Introduction

In Project 3: Wine Sales, data related to commercially available wines are examined and five models are fitted to predict the number of sample cases of wine that were purchased by wine distribution companies. The data set includes 12,795 observations and 16 variables including the target variable, labeled TARGET, and INDEX, which is an identifier variable and will not be examined in the analysis. The variables are mostly related to the chemical properties of the wine being sold. Each variable is explored based on its original state to understand size, shape, completeness and potential relationships with other variables. Data preparation is then conducted to address missing values, transform existing variables, and create new variables that attempt to augment the overall predictive power. Five separate models are fitted and evaluated based on measures of AIC and average squared error. The “best” model is then used to score out-of-sample data.

Data Exploration

Each observation in the data set represents an individual wine. There are 16 total variables – 14 variables contain information regarding each wine’s chemical properties. The other 2 variables are the target variable, TARGET, and an identifier variable called INDEX.

TARGET

The TARGET variable represents the number of sample cases of wine that were purchased by wine distribution companies. Figure 1 shows the distribution of TARGET with several zero values followed by what appears to be a bell curve distribution around the value of 4. With a mean of 3.03, standard deviation of 1.93, and a variance of 3.7249, TARGET is “over-dispersed” (the variance is greater than the mean) necessitating consideration of a negative binomial distribution instead of a Poisson distribution. The target variable has no missing values and 17 outliers.

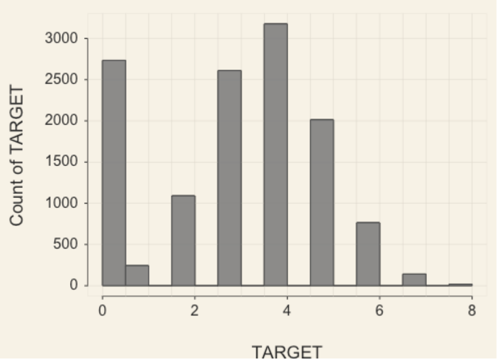


Figure 1: Histogram of TARGET

FixedAcidity

The "FixedAcidicity" variable represents the level of fixed acidity of each wine. With no missing values and 2,455 outliers, "FixedAcidity" has a mean value of 7.1 and a standard deviation of 6.3. As seen in Figure 2, a large percentage of observations fall in the middle. 6,714 observations (52%) are between 5 and 10 and 1,621 observations (12.7%) are negative – both occurrences warrant further investigation with regard to variable transformation in preparation for use within a model.

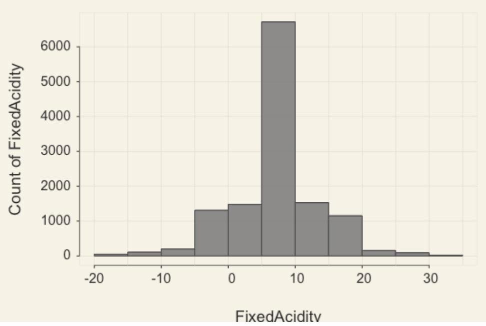


Figure 2: Histogram of FixedAcidity

VolatileAcidity

"VolatileAcidity" represents the level of volatile acid in each wine. Illustrated in Figure 3, most observations fall between 0 and 0.5. With no missing values, a mean of .32 and a standard deviation of .78, "VolatileAcidity" has 2,599 outliers and 2,827 negative values (22.1%) requiring deeper exploration – similar to that of "FixedAcidity".

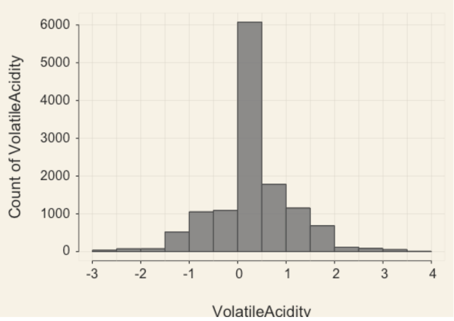


Figure 3: Histogram of VolatileAcidity

CitricAcid

"CitricAcid" represents the level of citric acid in each wine. The variable has no missing values with a mean of .31 and a standard deviation of .86. As Figure 4 illustrates, most observations fall between 0 and 0.5 with 2,688 outliers and 2,966 negative values (23.2% of total observations). Similar to the previously discussed variables, "CitricAcid" will require further examination before consideration of use within a model due to skewed distribution and negative values.

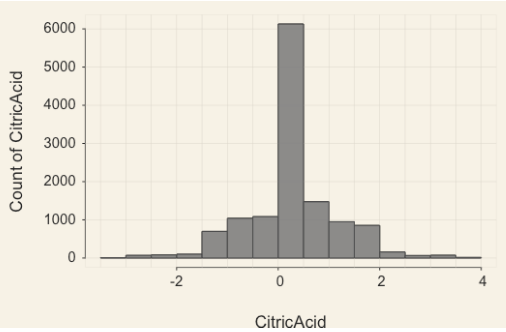


Figure 4: Histogram of CitricAcid

ResidualSugar

The “ResidualSugar” variable represents the level of residual sugar in each wine. With a mean of 5.4 and a standard deviation of 33.7, the variable is highly dispersed with 3,298 outliers. Figure 5 shows the distribution of “ResidualSugar” indicating the majority of values (50%) fall between 0 and 20. Given the degree of dispersion, existence of 3,136 negative values, and 616 missing values, data cleanup will be considered for the variable.

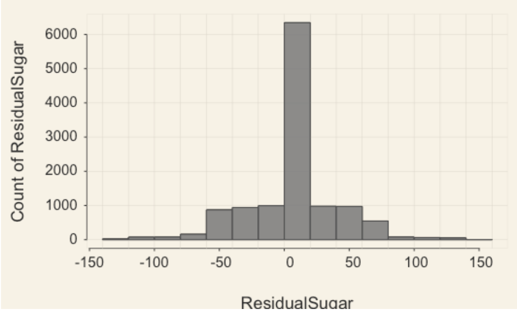


Figure 5: Histogram of ResidualSugar

Chlorides

The “Chlorides” variable represents the level of chloride content in each wine. With a mean of .05 and a standard deviation of .32, the variable has 3,021 outliers and 638 missing values. 49% of observations fall between 0 and 0.2 and are evenly distributed around this range. With 3,197 negative values (25% of the total), transformation is required.

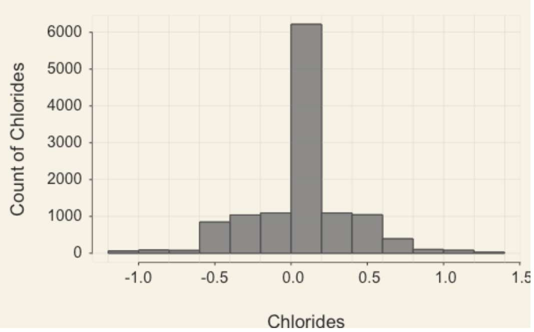


Figure 6: Histogram of Chlorides

FreeSulfurDioxide

“FreeSulfurDioxide” represents the level of sulfur dioxide content in each wine. The variable has 647 missing values with a mean of 30.9 and a standard deviation of 148.7 indicating a high degree of dispersion. The variable has 3,036 negative values (23.7% of all observations) warranting further examination of data quality. Figure 7 shows the distribution of FreeSulfurDioxide illustrating a similar shape as previously discussed variables.

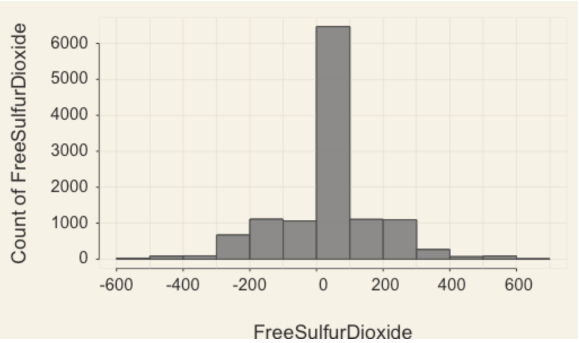


Figure 7: Histogram of FreeSulfurDioxide

TotalSulfurDioxide

"TotalSulfurDioxide" represents the level of total sulfur dioxide in each wine. With 682 missing values and 1,590 outliers, the variable is "over-dispersed" with a mean of 120.7 and a standard deviation of 231.9. Figure 8 shows the variable distribution is similar to aforementioned variables being close to a normal. Due to the missing values, outliers, and 2,504 negative values, the variable requires additional examination before being used in a model.

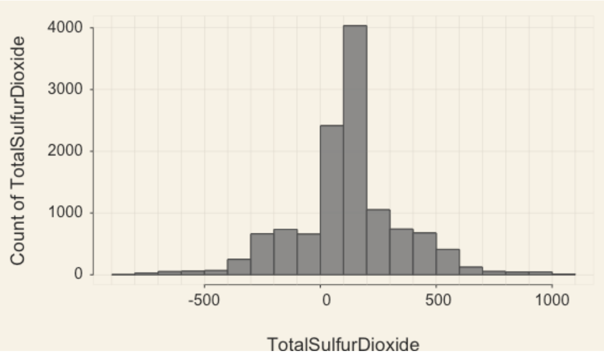


Figure 8: Histogram of TotalSulfurDioxide

Density

"Density" represents the density of each wine and has a mean of .99 and a standard deviation of .0265. The variable has no missing values with 3,823 outliers. "Density" has a similar distribution as previously described variables as pictured in Figure 9 – a large percentage of values in the middle with equal dispersion on either side. 53% of observations fall between .98 and 1. Though the variable has no negative values, additional consideration will be given before use within a model due to the skewed distribution.

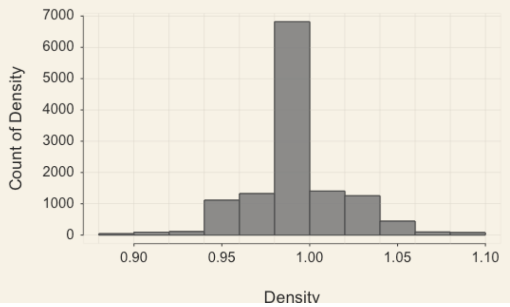


Figure 9: Histogram of Density

pH

"pH" represents the pH levels of each wine with a mean of 3.2 and a standard deviation of .68. With 395 missing values and 1,864 outliers, pH's distribution takes a similar shape as previously described variables. 47% of the observations fall between 3 and 3.5 as seen in Figure 10.

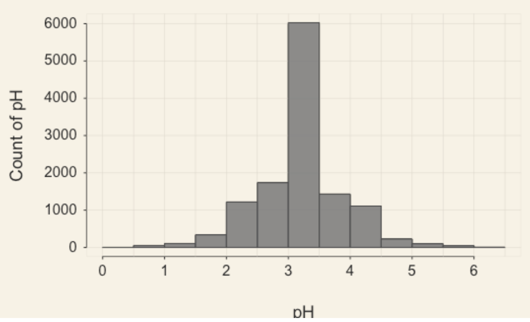


Figure 10: Histogram of pH

Sulphates

"Sulphates" represents the level of sulphate content of each wine with a mean of .53 and a standard deviation of .93. With 1,210 missing values, 2,361 negative values (18.5% of total) and 2,606 outliers, the variable requires additional examination before use within a model. Figure 11 shows the variable's distribution indicating a similar shape as other wine chemical property variables. 52% of observations fall between 0 and .5 with a positive skew (kurtosis of 1.755).

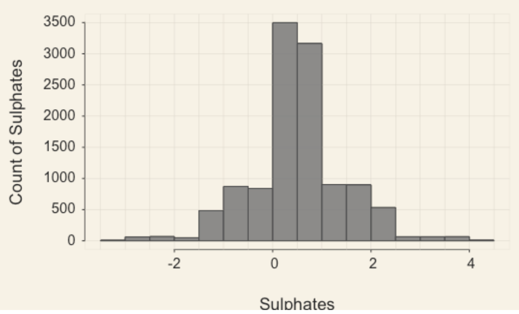


Figure 11: Histogram of Sulphates

Alcohol

The "Alcohol" variable represents the level of alcohol content in each wine. With a mean of 10.5 and a standard deviation of 3.7, the variable has 653 missing values and 928 outliers – requires further examination before use within a model. Figure 12 shows the variable's distribution taking, again, a similar shape as other wine properties with a slightly less "centralization" effect with only 51% of values falling between 8 and 12. The variable has 118 negative values (0.9% of total), which will be considered for potential variable transformation.

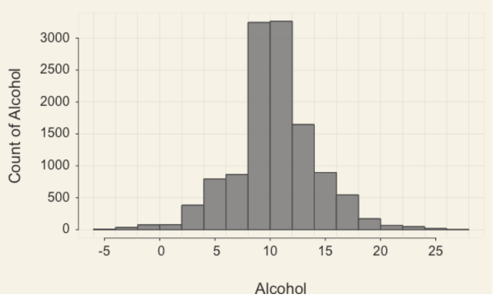


Figure 12: Histogram of Alcohol

LabelAppeal

"LabelAppeal" is a marketing score indicating the appeal of each wine's label design in the eyes of the consumer. On a scale of -2 to 2, higher values indicate that customers like the label design. Negative values indicate that customers don't like the design. With no missing values, the mean LabelAppeal value is just less than zero at -0.01 with a standard deviation of .89. Figure 13 shows the variable's distribution closely resembling a normal distribution. Negative values of the variable are specific to the scale of measurement, so this variable will not require transformation before consideration of use within a model.

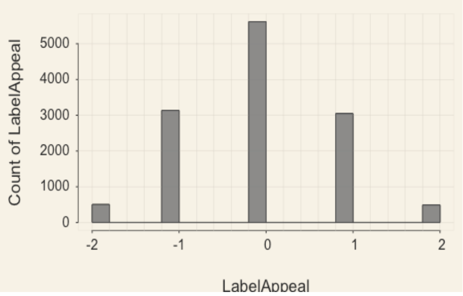


Figure 13: Histogram of LabelAppeal

AcidIndex

The “AcidIndex” variable is a value representing a proprietary method of testing each wine’s total acidity by using a weighted average. Figure 14 shows the shape of the variable’s distribution showing a positive skew with a kurtosis of 5.2. With no missing values and 1,151 outliers, the variable has a mean of 7.77 and a standard deviation of 1.32. Transformation will be considered before use in a model to adjust for the variable’s skewness.

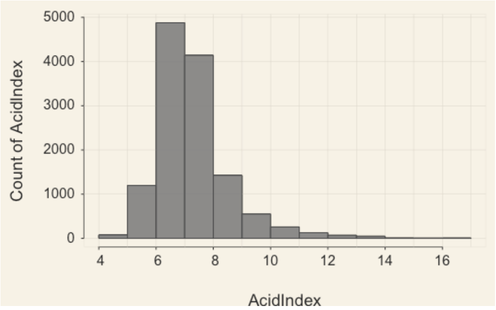


Figure 14: Histogram of AcidIndex

STARS

“STARS” variable represents values from wine a rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor. With a mean of 7.77 and a standard deviation of 1.32, the variable has no missing values with 1,151 outliers.

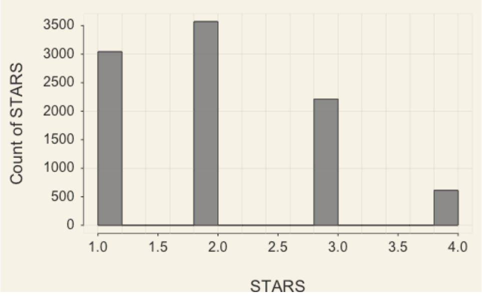


Figure 15: Histogram of STARS

Data Preparation

Many variables have “normal-like” distributions with negative values and a large number of outliers. As a remedy, transformed variables are created that take the square root of the original variables’ absolute value. Some variables do not suffer from this phenomenon but do require adjustment due to outliers. What follows is an examination of newly created variables that will be considered for use in models.

s_FixedAcidity

“FixedAcidity” has 1,621 negative values (12.7% of total observations) representing a “normal-like” distribution. A new variable is created, s_FixedAcidity, which takes the square root of the original variable’s absolute value preserving the original variable’s shape as seen in Figure 16. The transformation reduces the number of outliers from 2,455 to 1,374.

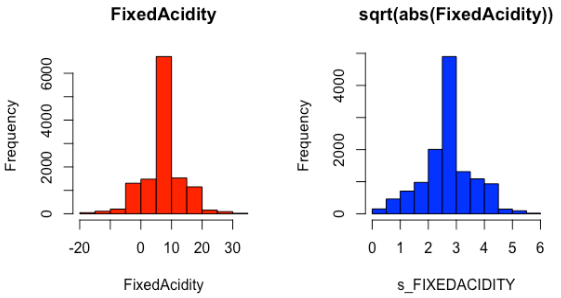


Figure 16: Histogram of FixedAcidity and s_FixedAcidity

VolatileAcidity

“VolatileAcidity” has 2,827 negative values (22.1% of total observations) also taking on a “normal-like” distribution. A newly created variable, s_VolatileAcidity, takes the square root of the absolute value of VolatileAcidity – both pictured in Figure 17. The newly transformed variable includes no negative values while maintaining a similar shape as the original variable while reducing the number of outliers from 2,599 to 119.

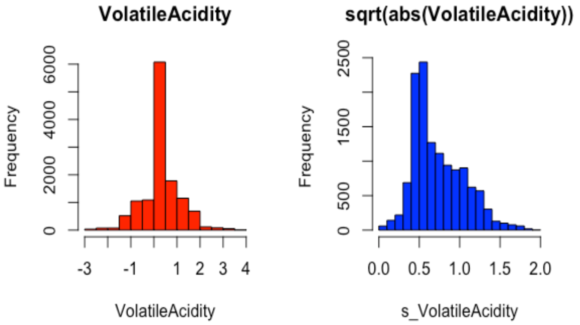


Figure 17: Histogram of VolatileAcidity and s_VolatileAcidity

CitricAcid

“CitricAcid” has 2,966 negative values (23.2% of total observations). A new variable, s_CitricAcid, takes the square root of the absolute value of the original variable resulting in a similar distribution as “CitricAcid” as illustrated in Figure 18 and reducing the number of outliers from 2,688 to 148.

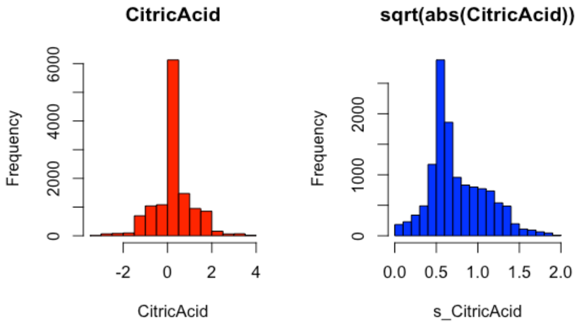


Figure 18: Histogram of CitricAcid and s_CitricAcid

ResidualSugar

“ResidualSugar” has 3,136 negative values (24.5% of total observations). Figure 19 shows the original variable alongside a new variable, s_ResidualSugar, which takes the square root of the absolute value of the original variable reducing the number of outliers from 3,298 to 0.

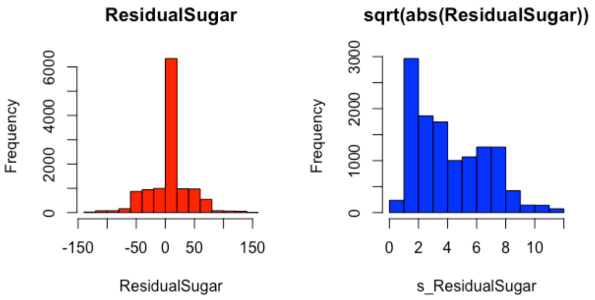


Figure 19: Histogram of ResidualSugar and s_ResidualSugar

Chlorides

“Chlorides” has 3,197 negative values (25% of total observations) and 3,021 outliers. A transformed variable, called s_Chlorides, takes the square root of the original variable’s absolute value resulting in 0 outliers and a similar distribution as evidenced in Figure 20.

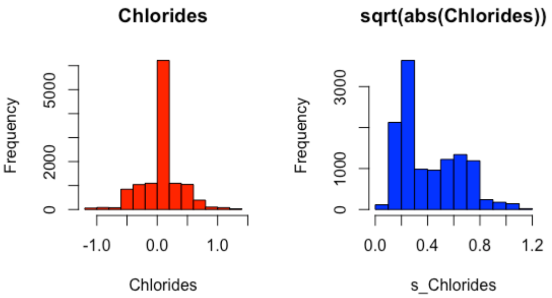


Figure 20: Histogram of Chlorides and s_Chlorides

FreeSulfurDioxide

“FreeSulfurDioxide” has 3,036 negative values (23.7% of total observations) and 3,712 outliers. s_FreeSulfurDioxide takes the square root of the original variable’s absolute value – both pictured in Figure 21 – reducing the number of outliers to 3 while maintaining much of the original variable’s distribution.

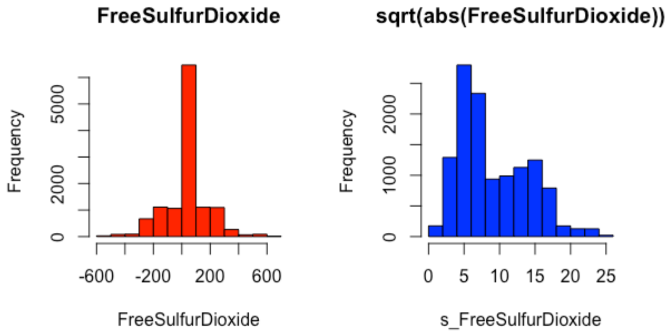


Figure 21: Histogram of FreeSulfurDioxide and s_FreeSulfurDioxide

TotalSulfurDioxide

"TotalSulfureDioxide" has 2,504 negative values (19.6% of total observations) and 1,590 outliers. s_TotalSulfureDioxide takes the square root of the original variable's absolute value – both pictured in Figure 22. The transformation removes all negative values and reduces the number of outliers from 1,590 to 256.

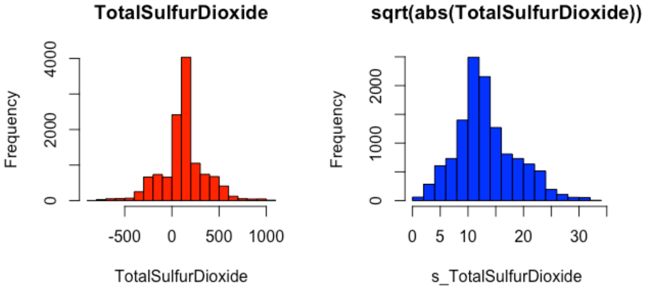


Figure 22: Histogram of TotalSulfurDioxide and s_TotalSulfurDioxide

Density

While "Density" has no negative values, a new variable is created to slightly reduce the number of outliers. s_Density takes the square root of "Density" reducing the number of outliers from 3,823 to 3,820. Figure 23 shows the distribution of both variables showing the transformation preserves the original variable's shape.

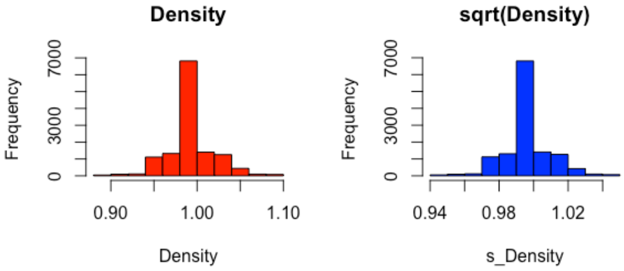


Figure 23: Histogram of Density and s_Density

Sulphates

"Sulphates" has 2,361 negative values (18.5% of total observations) with 2,606 outliers and a "normal-like" distribution. A new variable is created, s_Sulphates, which takes the square root of the original variable's absolute value reducing the number of outliers from 2,606 to 250 while preserving the original variable's distribution as pictured in Figure 24.

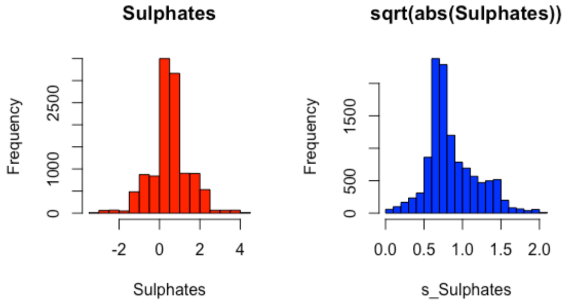


Figure 24: Histogram of Sulphates and s_Sulphates

Alcohol

"Alcohol" has 118 negative values (0.9% of total observations) with 928 outliers and a "normal-like" distribution. a_Alcohol takes the absolute value of "Alcohol", which reduces the number of outliers from 928 to 920 and removes all negative values.

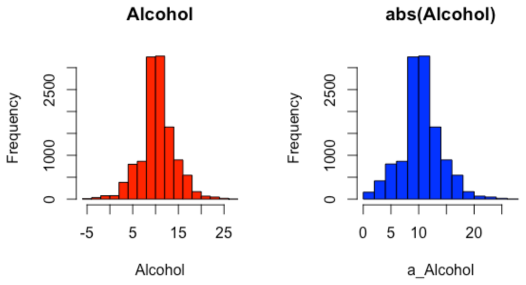


Figure 25: Histogram of Alcohol and a_Alcohol

AcidIndex

While "AcidIndex" has no negative values, the variable has strong positive skew (kurtosis = 5.19) with 1,151 outliers. s_AcidIndex takes the square root of "AcidIndex" reducing the kurtosis to 3.14 though maintaining the same number of outliers. Figure 26 shows both variables' distributions indicating s_AcidIndex has a slightly smaller skew.

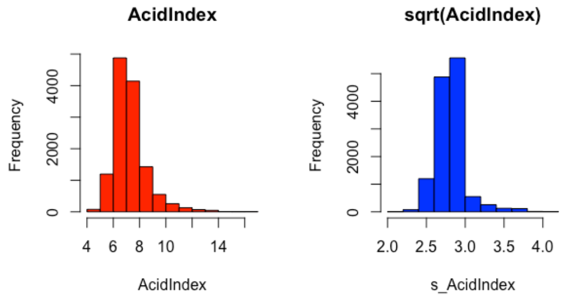


Figure 26: Histogram of AcidIndex and s_AcidIndex

Missing Values

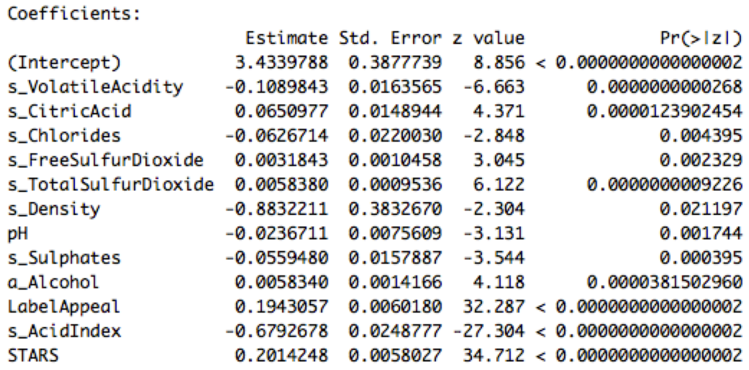
Finally, eight variables have missing values from the original data set: ResidualSugar (616), Chlorides (638), FreeSulfurDioxide (647), TotalSulfurDioxide (682), pH (395), Sulphates (1,210), Alcohol (653), and STARS (3,359). To address these missing values, the mice R package was used to impute missing values. Mice, short for Multivariate imputation by chained equations, "operates under the assumption that given the variables used in the imputation procedure, the missing data are Missing At Random (MAR), which means that the probability that a value is missing depends only on observed values and not on unobserved values" (Schafer & Graham, 2002). The observed values used for imputation are based off of the transformed variables as to impute the cleanest and most relevant variables.

Model Builds

Five models are fitted using the cleaned/transformed data set: Poisson (Model 1), Negative Binomial (Model 2), Zero Inflated Poisson (Model 3), Zero Inflated Negative Binomial (Model 4), and OLS (Model 5). What follows is an examination of each modeling approach and the resulting model output.

Poisson

Model 1 leverages the glm function in R to fit the clean data set using TARGET as the response variable, which is assumed to have a Poisson distribution. Variable selection was based on statistical significance (p value < .05) and evaluation of Akaike Information Criterion (AIC). This method led to the removal of s_FixedAcidity and s_ResidualSugar. Figure 27 provides a summary of Model 1 and its coefficients – 12 in total. The coefficients of LabelAppeal and STARS are both positive, which makes intuitive sense – higher label appeal and star ratings lead to higher sales. Each unit increase in label appeal leads to an increase in wine case sales by a factor of exp(0.1943057) = 1.21 or 21%, holding all other variables constant. Similarly, each unit increase in star rating leads to an increase in wine case sales by a factor of exp(0.2014248) = 1.22 or 22%, holding all other variables constant.

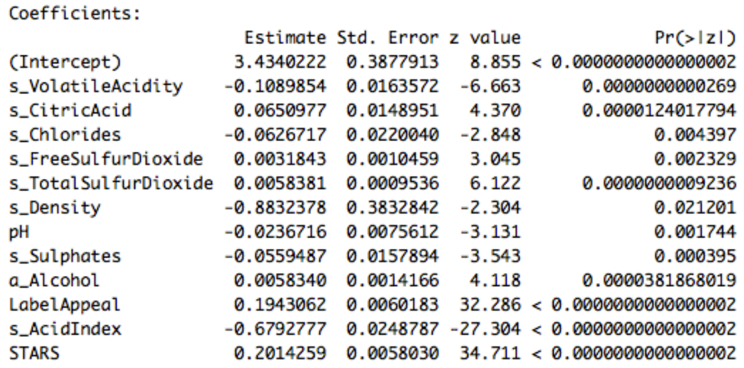


```
Coefficients:
                       Estimate Std. Error z value             Pr(>|z|)
(Intercept)           3.4339788  0.3877739   8.856 < 0.0000000000000002
s_VolatileAcidity    -0.1089843  0.0163565  -6.663       0.0000000000268
s_CitricAcid          0.0650977  0.0148944   4.371       0.0000123902454
s_Chlorides          -0.0626714  0.0220030  -2.848              0.004395
s_FreeSulfurDioxide   0.0031843  0.0010458   3.045              0.002329
s_TotalSulfurDioxide  0.0058380  0.0009536   6.122       0.0000000009226
s_Density            -0.8832211  0.3832670  -2.304              0.021197
pH                   -0.0236711  0.0075609  -3.131              0.001744
s_Sulphates          -0.0559480  0.0157887  -3.544              0.000395
a_Alcohol             0.0058340  0.0014166   4.118       0.0000381502960
LabelAppeal           0.1943057  0.0060180  32.287 < 0.0000000000000002
s_AcidIndex          -0.6792678  0.0248777 -27.304 < 0.0000000000000002
STARS                 0.2014248  0.0058027  34.712 < 0.0000000000000002
```

Figure 27: Poisson Model Summary

Negative Binomial

Model 2 fits a model with the cleaned data set using TARGET as the response variable, which is assumed to have a negative binomial distribution (TARGET's variance is assumed to be larger than its mean). Variables were selected based on statistical significance (p value < .05) and Aikaike Information Criterion (AIC). Similar to Poisson, s_FixedAcidity and s_ResidualSugar were removed from the model using this method of selection. Figure 28 provides the summary of Model 2 and its variable coefficients. Model 2 variable coefficients are virtually the same as the



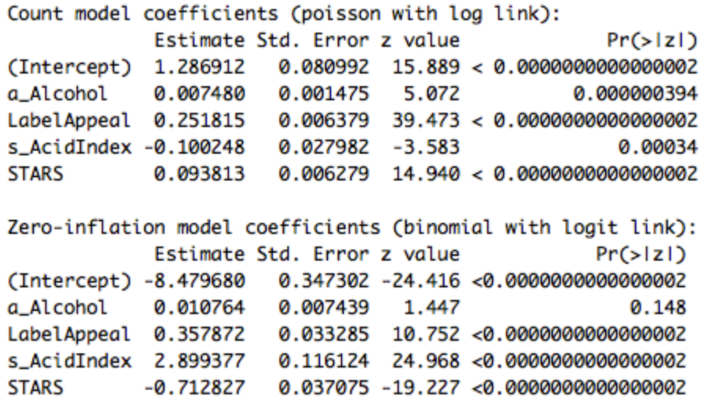
```
Coefficients:
                       Estimate Std. Error z value             Pr(>|z|)
(Intercept)           3.4340222  0.3877913   8.855 < 0.0000000000000002
s_VolatileAcidity    -0.1089854  0.0163572  -6.663       0.0000000000269
s_CitricAcid          0.0650977  0.0148951   4.370       0.0000124017794
s_Chlorides          -0.0626717  0.0220040  -2.848              0.004397
s_FreeSulfurDioxide   0.0031843  0.0010459   3.045              0.002329
s_TotalSulfurDioxide  0.0058381  0.0009536   6.122       0.0000000009236
s_Density            -0.8832378  0.3832842  -2.304              0.021201
pH                   -0.0236716  0.0075612  -3.131              0.001744
s_Sulphates          -0.0559487  0.0157894  -3.543              0.000395
a_Alcohol             0.0058340  0.0014166   4.118       0.0000381868019
LabelAppeal           0.1943062  0.0060183  32.286 < 0.0000000000000002
s_AcidIndex          -0.6792777  0.0248787 -27.304 < 0.0000000000000002
STARS                 0.2014259  0.0058030  34.711 < 0.0000000000000002
```

Figure 28: Negative Binomial Model Summary

variable coefficients in Model 1. While not surprising given the nature of Poisson and Negative Binomial distributions, it’s noteworthy that the standard errors and the variable p values are also very similar. Similar to Model 1, LabelAppeal and STARS are both significant resulting in similar types of increases in wine case sales with singular unit increases, holding all other variables constant.

Zero Inflated Poisson

Model 3 fits a model using the Zero Inflated Poisson (ZIP) distribution with TARGET as the response variable. The ZIP distribution accounts for large numbers of 0 values in the response variable. Model 3 variable selections followed a similar methodology with meaningfully different results as shown in Figure 29. 4 variables were included: a_Alcohol, LabelAppeal, s_AcidIndex, and STARS. The coefficients are consistent with an intuitive understanding of wine sales as well as consistent with Model 1 and Model 2 in that the coefficients are positively related to the response variable.



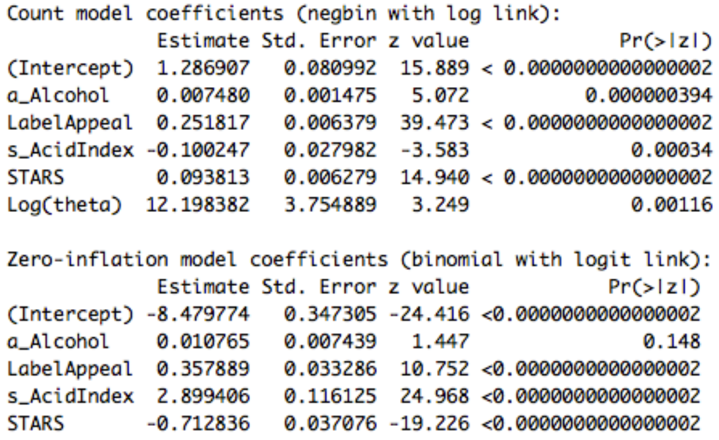
```
Count model coefficients (poisson with log link):
             Estimate Std. Error z value             Pr(>|z|)
(Intercept)  1.286912   0.080992  15.889 < 0.0000000000000002
a_Alcohol    0.007480   0.001475   5.072          0.000000394
LabelAppeal  0.251815   0.006379  39.473 < 0.0000000000000002
s_AcidIndex -0.100248   0.027982  -3.583              0.00034
STARS        0.093813   0.006279  14.940 < 0.0000000000000002

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value            Pr(>|z|)
(Intercept) -8.479680   0.347302 -24.416 <0.0000000000000002
a_Alcohol    0.010764   0.007439   1.447               0.148
LabelAppeal  0.357872   0.033285  10.752 <0.0000000000000002
s_AcidIndex  2.899377   0.116124  24.968 <0.0000000000000002
STARS       -0.712827   0.037075 -19.227 <0.0000000000000002
```

Figure 29: Zero Inflated Poisson Model Summary

Zero Inflated Negative Binomial

Model 4 fits the clean data set with a zero inflated negative binomial distribution using TARGET as the response variable. Similar to Model 3, Model 4 uses a_Alcohol, LabelAppeal, s_AcidIndex, and STARS with similar coefficients and similar standard errors – variables were selected using statistical significance test (p value < .05) and measures of model AIC. Through evaluation of coefficients and standard errors, Model 4 and Model 3 are virtually identical.



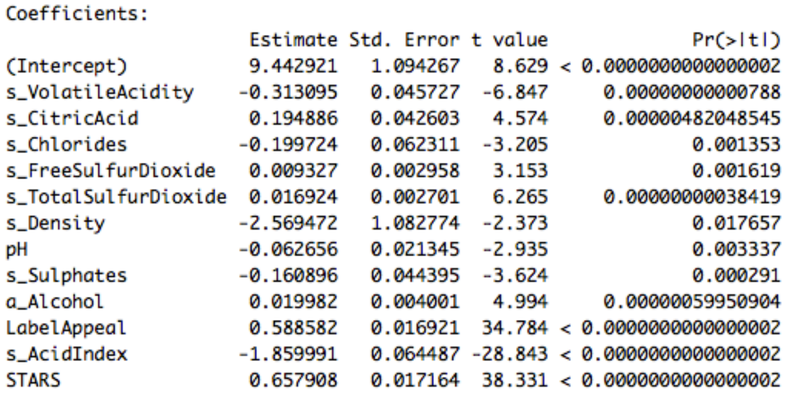
```
Count model coefficients (negbin with log link):
             Estimate Std. Error z value             Pr(>|z|)
(Intercept)  1.286907   0.080992  15.889 < 0.0000000000000002
a_Alcohol    0.007480   0.001475   5.072          0.000000394
LabelAppeal  0.251817   0.006379  39.473 < 0.0000000000000002
s_AcidIndex -0.100247   0.027982  -3.583              0.00034
STARS        0.093813   0.006279  14.940 < 0.0000000000000002
Log(theta)  12.198382   3.754889   3.249              0.00116

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value            Pr(>|z|)
(Intercept) -8.479774   0.347305 -24.416 <0.0000000000000002
a_Alcohol    0.010765   0.007439   1.447               0.148
LabelAppeal  0.357889   0.033286  10.752 <0.0000000000000002
s_AcidIndex  2.899406   0.116125  24.968 <0.0000000000000002
STARS       -0.712836   0.037076 -19.226 <0.0000000000000002
```

Figure 30: Zero Inflated Negative Binomial Model Summary

Ordinary Least Squares

Model 5 fits the clean data set with an OLS regression using TARGET as the response variable. Variables are selected using evaluation of statistical significance (p value < .05) and optimal adjusted r-squared values. LabelAppeal and STARS has a much larger impact in the OLS model than Poisson (LabelAppeal = .59 in Model 5 vs. .19 in Model 1, STARS = .66 in Model 5 vs. .2 in Model 1) though this does not directly translate into wine sales. In Model 1, we found that unit increases in LabelAppeal and STARS led to 21% and 22% respective increase in wine case sales. The coefficients in Model 5 indicate that unit increases in LabelAppeal and STARS leads to .59 and .66 respective unit increases in expected wine case sales.



```
Coefficients:
                      Estimate Std. Error t value              Pr(>|t|)
(Intercept)           9.442921   1.094267   8.629 < 0.0000000000000002
s_VolatileAcidity    -0.313095   0.045727  -6.847     0.00000000000788
s_CitricAcid          0.194886   0.042603   4.574     0.00000482048545
s_Chlorides          -0.199724   0.062311  -3.205             0.001353
s_FreeSulfurDioxide   0.009327   0.002958   3.153             0.001619
s_TotalSulfurDioxide  0.016924   0.002701   6.265     0.00000000038419
s_Density            -2.569472   1.082774  -2.373             0.017657
pH                   -0.062656   0.021345  -2.935             0.003337
s_Sulphates          -0.160896   0.044395  -3.624             0.000291
a_Alcohol             0.019982   0.004001   4.994     0.00000059950904
LabelAppeal           0.588582   0.016921  34.784 < 0.0000000000000002
s_AcidIndex          -1.859991   0.064487 -28.843 < 0.0000000000000002
STARS                 0.657908   0.017164  38.331 < 0.0000000000000002
```

Figure 30: OLS Model Summary

Model Selection

To select the "best" model, measure of Aikaike Information Criterion (AIC) and mean square error (MSE) will be examined with lower AIC and MSE measures indicating better model fit. Table 1 provides an overview AIC and MSE measures for each model. Model 3 (ZIP) has the smallest AIC at 45125.77 while Model 2 (Negative Binomial) has the largest AIC at 50389.41. Model 1 has the smallest MSE at 0.37736 while Model 3 is substantially higher at 2.638224. Due to the danger of over-fitting based on the training data set, Model 3 will be used to score out of sample data relying on the model's AIC measure to indicate out of sample performance.

Model	***AIC***	***MSE***
Model 1: Poisson	50387.29	0.37736
Model 2: Negative Binomial	50389.41	0.3773614
Model 3: Zero Inflated Poisson	45125.77	2.638224
Model 4: Zero Inflated Negative Binomial	45127.91	2.638224
Model 5: OLS	48807.7	2.649912

Table 1: Model Performance Summary

Conclusion

In Project 3: Wine Sales, data related to commercially available wines were examined and five models were fitted to predict the number of sample cases of wine using several statistical distribution methods. Each variable was examined and transformed to fix missing values, fix negative values, and reduce outliers. Automated variable selection was not used. Instead, manual variable selection was used based on statistical significance at the 95% confidence level and measures of AIC/adjusted r-squared. Select variable coefficients were examined based on intuitive understanding of wine sales. Using measures of AIC and MSE, the best model was selected regardless of variables included. Based on this methodology Model 3 was determined to be the "best" model and will be used to score out of sample data.