A Poisson Regression Prediction Model for the Number of Wine Cases Purchased

Julia Barnhart

Northwestern University

Introduction

The purpose of this study was to produce a statistical model that predicts the number of wine cases a wine distribution company will purchase after sampling a particular wine. This model enables wine manufacturers to gauge how many wine cases of a particular commercially-available wine will be ordered. This would allow them to make more accurate adjustments to their budget, sales, and marketing plans. The candidate models utilize available characteristics, mostly chemical properties of the wine, and were built using the popular statistical technique of Poisson and the Negative Binomial regression. The dataset itself was properly analyzed and cleaned using standard techniques in order to enhance the predictive power of the model. It was also randomly split into a separate training dataset (70% of the observations) and a test dataset (30% of the observations), which allowed for the predictive power of the model to be tested before final deployment. The model is reusable in future seasons, as additionally generated observations can be incorporated back into the dataset used to build the model, and the model itself can be adjusted accordingly.

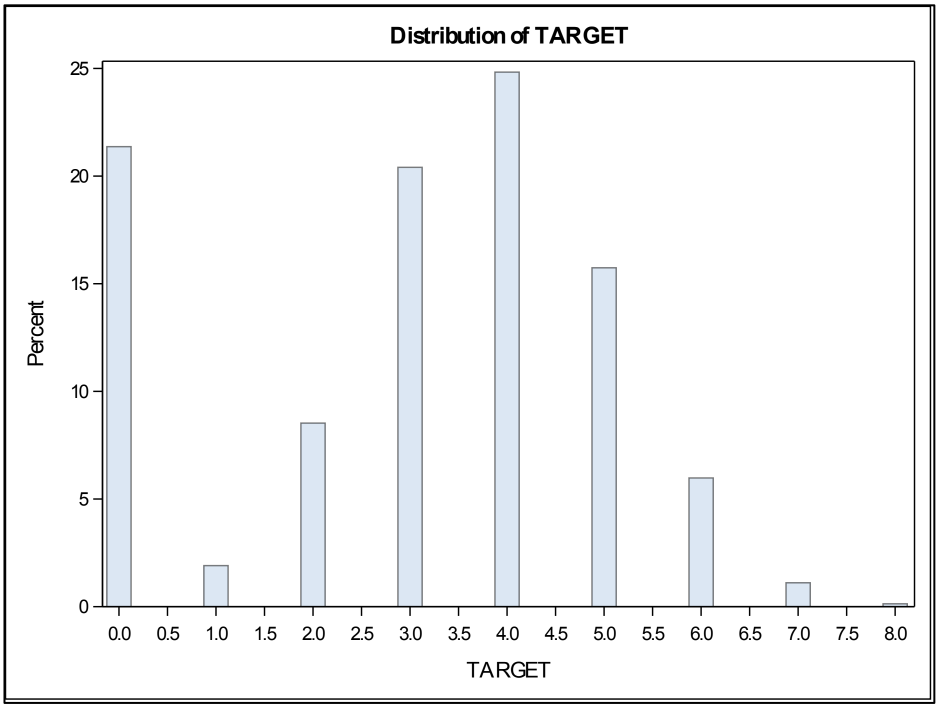
Data Exploration

Exploratory data analysis was performed in order to better understand the general characteristics of the available wine dataset and ascertain which of the 16 variables would make viable candidates as predictor variables in the Poisson and Negative Binomial Regression models. It was also performed to better understand the characteristics of the response variable, TARGET. The dataset contained 12,795 observations with 16 variables. The variable INDEX was immediately removed from predictor-variable candidacy, as it only serves an indexing role. The variable TARGET was designated as the response variable and therefore also removed from predictor-variable candidacy. All the variables are of type number, including the response variable. Many of the variables are chemical properties of the wine. However, the variable LabelAppeal is correlated to the overall appeal of a wine label on the wine bottle. A high number would suggest, according to marketing studies, high customer appeal. It could be postulated that distributors are more likely to purchase cases of wine with high label appeal. Similarly, the variable STARS represents a wine's rating on a scale of one to four, according to wine experts, with four being the highest. In general, a high number of stars correlate to higher sales. The positive effect of these two variables on the response variable may help gauge the model to its theoretical underpinnings during model selection. The available data dictionary of the variables is shown in Table 1 below.

#	Variable	Type	Definition
15	AcidIndex	Num	Proprietary method of testing total acidity of wine by using a weighted average
13	Alcohol	Num	Alcohol Content
7	Chlorides	Num	Chloride content of wine
5	CitricAcid	Num	Citric Acid Content
10	Density	Num	Density of Wine
3	FixedAcidity	Num	Fixed Acidity of Wine
8	FreeSulfurDioxide	Num	Sulfur Dioxide content of wine
1	INDEX	Num	Index
14	LabelAppeal	Num	Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design.
6	ResidualSugar	Num	Residual Sugar of wine
16	STARS	Num	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor; A high number of stars suggests high sales
12	Sulphates	Num	Sulfate content of wine
2	TARGET	Num	Number of Cases Purchased
9	TotalSulfurDioxide	Num	Total Sulfur Dioxide of Wine
4	VolatileAcidity	Num	Volatile Acid content of wine
11	pH	Num	pH of wine

Table 1: Data dictionary of all variables available in the wine dataset.

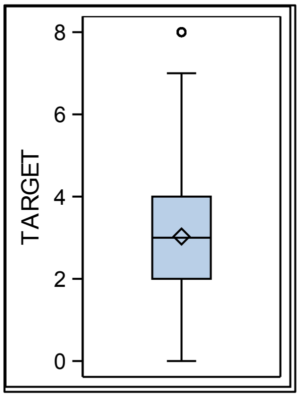
The response variable itself ranges from the values zero to eight, indicating that a distribution company can buy anywhere from zero to eight cases of a particular wine after a wine-tasting session. Theoretically, the upper bound is unbounded, while the lower bound cannot be less than zero (as realistically one cannot purchase a negative number of wine cases). However, the models built can only be used to predict the TARGET in this range. Additionally, only whole cases of wines can be purchased. Thus, the response variable is a discrete count. This property of the response variable makes Poisson and Negative Binomial regression models especially appropriate, although in this case, as the response variable can also be theoretically modeled as a continuous variable and the sample size is large, a simple OLS regression model may viable as well (Hoffmann, 2004, p. 105). The distribution of the response variable is displayed in Figure 1 below.



Graph 1: Histogram of the distribution of the response variable TARGET

The histogram above of the response variable clearly displays a significant concentration of values at zero. Eliminating the values at zero would enable the response variable to follow a normal distribution. In general, this may signify that there are two target populations: The box-and-whisker plot in Graph 2 below indicates that the value of eight may be an extreme observation. The mean number of wine cases bought is 3.029, the standard deviation is 1.936, and the variance is 3.711. In general, a Poisson distribution requires for the variance to equal the mean. However, a Negative

Binomial distribution requires that the variance is greater than the mean, which is the case here. This is something to take into consideration during model building and model selection. There are also no missing values for the response variable, so no observations were required to be deleted based on this criterion.



Graph 2: Box-and-Whisker plot of response variable TARGET.

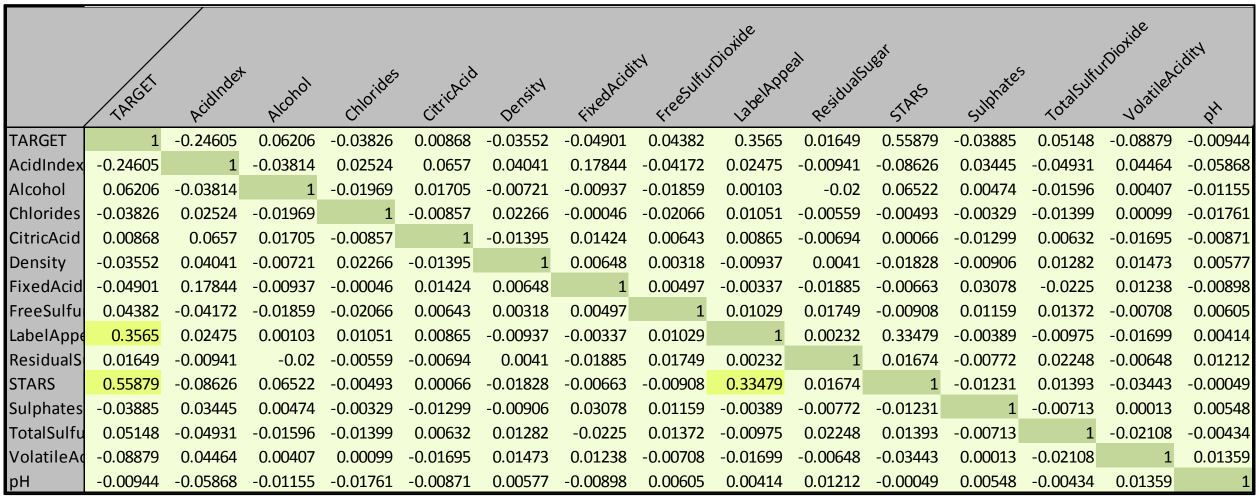
The basic statistics of the 15 candidate predictor variables were also analyzed, and displayed in Table 2 below. In general, eight candidate predictor variables had missing values, as marked in red under the column labeled “% Missing”. The most notable predictor variable was STARS, with 26.25% of the observations missing values. The predictor variables Alcohol, Chlorides, FreeSulfurDioxide, ResidualSugar, STARS, Sulphates, TotalSulfurDioxide, and pH were also marked as containing missing values. The percent of observations with missing values for these candidate predictor variables were under 10, and thus not initially alarming. As the SAS procedure PROC GENMOD (used to build the models) ignores any observations with a missing value for any variable involved in the model, these missing values pose a concern for the Poisson and Negative Binomial Regression. Usually observations with missing values are either removed altogether from the training and testing datasets, or are filled-in, depending on the appropriateness of the technique and the domain context.

Variable	N	N Miss	% Missing	Mean	Std Dev	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Minimum	Maximum
AcidIndex	12795	0	0	7.773	1.324	6.000	6.000	10.000	13.000	4.000	17.000
Alcohol	12142	653	5.10	10.489	3.728	0.100	4.100	16.700	20.300	-4.700	26.500
Chlorides	12157	638	4.99	0.055	0.318	-0.859	-0.489	0.598	0.957	-1.171	1.351
CitricAcid	12795	0	0	0.308	0.862	-2.180	-1.160	1.790	2.660	-3.240	3.860
Density	12795	0	0	0.994	0.027	0.917	0.949	1.040	1.070	0.888	1.099
FixedAcidity	12795	0	0	7.076	6.318	-10.900	-3.600	17.800	24.400	-18.100	34.400
FreeSulfurDioxide	12148	647	5.06	30.846	148.715	-388.000	-224.000	284.000	469.000	-555.000	623.000
LabelAppeal	12795	0	0	-0.009	0.891	-2.000	-1.000	1.000	2.000	-2.000	2.000
ResidualSugar	12179	616	4.81	5.419	33.749	-91.000	-52.701	62.701	99.201	-127.801	141.151
STARS	9436	3359	26.25	2.0418	0.903	1.000	1.000	4.000	4.000	1.000	4.000
Sulphates	11585	1210	9.46	0.5271	0.932	-2.130	-1.050	2.090	3.160	-3.130	4.240
TotalSulfurDioxide	12113	682	5.63	120.714	231.913	-531.000	-273.000	514.000	767.000	-823.000	1057.000
VolatileAcidity	12795	0	0	0.324	0.784	-1.865	-1.030	1.640	2.590	-2.790	3.680
pH	12400	395	3.09	3.208	0.680	1.320	2.060	4.370	5.125	0.480	6.130

Table 2: The number of missing values, mean, standard deviation, and percentiles of each candidate predictor variable.

Furthermore, the histograms of all the 15 variables were visually analyzed for overall distribution. The graphs are displayed in Appendix A. None of the distributions of the variables follow a normal one, with many values clustering starkly around the means. The variables AcidIndex, Alcohol, Chlorides, CitricAcid, Density, FixedAcidity, FreeSulfurDioxide, ResidualSugar, Sulphates, TotalSulfurDioxide, VolatileAcidity, and pH are continuous in nature, whereas the variables LabelAppeal and STARS are discrete (integers). The variable LabelAppeal can take on the values from the set {-2,-1,0,1,2}. The values of STARS can take on the values from the set {1,2,3,4}. The predictor variables Alcohol, Chlorides, CitricAcid, Density, FixedAcidity, FreeSulfurDioxide, ResidualSugar, Sulphates, TotalSulfurDioxide, VolatileAcidity, and pH may also have some extreme observations.

Each of the 15 candidate predictor variables were also measured for correlation with the response variable and with one another. Table 3 below displays the Pearson Product Momentum Correlation Coefficients. The correlation between LabelAppeal and the response variable (0.3563), STARS and the response variable (0.55879), and STARS and LabelAppeal (0.33479) seem to be very moderate. These are marked in yellow below. There seems to be no initial indicators for multicollinearity, as no predictor variable is highly correlated with another.

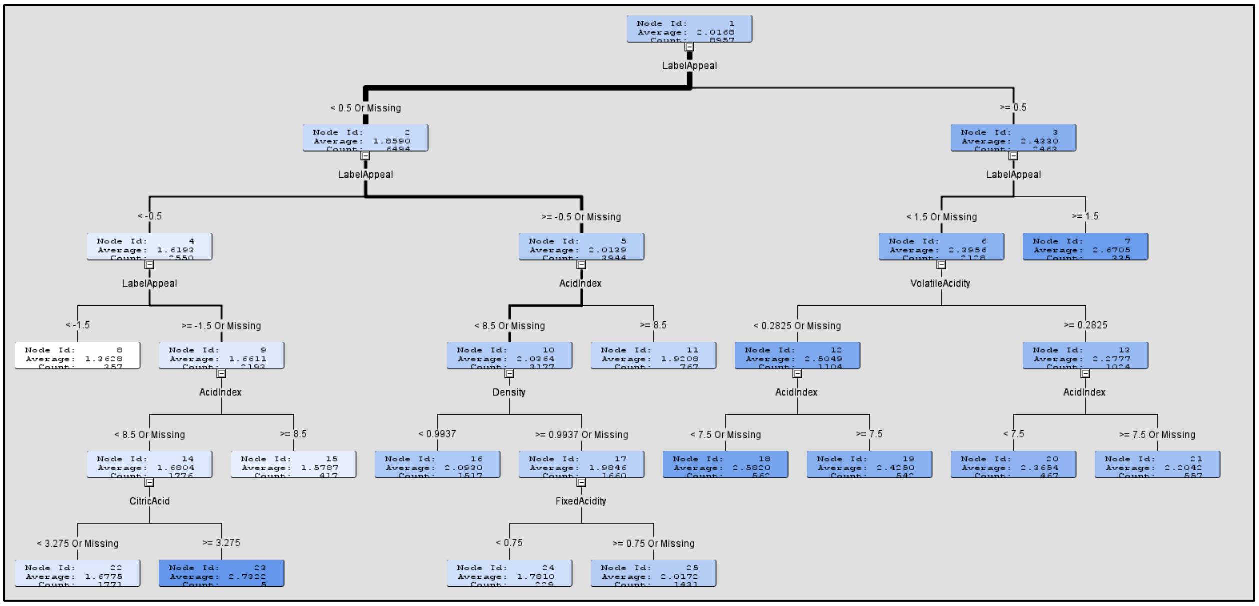


	TARGET	AcidIndex	Alcohol	Chlorides	CitricAcid	Density	FixedAcidity	FreeSulfurDioxide	LabelAppeal	ResidualSugar	STARS	Sulphates	TotalSulfurDioxide	VolatileAcidity	pH
TARGET	1	-0.24605	0.06206	-0.03826	0.00868	-0.03552	-0.04901	0.04382	0.3565	0.01649	0.55879	-0.03885	0.05148	-0.08879	-0.00944
AcidIndex	-0.24605	1	-0.03814	0.02524	0.0657	0.04041	0.17844	-0.04172	0.02475	-0.00941	-0.08626	0.03445	-0.04931	0.04464	-0.05868
Alcohol	0.06206	-0.03814	1	-0.01969	0.01705	-0.00721	-0.00937	-0.01859	0.00103	-0.02	0.06522	0.00474	-0.01596	0.00407	-0.01155
Chlorides	-0.03826	0.02524	-0.01969	1	-0.00857	0.02266	-0.00046	-0.02066	0.01051	-0.00559	-0.00493	-0.00329	-0.01399	0.00099	-0.01761
CitricAcid	0.00868	0.0657	0.01705	-0.00857	1	-0.01395	0.01424	0.00643	0.00865	-0.00694	0.00066	-0.01299	0.00632	-0.01695	-0.00871
Density	-0.03552	0.04041	-0.00721	0.02266	-0.01395	1	0.00648	0.00318	-0.00937	0.0041	-0.01828	-0.00906	0.01282	0.01473	0.00577
FixedAcid	-0.04901	0.17844	-0.00937	-0.00046	0.01424	0.00648	1	0.00497	-0.00337	-0.01885	-0.00663	0.03078	-0.0225	0.01238	-0.00898
FreeSulfu	0.04382	-0.04172	-0.01859	-0.02066	0.00643	0.00318	0.00497	1	0.01029	0.01749	-0.00908	0.01159	0.01372	-0.00708	0.00605
LabelAppe	0.3565	0.02475	0.00103	0.01051	0.00865	-0.00937	-0.00337	0.01029	1	0.00232	0.33479	-0.00389	-0.00975	-0.01699	0.00414
ResidualS	0.01649	-0.00941	-0.02	-0.00559	-0.00694	0.0041	-0.01885	0.01749	0.00232	1	0.01674	-0.00772	0.02248	-0.00648	0.01212
STARS	0.55879	-0.08626	0.06522	-0.00493	0.00066	-0.01828	-0.00663	-0.00908	0.33479	0.01674	1	-0.01231	0.01393	-0.03443	-0.00049
Sulphates	-0.03885	0.03445	0.00474	-0.00329	-0.01299	-0.00906	0.03078	0.01159	-0.00389	-0.00772	-0.01231	1	-0.00713	0.00013	0.00548
TotalSulfu	0.05148	-0.04931	-0.01596	-0.01399	0.00632	0.01282	-0.0225	0.01372	-0.00975	0.02248	0.01393	-0.00713	1	-0.02108	-0.00434
VolatileAc	-0.08879	0.04464	0.00407	0.00099	-0.01695	0.01473	0.01238	-0.00708	-0.01699	-0.00648	-0.03443	0.00013	-0.02108	1	0.01359
pH	-0.00944	-0.05868	-0.01155	-0.01761	-0.00871	0.00577	-0.00898	0.00605	0.00414	0.01212	-0.00049	0.00548	-0.00434	0.01359	1

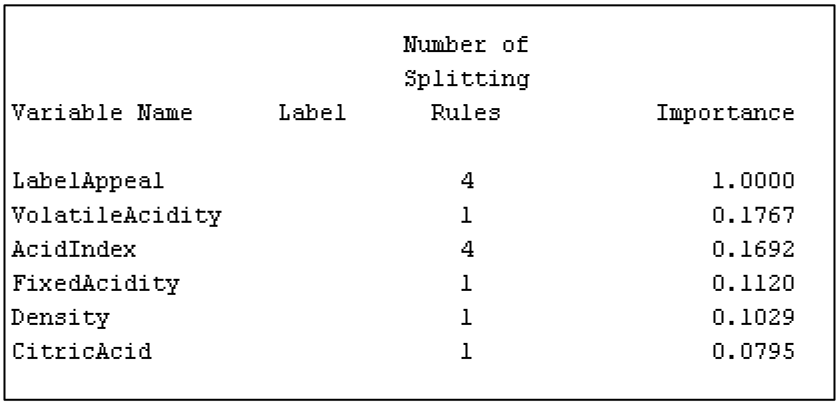
Table 3: Correlation matrix of all variables in wine dataset

Data Preparation

The exploratory data analysis identified missing values within the dataset. These missing values must be filled-in in order for the Poisson Regression and Negative Binomial model to use the observations. Decision trees were utilized in order to impute missing values with reasonable estimates, as displayed for the predictor variable STARS in Graph 3 below. The decision-tree is used to predict the best value of STARS for missing values given the values of all other predictor variables. The decision tree shows that the predictor variable LabelAppeal is the most important factor in determining the imputed value for STARS, as it is the first node. Table 4 below displays the variable important for determining the imputed values of STARS in descending order. A comprehensive display of the decision tree for each imputed variable is displayed in Appendix B.



Graph 3: Example of a decision tree built to impute missing values for predictor variable STARS



		Number of	
		Splitting	
Variable Name	Label	Rules	Importance
LabelAppeal		4	1.0000
VolatileAcidity		1	0.1767
AcidIndex		4	0.1692
FixedAcidity		1	0.1120
Density		1	0.1029
CitricAcid		1	0.0795

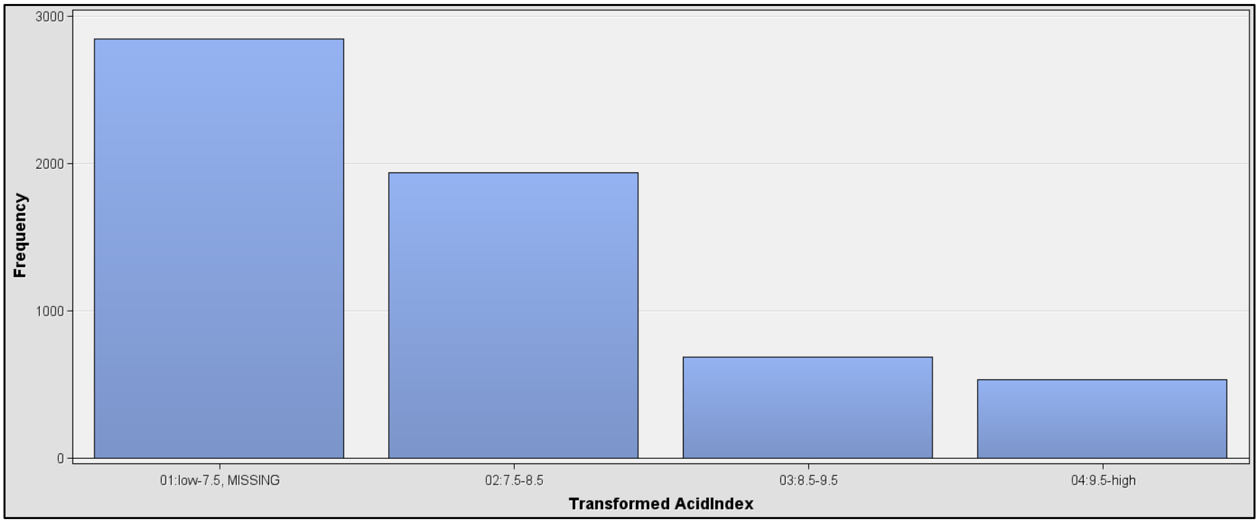
Table 4: Variables used in order of importance for imputation of missing values of STARS.

The imputation process created two new types of variables, imputed variables and missing flag variables. The imputed variables include the filled-in values, while the missing-flag variables denote the “missingness” of that particular variable. Sometimes this is a useful predictor variable in the model. This information is summarized in Table 5 below.

Original Variable	Imputed Variable	Missing Indicator	Transformed Variable
AcidIndex	x	x	OPT_AcidIndex
Alcohol	IMP_Alcohol	M_Alcohol	OPT_IMP_Alcohol
Chlorides	IMP_Chlorides	M_Chlorides	OPT_IMP_Chlorides
CitricAcid	x	x	OPT_CitricAcid
Density	x	x	OPT_Density
FixedAcidity	x	x	OPT_FixedAcidity
FreeSulfurDioxide	IMP_FreeSulfurDioxide	M_FreeSulfurDioxide	OPT_IMP_FreeSulfurDioxide
LabelAppeal	x	x	x
ResidualSugar	IMP_ResidualSugar	M_ResidualSugar	OPT_IMP_ResidualSugar
STARS	IMP_STARS	M_STARS	OPT_IMP_STARS
Sulphates	IMP_Sulphates	M_Sulphates	OPT_IMP_Sulphates
TotalSulfurDioxide	IMP_TotalSulfurDioxide	M_TotalSulfurDioxide	OPT_IMP_TotalSulfurDioxide
VolatileAcidity	x	x	OPT_VolatileAcidity
pH	IMP_pH	M_pH	OPT_IMP_pH

Table 5: Original predictor variables, imputed variables, missing-indicator variables, and transformed variables.

Lastly, SAS Enterprise Miner was used to optimally transform each variable using the “best” method, which includes transforming each predictor variable with functions such as squaring and binning and then selecting the best result using the R-squared metric. According to the result, the best transformation method was binning the variables using 4 bins. The transformed variables are displayed in Table 5 in the column labelled “Transformed Variable”. Graph 3 below displays the binning applied to the predictor variable AcidIndex in order to create the new variable OPT_AcidIndex. A complete display of all the binned variables can be found in Appedix D.



Graph 3: The predictor variable AcidIndex after binning.

Model Building

It is common practice to split the available dataset between a larger training dataset and a smaller testing dataset. It allows for the regression model to be built on the training dataset, while the test dataset is used to validate the model's predictive accuracy. Therefore, this dataset was split randomly into a training dataset comprising of 8,957 observations (70%) and a testing dataset comprising of 3,838 observations (30%). The regression models were built using the training dataset. Afterwards, each model was tested for predictive accuracy using the testing dataset and a preferred model was selected based on both the criteria of predictive accuracy and parsimony (least number of predictor variables).

Model 1: Handpicked Poisson Regression

The first model built was a Poisson regression model including both imputed variables and the missing flag variables. The model was completed after eight iterations, where each time the two most statistically-insignificant variables, according to the p-value of the Wald Chi-

Square statistic, were removed. This process was iterated until all remaining predictor variables were statistically significant (< 0.05). Table 7 below outlines which variables remained in each iteration. Due to the long names of the original variables, proxies were assigned and are displayed in Table 6 below. The goodness-of-fit table and maximum likelihood parameter estimates outputted during each iteration is displayed in Appendix C.

Proxy	Variable Name		Proxy	Variable Name
Var1	AcidIndex		Var12	IMP_TotalSulfurDioxide
Var2	IMP_Alcohol		Var13	VolatileAcidity
Var3	IMP_Chlorides		Var14	IMP_pH
Var4	CitricAcid		Var15	M_Alcohol
Var5	Density		Var16	M_Chlorides
Var6	FixedAcidity		Var17	M_FreeSulfurDioxide
Var7	IMP_FreeSulfurDioxide		Var18	M_ResidualSugar
Var8	LabelAppeal		Var19	M_STARS
Var9	IMP_ResidualSugar		Var20	M_Sulphates
Var10	IMP_STARS		Var21	M_TotalSulfurDioxide
Var11	IMP_Sulphates		Var22	M_pH

Table 6: Legend of the proxy names used in place of the original predictor variable names

Run	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11	Var12	Var13	Var14	Var15	Var16	Var17	Var 18	Var 19	Var 20	Var21	Var22
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
2	x	x	x	x	x		x	x	x	x	x	x	x	x	x		x	x	x	x	x	x
3	x	x	x	x	x		x	x		x	x	x	x	x	x			x	x	x	x	x
4	x	x	x	x	x		x	x		x	x	x	x	x				x	x	x		x
5	x	x	x		x		x	x		x	x	x	x	x				x	x	x		
6	x	x	x				x	x		x	x	x	x					x	x	x		

7	x	x	x				x	x		x		x	x					x	x			
8	x	x	x				x	x		x		x	x						x			

Table 7: The variables that remained in each of the eight iterations of the hand-picked Poisson regression model.

The final model contains nine predictor variables and one intercept, as displayed in Table 8 below.

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.4517	0.0487	1.3562	1.5472	887.95	<.0001
AcidIndex	1	-0.0791	0.0053	-0.0896	-0.0686	218.48	<.0001
IMP_Alcohol	1	0.0048	0.0017	0.0016	0.0081	8.36	0.0038
IMP_Chlorides	1	-0.0656	0.0196	-0.1040	-0.0272	11.19	0.0008
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.19	0.0073
LabelAppeal	1	0.1544	0.0075	0.1398	0.1690	428.89	<.0001
IMP_STARS	1	0.1769	0.0073	0.1626	0.1913	580.16	<.0001
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	8.23	0.0041
VolatileAcidity	1	-0.0286	0.0077	-0.0438	-0.0134	13.66	0.0002
M_STARS	1	-1.0028	0.0200	-1.0421	-0.9636	2507.80	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Table 8: The eight predictor variables and intercept used in the model *Handpicked Poisson*

The Poisson Regression model can be described by the following mathematical equation:

```
TARGET = exp (
               1.4517 -
               0.0791 * AcidIndex +
               0.0048 * IMP_Alcohol –
               0.0656 * IMP_Chlorides +
               0.0001 * IMP_FreeSulfurDioxide +
               0.1544 * LabelAppeal +
               0.1769 * IMP_STARS +
               0.0001 * IMP_TotalSulfurDioxide -
               0.0286 * VolatileAcidity –
               1.0028 * M_STARS
               )
```

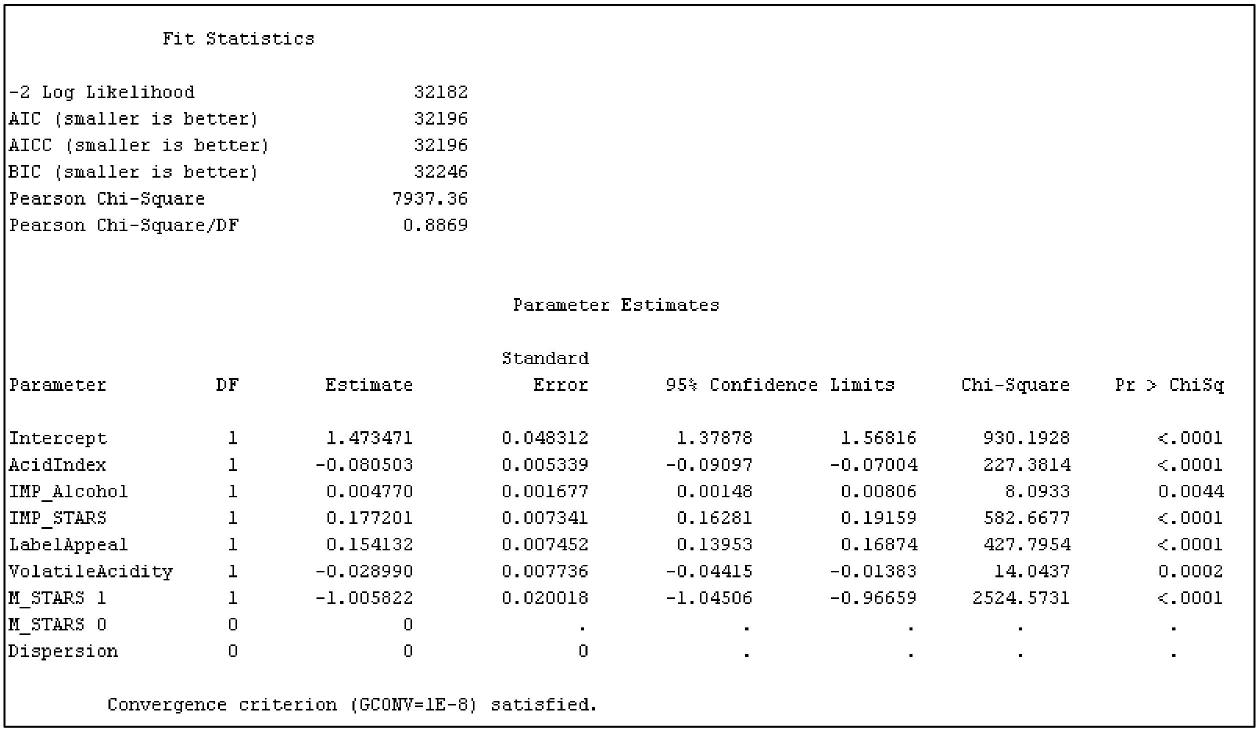
This model contains a mixture of both positive and negative coefficients. One could argue that having missing stars (M_STARS is true or "1") has a negative impact on the regression model, as the coefficient is negative. The regression parameters are interpreted as follows: for a one-unit change in the predictor variable, the difference in the logs of the expected counts of the response variable is expected to change by the respective regression coefficient, given that the other predictor variables are held constant. The model is quite parsimonious, containing only 10 predictor variables, including the intercept.

Model 2: Poisson Regression (using ANN) with Variable Transformation

This Poisson regression model was built using the following imputed, transformed variables (binned variables): M_Alcohol, M_Chlorides, M_FreeSulfurDioxide, M_ResidualSugars, M_STARS, M_Sulphates, M_TotalSulfurDioxide, M_pH, OPT_AcidINdex, OPT_CitricAcid, OPT_Density, OPT_FixedAcidity, OPT_IMP_Alcohol, OPT_IMP_Chlorides, OPT_IMP_FreeSulfurDioxide, OPT_IMP_ResidualSugar, OPT_IMP_STARS, OPT_IMP_Sulphates, OPT_IMP_TotalSulfurDioxide, OPT_IMP_pH, and OPT_VolatileAcidity. This model is difficult to describe, as it used a SAS Enterprise Miner's Neural Network to run the Poisson model. The target layer activation function was set to exponential, while the target layer error function was set to Poisson.

Model 3: SAS EM Variable Selection Negative Binomial Regression

The next model was a Negative Binomial Regression using SAS Enterprise Miner's variable selection feature in order to preselect candidate predictor variables. This feature uses the SAS procedure PROC DMINE, which is a forward stepwise selection process using R-squared, for variable selection. The final six variables chosen include: M_STARS, IMP_STARS, LabelAppeal, AcidIndex, VolatileAcidity, and IMP_Alcohol. This model should generally be used if extradisperson (where the variance is greater than the mean) is apparent (Hoffmann, 2004, p. 113). Table 8 below displays both the goodness-of-fit statistics and the maximum-likelihood parameter estimates for the model. All Wald Chi-Square statistics are statistically significant (< 0.05).



```
              Fit Statistics

-2 Log Likelihood                  32182
AIC (smaller is better)            32196
AICC (smaller is better)           32196
BIC (smaller is better)            32246
Pearson Chi-Square               7937.36
Pearson Chi-Square/DF             0.8869


                                        Parameter Estimates

                                           Standard
Parameter          DF       Estimate          Error       95% Confidence Limits      Chi-Square    Pr > ChiSq

Intercept           1       1.473471       0.048312       1.37878        1.56816       930.1928        <.0001
AcidIndex           1      -0.080503       0.005339      -0.09097       -0.07004       227.3814        <.0001
IMP_Alcohol         1       0.004770       0.001677       0.00148        0.00806         8.0933        0.0044
IMP_STARS           1       0.177201       0.007341       0.16281        0.19159       582.6677        <.0001
LabelAppeal         1       0.154132       0.007452       0.13953        0.16874       427.7954        <.0001
VolatileAcidity     1      -0.028990       0.007736      -0.04415       -0.01383        14.0437        0.0002
M_STARS 1           1      -1.005822       0.020018      -1.04506       -0.96659      2524.5731        <.0001
M_STARS 0           0              0              .             .              .              .             .
Dispersion          0              0              0             .              .              .             .

         Convergence criterion (GCONV=1E-8) satisfied.
```

Table 8: The six predictor variables and intercept used in the model *SAS EM Variable Selection Negative Binomial*

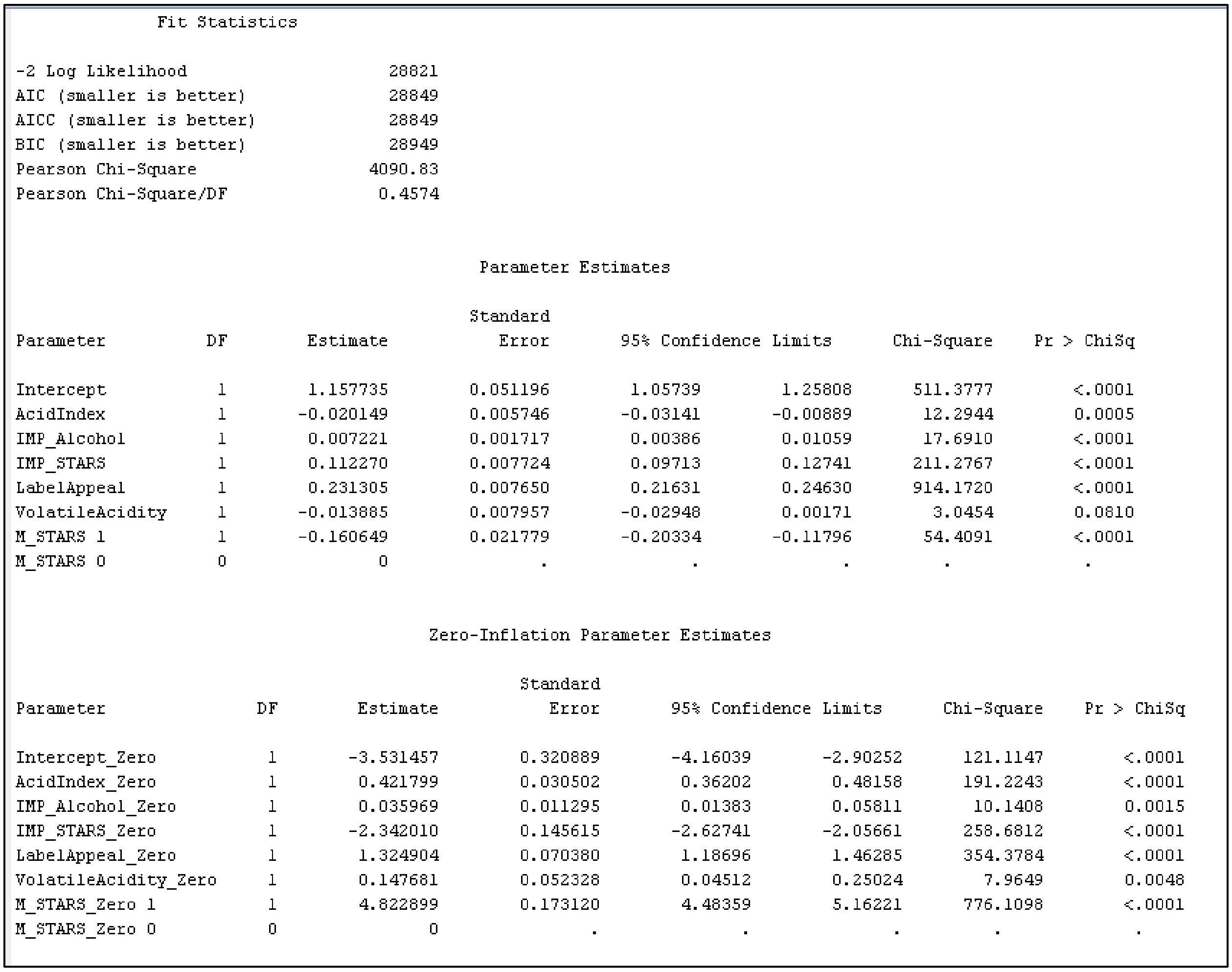
The Negative Binomial Regression Model can be described by the following mathematical equation:

```
TARGET = exp (
                    1.4735 -
                    0.0805 * AcidIndex +
                    0.0048 * IMP_Alcohol +
                    0.1772 * IMP_STARS +
                    0.1541 * LabelAppeal -
                    0.0290 * VolatileAcidity –
                    1.0058 * M_STARS
               )
```

This model contains a mixture of both positive and negative coefficients. One could argue that having missing stars (M_STARS is true or “1”) has a negative impact on the regression model, as the coefficient is negative. The regression parameters are interpreted as follows: for a one-unit change in the predictor variable, the difference in the logs of the expected counts of the response variable is expected to change by the respective regression coefficient, given that the other predictor variables are held constant. The model is quite parsimonious, containing only seven predictor variables, including the intercept.

Model 4: SAS EM Variable Selection Zero Inflated Poisson Regression

This model was built because there is an excess of zeroes that occurred in the distribution of the response variable. This model relates the information that some observations have a zero chance of occurring, while other observations differ in the count of events they experience. This would imply that there are two different mechanisms that govern the events. Relating to this particular study, some wine cases may not even be sold, while others cases just differ in count. Therefore, the model represents the probability of any wine cases being sold and then the probability of the number of wine cases (Hoffmann, 2004, p. 118). SAS Enterprise Miner’s variable selection feature was used in order to preselect candidate predictor variables. This feature uses the SAS procedure PROC DMINE, which is a forward stepwise selection process using R-squared, for variable selection. The final six variables chosen included: AcidIndex, IMP_Alcohol, IMP_STARS, LabelAppeal, M_STARS, and VolatileAcidity. The results are displayed in *Table 9* below.



```
               Fit Statistics

-2 Log Likelihood                 28821
AIC (smaller is better)           28849
AICC (smaller is better)          28849
BIC (smaller is better)           28949
Pearson Chi-Square              4090.83
Pearson Chi-Square/DF            0.4574


                                              Parameter Estimates

                                             Standard
Parameter          DF       Estimate            Error       95% Confidence Limits      Chi-Square    Pr > ChiSq

Intercept           1       1.157735         0.051196        1.05739        1.25808      511.3777        <.0001
AcidIndex           1      -0.020149         0.005746       -0.03141       -0.00889       12.2944        0.0005
IMP_Alcohol         1       0.007221         0.001717        0.00386        0.01059       17.6910        <.0001
IMP_STARS           1       0.112270         0.007724        0.09713        0.12741      211.2767        <.0001
LabelAppeal         1       0.231305         0.007650        0.21631        0.24630      914.1720        <.0001
VolatileAcidity     1      -0.013885         0.007957       -0.02948        0.00171        3.0454        0.0810
M_STARS 1           1      -0.160649         0.021779       -0.20334       -0.11796       54.4091        <.0001
M_STARS 0           0              0                .              .              .             .             .


                                         Zero-Inflation Parameter Estimates

                                                  Standard
Parameter               DF       Estimate            Error       95% Confidence Limits      Chi-Square    Pr > ChiSq

Intercept_Zero           1      -3.531457         0.320889       -4.16039       -2.90252      121.1147        <.0001
AcidIndex_Zero           1       0.421799         0.030502        0.36202        0.48158      191.2243        <.0001
IMP_Alcohol_Zero         1       0.035969         0.011295        0.01383        0.05811       10.1408        0.0015
IMP_STARS_Zero           1      -2.342010         0.145615       -2.62741       -2.05661      258.6812        <.0001
LabelAppeal_Zero         1       1.324904         0.070380        1.18696        1.46285      354.3784        <.0001
VolatileAcidity_Zero     1       0.147681         0.052328        0.04512        0.25024        7.9649        0.0048
M_STARS_Zero 1           1       4.822899         0.173120        4.48359        5.16221      776.1098        <.0001
M_STARS_Zero 0           0              0                .              .              .             .             .
```

Table 9: The parameter estimates for the logit and the Poisson models comprising the *SAS EM Variable Selection Zero Inflated Poisson.*

The Zero-Iinflated Poisson regression model generates two separate models: the first one is a logit model that predicts whether a wine case is bought at all (labelled “Zero-Inflation Parameter Estimates”). The parameter estimates predict the LOG ODDS that a wine distributor will not buy a case of wine. The second one is a Poisson model that predicts the counts if a wine case were bought (labelled “Parameter Estimates”). The overall model allows for both overdispersion and excess of zeros, which a standard Poisson model cannot predict.

The Zero-Inflated Poisson Regression Model can be described by the following mathematical equation:

EXP (TARGET_1) =

1.1577 –

0.0201 * AcidIndex +

0.0072 * IMP_Alcohol +

0.1123 * IMP_STARS +

0.2313 * LabelAppeal –

0.0139 * VolatileAcidity –

0.1606 * M_STARS (1)

TARGET_ZERO=

-3.5315 –

0.4218 * AcidIndex_Zero +

0.0340 * IMP_Alcohol_Zero -

2.3240 * IMP_STARS_Zero +

1.3249 * LabelAppeal_Zero +

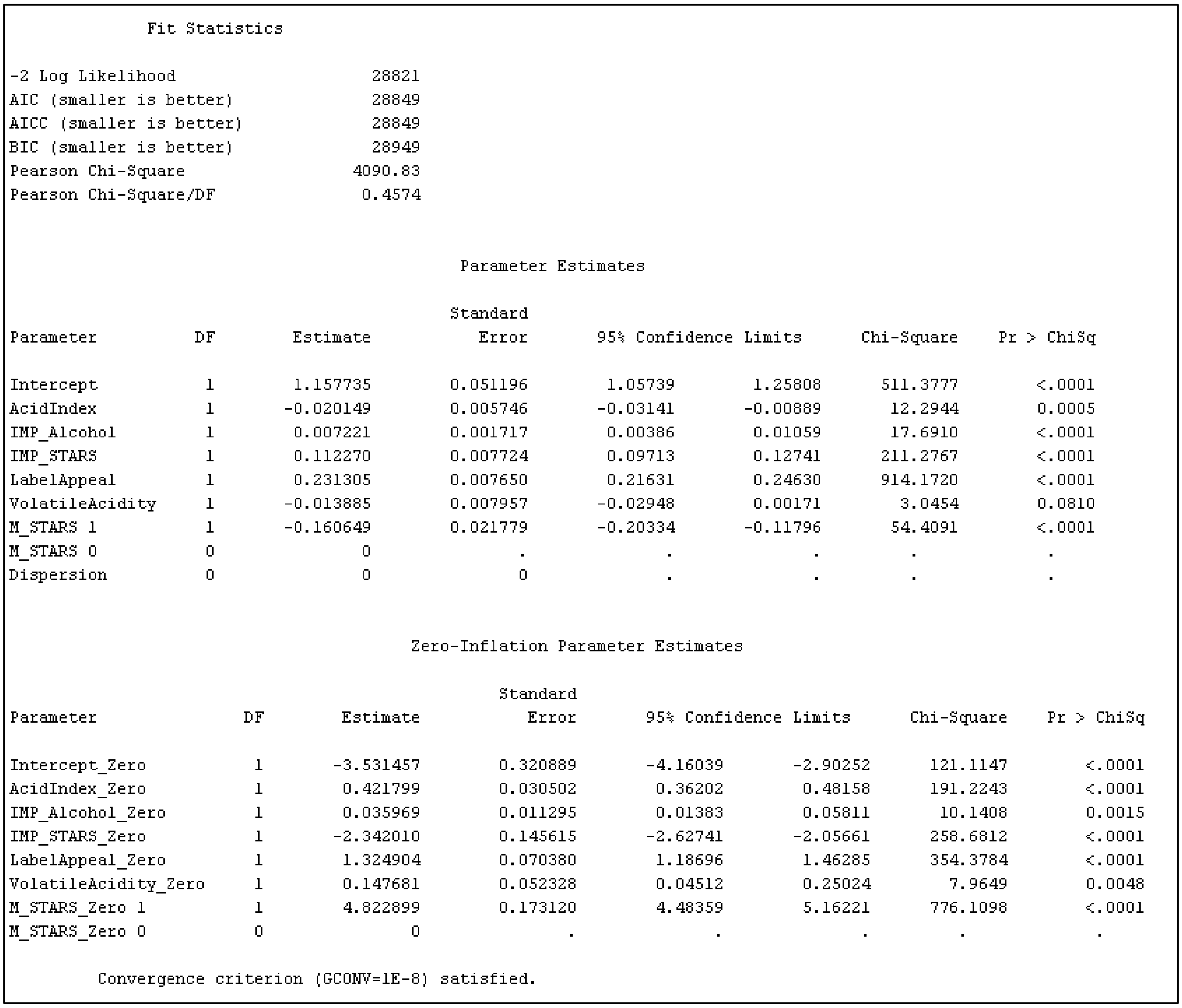
0.1477 * VolatileAcidity_Zero +

4.8229 * M_STARS_Zero (1)

FINAL_TARGET = EXP (TARGET_1) * (1- TARGET_ZERO)

Model 5: SAS EM Variable Selection Zero Inflated Negative Binomial Regression

SAS Enterprise Miner’s variable selection feature was used in order to preselect candidate predictor variables. This feature uses the SAS procedure PROC DMINE, which is a forward stepwise selection process using R-squared, for variable selection. The final six variables chosen included: AcidIndex, IMP_Alcohol, IMP_STARS, LabelAppeal, M_STARS, and VolatileAcidity. The results are displayed in *Table 10* below. Since the Zero-Inflated Poisson Regression Model converged on the same results as the *SAS EM Variable Selection Zero Inflated Poisson* model; the same mathematical equation can be used to describe the model.



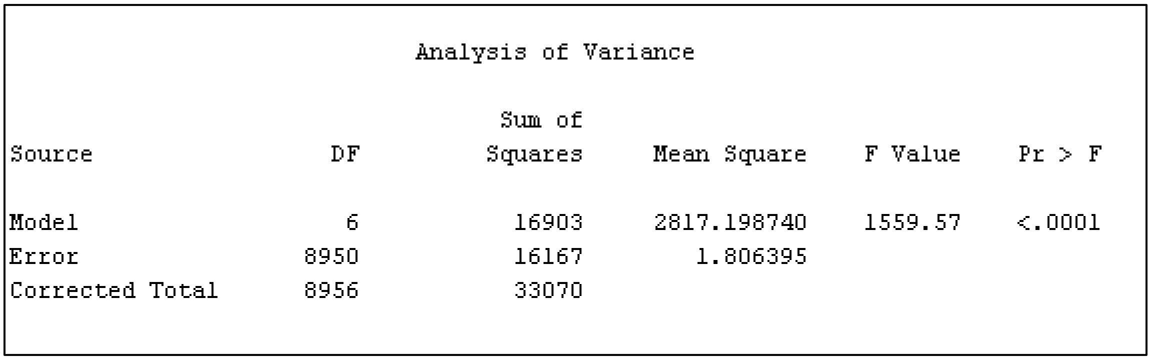
```
               Fit Statistics

-2 Log Likelihood                  28821
AIC (smaller is better)            28849
AICC (smaller is better)           28849
BIC (smaller is better)            28949
Pearson Chi-Square               4090.83
Pearson Chi-Square/DF             0.4574


                                              Parameter Estimates

                                             Standard
Parameter          DF        Estimate           Error       95% Confidence Limits      Chi-Square    Pr > ChiSq

Intercept           1        1.157735        0.051196        1.05739        1.25808      511.3777        <.0001
AcidIndex           1       -0.020149        0.005746       -0.03141       -0.00889       12.2944        0.0005
IMP_Alcohol         1        0.007221        0.001717        0.00386        0.01059       17.6910        <.0001
IMP_STARS           1        0.112270        0.007724        0.09713        0.12741      211.2767        <.0001
LabelAppeal         1        0.231305        0.007650        0.21631        0.24630      914.1720        <.0001
VolatileAcidity     1       -0.013885        0.007957       -0.02948        0.00171        3.0454        0.0810
M_STARS 1           1       -0.160649        0.021779       -0.20334       -0.11796       54.4091        <.0001
M_STARS 0           0               0               .              .              .             .             .
Dispersion          0               0               0              .              .             .             .


                                         Zero-Inflation Parameter Estimates

                                                  Standard
Parameter               DF        Estimate           Error       95% Confidence Limits      Chi-Square    Pr > ChiSq

Intercept_Zero           1       -3.531457        0.320889       -4.16039       -2.90252      121.1147        <.0001
AcidIndex_Zero           1        0.421799        0.030502        0.36202        0.48158      191.2243        <.0001
IMP_Alcohol_Zero         1        0.035969        0.011295        0.01383        0.05811       10.1408        0.0015
IMP_STARS_Zero           1       -2.342010        0.145615       -2.62741       -2.05661      258.6812        <.0001
LabelAppeal_Zero         1        1.324904        0.070380        1.18696        1.46285      354.3784        <.0001
VolatileAcidity_Zero     1        0.147681        0.052328        0.04512        0.25024        7.9649        0.0048
M_STARS_Zero 1           1        4.822899        0.173120        4.48359        5.16221      776.1098        <.0001
M_STARS_Zero 0           0               0               .              .              .             .             .

         Convergence criterion (GCONV=1E-8) satisfied.
```

Table 10: The parameter estimates for the logit and the Poisson models comprising the SAS EM Variable Selection Zero Inflated Negative Binomial.

Model 6: SAS EM Variable Selection Linear Regression using Maximum Likelihood

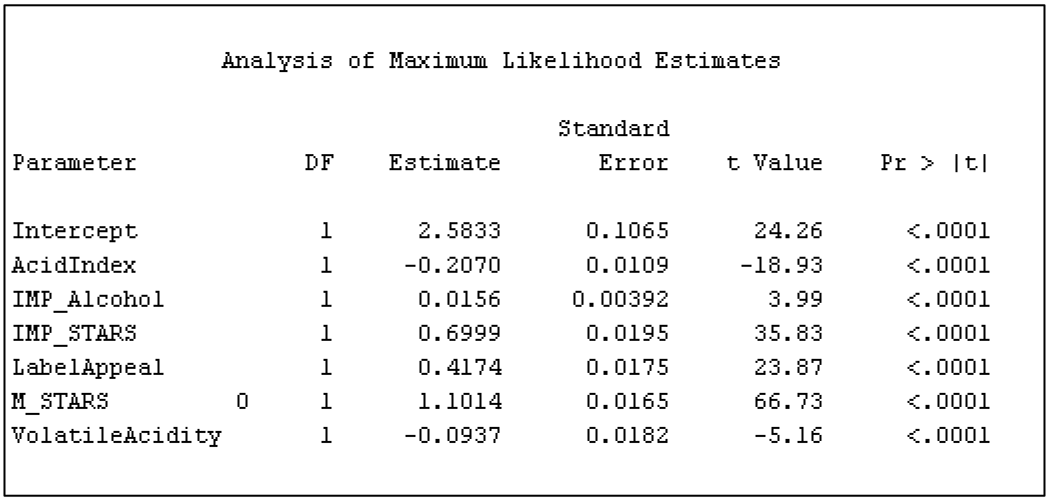
It may be appropriate to attempt to use a linear regression model, although the response variable is not continuous. This linear regression model used the Maximum Likelihood method (as opposed to Ordinary Least Squares) in order to estimate the parameters. SAS Enterprise Miner’s variable selection feature was used in order to preselect candidate predictor variables. This feature uses the SAS procedure PROC DMINE, which is a forward stepwise selection process using R-squared, for variable selection. The final six variables chosen included:

AcidIndex, IMP_Alcohol, IMP_STARS, LabelAppeal, M_STARS, and VolatileAcidity. Table 11 below displays the analysis of variance, while Table 12 below shows the six statistically-significant (< 0.05) parameter estimates and the intercept.



```
                              Analysis of Variance

                                    Sum of
Source                   DF        Squares     Mean Square    F Value    Pr > F

Model                     6          16903     2817.198740    1559.57    <.0001
Error                  8950          16167        1.806395
Corrected Total        8956          33070
```

Table 11: Analysis of Variance for the SAS EM Variable Selection Linear Regression using Maximum Likelihood model.



```
               Analysis of Maximum Likelihood Estimates

                                               Standard
Parameter              DF     Estimate           Error    t Value     Pr > |t|

Intercept               1       2.5833          0.1065      24.26       <.0001
AcidIndex               1      -0.2070          0.0109     -18.93       <.0001
IMP_Alcohol             1       0.0156         0.00392       3.99       <.0001
IMP_STARS               1       0.6999          0.0195      35.83       <.0001
LabelAppeal             1       0.4174          0.0175      23.87       <.0001
M_STARS         0       1       1.1014          0.0165      66.73       <.0001
VolatileAcidity         1      -0.0937          0.0182      -5.16       <.0001
```

Table 12: The parameter estimates for the SAS EM Variable Selection Linear Regression using Maximum Likelihood model.

The Linear Regression Model can be described by the following mathematical equation:

TARGET =

2.5833 –

0.2070 * AcidIndex +

0.0156 * IMP_Alcohol +

0.6999 * IMP_STARS +

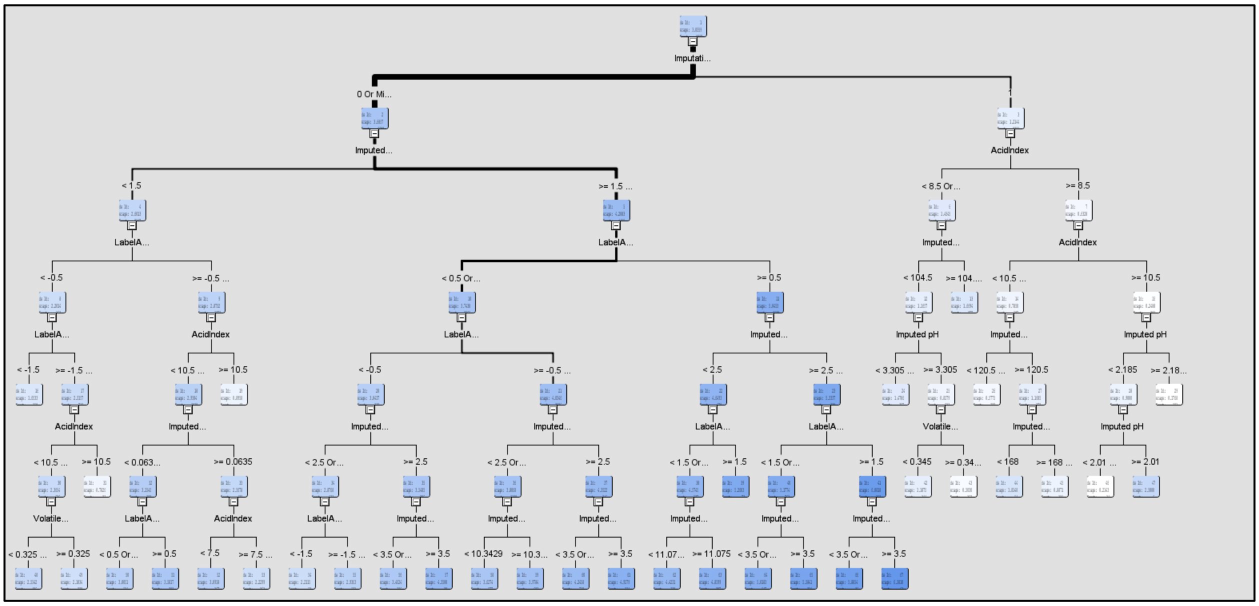
0.4174 * LabelAppeal +

1.1014 * M_STARS (0) -

-0.0937 * VolatileAcidity

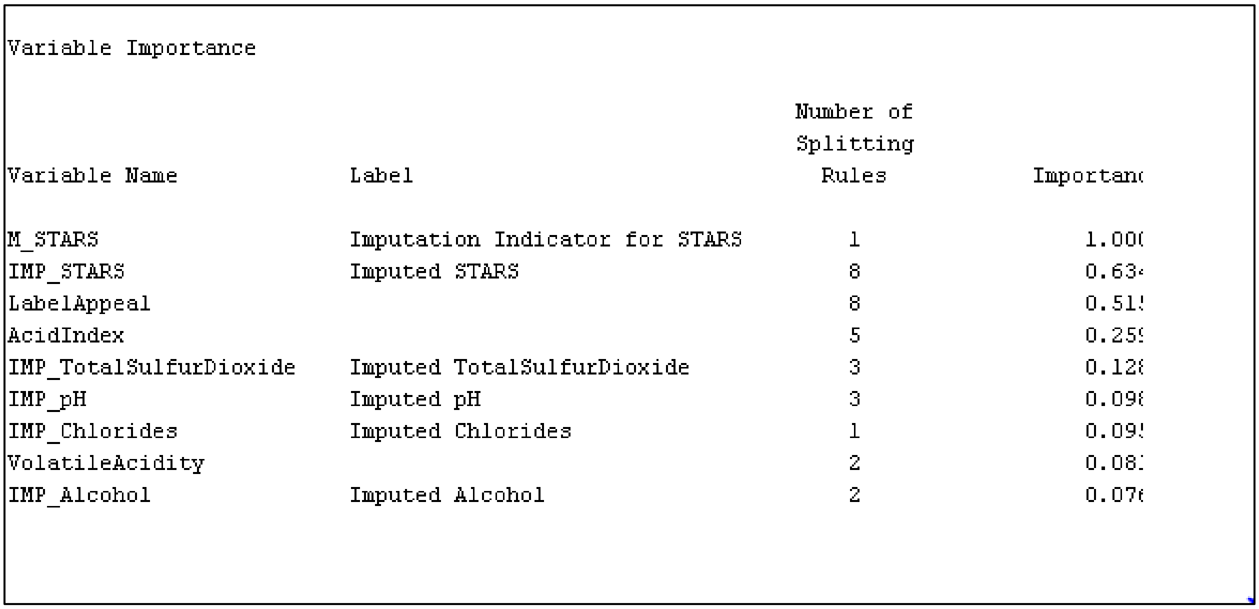
Model 7: Decision Tree

A decision tree was also built to predict the response variable using SAS Enterprise Miner. The tree structure is displayed in Graph 4 below.



Graph 4: Decision Tree built using SAS Enterprise Miner.

The variables used and the number of splitting rules is displayed in Table 13 below.



```
Variable Importance

                                                          Number of
                                                          Splitting
Variable Name             Label                             Rules           Importanc

M_STARS                   Imputation Indicator for STARS      1                 1.000
IMP_STARS                 Imputed STARS                       8                 0.634
LabelAppeal                                                   8                 0.515
AcidIndex                                                     5                 0.259
IMP_TotalSulfurDioxide    Imputed TotalSulfurDioxide          3                 0.128
IMP_pH                    Imputed pH                          3                 0.098
IMP_Chlorides             Imputed Chlorides                   1                 0.095
VolatileAcidity                                               2                 0.081
IMP_Alcohol               Imputed Alcohol                     2                 0.076
```

Table 13: The variables and the number of splitting rules used in the decision tree.

Model Selection

Model selection was based on three criteria: 1) model performance "in-sample" on the training dataset; 2) model performance "out-of-sample" on the testing dataset; 3) and parsimony and sensibleness. The first and second criteria were measured by the Average Square Error (ASE) for each model, as it can be used for both statistical and machine-learning models. The third criterion was measured by the number of variables (including intercept) in the model.

Model Name	**Number of Variables (not incl. intercept)**	**ASE Train**	**ASE Test**
Handpicked Poisson Regression	9	1.758	1.686
Poisson Regression (using ANN) with Variable Transformation	21	1.535	1.506
SAS EM Variable Selection Negative Binomial Regression	6	1.760	1.686

SAS EM Variable Selection Zero Inflated Poisson Regression	6	1.681	1.578
SAS EM Variable Selection Zero Inflated Negative Binomial Regression	6	1.681	1.578
SAS EM Variable Selection Linear Regression using Maximum Likelihood	6	1.805	1.743
Decision Tree	22	1.582	1.558

Table 14: A comparison of all seven models using training/test ASE and number of predictor variables.

The Decision Tree is the best model according to its low ASE value in both training and testing. However, according to this study, the champion model must be a Poisson model. The next best model would be the Poisson Regression (using ANN) with Variable Transformation one, according to its low ASE value in both training and testing. This one is going to be designated as the champion model due to its sheer performance. However, the best model that balances both performance and parsimony is either the SAS EM Variable Selection Zero Inflated Poisson Regression or SAS EM Variable Selection Zero Inflated Negative Binomial Regression, as they both converged on the same results. These models perform quite well and they have two-thirds less predictor variables. This makes it easier to explain to an audience and perhaps even more stable over time. It was interesting to note that the ASE scores were lower for testing than training, which was contra-intuitive to the author.

Lastly, it was interesting that across many models the population parameter estimate for LabelAppeal and IMP_STARS was positive, while M_STARS was negative. This implies that having high wine-label appeal and a higher number of stars is positively correlated with the purchases of that particular wine (case). However, having a missing star rating has a negative correlation. Interestingly, LabelAppeal was negative for the zero-inflated population parameter estimates of the Zero-Inflated Poisson/Negative Binomial models, while positive for the parameter estimates of the Logit portion of the entire model. This may indicate that label appeal has a negative correlation with the probability of purchasing any wine cases, but that it has a positive effect on the number of cases once the decision is made to buy at least one.

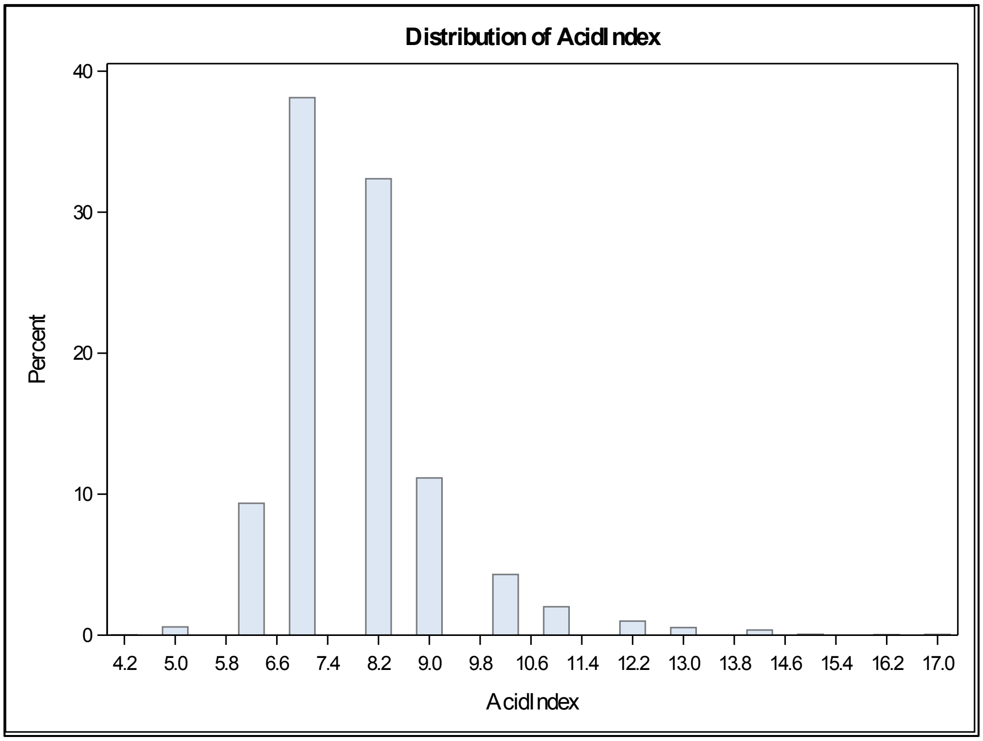
Conclusion

Data exploration and model building are very time-consuming tasks. Much of it is repetitive trial-and-error executed in iterative feedback loops. In order to improve the current proposed models, more time would be required for exploration into the binning of the continuous variables (maybe into bins other than four), transformations of numeric variables, and other various variable-selection methods. A hurdle model, for example, can also be attempted to see if it outperforms the two Zero-Inflation models. In general, the SAS EM Variable Selection Zero Inflated Poisson Regression and the SAS EM Variable Selection Zero Inflated Negative Binomial Regression models performed well and are quite parsimonious, containing just six variables. Therefore, perhaps these models are already approximating the best possible model for this dataset, if staying within the framework of Poisson and Negative Binomial models. In addition, a consultant with domain expert in wine manufacturing to parse over the models would be invaluable. Many of the predictor variables that are chemical properties of the wine contain negative values that are unfamiliar to the author. If any of these observations contain insensible measurements, then new models without those particular predictor variables would have to be built. Finally, the model should be updated each season to include new observations, as both the chemical properties of the wines and the marketing preferences of customers can seasonally change. Additionally, the predictive accuracy of the model can be tested with new predictor variables such as geographic location of the wine.

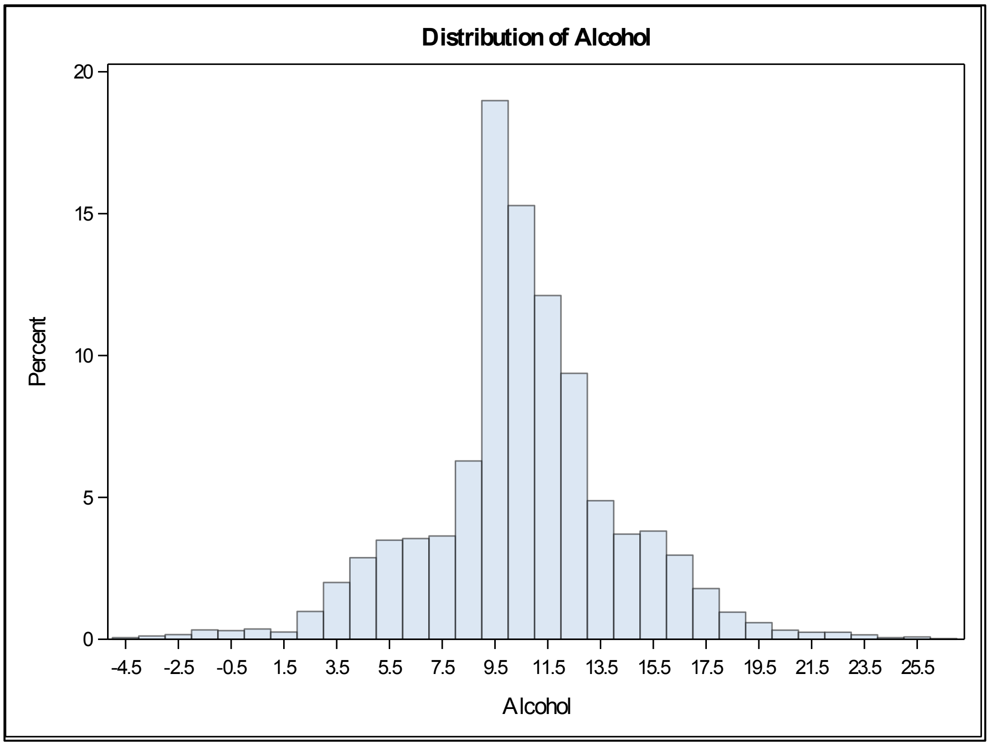
References

Hoffmann, J. P. (2004). Poisson and negative binomial regression models. *Generalized Linear Models: An Applied Approach* (pp. 101-120). Boston, Ma: Pearson Education, Inc.

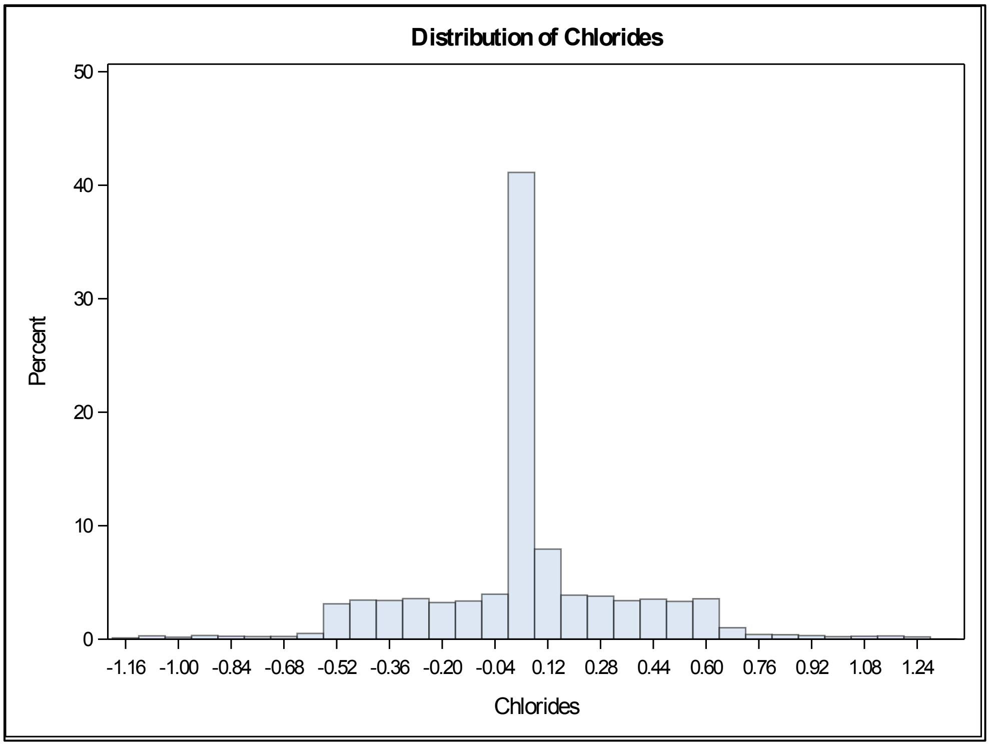
Appendix A



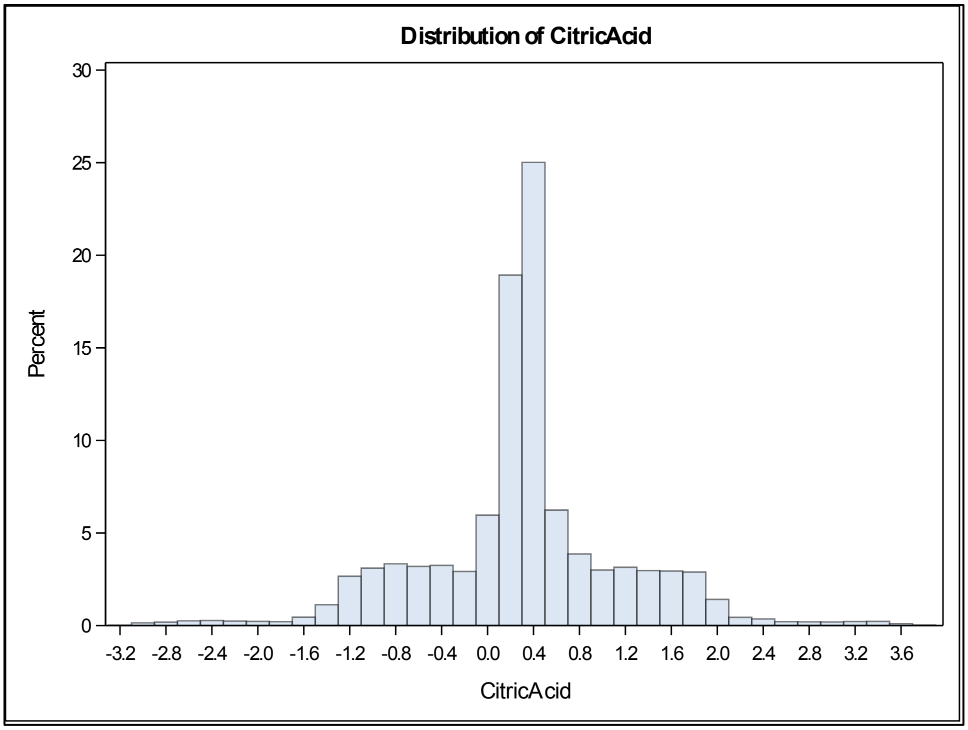
Graph A.1: Histogram of the distribution of the predictor variable AcidIndex



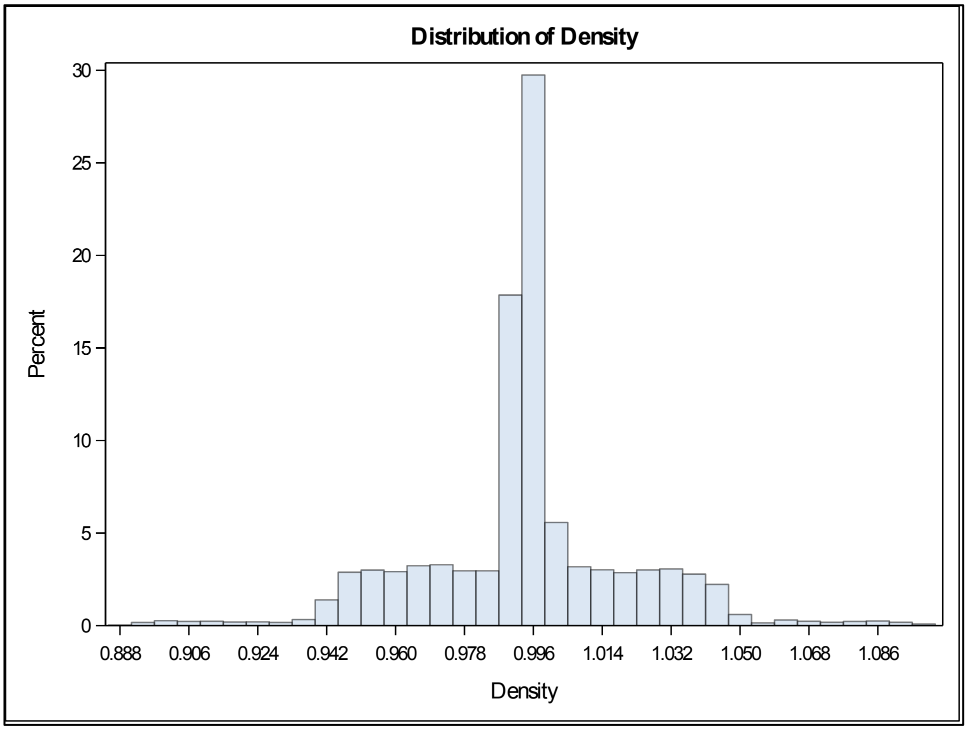
Graph A.2: Histogram of the distribution of the predictor variable Alcohol.



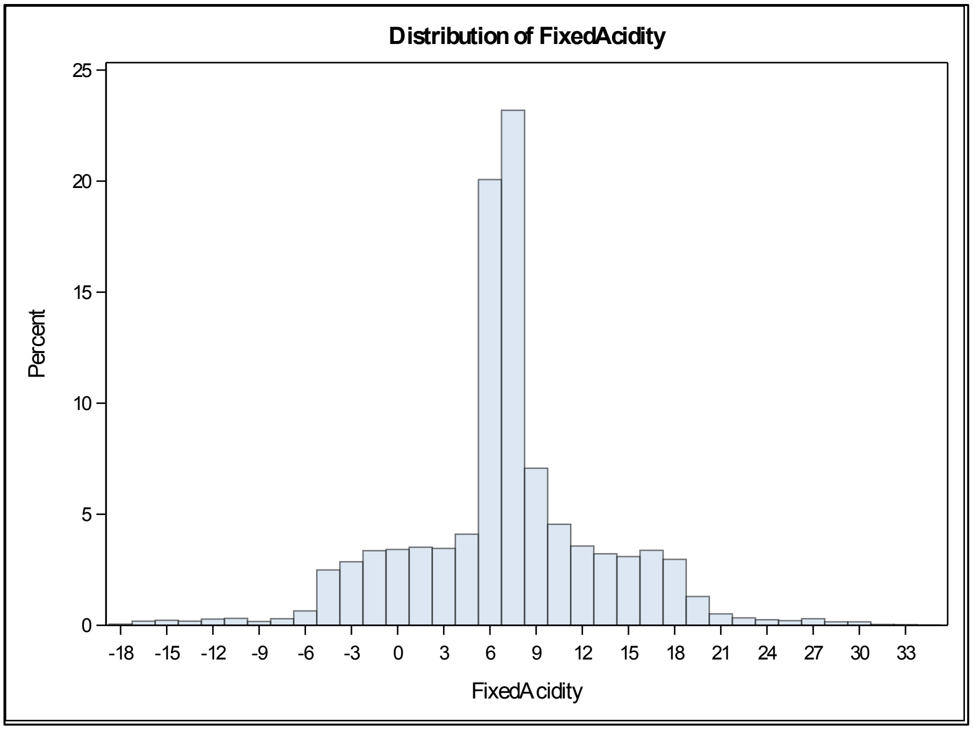
Graph A.3: Histogram of the distribution of the predictor variable Chlorides.



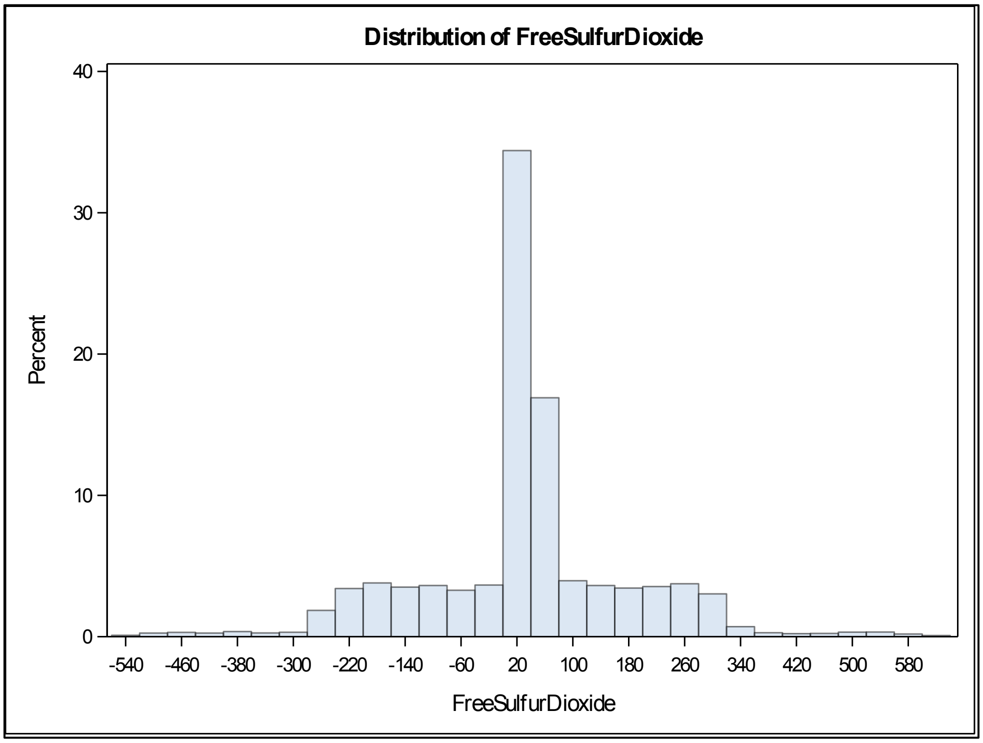
Graph A.4: Histogram of the distribution of the predictor variable CitricAcid.



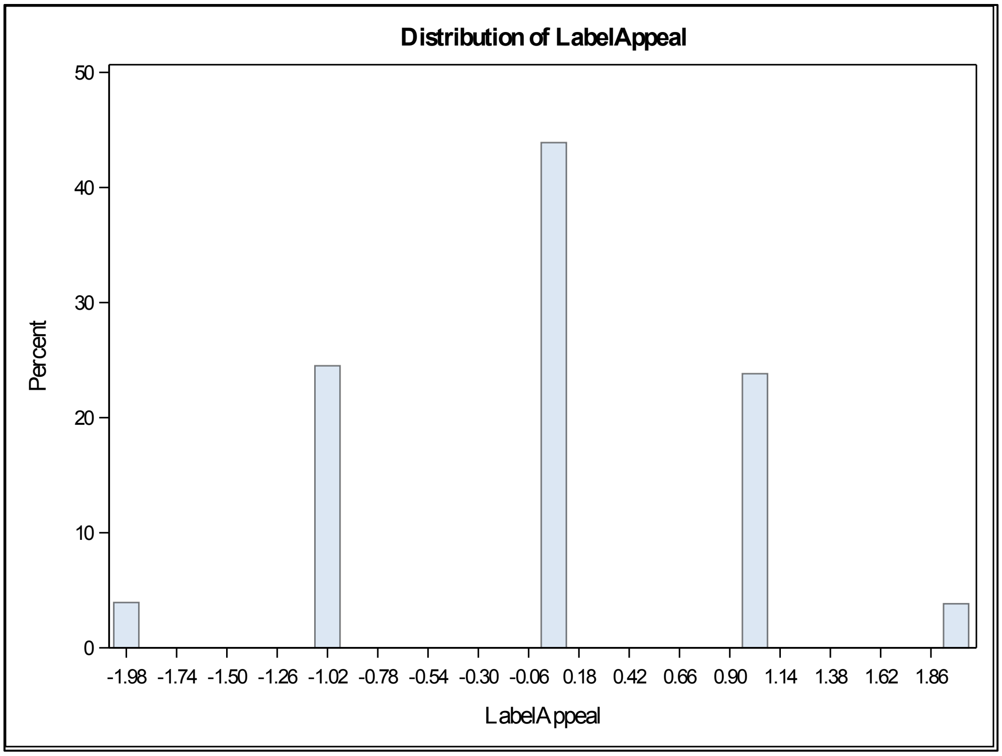
Graph A.5: Histogram of the distribution of the predictor variable Density.



Graph A.6: Histogram of the distribution of the predictor variable FixedAcidity.



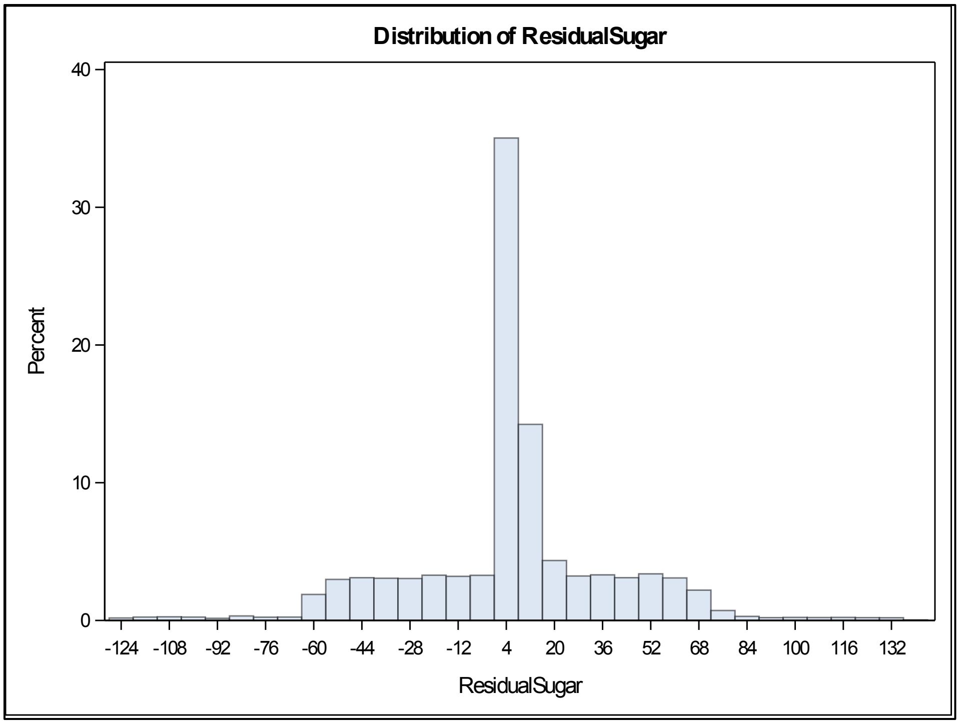
Graph A.7: Histogram of the distribution of the predictor variable FreeSulfurDioxide.



Graph A.8: Histogram of the distribution of the predictor variable LabelAppeal.

LabelAppeal	Frequency	Percent	Cumulative Percent
-2	504	3.94	3.94
-1	3136	24.51	28.45
0	5617	43.90	72.35
1	3048	23.82	96.17
2	490	3.83	100.00

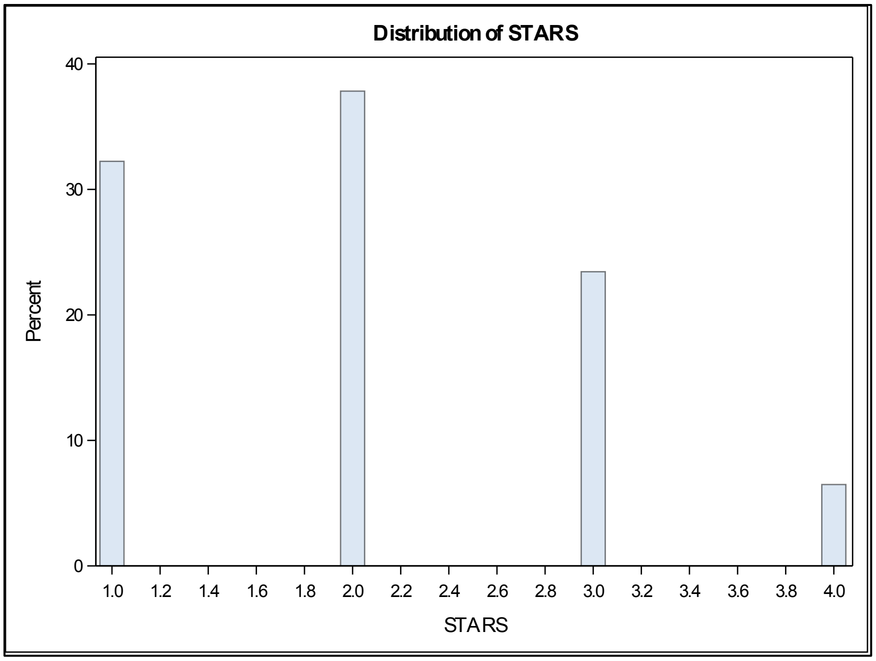
Table A.1: Frequency of the values of the candidate predictor variable LabelAppeal



Graph A.9: Histogram of the distribution of the predictor variable ResidualSugar.

Lowest		Highest	
Value	**Obs**	**Value**	**Obs**
-127.8	5832	136.50	7596
-127.1	2896	137.60	172
-126.2	11939	138.00	3810
-126.1	10799	140.65	11910
-125.7	9927	141.15	186

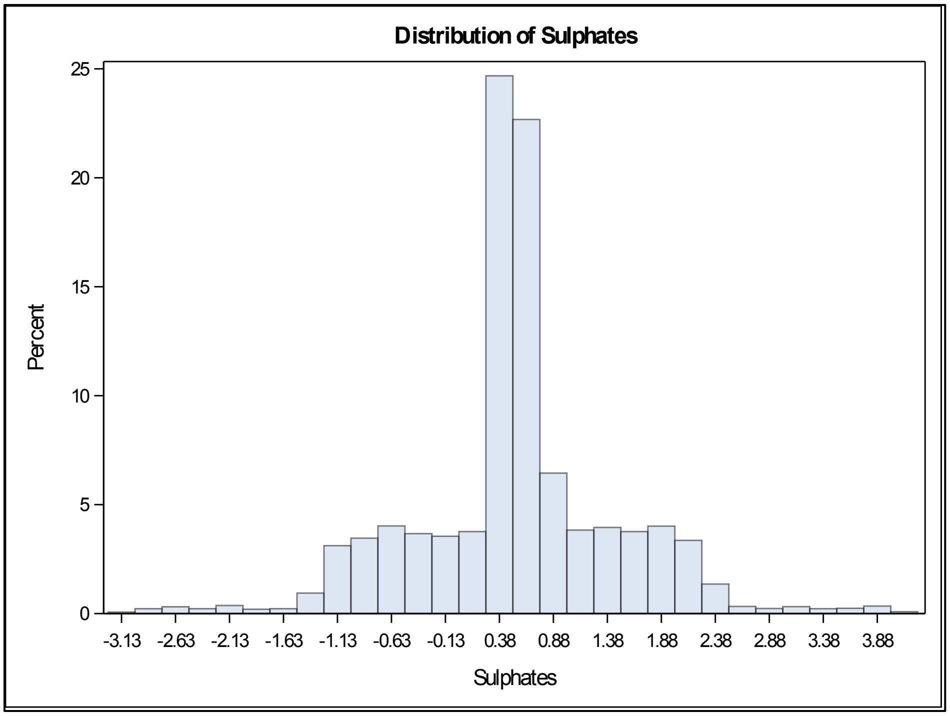
Table A.2: Extreme observations of the predictor variable ResidualSugar



Graph A.10: Histogram of the distribution of the predictor variable STARS.

STARS	Frequency	Percent	Cumulative Percent
missing	3359	26.25	26.25
1	3042	23.77	50.03
2	3570	27.90	77.93
3	2212	17.29	95.22
4	612	4.78	100.00

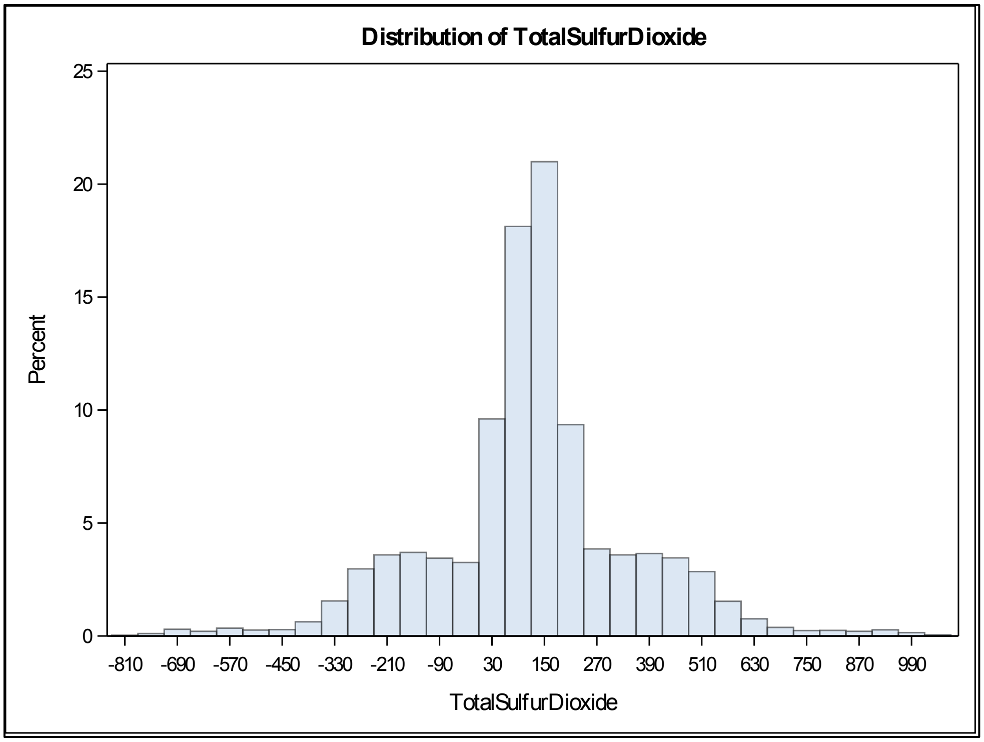
Table A.3: Frequency counts of the values of the predictor variable STARS



Graph A.11: Histogram of the distribution of the predictor variable Sulphates.

Lowest		Highest	
Value	Obs	Value	Obs
-3.13	10275	4.11	9450
-3.12	5209	4.16	6564
-3.12	27	4.19	5362
-3.10	11237	4.21	2943
-3.10	5627	4.24	6215

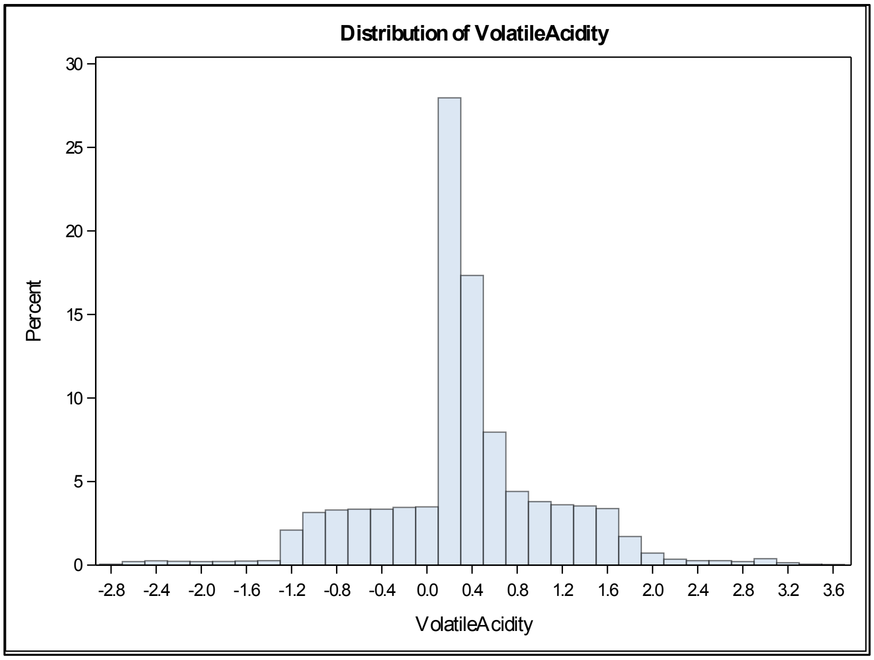
Table A.4: Extreme observations of the predictor variable Sulphates



Graph A.12: Histogram of the distribution of the predictor variable TotalSulfurDioxide.

Lowest		Highest	
Value	Obs	Value	Obs
-823	7147	1032	8635
-816	5189	1041	11791
-793	9553	1048	1208
-781	11742	1054	4099
-779	72	1057	4821

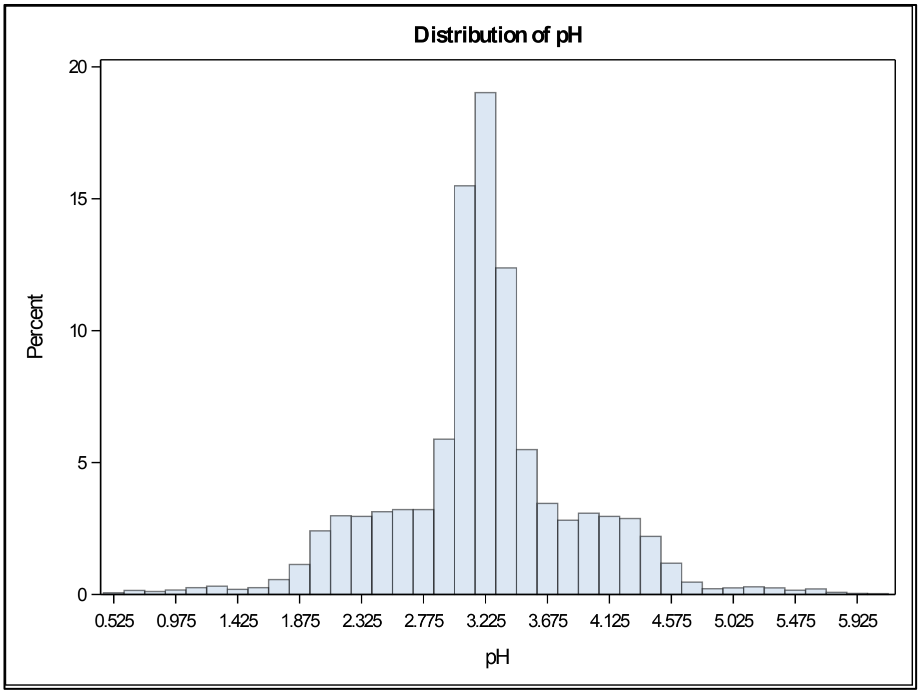
Table A.5: Extreme observations of the predictor variable TotalSulfurDioxide



Graph A.13: Histogram of the distribution of the predictor variable VolatileAcidity.

Lowest		Highest	
Value	Obs	Value	Obs
-2.790	12207	3.500	1676
-2.750	11824	3.550	5653
-2.745	12576	3.565	11568
-2.730	3441	3.590	8523
-2.720	3701	3.680	6081

Table A.6: Extreme observations of the predictor variable VolatileAcidity

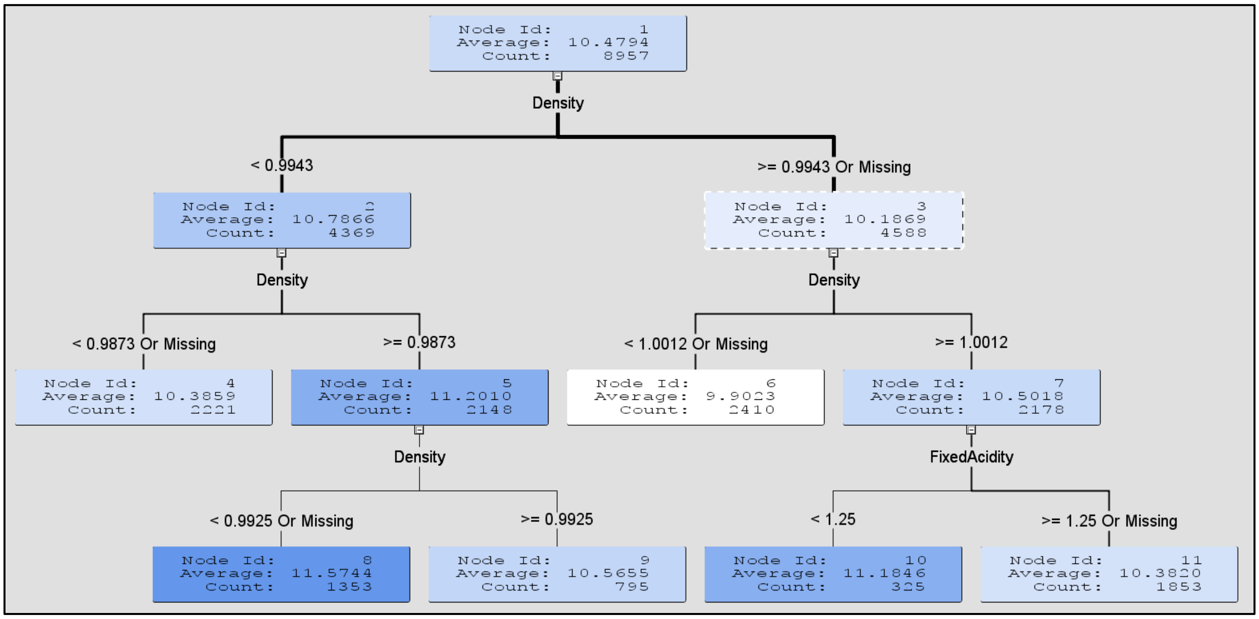


Graph A.14: Histogram of the distribution of the predictor variable pH.

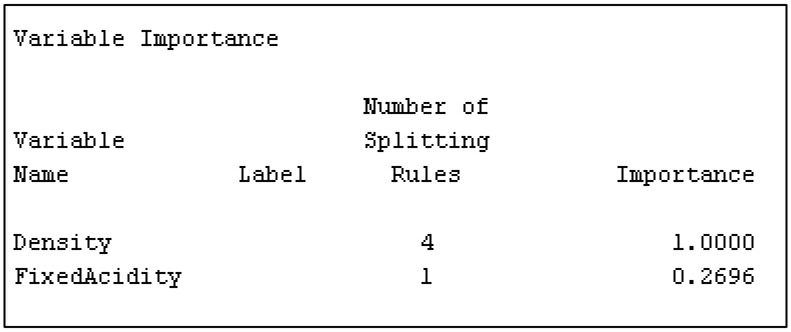
Lowest		Highest	
Value	Obs	Value	Obs
0.48	3803	5.94	8387
0.53	4651	6.02	15
0.54	7828	6.05	3701
0.54	7155	6.05	3704
0.58	9504	6.13	7463

Table A.7: Extreme observations of the predictor variable pH

Appendix B

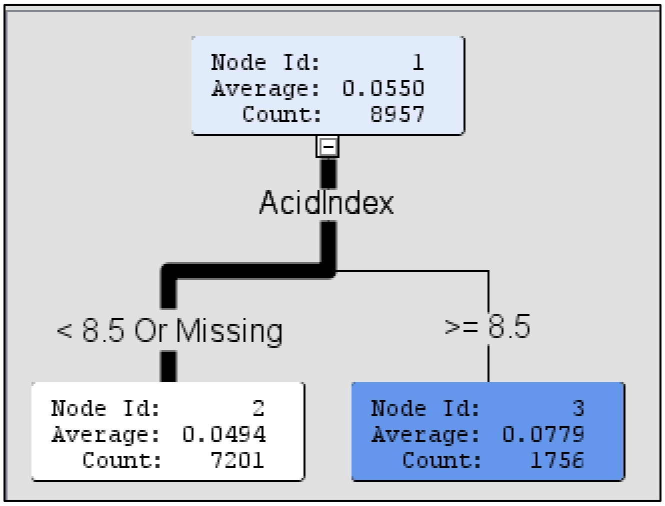


Graph B.1: Example of a decision tree built to impute missing values for predictor variable Alcohol

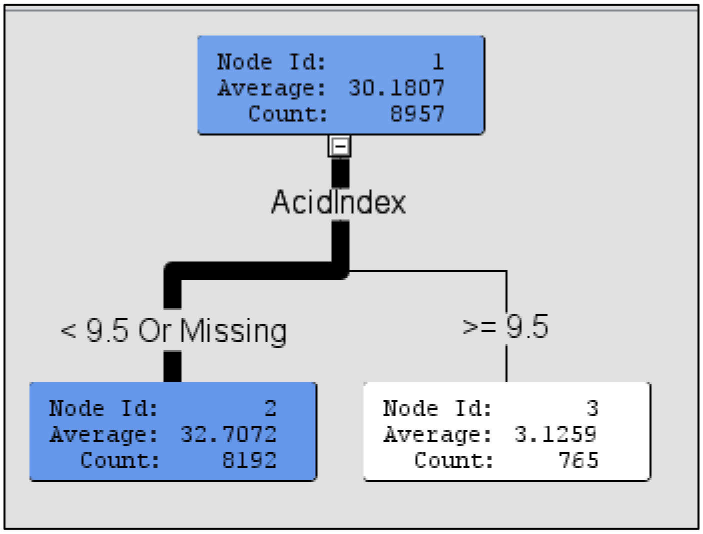


```
Variable Importance

                              Number of
Variable                      Splitting
Name             Label          Rules           Importance

Density                           4                 1.0000
FixedAcidity                      1                 0.2696
```

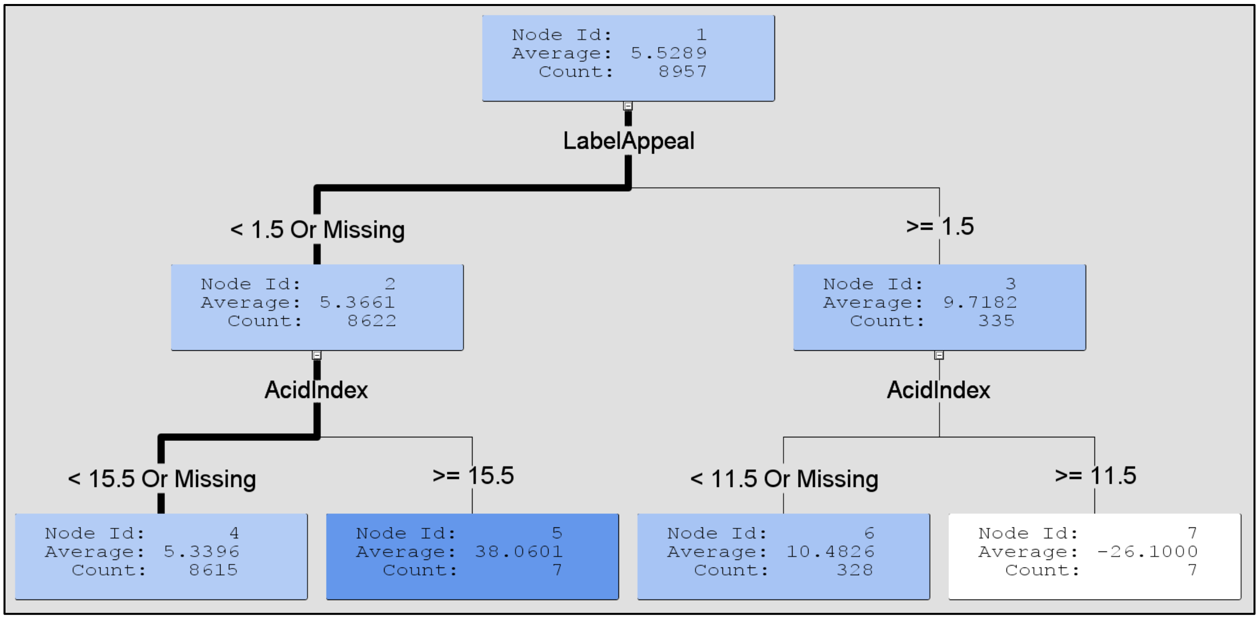
Table B.1: Variables used in order of importance for imputation of missing values of Alcohol.



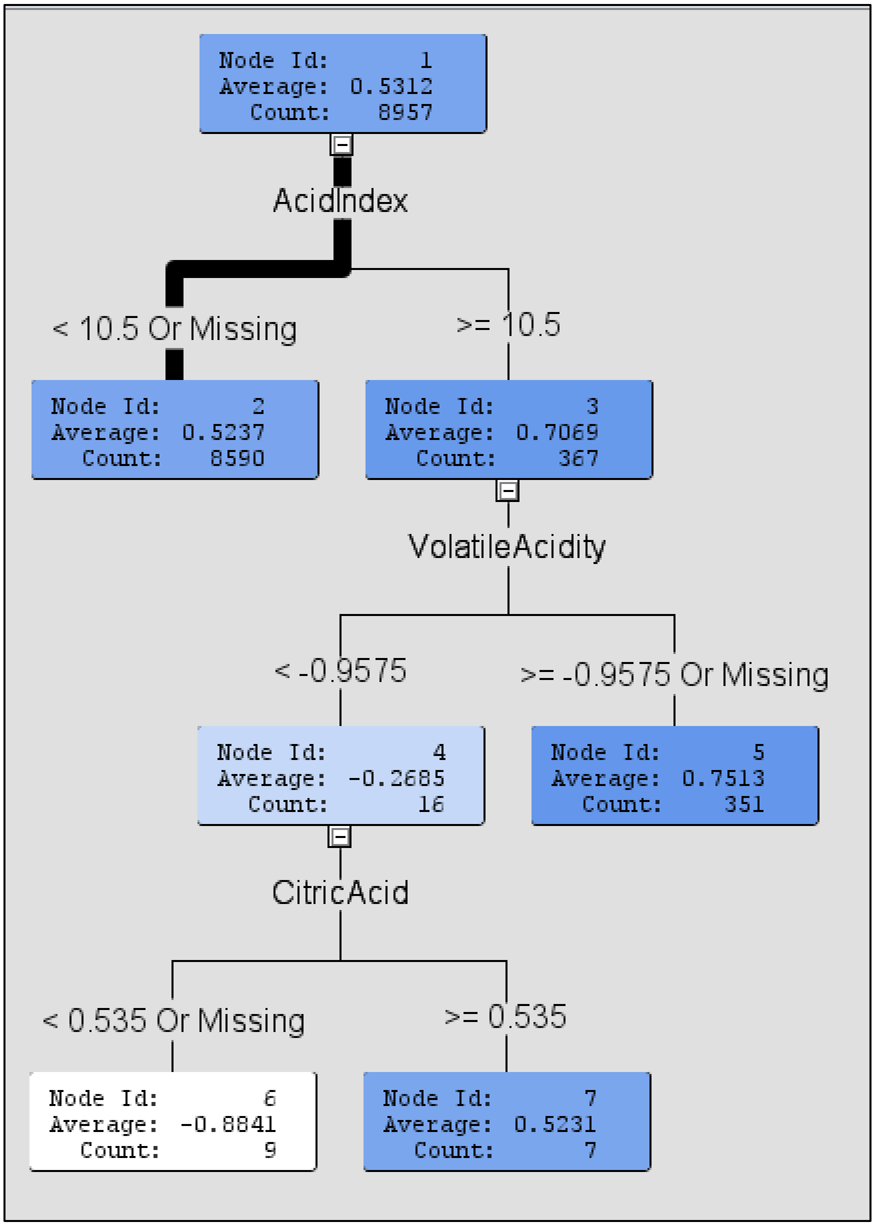
Graph B.2: Example of a decision tree built to impute missing values for predictor variable Chlorides



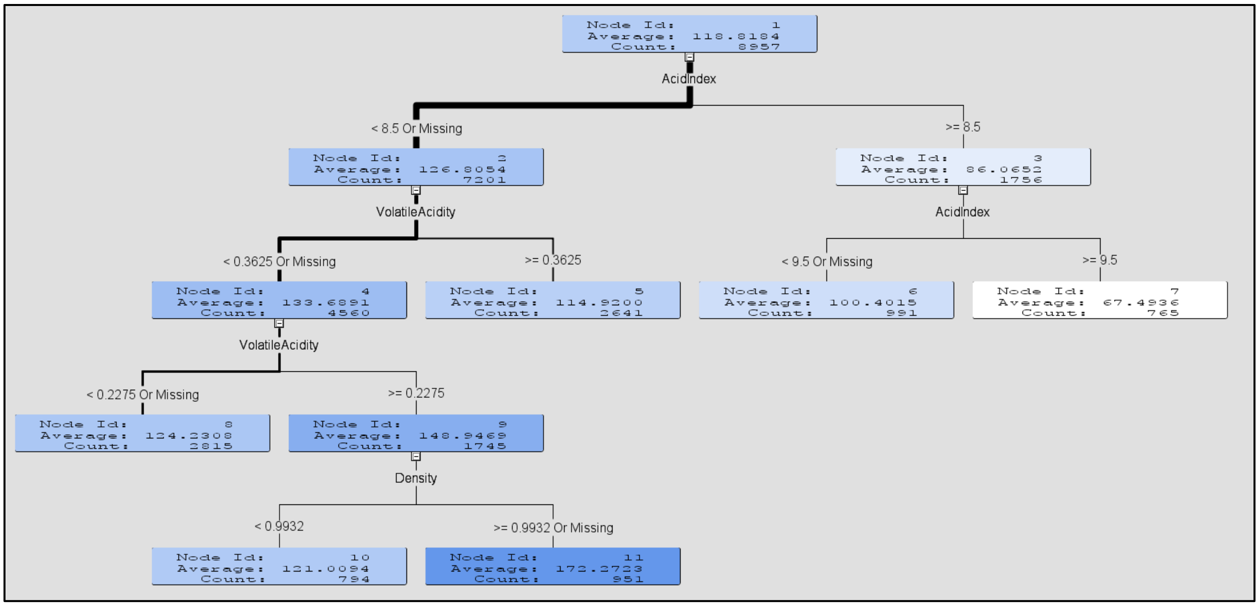
Graph B.3: Example of a decision tree built to impute missing values for predictor variable FreeSulfurDioxide



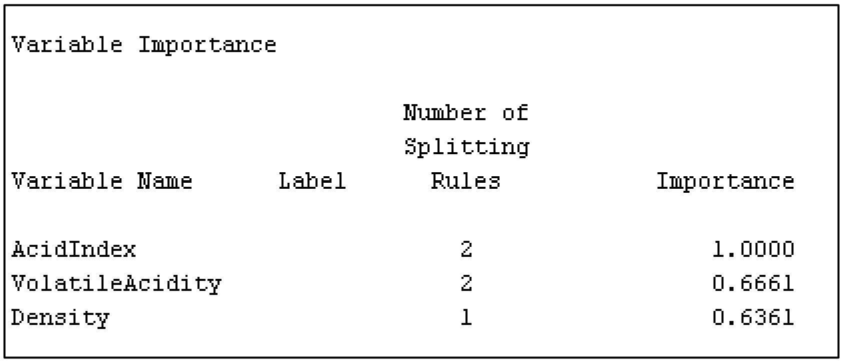
Graph B.4: Example of a decision tree built to impute missing values for predictor variable ResidualSugar



Graph B.5: Example of a decision tree built to impute missing values for predictor variable Sulphates

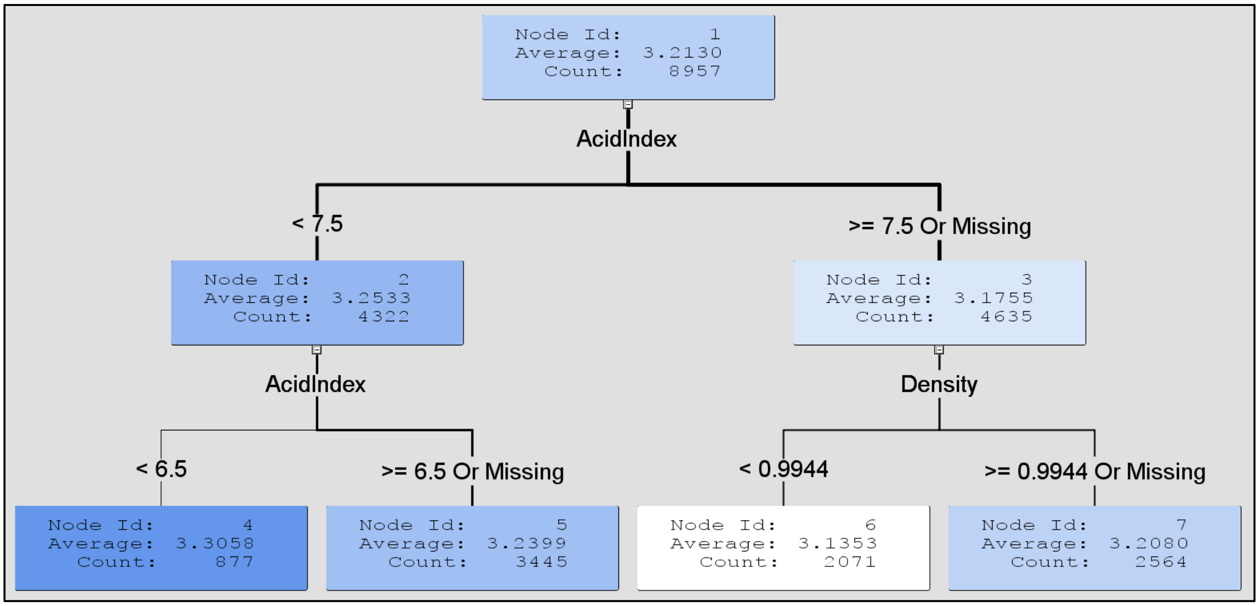


Graph B.6: Example of a decision tree built to impute missing values for predictor variable TotalSulfurDioxide.



```
Variable Importance

                                 Number of
                                 Splitting
Variable Name       Label          Rules              Importance

AcidIndex                            2                   1.0000
VolatileAcidity                      2                   0.6661
Density                              1                   0.6361
```

Table B.2: Variables used in order of importance for imputation of missing values of TotalSulfurDioxide



Graph B.7: Example of a decision tree built to impute missing values for predictor variable TotalSulfurDioxide.

Appendix C

Trial 1:

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	8934	9718.7668	1.0878
Scaled Deviance	8934	9718.7668	1.0878
Pearson Chi-Square	8934	7902.1443	0.8845
Scaled Pearson X2	8934	7902.1443	0.8845
Log Likelihood		6085.3459	
Full Log Likelihood		-16071.5520	
AIC (smaller is better)		32189.1039	
AICC (smaller is better)		32189.2275	
BIC (smaller is better)		32352.4083	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	1.7682	0.2324	1.3127	2.2238	57.89	<.0001
AcidIndex	1	-0.0796	0.0054	-0.0902	-0.0689	214.30	<.0001
IMP_Alcohol	1	0.0049	0.0017	0.0016	0.0082	8.41	0.0037
IMP_Chlorides	1	-0.0653	0.0196	-0.1038	-0.0269	11.09	0.0009
CitricAcid	1	0.0063	0.0071	-0.0075	0.0202	0.80	0.3711
Density	1	-0.2775	0.2283	-0.7250	0.1700	1.48	0.2242
FixedAcidity	1	0.0000	0.0010	-0.0019	0.0020	0.00	0.9649
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.21	0.0072
LabelAppeal	1	0.1548	0.0075	0.1402	0.1694	430.28	<.0001
IMP_ResidualSugar	1	-0.0001	0.0002	-0.0004	0.0003	0.10	0.7579
IMP_STARS	1	0.1763	0.0074	0.1618	0.1907	574.28	<.0001
IMP_Sulphates	1	-0.0095	0.0069	-0.0230	0.0039	1.93	0.1651
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	8.28	0.0040
VolatileAcidity	1	-0.0283	0.0077	-0.0434	-0.0131	13.30	0.0003

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
IMP_pH	1	-0.0106	0.0092	-0.0285	0.0074	1.33	0.2486
M_Alcohol	1	0.0172	0.0275	-0.0367	0.0711	0.39	0.5317
M_Chlorides	1	0.0026	0.0276	-0.0514	0.0566	0.01	0.9251
M_FreeSulfurDioxide	1	0.0150	0.0271	-0.0382	0.0682	0.30	0.5815
M_ResidualSugar	1	0.0396	0.0282	-0.0156	0.0949	1.98	0.1598
M_STARS	1	-1.0015	0.0200	-1.0407	-0.9622	2497.56	<.0001
M_Sulphates	1	-0.0283	0.0212	-0.0699	0.0133	1.77	0.1829
M_TotalSulfurDioxide	1	0.0179	0.0266	-0.0342	0.0699	0.45	0.5018
M_pH	1	-0.0397	0.0364	-0.1111	0.0317	1.19	0.2762
Scale	0	1.0000	0.0000	1.0000	1.0000		

Trial 2:

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	8936	9718.7777	1.0876
Scaled Deviance	8936	9718.7777	1.0876
Pearson Chi-Square	8936	7902.1953	0.8843
Scaled Pearson X2	8936	7902.1953	0.8843
Log Likelihood		6085.3405	
Full Log Likelihood		-16071.5574	
AIC (smaller is better)		32185.1148	
AICC (smaller is better)		32185.2182	
BIC (smaller is better)		32334.2188	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.7682	0.2324	1.3127	2.2237	57.89	<.0001
AcidIndex	1	-0.0796	0.0054	-0.0901	-0.0690	218.46	<.0001
IMP_Alcohol	1	0.0049	0.0017	0.0016	0.0082	8.41	0.0037
IMP_Chlorides	1	-0.0653	0.0196	-0.1038	-0.0269	11.09	0.0009
CitricAcid	1	0.0063	0.0071	-0.0075	0.0202	0.80	0.3713
Density	1	-0.2772	0.2283	-0.7247	0.1702	1.47	0.2246
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.21	0.0072
LabelAppeal	1	0.1548	0.0075	0.1402	0.1694	430.51	<.0001
IMP_ResidualSugar	1	-0.0001	0.0002	-0.0004	0.0003	0.10	0.7546
IMP_STARS	1	0.1763	0.0074	0.1618	0.1907	574.38	<.0001
IMP_Sulphates	1	-0.0095	0.0069	-0.0230	0.0039	1.92	0.1653
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	8.29	0.0040
VolatileAcidity	1	-0.0283	0.0077	-0.0434	-0.0131	13.30	0.0003
IMP_pH	1	-0.0106	0.0092	-0.0285	0.0074	1.33	0.2484
M_Alcohol	1	0.0172	0.0275	-0.0367	0.0711	0.39	0.5313
M_FreeSulfurDioxide	1	0.0150	0.0271	-0.0382	0.0682	0.30	0.5809
M_ResidualSugar	1	0.0397	0.0282	-0.0155	0.0949	1.98	0.1590
M_STARS	1	-1.0015	0.0200	-1.0407	-0.9622	2497.61	<.0001
M_Sulphates	1	-0.0283	0.0212	-0.0699	0.0133	1.78	0.1821
M_TotalSulfurDioxide	1	0.0179	0.0266	-0.0342	0.0699	0.45	0.5015
M_pH	1	-0.0397	0.0364	-0.1111	0.0317	1.19	0.2757
Scale	0	1.0000	0.0000	1.0000	1.0000		

Trial 3:

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	8938	9719.1762	1.0874
Scaled Deviance	8938	9719.1762	1.0874
Pearson Chi-Square	8938	7902.2015	0.8841
Scaled Pearson X2	8938	7902.2015	0.8841
Log Likelihood		6085.1412	
Full Log Likelihood		-16071.7567	
AIC (smaller is better)		32181.5134	
AICC (smaller is better)		32181.5984	
BIC (smaller is better)		32316.4170	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	1.7681	0.2324	1.3126	2.2236	57.88	<.0001
AcidIndex	1	-0.0796	0.0054	-0.0901	-0.0690	218.62	<.0001
IMP_Alcohol	1	0.0049	0.0017	0.0016	0.0082	8.47	0.0036
IMP_Chlorides	1	-0.0653	0.0196	-0.1037	-0.0268	11.07	0.0009
CitricAcid	1	0.0063	0.0071	-0.0075	0.0202	0.80	0.3704
Density	1	-0.2767	0.2283	-0.7241	0.1707	1.47	0.2255
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.18	0.0074
LabelAppeal	1	0.1546	0.0075	0.1400	0.1693	430.14	<.0001
IMP_STARS	1	0.1763	0.0074	0.1619	0.1907	575.29	<.0001
IMP_Sulphates	1	-0.0096	0.0069	-0.0230	0.0039	1.94	0.1639
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	8.25	0.0041
VolatileAcidity	1	-0.0283	0.0077	-0.0434	-0.0131	13.31	0.0003
IMP_pH	1	-0.0106	0.0092	-0.0285	0.0074	1.33	0.2481
M_Alcohol	1	0.0168	0.0275	-0.0371	0.0707	0.37	0.5417
M_ResidualSugar	1	0.0398	0.0282	-0.0155	0.0950	1.99	0.1584
M_STARS	1	-1.0013	0.0200	-1.0406	-0.9620	2497.26	<.0001
M_Sulphates	1	-0.0286	0.0212	-0.0702	0.0129	1.82	0.1770

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
M_TotalSulfurDioxide	1	0.0181	0.0266	-0.0340	0.0702	0.46	0.4956
M_pH	1	-0.0398	0.0364	-0.1112	0.0316	1.20	0.2743
Scale	0	1.0000	0.0000	1.0000	1.0000		

Trial 4:

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	8940	9720.0191	1.0873
Scaled Deviance	8940	9720.0191	1.0873
Pearson Chi-Square	8940	7902.5456	0.8840
Scaled Pearson X2	8940	7902.5456	0.8840
Log Likelihood		6084.7198	
Full Log Likelihood		-16072.1781	
AIC (smaller is better)		32178.3562	
AICC (smaller is better)		32178.4247	
BIC (smaller is better)		32299.0594	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.7669	0.2324	1.3114	2.2223	57.81	<.0001
AcidIndex	1	-0.0796	0.0054	-0.0901	-0.0690	218.53	<.0001
IMP_Alcohol	1	0.0049	0.0017	0.0016	0.0081	8.37	0.0038
IMP_Chlorides	1	-0.0653	0.0196	-0.1038	-0.0269	11.10	0.0009
CitricAcid	1	0.0063	0.0071	-0.0076	0.0202	0.79	0.3731
Density	1	-0.2740	0.2283	-0.7214	0.1734	1.44	0.2300
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.23	0.0072

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
LabelAppeal	1	0.1547	0.0075	0.1401	0.1693	430.43	<.0001
IMP_STARS	1	0.1764	0.0074	0.1620	0.1908	576.01	<.0001
IMP_Sulphates	1	-0.0095	0.0069	-0.0229	0.0040	1.91	0.1672
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	8.20	0.0042
VolatileAcidity	1	-0.0284	0.0077	-0.0436	-0.0132	13.45	0.0002
IMP_pH	1	-0.0105	0.0092	-0.0284	0.0075	1.30	0.2534
M_ResidualSugar	1	0.0400	0.0282	-0.0153	0.0952	2.01	0.1563
M_STARS	1	-1.0012	0.0200	-1.0405	-0.9620	2496.89	<.0001
M_Sulphates	1	-0.0285	0.0212	-0.0700	0.0131	1.80	0.1798
M_pH	1	-0.0395	0.0364	-0.1109	0.0319	1.17	0.2784
Scale	0	1.0000	0.0000	1.0000	1.0000		

Trial 5:

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	8942	9721.9880	1.0872
Scaled Deviance	8942	9721.9880	1.0872
Pearson Chi-Square	8942	7904.6368	0.8840
Scaled Pearson X2	8942	7904.6368	0.8840
Log Likelihood		6083.7353	
Full Log Likelihood		-16073.1625	
AIC (smaller is better)		32176.3251	
AICC (smaller is better)		32176.3788	
BIC (smaller is better)		32282.8280	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.7681	0.2323	1.3128	2.2235	57.92	<.0001
AcidIndex	1	-0.0792	0.0054	-0.0897	-0.0687	217.46	<.0001
IMP_Alcohol	1	0.0049	0.0017	0.0016	0.0082	8.43	0.0037
IMP_Chlorides	1	-0.0656	0.0196	-0.1040	-0.0271	11.18	0.0008
Density	1	-0.2776	0.2282	-0.7248	0.1697	1.48	0.2239
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.29	0.0070
LabelAppeal	1	0.1547	0.0075	0.1401	0.1693	430.57	<.0001
IMP_STARS	1	0.1764	0.0073	0.1620	0.1908	575.97	<.0001
IMP_Sulphates	1	-0.0094	0.0069	-0.0229	0.0040	1.89	0.1693
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	8.23	0.0041
VolatileAcidity	1	-0.0284	0.0077	-0.0436	-0.0132	13.48	0.0002
IMP_pH	1	-0.0105	0.0092	-0.0284	0.0075	1.31	0.2531
M_ResidualSugar	1	0.0400	0.0282	-0.0153	0.0952	2.01	0.1561
M_STARS	1	-1.0017	0.0200	-1.0410	-0.9625	2500.27	<.0001
M_Sulphates	1	-0.0283	0.0212	-0.0698	0.0133	1.77	0.1828
Scale	0	1.0000	0.0000	1.0000	1.0000		

Trial 6:

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	8944	9724.7815	1.0873
Scaled Deviance	8944	9724.7815	1.0873
Pearson Chi-Square	8944	7907.4557	0.8841
Scaled Pearson X2	8944	7907.4557	0.8841
Log Likelihood		6082.3386	
Full Log Likelihood		-16074.5593	
AIC (smaller is better)		32175.1186	
AICC (smaller is better)		32175.1593	
BIC (smaller is better)		32267.4211	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.4578	0.0489	1.3620	1.5537	888.70	<.0001
AcidIndex	1	-0.0790	0.0054	-0.0895	-0.0685	217.97	<.0001
IMP_Alcohol	1	0.0049	0.0017	0.0016	0.0082	8.44	0.0037
IMP_Chlorides	1	-0.0658	0.0196	-0.1042	-0.0273	11.25	0.0008
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.25	0.0071
LabelAppeal	1	0.1545	0.0075	0.1399	0.1691	429.72	<.0001
IMP_STARS	1	0.1764	0.0073	0.1620	0.1909	576.44	<.0001
IMP_Sulphates	1	-0.0095	0.0069	-0.0230	0.0039	1.93	0.1650
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	8.11	0.0044
VolatileAcidity	1	-0.0286	0.0077	-0.0437	-0.0134	13.63	0.0002
M_ResidualSugar	1	0.0394	0.0282	-0.0159	0.0946	1.95	0.1626
M_STARS	1	-1.0024	0.0200	-1.0417	-0.9632	2505.13	<.0001
M_Sulphates	1	-0.0288	0.0212	-0.0704	0.0128	1.85	0.1743
Scale	0	1.0000	0.0000	1.0000	1.0000		

Trial 7:

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	8946	9728.5821	1.0875
Scaled Deviance	8946	9728.5821	1.0875
Pearson Chi-Square	8946	7911.0428	0.8843
Scaled Pearson X2	8946	7911.0428	0.8843
Log Likelihood		6080.4383	
Full Log Likelihood		-16076.4596	
AIC (smaller is better)		32174.9192	
AICC (smaller is better)		32174.9488	
BIC (smaller is better)		32253.0213	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.4500	0.0487	1.3545	1.5456	885.33	<.0001
AcidIndex	1	-0.0790	0.0053	-0.0895	-0.0685	218.15	<.0001
IMP_Alcohol	1	0.0049	0.0017	0.0016	0.0081	8.37	0.0038
IMP_Chlorides	1	-0.0658	0.0196	-0.1042	-0.0274	11.27	0.0008
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.20	0.0073
LabelAppeal	1	0.1545	0.0075	0.1399	0.1691	429.59	<.0001
IMP_STARS	1	0.1767	0.0073	0.1623	0.1911	578.19	<.0001
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	8.15	0.0043
VolatileAcidity	1	-0.0287	0.0077	-0.0439	-0.0136	13.78	0.0002
M_ResidualSugar	1	0.0384	0.0282	-0.0168	0.0936	1.86	0.1728
M_STARS	1	-1.0031	0.0200	-1.0423	-0.9638	2508.77	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Trial 8:

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	8947	9730.4193	1.0876
Scaled Deviance	8947	9730.4193	1.0876
Pearson Chi-Square	8947	7912.3181	0.8844
Scaled Pearson X2	8947	7912.3181	0.8844
Log Likelihood		6079.5197	
Full Log Likelihood		-16077.3782	
AIC (smaller is better)		32174.7564	
AICC (smaller is better)		32174.7810	
BIC (smaller is better)		32245.7583	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	1.4517	0.0487	1.3562	1.5472	887.95	<.0001
AcidIndex	1	-0.0791	0.0053	-0.0896	-0.0686	218.48	<.0001
IMP_Alcohol	1	0.0048	0.0017	0.0016	0.0081	8.36	0.0038
IMP_Chlorides	1	-0.0656	0.0196	-0.1040	-0.0272	11.19	0.0008
IMP_FreeSulfurDioxid	1	0.0001	0.0000	0.0000	0.0002	7.19	0.0073
LabelAppeal	1	0.1544	0.0075	0.1398	0.1690	428.89	<.0001
IMP_STARS	1	0.1769	0.0073	0.1626	0.1913	580.16	<.0001
IMP_TotalSulfurDioxi	1	0.0001	0.0000	0.0000	0.0001	8.23	0.0041
VolatileAcidity	1	-0.0286	0.0077	-0.0438	-0.0134	13.66	0.0002
M_STARS	1	-1.0028	0.0200	-1.0421	-0.9636	2507.80	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Appendix D:

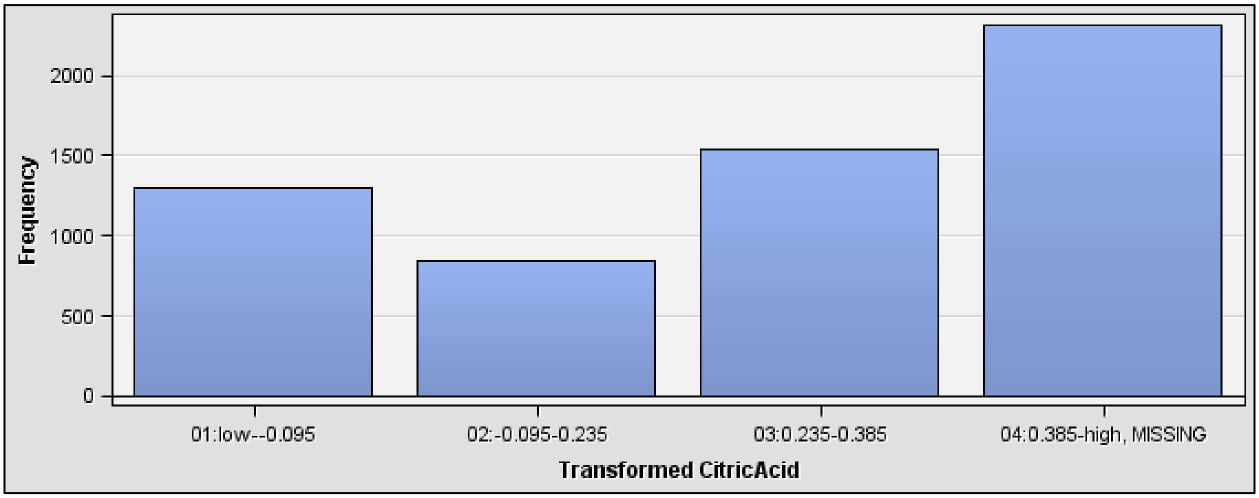


Table D.1: Binning of predictor variable CitricAcid

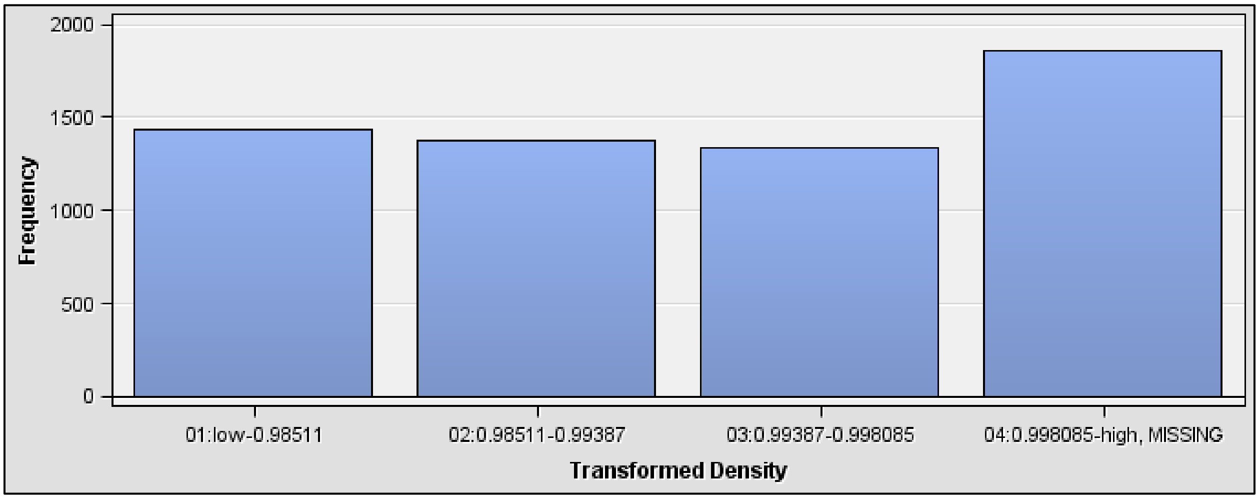


Table D.2: Binning of predictor variable Density

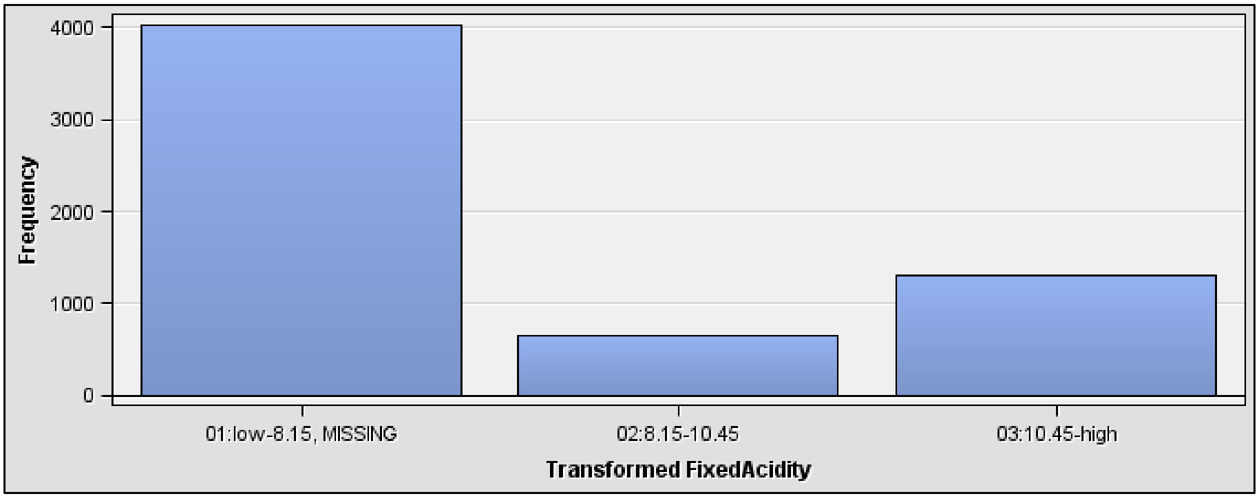


Table D.2: Binning of predictor variable FixedAcidity

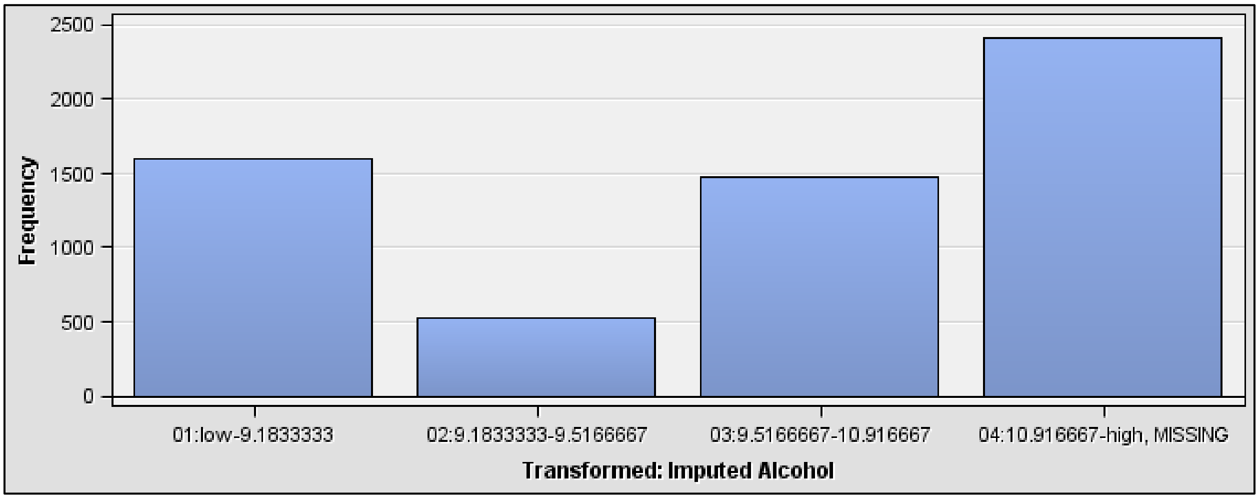


Table D.3: Binning of predictor variable Imp_Alcohol

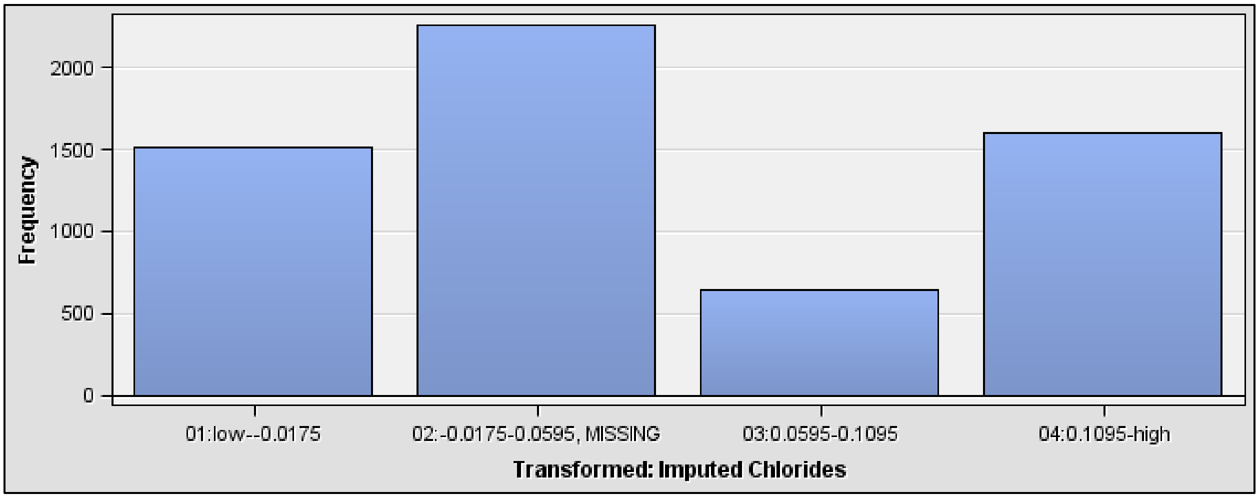


Table D.4: Binning of predictor variable Imp_Chlorides

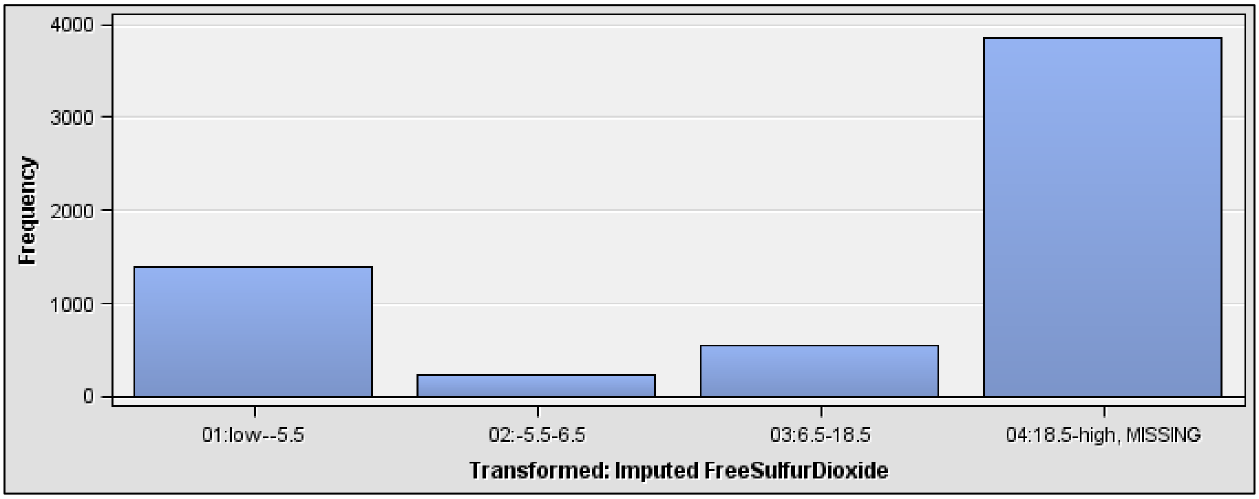


Table D.5: Binning of predictor variable Imp_FreeSulfurDioxide

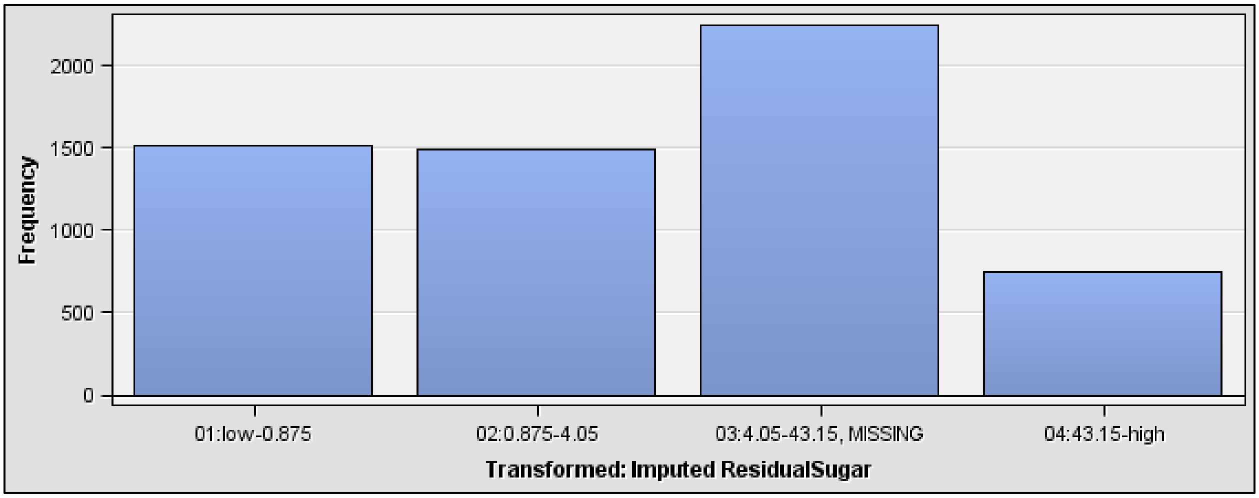


Table D.6: Binning of predictor variable Imp_ResidualSugar

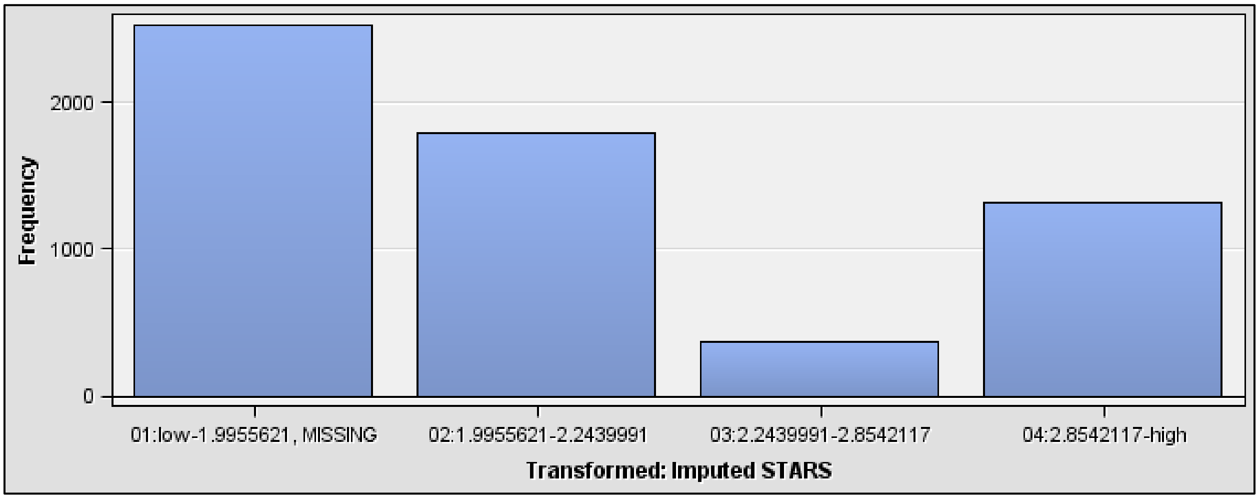


Table D.7: Binning of predictor variable Imp_STARS

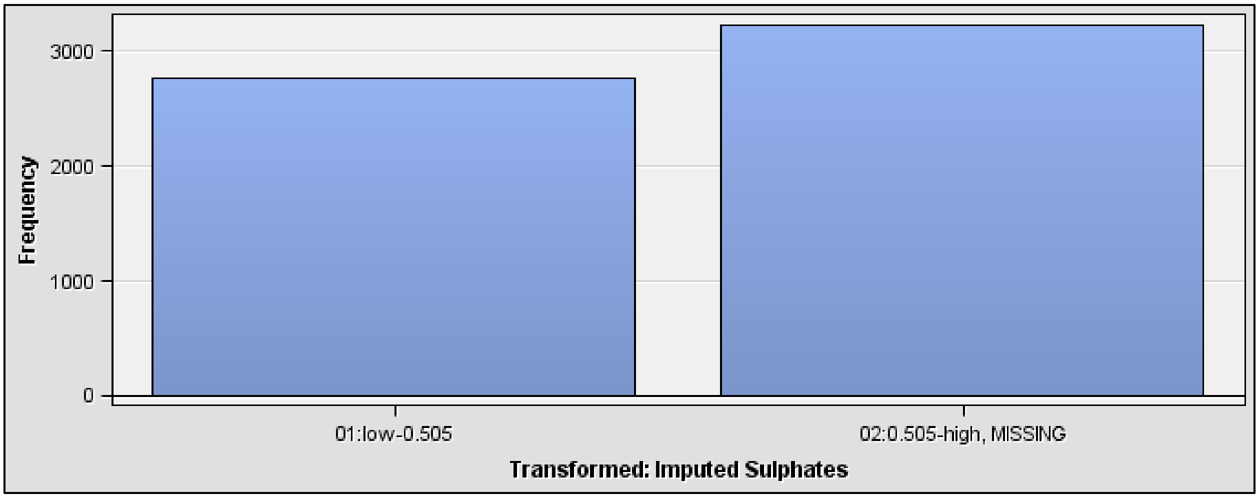


Table D.8: Binning of predictor variable Imp_Sulphates

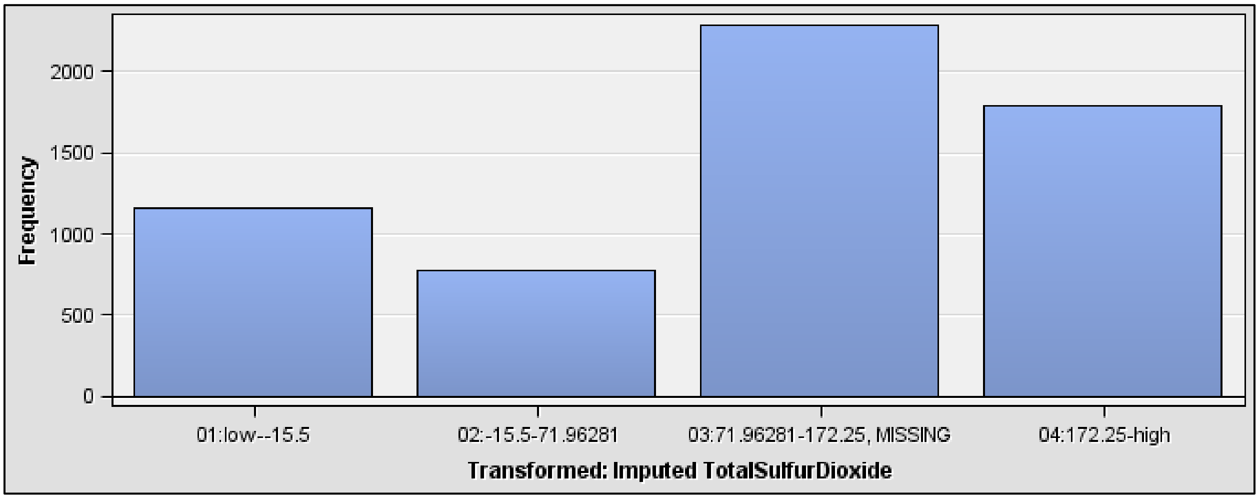


Table D.9: Binning of predictor variable Imp_TotalSulfurDioxide

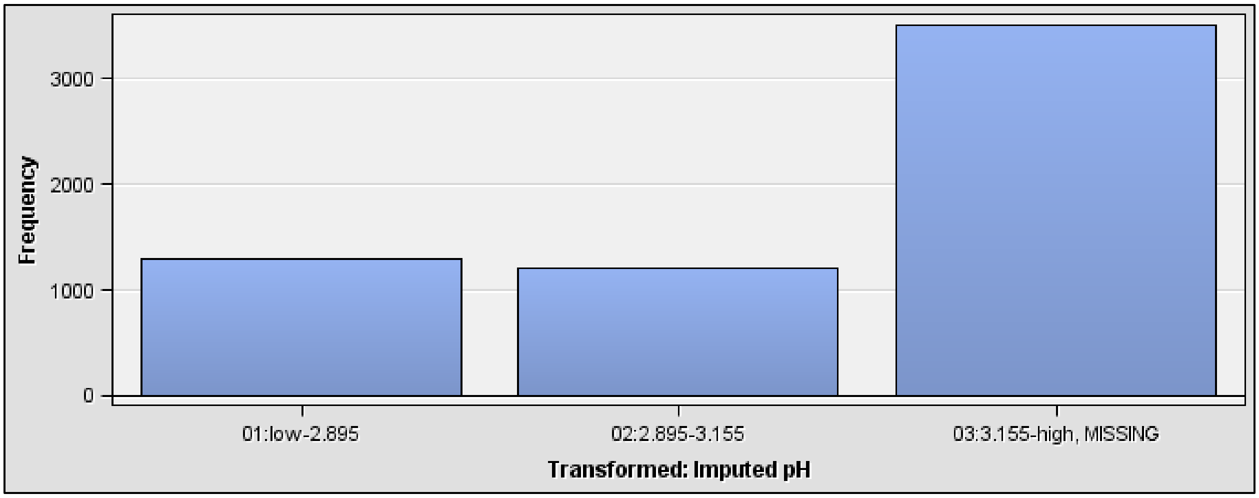


Table D.10: Binning of predictor variable Imp_pH

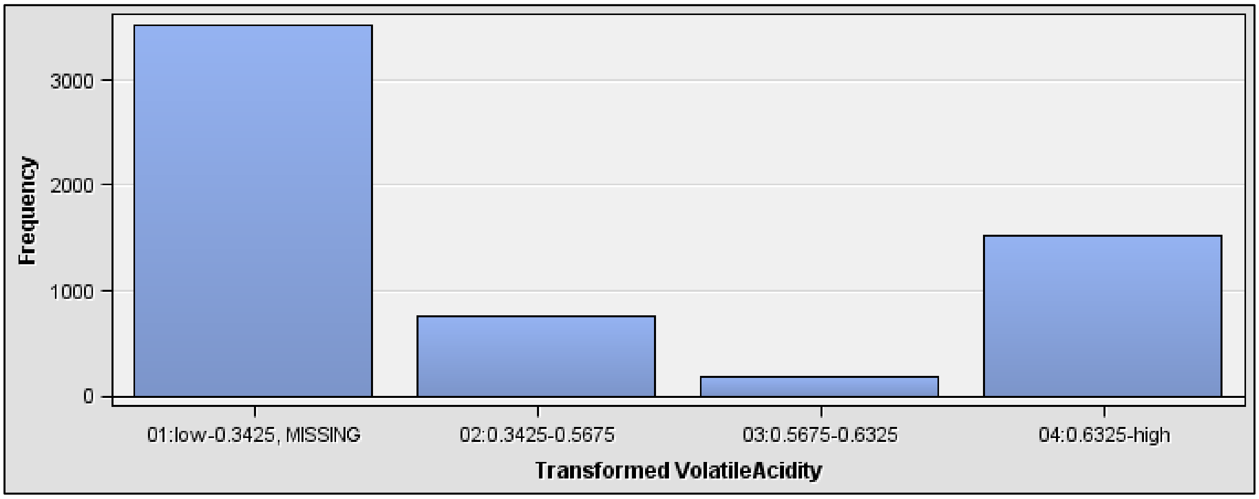


Table D.11: Binning of predictor variable Imp_VolatileAcidity