Unit 03 Wine Sales Project

Predict 411 Section 59

Winter 2016

Joshua Peng

Joshua Peng
Unit 03 Wine Sales Project

Hyperlinks

Bonus: For the bonus section, I am attempting to gain +30 points. I built 2 models (+10 for both models), interpreted the parameter coefficients, and compared them to my best model (model 6) in section 4. Select Models based on Goodness of Fit criteria, sum of absolute error, and sum of squared error over the first 30 observations.
1. I built a model using PROC GENMOD with a Zero Inflated Poisson distribution (zeromodel using complementary log-log link) with STARS0 and LabelAppeal as categorical variables.
2. I built a model using PROC GENMOD with a Zero Inflated Poisson distribution (zeromodel using probit link) with STARS0 and LabelAppeal as categorical variables.
3. As presented in the recorded session, we can gain +10 bonus points by confirming that using the output file from PROC GENMOD with PROC PLM will generate the same predicted values as with a SAS data step for a Zero Inflated Poisson Model (ZIP model) which are composed of 2 separate model processes.

Introduction

The wine data set contains 12,795 observations each of which represent information on 12,795 commercially available wines. There are 12 continuous variables related to the chemical properties of the wine being sold. There are 2 numerical variables for the marketing score based on the visual appeal of the label and wine rating based on number of stars. 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 stores around the United States. The more sample cases purchased, the more likely a wine is to be sold at a high end restaurant.

A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If it is possible to predict the number of cases, the wine manufacturer will be able to adjust their wine offerings with the goal to maximize sales. The purpose of this project is to build a model to predict the number of cases of wine that will be sold given certain properties of the wine. I will specifically work towards building Poisson and Negative Binomial models that will predict the target number of cases ordered for each wine.

1. Data Exploration

The Data Dictionary provided details the 14 variables related to the characteristics of the wine. The variables can be segmented into 2 groups: 12 continuous variables related to the chemical properties of the wine and 2 numerical variables related to the subjective perception of the wine. The variable, LabelAppeal, is the Marketing Score which indicates the visual appeal of the label design where high numbers suggest customers like the design and low numbers suggest customers do not like the design. The variable, STARS, is the wine rating by a team of experts between 1 to 4 stars (4 Stars = Excellent, 1 Star = Poor). STARS is a clear indication of the wine's popularity which should have a strong positive relationship with the number of cases.

Variable	**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. High numbers suggest customers like the label design. Negative numbers suggest customers don't like the design.	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 by a team of experts. 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	

Next, I use PROC MEANS to examine the descriptive statistics of the variables and open the wine data set to look at the first few rows of observations. The LabelAppeal marketing score is on a -2 to +2 scale and STARS wine rating is on a 1 to 4 scale. Since these variables take on a few possible values within a very small range, I may be able to use these two variables as categorical variables.

Addressing missing observations

I also notice that many variables have missing observations in the data set, most notably STARS with the most at 3359. For the continuous variables, I will impute the missing values with the mean value. For STARS, which I will also use as a categorical variable, I rounded the mean value to the nearest integer and created a binary indicator variable to flag when the data is missing. I will also try LabelAppeal as a categorical variable and see if that improves model performance.

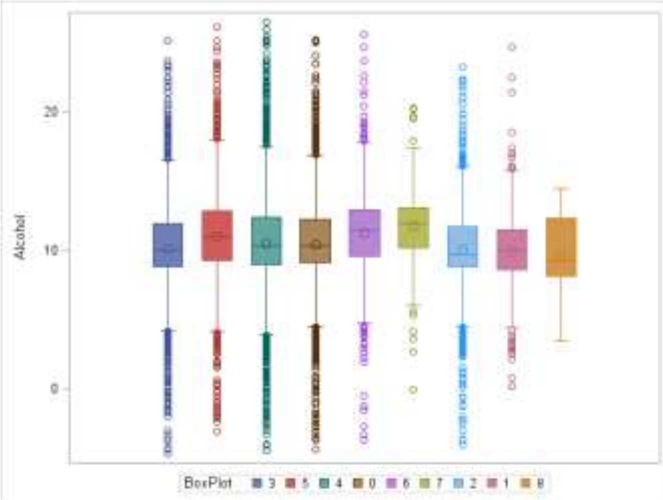
Variable	N	N Miss	Mean	Variance	Range	Minimum	Maximum
AcidIndex	12795	0	7.7727237	1.7527810	13.0000000	4.0000000	17.0000000
Alcohol	12142	653	10.4892363	13.8966348	31.2000000	-4.7000000	26.5000000
Chlorides	12157	638	0.0548225	0.1014214	2.5220000	-1.1710000	1.3510000
CitricAcid	12795	0	0.3084127	0.7431816	7.1000000	-3.2400000	3.8600000
Density	12795	0	0.9942027	0.000704247	0.2111500	0.8880900	1.0992400
FixedAcidity	12795	0	7.0757171	39.9126188	52.5000000	-18.1000000	34.4000000
FreeSulfurDioxide	12148	647	30.8455713	22116.02	1178.00	-555.0000000	623.0000000
LabelAppeal	12795	0	-0.0090660	0.7940400	4.0000000	-2.0000000	2.0000000
ResidualSugar	12179	616	5.4187331	1139.02	268.9500000	-127.8000000	141.1500000
STARS	9436	3359	2.0417550	0.8145785	3.0000000	1.0000000	4.0000000
Sulphates	11585	1210	0.5271118	0.8688650	7.3700000	-3.1300000	4.2400000
TotalSulfurDioxide	12113	682	120.7142326	53783.74	1880.00	-823.0000000	1057.00
VolatileAcidity	12795	0	0.3241039	0.6146783	6.4700000	-2.7900000	3.6800000
pH	12400	395	3.2076282	0.4619745	5.6500000	0.4800000	6.1300000
TARGET	12795	0	3.0290739	3.7108945	8.0000000	0	8.0000000

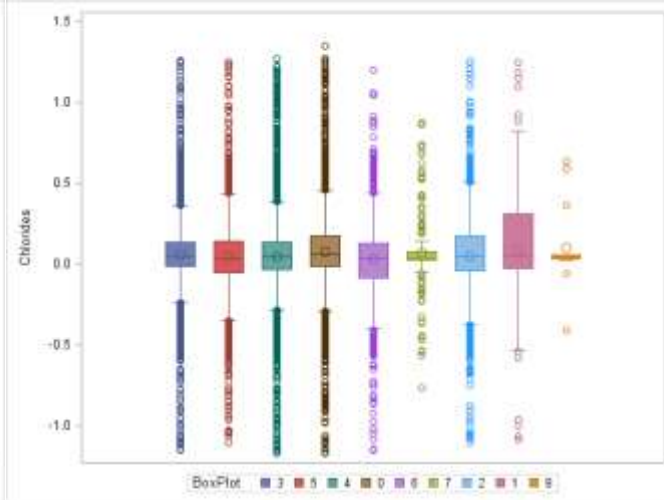
In the later stages of the project when I am selecting variables to include in the model, if I decide to include a variable that originally had missing values, I will consider bringing along with it the corresponding indicator variable because the missing observations make up a large portion of the data.

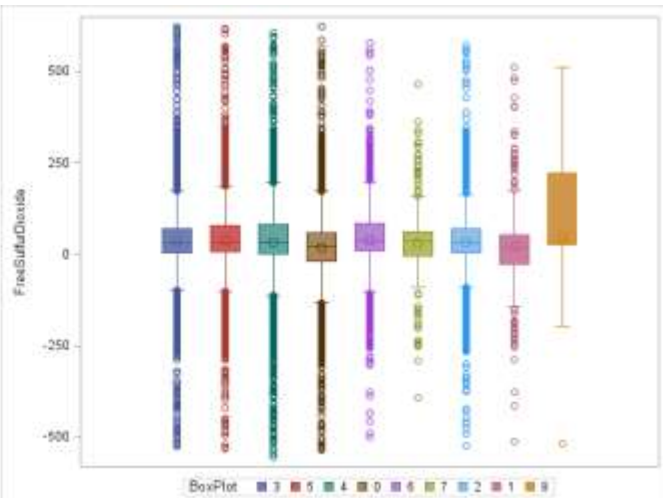
I want to contain information from the missing observations along with the original values of STARS together. I will create a new variable, STARS0, by replicating STARS and setting the missing values to 0 so that the range will be between 0 to 4. As a categorical variable, the “0” class will not signify the lowest wine rating, because the classes are not treated as ordered. However, as a continuous variable, the “0” class will indicate a very bad wine and perhaps this variable may turn out to be a better predictor than the original STARS.

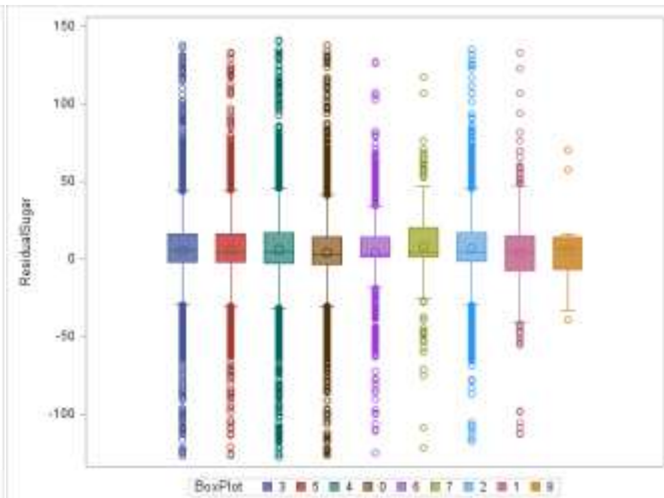
Quick look at box plots

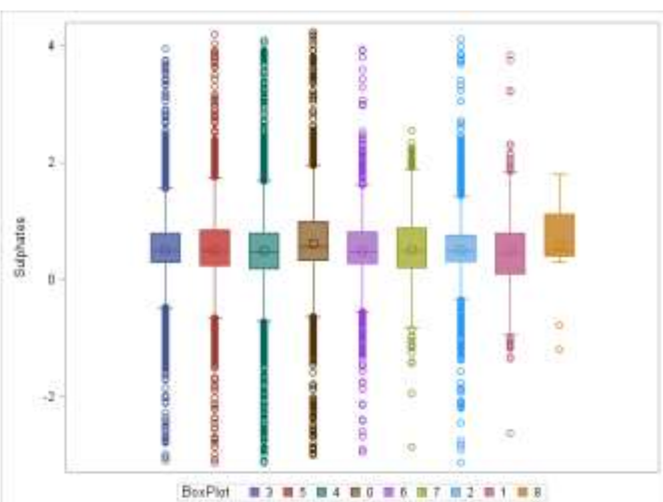
A quick glance at some box plots for Alcohol, Chlorides, FreeSulfurDioxide, ResidualSugar, Sulphates, TotalSulfurDioxide, and pH confirm and highlight the glaring problem that these variables have impossible negative values.

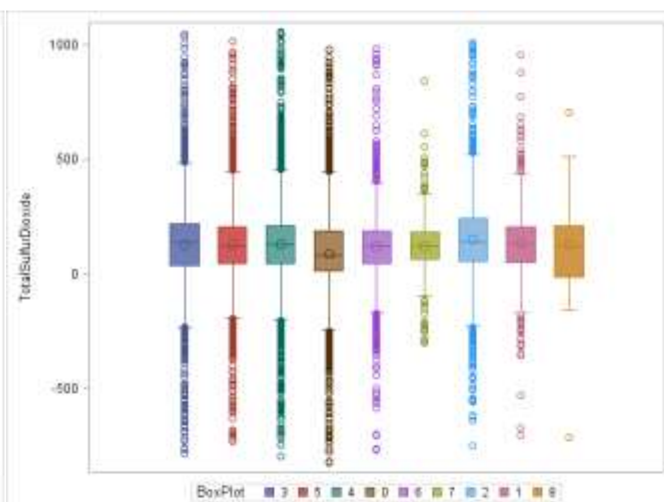

Alcohol
20
10
0
BoxPlot 3 5 4 0 6 7 2 1 8

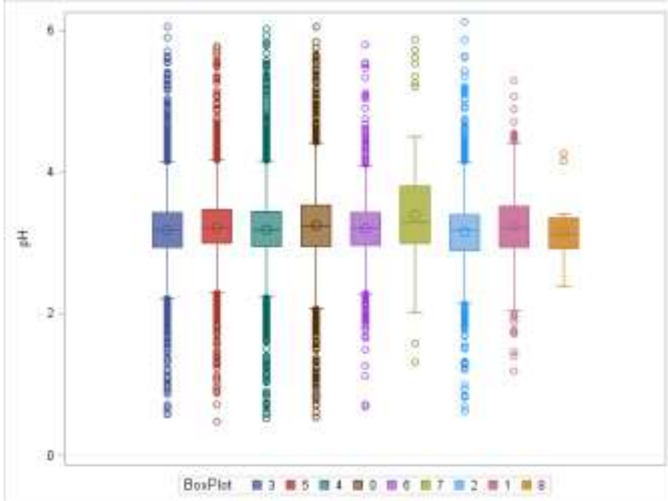

Chlorides
1.5
1.0
0.5
0.0
-0.5
-1.0
BoxPlot 3 5 4 0 6 7 2 1 8


FreeSulfurDioxide
500
250
0
-250
-500
BoxPlot 3 5 4 0 6 7 2 1 8


ResidualSugar
150
100
50
0
-50
-100
BoxPlot 3 5 4 0 6 7 2 1 8


Sulphates
4
2
0
-2
BoxPlot 3 5 4 0 6 7 2 1 8


TotalSulfurDioxide
1000
500
0
-500
BoxPlot 3 5 4 0 6 7 2 1 8



Addressing negative values and adding new variables
Many of the variables have negative values which do not make sense because they are count variables and measure a frequency, amount, or concentration of a particular substance which can only take on positive values, including: Alcohol, Chlorides, CitricAcid, FixedAcidity, FreeSulfurDioxide, ResidualSugar, Sulphates, TotalSulfurDioxide, and VolatileAcidity. The variable distributions may have been shifted downward by adding a negative constant. I will add the absolute value of the minimum negative value to all of the observations of variables with negative values to make sure the whole distribution is positive with a minimum value of 0. These new "reshifted" variables will have the prefix "rs_". There may have been a recording error such that several observations were mistakenly recorded as negative. For this case, I will take the absolute value of these variables with negative values to make sure all of the observations are positive. These new absolute value transformed variables will have the prefix "av_".

As I will discuss later when I examine the relationship of each variable with wine quality, I can derive new variables from existing variables that relate to wine quality and taste. Fixed acidity is measured as total acidity minus volatile acidity. Therefore, I can derive TotalAcidity by adding together FixedAcidity and VolatileAcidity.

Only a proportion of the sulfur dioxide added to a wine will be effective as an anti-oxidant. The rest will combine with other elements in the wine and cease to be useful. The part lost into the wine is said to be bound, the active part to be free. Therefore, I derive a new variable, BoundSulfurDioxide, from TotalSulfurDioxide and FreeSulfurDioxide by taking the difference between them.

It appears that the missing observations of TotalSulfurDioxide and FreeSulfurDioxide do not overlap as there are almost double the amount of missing observations for BoundSulfurDioxide. Additionally, the missing observations of FixedAcidity and VolatileAcidity do no overlap as well, as there are almost double the amount of missing observations for TotalAcidity. To address this discrepancy, I will derive another form of BoundSulfurDioxide called BoundSulfurDioxide2 that is computed after the missing values from TotalSulfurDioxide and FreeSulfurDioxide have been imputed with their mean values.

Variable	N	N Miss	Mean	Variance	Range	Minimum	Maximum
BoundSulfurDioxide	11512	1283	89.6924079	74761.76	2449.00	-1093.00	1356.00
FreeSulfurDioxide	12148	647	30.8455713	22116.02	1178.00	-555.0000000	623.0000000
TotalSulfurDioxide	12113	682	120.7142326	53783.74	1880.00	-823.0000000	1057.00
TotalAcidity	12795	0	7.3998210	40.6498870	53.7350000	-18.6100000	35.1250000
FixedAcidity	12795	0	7.0757171	39.9126188	52.5000000	-18.1000000	34.4000000
VolatileAcidity	12795	0	0.3241039	0.6146783	6.4700000	-2.7900000	3.6800000

New BoundSulfurDioxide Variables added:
- BoundSulfurDioxide = TotalSulfurDioxide - FreeSulfurDioxide
  - with its missing values replaced by its mean value
- BoundSulfurDioxide2 = TotalSulfurDioxide - FreeSulfurDioxide
  - after the missing values from TotalSulfurDioxide and FreeSulfurDioxide have had their missing values imputed with their mean values
- rs_BoundSulfurDioxide = abs(rs_TotalSulfurDioxide – rs_FreeSulfurDioxide)
- rs_BoundSulfurDioxide2 = BoundSulfurDioxide + abs(min(BoundSulfurDioxide))
- av_BoundSulfurDioxide = abs(abs(TotalSulfurDioxide) – abs(FreeSulfurDioxide))
- av_BoundSulfurDioxide2 = abs(BoundSulfurDioxide2)

Fortunately, there are no missing values in FixedAcidity or VolatileAcidity, however there are still negative values. To address this case, I derive other forms of TotalAcidity that are computed before FixedAcidity and VolatileAcidity have been adjusted for negative values through reshifting or absolute value transformations.

New TotalAcidity variables added:

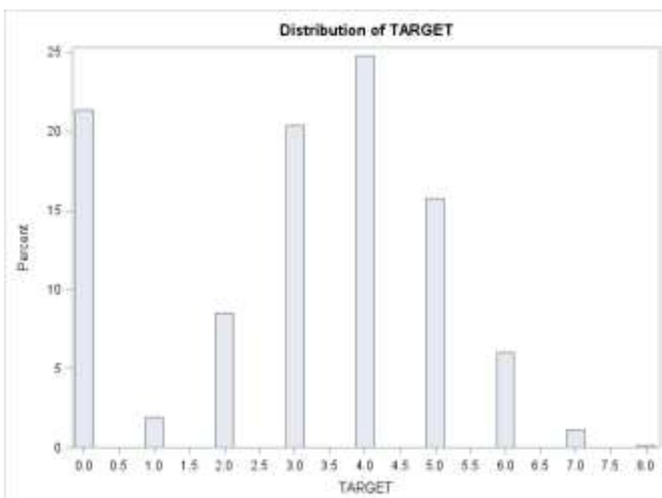
- TotalAcidity = FixedAcidity + VolatileAcidity
- rs_TotalAcidity = abs(rs_FixedAcidity + rs_VolatileAcidity)
- rs_TotalAcidity2 = TotalAcidity + abs(min(TotalAcidity))
- av_TotalAcidity = abs(abs(FixedAcidity) + abs(VolatileAcidity))
- av_TotalAcidity2 = abs(TotalAcidity)

Examining TARGET

Next, I examine the mean and variance of the TARGET variable which measures the number of cases purchased. The assumption that the mean (3.0290739) and variance (3.7108945) are equal for the Poisson distribution is violated, although the values are rather close in value. However, the assumption that the variation should be larger than the mean for the Negative Binomial distribution is satisfied which means that TARGET exemplifies overdispersion.

TARGET	
Mean	**Variance**
3.0290739	3.7108945

I examine the histogram of TARGET and find that the shape is zero inflated but otherwise resembles a normal distribution taking values between 0 to 8 with a central peak at 4. Although I would normally restrict my modeling approach based on the zero inflation present in TARGET, for the purposes of this assignment, I will build OLS regression, Poisson, and Negative Binomial models and examine the differences in performance.



Examining histograms and exploring the relationship of variables with wine quality

In the following section, I will not post the histogram of the reshifted transformed variables as the shape of the distribution will be exactly the same. The values of the reshifted variables are just shifted upward by a constant. The histogram of the original variable with imputed mean values will be on the left and the histogram of the absolute value transformed variable will be on the right.

Sulphates = Sulfites and av_Sulfites

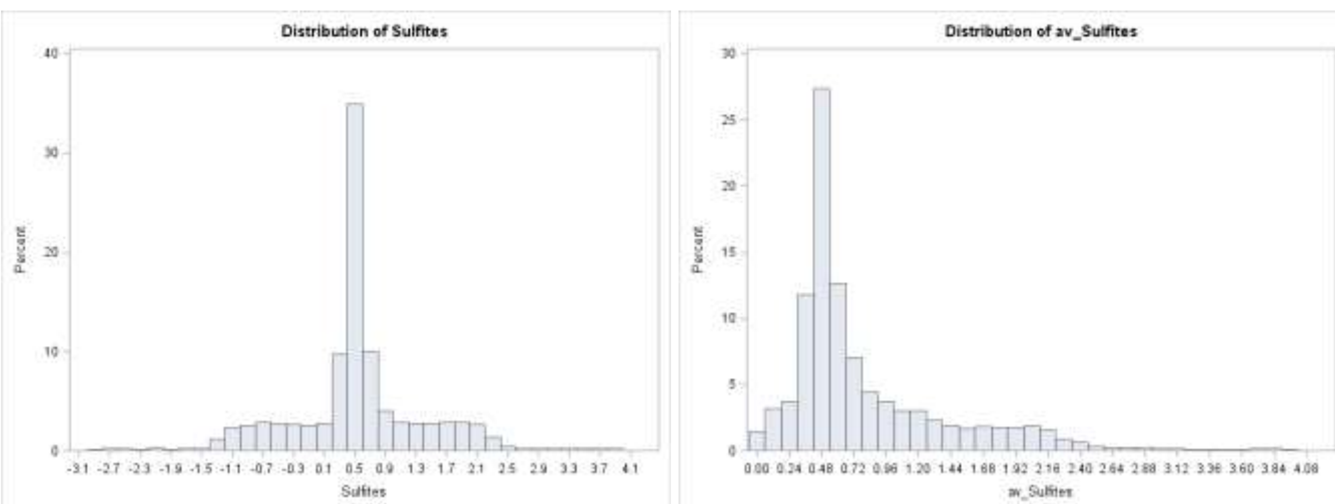
I believe this variable is incorrectly named and should actually be "Sulfites" instead of "Sulphates." Sulfur dioxide (SO2) is added at several points in the process of conventional vinification and is present in the finished wine in the form of sulphites (or sulfites if you are American). Sulfites are sulfur compounds that have a relationship with wine, but sulfate, a salt of sulfuric acid, is not. Sulfates are simple inorganic chemical compounds of sulfur which are not meant for consumption and play no role in the wine making process. Although there exist some approved sulfate additives for wine, there is abundantly more information about sulfites in wine. I will assume that the wine manufacturer that organized the data set intended to use "Sulfites" and will rename the variable accordingly. From this point on, I will refer to the variable "Sulphates" as "Sulfites."

All wines contain sulfur dioxide in various forms, collectively known as sulfites. Sulfur dioxide (SO2 for short) is by far the most important additive used in wine. Even in completely unsulfured wine it is present at concentrations of up to 10 milligrams per liter. Commercially-made wines contain from 10 to 20 times that amount. Its value derives from its ability to perform several critical functions such as preserving a wine's freshness and fruit character by virtue of anti-oxidative, anti-microbial and anti-enzymatic properties. Oxidation is the reaction of wine with oxygen. It can alter its color and odor, tending to make wines darker and dryer, and is often dismissed as a fault. Excessive oxidation does ruin wine. But controlled oxidization can add complexity, and is crucial to certain styles. Sulfur dioxide drastically inhibits the process of oxidation. playing an important part of the aging process. The judicious use of SO2 is required to make high-quality, shelf-stable wine.

There are four points at which SO2 is commonly used in conventional winemaking. It is applied in the form of metabisulfite to inhibit the action of wild yeasts and prevent oxidation during grape picking. So that the grapes can be preserved and not be rushed to the winery. It is added during grape crushing to prevent fermentation from beginning with wild yeasts before cultured yeasts can be added. Cultured yeasts are bred to be more resistant to SO2. It is added at any point during fermentation, but most commonly at the end to stop or prevent malolactic fermentation. A natural winemaker has to wait for this process to finish naturally. Lastly, it is added to prevent oxidation (or any other microbial action) in the bottled wine. In sweet wines there is the danger that fermentation will restart. A natural winemaker would only ever use SO2 at bottling, only in white wines, and only in very small quantities. Many natural winemakers use none at all.[1]

There are three main reasons you might not want sulfites added to your wine. Sulfites can cause potentially fatal allergic reactions and has been linked with numerous other health problems, including hangovers. Sulfite is an artificial ingredient and upon adding it to a wine, the winemaker can no longer claim that the wine is "natural." Sulfites have an unpleasant smell, like that of a struck match, and is detectable by your tongue at very low concentrations. Most people can detect sulfur dioxide in water at around 11 mg/l. In wine, the presence of alcohol and acids means that it is less obvious. For an experienced taster, accustomed to natural wine, SO2 becomes unpleasant at concentrations of around 20-30 mg/l, depending on the style of wine and the ratio of free to bound SO2. For most people the threshold is much higher, but most people have never tasted an unsulfured wine. They may well be able to taste the SO2, but are not accustomed to the taste.

The shape of the Sulfites histogram resembles a plateau with low kurtosis except there is a large central spike at the mean value of 0.5271118 which is accentuated with imputed mean values. The av_Sulfites histogram is positively skewed with a long right tail. The Kolmogorov-Smirnov test for normality for both histograms results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.



TotalSulfurDioxide and av_TotalSulfurDioxide
SO2 is a gas at room temperature. But when SO2 is free in wine, it can take 3 different forms:

$$H_2O + SO_2 \leftrightarrow H^+ + HSO_{3-} \leftrightarrow 2H^+ + {SO_3}^{2-}$$

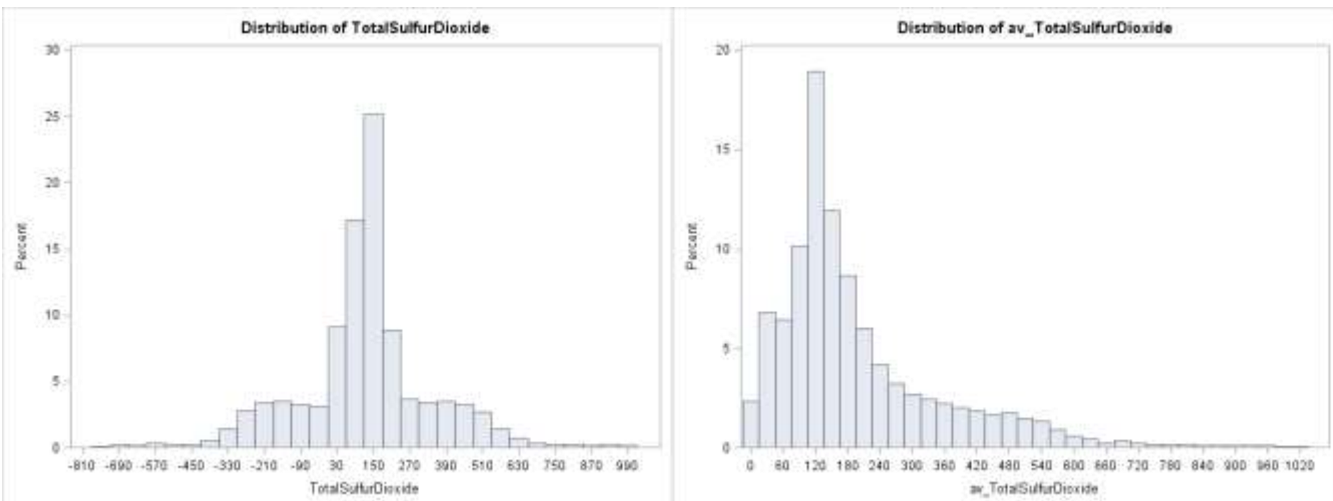
- molecular SO2 when in solution with water (H2O)
- bisulfite when it is a HSO3– ion
- sulfite when it is a SO32- ion

Total SO2 = free SO2 + bound SO2

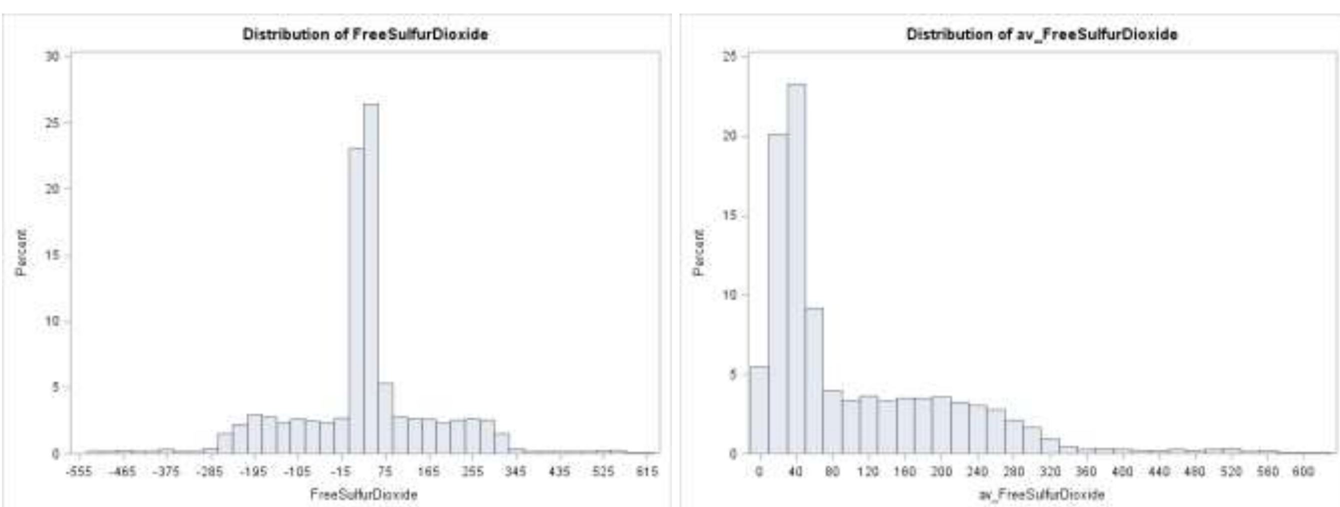
- free SO2: molecular SO2 + bisulfites + sulfites
- bound SO2: sulfites attached to either sugars, acetaldehyde or phenolic compounds

The free SO2 portion (not associated with wine molecules) is effectively the buffer against microbes and oxidation. Whereas the bound SO2 portion, which is associated with wine molecules, is the part which has already done its work and cannot be useful any longer in this context. Total SO2 should be kept below 110 ppm for table wines because, at higher levels, the wine can acquire off-flavors. For dessert and fortified wines that are very sweet, it may be necessary to exceed this limit to obtain adequate free SO2. The higher the level of total SO2 in the wine, the higher the ratio will be, because there are fewer unbound compounds available for reacting with additional sulfur dioxide as it is added. Sulfur dioxide is also more effective if it is added less often and in greater quantities because it will be more of a shock to the microbes.[2]

The shape of the TotalSulfurDioxide histogram resembles a plateau with low kurtosis except there is a large central spike at the mean value of 120.7142326 which is accentuated with imputed mean values. The av_TotalSulfurDioxide histogram is positively skewed with a large peak at the mean. The Kolmogorov-Smirnov test for normality for both histograms results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.



FreeSulfurDioxide and av_FreeSulfurDioxide
The shape of the FreeSulfurDioxide histogram resembles a plateau with low kurtosis except there is a large central spike at the mean value of 30.8455713 which is accentuated with imputed mean values. The av_FreeSulfurDioxide histogram is highly positively skewed with practically no left half/tail of the curve. The Kolmogorov-Smirnov test for normality for both histograms results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.

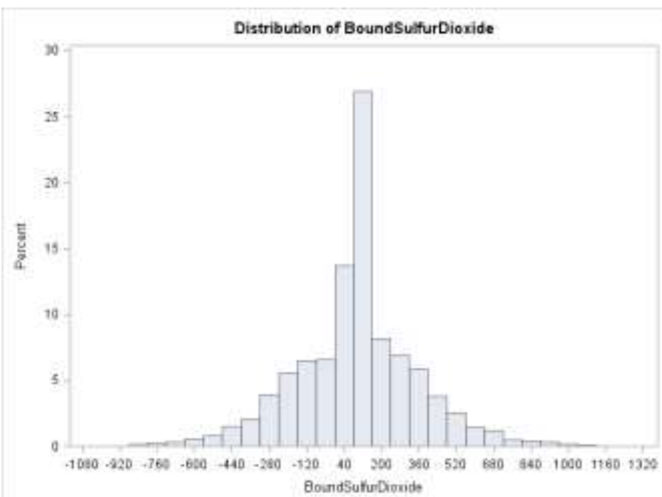


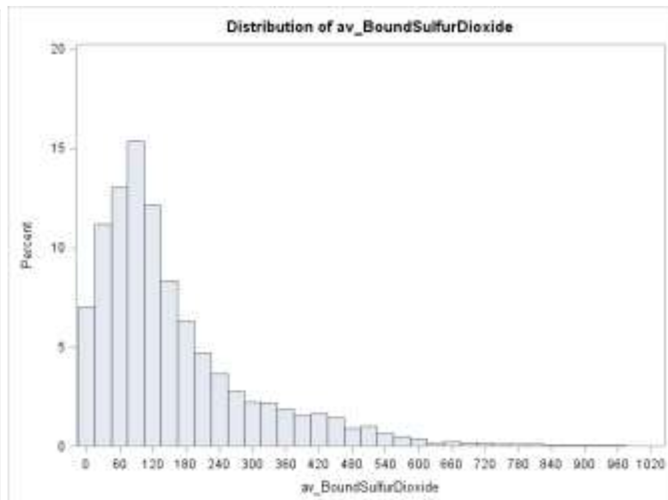
The amount of SO2 added to wine depends on the type of wine, the sensitivity of the taster, and the ratio between free and bound sulfur dioxide in the wine. Only a proportion of the SO2 added to a wine will be effective as an anti-oxidant. The rest will combine with other elements in the wine and cease to be useful. The part lost into the wine is said to be bound, the active part to be free. A good winemaker will try to get the highest proportion of free SO2 to bound that he can. At best, this will be about half the amount bound.

BoundSulfurDioxide and av_BoundSulfurDioxide

For white wines, a level of 0.8 ppm molecular SO2 will slow down the growth of yeast and will prevent the growth of most other microbes. This level of sulfur dioxide will bind up most of the acetaldehyde in a wine and reduce any oxidation aroma considerably. Therefore, 0.8 ppm is a good target level for molecular SO2 immediately prior to bottling and will provide the maximum protection for the finished wine. However, sensitive tasters will be able to detect a slight burnt match aroma at 0.8 ppm SO2. This is usually not a problem however because few consumers will be able to detect it. Additionally, if the wine is bottle-aged for a few months before consumption, the SO2 will decrease as more sulfites react with other chemical constituents in the wine and become bound. Thus, a wine bottled at 0.8 ppm will decrease to a lower level fairly quickly and there would be no detectable sulfur dioxide aroma.

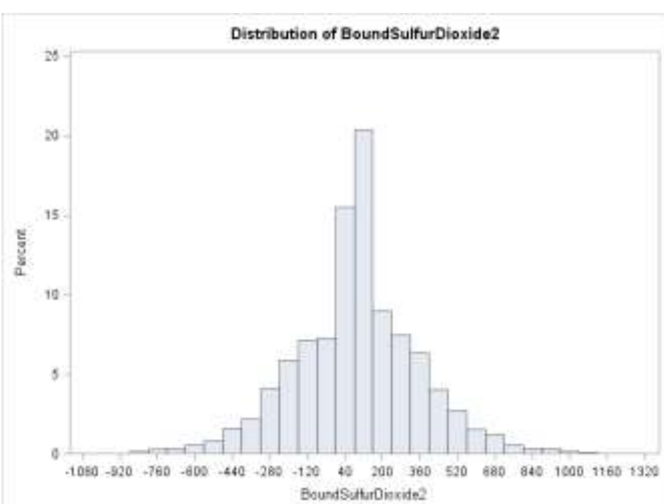
The shape of the BoundSulfurDioxide histogram resembles a normal distribution with low kurtosis except there is a large central spike at the mean value of 89.6924079 which is accentuated with imputed mean values. The av_BoundSulfurDioxide histogram is highly kurtotic and positively skewed with a long right tail. The Kolmogorov-Smirnov test for normality for both histograms results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.

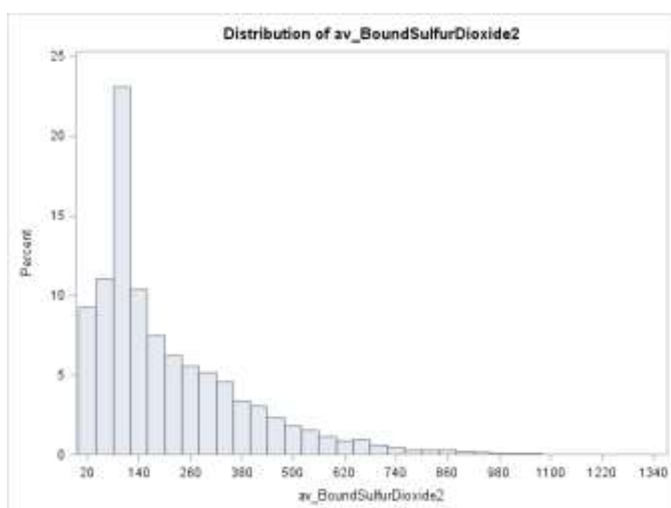




BoundSulfurDioxide2 and av_BoundSulfurDioxide2

The shape of the BoundSulfurDioxide2 histogram is similar to that of BoundSulfurDioxide resembling a normal distribution with low kurtosis except there is a large central spike at the mean value of 89.86866 which is accentuated with imputed mean values. The av_BoundSulfurDioxide2 histogram is positively skewed with a large spike at the mean with a long right tail. The Kolmogorov-Smirnov test for normality for both histograms results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.





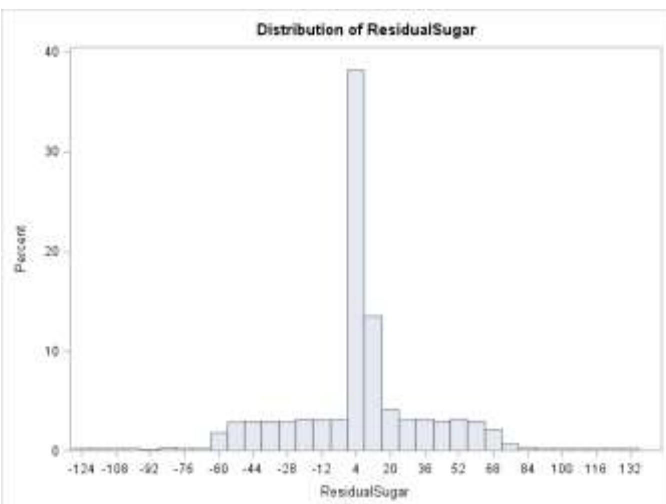
ResidualSugar and av_ResidualSugar

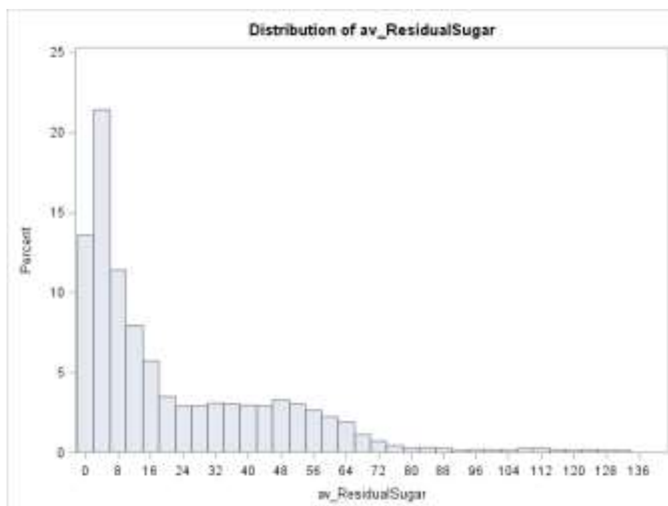
Residual sugar refers to any natural grape sugars that are left over after fermentation ceases whether on purpose or not. The juice of wine grapes starts out intensely sweet, and fermentation uses up that sugar as the yeasts feast upon it generating the by-products of CO2 gas and alcohol. Fermentation may stop due to alcohol toxicity. Different yeast strains can tolerate different levels of alcohol, so a weaker strain might die before eating all the sugar in the fermenting wine. In the case of a dessert wine, the sugars are concentrated when the grapes get shriveled, so there's a lot of sugar to ferment. When alcohol reaches the level of a normal dry wine, say 12 or 14%, the yeast might die, but plenty of uneaten sugar is left. In the case of a fortified wine, hard liquor is added to get a similar job done. Fermentation is also temperature-sensitive, happening faster at warm temperatures and slower in the cold, so it will stop if the temperature drops too much. A winemaker can chill a wine down until fermentation stops, then just get rid of the yeast.[3]

In addition to its obvious sweetening power, sugar also has a bonus effect: it can help wines age well. Wines with a little residual sugar can be the most exciting to taste as they evolve over time. The sugar compounds change shape and will be less directly perceivable, so the wines will dry out a bit. Residual sugars have a balancing relationship with acidity. They are on opposite sides of the balance, so if the wine has sugar you will probably want strong acidity, otherwise the wine will feel cloying. On the other hand, certain very high-acid wines can be far tastier with a few extra grams of residual sugar.

Sugar also has a balancing relationship with sulfur dioxide depending on the type of wine. Red wines do not need any added SO2 because they naturally contain anti-oxidants, acquired from their skins and stems during fermentation, but SO2 may be added anyway. White wines and rosés do not contain natural anti-oxidants because they are not left in contact with their skins after crushing. For this reason, they are more prone to oxidation and tend to be given larger doses of sulfur dioxide. Sweet wines get the largest doses of SO2 because sugar combines with and binds a high proportion of any SO2 added. To get the same level of free sulfur dioxide, the total concentration has to be higher than for dry wines. Dry wines are wines with no residual sugar which means they are not sweet.

The shape of the ResidualSugar histogram resembles a plateau with low kurtosis except there is a large central spike at the mean value of 5.4187331 which is accentuated with imputed mean values. The av_ResidualSugar histogram is highly positively skewed with practically no left half/tail of the curve. The Kolmogorov-Smirnov test for normality for both histograms results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.

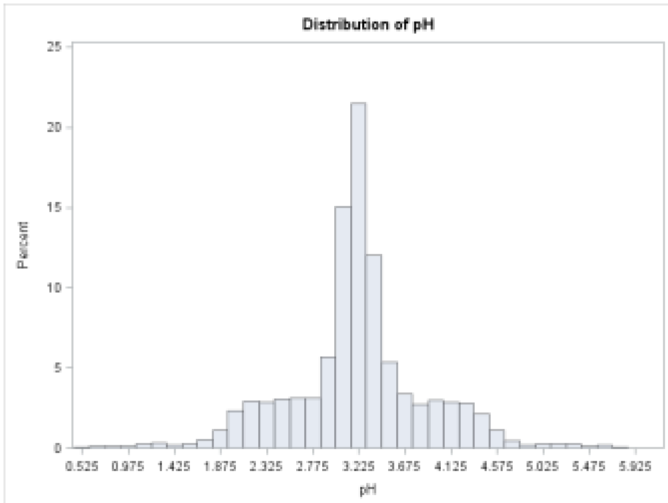




pH
pH is the measure of the degree of relative acidity versus the relative alkalinity of any liquid, on a scale of 0 to 14, with 7 being neutral. Winemakers use pH as a way to measure ripeness in relation to acidity. Low pH wines will taste tart and crisp, while higher pH wines are more susceptible to bacterial growth. Most wine pH's fall around 3 or 4; about 3.0 to 3.4 is desirable for white wines, while about 3.3 to 3.6 is best for red wines. Total acidity is another way of looking at similar things, this time measuring acidity by volume. The higher the pH, the lower the acidity, and the lower the pH, the higher the acidity. Most table wines will have a total acidity of about 0.6 to 0.7 percent.

While Total Acidity and pH may appear to be directly correlated as acidity indicators, they are not. The measurement of pH is the number of H+ ions in a solution using a logarithmic scale, with a lower number denoting a higher concentration of H+ ions. The measurement of acidic content is the acid’s potential to liberate H+ ions as it dissociates. While acid content affects pH, it is not directly predictive of pH or vice versa. This non-direct correlation is partially due to pH buffering caused by a number of compounds in wines, such as sugars, acids, and phenolic compounds. The addition of a given amount of acid to a wine may not reduce the pH as expected due to the wine's buffering capacity to maintain a stable pH.[4]

The shape of the pH histogram resembles a plateau with low kurtosis except there is a large central spike at the mean value of 3.2076282 which is even more accentuated with imputed mean values. Most wines have a pH between 3 to 4, but about 25% of the observations are well below this range. Lemon juice has a pH of 2 and full strength acid has a pH of 0, so many of these wines are fatal to drink. Many of these pH values observations appear to be impossible values or a result of recording error. Since the sample size is greater than 2000, the Kolmogorov-Smirnov test is used to assess normality. The p-value < 0.01, so the normality assumption does not hold, meaning the data does not follow a normal distribution.

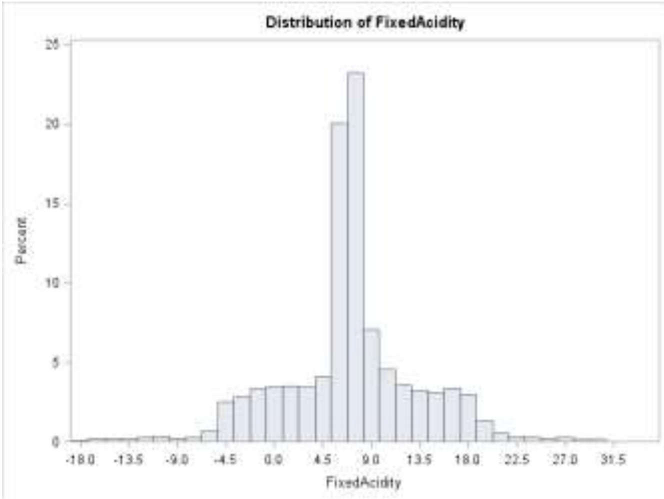


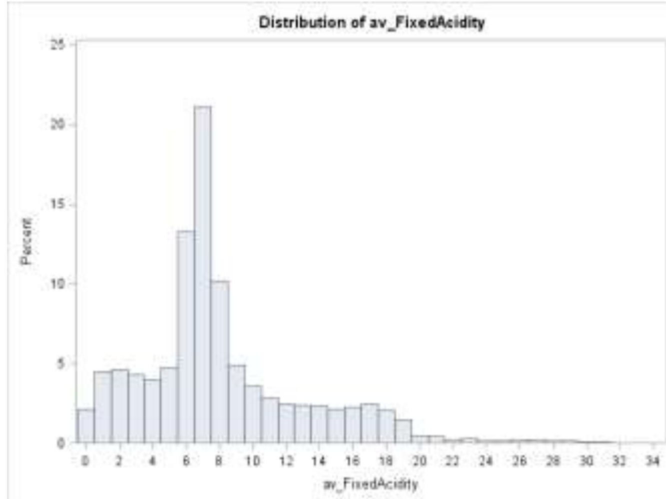
FixedAcidity and av_FixedAcidity
Acids are major wine constituents and contribute greatly to its taste. In fact, acids impart the sourness or tartness that is a fundamental feature in wine taste. Wines lacking in acid are flat tasting. Acidity affects taste, color, stability to oxidation, and consequently, the overall lifespan of a wine. The most abundant of these acids arise in the grapes themselves and carry over into the wine. However, there are also some acids that arise as a result of the fermentation process from either yeast and/or bacteria. Traditionally total acidity is divided into two groups, namely the volatile acids and the fixed (or nonvolatile) acids.

The predominant fixed acids found in wines are tartaric, malic, citric, and succinic acids. All of these fixed acids originate in grapes with the exception of succinic acid, which is produced by yeast during the fermentation process. Wines produced from cool climate grapes are high in acidity and thus taste sour. These high-acid wines can be treated to reduce the acidity, either by neutralizing agents, or by malolactic fermentation. Warm climate grapes can be low in acid, more or less depending on variety. In these areas tartaric acid, recycled from winemaking, is added to increase acidity and prevent wines from being flat.[5]

Tartaric and malic acids are produced by wine grapes as they develop. In warm climates, these acids are lost through the biochemical process of respiration. Therefore, grapes grown in warmer climates have lower acidity than grapes grown in cooler climates. Sugar production is the complete opposite of acid production. The warmer the climate the higher the sugar content of the grapes. In summary, warmer climates result in high sugar and low acid whereas cooler climates result in low sugar and high acid.

The shape of the FixedAcidity histogram resembles a plateau with low kurtosis except there is a large central spike at the mean value of 7.0757171. The av_FixedAcidity histogram very similarly shaped like a plateau with a large central spike. The Kolmogorov-Smirnov test for normality for both histograms results in a significant p-value (p < 0.01) which suggests that the data does not follow a normal distribution.





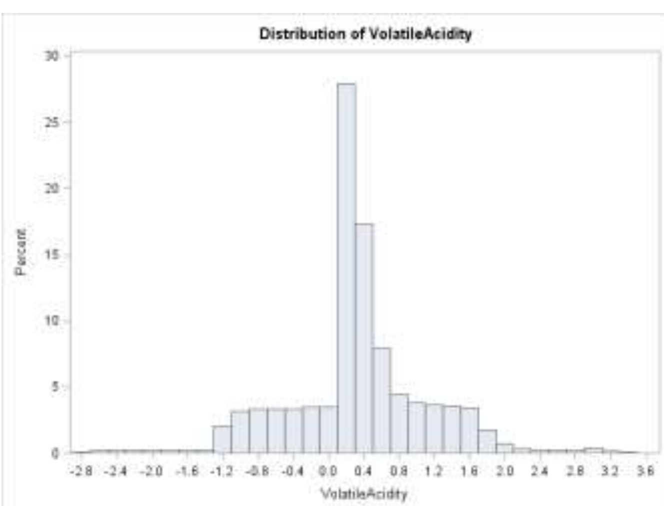
VolatileAcidity and av_VolatileAcidity
Volatile acidity refers to the steam distillable acids present in wine, primarily acetic acid but also lactic, formic, butyric, and propionic acids. Volatile acidity is closely associated with quality because it is an indication of spoilage. High levels of volatile acids are not desirable in wines. There are prevention and treatment methods to remove volatile acidity from a wine. The average level of acetic acid in a new dry table wine is less than 400 mg/L, though levels may range from undetectable up to 3 g/L. Acetic acid can be boiled off when heated. The amount of volatile acid is small with respect to total acidity. A volatile acidity measurement of 0.03-0.06% is produced during fermentation and is considered a normal level.

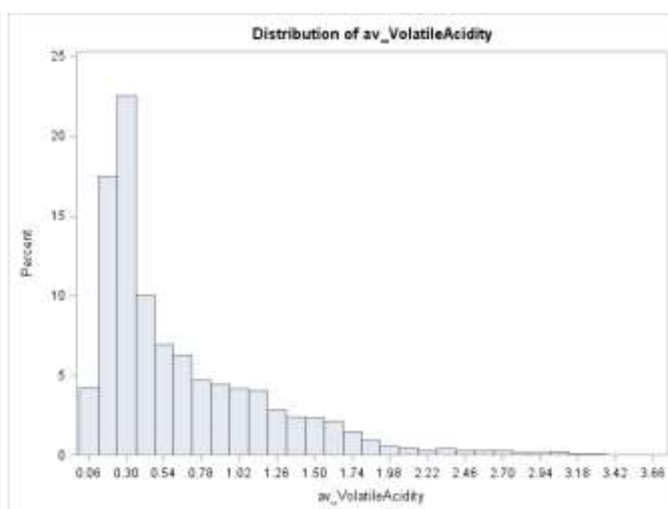
U.S. legal limits of Volatile Acidity:

- Red Table Wine 1.2 g/L
- White Table Wine 1.1 g/L

The aroma threshold for acetic acid in red wine varies from 600 mg/L and 900 mg/L, depending on the variety and style. While acetic acid is generally considered a spoilage product (vinegar), some winemakers seek a low or barely detectible level of acetic acid to add to the perceived complexity of a wine. In addition, the production of acetic acid will result in the concomitant formation of other, sometimes unpleasant, aroma compounds such as ethyl acetate and acetaldehyde. These compounds have a much lower sensory threshold than acetic acid. Both acetaldehyde and ethyl acetate are detectable at less than 200 mg/L in wine.[6]

The shape of the VolatileAcidity histogram resembles a plateau with low kurtosis except there is a large central spike at the mean value of 0.3241039. The Kolmogorov-Smirnov test for normality results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution. The av_VolatileAcidity histogram is highly positively skewed with practically no left half/tail of the curve. If the units of measurement are g/L, then both histograms have values above the U.S. legal limits of volatile acidity. This information should be factored into the decision of which wines to select. The Kolmogorov-Smirnov test for normality for both histograms results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.

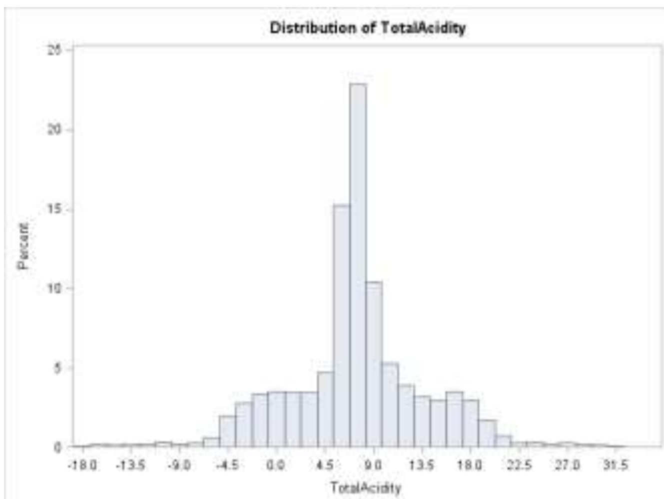


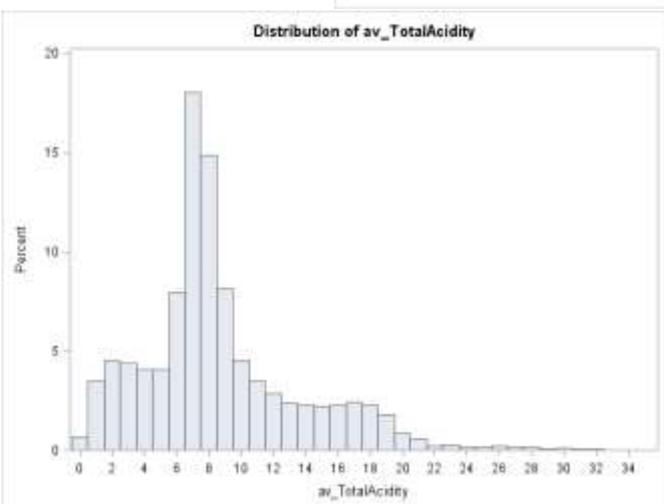


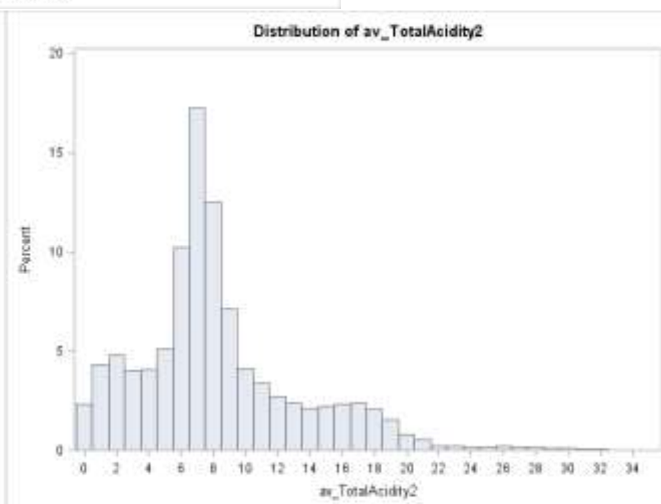
TotalAcidity, av_TotalAcidity, and av_TotalAcidity2
Total acidity takes into account all of the acids in wine. Interactions between the acids and the other chemical components are extremely complicated, yet each of these plays a role in the measurement of total acidity. The typical acidity measurements in wine are pH and total acidity. The pH measurement is used in the vineyard to assess the ripening pre-harvest to calculate sulfur dioxide requirements after fermentation, and to assess oxidation risk because high pH wines are generally more prone to oxidation. Total acidity is applied to sensory perception of a wine's acidity (i.e. tartness, sourness, and crispness). While pH and total acidity are related, pH is a measurement of the likelihood and speed of occurrence of pH dependent reactions, while total acidity is the best estimate of a wines perceived acidity.

Technically, total acidity is not the same as titratable acidity. It is actually very difficult to accurately measure total acidity because you need to be able to directly quantify organic acids so most winemakers measure titratable acidity. While total acidity only quantifies the molar weights of acids contained in a grape, must or wine; titratable acidity is an approximation of total acidity by titration with a strong base to a pH of 8.2. For this assignment, I am approximating total acidity by adding together fixed acidity and volatile acidity.[7]

The shape of the TotalAcidity histogram resembles a plateau with low kurtosis except there is a large central spike at the mean value of 0.3241039. The Kolmogorov-Smirnov test for normality results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution. The av_TotalAcidity and av_TotalAcidity2 histograms are slightly positively skewed with longer right tails. The Kolmogorov-Smirnov test for normality for all 3 histograms results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.



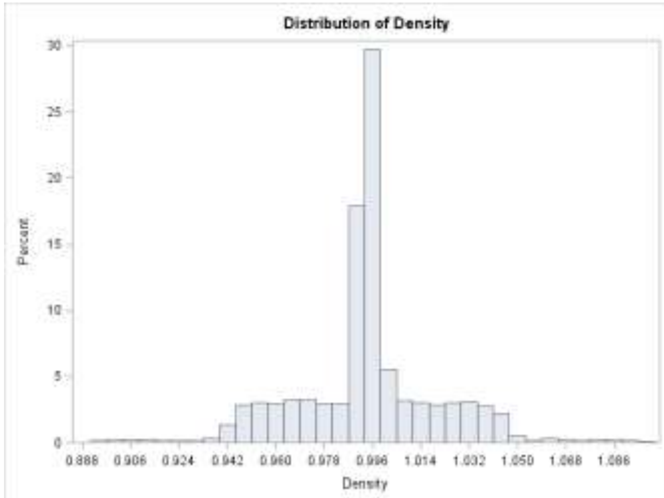




Density
Equipment such as hydrometers used to measure the density of wine give their readings in terms of specific gravity, which is the density of a liquid relative to pure water. Wines that are equally as dense as pure water have a specific gravity of 1. If a wine is denser than water, it will have a value over one. If its density is less than that of water, it will have a number between 0 and 1. The density of wine increases with more dissolved material and sugars make up most of the dissolved material. As yeast convert sugar to

alcohol, the density of the must decreases, both from the loss of sugar and from the increase in alcohol, which is less dense than water. In this case, wine density is an indirect measurement of sugar and alcohol content.[8,9]

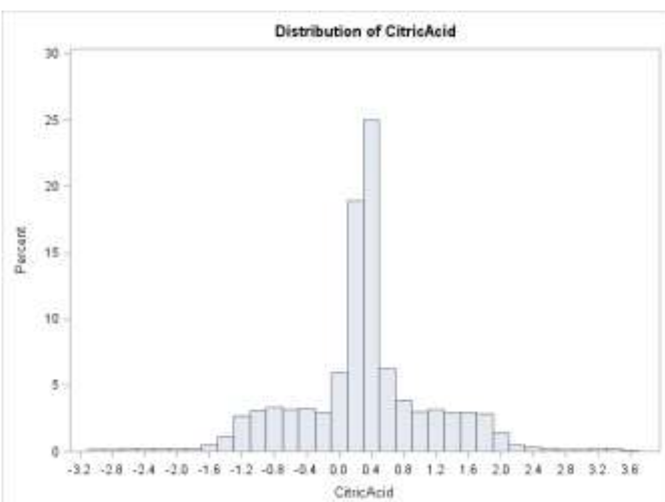
The shape of the Density histogram resembles a plateau with low kurtosis except there is a large central spike at the mean value of 0.9942027. The Kolmogorov-Smirnov test for normality results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.

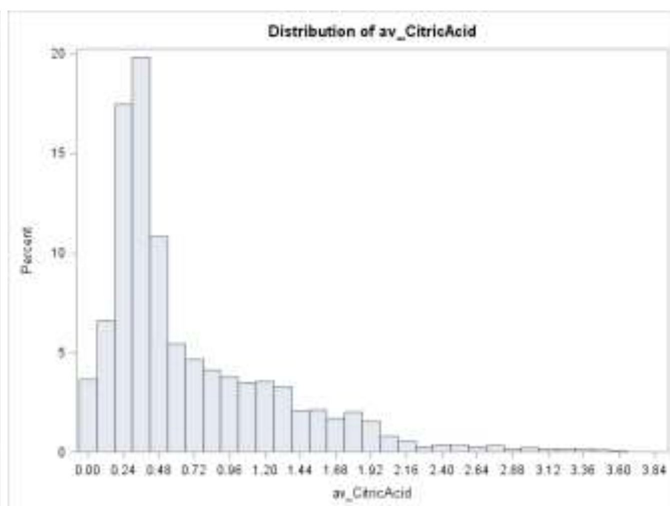


CitricAcid and av_CitricAcid

Citric acid is often added to wines to increase acidity, complement a specific flavor or prevent ferric hazes. It can be added to finished wines to increase acidity and give a fresh flavor. The disadvantage of adding citric acid is its microbial instability. Since bacteria use citric acid in their metabolism, it may increase the growth of unwanted microbes. Often to increase acidity of wine, winemakers will often add tartaric acid instead.[10]

The shape of the CitricAcid histogram resembles a plateau with low kurtosis except there is a large central spike at the mean value of 0.3084127. The histogram for av_CitricAcid is positively skewed with a long right tail. The Kolmogorov-Smirnov test for normality for both histograms results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.

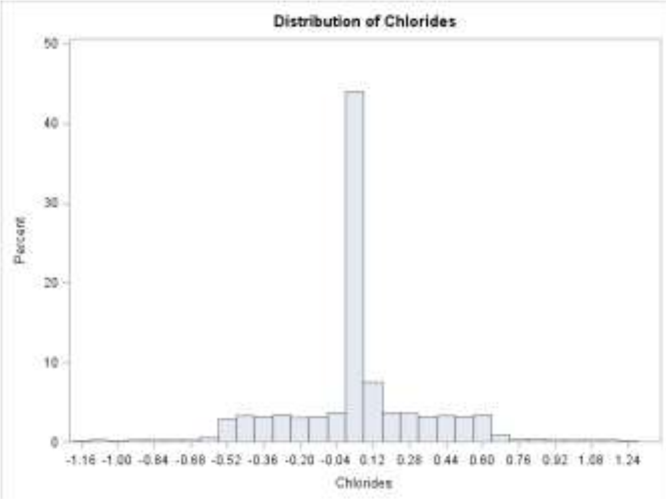


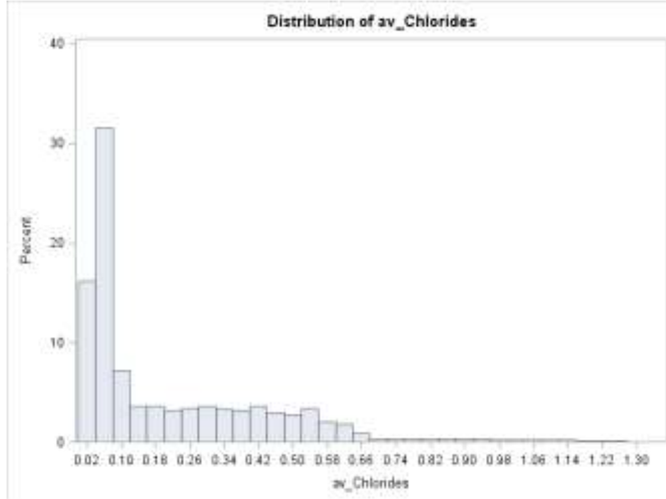


Chlorides and av_Chlorides

Wine contains from 2 to 4 g/L of salts of mineral acids and organic acids. These salts play a key role in the potential salty taste of a wine, with chlorides being a major contributor to saltiness. Moderate to large concentrations of chlorides and sodium might give the wine a salty flavor which may turn away potential consumers.[11]

The shape of the Chlorides histogram resembles a plateau with low kurtosis except there is a large central spike at the mean value of 0.0548225 which is accentuated with imputed mean values. The shape of the av_Chlorides histogram is highly positively skewed with practically no left half/tail of the curve. The Kolmogorov-Smirnov test for normality for both histograms results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.

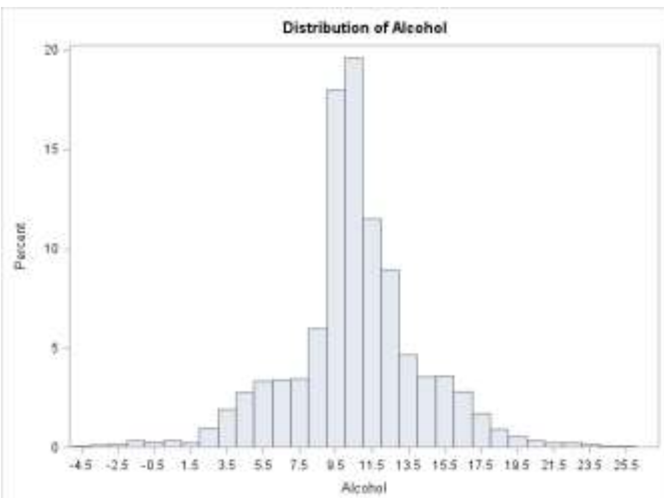


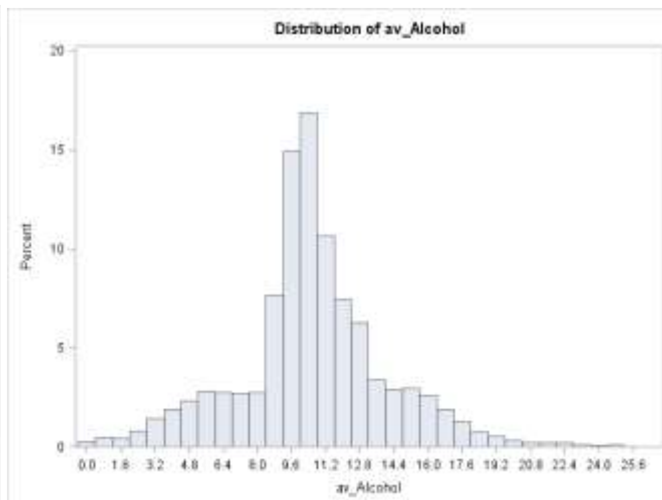


Alcohol and av_Alcohol

Recently, the alcohol content of wine has spiked considerably. There's pressure on winemakers from critics for intense flavors, and that means riper grapes. During the past few years, winemakers have been leaving grapes on the vines well after they would typically be picked, and that translates into fuller-bodied wines and more alcohol. Alcohol content of wine ranges normally between 5% to 21%. Wines are normally classified as very low (under 12.5%), moderately low (12.5% to 13.5%), high (13.5% to 14.5%), and very high (more than 14.5%).[12]

The shape of the Alcohol histogram resembles a plateau with low kurtosis except there is a large central spike at the mean value of 10.4892363 which is accentuated with imputed mean values. The av_Alcohol histogram is similarly shaped like a plateau with a large central spike. The Kolmogorov-Smirnov test for normality for both histograms results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.

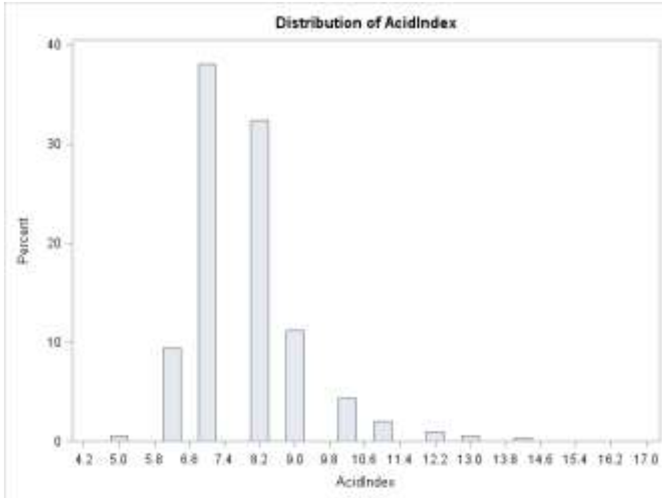




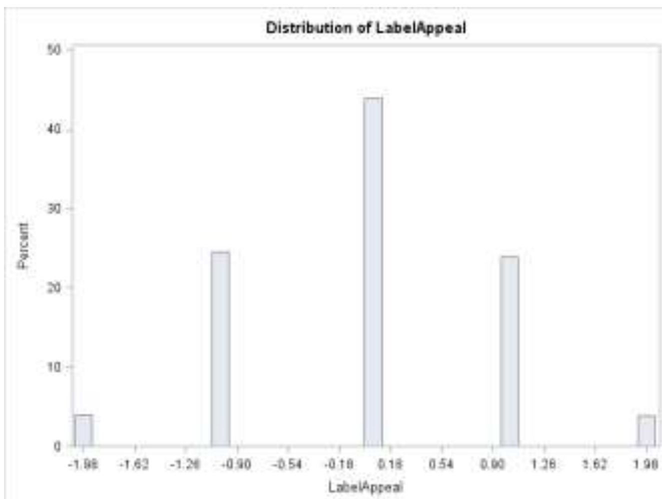
AcidIndex

Acid balance is a matter of taste and there is no set rule that determines the right acid balance. However, there are general guidelines to determine if acide balance is within the desired range for the type and style of wine. The formula for the AcidIndex (or Index of Acidity or Acid Taste Index) is to subtract pH from Total Acidity. Dry red wines should have an AcidIndex range of about 2 to 3, dry white wines about 2.7 to 3.7, and off-dry white wines about 3.8 to 4.8. AcidIndex numbers below these levels will result in flabby or soapy tasting wines while those far above them will taste sharp and acidic. Since the AcidIndex values of this data set are integers and peak between 7 and 8 with almost no values between 2 to 5, this may instead be a subjective rating of acidity from the wine consumer.[13]

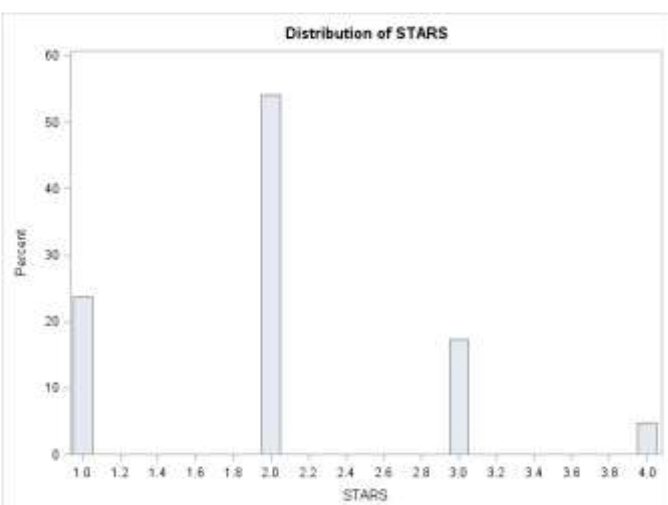
The shape of the AcidIndex histogram resembles a normal distribution which is slightly positively skewed with a peak at the mean value of 7.7727237. The Kolmogorov-Smirnov test for normality results in a significant p-value ($p < 0.01$) which suggests that the data does not follow a normal distribution.

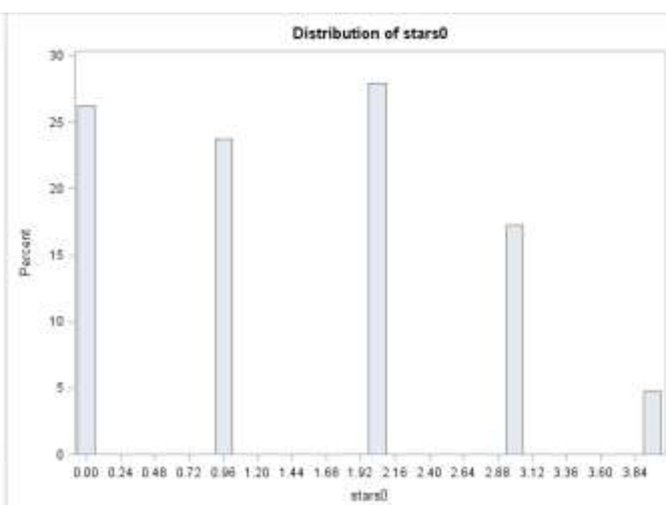


LabelAppeal
The histogram of LabelAppeal is shaped like a triangle pyramid with the peak at 0. Fortunately, there are no missing observations in LabelAppeal so there are no additional forms that need to be derived. I would expect that the most highly rated wine label designs (scores = 1 and 2) to be associated with a greater number of wine cases purchased and be the very highly correlated with TARGET.



STARS and STARS0
For both STARS and STARS0, most of the wines are rated with a score of 0. The missing values imputed with the mean (in STARS) or set to 0 (in STARS0) make up a very large portion of the responses. I would expect that the most highly rated wines (3 and 4 stars) to be associated with a greater number of wine cases purchased and be the most strongly correlated predictor variable with TARGET.





2. Data Preparation

Addressing missing observations
As mentioned in the last section, there are many variables have missing observations in the data set, most notably STARS with the most at 3359. For the continuous variables, I imputed the missing values with the mean value. For STARS, which I will also use as a categorical variable, I rounded the mean value to the nearest integer and created a binary indicator variable to flag when the data is missing.

Addressing negative values
Many of the variables have negative values which do not make sense because they are a frequency, amount, or concentration of a particular substance which can only take on positive values, including: Alcohol, Chlorides, CitricAcid, FixedAcidity, FreeSulfurDioxide, ResidualSugar, Sulphates, TotalSulfurDioxide, and VolatileAcidity. I added the absolute value of the minimum negative value to all of the observations of variables with negative values to make sure the whole distribution is positive with a minimum value of 0. These new "reshifted" variables were renamed with the prefix "rs_". I also applied the absolute value to variables with negative values to make sure all of the observations were positive. These new absolute value transformed variables were renamed with the prefix "av_".

Adding new variables
I renamed Sulphates to Sulfites. I added a "0" class for the missing values of STARS to make STARS0. I also derived BoundSulfurDioxide variables and TotalAcidity variables from existing variables.

New BoundSulfurDioxide Variables added:

- BoundSulfurDioxide = TotalSulfurDioxide - FreeSulfurDioxide
  - with its missing values replaced by its mean value
- BoundSulfurDioxide2 = TotalSulfurDioxide - FreeSulfurDioxide
  - after the missing values from TotalSulfurDioxide and FreeSulfurDioxide have had their missing values imputed with their mean values
- rs_BoundSulfurDioxide = abs(rs_TotalSulfurDioxide – rs_FreeSulfurDioxide)
- rs_BoundSulfurDioxide2 = BoundSulfurDioxide + abs(min(BoundSulfurDioxide))
- av_BoundSulfurDioxide = abs(abs(TotalSulfurDioxide) – abs(FreeSulfurDioxide))
- av_BoundSulfurDioxide2 = abs(BoundSulfurDioxide2)

New TotalAcidity variables added:

- TotalAcidity = FixedAcidity + VolatileAcidity
- rs_TotalAcidity = abs(rs_FixedAcidity + rs_VolatileAcidity)
- rs_TotalAcidity2 = TotalAcidity + abs(min(TotalAcidity))
- av_TotalAcidity = abs(abs(FixedAcidity) + abs(VolatileAcidity))
- av_TotalAcidity2 = abs(TotalAcidity)

Almost all of the absolute value transformed variables plus AcidIndex were positively skewed. Therefore, I added the natural logarithm and square root transform of all of the "av_" prefix variables plus AcidIndex in order to make the distribution appear more normal. The chart below details all of the new variables added.

Derived from original variables	Indicator variables for missing values	Reshifted
BoundSulfurDioxide	I_Alcohol	rs_Alcohol
BoundSulfurDioxide2	I_BoundSulfurDioxide	rs_Chlorides
TotalAcidity	I_Chlorides	rs_CitricAcid
STARS0	I_FreeSulfurDioxide	rs_FixedAcidity
	I_ResidualSugar	rs_FreeSulfurDioxide
	I_STARS	rs_ResidualSugar
	I_Sulfites	rs_Sulfites
	I_TotalSulfurDioxide	rs_TotalSulfurDioxide
	I_pH	rs_VolatileAcidity
		rs_BoundSulfurDioxide
		rs_BoundSulfurDioxide2
		rs_TotalAcidity
		rs_TotalAcidity2

Absolute Value	Natural Logarithm	Square Root
av_Alcohol	ln_av_Alcohol	sr_av_Alcohol
av_Chlorides	ln_av_Chlorides	sr_av_Chlorides
av_CitricAcid	ln_av_CitricAcid	sr_av_CitricAcid
av_FixedAcidity	ln_av_FixedAcidity	sr_av_FixedAcidity
av_FreeSulfurDioxide	ln_av_FreeSulfurDioxide	sr_av_FreeSulfurDioxide
av_ResidualSugar	ln_av_ResidualSugar	sr_av_ResidualSugar
av_Sulfites	ln_av_Sulfites	sr_av_Sulfites
av_TotalSulfurDioxide	ln_av_TotalSulfurDioxide	sr_av_TotalSulfurDioxide
av_VolatileAcidity	ln_av_VolatileAcidity	sr_av_VolatileAcidity
av_BoundSulfurDioxide	ln_av_BoundSulfurDioxide	sr_av_BoundSulfurDioxide
av_BoundSulfurDioxide2	ln_av_BoundSulfurDioxide2	sr_av_BoundSulfurDioxide2
av_TotalAcidity	ln_av_TotalAcidity	sr_av_TotalAcidity
av_TotalAcidity2	ln_av_TotalAcidity2	sr_av_TotalAcidity2
	ln_AcidIndex	sr_AcidIndex

3. Build Models

Next, I examine the correlations of all the variables with TARGET. The correlation table below lists the correlation coefficients by largest to smallest. It makes sense that the 3 subjective rating variables: STARS, LabelAppeal, and AcidIndex are among the most highly correlated variables to TARGET. Since they are so highly correlated with TARGET, I may not have to use them as categorical variables. I am not surprised that STARS and STARS0 are in the top 3 of most correlated with TARGET as both are subjective ratings of wine quality. However, it is surprising that I_STARS is the second most correlated with TARGET. Perhaps the wines that were not rated were, on average, actually low quality wines and were not highly purchased. In fact, Sulfur dioxide levels, acidity levels, and alcohol content are all truly subjective measures because they vary and are dependent upon the type of wine that most appeals to the consumer. I expect volatile, fixed, total acidity, and pH correlation coefficients to be in the same region of the chart. VolatileAcidity variables are the most correlated with TARGET and FixedAcidity variables are the least correlated of the 3 acidity variables. However, pH is very far away from the acidity variables appearing near the bottom of the chart, which means it is not a similar acidity measurement.

Correlation with TARGET					
Variable	**Correlation**				
STARS0	0.68538	ln_av_Chlorides	-0.05251	av_FreeSulfurDioxide	0.0236
i_stars	-0.57158	TotalSulfurDioxide	0.0501	BoundSulfurDioxide	0.02141
STARS	0.40013	rs_TotalSulfurDioxide	0.0501	rs_BoundSulfurDioxide2	0.02141
LabelAppeal	0.3565	FixedAcidity	-0.04901	ln_av_ResidualSugar	0.01931
AcidIndex	-0.24605	rs_FixedAcidity	-0.04901	BoundSulfurDioxide2	0.0192
sr_AcidIndex	-0.24311	ln_av_Alcohol	0.04897	rs_BoundSulfurDioxide	0.01618
ln_AcidIndex	-0.23847	ln_av_TotalAcidity	-0.04882	ResidualSugar	0.01607
VolatileAcidity	-0.08879	sr_av_FixedAcidity	-0.04841	rs_ResidualSugar	0.01607
rs_VolatileAcidity	-0.08879	sr_av_FreeSulfurDioxide	0.04323	sr_av_BoundSulfurDioxide2	0.01399
ln_av_TotalSulfurDioxide	0.08617	FreeSulfurDioxide	0.04269	av_CitricAcid	0.01395
ln_av_VolatileAcidity	-0.08405	rs_FreeSulfurDioxide	0.04269	i_sulfites	-0.0125
sr_av_VolatileAcidity	-0.08106	sr_av_Chlorides	-0.03852	ln_av_BoundSulfurDioxide	0.01209
ln_av_FreeSulfurDioxide	0.07774	ln_av_Sulfites	-0.03805	i_residualsugar	0.0112
av_VolatileAcidity	-0.07019	Chlorides	-0.03724	i_ph	-0.00997
av_TotalAcidity2	-0.06248	rs_Chlorides	-0.03724	pH	-0.00928
av_Alcohol	0.06173	Sulfites	-0.03691	sr_av_ResidualSugar	0.00922
Alcohol	0.06043	rs_Sulfites	-0.03691	CitricAcid	0.00868
rs_Alcohol	0.06043	sr_av_Sulfites	-0.03557	rs_CitricAcid	0.00868
av_TotalAcidity	-0.06035	Density	-0.03552	av_BoundSulfurDioxide2	0.00685
TotalAcidity	-0.05948	ln_av_FixedAcidity	-0.03409	i_totalsulfurdioxide	0.00617
rs_TotalAcidity	-0.05948	av_TotalSulfurDioxide	0.03334	av_BoundSulfurDioxide	-0.00531
rs_TotalAcidity2	-0.05948	av_Sulfites	-0.03127	i_boundsulfurdioxide	0.00469
sr_av_Alcohol	0.05845	sr_av_CitricAcid	0.03	i_chlorides	0.00269
sr_av_TotalAcidity	-0.05798	ln_av_CitricAcid	0.02781	av_ResidualSugar	0.00176
sr_av_TotalSulfurDioxide	0.05779	av_Chlorides	-0.02778	i_alcohol	0.00148
sr_av_TotalAcidity2	-0.05576	ln_av_BoundSulfurDioxide2	0.0262	sr_av_BoundSulfurDioxide	0.00103
av_FixedAcidity	-0.05298	ln_av_TotalAcidity2	-0.02602	i_freesulfurdioxide	-0.00015

The original data set included 14 predictor variables. I think a model with 9 or a little over half of the number of predictor variables should make the most accurate predictions while remaining parsimonious. I will only take one form of the variables with the highest correlation coefficients to prevent multicollinearity problems. Based on this chart, I can narrow down an initial list of candidate variables to include in my model: STARS0, LabelAppeal, AcidIndex, rs_VolatileAcidity, ln_av_TotalSulfurDioxide, ln_av_FreeSulfurDioxide, rs_TotalAcidity2, and av_Alcohol.

stars0
LabelAppeal
AcidIndex
VolatileAcidity
ln_av_TotalSulfurDioxide
ln_av_FreeSulfurDioxide
rs_TotalAcidity2
av_Alcohol
av_Fixed_Acidity

It will be interesting to potentially use 3 subjective variables and 6 physical variables. Additionally, none of these variables appear to be highly correlated to each other so there should not be any multicollinearity problems.

Pearson Correlation Coefficients, N = 12795
Prob > |r| under H0: Rho=0

STARS0	STARS0 1.00000	TARGET 0.68538 <.0001	LabelAppeal 0.26470 <.0001	AcidIndex -0.17093 <.0001	rs_VolatileAcidity -0.06228 <.0001	av_Alcohol 0.05847 <.0001	ln_av_TotalSulfurDioxide 0.05456 <.0001	rs_TotalAcidity2 -0.04275 <.0001
LabelAppeal	LabelAppeal 1.00000	TARGET 0.35650 <.0001	STARS0 0.26470 <.0001	AcidIndex 0.02475 0.0051	rs_VolatileAcidity -0.01699 0.0547	ln_av_TotalSulfurDioxide -0.01564 0.0769	rs_TotalAcidity2 -0.00542 0.5395	av_Alcohol 0.00238 0.7877
AcidIndex	AcidIndex 1.00000	TARGET -0.24605 <.0001	rs_TotalAcidity2 0.18230 <.0001	STARS0 -0.17093 <.0001	ln_av_TotalSulfurDioxide -0.09678 <.0001	rs_VolatileAcidity 0.04464 <.0001	av_Alcohol -0.03672 <.0001	LabelAppeal 0.02475 0.0051
rs_VolatileAcidity	rs_VolatileAcidity 1.00000	rs_TotalAcidity2 0.13523 <.0001	TARGET -0.08879 <.0001	STARS0 -0.06228 <.0001	AcidIndex 0.04464 <.0001	ln_av_TotalSulfurDioxide -0.02981 0.0007	LabelAppeal -0.01699 0.0547	av_Alcohol 0.00340 0.7008
ln_av_TotalSulfurDioxide	ln_av_TotalSulfurDioxide 1.00000	AcidIndex -0.09678 <.0001	TARGET 0.08617 <.0001	STARS0 0.05456 <.0001	rs_VolatileAcidity -0.02981 0.0007	av_Alcohol -0.02843 0.0013	rs_TotalAcidity2 -0.02606 0.0032	LabelAppeal -0.01564 0.0769
rs_TotalAcidity2	rs_TotalAcidity2 1.00000	AcidIndex 0.18230 <.0001	rs_VolatileAcidity 0.13523 <.0001	TARGET -0.05948 <.0001	STARS0 -0.04275 <.0001	ln_av_TotalSulfurDioxide -0.02606 0.0032	av_Alcohol -0.00893 0.3124	LabelAppeal -0.00542 0.5395
av_Alcohol	av_Alcohol 1.00000	TARGET 0.06173 <.0001	STARS0 0.05847 <.0001	AcidIndex -0.03672 <.0001	ln_av_TotalSulfurDioxide -0.02843 0.0013	rs_TotalAcidity2 -0.00893 0.3124	rs_VolatileAcidity 0.00340 0.7008	LabelAppeal 0.00238 0.7877
TARGET	TARGET 1.00000	STARS0 0.68538 <.0001	LabelAppeal 0.35650 <.0001	AcidIndex -0.24605 <.0001	rs_VolatileAcidity -0.08879 <.0001	ln_av_TotalSulfurDioxide 0.08617 <.0001	av_Alcohol 0.06173 <.0001	rs_TotalAcidity2 -0.05948 <.0001

The GENMOD procedures in SAS do not provide us with a method for automatic variable selection. Therefore, I will have to use PROC HPGENSELECT (available with SAS 9.4) to conduct automated variable selection for Poisson and Negative Binomial models. For linear regression models I will apply PROC REG to utilize the automated variable selection methods to find the best variables to include in my model. Then, I will examine different combinations of variables with different models and compare their respective performance in predicting TARGET.

With HPGENSELECT with poisson link function and logarithm distribution (link=poi, dist=log) using stepwise variable selection with an entry significance level of 0.05 (SLENTRY = 0.05) and stay significance level of 0.05 (SLSTAY = 0.05), the first 9 unique variables to be added and stay in the model are: STARS0, LabelAppeal, AcidIndex, ln_av_TotalSulfurDioxide, ln_av_VolatileAcidity, ln_av_FreeSulfurDioxide, av_Alcohol, BoundSulfurDioxide, and ln_av_chlorides. When I use a negative binomial link function keeping all other settings the same, the first 9 unique variables to be added and stay in the model are the same as with a poisson link function. When I use an identity link function keeping all other settings the same, and normal distribution, the first 9 unique variables remain the same except for av_BoundSulfurDioxide.

stars0	stars0
LabelAppeal	LabelAppeal
AcidIndex	AcidIndex

ln_av_TotalSulfurDioxide	ln_av_TotalSulfurDioxide
ln_av_VolatileAcidity	ln_av_VolatileAcidity
ln_av_FreeSulfurDioxide	ln_av_FreeSulfurDioxide
av_Alcohol	av_Alcohol
BoundSulfurDioxide	av_BoundSulfurDioxide
ln_av_chlorides	ln_av_chlorides

When I make STARS, STARS0, I_STARS, and LabelAppeal categorical variables and otherwise keep the same settings, the first 9 unique variables to be added and stay in the model are: STARS0, LabelAppeal, AcidIndex, VolatileAcidity, ln_av_TotalSulfurDioxide, av_BoundSulfurDioxide, av_Alcohol, ln_av_FreeSulfurDioxide, and ln_av_chlorides. When I use a negative binomial link function keeping all other settings the same, the first 9 unique variables are the same as with a poisson link function. When I use an identity link function keeping all other settings the same, and normal distribution, the first 9 unique variables remain the same except for ln_av_VolatileAcidity.

stars0	stars0
LabelAppeal	LabelAppeal
AcidIndex	AcidIndex
VolatileAcidity	ln_av_VolatileAcidity
ln_av_TotalSulfurDioxide	ln_av_TotalSulfurDioxide
av_BoundSulfurDioxide	av_BoundSulfurDioxide
av_Alcohol	av_Alcohol
ln_av_FreeSulfurDioxide	ln_av_FreeSulfurDioxide
ln_av_chlorides	ln_av_chlorides

Putting the tables of selected variables together and reordering the variables, I find that 6 of the variables are represented at least 4 times in 4 lists (highlighted in yellow) and 2 variables are represented at least 3 times in 3 lists (highlighted in blue).

Top 9 Selected from Correlation Table	**Top 9 from Stepwise variable selection with all quantitative variables**		**Top 9 from Stepwise variable selection with 4 categorical variables**	
stars0	stars0	stars0	stars0	stars0
LabelAppeal	LabelAppeal	LabelAppeal	LabelAppeal	LabelAppeal
AcidIndex	AcidIndex	AcidIndex	AcidIndex	AcidIndex
VolatileAcidity	ln_av_TotalSulfurDioxide	ln_av_TotalSulfurDioxide	VolatileAcidity	ln_av_VolatileAcidity
ln_av_TotalSulfurDioxide	ln_av_VolatileAcidity	ln_av_VolatileAcidity	ln_av_TotalSulfurDioxide	ln_av_TotalSulfurDioxide
ln_av_FreeSulfurDioxide	ln_av_FreeSulfurDioxide	ln_av_FreeSulfurDioxide	av_BoundSulfurDioxide	av_BoundSulfurDioxide
rs_TotalAcidity2	av_Alcohol	av_Alcohol	av_Alcohol	av_Alcohol
av_Alcohol	BoundSulfurDioxide	av_BoundSulfurDioxide	ln_av_FreeSulfurDioxide	ln_av_FreeSulfurDioxide
av_Fixed_Acidity	ln_av_chlorides	ln_av_chlorides	ln_av_chlorides	ln_av_chlorides

Normally, I would limit the selected variable subset to the 6 variables that appear 4 times in 4 lists. However, ln_av_VolatileAcidity is the first non-subjective physical measurement variable (along with ln_av_TotalSulfurDioxide) so it must be an important predictor variable. I do not choose to include av_BoundSulfurDioxide because it is formed from combining of TotalSulfurDioxide and FreeSulfurDioxide, both of which appear earlier/higher on the lists. Although ln_av_chlorides is represented 4 times in 4 lists, it very relatively low on the correlation table appearing at the top of the second column (in the chart above) while all of the other 7 variables appear in the closely together in the top half of the first column. In the table, av_alcohol, the last of the 7 variables grouped together, is #16, while ln_av_chlorides is #28. There is a big jump between these 2 variables. There are even 2 unique variable types, TotalAcidity (TotalAcidity2) and FixedAcidity that appear in the big gap between these 2 variables.

Correlation with TARGET (first 29)		
Position	**Variable**	**Correlation**
1	STARS0	0.68538
2	i_stars	-0.57158
3	STARS	0.40013
4	LabelAppeal	0.3565
5	AcidIndex	-0.24605
6	sr_AcidIndex	-0.24311
7	ln_AcidIndex	-0.23847
8	VolatileAcidity	-0.08879
9	rs_VolatileAcidity	-0.08879
10	ln_av_TotalSulfurDioxide	0.08617
11	ln_av_VolatileAcidity	-0.08405

12	sr_av_VolatileAcidity	-0.08106
13	ln_av_FreeSulfurDioxide	0.07774
14	av_VolatileAcidity	-0.07019
15	av_TotalAcidity2	-0.06248
16	av_Alcohol	0.06173
17	Alcohol	0.06043
18	rs_Alcohol	0.06043
19	av_TotalAcidity	-0.06035
20	TotalAcidity	-0.05948
21	rs_TotalAcidity	-0.05948
22	rs_TotalAcidity2	-0.05948
23	sr_av_Alcohol	0.05845
24	sr_av_TotalAcidity	-0.05798
25	sr_av_TotalSulfurDioxide	0.05779
26	sr_av_TotalAcidity2	-0.05576
27	av_FixedAcidity	-0.05298
28	ln_av_Chlorides	-0.05251

I will continue with the following subset of 7 variables, once with STARS0 and LabelAppeal as quantitative variables, and once with stars0 and LabelAppeal as categorical variables for each of the models.

Selected Variables
stars0
LabelAppeal
AcidIndex
ln_av_VolatileAcidity
ln_av_TotalSulfurDioxide
av_Alcohol
ln_av_FreeSulfurDioxide

For Poisson and Negative Binomial models using PROC GENMOD I continue to examine this subset of variables by attempting to explore comprehensively what variables make sense to incorporate. This becomes more difficult when working with the Zero Inflated variants of the model as I will need to produce frequency tables to examine which variables conditionally contribute to the probability that would result in a zero count in the TARGET. I expect to see very similar models from the Poisson and Negative Binomial approaches due to the TARGET variance being close to equal with the TARGET mean.

Model 1: GENMOD with Poisson distribution and all quantitative variables

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero, the logarithm of expected number of wine cases purchased would be 0.8933
- If a wine increased its stars0 rating by 1 point, the logarithm of expected number of wine cases purchased would be expected to increase by 0.3114.
- If a wine increased its LabelAppeal score by 1 point, the logarithm of expected number of wine cases purchased would be expected to increase by 0.1339.
- If a wine increased its AcidIndex score by 1 point, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0848.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0293.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0347.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0187.
- If a wine increased it absolute value transformed alcohol content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0028.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8933	0.0554	0.7847	1.0018	260.16	<.0001
STARS0	1	0.3114	0.0045	0.3025	0.3203	4725.53	<.0001
LabelAppeal	1	0.1330	0.0061	0.1211	0.1449	481.22	<.0001
AcidIndex	1	-0.0848	0.0045	-0.0936	-0.0760	355.82	<.0001
ln_av_VolatileAcidit	1	-0.0293	0.0057	-0.0406	-0.0180	25.96	<.0001
ln_av_TotalSulfurDio	1	0.0347	0.0060	0.0230	0.0464	33.82	<.0001
ln_av_FreeSulfurDiox	1	0.0187	0.0047	0.0095	0.0278	16.04	<.0001
av_Alcohol	1	0.0028	0.0014	-0.0000	0.0056	3.75	0.0528
Scale	0	1.0000	0.0000	1.0000	1.0000		

The Deviance, Log Likelihood, AIC, AICC, and BIC are all fairly high. I will need to compare these values with those of other models to pick the best model.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	14720.1931	1.1512
Scaled Deviance	13E3	14720.1931	1.1512
Pearson Chi-Square	13E3	10885.0897	0.8513
Scaled Pearson X2	13E3	10885.0897	0.8513
Log Likelihood		8266.0635	
Full Log Likelihood		23331.1074	
AIC (smaller is better)		46678.2148	
AICC (smaller is better)		46678.2261	
BIC (smaller is better)		46737.8693	

Model 2: GENMOD with Poisson distribution and STARS0 and LabelAppeal as categorical variables

In the case of the categorical variables with a Poisson distribution, the exponentiated coefficient is the multiplicative term relative to the base level for each variable. The exponentiated intercept is the baseline rate, and all other estimates will be relative to it.

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero (STARS0 at 0 and LabelAppeal at -2), the logarithm of expected number of wine cases purchased would be 0.1413.
- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:
  - 1 rating: 0.7676 increase in the logarithm of expected number of cases purchased
  - 2 rating: 1.0838 increase in the logarithm of expected number of cases purchased
  - 3 rating: 1.2051 increase in the logarithm of expected number of cases purchased
  - 4 rating: 1.3272 increase in the logarithm of expected number of cases purchased
- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
  - -1 rating: 0.2381 increase in the logarithm of expected number of cases purchased
  - 0 rating: 0.4274 increase in the logarithm of expected number of cases purchased
  - +1 rating: 0.5602 increase in the logarithm of expected number of cases purchased
  - +2 rating: 0.6962 increase in the logarithm of expected number of cases purchased
- If a wine increased its AcidIndex score by 1 point, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0778.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0262.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0283.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0156.
- If a wine increased it absolute value transformed alcohol content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0043.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.1413	0.0669	0.0101	0.2724	4.46	0.0347
STARS0	1	1	0.7676	0.0195	0.7293	0.8059	1543.96	<.0001
STARS0	2	1	1.0838	0.0182	1.0481	1.1196	3528.56	<.0001
STARS0	3	1	1.2051	0.0192	1.1674	1.2427	3938.99	<.0001
STARS0	4	1	1.3272	0.0243	1.2796	1.3748	2986.04	<.0001
STARS0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	0.2381	0.0380	0.1637	0.3126	39.30	<.0001
LabelAppeal	0	1	0.4274	0.0371	0.3547	0.5000	133.03	<.0001
LabelAppeal	1	1	0.5602	0.0377	0.4863	0.6341	220.90	<.0001
LabelAppeal	2	1	0.6962	0.0424	0.6131	0.7794	269.16	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	-0.0778	0.0045	-0.0867	-0.0690	296.48	<.0001
ln_av_VolatileAcidit		1	-0.0262	0.0057	-0.0374	-0.0149	20.75	<.0001
ln_av_TotalSulfurDio		1	0.0283	0.0060	0.0166	0.0400	22.34	<.0001
ln_av_FreeSulfurDiox		1	0.0156	0.0047	0.0064	0.0247	11.14	0.0008
av_Alcohol		1	0.0043	0.0014	0.0015	0.0071	8.84	0.0030
Scale		0	1.0000	0.0000	1.0000	1.0000		

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	13634.8913	1.0568
Scaled Deviance	13E3	13634.8913	1.0568
Pearson Chi-Square	13E3	11256.6553	0.8807
Scaled Pearson X2	13E3	11256.6553	0.8807
Log Likelihood		8008.7148	
Full Log Likelihood		-22788.4565	
AIC (smaller is better)		45604.9130	
AICC (smaller is better)		45604.9459	
BIC (smaller is better)		45709.3084	

Model 3: GENMOD with Negative Binomial distribution and all quantitative variables

The parameter estimates are the same as in model 1 but the Goodness of Fit criteria are different.

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero, the logarithm of expected number of wine cases purchased would be 0.8933
- If a wine increased its stars0 rating by 1 point, the logarithm of expected number of wine cases purchased would be expected to increase by 0.3114.
- If a wine increased its LabelAppeal score by 1 point, the logarithm of expected number of wine cases purchased would be expected to increase by 0.1339.
- If a wine increased its AcidIndex score by 1 point, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0848.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0293.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0347.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0187.
- If a wine increased it absolute value transformed alcohol content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0028.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8933	0.0554	0.7847	1.0018	260.16	<.0001
STARS0	1	0.3114	0.0045	0.3025	0.3203	4725.42	<.0001
LabelAppeal	1	0.1330	0.0061	0.1211	0.1449	481.21	<.0001
AcidIndex	1	-0.0848	0.0045	-0.0936	-0.0760	355.82	<.0001
ln_av_VolatileAcidit	1	-0.0293	0.0057	-0.0406	-0.0180	25.96	<.0001
ln_av_TotalSulfurDio	1	0.0347	0.0060	0.0230	0.0464	33.82	<.0001
ln_av_FreeSulfurDiox	1	0.0187	0.0047	0.0095	0.0278	16.04	<.0001
av_Alcohol	1	0.0028	0.0014	-0.0000	0.0056	3.75	0.0528
Dispersion	1	0.0000	0.0001	0.0000	1.94E158		

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	14720.1931	1.1512
Scaled Deviance	13E3	14720.1931	1.1512
Pearson Chi-Square	13E3	10885.0766	0.8513
Scaled Pearson X2	13E3	10885.0766	0.8513
Log Likelihood		8266.0639	
Full Log Likelihood		-23331.1074	
AIC (smaller is better)		46680.2148	
AICC (smaller is better)		46680.2289	
BIC (smaller is better)		46747.3261	

Model 4: GENMOD with Negative Binomial distribution and STARS0 and LabelAppeal as categorical variables

The parameter estimates are the same as in model 2 but the Goodness of Fit criteria are different.

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero (STARS0 at 0 and LabelAppeal at -2), the logarithm of expected number of wine cases purchased would be 0.1413.
- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:
  - 1 rating: 0.7676 increase in the logarithm of expected number of cases purchased
  - 2 rating: 1.0838 increase in the logarithm of expected number of cases purchased
  - 3 rating: 1.2051 increase in the logarithm of expected number of cases purchased
  - 4 rating: 1.3272 increase in the logarithm of expected number of cases purchased
- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
  - -1 rating: 0.2381 increase in the logarithm of expected number of cases purchased
  - 0 rating: 0.4274 increase in the logarithm of expected number of cases purchased
  - +1 rating: 0.5602 increase in the logarithm of expected number of cases purchased
  - +2 rating: 0.6962 increase in the logarithm of expected number of cases purchased
- If a wine increased its AcidIndex score by 1 point, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0778.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0262.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0283.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0156.
- If a wine increased it absolute value transformed alcohol content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0043.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.1413	0.0669	0.0101	0.2724	4.46	0.0347
STARS0	1	1	0.7676	0.0195	0.7293	0.8059	1543.96	<.0001
STARS0	2	1	1.0838	0.0182	1.0481	1.1196	3528.56	<.0001
STARS0	3	1	1.2051	0.0192	1.1674	1.2427	3938.98	<.0001
STARS0	4	1	1.3272	0.0243	1.2796	1.3748	2986.04	<.0001
STARS0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	0.2381	0.0380	0.1637	0.3126	39.30	<.0001
LabelAppeal	0	1	0.4274	0.0371	0.3547	0.5000	133.03	<.0001
LabelAppeal	1	1	0.5602	0.0377	0.4863	0.6341	220.90	<.0001
LabelAppeal	2	1	0.6962	0.0424	0.6131	0.7794	269.16	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	-0.0778	0.0045	-0.0867	-0.0690	296.48	<.0001
ln_av_VolatileAcidit		1	-0.0262	0.0057	-0.0374	-0.0149	20.75	<.0001
ln_av_TotalSulfurDio		1	0.0283	0.0060	0.0166	0.0400	22.34	<.0001
ln_av_FreeSulfurDiox		1	0.0156	0.0047	0.0064	0.0247	11.14	0.0008
av_Alcohol		1	0.0043	0.0014	0.0015	0.0071	8.84	0.0030
Dispersion		0	0.0000	0.0000	0.0000	0.0000		

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	13E3	13634.8913	1.0668
Scaled Deviance	13E3	13634.8913	1.0668
Pearson Chi-Square	13E3	11256.6463	0.8807
Scaled Pearson X2	13E3	11256.6463	0.8807
Log Likelihood		8808.7148	
Full Log Likelihood		-22788.4565	
AIC (smaller is better)		45606.9130	
AICC (smaller is better)		45606.9506	
BIC (smaller is better)		45718.7652	

Model 5: GENMOD with Zero Inflated Poisson distribution and all quantitative variables

I produce frequency tables and histograms of the 7 variables and find that STARS0, LabelAppeal, and AcidIndex all have large zero count and are zero inflated. I will incorporate these 3 variables into the zeromodel as they may conditionally contribute to the probability of observing a zero count in the target variable.

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero, the logarithm of expected number of wine cases purchased would be 1.1244.
- If a wine increased its stars0 rating by 1 point, the logarithm of expected number of wine cases purchased would be expected to increase by 0.1013.
- If a wine increased its LabelAppeal score by 1 point, the logarithm of expected number of wine cases purchased would be expected to increase by 0.2332.
- If a wine increased its AcidIndex score by 1 point, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0187.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0138.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0027.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0062.
- If a wine increased it absolute value transformed alcohol content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0069.

For the zero inflated parameter estimates, assuming that all other variables are held constant:

- If all of the predictor variables in the model are evaluated at zero, the log odds of the predicted number of wine cases purchased being zero would be 0.050545.
- If a wine increased its STARS0 rating by 1 point, the odds of the expected number wine cases purchased being zero would decrease by a factor of 0.093462 and by 90.65%.
- If a wine increased its LabelAppeal score by 1 point, odds of the expected number wine cases purchased would increase by a factor of 2.048525 and by 104.85%.

- If a wine increased its AcidIndex score by 1 point, odds of the expected number wine cases purchased would increase by a factor of 1.547901 and by 54.79%.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.1244	0.0581	1.0106	1.2383	374.96	<.0001
STARS0	1	0.1013	0.0052	0.0911	0.1115	378.92	<.0001
LabelAppeal	1	0.2332	0.0063	0.2209	0.2456	1370.52	<.0001
AcidIndex	1	-0.0187	0.0049	-0.0282	-0.0092	14.83	0.0001
ln_av_VolatileAcidit	1	-0.0138	0.0059	-0.0253	-0.0023	5.49	0.0191
ln_av_TotalSulfurDio	1	0.0027	0.0062	-0.0094	0.0148	0.20	0.6568
ln_av_FreeSulfurDiox	1	0.0062	0.0048	-0.0033	0.0156	1.64	0.2005
av_Alcohol	1	0.0069	0.0015	0.0040	0.0098	21.95	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9849	0.2017	-3.3801	-2.5896	219.06	<.0001
STARS0	1	-2.3702	0.0596	-2.4870	-2.2533	1580.49	<.0001
LabelAppeal	1	0.7172	0.0424	0.6342	0.8002	286.66	<.0001
AcidIndex	1	0.4369	0.0250	0.3879	0.4859	305.24	<.0001

Variable	Estimate	exp(β)	exp(β)-1
Intercept	-2.9849	0.050545	-0.94946
STARS0	-2.3702	0.093462	-0.90654
LabelAppeal	0.71712	2.048525	1.048525
AcidIndex	0.4369	1.547901	0.547901

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		40932.5745	
Scaled Deviance		40932.5745	
Pearson Chi-Square	13E3	5932.4742	0.4641
Scaled Pearson X2	13E3	5932.4742	0.4641
Log Likelihood		11130.8840	
Full Log Likelihood		-20466.2873	
AIC (smaller is better)		40956.5745	
AICC (smaller is better)		40956.5989	
BIC (smaller is better)		41046.0563	

Model 6: GENMOD with Zero Inflated Poisson distribution and STARS0 and LabelAppeal as categorical variables

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero (STARS0 at 0 and LabelAppeal at -2), the logarithm of expected number of wine cases purchased would be 0.4532.
- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:
  - 1 rating: 0.0632 increase in the logarithm of expected number of cases purchased
  - 2 rating: 0.1834 increase in the logarithm of expected number of cases purchased
  - 3 rating: 0.2816 increase in the logarithm of expected number of cases purchased
  - 4 rating: 0.3809 increase in the logarithm of expected number of cases purchased
- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
  - -1 rating: 0.4432 increase in the logarithm of expected number of cases purchased
  - 0 rating: 0.7311 increase in the logarithm of expected number of cases purchased
  - +1 rating: 0.9213 increase in the logarithm of expected number of cases purchased
  - +2 rating: 1.0785 increase in the logarithm of expected number of cases purchased
- If a wine increased its AcidIndex score by 1 point, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0191.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0135.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0028.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0059.

- If a wine increased it absolute value transformed alcohol content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0071.

For the zero inflated parameter estimates, assuming that all other variables are held constant:

- If all of the predictor variables in the model are evaluated at zero, the log odds of the predicted number of wine cases purchased being zero would be 0.005539.
- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:
  - 1 rating: decrease odds by a factor of 0.126413 and by 87.36% that the expected number of cases purchased will be zero.
  - 2 rating: decrease odds by a factor of 0.003014 and by 99.70% that the expected number of cases purchased will be zero.
  - 3 rating: decrease odds by a factor of 1.43e-11 and by 100% that the expected number of cases purchased will be zero.
  - 4 rating: decrease odds by a factor of 1.2e-11 and by 100% that the expected number of cases purchased will be zero.
- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
  - -1 rating: increase odds by a factor of 4.392506 and by 339.25% that the expected number of cases purchased will be zero.
  - 0 rating: increase odds by a factor of 9.272008 and by 827.20% that the expected number of cases purchased will be zero.
  - +1 rating: increase odds by a factor of 18.69021 and by 1769.02% that the expected number of cases purchased will be zero.
  - +2 rating: increase odds by a factor of 29.26815 and by 2826.82% that the expected number of cases purchased will be zero.
- If a wine increased its AcidIndex score by 1 point, the odds that the expected number wine cases purchased being zero would increase by a factor of 1.540643 and by 54.06%.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.4532	0.0706	0.3149	0.5915	41.24	<.0001
STARS0	1	1	0.0632	0.0212	0.0217	0.1047	8.92	0.0028
STARS0	2	1	0.1834	0.0198	0.1447	0.2221	86.19	<.0001
STARS0	3	1	0.2816	0.0207	0.2411	0.3222	185.32	<.0001
STARS0	4	1	0.3809	0.0256	0.3307	0.4311	221.18	<.0001
STARS0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	0.4432	0.0413	0.3622	0.5241	115.16	<.0001
LabelAppeal	0	1	0.7311	0.0404	0.6520	0.8102	328.08	<.0001
LabelAppeal	1	1	0.9213	0.0410	0.8409	1.0017	504.08	<.0001
LabelAppeal	2	1	1.0785	0.0456	0.9892	1.1678	560.39	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	-0.0191	0.0048	-0.0286	-0.0096	15.62	<.0001
ln_av_VolatileAcidit		1	-0.0136	0.0059	-0.0250	-0.0019	5.25	0.0219
ln_av_TotalSulfurDio		1	0.0028	0.0062	-0.0093	0.0148	0.20	0.6546
ln_av_FreeSulfurDiox		1	0.0059	0.0048	-0.0036	0.0154	1.52	0.2180
av_Alcohol		1	0.0071	0.0015	0.0042	0.0100	23.38	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.1960	0.3846	-5.9498	-4.4422	182.53	<.0001
STARS0	1	1	-2.0682	0.0751	-2.2153	-1.9211	759.37	<.0001
STARS0	2	1	-5.8044	0.3484	-6.4872	-5.1215	277.54	<.0001
STARS0	3	1	-24.9701	3706.405	-7289.39	7239.450	0.00	0.9946
STARS0	4	1	-25.1438	7089.507	-13920.3	13870.03	0.00	0.9972
STARS0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	1.4799	0.3297	0.8337	2.1260	20.15	<.0001
LabelAppeal	0	1	2.2270	0.3269	1.5863	2.8677	46.42	<.0001
LabelAppeal	1	1	2.9280	0.3320	2.2772	3.5788	77.76	<.0001
LabelAppeal	2	1	3.3765	0.3828	2.6262	4.1268	77.80	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	0.4322	0.0254	0.3824	0.4820	289.83	<.0001

Variable	Class	Estimate	exp(β)	exp(β)-1
Intercept		-5.196	0.005539	-0.99446

STARS0	1	-2.0682	0.126413	-0.87359
STARS0	2	-5.8044	0.003014	-0.99699
STARS0	3	-24.9701	1.43E-11	-1
STARS0	4	-25.1438	1.2E-11	-1
STARS0	0	0	1	0
LabelAppeal	-1	1.4799	4.392506	3.392506
LabelAppeal	0	2.227	9.272008	8.272008
LabelAppeal	1	2.928	18.69021	17.69021
LabelAppeal	2	3.3765	29.26815	28.26815
LabelAppeal	-2	0	1	0
AcidIndex		0.4322	1.540643	0.540643

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance		40796.8372	
Scaled Deviance		40796.8372	
Pearson Chi-Square	13E3	5743.6266	0.4497
Scaled Pearson X2	13E3	5743.6266	0.4497
Log Likelihood		11198.7527	
Full Log Likelihood		-20396.4186	
AIC (smaller is better)		40844.8372	
AICC (smaller is better)		40844.9312	
BIC (smaller is better)		41023.8007	

Model 7: GENMOD with Zero Inflated Negative Binomial distribution and all quantitative variables

The parameter estimates are the same as in model 5 but the Goodness of Fit criteria are different.

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero, the logarithm of expected number of wine cases purchased would be 1.1244.
- If a wine increased its stars0 rating by 1 point, the logarithm of expected number of wine cases purchased would be expected to increase by 0.1013.
- If a wine increased its LabelAppeal score by 1 point, the logarithm of expected number of wine cases purchased would be expected to increase by 0.2332.
- If a wine increased its AcidIndex score by 1 point, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0187.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0138.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0027.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0062.
- If a wine increased it absolute value transformed alcohol content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0069.

For the zero inflated parameter estimates, assuming that all other variables are held constant:

- If all of the predictor variables in the model are evaluated at zero, the log odds of the predicted number of wine cases purchased being zero would be 0.050545.
- If a wine increased its STARS0 rating by 1 point, the odds of the expected number wine cases purchased being zero would decrease by a factor of 0.093462 and by 90.65%.
- If a wine increased its LabelAppeal score by 1 point, odds of the expected number wine cases purchased would increase by a factor of 2.048525 and by 104.85%.
- If a wine increased its AcidIndex score by 1 point, odds of the expected number wine cases purchased would increase by a factor of 1.547901 and by 54.79%.

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	1.1244	0.0581	1.0106	1.2383	374.96	<.0001
STARS0	1	0.1013	0.0052	0.0911	0.1115	378.92	<.0001
LabelAppeal	1	0.2332	0.0063	0.2209	0.2456	1370.52	<.0001
AcidIndex	1	-0.0187	0.0049	-0.0282	-0.0092	14.83	0.0001
ln_av_VolatileAcidit	1	-0.0138	0.0059	-0.0253	-0.0023	5.49	0.0191
ln_av_TotalSulfurDio	1	0.0027	0.0062	-0.0094	0.0148	0.20	0.6568
ln_av_FreeSulfurDiox	1	0.0062	0.0048	-0.0033	0.0156	1.64	0.2005
av_Alcohol	1	0.0069	0.0015	0.0040	0.0098	21.95	<.0001
Dispersion	0	0.0000	0.0000	0.0000	0.0000		

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9849	0.2017	-3.3801	-2.5896	219.06	<.0001
STARS0	1	-2.3702	0.0596	-2.4870	-2.2533	1580.49	<.0001
LabelAppeal	1	0.7172	0.0424	0.6342	0.8002	286.66	<.0001
AcidIndex	1	0.4369	0.0250	0.3879	0.4859	305.24	<.0001

Variable	Estimate	exp(β)	exp(β)-1
Intercept	-2.9849	0.050545	-0.94946
STARS0	-2.3702	0.093462	-0.90654
LabelAppeal	0.71712	2.048525	1.048525
AcidIndex	0.4369	1.547901	0.547901

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		40932.5729	
Scaled Deviance		40932.5729	
Pearson Chi-Square	13E3	5932.4664	0.4641
Scaled Pearson X2	13E3	5932.4664	0.4641
Log Likelihood		-20466.2864	
Full Log Likelihood		-20466.2864	
AIC (smaller is better)		40958.5729	
AICC (smaller is better)		40958.6014	
BIC (smaller is better)		41055.5114	

Model 8: GENMOD with Zero Inflated Negative Binomial distribution and STARS0 and LabelAppeal as categorical variables

The parameter estimates are the same as in model 6, but the Goodness of Fit criteria are different.

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero (STARS0 at 0 and LabelAppeal at -2), the logarithm of expected number of wine cases purchased would be 0.4532.
- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:
  - 1 rating: 0.0632 increase in the logarithm of expected number of cases purchased
  - 2 rating: 0.1834 increase in the logarithm of expected number of cases purchased
  - 3 rating: 0.2816 increase in the logarithm of expected number of cases purchased
  - 4 rating: 0.3809 increase in the logarithm of expected number of cases purchased
- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
  - -1 rating: 0.4432 increase in the logarithm of expected number of cases purchased
  - 0 rating: 0.7311 increase in the logarithm of expected number of cases purchased
  - +1 rating: 0.9213 increase in the logarithm of expected number of cases purchased
  - +2 rating: 1.0785 increase in the logarithm of expected number of cases purchased
- If a wine increased its AcidIndex score by 1 point, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0191.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0135.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0028.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0059.

- If a wine increased it absolute value transformed alcohol content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0071.

For the zero inflated parameter estimates, assuming that all other variables are held constant:

- If all of the predictor variables in the model are evaluated at zero, the log odds of the predicted number of wine cases purchased being zero would be 0.005539.
- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:
  - 1 rating: decrease odds by a factor of 0.126413 and by 87.36% that the expected number of cases purchased will be zero.
  - 2 rating: decrease odds by a factor of 0.003014 and by 99.70% that the expected number of cases purchased will be zero.
  - 3 rating: decrease odds by a factor of 1.43e-11 and by 100% that the expected number of cases purchased will be zero.
  - 4 rating: decrease odds by a factor of 1.2e-11 and by 100% that the expected number of cases purchased will be zero.
- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
  - -1 rating: increase odds by a factor of 4.392506 and by 339.25% that the expected number of cases purchased will be zero.
  - 0 rating: increase odds by a factor of 9.272008 and by 827.20% that the expected number of cases purchased will be zero.
  - +1 rating: increase odds by a factor of 18.69021 and by 1769.02% that the expected number of cases purchased will be zero.
  - +2 rating: increase odds by a factor of 29.26815 and by 2826.82% that the expected number of cases purchased will be zero.
- If a wine increased its AcidIndex score by 1 point, the odds that the expected number wine cases purchased being zero would increase by a factor of 1.540643 and by 54.06%.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.4532	0.0706	0.3149	0.5915	41.24	<.0001
STARS0	1	1	0.0632	0.0212	0.0217	0.1047	8.92	0.0028
STARS0	2	1	0.1834	0.0198	0.1447	0.2221	86.19	<.0001
STARS0	3	1	0.2816	0.0207	0.2411	0.3222	185.32	<.0001
STARS0	4	1	0.3809	0.0256	0.3307	0.4311	221.18	<.0001
STARS0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	0.4432	0.0413	0.3622	0.5241	115.16	<.0001
LabelAppeal	0	1	0.7311	0.0404	0.6520	0.8102	328.08	<.0001
LabelAppeal	1	1	0.9213	0.0410	0.8409	1.0017	504.08	<.0001
LabelAppeal	2	1	1.0785	0.0456	0.9892	1.1678	560.39	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	-0.0191	0.0048	-0.0286	-0.0096	15.62	<.0001
ln_av_VolatileAcidit		1	-0.0136	0.0059	-0.0250	-0.0019	5.25	0.0219
ln_av_TotalSulfurDio		1	0.0028	0.0062	-0.0093	0.0148	0.20	0.6546
ln_av_FreeSulfurDiox		1	0.0059	0.0048	-0.0036	0.0154	1.52	0.2180
av_Alcohol		1	0.0071	0.0015	0.0042	0.0100	23.38	<.0001
Dispersion		0	0.0000	0.0000	0.0000	0.0000		

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.1960	0.3846	-5.9498	-4.4422	182.53	<.0001
STARS0	1	1	-2.0682	0.0751	-2.2153	-1.9211	759.37	<.0001
STARS0	2	1	-5.8044	0.3484	-6.4872	-5.1215	277.54	<.0001
STARS0	3	1	-17.9666	111.7275	-236.949	201.0154	0.03	0.8722
STARS0	4	1	-18.1389	213.5667	-436.722	400.4442	0.01	0.9323
STARS0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	1.4799	0.3297	0.8337	2.1260	20.15	<.0001
LabelAppeal	0	1	2.2270	0.3269	1.5863	2.8677	46.42	<.0001
LabelAppeal	1	1	2.9280	0.3320	2.2772	3.5788	77.76	<.0001
LabelAppeal	2	1	3.3765	0.3828	2.6262	4.1268	77.80	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	0.4322	0.0254	0.3824	0.4820	289.83	<.0001

Variable	Class	Estimate	exp(β)	exp(β)-1
Intercept		-5.196	0.005539	-0.99446

STARS0	1	-2.0682	0.126413	-0.87359
STARS0	2	-5.8044	0.003014	-0.99699
STARS0	3	-24.9701	1.43E-11	-1
STARS0	4	-25.1438	1.2E-11	-1
STARS0	0	0	1	0
LabelAppeal	-1	1.4799	4.392506	3.392506
LabelAppeal	0	2.227	9.272008	8.272008
LabelAppeal	1	2.928	18.69021	17.69021
LabelAppeal	2	3.3765	29.26815	28.26815
LabelAppeal	-2	0	1	0
AcidIndex		0.4322	1.540643	0.540643

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		40796.8374	
Scaled Deviance		40796.8374	
Pearson Chi-Square	13E3	5743.6264	0.4497
Scaled Pearson X2	13E3	5743.6264	0.4497
Log Likelihood		-20398.4187	
Full Log Likelihood		-20398.4187	
AIC (smaller is better)		40846.8374	
AICC (smaller is better)		40846.9392	
BIC (smaller is better)		41033.2576	

Model 9: Linear Regression with all quantitative variables

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero, the expected number of wine cases purchased would be 2.3049.
- If a wine increased its stars0 rating by 1 point, the expected number of wine cases purchased would be expected to increase by 0.9767.
- If a wine increased its LabelAppeal score by 1 point, the expected number of wine cases purchased would be expected to increase by 0.43282.
- If a wine increased its AcidIndex score by 1 point, the expected number of wine cases purchased would be expected to decrease by 0.20192.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the expected number of wine cases purchased would be expected to decrease by 0.08599.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the expected number of wine cases purchased would be expected to increase by 0.08612.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the expected number of wine cases purchased would be expected to increase by 0.04837.
- If a wine increased it absolute value transformed alcohol content by 1 unit, the expected number of wine cases purchased would be expected to increase by 0.01266.

Root MSE	1.32346	R-Square	0.5283
Dependent Mean	3.02907	Adj R-Sq	0.5280
Coeff Var	43.69198		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.30490	0.11918	19.34	<.0001
STARS0	1	0.97670	0.01044	93.52	<.0001
LabelAppeal	1	0.43282	0.01366	31.69	<.0001
AcidIndex	1	-0.20192	0.00906	-22.28	<.0001
ln_av_VolatileAcidity	1	-0.08599	0.01333	-6.45	<.0001
ln_av_TotalSulfurDioxide	1	0.08612	0.01313	6.56	<.0001
ln_av_FreeSulfurDioxide	1	0.04837	0.01039	4.65	<.0001
av_Alcohol	1	0.01266	0.00332	3.81	0.0001

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	2.3049	0.1191	2.0714	2.5384	374.26	<.0001
STARS0	1	0.9767	0.0104	0.9562	0.9972	8751.75	<.0001
LabelAppeal	1	0.4328	0.0137	0.4061	0.4596	1004.63	<.0001
AcidIndex	1	-0.2019	0.0091	-0.2197	-0.1842	496.57	<.0001
ln_av_VolatileAcidit	1	-0.0860	0.0133	-0.1121	-0.0599	41.64	<.0001
ln_av_TotalSulfurDio	1	0.0861	0.0131	0.0604	0.1118	43.07	<.0001
ln_av_FreeSulfurDiox	1	0.0484	0.0104	0.0280	0.0687	21.68	<.0001
av_Alcohol	1	0.0127	0.0033	0.0062	0.0192	14.56	0.0001
Scale	1	1.3230	0.0083	1.3069	1.3394		

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	22397.0977	1.7516
Scaled Deviance	13E3	12795.0000	1.0006
Pearson Chi-Square	13E3	22397.0977	1.7516
Scaled Pearson X2	13E3	12795.0000	1.0006
Log Likelihood		-21737.1311	
Full Log Likelihood		-21737.1311	
AIC (smaller is better)		43492.2622	
AICC (smaller is better)		43492.2763	
BIC (smaller is better)		43559.3735	

Model 10: Linear Regression and STARS0 and LabelAppeal as categorical variables

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero (STARS0 at 0 and LabelAppeal at -2), the expected number of wine cases purchased would be 1.2300.
- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:
    - 1 rating: 1.3677 increase in the expected number of cases purchased
    - 2 rating: 2.3944 increase in the expected number of cases purchased
    - 3 rating: 2.9659 increase in the expected number of cases purchased
    - 4 rating: 3.6576 increase in the expected number of cases purchased
- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
    - -1 rating: 0.3681 increase in the expected number of cases purchased
    - 0 rating: 0.8349 increase in the expected number of cases purchased
    - +1 rating: 1.2992 increase in the expected number of cases purchased
    - +2 rating: 1.8818 increase in the expected number of cases purchased
- If a wine increased its AcidIndex score by 1 point, the expected number of wine cases purchased would be expected to decrease by 0.1940.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the expected number of wine cases purchased would be expected to decrease by 0.0828.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the expected number of wine cases purchased would be expected to increase by 0.0786.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the expected number of wine cases purchased would be expected to increase by 0.0441.
- If a wine increased it absolute value transformed alcohol content by 1 unit, the expected number of wine cases purchased would be expected to increase by 0.0143.

Root MSE	1.30532	R-Square	0.5413
Dependent Mean	3.02907	Adj R-Sq	0.5408
Coeff Var	43.09308		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.23003	0.13103	9.39	<.0001
AcidIndex	1	-0.19397	0.00895	-21.67	<.0001
ln_av_VolatileAcidity	1	-0.08283	0.01315	-6.30	<.0001
ln_av_TotalSulfurDioxide	1	0.07861	0.01296	6.07	<.0001
ln_av_FreeSulfurDioxide	1	0.04412	0.01025	4.30	<.0001
av_Alcohol	1	0.01434	0.00328	4.38	<.0001
STARS0_1	1	1.36774	0.03289	41.59	<.0001
STARS0_2	1	2.39443	0.03201	74.79	<.0001
STARS0_3	1	2.96595	0.03703	80.10	<.0001
STARS0_4	1	3.65759	0.05917	61.82	<.0001
LabelAppeal_n1	1	0.36814	0.06283	5.86	<.0001
LabelAppeal_0	1	0.83488	0.06127	13.63	<.0001
LabelAppeal_p1	1	1.29924	0.06399	20.30	<.0001
LabelAppeal_p2	1	1.88184	0.08431	22.32	<.0001

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.2300	0.1310	0.9734	1.4867	88.22	<.0001
STARS0	1	1	1.3677	0.0329	1.3033	1.4322	1731.63	<.0001
STARS0	2	1	2.3944	0.0320	2.3317	2.4571	5600.36	<.0001
STARS0	3	1	2.9659	0.0370	2.8934	3.0385	6422.79	<.0001
STARS0	4	1	3.6576	0.0591	3.5417	3.7736	3825.40	<.0001
STARS0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	0.3681	0.0628	0.2451	0.4912	34.37	<.0001
LabelAppeal	0	1	0.8349	0.0612	0.7149	0.9549	185.87	<.0001
LabelAppeal	1	1	1.2992	0.0640	1.1739	1.4246	412.68	<.0001
LabelAppeal	2	1	1.8818	0.0843	1.7167	2.0470	498.72	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	-0.1940	0.0089	-0.2115	-0.1764	470.09	<.0001
ln_av_VolatileAcidit		1	-0.0828	0.0131	-0.1086	-0.0571	39.71	<.0001
ln_av_TotalSulfurDio		1	0.0786	0.0130	0.0532	0.1040	36.84	<.0001
ln_av_FreeSulfurDiox		1	0.0441	0.0102	0.0240	0.0642	18.54	<.0001
av_Alcohol		1	0.0143	0.0033	0.0079	0.0208	19.18	<.0001
Scale		1	1.3046	0.0082	1.2887	1.3207		

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	13E3	21777.0803	1.7039
Scaled Deviance	13E3	12795.0000	1.0011
Pearson Chi-Square	13E3	21777.0803	1.7039
Scaled Pearson X2	13E3	12795.0000	1.0011
Log Likelihood		21557.5320	
Full Log Likelihood		-21557.5320	
AIC (smaller is better)		43145.0639	
AICC (smaller is better)		43145.1015	
BIC (smaller is better)		43256.9161	

4. Select Models

As depicted in the previous section all parameter coefficients/estimates were very significant with p-values < 0.001 except for ln_av_VolatileAcidity, ln_av_TotalSulfurDioxide, ln_av_FreeTotalSulfurDioxide, and at times av_Alcohol. All models shared all 4 of these variables. All parameter coefficients signs (being positive or negative) were the same across all 10 models. The parameter coefficient signs were all intuitive. Both ln_av_VolatileAcidity and AcidIndex were negatively associated with TARGET while all of the other variables were positively associated with TARGET. Volatile Acidity is not desirable in wines and Acid Index may be a subjective rating of the acidity of the wine. If you think your wine tastes very acidic, then you probably will not enjoy the wine or order many cases of it.

The chart below details the metrics by which I can judge the 10 models with 7 variables. The odd number models utilize STARS0 and LabelAppeal as quantitative variables while the even number models utilize STARS0 and LabelAppeal as categorical variables. It appears that all of the even number models perform better on almost all metrics (overall lower deviance, lower log likelihood, lower

AIC, lower AICC, lower BIC, higher R-Squared, and higher Adjusted R-Squared values) than their odd number model counterparts. The only metric that is not consistently better is the Pearson Chi Square test statistic that tests that at least one of the predictors' regression coefficient is not equal to zero (so you would want a large Chi Square statistic), however this measure may not be an appropriate to compare against other models. Furthermore, model 10 has higher R-Squared and Adjusted R-Squared values than model 9 meaning that model 10 is a better fitting model to the data in predicting TARGET than model 9. For these reasons, I will choose between only even number models (models with STARS0 and LabelAppeal as categorical variables).

Of the all of the even number models, Model 6 performs the best because it has the lowest AIC, AICC, and BIC values. Model 6 does have very high deviance and log likelihood values, however, deviance and log likelihood are terms that make up the AIC, AICC, and BICC formulas. Deviance and log likelihood are used to calculate AIC, AICC, and BIC. Therefore, keeping AIC, AICC, and BIC low are ultimately more important than keeping deviance and log likelihood lower.

Model	Description	Deviance	Pearson Chi Square	Log Likelihood	AIC	AICC	BIC	R Squared	Adjusted R Squared
Model 1	Poisson with all quantitative variables	14720.1931	10885.0897	8266.0639	46678.2148	46678.2261	46678.8693		
Model 2	Poisson with 2 categorical variables	13634.8913	11256.6553	8808.7148	45604.9130	45604.9459	45709.3084		
Model 3	Negative Binomial with all quantitative variables	14720.1931	10885.0766	8266.0639	46680.2148	46680.2289	46747.3261		
Model 4	Negative Binomial with 2 categorical variables	13634.8913	11256.6463	8808.7148	45606.9130	45606.9506	45718.7652		
Model 5	Zero Inflated Poisson with all quantitative variables	40932.5745	5932.4742	11130.8840	40956.5745	40956.5989	41046.0563		
Model 6	Zero Inflated Poisson with 2 categorical variables	40796.8372	5743.6266	11198.7527	40844.8372	40844.9312	41023.8007		
Model 7	Zero Inflated Negative Binomial with all quantitative variables	40932.5729	5932.4664	-20466.2864	40958.5729	40958.6014	41055.5114		
Model 8	Zero Inflated Negative Binomial with 2 categorical variables	40796.8374	5743.6264	-20398.4187	40846.8374	40846.9392	41033.2573		
Model 9	Regression with all quantitative variables	22397.0977	22397.0977	-21737.1311	43492.2622	42492.2763	43559.3735	0.5283	0.5280
Model 10	Regression with 2 categorical variables	21777.0803	21777.0803	-21557.5320	43145.0639	43145.1015	43256.9161	0.5413	0.5408

The following matrix displays the first 20 observations of the wine training data set and the predicted values from each of the models. The bottom 2 rows of the table show the sum of absolute error and sum of squared error between the actual and predicted values. In even just the first 20 observations, model 6 is one of the top performing models with only a very miniscule difference with the best performing model.

	Actual Values	Predicted Values									
Obs	**TARGET**	**model 1**	**model 2**	**model 3**	**model 4**	**model 5**	**model 6**	**model 7**	**model 8**	**model 9**	**model 10**
1	3	3.06226	3.59694	3.06233	3.59694	3.54087	3.72045	3.54574	3.72045	3.40289	3.62811
2	3	4.09676	3.81408	4.20177	3.81408	3.29729	3.23896	3.29993	3.23896	4.30877	4.08321
3	5	3.65228	3.50689	3.74897	3.50689	3.42613	3.36562	3.42051	3.36562	4.06825	3.87384
4	3	2.28121	2.49443	2.2948	2.49443	2.55218	2.44195	2.55264	2.44195	2.356	2.47948
5	4	2.92057	3.49763	2.94997	3.49763	3.6354	3.83237	3.63521	3.83237	3.36061	3.59585
6	0	1.20486	0.94309	1.21933	0.94309	0.41385	0.44358	0.4062	0.44358	0.78463	0.61874
7	0	1.73359	1.28879	1.75234	1.28879	1.15509	1.2233	1.14013	1.2233	1.63505	1.41012
8	4	6.11354	5.37871	5.63827	5.37871	5.27774	5.24074	5.2702	5.24074	5.23051	5.0597
9	3	2.05234	1.52119	2.07788	1.52119	1.93526	2.04596	1.91951	2.04596	2.06804	1.83806
10	6	6.00909	4.84611	6.03422	4.84611	4.60285	4.76294	4.60576	4.76294	5.53294	5.06722
11	0	3.42439	3.93435	3.42912	3.93435	4.17753	3.99584	4.19008	3.99584	3.59172	3.77587
12	4	2.63702	3.19075	2.68596	3.19075	3.43263	3.64193	3.4303	3.64193	3.11843	3.35719
13	3	3.76596	4.38974	3.7999	4.38974	4.38695	4.41992	4.38735	4.41992	4.01136	4.25125
14	7	5.4369	5.38502	5.49798	5.38502	6.48342	6.00023	6.48553	6.00023	5.29292	5.30156
15	4	1.47016	1.11197	1.47469	1.11197	0.81273	0.89109	0.82551	0.89109	1.21481	1.01949

16	0	1.74446	1.27445	1.74634	1.27445	1.09054	1.18559	1.10554	1.18559	1.5988	1.36493
17	0	1.7193	1.26895	1.71904	1.26895	0.54131	0.58682	0.55885	0.58682	1.64647	1.45928
18	4	4.39497	4.30964	4.47374	4.30964	4.09348	4.24621	4.09568	4.24621	4.57221	4.3829
19	5	2.8845	3.49121	2.9344	3.49121	4.03797	4.28815	4.02729	4.28815	3.4569	3.72776
20	4	2.92664	3.26974	2.93771	3.26974	3.06003	3.02866	3.05953	3.02866	3.25899	3.45232
21	3	2.42094	2.62604	2.41631	2.62604	2.64446	2.53758	2.64995	2.53758	2.50741	2.62918
22	2	2.06254	2.31568	2.09593	2.31568	2.53355	2.41183	2.53169	2.41183	2.16574	2.30825
23	3	2.22555	2.60164	2.22277	2.60164	2.76274	2.74526	2.80961	2.74526	2.404	2.58083
24	4	1.93581	1.42797	1.94934	1.42797	0.84944	0.87861	0.84219	0.87861	1.97605	1.77844
25	4	3.51357	4.14991	3.59469	4.14991	3.7535	3.92394	3.75583	3.92394	3.81792	4.02264
26	0	1.6517	1.0878	1.54274	1.0878	1.77059	1.64863	1.61546	1.64863	1.21574	0.97361
27	4	3.60019	4.28368	3.65739	4.28368	4.43407	4.50505	4.46608	4.50505	3.92438	4.17804
28	6	7.83912	5.50281	6.81312	5.50281	5.72645	5.67084	5.71621	5.67084	5.89719	5.46931
29	4	2.86825	3.76407	3.1766	3.76407	3.83095	4.04721	3.83879	4.04721	3.60279	3.84021
30	3	1.97024	2.22539	1.98723	2.22539	2.54572	2.41082	2.54567	2.41082	2.07664	2.23271
Sum of Absolute Error		37.02637	32.57295	34.9215	32.57295	29.621	29.44818	29.49289	29.44818	31.22962	30.2732
Sum of Squared Error		63.37203	56.08298	57.27475	56.08298	56.92556	54.92481	56.54066	54.92481	50.61291	50.26625

The Best Model is Model 6: GENMOD with Zero Inflated Poisson distribution and STARS0 and LabelAppeal as categorical variables

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero (STARS0 at 0 and LabelAppeal at -2), the logarithm of expected number of wine cases purchased would be 0.4532.
- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:
  - 1 rating: 0.0632 increase in the logarithm of expected number of cases purchased
  - 2 rating: 0.1834 increase in the logarithm of expected number of cases purchased
  - 3 rating: 0.2816 increase in the logarithm of expected number of cases purchased
  - 4 rating: 0.3809 increase in the logarithm of expected number of cases purchased
- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
  - -1 rating: 0.4432 increase in the logarithm of expected number of cases purchased
  - 0 rating: 0.7311 increase in the logarithm of expected number of cases purchased
  - +1 rating: 0.9213 increase in the logarithm of expected number of cases purchased
  - +2 rating: 1.0785 increase in the logarithm of expected number of cases purchased
- If a wine increased its AcidIndex score by 1 point, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0191.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0135.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0028.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0059.
- If a wine increased it absolute value transformed alcohol content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0071.

For the zero inflated parameter estimates, assuming that all other variables are held constant:

- If all of the predictor variables in the model are evaluated at zero, the log odds of the predicted number of wine cases purchased being zero would be 0.005539.
- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:
  - 1 rating: decrease odds by a factor of 0.126413 and by 87.36% that the expected number of cases purchased will be zero.
  - 2 rating: decrease odds by a factor of 0.003014 and by 99.70% that the expected number of cases purchased will be zero.
  - 3 rating: decrease odds by a factor of 1.43e-11 and by 100% that the expected number of cases purchased will be zero.
  - 4 rating: decrease odds by a factor of 1.2e-11 and by 100% that the expected number of cases purchased will be zero.
- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
  - -1 rating: increase odds by a factor of 4.392506 and by 339.25% that the expected number of cases purchased will be zero.
  - 0 rating: increase odds by a factor of 9.272008 and by 827.20% that the expected number of cases purchased will be zero.

  - +1 rating: increase odds by a factor of 18.69021 and by 1769.02% that the expected number of cases purchased will be zero.
  - +2 rating: increase odds by a factor of 29.26815 and by 2826.82% that the expected number of cases purchased will be zero.
- If a wine increased its AcidIndex score by 1 point, the odds that the expected number wine cases purchased being zero would increase by a factor of 1.540643 and by 54.06%.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.4532	0.0706	0.3149	0.5915	41.24	<.0001
STARS0	1	1	0.0632	0.0212	0.0217	0.1047	8.92	0.0028
STARS0	2	1	0.1834	0.0198	0.1447	0.2221	86.19	<.0001
STARS0	3	1	0.2816	0.0207	0.2411	0.3222	185.32	<.0001
STARS0	4	1	0.3809	0.0256	0.3307	0.4311	221.18	<.0001
STARS0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	0.4432	0.0413	0.3622	0.5241	115.16	<.0001
LabelAppeal	0	1	0.7311	0.0404	0.6520	0.8102	328.08	<.0001
LabelAppeal	1	1	0.9213	0.0410	0.8409	1.0017	504.08	<.0001
LabelAppeal	2	1	1.0785	0.0456	0.9892	1.1678	560.39	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	-0.0191	0.0048	-0.0286	-0.0096	15.62	<.0001
ln_av_VolatileAcidit		1	-0.0136	0.0059	-0.0250	-0.0019	5.25	0.0219
ln_av_TotalSulfurDio		1	0.0028	0.0062	-0.0093	0.0148	0.20	0.6546
ln_av_FreeSulfurDiox		1	0.0059	0.0048	-0.0035	0.0154	1.52	0.2180
av_Alcohol		1	0.0071	0.0015	0.0042	0.0100	23.38	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.1960	0.3846	-5.9498	-4.4422	182.53	<.0001
STARS0	1	1	-2.0682	0.0751	-2.2153	-1.9211	759.37	<.0001
STARS0	2	1	-5.8044	0.3484	-6.4872	-5.1215	277.54	<.0001
STARS0	3	1	-24.9701	3706.405	-7289.39	7239.450	0.00	0.9946
STARS0	4	1	-25.1438	7089.507	-13920.3	13870.03	0.00	0.9972
STARS0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	1.4799	0.3297	0.8337	2.1260	20.15	<.0001
LabelAppeal	0	1	2.2270	0.3269	1.5863	2.8677	46.42	<.0001
LabelAppeal	1	1	2.9280	0.3320	2.2772	3.5788	77.76	<.0001
LabelAppeal	2	1	3.3765	0.3828	2.6262	4.1268	77.80	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	0.4322	0.0254	0.3824	0.4820	289.83	<.0001

Variable	Class	Estimate	exp(β)	exp(β)-1
Intercept		-5.196	0.005539	-0.99446
STARS0	1	-2.0682	0.126413	-0.87359
STARS0	2	-5.8044	0.003014	-0.99699
STARS0	3	-24.9701	1.43E-11	-1
STARS0	4	-25.1438	1.2E-11	-1
STARS0	0	0	1	0
LabelAppeal	-1	1.4799	4.392506	3.392506
LabelAppeal	0	2.227	9.272008	8.272008
LabelAppeal	1	2.928	18.69021	17.69021
LabelAppeal	2	3.3765	29.26815	28.26815
LabelAppeal	-2	0	1	0
AcidIndex		0.4322	1.540643	0.540643

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance		40796.8372	
Scaled Deviance		40796.8372	
Pearson Chi-Square	13E3	5743.6266	0.4497
Scaled Pearson X2	13E3	5743.6266	0.4497
Log Likelihood		11198.7527	
Full Log Likelihood		-20396.4186	
AIC (smaller is better)		40844.8372	
AICC (smaller is better)		40844.9312	
BIC (smaller is better)		41023.8007	

5. Model Deployment

The purpose of this assignment was to develop a model to predict the number of cases of wine that will be sold given certain properties of the wine. The wine training data set contained 12,795 observations and 14 variables. Two of the variables were subjective variables which I utilized as both quantitative and categorical variables during the modeling process. There were 12 continuous variables related to the chemical properties of the wine being sold. There were 2 numerical variables for the marketing score based on the visual appeal of the label and wine rating based on number of stars. 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 stores around the United States. The more sample cases purchased, the more likely a wine is to be sold at a high end restaurant. If it is possible to predict the number of cases, the wine manufacturer will be able to adjust their wine offerings with the goal to maximize sales. The purpose of this project was to build a model to predict the number of cases of wine that will be sold given certain properties of the wine. I built several Poisson and Negative Binomial distribution models to predict the target number of cases ordered for each wine. I compared 10 models of 7 variables each and found that the best model was a Zero Inflated Poisson distribution model with the STARS0 and LabelAppeal variables used as categorical variables.

Model 6: Zero Inflated Poisson distribution and STARS0 and LabelAppeal as categorical variables

```
TEMP = 0.4532
+ AcidIndex * -0.0632
+ ln_av_VolatileAcidity * -0.0135
+ ln_av_TotalSulfurDioxide * 0.0028
+ ln_av_FreeSulfurDioxide * 0.0059
+ av_Alcohol * 0.0071
+ (stars0 in ("1")) * 0.0632
+ (stars0 in ("2")) * 0.1834
+ (stars0 in ("3")) * 0.2816
+ (stars0 in ("4")) * 0.3809
+ (LabelAppeal in ("-1")) * 0.4432
+ (LabelAppeal in ("0")) * 0.7311
+ (LabelAppeal in ("1")) * 0.9213
+ (LabelAppeal in ("2")) * 1.0785;
P_SCORE_ZIP_ALL = exp(TEMP);
TEMP = -5.1960
+ AcidIndex * 0.4322
+ (stars0 in ("1")) * -2.0682
+ (stars0 in ("2")) * -5.8044
+ (stars0 in ("3")) * -24.9701
+ (stars0 in ("4")) * -25.1438
+ (LabelAppeal in ("-1")) * 1.4799
+ (LabelAppeal in ("0")) * 2.2270
+ (LabelAppeal in ("1")) * 2.9280
+ (LabelAppeal in ("2")) * 3.3765;
P_SCORE_ZERO = exp(TEMP)/(1+exp(TEMP));
P_TARGET = P_SCORE_ZIP_ALL * (1-P_SCORE_ZERO);
```

In order to use this model, please open the SAS program. Then, place “wine.sas7bdat” and “wine_test.sas7bdat” in the temporary working directory that is created after SAS is opened. Next, open the following SAS script file (“Joshua Peng Deploy Model.sas”) and run the script. The model predicted values computed with PROC GENMOD are stored in an output variable called “m6”.

The following is my SAS data step code for model deployment which creates a scored data file. After loading in the training and test data sets, the second portion of the code runs through all of the data transformation steps including imputing missing values, adding transformed variables, adding dummy variables, and adding indicator variables. Next, a data step uses the best model on the wine_test holdout test data set to generate predicted values listed under the variable “P_TARGET”. The same code below is in a data file entitled “Joshua Peng Deploy Model.sas.”

“Joshua Peng Deploy Model.sas”

```
* Loading in data;
data test; set wine_test;
* Imputing missing observations with mean value and adding new variables in test set;
data test0; set test;
    if missing(Alcohol) then alcohol = 10.4892363;
    if missing(FreeSulfurDioxide) then FreeSulfurDioxide = 30.8455713;
    stars0 = stars;
    if missing(stars) then stars0 = 0;
    if missing(TotalSulfurDioxide) then TotalSulfurDioxide = 120.7142326;
        av_Alcohol = abs(Alcohol);
        av_VolatileAcidity = abs(VolatileAcidity);
        av_FreeSulfurDioxide = abs(FreeSulfurDioxide);
        av_TotalSulfurDioxide = abs(TotalSulfurDioxide);
```

```
        if av_VolatileAcidity = 0 then ln_av_VolatileAcidity = 0;
                else ln_av_VolatileAcidity = log(av_VolatileAcidity);
        if av_FreeSulfurDioxide = 0 then ln_av_FreeSulfurDioxide = 0;
                else ln_av_FreeSulfurDioxide = log(av_FreeSulfurDioxide);
        if av_TotalSulfurDioxide = 0 then ln_av_TotalSulfurDioxide = 0;
                else ln_av_TotalSulfurDioxide = log(av_TotalSulfurDioxide);
run;
* Score test data with SAS data step;
data testscore; set test0;
TEMP = 0.4532
+ AcidIndex * -0.0632
+ ln_av_VolatileAcidity * -0.0135
+ ln_av_TotalSulfurDioxide * 0.0028
+ ln_av_FreeSulfurDioxide * 0.0059
+ av_Alcohol * 0.0071
+ (stars0 in (1)) * 0.0632
+ (stars0 in (2)) * 0.1834
+ (stars0 in (3)) * 0.2816
+ (stars0 in (4)) * 0.3809
+ (LabelAppeal in (-1)) * 0.4432
+ (LabelAppeal in (0)) * 0.7311
+ (LabelAppeal in (1)) * 0.9213
+ (LabelAppeal in (2)) * 1.0785;
P_SCORE_ZIP_ALL = exp(TEMP);
TEMP = -5.1960
+ AcidIndex * 0.4322
+ (stars0 in (1)) * -2.0682
+ (stars0 in (2)) * -5.8044
+ (stars0 in (3)) * -24.9701
+ (stars0 in (4)) * -25.1438
+ (LabelAppeal in (-1)) * 1.4799
+ (LabelAppeal in (0)) * 2.2270
+ (LabelAppeal in (1)) * 2.9280
+ (LabelAppeal in (2)) * 3.3765;
P_SCORE_ZERO = exp(TEMP)/(1+exp(TEMP));
P_TARGET = P_SCORE_ZIP_ALL * (1-P_SCORE_ZERO);
keep INDEX P_TARGET;
run;
```

If you want to generate the predicted values from models 2 (POI), 4 (NB), 6 (ZIP), 8 (ZINB), and 10 (REG) and have them merged in the same output file then you can run “Joshua Peng Deploy Model Merged.sas.”

```
* Loading in data;
data test; set wine_test;
* Imputing missing observations with mean value and adding new variables in test set;
data test0; set test;
    if missing(Alcohol) then alcohol = 10.4892363;
    if missing(FreeSulfurDioxide) then FreeSulfurDioxide = 30.8455713;
    stars0 = stars;
    if missing(stars) then stars0 = 0;
    if missing(TotalSulfurDioxide) then TotalSulfurDioxide = 120.7142326;
        av_Alcohol = abs(Alcohol);
        av_VolatileAcidity = abs(VolatileAcidity);
        av_FreeSulfurDioxide = abs(FreeSulfurDioxide);
        av_TotalSulfurDioxide = abs(TotalSulfurDioxide);
        if av_VolatileAcidity = 0 then ln_av_VolatileAcidity = 0;
                else ln_av_VolatileAcidity = log(av_VolatileAcidity);
        if av_FreeSulfurDioxide = 0 then ln_av_FreeSulfurDioxide = 0;
                else ln_av_FreeSulfurDioxide = log(av_FreeSulfurDioxide);
        if av_TotalSulfurDioxide = 0 then ln_av_TotalSulfurDioxide = 0;
                else ln_av_TotalSulfurDioxide = log(av_TotalSulfurDioxide);
        * Variable of reference: 0;
        if STARS0 in (0 1 2 3 4) then do;
                STARS0_1 = (STARS0 eq 1);
                STARS0_2 = (STARS0 eq 2);
                STARS0_3 = (STARS0 eq 3);
                STARS0_4 = (STARS0 eq 4);
        end;
        * Variable of reference: -2;
        if LabelAppeal in (-2 -1 0 1 2) then do;
                LabelAppeal_n1 = (LabelAppeal eq -1);
                LabelAppeal_0 = (LabelAppeal eq 0);
                LabelAppeal_p1 = (LabelAppeal eq 1);
                LabelAppeal_p2 = (LabelAppeal eq 2);
        end;
run;
* Score test data (POI Model 2) with SAS data step;
data testscore_poi; set test0;
```

```
TEMP = 0.1413
+ AcidIndex * -0.0778
+ ln_av_VolatileAcidity * -0.0262
+ ln_av_TotalSulfurDioxide * 0.0283
+ ln_av_FreeSulfurDioxide * 0.0156
+ av_Alcohol * 0.0043
+ (stars0 in (1)) * 0.7676
+ (stars0 in (2)) * 1.0838
+ (stars0 in (3)) * 1.2051
+ (stars0 in (4)) * 1.3272
+ (LabelAppeal in (-1)) * 0.2381
+ (LabelAppeal in (0)) * 0.4274
+ (LabelAppeal in (1)) * 0.5602
+ (LabelAppeal in (2)) * 0.6962;
P_TARGET_POI = exp(TEMP);
keep INDEX P_TARGET_POI;
run;
* Score test data (NB Model 4) with SAS data step;
data testscore_nb; set test0;
TEMP = 0.1413
+ AcidIndex * -0.0778
+ ln_av_VolatileAcidity * -0.0262
+ ln_av_TotalSulfurDioxide * 0.0283
+ ln_av_FreeSulfurDioxide * 0.0156
+ av_Alcohol * 0.0043
+ (stars0 in (1)) * 0.7676
+ (stars0 in (2)) * 1.0838
+ (stars0 in (3)) * 1.2051
+ (stars0 in (4)) * 1.3272
+ (LabelAppeal in (-1)) * 0.2381
+ (LabelAppeal in (0)) * 0.4274
+ (LabelAppeal in (1)) * 0.5602
+ (LabelAppeal in (2)) * 0.6962;
P_TARGET_NB = exp(TEMP);
keep INDEX P_TARGET_NB;
run;
* Score test data (ZIP Model 6) with SAS data step;
data testscore_zip; set test0;
TEMP = 0.4532
+ AcidIndex * -0.0632
+ ln_av_VolatileAcidity * -0.0135
+ ln_av_TotalSulfurDioxide * 0.0028
+ ln_av_FreeSulfurDioxide * 0.0059
+ av_Alcohol * 0.0071
+ (stars0 in (1)) * 0.0632
+ (stars0 in (2)) * 0.1834
+ (stars0 in (3)) * 0.2816
+ (stars0 in (4)) * 0.3809
+ (LabelAppeal in (-1)) * 0.4432
+ (LabelAppeal in (0)) * 0.7311
+ (LabelAppeal in (1)) * 0.9213
+ (LabelAppeal in (2)) * 1.0785;
P_SCORE_ZIP_ALL = exp(TEMP);
TEMP = -5.1960
+ AcidIndex * 0.4322
+ (stars0 in (1)) * -2.0682
+ (stars0 in (2)) * -5.8044
+ (stars0 in (3)) * -24.9701
+ (stars0 in (4)) * -25.1438
+ (LabelAppeal in (-1)) * 1.4799
+ (LabelAppeal in (0)) * 2.2270
+ (LabelAppeal in (1)) * 2.9280
+ (LabelAppeal in (2)) * 3.3765;
P_SCORE_ZERO = exp(TEMP)/(1+exp(TEMP));
P_TARGET_ZIP = P_SCORE_ZIP_ALL * (1-P_SCORE_ZERO);
keep INDEX P_TARGET_ZIP;
run;
* Score test data (ZINB Model 8) with SAS data step;
data testscore_zinb; set test0;
TEMP = 0.4532
+ AcidIndex * -0.0632
+ ln_av_VolatileAcidity * -0.0135
+ ln_av_TotalSulfurDioxide * 0.0028
+ ln_av_FreeSulfurDioxide * 0.0059
+ av_Alcohol * 0.0071
+ (stars0 in (1)) * 0.0632
+ (stars0 in (2)) * 0.1834
+ (stars0 in (3)) * 0.2816
+ (stars0 in (4)) * 0.3809
```

```
+ (LabelAppeal in (-1)) * 0.4432
+ (LabelAppeal in (0)) * 0.7311
+ (LabelAppeal in (1)) * 0.9213
+ (LabelAppeal in (2)) * 1.0785;
P_SCORE_ZINB_ALL = exp(TEMP);
TEMP = -5.1960
+ AcidIndex * 0.4322
+ (stars0 in (1)) * -2.0682
+ (stars0 in (2)) * -5.8044
+ (stars0 in (3)) * -24.9701
+ (stars0 in (4)) * -25.1438
+ (LabelAppeal in (-1)) * 1.4799
+ (LabelAppeal in (0)) * 2.2270
+ (LabelAppeal in (1)) * 2.9280
+ (LabelAppeal in (2)) * 3.3765;
P_SCORE_ZERO = exp(TEMP)/(1+exp(TEMP));
P_TARGET_ZINB = P_SCORE_ZINB_ALL * (1-P_SCORE_ZERO);
keep INDEX P_TARGET_ZINB;
run;
* Score test data (REG Model 10) with SAS data step;
data testscore_reg; set test0;
P_TARGET_REG = 1.23003
+ AcidIndex * -0.19397
+ ln_av_VolatileAcidity * -0.08283
+ ln_av_TotalSulfurDioxide * 0.07861
+ ln_av_FreeSulfurDioxide * 0.04412
+ av_Alcohol * 0.01434
+ stars0_1 * 1.36774
+ stars0_2 * 2.39443
+ stars0_3 * 2.96595
+ stars0_4 * 3.65759
+ LabelAppeal_n1 * 0.36814
+ LabelAppeal_0 * 0.83488
+ LabelAppeal_p1 * 1.29924
+ LabelAppeal_p2 * 1.88184;
keep INDEX P_TARGET_REG;
run;
* Score test data (ZIP with Cloglog link Model 11) with SAS data step;
data testscore_zip11; set test0;
TEMP = 0.4636
+ AcidIndex * -0.0209
+ ln_av_VolatileAcidity * -0.0135
+ ln_av_TotalSulfurDioxide * 0.0028
+ ln_av_FreeSulfurDioxide * 0.0060
+ av_Alcohol * 0.0071
+ (stars0 in (1)) * 0.0688
+ (stars0 in (2)) * 0.1853
+ (stars0 in (3)) * 0.2838
+ (stars0 in (4)) * 0.3833
+ (LabelAppeal in (-1)) * 0.4450
+ (LabelAppeal in (0)) * 0.7311
+ (LabelAppeal in (1)) * 0.9209
+ (LabelAppeal in (2)) * 1.0777;
P_SCORE_ZIP_ALL = exp(TEMP);
TEMP = -4.0081
+ AcidIndex * 0.2687
+ (stars0 in (1)) * -1.5432
+ (stars0 in (2)) * -5.1362
+ (stars0 in (3)) * -24.0085
+ (stars0 in (4)) * -25.1524
+ (LabelAppeal in (-1)) * 1.2427
+ (LabelAppeal in (0)) * 1.7738
+ (LabelAppeal in (1)) * 2.2531
+ (LabelAppeal in (2)) * 2.4212;
P_SCORE_ZERO = exp(TEMP)/(1+exp(TEMP));
P_TARGET_ZIP = P_SCORE_ZIP_ALL * (1-P_SCORE_ZERO);
keep INDEX P_TARGET_ZIP;
run;
* Score test data (ZIP with Probit link Model 12) with SAS data step;
data testscore_zip12; set test0;
TEMP = 0.4502
+ AcidIndex * -0.0189
+ ln_av_VolatileAcidity * -0.0135
+ ln_av_TotalSulfurDioxide * 0.0029
+ ln_av_FreeSulfurDioxide * 0.0060
+ av_Alcohol * 0.0071
+ (stars0 in (1)) * 0.0608
+ (stars0 in (2)) * 0.1819
+ (stars0 in (3)) * 0.2804
```

```
+ (stars0 in (4)) * 0.3797
+ (LabelAppeal in (-1)) * 0.4439
+ (LabelAppeal in (0)) * 0.7329
+ (LabelAppeal in (1)) * 0.9233
+ (LabelAppeal in (2)) * 1.0807;
P_SCORE_ZIP_ALL = exp(TEMP);
TEMP = -3.0205
+ AcidIndex * 0.2501
+ (stars0 in (1)) * -1.2351
+ (stars0 in (2)) * -3.0271
+ (stars0 in (3)) * -5.8146
+ (stars0 in (4)) * -5.5576
+ (LabelAppeal in (-1)) * 0.8673
+ (LabelAppeal in (0)) * 1.3150
+ (LabelAppeal in (1)) * 1.7290
+ (LabelAppeal in (2)) * 1.9903;
P_SCORE_ZERO = exp(TEMP)/(1+exp(TEMP));
P_TARGET_ZIP = P_SCORE_ZIP_ALL * (1-P_SCORE_ZERO);
keep INDEX P_TARGET_ZIP;
run;
*///merge all of the model results together///;
data testscore;
merge testscore_poi(in=ina) testscore_nb(in=inb)
   testscore_zip testscore_zinb testscore_reg;
by INDEX;
if ina;
run;
*///final datastep to retain index and model results///;
data testscore;
set testscore;
keep index p_target_poi p_target_nb p_target_zip p_target_zinb p_target_reg;
run;
```

6. Bonus

Model 11: GENMOD with Zero Inflated Poisson distribution (zeromodel using complementary log-log link) and STARS0 and LabelAppeal as categorical variables

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero (STARS0 at 0 and LabelAppeal at -2), the logarithm of expected number of wine cases purchased would be 0.4636.
- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:
    - 1 rating: 0.0688 increase in the logarithm of expected number of cases purchased
    - 2 rating: 0.1853 increase in the logarithm of expected number of cases purchased
    - 3 rating: 0.2838 increase in the logarithm of expected number of cases purchased
    - 4 rating: 0.3833 increase in the logarithm of expected number of cases purchased
- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
    - -1 rating: 0.4450 increase in the logarithm of expected number of cases purchased
    - 0 rating: 0.7311 increase in the logarithm of expected number of cases purchased
    - +1 rating: 0.9209 increase in the logarithm of expected number of cases purchased
    - +2 rating: 1.0777 increase in the logarithm of expected number of cases purchased
- If a wine increased its AcidIndex score by 1 point, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0209.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0135.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0028.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0060.
- If a wine increased it absolute value transformed alcohol content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0071.

For the zero inflated parameter estimates, assuming that all other variables are held constant:

- If all of the predictor variables in the model are evaluated at zero, the complementary log-log of the predicted number of wine cases purchased being zero are -4.0081.
- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:

- 1 rating: the probability that the expected number of cases purchased is zero decreases by 78.63%
- 2 rating: the probability that the expected number of cases purchased is zero decreases by 99.41%
- 3 rating: the probability that the expected number of cases purchased is zero decreases by 100%
- 4 rating: the probability that the expected number of cases purchased is zero decreases by 100%

- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
  - -1 rating: the probability that the expected number of cases purchased is zero increases by 246.50%
  - 0 rating: the probability that the expected number of cases purchased is zero increases by 489.32%
  - +1 rating: the probability that the expected number of cases purchased is zero increases by 851.72%
  - +2 rating: the probability that the expected number of cases purchased is zero increases by 1025.94%
- The probability that the expected number wine cases purchased is zero per each point increase in AcidIndex score increases by 30.83%

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.4636	0.0709	0.3246	0.6025	42.76	<.0001
stars0	1	1	0.0688	0.0212	0.0272	0.1103	10.54	0.0012
stars0	2	1	0.1853	0.0198	0.1464	0.2241	87.42	<.0001
stars0	3	1	0.2838	0.0207	0.2432	0.3244	187.53	<.0001
stars0	4	1	0.3833	0.0256	0.3331	0.4336	223.41	<.0001
stars0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	0.4450	0.0418	0.3631	0.5270	113.26	<.0001
LabelAppeal	0	1	0.7311	0.0409	0.6509	0.8112	319.65	<.0001
LabelAppeal	1	1	0.9209	0.0416	0.8395	1.0023	491.17	<.0001
LabelAppeal	2	1	1.0777	0.0460	0.9874	1.1679	547.74	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	-0.0209	0.0048	-0.0304	-0.0114	18.59	<.0001
ln_av_VolatileAcidit		1	-0.0135	0.0059	-0.0250	-0.0020	5.30	0.0214
ln_av_TotalSulfurDio		1	0.0028	0.0062	-0.0093	0.0149	0.20	0.6522
ln_av_FreeSulfurDiox		1	0.0060	0.0048	-0.0035	0.0154	1.53	0.2165
av_Alcohol		1	0.0071	0.0015	0.0043	0.0100	23.58	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.0081	0.3280	-4.6511	-3.3652	149.30	<.0001
stars0	1	1	-1.5432	0.0574	-1.6557	-1.4306	722.31	<.0001
stars0	2	1	-5.1362	0.3746	-5.8705	-4.4019	187.96	<.0001
stars0	3	1	-24.0085	3405.542	-6698.75	6650.731	0.00	0.9944
stars0	4	1	-24.1524	6497.252	-12758.5	12710.23	0.00	0.9970
stars0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	1.2427	0.3103	0.6344	1.8509	16.03	<.0001
LabelAppeal	0	1	1.7738	0.3081	1.1699	2.3776	33.15	<.0001
LabelAppeal	1	1	2.2531	0.3098	1.6460	2.8602	52.91	<.0001
LabelAppeal	2	1	2.4212	0.3271	1.7801	3.0622	54.80	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	0.2687	0.0149	0.2396	0.2980	326.03	<.0001

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance		40828.1463	
Scaled Deviance		40828.1463	
Pearson Chi-Square	13E3	5939.5240	0.4651
Scaled Pearson X2	13E3	5939.5240	0.4651
Log Likelihood		11183.0982	
Full Log Likelihood		-20414.0731	
AIC (smaller is better)		40876.1463	
AICC (smaller is better)		40876.2402	
BIC (smaller is better)		41055.1097	

Variable	Class	Estimate	exp(β)	exp(β)-1
Intercept		-4.0081	0.018168	-0.98183
STARS0	1	-1.5432	0.213696	-0.7863
STARS0	2	-5.1362	0.00588	-0.99412
STARS0	3	-24.0085	3.74E-11	-1
STARS0	4	-24.1524	3.24E-11	-1
STARS0	0	0	1	0

LabelAppeal	-1	1.2427	3.464956	2.464956
LabelAppeal	0	1.7738	5.893205	4.893205
LabelAppeal	1	2.2531	9.517193	8.517193
LabelAppeal	2	2.4212	11.25936	10.25936
LabelAppeal	-2	0	1	0
AcidIndex		0.2687	1.308263	0.308263

Model 12: GENMOD with Zero Inflated Poisson distribution (zeromodel using probit link) and STARS0 and LabelAppeal as categorical variables

The following interpretations assume that all other variables are held constant.

- Assuming that all variables are zero (STARS0 at 0 and LabelAppeal at -2), the logarithm of the expected number of wine cases purchased would be 0.4502.
- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:
  - 1 rating: 0.0608 increase in the logarithm of expected number of cases purchased
  - 2 rating: 0.1819 increase in the logarithm of expected number of cases purchased
  - 3 rating: 0.2804 increase in the logarithm of expected number of cases purchased
  - 4 rating: 0.3797 increase in the logarithm of expected number of cases purchased
- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
  - -1 rating: 0.4439 increase in the logarithm of expected number of cases purchased
  - 0 rating: 0.7329 increase in the logarithm of expected number of cases purchased
  - +1 rating: 0.9233 increase in the logarithm of expected number of cases purchased
  - +2 rating: 1.0807 increase in the logarithm of expected number of cases purchased
- If a wine increased its AcidIndex score by 1 point, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0189.
- If a wine increased its natural logarithm, absolute value transformed Volatile Acidity content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to decrease by 0.0135.
- If a wine increased its natural logarithm, absolute value transformed Total Sulfur Dioxide content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0029.
- If a wine increased its natural logarithm, absolute value transformed Free Sulfur Dioxide rating by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0060.
- If a wine increased it absolute value transformed alcohol content by 1 unit, the logarithm of expected number of wine cases purchased would be expected to increase by 0.0071.

For the zero inflated parameter estimates, assuming that all other variables are held constant:

- If all of the predictor variables in the model are evaluated at zero, the predicted probability that the number of wine cases purchased would be zero is F(-5.1960) = 1.01811E-07, where F is the cumulative distribution function of the standard normal.

However, interpretation of the coefficients in probit regression is not as straightforward as the interpretations of coefficients in linear regression or logit regression. The increase in probability attributed to a one-unit increase in a given predictor is dependent both on the values of the other predictors and the starting value of the given predictors. The probabilities do not change by a common difference or common factor.

- Given that STARS0 has a base level of 0 (lowest rating), we interpret obtaining a:
  - 1 rating: the predicted probability that the expected number of wine cases purchased will be zero decrease by a factor of 0.108397.
  - 2 rating: the predicted probability that the expected number of wine cases purchased will be zero decrease by a factor of 0.001235.
  - 3 rating: the predicted probability that the expected number of wine cases purchased will be zero decrease by a factor of 3.04e-09.
  - 4 rating: the predicted probability that the expected number of wine cases purchased will be zero decrease by a factor of 1.37e-08.
- Given that LabelAppeal has a base level of -2 (lowest rating), we interpret obtaining a:
  - -1 rating: the predicted probability that the expected number of wine cases purchased will be zero decrease by a factor of 0.807111.
  - 0 rating: the predicted probability that the expected number of wine cases purchased will be zero decrease by a factor of 0.905745.

    - +1 rating: the predicted probability that the expected number of wine cases purchased will be zero decrease by a factor of 0.958095.
    - +2 rating: the predicted probability that the expected number of wine cases purchased will be zero decrease by a factor of 0.976721.
- Increasing the AcidIndex score by 1 point decreases the predicted probability that the expected number of wine cases purchased will be zero by a factor 0.598745.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.4502	0.0705	0.3120	0.5884	40.78	<.0001
stars0	1	1	0.0606	0.0211	0.0194	0.1023	8.28	0.0040
stars0	2	1	0.1819	0.0197	0.1433	0.2206	85.09	<.0001
stars0	3	1	0.2804	0.0207	0.2399	0.3209	184.05	<.0001
stars0	4	1	0.3797	0.0256	0.3295	0.4298	219.91	<.0001
stars0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	0.4439	0.0412	0.3632	0.5247	116.10	<.0001
LabelAppeal	0	1	0.7329	0.0403	0.6540	0.8118	331.35	<.0001
LabelAppeal	1	1	0.9233	0.0409	0.8431	1.0036	508.73	<.0001
LabelAppeal	2	1	1.0807	0.0455	0.9916	1.1698	565.06	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	-0.0189	0.0048	-0.0284	-0.0094	15.29	<.0001
ln_av_VolatileAcidit		1	-0.0135	0.0059	-0.0250	-0.0020	5.28	0.0216
ln_av_TotalSulfurDio		1	0.0029	0.0062	-0.0092	0.0149	0.22	0.6415
ln_av_FreeSulfurDiox		1	0.0060	0.0048	-0.0034	0.0154	1.55	0.2134
av_Alcohol		1	0.0071	0.0015	0.0042	0.0100	23.33	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.0205	0.2133	-3.4386	-2.6024	200.53	<.0001
stars0	1	1	-1.2351	0.0430	-1.3195	-1.1507	823.40	<.0001
stars0	2	1	-3.0271	0.1401	-3.3016	-2.7526	467.14	<.0001
stars0	3	1	-5.8146	4.2613	-14.1667	2.5375	1.86	0.1724
stars0	4	1	-5.5576	7.5296	-20.3152	9.2000	0.54	0.4604
stars0	0	0	0.0000	0.0000	0.0000	0.0000	.	.
LabelAppeal	-1	1	0.8673	0.1831	0.5084	1.2261	22.43	<.0001
LabelAppeal	0	1	1.3150	0.1812	0.9598	1.6702	52.65	<.0001
LabelAppeal	1	1	1.7290	0.1842	1.3679	2.0901	88.07	<.0001
LabelAppeal	2	1	1.9903	0.2116	1.5767	2.4048	88.56	<.0001
LabelAppeal	-2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex		1	0.2501	0.0142	0.2222	0.2780	308.59	<.0001

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance		40789.7561	
Scaled Deviance		40789.7561	
Pearson Chi-Square	13E3	5736.0525	0.4491
Scaled Pearson X2	13E3	5736.0525	0.4491
Log Likelihood		11202.2932	
Full Log Likelihood		-20394.8780	
AIC (smaller is better)		40837.7561	
AICC (smaller is better)		40837.8500	
BIC (smaller is better)		41016.7195	

Variable	Class	Estimate	F(β)
Intercept		-3.0205	0.001262
STARS0	1	-1.2351	0.108397
STARS0	2	-3.0271	0.001235
STARS0	3	-5.8146	3.04E-09
STARS0	4	-5.5576	1.37E-08
STARS0	0	0	0.5
LabelAppeal	-1	0.8673	0.807111
LabelAppeal	0	1.315	0.905745
LabelAppeal	1	1.729	0.958095
LabelAppeal	2	1.9903	0.976721
LabelAppeal	-2	0	0.5

AcidIndex		0.2501	0.598745

It appears that model 12 actually outperforms model 6 in all metrics (lower deviance, lower log likelihood, lower AIC, lower AICC, and lower BIC). All 3 models depicted here perform very well in comparison to other 9 models. Changing the zeromodel link function does not drastically affect the model performance but in this case, it helped to generate a model that was slightly better than model 6 which I determined to be the best.

Model	Description	Deviance	Pearson Chi Square	Log Likelihood	AIC	AICC	BIC
Model 6	Zero Inflated Poisson with 2 categorical variables, zeromodel with default **logit link**	40796.8372	5743.6266	11198.7527	40844.8372	40844.9312	41023.8007
Model 11	Zero Inflated Poisson with 2 categorical variables, zeromodel with **complementary log-log link**	40828.1463	5939.5240	11183.0982	40876.1463	40876.2402	41055.1097
Model 12	Zero Inflated Poisson with 2 categorical variables, zeromodel with **probit link**	40789.7561	5736.0525	11202.2932	40837.7561	40837.8500	41016.7195

For the first 30 observations, it appears that model 11 performs the best. The Sum of Absolute Error and Sum of Squared Error are relatively similar for all 3 models. In the future, when I am working for Zero Inflated Poisson distributions I will consider changing the zeromodel link function as it may improve performance and prediction accuracy.

	Actual Values	Predicted Values		
Obs	**TARGET**	**m6**	**m11**	**m12**
1	3	3.72045	3.71386	3.72621
2	3	3.23896	3.24816	3.23529
3	5	3.36562	3.37015	3.36205
4	3	2.44195	2.39501	2.4609
5	4	3.83237	3.82343	3.83616
6	0	0.44358	0.39109	0.45738
7	0	1.2233	1.27962	1.23687
8	4	5.24074	5.24389	5.24232
9	3	2.04596	2.01865	2.00742
10	6	4.76294	4.76026	4.76286
11	0	3.99584	3.8912	3.9944
12	4	3.64193	3.63345	3.63804
13	3	4.41992	4.41853	4.4249
14	7	6.00023	5.99072	6.00276
15	4	0.89109	0.93571	0.91726
16	0	1.18559	1.24001	1.19886
17	0	0.58682	0.50752	0.60119
18	4	4.24621	4.24285	4.24695
19	5	4.28815	4.32823	4.18684
20	4	3.02866	3.02902	3.02997
21	3	2.53758	2.48926	2.55754
22	2	2.41183	2.36239	2.42842
23	3	2.74526	2.72721	2.73293
24	4	0.87861	0.85055	0.90173
25	4	3.92394	3.9226	3.92967
26	0	1.64863	1.63868	1.62426
27	4	4.50505	4.50401	4.50965
28	6	5.67084	5.67462	5.66993
29	4	4.04721	4.03875	4.05106
30	3	2.41082	2.36842	2.42196
Sum of Absolute Error		29.44818	29.40327	29.55974
Sum of Squared Error		54.92481	54.24441	54.85597

I've posted the SAS Data Step below for these bonus models. The resulting sas7bdat files are entitled "JoshuaPeng_testscore_zip11.sas7bdat" for the ZIP model with Cloglog link and "JoshuaPeng_testscore_zip12.sas7bdat" for the ZINB model with Probit link.

```
* Score test data (ZIP with Cloglog link Model 11) with SAS data step;
data testscore_zip11; set test0;
TEMP = 0.4636
+ AcidIndex * -0.0209
+ ln_av_VolatileAcidity * -0.0135
+ ln_av_TotalSulfurDioxide * 0.0028
+ ln_av_FreeSulfurDioxide * 0.0060
+ av_Alcohol * 0.0071
+ (stars0 in (1)) * 0.0688
+ (stars0 in (2)) * 0.1853
+ (stars0 in (3)) * 0.2838
+ (stars0 in (4)) * 0.3833
+ (LabelAppeal in (-1)) * 0.4450
+ (LabelAppeal in (0)) * 0.7311
+ (LabelAppeal in (1)) * 0.9209
+ (LabelAppeal in (2)) * 1.0777;
P_SCORE_ZIP_ALL = exp(TEMP);
TEMP = -4.0081
+ AcidIndex * 0.2687
+ (stars0 in (1)) * -1.5432
+ (stars0 in (2)) * -5.1362
+ (stars0 in (3)) * -24.0085
+ (stars0 in (4)) * -25.1524
+ (LabelAppeal in (-1)) * 1.2427
+ (LabelAppeal in (0)) * 1.7738
+ (LabelAppeal in (1)) * 2.2531
+ (LabelAppeal in (2)) * 2.4212;
P_SCORE_ZERO = exp(TEMP)/(1+exp(TEMP));
P_TARGET_ZIP = P_SCORE_ZIP_ALL * (1-P_SCORE_ZERO);
keep INDEX P_TARGET_ZIP;
run;
* Score test data (ZIP with Probit link Model 12) with SAS data step;
data testscore_zip12; set test0;
TEMP = 0.4502
+ AcidIndex * -0.0189
+ ln_av_VolatileAcidity * -0.0135
+ ln_av_TotalSulfurDioxide * 0.0029
+ ln_av_FreeSulfurDioxide * 0.0060
+ av_Alcohol * 0.0071
+ (stars0 in (1)) * 0.0608
+ (stars0 in (2)) * 0.1819
+ (stars0 in (3)) * 0.2804
+ (stars0 in (4)) * 0.3797
+ (LabelAppeal in (-1)) * 0.4439
+ (LabelAppeal in (0)) * 0.7329
+ (LabelAppeal in (1)) * 0.9233
+ (LabelAppeal in (2)) * 1.0807;
P_SCORE_ZIP_ALL = exp(TEMP);
TEMP = -3.0205
+ AcidIndex * 0.2501
+ (stars0 in (1)) * -1.2351
+ (stars0 in (2)) * -3.0271
+ (stars0 in (3)) * -5.8146
+ (stars0 in (4)) * -5.5576
+ (LabelAppeal in (-1)) * 0.8673
+ (LabelAppeal in (0)) * 1.3150
+ (LabelAppeal in (1)) * 1.7290
+ (LabelAppeal in (2)) * 1.9903;
P_SCORE_ZERO = exp(TEMP)/(1+exp(TEMP));
P_TARGET_ZIP = P_SCORE_ZIP_ALL * (1-P_SCORE_ZERO);
keep INDEX P_TARGET_ZIP;
run;
```

Comparing PROC PLM vs. SAS Data Step for Zero Inflated Poisson/Negative Binomial models

The Zero Inflated Poisson (ZIP) model (and also the Zero Inflated Negative Binomial (ZINB) model) are composed of 2 model processes. Because of this important point, I cannot generate the same, correct predicted values using PROC PLM on the PROC GENMOD stored output file as with a SAS data step. I have to write a separate SAS data step to obtain the correct predicted target values. The following are the SAS data steps to generate my ZIP and ZINB models. The following is my SAS code to compare the PROC PLM and SAS data step predicted value output of my best ZIP and ZINB models.

```
* Loading in data;
data train; set wine;
data test; set wine_test;
* Imputing missing observations with mean value and adding new variables in training set;
data train0; set train;
    if missing(Alcohol) then alcohol = 10.4892363;
```

```
    if missing(FreeSulfurDioxide) then FreeSulfurDioxide = 30.8455713;
    stars0 = stars;
    if missing(stars) then stars0 = 0;
    if missing(TotalSulfurDioxide) then TotalSulfurDioxide = 120.7142326;
        av_Alcohol = abs(Alcohol);
        av_VolatileAcidity = abs(VolatileAcidity);
        av_FreeSulfurDioxide = abs(FreeSulfurDioxide);
        av_TotalSulfurDioxide = abs(TotalSulfurDioxide);
        if av_VolatileAcidity = 0 then ln_av_VolatileAcidity = 0;
                else ln_av_VolatileAcidity = log(av_VolatileAcidity);
        if av_FreeSulfurDioxide = 0 then ln_av_FreeSulfurDioxide = 0;
                else ln_av_FreeSulfurDioxide = log(av_FreeSulfurDioxide);
        if av_TotalSulfurDioxide = 0 then ln_av_TotalSulfurDioxide = 0;
                else ln_av_TotalSulfurDioxide = log(av_TotalSulfurDioxide);
        * Variable of reference: 0;
        if STARS0 in (0 1 2 3 4) then do;
                STARS0_1 = (STARS0 eq 1);
                STARS0_2 = (STARS0 eq 2);
                STARS0_3 = (STARS0 eq 3);
                STARS0_4 = (STARS0 eq 4);
        end;
        * Variable of reference: -2;
        if LabelAppeal in (-2 -1 0 1 2) then do;
                LabelAppeal_n1 = (LabelAppeal eq -1);
                LabelAppeal_0 = (LabelAppeal eq 0);
                LabelAppeal_p1 = (LabelAppeal eq 1);
                LabelAppeal_p2 = (LabelAppeal eq 2);
        end;
run;
* Imputing missing observations with mean value and adding new variables in test set;
data test0; set test;
    if missing(Alcohol) then alcohol = 10.4892363;
    if missing(FreeSulfurDioxide) then FreeSulfurDioxide = 30.8455713;
    stars0 = stars;
    if missing(stars) then stars0 = 0;
    if missing(TotalSulfurDioxide) then TotalSulfurDioxide = 120.7142326;
        av_Alcohol = abs(Alcohol);
        av_VolatileAcidity = abs(VolatileAcidity);
        av_FreeSulfurDioxide = abs(FreeSulfurDioxide);
        av_TotalSulfurDioxide = abs(TotalSulfurDioxide);
        if av_VolatileAcidity = 0 then ln_av_VolatileAcidity = 0;
                else ln_av_VolatileAcidity = log(av_VolatileAcidity);
        if av_FreeSulfurDioxide = 0 then ln_av_FreeSulfurDioxide = 0;
                else ln_av_FreeSulfurDioxide = log(av_FreeSulfurDioxide);
        if av_TotalSulfurDioxide = 0 then ln_av_TotalSulfurDioxide = 0;
                else ln_av_TotalSulfurDioxide = log(av_TotalSulfurDioxide);
        * Variable of reference: 0;
        if STARS0 in (0 1 2 3 4) then do;
                STARS0_1 = (STARS0 eq 1);
                STARS0_2 = (STARS0 eq 2);
                STARS0_3 = (STARS0 eq 3);
                STARS0_4 = (STARS0 eq 4);
        end;
        * Variable of reference: -2;
        if LabelAppeal in (-2 -1 0 1 2) then do;
                LabelAppeal_n1 = (LabelAppeal eq -1);
                LabelAppeal_0 = (LabelAppeal eq 0);
                LabelAppeal_p1 = (LabelAppeal eq 1);
                LabelAppeal_p2 = (LabelAppeal eq 2);
        end;
run;
* Generating GENMOD ZIP model and storing output file as m6;
proc genmod data=train0;
class stars0 (ref="0") labelappeal (ref="-2");
model target = stars0 labelappeal acidindex ln_av_volatileacidity ln_av_totalsulfurdioxide
                          ln_av_freesulfurdioxide av_alcohol / link=log dist=zip;
zeromodel stars0 labelappeal acidindex / link=logit;
store out=m6;
run;
* Scoring ZIP test data with PROC PLM;
proc plm source=m6;
        score data=test0 out=testscore_zip0 pred=p_1 / ilink;
run;
* Generating GENMOD ZINB model and storing output file as m8;
proc genmod data=train0;
class stars0 (ref="0") labelappeal (ref="-2");
model target = stars0 labelappeal acidindex ln_av_volatileacidity ln_av_totalsulfurdioxide
                          ln_av_freesulfurdioxide av_alcohol / link=log dist=zinb;
zeromodel stars0 labelappeal acidindex / link=logit;
```

```
store out=m8;
run;
* Scoring ZINB test data with PROC PLM;
proc plm source=m8;
        score data=test0 out=testscore_zinb0 pred=p_1 / ilink;
run;
* Keeping only INDEX and P_TARGET_ZIP;
data testscore_zip0; set testscore_zip0;
P_TARGET_ZIP = p_1;
keep INDEX P_TARGET_ZIP;
run;
* Keeping only INDEX and P_TARGET_ZINB;
data testscore_zinb0; set testscore_zinb0;
P_TARGET_ZINB = p_1;
keep INDEX P_TARGET_ZINB;
run;
* Score test data (ZIP Model 6) with SAS data step;
data testscore_zip; set test0;
TEMP = 0.4532
+ AcidIndex * -0.0632
+ ln_av_VolatileAcidity * -0.0135
+ ln_av_TotalSulfurDioxide * 0.0028
+ ln_av_FreeSulfurDioxide * 0.0059
+ av_Alcohol * 0.0071
+ (stars0 in (1)) * 0.0632
+ (stars0 in (2)) * 0.1834
+ (stars0 in (3)) * 0.2816
+ (stars0 in (4)) * 0.3809
+ (LabelAppeal in (-1)) * 0.4432
+ (LabelAppeal in (0)) * 0.7311
+ (LabelAppeal in (1)) * 0.9213
+ (LabelAppeal in (2)) * 1.0785;
P_SCORE_ZIP_ALL = exp(TEMP);
TEMP = -5.1960
+ AcidIndex * 0.4322
+ (stars0 in (1)) * -2.0682
+ (stars0 in (2)) * -5.8044
+ (stars0 in (3)) * -24.9701
+ (stars0 in (4)) * -25.1438
+ (LabelAppeal in (-1)) * 1.4799
+ (LabelAppeal in (0)) * 2.2270
+ (LabelAppeal in (1)) * 2.9280
+ (LabelAppeal in (2)) * 3.3765;
P_SCORE_ZERO = exp(TEMP)/(1+exp(TEMP));
P_TARGET_ZIP = P_SCORE_ZIP_ALL * (1-P_SCORE_ZERO);
keep INDEX P_TARGET_ZIP;
run;
* Score test data (ZINB Model 8) with SAS data step;
data testscore_zinb; set test0;
TEMP = 0.4532
+ AcidIndex * -0.0632
+ ln_av_VolatileAcidity * -0.0135
+ ln_av_TotalSulfurDioxide * 0.0028
+ ln_av_FreeSulfurDioxide * 0.0059
+ av_Alcohol * 0.0071
+ (stars0 in (1)) * 0.0632
+ (stars0 in (2)) * 0.1834
+ (stars0 in (3)) * 0.2816
+ (stars0 in (4)) * 0.3809
+ (LabelAppeal in (-1)) * 0.4432
+ (LabelAppeal in (0)) * 0.7311
+ (LabelAppeal in (1)) * 0.9213
+ (LabelAppeal in (2)) * 1.0785;
P_SCORE_ZINB_ALL = exp(TEMP);
TEMP = -5.1960
+ AcidIndex * 0.4322
+ (stars0 in (1)) * -2.0682
+ (stars0 in (2)) * -5.8044
+ (stars0 in (3)) * -24.9701
+ (stars0 in (4)) * -25.1438
+ (LabelAppeal in (-1)) * 1.4799
+ (LabelAppeal in (0)) * 2.2270
+ (LabelAppeal in (1)) * 2.9280
+ (LabelAppeal in (2)) * 3.3765;
P_SCORE_ZERO = exp(TEMP)/(1+exp(TEMP));
P_TARGET_ZINB = P_SCORE_ZINB_ALL * (1-P_SCORE_ZERO);
keep INDEX P_TARGET_ZINB;
run;
```

As you can see, the first 37 observations are different when using a SAS data step vs. PROC PLM.

ZIP model predicted values using SAS data step			ZIP model predicted values using PROC PLM		
	INDEX	P_TARGET_ZIP		INDEX	P_TARGET_ZIP
1	3	1.8671618371	1	3	1.4335621597
2	9	4.0974513763	2	9	3.1457019923
3	10	2.6785762867	3	10	1.882843409
4	18	2.3908746169	4	18	1.6807551706
5	21	0.6073423681	5	21	0.3909170024
6	30	5.88883523	6	30	4.1396367473
7	31	3.7687963839	7	31	2.2209590475
8	37	1.2483716791	8	37	0.9170655547
9	39	0.2687805708	9	39	0.158408281
10	47	1.5867536185	10	47	1.1657420535
11	60	2.7941281297	11	60	1.9641537448
12	62	0.4417680054	12	62	0.2843573608
13	63	3.8690836823	13	63	2.6027799258
14	64	1.1696120034	14	64	0.8222056506
15	68	1.094967202	15	68	0.7365120717
16	75	2.7506903387	16	75	1.6943016886
17	76	2.3422062011	17	76	1.646553982
18	83	0.1111897831	18	83	0.0627077959
19	87	3.8347095492	19	87	2.8171329966
20	92	5.3438684028	20	92	4.287118152
21	98	1.8993882643	21	98	1.335277219
22	106	1.5250290882	22	106	1.120256357
23	107	0.7275459659	23	107	0.5114936562
24	113	2.3676649086	24	113	1.7392963986
25	120	3.7787497958	25	120	2.7761967379
26	123	5.407681659	26	123	3.9730006892
27	125	2.8474530801	27	125	1.9155344723
28	126	5.9281825333	28	126	4.5508411868
29	128	4.666658614	29	128	3.2804583701
30	129	2.4065842991	30	129	1.7678334952
31	131	4.3112601555	31	131	3.0305147311
32	135	0.8689210593	32	135	0.6108317168
33	141	4.4131362468	33	141	3.2420819775
34	147	3.1793472434	34	147	2.3354538233
35	148	1.0753289308	35	148	0.7232954264
36	151	3.8820306442	36	151	2.8521482477
37	156	3.247546611	37	156	2.3858098906

ZINB model predicted values using SAS data step			ZINB model predicted values using PROC PLM		
	INDEX	P_TARGET_ZINB		INDEX	P_TARGET_ZINB
1	3	1.4335621597	1	3	1.8671618362
2	9	3.1457019923	2	9	4.0974513783
3	10	1.882843409	3	10	2.678576289
4	18	1.6807551706	4	18	2.3908746178
5	21	0.3909170024	5	21	0.6073423726
6	30	4.1396367473	6	30	5.8888349641
7	31	2.2209590475	7	31	3.7687958545
8	37	0.9170655547	8	37	1.2483716856
9	39	0.158408281	9	39	0.2687805747
10	47	1.1657420535	10	47	1.5867536199
11	60	1.9641537448	11	60	2.7941281276
12	62	0.2843573608	12	62	0.4417680117
13	63	2.6027799258	13	63	3.8690836761
14	64	0.8222056506	14	64	1.1696120057
15	68	0.7365120717	15	68	1.0949671991
16	75	1.6943016886	16	75	2.7506903422
17	76	1.646553982	17	76	2.3422061998
18	83	0.0627077959	18	83	0.1111897861
19	87	2.8171329966	19	87	3.8347095514
20	92	4.287118152	20	92	5.3438683184
21	98	1.335277219	21	98	1.899388264
22	106	1.120256357	22	106	1.5250290926
23	107	0.5114936562	23	107	0.7275459731
24	113	1.7392963986	24	113	2.3676649127
25	120	2.7761967379	25	120	3.778749792
26	123	3.9730006892	26	123	5.407681655
27	125	1.9155344723	27	125	2.8474530818
28	126	4.5508411868	28	126	5.9281824159
29	128	3.2804583701	29	128	4.6666585087
30	129	1.7678334952	30	129	2.4065843012
31	131	3.0305147311	31	131	4.3112600399
32	135	0.6108317168	32	135	0.8689210662
33	141	3.2420819775	33	141	4.4131361697
34	147	2.3354538233	34	147	3.1793472423
35	148	0.7232954264	35	148	1.0753289281
36	151	2.8521482477	36	151	3.8820306468
37	156	2.3858098906	37	156	3.2475466091

The full SAS code to compare PROC PLM and the SAS Data Step along with the SAS Data Steps for generating predicted values for the ZIP models with Cloglog link and Probit link are fully included in “Joshua Peng Deploy Model Bonus.sas.”

7. Conclusion

The purpose of this assignment was to develop a model to predict the number of cases of wine that will be sold given certain properties of the wine. The wine training data set contained 12,795 observations and 14 variables. Two of the variables were subjective variables which I utilized as both quantitative and categorical variables during the modeling process. There were 12 continuous variables related to the chemical properties of the wine being sold. There were 2 numerical variables for the marketing score based on the visual appeal of the label and wine rating based on number of stars. The target variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. The purpose of this project was to build a model to predict the number of cases of wine that will be sold given certain properties of the wine. I first examined all of the variables and found that many had negative values which should not be possible since these variables measure the physical amount or level of a substance (count variables). For variables with negative values, I added reshifted (adding the absolute value of the minimum negative value) variables and the absolute value transformed counterparts. I also computed BoundSulfurDioxide from TotalSulfurDioxide and FreeSulfurDioxide and TotalAcidity from VolatileAcidity and FixedAcidity. For variables with missing observations, I imputed missing observations with the mean value. I added STARS0 which was the same as STARS but had the missing observations in its own class equal to zero. I generated the correlation table with TARGET and ran automatic variable selection methods with PROC HPGENSELECT in order to narrow down my set of variables to the 7 best variables. I built several Poisson and Negative Binomial distribution models with and without the zero inflation model to predict the target number of cases ordered for each wine. I also built a linear regression model to compare with all of the other models. I compared 12 models of 7 variables each and found that the best model was a Zero Inflated Poisson distribution model with the STARS0 and LabelAppeal variables used as categorical variables.

I believe this data set could be improved with a wine type variable. All of the physical characteristics such as density, sulfite content, acidity, chlorides, residual sugar, pH, and alcohol all vary with wine type, whether it is a red wine, white wine, rosé wine, dry white wine, dry red wine, sweet white wine, sweet red wine, sherry grape wines, etc. Even consumers have preferences for different wines, and their personal affinity for a certain type of wine will influence and affect their wine rating (STARS). It would be interesting to see the differences in physical characteristics between red and white wines and determine which type of wine is most preferred among the two major types. Other variables that would also be interesting to look at would be wine age, country of origin, and color density.

8. References

1. Sulphur in the winery. 2016; http://www.morethanorganic.com/sulphur-in-the-winery. Accessed February 17, 2016.
2. Carel M. Sulfur Dioxide (SO2) in wine | Wine From Here. 2011; http://winobrothers.com/2011/10/11/sulfur-dioxide-so2-in-wine/. Accessed February 17, 2016.
3. Grubbs S. Wine Jargon: What is Residual Sugar? | Serious Eats. 2013; http://drinks.seriouseats.com/2013/04/wine-jargon-what-is-residual-sugar-riesling-fermentation-steven-grubbs.html. Accessed February 17, 2016.
4. Pandell AJ. The Acidity of Wine. 2011; http://www.wineperspective.com/the_acidity_of_wine.htm. Accessed February 17, 2016.
5. Nierman D. Fixed Acidity — Waterhouse Lab. 2014; http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity. Accessed February 17, 2016.
6. Neeley E. Volatile Acidity — Waterhouse Lab. 2004; http://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity. Accessed February 17, 2016.
7. Leonardelli MJ. Acidity in Wine: The importance of management through measurement. 2013(February 17, 2016). http://gwi.missouri.edu/publications/2013spring.pdf.
8. Colby C. Calibrate your Winemaking Tools - WineMaker Magazine. 2008; https://winemakermag.com/127-calibrate-your-winemaking-tools. Accessed February 17, 2016.
9. Henrie J. Taking the Fear Out of Must Analysis - WineMaker Magazine. 1999; https://winemakermag.com/654-taking-the-fear-out-of-must-analysis. Accessed February 17, 2016.
10. Citric Acid | Viticulture & Enology. 2014; http://wineserver.ucdavis.edu/industry/enology/methods_and_techniques/reagents/citric_acid.html. Accessed February 17, 2016.
11. Coli MS, Rangel AGP, Souza ES, Oliveira MF, Chiaradia ACN. Chloride concentration in red wines: influence of terroir and grape type. *Food Science and Technology (Campinas).* 2015;35:95-99.
12. A Guide to the Alcohol Content in Wine - Real Simple. 2016; http://www.realsimple.com/holidays-entertaining/entertaining/food-drink/alcohol-content-wine. Accessed February 17, 2016.
13. Taylor R. The Truth About Grape Growing and Wine Making. 2013.