Homework 3 – PREDICT 411

Nate Bitting

Introduction and Data Exploration

For this assignment we will be working with the wine dataset. This dataset consists of the chemical properties of over 12,000 commercially available wines. The objective is to provide samples of these various wines to high-end restaurants throughout the United States. We would like to build a predictive model to aid the wine manufacturers by identifying which chemical properties yield the highest number of cases of these wine sold.

Table 1: Wine Dataset Variable List

Alphabetic List of Variables and Attributes			
#	Variable	Type	Len
15	AcidIndex	Num	8
13	Alcohol	Num	8
7	Chlorides	Num	8
5	CitricAcid	Num	8
10	Density	Num	8
3	FixedAcidity	Num	8
8	FreeSulfurDioxide	Num	8
1	INDEX	Num	8
14	LabelAppeal	Num	8
6	ResidualSugar	Num	8
16	STARS	Num	8
12	Sulphates	Num	8
2	TARGET	Num	8
9	TotalSulfurDioxide	Num	8
4	VolatileAcidity	Num	8
11	pH	Num	8

As you can see in Table 1, we are working with 14 numeric variables that will be used to predict the number of cases sold, denoted as the TARGET variable. It is important to note that while all variables are numeric, some are actually categorical variables including STARS, LabelAppeal, and AcidIndex.

Table 2: 10 Observations of Wine Dataset (does not include all variables to save space)

Obs	INDEX	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH
1	1	3	3.2	1.160	-0.98	54.20	-0.567	.	268	0.99280	3.33
2	2	3	4.5	0.160	-0.81	26.10	-0.425	15	-327	1.02792	3.38
3	4	5	7.1	2.640	-0.88	14.80	0.037	214	142	0.99518	3.12
4	5	3	5.7	0.385	0.04	18.80	-0.425	22	115	0.99640	2.24
5	6	4	8.0	0.330	-1.26	9.40	.	-167	108	0.99457	3.12
6	7	0	11.3	0.320	0.59	2.20	0.556	-37	15	0.99940	3.20
7	8	0	7.7	0.290	-0.40	21.50	0.060	287	156	0.99572	3.49
8	11	4	6.5	-1.220	0.34	1.40	0.040	523	551	1.03236	3.20
9	12	3	14.8	0.270	1.05	11.25	-0.007	-213	.	0.99620	4.93
10	13	6	5.5	-0.220	0.39	1.80	-0.277	62	180	0.94724	3.09

In Table 2, you can see a few observations just to get a better understanding of what we are working with for this assignment. You will notice these are some variables with missing values and others with negative values. We will need to explore this dataset to better understand any data cleansing we will need to perform. Also, given this is the first time we have ever worked with wine data before, we had to conduct some preliminary research about wine chemistry to understand what the appropriate ranges were for each chemical property included in the dataset. In Table 3, we have provided a list of all findings from that research.

Table 3: Preliminary Wine Chemistry Research Findings

Variable	Findings	Source
Fixed Acidity	Safe, gives a lot of the sour taste to grapes. Typically ranges from 1.5 to 14.5 in wine.	http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity
Volatile Acidity	US Legal Limits: • Red 1.2 g/L • White 1.1 g/L	http://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity
Residual Sugar	Ranges: • Dry wines: 0.2% - 0.3% • Off-Dry wines: 1.0% - 5.0% • Sweet Dessert wines: 5.0% - 15.0%	https://winemakermag.com/501-measuring-residual-sugar-techniques
Free Sulfur Dioxide	Ranges from 33-75% of Total Sulfur Dioxide	http://www.sdaws.org/News/Articles/documents/Sulfur%20Dioxide.pdf
Total Sulfur Dioxide	US Legal Maximum of 350. Try to keep below 100 to maintain fruitiness in wine. Wine average 80 mg/L (10 mg in the typical glass of wine). Ranges from 16 to 103: • Red 22 ± 6 • White 100 ± 3	http://www.piwine.com/media/home-wine-making-basics/using_sulfur_dioxide.pdf
pH	Ranges for wine (most fall between 3 - 4): • Sparkling <3.0 • White 3.0 – 3.5 • Most Reds 3.3 – 3.8 • Fortified up to 4.0	http://winemakersacademy.com/importance-ph-wine-making/
Alcohol	Ranges for wine (5 – 23%): • Whites: 5-15% • Reds: 12-16% • Fortified: 17-23% • Most wines fall between 12-15%	http://web2.slc.qc.ca/jmc/w05/Wine/results.htm

The findings in Table 3 provide us with valuable information we will need to rely on as we explore the dataset provided for this assignment. We noticed in Table 2, negative and/or very large values exist in some of these variables. Some of those large values seem to exceed some of the ranges provided in Table 3. We will cover how we plan to address these issues during data preparation section.

Missing Values

As we noticed earlier in Table 2, some of the variables had missing values that we will need to address before we move into the model building phase.

Table 4: Summary from SAS Means Procedure

Variable	N	N Miss	Mean	Median	Minimum	5th Pctl	50th Pctl	75th Pctl	90th Pctl	95th Pctl	99th Pctl	Maximum
TARGET	12795	0	3.03	3.00	0.00	0.00	3.00	4.00	5.00	6.00	7.00	8.00
FixedAcidity	12795	0	7.08	6.90	-18.10	-3.60	6.90	9.50	15.60	17.80	24.40	34.40
VolatileAcidity	12795	0	0.32	0.28	-2.79	-1.03	0.28	0.64	1.35	1.64	2.59	3.68
CitricAcid	12795	0	0.31	0.31	-3.24	-1.16	0.31	0.58	1.43	1.79	2.66	3.86
ResidualSugar	12179	616	5.42	3.90	-127.80	-52.70	3.90	15.90	49.80	62.70	99.20	141.15
Chlorides	12157	638	0.05	0.05	-1.17	-0.49	0.05	0.15	0.48	0.60	0.96	1.35
FreeSulfurDioxide	12148	647	30.85	30.00	-555.00	-224.00	30.00	70.00	230.00	284.00	469.00	623.00
TotalSulfurDioxide	12113	682	120.71	123.00	-823.00	-273.00	123.00	208.00	422.00	514.00	767.00	1057.00
Density	12795	0	0.99	0.99	0.89	0.95	0.99	1.00	1.03	1.04	1.07	1.10
pH	12400	395	3.21	3.20	0.48	2.06	3.20	3.47	4.10	4.37	5.13	6.13
Sulphates	11585	1210	0.53	0.50	-3.13	-1.05	0.50	0.86	1.77	2.09	3.16	4.24
Alcohol	12142	653	10.49	10.40	-4.70	4.10	10.40	12.40	15.20	16.70	20.30	26.50
LabelAppeal	12795	0	-0.01	0.00	-2.00	-1.00	0.00	1.00	1.00	1.00	2.00	2.00
AcidIndex	12795	0	7.77	8.00	4.00	6.00	8.00	8.00	9.00	10.00	13.00	17.00
STARS	9436	3359	2.04	2.00	1.00	1.00	2.00	3.00	3.00	4.00	4.00	4.00

N = number of observations;
N Miss = number of missing observations
Mean = mean value for each numeric variable

In Table 4, we notice there are several variables with missing values as denoted by the N Miss column. The variables with missing values include Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, pH, Sulphates, Alcohol, and STARS. We will cover the appropriate transformations and/or imputation for the missing values that exist in the dataset.

To better understand the distribution for the categorical variables and the target variable we have built frequency tables as shown in Table 5.

Table 5: Frequency Table for TARGET, LabelAppeal, STARS, and AcidIndex FREQ

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
1	3042	32.24	3042	32.24
2	3570	37.83	6612	70.07
3	2212	23.44	8824	93.51
4	612	6.49	9436	100.00
Frequency Missing = 3359				

AcidIndex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
4	3	0.02	3	0.02
5	75	0.59	78	0.61
6	1197	9.36	1275	9.96
7	4878	38.12	6153	48.09
8	4142	32.37	10295	80.46
9	1427	11.15	11722	91.61
10	551	4.31	12273	95.92
11	258	2.02	12531	97.94
12	128	1.00	12659	98.94
13	69	0.54	12728	99.48
14	47	0.37	12775	99.84
15	8	0.06	12783	99.91
16	5	0.04	12788	99.95
17	7	0.05	12795	100.00

As you can see, the TARGET variable ranges from 0 to 8 cases sold with the majority falling between 3-5 cases. LabelAppeal appears to have the majority of observations with a value of zero, indicating this variable might be a good candidate to be used in the zeromodel for a zero-inflated model. You will also notice 3,359 missing values for the STARS variable, which is most likely due to the fact that users have not yet rated those wine samples yet. Lastly, the majority of observations fall between 7 and 8 based on the distribution in the AcidIndex frequency table above. In order to explore the categorical variables in relation to the TARGET variable, we have constructed a crosstab frequency table for LabelAppeal, AcidIndex and STARS as shown in Table 6.

Table 6: Crosstab of Categorical Variables vs TARGET FLAG Frequency %

LabelAppeal	0	1	2	3	4	5	6	7	8	Grand Total
-2	4%	56%	16%	3%	0%	0%	0%	0%	0%	4%
-1	25%	36%	69%	43%	13%	4%	0%	0%	0%	25%
0	44%	8%	14%	52%	62%	38%	20%	3%	0%	44%
1	24%	0%	1%	3%	24%	52%	56%	56%	12%	24%
2	4%	0%	0%	0%	0%	5%	24%	42%	88%	4%

AcidIndex	0	1	2	3	4	5	6	7	8	Grand Total
4	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
5	0%	0%	0%	1%	1%	1%	1%	1%	0%	1%
6	6%	7%	8%	10%	10%	12%	11%	15%	0%	9%
7	27%	36%	42%	38%	42%	43%	46%	40%	24%	38%
8	29%	36%	32%	35%	34%	31%	31%	31%	71%	32%
9	16%	11%	12%	11%	10%	9%	7%	10%	6%	11%
10	9%	5%	4%	4%	2%	3%	1%	2%	0%	4%
11	6%	2%	1%	1%	1%	1%	0%	1%	0%	2%
12	3%	1%	0%	0%	0%	0%	1%	1%	0%	1%
13	2%	0%	0%	0%	0%	0%	0%	0%	0%	1%
14	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
15	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
16	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
17	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

STARS	0	1	2	3	4	5	6	7	8	Grand Total
0	75%	52%	31%	18%	8%	5%	4%	6%	12%	26%
1	22%	40%	43%	35%	23%	11%	3%	0%	0%	24%
2	3%	8%	23%	36%	42%	36%	26%	8%	0%	28%
3	0%	0%	3%	11%	24%	37%	41%	40%	24%	17%
4	0%	0%	0%	0%	3%	12%	26%	46%	65%	5%

Key Takeaways from Table 6

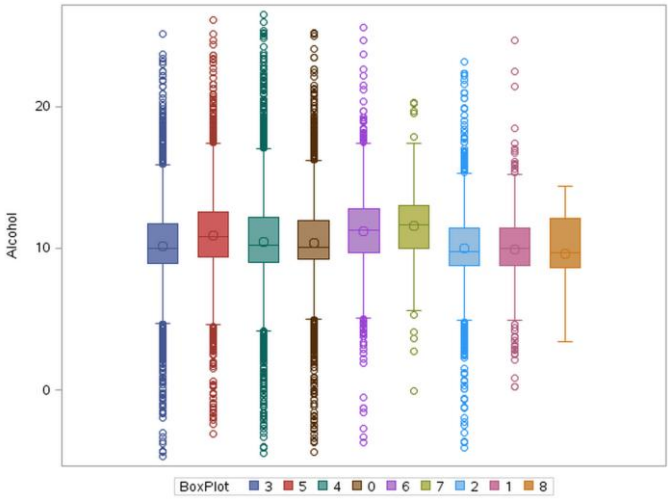
- LabelAppeal does have a positive correlation to greater wine sales as we notice the higher the Label Appeal, the percentage of observations tends to move to the right with a higher number of cases sold. This indicates that having a good label design could potentially be helpful in the increase of cases sold.
- The majority of observations form the dataset fall within the 7 to 8 range on the Acid Index. The data are very sparse for the ranges below 6 and above 9.
- STARS is the rating for the various wines provided in the dataset. We also can see a positive correlation with STARS to the TARGET variable as with each 1 unit increase in STARS rating, the number of cases sold tends to increase.

Assessment of Outliers

In order to understand any extreme observations that may exist in the data, we have constructed boxplots for each predictor variable as shown in Table 7.

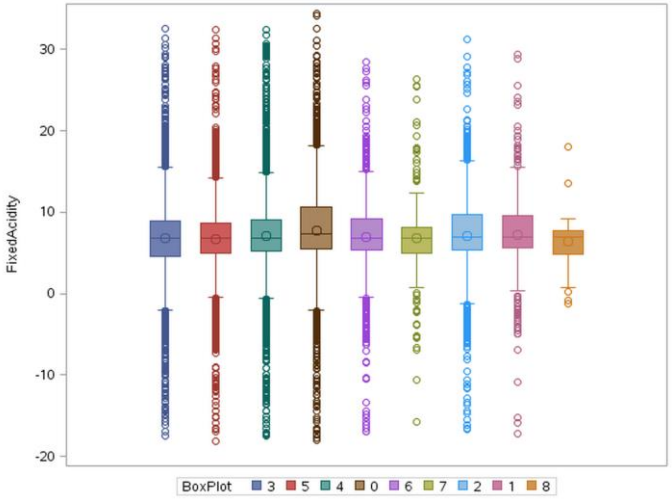
Table 7: Boxplots to explore outliers

6.1 Boxplot for Alcohol



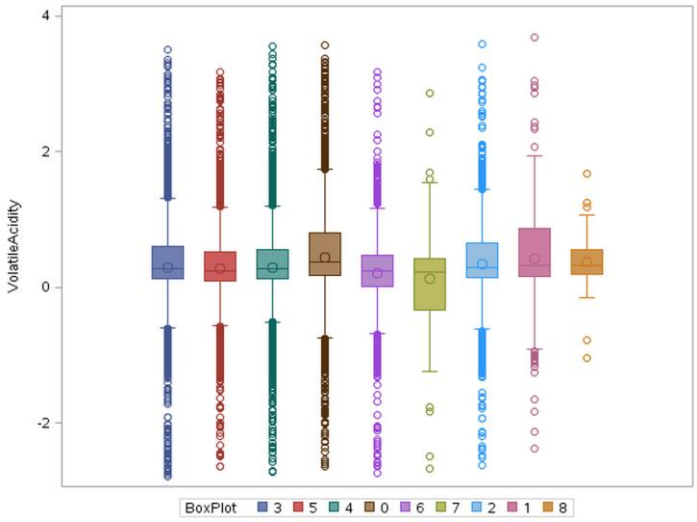
There does seem to be a positive correlation with relation to alcohol % and the number of cases sold. We notice the majority of the alcohol values tend to be higher as the number of cases sold increases. We also notice there are many extreme observations that exist with the alcohol variable. Lastly, we can see values falling below zero, which is something we must address before moving into the model building phase.

6.2 Boxplot for FixedAcidity



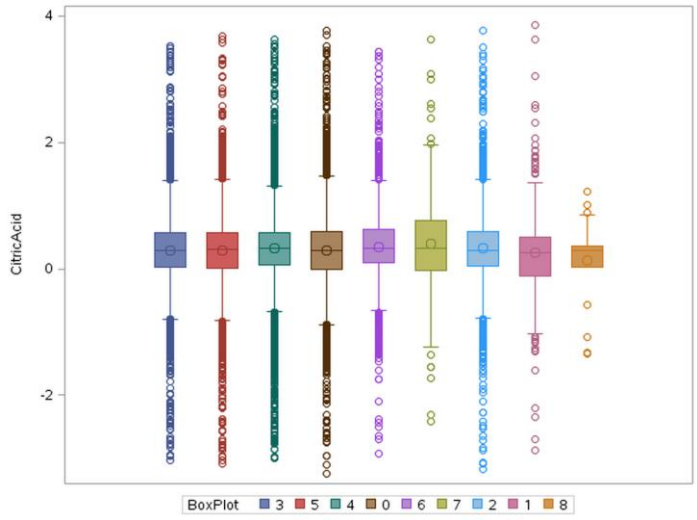
There appears to be a negative correlation with FixedAcidity to the number of cases sold. WE notice that the fixed acidity tends to decrease as the number of cases sold increases. We also can clearly see signs of outliers that exist with this variable. We also notice negative values in the data, which cannot be possible.

6.3 Boxplot for VolatileAcidity



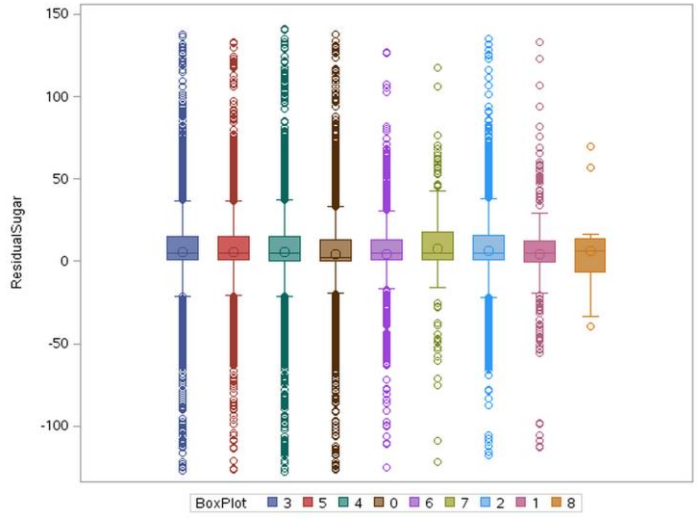
The majority of values for VolatileAcidity appear to fall between 0 and 1, however there are some values that are negative that must be addressed. There seems to be quite a few extreme observations for this variable.

6.4 Boxplot for CitricAcid



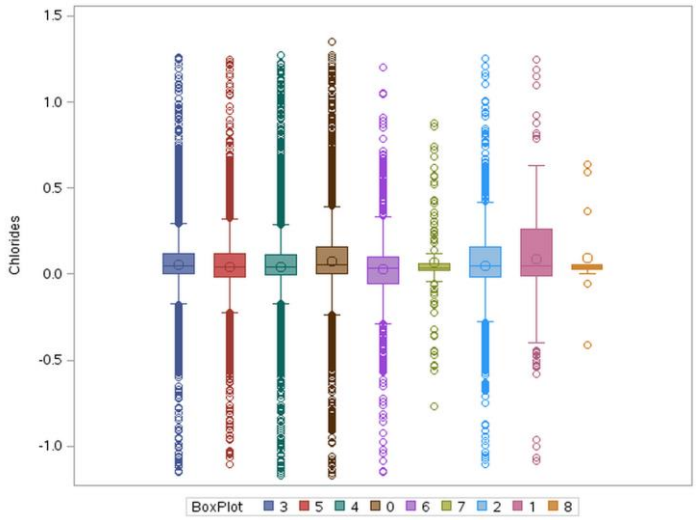
There does not seem to be any clear correlation between chlorides and number of cases sold. We do, however, see quite a few outliers and negative values that exist in the data. We must address this before moving into the model building phase.

6.5 Boxplot for ResidualSugar



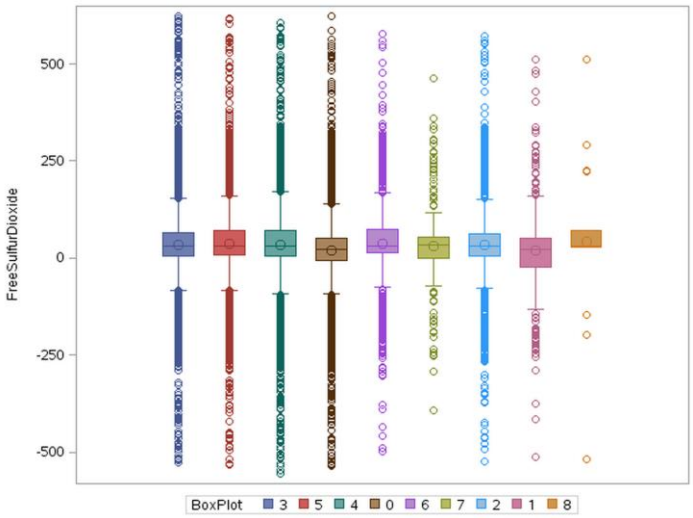
As we learned earlier in Table 3, Residual Sugar usually ranges from 0.2% to 15% and we can clearly see very large values exceeding or coming in below this range. We must assume outliers exist in the data and we will definitely need to address this in the data preparation phase.

6.6 Boxplot for Chlorides



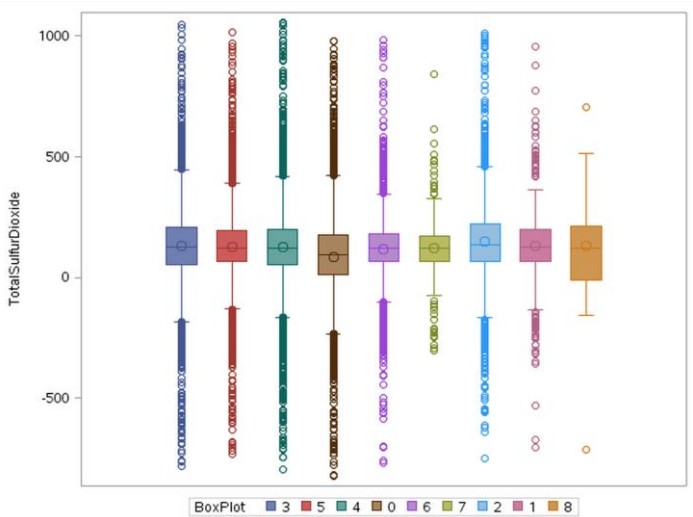
We notice negative values with this variable, which is not possible. We will have to address this in the data preparation phase.

6.7 Boxplot for FreeSulfurDioxide



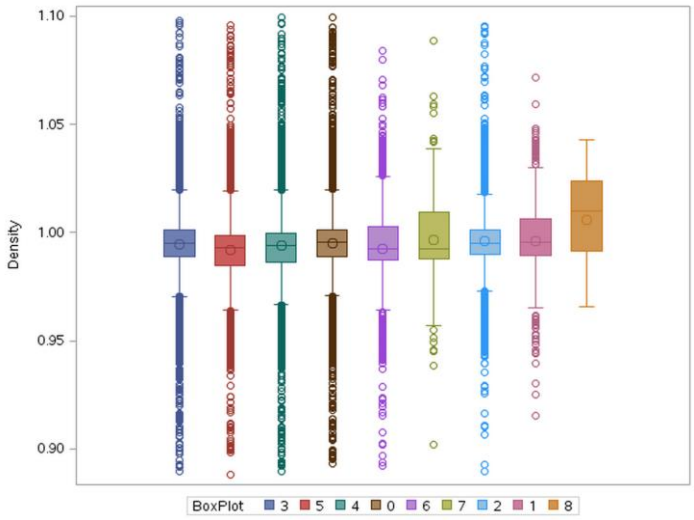
There doesn't appear to be a linear relationship between FreeSulfurDioxide and the number of cases sold. We also learned in Table 3 that Free SO2 tends to be 33-75% of total SO2 and total SO2 legally cannot exceed 350. Clearly we have values in this variable that exceed this limit. We also have negative values that must be addressed as well.

6.8 Boxplot for TotalSulfurDioxide



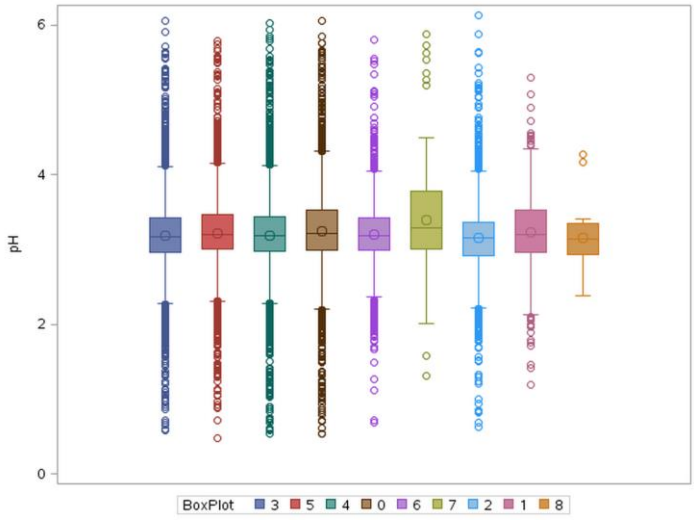
As we mentioned with Free SO2, Total SO2 cannot exceed 350 and yet we see values exceeding 1,000 and negative values as well. Cleaning of this variable is essential if we want to use it as a predictor in our models.

6.9 Boxplot for Density



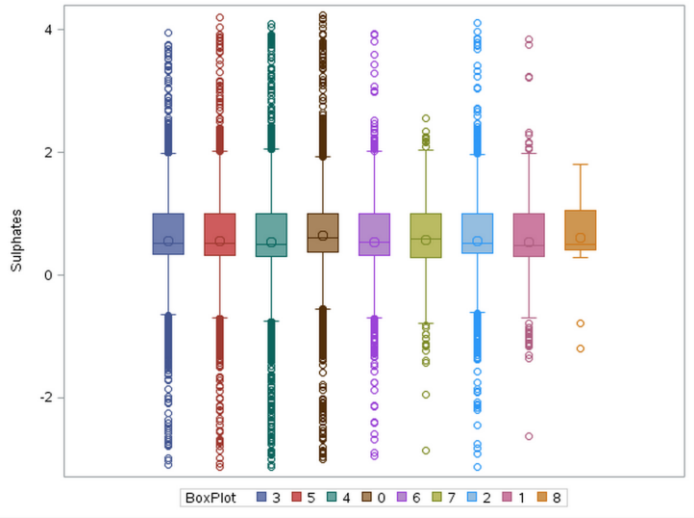
This variable does appear to contain outliers, but is in the acceptable range.

6.10 Boxplot for pH



As we learned in Table 3, pH levels for wine primarily fall between 3 and 4, but we can clearly see values outside of that range. We must address this variable before moving into the model building phase.

6.11 Boxplot for Sulphates



This variable appears to be constant, regardless of the number of cases sold. We do see negative values, which cannot happen, so we must address this if we want to use this as a predictor in our models.

Data Preparation

Address Missing Values

As we have covered in previous sections, there are several variables with missing values that must be addressed before moving into the model building phase. The following list provides you with an overview of the variables and how many missing values exist for each variable.

Results of assessing predictors with missing values

- ResidualSugar: Mean is 5.42 and 616 values are missing
- Chlorides: Mean is 0.05 and 638 values are missing
- FreeSulfurDioxide: Mean is 30 and 647 values are missing
- TotalSulfurDioxide: Mean is 123 and 682 values are missing
- pH: Mean is 3.21 and 395 values are missing
- Sulphates: Mean is 0.53 and 1210 values are missing
- Alcohol: Mean is 10.49 and 653 values are missing
- STARS: Mean is 2.04 and 3359 values are missing

For the variables listed above, we will impute the mean in place of the missing values for all variables with the exception of the STARS variable. For the STARS variable, we will assume that the missing values have not yet been rated yet and will impute these missing values with the value of zero.

Variable Transformations

As we saw with the boxplots in Table 7, there were many variables that had negative values or were outside the appropriate ranges we derived from our preliminary research in Table 3. In Table 8 below, we will go through each variable with messy data and cover the transformations performed to cleanse the data.

Table 8: Variable Transformations for Variables with Outliers

Variable	Transformation
FixedAcidity	Given there were negative values, we imputed the values with the absolute values for this variable. This variable usually does not exceed 14.5, so we chose to treat any variables outside of this range as a missing value and imputed it with the median of 7.
VolatileAcidity	Based on what we saw in Table 3, it is illegal to exceed 1.2 g/L, so we chose to treat any observation exceeding this limit as a missing value and imputed it with the median of 0.41. We also imputed the negative values with the absolute values.
CitricAcid	To address the negative values, we simply imputed using the absolute values and then simply did a trimming at the 95th percentile.
TotalSulfurDioxide	We addressed the negative values by imputing with the absolute values. Also, based on what we learned in Table 3, total SO2 cannot exceed 350 in the US and the average wine is in the 80 mg/L range. We chose to impute any value exceeding 250 with the median of 154.
FreeSulfurDioxide	This variable also contained negative values, so we imputed those values using the absolute values. Also, we know that Free SO2 cannot exceed Total SO2 and usually ranges from 33-75% of the total SO2. For the observations where Free SO2 exceeded Total SO2, we simply took the total SO2 and divided it by 0.54, the median % of free SO2 to total SO2.

ResidualSugar	To address the negative values we chose to impute with the absolute values. We also learned that anything exceeding 25 is suspect and decided to treat them as missing values. We then imputed those values using the mean of 5.42.
Alcohol	We also learned in Table 3 that the maximum alcohol % is around the 23% range, thus indicating bad data that exists with this variable. First, we imputed the negative values using the absolute values. We then imputed anything exceeding 23 with the mean of 10.49.
Chlorides	We first handled the negative values by imputing them with the absolute values. For missing values, we simply imputed with the mean of 0.22.
Sulphates	As with the previous variables, we address the negative values by imputing with the absolute values. We also imputed the missing values with the mean of 0.53.
pH	We imputed the missing values with the mean of 3.21. We also learned in Table 3 that the majority of wine falls between 3 and 4. Anything outside of this range we treated as a missing value and imputed it with the mean of 3.21.
STARS	As we mentioned earlier, we simply imputed any missing value with a zero value, indicating it has not yet been rated.

Model Building

For this assignment, we are dealing with a dependent variable that represents the count of number of cases sold. Given we are dealing with a count variable, we will explore the use of appropriate models for this type of distribution including Poisson, Negative Binomial, Zero-Inflated Poisson, Zero-Inflated Negative Binomial, and an OLS regression model just for good measure. In this section we will provide the output of each of the five models and provide any insights gleaned from each model.

Model A: Poisson Regression

The first model we chose to explore is the Poisson regression model as this model is often used when the dependent variable is a count variable. Given this was the first model we decided to use for this assignment, we went through iterative trial and error to determine which variables were statistically significant for inclusion in the model. As a result we landed on the model shown in Table 9 that excludes the following variables: Sulphates, Chlorides, FreeSulfurDioxide, CitricAcid, and Density. We chose to exclude those variables as they had p-values exceeding 0.05.

Table 9: Poisson Regression Model

Model Information	
Data Set	WORK.WINE_2
Distribution	Poisson
Link Function	Log
Dependent Variable	TARGET

Number of Observations Read	12795
Number of Observations Used	12795

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	13536.6766	1.0602
Scaled Deviance	13E3	13536.6766	1.0602
Pearson Chi-Square	13E3	11189.2756	0.8764
Scaled Pearson X2	13E3	11189.2756	0.8764
Log Likelihood		8857.8221	
Full Log Likelihood		-22739.3492	
AIC (smaller is better)		45532.6983	
AICC (smaller is better)		45532.8168	
BIC (smaller is better)		45734.0322	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.0103	0.4538	0.1209	1.8997	4.96	0.0260
residualsugar1		1	0.0020	0.0010	0.0000	0.0039	3.89	0.0486
totalSulfurDioxide1		1	0.0003	0.0001	0.0002	0.0005	18.08	<.0001
pH		1	-0.0850	0.0227	-0.1294	-0.0405	14.04	0.0002
LabelAppeal	-2	1	-0.7001	0.0424	-0.7833	-0.6169	272.07	<.0001
LabelAppeal	-1	1	-0.4582	0.0250	-0.5072	-0.4092	335.88	<.0001
LabelAppeal	0	1	-0.2692	0.0229	-0.3141	-0.2244	138.62	<.0001
LabelAppeal	1	1	-0.1351	0.0232	-0.1805	-0.0896	33.93	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex	4	1	1.2014	0.5481	0.1271	2.2756	4.80	0.0284
AcidIndex	5	1	1.0889	0.4517	0.2037	1.9742	5.81	0.0159
AcidIndex	6	1	1.1196	0.4478	0.2420	1.9972	6.25	0.0124
AcidIndex	7	1	1.0823	0.4476	0.2052	1.9595	5.85	0.0156
AcidIndex	8	1	1.0505	0.4476	0.1733	1.9277	5.51	0.0189
AcidIndex	9	1	0.9396	0.4478	0.0620	1.8172	4.40	0.0359
AcidIndex	10	1	0.7847	0.4485	-0.0943	1.6637	3.06	0.0802
AcidIndex	11	1	0.4255	0.4510	-0.4585	1.3095	0.89	0.3455
AcidIndex	12	1	0.4119	0.4551	-0.4800	1.3039	0.82	0.3654
AcidIndex	13	1	0.5822	0.4572	-0.3139	1.4783	1.62	0.2029
AcidIndex	14	1	0.4713	0.4663	-0.4426	1.3852	1.02	0.3121
AcidIndex	15	1	0.9132	0.5126	-0.0916	1.9180	3.17	0.0749
AcidIndex	16	1	0.2463	0.6327	-0.9937	1.4863	0.15	0.6970
AcidIndex	17	0	0.0000	0.0000	0.0000	0.0000	.	.
STARS	0	1	-1.3144	0.0243	-1.3621	-1.2668	2923.10	<.0001
STARS	1	1	-0.5592	0.0217	-0.6016	-0.5167	666.51	<.0001
STARS	2	1	-0.2405	0.0199	-0.2795	-0.2015	146.03	<.0001
STARS	3	1	-0.1231	0.0202	-0.1627	-0.0835	37.14	<.0001
STARS	4	0	0.0000	0.0000	0.0000	0.0000	.	.
volatileAcidity1		1	-0.0529	0.0186	-0.0893	-0.0165	8.11	0.0044
alcohol1		1	0.0046	0.0015	0.0017	0.0075	9.67	0.0019
Scale		0	1.0000	0.0000	1.0000	1.0000		

As we can see in Table 9, the ratio of deviance to degrees of freedom is fairly close to one, indicating dispersion to be a non-issue. This gives us a good confidence that this model fits the data well. Some of the other important model fit statistics to note include the AIC and BIC being at 45532.6983 and 45734.0322, respectively.

An assessment of the coefficients can be done by leveraging the following formula: 100 * (exp(*B*) – 1), which yields the percentage change for every unit increase of each variable. For example, if we look at VolatileAcidity, for every unit increase in VolatileAcidity, results in a 41% decrease of number of cases sold. Similarly, every unit increase of pH results in a decrease of 8% of number of cases sold. This indicates that a higher pH level has a negative correlation with number of cases sold. This makes intuitive sense because the higher the pH the less acidic the wine becomes, thus becoming flat. Another observation is for the LabelAppeal of -2, indicating an unfavorable rating for the Label. For this dummy variable, we see that it results in a 50% decrease in number of cases sold for wines with a rating of -2 versus those that do not, holding all other variables constant. Again, this makes intuitive sense and would support our notion that an appealing label will aid in the increase of number of cases sold.

Model B: Negative Binomial Regression Model

The next useful model for count dependent variables is the negative binomial model. Table 10 contains the output of a negative binomial model leveraging the same predictor variables as Model A.

Table 10: Negative Binomial Regression Model

Model Information	
Data Set	WORK.WINE_2
Distribution	Negative Binomial
Link Function	Log
Dependent Variable	TARGET

Number of Observations Read	12795
Number of Observations Used	12795

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	13E3	13536.6766	1.0602
Scaled Deviance	13E3	13536.6766	1.0602
Pearson Chi-Square	13E3	11189.2667	0.8764
Scaled Pearson X2	13E3	11189.2667	0.8764
Log Likelihood		8857.8221	
Full Log Likelihood		-22739.3492	
AIC (smaller is better)		45534.6983	
AICC (smaller is better)		45534.8256	
BIC (smaller is better)		45743.4890	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.0103	0.4538	0.1209	1.8997	4.96	0.0260
residualsugar1		1	0.0020	0.0010	0.0000	0.0039	3.89	0.0486
totalSulfurDioxide1		1	0.0003	0.0001	0.0002	0.0005	18.08	<.0001
pH		1	-0.0850	0.0227	-0.1294	-0.0405	14.04	0.0002
LabelAppeal	-2	1	-0.7001	0.0424	-0.7833	-0.6169	272.07	<.0001
LabelAppeal	-1	1	-0.4582	0.0250	-0.5072	-0.4092	335.88	<.0001
LabelAppeal	0	1	-0.2692	0.0229	-0.3141	-0.2244	138.62	<.0001
LabelAppeal	1	1	-0.1351	0.0232	-0.1805	-0.0896	33.93	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex	4	1	1.2014	0.5481	0.1271	2.2756	4.80	0.0284
AcidIndex	5	1	1.0889	0.4517	0.2037	1.9742	5.81	0.0159
AcidIndex	6	1	1.1196	0.4478	0.2420	1.9972	6.25	0.0124
AcidIndex	7	1	1.0823	0.4476	0.2052	1.9595	5.85	0.0156
AcidIndex	8	1	1.0505	0.4476	0.1733	1.9277	5.51	0.0189
AcidIndex	9	1	0.9396	0.4478	0.0620	1.8172	4.40	0.0359
AcidIndex	10	1	0.7847	0.4485	-0.0943	1.6637	3.06	0.0802
AcidIndex	11	1	0.4255	0.4510	-0.4585	1.3095	0.89	0.3455
AcidIndex	12	1	0.4119	0.4551	-0.4800	1.3039	0.82	0.3654
AcidIndex	13	1	0.5822	0.4572	-0.3139	1.4783	1.62	0.2029
AcidIndex	14	1	0.4713	0.4663	-0.4426	1.3852	1.02	0.3121
AcidIndex	15	1	0.9132	0.5126	-0.0916	1.9180	3.17	0.0749
AcidIndex	16	1	0.2463	0.6327	-0.9937	1.4863	0.15	0.6970
AcidIndex	17	0	0.0000	0.0000	0.0000	0.0000	.	.
STARS	0	1	-1.3144	0.0243	-1.3621	-1.2668	2923.09	<.0001
STARS	1	1	-0.5592	0.0217	-0.6016	-0.5167	666.51	<.0001
STARS	2	1	-0.2405	0.0199	-0.2795	-0.2015	146.03	<.0001
STARS	3	1	-0.1231	0.0202	-0.1627	-0.0835	37.14	<.0001
STARS	4	0	0.0000	0.0000	0.0000	0.0000	.	.
volatileAcidity1		1	-0.0529	0.0186	-0.0893	-0.0165	8.11	0.0044
alcohol1		1	0.0046	0.0015	0.0017	0.0075	9.67	0.0019
Dispersion		1	0.0000	0.0001	0.0000	8.01E123		

The AIC and BIC for Model B is very similar to that of Model A coming in at 45534.6983 and 45743.4890. There seems to be no discernable difference between both models. We also can see that the ratio of deviance to degrees of freedom is exactly the same, indication dispersion is not an issue with the model. Based on the Chapter 3 of the Allison SAS textbook, a negative binomial model is almost always better than the Poisson model as it allows for some overdispersion

in the model. However, in this particular case, overdispersion does not appear to be an issue. This makes sense as the resulting AIC and BIC are very similar between models A and B.

Model C: Zero-Inflated Poisson Regression

As we discussed earlier in the data exploration phase, the only variable that contains a significant number of zero values is LabelAppeal, roughly 44% of the observations. Due to this fact, we chose to use the LabelAppeal variable in our zeromodel for both of the zero-inflated models we will be using for this analysis.

Table 11: Zero-Inflated Poisson Regression Model

Model Information	
Data Set	WORK.WINE_2
Distribution	Zero Inflated Poisson
Link Function	Log
Dependent Variable	TARGET

Number of Observations Read	12795
Number of Observations Used	12795

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		45047.8716	
Scaled Deviance		45047.8716	
Pearson Chi-Square	13E3	9153.0889	0.7172
Scaled Pearson X2	13E3	9153.0889	0.7172
Log Likelihood		9073.2355	
Full Log Likelihood		-22523.9358	
AIC (smaller is better)		45111.8716	
AICC (smaller is better)		45112.0371	
BIC (smaller is better)		45350.4895	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.9388	0.4694	0.0188	1.8588	4.00	0.0455
residualsugar1		1	0.0012	0.0010	-0.0008	0.0032	1.47	0.2249
totalSulfurDioxide1		1	0.0003	0.0001	0.0001	0.0004	10.49	0.0012
pH		1	-0.0628	0.0232	-0.1083	-0.0172	7.29	0.0069
LabelAppeal	-2	1	-0.8828	0.0437	-0.9683	-0.7972	408.91	<.0001
LabelAppeal	-1	1	-0.5863	0.0269	-0.6390	-0.5337	476.65	<.0001
LabelAppeal	0	1	-0.3387	0.0240	-0.3858	-0.2917	199.35	<.0001
LabelAppeal	1	1	-0.1694	0.0242	-0.2168	-0.1220	49.10	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex	4	1	1.2180	0.5672	0.1064	2.3296	4.61	0.0318
AcidIndex	5	1	1.1186	0.4672	0.2029	2.0342	5.73	0.0167
AcidIndex	6	1	1.1516	0.4633	0.2435	2.0596	6.18	0.0129
AcidIndex	7	1	1.1182	0.4631	0.2106	2.0258	5.83	0.0157
AcidIndex	8	1	1.0959	0.4631	0.1883	2.0035	5.60	0.0180
AcidIndex	9	1	0.9968	0.4633	0.0887	1.9048	4.63	0.0314
AcidIndex	10	1	0.8439	0.4641	-0.0657	1.7535	3.31	0.0690
AcidIndex	11	1	0.4855	0.4669	-0.4296	1.4006	1.08	0.2984
AcidIndex	12	1	0.4728	0.4715	-0.4513	1.3969	1.01	0.3160
AcidIndex	13	1	0.7142	0.4749	-0.2166	1.6451	2.26	0.1326
AcidIndex	14	1	0.5896	0.4866	-0.3642	1.5434	1.47	0.2257
AcidIndex	15	1	1.0195	0.5406	-0.0401	2.0791	3.56	0.0593
AcidIndex	16	1	0.3272	0.6665	-0.9791	1.6335	0.24	0.6235
AcidIndex	17	0	0.0000	0.0000	0.0000	0.0000	.	.
STARS	0	1	-1.0972	0.0292	-1.1545	-1.0400	1409.41	<.0001
STARS	1	1	-0.3960	0.0229	-0.4408	-0.3512	300.22	<.0001
STARS	2	1	-0.1984	0.0200	-0.2376	-0.1592	98.33	<.0001
STARS	3	1	-0.1084	0.0202	-0.1481	-0.0688	28.73	<.0001
STARS	4	0	0.0000	0.0000	0.0000	0.0000	.	.
volatileAcidity1		1	-0.0459	0.0190	-0.0830	-0.0087	5.85	0.0156
alcohol1		1	0.0054	0.0015	0.0024	0.0084	12.71	0.0004
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	-1.7958	0.1643	-2.1178	-1.4737	119.45	<.0001
LabelAppeal	-2	1	-20.3461	5381.427	-10567.8	10527.06	0.00	0.9970
LabelAppeal	-1	1	-1.9140	0.3843	-2.6673	-1.1607	24.80	<.0001
LabelAppeal	0	1	-0.7382	0.1849	-1.1006	-0.3757	15.94	<.0001
LabelAppeal	1	1	-0.3650	0.1843	-0.7262	-0.0038	3.92	0.0476
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.

Table 11 provides the output of the zero-inflated Poisson model leveraging the same values as the previous models. We can see the ratio of the chi-square to degrees of freedom is less than one, indication a good fit to the data. We also notice an improvement in both AIC and BIC compared with the first two models as well coming in at 45111.8716 and

45350.48495. Residual Sugar has a p-value exceeding 0.05 and its coefficient close to zero, indicating it is not very significant in the model. An assessment of the coefficients is also in line with what we saw with the first two models.

Model D: Zero-Inflated Negative Binomial Regression

As with the previous zero-inflated model, we chose to leverage the LabelAppeal in our zeromodel given the high frequency of zeros in the dataset for that particular variable. Table 12 below shows the output of the zero-inflated negative binomial model.

Table 12: Zero-Inflated Negative Binomial Regression Model

Model Information	
Data Set	WORK.WINE_2
Distribution	Zero Inflated Negative Binomial
Link Function	Log
Dependent Variable	TARGET

Number of Observations Read	12795
Number of Observations Used	12795

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance		45047.8699	
Scaled Deviance		45047.8699	
Pearson Chi-Square	13E3	9153.0816	0.7172
Scaled Pearson X2	13E3	9153.0816	0.7172
Log Likelihood		-22523.9349	
Full Log Likelihood		-22523.9349	
AIC (smaller is better)		45113.8699	
AICC (smaller is better)		45114.0457	
BIC (smaller is better)		45359.9446	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.9388	0.4694	0.0188	1.8588	4.00	0.0455
residualsugar1		1	0.0012	0.0010	-0.0008	0.0032	1.47	0.2249
totalSulfurDioxide1		1	0.0003	0.0001	0.0001	0.0004	10.49	0.0012
pH		1	-0.0628	0.0232	-0.1083	-0.0172	7.29	0.0069
LabelAppeal	-2	1	-0.8828	0.0437	-0.9683	-0.7972	408.91	<.0001
LabelAppeal	-1	1	-0.5863	0.0269	-0.6390	-0.5337	476.65	<.0001
LabelAppeal	0	1	-0.3387	0.0240	-0.3858	-0.2917	199.35	<.0001
LabelAppeal	1	1	-0.1694	0.0242	-0.2168	-0.1220	49.10	<.0001
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.
AcidIndex	4	1	1.2180	0.5672	0.1064	2.3296	4.61	0.0318
AcidIndex	5	1	1.1186	0.4672	0.2029	2.0342	5.73	0.0167
AcidIndex	6	1	1.1516	0.4633	0.2435	2.0596	6.18	0.0129
AcidIndex	7	1	1.1182	0.4631	0.2106	2.0258	5.83	0.0157
AcidIndex	8	1	1.0959	0.4631	0.1883	2.0035	5.60	0.0180
AcidIndex	9	1	0.9968	0.4633	0.0887	1.9048	4.63	0.0314
AcidIndex	10	1	0.8439	0.4641	-0.0657	1.7535	3.31	0.0690
AcidIndex	11	1	0.4855	0.4669	-0.4296	1.4006	1.08	0.2984
AcidIndex	12	1	0.4728	0.4715	-0.4513	1.3969	1.01	0.3160
AcidIndex	13	1	0.7142	0.4749	-0.2166	1.6451	2.26	0.1326
AcidIndex	14	1	0.5896	0.4866	-0.3642	1.5434	1.47	0.2257
AcidIndex	15	1	1.0195	0.5406	-0.0401	2.0791	3.56	0.0593
AcidIndex	16	1	0.3272	0.6665	-0.9791	1.6335	0.24	0.6235
AcidIndex	17	0	0.0000	0.0000	0.0000	0.0000	.	.
STARS	0	1	-1.0972	0.0292	-1.1545	-1.0400	1409.41	<.0001
STARS	1	1	-0.3960	0.0229	-0.4408	-0.3512	300.22	<.0001
STARS	2	1	-0.1984	0.0200	-0.2376	-0.1592	98.33	<.0001
STARS	3	1	-0.1084	0.0202	-0.1481	-0.0688	28.73	<.0001
STARS	4	0	0.0000	0.0000	0.0000	0.0000	.	.
volatileAcidity1		1	-0.0459	0.0190	-0.0830	-0.0087	5.85	0.0156
alcohol1		1	0.0054	0.0015	0.0024	0.0084	12.71	0.0004
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	-1.7958	0.1643	-2.1178	-1.4737	119.45	<.0001
LabelAppeal	-2	1	-11.3117	58.7623	-126.484	103.8602	0.04	0.8474
LabelAppeal	-1	1	-1.9140	0.3843	-2.6673	-1.1607	24.80	<.0001
LabelAppeal	0	1	-0.7382	0.1849	-1.1006	-0.3757	15.94	<.0001
LabelAppeal	1	1	-0.3650	0.1843	-0.7262	-0.0038	3.92	0.0476
LabelAppeal	2	0	0.0000	0.0000	0.0000	0.0000	.	.

Model D has almost identical output of the zero-inflated Poisson model, which makes intuitive sense given the fact overdispersion is not a major issue with these models. The chi-square to degrees of freedom ratio is below zero, indicating good model fit and the AIC and BIC are very similar to the previous model with 45113.8699 and 45359.9446,

respectively. Lastly, as with the previous model, Residual Sugar is no longer significant with the model given its p-value exceeds 0.05. We also see a very consistent trend with the coefficient values between both Models C and D as well in that they make intuitive sense based on what the data dictionary provided as a theoretical effect.

Model E: Ordinary Least Squares Regression

The last model we chose to use for this assignment is the ordinary least squares regression model. We know an OLS regression model is not ideal for count variables because it violates some of the statistical assumptions of OLS regression. Table 13 contains the output of the OLS regression model we built for the wine dataset.

Table 13: OLS Regression Model

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	25721	1513.02081	888.58	<.0001
Error	12777	21756	1.70273		
Corrected Total	12794	47477			

Root MSE	1.30489	R-Square	0.5418
Dependent Mean	3.02907	Adj R-Sq	0.5412
Coeff Var	43.07879		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	5.83422	0.19059	30.61	<.0001	0
residualsugar1	1	0.00572	0.00228	2.51	0.0122	1.01221
totalSulfurDioxide1	1	0.00113	0.00018470	6.14	<.0001	1.01946
pH	1	-0.26433	0.05056	-5.23	<.0001	1.01156
alcohol1	1	0.01447	0.00336	4.31	<.0001	1.01073
label_neg2	1	-1.89001	0.08428	-22.42	<.0001	2.01990
label_neg1	1	-1.52118	0.06469	-23.51	<.0001	5.81840
label_1	1	-0.59037	0.06369	-9.27	<.0001	5.53084
label_0	1	-1.05441	0.06217	-16.96	<.0001	7.15300
acidindex45	1	0.97674	0.15356	6.36	<.0001	1.07362
acidindex6	1	1.05499	0.05588	18.88	<.0001	1.98956
acidindex7	1	0.94134	0.04518	20.83	<.0001	3.61860
acidindex8	1	0.83087	0.04560	18.22	<.0001	3.42037
acidindex9	1	0.52533	0.05313	9.89	<.0001	2.10221
stars0	1	-3.64919	0.05917	-61.67	<.0001	5.09333
stars1	1	-2.29690	0.05968	-38.49	<.0001	4.85009
stars2	1	-1.25747	0.05814	-21.63	<.0001	5.10923
stars3	1	-0.69862	0.05996	-11.65	<.0001	3.86248

As you can see in Table 13, the R-square is 0.54 and the P-value for the Overall F-test is below 0.05. We can also see that all of the p-values for the coefficients are below 0.05 as well. We chose to include the variance inflation factor values to see if any multicollinearity existed among the variables, however, they are all below 9 indicating multicollinearity is not a problem with this model. The analysis of coefficients is also consistent with what we saw in all of the previous models as well.

Model Selection

After reviewing the output for each of the models in the previous section, we now must decide on our method for final model selection. In the section below, we will consolidate all of the model fit statistics and other diagnoses used to determine which model is the superior model for the data we are working with.

The criteria we will use to assess model performance are as follows:

- Lowest AIC
- Lowest BIC
- Deviance to Degrees of Freedom ratio close to 1
- Chi-Square to Degrees of Freedom ratio below 1
- Intuitive values for coefficients

Table 12: Model Selection Criteria

	Poisson	Negative Binomial	Zero-Inflated Poisson	Zero-Inflated Negative Binomial
AIC	45532.6983	45534.6983	45111.8716	45113.8699
BIC	45734.0322	45743.4890	45350.4895	45359.9446
Deviance to DF	1.0602	1.0602	N/A	N/A
Chi-Square to DF	0.8764	0.8764	0.7172	0.7172
Intuitive Coefficients	Yes	Yes	Yes	Yes

We chose to exclude the OLS regression model for consideration in our final model given the statistical assumptions violated when dealing with a count variable as the dependent variable. However, the OLS regression model as useful in determining no multicollinearity exists among the predictor variables.

Based on the criteria we highlighted previously, the model with the best performance appears to be the zero-inflated Poisson (ZIP) model. The basis for this determination is due to the ZIP model having the lowest AIC and BIC and the chi-square to degrees of freedom ratio being below one. We also chose this model as it aided us in providing special treatment for the high frequency of zero values in the LabelAppeal variable. The zero-inflated negative binomial model was very similar to the ZIP model, but overdispersion was not an issue so going with the ZIP model made the most intuitive sense.

Conclusion

The final step in this assignment was to create a data step to handle out-of-sample test data that can be ran through our final ZIP model for predicting the number of cases sold. We constructed the necessary variable transformations, handling of missing values, and trimming that ensures no observations are removed during the data processing step for new observations.

As we discussed during the introduction of this assignment, the primary objective was to leverage the chemical properties of wines to estimate the number of cases sold. Our final model will aid the wine manufacturer to know which chemical properties will increase the likelihood of selling more cases of wine in high-end restaurants. The learnings from the data exploration and model building phases provided invaluable insight we can use to optimize the sales of wine in the future. We can continue to build and refine the model as we get more wine sales data in the future, which will only aid us in the ability to increase predictive accuracy as we go along.