Modeling Assigment #4

Brandon Moretz

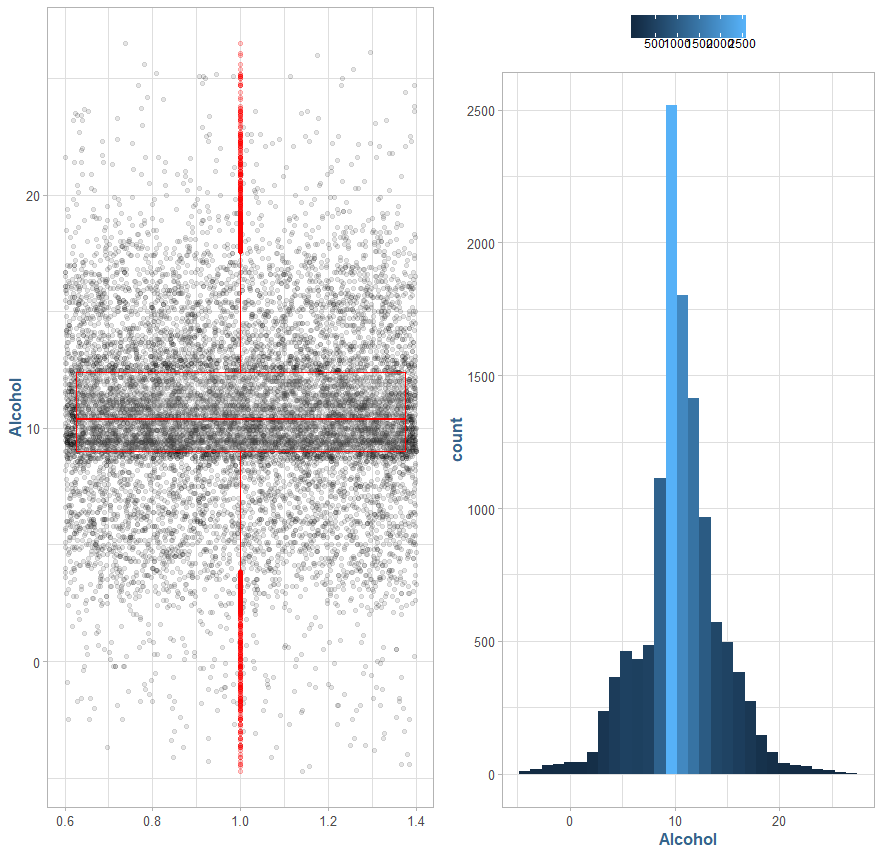
### Introduction

In this lab we are going the explore a data set consisting of various attributes of approximately twelve thousand commercially available wines 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 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, we will attempt to quantity what makes a wine ‘good’ based upon its chemical composition and characteristics.

We will start with some preliminary exploratory data analysis to look for distribution characteristics of the independent variables and their statistical properties. From there, we will divide the data set into different sets based upon the type of statistical model we will ultimately be producing and the various response variables we are interested in. There are three response variables we will attempt to explain throughout this lab: purchased, cases sold, and stars (rating), which are although somewhat related variables, they require different statistical procedures to model correctly.

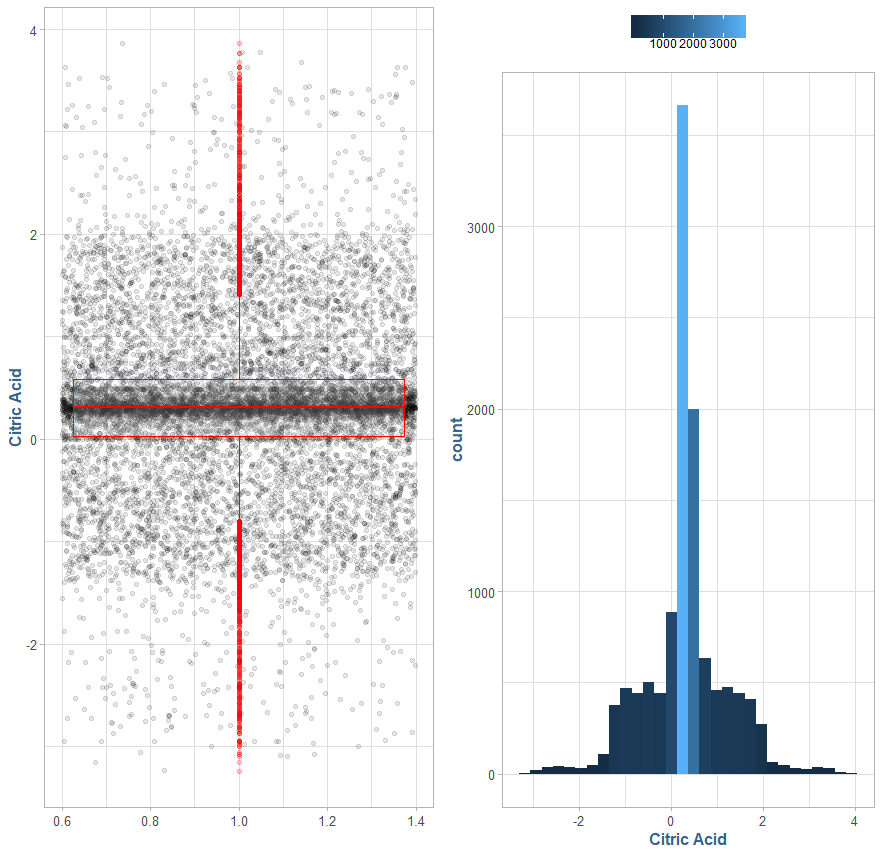
### Exploratory data analysis

The first step in this analysis project is to take a clean sweep to the data and look for 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. For example, the Alcohol variable 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).

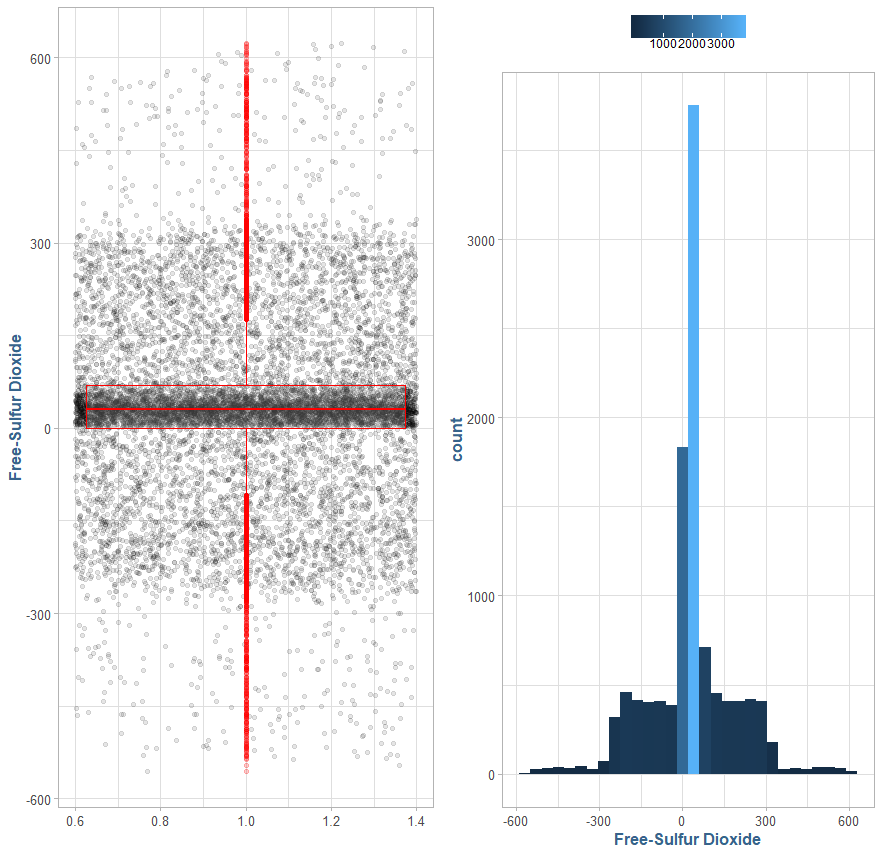


Given this measurement, we are going to remove the 771 bottles with missing or negative values.

Other 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). The Citric Acid, Sulphates and Volatile Acidity variables for example has this quality when we look at it closely. This is the only logical explanation I can see for over half of the values being in the negative range for so many variables that represent measurements of a quantity.

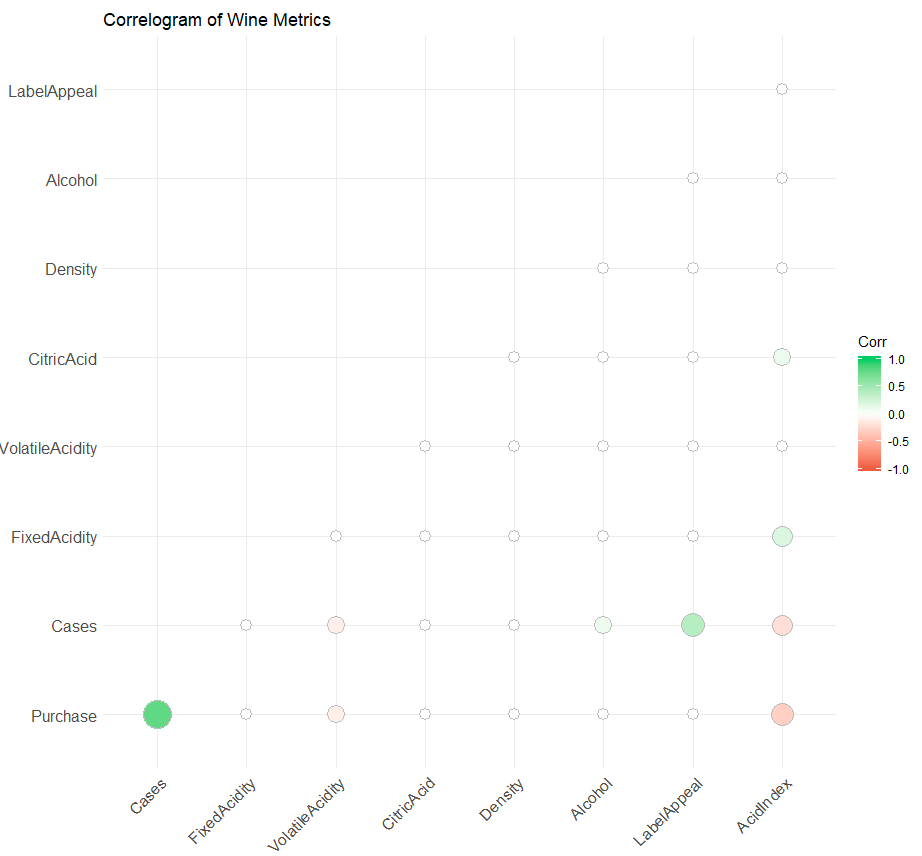


Some of the other variables have similar properties (Free-Sulfur Dioxide, Chlorides and Residual Sugar), however, they appear to be on slightly different scales. We will standardize these variables and append them to our data set; this process won’t change the actual properties of the values underlying distribution, 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.

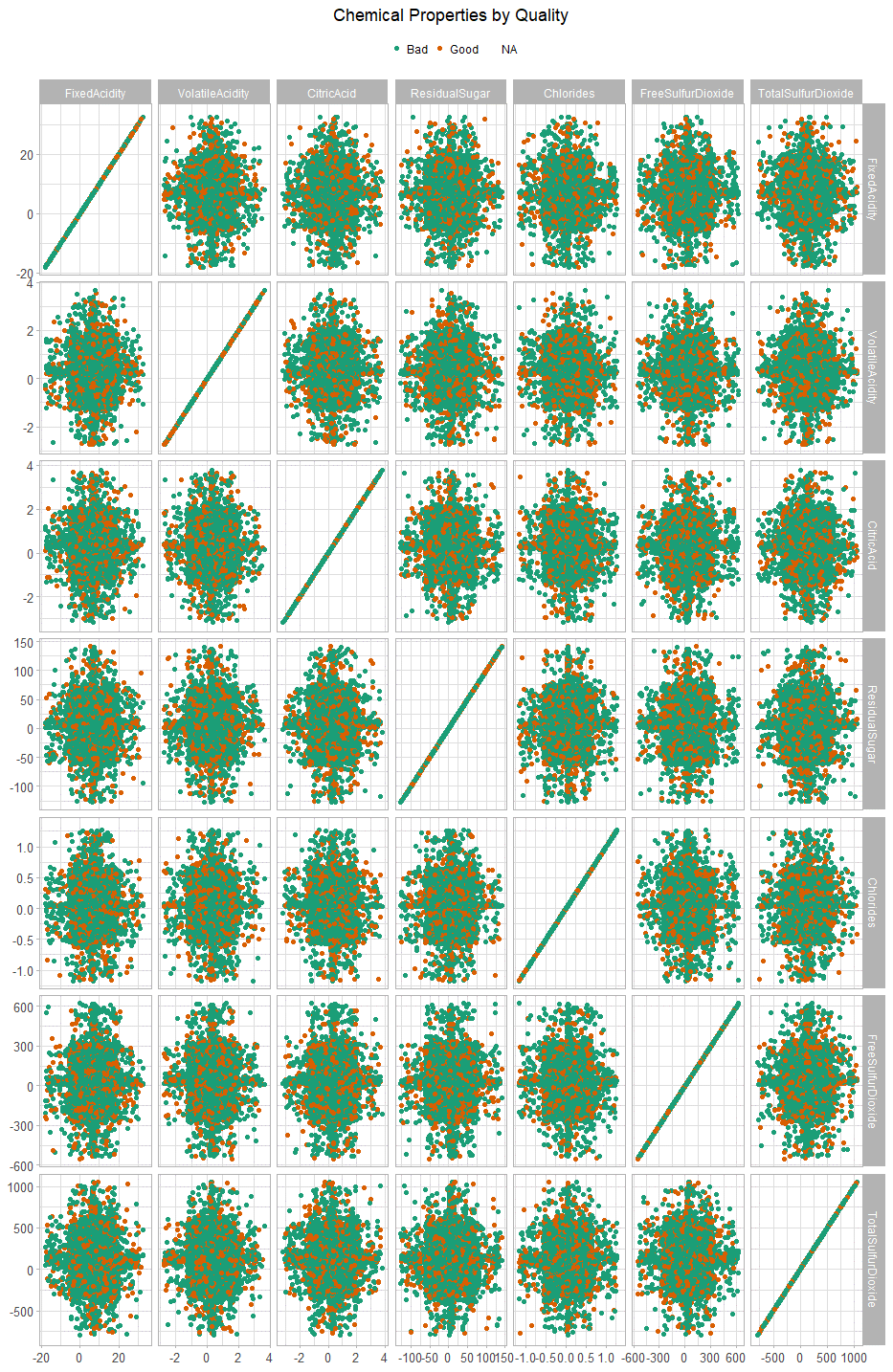


The label appeal variable has the scale -2 to 2, which we will adjust by +3 so that it is on a standard 1-5 scale. We will also create a simpler metric in terms of quality based on the STARS variable, which is a binary ‘good’ or ‘bad’, which is ‘good’ if the STARS rating is greater than or equal to 3.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

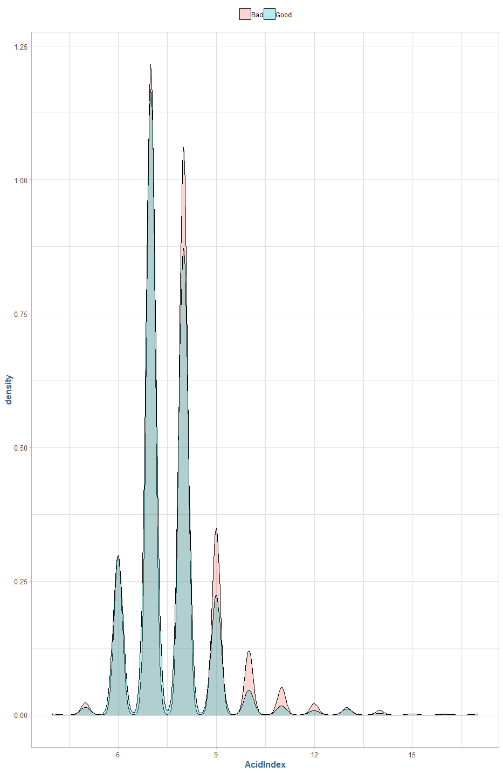
Before we dive into some more detailed bivariate EDA analysis, we will look for some correlations amongst the dataset which will hopefully expand our analysis by raising some additional questions. First, we will look to see if there are any correlations to the STARS rating variable. The most obvious 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. A bit more surprisingly however is the relatively strong correlation to the label appeal, at 33%, which indicates that the better the presentation the higher the rating. We will explore this in more depth later. The correlation to STARS table can be found in the [appendix](#_APPENDIX), and a visual breakdown can be seen in the following chart:



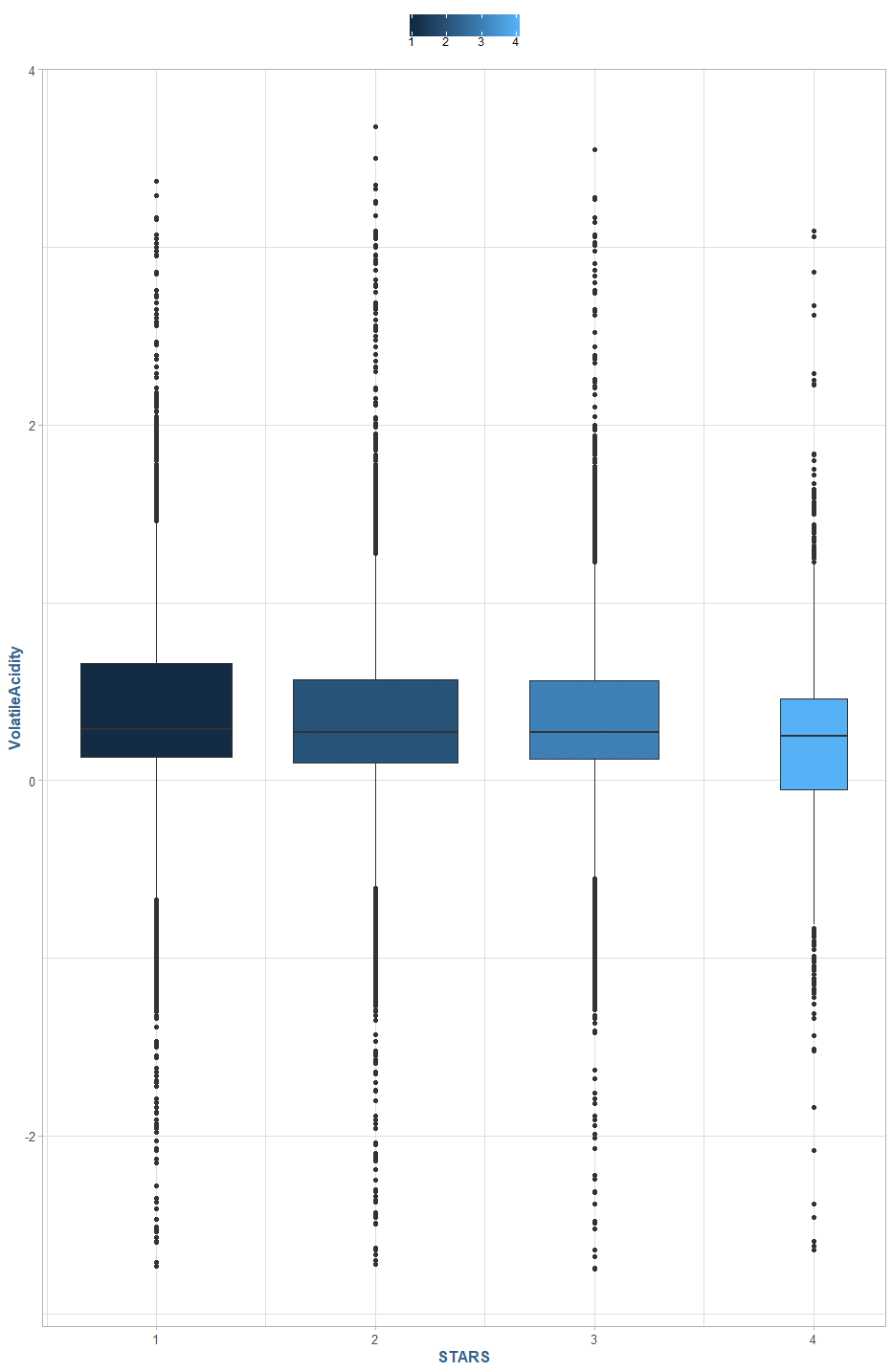
Next, we will look for relationships between the high-level quality variable we defined to look for any chemical relationships that may help determine what makes for a ‘good’ wine. In the following diagram, we can see a detailed breakout for each chemical property and the quality variable:



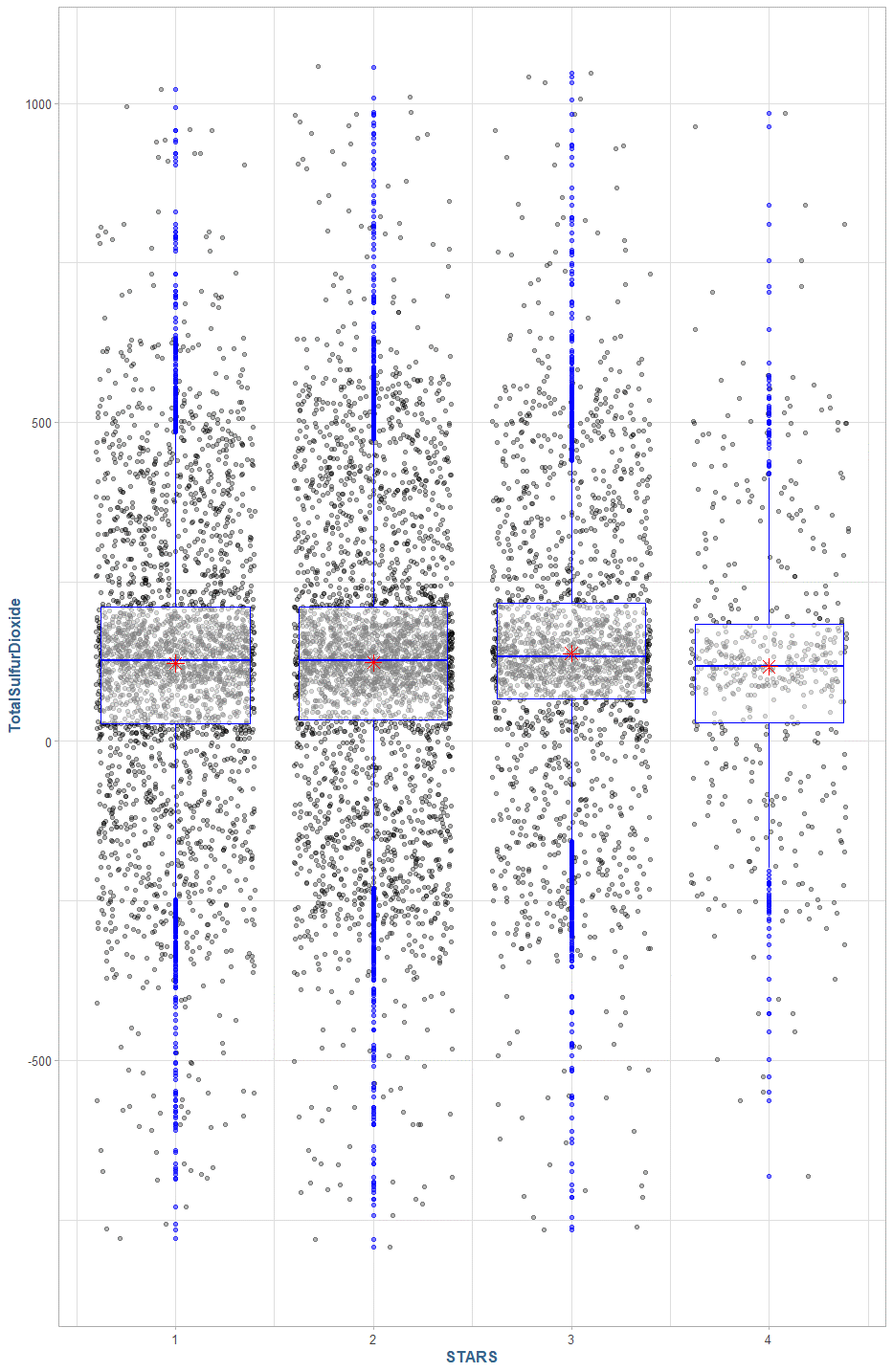
Interestingly, the acid index property seems to have a negative effect on wine quality as the acidity increases. We can see an overlay of ‘good’ and ‘bad’ wines by acid index in the following diagram:



There also seems to be a lower concentration of volatile acidity in higher rated wines, as we can see in the following diagram:

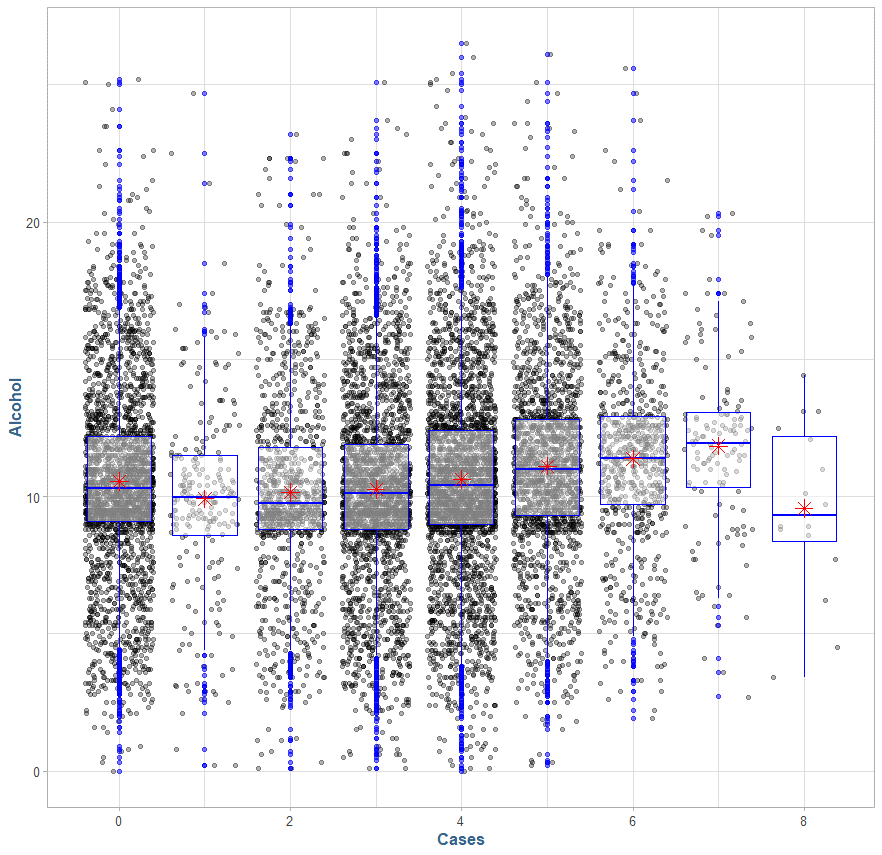


Additionally, there seems to be slightly lower levels of sulfur dioxide in higher rated wines:



The rest of the chemical composition properties seem to be relatively equally distributed across the various rating groups and there is little visual evidence to suggest there are meaningful differences. However, we will explore these relationships further in the modeling and research section with additional statistical methods.

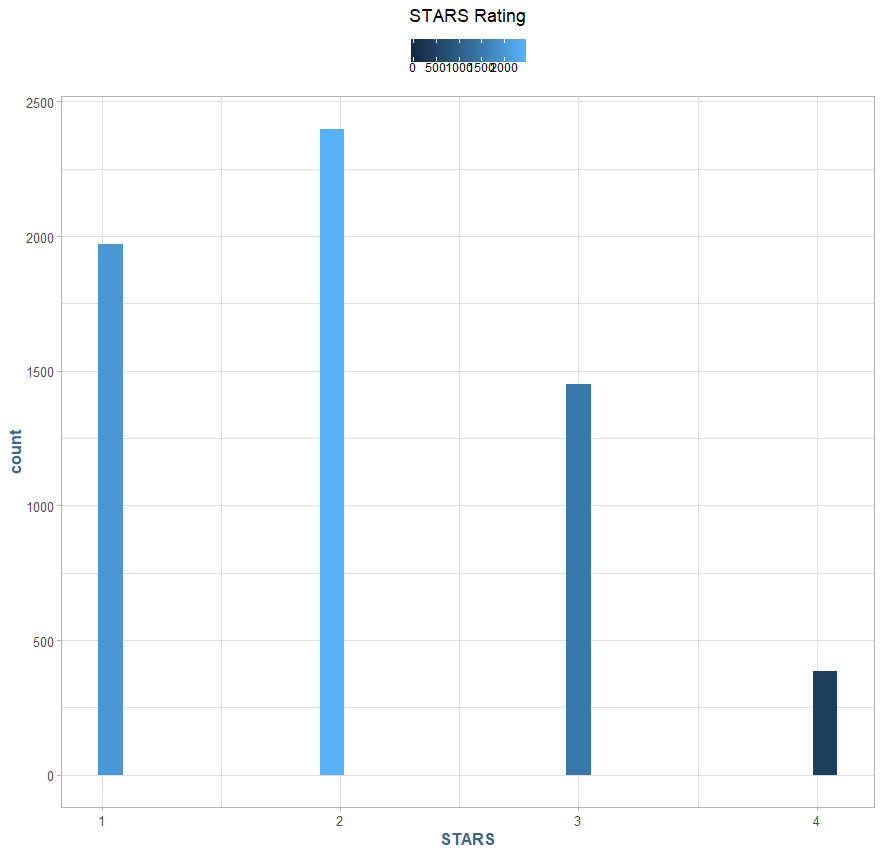
Another relationship we should note is between the number of cases sold and the alcohol content of the wine. It appears that there is a moderate relationship between the higher alcohol content wines and the number of cases sold. There is a drop off in the 8-case range, however, there is also just a relatively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



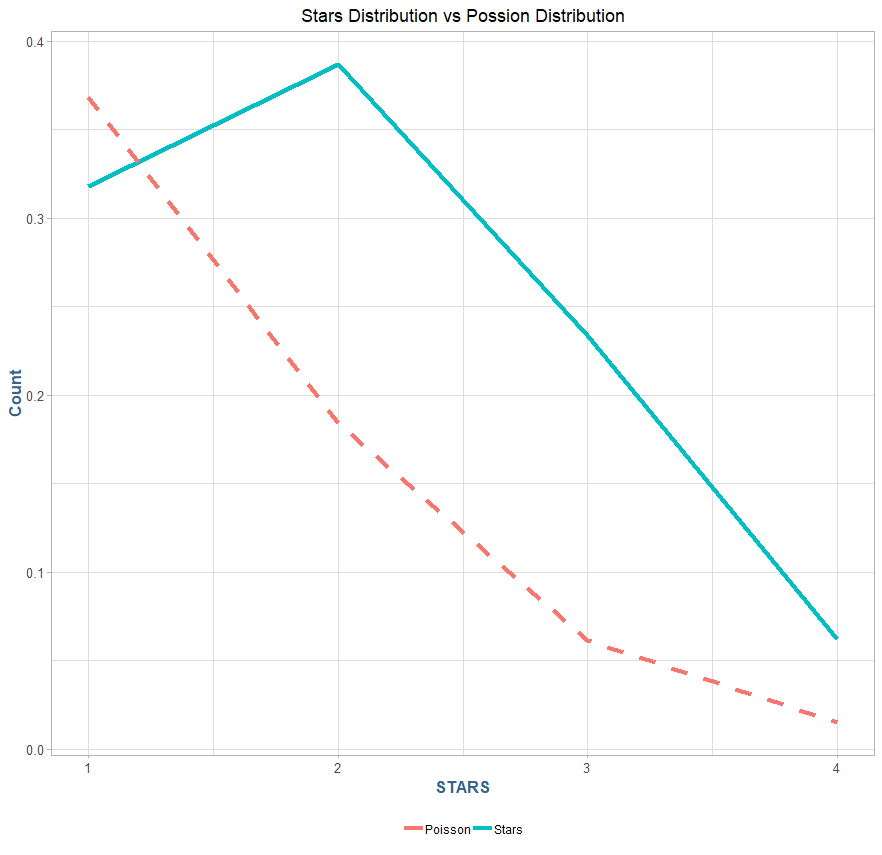
### STARS Model

In this section we are going to explore the data further has it relates to the stars rating received by the wine. For this model, we are going to further refine our data set to exclude any wines that do not have a star rating (3,150 bottles dropped for this analysis total). After that, we will split our cleaned data into a standard 70/30 split so we can evaluate our model against the 30% hold-out set later.

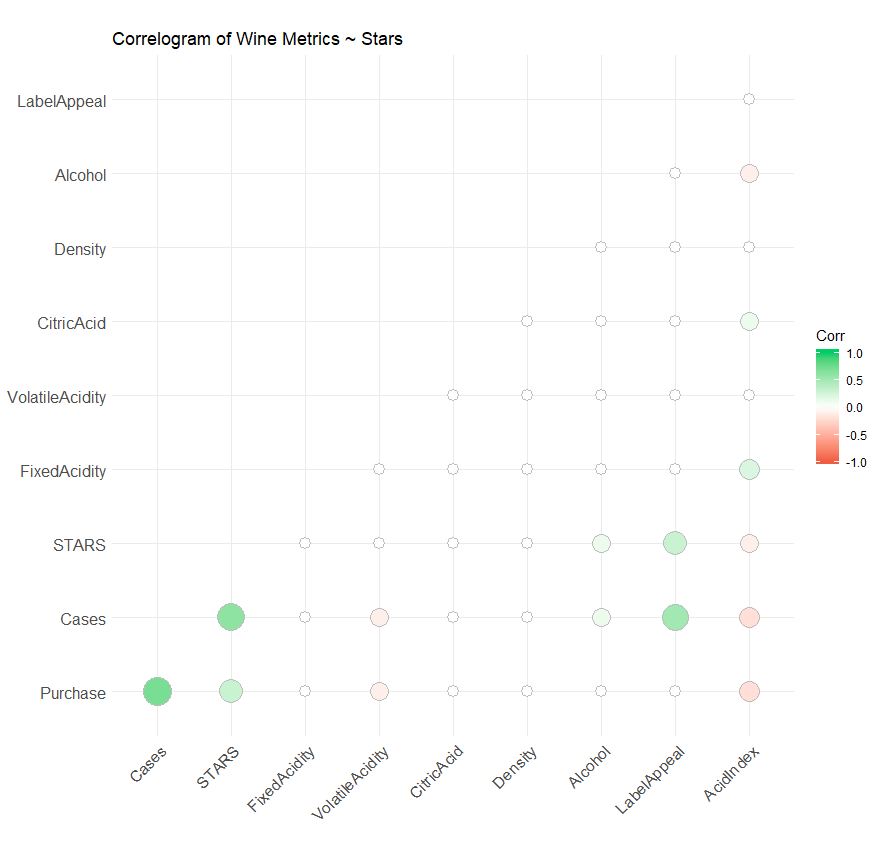
First, we will take a look at the distribution of the stars rating to get an overall idea how the scores are distributed across the bottles of wine:

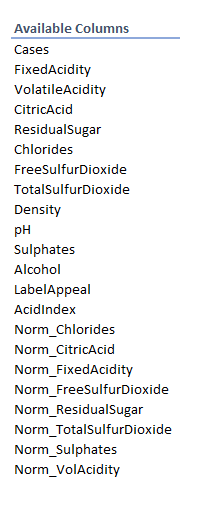


The distribution is skewed to the right, where the majority (approximately 70%) of our wines have a “bad” rating of one or two. These star counts do not exactly follow a Poisson distribution, it is relatively close in general form:



We will favor the Poisson distribution over the Gaussian distribution as we are predicting discrete scale data instead of continuous. We will again run a correlation matrix on the stars data 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, we see slightly negative relationships to volatile acidity and acidity index, and a positive correlation to cases and label appeal, as we would have expected from our prior analysis. Given that “cases” is the standout variable here in terms of being associated with the stars rating we would expect our variable selection techniques will favor this variable.

Additionally, since the quality variable is derived from the stars rating for the wine, we will exclude that column from the set of available columns for auto-selection methods.

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

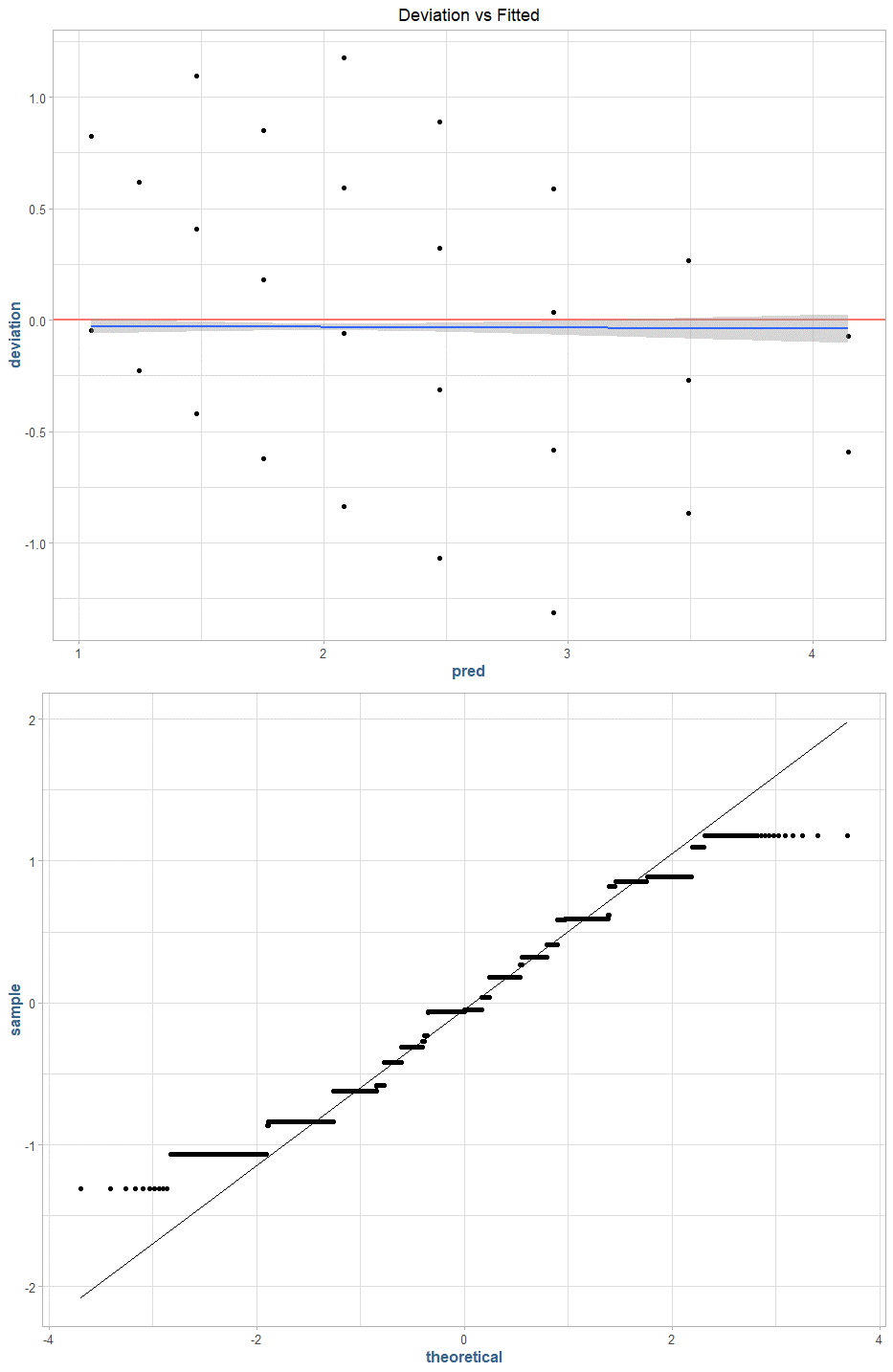
Ŷ = 0.05 + 0.17β1

Where β1 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:

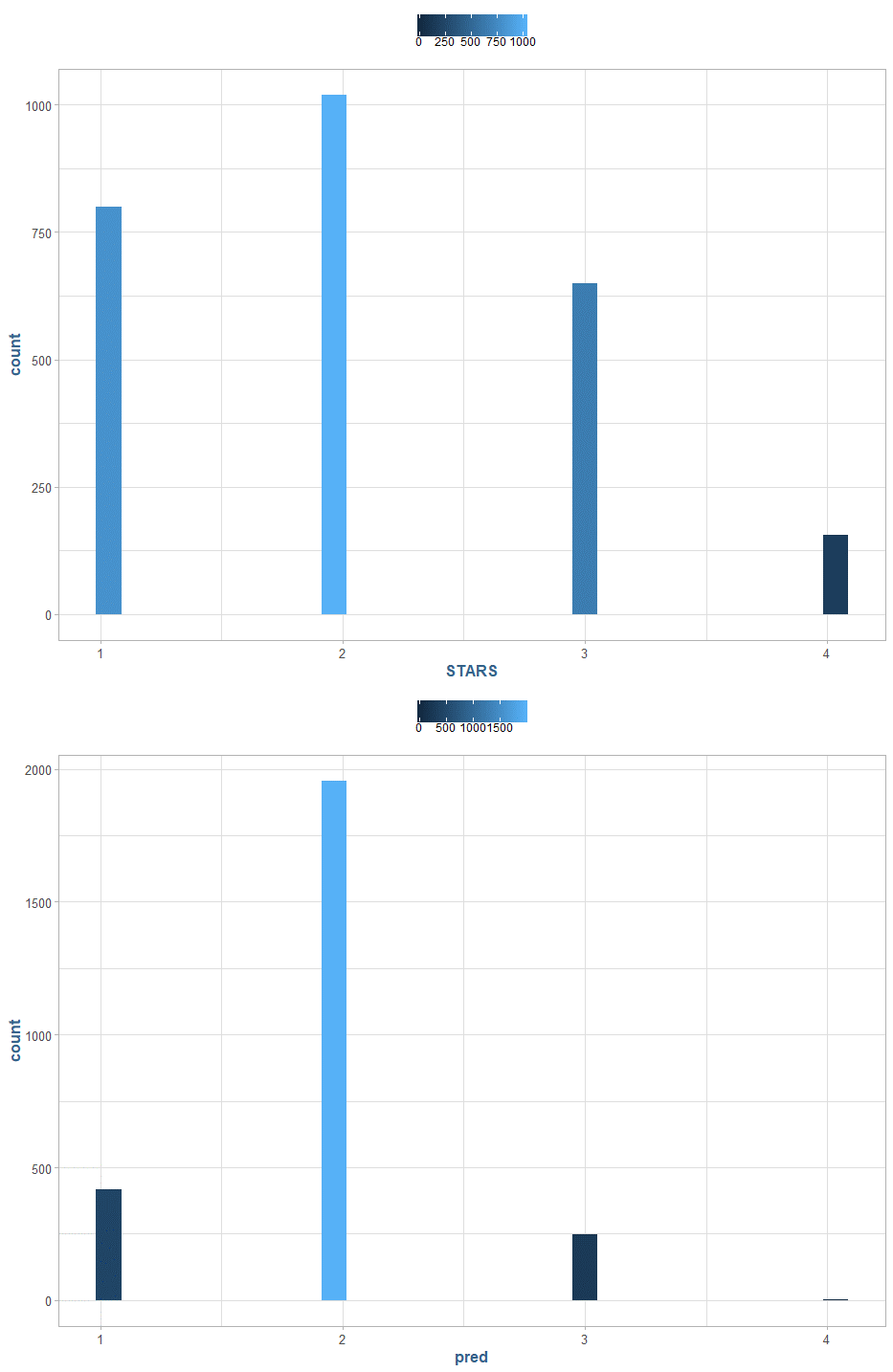
AIC: **12,612**, Null deviance: **1,798**, Residual deviance: **1,213**

MSE**: 0.74**, R2: **31%**, MAE: **.59**

Visually, we can look at the deviation vs 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. We can look at the actual ratings distribution in the test set vs the predictions made on the test set:



We can see there is clearly a bias prediction for bottles of wine with a two-star rating, or a over convergence to the mean in the predictions. Also, if we break down the accuracy level by star rating produced by the model:



We see that there is a huge 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, we are going to re-visit the entire modeling process and 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**

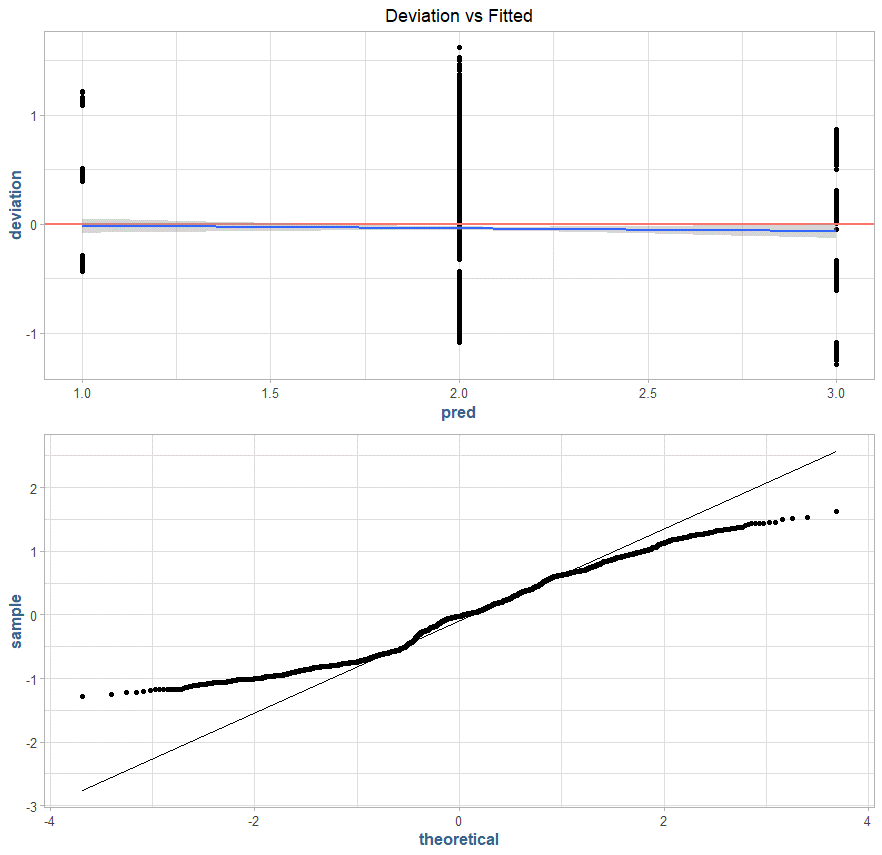
Ŷ = 0.46 + 0.007β1 + 0.16β2 - 0.04β3

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.

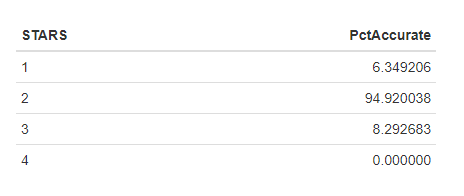
AIC: **12,783**, Null deviance: **1,739**, Residual deviance: **1,549**

MSE**: 0.84**, R2: **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:

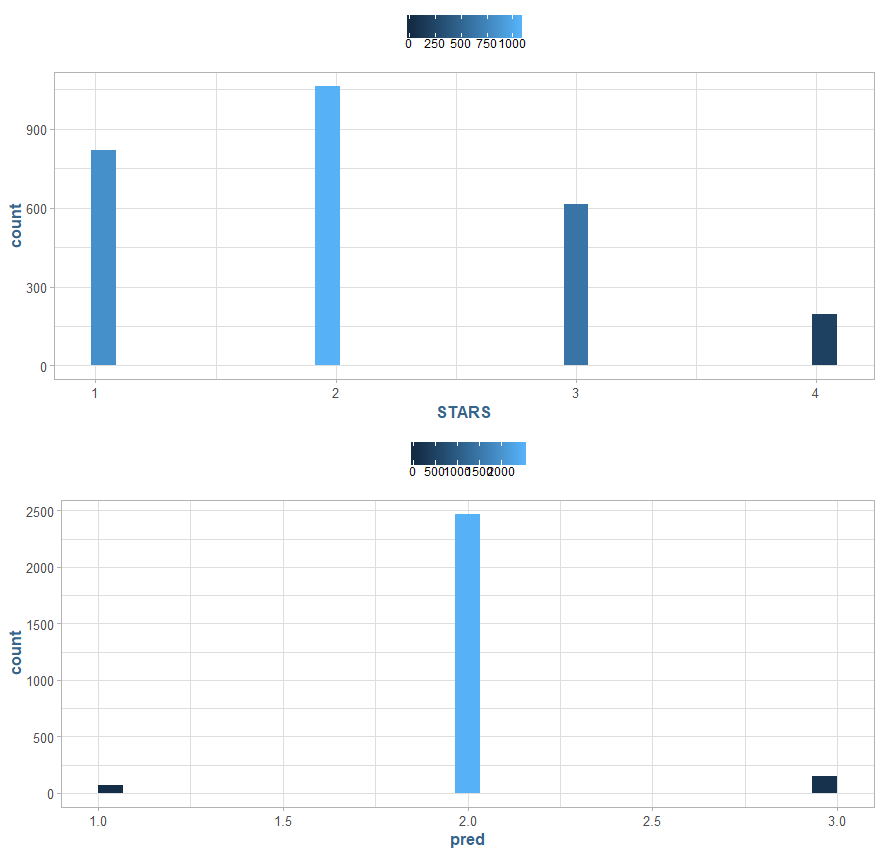


The second model correctly predicted the rating 41.7% for all the bottles in the out of sample test set. By category, we see that there is a significant breakdown in the model when it comes to four start bottles of wine:



However, the results for predicting an average two-star bottle of wine is dead on, which is not unlike the results we 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 we see the distribution of the predictions for the second iteration of the model:



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

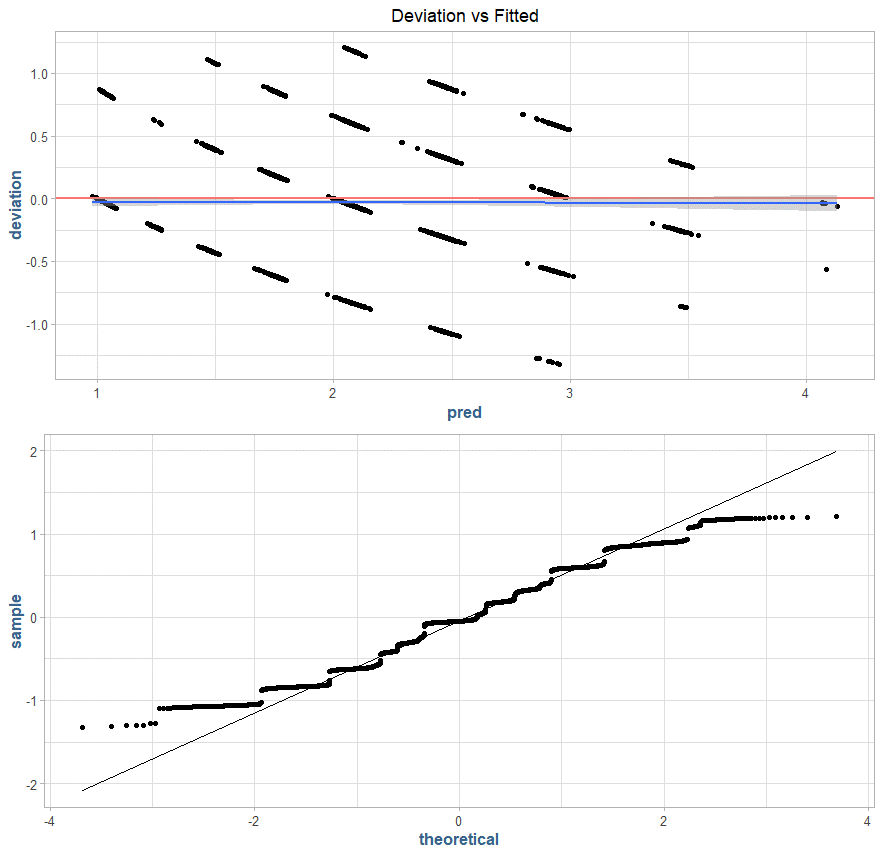
**STARS Model 3**

Ŷ = 0.10 + 0.17β1 + 0.002β2 - 0.002β3- 0.007β4

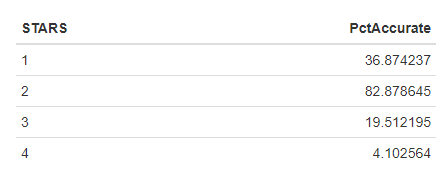
Where a one unit increase in cases sold represents a .17 increase in rating, a one unit increase in alcohol content by volume (ABV) results in a 0.002 increase in stars rating, a one unit increase in label appeal results in a 0.002 decrease to the stars rating, and a one unit decrease in acidity index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

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

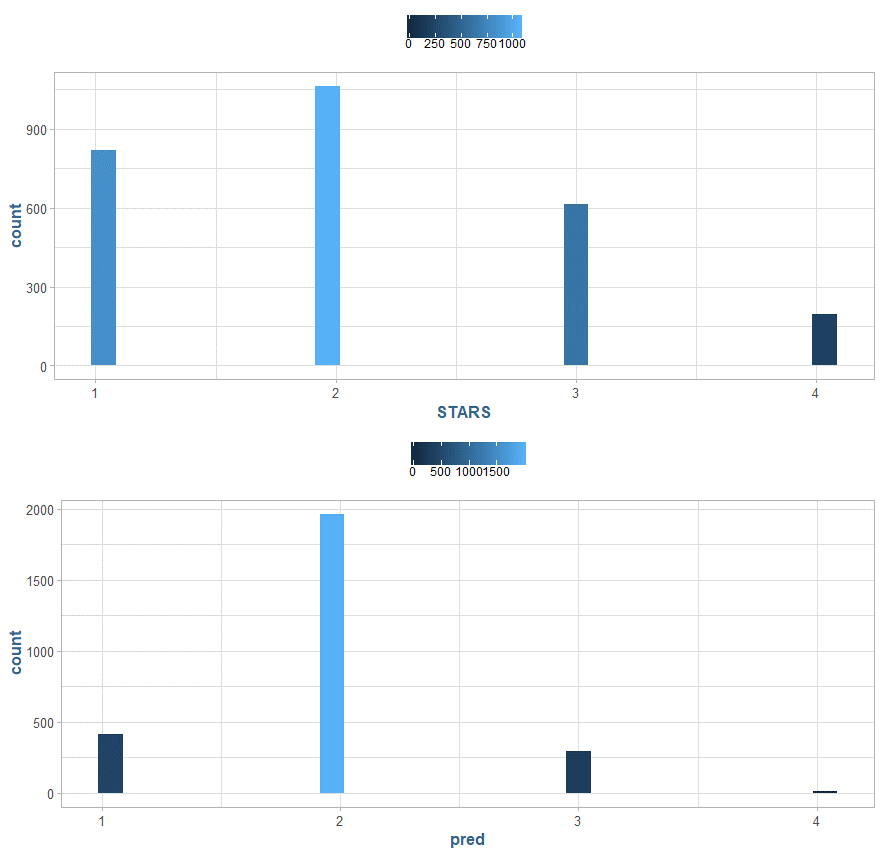
Which is a slight statistical improvement over the prior models. Visually we can see the diagnostics of the model is a slight improvement as well:



However, more importantly our final model has an overall prediction accuracy of 48.7% and we see substantial increases in prediction 3 of the 4 areas in the stars rating categories:



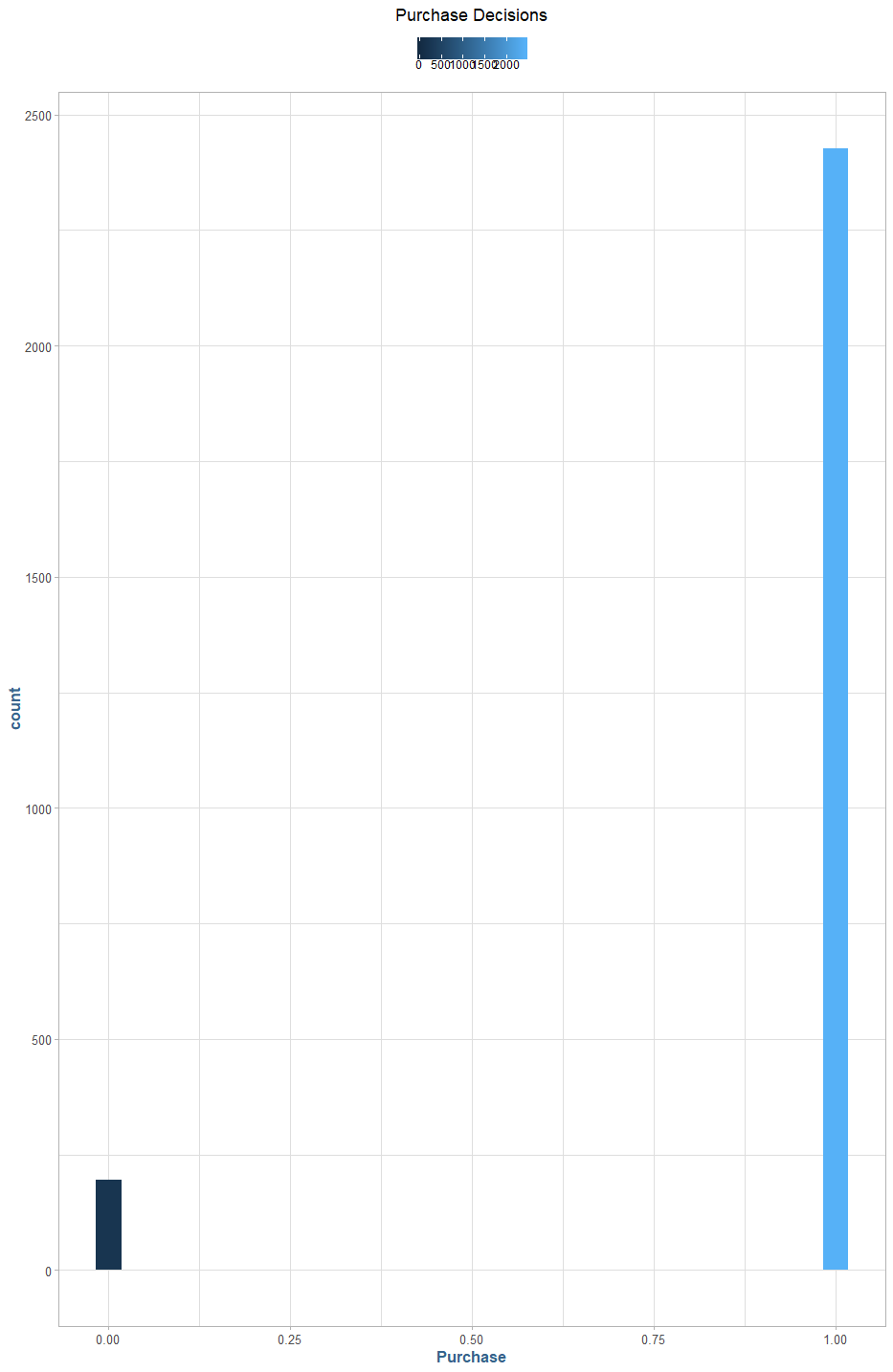
While the model is clearly not a home-run hitter, it is a substantial improvement over the prior two iterations. The prediction distributions are still skewed to the two-star wines, however, there is improvement:



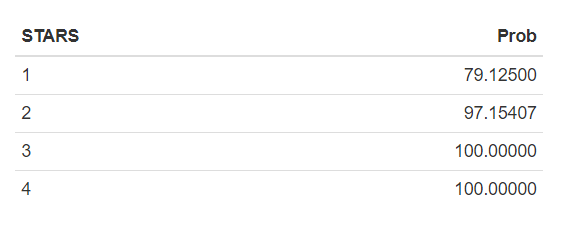
### Purchase model

Now we are going to look 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 amount of cases of this wine that have been purchased. Since there is a dependency on the cases and purchase variable, we are going to exclude this variable from our analysis here given that a model based upon data that already has the outcome backed in it would not be so 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%:



We also note that any wine we observe with over 2 stars has a 100% likelihood of getting purchased:



### cases model

### Conclusion

### APPENDIX

Correlations to STARS

