Homework 3: Wine Sales Project
Jack Doig
PREDICT 411 Section 59

Contents

Introduction

The purpose of this paper is to analyse data on wine characteristics so as to build a model to predict the wine cases that will be ordered. The target variable is therefore a count of cases that are bought by wine distribution centres after sampling a wine. With accurate prediction of wine sales, the wine manufacturer will be able to optimize inventory, maximise profits and minimize waste.

The data represents approximately 12,000 wines, with each record containing largely chemical information, but also marketing information such as label appeal and stars. The predictor variables are largely continuous variables.

This task will involve firstly exploring and possibly correcting the data through imputation such that it offers as much predictive intelligence as possible, and then developing and scoring several Poisson, standard OLS regression and Negative Binomial models.

In order to determine the best model, a variety of different methods will be employed, varying the predictors for the Poisson and Negative Binomial regressions, and using Stepwise, Forward and backward variable selection for OLS regression, while each model judged on a balance of in-sample and out-of-sample measures, interpretability and simplicity.

The final model will then be analysed so as to understand its accuracy in predicting insurance claims, and further analysis undertaken if required.

1. Data Exploration

a) Data Summary and business of wine chemistry

There are approximately 13000 rows in the data set. There are a 14 continuous predictor variables.

The predictor data can be described in 2 broad categories. There is a category of information about the chemical composition of the wine, for example the ph level, citric acid and density of the wine. Secondly, there is information about the marketing of the wine, which includes two predictors: the Stars, as rated by a team of experts, and the label appeal, or how attractive the label is.

Within the chemical predictors, 5 deal with measures of acidity, 3 predictors deal with Sulphites and Sulphate, while others concern other chemical characteristics. A little more on the chemical composition predictors:

Total Acidity (which is measured in the data as an index) is typically divided into 2 groups, **volatile and fixed acidity**, and are a result of the fermentation process. Volatile acidity refers to the steam distillable acids in wine. Acidity contributes to the fundamental taste of wine, and also affects colour and the stability of wine, or the overall lifespan. **Citric Acid** is often added to the winemaking process to boost the wine's acidity and therefore its flavour. Regarding the **pH level**, most wines fall between 3 and 4. Currently Winemakers are pushing pH levels upwards, whereas previously pH levels above 3.6 were unusual.

Sulphur Dioxide, while produced naturally in wine, is often added to the winemaking process to stabilise the wine and allow it to age past a few months, preventing the growth of bacteria in the wine. Secondly it acts as an antioxidant, maintaining the fruitiness of the wine. Excessive use of Sulphur dioxide can ruin the fermentation process and result in an undesirable aroma and taste, furthermore it has been linked with allergic reactions and hangovers. **Free Sulphur Dioxide** is an anti-oxidant, and a good winemaker will try to maximise the ratio of free Sulphur Dioxide to bound Sulphur Dioxide.

In summary, we would expect a good wine to be correlated with a mid-level of sulphites, but a high ratio of free Sulphur Dioxide to bounded Sulphur dioxide.

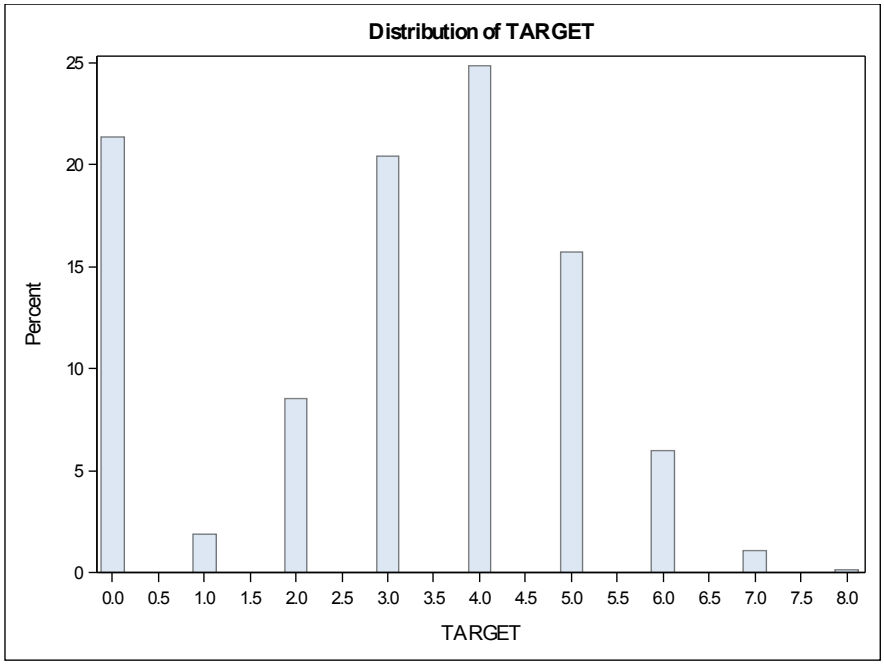
Chloride, sodium chloride, or salt tends to depend on the region the wine is coming from, as each region will have a different level of chlorides in the water supply. It may give the wine a salty flavour. An excess of chlorides may turn away customers and result in lower sales.

Residual Sugar, often measure in grams per litre, refers to the amount of sugar in the wine after the completion of the fermentation process. The sweet or dryness of the wine is controlled by the residual sugar, the alcohol content and the acidity level of the wine.

The **alcohol** content of wine typically ranges between 8-15%, and is unlikely to be correlated with sales or quality, as wine from warmer climates (eg. California) is likely to have a high alcohol content than cooler climates like Germany, similarly sweeter wines tend to have a higher alcohol content.

b) Descriptive Statistics: Target Variables

There are no missing values among the target variable, representing how many cases of wine have been sold. The range of values is from 0 to 8, and the distribution is as follows:



As can be seen in the distribution above, the target value is 0 for approximately 22% of the cases. There is then an approximate normal distribution of the remaining data, with the average approximately 4 and the maximum of 8.

c) Data integrity check: Understanding the Predictor variables in isolation

Aside from the target variable, there are 14 predictor variables. The table below summarizes the continuous predictor variables

Variable	N	Mean	Std Dev	Minimum	Maximum
TARGET	12795	3.029074	1.926368	0	8
FixedAcidity	12795	7.075717	6.317644	-18.1	34.4
VolatileAcidity	12795	0.324104	0.784014	-2.79	3.68
CitricAcid	12795	0.308413	0.86208	-3.24	3.86
ResidualSugar	12179	5.418733	33.74938	-127.8	141.15
Chlorides	12157	0.054823	0.318467	-1.171	1.351
FreeSulfurDioxide	12148	30.84557	148.7146	-555	623
TotalSulfurDioxide	12113	120.7142	231.9132	-823	1057
Density	12795	0.994203	0.026538	0.88809	1.09924
pH	12400	3.207628	0.679687	0.48	6.13
Sulphates	11585	0.527112	0.932129	-3.13	4.24
Alcohol	12142	10.48924	3.727819	-4.7	26.5
LabelAppeal	12795	-0.00907	0.891089	-2	2
AcidIndex	12795	7.772724	1.323926	4	17
STARS	9436	2.041755	0.90254	1	4

This table highlights the missing variables under the column N (discussed further in the next section). It also highlights possible unrealistic values which may need to be imputed or treated as a separate variable.

There is no reason to think that these variables are scaled (which may explain the negative values), and furthermore the scale of the variables is inconsistent. I will therefore treat the negative values as bad data which will need to be imputed or dealt with in some way. This will be dealth with in the data preparation section.

d) Missing variables

There are number of missing variables which will need to be examined and either possibly corrected or treated as a separate category. These are summarized below:

Variable	Number Missing	Percentage missing
ResidualSugar	616	4.81%
Chlorides	638	4.99%
FreeSulfurDioxide	647	5.06%
TotalSulfurDioxide	682	5.33%
pH	395	3.09%
Sulphates	1210	9.46%
Alcohol	653	5.10%
STARS	3359	26.25%

There are a number of approaches to take in order to correct these missing variables. They can be removed as rows or columns, they can form a separate category, or they can be imputed using an average, a distribution, or a more sophisticated approach such as a regression on another predictor. The best approach will be determined by looking at the correlations and descriptive statistics and then executed in the data preparation section.

e) Unrealistic Outliers

The quality of the data, when considered as isolated columns, contains a number of negative values that will need to be dealt with in some way. As discussed above there is no reason to believe the data has been scaled, so I will take these negative values as bad data.

Variable	# negative values	Percentage
FixedAcidity	1621	12.67%
VolatileAcidity	2827	22.09%
CitricAcid	2966	23.18%
ResidualSugar	3136	24.51%
Chlorides	3197	24.99%
FreeSulfurDioxide	3036	23.73%
TotalSulfurDioxide	2504	19.57%
Sulphates	2361	18.45%
Alcohol	118	0.92%
LabelAppeal	3640	28.45%

Among the maximum values of the predictor variables, I am comfortable that these are within the realms

f) Correlations in the raw data

Correlations between variables might be useful for dimensionality reduction or for imputing missing data, and secondly correlations between the categorical variables and the target variable might indicate a relationship.

There are some interesting inferences we can make, which may help later in validating the model, checking it for consistency with real-world logic, suggest ways to come up with derived variables or reduce dimensionality and possibly helping in making it more parsimonious.

The table of correlations of the predictor variables, based on the raw data, is below.

	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol	LabelAppeal	AcidIndex	STARS
TARGET	1	-0.04901	-0.08879	0.00868	0.01649	-0.03826	0.04382	0.05148	-0.03552	-0.00944	-0.03885	0.06206	0.3565	-0.24605	0.55879
FixedAcidity	-0.04901	1	0.01238	0.01424	-0.01885	-0.00046	0.00497	-0.0225	0.00648	-0.00898	0.03078	-0.00937	-0.0034	0.17844	-0.00663
VolatileAcidity	-0.08879	0.01238	1	-0.01695	-0.00648	0.00099	-0.00708	-0.02108	0.01473	0.01359	0.00013	0.00407	-0.017	0.04464	-0.03443
CitricAcid	0.00868	0.01424	-0.01695	1	-0.00694	-0.00857	0.00643	0.00632	-0.01395	-0.00871	-0.01299	0.01705	0.00865	0.0657	0.00066
ResidualSugar	0.01649	-0.01885	-0.00648	-0.00694	1	-0.00559	0.01749	0.02248	0.0041	0.01212	-0.00772	-0.02	0.00232	-0.00941	0.01674
Chlorides	-0.03826	-0.00046	0.00099	-0.00857	-0.00559	1	-0.02066	-0.01399	0.02266	-0.01761	-0.00329	-0.01969	0.01051	0.02524	-0.00493
FreeSulfurDioxide	0.04382	0.00497	-0.00708	0.00643	0.01749	-0.02066	1	0.01372	0.00318	0.00605	0.01159	-0.01859	0.01029	-0.04172	-0.00908
TotalSulfurDioxide	0.05148	-0.0225	-0.02108	0.00632	0.02248	-0.01399	0.01372	1	0.01282	-0.00434	-0.00713	-0.01596	-0.0098	-0.04931	0.01393
Density	-0.03552	0.00648	0.01473	-0.01395	0.0041	0.02266	0.00318	0.01282	1	0.00577	-0.00906	-0.00721	-0.0094	0.04041	-0.01828
pH	-0.00944	-0.00898	0.01359	-0.00871	0.01212	-0.01761	0.00605	-0.00434	0.00577	1	0.00548	-0.01155	0.00414	-0.05868	-0.00049
Sulphates	-0.03885	0.03078	0.00013	-0.01299	-0.00772	-0.00329	0.01159	-0.00713	-0.00906	0.00548	1	0.00474	-0.0039	0.03445	-0.01231
Alcohol	0.06206	-0.00937	0.00407	0.01705	-0.02	-0.01969	-0.01859	-0.01596	-0.00721	-0.01155	0.00474	1	0.00103	-0.03814	0.06522
LabelAppeal	0.3565	-0.00337	-0.01699	0.00865	0.00232	0.01051	0.01029	-0.00975	-0.00937	0.00414	-0.00389	0.00103	1	0.02475	0.33479
AcidIndex	-0.24605	0.17844	0.04464	0.0657	-0.00941	0.02524	-0.04172	-0.04931	0.04041	-0.05868	0.03445	-0.03814	0.02475	1	-0.08626
STARS	0.55879	-0.00663	-0.03443	0.00066	0.01674	-0.00493	-0.00908	0.01393	-0.01828	-0.00049	-0.01231	0.06522	0.33479	-0.08626	1

As can be seen in the table above, there are few strong correlations, either with the target variable, or with between the predictor variables. The exceptions are:

- There is a strong correlation between Stars and the target variable, sales (0.55)
- There is a reasonable correlation between LabelAppeal and Sales (0.35)
- There is a reasonable negative correlation between AcidIndex and Sales (-0.24)
- Finally, Stars and LabelAppeal has a reasonable correlation (0.33)

2. Data Preparation

All imputation of missing or unrealistic variables have been rounded to the nearest whole number.

a) Missing variables

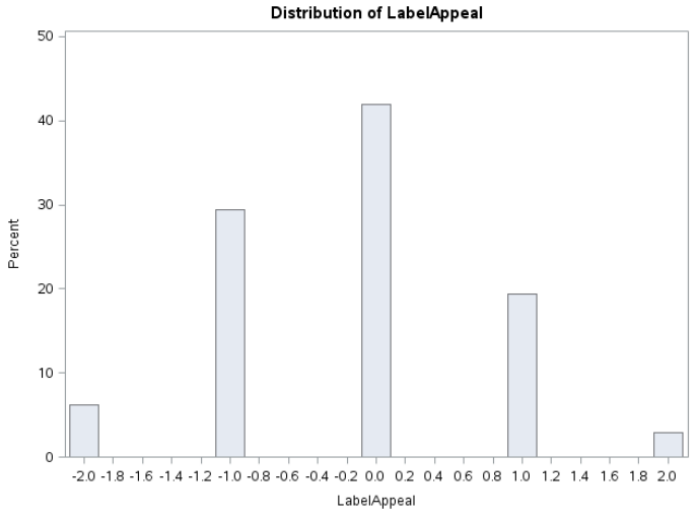
The list of missing variables and my approach to imputing them is outlined below. With the low level of correlation within the predictor variables, imputation by regression with another predictor is not a realistic option.

The summary of actions taken is outlined below.

Variable	Number Missing	Percentage missing	Corrective Action
ResidualSugar	616	4.81%	Average Value
Chlorides	638	4.99%	Average Value
FreeSulfurDioxide	647	5.06%	Average Value
TotalSulfurDioxide	682	5.33%	Average Value
pH	395	3.09%	Average Value
Sulphates	1210	9.46%	Average Value
Alcohol	653	5.10%	Average Value
STARS	3359	26.25%	New category of variable

The rationale for adding a new category for the missing STARS variable is that it seems that a non-rated wine would have a real-world significance: it may be too rare, or the wine too low-market or poor, to achieve a rating. To this end I have concluded, in spite of the correlation with label appeal, to create a new category.

Note, as a counterpoint, this new category does not show any clear pattern within the Label Appeal predictor, as shown in the graph below. This would suggest (but not confirm) that the missing STARS variables are not part of a separate category:



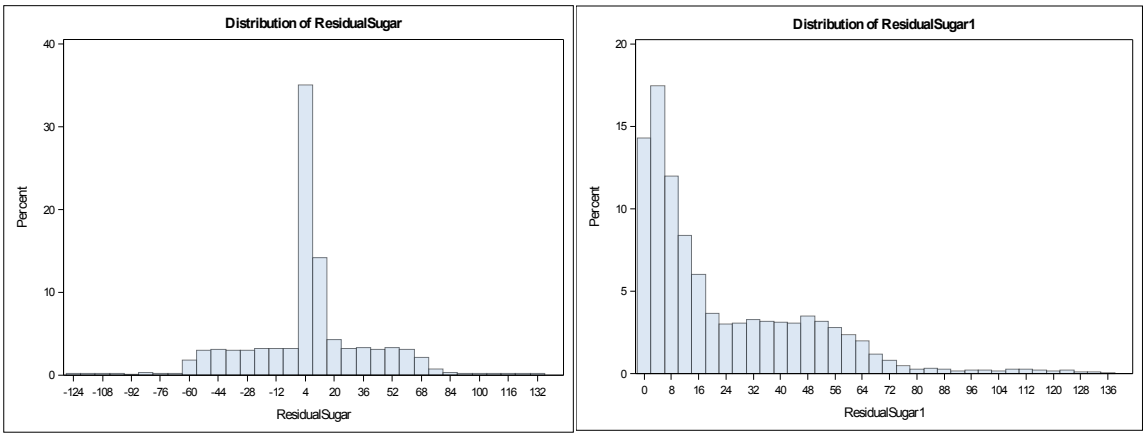
b) Correcting variables

The main correction deals with the negative values. The summary of the negative values and the corrective action is below

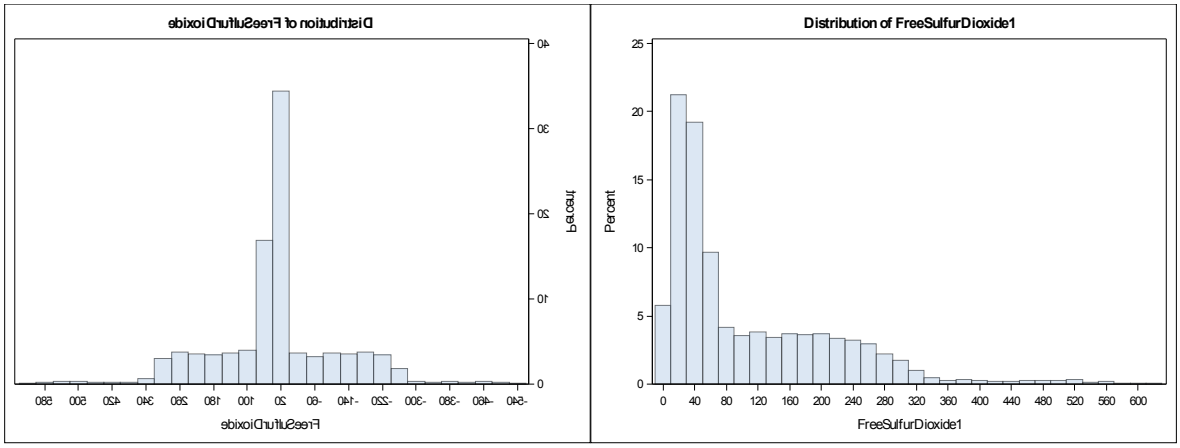
Variable	# negative values	Percentage	Corrective Action
FixedAcidity	1621	12.67%	Absolute Value
VolatileAcidity	2827	22.09%	Absolute Value
CitricAcid	2966	23.18%	Absolute Value
ResidualSugar	3136	24.51%	Absolute Value
Chlorides	3197	24.99%	Absolute Value
FreeSulfurDioxide	3036	23.73%	Absolute Value
TotalSulfurDioxide	2504	19.57%	Absolute Value
Sulphates	2361	18.45%	Absolute Value
Alcohol	118	0.92%	Absolute Value
LabelAppeal	3640	28.45%	Adding 3 to all values

As indicated above, many of the variables appear to be within a reasonable range of the positive-valued population when the absolute value is taken. I am therefore assuming that a number of records in this set have been incorrectly labelled as negative.

An example of this pattern is shown below, the first graph shows the distribution of the raw data for **Residual Sugar**, while the second shows the distribution after the data transform.

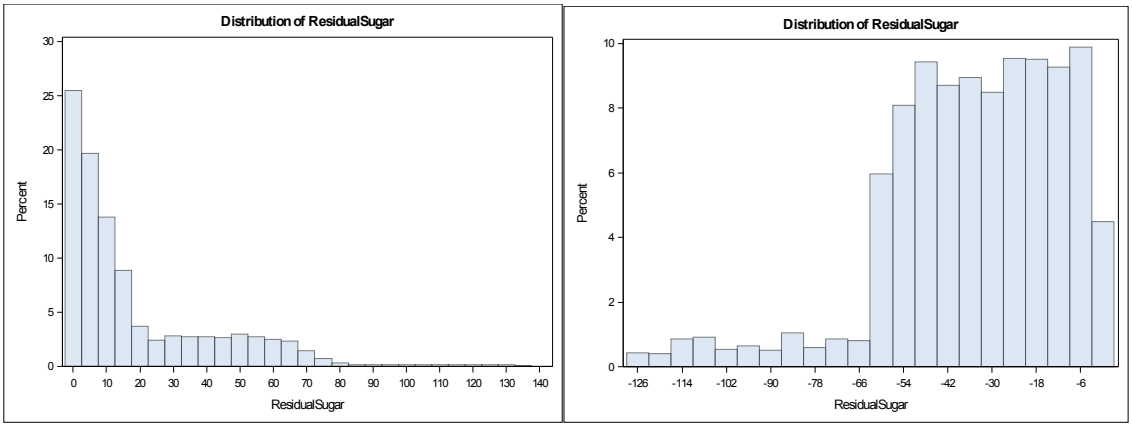


The case below concerns the distribution of **free Sulphur Dioxide,** again with the second graph showing the distribution after the transform



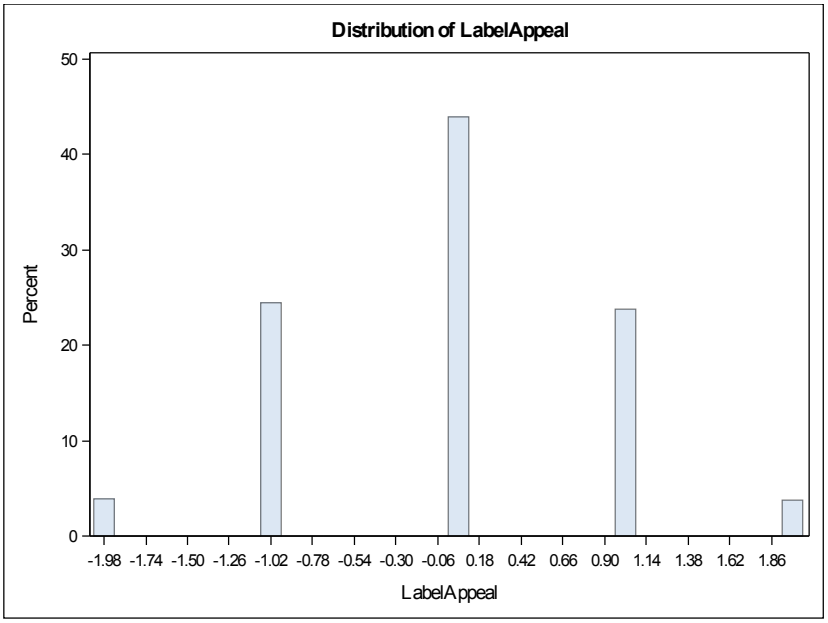
I would argue, in both of these cases, the second graph is more consistent with the real-world expectation. For the distribution of Residual Sugar, or a (albeit weakly correlated) indicator of the sweetness of the wine. For the free sulphur dioxide, the corrected distribution looks more natural: there is naturally occurring free Sulphur dioxide, while winemakers will add it to high quality wines in addition to improve the wine.

To further support this transform, below are the distributions of the values for the negative and positive values of Residual Sugar. The negative values drop away at a very similar point (by absolute value) in both graphs.



The second pattern, adding 3 to the labelAppeal record (and therefore concluding the distribution is accurate but is distributed around zero) is as follows. Adding 3 simply makes the data more interpretable.

- The raw data shows a strong correlation with the stars, and other methods (such as taking the absolute value of the negative values) results in a significant reduction in this correlation
- The raw data approximates a normal distribution as I would expect (as below), and not one that is skewed in any way as a result of incorrect data.



c) Derived variables

There is the possibility of deriving variables for the purpose of reducing dimensionality, making any model easier to understand, and to possibility reduce mutli-colinearity and therefore reduce it side effects during model building. Judging by the lack of correlation within the raw predictors however, there is no value in pursuing PCA or similar in order to do this. This is largely supported by the business meaning of the predictors; there is little in my investigations of wine chemistry to suggest clear correlation amongst the predictors.

There was one suggestion that a good wine would maximise the ratio of FreeSulphurDioxide to bounded Sulphur Dioxide. I explored the possibility of this being a strong predictor of wine sales using the following formula:

```
BoundedSulfurDioxide = TotalSulfurDioxide-FreeSulfurDioxide;

RatioFreeSDToBoundedSD  = FreeSulfurDioxide/BoundedSulfurDioxide ;
```

I found that there was almost no correlation with the TARGET variable however, so I discarded this derived variable during the data exploration process.

Given the new category of the STARS predictor, I have used dummy variables so as to represent this feature so as it can be used as part of any OLS regression I choose to do. The SAS functions for Poisson and negative binomial regression appear to handle categorical variables.

d) Mathematical transformations

I have explored square roots, squares and natural logs of the predictor and the response variables in order to see if there was a noticeable increase in correlation with the response variable.

Log transformation served to improve correlation significantly with the predictor variable for **Volatile Acidity and free sulphur dioxide**, while existing strong correlations were either weakened or only marginally improved. Therefore these two log transformations would be retained for future

use in modelling however this improvement is seen before the absolute values of this predictor are taken, and would result in negative values not being predicted.

Square transformation improved the correlation of the target variable with STARS (from 0.56 to 0.61) however in the interest of parsimony, and that STARS is a categorical variable, I ignored this marginal improvement. No other variable was worthy of pursuing under this transformation.

Square root similarly offered no significant improvement that was not already captured by the log transform.

3. Build Models

I have chosen to use the following ways of deriving a set of possible models:

a) All original variables (14) in a Possion regression, and then compared to minimum of strongly correlated predictors (2), and then compared to a medium level of the best correlated predictors (5)
b) All original variables in a Negative Binomial regression, and then compared to minimum of strongly correlated predictors, and then compared to a medium level of predictors.
c) All original variables in a Zero Inflated Poisson regression, and then compared to minimum of strongly correlated predictors, and then compared to a medium level of predictors. In all cases STARS is used as the zeromodel predictor as it is the most closely correlated predictor with the TARGET variable
d) All original variables in a Zero Inflated Negative Binomial regression, and then compared to minimum of strongly correlated predictors, and then compared to a medium level of predictors. In all cases STARS is used as the zeromodel predictor as it is the most closely correlated predictor with the TARGET variable
e) Finally, a standard OLS regression using Stepwise, forward and backward variable selection methods

The selection criteria for these models I intend to look at the AIC, AICC, BIC the number of predictors, as well as Root mean squared error (RMSE) performance on an out of sample test.

I chose 20% of the data (randomly) to make up the out-of-sample test.

Model type	Selection method / Variables	# variables in model	AIC	AICC	BIC	Out-of-sample RMSE
Poisson	Complex	14	36466.3392	36466.5591	36705.0501	1.311017
	Medium	5	36468.5139	36468.6314	36642.1219	1.311621
	Low	2	36842.1130	36842.1306	36907.2160	1.313941
Negative Binomial	Complex	14	36468.3392	36468.5725	36714.2838	1.311017
	Medium	5	36470.5139	36470.6412	36651.3556	1.311621
	Low	2	36844.1130	36844.1345	36916.4496	1.313941
Zero Inflated Poisson	Complex	14	33341.5017	33341.7924	33616.3810	1.305949
	Medium	5	33345.9421	33346.1126	33555.7185	1.306066
	Low	2	33371.5744	33371.6155	33472.8458	1.316209
Zero Inflated Negative Binomial	Complex	14	33343.5019	33343.8080	33625.6149	1.305949
	Medium	5	33347.9424	33348.1247	33564.9524	1.306066
	Low	2	33373.5746	33373.6215	33482.0795	1.316209
OLS Regression	Stepwise	11	5532.94		5534.97	1.30955
	Forward	15	5536.61		5538.66	1.30953
	Backward	11	5532.94		5534.97	1.30955
	All	18*	5540.29		5542.36	1.30964

*this is higher as the STARS variable is treated as categorical

4. Select Models

My objective in choosing the model are as follows:

- The model should be meaningful, the predictor variables should contribute as one would expect for the real-world meaning of the predictors
- The model should perform well in in-sample testing based on AIC, AIC, and the BIC
- In a choice between similarly performing complex versus parsimonious, the parsimonious model is preferred.
- The model should perform well on out-of-sample testing, measured by the RMSE

The best performing model, on the basis of the measures in the section above, is the medium complexity model using the zero inflated negative binomial model. While the complex version of this model performed marginally better on the measures of fit and out of sample test set, the more simple version is preferred as the betas are more consistent with real world understanding.

The betas in the model generally make sense and are consistent with the findings of the data step:

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	1.5828	0.544	0.5165	2.6491	8.46	0.0036
FreeSulfurDioxide1		1	0	0.0001	-0.0001	0.0001	0	0.9619
VolatileAcidity1		1	-0.0154	0.0106	-0.0361	0.0054	2.1	0.1468
LabelAppeal1	1	1	-0.9758	0.049	-1.0718	-0.8798	396.69	<.0001
LabelAppeal1	2	1	-0.6085	0.0287	-0.6647	-0.5523	450.82	<.0001
LabelAppeal1	3	1	-0.3398	0.026	-0.3908	-0.2888	170.44	<.0001
LabelAppeal1	4	1	-0.157	0.0263	-0.2085	-0.1056	35.76	<.0001
LabelAppeal1	5	0	0	0	0	0	.	.
AcidIndex	4	1	0.1866	0.705	-1.1951	1.5684	0.07	0.7912
AcidIndex	5	1	0.3462	0.5483	-0.7285	1.4209	0.4	0.5278
AcidIndex	6	1	0.3594	0.5446	-0.7079	1.4268	0.44	0.5092
AcidIndex	7	1	0.3324	0.5444	-0.7346	1.3993	0.37	0.5415
AcidIndex	8	1	0.3178	0.5444	-0.7492	1.3847	0.34	0.5594
AcidIndex	9	1	0.2715	0.5446	-0.796	1.3389	0.25	0.6182
AcidIndex	10	1	0.1916	0.5456	-0.8777	1.2609	0.12	0.7254
AcidIndex	11	1	0.0081	0.5502	-1.0702	1.0865	0	0.9882
AcidIndex	12	1	0.097	0.5556	-0.992	1.186	0.03	0.8614
AcidIndex	13	1	0.2235	0.5552	-0.8646	1.3116	0.16	0.6873
AcidIndex	14	1	0.2392	0.5737	-0.8853	1.3637	0.17	0.6767
AcidIndex	15	1	0.3677	0.6226	-0.8525	1.5879	0.35	0.5548
AcidIndex	16	1	-19.4828	7407.953	-14538.8	14499.84	0	0.9979
AcidIndex	17	0	0	0	0	0	.	.
STARS	0	1	-0.3843	0.0288	-0.4408	-0.3278	177.49	<.0001
STARS	1	1	-0.3294	0.0252	-0.3788	-0.28	170.76	<.0001

STARS	**2**	1	-0.2055	0.0229	-0.2503	-0.1606	80.77	<.0001
STARS	**3**	1	-0.1021	0.023	-0.1472	-0.057	19.68	<.0001
STARS	**4**	0	0	0	0	0	.	.

The betas make sense, as reconciled with the learnings during the data exploration. The most significant predictors are the Stars and the Label Appeal, with a reduction in each reducing the impact on the target variable. The acid Index suggests two optimal points of acidity, around 6 or 15, with anything over 15 having a drastic impact on sales.

Note for completeness, the estimate of the zero inflation parameter are below:

Analysis Of Maximum Likelihood Zero Inflation Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-16.2029	152.5076	-315.112	282.7064	0.01	0.9154
STARS	**0**	1	16.5102	152.5076	-282.399	315.4195	0.01	0.9138
STARS	**1**	1	14.5434	152.5076	-284.366	313.4527	0.01	0.924
STARS	**2**	1	10.2747	152.5104	-288.64	309.1896	0	0.9463
STARS	**3**	1	0	171.5596	-336.251	336.2505	0	1
STARS	**4**	0	0	0	0	0	.	

Conclusion

In doing this analysis, I have explored and the wine data set in the context of the real-world domain. Generally the quality of the data was good, however there were some interesting patterns in the data, for example the zero-inflation of the target variable, or the number of negative values in the data where they would not normally be expected which needed to be corrected.

There appeared to be little value in pursuing mathematical transformations of the data. There appeared to be some gain in correlation, however there is a trade-off in terms of interpretability which I was unwilling to make for these marginal improvements.

Poisson and Negative Bionomial seemed to perform more poorly than Zero-inflated models on this dataset, this is consistent with the findings about the target variable in the data exploration step. The final model was a winner because of the mode limited number of predictors needed to achieve the same, or very slightly worse, performance than the best performing model.

The final model had an error rate (RMSE) of 1.306 on the out of sample test, contained 5 variables (of which 2 are categorical) and were consistent with the real-world intuition and findings in the data set during data exploration.

Therefore the model is reasonably accurate, a number of steps have been explored to arrive at this accuracy, and the model has a high level of interpretability, consistent with real-world intuition.

References:

- Waterhouse Lab, Fixed Acidity: http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity
- Wikipedia: Acids in Wine: https://en.wikipedia.org/wiki/Acids_in_wine
- Chloride concentration in red wines: influence of terroir and grape type : http://www.scielo.br/scielo.php?pid=S0101-20612015000100095&script=sci_arttext
- Wikipedia: Sweetness of Wine: https://en.wikipedia.org/wiki/Sweetness_of_wine
- Waterhouse Lab, Sulphites in Wine: http://waterhouse.ucdavis.edu/whats-in-wine/sulfites-in-wine
- MorethanOrganic.com, Sulphites in Wine: http://www.morethanorganic.com/sulphur-in-the-bottle