Introduction

The purpose of this assignment is to practice using Poisson regression models to predict count target variables. The dataset that we will be using consists of approximately 12,000 records of wine that are available for sale in the United States. Our goal is to predict the number of cases sold for each type of wine based on independent variables ranging from chemical properties of the wine to more qualitative factors such as label appeal and expert rating. We will be comparing the performance of a variety of different regression models – Poisson, Negative Binomial, Zero Inflated Poisson, Zero Inflated Negative Binomial, and OLS regression – to see if one model works better at predicting counts with our data. Given that I do not have a chemistry background and am only a novice wine-drinker, part of this assignment was also spent researching wine and the qualities provided in the dataset. This will be discussed further in the next section.

Domain Knowledge Research

Before diving into the data exploration phase of the assignment, I spent some time researching the wine properties available in our dataset. My goal was to develop some domain expertise that can be applied in the proceeding sections and ultimately improve model performance. My findings were rather useful as I proceeded to explore the data. For instance, I learned that alcohol in the United States ranges from 5% - 21%[1], pH ranges from 2.9 – 3.9[2], Total Sulfur Dioxide can be no more than 350 mg/L[3], and Volatile Acidity can be no more than 1.2 g/L (1.1 g/L for white wine, but since color is not available I decided to use the higher limit)[4]. While hard limits were not available for the other variables, I was able to find their typical ranges – Chlorides are usually less than 60 mg/L[5], Citric Acid ranges from 0-500 mg/L[6], Density is close to that of water (1Kg/L)[7], Fixed Acidity ranges from 0-

---

[1] Alcohol Content in Wine and Other Drinks (Infographic) | Wine Folly. (2013, December 23). Retrieved

[2] Acids in wine. (n.d.). Retrieved May 23, 2015, from http://en.wikipedia.org/wiki/Acids_in_wine

[3] Sulfites. (n.d.). Retrieved May 22, 2015, from http://waterhouse.ucdavis.edu/whats-in-wine/sulfites-in-wine

[4] Volatile Acidity. (n.d.). Retrieved May 23, 2015, from http://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity

[5] The level of sodium and chloride ions in wines. (n.d.). Retrieved May 22, 2015, from http://webcache.googleusercontent.com/search?q=cache:yEADr2OrCPQJ:www.oiv.int/oiv/files/6%20-%20Domaines%20scientifiques/6%20-%204%20Methodes%20d%20analyses/6-4-1/EN/OIV-MA-D1-03.pdf &cd=1&hl=en&ct=clnk

[6] Fixed Acidity. (n.d.). Retrieved May 22, 2015, from http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity

[7] Wine Density. (n.d.). Retrieved May 24, 2015, from http://www.answers.com/Q/What_is_the_density_of_wine

8000 mg/L[6], Free Sulfur Dioxide is usually 1/3 to 2/3 of Total Sulfur Dioxide[8], Sulphates can be no higher than 2-3 g[9], and anything with Residual Sugar greater than 45 g/L is considered sweet[10]. Rather than merely using the data's distributions to find and address outliers, I will also be leveraging this domain expertise.

Exploratory data analysis

For the first part of this assignment, we examine the wine data to learn about the variable distributions and see if there are any values in the data that may need to be addressed before the modeling process. By analyzing the PROC MEANS output seen in Figure 1, we have already learned a great deal about our dataset. Beginning with the target variable, one can see that the mean is less than the standard deviation. This situation is know as overdispersion and means that our data does not actually fit a Poisson distribution or Negative Binomial distribution since both assume that the mean is equal to the variance (equidispersion). As we have learned from Tukey (and our class slide deck), however, it is worth continuing since statistics are robust and may give good results even when underlying assumptions are violated.

Moving on the predictor variables, a variety of issues become apparent. First we see that eight of our variables – Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, pH, Sulphates, Alcohol, and STARS have missing values, so we will need to address imputation in the data preparation phase of the assignment. Looking at the minimum values, we can also see that many of these variables do not pass the "sniff test." For instance, we see negative values in Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Sulphates, Alcohol, and Label Appeal. Besides label appear (an ugly label could hurt sales), the rest of the fields should not have negative values and need to be adjusted. The minimum pH value of zero also looks suspicious, given that we know wine pH ranges from 2.9 to 3.9. Looking at the maximum values, we run into further outliers. At 34 g/L, the maximum Fixed Acidity seems much higher than what we'd expect. Volatile Acidity of 4 g/L is above the legal limit, so we know that there must be issues with the data quality. Citric Acid values of 4 g/L are much higer than the upper limit of 500 mg/L that we would expect. Residual Sugar above 45 g/L is considered sweet, so 141 g/L seems unusually high. Chlorides are usually less

---

[8] Sulfur Dioxide. (n.d.). Retrieved May 24, 2015, from http://www.santarosa.edu/~jhenderson/Sulfur Dioxide.pdf

[9] The Complete Book on Wine Production. (n.d.). Retrieved May 23, 2015, from https://books.google.co.uk/books?id=0eY9AQAAQBAJ&pg=PA228&lpg=PA228&dq=Sulfate content of wine&source=bl&ots=Udv_O_6gWN&sig=TkTrF3FHN8lNzjQKMmDASSOY9Yo&hl=en&sa=X&ei=6sBhVeeZBbCv7AbC34GYAw&redir_esc=y#v=onepage&q=legal limits&f=false

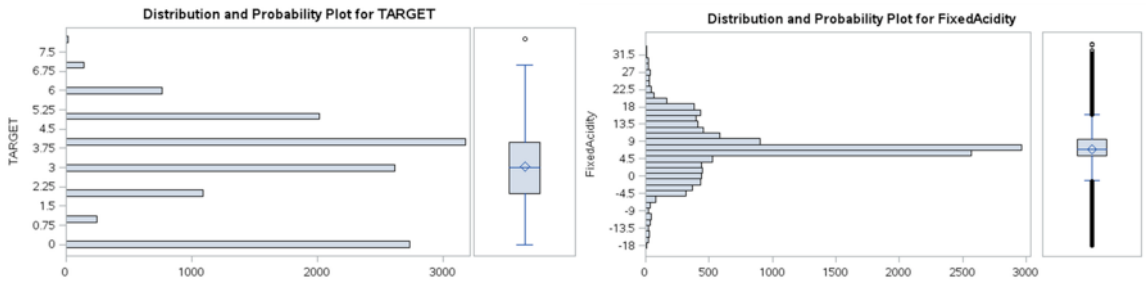
[10] Sweetness of wine. (n.d.). Retrieved May 20, 2015, from http://en.wikipedia.org/wiki/Sweetness_of_wine

than 60 mg/L, so 1 g/L also looks to be high. Total Sulfur Dioxide has a legal limit of 350 mg/L so 1,057 mg/L indicates that there are issue with the data. We already know that the pH of wine ranges from 2.9 – 3.9, so 6 pH is much higher than we would expect. Sulphates can be no higher than 2-3 g, so we know that 4 g is not an acceptable value. Alcohol can be no higher than 21%, so again know that the maximum value is most likely due to data issues. All of these values will have to be addressed in the data preparation stage.

Variable	N Miss	Mean	Std Dev	Variance	Median	Maximum	Minimum
TARGET	0	3	2	4	3	8	0
FixedAcidity	0	7	6	40	7	34	-18
VolatileAcidity	0	0	1	1	0	4	-3
CitricAcid	0	0	1	1	0	4	-3
ResidualSugar	616	5	34	1139	4	141	-128
Chlorides	638	0	0	0	0	1	-1
FreeSulfurDioxide	647	31	149	22116	30	623	-555
TotalSulfurDioxide	682	121	232	53784	123	1057	-823
Density	0	1	0	0	1	1	1
pH	395	3	1	0	3	6	0
Sulphates	1210	1	1	1	1	4	-3
Alcohol	653	10	4	14	10	27	-5
LabelAppeal	0	-0	1	1	0	2	-2
AcidIndex	0	8	1	2	8	17	4
STARS	3359	2	1	1	2	4	1

Figure 1 – PROC MEANS output

After looking at the values of the variables, I proceeded to examine their distributions. Beginning with the target variable, we can see that the distribution looks to be fairly normal with the exception of the zero values. Given that zero counts represent approximately 21% of the dataset, I believe that Zero Inflated Poisson and Zero Inflated Negative Binomial will be very useful in working with data. After seeing all of the issues with data values, I expected there to be many issues with the distributions of the predictor variables. As you can see in Figure 2, however, most of the distributions appear to be rather normal (with the exception of Acid Index, which is skewed to the left). Based on the accompanying box plots, the predictor variables seem to have many outliers that extend well beyond the whiskers. Given the analyses already conducted, these extreme values expected, and we will definitely need to address them in the data preparation stage.



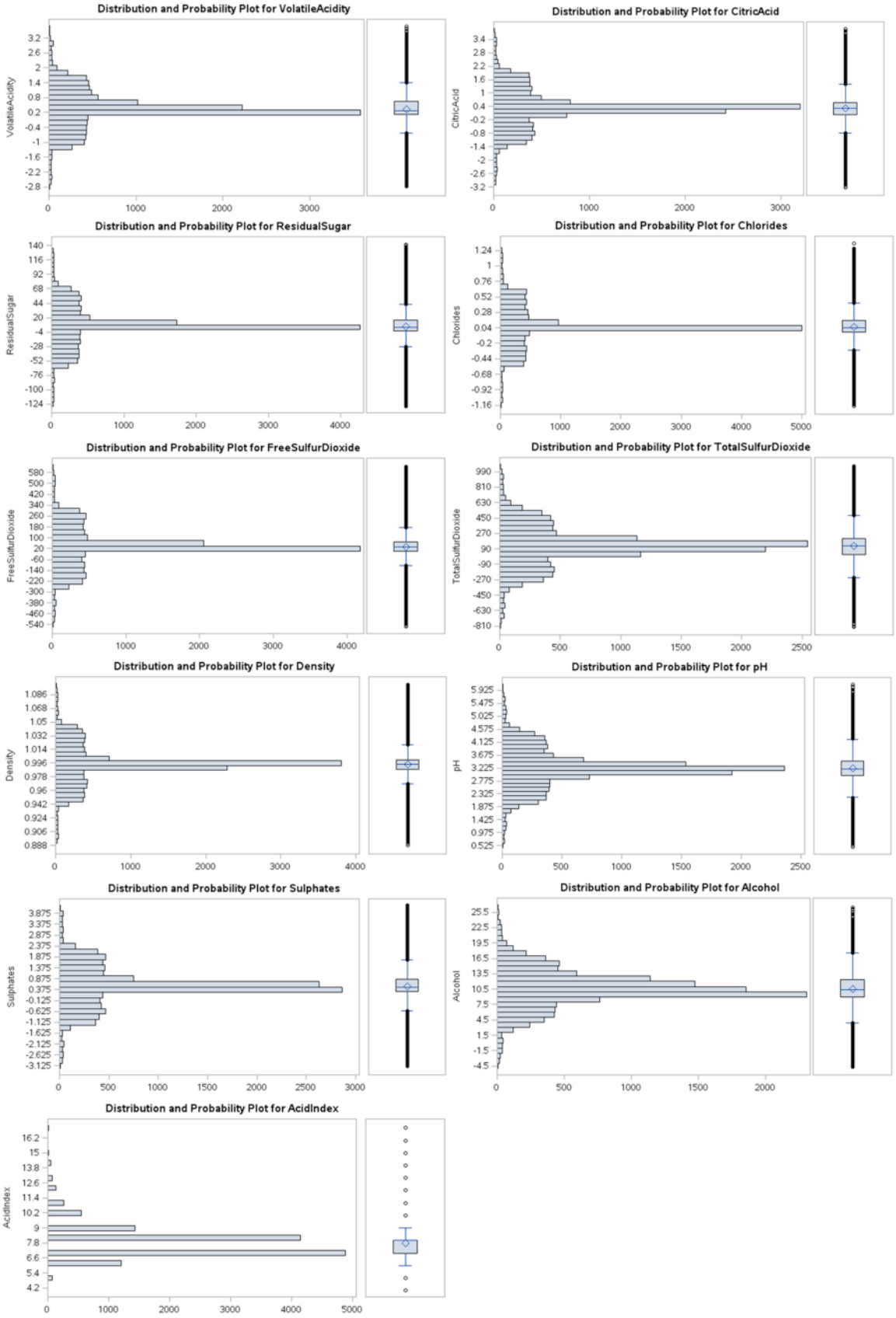

Distribution and Probability Plot for VolatileAcidity
VolatileAcidity
Distribution and Probability Plot for CitricAcid
CitricAcid
Distribution and Probability Plot for ResidualSugar
ResidualSugar
Distribution and Probability Plot for Chlorides
Chlorides
Distribution and Probability Plot for FreeSulfurDioxide
FreeSulfurDioxide
Distribution and Probability Plot for TotalSulfurDioxide
TotalSulfurDioxide
Distribution and Probability Plot for Density
Density
Distribution and Probability Plot for pH
pH
Distribution and Probability Plot for Sulphates
Sulphates
Distribution and Probability Plot for Alcohol
Alcohol
Distribution and Probability Plot for AcidIndex
AcidIndex

Figure 2 – PROC UNIVARIATE variable distributions

Moving on to the discrete variables, we can see from the PROC FREQ output in Figure 3 that four cases of wine sold is our largest bucket (we have already seen that the second largest bucket is zero cases sold). We can also see that the largest bucket for Label Appeal is zero and that the largest grouping of wine received a two-star expert rating. I also tried binning the values to see if there were any interesting insights - target_high checks if a wine sold more than three cases, label_appeal_high checks if the appeal is greater than zero and stars_high checks if the wine has more than two stars. As you can see in the bottom two tables of Figure 3, the results are not what one would initially suspect. For wines that sold higher numbers of cases, most did not have one of the higher star rankings. Likewise, the high-selling wines do not appear to have the more appealing labels. Perhaps we can infer from this analysis that "classy" wines do not have sales volumes as high as Two-Buck Chuck. We don't have bottle price in this dataset, but I would speculate that the high-volume wines have a much lower price point.

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

LabelAppeal	Frequency	Percent	Cumulative Frequency	Cumulative Percent
-2	504	3.94	504	3.94
-1	3136	24.51	3640	28.45
0	5617	43.90	9257	72.35
1	3048	23.82	12305	96.17
2	490	3.83	12795	100.00

STARS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	3359	26.25	3359	26.25
1	3042	23.77	6401	50.03
2	3570	27.90	9971	77.93
3	2212	17.29	12183	95.22
4	612	4.78	12795	100.00

Table of target_high by label_appeal_high

target_high	label_appeal_high 0	1	Total
0	5833 45.59 87.32 63.01	847 6.62 12.68 23.94	6680 52.21
1	3424 26.76 55.99 36.99	2691 21.03 44.01 76.06	6115 47.79
Total	9257 72.35	3538 27.65	12795 100.00

Table of target_high by stars_high

target_high	stars_high 0	1	Total
0	6356 49.68 95.15 63.74	324 2.53 4.85 11.47	6680 52.21
1	3615 28.25 59.12 36.26	2500 19.54 40.88 88.53	6115 47.79
Total	9971 77.93	2824 22.07	12795 100.00

Figure 3 – PROC FREQ output for discrete variables

The correlation analysis for this assignment is rather short given that most of the variables in our dataset are for chemical qualities of the wine and we do not have a baseline understanding of how they should correlate with the number of wine cases sold. However, it does make sense that expert star rating and label have the highest correlation with the target variable since those are factor that consumers regularly consider when purchasing wine. I find these results especially interesting given the binning analysis previously conducted. Although having a one of the best labels or highest ratings does not mean that the wine will be a high seller, both of these act as positive factors in relation so wine sales.

Obs	_NAME_	TARGET
1	TARGET	1.00000
2	target_high	0.81024
3	STARS	0.55879
4	stars_high	0.47124
5	LabelAppeal	0.35650
6	label_appeal_high	0.28349
7	Alcohol	0.06206
8	TotalSulfurDioxide	0.05148
9	FreeSulfurDioxide	0.04382
10	ResidualSugar	0.01649
11	CitricAcid	0.00868
12	pH	-0.00944
13	Density	-0.03552
14	Chlorides	-0.03826
15	Sulphates	-0.03885
16	FixedAcidity	-0.04901
17	VolatileAcidity	-0.08879
18	AcidIndex	-0.24605

Figure 4 – Correlation Analysis

Data Preparation

After completing the exploratory data analysis, we can see that this is a very dirty dataset that will require extensive cleaning before be can build our models. Starting with imputing, I decided to give all of the records without an expert score a value of zero. This seems reasonable since it is a valid score on the ratings spectrum and it is the only value to not appear in our data. This leads me to believe that zero scores were substituted for blank values. For the remaining seven variables with missing values (Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, pH, Sulphates, and Alcohol), I decided to impute the values using the medians of our training dataset. As we saw in the PROC MEANS table (Figure 1), the variables we are trying to impute appear to have very similar mean and median values. I ultimately decided the median values would be a better choice to use for this imputing process given that they are less sensitive to outliers. As we learned in Unit 1, it is best practice to create new variables when we are imputing rather than replace the existing values, so I created new variables (using the IMP_ prefix) for this purpose. Additionally, I followed the best practice of creating corresponding flag variables (with the m_ prefix) to indicate when a value was missing.

Next we need to handle the variable with unexplained negative values (Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Sulphates, and Alcohol). As we saw during the exploratory data analysis, most of these values seem to be due to data entry errors, so I will be taking the absolute values of these ten variables to ensure that they are all positive. The next issue that needs to be addressed is the outliers in our data. This is where we will be able to leverage all of the domain knowledge collected earlier in the assignment (See Domain Knowledge Research section for rationale of the proceeding decisions). Based on this information, I will be trimming pH into a range of 2.9–3.9, alcohol content into a range of 5%-21%, Sulphates to an upper maximum of 3 g, Total Sulfur Dioxide to an upper maximum of 350 mg/L, and Volatile Acidity to an upper maximum of 1.2 g/L. Chloride, Citric Acid and Fixed Acidity all have values that seem to go much higher than their expected ranges (<60 mg/L, 0-500 mg/L, and 0-8000 mg/L respectively), so I will try attempt to trim at the 75th percentile.

Model Building

Model 1 – Poisson Regression

The first model that I created used Poisson Regression. What I found really interesting about this model was that the coefficients for all classes of stars and label appeal were negative. I would have expected the beta values to become positive as the star ratings moved higher and the label appeal became positive. It also appears that several of our variables are not statistically significant (Residual Sugar, Free Sulfur Dioxide, Density, Sulphates, Acid Index=10-16, and Fixed Acidity).

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	13525.5385	1.0598
Scaled Deviance	13E3	13525.5385	1.0598
Pearson Chi-Square	13E3	11182.0159	0.8762
Scaled Pearson X2	13E3	11182.0159	0.8762
Log Likelihood		8863.3912	
Full Log Likelihood		-22733.7801	
AIC (smaller is better)		45533.5602	
AICC (smaller is better)		45533.7360	
BIC (smaller is better)		45779.6349	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.1851	0.4945	0.2159	2.1543	5.74	0.0166
abs_residualsugar		1	-0.0000	0.0002	-0.0004	0.0004	0.02	0.8832
T_75_chlorides		1	-0.2734	0.0954	-0.4603	-0.0864	8.22	0.0042
abs_freesulfurdioxid		1	0.0000	0.0000	-0.0001	0.0001	0.54	0.4622
T_totalsulfurdioxide		1	0.0001	0.0000	0.0000	0.0002	4.50	0.0340
Density		1	-0.2887	0.1920	-0.6650	0.0875	2.26	0.1326
T_ph		1	-0.0380	0.0151	-0.0675	-0.0084	6.35	0.0118
T_sulphates		1	-0.0125	0.0085	-0.0291	0.0041	2.17	0.1411
LabelAppeal	-2	1	-0.6981	0.0425	-0.7813	-0.6148	270.32	<.0001
LabelAppeal	-1	1	-0.4589	0.0250	-0.5079	-0.4098	336.65	<.0001
LabelAppeal	0	1	-0.2683	0.0229	-0.3131	-0.2235	137.55	<.0001
LabelAppeal	1	1	-0.1362	0.0232	-0.1816	-0.0907	34.47	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex	4	1	1.2129	0.5487	0.1375	2.2884	4.89	0.0271
AcidIndex	5	1	1.0639	0.4518	0.1783	1.9494	5.54	0.0185
AcidIndex	6	1	1.0844	0.4480	0.2064	1.9624	5.86	0.0155
AcidIndex	7	1	1.0502	0.4477	0.1727	1.9277	5.50	0.0190
AcidIndex	8	1	1.0164	0.4477	0.1389	1.8939	5.15	0.0232
AcidIndex	9	1	0.9043	0.4479	0.0265	1.7821	4.08	0.0435
AcidIndex	10	1	0.7437	0.4486	-0.1355	1.6229	2.75	0.0973
AcidIndex	11	1	0.3812	0.4511	-0.5030	1.2654	0.71	0.3981
AcidIndex	12	1	0.3678	0.4552	-0.5243	1.2599	0.65	0.4190
AcidIndex	13	1	0.5259	0.4573	-0.3705	1.4222	1.32	0.2502
AcidIndex	14	1	0.4302	0.4664	-0.4839	1.3442	0.85	0.3563
AcidIndex	15	1	0.8702	0.5127	-0.1348	1.8751	2.88	0.0897
AcidIndex	16	1	0.2079	0.6328	-1.0323	1.4481	0.11	0.7425
AcidIndex	17	0	0.0000	0.0000	0.0000	0.0000	.	.
IMP_stars	0	1	-1.3119	0.0243	-1.3596	-1.2642	2908.00	<.0001
IMP_stars	1	1	-0.5571	0.0217	-0.5995	-0.5146	660.91	<.0001
IMP_stars	2	1	-0.2395	0.0199	-0.2786	-0.2005	144.52	<.0001
IMP_stars	3	1	-0.1212	0.0202	-0.1608	-0.0816	35.95	<.0001
IMP_stars	4	0	0.0000	0.0000	0.0000	0.0000	.	.
T_75_fixedacidity		1	0.0003	0.0032	-0.0060	0.0066	0.01	0.9210
T_volatileacidity		1	-0.0502	0.0134	-0.0765	-0.0239	13.98	0.0002
T_75_citricacid		1	0.1158	0.0304	0.0562	0.1754	14.48	0.0001
T_alcohol		1	0.0051	0.0016	0.0020	0.0081	10.38	0.0013
Scale		0	1.0000	0.0000	1.0000	1.0000		

Figure 5 – Poisson Regression Model Output

Model 2 – Negative Binomial Regression

The next model we built used Negative Binomial Regression. This model appears to be very similar to the Poisson Regression model in regards to its beta values, as well as the variables that are not statistically significant. AIC values are very similar to that of the Poisson model as well.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	13525.5385	1.0598
Scaled Deviance	13E3	13525.5385	1.0598
Pearson Chi-Square	13E3	11182.0070	0.8762
Scaled Pearson X2	13E3	11182.0070	0.8762
Log Likelihood		8863.3912	
Full Log Likelihood		-22733.7801	
AIC (smaller is better)		45535.5602	
AICC (smaller is better)		45535.7467	
BIC (smaller is better)		45789.0917	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.1851	0.4945	0.2159	2.1543	5.74	0.0166
abs_residualsugar		1	-0.0000	0.0002	-0.0004	0.0004	0.02	0.8832
T_75_chlorides		1	-0.2734	0.0954	-0.4603	-0.0864	8.22	0.0042
abs_freesulfurdioxid		1	0.0000	0.0000	-0.0001	0.0001	0.54	0.4622
T_totalsulfurdioxide		1	0.0001	0.0000	0.0000	0.0002	4.50	0.0340
Density		1	-0.2887	0.1920	-0.6650	0.0875	2.26	0.1326
T_ph		1	-0.0380	0.0151	-0.0675	-0.0084	6.35	0.0118
T_sulphates		1	-0.0125	0.0085	-0.0291	0.0041	2.17	0.1411
LabelAppeal	-2	1	-0.6981	0.0425	-0.7813	-0.6148	270.31	<.0001
LabelAppeal	-1	1	-0.4589	0.0250	-0.5079	-0.4098	336.65	<.0001
LabelAppeal	0	1	-0.2683	0.0229	-0.3131	-0.2235	137.55	<.0001
LabelAppeal	1	1	-0.1362	0.0232	-0.1816	-0.0907	34.47	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex	4	1	1.2129	0.5487	0.1375	2.2884	4.89	0.0271
AcidIndex	5	1	1.0639	0.4518	0.1783	1.9494	5.54	0.0185
AcidIndex	6	1	1.0844	0.4480	0.2064	1.9624	5.86	0.0155
AcidIndex	7	1	1.0502	0.4477	0.1727	1.9277	5.50	0.0190
AcidIndex	8	1	1.0164	0.4477	0.1389	1.8939	5.15	0.0232
AcidIndex	9	1	0.9043	0.4479	0.0265	1.7821	4.08	0.0435
AcidIndex	10	1	0.7437	0.4486	-0.1355	1.6229	2.75	0.0973
AcidIndex	11	1	0.3812	0.4511	-0.5030	1.2654	0.71	0.3981
AcidIndex	12	1	0.3678	0.4552	-0.5243	1.2599	0.65	0.4190
AcidIndex	13	1	0.5259	0.4573	-0.3705	1.4222	1.32	0.2502
AcidIndex	14	1	0.4302	0.4664	-0.4839	1.3442	0.85	0.3563
AcidIndex	15	1	0.8702	0.5127	-0.1348	1.8751	2.88	0.0897
AcidIndex	16	1	0.2079	0.6328	-1.0323	1.4481	0.11	0.7425
AcidIndex	17	0	0.0000	0.0000	0.0000	0.0000	.	.
IMP_stars	0	1	-1.3119	0.0243	-1.3596	-1.2642	2908.00	<.0001
IMP_stars	1	1	-0.5571	0.0217	-0.5995	-0.5146	660.91	<.0001
IMP_stars	2	1	-0.2395	0.0199	-0.2786	-0.2005	144.52	<.0001
IMP_stars	3	1	-0.1212	0.0202	-0.1608	-0.0816	35.95	<.0001
IMP_stars	4	0	0.0000	0.0000	0.0000	0.0000	.	.
T_75_fixedacidity		1	0.0003	0.0032	-0.0060	0.0066	0.01	0.9210
T_volatileacidity		1	-0.0502	0.0134	-0.0765	-0.0239	13.98	0.0002
T_75_citricacid		1	0.1158	0.0304	0.0562	0.1754	14.48	0.0001
T_alcohol		1	0.0051	0.0016	0.0020	0.0081	10.38	0.0013
Dispersion		0	0.0000	0.0000	0.0000	0.0000		

Figure 6 – Negative Binomial Regression Output

Model 3 – Zero Inflated Poisson

The next model built was using Zero Inflated Poisson Regression. The top parameter table was for the Poisson model, and the bottom table was for the logistic regression model to determine if a case was sold. The beta values for label appeal and star rating are still negative in the Poisson model. However, have a star value of 0-2 seems to have a rather significant positive impact on the logistic model. The AIC value also seems to be lower than the fist two models.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		40610.6322	
Scaled Deviance		40610.6322	
Pearson Chi-Square	13E3	5660.0170	0.4447
Scaled Pearson X2	13E3	5660.0170	0.4447
Log Likelihood		11291.8552	
Full Log Likelihood		-20305.3161	
AIC (smaller is better)		40742.6322	
AICC (smaller is better)		40743.3270	
BIC (smaller is better)		41234.7816	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.8736	0.5014	0.8909	2.8563	13.96	0.0002
abs_residualsugar		1	-0.0000	0.0002	-0.0005	0.0004	0.05	0.8187
T_75_chlorides		1	-0.1288	0.0970	-0.3189	0.0612	1.77	0.1839
abs_freesulfurdioxid		1	-0.0000	0.0000	-0.0001	0.0001	0.06	0.8027
T_totalsulfurdioxide		1	-0.0000	0.0001	-0.0001	0.0001	0.63	0.4257
Density		1	-0.2514	0.1979	-0.6393	0.1365	1.61	0.2040
T_ph		1	0.0084	0.0154	-0.0218	0.0386	0.30	0.5862
T_sulphates		1	0.0058	0.0086	-0.0111	0.0227	0.45	0.5003
LabelAppeal	-2	1	-1.0775	0.0454	-1.1666	-0.9885	562.33	<.0001
LabelAppeal	-1	1	-0.6377	0.0257	-0.6880	-0.5875	618.03	<.0001
LabelAppeal	0	1	-0.3496	0.0232	-0.3950	-0.3041	227.27	<.0001
LabelAppeal	1	1	-0.1595	0.0234	-0.2054	-0.1136	46.45	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex	4	1	0.1364	0.5584	-0.9581	1.2310	0.06	0.8070
AcidIndex	5	1	0.1575	0.4528	-0.7299	1.0450	0.12	0.7279
AcidIndex	6	1	0.1854	0.4489	-0.6944	1.0653	0.17	0.6796
AcidIndex	7	1	0.1588	0.4486	-0.7205	1.0381	0.13	0.7234
AcidIndex	8	1	0.1433	0.4486	-0.7360	1.0226	0.10	0.7494
AcidIndex	9	1	0.1072	0.4488	-0.7725	0.9869	0.06	0.8112
AcidIndex	10	1	0.0367	0.4496	-0.8446	0.9180	0.01	0.9349
AcidIndex	11	1	0.0430	0.4524	-0.8437	0.9298	0.01	0.9242
AcidIndex	12	1	0.1203	0.4570	-0.7754	1.0160	0.07	0.7924
AcidIndex	13	1	0.1763	0.4593	-0.7238	1.0765	0.15	0.7010
AcidIndex	14	1	0.2044	0.4689	-0.7146	1.1234	0.19	0.6628
AcidIndex	15	1	0.2591	0.5186	-0.7574	1.2756	0.25	0.6174
AcidIndex	16	1	0.3589	0.6335	-0.8827	1.6005	0.32	0.5710
AcidIndex	17	0	0.0000	0.0000	0.0000	0.0000	.	.
IMP_stars	0	1	-0.3764	0.0256	-0.4265	-0.3262	216.08	<.0001
IMP_stars	1	1	-0.3159	0.0222	-0.3593	-0.2725	203.39	<.0001
IMP_stars	2	1	-0.1942	0.0200	-0.2335	-0.1550	94.17	<.0001
IMP_stars	3	1	-0.0984	0.0202	-0.1381	-0.0588	23.70	<.0001
IMP_stars	4	0	0.0000	0.0000	0.0000	0.0000	.	.
T_75_fixedacidity		1	0.0027	0.0033	-0.0037	0.0092	0.68	0.4092
T_volatileacidity		1	-0.0195	0.0137	-0.0463	0.0072	2.05	0.1523
T_75_citricacid		1	0.0387	0.0316	-0.0231	0.1006	1.51	0.2197
T_alcohol		1	0.0081	0.0016	0.0050	0.0113	25.83	<.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	-8.9067	783.4239	-1544.39	1526.576	0.00	0.9909
abs_residualsugar		1	-0.0006	0.0014	-0.0033	0.0021	0.18	0.6696
T_75_chlorides		1	1.3748	0.6703	0.0610	2.6886	4.21	0.0403
abs_freesulfurdioxid		1	-0.0006	0.0003	-0.0012	0.0001	3.18	0.0746
T_totalsulfurdioxide		1	-0.0016	0.0003	-0.0022	-0.0009	23.39	<.0001
Density		1	0.4877	1.3190	-2.0975	3.0730	0.14	0.7116
T_ph		1	0.5516	0.1003	0.3550	0.7482	30.25	<.0001
T_sulphates		1	0.1722	0.0557	0.0631	0.2813	9.57	0.0020
LabelAppeal	-2	1	-3.4163	0.3811	-4.1633	-2.6694	80.36	<.0001
LabelAppeal	-1	1	-1.9156	0.2185	-2.3439	-1.4873	76.85	<.0001
LabelAppeal	0	1	-1.1687	0.2117	-1.5836	-0.7539	30.49	<.0001
LabelAppeal	1	1	-0.4389	0.2183	-0.8668	-0.0110	4.04	0.0444
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex	4	1	-14.0199	361.1163	-721.795	693.7551	0.00	0.9690
AcidIndex	5	1	-13.0399	361.1130	-720.808	694.7285	0.00	0.9712
AcidIndex	6	1	-13.2002	361.1127	-720.968	694.5676	0.00	0.9708
AcidIndex	7	1	-13.0384	361.1127	-720.806	694.7294	0.00	0.9712
AcidIndex	8	1	-12.8061	361.1127	-720.574	694.9617	0.00	0.9717
AcidIndex	9	1	-12.1355	361.1127	-719.903	695.6323	0.00	0.9732
AcidIndex	10	1	-11.8812	361.1127	-719.649	695.8866	0.00	0.9738
AcidIndex	11	1	-10.4087	361.1127	-718.177	697.3592	0.00	0.9770
AcidIndex	12	1	-10.3996	361.1128	-718.168	697.3685	0.00	0.9770
AcidIndex	13	1	-9.8381	361.1130	-717.607	697.9303	0.00	0.9783
AcidIndex	14	1	-10.7529	361.1129	-718.521	697.0154	0.00	0.9762
AcidIndex	15	1	-11.4232	361.1140	-719.194	696.3472	0.00	0.9748
AcidIndex	16	1	1.7849	440.6228	-861.820	865.3898	0.00	0.9968
AcidIndex	17	0	0.0000	0.0000	0.0000	0.0000	.	.
IMP_stars	0	1	20.6158	695.2329	-1342.02	1383.247	0.00	0.9763
IMP_stars	1	1	18.5237	695.2329	-1344.11	1381.155	0.00	0.9787
IMP_stars	2	1	14.7764	695.2330	-1347.86	1377.408	0.00	0.9830
IMP_stars	3	1	-6.1503	820.3945	-1614.09	1601.793	0.00	0.9940
IMP_stars	4	0	0.0000	0.0000	0.0000	0.0000	.	.
T_75_fixedacidity		1	0.0233	0.0219	-0.0197	0.0663	1.13	0.2883
T_volatileacidity		1	0.3215	0.0897	0.1457	0.4972	12.85	0.0003
T_75_citricacid		1	-0.7226	0.1957	-1.1061	-0.3391	13.64	0.0002
T_alcohol		1	0.0273	0.0106	0.0065	0.0482	6.61	0.0101

Figure 7 – Zero Inflated Poisson Regression Output

Model 4 – Zero Inflated Negative Binomial

The next model used is Zero Inflated Negative Binomial regression. I received a warning that convergence for this model was questionable based on the relative Hessian convergence criterion, so I am not sure how reliable its performance will be. We can see again that the label appeal variable has negative beta values. This model seems to have an AIC score that is slightly higher than the previous model.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		40719.7716	
Scaled Deviance		40719.7716	
Pearson Chi-Square	13E3	5874.6323	0.4604
Scaled Pearson X2	13E3	5874.6323	0.4604
Log Likelihood		-20359.8858	
Full Log Likelihood		-20359.8858	
AIC (smaller is better)		40793.7716	
AICC (smaller is better)		40793.9920	
BIC (smaller is better)		41069.6736	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.7857	0.2103	1.3735 2.1980	72.08	<.0001
abs_residualsugar		1	-0.0000	0.0002	-0.0005 0.0004	0.05	0.8280
T_75_chlorides		1	-0.1188	0.0970	-0.3090 0.0713	1.50	0.2207
abs_freesulfurdioxid		1	-0.0000	0.0000	-0.0001 0.0001	0.04	0.8356
T_totalsulfurdioxide		1	-0.0000	0.0001	-0.0001 0.0001	0.70	0.4021
Density		1	-0.2637	0.1979	-0.6516 0.1243	1.77	0.1828
T_ph		1	0.0089	0.0154	-0.0213 0.0390	0.33	0.5652
T_sulphates		1	0.0065	0.0086	-0.0104 0.0234	0.57	0.4518
LabelAppeal	-2	1	-1.0818	0.0452	-1.1704 -0.9932	572.16	<.0001
LabelAppeal	-1	1	-0.6406	0.0255	-0.6906 -0.5906	630.34	<.0001
LabelAppeal	0	1	-0.3499	0.0230	-0.3950 -0.3048	230.96	<.0001
LabelAppeal	1	1	-0.1586	0.0233	-0.2043 -0.1130	46.37	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000 0.0000	.	.
AcidIndex		1	-0.0195	0.0051	-0.0294 -0.0096	14.92	0.0001
IMP_stars		1	0.1003	0.0052	0.0901 0.1105	373.15	<.0001
T_75_fixedacidity		1	0.0020	0.0033	-0.0045 0.0084	0.36	0.5495
T_volatileacidity		1	-0.0191	0.0137	-0.0459 0.0077	1.96	0.1618
T_75_citricacid		1	0.0392	0.0316	-0.0227 0.1010	1.54	0.2147
T_alcohol		1	0.0080	0.0016	0.0049 0.0111	25.03	<.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	-4.4472	1.4007	-7.1925 -1.7019	10.08	0.0015
abs_residualsugar		1	-0.0007	0.0014	-0.0035 0.0020	0.26	0.6125
T_75_chlorides		1	1.4014	0.6752	0.0780 2.7249	4.31	0.0379
abs_freesulfurdioxid		1	-0.0006	0.0003	-0.0012 0.0001	3.00	0.0833
T_totalsulfurdioxide		1	-0.0017	0.0003	-0.0023 -0.0010	25.13	<.0001
Density		1	0.2737	1.3233	-2.3199 2.8673	0.04	0.8361
T_ph		1	0.5759	0.1008	0.3784 0.7734	32.65	<.0001
T_sulphates		1	0.1881	0.0556	0.0792 0.2970	11.47	0.0007
LabelAppeal	-2	1	-3.4427	0.3848	-4.1969 -2.6884	80.03	<.0001
LabelAppeal	-1	1	-1.8569	0.2083	-2.2652 -1.4487	79.46	<.0001
LabelAppeal	0	1	-1.0746	0.1999	-1.4664 -0.6827	28.89	<.0001
LabelAppeal	1	1	-0.3761	0.2058	-0.7794 0.0271	3.34	0.0675
LabelAppeal	2	0	0.0000	0.0000	0.0000 0.0000	.	.
AcidIndex		1	0.4422	0.0271	0.3890 0.4953	265.87	<.0001
IMP_stars		1	-2.3773	0.0602	-2.4953 -2.2593	1559.38	<.0001
T_75_fixedacidity		1	0.0301	0.0220	-0.0130 0.0732	1.88	0.1705
T_volatileacidity		1	0.3673	0.0902	0.1905 0.5440	16.58	<.0001
T_75_citricacid		1	-0.7218	0.1976	-1.1091 -0.3346	13.35	0.0003
T_alcohol		1	0.0291	0.0107	0.0081 0.0501	7.41	0.0065

Figure 8 – Zero Inflated Negative Binomial Regression Output

Model 5 – OLS Regression

Last, since have learned that is always worthwhile to try linear regression when trying to predict counts, I created a model using OLS regression and stepwise variable selection. Interestingly enough, this model actually has the lowest AIC score of all the models. I also found it interesting that the model chose some of the imputation flags to include in the model (m_ prefix). When I included them in my other models, the AIC score always decreased.

Observations	12795
Parameters	14
Error DF	12781
MSE	1.7139
R-Square	0.5386
Adj R-Square	0.5382
AIC	6907.3
BIC	6909.4
CP	14

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4.70602	0.45750	10.29	<.0001	0
T_75_chlorides	1	-0.80611	0.22016	-3.66	0.0003	1.00910
T_totalsulfurdioxide	1	0.00039303	0.00011142	3.53	0.0004	1.00586
Density	1	-0.80734	0.43672	-1.85	0.0645	1.00266
T_ph	1	-0.12953	0.03440	-3.77	0.0002	1.02880
T_sulphates	1	-0.03179	0.01910	-1.66	0.0960	1.00248
LabelAppeal	1	0.46582	0.01366	34.10	<.0001	1.10615
AcidIndex	1	-0.20683	0.00906	-22.83	<.0001	1.07415
IMP_stars	1	0.77885	0.01566	49.72	<.0001	2.57847
T_volatileacidity	1	-0.16598	0.03034	-5.47	<.0001	1.00784
T_75_citricacid	1	0.34808	0.06801	5.12	<.0001	1.01716
T_alcohol	1	0.01473	0.00358	4.11	<.0001	1.00659
m_ph	1	-0.12673	0.06760	-1.87	0.0609	1.02076
m_stars	1	-0.68379	0.04083	-16.75	<.0001	2.40949

Figure 9 – OLS Regression Output

BONUS:
I completed the 10-point bonus assignment to create a data step to check the results of the Zero Inflated Poisson Model. As you can see from output in Figure 10, the results from the data step are nearly identical to the values generated by PROC PLM – the small differences can be attributed to rounding.

Obs	P_SCORE_ZIP	p_target_zip
1	3.08597	3.08683
2	3.22807	3.22926
3	2.95861	2.95915
4	3.08597	3.08683
5	2.83824	2.83850
6	3.04396	3.04471
7	2.95688	2.95741
8	3.04396	3.04471
9	2.83824	2.83850
10	3.04109	3.04182

Figure 10 – Bonus Assignment – PROC PRINT output

Model Selection

As usual, it is difficult to select my preferred model. When examining the predicted values for the test data, we can see in Figure 11 that the predicted values for Poisson and Negative Binomial are identical (similar mathematical properties). Up until I ran the linear regression model, the Zero Inflated Poisson model had the lowest AIC value. I am inclined to choose that model over the OLS model since I feel like it does a better job at handling the situations where zero cases of wine were sold - I ran into issues with the OLS model where is predicted negative cases sold when applied to the test dataset.

Obs	INDEX	p_target_poi	p_target_nb	p_target_zip	p_target_zinb	p_target_reg
1	3	1.13032	1.13032	1.59428	1.69130	1.19025
2	9	4.18693	4.18693	4.23641	4.15924	4.02893
3	10	2.48938	2.48938	2.36967	2.40670	2.42203
4	18	2.33917	2.33917	2.40853	2.45291	2.36265
5	21	0.98837	0.98837	0.78560	0.71844	0.87178
6	30	5.94184	5.94184	5.78839	5.77566	5.76325
7	31	2.20270	2.20270	4.09116	3.81678	3.43667
8	37	1.54898	1.54898	1.44413	1.39707	1.99467
9	39	0.66650	0.66650	0.17270	0.22581	0.41677
10	47	1.38156	1.38156	1.43963	1.37802	1.56856

Figure 11 – PROC PRINT - Comparison of predicted values for all models created

Model Deployment

As in the previous assignment, I will skip the process of writing out the entire data step for my Zero Inflated Poisson Model (given that I already created one with a smaller subset of the variables for the bonus points) and will comment on the process of scoring new records as they become available. We begin by applying all of the data preparation steps (imputation, absolute value, and trimming) for the new variables. We then proceed by multiplying the newly created fields by the beta values seen in the middle table of Figure 7 (the Poisson Regression model). After adding all of the values up like a standard formula, we simply take the exponent of that value (exp(value)). Moving on the lower table in Figure 7 (the logistic regression model), we again multiplying the newly created fields by the beta values and add everything up. At that point we need to take the exponent of that sum and divide it by the one plus the exponent of that sum ((exp(value) / (1+exp(value))). As we learned in the slide deck, they two numbers represent the predicted count and whether the count is greater than zero respectively, so multiplying these two numbers together will give us our final predicted value for the Zero Inflated Poisson model.

Conclusion

To conclude, we have learned about various modeling options when trying to predict a count value. Based on my results, it seems as though zero inflated models do a better job of handling the cases where the majority of the target variables have a value of zero. We also learned that equidispersion is not present, Poisson and Negative Binomial models can still be useful. Lastly, I learned how important domain knowledge is in a modeling exercise. If I had not done my due diligence of learning about wine properties prior to the exploratory data analysis, I would not have had any idea as to what values actually made sense and what were outliers. I plan to all that to my list of standard operating procedures for all modeling exercise going forward.