Michael Venit

MSDS 410

Modeling Assignment #4

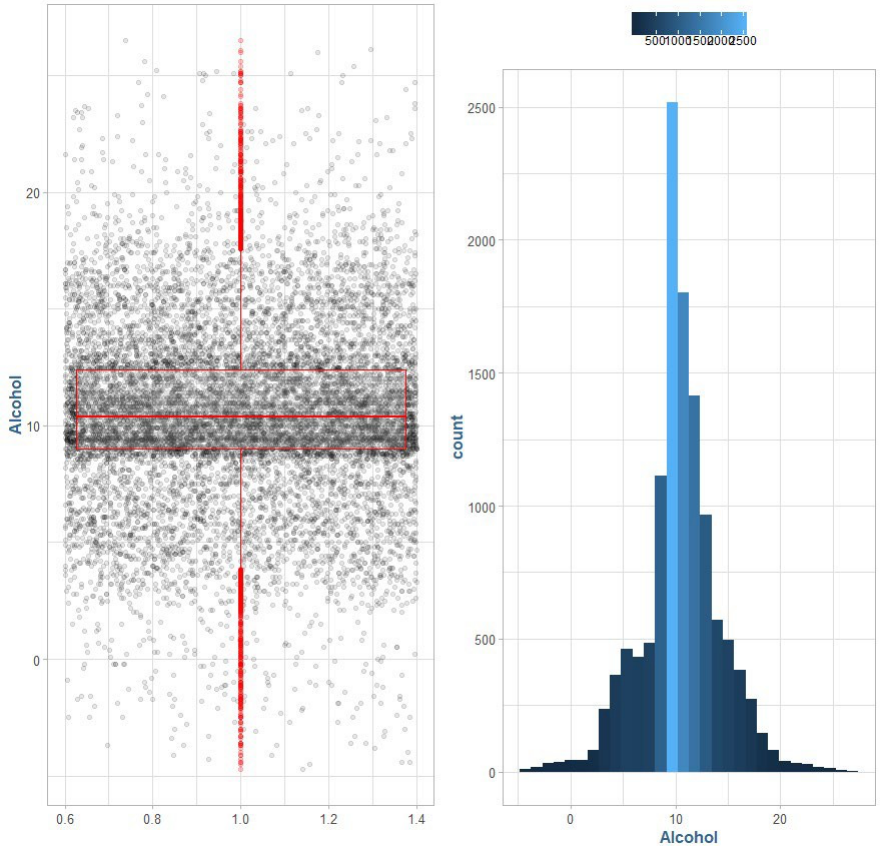
Introduction

In this lab a data set consisting of various attributes of approximately twelve thousand commercially available wines will be explored in order to predict if they will be sold, if so, how many cases will be sold, and the number of stars an expert would give this wine using a Vivino style rating system. The ultimate purpose of this data analysis and the resulting model is to provide a large wine distributor client our recommendations on what kind of wines will be ordered and in what amounts so that we can be operationally prepared to manage the supply chain and resulting required logistics. Additionally, quantifying what makes a wine 'good' based upon its chemical composition and characteristics will be attempted.

Some preliminary exploratory data analysis will be necessary to look for distribution characteristics of the independent variables and their statistical properties. From there, the data set will be divided into different sets based upon the type of statistical model that will ultimately be produced and the various response variables of interested. There are three response variables for explanation throughout this lab: purchased, cases sold, and stars (rating). These are somewhat related variables; thought they require different statistical procedures to model correctly.

EDA

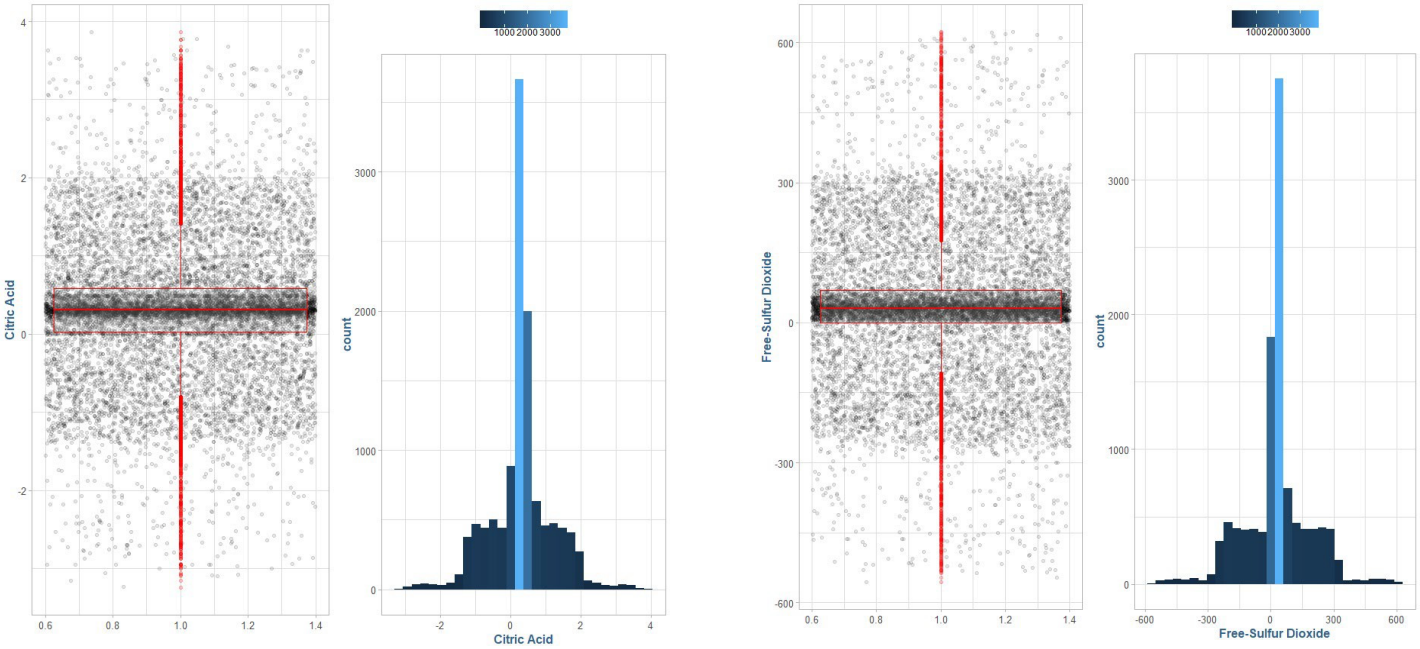
First, it is essential to perform a clean sweep on the data to find any bad encodings or invalid values. Inspection of the continuous variables leads to some interesting questions on exactly what scales of representation these measurements are in. Alcohol, for example, appears to be in a measurement of percent volume, as we see the values lie in the range of what we would expect the alcohol content to be by volume (ABV).



Given this measurement, the 771 bottles with missing or negative values will be removed.

The additional continuous variables in this data set appear to be on a normalized scale (approximately -4, 4), given that they are measurements of the contents of various chemicals and

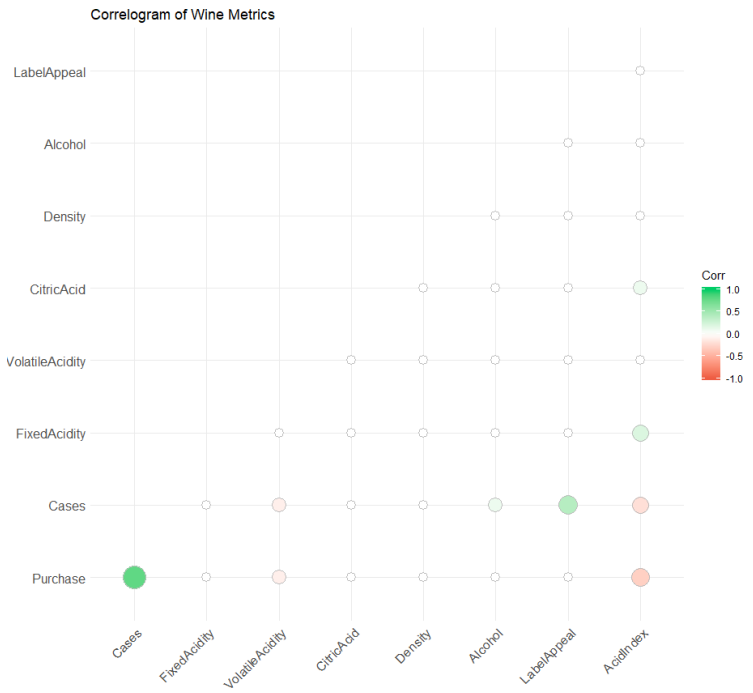
the distributions are centered near zero with equal parts above and below the mean (in the positive and negative direction). Citric Acid, Sulphates and Volatile Acidity embody this quality when looked at closely. Half of the values are in the negative range for so many variables that represent measurements of a quantity, leading to this assumption.



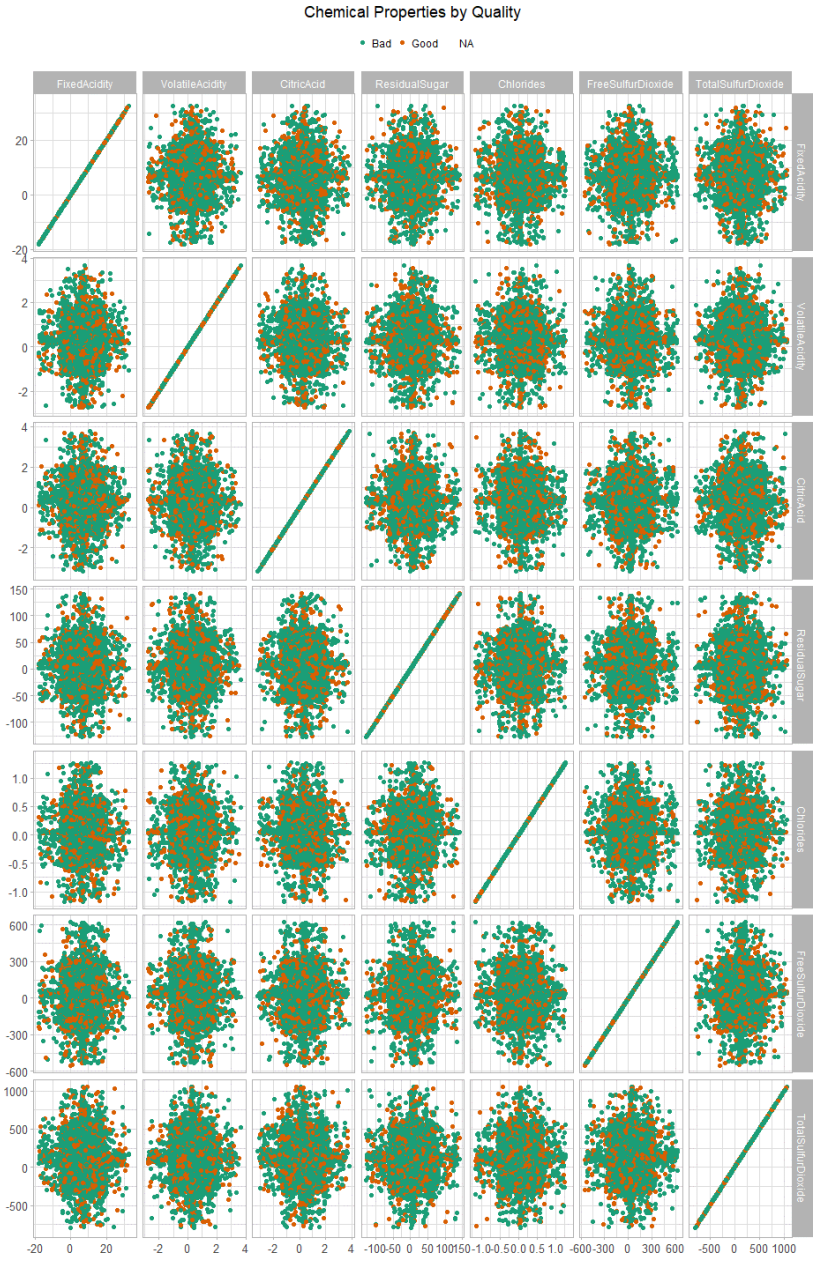
Some of the other variables have similar properties (i.e. Free-Sulfur Dioxide, Chlorides and Residual Sugar). However, these features present miniscule differences in scales. Hence, these will be standardized appended to our data set. By performing standardization, the actual properties of the values underlying distribution, will go unchanged and it will simply help later with coefficient interpretations related to the chemical properties by having all the variables on a similar scale and value range.

The label appeal variable has the scale -2 to 2, which will be adjusted by +3 so that it is on a standard 1-5 scale. A simpler metric in terms of quality based on the STARS variable will be created, which is a binary ‘good’ or ‘bad’. The ‘good’ designation will be assigned if the STARS rating is greater than or equal to 3. This will help uncover basic relationships by condensing the rating scale, then once we find some basic patterns, we can explore them in-depth with the more complete rating variable.

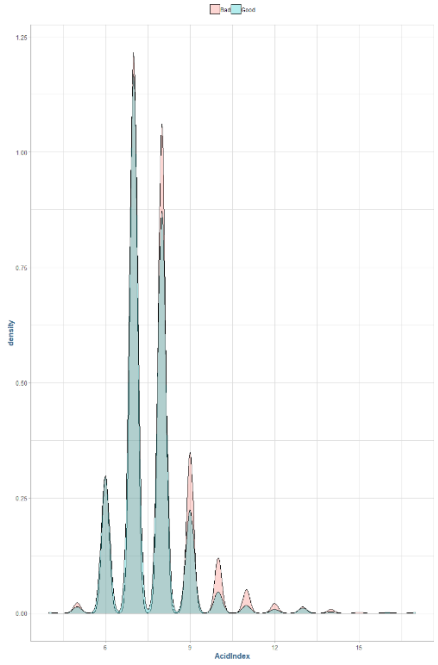
It is important to also look for some correlation amongst the features in the dataset which will hopefully expand the analysis by raising some additional questions. First, it should be known if there are any correlations to the STARS rating variable. The most obvious correlation to the STARS variable, at 29%, is the purchase variable, which we would intuitively expect given that the higher the quality, the more it sells. Surprisingly, there exists a relatively strong correlation to the label appeal, at 33%, which indicates that the better the presentation the higher the rating and is something that will be explored more in depth later.



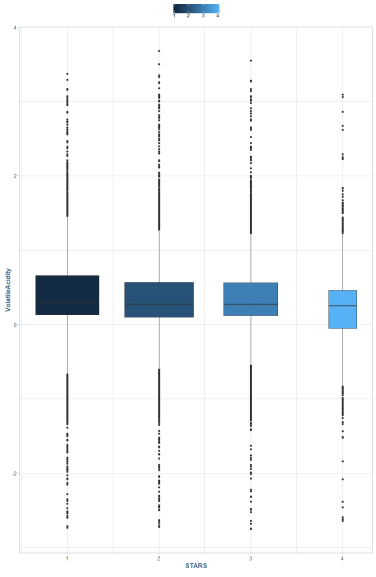
Next, it is imperative to find relationships between the high-level quality variable we defined to look for any chemical relationships that may help determine what makes for a ‘good’ wine. In the following visualization below, we can see a detailed breakout for each chemical property and the quality variable:



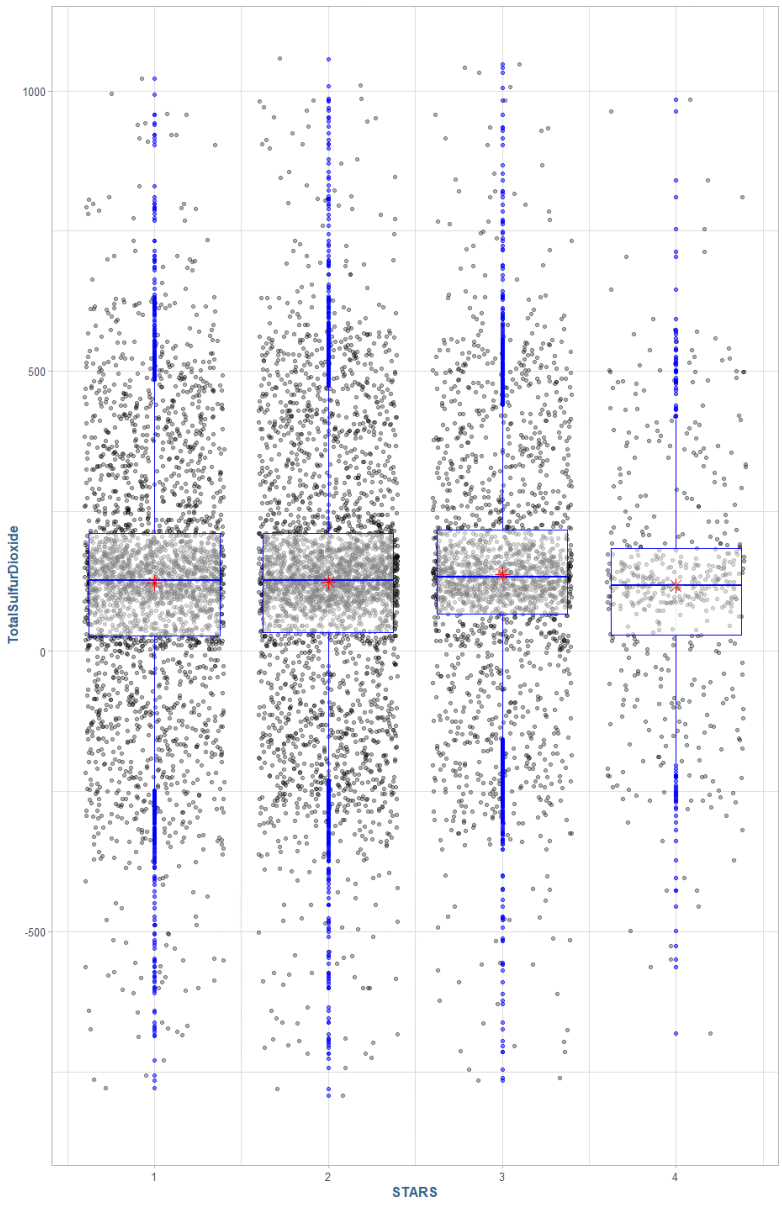
A noteworthy finding from this diagram shows that the acid index property appears to have a negative effect on wine quality as the acidity increases. We can see an overlay of ‘good’ and ‘bad’ wines by acid index in the following diagram:



There also looks to be a lower concentration of volatile acidity in higher rated wines, as can be seen in the following diagram below:

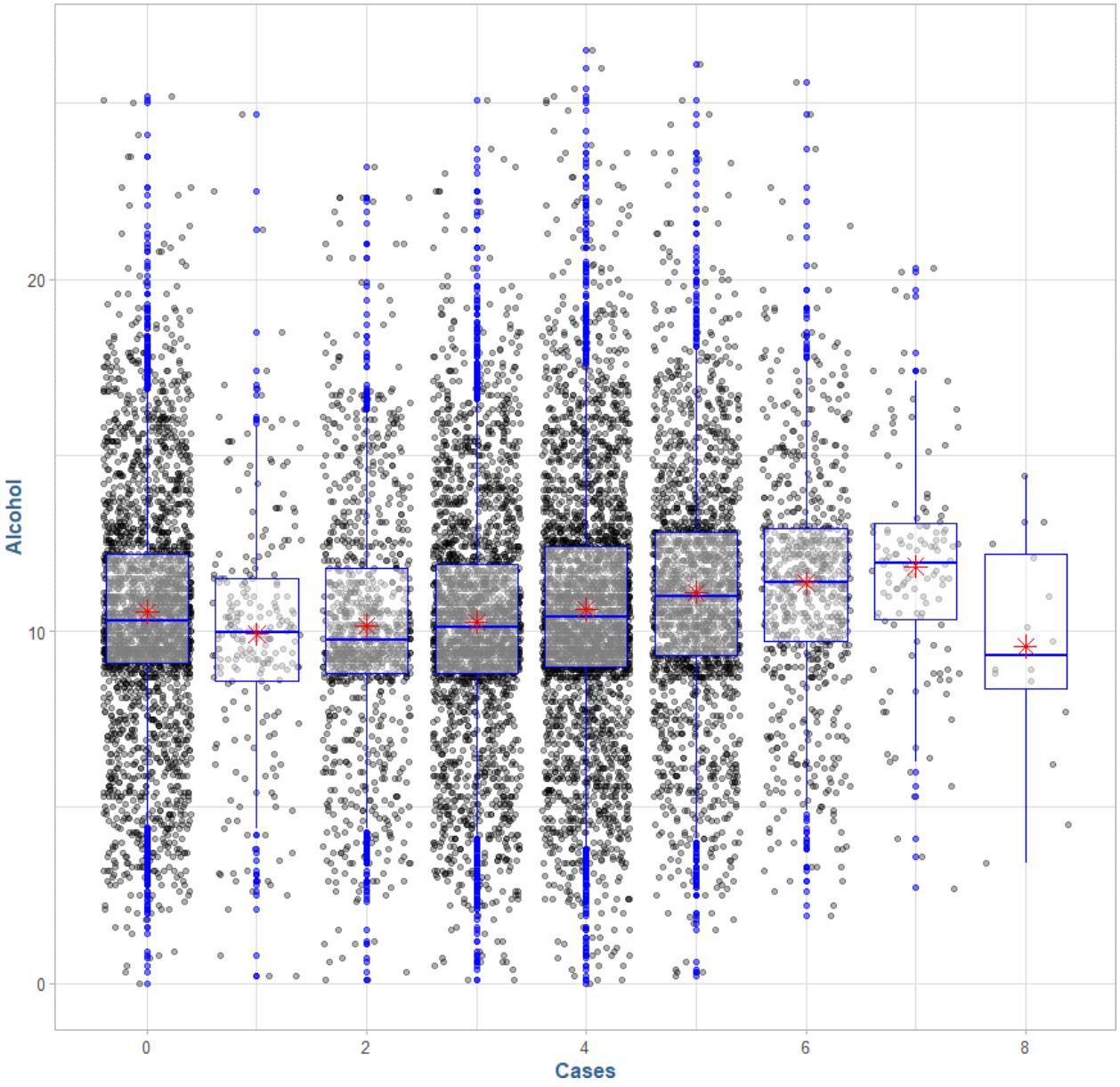


Additionally, there are slightly lower levels of sulfur dioxide in higher rated wines:



The rest of the chemical composition properties appear to be equally distributed across their various rating groups and there is little visual evidence to suggest there are meaningful differences.

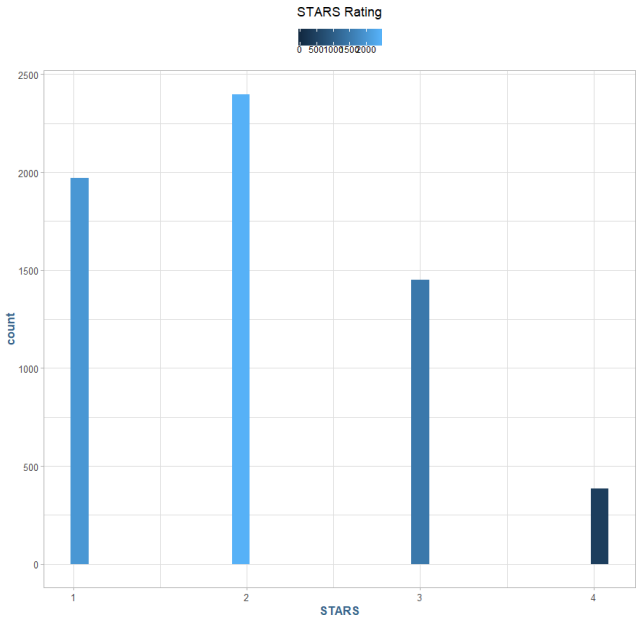
Another relationship that should be noted is between the number of cases sold and the alcohol content of the wine. It appears that there is a moderate relationship between the higher alcohol content wines and the number of cases sold. There is a drop off in the 8-case range, however, there is also just a relatively small sample size to draw upon as not many wines sell that many cases.

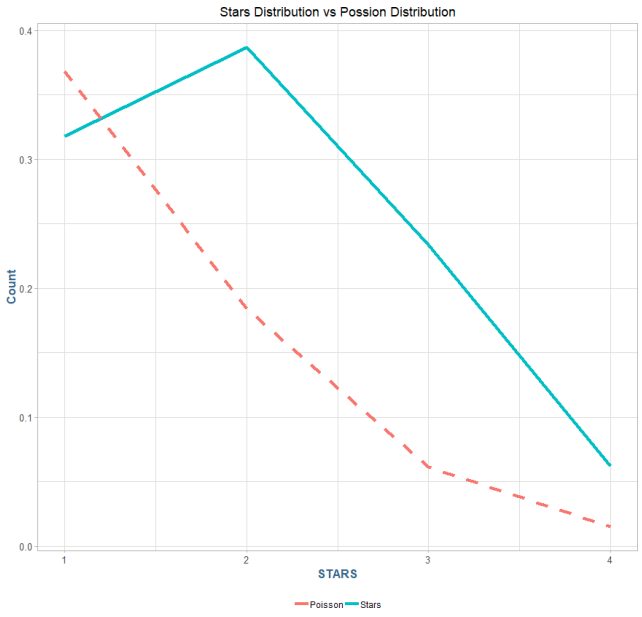


Stars Model

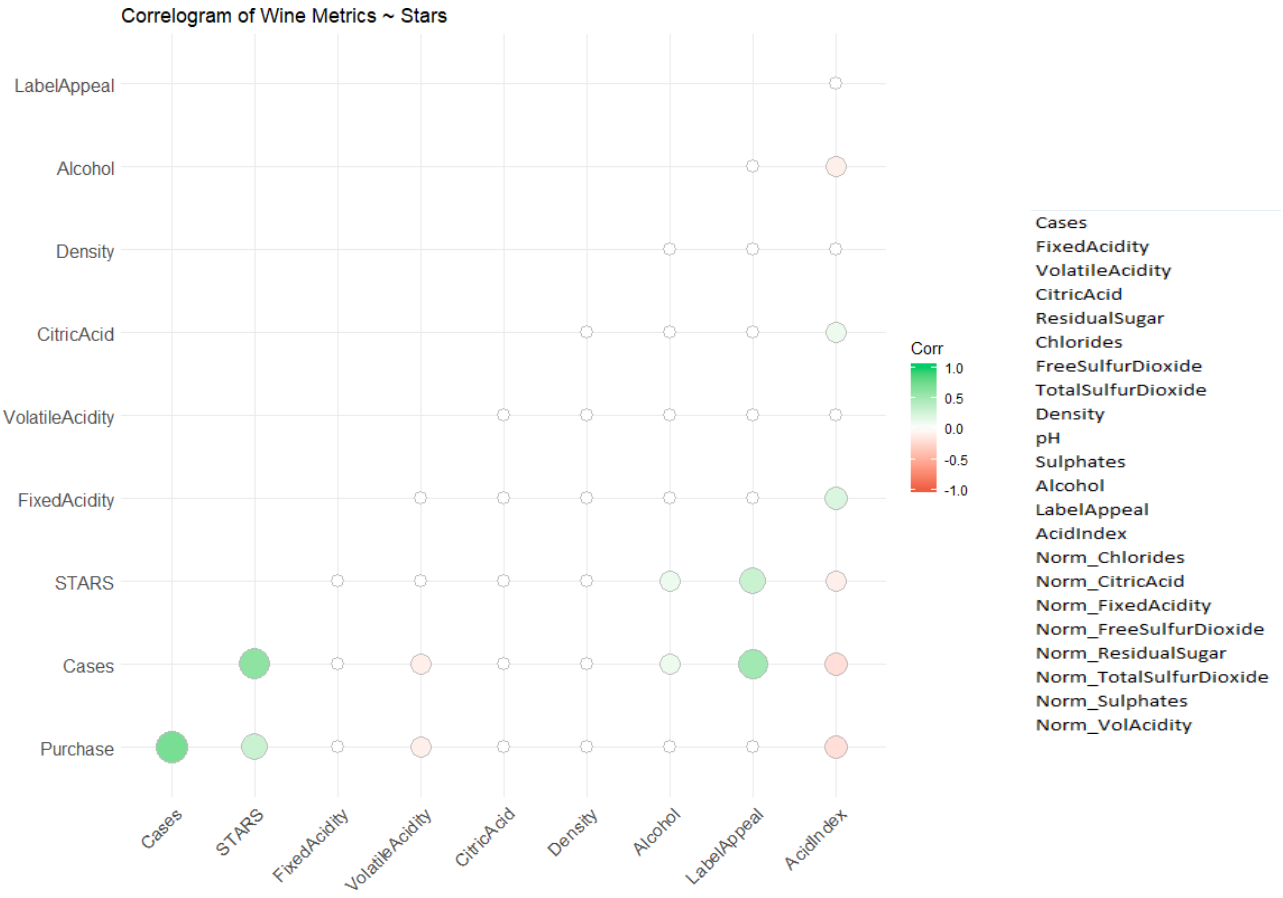
In this section I have chosen to explore the data further as it relates to the stars rating received by the wine. For this model, I would like to further refine the data set to exclude any wines that do not have a star rating (3,150 bottles dropped for this analysis total). Following that step, the cleaned data will then be partitioned into a standard 70/30 split to evaluate the model against the 30% hold-out set later

First, observing the distribution of the stars rating to get an overall idea how the scores are distributed across the bottles of wine is essential. The distribution is skewed to the right, where the majority (approximately 70%) of our wines have a "bad" rating of one or two. These star counts do not exactly follow a Poisson distribution, it is relatively close in general form as can be seen below:





I have chosen to favor the Poisson distribution over the Gaussian distribution as the goal is to predict discrete scale data instead of continuous. A correlation matrix on the stars data was assessed again, given that we have pulled a smaller subset of the wines and removed any wines without a star rating, so this could yield some additional relationships:



Looking across the STARS row, there is slightly negative relationships to volatile acidity and acidity index, while there exists a positive correlation to cases and label appeal, as we would have expected from our prior analysis. Given that “cases” is the notable variable here, in terms of being associated with the stars rating, it would be expected that various feature selection techniques will favor this variable.

Using forward, backward and step-wise selection methods on generalized linear models in the Poisson family, we see that all three techniques select the same "optimal" model: a single variable model based upon cases.

STARS Model 1

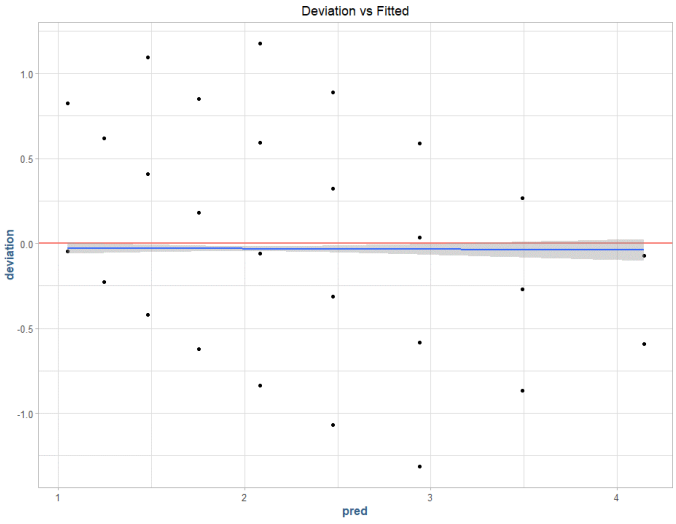
$$\hat{Y} = 0.05 + 0.17\beta_1$$

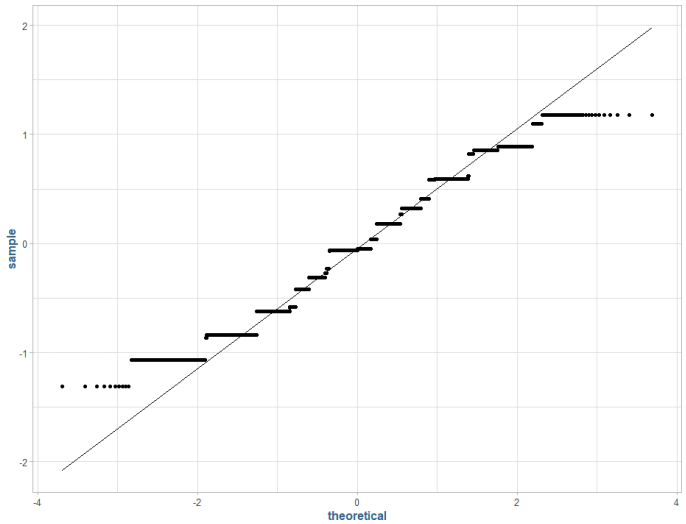
Where β_1 is the number of cases of the wine has sold. This means that for each 1 unit of cases sold, the estimated rating of the wine increases by .17 points. For the diagnostics of this model, which are the same since all techniques produced the same model, are:

AIC: **12,612**, Null deviance: **1,798**, Residual deviance: **1,213**

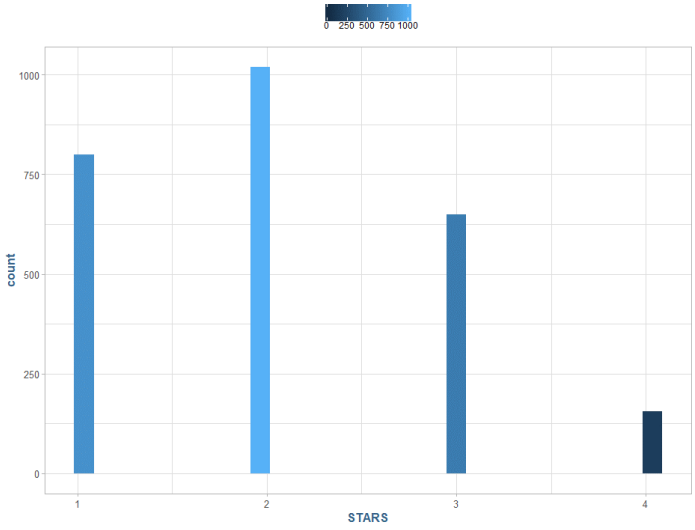
MSE**: 0.74**, R_2: **31%**, MAE: **.59**

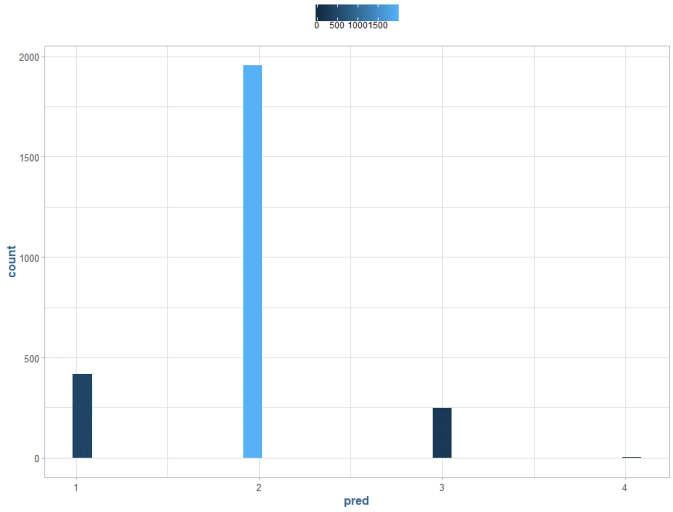
Visually, the following observes the deviation v. fitted and a QQ-plot of the deviations from the model on the training data:





As for the out of sample predictions, the model correctly predicted the stars rating for approximately 49.3% of the bottles of wine we kept out of sample in the hold-out set. Below is a look at the actual ratings distribution in the test set v. the predictions made on the test set:





There is clearly a bias prediction for bottles of wine with a two-star rating, or an over convergence to the mean in the predictions. Also, if the accuracy level were broken down by star rating produced by the model:

STARS	PctAccurate
1	39.682540
2	84.289746
3	16.747967
4	1.025641

There seems to be a substantial deviation between two-star wines and the rest of the sample. Since this model can be considered somewhat brittle due to the reliance on one single factor, the number of cases sold, this will be re-visited in the entire modeling process and I will hold out the cases sold variable to look for additional statistical measures.

The second round of variable selection again yielded the same parameters and identical models in forward, backward and stepwise techniques.

STARS Model 2

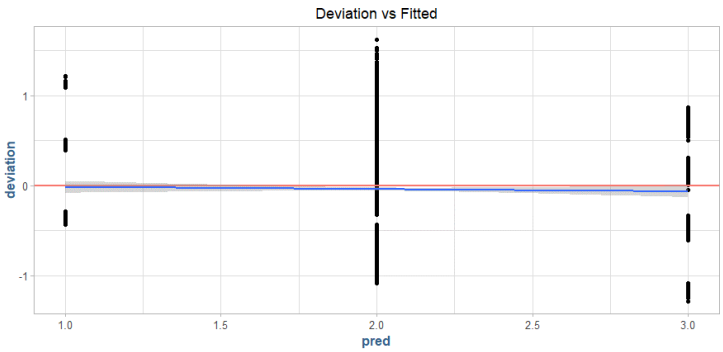
$$\hat{Y} = 0.46 + 0.007\beta_1 + 0.16\beta_2 - 0.04\beta_3$$

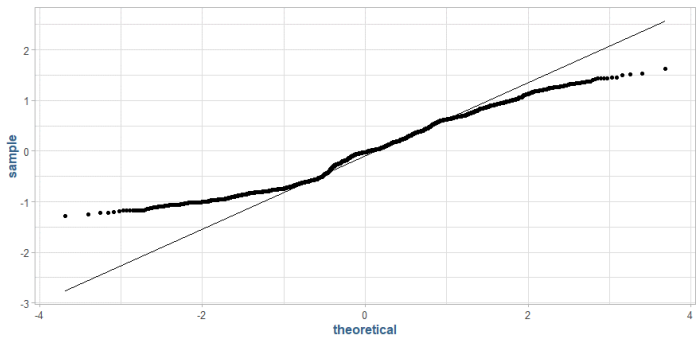
Where a one unit increase in alcohol content by volume (ABV) results in a 0.007 increase in stars rating, a one unit increase in label appeal results in a 0.16 increase to the stars rating, and a one unit decrease in acidity index results in a 0.04 increase in stars rating.

AIC: **12,783**, Null deviance: **1,739**, Residual deviance: **1,549**

MSE**: 0.84**, R_2: **11%**, MAE: **.69**

By all accounts, this is indeed an inferior model than the one that relies on the number of cases sold alone. Visually, the deviations vs the predictions and QQ-plot of the deviations can be seen below:



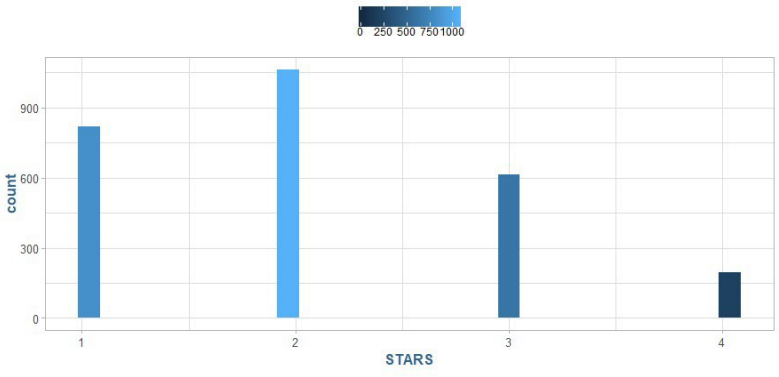


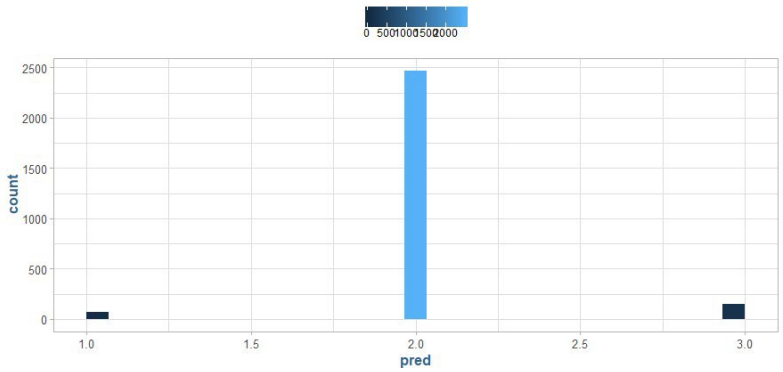
The second model correctly predicted the rating 41.7% for all the bottles in the out of sample test set. By category, a significant breakdown in the model can be observed when it comes to four-star bottles of wine:

STARS	PctAccurate
1	6.349206
2	94.920038
3	8.292683
4	0.000000

However, the results for predicting an average two-star bottle of wine is extremely accurate, which is not unlike the results achieved from in the previous iteration. Most notably, this model predicted absolutely none of the four-star wines correctly, which depending on the business case could be our most important metric.

Below, the distribution of the predictions can be seen for the second iteration of the model:





The final attempt at the stars rating will attempt to combine the previous two models in order to get a more robust modeling procedure.

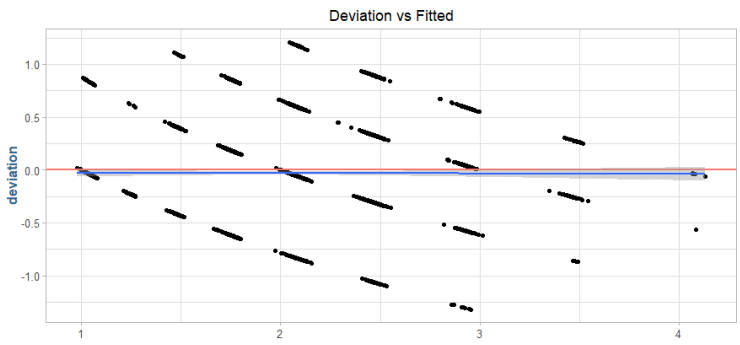
STARS Model 3

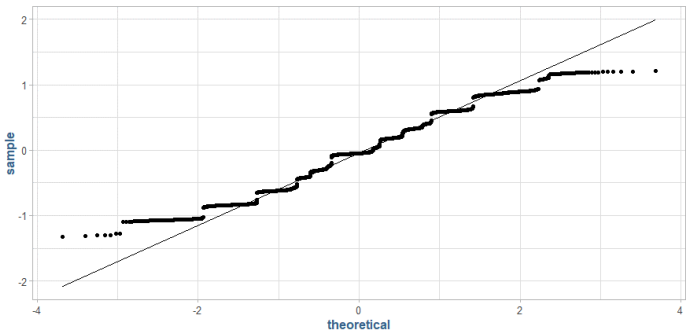
$$\hat{Y} = 0.10 + 0.17\beta_1 + 0.002\beta_2 - 0.002\beta_3 - 0.007\beta_4$$

Where a one unit increase in cases sold represents a .17 increase in rating, a one unit increase in alcohol content by volume (ABV) results in a 0.002 increase in stars rating, a one unit increase in label appeal results in a 0.002 decrease to the stars rating, and a one unit decrease in acidity index results in a 0.007 increase in stars rating.

AIC: **12,414**, Null deviance: **1,739**, Residual deviance: **1,178**

Which is a slight statistical improvement over the prior models. Visually, the diagnostics of the model show a slight improvement as well:

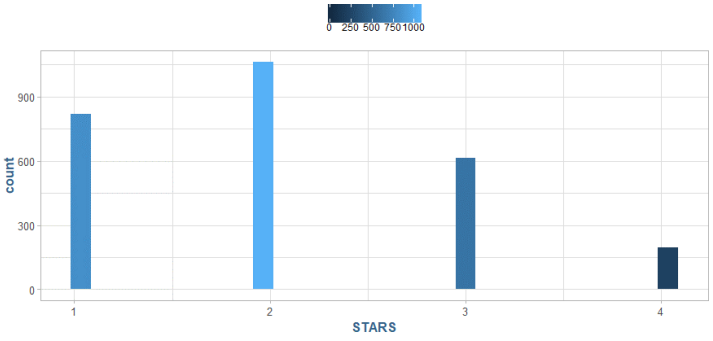


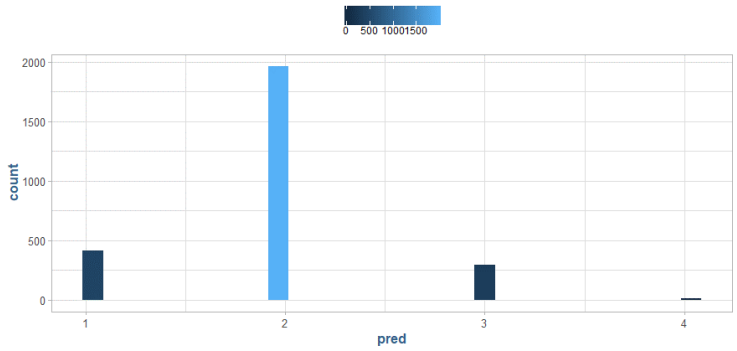


More importantly, the final model has an overall prediction accuracy of 48.7%, which is a substantial increase in prediction 3 of the 4 areas in the stars rating categories:

STARS	PctAccurate
1	36.874237
2	82.878645
3	19.512195
4	4.102564

While the model is clearly not the best it can possibly be, it is a considerable improvement over the prior two iterations. The prediction distributions are still skewed to the two-star wines, however, there is improvement:



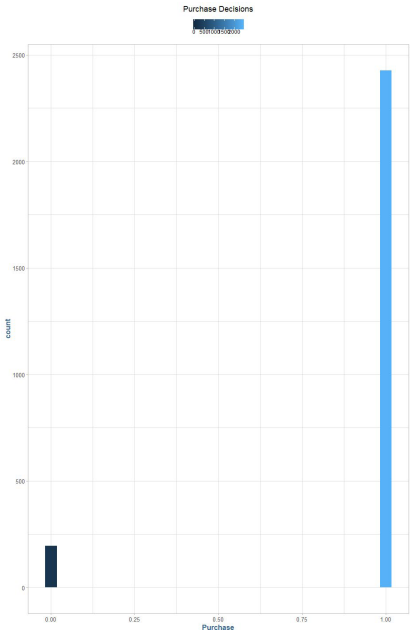


Purchase Model

Now I would like to model the purchased decision variable using the data set at hand. We first note that this variable is a dichotomous yes/no decision variable, and there is another related variable already present in the data set, which is cases, or the number of cases of this wine that has been purchased. Since there is a dependency on the cases and purchase variables, these will

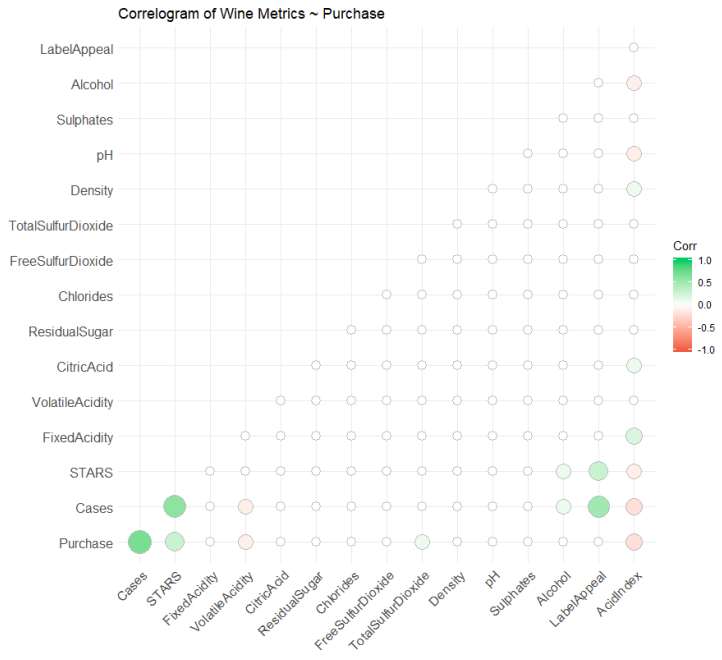
be excluded from the analysis, given that a model based upon data that already has the outcome backed in it would not be useful.

Looking at the purchase variable, we notice that there is an extremely high likelihood of a wine getting purchased according to this data set, at approximately 92.5%, while also noting that any wine we observe with over 2 stars has a 100% likelihood of getting purchased:



STARS	Prob
1	79.12500
2	97.15407
3	100.00000
4	100.00000

A quick survey of the variable correlations to the purchase decision variable would be useful to revisit:



From the visualization above, there is an observed strong correlation to cases, which implies a purchase. However, we also see positive correlations with stars rating, residual sugar, total sulfur dioxide, and negative correlations to volatile and total acidity.

For the purchase model I will employ similar feature selection techniques that was performed with the stars rating model, namely forward, backward and stepwise variable selection. I will change the distribution in the models to the binomial family due to the binary nature of the response. The results of the variable selection process again yield identical models across the data set, and result in the same AIC/BIC statistical fit measures which can be seen below:

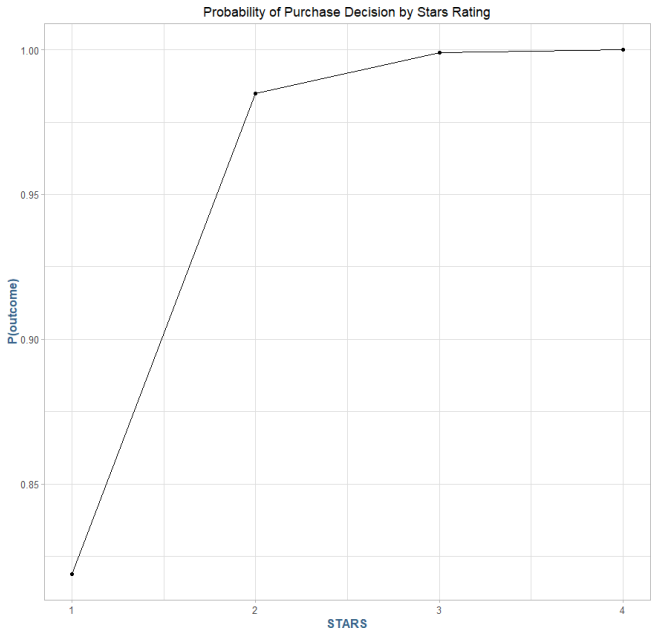
Model	AIC	BIC
Forward Selection	1659.905	1711.056
Backward Selection	1659.905	1711.056
Stepwise Selection	1659.905	1711.056

Purchase Model

$$\hat{Y} = 3.55 + 2.666\beta_1 - 0.178\beta_2 + 0.004\beta_3 + 0.001\beta_4 - 0.038\beta_5 - 0.364\beta_6 - 0.433\beta_7$$

Where the intercept term is an obvious placeholder value for the rest of the model due to the high overall likelihood of a wine being purchased at 92.5% in the sample, and the coefficient here denotes a baseline probability of (*exp*(3.55) =34.8) 34.8%. The first coefficient is the effect the stars rating has on the purchase decision, which increases the probability of a wine being purchases by (*exp*(2.666) = 14.38) 14.3% per level increase in stars rating the wine has. The next coefficient reflects the negative impact volatile acidity has on the purchase decision, which is denoted by (exp(-0.178 = .838), which is a 16.28% decrease per standard unit increase in the volatile acidity scale (normalized values).

Residual sugar is the next coefficient, which has a small impact (*exp*(0.004) = 1.0036) at .36% per unit increase in residual sugar. Total sulfur dioxide is also a small impacting coefficient at (exp(0.001) = 1.001) .11% increase in probability per one-unit increase. Alcohol by volume, label appeal and acidity index all have a negative impact on the probability of an alcohol being purchased at -3.72% per unit increase in ABV, -30.5% per unit increase in label appeal and -35.17% per unit increase in acidity index, respectively.



The probability of purchase decision based upon the stars rating of the wine can be seen above, while holding all other variables constant with their mean values.

For out of sample performance, the probability of getting an accurate prediction for the purchase decision variable can be observed, to find that true purchases are predicted correctly 99.6% of the time on the test set, whereas the do-not purchase value is only correctly predicted 8.9% of the time:

Purchase	PctAccurate
0	8.889
1	99.603

This is most likely due to the overwhelmingly strong presence of “yes” purchase decisions in the data set. In a business context we are more than likely going to care about the percent of accurate result by star ratings, given that it would be more useful in planning inventory and logistics to

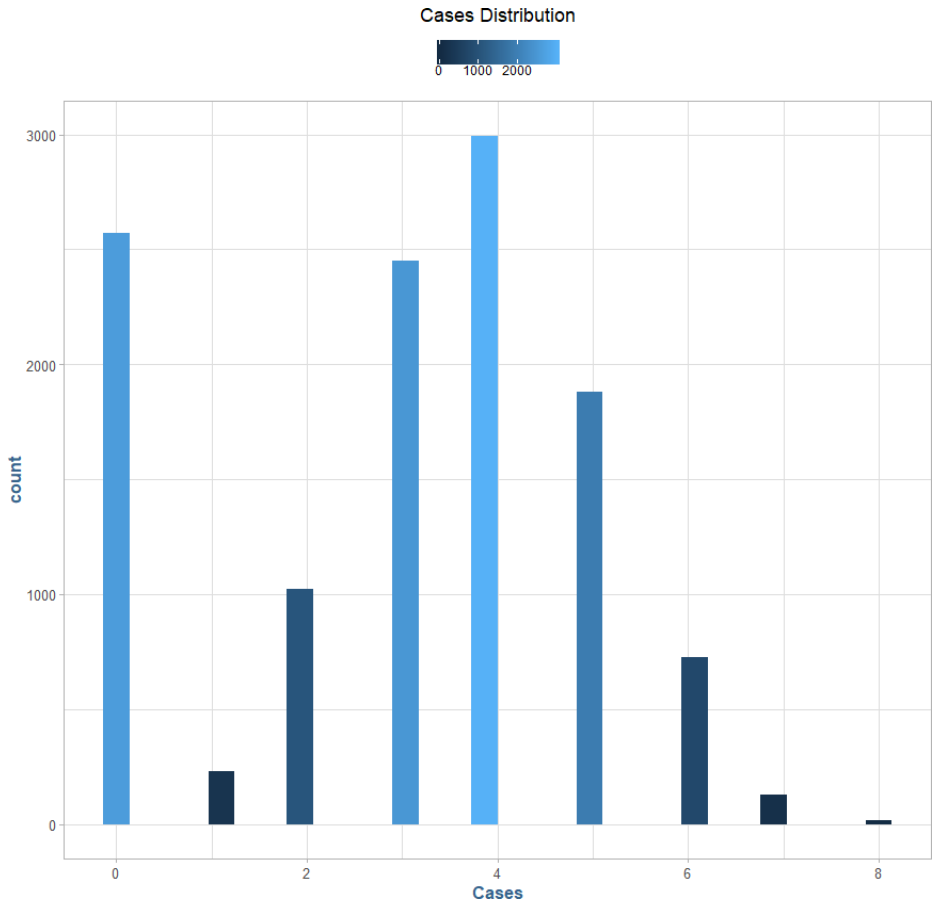
have this information as price of the wine. Although not listed in the data, it would intuitively be highly correlated to the rating of the wine. On the out of sample test data, the model performed extremely well for predicting purchases of three and four-star wines, which would be the top priority:

STARS	PctAccurate
1	80.239
2	97.526
3	100.000
4	100.000

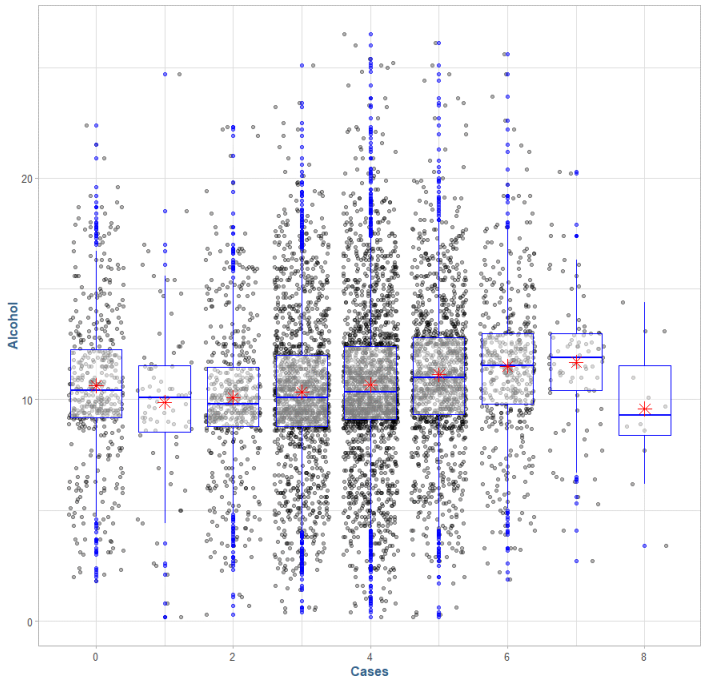
The results for one and two-star wines are not bad, although the one start wines could stand to be improved in further revisions.

Cases Model

The last variable to be examined in this report is the cases feature, which represents how many cases of each wine was ordered. The desired response variable here follows a zero-inflated binomial distribution, which can see visually below:



It can clearly be seen that the dominate presence of zero values in this data set, with the rest of the values being relatively evenly distributed around the mean, mu = 4. Looking at the variables that have a high correlation to number of cases sold, we can see a clear pattern in the alcohol content of the wine to the number of cases sold:



It is noted that there is not a high degree of any correlation between any of the chemical composition properties of the wines to the number of cases sold. For this modeling process, I declined to use the automated variable selection procedures that were employed in the previous two models, and instead start with a "full" model and reduce the coefficients by examining statistical significance and prediction results, dropping variables that do not meet significance criteria.

Cases Model

$\hat{Y}$ =*number of cases (Poisson)*

$$\mathbf{0.483 + 0.116\beta_1 + 0.007\beta_2 + 0.215\beta_3 - 0.018\beta_4}$$

Is Zero (Binomial)?

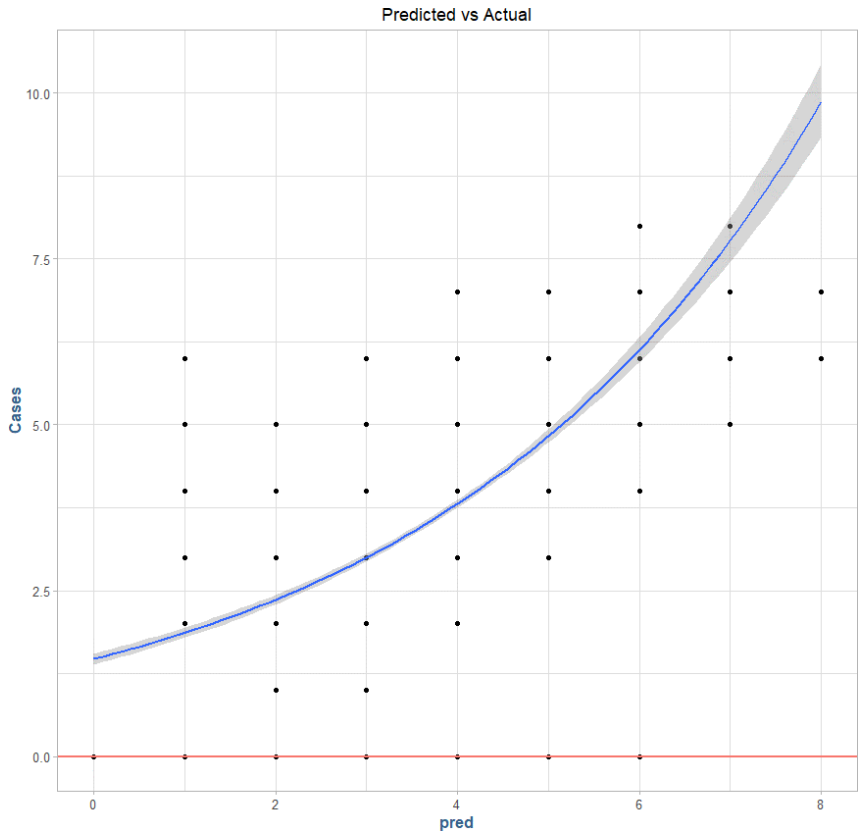
$$\mathbf{-5.011 - 3.830\beta_6 - 0.065\beta_7 - 0.534\beta_8}$$

Where the first part of the model denotes the probability of being zero from the binomial distribution, and the second part determines the number of cases if the binomial model determines the probably of it being 0 is less than 50%.

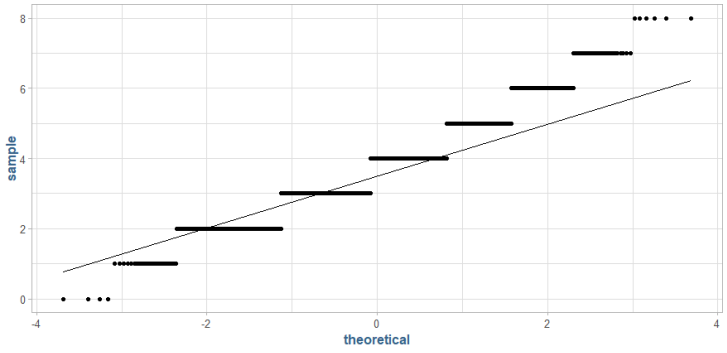
For the cases distribution, we can interpret the intercept of the count (Poisson) model of here as a placeholder value for the rest of the model. With the Poisson model, all the coefficients must be interpreted together to get the estimated count. Here, stars rating can be interpreted as a 12% increase per one unit in rating, .75% increase per unit of alcohol by volume in the wine, 24% increase per one unit of label appeal, a 1.75% decrease per unit increase on the acidity index.

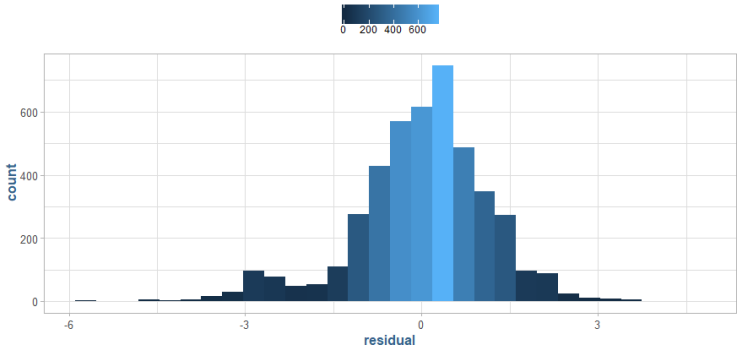
The intercept of the binomial model (-5.011) represents a baseline probability of 99.9% of the number of cases not being zero, with that decreasing 97.83% per one unit increase in stars rating on the bottle, 6.75% per one unit increase in alcohol by volume of the bottle, and 24.03% per unit increase in label appeal. We decrease the probability of being non-zero by 1.75% per unit increase in the acidic index.

Visually, the cases vs the predicted cases can be seen in the following diagram:



And the output from the residuals, which follow the Poisson as we would hope:





Additionally, the model performance on the test data is 40% of the out of sample cases were correctly predicted.

Conclusion

For the purpose of this assignment, I explored a data set containing various attributes of bottles of wine with the goal of producing three distinct models; predicting an ordinal discrete variable (the stars rating), predicting a dichotomous variable that represents a binary yes/no (decision on purchase), and lastly a variable that is a heavily skewed toward zero Poisson (the number of cases a particular wine will sell). This analysis was a difficult one due to the chemical properties data being presented in a pre-normalized form, making it difficult to detect any relationships that might exist here. There appears to be few general correlations amongst the variable in the data set, although the ones that we might expect intuitively did show statistical significance, such as label appeal and cases sold, stars rating and purchase / cases.

As for the three models, they all ended up producing wildly varying degrees of accuracy and robustness. The first model that focused on predicting the stars rating had some very promising results with category one and two wines, however, left much to be desired for the three and four-star wines. The purchase model was overall our best performing model, yielding 100% out-of-sample accuracy for some subsets of the data and a minimum hit rate of 80% in others. The cases model ended up being our overall worst performing model, only able to achieve 40% accuracy on our out-of-sample data set, which is suboptimal for a production setting in this case.

Overall, this was an interesting and difficult modeling task. Starting with the exploratory data analysis which was one of the more difficult ones I have performed due to the hard to explain variable values, difficult to detect outliers and generally low correlations. But I was able to press forward and produce three reasonable models for the task at hand. Thank you for a great quarter!

Appendix

Correlations to STARS

	Correlation to STARS
LabelAppeal	0.332991663
Purchase	0.286085136
Alcohol	0.064522760
ResidualSugar	0.015092521
Norm_ResidualSugar	0.015092521
TotalSulfurDioxide	0.014257996
Norm_TotalSulfurDioxide	0.014257996
Norm_CitricAcid	0.003193647
CitricAcid	0.003193647
Norm_FixedAcidity	-0.003177822
FixedAcidity	-0.003177822
Chlorides	-0.003722706
Norm_Chlorides	-0.003722706
pH	-0.005588239
Sulphates	-0.010807207
Norm_Sulphates	-0.010807207
FreeSulfurDioxide	-0.011024276
Norm_FreeSulfurDioxide	-0.011024276
Density	-0.022879954
Norm_VolAcidity	-0.029881237
VolatileAcidity	-0.029881237
AcidIndex	-0.088784114

Correlations to Cases