.887 -0.009066    0.9566

Value         -2    -1     0     1     2
Frequency    504  3136  5617  3048   490
Proportion 0.039 0.245 0.439 0.238 0.038
-------------------------------------------------------------------------------------------------------------------
-----
AcidIndex
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
```

```
   12795        0       14    0.908    7.773    1.316        6        7        7        8        8        9       10

Value          4     5     6     7     8     9    10    11    12    13    14    15    16    17
Frequency      3    75  1197  4878  4142  1427   551   258   128    69    47     8     5     7
Proportion 0.000 0.006 0.094 0.381 0.324 0.112 0.043 0.020 0.010 0.005 0.004 0.001 0.000 0.001
-------------------------------------------------------------------------------------------------------------------------
-----
STARS
       n  missing distinct     Info     Mean      Gmd
    9436     3359        4    0.899    2.042   0.9777

Value          1     2     3     4
Frequency   3042  3570  2212   612
Proportion 0.322 0.378 0.234 0.065
-------------------------------------------------------------------------------------------------------------------------
-----
```