Wine Sales Project Report

By Mimi Trinh

INTRODUCTION

A large wine manufacturer is conducting a study to predict the number of wine cases ordered based upon the wine characteristics. Specifically, if they can predict the number of cases ordered, they can adjust the wine offering to maximize sales. The more sample cases purchased by wine distribution companies, the more likely is a wine to be sold at a high-end restaurant since these cases would be used to provide tasting samples to restaurants and wine stores around the country.

A dataset of approximately 12,000 commercially available wines is used to analyze the problem. The variables are mostly related to the chemical properties of the wine being sold. Some of the variables have negative values when they technically shouldn't because the original data was modified for proprietary reasons, so this issue will be ignored in the study. The target variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine.

The wine sales project will include four main sections 1) data exploration 2) data preparation 3) model development 4) model evaluation. Several models are developed, and based on multiple metrics, a champion model is chosen to predict the number of cases of wine that will be sold given certain properties of the wine.

RESULTS

Section 1: Data Exploration

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

The wine dataset used in this study has 12,795 observations with 16 variables.

1. INDEX: identification variable
2. TARGET: number of cases purchased
3. Fixed Acidity: fixed acidity of wine
4. Volatile Acidity: volatile acid content of wine
5. Citric Acid: citric acid content
6. Residual Sugar: residual sugar of wine
7. Chlorides: chloride content of wine
8. Free Sulfur Dioxide: sulfur dioxide content of wine
9. Total Sulfur Dioxide: total sulfur dioxide content of wine
10. Density: density of wine
11. pH: pH of wine
12. Sulphates: sulfate content of wine
13. Alcohol: alcohol content
14. Label Appeal: marketing score indicating the appeal of label design for customers. High numbers suggest customers like the label design. Negative numbers suggest customers don't like the design
15. Acid Index: proprietary method of testing total acidity of wine by using a weighted average
16. STARS: wine rating by a team of experts with 4 stars = excellent and 1 star = poor

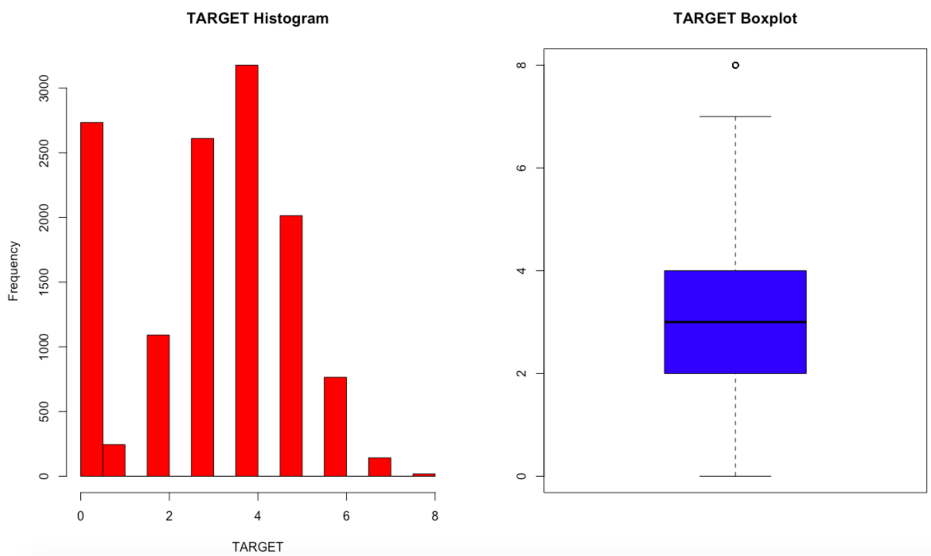
```
> summary(wine)
     INDEX           TARGET       FixedAcidity     VolatileAcidity     CitricAcid
 Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400
 1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300
 Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100
 Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084
 3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800
 Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600

 ResidualSugar        Chlorides       FreeSulfurDioxide TotalSulfurDioxide    Density
 Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00   Min.   :-823.0     Min.   :0.8881
 1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00   1st Qu.:  27.0     1st Qu.:0.9877
 Median :   3.900   Median : 0.0460   Median :  30.00   Median : 123.0     Median :0.9945
 Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85   Mean   : 120.7     Mean   :0.9942
 3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00   3rd Qu.: 208.0     3rd Qu.:1.0005
 Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00   Max.   :1057.0     Max.   :1.0992
 NA's   :616        NA's   :638       NA's   :647       NA's   :682
       pH          Sulphates          Alcohol       LabelAppeal          AcidIndex
 Min.   :0.480   Min.   :-3.1300   Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000
 1st Qu.:2.960   1st Qu.: 0.2800   1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000
 Median :3.200   Median : 0.5000   Median :10.40   Median : 0.000000   Median : 8.000
 Mean   :3.208   Mean   : 0.5271   Mean   :10.49   Mean   :-0.009066   Mean   : 7.773
 3rd Qu.:3.470   3rd Qu.: 0.8600   3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000
 Max.   :6.130   Max.   : 4.2400   Max.   :26.50   Max.   : 2.000000   Max.   :17.000
 NA's   :395     NA's   :1210      NA's   :653
     STARS
 Min.   :1.000
 1st Qu.:1.000
 Median :2.000
 Mean   :2.042
 3rd Qu.:3.000
 Max.   :4.000
 NA's   :3359
```

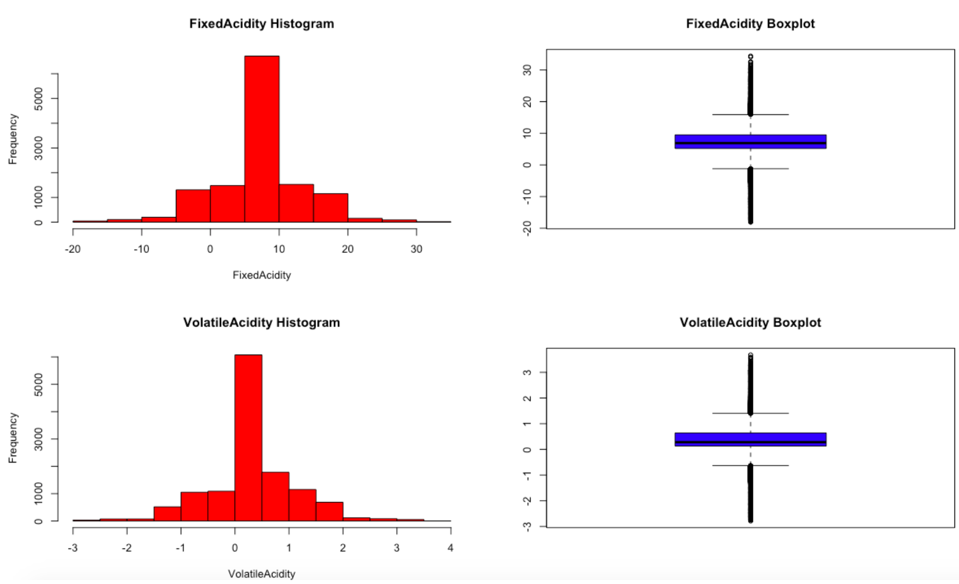
The summary above shows the mean, median, and five-number summary of each variable. It also shows that there are missing value issues with the following variables that need to be imputed in section 2 of the report: ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol, STARS.

```
> skewness(wine$TARGET)
[1] -0.3263393
> skewness(wine$FixedAcidity)
[1] -0.02258861
> skewness(wine$VolatileAcidity)
[1] 0.02038235
> skewness(wine$CitricAcid)
[1] -0.05031294
> skewness(wine$ResidualSugar,na.exclude(wine$ResidualSugar))
[1] -0.05312945
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
> skewness(wine$Chlorides,na.exclude(wine$Chlorides))
[1] 0.03043093
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
> skewness(wine$FreeSulfurDioxide,na.exclude(wine$FreeSulfurDioxide))
[1] 0.0063938
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
> skewness(wine$TotalSulfurDioxide,na.exclude(wine$TotalSulfurDioxide))
[1] -0.00718024
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
> skewness(wine$Density)
[1] -0.01869596
> skewness(wine$pH,na.exclude(wine$pH))
[1] 0.04429337
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
> skewness(wine$Sulphates,na.exclude(wine$Sulphates))
[1] 0.005912661
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
> skewness(wine$Alcohol,na.exclude(wine$Alcohol))
[1] -0.03071963
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
> skewness(wine$LabelAppeal)
[1] 0.008430445
> skewness(wine$AcidIndex)
[1] 1.648689
> skewness(wine$STARS,na.exclude(wine$STARS))
[1] 0.4473064
Warning message:
In if (na.rm) x <- x[!is.na(x)] :
  the condition has length > 1 and only the first element will be used
```

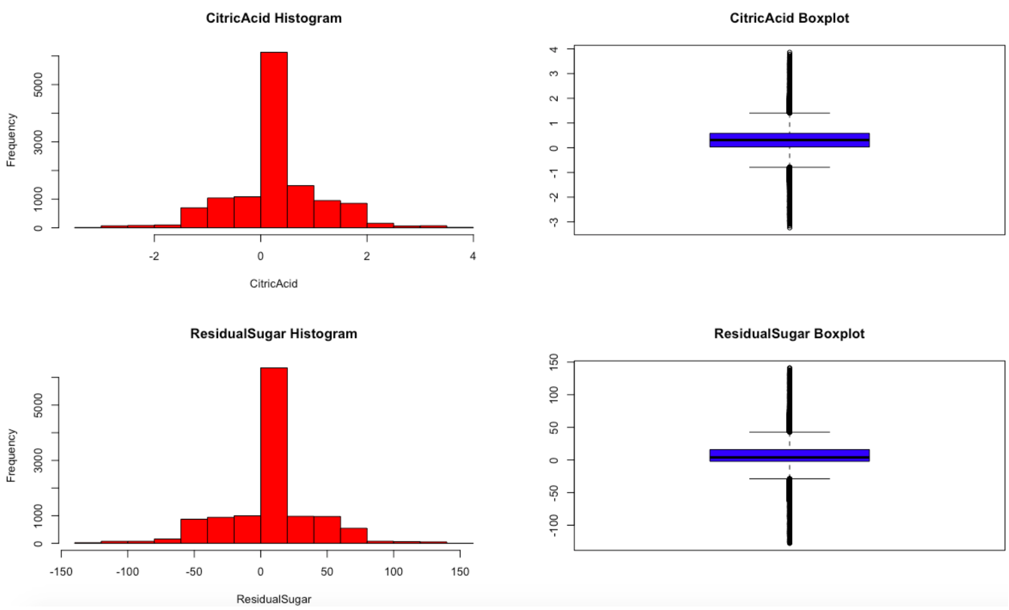
In order to identify variables with outlier issues, we use skewness number. If a variable has the skewness number greater than 1 or less than -1, that variable has outlier issue. The output above shows that only AcidIndex has outlier issue. Variables with missing values will have a warning message output, as shown above.



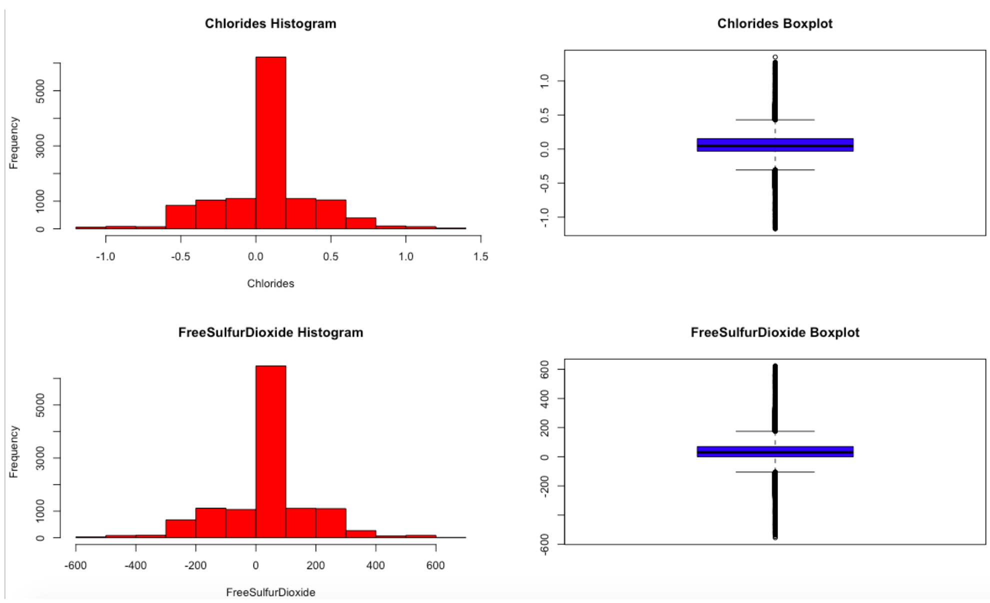
The histogram and boxplot above show that the target variable is a discrete count variable with many 0's indicating that the particular type of wine is not sold. Also, there's no outlier issue in this variable. The fact that the variable is a discrete count variable with many 0's and the histogram above indicate that Poisson regression is an appropriate analysis for this project. We will explore this topic in section 3 and 4 of the report.



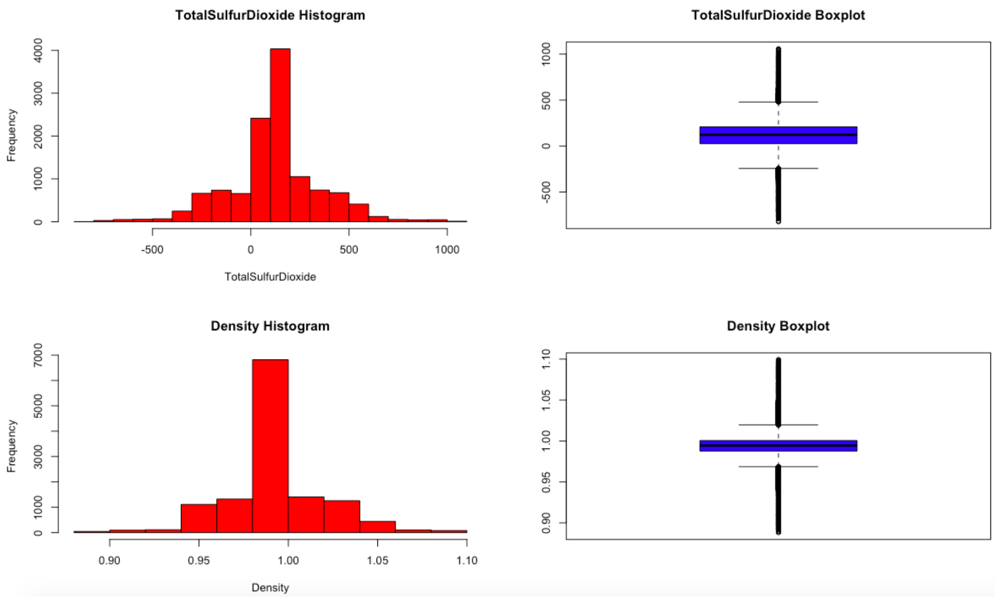
The histograms and boxplots above show that Fixed Acidity and Volatile Acidity may not have normality issue since they have outliers on both tails, but they have a lot of outlier issues, which will be addressed in section 2 of the report.



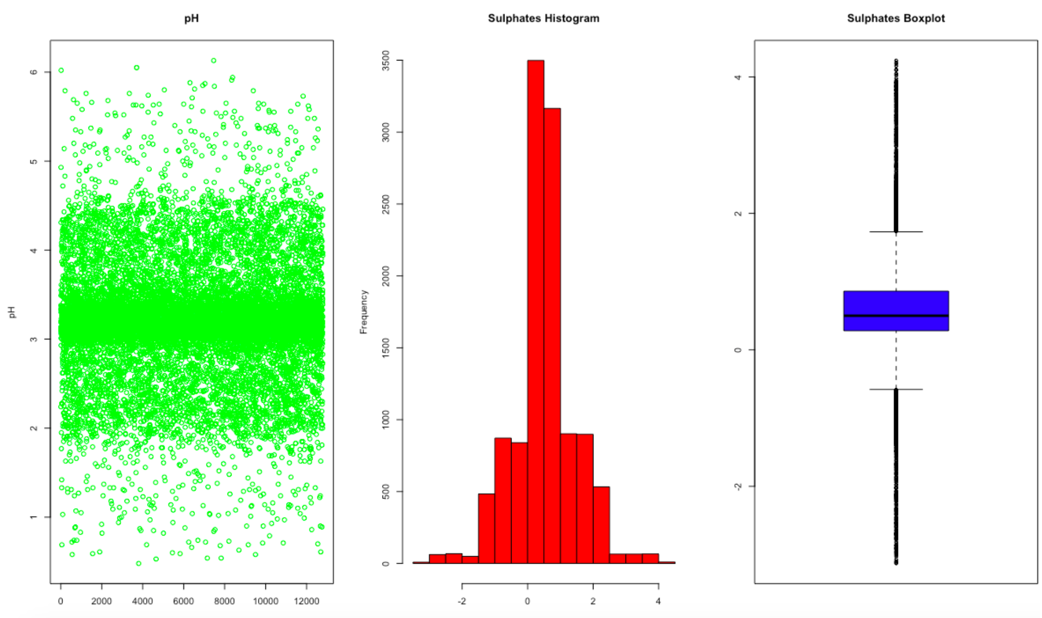
The histograms and boxplots above show that Citric Acid and Residual Sugar may not have normality issue since they have outliers on both tails, but they have a lot of outlier issues, which will be addressed in section 2 of the report.



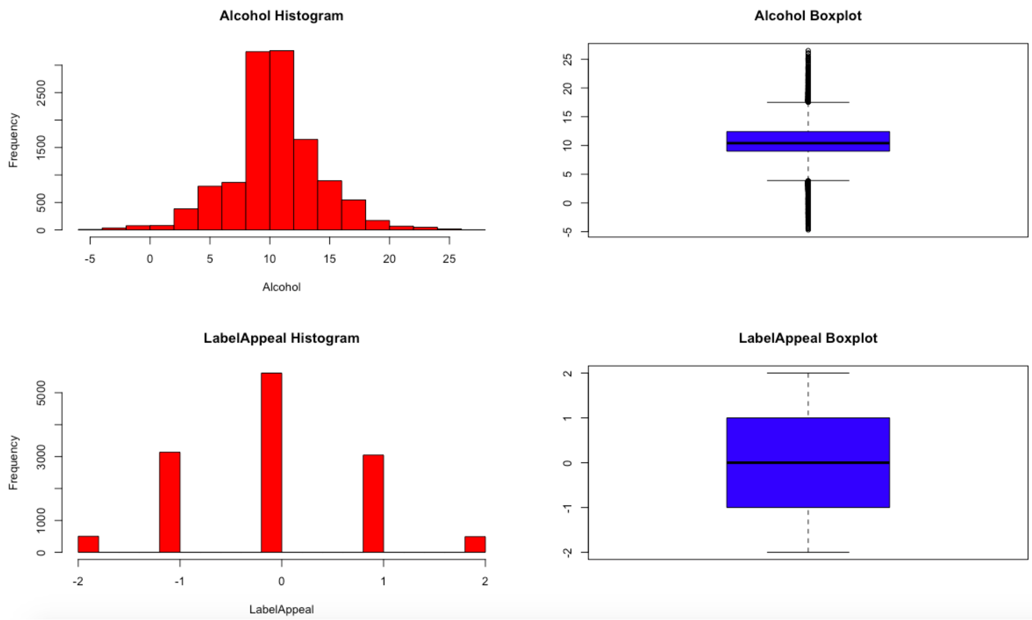
The histograms and boxplots above show that Chlorides and Free Sulfur Dioxide may not have normality issue since they have outliers on both tails, but they have a lot of outlier issues, which will be addressed in section 2 of the report.



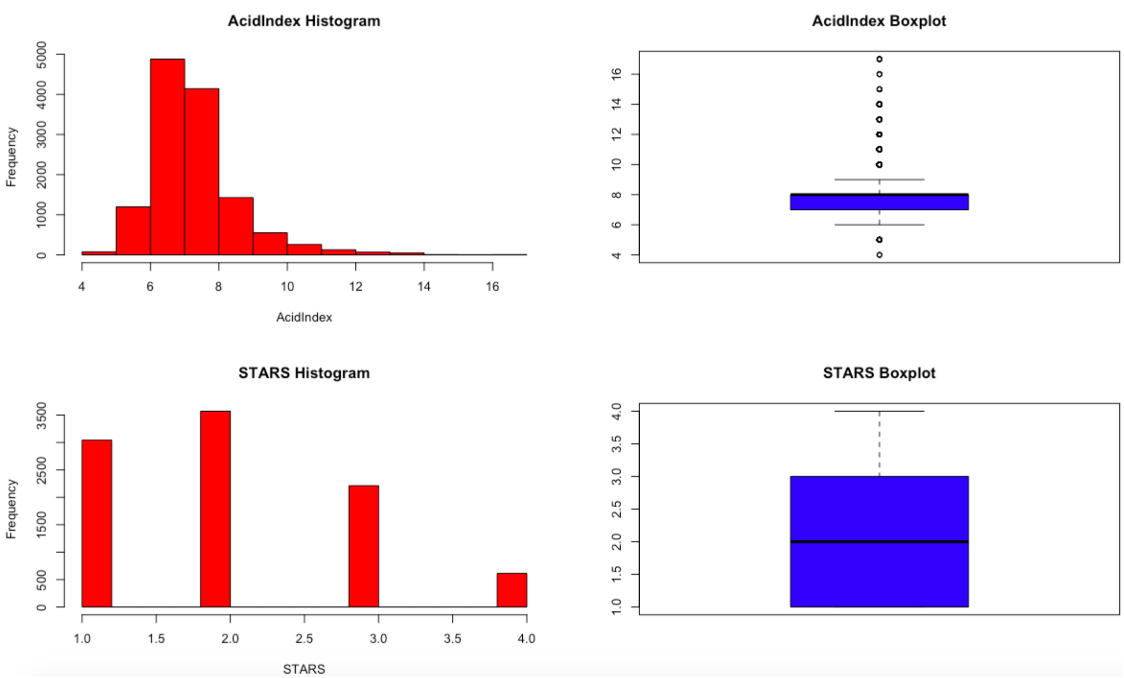
The histograms and boxplots above show that Total Sulfur Dioxide and Density may not have normality issue since they have outliers on both tails, but they have a lot of outlier issues, which will be addressed in section 2 of the report.



The histograms and boxplots above show that pH and Sulphates may not have normality issue since they have outliers on both tails, but they have a lot of outlier issues, which will be addressed in section 2 of the report.

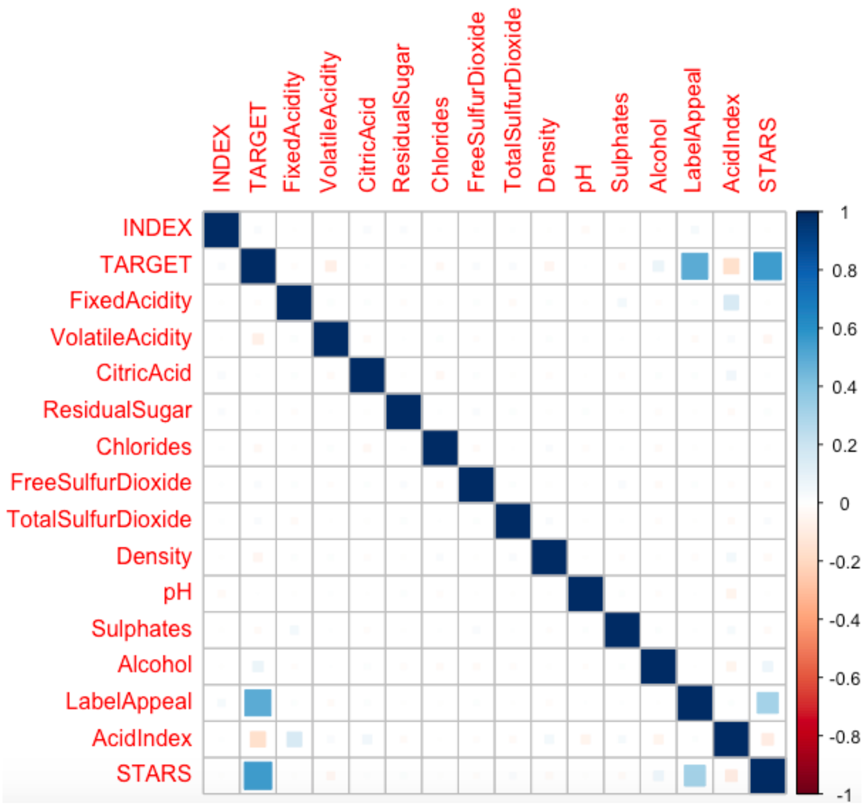


The histograms and boxplots above show that Alcohol may not have normality issue since it has outliers on both tails, but it has a lot of outlier issues, which will be addressed in section 2 of the report. Label Appeal doesn't have normality or outlier issue.



The histogram and boxplots above show that Acid Index may have a light outlier issue skewing toward the right whereas STARS has no normality or outlier issue.

When a variable has outliers on both tails, it still maintains normality since the skewness number is typically close to 0. However, the fact that there are many outliers on both ends may affect the coefficients in the predictive models. The outliers on both tails are not deal-breaker with the regression assumptions, but they can cause problems with the regression formulas. Thus, the outlier issues will be addressed in section 2 of the report prior to the model development stage.



The correlation plot above shows that Label Appeal and STARS have a strong positive relationship with the target variable whereas the remaining predictors don't have direct relationship with the target variable. Also, there's no correlation among the predictors, which reduce the risk of multicollinearity issue in the models.

Section 2: Data Preparation

The first part of this section addresses the missing value issues whereas the second part addresses the outlier issues. For every predictor with missing values, two new variables are created: flag variable beginning with an "M" with binary values (1=missing value and 0=known value) and imputed variable beginning with an "IMP" to replace the missing values with the mean. Typically, we replace missing values with the mean if there's no outlier issue and median if there's outlier issue. Although there are outliers, these outliers take place on both tails, so it's appropriate in this case to replace missing values with the mean. The following variables have missing values and thus have two additional variables in the dataset: flag and imputed variables.

- Residual Sugar
- Chlorides
- Free Sulfur Dioxide

- Total Sulfur Dioxide
- pH
- Alcohol
- Sulphates
- STARS

```
> skewness(wine$AcidIndex)
[1] 1.648689
> summary(wine$AcidIndex)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.000   7.000   8.000   7.773   8.000  17.000
```

The output above shows that Acid Index has outlier issue since its skewness number is greater than 1.

```
> skewness(wine$AcidIndex)
[1] 0.5435572
> summary(wine$AcidIndex)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.000   7.000   8.000   7.699   8.000  10.000
```

After a 95% trim, the output above shows that we have successfully addressed the skewness number of Acid Index.

In addition, as mentioned earlier, the existence of outliers on both tails of the majority of predictors is not deal-breaker in the regression assumptions, but it may affect the regression formulas. If we leave the variables the way they are with these outliers, we may run into issues with the coefficients later on during the model development process. If we trim these variables now, we may unnecessarily modify the original data, which should be avoided unless necessary.

Since there are pros and cons for both options to keep the variables the same or trim them, we won't trim them now in this section. Instead we will continue with the project to develop models. After the models are developed, we will examine the coefficients of the models to determine whether we should go back to trim the variables and rerun the models. If the betas don't make sense in real life (such as a negative beta for Label Appeal or STARS), we know that the coefficients are unreliable. Then we will go back to this stage to trim the variables and rerun the predictive models.

By replacing the original with the imputed variables for predictors with missing value. The following dataset is put together with no duplicate variable and no missing value to be used in the data development process.

```
> str(wine)
'data.frame':	12795 obs. of  24 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 ...
 $ Density               : num  0.993 1.028 0.995 0.996 0.995 ...
 $ LabelAppeal           : int  0 -1 -1 -1 0 0 0 1 0 0 ...
 $ AcidIndex             : num  8 7 8 6 9 10 8 7 6 8 ...
 $ M_ResidualSugar       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ IMP_ResidualSugar     : num  54.2 26.1 14.8 18.8 9.4 ...
 $ M_Chlorides           : num  0 0 0 0 1 0 0 0 0 0 ...
 $ IMP_Chlorides         : num  -0.567 -0.425 0.037 -0.425 0.0548 ...
 $ M_FreeSulfurDioxide   : num  1 0 0 0 0 0 0 0 0 0 ...
 $ IMP_FreeSulfurDioxide : num  30.8 15 214 22 -167 ...
 $ M_TotalSulfurDioxide  : num  0 0 0 0 0 0 0 0 1 0 ...
 $ IMP_TotalSulfurDioxide: num  268 -327 142 115 108 ...
 $ M_pH                  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ IMP_pH                : num  3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
 $ M_Sulphates           : num  0 0 0 0 0 0 0 1 0 0 ...
 $ IMP_Sulphates         : num  -0.59 0.7 0.48 1.83 1.77 ...
 $ M_Alcohol             : num  0 1 0 0 0 0 0 0 0 0 ...
 $ IMP_Alcohol           : num  9.9 10.5 22 6.2 13.7 ...
 $ M_STARS               : num  0 0 0 0 0 1 1 0 1 0 ...
 $ IMP_STARS             : num  2 3 3 1 2 ...
```

```
> summary(wine)
     INDEX           TARGET       FixedAcidity     VolatileAcidity     CitricAcid     
 Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400  
 1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300  
 Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100  
 Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084  
 3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800  
 Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600  
    Density         LabelAppeal          AcidIndex      M_ResidualSugar   IMP_ResidualSugar
 Min.   :0.8881   Min.   :-2.000000   Min.   : 6.000   Min.   :0.00000   Min.   :-127.800  
 1st Qu.:0.9877   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:0.00000   1st Qu.:   0.900  
 Median :0.9945   Median : 0.000000   Median : 8.000   Median :0.00000   Median :   4.900  
 Mean   :0.9942   Mean   :-0.009066   Mean   : 7.699   Mean   :0.04814   Mean   :   5.419  
 3rd Qu.:1.0005   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:0.00000   3rd Qu.:  14.900  
 Max.   :1.0992   Max.   : 2.000000   Max.   :10.000   Max.   :1.00000   Max.   : 141.150  
  M_Chlorides      IMP_Chlorides       M_FreeSulfurDioxide IMP_FreeSulfurDioxide
 Min.   :0.00000   Min.   :-1.17100   Min.   :0.00000     Min.   :-555.00      
 1st Qu.:0.00000   1st Qu.: 0.00000   1st Qu.:0.00000     1st Qu.:   5.00      
 Median :0.00000   Median : 0.04800   Median :0.00000     Median :  30.85      
 Mean   :0.04986   Mean   : 0.05482   Mean   :0.05057     Mean   :  30.85      
 3rd Qu.:0.00000   3rd Qu.: 0.12800   3rd Qu.:0.00000     3rd Qu.:  64.00      
 Max.   :1.00000   Max.   : 1.35100   Max.   :1.00000     Max.   : 623.00      
 M_TotalSulfurDioxide IMP_TotalSulfurDioxide      M_pH             IMP_pH       M_Sulphates     
 Min.   :0.0000       Min.   :-823.0         Min.   :0.00000   Min.   :0.480   Min.   :0.00000  
 1st Qu.:0.0000       1st Qu.:  34.0         1st Qu.:0.00000   1st Qu.:2.970   1st Qu.:0.00000  
 Median :0.0000       Median : 120.7         Median :0.00000   Median :3.208   Median :0.00000  
 Mean   :0.0533       Mean   : 120.7         Mean   :0.03087   Mean   :3.208   Mean   :0.09457  
 3rd Qu.:0.0000       3rd Qu.: 198.0         3rd Qu.:0.00000   3rd Qu.:3.450   3rd Qu.:0.00000  
 Max.   :1.0000       Max.   :1057.0         Max.   :1.00000   Max.   :6.130   Max.   :1.00000  
 IMP_Sulphates       M_Alcohol         IMP_Alcohol        M_STARS          IMP_STARS    
 Min.   :-3.1300   Min.   :0.00000   Min.   :-4.70   Min.   :0.0000   Min.   :1.000  
 1st Qu.: 0.3400   1st Qu.:0.00000   1st Qu.: 9.10   1st Qu.:0.0000   1st Qu.:2.000  
 Median : 0.5271   Median :0.00000   Median :10.49   Median :0.0000   Median :2.000  
 Mean   : 0.5271   Mean   :0.05104   Mean   :10.49   Mean   :0.2625   Mean   :2.042  
 3rd Qu.: 0.7700   3rd Qu.:0.00000   3rd Qu.:12.20   3rd Qu.:1.0000   3rd Qu.:2.042  
 Max.   : 4.2400   Max.   :1.00000   Max.   :26.50   Max.   :1.0000   Max.   :4.000  
```

Section 3: Model Development

In this section, five models below are developed, and one winning model is chosen based on multiple metrics in section 4 of the report.

1. Ordinary least square (OLS) multiple linear regression using stepwise variable automatic selection method
2. Poisson regression
3. Negative binomial regression
4. Zero-inflated Poisson regression (ZIP)
5. Zero-inflated negative binominal regression (ZINB)

Model #1: OLS Multiple Linear Regression

```
> summary(model1)

Call:
lm(formula = wine$TARGET ~ VolatileAcidity + Density + LabelAppeal +
    AcidIndex + IMP_Chlorides + IMP_FreeSulfurDioxide + IMP_TotalSulfurDioxide +
    IMP_pH + IMP_Sulphates + IMP_Alcohol + M_STARS + IMP_STARS,
    data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7043 -0.8554  0.0284  0.8536  6.1955

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.677e+00  4.481e-01  10.438  < 2e-16 ***
VolatileAcidity        -9.868e-02  1.484e-02  -6.647 3.10e-11 ***
Density                -8.544e-01  4.379e-01  -1.951 0.051088 .
LabelAppeal             4.637e-01  1.369e-02  33.862  < 2e-16 ***
AcidIndex              -2.337e-01  1.114e-02 -20.986  < 2e-16 ***
IMP_Chlorides          -1.139e-01  3.744e-02  -3.043 0.002350 **
IMP_FreeSulfurDioxide   2.946e-04  8.022e-05   3.673 0.000241 ***
IMP_TotalSulfurDioxide  2.333e-04  5.153e-05   4.527 6.03e-06 ***
IMP_pH                 -3.140e-02  1.739e-02  -1.806 0.070947 .
IMP_Sulphates          -3.224e-02  1.310e-02  -2.462 0.013841 *
IMP_Alcohol             1.240e-02  3.206e-03   3.867 0.000111 ***
M_STARS                -2.288e+00  2.699e-02 -84.752  < 2e-16 ***
IMP_STARS               7.809e-01  1.571e-02  49.715  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.313 on 12782 degrees of freedom
Multiple R-squared:  0.5361,	Adjusted R-squared:  0.5357
F-statistic:  1231 on 12 and 12782 DF,  p-value: < 2.2e-16

> AIC(model1)
[1] 43287.38
```

The output above has p-value less than 0.05, which means that the model is statistically significant at 95% confidence level. It has adjusted R-squared of 0.5357, which means that 53.57% of variation in the target variable can be explained by the model. Both Label Appeal and STARS have positive coefficients, which make sense in real life that as the wine bottles have better label design and ratings by wine experts, more cases are sold. Thus, it seems like the outlier issues in section 2 don't affect the regression formulas/coefficients in this model development process. Therefore, there's no need to go back to section 2 to trim the outliers in the predictors. We can proceed with this section 3 of model development.

Model #2: Poisson Regression

By putting all predictors in the Poisson regression analysis, we have the following result.

```
> summary(model2)

Call:
glm(formula = wine$TARGET ~ ., family = poisson(link = "log"),
    data = wine)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1621  -0.6500   0.0145   0.4562   3.7798

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.821e+00  1.967e-01   9.258  < 2e-16 ***
INDEX                   4.727e-07  1.095e-06   0.432 0.665916
FixedAcidity           -2.602e-04  8.192e-04  -0.318 0.750733
VolatileAcidity        -3.169e-02  6.522e-03  -4.859 1.18e-06 ***
CitricAcid              4.831e-03  5.896e-03   0.819 0.412507
Density                -2.990e-01  1.918e-01  -1.559 0.118952
LabelAppeal             1.586e-01  6.130e-03  25.864  < 2e-16 ***
AcidIndex              -8.546e-02  5.261e-03 -16.245  < 2e-16 ***
M_ResidualSugar         2.273e-02  2.340e-02   0.971 0.331310
IMP_ResidualSugar       9.001e-05  1.549e-04   0.581 0.561180
M_Chlorides            -4.144e-04  2.329e-02  -0.018 0.985801
IMP_Chlorides          -3.584e-02  1.647e-02  -2.176 0.029583 *
M_FreeSulfurDioxide     1.763e-02  2.320e-02   0.760 0.447453
IMP_FreeSulfurDioxide   1.009e-04  3.511e-05   2.873 0.004066 **
M_TotalSulfurDioxide    1.920e-02  2.245e-02   0.855 0.392552
IMP_TotalSulfurDioxide  8.328e-05  2.276e-05   3.658 0.000254 ***
M_pH                   -3.859e-02  2.991e-02  -1.290 0.196901
IMP_pH                 -1.255e-02  7.648e-03  -1.641 0.100740
M_Sulphates            -1.163e-02  1.757e-02  -0.662 0.507861
IMP_Sulphates          -1.220e-02  5.754e-03  -2.121 0.033924 *
M_Alcohol               1.637e-02  2.305e-02   0.710 0.477768
IMP_Alcohol             3.582e-03  1.409e-03   2.543 0.011003 *
M_STARS                -1.037e+00  1.697e-02 -61.133  < 2e-16 ***
IMP_STARS               1.883e-01  6.093e-03  30.910  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Only a couple predictors are significant in the model based on the output above using p-value for each variable. By removing insignificant variables and rerun the Poisson regression, we have the following result.

```
> summary(model2)

Call:
glm(formula = wine$TARGET ~ wine$VolatileAcidity + wine$LabelAppeal +
    wine$AcidIndex + wine$IMP_Chlorides + wine$IMP_FreeSulfurDioxide +
    wine$IMP_TotalSulfurDioxide + wine$IMP_Sulphates + wine$IMP_Alcohol +
    wine$M_STARS + wine$IMP_STARS, family = poisson(link = "log"),
    data = wine)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1916  -0.6451   0.0135   0.4543   3.7735

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  1.485e+00  4.556e-02  32.600  < 2e-16 ***
wine$VolatileAcidity        -3.194e-02  6.518e-03  -4.900  9.6e-07 ***
wine$LabelAppeal             1.584e-01  6.128e-03  25.851  < 2e-16 ***
wine$AcidIndex              -8.509e-02  5.180e-03 -16.429  < 2e-16 ***
wine$IMP_Chlorides          -3.602e-02  1.646e-02  -2.188 0.028652 *
wine$IMP_FreeSulfurDioxide   1.017e-04  3.508e-05   2.900 0.003730 **
wine$IMP_TotalSulfurDioxide  8.290e-05  2.273e-05   3.647 0.000265 ***
wine$IMP_Sulphates          -1.219e-02  5.752e-03  -2.120 0.034042 *
wine$IMP_Alcohol             3.602e-03  1.407e-03   2.560 0.010472 *
wine$M_STARS                -1.038e+00  1.696e-02 -61.220  < 2e-16 ***
wine$IMP_STARS               1.887e-01  6.090e-03  30.994  < 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: 13831  on 12784  degrees of freedom
AIC: 45795

Number of Fisher Scoring iterations: 6
```

Similar to model #1, both Label Appeal and STARS have positive coefficients, so there's no need go to back to section 2 to trim the variables with outlier issues. It's also interesting to notice that both M_STARS and IMP_STARS are significant variables in the model while they have different signs: negative in M_STARS and positive in IMP_STARS. Thus, we can expect to sell more wine cases on wine with known values in star ratings (M_STARS=0) than those with missing values (M_STARS=1). In addition, the higher the star ratings (IMP_STARS), the more cases of wine are sold.

Model #3: Negative Binomial Regression

By putting all predictors in the negative binomial regression analysis, we have the following result.

```
> summary(model3)

Call:
glm.nb(formula = wine$TARGET ~ ., data = wine, init.theta = 40527.13939,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1620  -0.6500   0.0145   0.4561   3.7796

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              1.821e+00  1.967e-01   9.258  < 2e-16 ***
INDEX                    4.727e-07  1.095e-06   0.432 0.665958
FixedAcidity            -2.603e-04  8.193e-04  -0.318 0.750726
VolatileAcidity         -3.169e-02  6.523e-03  -4.859 1.18e-06 ***
CitricAcid               4.831e-03  5.896e-03   0.819 0.412523
Density                 -2.991e-01  1.918e-01  -1.559 0.118961
LabelAppeal              1.586e-01  6.131e-03  25.862  < 2e-16 ***
AcidIndex               -8.547e-02  5.261e-03 -16.245  < 2e-16 ***
M_ResidualSugar          2.274e-02  2.340e-02   0.971 0.331328
IMP_ResidualSugar        9.002e-05  1.549e-04   0.581 0.561163
M_Chlorides             -4.146e-04  2.329e-02  -0.018 0.985794
IMP_Chlorides           -3.584e-02  1.648e-02  -2.176 0.029584 *
M_FreeSulfurDioxide      1.763e-02  2.320e-02   0.760 0.447464
IMP_FreeSulfurDioxide    1.009e-04  3.511e-05   2.873 0.004067 **
M_TotalSulfurDioxide     1.920e-02  2.246e-02   0.855 0.392561
IMP_TotalSulfurDioxide   8.328e-05  2.276e-05   3.658 0.000254 ***
M_pH                    -3.860e-02  2.991e-02  -1.290 0.196897
IMP_pH                  -1.255e-02  7.649e-03  -1.641 0.100731
M_Sulphates             -1.163e-02  1.757e-02  -0.662 0.507847
IMP_Sulphates           -1.221e-02  5.755e-03  -2.121 0.033924 *
M_Alcohol                1.637e-02  2.305e-02   0.710 0.477778
IMP_Alcohol              3.582e-03  1.409e-03   2.542 0.011008 *
M_STARS                 -1.037e+00  1.697e-02 -61.131  < 2e-16 ***
IMP_STARS                1.883e-01  6.094e-03  30.908  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for Negative Binomial(40527.14) family taken to be 1)

    Null deviance: 22860  on 12794  degrees of freedom
Residual deviance: 13819  on 12771  degrees of freedom
AIC: 45812

Number of Fisher Scoring iterations: 1


              Theta:  40527
          Std. Err.:  34507
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -45762.3
```

Using p-values of the output above, only a couple predictors are significant. By removing insignificant variables and rerun the model, we have the following result.

```
> summary(model3)

Call:
glm.nb(formula = wine$TARGET ~ wine$VolatileAcidity + wine$LabelAppeal +
    wine$AcidIndex + wine$IMP_Chlorides + wine$IMP_FreeSulfurDioxide +
    wine$IMP_TotalSulfurDioxide + wine$IMP_Sulphates + wine$IMP_Alcohol +
    wine$M_STARS + wine$IMP_STARS, data = wine, init.theta = 40508.37626,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1915  -0.6451   0.0135   0.4542   3.7733

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 1.485e+00  4.556e-02  32.599  < 2e-16 ***
wine$VolatileAcidity       -3.194e-02  6.518e-03  -4.900 9.61e-07 ***
wine$LabelAppeal            1.584e-01  6.128e-03  25.849  < 2e-16 ***
wine$AcidIndex             -8.510e-02  5.180e-03 -16.429  < 2e-16 ***
wine$IMP_Chlorides         -3.602e-02  1.646e-02  -2.188 0.028654 *
wine$IMP_FreeSulfurDioxide  1.017e-04  3.508e-05   2.900 0.003731 **
wine$IMP_TotalSulfurDioxide 8.290e-05  2.273e-05   3.647 0.000265 ***
wine$IMP_Sulphates         -1.219e-02  5.753e-03  -2.120 0.034043 *
wine$IMP_Alcohol            3.602e-03  1.407e-03   2.560 0.010477 *
wine$M_STARS               -1.038e+00  1.696e-02 -61.218  < 2e-16 ***
wine$IMP_STARS              1.887e-01  6.090e-03  30.992  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(40508.38) family taken to be 1)

    Null deviance: 22860  on 12794  degrees of freedom
Residual deviance: 13830  on 12784  degrees of freedom
AIC: 45797

Number of Fisher Scoring iterations: 1

              Theta:  40508
          Std. Err.:  34489
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -45773.42
```

Based on the output above, model #3 has the same list of significant variables as model #2 with similar coefficients. Some of the betas are slightly different such as Acid Index. However, AIC’s for both models are different, which means that they’re still different models.

Model #4: ZIP

By putting all predictors in the ZIP model, we have the following result.

```
> summary(model4)

Call:
zeroinfl(formula = wine$TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + Density +
    LabelAppeal + AcidIndex + M_ResidualSugar + IMP_ResidualSugar + M_Chlorides + IMP_Chlorides +
    M_FreeSulfurDioxide + IMP_FreeSulfurDioxide + M_TotalSulfurDioxide + IMP_TotalSulfurDioxide +
    M_pH + IMP_pH + M_Sulphates + IMP_Sulphates + M_Alcohol + IMP_Alcohol + M_STARS +
    IMP_STARS, data = wine)

Pearson residuals:
      Min        1Q    Median        3Q       Max
-2.340615 -0.415120 -0.002201  0.378937  5.666990

Count model coefficients (poisson with log link):
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.468e+00  2.025e-01   7.249 4.18e-13 ***
FixedAcidity            3.621e-04  8.409e-04   0.431   0.6667
VolatileAcidity        -1.240e-02  6.713e-03  -1.847   0.0648 .
CitricAcid              1.047e-03  6.019e-03   0.174   0.8619
Density                -2.714e-01  1.978e-01  -1.372   0.1700
LabelAppeal             2.322e-01  6.320e-03  36.736  < 2e-16 ***
AcidIndex              -2.478e-02  5.545e-03  -4.469 7.87e-06 ***
M_ResidualSugar         2.306e-02  2.391e-02   0.964   0.3348
IMP_ResidualSugar      -6.410e-05  1.586e-04  -0.404   0.6861
M_Chlorides             2.092e-03  2.387e-02   0.088   0.9302
IMP_Chlorides          -2.287e-02  1.689e-02  -1.354   0.1758
M_FreeSulfurDioxide     4.525e-03  2.367e-02   0.191   0.8484
IMP_FreeSulfurDioxide   2.613e-05  3.542e-05   0.738   0.4606
M_TotalSulfurDioxide   -3.620e-03  2.308e-02  -0.157   0.8754
IMP_TotalSulfurDioxide -1.565e-05  2.261e-05  -0.692   0.4890
M_pH                   -6.799e-03  3.068e-02  -0.222   0.8246
IMP_pH                  5.134e-03  7.848e-03   0.654   0.5130
M_Sulphates            -7.363e-03  1.798e-02  -0.410   0.6822
IMP_Sulphates          -1.869e-04  5.914e-03  -0.032   0.9748
M_Alcohol               2.735e-04  2.363e-02   0.012   0.9908
IMP_Alcohol             6.926e-03  1.439e-03   4.814 1.48e-06 ***
M_STARS                -1.863e-01  1.858e-02 -10.030  < 2e-16 ***
IMP_STARS               1.043e-01  6.409e-03  16.270  < 2e-16 ***
Zero-inflation model coefficients (binomial with logit link):
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -3.5568158  1.3643193  -2.607 0.009133 **
FixedAcidity            0.0034130  0.0054349   0.628 0.530018
VolatileAcidity         0.1903772  0.0431724   4.410 1.04e-05 ***
CitricAcid             -0.0146805  0.0393847  -0.373 0.709337
Density                 0.9186510  1.2914077   0.711 0.476863
LabelAppeal             0.7297796  0.0422482  17.274  < 2e-16 ***
AcidIndex               0.4928941  0.0319419  15.431  < 2e-16 ***
M_ResidualSugar         0.0075088  0.1571279   0.048 0.961885
IMP_ResidualSugar      -0.0010982  0.0010275  -1.069 0.285151
M_Chlorides             0.0947962  0.1594084   0.595 0.552061
IMP_Chlorides           0.0917016  0.1081743   0.848 0.396594
M_FreeSulfurDioxide    -0.1329464  0.1528819  -0.870 0.384518
IMP_FreeSulfurDioxide  -0.0007608  0.0002385  -3.189 0.001426 **
M_TotalSulfurDioxide   -0.1728922  0.1532980  -1.128 0.259397
IMP_TotalSulfurDioxide -0.0009519  0.0001506  -6.322 2.58e-10 ***
M_pH                    0.2903890  0.1925911   1.508 0.131606
IMP_pH                  0.2032362  0.0504837   4.026 5.68e-05 ***
M_Sulphates             0.1067880  0.1123442   0.951 0.341836
IMP_Sulphates           0.1375941  0.0382478   3.597 0.000321 ***
M_Alcohol              -0.1613730  0.1529210  -1.055 0.291302
IMP_Alcohol             0.0278925  0.0094048   2.966 0.003019 **
M_STARS                 6.0552029  0.3568026  16.971  < 2e-16 ***
IMP_STARS              -3.8367600  0.3425468 -11.201  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 56
Log-likelihood: -2.04e+04 on 46 Df
```

Using p-values in the output above, by eliminating insignificant variables in both tables and rerunning the ZIP model, we have the following result.

```
Call:
zeroinfl(formula = wine$TARGET ~ wine$VolatileAcidity + wine$LabelAppeal + wine$AcidIndex +
    wine$IMP_Alcohol + wine$M_STARS + wine$IMP_STARS + wine$IMP_FreeSulfurDioxide +
    wine$IMP_TotalSulfurDioxide + wine$IMP_pH + wine$IMP_Sulphates, data = wine)

Pearson residuals:
      Min        1Q    Median        3Q       Max
-2.335334 -0.418925 -0.002208  0.379241  5.785516

Count model coefficients (poisson with log link):
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  1.199e+00  5.526e-02  21.688  < 2e-16 ***
wine$VolatileAcidity        -1.239e-02  6.706e-03  -1.848   0.0646 .
wine$LabelAppeal             2.321e-01  6.319e-03  36.729  < 2e-16 ***
wine$AcidIndex              -2.481e-02  5.477e-03  -4.529 5.92e-06 ***
wine$IMP_Alcohol             6.985e-03  1.437e-03   4.860 1.18e-06 ***
wine$M_STARS                -1.865e-01  1.857e-02 -10.046  < 2e-16 ***
wine$IMP_STARS               1.046e-01  6.407e-03  16.325  < 2e-16 ***
wine$IMP_FreeSulfurDioxide   2.611e-05  3.539e-05   0.738   0.4606
wine$IMP_TotalSulfurDioxide -1.618e-05  2.260e-05  -0.716   0.4742
wine$IMP_pH                  5.340e-03  7.842e-03   0.681   0.4959
wine$IMP_Sulphates          -2.080e-05  5.911e-03  -0.004   0.9972

Zero-inflation model coefficients (binomial with logit link):
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.6606870  0.4888595  -5.443 5.25e-08 ***
wine$VolatileAcidity         0.1909649  0.0430345   4.437 9.10e-06 ***
wine$LabelAppeal             0.7302642  0.0421919  17.308  < 2e-16 ***
wine$AcidIndex               0.4977355  0.0311876  15.959  < 2e-16 ***
wine$IMP_Alcohol             0.0279718  0.0094032   2.975 0.002933 **
wine$M_STARS                 6.0565220  0.3562297  17.002  < 2e-16 ***
wine$IMP_STARS              -3.8377597  0.3421020 -11.218  < 2e-16 ***
wine$IMP_FreeSulfurDioxide  -0.0007813  0.0002383  -3.279 0.001043 **
wine$IMP_TotalSulfurDioxide -0.0009605  0.0001504  -6.386 1.70e-10 ***
wine$IMP_pH                  0.2042039  0.0503938   4.052 5.07e-05 ***
wine$IMP_Sulphates           0.1345879  0.0381229   3.530 0.000415 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ZIP and ZINB results have two tables in the output. The second table “zero-inflation model” uses logit or probit regression to predict whether the target variable is 0 or not 0 (is the wine sold or not sold?). If the target variable is predicted to be non-0, the first table “count model” uses Poisson or NB to predict the number of wine cases sold.

Label Appeal has positive coefficient in both tables, which means that better label design yields a higher probability of the wine being sold and also more wine cases being sold. However, it’s interesting that M_STARS and IMP_STARS have different signs in the two tables. Positive M_STARS in the second table means that wine with missing values in star ratings (M_STARS=1) is more likely to be sold (target=non-0). Negative M_STARS in the first table means that wine with known values in star ratings (M_STARS=0) has more wine cases sold. Negative IMP_STARS in the second table means that wine with lower star ratings is more likely to be sold (target=non-0). Perhaps wine distribution companies are looking for hidden gems in the industry by taking the risk on lower-rated wine. Positive IMP_STARS in the first table means that as the star ratings go up, more wine cases are sold, which make sense in real life.

Most of the coefficients in the first table have the same sign (negative or positive) as the result in model #2 with different absolute value. However, Total Sulfur Dioxide is an exception with different sign and absolute value between this model and model #2.

Model #5: ZINB

By putting all predictors in the ZINB model, we have the following result.

```
> summary(model5)

Call:
zeroinfl(formula = wine$TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + Density +
    LabelAppeal + AcidIndex + M_ResidualSugar + IMP_ResidualSugar + M_Chlorides + IMP_Chlorides +
    M_FreeSulfurDioxide + IMP_FreeSulfurDioxide + M_TotalSulfurDioxide + IMP_TotalSulfurDioxide +
    M_pH + IMP_pH + M_Sulphates + IMP_Sulphates + M_Alcohol + IMP_Alcohol + M_STARS +
    IMP_STARS, data = wine, dist = "negbin", EM = TRUE)

Pearson residuals:
      Min        1Q    Median        3Q       Max
-2.340632 -0.415125 -0.002201  0.378936  5.667016

Count model coefficients (negbin with log link):
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.468e+00  2.025e-01   7.249 4.19e-13 ***
FixedAcidity            3.621e-04  8.409e-04   0.431  0.66674
VolatileAcidity        -1.240e-02  6.713e-03  -1.847  0.06480 .
CitricAcid              1.047e-03  6.019e-03   0.174  0.86191
Density                -2.714e-01  1.978e-01  -1.372  0.17004
LabelAppeal             2.322e-01  6.320e-03  36.736  < 2e-16 ***
AcidIndex              -2.478e-02  5.545e-03  -4.469 7.86e-06 ***
M_ResidualSugar         2.306e-02  2.391e-02   0.965  0.33479
IMP_ResidualSugar      -6.409e-05  1.586e-04  -0.404  0.68608
M_Chlorides             2.091e-03  2.387e-02   0.088  0.93019
IMP_Chlorides          -2.287e-02  1.689e-02  -1.354  0.17578
M_FreeSulfurDioxide     4.525e-03  2.367e-02   0.191  0.84839
IMP_FreeSulfurDioxide   2.613e-05  3.542e-05   0.738  0.46060
M_TotalSulfurDioxide   -3.622e-03  2.308e-02  -0.157  0.87532
IMP_TotalSulfurDioxide -1.564e-05  2.261e-05  -0.692  0.48907
M_pH                   -6.801e-03  3.068e-02  -0.222  0.82459
IMP_pH                  5.134e-03  7.848e-03   0.654  0.51303
M_Sulphates            -7.363e-03  1.798e-02  -0.410  0.68217
IMP_Sulphates          -1.872e-04  5.914e-03  -0.032  0.97475
M_Alcohol               2.744e-04  2.363e-02   0.012  0.99073
IMP_Alcohol             6.926e-03  1.439e-03   4.814 1.48e-06 ***
M_STARS                -1.863e-01  1.858e-02 -10.030  < 2e-16 ***
IMP_STARS               1.043e-01  6.409e-03  16.270  < 2e-16 ***
Log(theta)              1.232e+01  3.914e+00   3.148  0.00164 **
```

```
Zero-inflation model coefficients (binomial with logit link):
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -3.5577787  1.3643632  -2.608 0.009117 **
FixedAcidity            0.0034130  0.0054349   0.628 0.530025
VolatileAcidity         0.1903755  0.0431727   4.410 1.04e-05 ***
CitricAcid             -0.0146764  0.0393850  -0.373 0.709418
Density                 0.9199998  1.2914185   0.712 0.476220
LabelAppeal             0.7297887  0.0422487  17.274  < 2e-16 ***
AcidIndex               0.4928952  0.0319422  15.431  < 2e-16 ***
M_ResidualSugar         0.0075311  0.1571289   0.048 0.961773
IMP_ResidualSugar      -0.0010982  0.0010275  -1.069 0.285151
M_Chlorides             0.0947874  0.1594095   0.595 0.552100
IMP_Chlorides           0.0916944  0.1081752   0.848 0.396635
M_FreeSulfurDioxide    -0.1329500  0.1528829  -0.870 0.384508
IMP_FreeSulfurDioxide  -0.0007608  0.0002385  -3.189 0.001426 **
M_TotalSulfurDioxide   -0.1729121  0.1532993  -1.128 0.259346
IMP_TotalSulfurDioxide -0.0009519  0.0001506  -6.322 2.58e-10 ***
M_pH                    0.2903880  0.1925929   1.508 0.131610
IMP_pH                  0.2032389  0.0504840   4.026 5.68e-05 ***
M_Sulphates             0.1067911  0.1123450   0.951 0.341826
IMP_Sulphates           0.1375940  0.0382481   3.597 0.000321 ***
M_Alcohol              -0.1613623  0.1529220  -1.055 0.291337
IMP_Alcohol             0.0278939  0.0094049   2.966 0.003018 **
M_STARS                 6.0556588  0.3569399  16.965  < 2e-16 ***
IMP_STARS              -3.8371855  0.3426783 -11.198  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 224194.5129
Number of iterations in BFGS optimization: 1
Log-likelihood: -2.04e+04 on 47 Df
```

Using p-values of the output above, we see that only a couple predictors are significant. By eliminating all insignificant predictors in both tables and rerunning the model, we have the following result.

```
> summary(model5)

Call:
zeroinfl(formula = wine$TARGET ~ wine$VolatileAcidity + wine$LabelAppeal + wine$AcidIndex +
    wine$IMP_Alcohol + wine$M_STARS + wine$IMP_STARS + wine$IMP_FreeSulfurDioxide +
    wine$IMP_TotalSulfurDioxide + wine$IMP_pH + wine$IMP_Sulphates, data = wine, dist = "negbin",
    EM = TRUE)

Pearson residuals:
      Min        1Q    Median        3Q       Max
-2.335344 -0.418920 -0.002206  0.379241  5.785566

Count model coefficients (negbin with log link):
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 1.199e+00  5.526e-02  21.688  < 2e-16 ***
wine$VolatileAcidity       -1.239e-02  6.706e-03  -1.848  0.06458 .
wine$LabelAppeal            2.321e-01  6.319e-03  36.729  < 2e-16 ***
wine$AcidIndex             -2.481e-02  5.477e-03  -4.529 5.92e-06 ***
wine$IMP_Alcohol            6.985e-03  1.437e-03   4.860 1.18e-06 ***
wine$M_STARS               -1.866e-01  1.857e-02 -10.046  < 2e-16 ***
wine$IMP_STARS              1.046e-01  6.407e-03  16.325  < 2e-16 ***
wine$IMP_FreeSulfurDioxide  2.611e-05  3.539e-05   0.738  0.46056
wine$IMP_TotalSulfurDioxide -1.617e-05 2.260e-05  -0.716  0.47423
wine$IMP_pH                 5.340e-03  7.842e-03   0.681  0.49594
wine$IMP_Sulphates         -2.108e-05  5.911e-03  -0.004  0.99716
Log(theta)                  1.232e+01  3.869e+00   3.184  0.00145 **
```

```
Zero-inflation model coefficients (binomial with logit link):
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.6603457  0.4889413  -5.441 5.30e-08 ***
wine$VolatileAcidity        0.1909636  0.0430348   4.437 9.10e-06 ***
wine$LabelAppeal            0.7302710  0.0421923  17.308  < 2e-16 ***
wine$AcidIndex              0.4977362  0.0311878  15.959  < 2e-16 ***
wine$IMP_Alcohol            0.0279730  0.0094033   2.975 0.002932 **
wine$M_STARS                6.0569125  0.3563470  16.997  < 2e-16 ***
wine$IMP_STARS             -3.8381241  0.3422143 -11.216  < 2e-16 ***
wine$IMP_FreeSulfurDioxide -0.0007813  0.0002383  -3.279 0.001043 **
wine$IMP_TotalSulfurDioxide -0.0009605  0.0001504  -6.386 1.70e-10 ***
wine$IMP_pH                 0.2042026  0.0503942   4.052 5.08e-05 ***
wine$IMP_Sulphates          0.1345887  0.0381231   3.530 0.000415 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 223701.1953
Number of iterations in BFGS optimization: 1
Log-likelihood: -2.041e+04 on 23 Df
```

Similar to our observations between models #2 and #3, using the output above, models #4 and #5 have the same list of predictors with similar betas. However, they have different AIC values, so they are still different models.

```
> mean(wine$TARGET)
[1] 3.029074
> var(wine$TARGET)
[1] 3.710895
```

As shown above, the mean and variance of the target variable are almost the same, which explains why the results of models #2 and #3, models #4 and #5 are similar. When the mean is the same as the variance, Poisson and NB give the same result and yield the same model.

Section 4: Model Evaluation

To compare the five models in section 3, we use the following metrics.

- AIC
- MSE (mean squared error)
- SSE (sum of square of error)

To calculate MSE and SSE for each model, we apply the predictive formulas of each model to the train dataset to generate five new variables with the forecasted results for the target variable.

	AIC	MSE	SSE
model 1	43287.38	1.7213	22024.10
model 2	45795.01	1.7349	7358.64
model 3	45797.42	1.7349	7358.64
model 4	40853.97	1.6205	20734.84
model 5	40856.10	1.6205	20734.85

Quantitatively, the best model is the one with the lowest AIC, lowest MSE, and lowest SSE. Based on the table above, using AIC, model #4 is the winner. Using MSE, models #4 and #5 tie as winner. Using SSE, models #2 and #3 tie as the winner. Since the analyst doesn't have any industry knowledge, quantitative reasoning is the only factor used to evaluate the models. Based on the table result above, model #4 ZIP is the champion model chosen.

CONCLUSION

In conclusion, the wine sales project starts with the data exploration section to get to know the data. In this stage, we examine the size (number of observations and variables) of the dataset, identify variables with missing value and outlier issues, and get the descriptive statistics of each variable such as five-number summary, mean, and median. The second section data preparation has two parts: the first part addresses missing value issues, and the second part addresses outlier issues. Majority of the variables have outliers on both ends. Though this is not a deal-breaker on regression assumptions, the existence of outliers on both tails might affect the regression formulas. There are pros and cons for both options of keeping the variables the same or trimming them. So, we keep the variables the same and proceed to the next section. The final decision will be made as we get to the model development phase by examining the betas of the models.

The project then proceed to the third section of model development where five models are created:

1. OLS multiple linear regression
2. Poisson regression
3. Negative binomial regression
4. Zero-inflated Poisson regression
5. Zero-inflated negative binomial regression

By examining the coefficients of these models, especially Label Appeal and STARS, we conclude that there's no need to go back to section 2 to trim the data. Thus, we move forward with the last section of the project: model evaluation. By using three metrics AIC, MSE, and SSE, model #4 zero-inflated Poisson regression is chosen as the champion model. A stand-alone scoring program is developed using this model #4 to forecast the number of wine cases sold for future dataset.

Researchers who want to continue to study this project are encouraged to add more variables, gather more data, and speak to industry experts to develop better models than the ones created here. The last step is the most important recommendation since the data analyst of this project doesn't have any industry knowledge about wine, its characteristics, and chemical components.