Allison Roeser
Kaggle Name: Allison Roeser
Predict 411
Section 55
August 20, 2017

Poisson Regression

Wine Distribution: Predicting Sales of Sampling Cases

BINGO BONUS:
90 points requested

- (20 Points) Develop a LOGISTIC / POISSON model (if you like it, you may select this as your champion) **20 points. Pages 14, 15, and 16**
- (20 Points) Use decision tree software such as Angoss or Weka or something else for variable selection or missing value imputation (the more use you make of decision trees, the more points you will receive). Be sure to carefully present your decision tree output so that I can see what you did. **20 points. Page 6,7, and 8**
- (20 Points) Build a decision tree model to predict the number of cases sold. You cannot use this as your champion model, but comment on it and compare it to your Poisson models. **20 points. Pages 19 and 20**
- (20 Points) Recreate as much of the program as you can in "R" **20 points. Random Forest on Page 5. R code attached**
- (10 Points) Use SAS Macros or use, in my opinion, good programming technique **10 points. Page 22**
- (?? Points) Roll the dice … think of something creative and run with it. I might give you points.

Introduction

The purpose of this study was to create a predictive model that could be used to forecast the number cases of wine that a wine distributer would purchase after sampling a particular wine. A wine distributor would usually purchase between one and four cases of wine to use for tasting samples at wine stores and high end restaurants. The wines that have the highest number of sample cases purchased, will have the best chance of becoming good selling wines in restaurants and wine stores. By understanding what characteristics create a high sample order wine, the wine manufacture in the study can adjust their wine assortment appropriately to maximize sales.

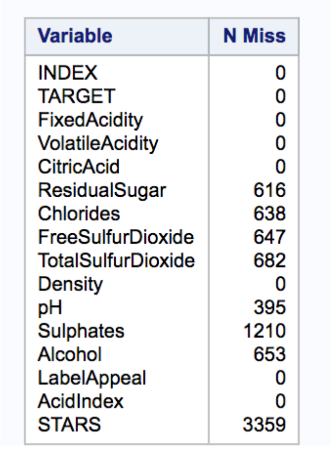
The dataset contains information on the chemical properties, label aesthetics, and stars awarded for 12,000 wines. The dataset also includes a Target value, which is the number of cases the distributor actually bought after sampling the wine. This Target value is what the model is being used to predict.

The dataset was cleaned before the potential models were created. The observations with missing values were flagged, the missing values were imputed with median values, and extreme outliers were capped at the 5% / 95% threshold. Various methods were used to build the models including: Linear Regression, Poisson, Negative Binomial, Zero Inflated Poisson, Hurdle, and Ensemble modeling. The models were compared to identify the champion model that was most accurate, understandable, and useful.

Data Exploration

Overview of Numeric Variables

The dataset contains sixteen numeric variables, which are listed in the table below.



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

Table 1: Numeric variables

The first variable, INDEX, is used to identify observations. This variable is not needed and will be dropped from the dataset. The next variable, TARGET, measures how many cases of wine were purchased by a distributor after the wine was sampled. This is the variable that the model is being built to predict. TARGET is a zero-inflated variable, as shown in the histogram below.

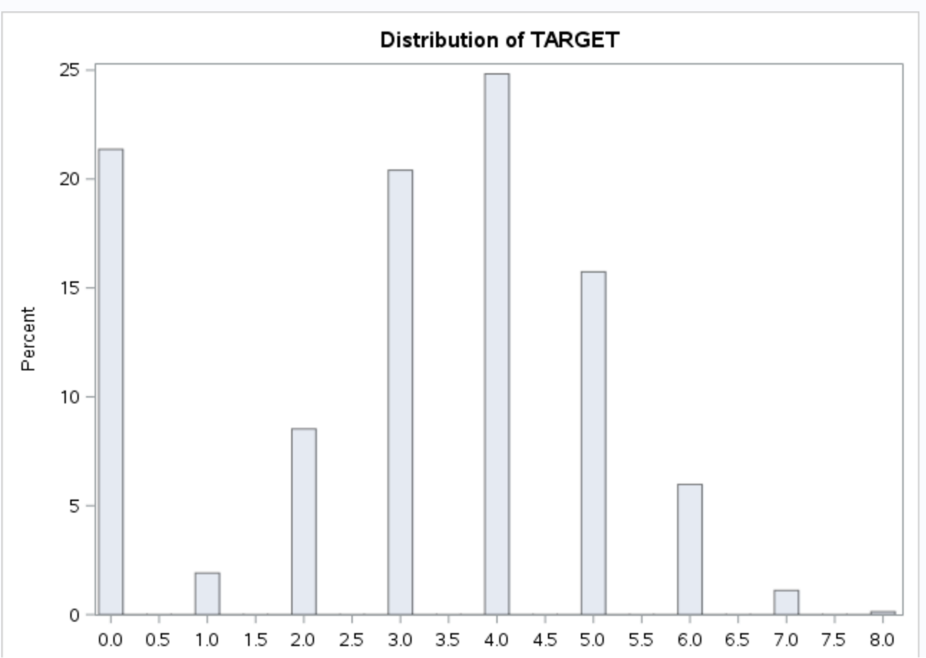
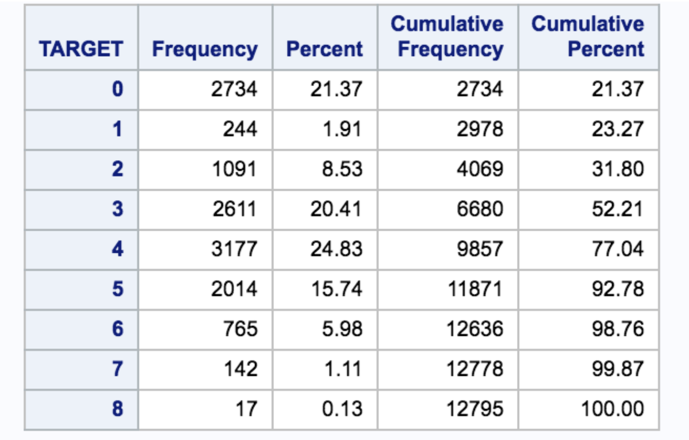


Figure 1: Histogram of TARGET

21.37% of the observations have a TARGET value of zero. If the zero values are removed from the data, the remaining values show a normal distribution. The frequency table below shows that the non-zero values for TARGET range from one to eight, and that the plurality of records have a TARGET value of four. The zero-inflated distribution of TARGET will require a special model to account for the large number of zero values.



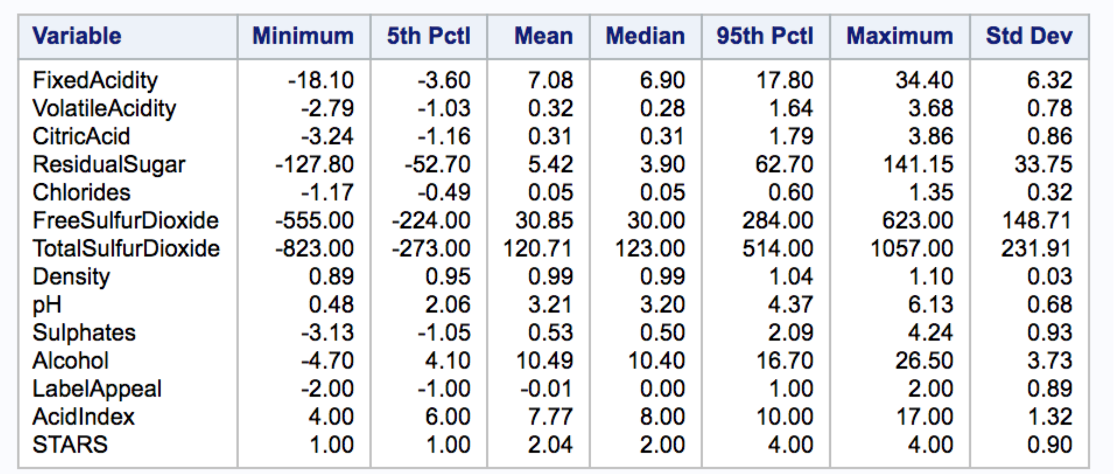
TARGET	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2734	21.37	2734	21.37
1	244	1.91	2978	23.27
2	1091	8.53	4069	31.80
3	2611	20.41	6680	52.21
4	3177	24.83	9857	77.04
5	2014	15.74	11871	92.78
6	765	5.98	12636	98.76
7	142	1.11	12778	99.87
8	17	0.13	12795	100.00

Table 2: Frequency table for TARGET

The remaining fourteen variables listed in Table 1 are the potential predictor variables. The majority of these variables are numerical measurements of the chemical composition of the wine. LabelAppeal and STARS are the only variables that do not relate to the chemical properties of the wine.

<u>Outliers</u>

Table 4 displays the Summary Statistics for the numeric, potential predictor variables. The mean and median values are similar for most variables. The notable exception is for ResidualSugar where the mean is 5.42 and the median is 3.90. This is caused by extreme outliers on both sides of the distribution.



Variable	Minimum	5th Pctl	Mean	Median	95th Pctl	Maximum	Std Dev
FixedAcidity	-18.10	-3.60	7.08	6.90	17.80	34.40	6.32
VolatileAcidity	-2.79	-1.03	0.32	0.28	1.64	3.68	0.78
CitricAcid	-3.24	-1.16	0.31	0.31	1.79	3.86	0.86
ResidualSugar	-127.80	-52.70	5.42	3.90	62.70	141.15	33.75
Chlorides	-1.17	-0.49	0.05	0.05	0.60	1.35	0.32
FreeSulfurDioxide	-555.00	-224.00	30.85	30.00	284.00	623.00	148.71
TotalSulfurDioxide	-823.00	-273.00	120.71	123.00	514.00	1057.00	231.91
Density	0.89	0.95	0.99	0.99	1.04	1.10	0.03
pH	0.48	2.06	3.21	3.20	4.37	6.13	0.68
Sulphates	-3.13	-1.05	0.53	0.50	2.09	4.24	0.93
Alcohol	-4.70	4.10	10.49	10.40	16.70	26.50	3.73
LabelAppeal	-2.00	-1.00	-0.01	0.00	1.00	2.00	0.89
AcidIndex	4.00	6.00	7.77	8.00	10.00	17.00	1.32
STARS	1.00	1.00	2.04	2.00	4.00	4.00	0.90

Table 3: Summary Statistics for predictor variables

AcidIndex is another variable that has extreme outliers, as seen in the distribution and QQ plots below. For high values of AcidIndex, the points pull away from the line on the QQ plot. This shows that the variable is not normally distributed. The Data Preparation section will discuss how outliers were reduced by capping the minimum and the maximum at the 5% and 95% threshold values.

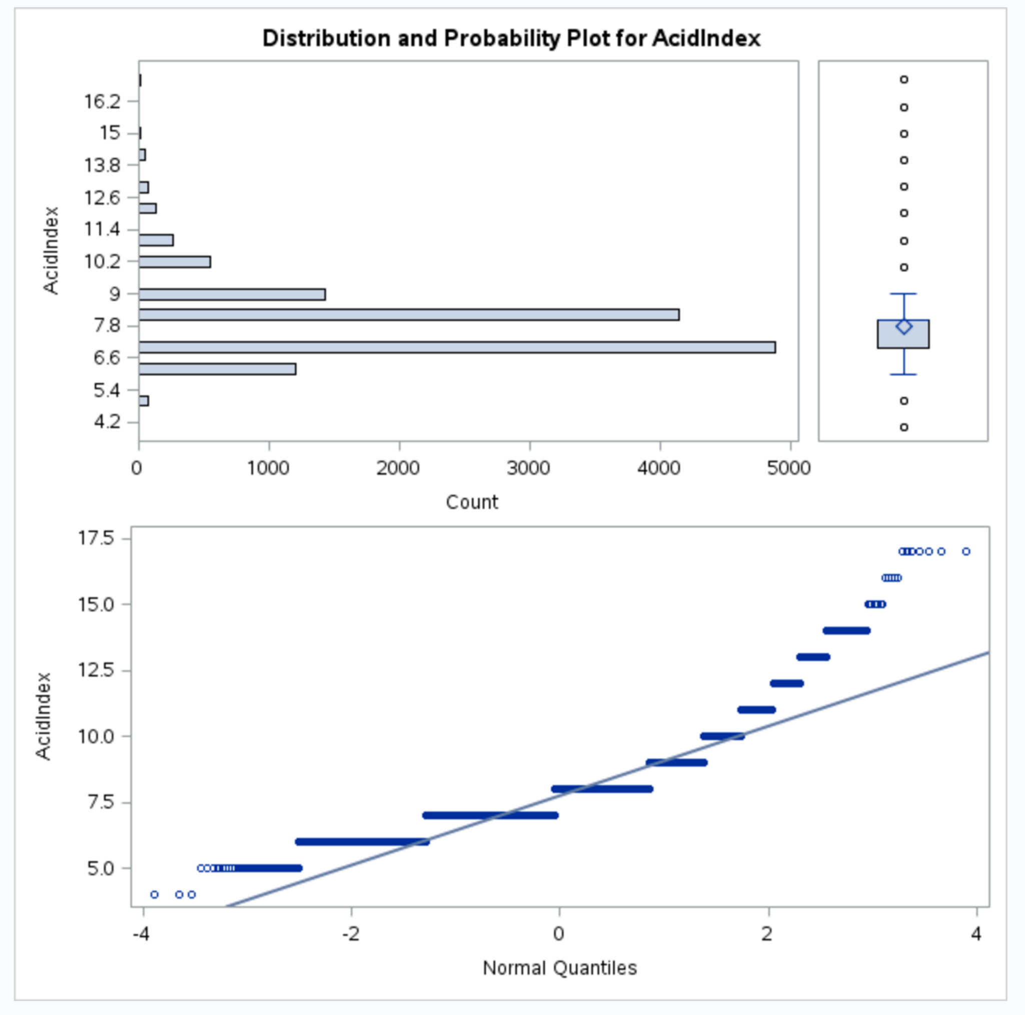
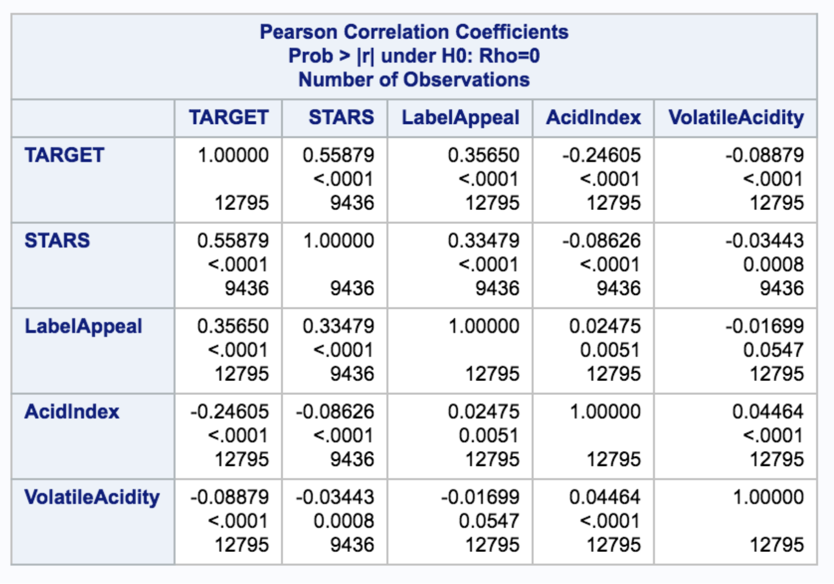


Figure 2: Distribution and QQ plot of AcidIndex

Correlation Between the Variables

The predictor variables that have the highest correlation with TARGET are STARS, LabelAppeal, AcidIndex, and VolatileAcidity. The table below provides a breakdown of how these variables are correlated with TARGET and with one another. The majority of the chemical composition variables have very low correlations with TARGET. These correlations show that reviews (STARS) and label aesthetics are the primary drivers in wine case sales. These variables will be especially important in the models. STARS and LabelAppeal also have a moderate correlation with each other (0.3350). This may mean that a wine that has an appealing label is more likely to earn a high review (high STARS count). Some multicollinearity between these variables may be present in the models and should be investigated.



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	TARGET	STARS	LabelAppeal	AcidIndex	VolatileAcidity
TARGET	1.00000 12795	0.55879 <.0001 9436	0.35650 <.0001 12795	-0.24605 <.0001 12795	-0.08879 <.0001 12795
STARS	0.55879 <.0001 9436	1.00000 9436	0.33479 <.0001 9436	-0.08626 <.0001 9436	-0.03443 0.0008 9436
LabelAppeal	0.35650 <.0001 12795	0.33479 <.0001 9436	1.00000 12795	0.02475 0.0051 12795	-0.01699 0.0547 12795
AcidIndex	-0.24605 <.0001 12795	-0.08626 <.0001 9436	0.02475 0.0051 12795	1.00000 12795	0.04464 <.0001 12795
VolatileAcidity	-0.08879 <.0001 12795	-0.03443 0.0008 9436	-0.01699 0.0547 12795	0.04464 <.0001 12795	1.00000 12795

Table 4: Variables with highest correlation to TARGET

Random Forest (built with R): BINGO BONUS

The Random Forest model shown below is useful for understanding which predictor variables may be especially important in the model. STARS and LabelAppeal have Node Purity scores that are significantly higher than for the variables that describe the chemical composition of the wine. STARS and LabelAppeal were also the variables with the larges correlation coefficients.

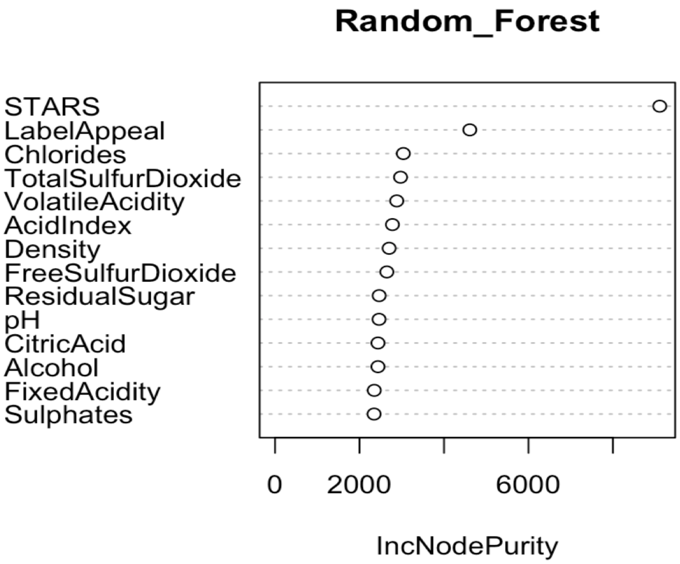


Figure 3: Random Forest model

Decision Tree: BINGO BONUS

A Decision Tree was constructed in Enterprise Miner to further understand which predictor variables have the greatest impact on TARGET. A pruned version of the tree is shown in the figure below. This tree shows that STARS, AcidIndex, and LabelAppeal have the greatest impact on the direction of branching. It has become quite clear that these will be important variables in the model.

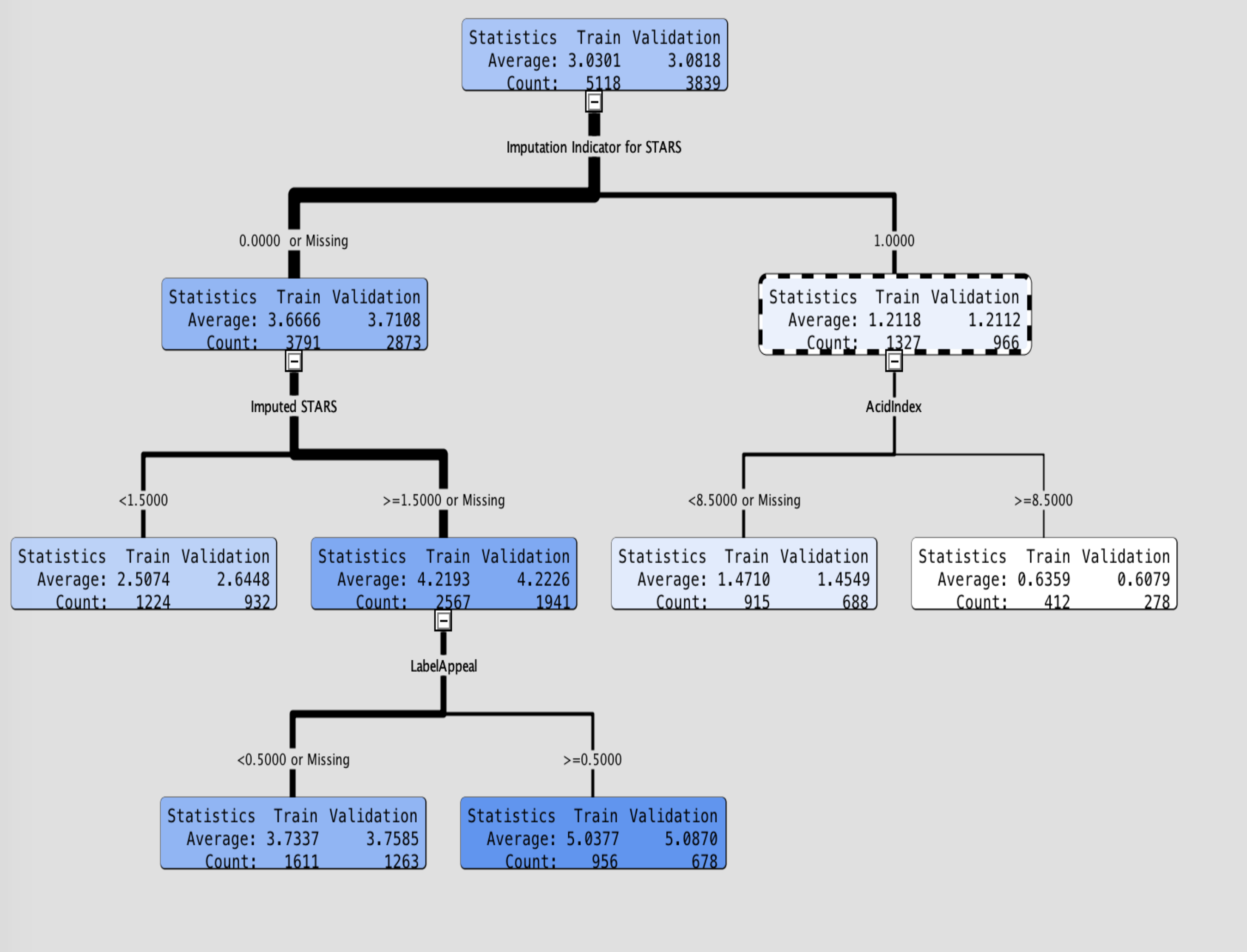
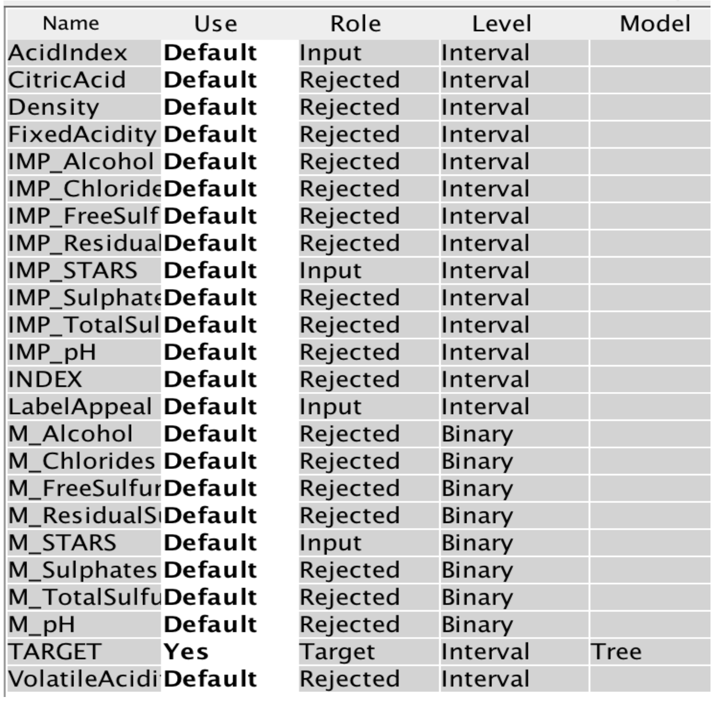


Figure 4: Decision Tree

Enterprise Miner Variable Selection: BINGO BONUS

Enterprise Miner was also used for variable selection. The table below provides the output. Most of the potential predictor variables were rejected. The only variables that remain are AcidIndex, IMP_STARS, LabelAppeal, and M_STARS. This result mirrors the results shown in the decision tree above. It has become clear that these are the most important variables in the dataset.



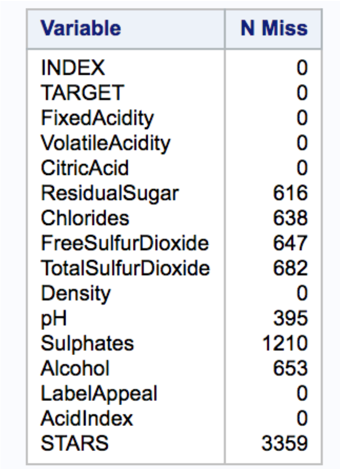
Name	Use	Role	Level	Model
AcidIndex	Default	Input	Interval	
CitricAcid	Default	Rejected	Interval	
Density	Default	Rejected	Interval	
FixedAcidity	Default	Rejected	Interval	
IMP_Alcohol	Default	Rejected	Interval	
IMP_Chloride	Default	Rejected	Interval	
IMP_FreeSulf	Default	Rejected	Interval	
IMP_Residual	Default	Rejected	Interval	
IMP_STARS	Default	Input	Interval	
IMP_Sulphate	Default	Rejected	Interval	
IMP_TotalSul	Default	Rejected	Interval	
IMP_pH	Default	Rejected	Interval	
INDEX	Default	Rejected	Interval	
LabelAppeal	Default	Input	Interval	
M_Alcohol	Default	Rejected	Binary	
M_Chlorides	Default	Rejected	Binary	
M_FreeSulfur	Default	Rejected	Binary	
M_ResidualS	Default	Rejected	Binary	
M_STARS	Default	Input	Binary	
M_Sulphates	Default	Rejected	Binary	
M_TotalSulfu	Default	Rejected	Binary	
M_pH	Default	Rejected	Binary	
TARGET	Yes	Target	Interval	Tree
VolatileAcidi	Default	Rejected	Interval	

Table 5: Enterprise Miner variable selection results

Data Preparation

Missing Values

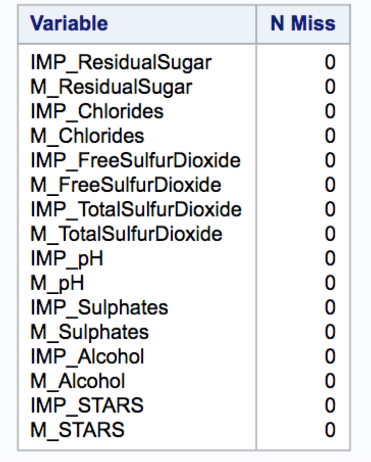
Eight numeric variables had missing values in the original dataset.



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

Table 6: Variables with missing observations

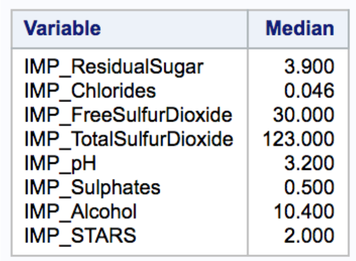
To maintain the integrity of the original dataset, the contents of the original variables was not change. Variables that contained missing values were copied to new variables call IMP_variable. The missing values were imputed into these variable copies. The imputed variables were identical to the original variables with the exception that the missing values had been fixed. Flag variables were also created for each variable that included missing values. The flag variables were named M-variable. If a variable was missing, the flag variable in that observation would be set to one. If the variable was not missing, the flag variable value would be set to zero. The table below shows the five original numeric variables that contained missing values, the five flag variables, and the five imputed variables. The original variables still contain missing values. However, the flag and imputed variables do not have any missing values.



Variable	N Miss
IMP_ResidualSugar	0
M_ResidualSugar	0
IMP_Chlorides	0
M_Chlorides	0
IMP_FreeSulfurDioxide	0
M_FreeSulfurDioxide	0
IMP_TotalSulfurDioxide	0
M_TotalSulfurDioxide	0
IMP_pH	0
M_pH	0
IMP_Sulphates	0
M_Sulphates	0
IMP_Alcohol	0
M_Alcohol	0
IMP_STARS	0
M_STARS	0

Table 7: Imputed and Missing Flag variables showing no missing values

The missing values were imputed with the median value for the variable. The median was chosen over the mean because a number of the variables in the dataset contain significant outliers. Outliers will impact the mean, but they do not impact the median. The median values imputed for each variable are listed in the table below.

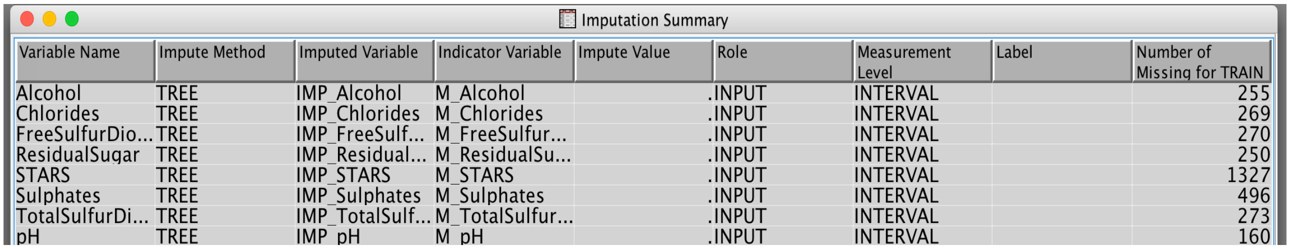


Variable	Median
IMP_ResidualSugar	3.900
IMP_Chlorides	0.046
IMP_FreeSulfurDioxide	30.000
IMP_TotalSulfurDioxide	123.000
IMP_pH	3.200
IMP_Sulphates	0.500
IMP_Alcohol	10.400
IMP_STARS	2.000

Table 8: Median value for each variable

Missing Values with Enterprise Miner: BINGO BONUS

An alternative method for imputing missing values and creating flag variables was performed with Enterprise Miner. Enterprise Miner used a decision tree to imput the variables, as shown in the table below.



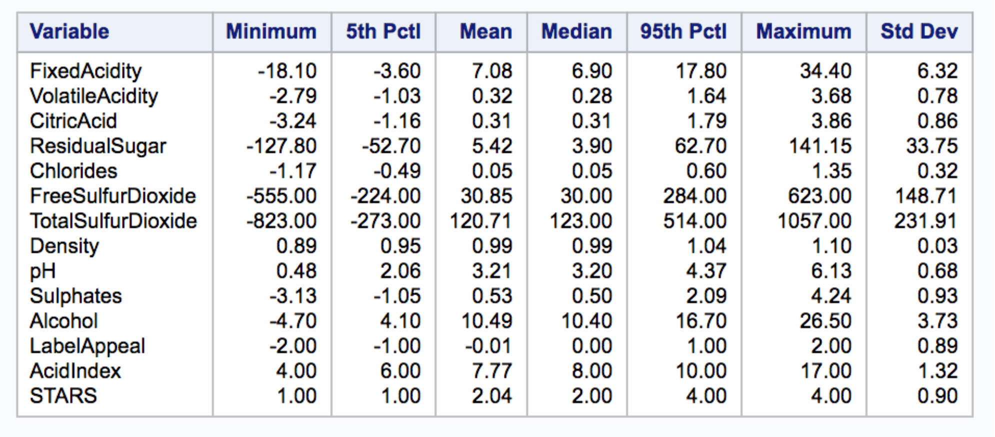
Imputation Summary

Variable Name	Impute Method	Imputed Variable	Indicator Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Alcohol	TREE	IMP_Alcohol	M_Alcohol	.	INPUT	INTERVAL		255
Chlorides	TREE	IMP_Chlorides	M_Chlorides	.	INPUT	INTERVAL		269
FreeSulfurDio...	TREE	IMP_FreeSulf...	M_FreeSulfur...	.	INPUT	INTERVAL		270
ResidualSugar	TREE	IMP_Residual...	M_ResidualSu...	.	INPUT	INTERVAL		250
STARS	TREE	IMP_STARS	M_STARS	.	INPUT	INTERVAL		1327
Sulphates	TREE	IMP_Sulphates	M_Sulphates	.	INPUT	INTERVAL		496
TotalSulfurDi...	TREE	IMP_TotalSulf...	M_TotalSulfur...	.	INPUT	INTERVAL		273
pH	TREE	IMP_pH	M_pH	.	INPUT	INTERVAL		160

Table 9: Imputation Summary from Enterprise Miner

Outliers

Many of the numeric variables have minimums and maximums that are 5 or 6 standard deviations from the mean. These outliers were capped at the 5% and 95% thresholds. The table below shows the Summary Statistics that were discussed in detail in the Data Exploration section.



Variable	Minimum	5th Pctl	Mean	Median	95th Pctl	Maximum	Std Dev
FixedAcidity	-18.10	-3.60	7.08	6.90	17.80	34.40	6.32
VolatileAcidity	-2.79	-1.03	0.32	0.28	1.64	3.68	0.78
CitricAcid	-3.24	-1.16	0.31	0.31	1.79	3.86	0.86
ResidualSugar	-127.80	-52.70	5.42	3.90	62.70	141.15	33.75
Chlorides	-1.17	-0.49	0.05	0.05	0.60	1.35	0.32
FreeSulfurDioxide	-555.00	-224.00	30.85	30.00	284.00	623.00	148.71
TotalSulfurDioxide	-823.00	-273.00	120.71	123.00	514.00	1057.00	231.91
Density	0.89	0.95	0.99	0.99	1.04	1.10	0.03
pH	0.48	2.06	3.21	3.20	4.37	6.13	0.68
Sulphates	-3.13	-1.05	0.53	0.50	2.09	4.24	0.93
Alcohol	-4.70	4.10	10.49	10.40	16.70	26.50	3.73
LabelAppeal	-2.00	-1.00	-0.01	0.00	1.00	2.00	0.89
AcidIndex	4.00	6.00	7.77	8.00	10.00	17.00	1.32
STARS	1.00	1.00	2.04	2.00	4.00	4.00	0.90

Table 10: Summary Statistics for predictor variables

For example, the Distribution and QQ plots for AcidIndex before the outliers were capped are shown below. There are many high value outliers that create a long tail and a skewed distribution. In the QQ Plot the points move away from the center guidance line.

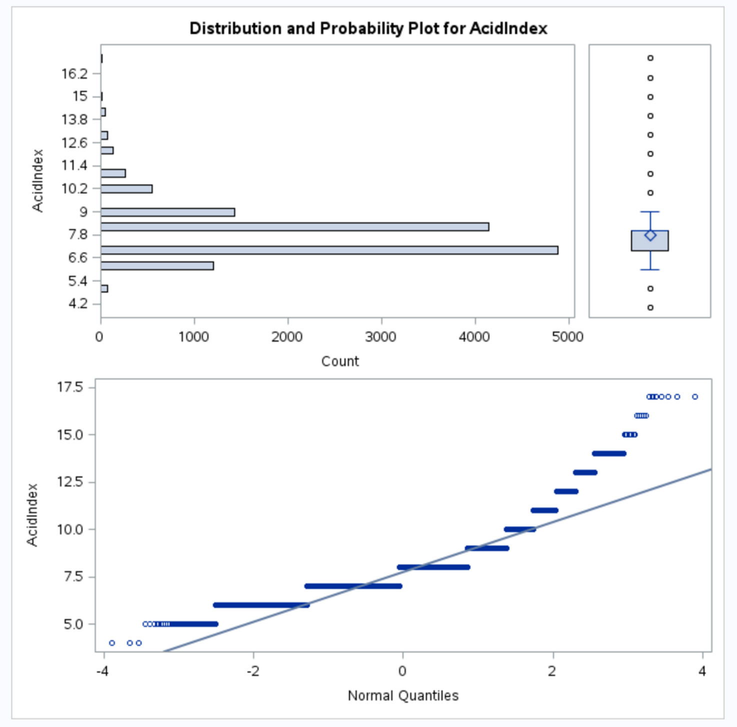


Figure 5: Distribution and QQ Plots for AcidIndex before outliers are removed

The figure below shows the Distribution and QQ plots for IMP_AcidIndex after the outliers have been removed and the missing values imputed with the median value. The distribution is more normal (with the exception of the spikes at both ends and in the middle from the imputed values), and there are no outlier dots present in the boxplot. The QQ plot follows the linear guideline more closely.

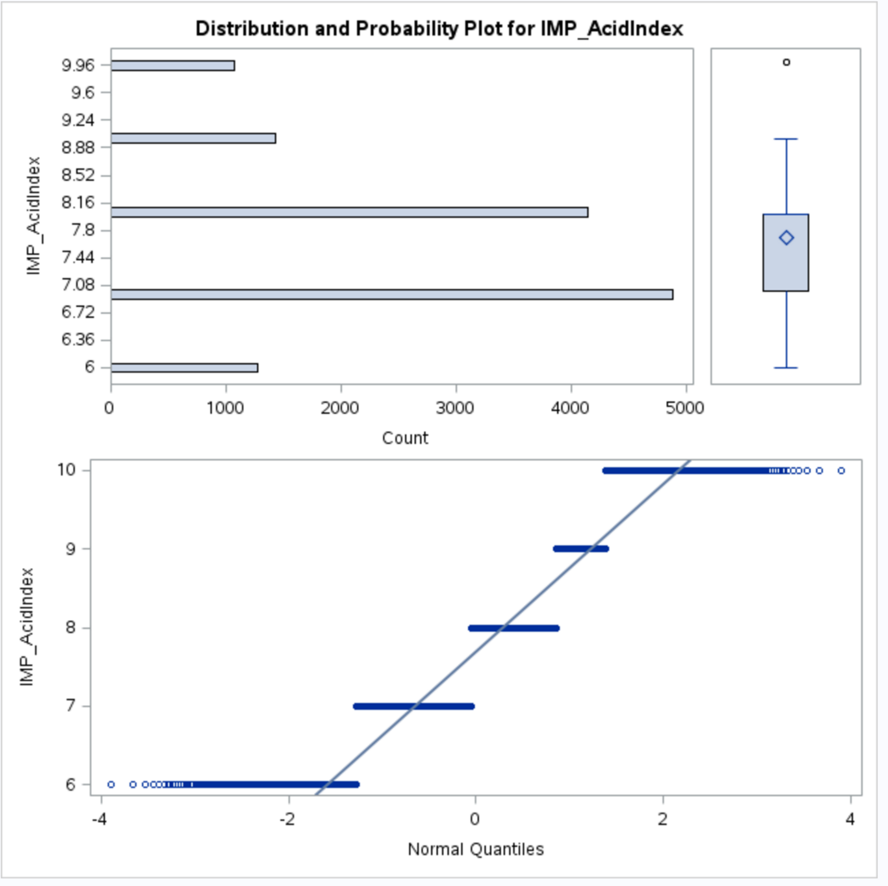
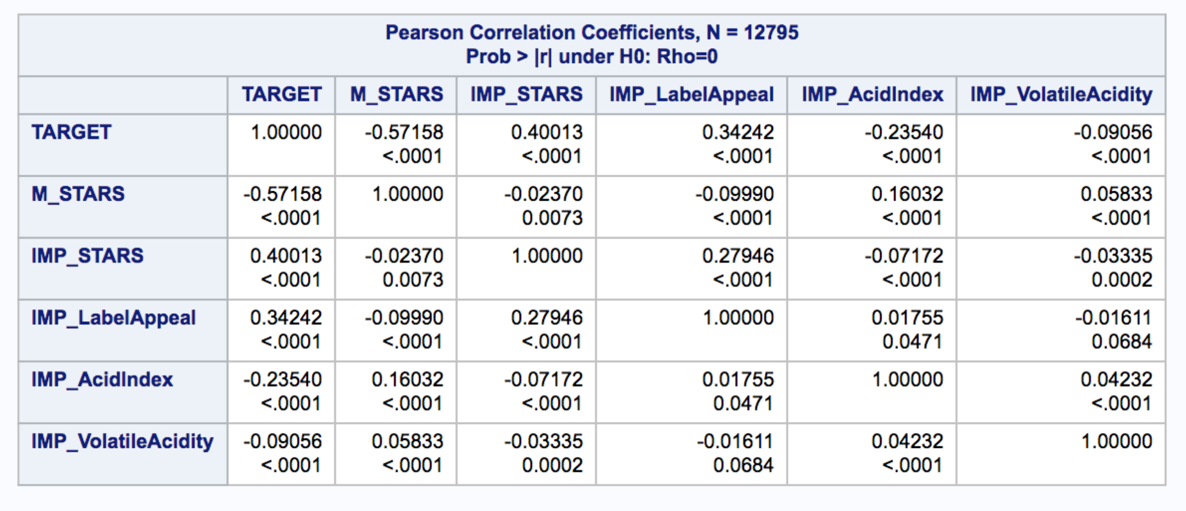


Figure 6: Distribution and QQ Plots for AcidIndex after outliers are removed

The table of correlation coefficients after the data is cleaned is presented in the table below. IMP_STARS, IMP_LabelAppeal, and IMP_AcidIndex still have strong correlations with TARGET. The missing flag variable for STARS (M_STARS) now has the highest correlation with TARGET. This means that not having any Star rating is highly predictive of low wine case sales.



Pearson Correlation Coefficients, N = 12795 Prob > |r| under H0: Rho=0						
	TARGET	M_STARS	IMP_STARS	IMP_LabelAppeal	IMP_AcidIndex	IMP_VolatileAcidity
TARGET	1.00000	-0.57158 <.0001	0.40013 <.0001	0.34242 <.0001	-0.23540 <.0001	-0.09056 <.0001
M_STARS	-0.57158 <.0001	1.00000	-0.02370 0.0073	-0.09990 <.0001	0.16032 <.0001	0.05833 <.0001
IMP_STARS	0.40013 <.0001	-0.02370 0.0073	1.00000	0.27946 <.0001	-0.07172 <.0001	-0.03335 0.0002
IMP_LabelAppeal	0.34242 <.0001	-0.09990 <.0001	0.27946 <.0001	1.00000	0.01755 0.0471	-0.01611 0.0684
IMP_AcidIndex	-0.23540 <.0001	0.16032 <.0001	-0.07172 <.0001	0.01755 0.0471	1.00000	0.04232 <.0001
IMP_VolatileAcidity	-0.09056 <.0001	0.05833 <.0001	-0.03335 0.0002	-0.01611 0.0684	0.04232 <.0001	1.00000

Table 11: Variables with highest correlation to TARGET after imputation of missing values

Transformations

After removing outliers and imputing missing values, IMP_VolatileAcidity still did not have a normal distribution, as shown in the figure below.

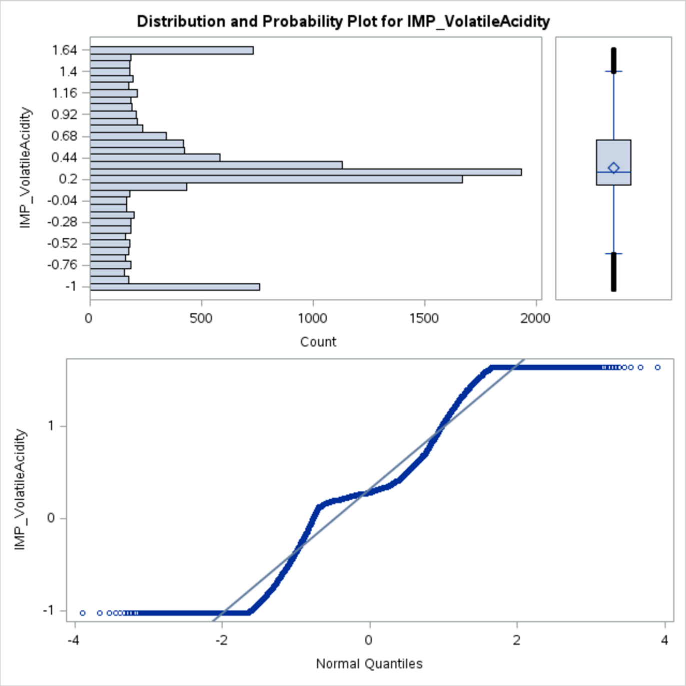


Figure 7: Distribution and QQ Plots for IMP_VolatileAcidity show a non-normal distribution

Both square root and logarithmic transformations were tried. The logarithmic transformation produced the best results, as shown in the figure below.

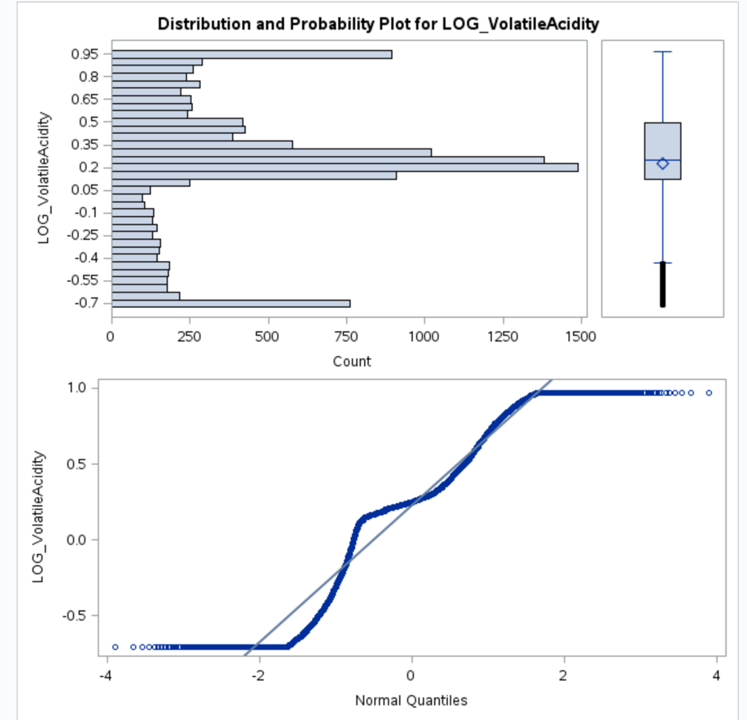


Figure 8: Distribution and QQ Plots for LOG_VolatileAcidity

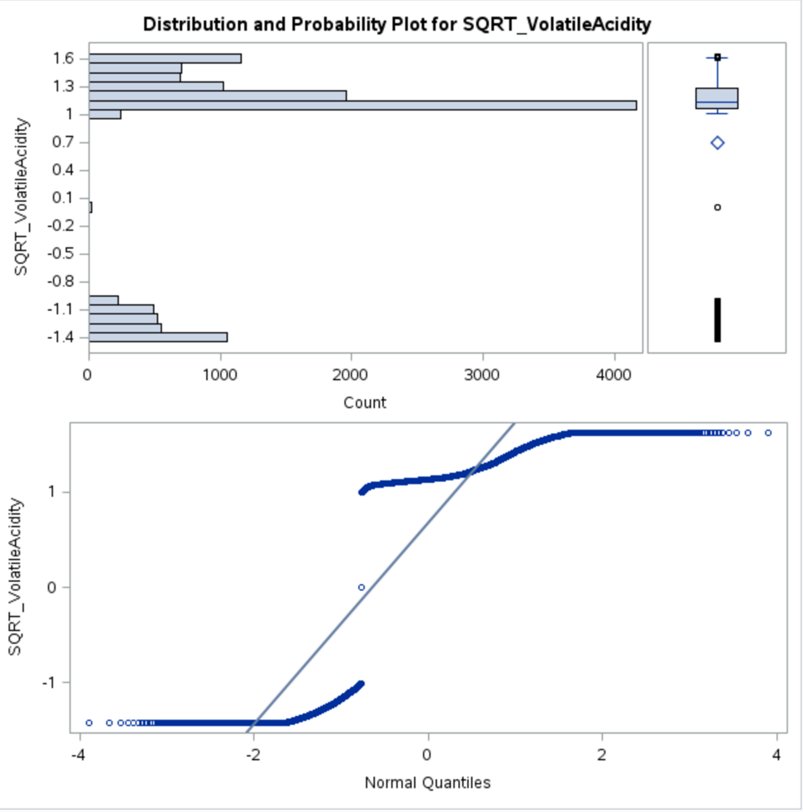


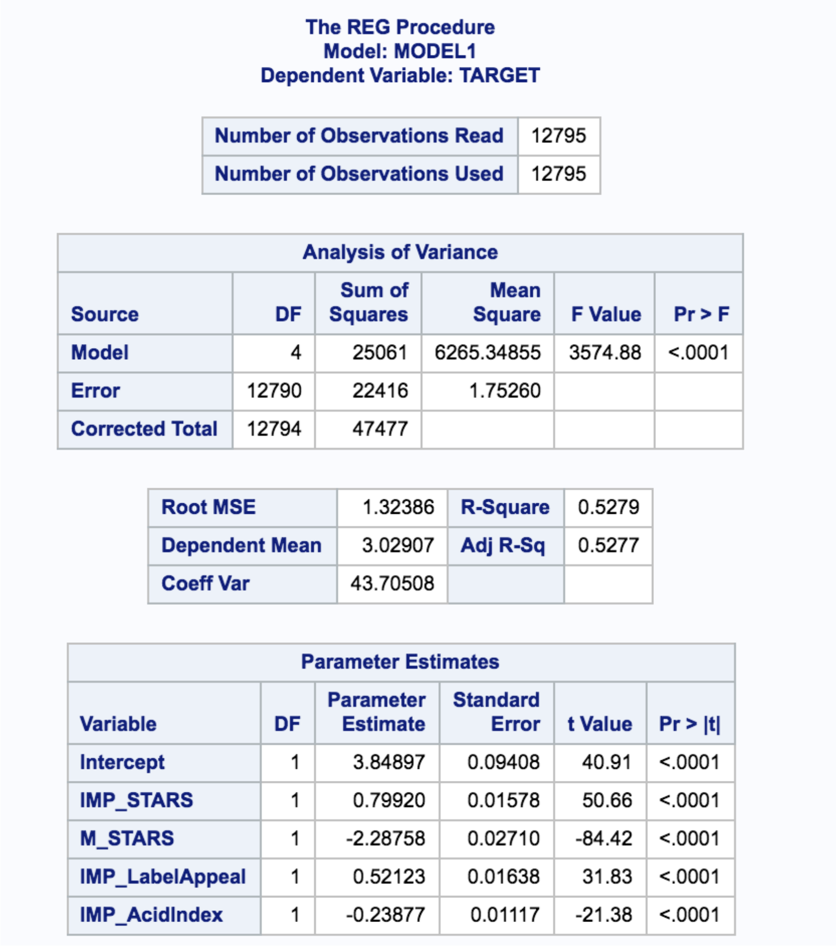
Figure 9: Distribution and QQ Plots for SQRT_VolatileAcidity

Model Building

Model 1: Linear Regression

The first model is a Linear Regression. Usually, the Linear Regression is not a good choice for modeling a zero-inflated variable. However, it is a good baseline to start. The model was built using the four predictor variables that had the highest correlation with TARGET: M_STARS, IMP_STARS, IMP_LabelAppeal, and IMP_AcidIndex.

The Adjusted R Squared value is fairly strong at 52.8%. The Root MSE is 1.324. The linear regression model could be a potentially strong model despite not being optimized for zero-inflated variable modeling.



The REG Procedure
Model: MODEL1
Dependent Variable: TARGET

Number of Observations Read	12795
Number of Observations Used	12795

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	25061	6265.34855	3574.88	<.0001
Error	12790	22416	1.75260		
Corrected Total	12794	47477			

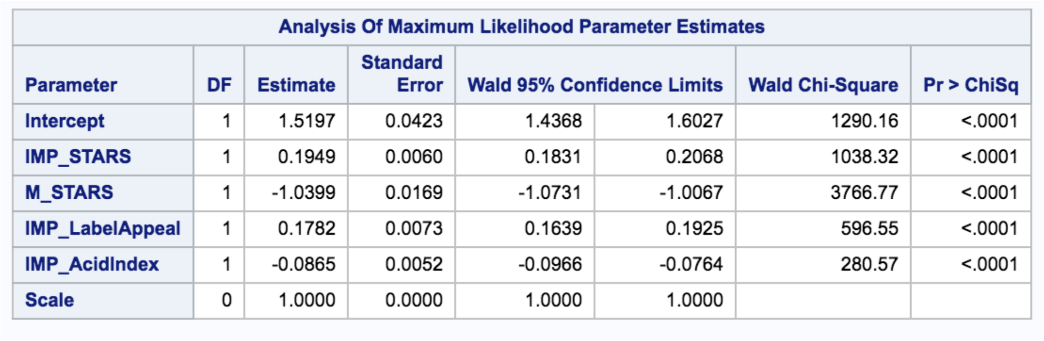
Root MSE	1.32386	R-Square	0.5279
Dependent Mean	3.02907	Adj R-Sq	0.5277
Coeff Var	43.70508		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.84897	0.09408	40.91	<.0001
IMP_STARS	1	0.79920	0.01578	50.66	<.0001
M_STARS	1	-2.28758	0.02710	-84.42	<.0001
IMP_LabelAppeal	1	0.52123	0.01638	31.83	<.0001
IMP_AcidIndex	1	-0.23877	0.01117	-21.38	<.0001

Table 12: Regression Model

Model 2: Poisson

A Poisson model is a better alternative to modeling “counting” variables such as TARGET. A Poisson model does not require a linear relationship between the inputs and the Target. It also will not allow for a negative prediction. Poisson models are frequently used to model “count” variables, unlike linear regression which is used to predict continuous variables. The intercept and coefficients for the Poisson model are shown below.

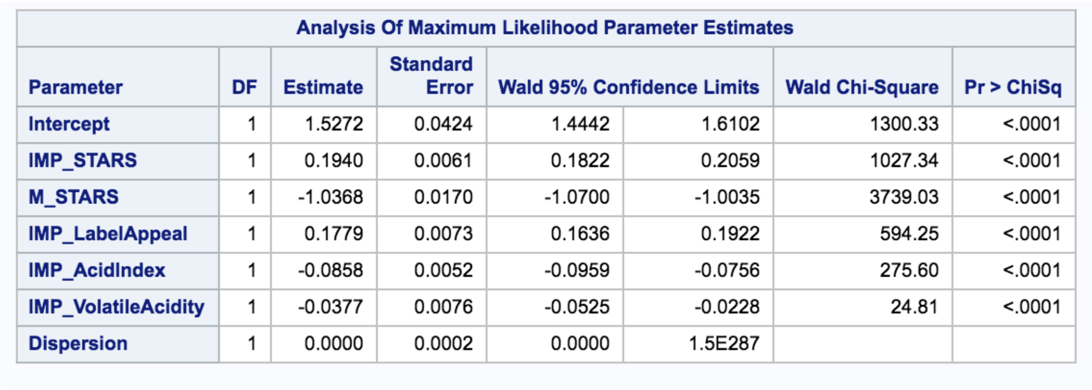


Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.5197	0.0423	1.4368	1.6027	1290.16	<.0001
IMP_STARS	1	0.1949	0.0060	0.1831	0.2068	1038.32	<.0001
M_STARS	1	-1.0399	0.0169	-1.0731	-1.0067	3766.77	<.0001
IMP_LabelAppeal	1	0.1782	0.0073	0.1639	0.1925	596.55	<.0001
IMP_AcidIndex	1	-0.0865	0.0052	-0.0966	-0.0764	280.57	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Table 13: Poisson Model

Model 3: Negative Binomial

The Negative Binomial Regression actually produces the same model as the Poisson Regression, given this data. In order to have an alternative model, an additional variable (IMP_VolatileAcidity) was added to the Negative Binomial Regression. IMP_VolatileAcidity is statistically significant. However, the coefficient is much lower than for the other variables. IMP_VolatileAcidity has less impact on the final prediction value. The 95% confidence limit range is also wide. The confidence range is almost as large as the magnitude of the intercept itself. IMP_VolatileAcidity may not be required in the model.



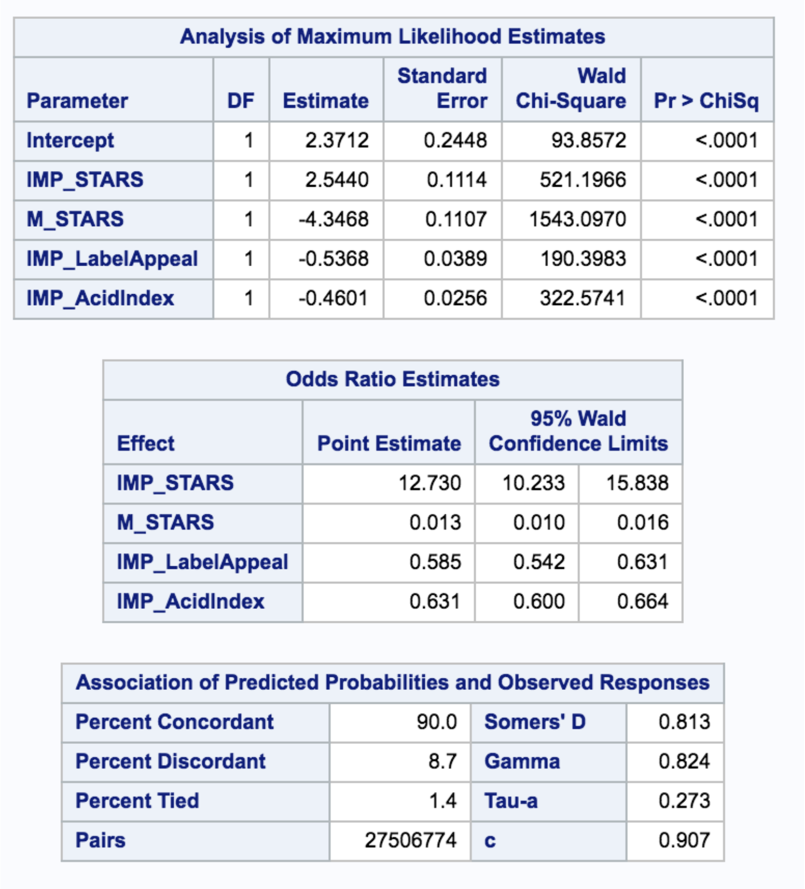
Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.5272	0.0424	1.4442	1.6102	1300.33	<.0001
IMP_STARS	1	0.1940	0.0061	0.1822	0.2059	1027.34	<.0001
M_STARS	1	-1.0368	0.0170	-1.0700	-1.0035	3739.03	<.0001
IMP_LabelAppeal	1	0.1779	0.0073	0.1636	0.1922	594.25	<.0001
IMP_AcidIndex	1	-0.0858	0.0052	-0.0959	-0.0756	275.60	<.0001
IMP_VolatileAcidity	1	-0.0377	0.0076	-0.0525	-0.0228	24.81	<.0001
Dispersion	1	0.0000	0.0002	0.0000	1.5E287		

Table 14: Negative Binomial Model

Model 4: Hurdle (first version): BINGO BONUS

A hurdle model is the combination of two different models. The first model is a Logistic Regression to model the probability of the TARGET variable being non-zero. A TARGET_FLAG variable is created for this purpose. When TARGET is non-zero, TARGET_FLAG is set to one. The Logistic Regression equation predicts the probability of TARGET_FLAG being 1. The results for the Logistic Regression model are shown below.

One problem with this model is that the sign of IMP_LabelAppeal is opposite of the expected direction. As the Label Appeal increases, the probability of buying should increase (not decrease). The correlation coefficient shows that this variable is positively correlated with TARGET. Some multicollinearity with IMP_STARS and M_STARS may be causing this result.



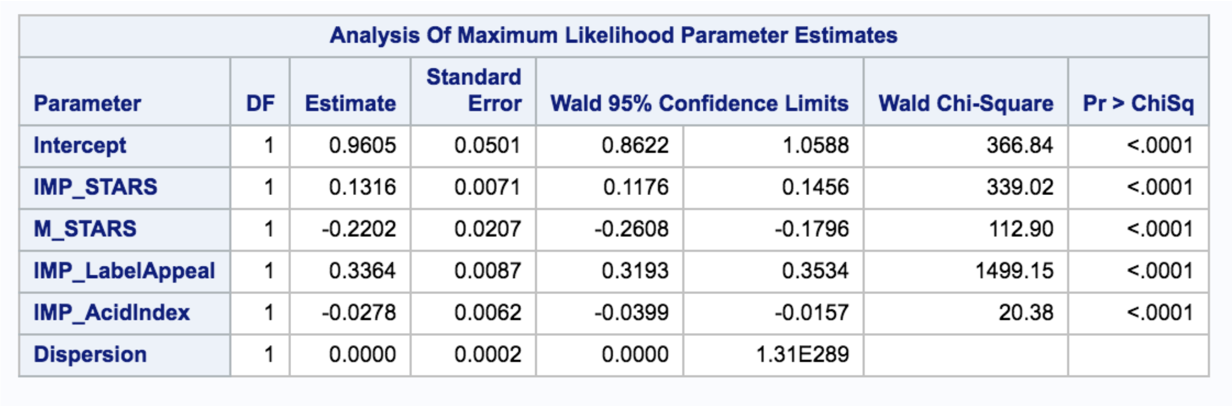
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.3712	0.2448	93.8572	<.0001
IMP_STARS	1	2.5440	0.1114	521.1966	<.0001
M_STARS	1	-4.3468	0.1107	1543.0970	<.0001
IMP_LabelAppeal	1	-0.5368	0.0389	190.3983	<.0001
IMP_AcidIndex	1	-0.4601	0.0256	322.5741	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
IMP_STARS	12.730	10.233	15.838
M_STARS	0.013	0.010	0.016
IMP_LabelAppeal	0.585	0.542	0.631
IMP_AcidIndex	0.631	0.600	0.664

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	90.0	Somers' D	0.813
Percent Discordant	8.7	Gamma	0.824
Percent Tied	1.4	Tau-a	0.273
Pairs	27506774	c	0.907

Table 15: Logistic Regression portion of first Hurdle model

The second portion of the hurdle model is a Poisson model to predict the number of cases of wine sold, given a that a purchase event occurs. A variable called TARGET_AMT is created that provides the number of nonzero cases sold. If cases sold is zero, TARGET_AMT is set to missing. The Poisson regression for predicting TARGET_AMT is shown below.



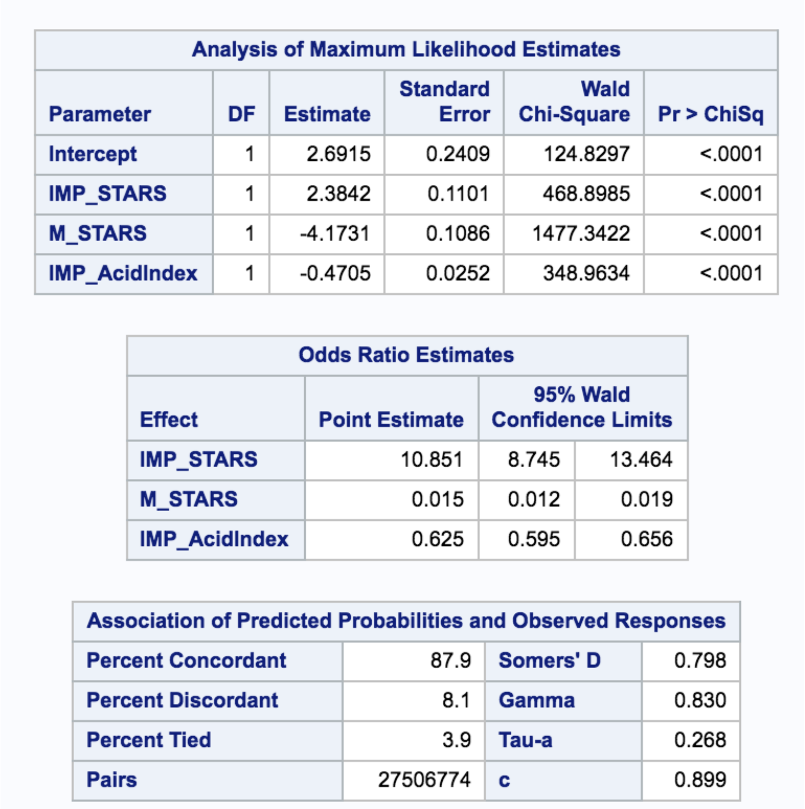
Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.9605	0.0501	0.8622	1.0588	366.84	<.0001
IMP_STARS	1	0.1316	0.0071	0.1176	0.1456	339.02	<.0001
M_STARS	1	-0.2202	0.0207	-0.2608	-0.1796	112.90	<.0001
IMP_LabelAppeal	1	0.3364	0.0087	0.3193	0.3534	1499.15	<.0001
IMP_AcidIndex	1	-0.0278	0.0062	-0.0399	-0.0157	20.38	<.0001
Dispersion	1	0.0000	0.0002	0.0000	1.31E289		

Table 16: Poisson Regression portion of Hurdle model

The results from the two models are multiplied together to provide the final prediction.

Model 5: Hurdle (second version): BINGO BONUS

Another Hurdle model was created using the Poisson Regression model from above to predict TARGET AMT and a new Logistic Regression model to predict TARGET_FLAG. This model did not include IMP_LabelAppeal. The signs for the coefficients are all as expected. The magnitude of the intercept and remaining coefficients was not greatly impacted by the removal of IMP_LabelAppeal. The model is a strong predictive model with 87.9% of the pairs concordant.



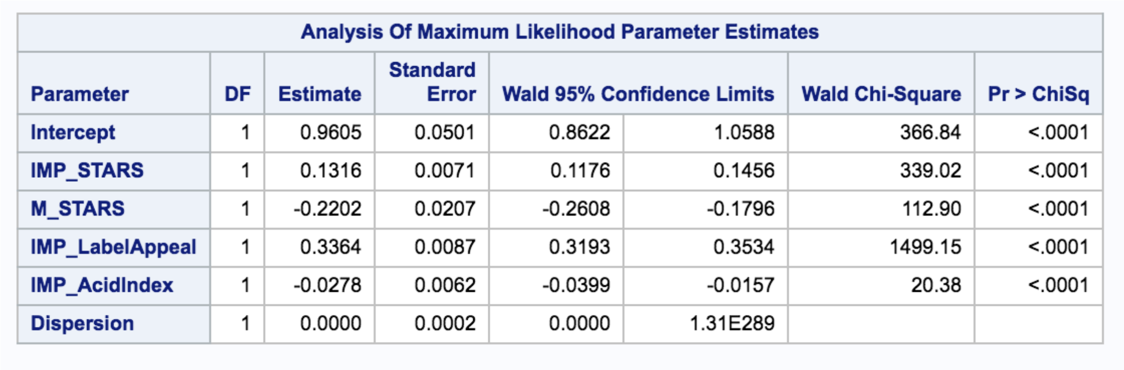
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.6915	0.2409	124.8297	<.0001
IMP_STARS	1	2.3842	0.1101	468.8985	<.0001
M_STARS	1	-4.1731	0.1086	1477.3422	<.0001
IMP_AcidIndex	1	-0.4705	0.0252	348.9634	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
IMP_STARS	10.851	8.745	13.464
M_STARS	0.015	0.012	0.019
IMP_AcidIndex	0.625	0.595	0.656

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	87.9	Somers' D	0.798
Percent Discordant	8.1	Gamma	0.830
Percent Tied	3.9	Tau-a	0.268
Pairs	27506774	c	0.899

Table 17: Logistic Regression portion of Hurdle model

The Poisson regression model to predict the TARGET_AMT portion remains the same.



Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.9605	0.0501	0.8622	1.0588	366.84	<.0001
IMP_STARS	1	0.1316	0.0071	0.1176	0.1456	339.02	<.0001
M_STARS	1	-0.2202	0.0207	-0.2608	-0.1796	112.90	<.0001
IMP_LabelAppeal	1	0.3364	0.0087	0.3193	0.3534	1499.15	<.0001
IMP_AcidIndex	1	-0.0278	0.0062	-0.0399	-0.0157	20.38	<.0001
Dispersion	1	0.0000	0.0002	0.0000	1.31E289		

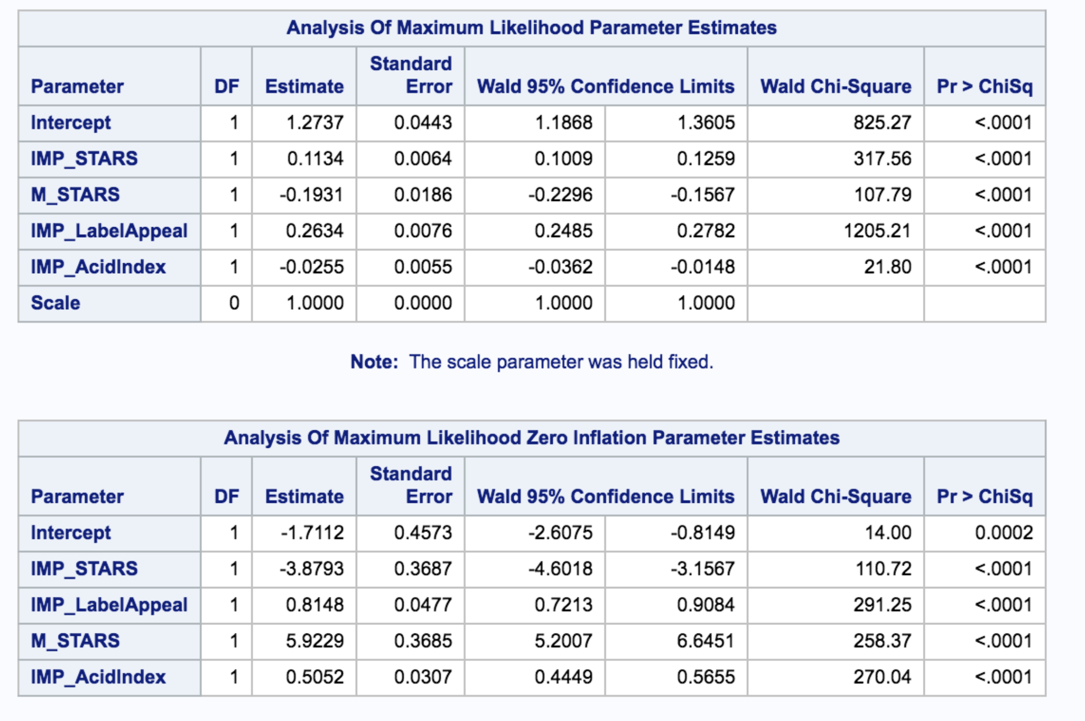
Table 18: Poisson Regression portion of Hurdle model

Model 5 is a stronger model than Model 4 because the direction of each coefficient is explainable and sensible. Model 5 is a more useful model.

Model 6: Zero Inflated Poisson

The Zero Inflated Poisson model is a better alternative to modeling counting variables with a distribution like that of TARGAT. ZIP models have the attributes of a general Poisson Model such as no negative prediction and no requirement for a linear relationship between independent and dependent variables. However, ZIP models are specially made for situations in which the dependent variable has a large spike in zeros.

The first table of the ZIP model predicts the amount of wine that a distributor will buy if they decide to purchase at all. This is similar to the second portion of the Hurdle model above. And the second table of the ZIP model predicts the Log Odds probability that the distributor will NOT buy. The ZIP modeling process is slightly more complicated than the previous models. However, it is a model that is specially tailored to the kind of distribution seen in the TARGET variable.



Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2737	0.0443	1.1868	1.3605	825.27	<.0001
IMP_STARS	1	0.1134	0.0064	0.1009	0.1259	317.56	<.0001
M_STARS	1	-0.1931	0.0186	-0.2296	-0.1567	107.79	<.0001
IMP_LabelAppeal	1	0.2634	0.0076	0.2485	0.2782	1205.21	<.0001
IMP_AcidIndex	1	-0.0255	0.0055	-0.0362	-0.0148	21.80	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

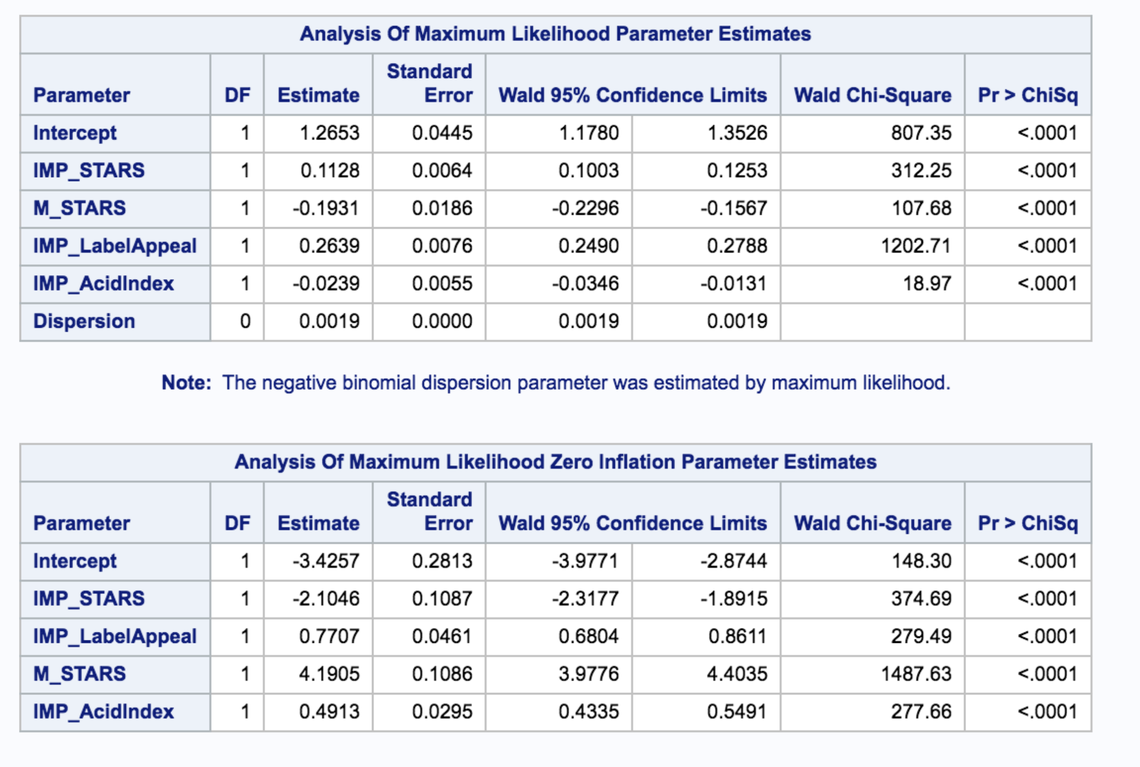
Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.7112	0.4573	-2.6075	-0.8149	14.00	0.0002
IMP_STARS	1	-3.8793	0.3687	-4.6018	-3.1567	110.72	<.0001
IMP_LabelAppeal	1	0.8148	0.0477	0.7213	0.9084	291.25	<.0001
M_STARS	1	5.9229	0.3685	5.2007	6.6451	258.37	<.0001
IMP_AcidIndex	1	0.5052	0.0307	0.4449	0.5655	270.04	<.0001

Table 19: ZIP model

Model 7: Zero Inflated Negative Binomial

The ZINB is similar to the ZIP model. The only difference is that it is based on a Negative Binomial Regression rather than a Poisson regression. The results from the first portion of the ZINB, which is used to predict the value of TARGET (if a purchase is made) is very similar to the previous problem. However, the second portion that predicts the probability that someone will NOT buy is significantly different than for the ZIP model. The magnitude of the intercept and coefficients is twice that of the previous model. Changes in the predictor values have a greater impact in this model than in the previous ZIP version.



Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2653	0.0445	1.1780	1.3526	807.35	<.0001
IMP_STARS	1	0.1128	0.0064	0.1003	0.1253	312.25	<.0001
M_STARS	1	-0.1931	0.0186	-0.2296	-0.1567	107.68	<.0001
IMP_LabelAppeal	1	0.2639	0.0076	0.2490	0.2788	1202.71	<.0001
IMP_AcidIndex	1	-0.0239	0.0055	-0.0346	-0.0131	18.97	<.0001
Dispersion	0	0.0019	0.0000	0.0019	0.0019		

Note: The negative binomial dispersion parameter was estimated by maximum likelihood.

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.4257	0.2813	-3.9771	-2.8744	148.30	<.0001
IMP_STARS	1	-2.1046	0.1087	-2.3177	-1.8915	374.69	<.0001
IMP_LabelAppeal	1	0.7707	0.0461	0.6804	0.8611	279.49	<.0001
M_STARS	1	4.1905	0.1086	3.9776	4.4035	1487.63	<.0001
IMP_AcidIndex	1	0.4913	0.0295	0.4335	0.5491	277.66	<.0001

Table 18: ZINB model

Model 8: Ensemble

The ensemble model is an average of the results from the four different kinds of models: Linear Regression, Poisson, Hurdle, and Zero Inflated Poisson, as shown in the equation below.

```
P_ENSEMBLE = (P_REGRESSION +
              P_GENMOD_POI +
              P_HURDLE_2   +
              P_GENMOD_ZIP)/4;
```

Taking an average of the four models can be used to minimize the error from any one model. Each model has unique benefits. Thus, a combination model can capture aspects of each model.

Model 9: Decision Tree: BINGO BONUS

A decision tree model was also created to predict the number of wine cases sold. The full decision tree from Enterprise Miner is shown below.

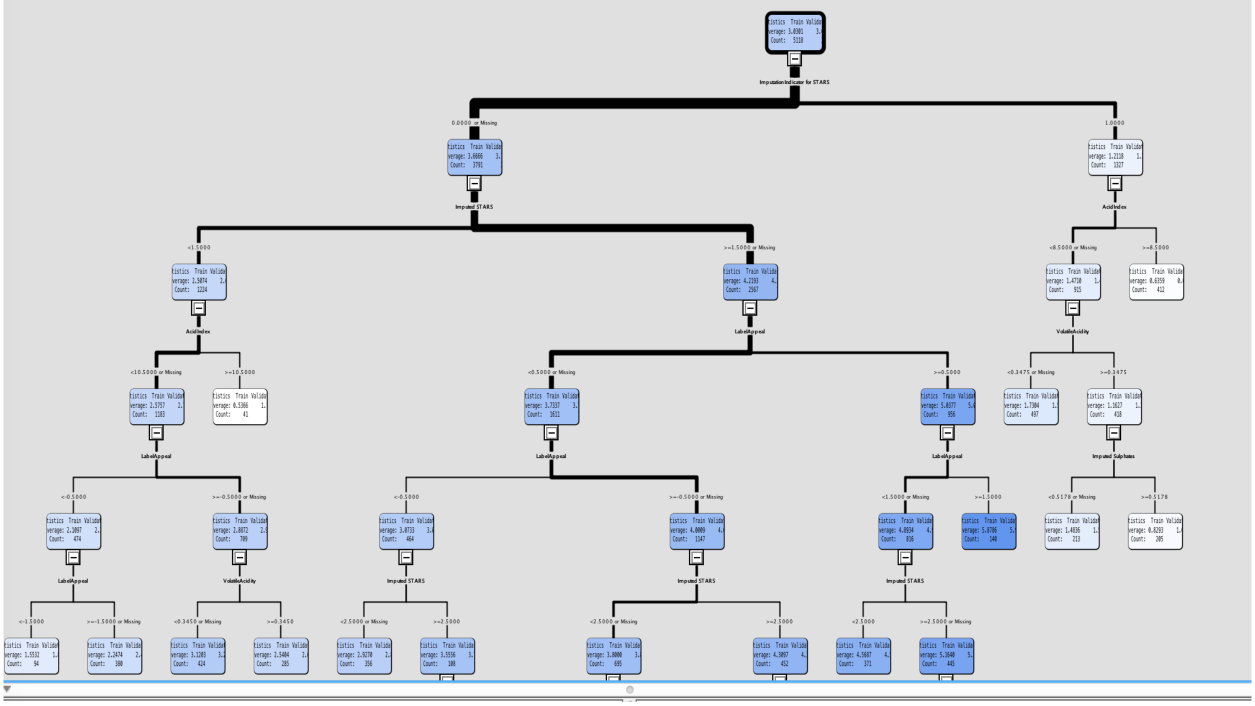


Figure 10: Decision Tree

A pruned tree is also included to the show the most important branches of the decision tree. The results from the decision tree are similar to the results from the various regression models tested. The tree verifies the importance of the predictor variables selected in the other models.

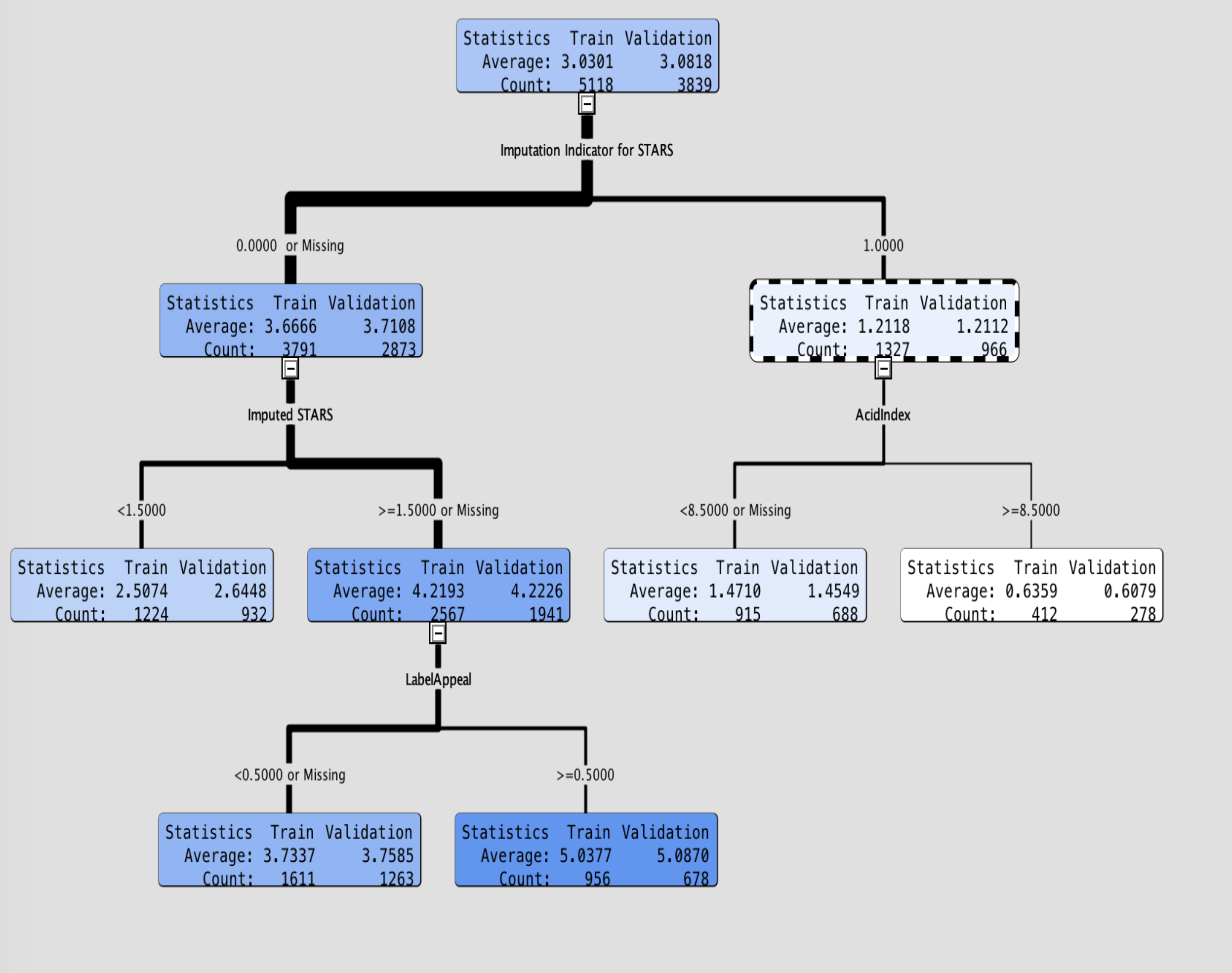
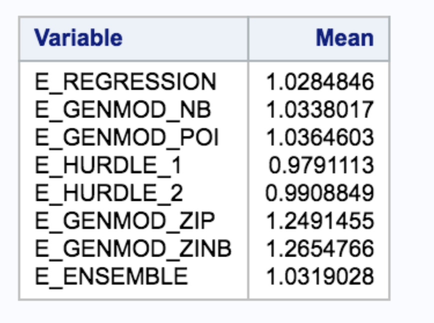


Figure 11: Pruned Decision Tree

Model Selection

The mean error for each model is listed below. The Hurdle models have the lowest error rates. The first hurdle model (Model 4) has the lowest mean absolute error at 0.979. However, this model was problematic because of the sign for IMP_LabelAppeal in the Logistic Regression part of the model. The second hurdle model (Model 5) is a better choice because the signs for the coefficients are all correct. The mean absolute error rate for the second hurdle model (Model 5) is only slightly greater at 0.991.



Variable	Mean
E_REGRESSION	1.0284846
E_GENMOD_NB	1.0338017
E_GENMOD_POI	1.0364603
E_HURDLE_1	0.9791113
E_HURDLE_2	0.9908849
E_GENMOD_ZIP	1.2491455
E_GENMOD_ZINB	1.2654766
E_ENSEMBLE	1.0319028

Table 19: Mean error for each model

The full model is shown below. This model, Model 5, is the champion model.

```
P_LOGIT_PROB = 2.6915                            +
                IMP_STARS         *(2.3842)    +
                M_STARS           *(-4.1731)   +
                IMP_AcidIndex     *(-0.4705)
                ;
if P_LOGIT_PROB > 1000 then P_LOGIT_PROB = 1000;
if P_LOGIT_PROB < -1000 then P_LOGIT_PROB = -1000;
P_LOGIT_PROB = exp(P_LOGIT_PROB) / (1+exp(P_LOGIT_PROB));
```

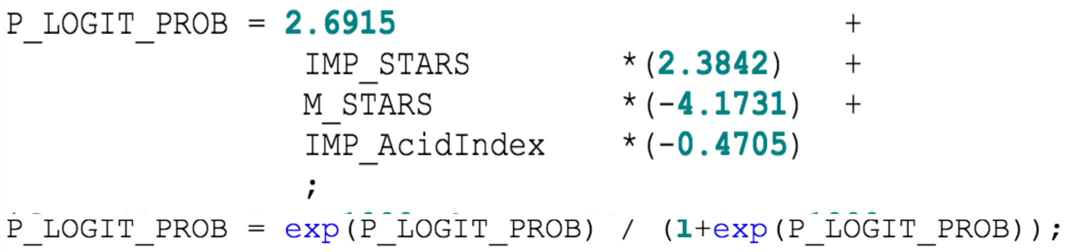
```
P_GENMOD_HURDLE =
                0.9605                         +
                IMP_STARS         *(0.1316)   +
                M_STARS           *(-0.2202)  +
                IMP_LabelAppeal   *(0.3364)   +
                IMP_AcidIndex     *(-0.0278)
                ;
P_GENMOD_HURDLE = exp(P_GENMOD_HURDLE);
```

```
P_HURDLE_2 = P_LOGIT_PROB * (P_GENMOD_HURDLE+1);
```

Conclusion

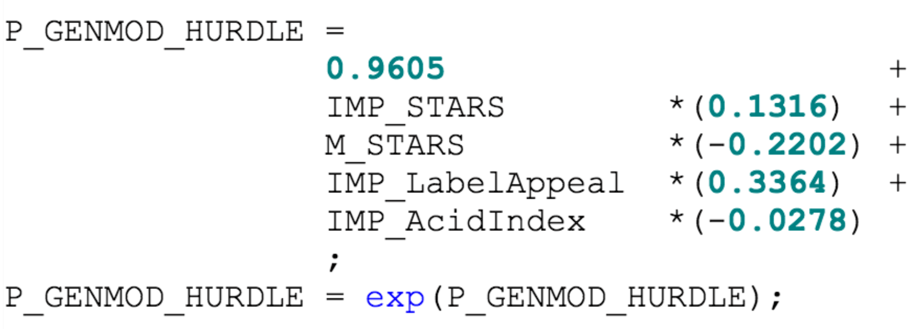
The aim of this study was to predict the number of wine cases of a particular wine that a wine distributor would purchase after sampling the wine. The primary aim was to understand what attributes of the wine and bottle led to high sales. Various predicative models were built to answer this question.

The champion model selected is a hurdle model. This type of model is actually the combination of two models. The first model is a Logistic Regression that predicts the probability that wine will be purchased. This model is:



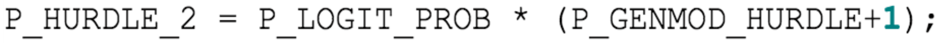
```
P_LOGIT_PROB = 2.6915                            +
                IMP_STARS         *(2.3842)    +
                M_STARS           *(-4.1731)   +
                IMP_AcidIndex     *(-0.4705)
                ;
P_LOGIT_PROB = exp(P_LOGIT_PROB) / (1+exp(P_LOGIT_PROB));
```

The second model is a Poisson model that predicts the value of the wine sales given that a purchase decision is made. This model is:



```
P_GENMOD_HURDLE =
                0.9605                        +
                IMP_STARS          *(0.1316)  +
                M_STARS            *(-0.2202) +
                IMP_LabelAppeal    *(0.3364)  +
                IMP_AcidIndex      *(-0.0278)
                ;
P_GENMOD_HURDLE = exp(P_GENMOD_HURDLE);
```

The results from the two models are then multiplied together to provide the final prediction of the number of wine cases sold.

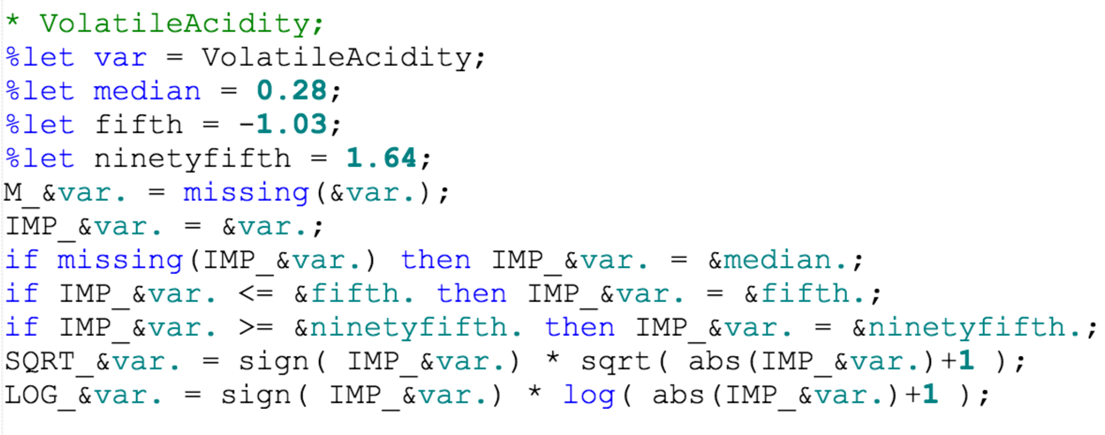


```
P_HURDLE_2 = P_LOGIT_PROB * (P_GENMOD_HURDLE+1);
```

The inputs to the model shows that the most important predictors of wine case sales are the number of Stars a wine receives (or the absence of Stars), the appeal of the label and the acidity of the wine. The data contained many chemical properties of the wine that had little impact on the number of cases sold. If a wine manufacturer is looking to increase wine sales, it would likely be a good idea to focus on getting as many wines as possible to be reviewed and to improve the quality of the label aesthetics. When considering the actual chemical composition of the wine, the most important consideration is acidity. Higher levels of acidity decrease wine sales.

BINGO BONUS: Macros

Example of Macros used in program:



```
* VolatileAcidity;
%let var = VolatileAcidity;
%let median = 0.28;
%let fifth = -1.03;
%let ninetyfifth = 1.64;
M_&var. = missing(&var.);
IMP_&var. = &var.;
if missing(IMP_&var.) then IMP_&var. = &median.;
if IMP_&var. <= &fifth. then IMP_&var. = &fifth.;
if IMP_&var. >= &ninetyfifth. then IMP_&var. = &ninetyfifth.;
SQRT_&var. = sign( IMP_&var.) * sqrt( abs(IMP_&var.)+1 );
LOG_&var. = sign( IMP_&var.) * log( abs(IMP_&var.)+1 );
```