Wine Sales Predictions

Introduction

In this document we explore, analyze and model a data set containing information on approximately 13,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response 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 throughout the country. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. *The* objective is to build a regression model to predict the number of cases of wine that will be sold given certain properties of the wine.

Data Exploration

The data set we are exploring contains empirical and qualitative data related to samples of commercially available wines. The data set contains 12795 observations, each representing chemical, marketing and rating information for a specific wine. For each wine we are provided with 14 attributes that could potentially be used as predictor variables and one response variable, TARGET, which indicates the total number of sample cases of wine purchased by wine resellers subsequent to their sampling of an individual wine. As such, the response variable serves as a useful indicator of a wine's potential future sales volume. All the features included in the dataset are part of wine composition. Figure 1 shows the influence of these components in the wine. Although dataset is missing some features, we will leverage the available ones.

Figure 1: Wine Composition

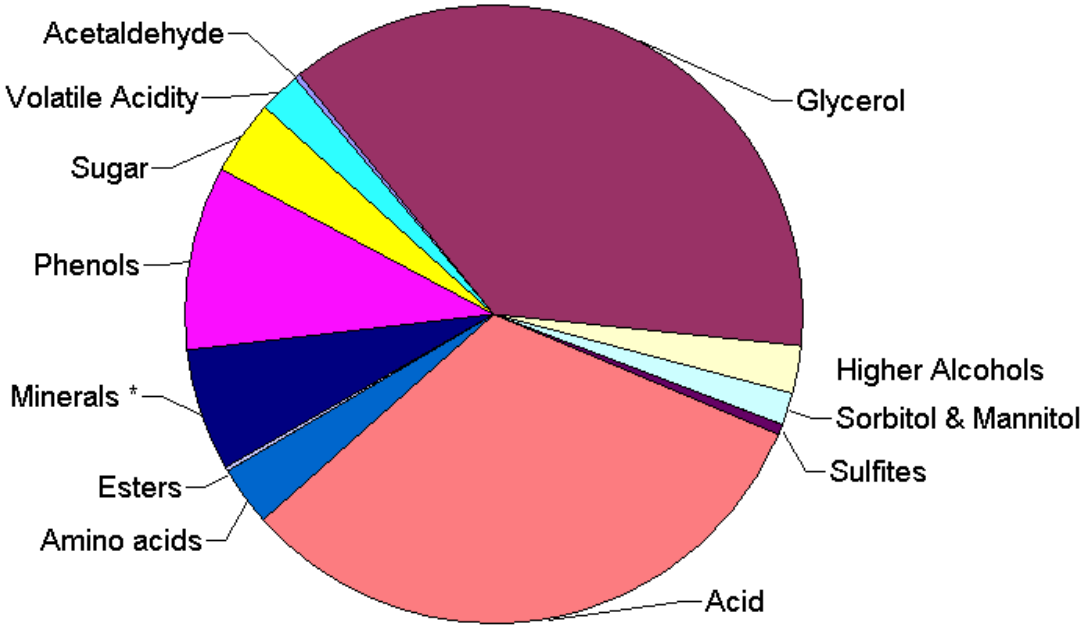


Table 1: Data Dictionary with Theoretical Effect

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
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 customes 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 conten of wine	
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	
VolatileAcidity	Volatile Acid content of wine	
pH	pH of wine	

Table 1 shows the data dictionary for each of the features included in the dataset. It also lists their theoretical effects on the response variable. It is obvious that the 'stars' is the highly influential feature within the dataset as high number of stars suggests high sales. Same applies to labelappel as well. Many consumers purchase based on the visual of the wine label.

Next we will look at the statistics summary of the dataset. Some of the variables are showing high scales compared to most of the other variables in the dataset, so we may have to transform these variables during the preparation phase. The 'labelappeal', 'acidindex' and the 'stars' variables appear to be categorical data and the rest of the variables in the dataset are continuous.

Table 2: Data Statistics

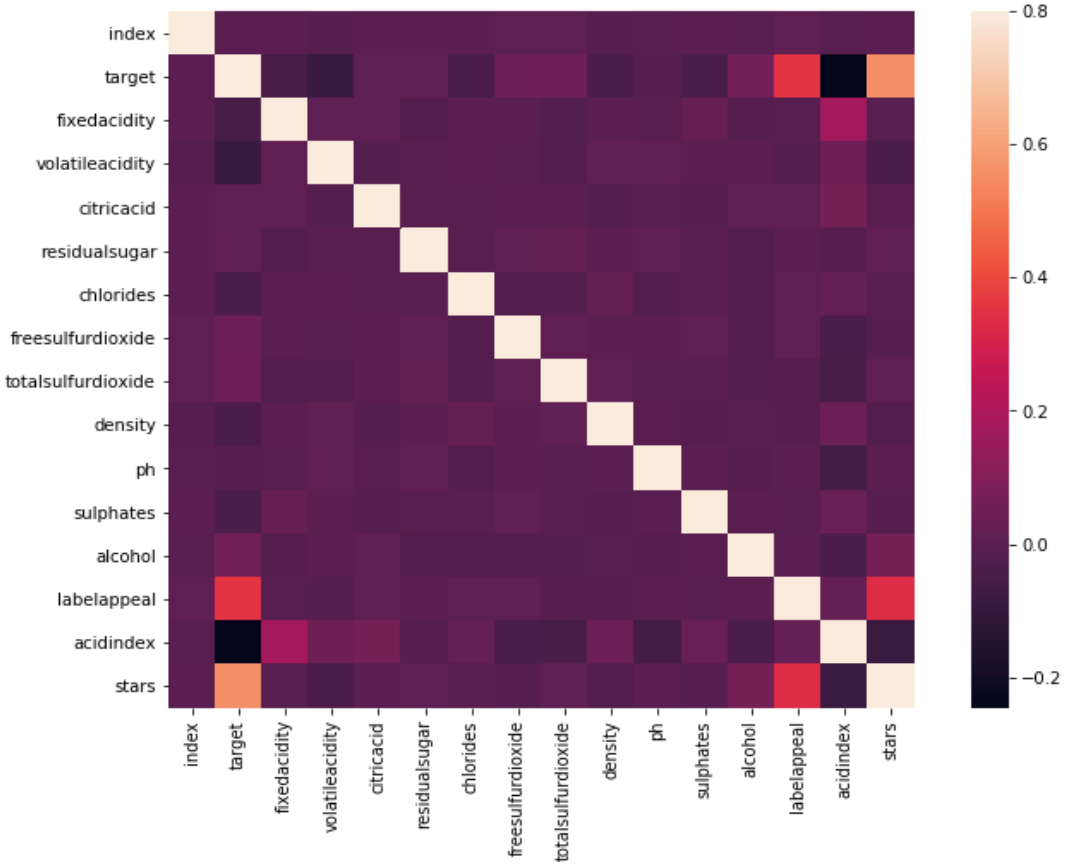
	count	mean	std	min	25%	50%	75%	max
index	12795.0	8069.980305	4656.905107	1.00000	4037.50000	8110.00000	12106.500000	16129.00000
target	12795.0	3.029074	1.926368	0.00000	2.00000	3.00000	4.000000	8.00000
fixedacidity	12795.0	7.075717	6.317643	-18.10000	5.20000	6.90000	9.500000	34.40000
volatileacidity	12795.0	0.324104	0.784014	-2.79000	0.13000	0.28000	0.640000	3.68000
citricacid	12795.0	0.308413	0.862080	-3.24000	0.03000	0.31000	0.580000	3.86000
residualsugar	12179.0	5.418733	33.749379	-127.80000	-2.00000	3.90000	15.900000	141.15000
chlorides	12157.0	0.054822	0.318467	-1.17100	-0.03100	0.04600	0.153000	1.35100
freesulfurdioxide	12148.0	30.845571	148.714558	-555.00000	0.00000	30.00000	70.000000	623.00000
totalsulfurdioxide	12113.0	120.714233	231.913211	-823.00000	27.00000	123.00000	208.000000	1057.00000
density	12795.0	0.994203	0.026538	0.88809	0.98772	0.99449	1.000515	1.09924
ph	12400.0	3.207628	0.679687	0.48000	2.96000	3.20000	3.470000	6.13000
sulphates	11585.0	0.527112	0.932129	-3.13000	0.28000	0.50000	0.860000	4.24000
alcohol	12142.0	10.489236	3.727819	-4.70000	9.00000	10.40000	12.400000	26.50000
labelappeal	12795.0	-0.009066	0.891089	-2.00000	-1.00000	0.00000	1.000000	2.00000
acidindex	12795.0	7.772724	1.323926	4.00000	7.00000	8.00000	8.000000	17.00000
stars	9436.0	2.041755	0.902540	1.00000	1.00000	2.00000	3.000000	4.00000

Next, we will look at the missing values within the dataset, shown in Table 3. Out of 14 variables, majority of them missing the data (Chlorides, ResidualSugar, FreeSulfurDioxide, CitricAcid, VolatileAcidity, TotalSulfur, DioxideSulphates, FixedAcidity and Alcohol). Also data has the lot of negative values which does not make sense for most of the variables, except labelappeal which ranges -2 to 2. We will impute these variables during the transformation and may have to convert most of the negative values to the positive. Also, we notice that some of the variables, such as FreeSufurDioxide and TotalSulfurDioxide, have very high and very low values. These extreme values might be indicative of outliers, which is something we will have to look for when we examine distribution plots. Figure 2 shows the correlation matrix for the dataset.

Table 3: Missing Values

index	0
target	0
fixedacidity	0
volatileacidity	0
citricacid	0
residualsugar	616
chlorides	638
freesulfurdioxide	647
totalsulfurdioxide	682
density	0
ph	395
sulphates	1210
alcohol	653
labelappeal	0
acidindex	0
stars	3359
dtype: int64	

Figure 2: Correlation Matrix



Let's at each chemical component that is in the dataset to understand their influence on the wine taste.

Fixed Acidity

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 usually "flat" in taste. Chemically the acids influence the color, stability to oxidation, and the overall lifespan of a wine. The acids may arise in the grapes themselves and carry over into wines or they may arise from the fermentation process. There are two types of acidity, volatile acidity or fixed acidity. The predominant fixed acids found in wines are tartaric, malic, citric, and succinic.

Figure 3: Fixed Acidity

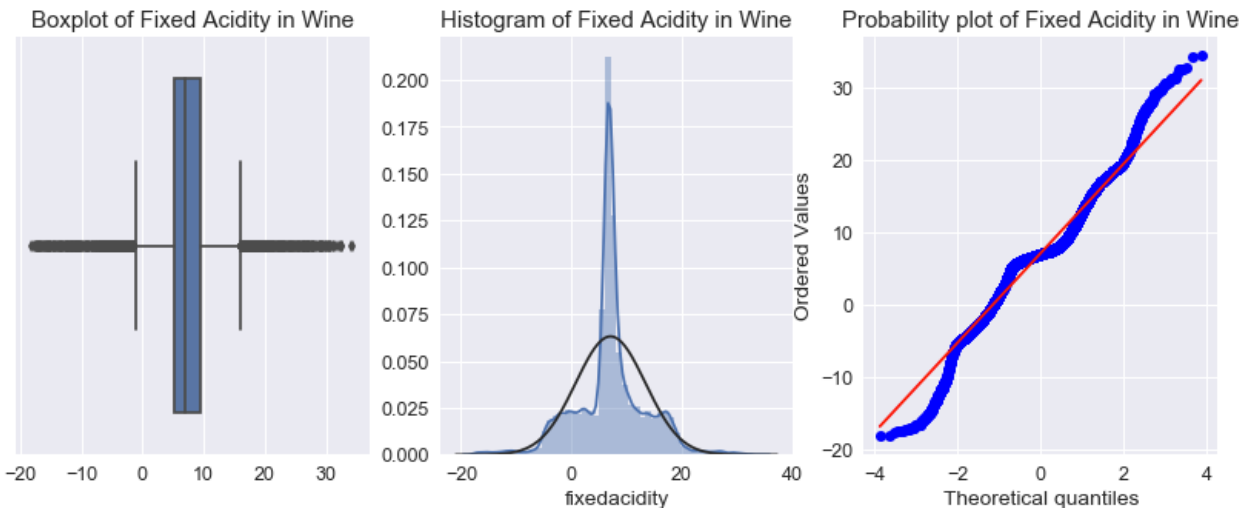
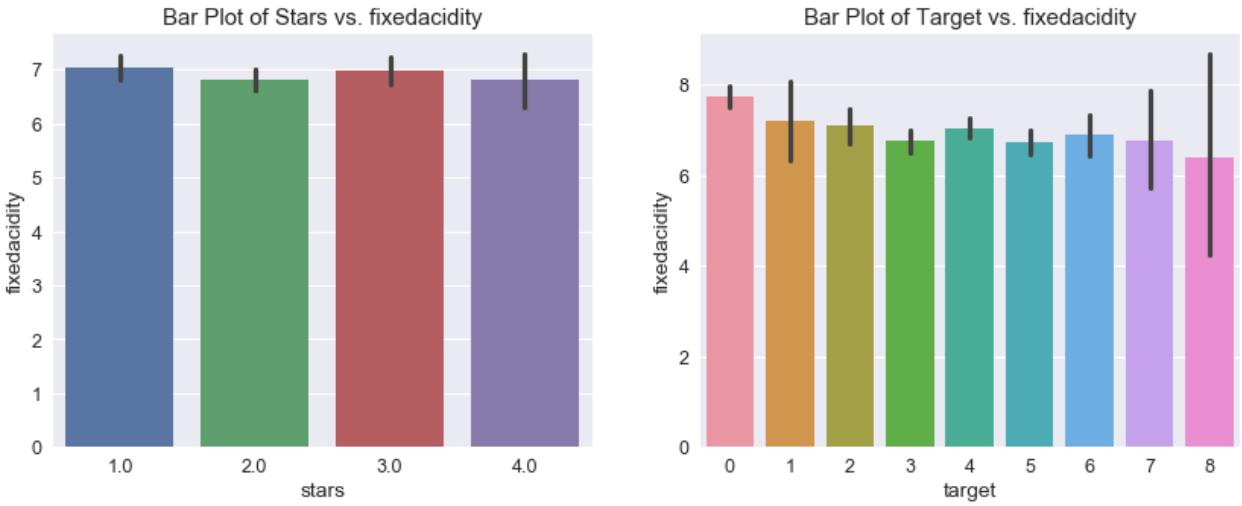


Figure 2 shows the data distribution for the '***fixedacidity'*** show the high peak that is centered in the distribution. Also most of the data seems to fall outside of the quartile range. Figure 3 shows the influence of fixedacidity against the stars, which essentially rates the taste of the wine. We can deduct that the influence on the stars seems to be low and consistent as most of the wines try to include some sort of fixed acidity to influence the taste. It just shows that even the wines that receive one star tries to make them not flat.

Figure 4: Fixed Acidity vs. Stars & Target



Volatile Acidity

Volatile acidity refers to the steam distillable acids present in wine, primarily acetic acid but also lactic, formic, butyric, and propionic acids. Commonly, these acids are measured by Cash Still, though now they can be measured by gas chromatography, HPLC or enzymatic methods. 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 3g/L. 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. Figure 5 shows the distribution of the volatile acidity, the spike is probably related to the number of outliers within the data.

Figure 5: Volatile Acidity

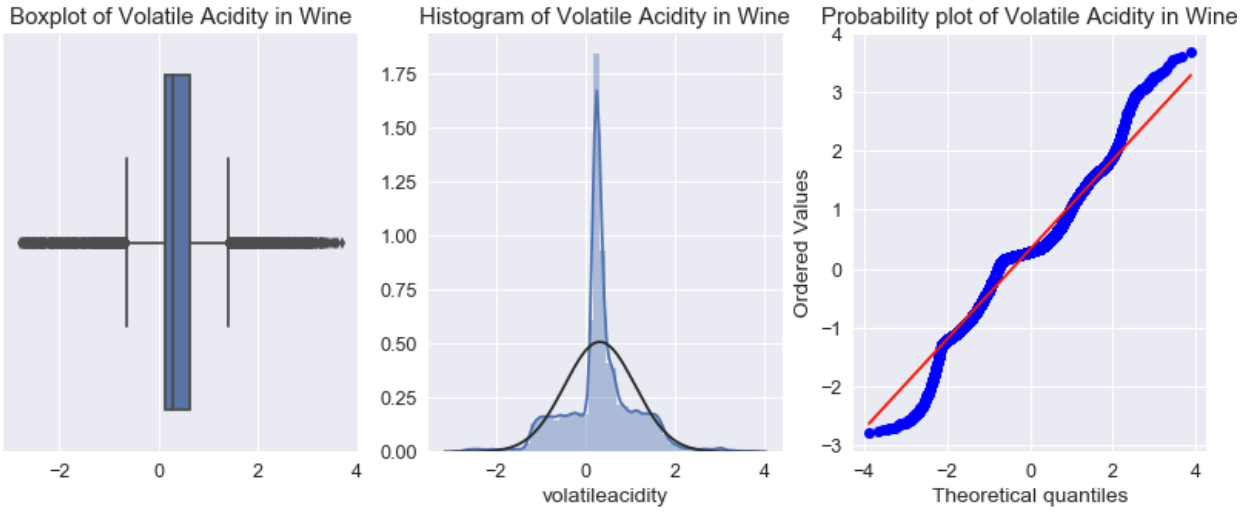
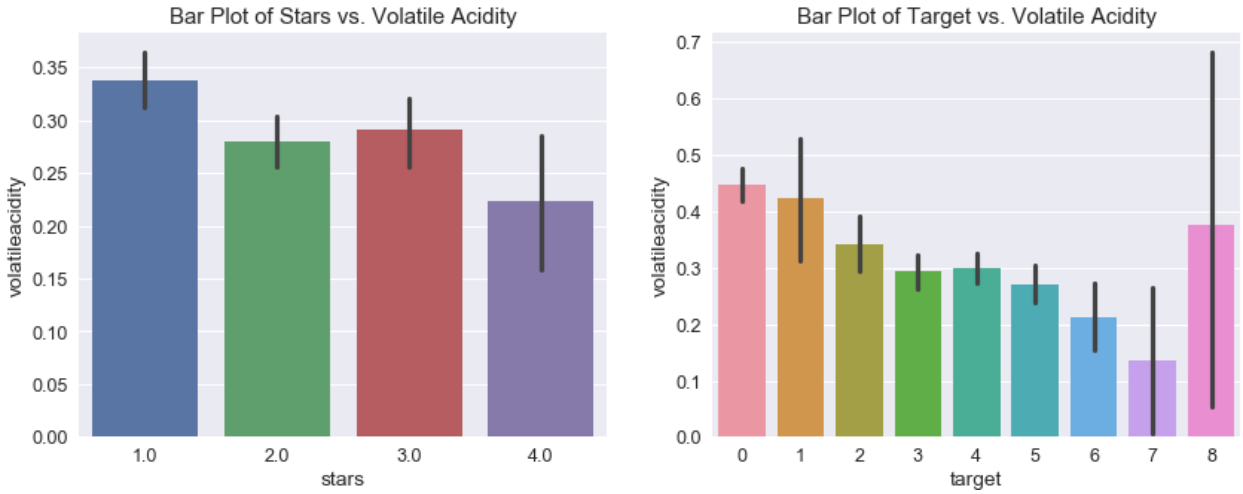


Figure 6 shows the box and bar plot for the 'volatileacidity' and 'stars'. As it shows the star rating seems to be going down as the volatile acidity goes down. Having negligible amount of acetic acid (type of volatile acid) will increase the complexity of the wine.

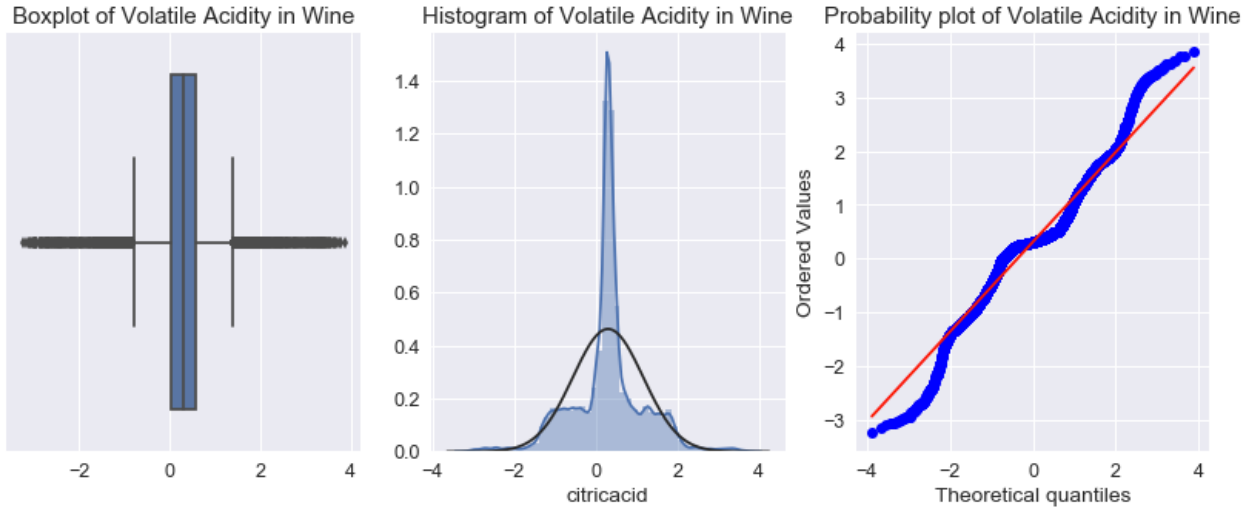
Figure 6: Volatile Acidity vs. Stars & Target



Citric Acid

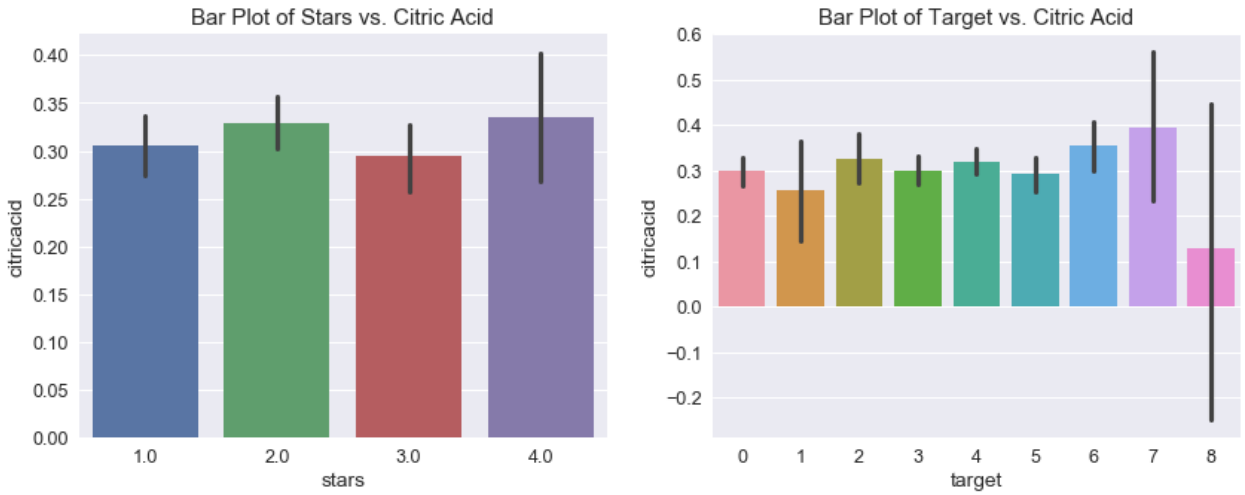
This is a type of fixed acid within the wine. The predominant fixed acids found in wines are tartaric, malic, citric, and succinic. Their respective levels found in wine can vary greatly but in general one would expect to see 1,000 to 4,000 mg/L tartaric acid, 0 to 8,000 mg/L malic acid, 0 to 500 mg/L citric acid, and 500 to 2,000 mg/L succinic acid. All of these acids originate in grapes with the exception of succinic acid, which is produced by yeast during the fermentation process. Grapes also contain ascorbic acid (Vitamin C), but this is lost during fermentation. It is also legal to add fumaric acid as a preservative.

Figure 7: Citric Acid



Like other variables in the dataset, citric acid has way too many negative values and the outliers also skewing the distribution. Both box and bar plots show that there is not much effect on the stars variable, at least not in predictable way.

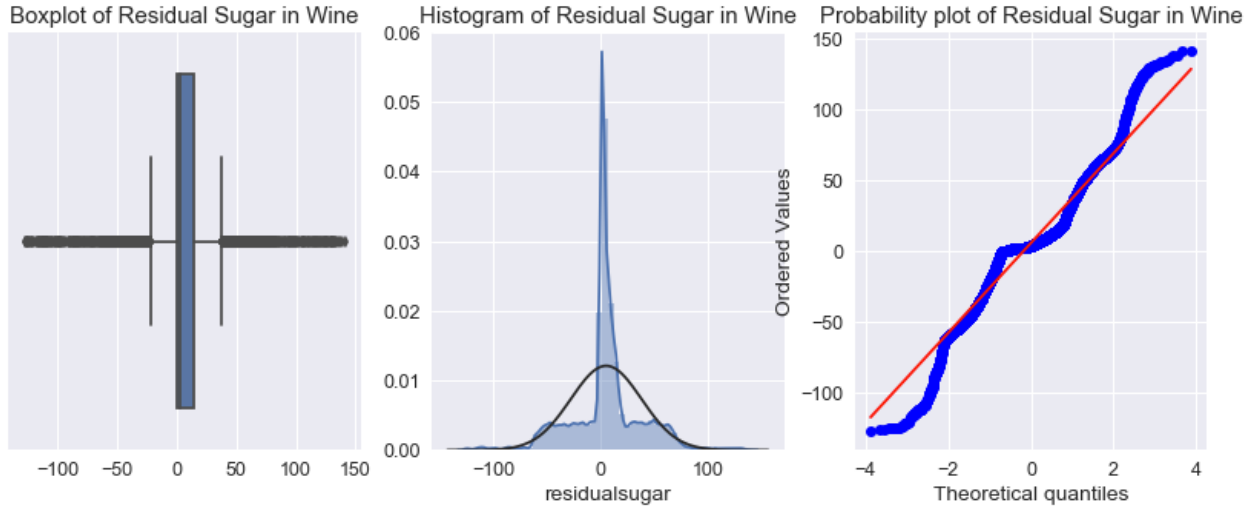
Figure 8: Citric Acid vs. Stars & Target



Residual Sugar

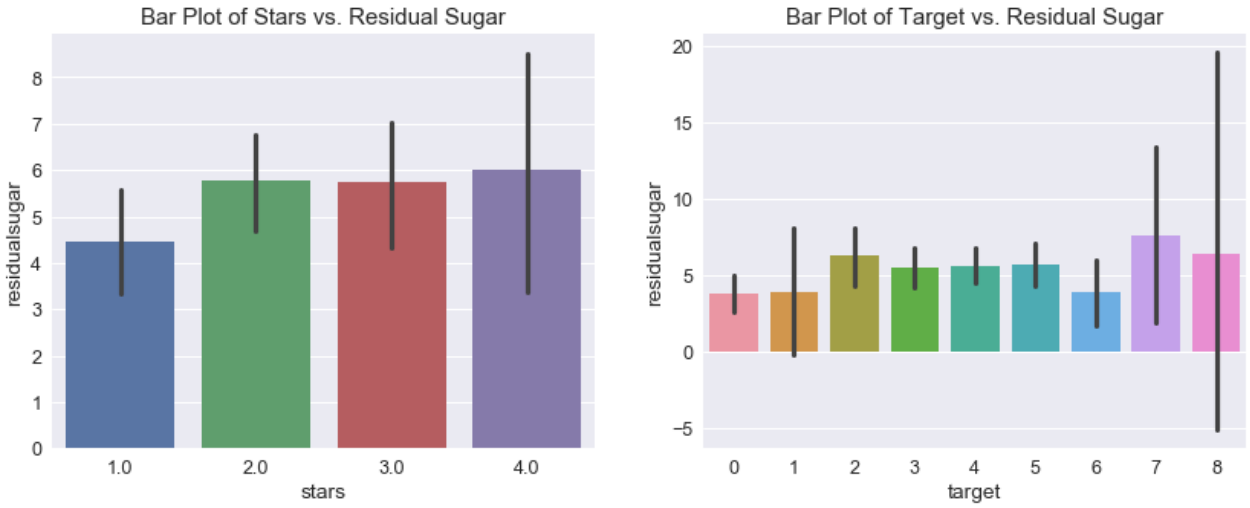
Residual sugar is the sugar from the grapes that's left over after fermentation; more residual sugar makes a sweet wine, and the absence of residual sugar makes a dry wine. The term "dry", in reference to fermentation technically means that there is less than 1% residual sugar in the wine. During fermentation the yeast consumes the sugar in the grape juice producing alcohol and CO2. The yeast will continue this process until all of the grape sugar has been used up at which point having no food source, the yeast cells die and become the lees. Wine is fermented to dryness because, among other things, leaving sugar in it would make it microbially unstable.

Figure 9: Residual Sugar



As shown in both figure 9 and 10, the data seems to be all over the place. As seen in bar plot, this variable seems to have an impact on the taste to some extent and then evens out after that.

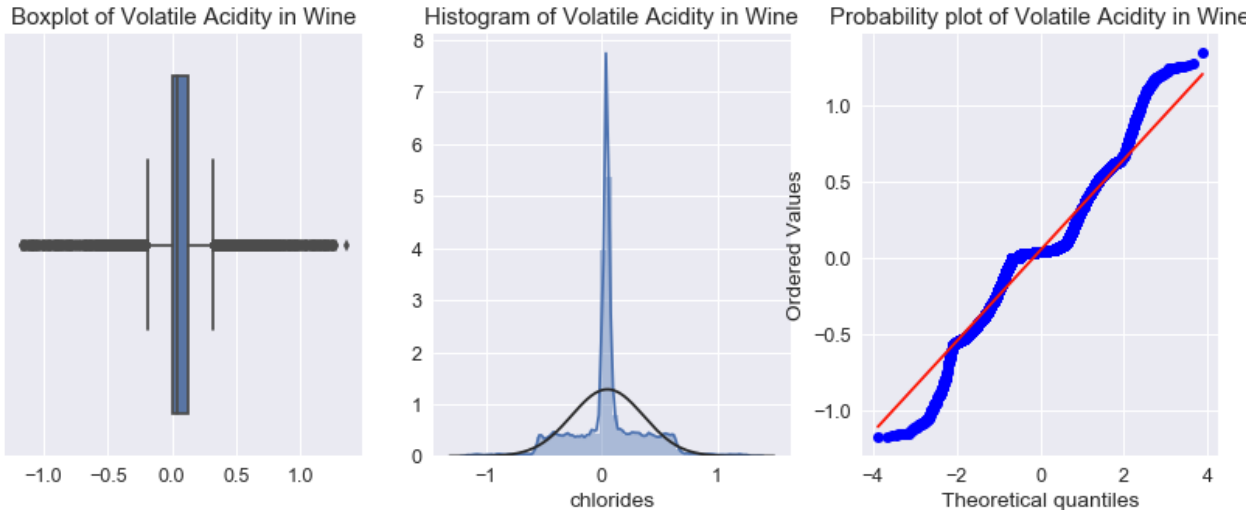
Figure 10: Residual Sugar vs. Stars & Target



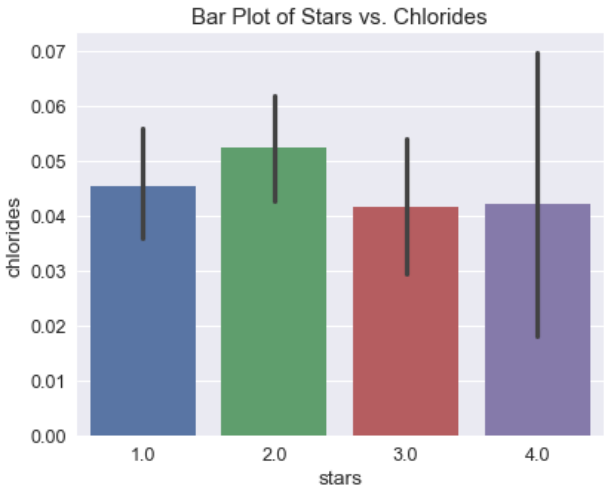
Chlorides

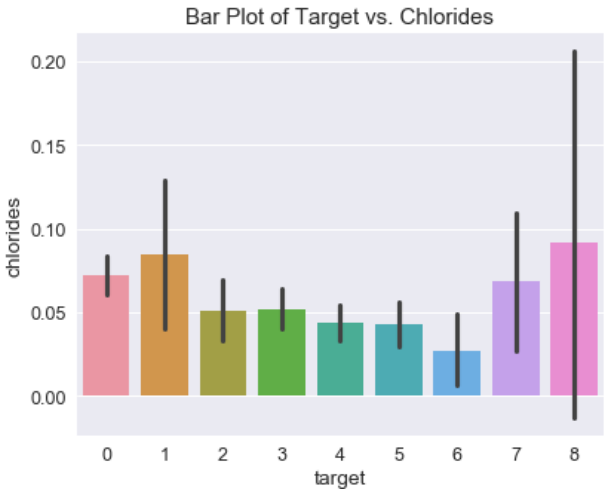
Salty is not a common wine descriptor. That it's also not a positive one probably goes without saying. Salinity is a concern in dry locations when frequent irrigation increases soil salinity, which increases wine salinity. The amount of chloride in wine is influenced by both the terroir and type of grape, and the importance of quantification lies in the fact that wine flavor is strongly impacted by this particular ion, which, in high concentration, gives the wine an undesirable salty taste and significantly decreases its market appeal. The values here are not that extreme.

Figure 11: Chlorides



Chlorides vs. Stars & Targets





FreeSulfurdioxide

Sulfur dioxide (SO2) is frequently added to must and juice as a preservative to prevent bacterial growth and slow down the process of oxidation by inhibiting oxidative enzymes. SO2 also improves the taste and retains the wine's fruity flavors and freshness of aroma. Two classes of sulfites are found in wine: free and bound. The free sulfites are those available to react and thus exhibit both germicidal and antioxidant properties. This variable has high numbers compared to the other variables, so we may need to perform some sort of log transformation here.

Figure 12: FreeSulfurdioxide

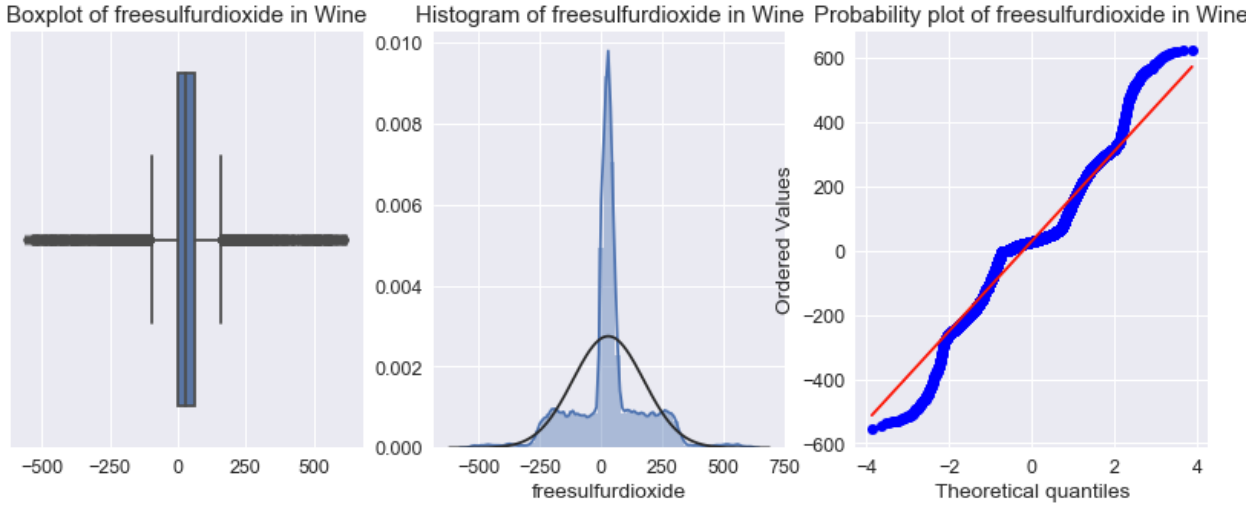
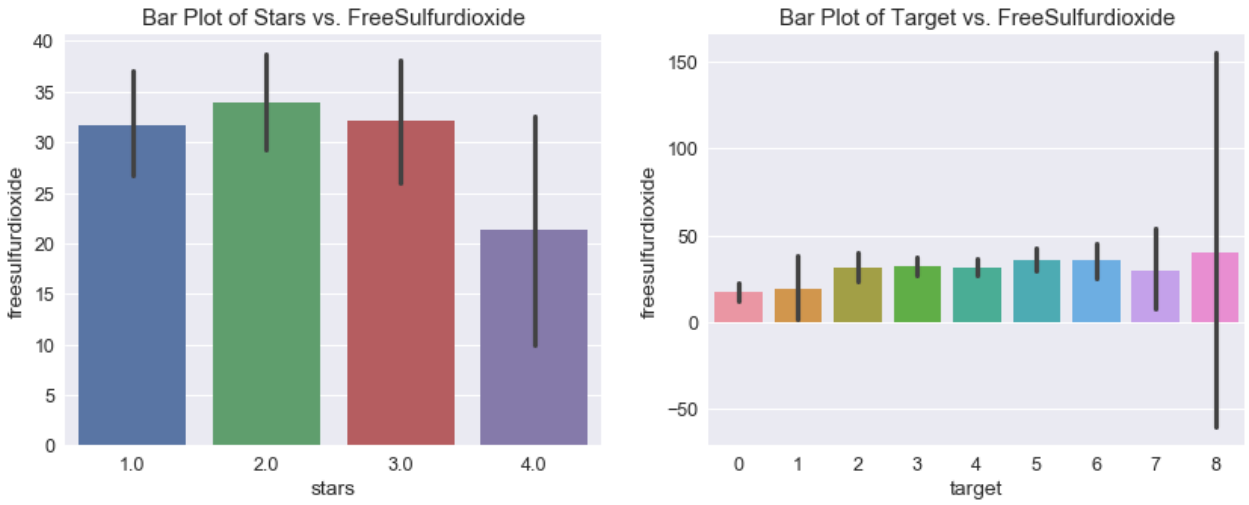


Figure 13: FreeSulfurdioxide vs. Stars & Target



TotalSulfurdioxide

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. In this case as well we see that the variable has high values and the outliers seems to be in wide range.

Figure 14: TotalSulfurdioxide

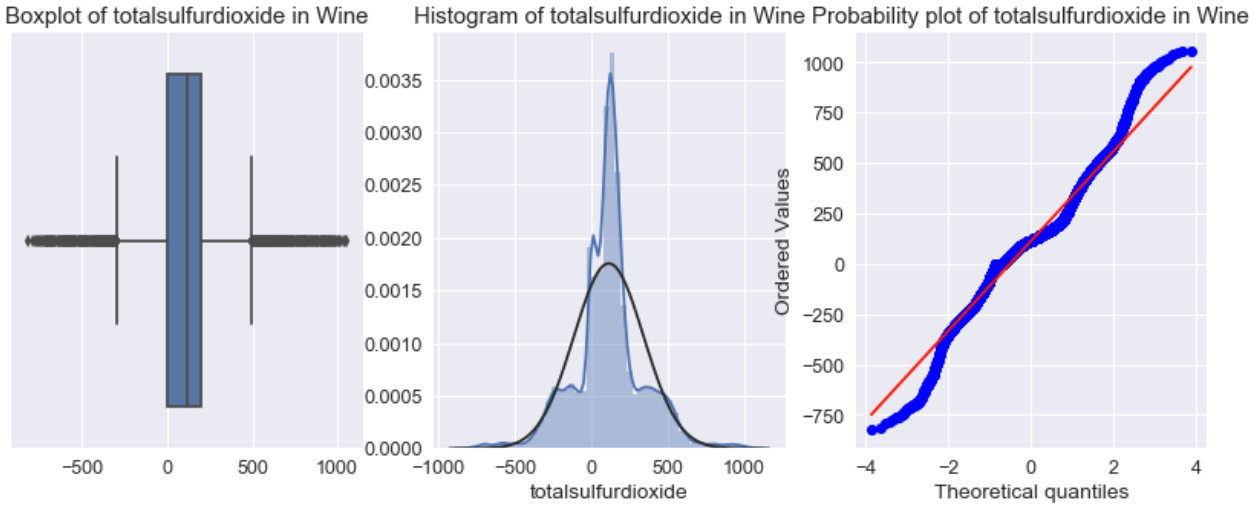
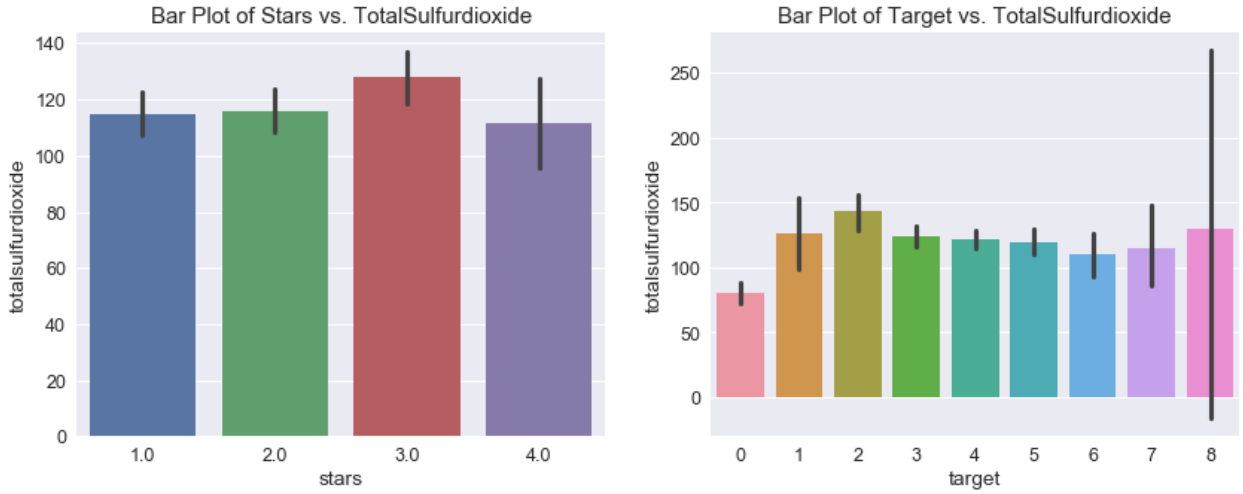


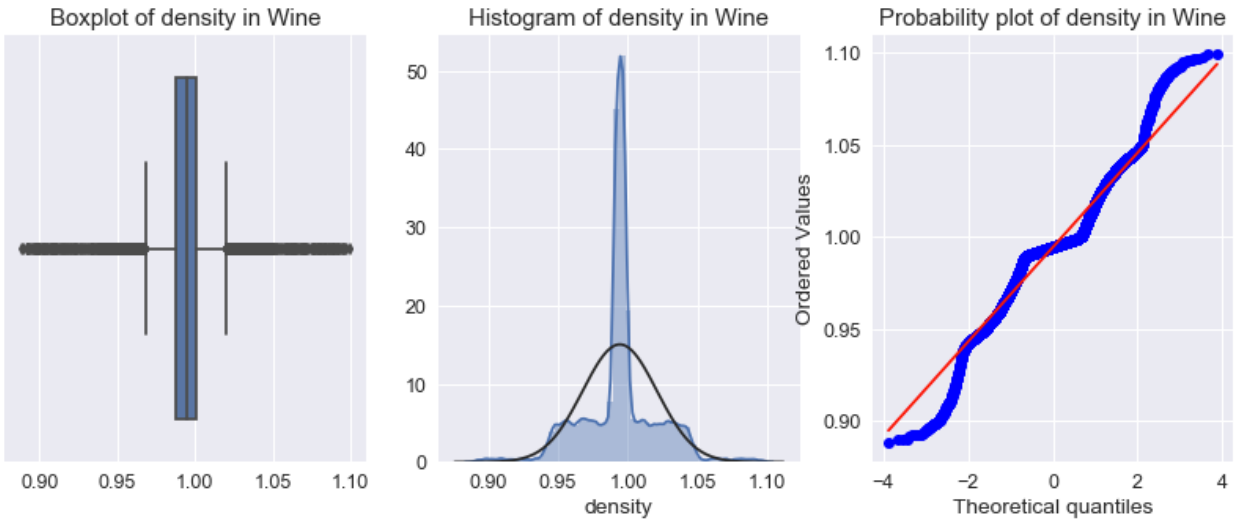
Figure 15: TotalSulfurdioxide vs. Stars & Target



Density

Density is defined as the mass, or weight, per volume of a material. In the case of liquids, density is often measured in units of g/mL. The density of wine is primarily determined by the concentration of alcohol, sugar, glycerol, and other dissolved solids. Although data in this case is not extreme as we have seen past couple of variables, but it is consistent with the rest of the variables in terms of skewness.

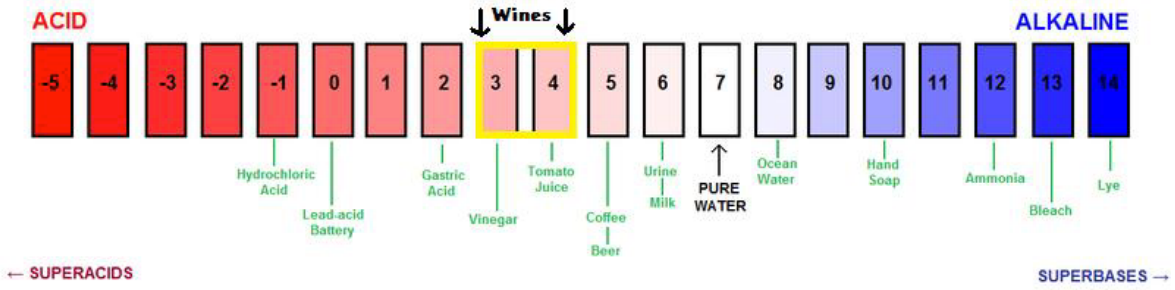
Figure 16: Density



ph

The pH of a wine is critical not only to its flavor but to nearly every aspect of the wine. The pH is a logarithmic scale that measures the concentration of free hydrogen ions floating in wine. The stronger the acid the more hydrogen ions you'll have so in essence it is a measurement of how strong an acid is. The pH value affects nearly every aspect of the wine. The pH affect flavor, aroma, color, tartrate precipitation, carbon dioxide absorption, malolactic fermentation, stability, ageablity, and fermentation rate. It can also affect the many chemical reactions that

take place in a wine during and after fermentation. pH value in most wines fall between 3.0 and 3.6.



https://commons.wikimedia.org/wiki/File:PH_scale_with_wine_highlighted.jpg

The effect of the ph on the stars seems to be consistent and also there aren't many negative values for this variable. Although there are many zero values, we should be able to transform and use this the model.

Figure 17: ph

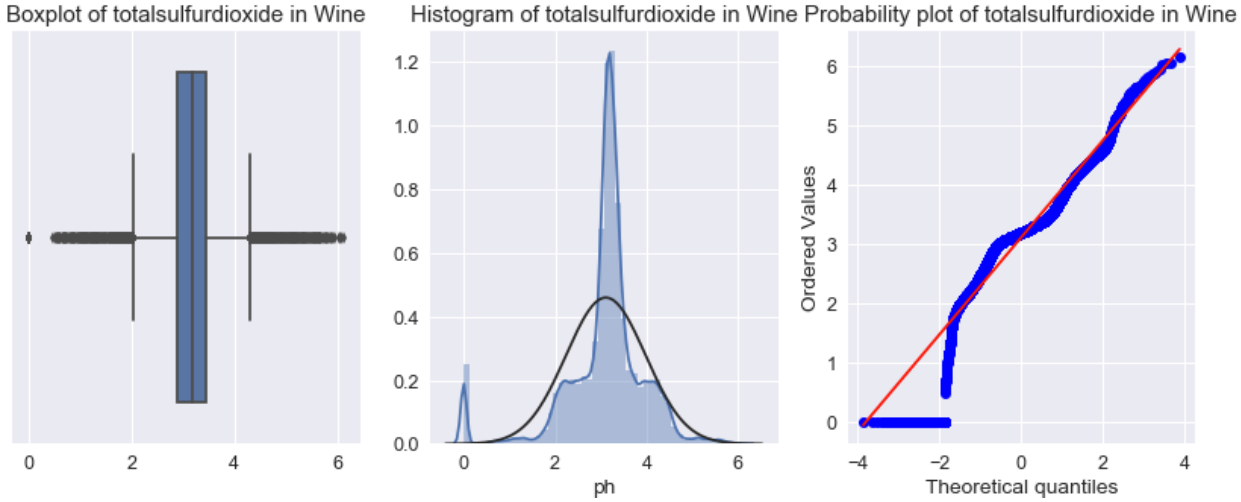
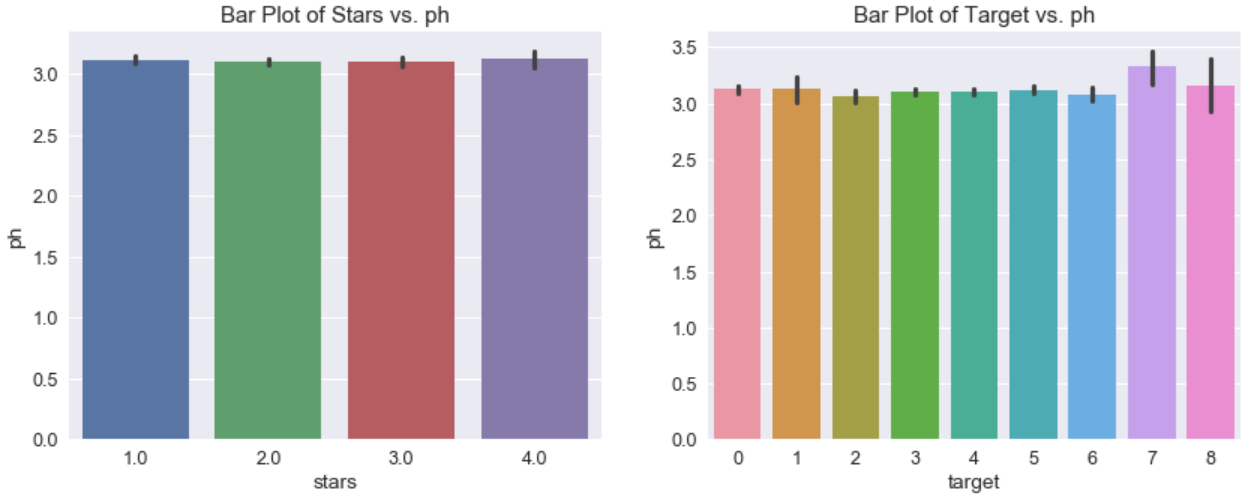


Figure 18: ph vs. Stars & Target



Sulfates

Sulfites or sulfur dioxide is a fruit preservative widely used in dried fruits as well as wine. Sulphur dioxide has an unpleasant smell, like that of a struck match, detectable at very low concentrations. Sulphur dioxide can cause potentially fatal allergic reactions and has been linked with numerous other health problems, including hangover. The levels in wine average 80 mg/liter, or about 10 mg in a typical glass of wine, with slightly higher amounts in white versus red. Wines with lower acidity need more sulfur than higher acidity wines. At pH 3.6 and above, the sulfites needed is much higher because it's an exponential ratio. Wines with more color (i.e. red wines) need less sulfur than clear wines (i.e. white wines). Wines with higher sugar content tend to need more sulfur to prevent secondary fermentation of the remaining sugar.

Figure 19: Sulfates

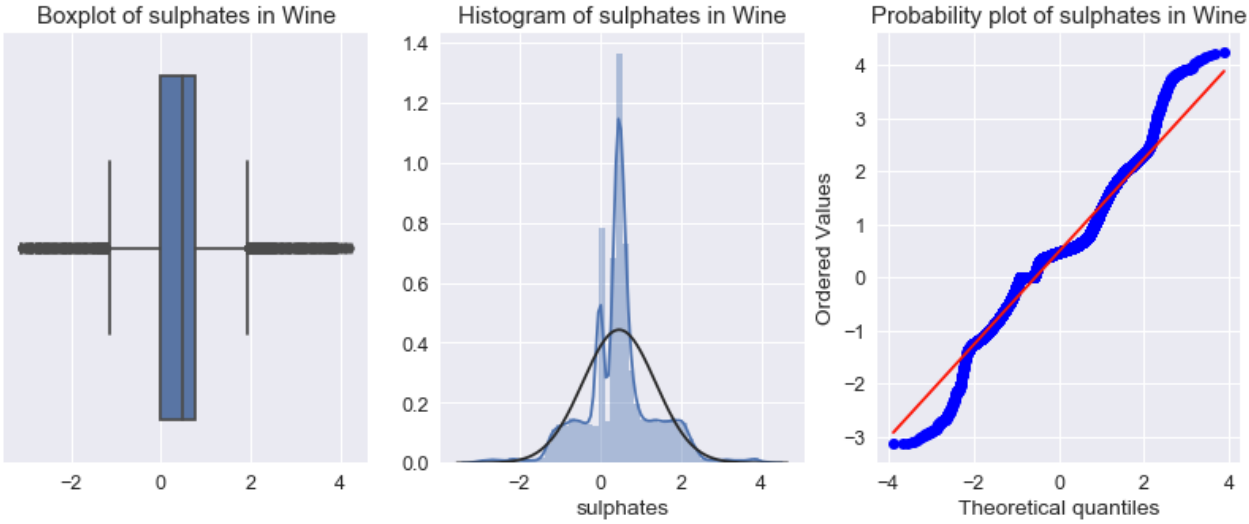
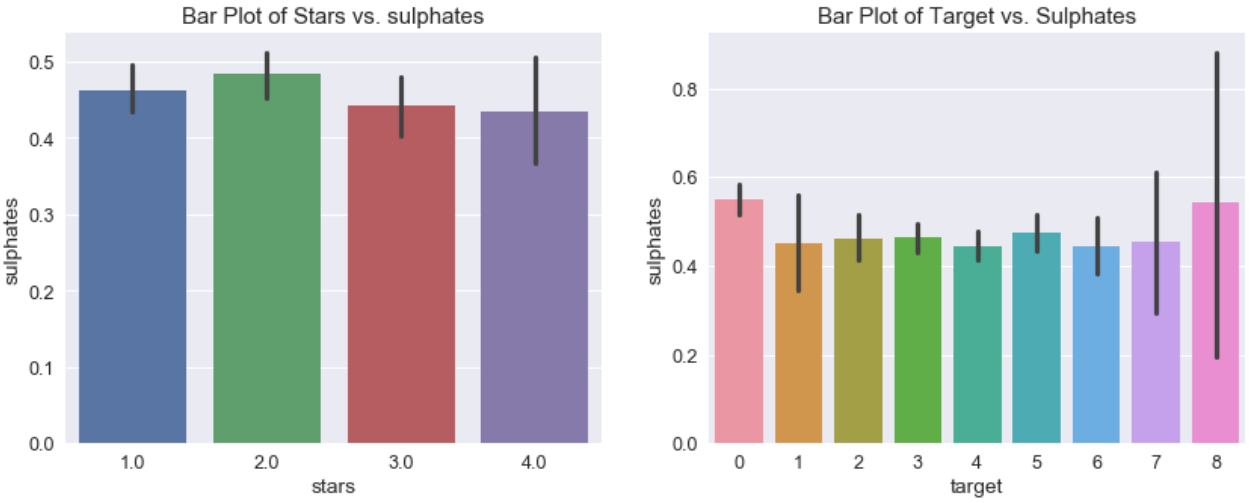


Figure 20: Label Appeal vs. Stars



Alcohol

The amount of alcohol produced during fermentation is dependent of the genus, specie and strain of yeast, the specific nutrient status such as amino acids and composition of the must and temperature, aeration and pH during fermentation. Higher alcohols can have an aromatic effect in wines and some higher alcohols can be considered positive and others can be considered negative to the aromatic wine profile. However, due to the concentration that are found in wines and its high threshold, higher alcohols does not have many sensory effects in wine. In this case also we many zero's, however this may be the only variable that had good distribution for the data.

Figure 21: Alcohol

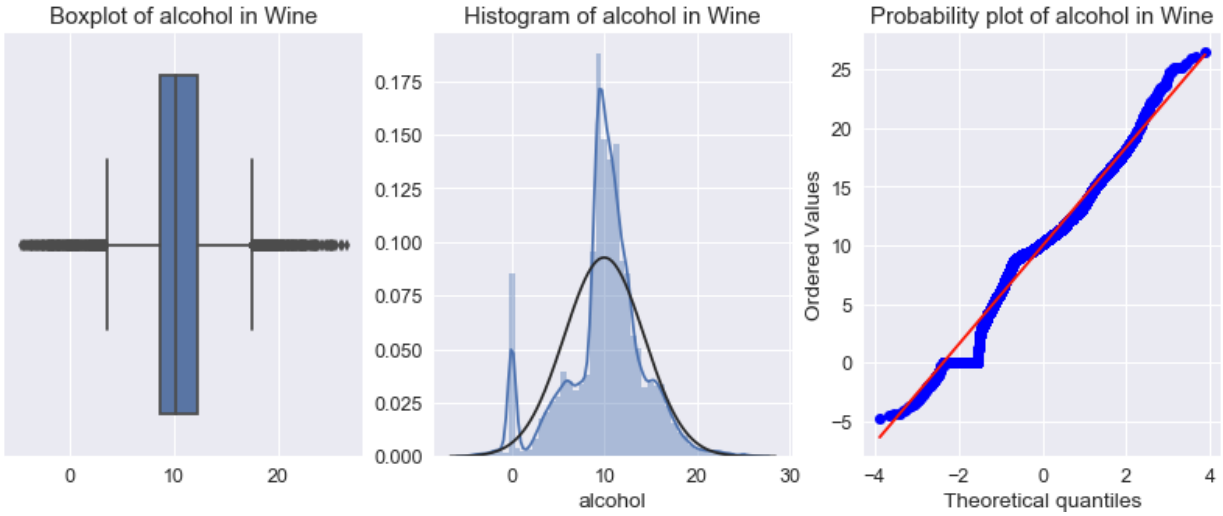
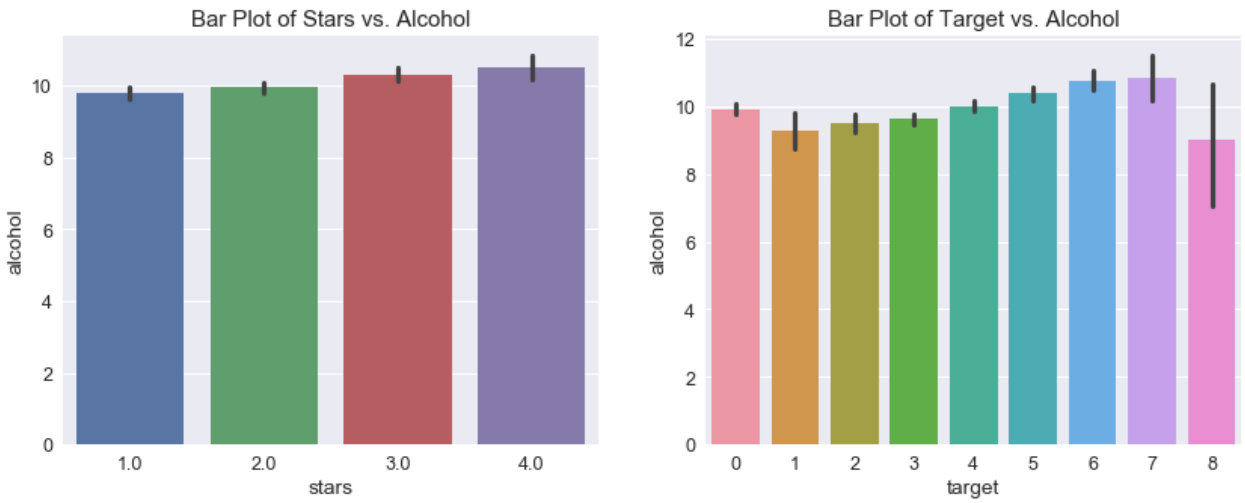


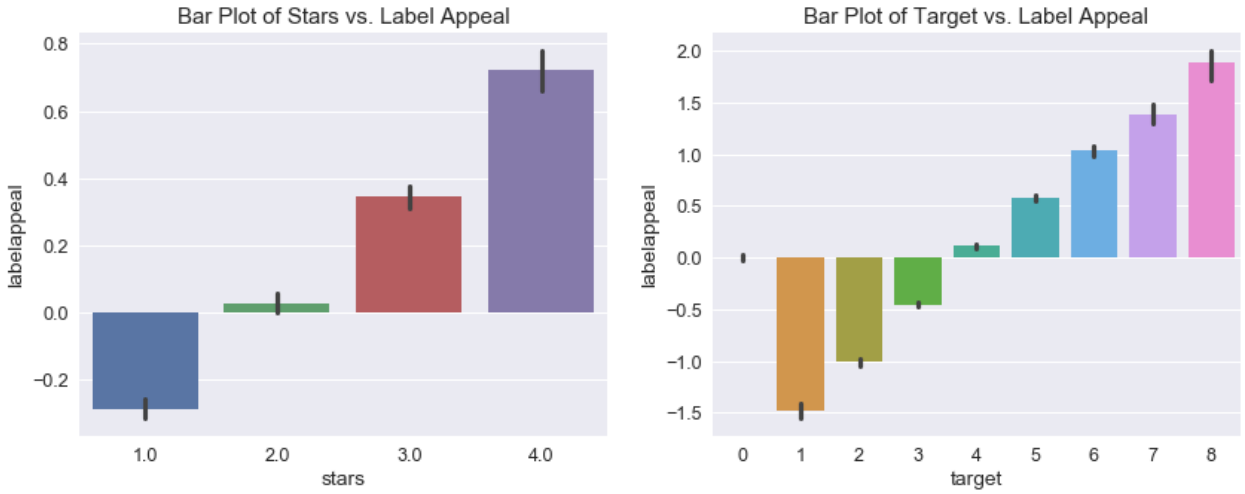
Figure 22: Alcohol vs. Stars & Alcohol



Label Appeal

This variable clearly has the effect on the Stars variable. As these two variable have influence on the response variable, we should try and include them in the final model. Figure 22 shows an interesting compraision between these two variables. When Label Appeal is less than 0, star rating seems to be at the lowest. We will consider this fact during the transformation. Figure 23 shows the comparison between the label appeal and the target. As we expected, there is a clear linear relationship between these two variables.

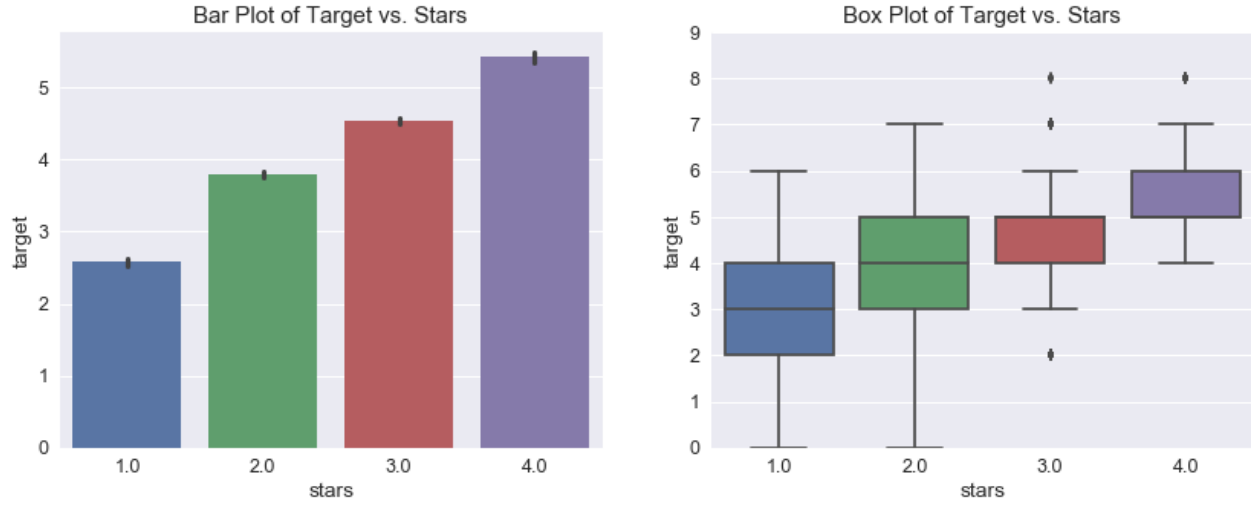
Figure 23: Label Appeal vs. Stars & Target



Stars

As we have seen above, Stars seems to be the most influential variable on the prediction of sales along with the label. As figure 24 shows the number of sales increases as the number of Stars going up. However, there are many missing values for this variable and we will transform them during the preparation phase.

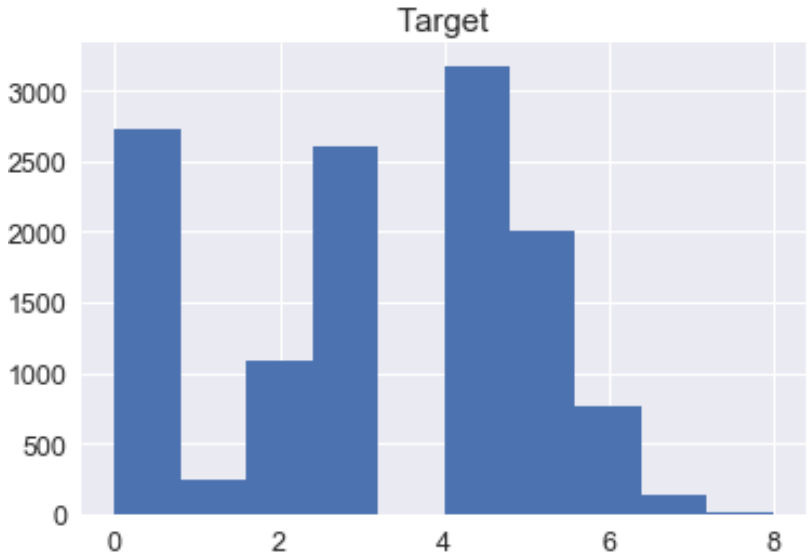
Figure 24: Stars vs. Target



Target

Finally, we look at the distribution for the target variable. There is a big spike at the zero, which indicates that this dataset is a good candidate for the zero inflated models. We will consider this fact during the model building.

Figure 25: Target



Data Preparation

During the exploration phase, we have discovered that there are many missing values for eight features within the dataset. There are also lot of negative values that needs to be corrected. I will divide this phase into three categories, Missing Values, Negative Values and the Outliers.

Missing Values

As part of the data preparation, first thing I have performed is to fix the missing values. I have decided to use the median value to replace all the missing values for each continuous variable. In most cases, both the median and the mean variables are similar. Stars variable is the exception here as it is categorical and replacing with mean or median does not make sense. So I have decided to replace all the Stars records that have labelAppeal less than 0 with 0 value and the rest of the missing values with 1. Table 4 shows the frequency table for stars vs. labelappeal.

Table 4: Frequency Table LabelAppeal vs. Stars

labelappeal / stars	-2	-1	0	1	2
1.0	203	1008	1334	448	49
2.0	70	849	1669	873	109
3.0	21	262	1011	766	152
4.0	0	29	192	310	81

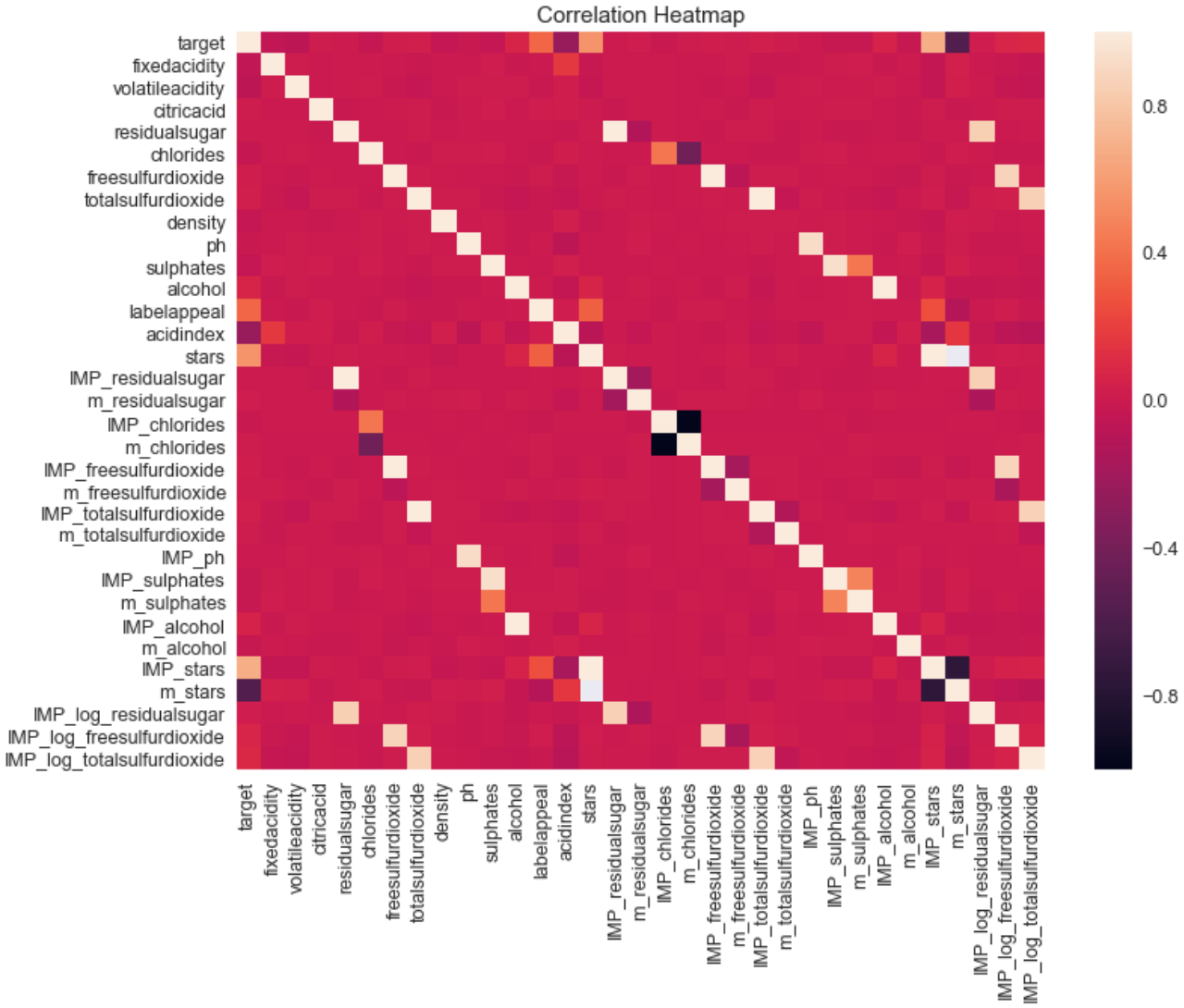
Negative Values

I have decided to simply convert all the negative values to positive. Although there are other techniques available to impute such data, I went with the simplest approach so that it will not add any bias to the data.

Outliers & log transformation

For all the outliers, instead of dropping them I chose to use the truncating strategy based off of quantiles. For the variables listed above, if any values exceeded the 99th percentile, then they were replaced with the value of the 99th percentile. Likewise, for values less than the 1th percentile. Instead of replacing the data in the existing variables, I chose to create new variables with IMP_* for the imputed value and m_* to represent the existence of the data. Finally, with missing values imputed and outliers mostly fixed, it was important to remember to do these same actions with the test data. As such, I imputed missing data with the medians from the training data and truncating the variables using the original 99th and 1th percentiles of the same variables from the training set. I have created three new variables, IMP_log_residualsugar, IMP_log_freesulfurdioxide and IMP_log_totalsulfurdioxide mainly because they high numbers compared to rest of the dataset. And performing log transformation on these variables scales the data back to same scale.

Figure 26: Correlation Matrix with Imputed data



One thing that pops up from the above heat map is that IMP_stars has positive correlation with the target variable, however m_stars seems to be negatively correlated. We need to consider this when building the model.

Models

For this section, I have built five prediction models. This includes a linear regression model and four generalized linear regression models of the following forms; Poisson, Negative Binomial, Zero Inflated Poisson and Zero Inflated Negative Binomial. For each model, we use a stepwise selection technique in order to determine which variables are included in each specification.

Model 1: Linear Regression

First model I have created is using the linear regression. I have selected the features based on F-regression results. The following table shows the summary of the linear regression. AIC and BIC values seems to be doing well compared to how simple this model is.

Table 5: Linear Regression

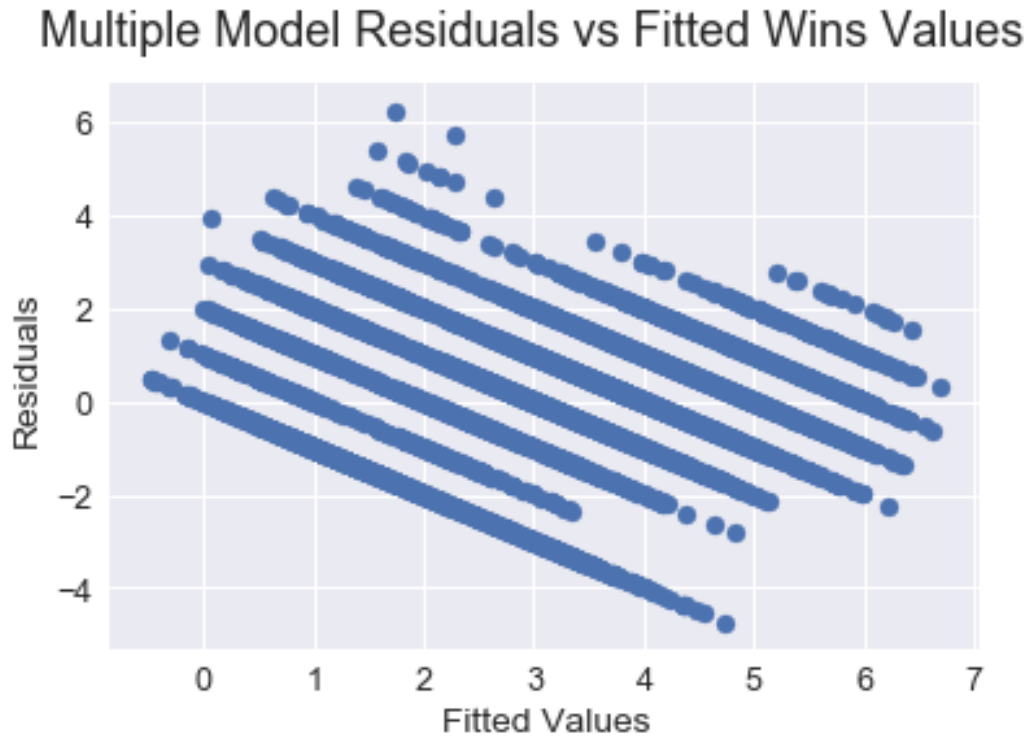
OLS Regression Results

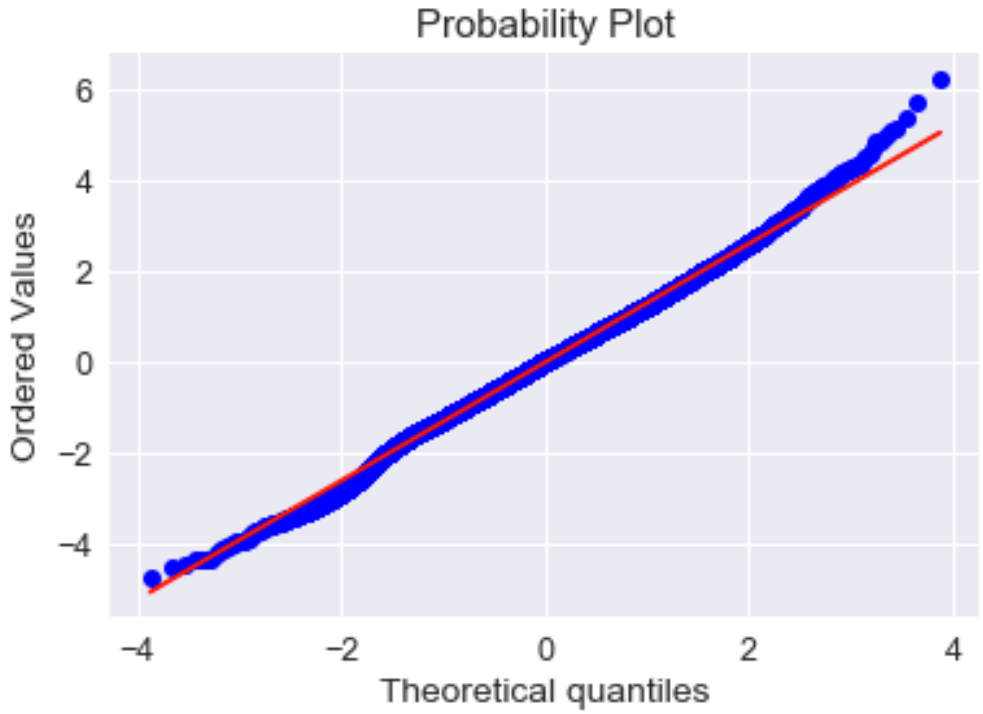
Dep. Variable:	target	R-squared:	0.537
Model:	OLS	Adj. R-squared:	0.536
Method:	Least Squares	F-statistic:	1139.
Date:	Sat, 03 Mar 2018	Prob (F-statistic):	0.00
Time:	22:24:17	Log-Likelihood:	-21620.
No. Observations:	12795	AIC:	4.327e+04
Df Residuals:	12781	BIC:	4.337e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.0107	0.135	22.290	0.000	2.746	3.275
IMP_log_citricacid	0.0823	0.038	2.177	0.029	0.008	0.156
volatileacidity	-0.1148	0.021	-5.353	0.000	-0.157	-0.073
fixedacidity	-0.0012	0.002	-0.513	0.608	-0.006	0.003
m_residualsugar	0.0820	0.045	1.817	0.069	-0.006	0.171
IMP_log_freesulfurdioxide	0.0523	0.011	4.935	0.000	0.032	0.073
m_freesulfurdioxide	0.1236	0.049	2.526	0.012	0.028	0.219
IMP_log_totalsulfurdioxide	0.0922	0.014	6.789	0.000	0.066	0.119
m_totalsulfurdioxide	0.0858	0.050	1.712	0.087	-0.012	0.184
IMP_alcohol	0.0141	0.003	4.195	0.000	0.007	0.021
labelappeal	0.4637	0.014	33.886	0.000	0.437	0.491
acidindex	-0.2281	0.011	-20.178	0.000	-0.250	-0.206
IMP_stars	0.7792	0.016	49.634	0.000	0.748	0.810
m_stars	-1.4751	0.031	-47.772	0.000	-1.536	-1.415

Residual plot shows clear linear limits at lower and upper end of the distribution and is not truly randomized. This model suffers from estimation of dependent variable in some extreme cases.

Figure 27: Model 1 Residuals and QQ Plot






```
Mean Abosolute Error:  1.0290982750127364
```

Model 2: Poisson Regression

Next, I have implemented Poisson based generalized linear model. The initial iteration included all possible predictor variables. F-test results enabled the removal of a total of nine variables during subsequent modeling iterations, producing a model comprised of twelve predictors.

Figure 28: Poisson

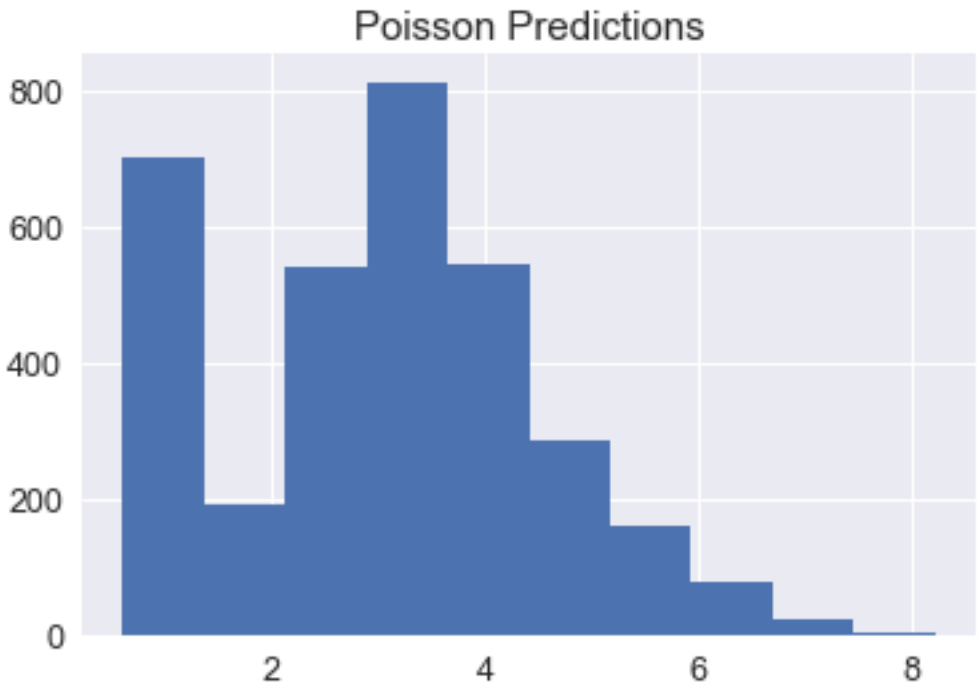
Model:	Poisson	Pseudo R-squared:	0.165
Dependent Variable:	target	AIC:	45779.0043
Date:	2018-03-03 23:29	BIC:	45875.9429
No. Observations:	12795	Log-Likelihood:	-22877.
Df Model:	12	LL-Null:	-27401.
Df Residuals:	12782	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	9.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
IMP_log_citricacid	0.0266	0.0165	1.6132	0.1067	-0.0057	0.0590
volatileacidity	-0.0388	0.0096	-4.0435	0.0001	-0.0577	-0.0200
m_residualsugar	0.0289	0.0193	1.4933	0.1354	-0.0090	0.0668
IMP_log_freesulfurdioxide	0.0190	0.0048	3.9470	0.0001	0.0096	0.0284
m_freesulfurdioxide	0.0423	0.0213	1.9880	0.0468	0.0006	0.0839
IMP_log_totalsulfurdioxide	0.0345	0.0062	5.5317	0.0000	0.0223	0.0467
m_totalsulfurdioxide	0.0306	0.0217	1.4098	0.1586	-0.0119	0.0731
IMP_alcohol	0.0041	0.0015	2.7931	0.0052	0.0012	0.0070
labelappeal	0.1582	0.0061	25.8109	0.0000	0.1462	0.1702
acidindex	-0.0839	0.0052	-16.1336	0.0000	-0.0941	-0.0737
IMP_stars	0.1883	0.0061	30.8840	0.0000	0.1763	0.2002
m_stars	-0.8405	0.0183	-45.9212	0.0000	-0.8764	-0.8047
intercept	1.2239	0.0612	20.0124	0.0000	1.1040	1.3437

AIC and BIC values seems to be bit high compared to the model 1 and RMSE value also is higher than the model 1.

Figure 28: Poisson AIC & BIC

```
AIC 45779.004333
BIC 45875.942860
RMSE value: 3.377104
```



Model 3: Negative Binomial Regression

Negative Binomial regression modeling is usually a more effective approach than Poisson regression modeling when the mean and variance of the response variable are *not* equivalent. As we saw earlier, the mean and variance of the 'TARGET' response variable are not, in fact, equivalent, though their values do appear reasonably proximal. AIC and BIC for this model also does not look good as the model 1. RMSE is also much higher than the previous models.

```
Mean value for Target:  3.0290738569753812
Variance value for Target:  3.710894522839234
```

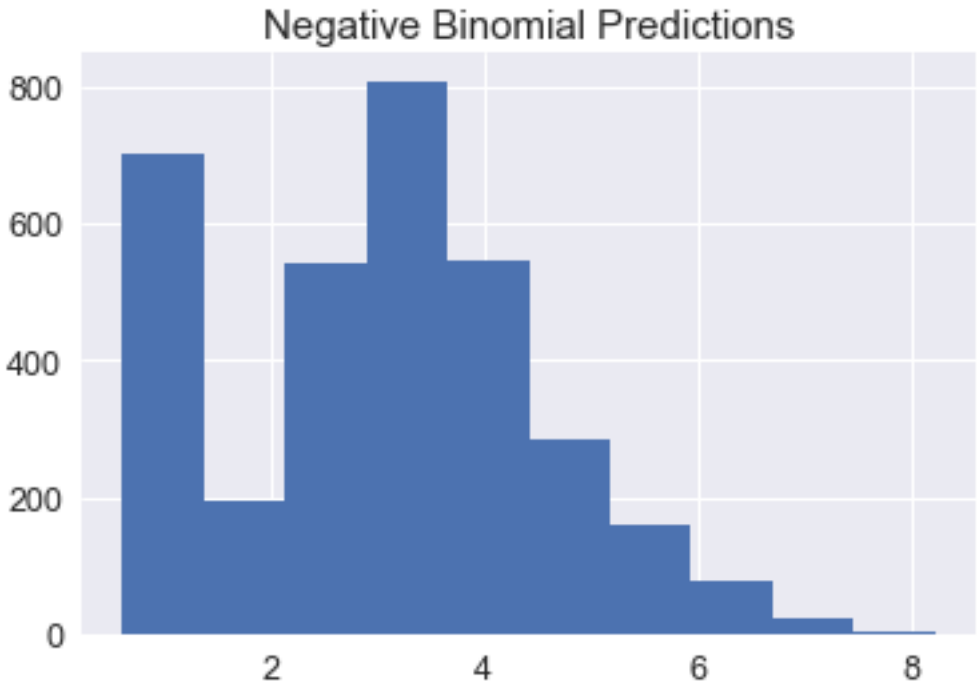
Figure 29: Negative Binomial

Model:	NegativeBinomial	Pseudo R-squared:	0.158
Dependent Variable:	target	AIC:	45781.0065
Date:	2018-03-04 00:12	BIC:	45885.4019
No. Observations:	12795	Log-Likelihood:	-22877.
Df Model:	12	LL-Null:	-27165.
Df Residuals:	12782	LLR p-value:	0.0000
Converged:	0.0000	Scale:	1.0000

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
IMP_log_citricacid	0.0267	0.0165	1.6165	0.1060	-0.0057	0.0591
volatileacidity	-0.0388	0.0096	-4.0421	0.0001	-0.0576	-0.0200
m_residualsugar	0.0289	0.0193	1.4935	0.1353	-0.0090	0.0668
IMP_log_freesulfurdioxide	0.0190	0.0048	3.9512	0.0001	0.0096	0.0284
m_freesulfurdioxide	0.0422	0.0213	1.9865	0.0470	0.0006	0.0839
IMP_log_totalsulfurdioxide	0.0345	0.0062	5.5392	0.0000	0.0223	0.0468
m_totalsulfurdioxide	0.0305	0.0217	1.4082	0.1591	-0.0120	0.0730
IMP_alcohol	0.0041	0.0015	2.7969	0.0052	0.0012	0.0070
labelappeal	0.1582	0.0061	25.8092	0.0000	0.1462	0.1702
acidindex	-0.0839	0.0052	-16.1247	0.0000	-0.0941	-0.0737
IMP_stars	0.1883	0.0061	30.8903	0.0000	0.1764	0.2003
m_stars	-0.8404	0.0183	-45.9142	0.0000	-0.8763	-0.8045
intercept	1.2230	0.0612	19.9981	0.0000	1.1031	1.3429
alpha	0.0000	0.0000	0.0085	0.9932	-0.0000	0.0000

Figure 30: Negative Binomial AIC & BIC

```
AIC 45781.006523
BIC 45885.401860
RMSE value: 3.377071
```



Model 4: Hurdle Model

For the final model, I have combined a logistic regression and a Poisson model to create the hurdle model. Figure 31 shows the results from the logistic regression results. From the logistic regression and Poisson formulas that compose Model C Hurdle, we see that only a few coefficient signs changed, and some of the magnitudes changed. We cannot comment much on these changes, aside from the fact that it is interesting that LabelAppeal has a negative coefficient in the logistic regression formula, whereas it has always had a positive coefficient in previous models. As shown in figure 33 below, AIC and BIC values seems to be improved, however this model performed very poorly in Kaggle submission.

Figure 31: Logistic regression results

Logit Regression Results

Dep. Variable:	FLAG	**No. Observations:**	12795
Model:	Logit	**Df Residuals:**	12787
Method:	MLE	**Df Model:**	7
Date:	Sat, 03 Mar 2018	**Pseudo R-squ.:**	0.3444
Time:	23:31:21	**Log-Likelihood:**	-4352.1
converged:	True	**LL-Null:**	-6637.9
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.9172	0.298	-3.079	0.002	-1.501	-0.333
volatileacidity	-0.1843	0.047	-3.900	0.000	-0.277	-0.092
IMP_log_freesulfurdioxide	0.0952	0.023	4.186	0.000	0.051	0.140
IMP_log_totalsulfurdioxide	0.2330	0.029	8.083	0.000	0.176	0.289
IMP_alcohol	-0.0158	0.008	-2.092	0.036	-0.031	-0.001
labelappeal	-0.4478	0.031	-14.615	0.000	-0.508	-0.388
acidindex	-0.4494	0.024	-18.833	0.000	-0.496	-0.403
IMP_stars	3.4792	0.110	31.673	0.000	3.264	3.695

Figure 32: Poisson regression results

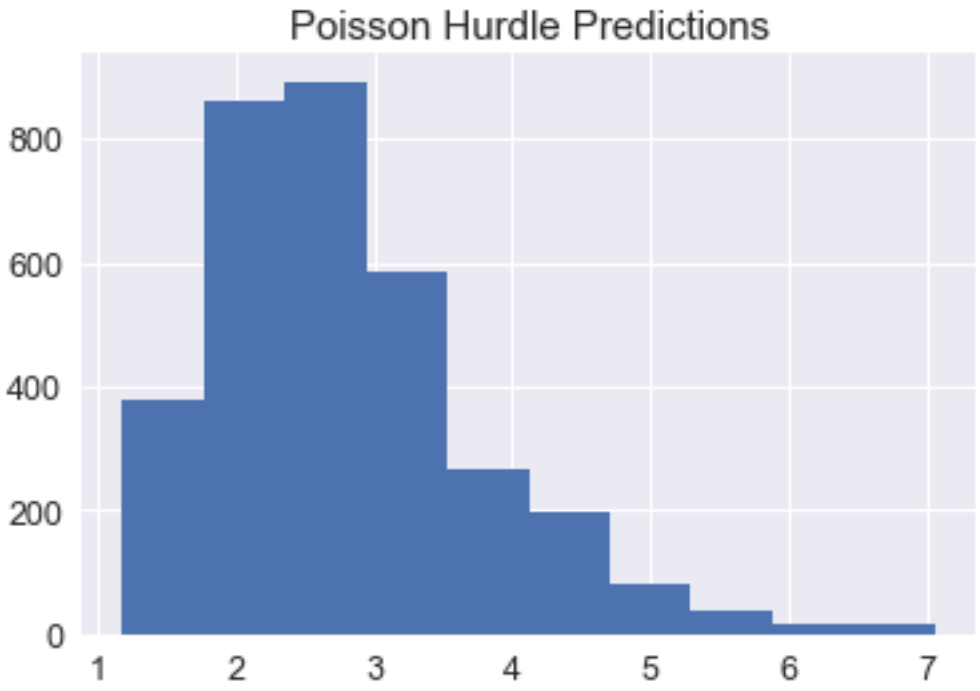
Generalized Linear Model Regression Results

Dep. Variable:	target	**No. Observations:**	10061
Model:	GLM	**Df Residuals:**	10053
Model Family:	Poisson	**Df Model:**	7
Link Function:	log	**Scale:**	1.0
Method:	IRLS	**Log-Likelihood:**	-15648.
Date:	Sat, 03 Mar 2018	**Deviance:**	2960.2
Time:	23:31:25	**Pearson chi2:**	2.62e+03
No. Iterations:	4		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.8672	0.071	12.229	0.000	0.728	1.006
volatileacidity	-0.0143	0.011	-1.288	0.198	-0.036	0.007
IMP_log_freesulfurdioxide	0.0054	0.006	0.977	0.328	-0.005	0.016
IMP_log_totalsulfurdioxide	-0.0056	0.007	-0.769	0.442	-0.020	0.009
IMP_alcohol	0.0095	0.002	5.543	0.000	0.006	0.013
labelappeal	0.2966	0.007	41.175	0.000	0.282	0.311
acidindex	-0.0278	0.006	-4.500	0.000	-0.040	-0.016
IMP_stars	0.1309	0.007	19.533	0.000	0.118	0.144

Figure 33: Poisson Hurdle Predictions

```
AIC 31311.678131
BIC -89692.454792
RMSE value: 2.571753
```



Model Selection

The following table shows the comparison for all the models that are described above. Based on all the four models summary, I chose the model 1, Linear Regression as the best performing one based on its RMSE value and also based on AIC & BIC values. Even though this assignment is meant for zero inflated mode, I couldn't produce best model using Poisson or Negative Binomial methods. And Model 1 (Linear Regression) also performed very well in the Kaggle submissions.

Table 8: Model Comparison

	Model 1 (Linear Regression)	Model 2 (Poisson)	Model 3 (Binomial)	Model 4 (Hurdle)
AIC	43268	45779	45781	31311
BIC	43372	45875	45885	-89692
RMSE	1.31	3.37	3.37	2.57

Figure 34: Linear Regression

OLS Regression Results

Dep. Variable:	target	R-squared:	0.537
Model:	OLS	Adj. R-squared:	0.536
Method:	Least Squares	F-statistic:	1139.
Date:	Sat, 03 Mar 2018	Prob (F-statistic):	0.00
Time:	22:24:17	Log-Likelihood:	-21620.
No. Observations:	12795	AIC:	4.327e+04
Df Residuals:	12781	BIC:	4.337e+04
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.0107	0.135	22.290	0.000	2.746	3.275
IMP_log_citricacid	0.0823	0.038	2.177	0.029	0.008	0.156
volatileacidity	-0.1148	0.021	-5.353	0.000	-0.157	-0.073
fixedacidity	-0.0012	0.002	-0.513	0.608	-0.006	0.003
m_residualsugar	0.0820	0.045	1.817	0.069	-0.006	0.171
IMP_log_freesulfurdioxide	0.0523	0.011	4.935	0.000	0.032	0.073
m_freesulfurdioxide	0.1236	0.049	2.526	0.012	0.028	0.219
IMP_log_totalsulfurdioxide	0.0922	0.014	6.789	0.000	0.066	0.119
m_totalsulfurdioxide	0.0858	0.050	1.712	0.087	-0.012	0.184
IMP_alcohol	0.0141	0.003	4.195	0.000	0.007	0.021
labelappeal	0.4637	0.014	33.886	0.000	0.437	0.491
acidindex	-0.2281	0.011	-20.178	0.000	-0.250	-0.206
IMP_stars	0.7792	0.016	49.634	0.000	0.748	0.810
m_stars	-1.4751	0.031	-47.772	0.000	-1.536	-1.415

Model Equation

As explained in the EDA section, all the missing values are imputed with the median or mean value. I have performed log operation on IMP_log_totalsulfurdioxide and IMP_log_freesulfurdioxide to scale down the values of these two features. IMP_residualsugar and IMP_alcohol have been imputed to replace missing values with their median values. IMP_stars have been imputed to use either 1 or 0 based on the 'LabelAppeal' value.

P_TARGET = 3.0107 + 0.0823 * IMP_log_citricacid - 0.1148 * volatileacidity - 0.0012 * fixedacidity + 0.0820 * m_residualsugar + 0.0523 * IMP_log_freesulfurdioxide + 0.1236 * m_freesulfurdioxide + 0.0922 * IMP_log_totalsulfurdioxide + 0.0858 * m_totalsulfurdioxide + 0.0141 * IMP_alcohol + 0.4637 * labelappeal - 0.2281 * acidindex + 0.7792 * IMP_stars - 1.4751 * m_stars

Conclusion

The goal for this assignment was to create a model that best predicted how many cases of wine would be purchased based on its characteristics. I have created numerous Poisson, Negative Binomial, Zero Inflated Poisson, Zero Inflated Negative Binomial and Linear Regression models based on the Wine data set provided. I have prepared the data using imputation and binning, and used stepwise automated variable selection to help choose variables for our models. The best model surpassed the other models by having the lowest ME and RMSE values. Out of all the models produced, I have chosen Model 1 as the best predicting model based on its RMSE and AIC/BIC values.