**Wine Assignment**

Lauren Camero

**Bingo Bonus:** (20 Points) Develop a LOGISTIC / POISSON model

**Introduction:**

In this report, we have completed a data quality check, an initial exploratory data analysis, and built models to predict the number of sample cases of wine that were purchased by wine distribution companies after providing tasting samples to restaurants and wine stores.

**Data:**

We used a data set consisting of 12,000 records and different wine characteristics that a wine manufacturer can use to predict how many cases of each wine type will be sold. The table below defines the variables given.

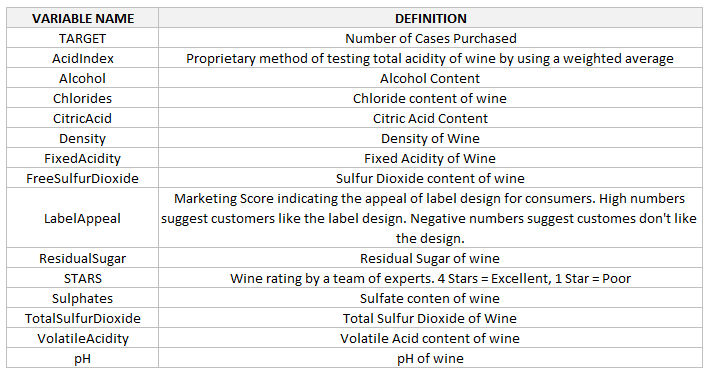


Figure 1: definition of variables

**Exploratory Data Analysis:**

We began the exploratory data analysis by observing boxplots to see if the continuous variables were highly skewed in one direction or contained extreme outliers. The figure below illustrates 14 variables. Visually, we only detected questionable data in Stars after looking at the boxplots and seeing outliers very far away from the normal range.

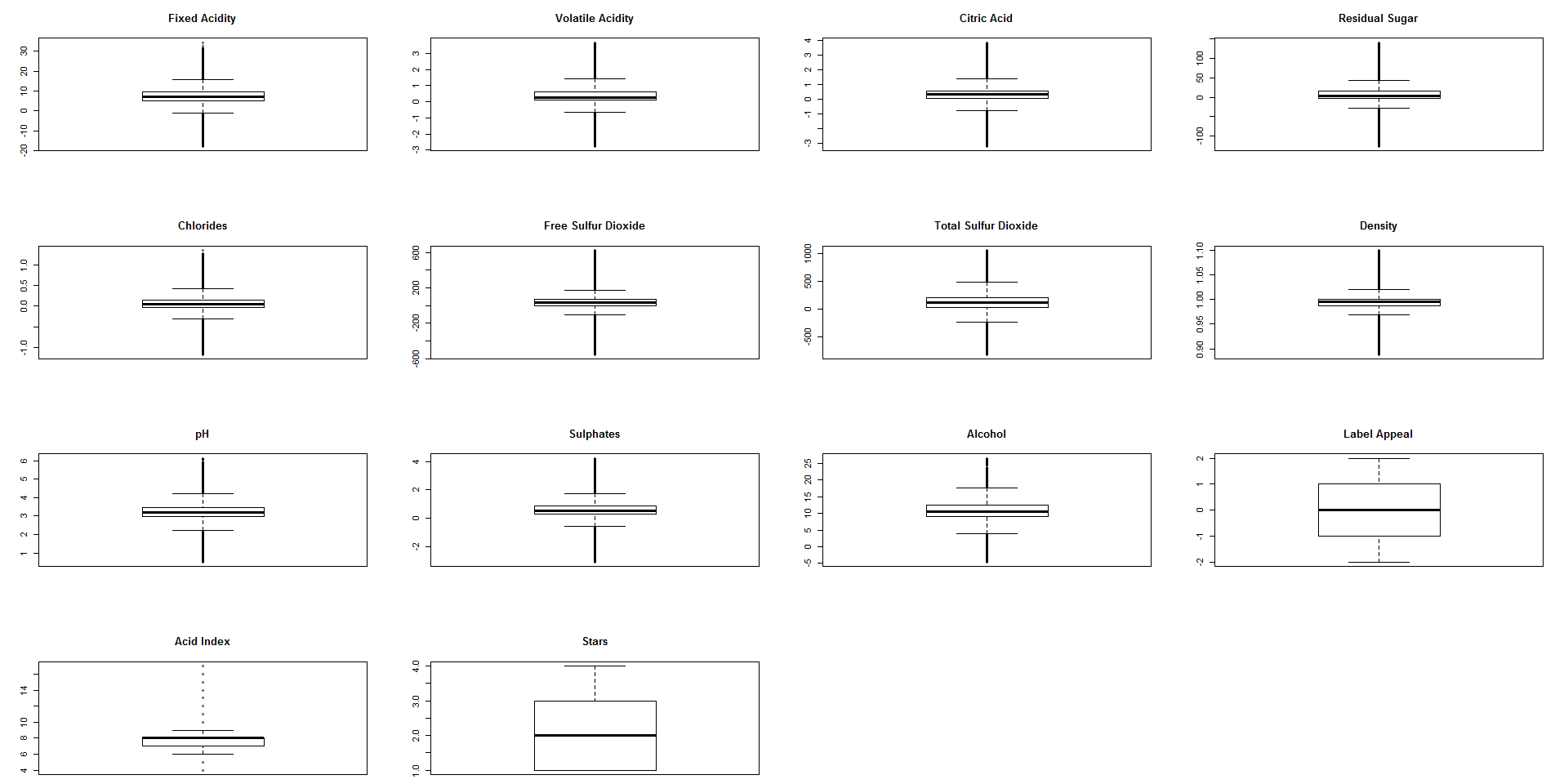


Figure 2: boxplot of continuous variables

Additionally, we created histograms for the Star feature that seemed to have the largest positive skew that would require transformation. After running the histogram, we see that Star is a categorical variable. Since none of the variables were skewed positively or negatively, we did not need to transform the variables using the log or square root.

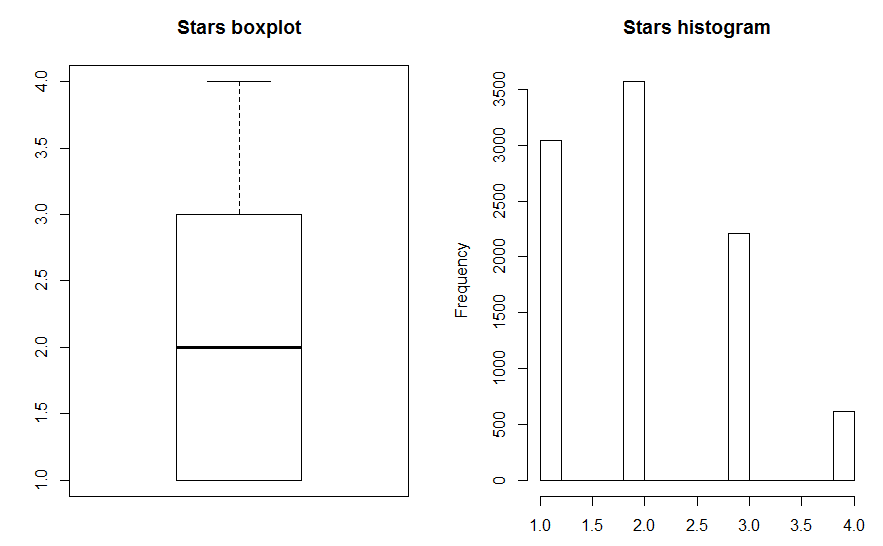


Figure 3: histogram of Star feature

We additionally ran the summary statistics for the dataset and found the variables below had missing data that would require imputations. Since there are too many missing records, we found another way to correct the missing data.

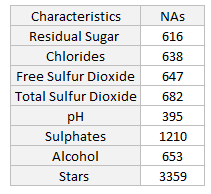


Figure 4: Missing data check

**Data Transformation:**

We imputed the missing values by first creating flags to let us know that the records were imputed, then taking the mean of the variable and replacing every missing with record with that mean.

**Build Models**

Before building the models, we assess the multicollinearity of the variables within the model by plotting the correlations of the variables below. We see a strong positive correlation with our response variable and label appeal and a strong negative correlation between the sample number of cases purchased and the acid index on the wine. We also see a strong correlation between the label appeal and its star rating. It is surprising that there are no other strong relationships between variables in this dataset.

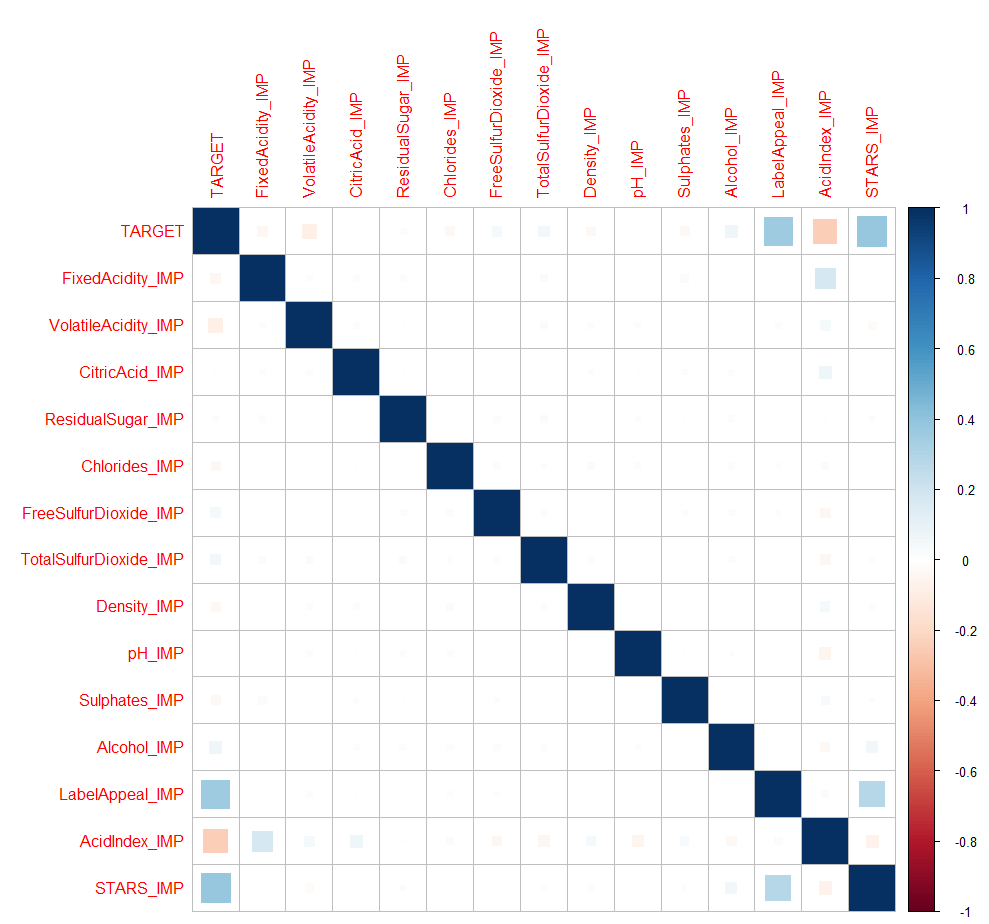


Figure 5: correlations matrix between variables

Using this information to pick our initial set of variables, we set up a linear regression using the formula below and chose 4 different model approaches: manual, forward, backward, and stepwise.

+++ … +

In addition, we run a Poisson, Zero-Inflated Poisson, Negative Binomial, Zero-Inflated Negative Binomial Regressions to find the best result.

**Model 1: Manual Selection**

We first used a manual selection of variables. We picked the predictive variables by observing the correlation matrix and taking the variables that were independent and observing their outputs. Below is a list of the variables we included in the model. We excluded the variables that were highly correlated with other variables or visually illustrated no correlation with our response variable. We also ran a manual model with more variables and found that it did not produce the best results. We used this model’s summary outputs to see the most significant variables and only included those in out manual model going forward. We see the model results for the manual model with only remaining significant variables.

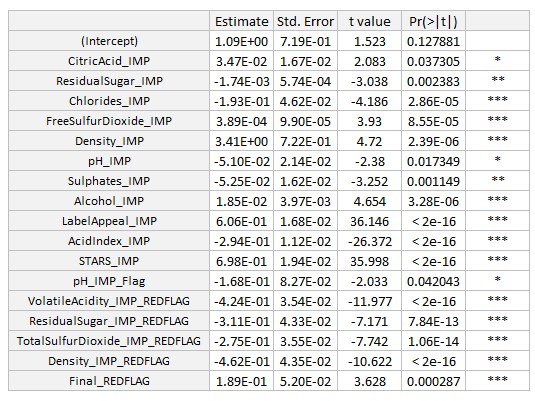


Figure 6: summary results of manual model

**Model 2: Stepwise Backward Selection**

The next model we built was the backwardstepwise selection model. We added back in the remaining variables we had removed in the manual model. The backward stepwise model went through 8 iterations to end up with the following results.

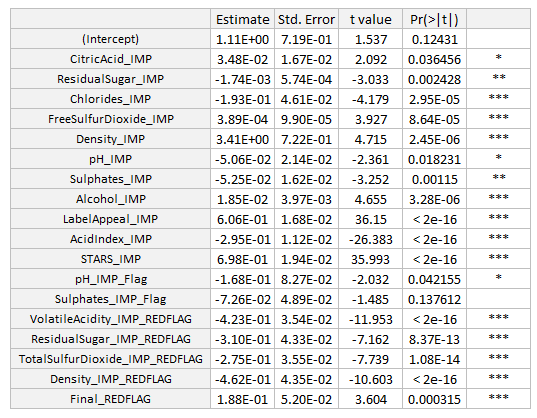


Figure 7: summary of backward model

**Model 3: Stepwise Forward Selection**

The third model we ran used the same linear model variables as the forward selection process. This method only required one step.

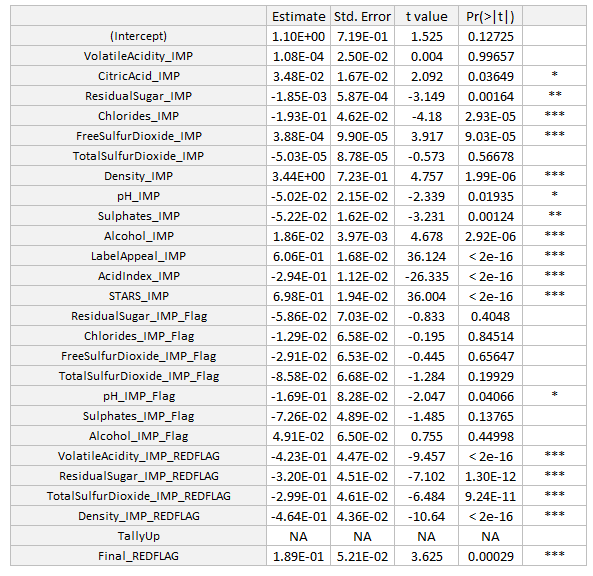


Figure 8: summary of forward stepwise mode

**Model 4: Stepwise Forward & Backward Selection**

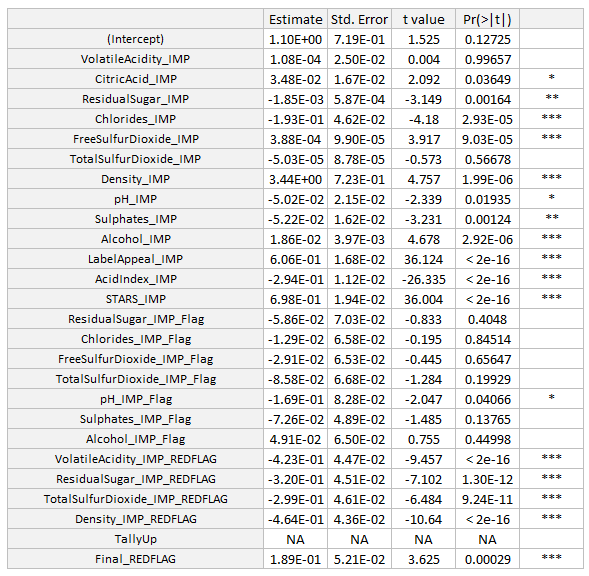
Following the forward model selection, we used a stepwise model selection process using both directions. This took 9 steps and left us with the remaining variables shown in the summary output below. Linear regressions are not usually the best choice for counting models since the relationship between inputs and targets are not linear. 

Figure 9: stepwise both direction GLM

**Model 5: Poisson Regression**

We next started exploring poisson models since we are counting the number of cases in our prediction. Using the same variables selected from the first manual model, we ran the remaining regression models starting with the Poisson Regression.

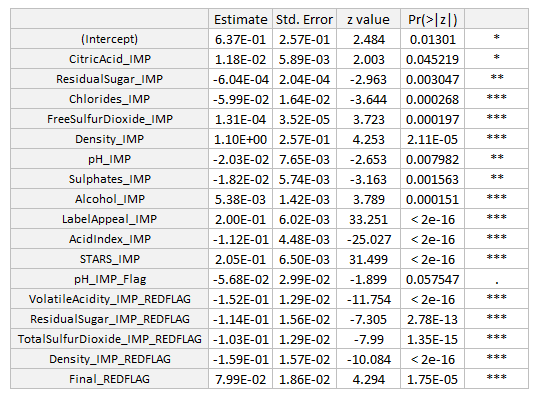


Figure 10: poisson regression summary results

**Model 5: Zero inflated Poisson Regression**

We next ran a zero inflated poisson regression and ended up with the following results. The zero inflated model is best when there a large amount of spikes in the 0 values. Given the graph below, we would assume that one of the zero inflated distributions would be the best performing model for this dataset.

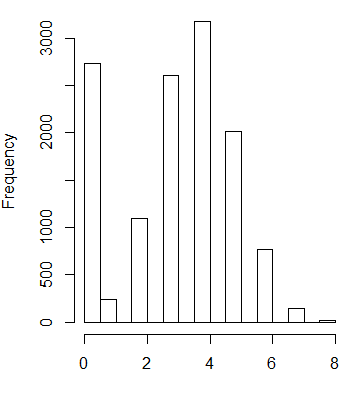


Figure : histogram of response variable

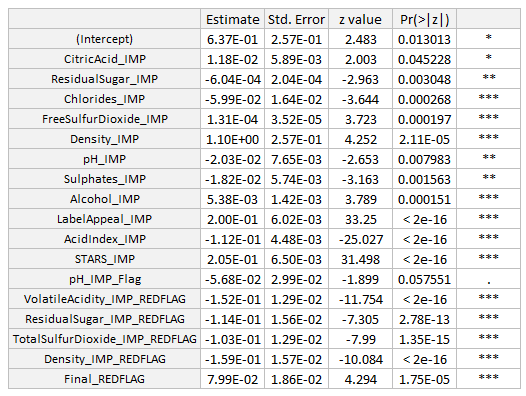


Figure 12: Zero inflated poisson regression

**Model 5: Negative Binomial Regression**

Similarly, the negative binomial regression produced the following results. The negative binomial distribution model is the best counting model when the variance is greater than the mean.

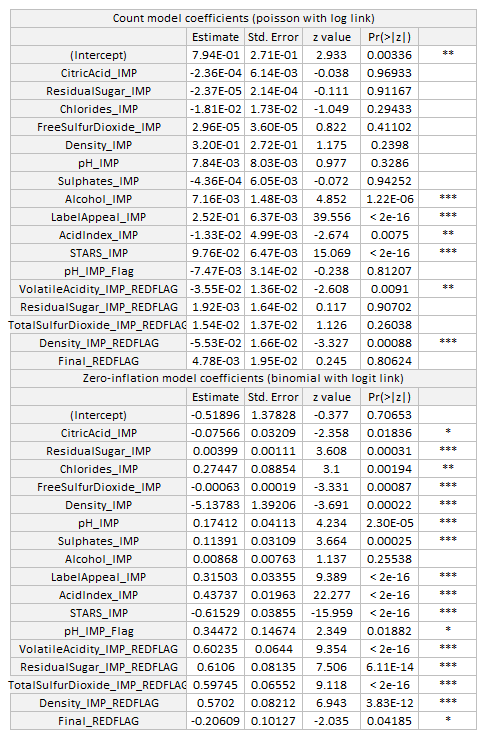


Figure 13: Zero Inflated Binomial regression summary results

**Model 6: Zero Inflated Negative Binomial Regression**

Finally, we run the Zero Inflated Negative Binomial Regression. The zero inflated negative binomial regression uses the same methodology as the zero inflated poisson but uses a different distribution.

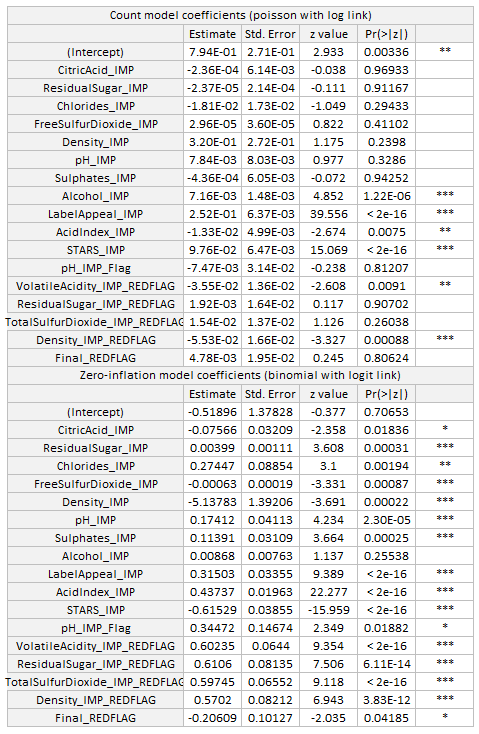


Figure 14: Zero Inflated Negative Binomial Regression

**Model Selection:**

We used a mixture of results from AIC, BIC, and MSE to pick our best model. The negative binomial and poisson regressions had similar results. This occurs when the variance is not significantly greater than the mean.

Since the metrics we used to pick the best model did not give one result, we chose to use the AIC as the primary metric to pick the champion model and went with the Zero Inflated Poisson Regression.

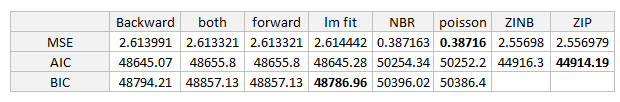


Figure 15: metric comparison between models

**Conclusion:**

After running six different models, we picked the zero inflated poisson regression model as out champion model to predict the number of cases purchased after wine tasting. This was not surprising after running an exploratory data analysis and seeing that the response variable followed a somewhat normal distribution besides a high frequency at 0 cases.