Predicting Wine Sales

Scott Herman

March, 5th 2017

Prepared for Predict-411: Generalized Linear Models
Northwestern University Masters in Science, Predictive Analytics
Kaggle File Submission: **wine_test_score_Herman.csv**

Introduction

The purpose of this analysis is to predict the number of wine cases sold by a wine manufacturing company based upon each wine's chemical characteristics. The data set contains informaiton on approximately 12,000 commercially available wines. The target variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stroes around the United States, and if the manaufacturer can predict the number of cases, then they will be able to adjust their wine offering to maximize sales.

Our initial data set includes a total of 14 predictor variables, each describing the various properties associated with a given wine. This set of predictors includes both quantitative and qualitative descriptive metrics, which will be utilized to develop Poisson, Negative Binomial, and OLS Regression preditive models. The table below gives the definition for each of these variables.

Table 1: Data Definitions

Variable	Description
AcidIndex	Proprietary method of testing total acidity of wine by using a weighted average
Alcohol	Alcohol Content
Chlorides	Chloride content of wine
CitricAcid	Citric Acid Content
Density	Density of Wine
FixedAcidity	Fixed Acidity of Wine
FreeSulfurDioxide	Sulfur Dioxide content of wine
LabelAppeal	Marketing Score indicating the appeal of label design for consumers.
ResidualSugar	Residual Sugar of wine
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor
Sulphates	Sulfate conten of wine
TotalSulfurDioxide	Total Sulfur Dioxide of Wine
VolatileAcidity	Volatile Acid content of wine
pH	pH of wine

Data Exploration

This analysis begins by understanding the variables within our data set along with their corresponding observations in an attempt to identify the structure and quality of our data set. First, we will examine the distribution of our target variable, along with the mean and variance to identify whether or not equidispersion is present. Then, we will plot the distributions of our predictor varibles, check for any missing of influential observations, and examine their correlation to our response.

The histogram below, reveals the distribution of our target variable. This visual indicates that Target is actually categorical in nature, and possesses 9 different levels, or categories, ranging from 0 to 8. Additionally, this shows that Target does appear to possess normality, while also showing signs of being zero-inflated.

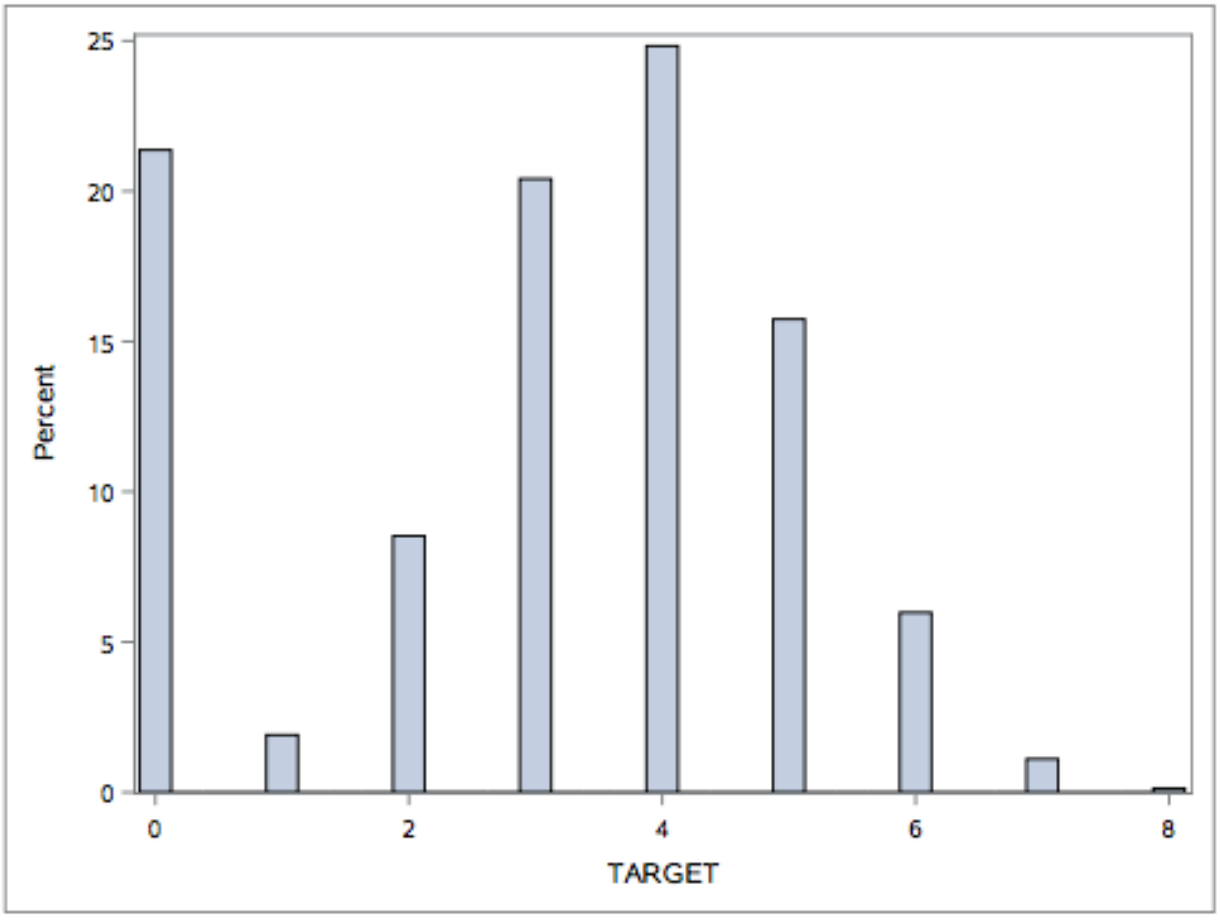


Figure 1: Histogram of Target Variable

Next, we take a look at the mean and variance of target which is given in Table 2, below. This shows that the variance is larger than the mean for our response, which violates the necessary assumptions for the Poisson distribution, but does meet the requirements for the Negative Binomial distribution.

Table 2: Mean and Variance of Target

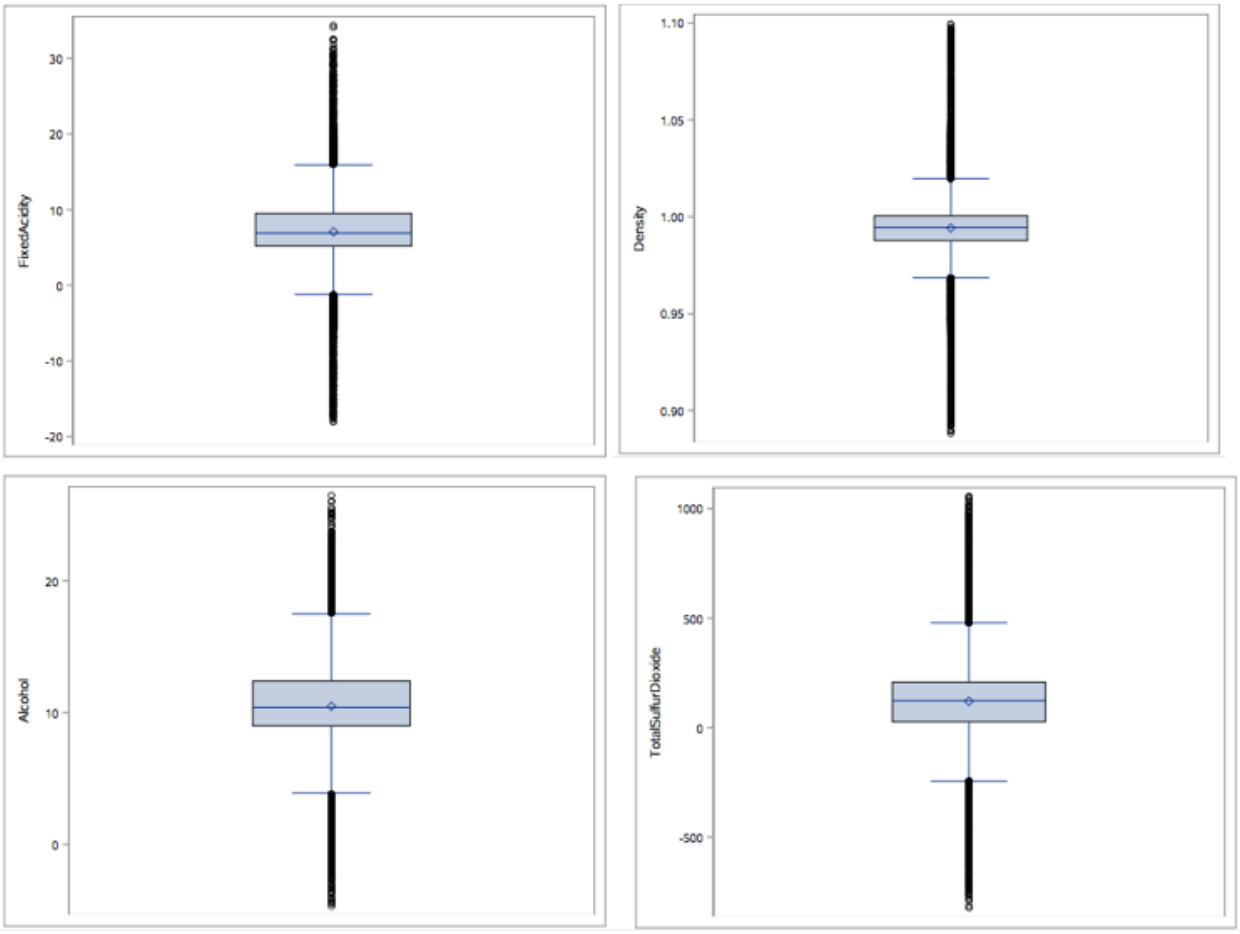
Mean	Variance
3.0290739	3.7108945

Now that we've examined our response variable, we move on to exporing our set of predictor varibles. We can see in the table given below, that our data set contains a number of variables with missing values. This indicates that we have a total of seven variables with missing records; Alcohol, Chlorides, FreeSulfurDioxide, pH, ResidualSugar, Sulphates, TotalSulfurDioxide and Stars all have missing values that will need to be accounted for in order to use these variables in the model. These observations will need to be further examined to determine how to make the proper corrections in the our data preparation step.

Table 3: Missing Variables

Variable	N	N Missing
AcidIndex	12795	0
Alcohol	12142	653
Chlorides	12157	638
CitricAcid	12795	0
Density	12795	0
FixedAcidity	12795	0
FreeSulfurDioxide	12148	647
LabelAppeal	12795	0
pH	12400	335
ResidualSugar	12179	616
STARS	9436	3359
Sulphates	11585	1210
TotalSulfurDioxide	12113	682
VolatileAcidity	12795	0

In addition to these missing observations, it also appears that there are a number of variables with a large percentage of outlier values at both extremes. The boxplots below showcase the outlier values in each of the variables given below. The results from these visuals lead us to believe that these values should also be addressed in our data preparation, as we may want to adjust their high and low end values.


FixedAcidity
30
20
10
0
-10
-20
Density
1.10
1.05
1.00
0.95
0.90
Alcohol
20
10
0
TotalSulfurDioxide
1000
500
0
-500

Data Preparation

In exploring our data set it was determined that our data set contains a number of missing observations and outlier values that will need to be addressed. This step of our modeling process aims to identify how and where to properly adjust these values in a way that will maintain the consistency of the values in our original data set, while also allowing for increased accuracy in our actual model results. First, we will address our missing values.

In examining the missing data we identified that the missing values for STARS showed a strong negative correlation of -0.57158 and a significant p-value of <0.0001. to our target for wine sales. This allows us to assume that M_STARS may be a strong predictor and have decided to keep this variable in our data set. The remaining set of missing variables did not appear to have the same potential for predictive in power in explaining our resoponse, so we have decided to drop these variables to be used in our final model.

Next, we move to our treatment of outlier values which we identified in the following variables: FixedAcidity, Volatiile Acidity, ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, Density, pH, Sulphates, and Alcohol. Each of these variables showed a large quantity of extreme values, and we have decided to trim these variable values by 10% of their maximium and minimum values. We can see in the set of boxplots below that this technique was moderately successful in decreasing our outlier values. Density stil appears to contain some potential outliers, but for the sake of model complexity we have decided to keep these values for now, and move forward with the model building process.

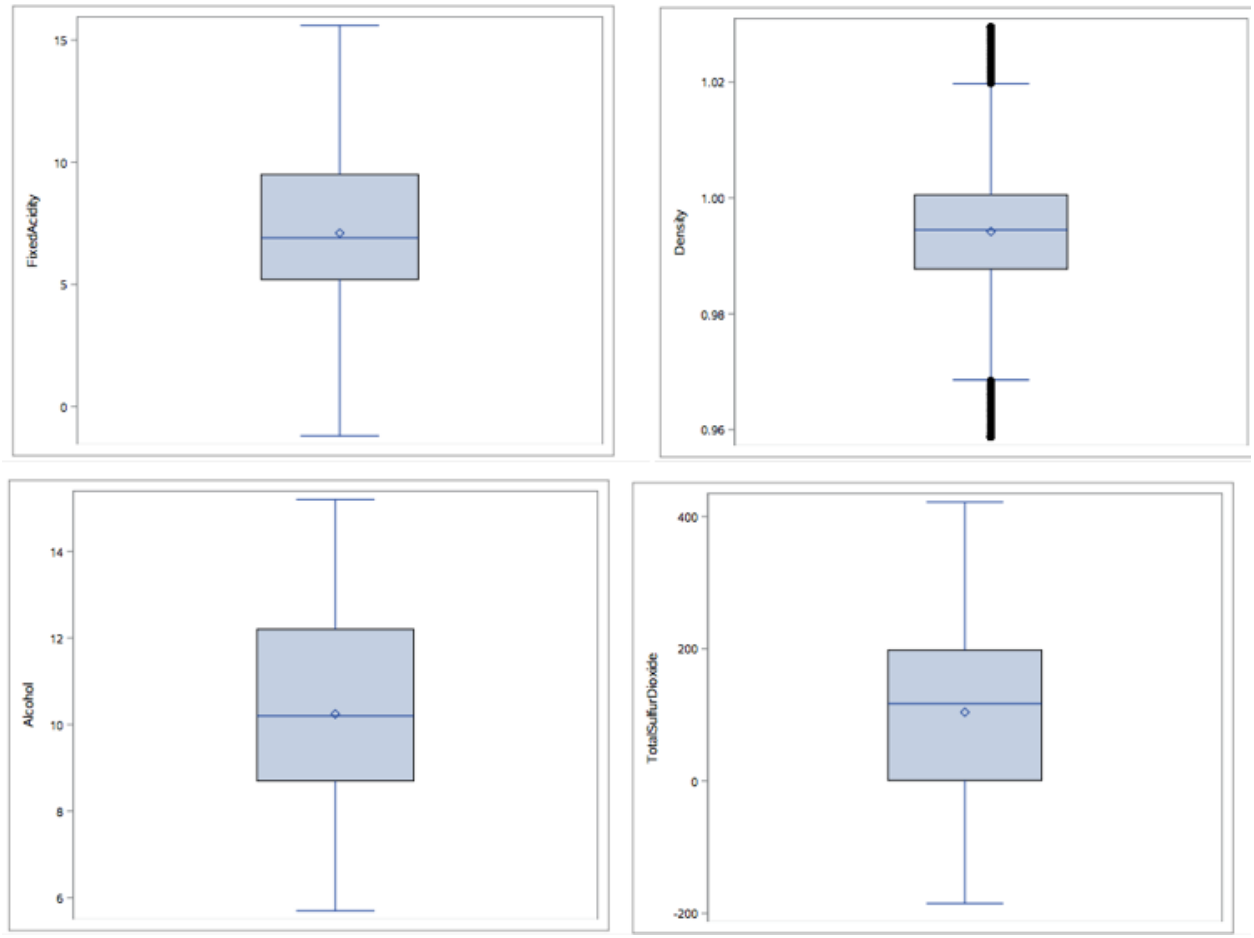


Figure 2: Transformed Variable Boxplot Distributions

After making these changes, our final prediction set contains a total of 15 variables. The summary statistics and response correaltions for our updated set of predictor variables is given in the table below.

Variable	N	N Miss	Mean	Median	Variance	Std Dev
TARGET	12795	0	3.0291	3.0000	3.7109	1.9264
FixedAcidity	12795	0	7.0757	6.9000	39.9126	6.3176
VolatileAcidity	12795	0	0.3241	0.2800	0.6147	0.7840
CitricAcid	12795	0	0.3084	0.3100	0.7432	0.8621
Density	12795	0	0.9942	0.9945	0.0007	0.0265
LabelAppeal	12795	0	-0.0091	0.0000	0.7940	0.8911
AcidIndex	12795	0	7.7727	8.0000	1.7528	1.3239
IMP_RES	12795	0	5.4187	4.9000	1084.1795	32.9269
IMP_CHLORIDES	12795	0	0.0548	0.0480	0.0964	0.3104
IMP_FREE_SD	12795	0	30.8456	30.8456	20997.5997	144.9055
IMP_TOTAL_SD	12795	0	120.7142	120.7142	50916.7285	225.6474
IMP_pH	12795	0	3.2076	3.2076	0.4477	0.6691
IMP_SULPHATES	12795	0	0.5271	0.5271	0.7867	0.8870
IMP_Alcohol	12795	0	10.4892	10.4892	13.1874	3.6314
IMP_STARS	12795	0	2.0308	2.0000	0.6011	0.7753
M_STARS	12795	0	0.2625	0.0000	0.1936	0.4400

Table 4: Predictor Variable Correlations to Cases of Wine Sold

Variable	Correlation	$Pr >
m_stars	-0.57158	< 0.0001
IMP_STARS	0.40013	< 0.0001
LabelAppeal	0.35650	< 0.0001
acidindex	-0.24605	< 0.0001
volatileacidity	-0.08879	< 0.0001
IMP_ALCOHOL	0.06043	< 0.0001
imp_total_sd	0.05010	< 0.0001
fixedacidity	-0.04901	< 0.0001
imp_free_sd	0.04269	< 0.0001
imp_chlorides	-0.03724	< 0.0001
imp_sulphates	-0.03691	< 0.0001
density	-0.03552	< 0.0001
imp_res	0.01607	0.0691
imp_ph	-0.00928	0.2939
citricacid	0.00868	0.3260

In addition to analyzing the summary and correlation results a number of additional plots were produced to determine the visual relationship to one of the nine different levels for our categorical/ordinal response.

Model Development

After completing our data preparation, we can move on to selecting the variables to be utilized in our final models. In this stage of the process, we will utilize five separate types of modeling procedures and analyze

the results from each to determine the most accurate model to be selected for our scoring file. Our rationale behind utilizing each of these modeling procedures are highlighted below, along with our assessment of how well our data aligns with the assumption requirements:

- **Poisson Distribution**:
  - Appropriate when a large proportion of response values are less than or equal to zero, which our Target for wine sales appears to contain.
  - This technique assumes that there is equidispersion present. Our dataset does not appear to meet this assumption as the variance is higher than the mean in Target.
  - Utilzes the Log Identity function to predict counts in each of the nine levels of our resposne.
- **Negative Binomial Distribution**:
  - This technique assumes that the variance is greater than the mean, which are confirmed in our data set.
  - This type of model is assumed to lead to more precise coefficients and standard errors present in our results.
  - The model results of this model should closely resemble what we observe in the Poisson model. The differences between these results will appear in comparing the standard errors rather than the estimated counts.
- **Zero-Inflated Poisson Distribution**:
  - A Zero-Inflated Poisson Distribution is especially useful when there are an excess of zero-values in our dependent variable. This assumption also appears to be valid within our response.
  - However, our target values appear to possess extradispersion rather than an equal mean and variance.
- **Zero-Inflated Negative Binomial Distribution**:
  - This is also appropriate when there are an excess of zero-values present. Again, this was confirmed in our data.
  - Our data also meets the assumption of extradispersion which was also already identified.
- **OLS Regression**:
  - This type of modeling technique is most approriate for predicting probabilities or continuous values that are normally distributed.
  - Since we know that Target has a number of zero-values, we can interpret this as having a zero probability of occurence. This suggests that it may be best to develop this type of model in a two-step, hurdle approach.

The final set of predictors to be included in our final model are summarized in the table shown in the below section. The PROC GENMOD SAS Procedure was used with each modeling procedure listed above with a number of different combinations of variables. Ulimately, we felt that there were five total variables which indicated the strongest potential of predictive accuracy, while also limiting the complexity of our final model. We hope that this will give us the strongest ability to estimate the actual number of wine caes sold while also simplifying the interpretation of our model results.

Table 5: Final Set of Predictors to be Utilized in Model Development

Variable	Correlation	$Pr >
M_STARS	-0.57158	< 0.0001
IMP_STARS	0.40013	< 0.0001
LabelAppeal	0.35650	< 0.0001
acidindex	-0.24605	< 0.0001
IMP_ALCOHOL	0.06043	< 0.0001

The model formula and results from each of our five models are given below. We will discuss and summarize the results from each at the end of this section, where we will identify the strongest model for selection.

Poisson Distrubtion Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \epsilon$$

Where:

Table 6: Poisson Model Variables

In Model	In Data
Y is	target
X_1 is	LabelAppeal
X_2 is	acid_index
X_3 is	IMP_ALCOHOL
X_4 is	IMP_STARS
X_5 is	M_STARS

Table 7: Poisson Model Analysis Of Maximum Likelihood Parameter Estimates

Parameter	Set	Estimate	Std Error	Wald Chi-Square	Pr > ChiSq
intercept		1.3283	0.0518	656.79	<.0001
LabelAppeal	-2	-0.6958	0.0424	269.03	<.0001
LabelAppeal	-1	-0.4597	0.0250	338.98	<.0001
LabelAppeal	0	-0.2702	0.0228	139.87	<.0001
LabelAppeal	1	-0.1377	0.0232	35.38	<.0001
LabelAppeal	2	0.0000	0.000	.	.
AcidIndex		-0.0809	0.0045	328.69	<.0001
IMP_ALCOHOL		.0040	0.0017	682.89	<.0001
IMP_STARS	1	-0.2409	0.0216	149.78	<.0001
IMP_STARS	2	-0.1207	0.0199	35.77	<.0001
IMP_STARS	3	0.00000	0.0202	.	.
IMP_STARS	4	0.00000	0.000	.	.
M_STARS	0	1.0923	0.0182	3597.47	<.0001
M_STARS	1	0.0000	0.000	.	.
Scale	0	1.0000	0.0000	.	.

Table 8: Poisson Model Criteria for Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	1.30E+004	13695.1767	1.0713
Scaled Deviance	1.30E+004	13695.1767	1.0713
Pearson Chi-Square	1.30E+004	11320.7636	0.8855
Scaled Pearson X2	1.30E+004	11320.7636	0.8855
Log Likelihood		8778.5721	
Full Log Likelihood		-22818.5992	
AIC (smaller is better)		45659.1984	
AICC (smaller is better)		45659.2191	
BIC (smaller is better)		45741.223	

Negative Binomial Distrubtion Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \epsilon$$

Where:

Table 9: Negative Binomial Model Variables

In Model	In Data
Y is	target
X_1 is	LabelAppeal
X_2 is	acid_index
X_3 is	IMP_ALCOHOL
X_4 is	IMP_STARS
X_5 is	M_STARS

Table 10: Negative Binomial Distribution Model Analysis Of Maximum Likelihood Parameter Estimates

Parameter	Set	Estimate	Std Error	Wald Chi-Square	Pr > ChiSq
intercept		1.3283	0.0518	656.79	<.0001
LabelAppeal	-2	-0.6965	0.0424	269.03	<.0001
LabelAppeal	-1	-0.4593	0.0250	338.98	<.0001
LabelAppeal	0	-0.2698	0.0228	139.87	<.0001
LabelAppeal	1	-0.1369	0.0232	35.38	<.0001
LabelAppeal	2	0.0000	0.000	.	.
AcidIndex		-0.0809	0.0045	328.69	<.0001
IMP_ALCOHOL		.0040	0.0017	682.89	<.0001
IMP_STARS	1	-0.5621	0.0216	149.78	<.0001
IMP_STARS	2	-0.2409	0.0199	35.77	<.0001
IMP_STARS	3	-0.1199	0.0202	.	.
IMP_STARS	4	0.0000	0.0000	.	.
M_STARS	0	1.0923	0.0182	3597.47	<.0001
M_STARS	1	0.0000	0.0000	.	.
Dispersion	0	0.0001	0.0000	.	.

Table 11: Negative Binomial Model Criteria for Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	1.30E+004	13695.1767	1.0713
Scaled Deviance	1.30E+004	13695.1767	1.0713
Pearson Chi-Square	1.30E+004	11320.7545	0.8855
Scaled Pearson X2	1.30E+004	11320.7545	0.8855
Log Likelihood		8778.5721	
Full Log Likelihood		-22818.5992	
AIC (smaller is better)		45661.1984	
AICC (smaller is better)		45661.2228	
BIC (smaller is better)		45750.6801	

Zero-Inflated Poisson Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \epsilon$$

Where:

Table 12: Zero-Inflated Poisson Variables

In Model	In Data
Y is	target
X_1 is	LabelAppeal
X_2 is	acid_index
X_3 is	IMP_ALCOHOL
X_4 is	IMP_STARS
X_5 is	M_STARS

Table 13: Zero-Inflated Poisson Model Analysis Of Maximum Likelihood Parameter Estimates

Parameter	Set	Estimate	Std Error	Wald Chi-Square	Pr > ChiSq
intercept		1.7861	0.0541	1088.44	<.0001
LabelAppeal	-2	-0.9682	0.0439	487.07	<.0001
LabelAppeal	-1	-0.6001	0.0260	534.35	<.0001
LabelAppeal	0	-0.3393	0.0236	207.25	<.0001
LabelAppeal	1	-0.1561	0.0238	43.18	<.0001
LabelAppeal	2	0.0000	0.000	.	.
AcidIndex		-0.0201	0.0049	18.07	<.0001
IMP_ALCOHOL		.0076	0.0018	17.78	<.0001
IMP_STARS	1	-0.4107	0.0231	316.21	<.0001
IMP_STARS	2	-0.1970	0.0200	97.38	<.0001
IMP_STARS	3	-0.1034	0.0202	26.23	.
IMP_STARS	4	0.0000	0.0000	.	.
M_STARS	0	0.1846	0.0196	88.51	<.0001
M_STARS	1	0.0000	0.0000	.	.
Scale	0	1.0000	0.0000	.	.

Table 14: Zero-Inflated Model Criteria for Asessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	1.30E+004	41909.8340	
Scaled Deviance	1.30E+004	41909.8340	
Pearson Chi-Square	1.30E+004	6093.1291	0.4767
Scaled Pearson X2	1.30E+004	6093.1291	0.4767
Log Likelihood		10642.2543	
Full Log Likelihood		-20954.9170	
AIC (smaller is better)		41937.8340	
AICC (smaller is better)		41937.8669	
BIC (smaller is better)		42042.2294	

Zero-Inflated Negative Binomial Distrubiton Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \epsilon$$

Where:

Table 15: Zero-Inflated Negative Binomial Model Variables

In Model	In Data
Y is	target
X_1 is	LabelAppeal
X_2 is	acid_index
X_3 is	IMP_ALCOHOL
X_4 is	IMP_STARS
X_5 is	M_STARS

Table 16: Zero-Inflated Negative Binomial Distribution Model Analysis Of Maximum Likelihood Parameter Estimates

Parameter	Set	Estimate	Std Error	Wald Chi-Square	Pr > ChiSq
intercept		1.7807	0.0541	1078.73	<.0001
LabelAppeal	-2	-0.9734	0.0439	490.14	<.0001
LabelAppeal	-1	-0.6035	0.0260	537.60	<.0001
LabelAppeal	0	-0.3411	0.0236	208.20	<.0001
LabelAppeal	1	-0.1568	0.0238	43.26	<.0001
LabelAppeal	2	0.0000	0.000	.	.
AcidIndex		-0.0201	0.0049	16.53	<.0001
IMP_ALCOHOL		.0077	0.0018	18.09	<.0001
IMP_STARS	1	-0.4004	0.0231	301.85	<.0001
IMP_STARS	2	-0.1957	0.0200	95.18	<.0001
IMP_STARS	3	-0.1031	0.0202	25.81	.
IMP_STARS	4	0.0000	0.0000	.	.
M_STARS	0	0.1831	0.0196	86.75	<.0001
M_STARS	1	0.0000	0.000	.	.
Scale	0	1.0000	0.0000	.	.

Table 17: Zero-Inflated Negative Binomial Model Criteria for Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	1.30E+004	41966.8578	
Scaled Deviance	1.30E+004	41966.8578	
Pearson Chi-Square	1.30E+004	5987.0665	0.4684
Scaled Pearson X2	1.30E+004	5987.0665	0.4684
Log Likelihood		-20983.4289	
Full Log Likelihood		-20983.4289	
AIC (smaller is better)		41996.8578	
AICC (smaller is better)		41996.8954	
BIC (smaller is better)		42108.7099	

OLS Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 \epsilon$$

Where:

Table 18: OLS Regression Variables

In Model	In Data
Y is	target
X_1 is	LabelAppeal
X_2 is	acid_index
X_3 is	IMP_ALCOHOL
X_4 is	IMP_STARS
X_5 is	M_STARS

Table 19: OLS Regression Model Parameter Estimates

Parameter	Set	Estimate	Standard Error	t Value	P value
intercept		3.49983	0.08917	39.25	<.0001
LabelAppeal	1	0.46555	0.01371	33.95	<.0001
AcidIndex	1	-0.20546	0.00895	-22.95	<.0001
IMP_ALCOHOL	1	.01300	0.00399	3.26	<.0001
IMP_STARS	1	-0.78364	0.01572	49.84	0.0011
M_STARS	1	-2.26333	0.02698	-83.89	<.0001

Table 20: OLS Regression Model Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	25372	5074.42851	2935.84	<.0001
Error	12789	22105	1.72844		
Corrected Total	12794	47477			
	Root MSE	1.31470	R-Square	0.5344	
	Dependent Mean	3.20907	Adj R-Square	0.5342	
	Coeff Var	43.40279			

Model Results

The results above allow us to analyze the summary of fit results and by providing a number of various metrics for comparison. This also allows us to analyze the coefficients and their weighted impacted to the response in each different model. In this step, will identify any notable difference between models and confirm that each makes sense in the context of predicting the cases of wine sold.

In comparing our first two models, we found that the results from the Poisson and Negative Binomial model were very similar. Each of the coefficient estimates were nearly identical in value and direction. Additional we saw that the standard error figures were also close to equal in value. We will need to review the AIC, AICC and BIC to further examine the strength of these models.

The Zero-Inflated Poisson distribution yielded similar results to the Zero-Inflated Negative Binomial model. The coefficient estimates for each variable were similar, but not as close as the comparison in the poisson and neg bin model. It appears the Zero-Inflated Poisson model showed a stronger Wald Chi-Sqare for nearly each different variable input. Although these numbers are also comparable to the Zero-Inflated Negative Binomial model.

We also identified that each variable showed a statistically significant correlation to our target variable in every model procedure discussed above. This might allow us to conclude that our target variables are good predictors, but that will be confirmed when we score the model against the actual results. Additionally, in reviewing our OLS Regression model results we can see this model shows statistical significance, and a fairly strong a R-Score of 0.5344. This can be interpreted as this model explains roughly 53.44% of the variability of our resonse value results. Although this result is fairly strong, we will likely decide not to move forward in selecting this type of model due to the fact that the distributions found within our data failed to meet a number of assumptions necessary to ensure accuracy within the results.

Next, The table below gives the AIC, AICC, and BIC scores which help us understand the initial strength of the first four models we developed. When we review these values, it is important to note that a lower score indicates a better likelyhood of accurate model results. Although there is no one single figure that will tell us what model will work best in the real world, these scores provide a generally benchmark do equally compare multiple models against each other. These model scores will be compared against our first four models that were built. The OLS Regression model will yield a different set of summary fit statistics, so we will compare the R-Square value and determine whether or not this provides higher accuracy than the other four models.

Table 21: Comparison Summary of Model Scoring Criteria Results

Modeling Procedue	AIC	AICC	BIC
Poisson	45659.1984	45659.2191	45741.223
Negative Binomial	45661.1984	45661.2228	45750.6801
Zero-Inflated Poisson	41937.8340	41937.8669	42042.2294
Zero-Inflated Negative Binomial	41996.8578	41996.8954	42108.7099

Model Selection

In comparing the results from our first four models, we saw similar coefficient estimates, standard error rates, and Wald Chi-Square scores between the Poisson and Negative Binomial procedures, and separately in comparison of the Zero-Inflated Poisson and Zero-Negative Binomial models. Although the first two models meet enough assumptions of these procedures to promote valid results, their Maximum Likelyhood Scores were a bit higher than the Zero-Inflated models. In addition to containing stronger scores here, we also feel that the distributions of our initial data set better match the requirements for a Zero-Inflated model. This is especially true given that we identified that our reponse contains zero-values, representing no wine sales, in over 20% of the observations. In the final comparison of these two models, we can see that the Zero-Inflated Poisson model possesses lower AIC, AICC and BIC scores than the Zero-Inflated Negative Binomial Distribution model. The differnces in these score values are minimal which probably means they will yield to similar results. However, we believe that the distributions found within our data set better meet the assumptions of the Zero-Inflated Negative Binomial model. In going back to our exploratory data analysis, we identified that the variance of our target exceeded the mean which ultimately drove us to select this model.

Conclusion

The results from this research exploration enabled us to measure and compare the results from five different modeling procedures. The summary-of-fit and maximum likleyhood scores were comparable across multiple models, despite the fact that they didn't necesarrily meet every single assumption of that technique. These procedures also revealed that there are two qualitative varaibles showing significant and strong correlation to the quantity of wine sold in each of the five models. As a wine manufacturer, these results can be utilized to better market their product to increase sales. For example, in each of these models, we saw that LabelAppeal showed the strongest correlation to the success of wine sales. If we identify that our wine products are receving negative ratings from consumers, we may want to develop and test new packaging colors and types. The results from this analysis may allow us to confidently recommend an increased investment in packaging innovation. Further research on this subject would be needed in validating these decisions.

BINGO BONUS

- **This document was created in R-Markdown using the knitr and pandoc packages**:

- This type of formatting is useful in data science becuase it allows the researcher to weave their r-code right into their actual written analysis. Although there is no code specifically shown within the document output for this particular report, this type of programming provides a method for creating reproduceable reports while also providing a professional and well formatted PDF output.
- **R Coding and Decision Trees**:
- For example, I can run the following piece of code to showcase the code utilized for a particular report output. This would be great for collaborating across different teams and provides a more seamless experience when multiple people are working on the same type of project and need to share code. This would decrease time wasted in copying and pasting, or debugging newly produced code everytime.

```
library(stringr)
library(ggplot2)
library(lattice)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(plotly)
```

```
## Warning: package 'plotly' was built under R version 3.4.1
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
library(moments)
library(tidyverse)
```

```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## as.difftime(): lubridate, base
## date():        lubridate, base
## filter():      dplyr, plotly, stats
## intersect():   lubridate, base
## lag():         dplyr, stats
## setdiff():     lubridate, base
## union():       lubridate, base
```

```
library(car)
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(rpart)
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 3.4.1
```

```
library(party)
```

```
## Warning: package 'party' was built under R version 3.4.1
```

```
## Loading required package: grid
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: modeltools
```

```
## Loading required package: stats4
```

```
## Loading required package: strucchange
```

```
## Warning: package 'strucchange' was built under R version 3.4.1
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
##
## Attaching package: 'strucchange'
```

```
## The following object is masked from 'package:stringr':
##
##     boundary
```

```
WINE<- read.csv('wine.csv',stringsAsFactors = FALSE)
summary(WINE)
```

```
##     ï..INDEX         TARGET       FixedAcidity     VolatileAcidity
##  Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900
##  1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300
##  Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800
##  Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241
##  3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400
##  Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800
##
##    CitricAcid      ResidualSugar        Chlorides       FreeSulfurDioxide
##  Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00
##  1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00
##  Median : 0.3100   Median :   3.900   Median : 0.0460   Median :  30.00
##  Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85
##  3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00
##  Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00
##                    NA's   :616        NA's   :638       NA's   :647
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
##  1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800
##  Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000
##  Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
##  3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600
##  Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
##  NA's   :682                         NA's   :395     NA's   :1210
##     Alcohol       LabelAppeal          AcidIndex          STARS
##  Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median :10.40   Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##  NA's   :653                                          NA's   :3359
```

```
str(WINE)
```

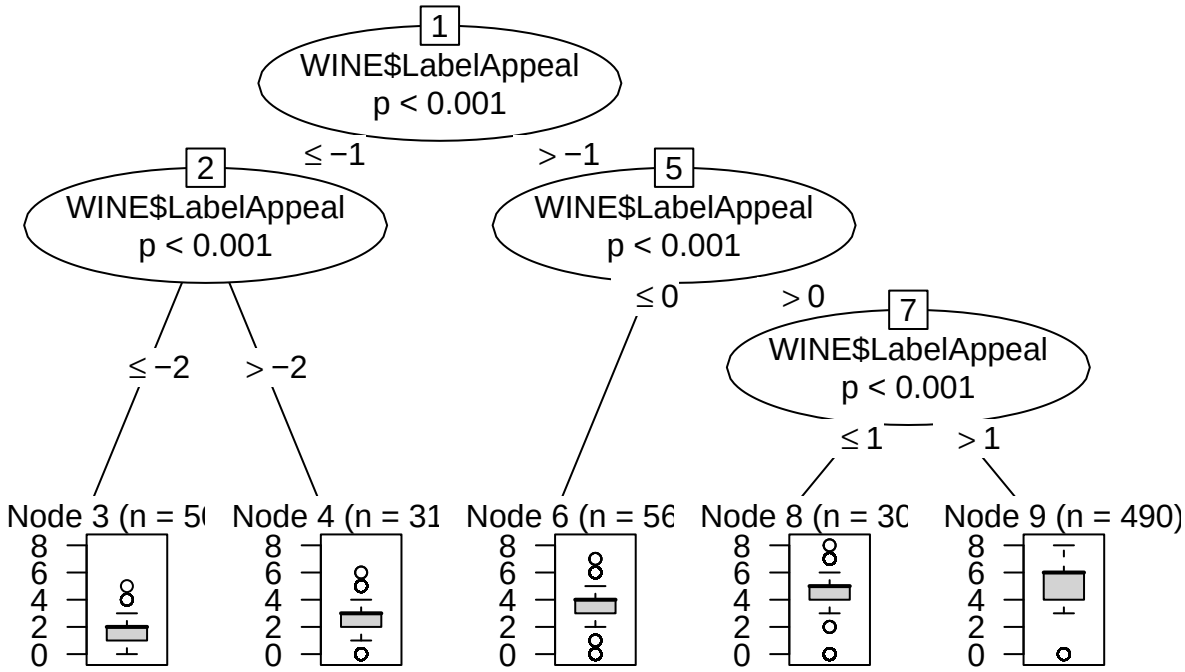
```
## 'data.frame':    12795 obs. of  16 variables:
##  $ ï..INDEX          : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET            : int  3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity      : num  3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity   : num  1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
##  $ CitricAcid        : num  -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
```

```
##  $ ResidualSugar     : num  54.2 26.1 14.8 18.8 9.4 ...
##  $ Chlorides         : num  -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...
##  $ FreeSulfurDioxide : num  NA 15 214 22 -167 -37 287 523 -213 62 ...
##  $ TotalSulfurDioxide: num  268 -327 142 115 108 15 156 551 NA 180 ...
##  $ Density           : num  0.993 1.028 0.995 0.996 0.995 ...
##  $ pH                : num  3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
##  $ Sulphates         : num  -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...
##  $ Alcohol           : num  9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
##  $ LabelAppeal       : int  0 -1 -1 -1 0 0 0 1 0 0 ...
##  $ AcidIndex         : int  8 7 8 6 9 11 8 7 6 8 ...
##  $ STARS             : int  2 3 3 1 2 NA NA 3 NA 4 ...
```

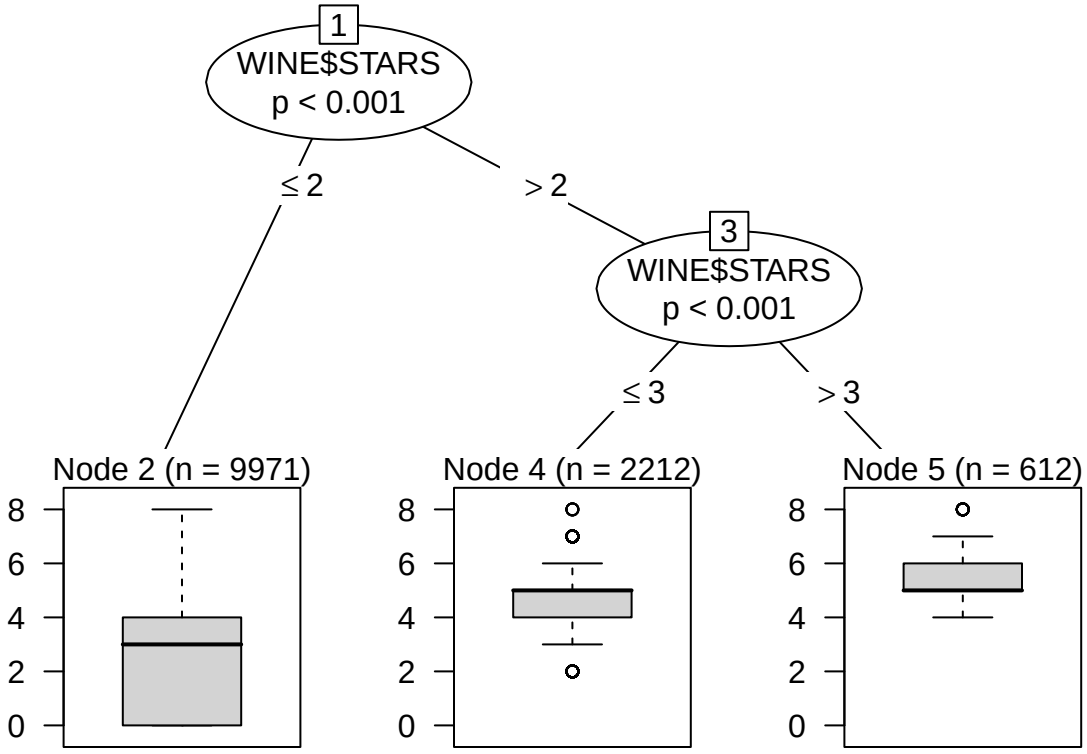
```
#Impute Variables with NA Values
WINE$ResidualSugar<- recode(WINE$ResidualSugar,"NA=5.4187331")
WINE$Chlorides<- recode(WINE$Chlorides,"NA=0.0548225")
WINE$FreeSulfurDioxide<- recode(WINE$FreeSulfurDioxide,"NA=30.8455713")
WINE$TotalSulfurDioxide<- recode(WINE$TotalSulfurDioxide,"NA=120.7142326")
WINE$pH<-recode(WINE$pH,"NA=3.2076282")
WINE$Sulphates<- recode(WINE$Sulphates,"NA=0.5271118")
WINE$Alcohol<- recode(WINE$Alcohol,"NA=10.4892363")
WINE$STARS<- recode(WINE$STARS,"NA=2")

summary(WINE)
```

```
##     ï..INDEX         TARGET       FixedAcidity     VolatileAcidity
##  Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900
##  1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300
##  Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800
##  Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241
##  3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400
##  Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800
##    CitricAcid      ResidualSugar        Chlorides        FreeSulfurDioxide
##  Min.   :-3.2400   Min.   :-127.800   Min.   :-1.17100   Min.   :-555.00
##  1st Qu.: 0.0300   1st Qu.:   0.900   1st Qu.: 0.00000   1st Qu.:   5.00
##  Median : 0.3100   Median :   4.900   Median : 0.04800   Median :  30.85
##  Mean   : 0.3084   Mean   :   5.419   Mean   : 0.05482   Mean   :  30.85
##  3rd Qu.: 0.5800   3rd Qu.:  14.900   3rd Qu.: 0.12800   3rd Qu.:  64.00
##  Max.   : 3.8600   Max.   : 141.150   Max.   : 1.35100   Max.   : 623.00
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
##  1st Qu.:  34.0     1st Qu.:0.9877   1st Qu.:2.970   1st Qu.: 0.3400
##  Median : 120.7     Median :0.9945   Median :3.208   Median : 0.5271
##  Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
##  3rd Qu.: 198.0     3rd Qu.:1.0005   3rd Qu.:3.450   3rd Qu.: 0.7700
##  Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
##     Alcohol       LabelAppeal          AcidIndex          STARS
##  Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.: 9.10   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:2.000
##  Median :10.49   Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.031
##  3rd Qu.:12.20   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:2.000
##  Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000
```

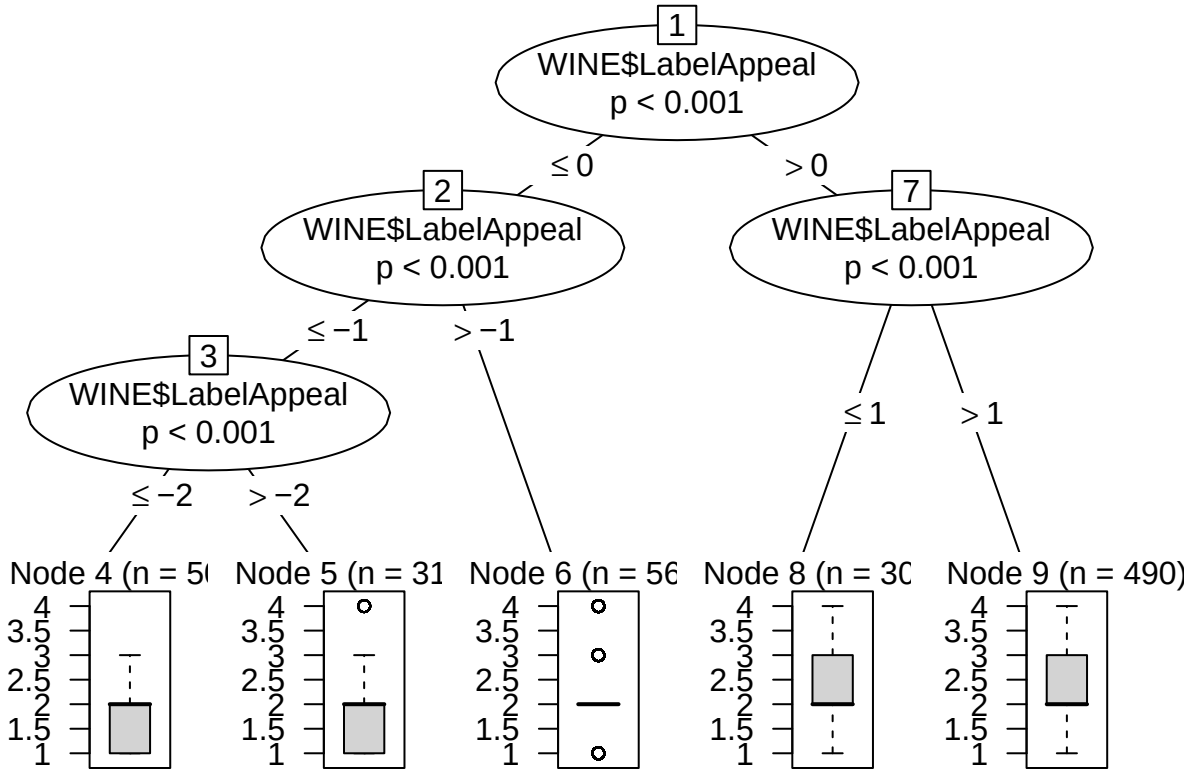
```
fit<- ctree(WINE$TARGET ~ WINE$LabelAppeal)
plot(fit, main="")
```



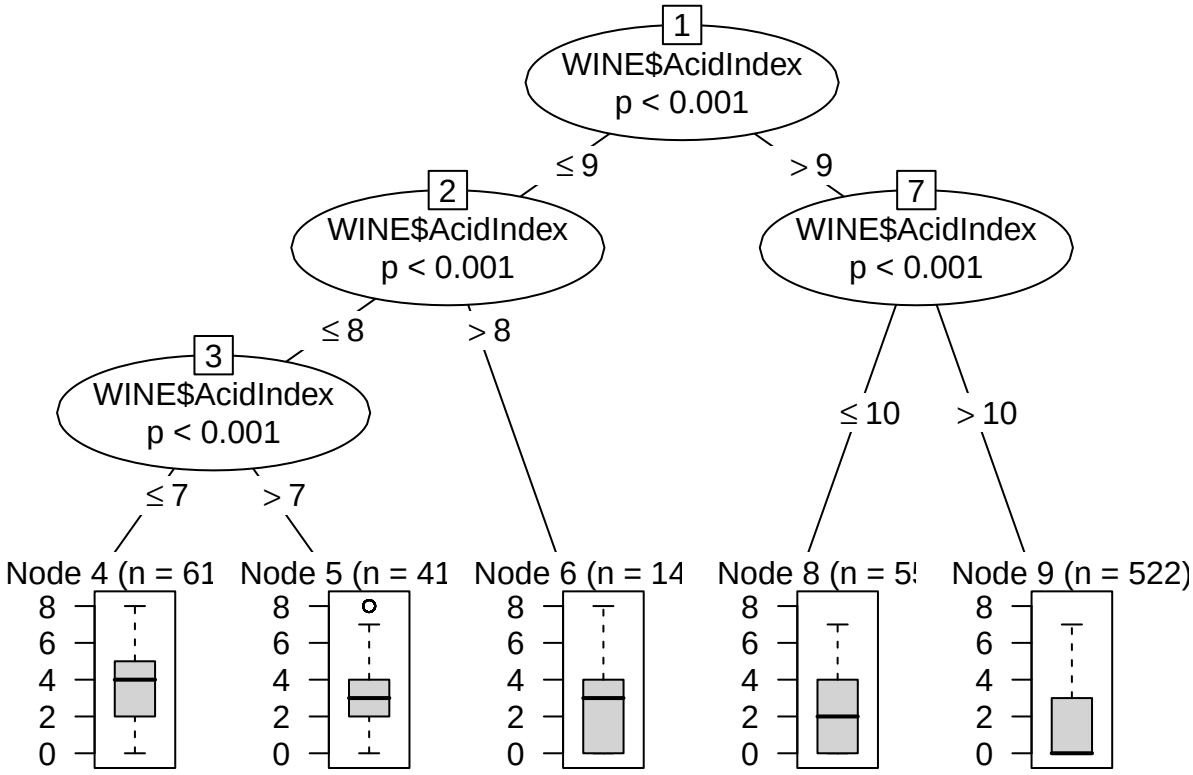

```
fit<- ctree(WINE$TARGET ~ WINE$STARS)
plot(fit)
```



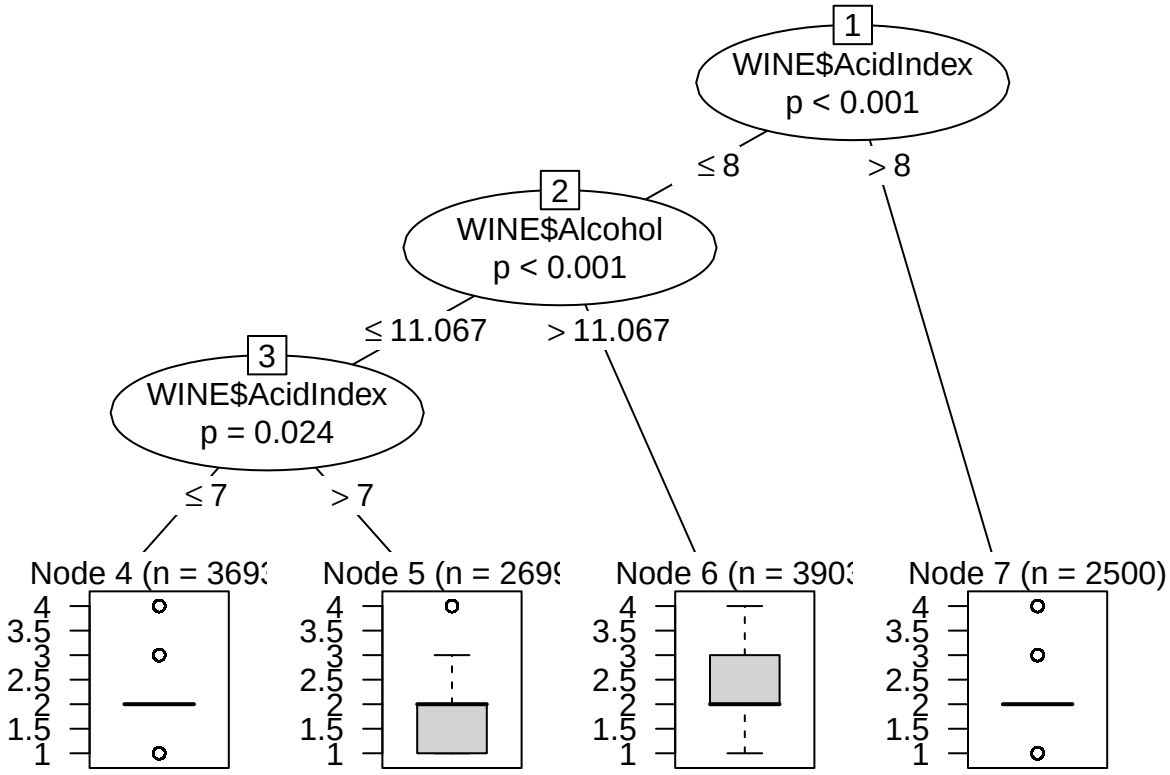

```
fit<- ctree(WINE$STARS ~ WINE$LabelAppeal)
plot(fit)
```




```
fit<- ctree(WINE$TARGET~ WINE$AcidIndex)
plot(fit)
```




```
fit<- ctree(WINE$STARS ~ WINE$AcidIndex + WINE$Alcohol)
plot(fit)
```



APPENDIX: SAS CODE

```
$ libname mydata '/sscc/home/s/sth932/WINE';

proc contents data=mydata.wine;
run;

proc means data=mydata.wine p10 mean median p90;
run;

proc sgplot data=mydata.wine;
vbox FixedAcidity;
run;

proc sgplot data=mydata.wine;
vbox FixedAcidity/group=target;
run;

proc sgplot data=mydata.wine;
vbox Density;
run;

proc sgplot data=mydata.wine;
vbox Density/group=target;
run;

proc sgplot data=mydata.wine;
vbox ALCOHOL;
```

```
run;

proc sgplot data=mydata.wine;
vbox ALCOHOL/group=target;
run;

proc sgplot data=mydata.wine;
vbox TOTALSULFURDIOXIDE;
run;

proc sgplot data=mydata.wine;
vbox TOTALSULFURDIOXIDE/group=target;
run;

proc sgplot data=mydata.wine;
vbox labelappeal;
run;

proc sgplot data=mydata.wine;
vbox labelappeal/group=target;
run;

data one (DROP=INDEX);
set mydata.wine;

if fixedacidity > 15.6 then do;
fixedacidity = 15.6;
end;

if fixedacidity < -1.2 then do;
fixedacidity = -1.2;
end;

if VolatileAcidity < -0.72 then do;
VolatileAcidity= -0.72;
end;

if VolatileAcidity > 1.35 then do;
VolatileAcidity= 1.35;
end;

if ResidualSugar>.481 then do;
ResidualSugar= .481;
end;

if ResidualSugar <-39.7 then do;
ResidualSugar=-39.7 ;
end;

if Chlorides< -.372 then do;
Chlorides= -.372;
end;

if Chlorides > .481 then do;
Chlorides=.481;
end;

if FreeSulfurDioxide< -171.0 then do;
FreeSulfurDioxide=-171.0;
end;
```

```
if FreeSulfurDioxide > 230 then do;
FreeSulfurDioxide= 230;
end;

if totalsulfurdioxide <-185 then do;
totalsulfurdioxide=-185;
end;

if totalsulfurdioxide > 422 then do;
totalsulfurdioxide=422;
end;

if density< .9587 then do;
density=.9587;
end;

if density > 1.0295then do;
density=1.0295;
end;

if pH<2.31 then do;
pH=2.31;
end;

if pH> 4.1 then do;
pH=4.1;
end;

if sulphates<-0.7 then do;
sulphates=-0.7;
end;

if sulphates > 1.77then do;
sulphates=1.77;
end;

if alcohol < 5.7 then do;
alcohol=5.7;
end;

if alcohol >15.2 then do;
alcohol=15.2;
end;

proc contents data=one;
run;

proc means data=one n nmiss mean median max std stderr var qrange;
run;

DATA two;
set one;

IMP_RES=RESIDUALSUGAR;
M_RES=0;
if RESIDUALSUGAR="." then do;
IMP_RES=5.4187331;
M_RES=1;
end;
```

```
IMP_CHLORIDES=CHLORIDES;
M_CHLORIDES=0;
if CHLORIDES="." then do;
IMP_CHLORIDES=0.0548225;
M_CHLORIDES=1;
end;

IMP_FREE_SD=FREESULFURDIOXIDE;
M_FREE_SD=0;
if FREESULFURDIOXIDE="." then do;
IMP_FREE_SD=30.8455713;
M_FREE_SD=1;
end;

IMP_TOTAL_SD=TOTALSULFURDIOXIDE;
M_TOTAL_SD=0;
if TOTALSULFURDIOXIDE="." then do;
IMP_TOTAL_SD=120.7142326;
M_TOTAL_SD=1;
end;

IMP_pH=pH;
M_pH=0;
if pH="." then do;
IMP_pH=3.2076282;
M_pH=1;
end;

IMP_SULPHATES=SULPHATES;
M_SULPHATES=0;
if SULPHATES="." then do;
IMP_SULPHATES=0.5271118;
M_SULPHATES=1;
end;

IMP_Alcohol=Alcohol;
M_Alcohol=0;
if Alcohol="." then do;
IMP_Alcohol=10.4892363;
M_Alcohol=1;
end;

IMP_STARS = STARS;
M_STARS = 0;
if STARS='.' then do;
IMP_STARS=2;
M_STARS=1;
end;

proc means data=two n nmiss mean median var std ndec=4;
run;

proc sgplot data=two;
vbox FixedAcidity;
run;

proc sgplot data=two;
vbox FixedAcidity/group=target;
```

```
run;

proc sgplot data=two; vbox Density; run;

proc sgplot data=two; vbox Density/group=target; run;

proc sgplot data=two; vbox ALCOHOL; run;

proc sgplot data=two; vbox ALCOHOL/group=target; run;

proc sgplot data=two; vbox TOTALSULFURDIOXIDE; run;

proc sgplot data=two; vbox TOTALSULFURDIOXIDE/group=target; run;

proc sgplot data=two; vbox labelappeal; run;

proc sgplot data=two; vbox labelappeal/group=target; run;

proc corr data=two rank; var FixedAcidity Density LabelAppeal AcidIndex IMP_TOTAL_SD IMP_Alcohol
IMP_STARS M_STARS; with TARGET; run;

proc genmod data=two; class labelappeal imp_stars M_STARS; model target = LabelAppeal AcidIndex
IMP_Alcohol IMP_STARS M_STARS/ link=log dist=poi; output out=two p=pr1;

proc genmod data=two; class labelappeal imp_stars M_STARS; model target = LabelAppeal AcidIndex
IMP_Alcohol IMP_STARS M_STARS/ link=log dist=nb; output out=two p=nbr1;

proc genmod data=two; class labelappeal imp_stars M_STARS; model target = LabelAppeal AcidIndex
IMP_Alcohol IMP_STARS M_STARS/ link=log dist=ZIP; zeromodel acidindex m_stars/link=logit; output
out=two p=zip1;

proc genmod data=two; class labelappeal imp_stars M_STARS; model target = LabelAppeal AcidIndex
IMP_Alcohol IMP_STARS M_STARS/ link=log dist=ZIP; zeromodel acidindex m_stars/link=logit; output
out=two p=zip1 pzero=zzip1;

proc genmod data=two; class labelappeal imp_stars M_STARS; model target = LabelAppeal AcidIndex
IMP_Alcohol IMP_STARS M_STARS/ link=log dist=ZINB; zeromodel acidindex m_stars/link=logit;
output out=two p=zinb1 pzero=zzinb1;

proc reg data=two; model target =LabelAppeal AcidIndex IMP_Alcohol IMP_STARS M_STARS; output
out=two p=yhat; run;

proc genmod data=two; class labelappeal imp_stars m_stars; model target = LabelAppeal AcidIndex
IMP_Alcohol IMP_STARS M_STARS/ link=identity dist=normal; output out=two p=ols1;

proc print data=two (obs=20); var target pr1 nbr1 zip1 zinb1 yhat ols1; run;
```