**Noé Flores**
**Wine Sales:**

**Introduction:**
The wine production industry and retailers of wine have become beneficiaries of predictive modeling. Wine producers are interested in knowing what makes a particular wine product a top seller. Is it all in the chemistry of what the wine is composed of or are there external variables, such as labeling that could play a significant role? We will attempt to answer this question by performing an initial exploratory analysis of our data to gather information on the various predictor variables available to us. We will follow through with any data preparation needed to ensure we have optimal variables to predict which characteristics of wine composition are most predictive with respect to future wine sales. Finally, using Poisson and Negative binomial models, we will build and compare several models to determine which has the highest predictive value, interpretability, and practicality based on AIC and BIC. These models will be constructed and compared with our end goal in mind to select the best possible fitting and most intuitive model.

**Data Exploration:**
An examination of the contents of our dataset produces the results below. There are 12796 records of commercially available wines and 16 variables, of which most are dedicated to the chemical properties of the wine, but some deal with qualitative measures as well. The first variable is (TARGET), which indicates the total number of sample cases of wine that were purchased by wine distributors and initial sampling of the wine. The more sample cases that were purchased, the higher the likelihood of a wine distributor's ability to sell their wine at high-end restaurants.

***Figure 1: Summary table of Data Contents.***

| Wine Data | Contents |
|---|---|
| **Observations** | 12796 |
| **Variables** | 16 |
| **Indexes** | 0 |
| **Observation Length** | 128 |
| **Deleted Observations** | 0 |
| **Compressed** | NO |
| **Sorted** | NO |

***Figure 2: Summary Data Dictionary and Variable Theoretical Effects.***

| VARIABLE NAME | DEFINITION | THEORETICAL EFFECT |
|---|---|---|
| **TARGET** | Number of Cases Purchased | None |
| **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. | Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales. |
| **ResidualSugar** | Residual Sugar of wine | |
| **STARS** | Wine rating: 4 Stars = Excellent, 1 Star = Poor | A high number of stars suggests high sales |
| **Sulphates** | Sulfate content of wine | |
| **TotalSulfurDioxide** | Total Sulfur Dioxide of Wine | |
| **VolatileAcidity** | Volatile Acid content of wine | |
| **pH** | pH of wine | |

Figure .2 contains a breakdown of the observations and shows the 14 measurable variables and their theoretical impact on the eventual sales of wine. As you can see, from the table above, we don't have a very robust grasp on what the theoretical effect on wine sales would be for the majority of our variables. While label appeal and Stars are somewhat self-explanatory, we will focus our efforts on trying to determine what effect if any, the other variables have on the wine sales.

We have some basic summary statistics of our data set in Figure.3. We use summary statistics to examine the data and look for any possible issues such as missing data. One rather massive observation is the number of missing values for (STARS). The central tendency information is useful to get a sense of how the variable data is spread. The mean and median help us understand the distribution of our variables. If we see numbers that are relatively close to each other, it's indicative that the data is likely to be closely divided around the mean, or close to symmetrical. If the data is spread further apart, it's indicative that the distribution for that particular variable is not normally distributed. These measures also help point out potential outliers. One of our variables, (TotalSulferDioxide), has a range of 1,880 which is unusually large when compared to the other variables. The same could be said of the range for (FreeSulferDioxide).

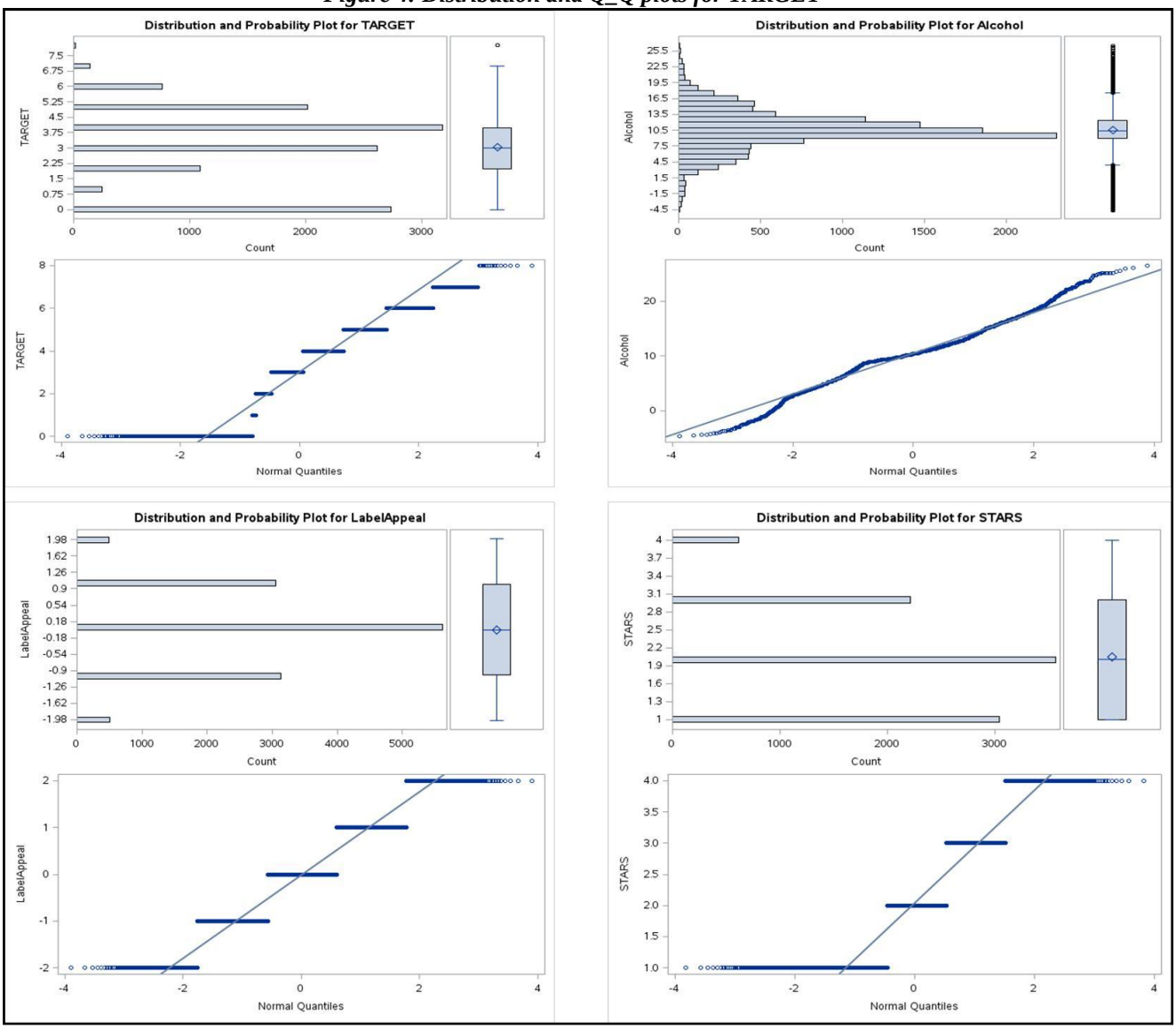
***Figure 3: Summary Statistics.***

| Variable | Minimum | Maximum | Range | Mean | Median | Std Dev | Variance | N | N Miss |
|---|---|---|---|---|---|---|---|---|---|
| **TARGET** | 0 | 8 | 8 | 3.03 | 3 | 1.93 | 3.71 | 12795 | 0 |
| **AcidIndex** | 4 | 17 | 13 | 7.77 | 8 | 1.32 | 1.75 | 12795 | 0 |
| **Alcohol** | -4.7 | 26.5 | 31.2 | 10.49 | 10.4 | 3.73 | 13.90 | 12142 | 653 |
| **Chlorides** | -1.17 | 1.35 | 2.52 | 0.05 | 0.046 | 0.32 | 0.10 | 12157 | 638 |
| **CitricAcid** | -3.24 | 3.86 | 7.1 | 0.31 | 0.31 | 0.86 | 0.74 | 12795 | 0 |
| **Density** | 0.89 | 1.10 | 0.21 | 0.99 | 0.99 | 0.03 | 0.00 | 12795 | 0 |
| **FixedAcidity** | -18.1 | 34.4 | 52.5 | 7.08 | 6.9 | 6.32 | 39.91 | 12795 | 0 |
| **FreeSulfurDioxide** | -555 | 623 | **1,178** | 30.85 | 30 | 148.71 | 22116.02 | 12148 | 647 |
| **LabelAppeal** | -2 | 2 | 4 | -0.01 | 0 | 0.89 | 0.79 | 12795 | 0 |
| **ResidualSugar** | -127.8 | 141.15 | 268.95 | 5.42 | 3.9 | 33.75 | 1139.02 | 12179 | 616 |
| **STARS** | 1 | 4 | 3 | 2.04 | 2 | 0.90 | 0.81 | 9436 | **3,359** |
| **Sulphates** | -3.13 | 4.24 | 7.37 | 0.53 | 0.5 | 0.93 | 0.87 | 11585 | 1,210 |
| **TotalSulfurDioxide** | -823 | 1057 | **1,880** | 120.71 | 123 | 231.91 | 53783.74 | 12113 | 682 |
| **VolatileAcidity** | -2.79 | 3.68 | 6.47 | 0.32 | 0.28 | 0.78 | 0.61 | 12795 | 0 |
| **pH** | 0.48 | 6.13 | 5.65 | 3.21 | 3.2 | 0.68 | 0.46 | 12400 | 395 |

To get a better understanding of our data, we examined the histogram and Q_Q plots for our variables. In Figure.4 we have four of the variables and the accompanying Q_Q plots. On the upper left-hand side, we can see that the (TARGET) appears to have a fairly normal distribution, but we also see some clear signs of zero spiking. This makes sense given the nature of the question we are trying to answer in this analysis. On the top right of Figure.4, Alcohol appears to have a fairly normal distribution with some clear center spikes and the shape of the curve indicating a smaller standard deviation. The majority of our variables follow this pattern. However, we do have some exceptions.

On the bottom of Figure.4, we see the information for (LabelAppeal) and (STARS). These variables are considered to be ordinal, and are used to rank and order preferences, such as one's preference for one wine over the other or one label versus another. These are more qualitative than quantitative, and we may need to consider modeling these variables differently.

***Figure 4: Distribution and Q_Q plots for TARGET***



Our ambition is to build a predictive model for the number of cases of wine that will be sold, given specific properties and characteristics of the wine. We can quickly attempt to identify possible predictor variables for Target sales by examining the correlation of those particular wine characteristics and properties to Target Sales. The correlation information is displayed below.

***Figure 5: Summary Correlation matrix to TARGET_WINS.***

| Correlation to Target | Correlation | P-Value | Observations |
|---|---|---|---|
| AccidIndex | -0.25 | <.0001 | 12795 |
| Alcohol | 0.06 | <.0001 | 12142 |
| Chlorides | -0.04 | <.0001 | 12157 |
| CitricAcid | 0.01 | **0.326** | 12795 |
| Density | -0.04 | <.0001 | 12795 |
| FixedAcidity | -0.05 | <.0001 | 12795 |
| FreeSulferDioxide | 0.04 | <.0001 | 12148 |
| LabelAppeal | 0.36 | <.0001 | 12795 |
| ResidualSugar | 0.02 | 0.069 | 12179 |
| Stars | 0.56 | <.0001 | 9436 |
| Sulphates | -0.04 | <.0001 | 11585 |
| TotalSulferDioxide | 0.05 | <.0001 | 12113 |
| VolatileAcidity | -0.09 | <.0001 | 12795 |
| pH | -0.01 | **0.293** | 12400 |

Stars, Label-Appeal, and Alcohol, stand out as the most influential positive predictors for our Target. Acid-Index, on the other hand, has the strongest negative correlation to Target. For the majority of our variables, the P-Values indicate that they are statistically significant. The threshold typically used for statistical significance is .05. Residual Sugar finds itself slightly above the .05 threshold at a P-Value of .069, but metrics of sugar content in wine are likely essential factors to consider when trying to determine which properties make wine most appealing to consumers. We did find two variables that were clearly above and beyond our desired statistical significance level. CitricAcid and pH are highlighted in Figure.5, and it's clear that their P-Values significantly deviate from our threshold of .05.

**Data Preparation:**

Our exploratory analysis revealed that we had several variables we need to address and prepare before we can proceed. The variables listed below have null or missing values and in the case of (FreeSulferDioxide) and (TotalSulferDioxide), reasonably extreme ranges. Also of note, (STARS) has a relatively significant amount of missing values.

***Figure 6: Summary of Variables with missing data.***

| Variable | Minimum | Maximum | Range | Mean | N | N Miss |
|---|---|---|---|---|---|---|
| Alcohol | -4.7 | 26.5 | 31.2 | 10.49 | 12142 | **653** |
| Chlorides | -1.17 | 1.35 | 2.522 | 0.05 | 12157 | **638** |
| CitricAcid | -3.24 | 3.86 | 7.1 | 0.31 | 12795 | |
| FreeSulfurDioxide | -555 | 623 | **1178** | 30.85 | 12148 | **647** |
| ResidualSugar | -127.8 | 141.15 | 268.95 | 5.42 | 12179 | **616** |
| STARS | 1 | 4 | 3 | 2.04 | 9436 | **3359** |
| Sulphates | -3.13 | 4.24 | 7.37 | 0.53 | 11585 | **1210** |
| TotalSulfurDioxide | -823 | 1057 | **1880** | 120.71 | 12113 | **682** |
| pH | 0.48 | 6.13 | 5.65 | 3.21 | 12400 | **395** |

You can also see in the table in Figure.6 above that I have included (CitricAcid), even though it doesn't have any missing variables. The reason for this is to draw attention to the fact that we will be removing that particular variable along with pH from our future models. Each of these variables lacked to exhibit any statistical significance with respect to our Target variable. For this reason, we felt it was prudent to remove these variables completely.

**Missing Values:**
The variables with missing values in the dataset will be replaced by their respective means. The large number of observations for each variable, (greater than 9400) indicate that simple imputation should suffice.

**Outliers:**
A conventional method used to deal with outliers in data is referred to as the two standard deviation rule. This rule advocate taking the standard deviation of our variable, multiplying it by 2, and subsequently adding and subtracting that total from the variable's mean. We incorporated this method on the variables (ResidualSugar), (FreeSulfurDioxide) and (TotalSulfurDioxide). The results of those variable imputations are displayed in Figure.7.

***Figure 7: Variable Imputation.***

| Variable | Minimum | Maximum | Range | Mean | N | N Miss |
|---|---|---|---|---|---|---|
| TARGET | 0 | 8 | 8 | 3.03 | 12795 | 0 |
| AcidIndex | 4 | 17 | 13 | 7.77 | 12795 | 0 |
| imp_Alcohol | -4.7 | 26.5 | 31.2 | 10.49 | 12795 | 0 |
| imp_Chlorides | -1.17 | 1.35 | 2.52 | 0.05 | 12795 | 0 |
| Density | 0.89 | 1.10 | 0.21 | 0.99 | 12795 | 0 |
| FixedAcidity | -18.1 | 34.4 | 52.5 | 7.08 | 12795 | 0 |
| imp_FreeSulfurDioxide | -267 | 328 | 595 | 30.77 | 12795 | 0 |
| LabelAppeal | -2 | 2 | 4 | -0.01 | 12795 | 0 |
| imp_ResidualSugar | -62 | 73 | 135 | 5.48 | 12795 | 0 |
| imp_STARS | 1 | 4 | 3 | 2.03 | 12795 | 0 |
| imp_Sulphates | -3.13 | 4.24 | 7.37 | 0.53 | 12795 | 0 |
| imp_TotalSulfurDioxide | -343 | 583 | 926 | 120.68 | 12795 | 0 |
| VolatileAcidity | -2.79 | 3.68 | 6.47 | 0.32 | 12795 | 0 |
| VolatileAcidity | -2.79 | 3.68 | 6.47 | 0.32 | 12795 | 0 |

**New Variable:**
We created a new variable called Alcohol Indicator, (Alcohol_Ind). This will serve as separation for wines with alcohol levels above and below the mean level of 10.5.

## Transformations:
The final change performed on our data comes in the form of a logistic transformation for our Density variable. Logistic transformations are typically used on independent variables in an effort to scale it and make it more normally distributed. This adjustment is also displayed in Figure.7.

**Modeling:**
We will be developing a variety of models to asses our wine data, starting with a linear regression model that incorporates an automated stepwise selection process. Linear regression models are a flexible regression tool that can give us a useful base model for comparison. The bulk of the models constructed will be Poisson and Negative Binomial regression models. Poisson and Negative Binomial regression are more appropriate when we are dealing with Count data, or trying to answer how many times an event will happen, such as wine sales.

***Model#1 Regression Model Imputed with Mean Values for missing Data: Stepwise selection:***
This model began with all of the variables in our dataset inclusive of the new variables described in the previous section. We incorporated an automated variable selection procedure called Stepwise selection, which adds or removes predictor variables in steps based on the impact to the statistical significance of the model.

***Figure.8 Parameter estimates for OLS Regression: Model #1***

| Variable | Parameter estimates | Standard Error | Type II SS | F Value | Pr > F |
|---|---|---|---|---|---|
| Intercept | 4.11 | 0.44 | 148.81 | 86.9 | <.0001 |
| VolatileAcidity | -0.10 | 0.01 | 72.67 | 42.44 | <.0001 |
| Density | -0.74 | 0.44 | 4.95 | 2.89 | 0.0892 |
| LabelAppeal | 0.47 | 0.01 | 2003.67 | 1170.04 | <.0001 |
| AcidIndex | -0.20 | 0.01 | 834.25 | 487.16 | <.0001 |
| Imp_Alcohol_ind | 0.15 | 0.02 | 68.97 | 40.27 | <.0001 |
| Imp_Chlorides | -0.12 | 0.04 | 16.29 | 9.51 | 0.002 |
| Imp_FreeSulfurDioxide | 0.00 | 0.00 | 23.85 | 13.93 | 0.0002 |
| M_STARS | -2.24 | 0.03 | 11878.00 | 6936.04 | <.0001 |
| Imp_STARS | 0.78 | 0.02 | 4200.51 | 2452.89 | <.0001 |
| Imp_Sulphates | -0.03 | 0.01 | 9.67 | 5.65 | 0.0175 |
| Imp_TotalSulfurDioxide | 0.00 | 0.00 | 34.63 | 20.22 | <.0001 |

***Figure 9: Model #1 Fit Diagnostics.***

| Model Fit Statistics | |
|---|---|
| Criterion | |
| AIC | 6894.94 |
| BIC | 6896.96 |

The parameter estimates produced from the stepwise regression are displayed in Figure.8. For the most part these input variables coincide with the effect we believe they should have on the desirability of wine. We can observe that (LabelAppeal), (Imp_STARS), and (Imp_Alcohol_Ind) all appear to have a positive impact on the target. The variables for flagging a missing star rating (M_STARS), also seems to have a powerful effect on the Target. Since we will be utilizing AIC and BIC as metrics to assist us in determining which model to choose at the end of our analysis, we included those metrics for this specific model in Figure.9.

One caveat to this opening model is that we must remember that we are dealing with count data (Number of cases of wine sold). A major problem in utilizing a regression model for count data is that the distributions have the potential to be skewed.

***Model#2 Poisson Regression :***

This model contains all of the variables and data adjustments in our dataset, but we will now switch to a Poisson Regression Model. As a reminder, Poisson regression is useful when dealing with "count" data. We will also treat (LabelAppeal) and (Imp_STARS) at the class level, meaning we will be able to model the effect the varying levels for each variable.

***Figure.10 Parameter estimates for Poisson Regression: Model #2***

| Parameter | | DF | Parameter Estimates | Standard Error | Wald 95% | Confidence limits | Wald chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|---|
| Intercept | | 1 | -4.36 | 5.67 | -15.47 | 6.74 | 0.59 | 0.4415 |
| FixedAcidity | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.9625 |
| VolatileAcidity | | 1 | -0.03 | 0.01 | -0.04 | -0.02 | 21.88 | <.0001 |
| Density | | 1 | -18.54 | 14.87 | -47.69 | 10.60 | 1.55 | 0.2124 |
| LabelAppeal | -2 | 1 | -0.70 | 0.04 | -0.78 | -0.62 | 270.88 | <.0001 |
| LabelAppeal | -1 | 1 | -0.46 | 0.03 | -0.51 | -0.41 | 339.97 | <.0001 |
| LabelAppeal | 0 | 1 | -0.27 | 0.02 | -0.32 | -0.23 | 139.93 | <.0001 |
| LabelAppeal | 1 | 1 | -0.14 | 0.02 | -0.18 | -0.09 | 35.13 | <.0001 |
| LabelAppeal | 2 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | . | . |
| AcidIndex | | 1 | -0.08 | 0.00 | -0.09 | -0.07 | 296.00 | <.0001 |
| Imp_Alcohol | | 1 | 0.00 | 0.00 | -0.01 | 0.00 | 0.71 | 0.3999 |
| Imp_Alcohol_ind | | 1 | 0.06 | 0.01 | 0.03 | 0.09 | 15.95 | <.0001 |
| Imp_Chlorides | | 1 | -0.04 | 0.02 | -0.07 | -0.01 | 5.31 | 0.0213 |
| Imp_FreeSulfurDioxide | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 7.65 | 0.0057 |
| Imp_ResidualSugar | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.4938 |
| M_STARS | | 1 | -1.09 | 0.02 | -1.12 | -1.05 | 3545.18 | <.0001 |
| Imp_STARS | 1 | 1 | -0.56 | 0.02 | -0.60 | -0.51 | 655.47 | <.0001 |
| Imp_STARS | 2 | 1 | -0.24 | 0.02 | -0.28 | -0.20 | 140.90 | <.0001 |
| Imp_STARS | 3 | 1 | -0.12 | 0.02 | -0.16 | -0.08 | 34.61 | <.0001 |
| Imp_STARS | 4 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | . | . |
| Imp_Sulphates | | 1 | -0.01 | 0.01 | -0.02 | 0.00 | 4.54 | 0.0331 |
| Imp_TotalSulfurDioxide | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 12.60 | 0.0004 |
| log_Density | | 1 | 36.47 | 29.63 | -21.60 | 94.53 | 1.52 | 0.2183 |
| Scale | | 0 | 1.00 | 0.00 | 1.00 | 1.00 | | |

Immediately we see that Model#2 is more involved than a standard regression model. From this model, we can determine the effect (LabelAppeal) and (Imp_STARS) have at each specific level for a unit increase or decrease at a specified level. It also remains clear that the omission of a star rating has a significant impact on our model as well. We do have some parameters that aren't showing a great deal of statistical significance, and this is something we may need to consider in future models.

***Figure 11: Model #2: Fit Diagnostics.***

| Model Fit Statistics Criterion | |
|---|---|
| AIC | 45604.85 |
| BIC | 45761.44 |

The fit diagnostics are displayed in Figure.11. As we continue to build our models, the criteria will partially center around selecting a model with a lower AIC and BIC value.

***Model#3 Negative Binomial Regression :***

This model contains all of the variables and data adjustments in our dataset but incorporates a Negative Binomial Regression Model. If you look closely at the two models, you will see that they are identical. This isn't very surprising, given that a Poisson distribution is a special case of a Negative Binomial distribution.

***Figure.12 Parameter estimates for Negative Binomial Regression: Model #3***

| Parameter | | DF | Parameter Estimates | Standard Error | Wald 95% | Confidence limits | Wald chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|---|
| Intercept | | 1 | -4.36 | 5.67 | -15.47 | 6.74 | 0.59 | 0.4415 |
| FixedAcidity | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.9625 |
| VolatileAcidity | | 1 | -0.03 | 0.01 | -0.04 | -0.02 | 21.88 | <.0001 |
| Density | | 1 | -18.54 | 14.87 | -47.69 | 10.60 | 1.55 | 0.2124 |
| LabelAppeal | -2 | 1 | -0.70 | 0.04 | -0.78 | -0.62 | 270.88 | <.0001 |
| LabelAppeal | -1 | 1 | -0.46 | 0.03 | -0.51 | -0.41 | 339.97 | <.0001 |
| LabelAppeal | 0 | 1 | -0.27 | 0.02 | -0.32 | -0.23 | 139.93 | <.0001 |
| LabelAppeal | 1 | 1 | -0.14 | 0.02 | -0.18 | -0.09 | 35.13 | <.0001 |
| LabelAppeal | 2 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | . | . |
| AcidIndex | | 1 | -0.08 | 0.00 | -0.09 | -0.07 | 296.00 | <.0001 |
| Imp_Alcohol | | 1 | 0.00 | 0.00 | -0.01 | 0.00 | 0.71 | 0.3999 |
| Imp_Alcohol_ind | | 1 | 0.06 | 0.01 | 0.03 | 0.09 | 15.95 | <.0001 |
| Imp_Chlorides | | 1 | -0.04 | 0.02 | -0.07 | -0.01 | 5.31 | 0.0213 |
| Imp_FreeSulfurDioxide | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 7.65 | 0.0057 |
| Imp_ResidualSugar | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.47 | 0.4938 |
| M_STARS | | 1 | -1.09 | 0.02 | -1.12 | -1.05 | 3545.18 | <.0001 |
| Imp_STARS | 1 | 1 | -0.56 | 0.02 | -0.60 | -0.51 | 655.47 | <.0001 |
| Imp_STARS | 2 | 1 | -0.24 | 0.02 | -0.28 | -0.20 | 140.90 | <.0001 |
| Imp_STARS | 3 | 1 | -0.12 | 0.02 | -0.16 | -0.08 | 34.61 | <.0001 |
| Imp_STARS | 4 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | . | . |
| Imp_Sulphates | | 1 | -0.01 | 0.01 | -0.02 | 0.00 | 4.54 | 0.0331 |
| Imp_TotalSulfurDioxide | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 12.60 | 0.0004 |
| log_Density | | 1 | 36.47 | 29.63 | -21.60 | 94.53 | 1.52 | 0.2183 |
| Scale | | 0 | 1.00 | 0.00 | 1.00 | 1.00 | | |

The fit diagnostics displayed in Figure.13 for the Negative Binomial Distribution are similar to the Poisson model albeit slightly worse.

***Figure 13: Model #3: Fit Diagnostics.***

| Model Fit Statistics<br>Criterion | |
|---|---|
| AIC | 45606.85 |
| BIC | 45770.90 |

Given that the Poisson distribution results produced identical variable inputs to the Negative Binomial Model results, we will adjust the model input variables in an attempt to build a different and hopefully better fitting model. We will employ the predictor variables determined by our Stepwise regression analysis, given the stated purpose of Stepwise selection to return coefficients that increase a model's statistical significance. Since we know Poisson distribution to be a special case of the negative binomial distribution, and we know that the mean and variance of our Target are not quite equal, ( Target variable Mean = 3.03 Variance = 3.71), we will run a Negative Binomial Regression model with the Stepwise selected variable inputs.

***Model#4 Negative Binomial Regression with Adjusted Variable Inputs :***

This model contains the variables selected by the automated stepwise selection process in our regression model. Overall, this model appears to have produced more intuitive parameter inputs. (Label appeal) is a bit strange in that it creates negative coefficients, but they improve or get less negative as the label rating improves from (negative 2) to (positive 2). That pattern is closely replicated for stars.

***Figure.14 Parameter estimates for Negative Binomial Regression Version 2: Model #4***

| Parameter | | DF | Parameter Estimates | Standard Error | Wald 95% | Confidence limits | Wald chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|---|
| Intercept | | 1 | 2.6094 | 0.1954 | 2.2263 | 2.9925 | 178.25 | <.0001 |
| VolatileAcidity | | 1 | -0.0306 | 0.0065 | -0.0434 | -0.0178 | 22.03 | <.0001 |
| Density | | 1 | -0.2415 | 0.1918 | -0.6174 | 0.1344 | 1.59 | 0.208 |
| LabelAppeal | -2 | 1 | -0.6984 | 0.0424 | -0.7816 | -0.6153 | 270.91 | <.0001 |
| LabelAppeal | -1 | 1 | -0.4607 | 0.025 | -0.5096 | -0.4117 | 340.35 | <.0001 |
| LabelAppeal | 0 | 1 | -0.2708 | 0.0229 | -0.3156 | -0.226 | 140.44 | <.0001 |
| LabelAppeal | 1 | 1 | -0.1377 | 0.0232 | -0.1831 | -0.0923 | 35.34 | <.0001 |
| LabelAppeal | 2 | 0 | 0 | 0 | 0 | 0 | . | . |
| AcidIndex | | 1 | -0.0784 | 0.0045 | -0.0872 | -0.0696 | 302.84 | <.0001 |
| Imp_Alcohol_ind | | 1 | 0.0487 | 0.0102 | 0.0286 | 0.0688 | 22.64 | <.0001 |
| Imp_Chlorides | | 1 | -0.0374 | 0.0165 | -0.0697 | -0.0052 | 5.17 | 0.0229 |
| Imp_FreeSulfurDioxide | | 1 | 0.0001 | 0 | 0 | 0.0002 | 7.73 | 0.0054 |
| Imp_STARS | 1 | 1 | -0.5552 | 0.0217 | -0.5977 | -0.5127 | 655.68 | <.0001 |
| Imp_STARS | 2 | 1 | -0.2363 | 0.0199 | -0.2753 | -0.1973 | 140.81 | <.0001 |
| Imp_STARS | 3 | 1 | -0.1187 | 0.0202 | -0.1583 | -0.0791 | 34.52 | <.0001 |
| Imp_STARS | 4 | 0 | 0 | 0 | 0 | 0 | . | . |
| M_STARS | | 1 | -1.0856 | 0.0182 | -1.1214 | -1.0499 | 3545.6 | <.0001 |
| Imp_Sulphates | | 1 | -0.0123 | 0.0057 | -0.0236 | -0.001 | 4.59 | 0.0322 |
| Imp_TotalSulfurDioxide | | 1 | 0.0001 | 0 | 0 | 0.0001 | 12.77 | 0.0004 |
| Dispersion | | 0 | 0 | 0 | 0 | 0 | | |

The fit diagnostics for the second version of our Negative Binomial Distribution show a slight improvement over the previous results. Both AIC and BIC are lower, which is the desired result.

***Figure 15: Model #4: Fit Diagnostics.***

| Model Fit Statistics<br>Criterion | |
|---|---|
| AIC | 45601.20 |
| BIC | 45735.72 |

***Model#5: Zero Inflated Poisson distribution Regression.***

Zero-inflated Poisson regression is another modeling method utilized in count data. The Zero-Inflated Poisson Distribution model, (ZIP) has two parts, a Poisson count model and the Logit model for predicting the zeros.

***Figure.16 Parameter estimates for ZIP Model #5***

| Analysis of maximum Likelihood Parameters | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parameter** | | **DF** | **Parameter Estimates** | **Standard Error** | **Wald 95%** | **Confidence limits** | **Wald chi-Square** | **Pr > ChiSq** |
| **Intercept** | | 1 | 2.21 | 0.20 | 1.81 | 2.60 | 119.30 | <.0001 |
| **VolatileAcidity** | | 1 | -0.02 | 0.01 | -0.03 | -0.01 | 7.76 | 0.0053 |
| **Density** | | 1 | -0.25 | 0.20 | -0.64 | 0.14 | 1.60 | 0.2058 |
| **LabelAppeal** | **-2** | 1 | -1.04 | 0.05 | -1.13 | -0.95 | 532.64 | <.0001 |
| **LabelAppeal** | **-1** | 1 | -0.63 | 0.03 | -0.68 | -0.58 | 585.49 | <.0001 |
| **LabelAppeal** | **0** | 1 | -0.35 | 0.02 | -0.39 | -0.30 | 221.22 | <.0001 |
| **LabelAppeal** | **1** | 1 | -0.16 | 0.02 | -0.20 | -0.11 | 44.81 | <.0001 |
| **LabelAppeal** | **2** | 0 | 0.00 | 0.00 | 0.00 | 0.00 | . | . |
| **AcidIndex** | | 1 | -0.02 | 0.01 | -0.03 | -0.01 | 21.00 | <.0001 |
| **Imp_Alcohol_ind** | | 1 | 0.07 | 0.01 | 0.05 | 0.10 | 49.59 | <.0001 |
| **Imp_Chlorides** | | 1 | -0.02 | 0.02 | -0.06 | 0.01 | 2.00 | 0.1577 |
| **Imp_FreeSulfurDioxide** | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 1.49 | 0.222 |
| **Imp_STARS** | **1** | 1 | -0.40 | 0.02 | -0.44 | -0.35 | 300.79 | <.0001 |
| **Imp_STARS** | **2** | 1 | -0.19 | 0.02 | -0.23 | -0.15 | 87.60 | <.0001 |
| **Imp_STARS** | **3** | 1 | -0.10 | 0.02 | -0.14 | -0.06 | 23.40 | <.0001 |
| **Imp_STARS** | **4** | 0 | 0.00 | 0.00 | 0.00 | 0.00 | . | . |
| **M_STARS** | | 1 | -0.19 | 0.02 | -0.23 | -0.15 | 92.61 | <.0001 |
| **Imp_Sulphates** | | 1 | 0.00 | 0.01 | -0.01 | 0.01 | 0.24 | 0.625 |
| **Imp_TotalSulfurDioxide** | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.7926 |
| **Scale** | | 0 | 1.00 | 0.00 | 1.00 | 1.00 | | |

***Figure.17 Maximum Likelihood Parameter estimates Model #5***

| Analysis of maximum Likelihood Zero Inflation Parameter estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parameter** | | **DF** | **Parameter Estimates** | **Standard Error** | **Wald 95%** | **Confidence limits** | **Wald chi-Square** | **Pr > ChiSq** |
| **Intercept** | | 1 | -6.80 | 0.31 | -7.41 | -6.19 | 471.56 | <.0001 |
| **LabelAppeal** | **-2** | 1 | -2.93 | 0.36 | -3.63 | -2.23 | 67.74 | <.0001 |
| **LabelAppeal** | **-1** | 1 | -1.52 | 0.20 | -1.90 | -1.14 | 60.54 | <.0001 |
| **LabelAppeal** | **0** | 1 | -0.76 | 0.19 | -1.12 | -0.40 | 16.80 | <.0001 |
| **LabelAppeal** | **1** | 1 | -0.20 | 0.19 | -0.57 | 0.17 | 1.15 | 0.2836 |
| **LabelAppeal** | **2** | 0 | 0.00 | 0.00 | 0.00 | 0.00 | . | . |
| **AcidIndex** | | 1 | 0.47 | 0.03 | 0.42 | 0.53 | 318.15 | <.0001 |
| **Imp_Alcohol_ind** | | 1 | 0.22 | 0.07 | 0.07 | 0.36 | 8.47 | 0.0036 |
| **Imp_FreeSulfurDioxide** | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 4.80 | 0.0284 |
| **M_STARS** | | 1 | 4.01 | 0.10 | 3.80 | 4.21 | 1468.21 | <.0001 |

Model#5 has thus far produced the best goodness of fit results, showing an improvement in both the AIC and BIC metrics. Analysis of the ZIP portion of the model indicates that the predictors are statistically significant.

***Figure 18: Model #5 FitDiagnostics.***

| Model Fit Statistics Criterion | |
|---|---|
| AIC | 41637.56 |
| BIC | 41831.43 |

***Model#6: Zero Inflated Negative Binomial distribution Regression.***

Model #6 is similar to the previous model, but we will now use a negative binomial distribution as opposed to the Poisson distribution employed previously in Model#5. This model has the same input variables as our previous model.

***Figure.19 Parameter estimates for ZIP Model #6***

| Analysis of maximum Likelihood Parameters | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Parameter | | DF | Parameter Estimates | Standard Error | Wald 95% | Confidence limits | Wald chi-Square | Pr > ChiSq |
| Intercept | | 1 | 2.20 | 0.20 | 1.80 | 2.59 | 117.45 | <.0001 |
| VolatileAcidity | | 1 | -0.02 | 0.01 | -0.03 | 0.00 | 7.12 | 0.0076 |
| Density | | 1 | -0.25 | 0.20 | -0.64 | 0.14 | 1.62 | 0.203 |
| LabelAppeal | -2 | 1 | -1.05 | 0.05 | -1.14 | -0.96 | 538.11 | <.0001 |
| LabelAppeal | -1 | 1 | -0.63 | 0.03 | -0.68 | -0.58 | 588.42 | <.0001 |
| LabelAppeal | 0 | 1 | -0.35 | 0.02 | -0.40 | -0.30 | 220.39 | <.0001 |
| LabelAppeal | 1 | 1 | -0.16 | 0.02 | -0.20 | -0.11 | 44.65 | <.0001 |
| LabelAppeal | 2 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | . | . |
| AcidIndex | | 1 | -0.02 | 0.00 | -0.03 | -0.01 | 18.55 | <.0001 |
| Imp_Alcohol_ind | | 1 | 0.08 | 0.01 | 0.05 | 0.10 | 50.23 | <.0001 |
| Imp_Chlorides | | 1 | -0.02 | 0.02 | -0.06 | 0.01 | 1.99 | 0.1588 |
| Imp_FreeSulfurDioxide | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 1.28 | 0.2572 |
| Imp_STARS | 1 | 1 | -0.39 | 0.02 | -0.43 | -0.34 | 283.79 | <.0001 |
| Imp_STARS | 2 | 1 | -0.19 | 0.02 | -0.23 | -0.15 | 85.53 | <.0001 |
| Imp_STARS | 3 | 1 | -0.10 | 0.02 | -0.14 | -0.06 | 23.08 | <.0001 |
| Imp_STARS | 4 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | . | . |
| M_STARS | | 1 | -0.19 | 0.02 | -0.22 | -0.15 | 89.02 | <.0001 |
| Imp_Sulphates | | 1 | 0.00 | 0.01 | -0.01 | 0.01 | 0.19 | 0.6648 |
| Imp_TotalSulfurDioxide | | 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.8743 |
| Dispersion | | 0 | 0.00 | 0.00 | 0.00 | 0.00 | | |

***Figure.20 Maximum Likelihood Parameter estimates Model #6***

| Analysis of maximum Likelihood Zero Inflation Parameter estimates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Parameter** | | **DF** | **Parameter Estimates** | **Standard Error** | **Wald 95%** | **Confidence limits** | **Wald chi-Square** | **Pr > ChiSq** |
| **Intercept** | | 1 | -6.484 | 0.295 | -7.061 | -5.907 | 484.320 | <.0001 |
| **LabelAppeal** | **-2** | 1 | -2.665 | 0.323 | -3.299 | -2.032 | 67.970 | <.0001 |
| **LabelAppeal** | **-1** | 1 | -1.366 | 0.185 | -1.728 | -1.004 | 54.640 | <.0001 |
| **LabelAppeal** | **0** | 1 | -0.639 | 0.174 | -0.981 | -0.298 | 13.440 | 0.0002 |
| **LabelAppeal** | **1** | 1 | -0.162 | 0.177 | -0.510 | 0.185 | 0.840 | 0.3598 |
| **LabelAppeal** | **2** | 0 | 0.000 | 0.000 | 0.000 | 0.000 | . | . |
| **AcidIndex** | | 1 | 0.455 | 0.025 | 0.405 | 0.504 | 324.210 | <.0001 |
| **Imp_Alcohol_ind** | | 1 | 0.199 | 0.071 | 0.060 | 0.337 | 7.900 | 0.0049 |
| **Imp_FreeSulfurDioxid** | | 1 | -0.001 | 0.000 | -0.001 | 0.000 | 5.160 | 0.0231 |
| **M_STARS** | | 1 | 3.754 | 0.090 | 3.578 | 3.931 | 1738.120 | <.0001 |

Model#6 managed to produce goodness of fit metrics that were firmly in line with our ZIP Poisson regression model. There is a slight difference in AIC and BIC, but overall the models are closely related in terms of performance.

***Figure 21: Model #6 FitDiagnostics.***

| **Model Fit Statistics** | |
|---|---|
| **Criterion** | |
| **AIC** | **41700.97** |
| **BIC** | **41902.31** |

***Model#7: Logistic Regression.***

We wanted to test the results of building a Logistic regression model for our wine data. Logistic regression can be a powerful way of modeling binomial outcomes, such as selling wine, or not selling wine. There are however some assumptions built into Logistics models pertaining to the predictor variables which aren't entirely satisfied here.

***Figure.22 Parameter estimates for Logistic Model #7***

| Parameter | | DF | Parameter Estimates | Standard Error | Wald 95% | Confidence limits | Wald chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|---|
| Intercept | | 1 | 1.734 | 0.054 | 1.627 | 1.840 | 1019.88 | <.0001 |
| VolatileAcidity | | 1 | -0.013 | 0.008 | -0.028 | 0.002 | 2.89 | 0.089 |
| LabelAppeal | -2 | 1 | -1.465 | 0.058 | -1.579 | -1.352 | 641.69 | <.0001 |
| LabelAppeal | -1 | 1 | -0.810 | 0.029 | -0.866 | -0.754 | 801.95 | <.0001 |
| LabelAppeal | 0 | 1 | -0.435 | 0.025 | -0.485 | -0.385 | 293.2 | <.0001 |
| LabelAppeal | 1 | 1 | -0.193 | 0.026 | -0.243 | -0.143 | 56.92 | <.0001 |
| LabelAppeal | 2 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | . | . |
| AcidIndex | | 1 | -0.020 | 0.005 | -0.030 | -0.009 | 13.17 | 0.0003 |
| Imp_Alcohol_ind | | 1 | 0.103 | 0.012 | 0.079 | 0.126 | 74.35 | <.0001 |
| Imp_FreeSulfurDioxide | | 1 | 0.000 | 0.000 | 0.000 | 0.000 | 0.65 | 0.4218 |
| Imp_STARS | 1 | 1 | -0.367 | 0.025 | -0.415 | -0.318 | 219.2 | <.0001 |
| Imp_STARS | 2 | 1 | -0.229 | 0.022 | -0.272 | -0.185 | 105.19 | <.0001 |
| Imp_STARS | 3 | 1 | -0.122 | 0.023 | -0.166 | -0.077 | 29.14 | <.0001 |
| Imp_STARS | 4 | 0 | 0.000 | 0.000 | 0.000 | 0.000 | . | . |
| M_STARS | | 1 | -0.200 | 0.022 | -0.243 | -0.156 | 81.39 | <.0001 |
| Imp_Sulphates | | 1 | 0.001 | 0.007 | -0.012 | 0.014 | 0.03 | 0.8616 |
| Imp_TotalSulfurDioxi | | 1 | 0.000 | 0.000 | 0.000 | 0.000 | 1.38 | 0.2404 |
| Scale | | 0 | 1.000 | 0.000 | 1.000 | 1.000 | | |

The results of this logistic model appear to return some decent fitting coefficients irrespective of our initial concerns. They all seem to line up with what we would expect the effect each individual variable would have on wine sales. The overall fit diagnostics also returned better than expected results, coming in with the lowest AIC and BIC values.

***Figure 23: Model #7 FitDiagnostics.***

| Model Fit Statistics<br>Criterion | |
|---|---|
| AIC | 31147.91 |
| BIC | 31256.16 |

**Model Selection:**
We were able to run seven different models based on variable adjustments, different compositions, manual selection procedures and automated selection processes. The table below provides some goodness of fit summary metrics we discussed during the modeling process.

***Figure.24 Model Comparison***

| Model | Number of Variables | AIC | BIC |
|---|---|---|---|
| **(1)** OLS Regression Model With Stepwise Selection | 11 | 6894.94 | 6896.96 |
| **(2)** Poisson Regression Model All Variables | 22 | 45604.85 | 45761.44 |
| **(3)** Negative Binomial Regression Model All Variables | 22 | 45604.85 | 45761.44 |
| **(4)** Negative Binomial Regression Version 2 | 18 | 45601.20 | 45735.72 |
| **(5)** ZIP Poisson Distribution Regression model | 18 | **41637.56** | **41831.43** |
| **(6)** ZIP Negative Binomial Distribution Regression model | 18 | 41700.97 | 41902.31 |
| **(7)** Logistic Regression Model | 22 | **31147.91** | **31256.16** |

We initially mentioned that we would base our model selection on the AIC, and BIC goodness of fit metrics. The AIC is an estimate of a constant, plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model. BIC is an estimate of a function of the probability of a model being accurate. What that all means is that in each case we want to find the model with the lowest possible values. The information in Figure.15, presents us with a bit of complex decision. While the regression model and the logistic model each have lower values for AIC and BIC, we know that these models aren't optimal for modeling the type of data we have in this analysis. Given that Logistics models can be advantageous when you are attempting to predict one outcome versus another, we were really tempted to choose this model, but we're making a "better safe than sorry" decision and choosing to proceed with the ZIP Poisson Regression Model highlighted above in Figure.15. This model produced the lowest AIC and BIC values amongst the Poisson and Negative Binomial models in this analysis.

## Conclusion:

After going through and analyzing the data with various modeling techniques, we generated a total of seven different models for our wine data. We preceded this by examining our data for inconsistencies, such as missing observations or extreme outliers, and made imputations that were in line with the size of our data set and our goals of the analysis. This modeling process was challenging, and the end results and decisions left us with some questions with respect to the model we chose. One model had the better goodness of fit, while the other was professionally considered more appropriate. Fortune favors the bold, so looking forward we can incorporate multiple modeling techniques and compare actual results of both to come to the best possible decision for future utilization.

## Appendix: SAS Code

```
/****Load data***/
%let ME =  noeflores20160;
%let PATH = /home/&ME./my_courses/donald.wedding/c_8888/PRED411/UNIT03/HW;
%let NAME = HW;
%let LIB = &NAME..;

libname &NAME. "&PATH.";
%let INFILE = HW.WINE;
%let INFILE2 = HW.WINE_TEST;
%let OUTFILE      = OUTFILE_FLAG;

%let TEMPFILE    = TEMPFILE;
%let FIXFILE      = FIXFILE;
%let VARLIST      = VARLIST;

proc print data=&INFILE. (obs=10);
run;

/***Check contents of data***/
proc contents data=&INFILE order =varnum;
run;

data mydata;
set &INFILE.;
drop INDEX;
run;

proc print data = mydata (obs=10);
run;

/***Check contents of data***/
proc contents data=mydata order =varnum;
run;

/*****proc means on numeric variables******/
proc means data= mydata N NMISS MEAN MIN MAX MEDIAN std;
var _numeric_ ;
run;

/*****proc means without TARGET_FLAG********/
proc means data= mydata MIN MAX Range MEAN MEDIAN Var N NMISS std;
var
Target
AcidIndex
Alcohol
Chlorides
CitricAcid
Density
FixedAcidity
FreeSulfurDioxide
LabelAppeal
ResidualSugar
STARS
Sulphates
TotalSulfurDioxide
VolatileAcidity
pH;
run;

/****Check for outliers******/
```

```
proc univariate normal plot data=mydata;
var
AcidIndex
Alcohol
Chlorides
CitricAcid
Density
FixedAcidity
FreeSulfurDioxide
LabelAppeal
ResidualSugar
STARS
Sulphates
TotalSulfurDioxide
VolatileAcidity
pH;
RUN;

proc univariate normal plot data=mydata;
var _numeric_;
histogram;
run;

/****Run Univariate for specific variables******/
proc univariate normal plot data=mydata;
var target alcohol labelappeal stars;
histogram;
run;

/****Run Univariate for Target******/
proc univariate normal plot data=mydata;
var target;
histogram;
run;

/***Proc Corr to check for initial correlations***/
proc corr data=mydata;
var _numeric_;
with TARGET;
run;

/*********Data Preparation*************************/
proc means data= mydata MIN MAX Range MEAN N NMISS;
var
Alcohol
Chlorides
CitricAcid
FreeSulfurDioxide
ResidualSugar
STARS
Sulphates
TotalSulfurDioxide
pH;
run;

/****Run Univariate for Sulfers and Sugar ******/
proc univariate normal plot data=mydata;
var FreeSulfurDioxide TotalSulfurDioxide;
histogram;
run;
```

```
/**************Data Imputation********************/
data imputed;
set mydata;

TARGET_FLAG = ( TARGET > 0 );
TARGET_AMT = TARGET - 1;
if TARGET_FLAG = 0 then TARGET_AMT = .;

Imp_Alcohol = Alcohol;
M_Alcohol = 0;
if missing(Imp_Alcohol) then do;
Imp_Alcohol = 10.49;
M_Alcohol = 1;
end;

if Imp_Alcohol >= 10.50 then Imp_Alcohol_ind = 2;
if Imp_Alcohol < 10.50 then Imp_Alcohol_ind = 1;

Imp_Chlorides = Chlorides;
M_Chlorides = 0;
if missing(Imp_Chlorides) then do;
Imp_Chlorides = 0.05;
M_Chlorides = 1;
end;

Imp_FreeSulfurDioxide = FreeSulfurDioxide;
M_FreeSulfurDioxide = 0;
if missing (Imp_FreeSulfurDioxide) then do;
Imp_FreeSulfurDioxide = 30.85;
M_FreeSulfurDioxide = 1;
end;
if Imp_FreeSulfurDioxide < -267
then Imp_FreeSulfurDioxide = -267;
if Imp_FreeSulfurDioxide > 328
then Imp_FreeSulfurDioxide = 328;


Imp_ResidualSugar = ResidualSugar;
M_ResidualSugar = 0;
if missing(Imp_ResidualSugar) then do;
Imp_ResidualSugar = 5.42;
M_ResidualSugar = 1;
end;
if Imp_ResidualSugar < -62
then Imp_ResidualSugar = -62;
if Imp_ResidualSugar > 73
then Imp_ResidualSugar = 73;

Imp_STARS = STARS;
M_STARS = 0;
if missing(Imp_STARS) then do;
Imp_STARS = 2.0;
M_STARS = 1;
end;

Imp_Sulphates = Sulphates;
M_Sulphates = 0;
if missing(Imp_Sulphates) then do;
Imp_Sulphates = 0.53;
M_Sulphates = 1;
end;
```

```
Imp_TotalSulfurDioxide = TotalSulfurDioxide;
M_TotalSulfurDioxide = 0;
if missing (Imp_TotalSulfurDioxide) then do;
Imp_TotalSulfurDioxide = 120.71;
M_TotalSulfurDioxide = 1;
end;
if Imp_TotalSulfurDioxide < -343
then Imp_TotalSulfurDioxide = -343;
if Imp_TotalSulfurDioxide > 583
then Imp_TotalSulfurDioxide = 583;

drop residualsugar;
drop chlorides;
drop FreeSulfurDioxide;
drop TotalSulfurDioxide;
drop pH;
drop CitricAcid;
drop sulphates;
drop Alcohol;
drop STARS;

log_Density = log(Density+1);

run;
quit;

proc print data = imputed (obs =10);
run;

/********Check to ensure that missing values are corrected******/
proc means data=imputed Min Max RANGE Mean n nmiss;
var _numeric_;
run;

/*********Double check Correlations*********/
proc corr data=imputed;
var
FixedAcidity
VolatileAcidity
Density
LabelAppeal
AcidIndex
Imp_Alcohol
M_Alcohol
Imp_Alcohol_ind
Imp_Chlorides
M_Chlorides
Imp_FreeSulfurDioxide
M_FreeSulfurDioxide
Imp_ResidualSugar
M_ResidualSugar
Imp_STARS
M_STARS
Imp_Sulphates
M_Sulphates
Imp_TotalSulfurDioxide
M_TotalSulfurDioxide
log_Density;
with TARGET;
run;

proc freq data=imputed;
```

```
   tables target_flag*M_STARS;

proc freq data=imputed;
   tables target*labelappeal;

proc freq data=imputed;
   tables target*Imp_Alcohol_Ind;

proc sgscatter data=imputed;
compare x=LabelAppeal
y= TARGET;
run;

proc sgscatter data=imputed;
compare x=Imp_STARS
y= TARGET;
run;

proc freq data=imputed;
table TARGET_FLAG /missing;
run;

proc univariate data = imputed noprint;
histogram TARGET TARGET_AMT;
run;

proc means data = imputed  nmiss mean median min max;
class TARGET_FLAG;
var _numeric_;
run;

proc freq data= imputed;
table (Imp_STARS LABELAPPEAL)*TARGET_FLAG /missing;
run;

/**************Model Bulding*****************/
data modelfile;
set imputed;
run;

proc print data = modelfile (obs =10);
run;

/************** Model #1 Regression with Stepwise selection**************/
proc reg data=modelfile;
model target =
FixedAcidity
VolatileAcidity
Density
LabelAppeal
AcidIndex
Imp_Alcohol
Imp_Alcohol_ind
Imp_Chlorides
Imp_FreeSulfurDioxide
Imp_ResidualSugar
M_STARS
Imp_STARS
Imp_Sulphates
Imp_TotalSulfurDioxide
log_Density /selection = stepwise aic bic;
output out = wine p=step_Regression;
```

```
run;

/**************Regression Model with R-Square selection with AIC and BIC*************/
proc reg data=modelfile;
model target =
FixedAcidity
VolatileAcidity
Density
LabelAppeal
AcidIndex
Imp_Alcohol
Imp_Alcohol_ind
Imp_Chlorides
Imp_FreeSulfurDioxide
Imp_ResidualSugar
M_STARS
Imp_STARS
Imp_Sulphates
Imp_TotalSulfurDioxide
log_Density /selection = rsquare aic bic;
run;

/************ Model #2 Genmod with Poisson Distribution***************/
proc genmod data=modelfile;
class labelappeal Imp_STARS;
model target =
FixedAcidity
VolatileAcidity
Density
LabelAppeal
AcidIndex
Imp_Alcohol
Imp_Alcohol_ind
Imp_Chlorides
Imp_FreeSulfurDioxide
Imp_ResidualSugar
M_STARS
Imp_STARS
Imp_Sulphates
Imp_TotalSulfurDioxide
log_Density / link=log dist=pois1;
output out=wine p=pois1;

/****** Model #3 Genmod with negative binomial Distribution*******/
proc genmod data=modelfile;
class labelappeal Imp_STARS;
model target =
FixedAcidity
VolatileAcidity
Density
LabelAppeal
AcidIndex
Imp_Alcohol
Imp_Alcohol_ind
Imp_Chlorides
Imp_FreeSulfurDioxide
Imp_ResidualSugar
M_STARS
Imp_STARS
Imp_Sulphates
Imp_TotalSulfurDioxide
log_Density/ link=log dist=nb;
```

```
output out=wine p=negbi1;

/****** Model #4 Genmod with negative binomial Distribution Version 2*******/
proc genmod data=modelfile;
class labelappeal Imp_STARS;
model target =
VolatileAcidity
Density
LabelAppeal
AcidIndex
Imp_Alcohol_ind
Imp_Chlorides
Imp_FreeSulfurDioxide
Imp_STARS
M_STARS
Imp_Sulphates
Imp_TotalSulfurDioxide/ link=log dist=nb;
output out=wine p=negbiV2;

/******* Model #5 Genmod with Zero Inflated Poisson Distribution**********/
proc genmod data = modelfile;
class labelappeal Imp_STARS;
model target =
VolatileAcidity
Density
LabelAppeal
AcidIndex
Imp_Alcohol_ind
Imp_Chlorides
Imp_FreeSulfurDioxide
Imp_STARS
M_STARS
Imp_Sulphates
Imp_TotalSulfurDioxide/ link=log dist=ZIP;
zeromodel
LabelAppeal
AcidIndex
Imp_Alcohol_ind
Imp_FreeSulfurDioxide
M_STARS/ link=logit;
            ;
output out=wine p=zippois1;
run;

/******* Model #6 Genmod with Zero Inflated Negative Binomial Distribution**********/
proc genmod data = modelfile;
class labelappeal Imp_STARS;
model target =
VolatileAcidity
Density
LabelAppeal
AcidIndex
Imp_Alcohol_ind
Imp_Chlorides
Imp_FreeSulfurDioxide
Imp_STARS
M_STARS
Imp_Sulphates
Imp_TotalSulfurDioxide/ link=log dist=ZINB;
zeromodel
LabelAppeal
AcidIndex
```

```
Imp_Alcohol_ind
Imp_FreeSulfurDioxide
M_STARS/ link=logit;
                ;
output out=wine p=zipbin1;
run;

/************* Model #7 PROC LOGISTIC/Poisson******************/
proc logistic data=modelfile;
class Imp_STARS LabelAppeal;
model TARGET_FLAG(ref="0") =
VolatileAcidity
Density
LabelAppeal
AcidIndex
Imp_Alcohol_ind
Imp_Chlorides
Imp_FreeSulfurDioxide
Imp_STARS
M_STARS
Imp_Sulphates
Imp_TotalSulfurDioxide;
output out=logfile p=X_Log;
run;

proc genmod data=logfile;
class Imp_STARS LabelAppeal;
model TARGET_AMT =
VolatileAcidity
LabelAppeal
AcidIndex
Imp_Alcohol_ind
Imp_FreeSulfurDioxide
Imp_STARS
M_STARS
Imp_Sulphates
Imp_TotalSulfurDioxide/ link=log dist=poi;
                                                ;
output out=wine p=logistic;
run;

/******************** Stand Alone scoring code**********************/
Data testing;
  set &INFILE2.;

data testing_fixed;
    set testing;

Imp_Alcohol = Alcohol;
M_Alcohol = 0;
if missing(Imp_Alcohol) then do;
Imp_Alcohol = 10.49;
M_Alcohol = 1;
end;

if Imp_Alcohol >= 10.50 then Imp_Alcohol_ind = 2;
if Imp_Alcohol < 10.50 then Imp_Alcohol_ind = 1;

Imp_Chlorides = Chlorides;
M_Chlorides = 0;
if missing(Imp_Chlorides) then do;
Imp_Chlorides = 0.05;
```

```
M_Chlorides = 1;
end;

Imp_FreeSulfurDioxide = FreeSulfurDioxide;
M_FreeSulfurDioxide = 0;
if missing (Imp_FreeSulfurDioxide) then do;
Imp_FreeSulfurDioxide = 30.85;
M_FreeSulfurDioxide = 1;
end;
if Imp_FreeSulfurDioxide < -267
then Imp_FreeSulfurDioxide = -267;
if Imp_FreeSulfurDioxide > 328
then Imp_FreeSulfurDioxide = 328;

Imp_ResidualSugar = ResidualSugar;
M_ResidualSugar = 0;
if missing(Imp_ResidualSugar) then do;
Imp_ResidualSugar = 5.42;
M_ResidualSugar = 1;
end;
if Imp_ResidualSugar < -62
then Imp_ResidualSugar = -62;
if Imp_ResidualSugar > 73
then Imp_ResidualSugar = 73;

Imp_STARS = STARS;
M_STARS = 0;
if missing(Imp_STARS) then do;
Imp_STARS = 2.0;
M_STARS = 1;
end;

Imp_Sulphates = Sulphates;
M_Sulphates = 0;
if missing(Imp_Sulphates) then do;
Imp_Sulphates = 0.53;
M_Sulphates = 1;
end;

Imp_TotalSulfurDioxide = TotalSulfurDioxide;
M_TotalSulfurDioxide = 0;
if missing (Imp_TotalSulfurDioxide) then do;
Imp_TotalSulfurDioxide = 120.71;
M_TotalSulfurDioxide = 1;
end;
if Imp_TotalSulfurDioxide < -343
then Imp_TotalSulfurDioxide = -343;
if Imp_TotalSulfurDioxide > 583
then Imp_TotalSulfurDioxide = 583;

drop residualsugar;
drop chlorides;
drop FreeSulfurDioxide;
drop TotalSulfurDioxide;
drop pH;
drop CitricAcid;
drop sulphates;
drop Alcohol;
drop STARS;

log_Density = log(Density+1);
```

```
data testing_score;
   set testing_fixed;

   TEMP = -6.8
   + (LabelAppeal in (-2)) * -2.93
   + (LabelAppeal in (-1)) * -1.52
   + (LabelAppeal in (0))  * -0.76
   + (LabelAppeal in (1))  * -0.20
   + (LabelAppeal in (2))  *  0.00
   +  AcidIndex            *  0.47
   + Imp_alcohol_ind       *  0.22
   + Imp_FreeSulfurDioxide *  0.00
   + (M_STARS in (0))      *  4.01;
   P_SCORE_ZERO = exp(TEMP) / (1 + exp(TEMP));

   temp = 2.21
   + VolatileAcidity * -0.02
   + Density         * -0.25
   + (LabelAppeal in (-2))  * -1.04
   + (LabelAppeal in (-1))  * -0.63
   + (LabelAppeal in (0))   * -0.35
   + (LabelAppeal in (1))   * -0.16
   + (LabelAppeal in (2))   *  0.00
   + AcidIndex              * -0.02
   + Imp_Alcohol_ind        *  0.07
   + Imp_Chlorides          * -0.02
   + Imp_FreeSulfurDioxide  *  0.00
   + (Imp_STARS in (1))     * -0.40
   + (Imp_STARS in (2))     * -0.19
   + (Imp_STARS in (3))     * -0.10
   + (Imp_STARS in (4))     *  0.00
   + (M_STARS in (0))       * -0.19
   + Imp_Sulphates          *  0.00
   + Imp_TotalSulfurDioxide *  0.00;

   P_SCORE_ZIP_ALL = exp(TEMP);

   P_TARGET_INIT = P_SCORE_ZIP_ALL * (1 - P_SCORE_ZERO);
   P_TARGET = round(P_TARGET_INIT,1);

   keep index P_TARGET P_TARGET_ROUND;

/*******Please print******/
proc print data = testing_score;
run;

/*******Save to Folder********/
libname NOELIB "/home/noeflores20160/sasuser.v94";

data NOELIB.NOEFILE;
set testing_score;
run;
```