Matthew Dobbin 12/11/2017

Assignment #3
Predict 411 Fall 2017 Section 55

Introduction
The purpose of this assignment is to analyse a data set that contains information on approximately 12,000 commercially available wines and then to produce a model that predicts the number of cases of wine that will be sold given certain properties of the wine. The variables in the data set are mostly related to the chemical properties of the wine being sold.

A variety of different models will be trialled including Poisson distribution and Negative Binomial generalised linear models. The models will be compared against each other and the best model will be selected based on model validation metrics.

1.0 Data Exploration
The wine data set contains 12795 observations and 15 variables (excluding INDEX). The target variable, TARGET, represents the number cases of wine that were purchased by wine distribution companies after sampling a wine.

The variables in the wine data set can be viewed in Table 1. The data set is a mix of discrete and continuous variables. The theoretical effect for label appearance is that higher numbers suggest better sales. Many consumers purchase based on the visual appeal of the wine label design. Higher numbers in the STARS variable would also suggest higher wine sales.

Table 1 - Variables in the wine data set.

Variable	Definition
INDEX	Identification Variable (do not use)
TARGET	Number of Cases Purchased
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 customers don't like the design.
ResidualSugar	Residual Sugar of wine
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor
Sulphates	Sulfate content of wine
TotalSulfurDioxide	Total Sulfur Dioxide of Wine
VolatileAcidity	Volatile Acid content of wine
pH	pH of wine

1.1 Summary Statistics

Table 2 shows the summary statistics for the variables in the data set. The "miss" column represents the number of missing records for the variable. The variables Alcohol, Chlorides, FreeSulfurDioxide, ResidualSugar, STARS, Sulphates, TotalSulfurDioxide and pH have missing values. The minimum column identified 10 variables that had negative values. The missing values and negative values will be further investigated and treated in Section 2.

Table 2 - Summary statistics for the numerical variables.

Variable	n	miss	mean	sd	skew	krts	min	max	IQR
TARGET	12795	0	3.03	1.93	-0.33	-0.88	0.00	8.00	2.00
AcidIndex	12795	0	7.77	1.32	1.65	5.19	4.00	17.00	1.00
Alcohol	12142	653	10.49	3.73	-0.03	1.54	-4.70	26.50	3.40
Chlorides	12157	638	0.05	0.32	0.03	1.79	-1.17	1.35	0.18
CitricAcid	12795	0	0.31	0.86	-0.05	1.84	-3.24	3.86	0.55
Density	12795	0	0.99	0.03	-0.02	1.90	0.89	1.10	0.01
FixedAcidity	12795	0	7.08	6.32	-0.02	1.68	-18.10	34.40	4.30
FreeSulfurDioxide	12148	647	30.85	148.71	0.01	1.84	-555.0	623.00	70.00
LabelAppeal	12795	0	-0.01	0.89	0.01	-0.26	-2.00	2.00	2.00
ResidualSugar	12179	616	5.42	33.75	-0.05	1.89	-127.8	141.15	17.90
STARS	9436	3359	2.04	0.90	0.45	-0.69	1.00	4.00	2.00
Sulphates	11585	1210	0.53	0.93	0.01	1.76	-3.13	4.24	0.58
TotalSulfurDioxide	12113	682	120.71	231.91	-0.01	1.68	-823.	1057.0	181.00
VolatileAcidity	12795	0	0.32	0.78	0.02	1.83	-2.79	3.68	0.51
pH	12400	395	3.21	0.68	0.04	1.65	0.48	6.13	0.51

1.2 TARGET Distribution

Table 2 shows that the mean value and standard deviation for TARGET is 3.0 and 1.9 respectively. The variance of 3.7 is calculated by squaring the standard deviation. Poisson distribution assumes that the mean value equals the variance. An alternative to the Poisson regression model is the negative binomial regression model. It can be used as a substitute for the Poisson model when the variance is larger than the mean (Hoffmann, 2004). Both a Poisson model and a negative binomial regression model will be trialled in Section 3.

Figure 1 shows the distribution of the target variable TARGET. The x-axis represents the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. The graph reveals a high count of zero cases being purchased, approximately 2700. Due to this, a zero inflated model will also be trialled in Section 3 and may result in the best fit.

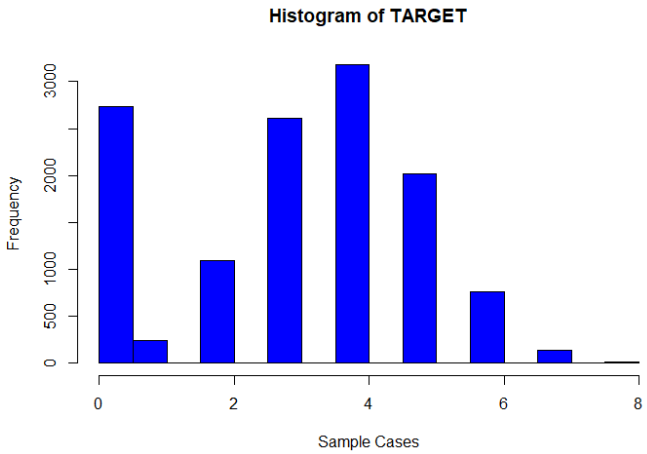


Figure 1 - Histogram of the dependant variable TARGET.

2.0 Data Preparation

The exploratory data analysis conducted in Section 1 identified missing values, negative values and potential outliers in the data set. For the variables with missing values, new variables with the prefix IMP were created to highlight the fact that the missing values were imputed using an average value. Flag variables were also created using the prefix M, which used a '0' if it was the original value or a '1' if it had been imputed. The following sections detail how these values were treated so that an effective analysis could be completed.

As seen in Section 1, there were many variables with negative values. Interestingly, the range of the absolute values of the negative values was very similar to the range of the positive values. The assumption has been made that there was a sign error when entering the data for the negative values. The absolute values of the negative numbers will be used.

2.1 Acid Index

The data dictionary states that the AcidIndex is a proprietary method of testing total acidity of wine by using a weighted average. The boxplot shown in Figure 2 reveals potential upper limit outliers. Research online found that an acid index (may not be the same test) provides a quantitative measure of the relationship between total acid, pH and acid taste. It is calculated by subtracting the pH from the Total acid which is in grams per litre (Jackisch, 1985).

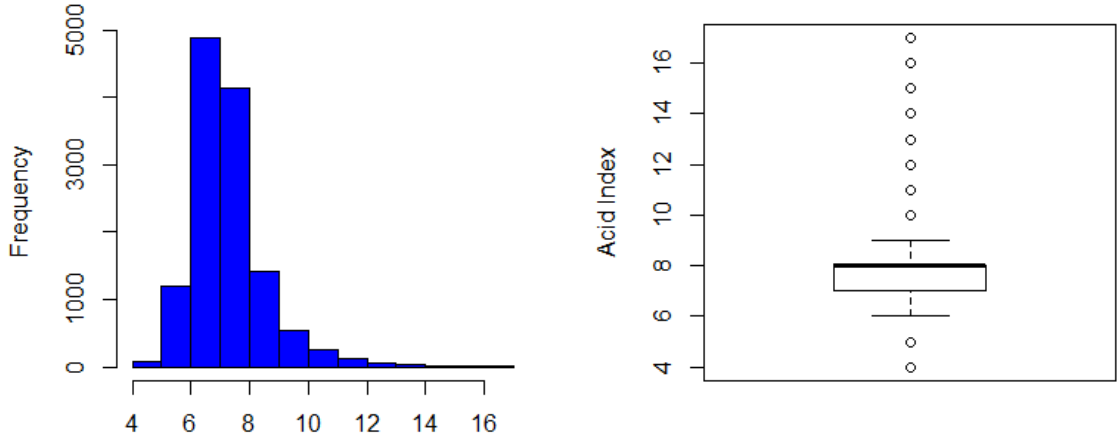


Figure 2 - Histogram and boxplot for the predictor variable AcidIndex.

Dry red wines on average have acid index values of 2.5, while dry white wines averaged 3.8. Sweeter wines tend to have high acid index values. Icewines can have acid index values as high as 12 (Collings, 2007). The maximum limit was set at 14. The values above the cutoff were treated as NA and then replaced with the mean value. The variables IMP_AcidIndex and M_AcidIndex were created.

2.2 Alcohol

The variable Alcohol represents the alcohol content of the wines. Alcohol content for wines can range between 5% to 25%. The data set contains wines with Alcohol content of zero. This is possible as it could represent non-alcoholic wines. The box plot in Figure 3 reveals negative values which is not expected. The absolute values will be used. The variable Alcohol also has 653 missing values. The missing values were replaced with the absolute mean value and the variables IMP_Alcohol and M_Alcohol were created.

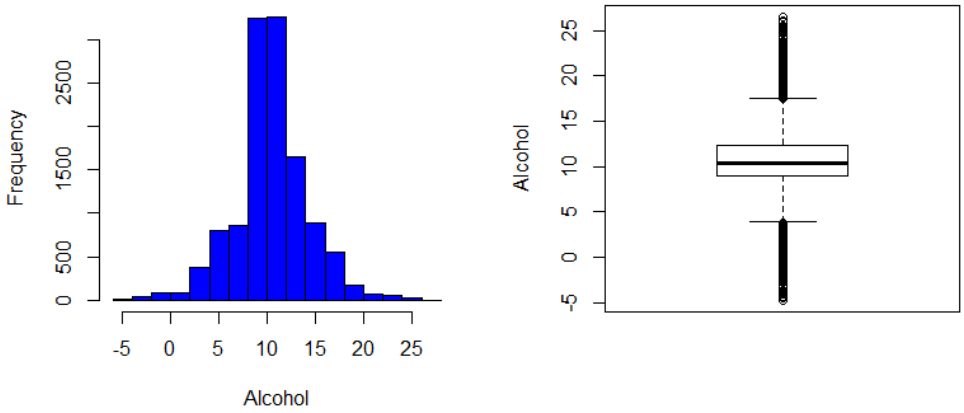


Figure 3 - Histogram and boxplot for the predictor variable Alcohol.

2.3 Chlorides

The variable Chlorides represents the chloride content of the wine. Wine compliance in Australia allows a maximum of 1g/L for Sodium Chloride (Australian Government, 2016). The boxplot in Figure 4 reveals negative numbers which are not expected. Using absolute values, the majority of the data points are below 1g/L. The variable Chlorides also has 638 missing values. The missing values were replaced with the mean of the absolute values and the variables IMP_Chlorides and M_Chlorides were created.

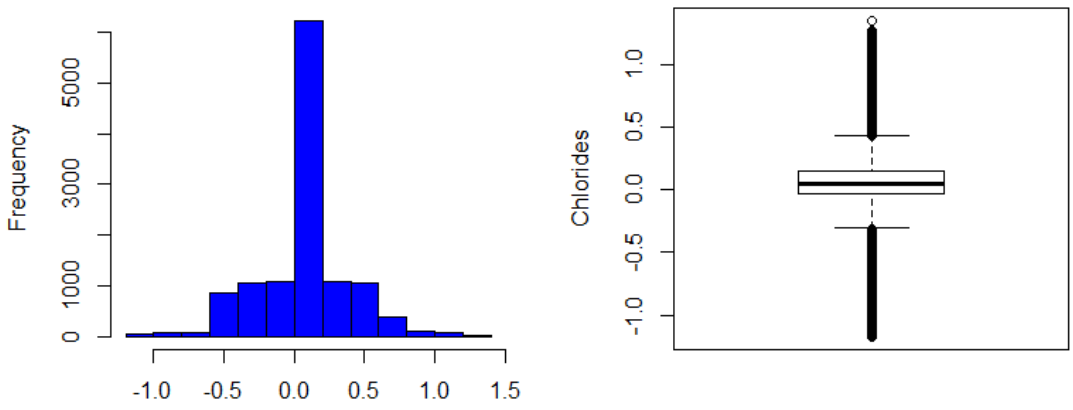


Figure 4 - Histogram and boxplot for the predictor variable Chlorides.

2.4 Citric Acid

The boxplot in Figure 5 reveals negative numbers that are not expected. The absolute values will be used. The variables IMP_CitricAcid and M_CitricAcid were created.

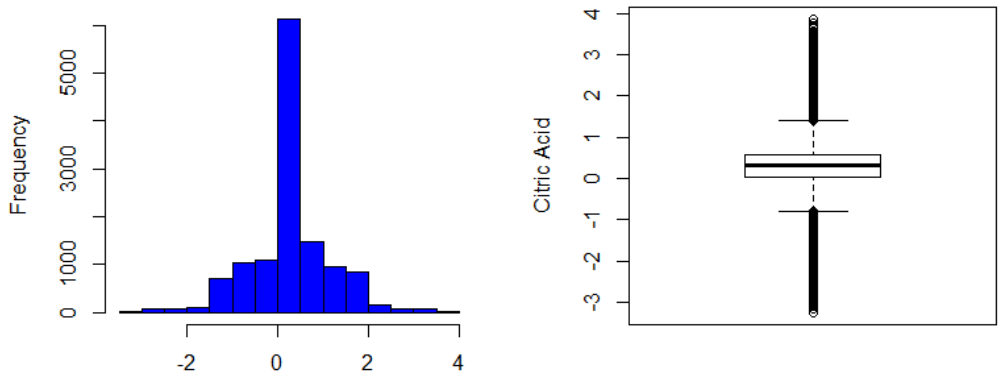


Figure 5 - Histogram and boxplot for the predictor variable CitricAcid.

2.5 Density

The variable Density did not have any missing values or negative numbers. The boxplot in Figure 6 does not reveal any extreme outliers. Potentially the value of one represents 1g/cm^3 which is the density of water. As wine contains ethanol which has a density of 0.789g/cm^3 it would be expected that the density of wine would be slightly less than one.

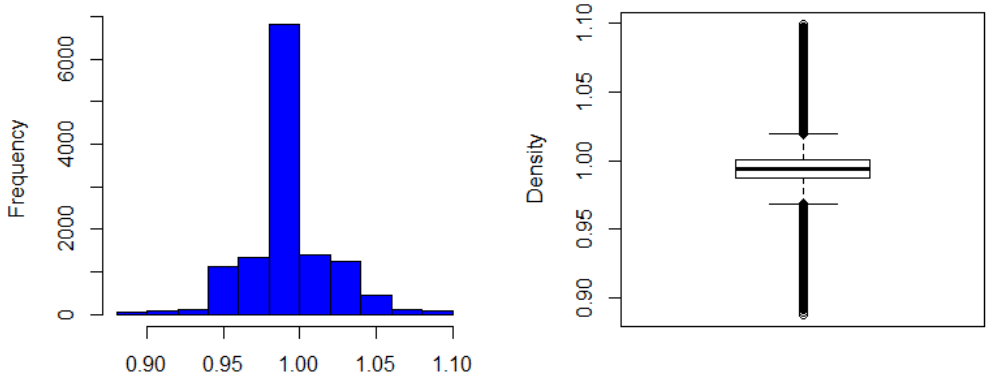


Figure 6 - Histogram and boxplot for the predictor variable Density.

2.6 Fixed Acidity

The variable FixedAcidity did not have any missing values. Total acidity is divided into two groups, fixed acidity and volatile acidity. Fixed acids found in wines are generally tartaric, malic, citric and succinic. The levels found in wine can vary greatly (Nierman, 2004). The boxplot in Figure 7 reveals unexpected negative numbers. The absolute values will be used. The variables IMP_FixedAcidity and M_FixedAcidity were created.

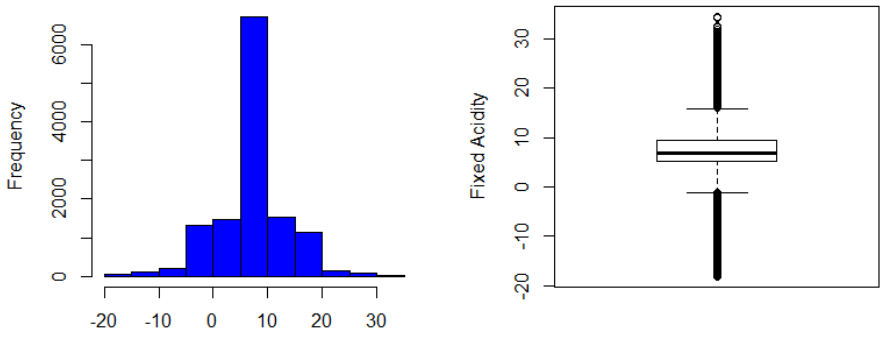


Figure 7 - Histogram and boxplot for the predictor variable FixedAcidity.

2.7 Free Sulfur Dioxide

The variable FreeSulfurDioxide represents the portion of sulphur dioxide that is not bound to sulphur dioxide binding compounds present in wine (Monash , n.d.). Sulphur dioxide is limited to 300mg/L in Australia (Australian Government, 2016). A maximum cutoff point will be set to 500. The boxplot in Figure 8 reveals negative numbers. The absolute values will be used. The variable FreeSulfurDioxide was missing 647 values. The missing values were replaced with the mean of the absolute values and the variables IMP_FreeSulfurDioxide and M_FreeSulfurDioxide were created.

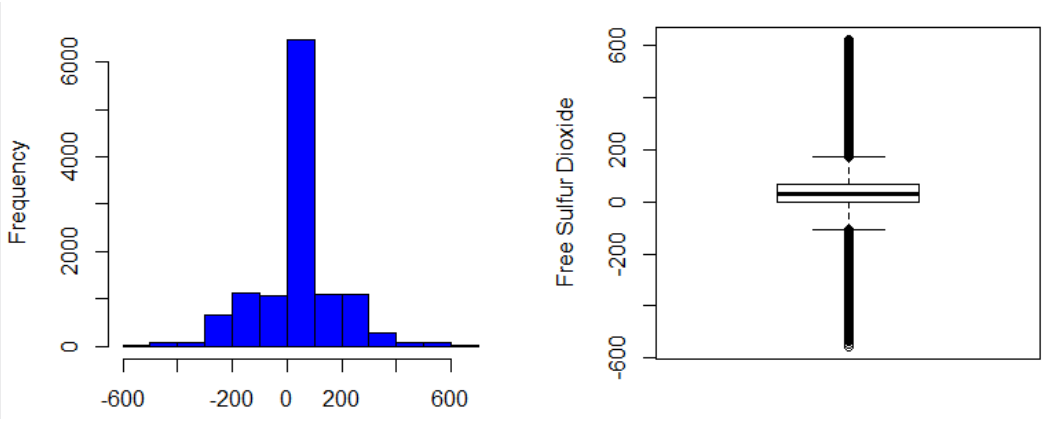


Figure 8 - Histogram and boxplot for the predictor variable FreeSulfurDioxide.

2.8 Label Appeal

The variable LabelAppeal represents the marketing score indicating the appeal of the wine bottle label. The boxplot shown in Figure 9 shows that it could be a reasonable predictor of sample cases. LabelAppeal is a discrete variable but for modelling purposes it has been treated as a continuous variable.

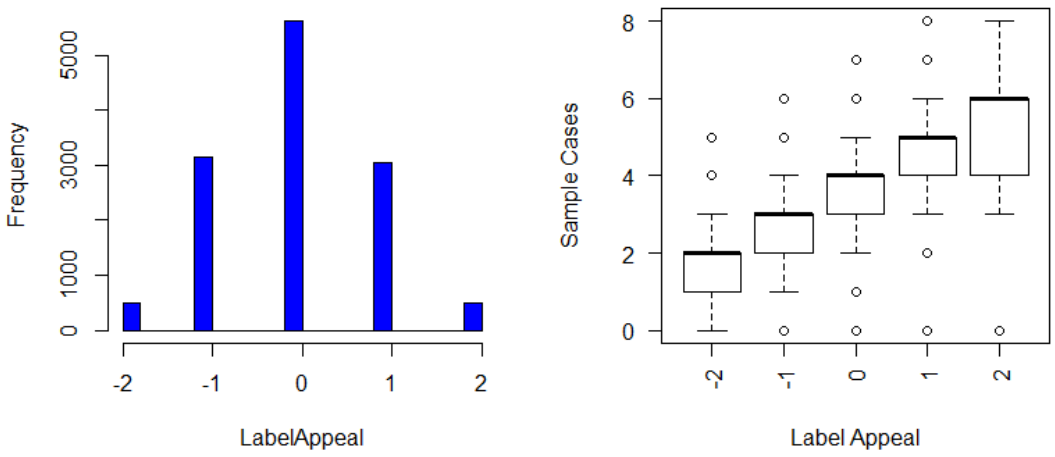


Figure 9 - Histogram and boxplot for the predictor variable LabelAppeal.

2.9 Residual Sugar

The variable ResidualSugar represents the sugar quantity that is left over in the wine after the fermentation process. The boxplot in Figure 10 reveals unexpected negative values. The absolute values will be used. The variable ResidualSugar had 616 missing values. The missing values were replaced with the mean of the absolute values and the variables IMP_ResidualSugar and M_ResidualSugar were created.

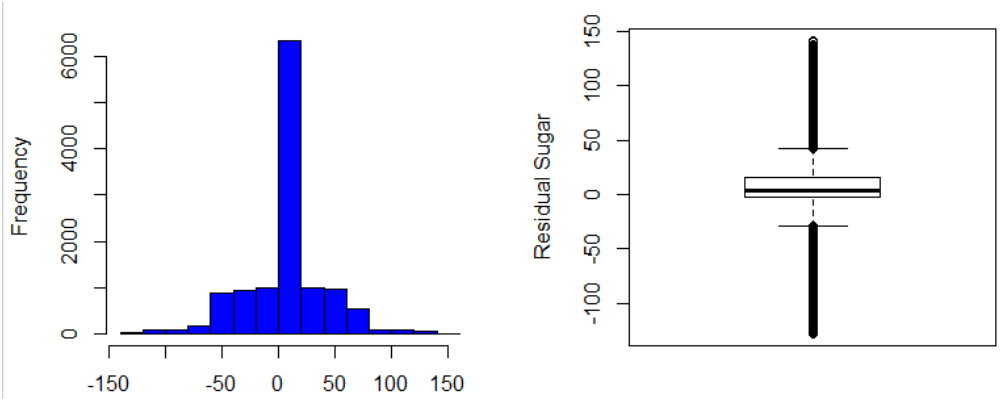


Figure 10 - Histogram and boxplot for the predictor variable ResidualSugar.

2.10 STARS

The variable STARS represents the rating given to the wine by a team of experts. One star represents a poor wine, four stars represents an excellent wine. STARS is a discrete variable but for this case it has been treated as a continuous variable. The variable STARS had 3359 missing values which is approximately 25% of the values were missing. The boxplot shown in Figure 11 shows that STARS could be a reasonable predictor of the number of cases sold. The missing values were replaced with the mean value and the variables IMP_STARS and M_STARS were created.

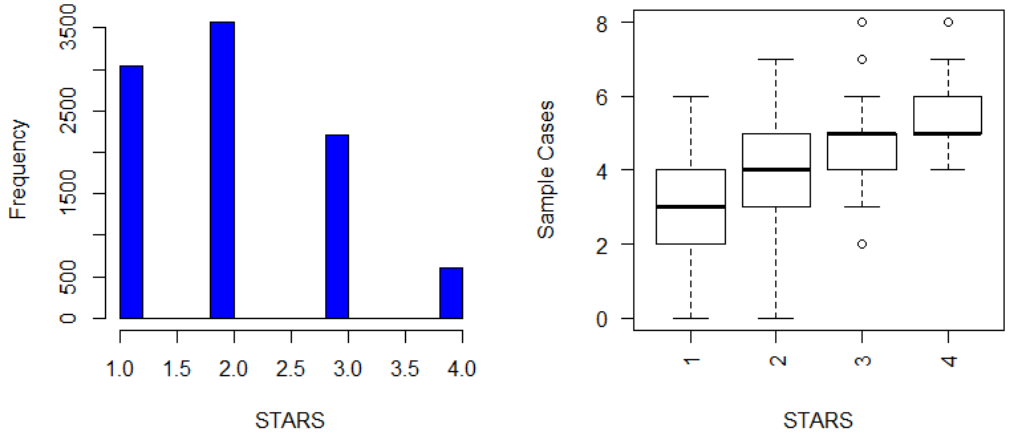


Figure 11 - Histogram and boxplot for the predictor variable STARS.

2.11 Sulphates

The variable Sulphates had 1210 missing values. The boxplot in Figure 12 reveals unexpected negative values. The absolute values will be used. The missing values were replaced with the mean of the absolute values and the variables IMP_ Sulphates and M_ Sulphates were created.

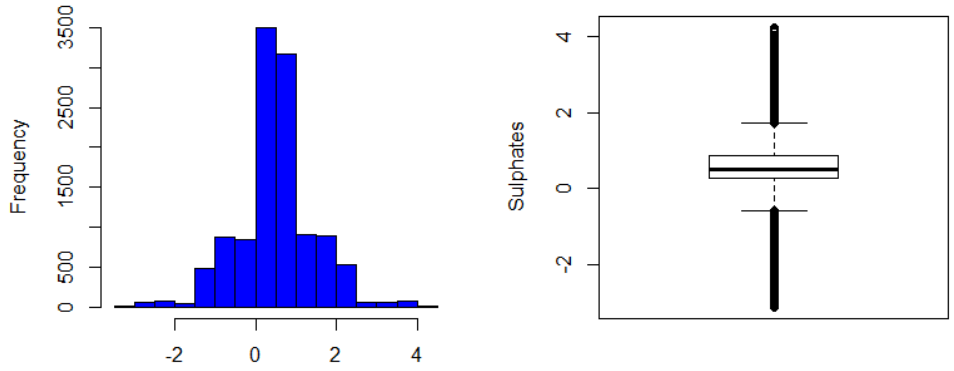


Figure 12 - Histogram and boxplot for the predictor variable Sulphates.

2.12 Total Sulfur Dioxide

The variable TotalSulfurDioxide represents the total sulfur dioxide content in the wine. Sulfur dioxide in wines typically range from 10-350 PPM (Wine Folly, 2014). The maximum cutoff will be set at 500PPM. The boxplot in Figure 13 reveals unexpected negative values. The absolute values will be used. The variable TotalSulfurDioxide had 682 missing values. The missing values were replaced with the mean of the absolute values and the variables IMP_TotalSulfurDioxide and M_TotalSulfurDioxide were created.

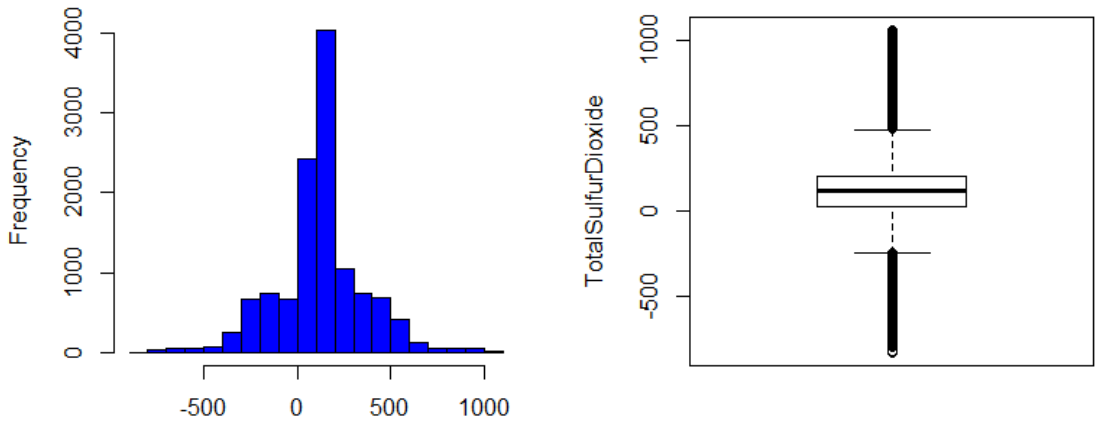


Figure 13 - Histogram and boxplot for the predictor variable TotalSulfurDioxide

2.13 Volatile Acidity

Volatile acidity refers to the steam distillable acids in the wine. These may include acetic and lactic acid. The acetic acid level may range from undetectable up to 3g/L (Waterhouse Lab, 2015). The boxplot in Figure 14 reveals negative numbers which is unexpected. Using absolute values the majority of the data points fall between the expected range. The variables IMP_VolatileAcidity and M_VolatileAcidity were created.

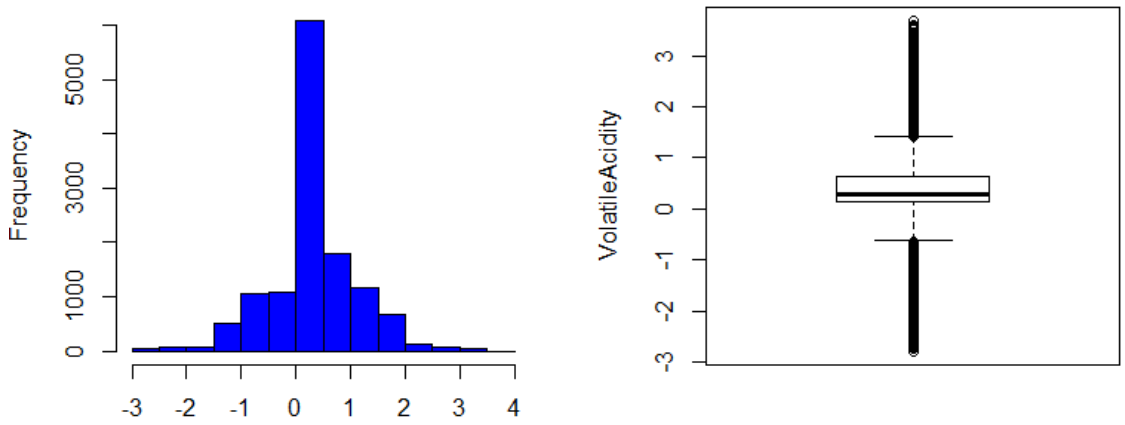


Figure 14 - Histogram and boxplot for the predictor variable VolatileAcidity.

2.14 pH

The pH of wines varies depending on the wine type. For most wines, the pH ranges from 2.5 to about 4.5 (Wine Folly, 2015). The boxplot in Figure 15 show numbers outside the expected range. The lower limit cutoff was set to two and the upper limit cut off was set to five. The values outside the cutoff range were treated as NA and then replaced with the mean value. The variable pH had 395 missing values. The missing values were replaced with the mean. The variables IMP_pH and M_pH were created.

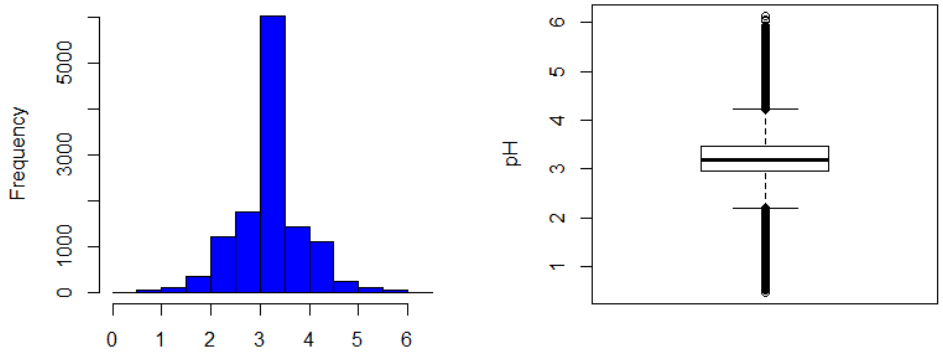


Figure 15 - Histogram and boxplot for the predictor variable pH.

2.15 Treated Variables Summary

Table 3 lists the 26 predictor variables that were used in the predictive models in Section 3.0. The original variables that had missing values, negative values or values outside the expected range were removed and replaced with the imputed variables.

Table 3 - List of treated variables to be used in the regression models.

IMP_AcidIndex	IMP_CitricAcid	IMP_FreeSulfurDioxide	IMP_Sulphates
M_AcidIndex	M_CitricAcid	M_FreeSulfurDioxide	M_Sulphates
IMP_Alcohol	Density	IMP_ResidualSugar	IMP_TotalSulfurDioxide
M_Alcohol	LabelAppeal	M_ResidualSugar	M_TotalSulfurDioxide
IMP_Chlorides	IMP_FixedAcidity	IMP_STARS	IMP_VolatileAcidity
M_Chlorides	M_FixedAcidity	M_STARS	M_VolatileAcidity
IMP_pH	M_pH		

2.16 Correlation Matrix

A correlation matrix was created for the variables and can be seen in Appendix 1. From the matrix it can be seen that the variables LabelAppeal and IMP_STARS are the most positively correlated variables with the dependent variable TARGET. M_STARS and IMP_Acid are the most negatively correlated variables with the dependent variable TARGET.

3.0 Building Models

The overall goal is to build a model that predicts the number of cases of wine that will be sold given certain properties of the wine. There are a variety of different models that could be used. These include OLS, Poisson distribution, negative binomial distribution, zero inflated Poisson distribution and zero inflated negative binomial distribution regression models.

3.1 Model 1 OLS Regression

The first model built used a simple OLS regression method which included all the predictor variables. The summary statistics for the model can be viewed in Appendix 2. The variables that appear to be the most influential in predicting the number of cases of wine sold are IMP_AcidIndex, IMP_Alcohol, LabelAppeal, IMP_STARS, M_STARS and IMP_VolatileAcidity. The VIF values for the model were checked and they were all less than two. Therefore, multicollinearity was not considered to be an issue.

A second OLS regression model (Model 1.1) was built using only the influential variables. The summary statistics for the model can be viewed in Table 4.

Table 4 - OLS Regression Model 1.1 summary statistics.

```
Call:
lm(formula = TARGET ~ IMP_AcidIndex + IMP_Alcohol + IMP_STARS +
    M_STARS + LabelAppeal + IMP_VolatileAcidity, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7503 -0.8480  0.0254  0.8567  6.1697

Coefficients:
                    Estimate Std. Error t value             Pr(>|t|)
(Intercept)         3.602799   0.088606  40.661 < 0.0000000000000002
IMP_AcidIndex      -0.209350   0.009217 -22.713 < 0.0000000000000002
IMP_Alcohol         0.013122   0.003292   3.986           0.0000676435
IMP_STARS           0.781244   0.015708  49.737 < 0.0000000000000002
M_STARS            -2.293548   0.026968 -85.047 < 0.0000000000000002
LabelAppeal         0.464045   0.013695  33.885 < 0.0000000000000002
IMP_VolatileAcidity -0.118661   0.020928  -5.670           0.0000000146

Residual standard error: 1.313 on 12788 degrees of freedom
Multiple R-squared:  0.5355,   Adjusted R-squared:  0.5353
F-statistic:  2457 on 6 and 12788 DF,  p-value: < 0.00000000000000022
```

The variables LabelAppeal and IMP_STARS have positive associations. This is expected as the higher the label appeal or number of stars you would expect higher wine sales. IMP_Alcohol has a positive association. IMP_AcidIndex and IMP_VolatileAcidity have negative associations. This makes logical sense, as the wine becomes more acidic the wine is less enjoyable. M_STARS has a negative association. This would mean that bottles of wine that have not received a star rating from experts have a negative effect on wine sales.

A plot of the residuals by the predicted values shown in Figure 16 reveals that there were a few negative values predicted (minimum value -0.955). This is not a desired attribute.

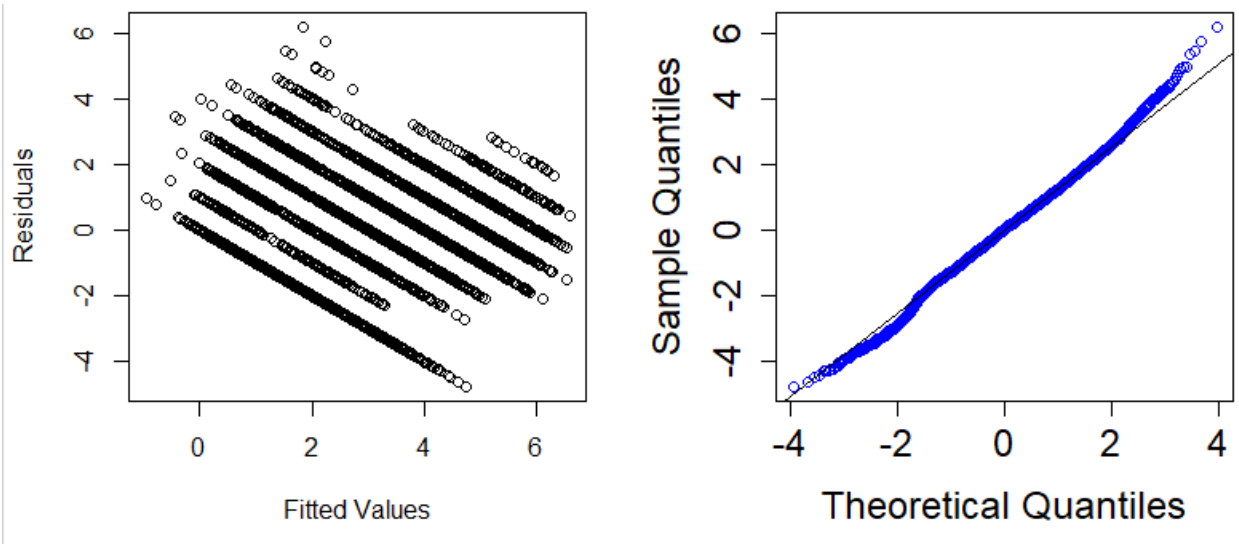


Figure 16 - Plot of Model 1.1 residuals by the predicted values and a normal probability plot of the residuals.

3.2 Model 2 Poisson Distribution Regression

A Poisson distribution regression model was trialled. The coefficients and summary statistics for the model can be seen in Table 5. The associations are inline with the expected as they are the same as Model 1.1.

Table 5 - Model 2 Poisson distribution regression model summary statistics.

```
Call:
glm(formula = TARGET ~ IMP_AcidIndex + IMP_Alcohol + IMP_STARS +
    M_STARS + LabelAppeal + IMP_VolatileAcidity, family = "poisson",
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2087  -0.6464   0.0104   0.4530   3.7703

Coefficients:
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)         1.485947   0.041682  35.650 < 0.0000000000000002
IMP_AcidIndex      -0.082091   0.004563 -17.992 < 0.0000000000000002
IMP_Alcohol         0.003728   0.001443   2.583              0.00979
IMP_STARS           0.188535   0.006086  30.977 < 0.0000000000000002
M_STARS            -1.037390   0.016950 -61.202 < 0.0000000000000002
LabelAppeal         0.158381   0.006125  25.860 < 0.0000000000000002
IMP_VolatileAcidity -0.040747   0.009389  -4.340            0.0000143

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22861  on 12794  degrees of freedom
Residual deviance: 13815  on 12788  degrees of freedom
AIC: 45771

Number of Fisher Scoring iterations: 6
```

Poisson regression uses the Log transformation. Taking the exponential of the coefficients will give the predicted counts. Figure 17 provides a plot of the residuals versus the fitted values and a normal probability plot of the residuals.

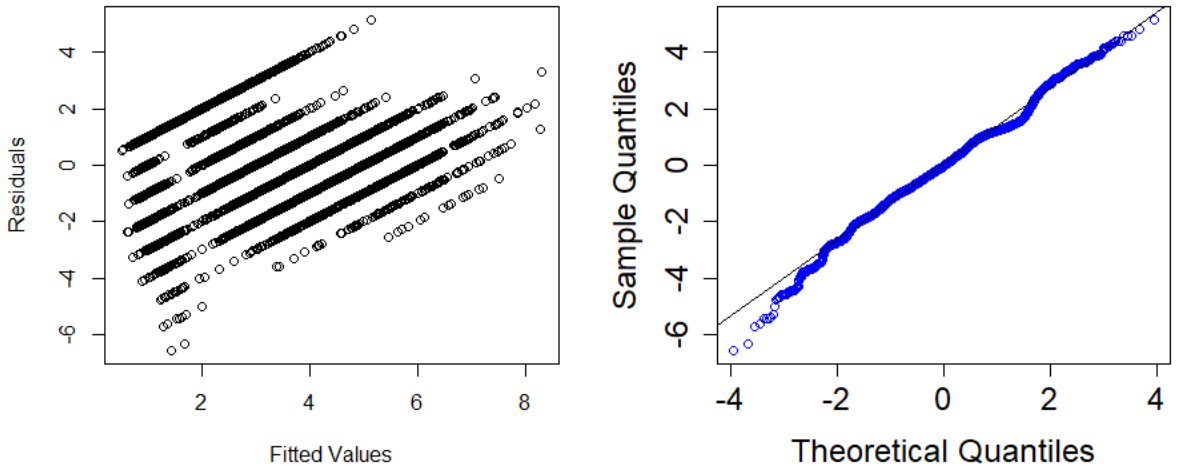


Figure 17 - Plot of Model 2 residuals by the predicted values and a normal probability plot of the residuals.

3.3 Model 3 Negative Binomial Distribution Regression

A negative binomial distribution regression model was trialled as the variance of the target variable exceeded the mean. The coefficients and summary statistics for the model can be seen in Table 6. The coefficients and standard errors are almost identical to the Poisson distribution model.

Table 6 - Model 3 Negative binomial regression model summary statistics.

```
glm.nb(formula = TARGET ~ IMP_AcidIndex + IMP_Alcohol + IMP_STARS +
    M_STARS + LabelAppeal + IMP_VolatileAcidity, data = mydata,
    init.theta = 40481.93148, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2086  -0.6464   0.0104   0.4530   3.7702

Coefficients:
                     Estimate Std. Error z value             Pr(>|z|)
(Intercept)          1.485963   0.041683  35.649 < 0.0000000000000002
IMP_AcidIndex       -0.082094   0.004563 -17.991 < 0.0000000000000002
IMP_Alcohol          0.003728   0.001443   2.583              0.00979
IMP_STARS            0.188537   0.006087  30.976 < 0.0000000000000002
M_STARS             -1.037389   0.016950 -61.201 < 0.0000000000000002
LabelAppeal          0.158380   0.006125  25.859 < 0.0000000000000002
IMP_VolatileAcidity -0.040748   0.009390  -4.340            0.0000143

(Dispersion parameter for Negative Binomial(40481.93) family taken to be 1)

    Null deviance: 22860  on 12794  degrees of freedom
Residual deviance: 13814  on 12788  degrees of freedom
AIC: 45773
Number of Fisher Scoring iterations: 1

              Theta:  40482
          Std. Err.:  34436
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -45757.24
```

The negative binomial transformation is similar to the Poisson regression in that it uses the natural log transformation. Poisson regression is a special case of the negative binomial regression where the mean and the variance are equal. As the mean and variance are similar, 3.0 and 3.7 respectively, it could be expected that the two models are similar. Figure 18 provides a plot of the residuals versus the fitted values and a normal probability plot of the residuals.

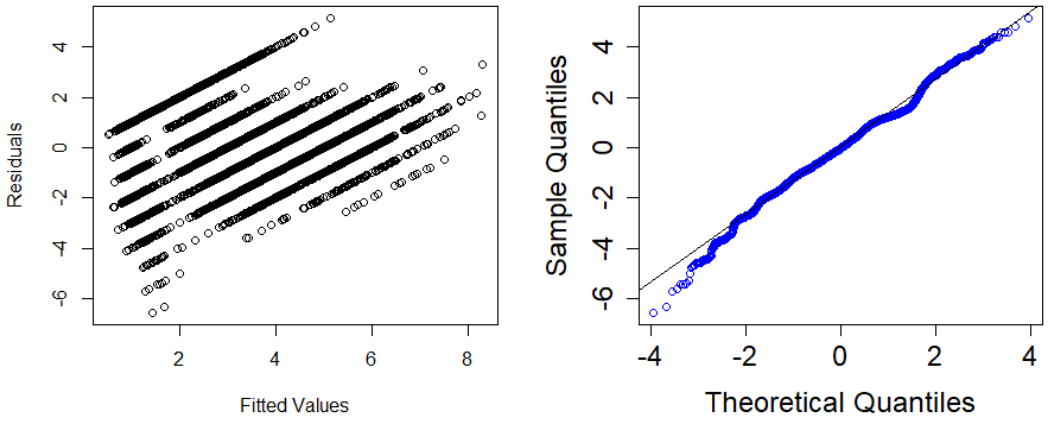


Figure 18 - Plot of Model 3 residuals by the predicted values and a normal probability plot of the residuals.

3.4 Model 4 Zero Inflated Poisson Regression

A zero inflated Poisson (ZIP) distribution regression model was trialled. There are two outputs from a ZIP model. The first is the coefficients predicting the number of cases of wine that will be purchased if a purchase is made (count model). The coefficients are transformed to counts using the exponential function.

The second output is the coefficients predicting the probability that the number of wine cases purchased will be zero (zero inflation model). The zero inflated coefficients predict LOG ODDS and need to be converted to ODDS and then to probability.

As in Model 1, a full model was fitted and the summary statistics were analysed to determine the most influential predictors. The full model statistics can be viewed in Appendix 3. The coefficients and summary statistics for the ZIP model with the resulting selected predictor variables can be seen in Table 7. The associations of the count model are as expected. In the zero inflated model the negative association of the IMP_STARS coefficient and the positive association of M_STARS is unexpected.

Table 7 - Model 4 Zero inflated Poisson regression model summary statistics.

```
Call:
zeroinfl(formula = TARGET ~ IMP_AcidIndex + IMP_Alcohol + LabelAppeal + IMP_
STARS + M_STARS |
    IMP_AcidIndex + LabelAppeal + IMP_STARS + M_STARS, data = mydata)

Pearson residuals:
      Min        1Q    Median        3Q       Max
-2.348386 -0.427800  0.001878  0.386185  6.080780

Count model coefficients (poisson with log link):
               Estimate Std. Error z value             Pr(>|z|)
(Intercept)    1.182816   0.043682  27.078 < 0.0000000000000002
IMP_AcidIndex -0.020987   0.004903  -4.280           0.00001867
IMP_Alcohol    0.006860   0.001468   4.672           0.00000298
LabelAppeal    0.232271   0.006315  36.784 < 0.0000000000000002
IMP_STARS      0.105005   0.006406  16.391 < 0.0000000000000002
M_STARS       -0.187246   0.018555 -10.092 < 0.0000000000000002

Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value             Pr(>|z|)
(Intercept)   -1.21903    0.41458   -2.94              0.00328
IMP_AcidIndex  0.43568    0.02564   16.99 < 0.0000000000000002
LabelAppeal    0.71827    0.04168   17.23 < 0.0000000000000002
IMP_STARS     -3.85689    0.34281  -11.25 < 0.0000000000000002
M_STARS        6.07430    0.35669   17.03 < 0.0000000000000002

Number of iterations in BFGS optimization: 20
Log-likelihood: -2.045e+04 on 11 Df
```

The expected count of wine cases purchased is calculated by multiplying the number of cases of wine purchased (assuming wine was purchased) by the probability that wine is purchased. Figure 19 provides a plot of the residuals versus the fitted values and a normal probability plot of the residuals.

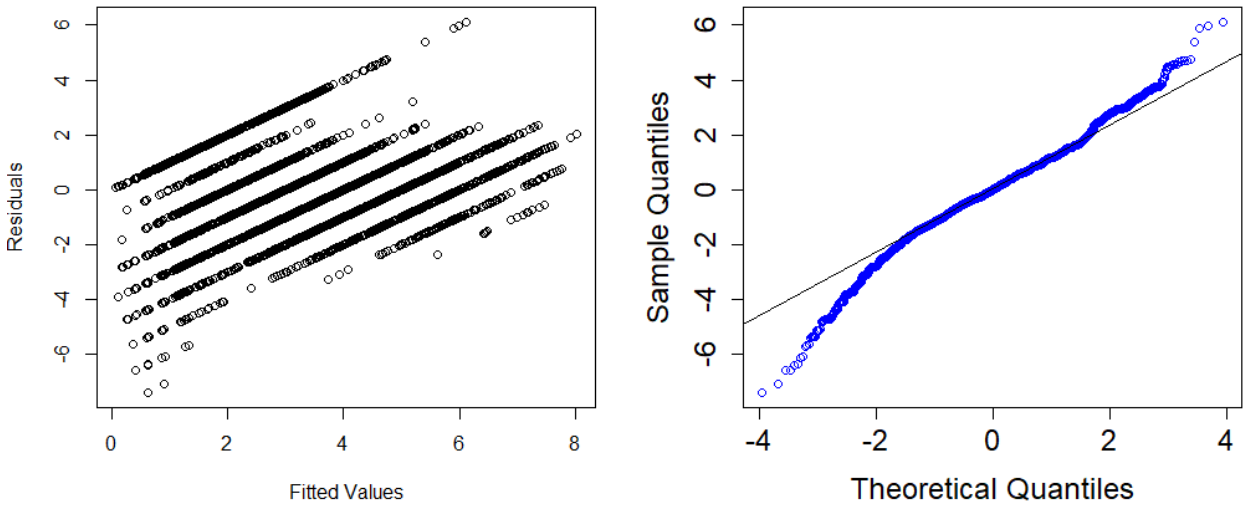


Figure 19 - Plot of Model 4 residuals by the predicted values and a normal probability plot of the residuals.

3.5 Model 5 Zero Inflated Negative Binomial Regression

A zero inflated negative binomial (ZINB) regression model was trialled. It uses the same method as the ZIP regression except that is utilizes the negative binomial distribution instead of Poisson. The coefficients and summary statistics for the model can be seen in Table 8. The summary statistics for the ZINB are almost identical to the ZIP model.

Table 8 - Model 5 Zero inflated negative binomial regression model summary statistics.

```
Call:
zeroinfl(formula = TARGET ~ IMP_AcidIndex + IMP_Alcohol + LabelAppeal + IMP_
STARS + M_STARS | IMP_AcidIndex +
    LabelAppeal + IMP_STARS + M_STARS, data = mydata, dist = "negbin", EM =
TRUE)
Pearson residuals:
     Min       1Q   Median       3Q      Max
-2.34838 -0.42780  0.00188  0.38618  6.08072

Count model coefficients (negbin with log link):
               Estimate Std. Error z value             Pr(>|z|)
(Intercept)    1.182818   0.043682  27.078 < 0.0000000000000002
IMP_AcidIndex -0.020987   0.004903  -4.280            0.00001866
IMP_Alcohol    0.006860   0.001468   4.672            0.00000299
LabelAppeal    0.232271   0.006315  36.783 < 0.0000000000000002
IMP_STARS      0.105006   0.006406  16.391 < 0.0000000000000002
M_STARS       -0.187246   0.018555 -10.092 < 0.0000000000000002
Log(theta)    12.316042   3.890777   3.165              0.00155

Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value             Pr(>|z|)
(Intercept)   -1.21873    0.41465  -2.939              0.00329
IMP_AcidIndex  0.43568    0.02564  16.994 < 0.0000000000000002
LabelAppeal    0.71827    0.04168  17.233 < 0.0000000000000002
IMP_STARS     -3.85717    0.34289 -11.249 < 0.0000000000000002
M_STARS        6.07460    0.35679  17.026 < 0.0000000000000002

Theta = 223248.6946
Number of iterations in BFGS optimization: 1
Log-likelihood: -2.045e+04 on 12 Df
```

4.0 Model Comparison

This section of the report compares the model validation metrics of the six different predictive models. Akaike's information criterion (AIC), Bayesian information criterion (BIC) have previously been used to compare the different models. AIC provides a method to compare models of different sizes. It is based on deviance and has a penalty term for more complicated models (larger number of predictor variables). When looking at AIC and BIC, the model with the smaller value is deemed the better model. These values can be easily calculated for the Models 1-3. To compare all the models, the mean absolute error (MAE) has been used.

Table 9 compares the MAE for the six different predictive models that were built. Surprisingly, the OLS regression model has lower AIC and MAE compared to the Poisson and Negative Binomial regression models. The ZIP and ZINB models resulted in the lowest MAE.

Table 9 - Comparison of the six predictive model key validation statistics.

Model Type	Adjusted R-Squared	AIC	BIC	MAE
Model 1 – Full OLS Regression	0.5367	43273.65	43482.44	1.027
Model 1.1 - OLS Regression	0.5355	43291.69	43351.34	1.029
Model 2 – Poisson	N/A	45770.83	45823.03	1.033
Model 3 – Negative Binomial	N/A	45773.24	45832.89	1.033
Model 4 – Zero Inflated Poisson	N/A	N/A	N/A	0.976
Model 5 – Zero Inflated Negative Binomial	N/A	N/A	N/A	0.976

Histograms shown in Figure 20 display the distribution of the fitted values for the predictive models. These have been compared against the distribution of the target variable. The Poisson and Negative Binomial models did not predict any values of zero for the target variable. The ZIP and ZINB models have a distribution that most closely matches the variable TARGET. They both predict values of zero, however the spike is not as large as the variable TARGET.

Conclusion

Based on the metrics discussed in Section 4, the ZIP model or the ZINB model could both be selected as the best most to predict the number of cases of wine that will be sold given certain properties of the wine. Due to the need to export a single model, the ZIP model was selected. The model coefficients and the equations for calculating the predicted target value is shown in Table 10.

Table 10 - ZIP Model coefficients and equations for calculating predicted TARGET.

```
Count Model (CM) = 1.182816 - 0.020987*IMP_AcidIndex + 0.006860*IMP_Alcohol +
                   0.232271*LabelAppeal + 0.105005*IMP_STARS - 0.187246*M_STARS

Zero Inflated Model (ZIM) =  -1.21903 + 0.43568*IMP_AcidIndex +  0.71827*LabelAppeal
                             -3.85689*IMP_STARS + 6.07430*M_STARS

P_SCORE_ZIP_ALL = exp(CM)
P_SCORE_ZERO = exp(ZIM)/(1+exp(ZIM))
P_SCORE_ZIP = P_SCORE_ZIP_ALL*(1-P_SCORE_ZERO)
P_TARGET = round(P_SCORE_ZIP)
```

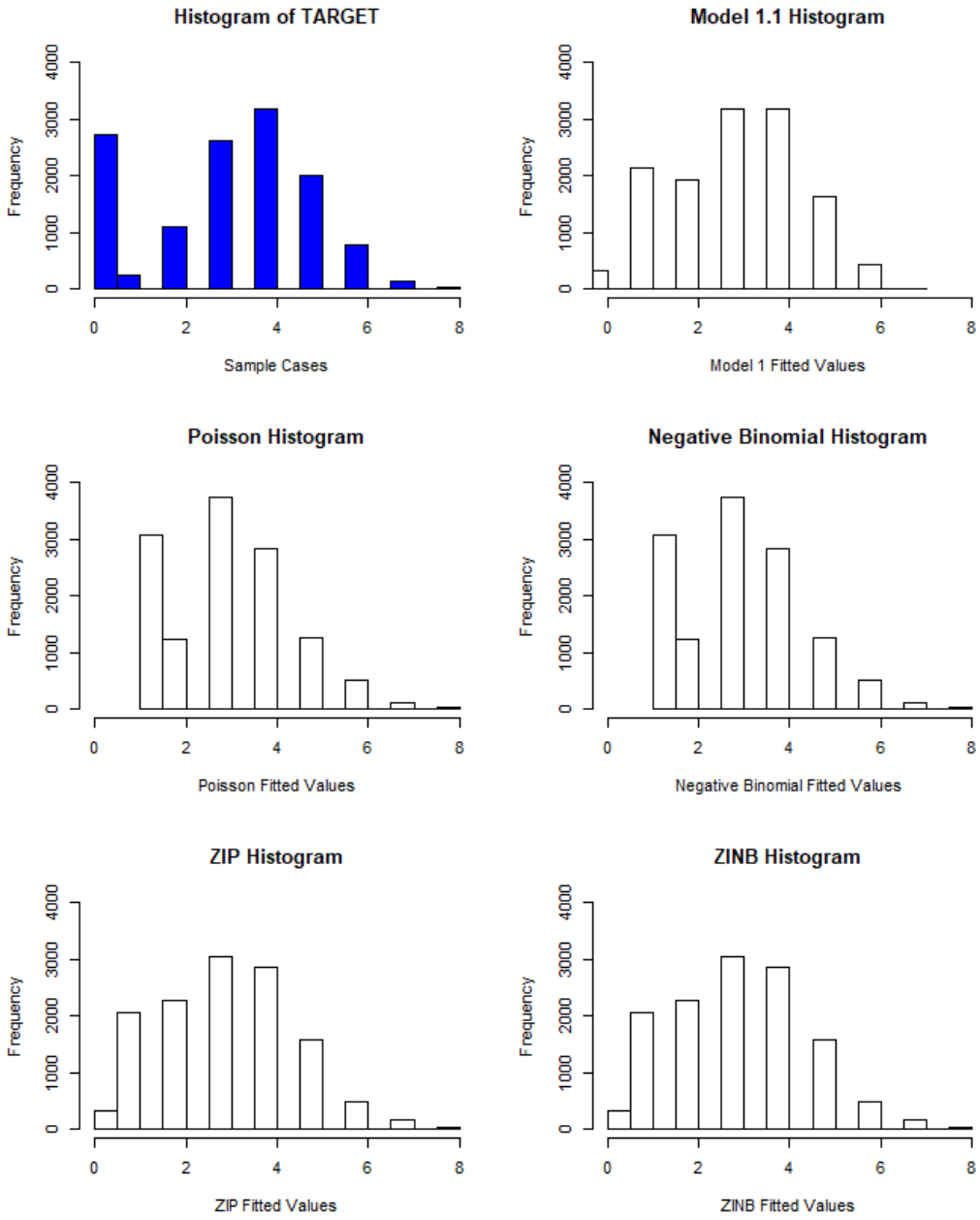


Figure 20 - Comparison of the fitted value distribution for the 5 models and the target variable.

References

Australian Government. (2016, June). *Wine Australia for Australian Wine produces*. Retrieved from https://www.wineaustralia.com/getmedia/81cbe0c6-491b-46ed-8b82-4f5af51c44d4/Wine-Australia-Compliance-Guide-June-2016.pdf

Collings, B. (2007, June 17). *Acid/pH ADJUSTMENTS*. Retrieved from http://www.bcawa.ca/winemaking/acidph.htm

Hoffmann, J. P. (2004). Generalized linear Models An Applied Approach. Boston: Pearson Education.

Jackisch, P. (1985). Modern Winemaking. Ithaca: Cornell University Press.

Monash . (n.d.). *Free Sulphur Dioxide*. Retrieved from Monash Scientific: http://www.monashscientific.com.au/FSO2.htm

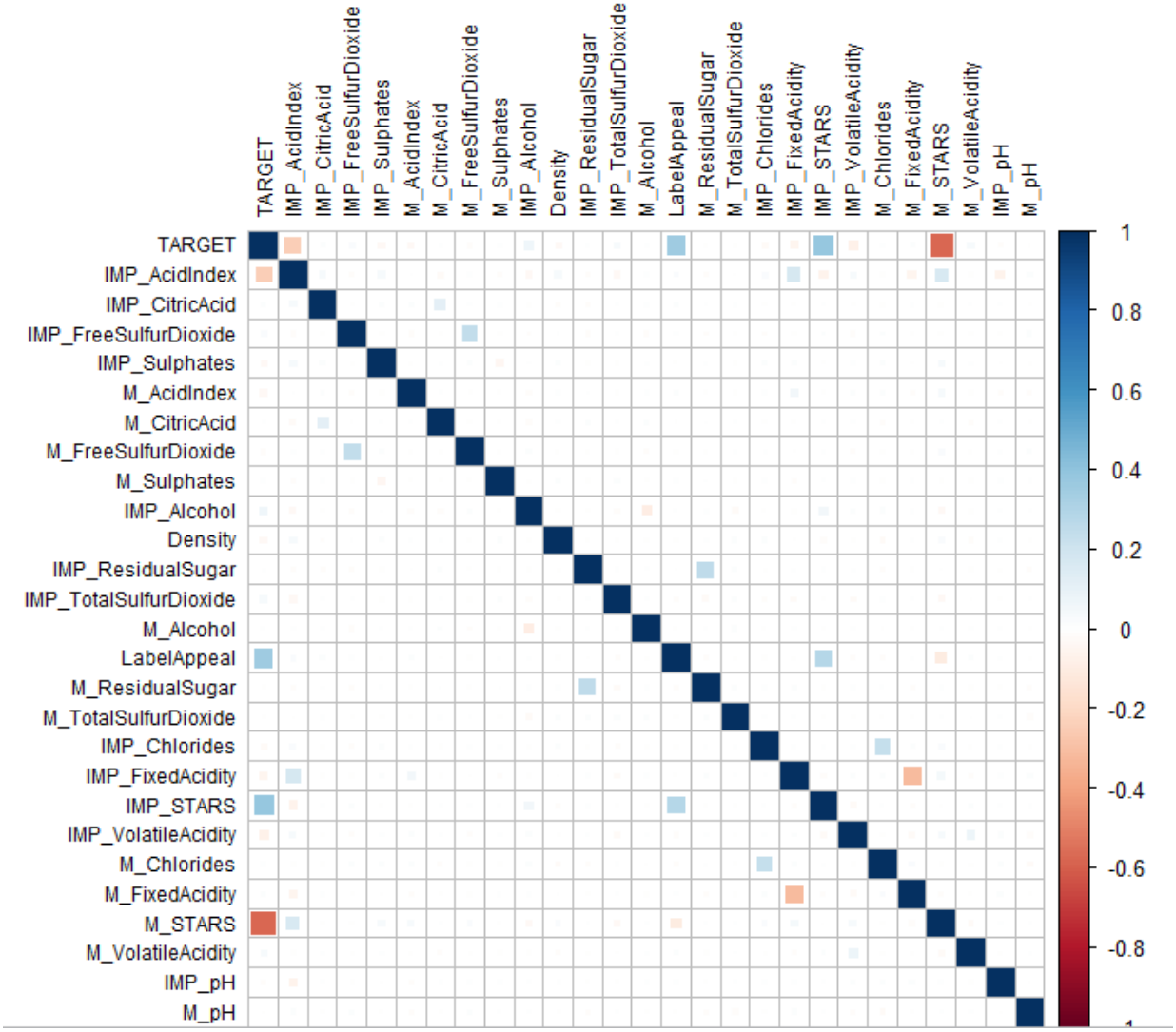
Nierman, D. (2004). *Fixed Acidity*. Retrieved from UC Davis: http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity

Waterhouse Lab. (2015, November 1). *Volatile acidity*. Retrieved from UC Davis: http://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity

Wine Folly. (2014, January 15). *The Bottom Line on Sulfites in Wine*. Retrieved from Wine Folly: http://winefolly.com/tutorial/sulfites-in-wine/

Wine Folly. (2015, December 9). *Understanding Acidity in Wine*. Retrieved from http://winefolly.com/review/understanding-acidity-in-wine/

Appendix 1 – Correlation Matrix



Appendix 2 – OLS Regression Full Model

```
Call:
lm(formula = TARGET ~ ., data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6428 -0.8530  0.0297  0.8511  6.2223

Coefficients:
                        Estimate  Std. Error t value             Pr(>|t|)
(Intercept)           4.51139200  0.44969606  10.032 < 0.0000000000000002
IMP_AcidIndex        -0.20916310  0.00939212 -22.270 < 0.0000000000000002
IMP_CitricAcid        0.03094280  0.01929329   1.604               0.10878
IMP_FreeSulfurDioxide 0.00026567  0.00012478   2.129               0.03327
IMP_Sulphates        -0.03363130  0.01862971  -1.805               0.07106
M_AcidIndex          -0.92071584  0.29435588  -3.128               0.00176
M_CitricAcid          0.01066380  0.02769445   0.385               0.70021
M_FreeSulfurDioxide  -0.03947210  0.02625731  -1.503               0.13279
M_Sulphates           0.01487681  0.02588507   0.575               0.56549
IMP_Alcohol           0.01383491  0.00330680   4.184         0.0000288631
Density              -0.78723294  0.43779170  -1.798               0.07217
IMP_ResidualSugar    -0.00007931  0.00049286  -0.161               0.87215
IMP_TotalSulfurDioxide 0.00032163  0.00010739   2.995               0.00275
M_Alcohol             0.07124088  0.04896608   1.455               0.14572
LabelAppeal           0.46529778  0.01368690  33.996 < 0.0000000000000002
M_ResidualSugar       0.01515107  0.02635597   0.575               0.56539
M_TotalSulfurDioxide  0.01220634  0.02525879   0.483               0.62893
IMP_Chlorides        -0.09689612  0.05223811  -1.855               0.06363
IMP_FixedAcidity      0.00002954  0.00248647   0.012               0.99052
IMP_STARS             0.77967377  0.01569397  49.680 < 0.0000000000000002
IMP_VolatileAcidity  -0.11971492  0.02098053  -5.706          0.0000000118
M_Chlorides           0.02816374  0.02605097   1.081               0.27967
M_FixedAcidity       -0.01145687  0.03679807  -0.311               0.75554
M_STARS              -2.28008571  0.02700359 -84.436 < 0.0000000000000002
M_VolatileAcidity     0.07192260  0.02804615   2.564               0.01035
IMP_pH               -0.06650359  0.02175684  -3.057               0.00224
M_pH                 -0.00057004  0.04277770  -0.013               0.98937

Residual standard error: 1.311 on 12768 degrees of freedom
Multiple R-squared:  0.5376,   Adjusted R-squared:  0.5367
F-statistic:   571 on 26 and 12768 DF,  p-value: < 0.00000000000000022
```

Appendix 3 – ZIP Full Model

```
zeroinfl(formula = TARGET ~ . | ., data = mydata)
Pearson residuals:
      Min        1Q    Median        3Q       Max
-2.301118 -0.417048 -0.002621  0.377873  5.808164
Count model coefficients (poisson with log link):
                         Estimate   Std. Error z value           Pr(>|z|)
(Intercept)             1.455536175  0.203989765   7.135    0.0000000000009
IMP_AcidIndex          -0.020477229  0.004978820  -4.113    0.0000390774556
IMP_CitricAcid         -0.001558354  0.008632631  -0.181              0.857
IMP_FreeSulfurDioxide   0.000006649  0.000055228   0.120              0.904
IMP_Sulphates           0.005627202  0.008410976   0.669              0.503
M_AcidIndex             0.069855420  0.200261104   0.349              0.727
M_CitricAcid            0.000002342  0.012417372   0.000              1.000
M_FreeSulfurDioxide    -0.005300768  0.011781992  -0.450              0.653
M_Sulphates            -0.000045330  0.011561405  -0.004              0.997
IMP_Alcohol             0.007309195  0.001482503   4.930    0.0000008210065
Density                -0.280667639  0.197856341  -1.419              0.156
IMP_ResidualSugar      -0.000035844  0.000221581  -0.162              0.871
IMP_TotalSulfurDioxide -0.000034978  0.000048881  -0.716              0.474
M_Alcohol               0.009413223  0.021963247   0.429              0.668
LabelAppeal             0.232151114  0.006321581  36.724 < 0.00000000000000
M_ResidualSugar         0.012212262  0.011811161   1.034              0.301
M_TotalSulfurDioxide   -0.001581054  0.011339088  -0.139              0.889
IMP_Chlorides          -0.023117486  0.023778716  -0.972              0.331
IMP_FixedAcidity        0.000426372  0.001125228   0.379              0.705
IMP_STARS               0.104617302  0.006408023  16.326 < 0.00000000000000
IMP_VolatileAcidity    -0.014780116  0.009638114  -1.534              0.125
M_Chlorides             0.008898296  0.011720948   0.759              0.448
M_FixedAcidity         -0.002283867  0.016304596  -0.140              0.889
M_STARS                -0.187388743  0.018575472 -10.088 < 0.00000000000000
M_VolatileAcidity       0.009906046  0.012424459   0.797              0.425
IMP_pH                  0.001709263  0.009898843   0.173              0.863
M_pH                   -0.004957578  0.019182362  -0.258              0.796
Zero-inflation model coefficients (binomial with logit link):
                       Estimate Std. Error z value             Pr(>|z|)
(Intercept)            -2.7892756  1.3709239  -2.035             0.041892
IMP_AcidIndex           0.4433517  0.0266358  16.645 < 0.0000000000000002
IMP_CitricAcid         -0.1125431  0.0574725  -1.958             0.050206
IMP_FreeSulfurDioxide  -0.0007002  0.0003762  -1.861             0.062753
IMP_Sulphates           0.1950157  0.0529258   3.685             0.000229
M_AcidIndex             2.1314721  0.8257364   2.581             0.009843
M_CitricAcid           -0.0879769  0.0814553  -1.080             0.280114
M_FreeSulfurDioxide     0.0948780  0.0764845   1.240             0.214795
M_Sulphates            -0.0408067  0.0758852  -0.538             0.590755
IMP_Alcohol             0.0271068  0.0096881   2.798             0.005143
Density                 0.3591545  1.2980384   0.277             0.782018
IMP_ResidualSugar      -0.0006489  0.0014472  -0.448             0.653850
IMP_TotalSulfurDioxide -0.0012328  0.0003154  -3.909            0.0000928
M_Alcohol              -0.1012136  0.1415610  -0.715             0.474620
LabelAppeal             0.7265529  0.0423959  17.137 < 0.0000000000000002
M_ResidualSugar         0.0104373  0.0769112   0.136             0.892053
M_TotalSulfurDioxide   -0.0290501  0.0739663  -0.393             0.694506
IMP_Chlorides           0.0348553  0.1516829   0.230             0.818254
IMP_FixedAcidity        0.0028523  0.0072370   0.394             0.693489
IMP_STARS              -3.8159856  0.3327619 -11.468 < 0.0000000000000002
IMP_VolatileAcidity     0.2159909  0.0593544   3.639             0.000274
M_Chlorides             0.0897098  0.0761831   1.178             0.238974
M_FixedAcidity          0.0218774  0.1093877   0.200             0.841481
M_STARS                 6.0401861  0.3464254  17.436 < 0.0000000000000002
M_VolatileAcidity      -0.1711812  0.0843911  -2.028             0.042517
IMP_pH                  0.2684933  0.0638922   4.202            0.0000264
M_pH                   -0.1275748  0.1261910  -1.011             0.312033
```