# 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).
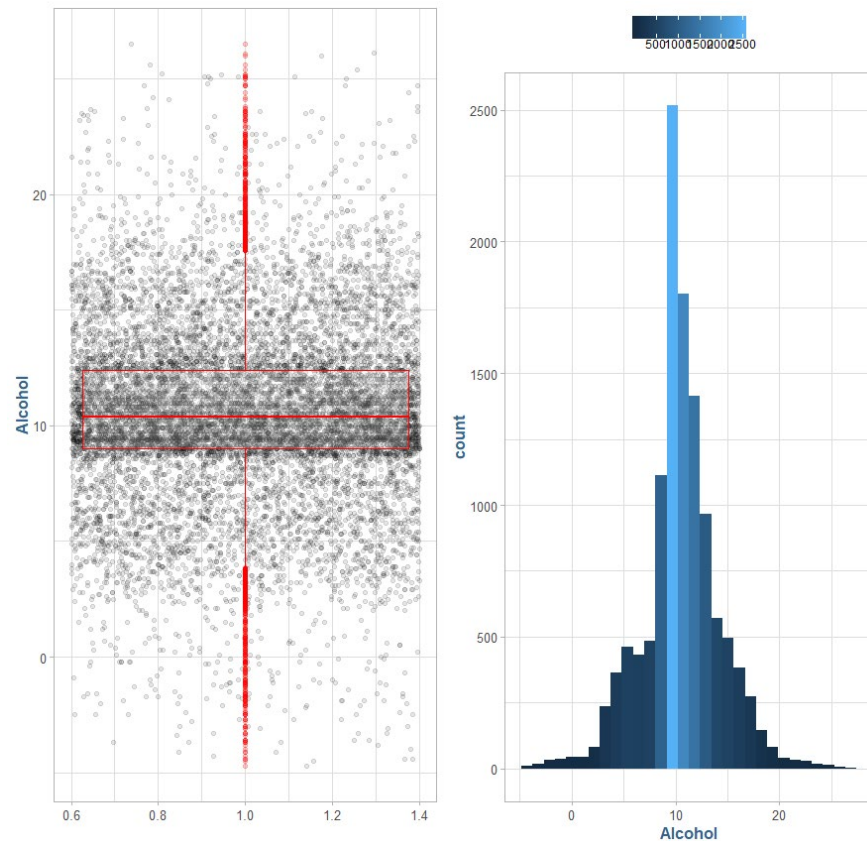


Given this measurement, we are going to remove the 771 bottles with missing or negative values.

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.
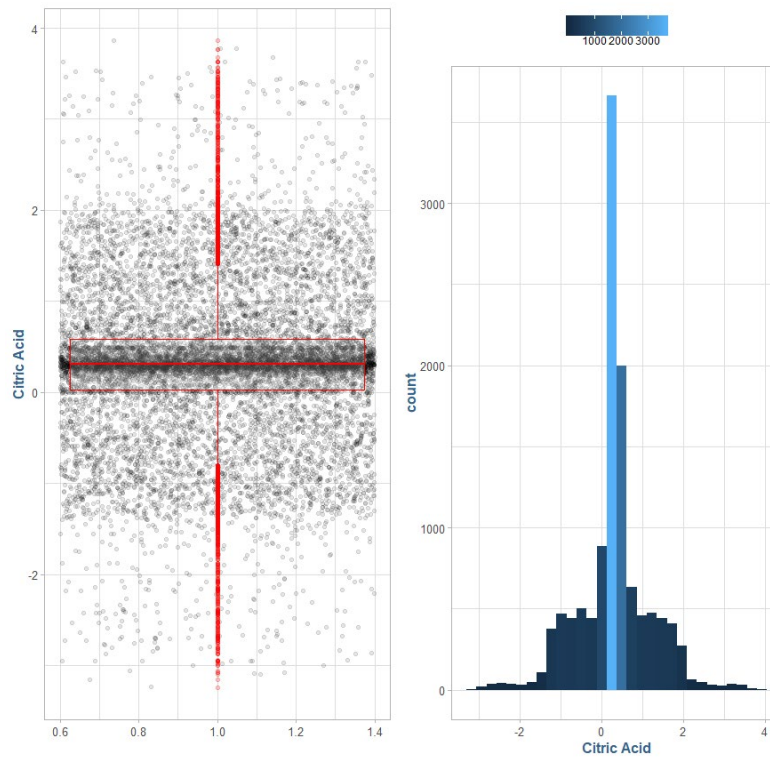
Some of the other variables have similar properties (Free-Sulfur Dioxide, Chlorides and Residual Sugar), however, they appear to be on slightly different scales. We will standardize these variables and append them to our data set; this process won't change the actual properties of the values underlying 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.
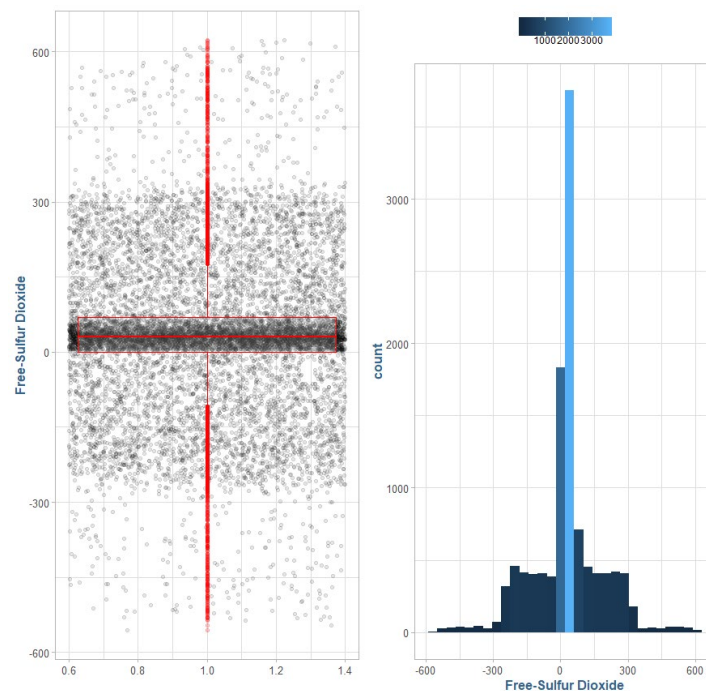
The label appeal variable has the scale -2 to 2, which we will adjust by +3 so that it is on a standard 1-5 scale. We will also create a simpler metric in terms of quality based on the STARS variable, which is a binary 'good' or 'bad', which is 'good' if the STARS rating is greater than or equal to 3. This will help uncover basic relationships by condensing 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, and a visual breakdown can be seen in the following chart:



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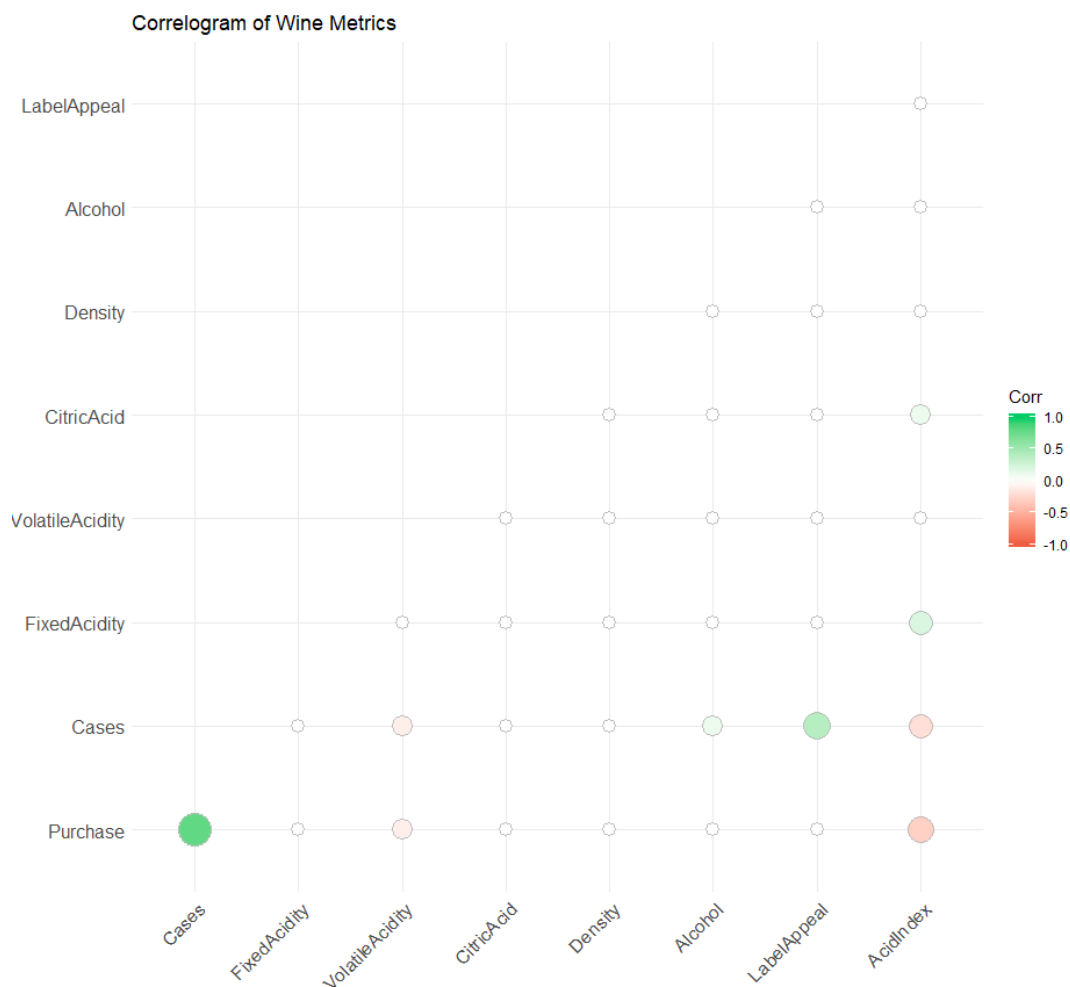
Next, we will look for relationships between the high-level quality variable we defined to look for any chemical relationships that may help determine what makes for a 'good' wine. In the following diagram, we can see a detailed breakout for each chemical property and the quality variable:



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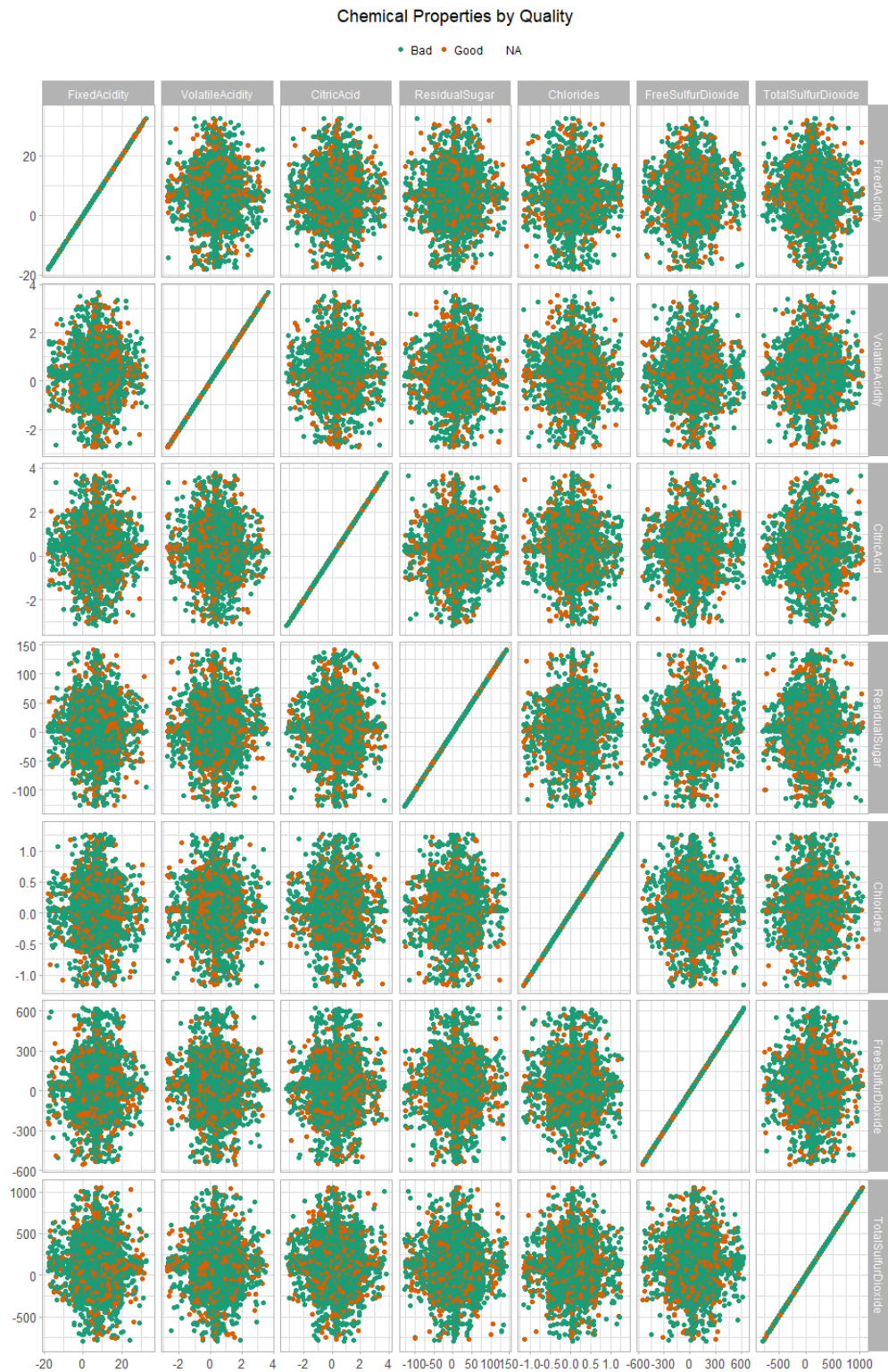
Interestingly, the acid index property seems to have a negative effect on wine quality as the acidity increases. We can see an overlay of 'good' and 'bad' wines by acid index in the following diagram:
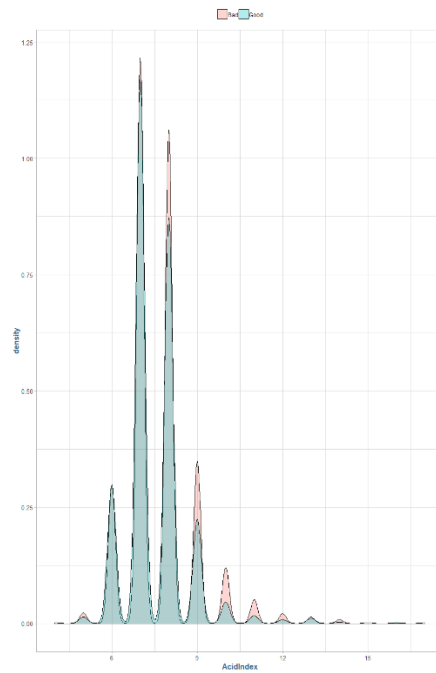


There also seems to be a lower concentration of volatile acidity in higher rated wines, as we can see in the following diagram:



6
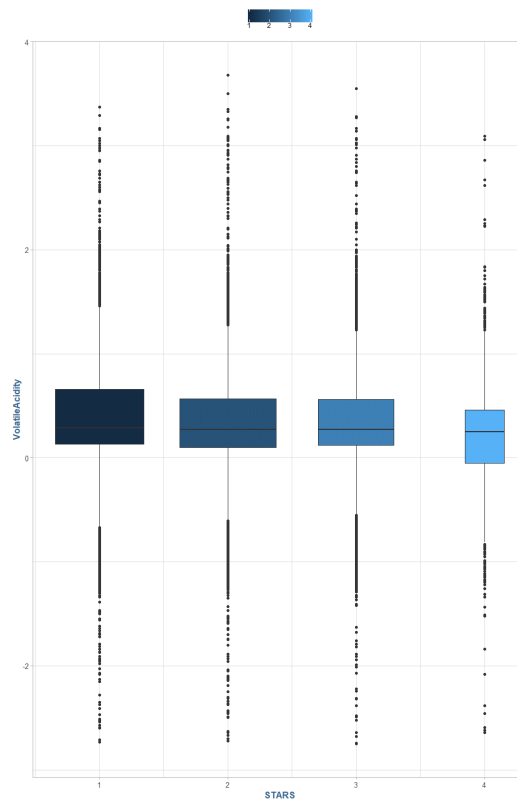
Additionally, there seems to be slightly lower levels of sulfur dioxide in higher rated wines:
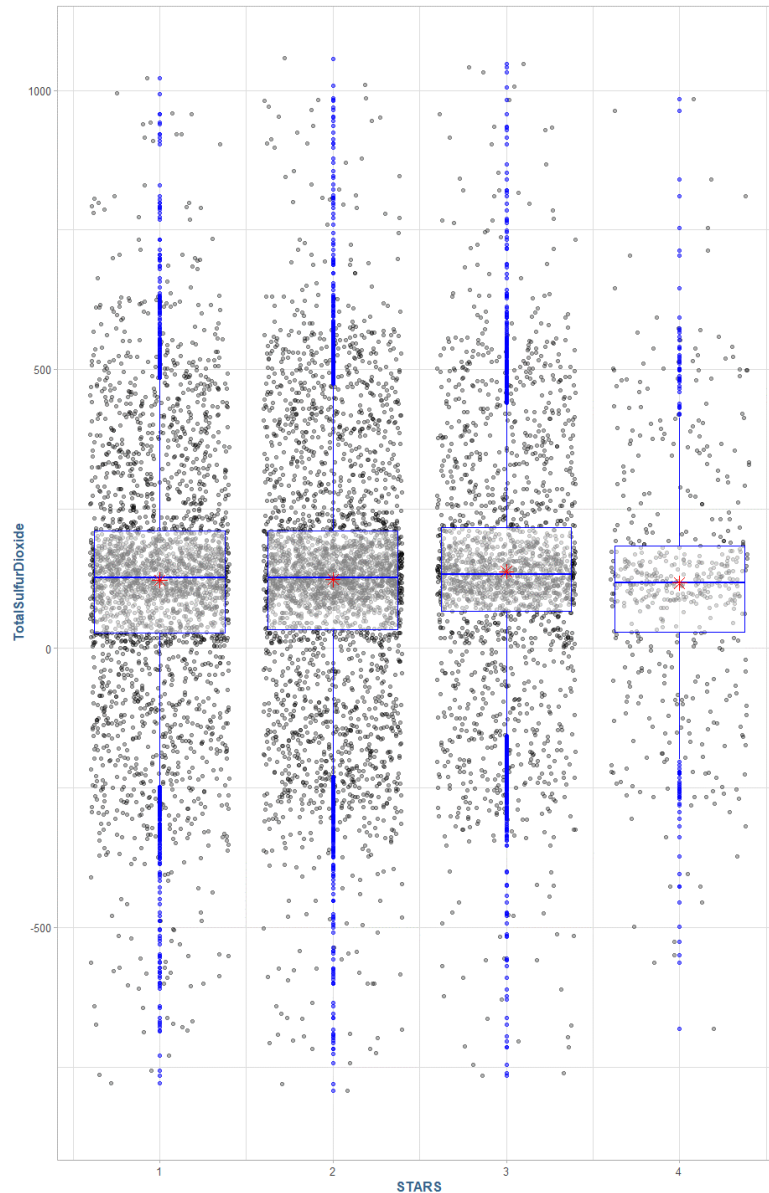


The rest of the chemical composition properties seem to be relatively equally distributed across the various rating groups and there is little visual evidence to suggest there are meaningful differences. However, we will explore these relationships further in the modeling and research section with additional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

7
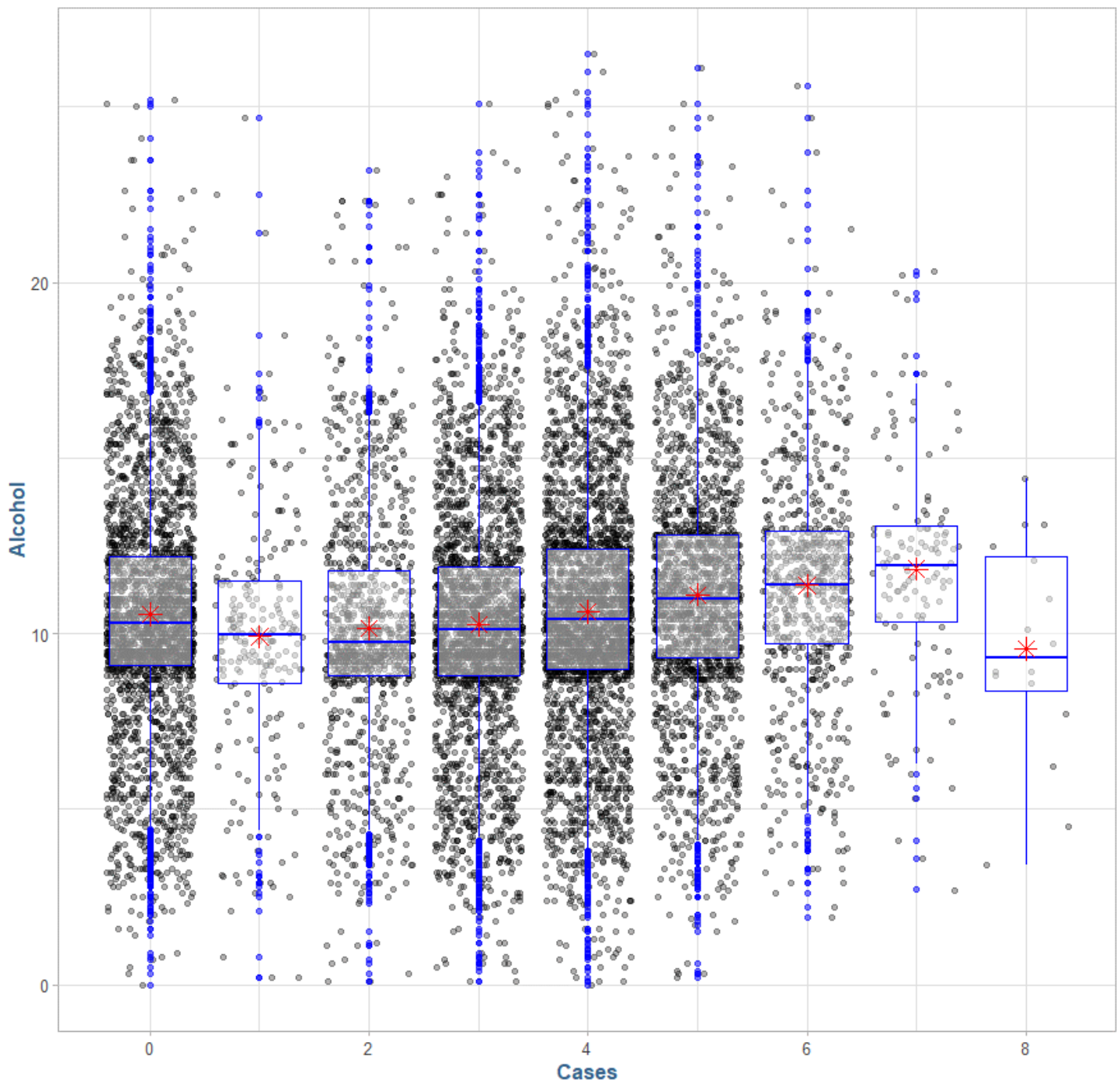
## STARS MODEL

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:
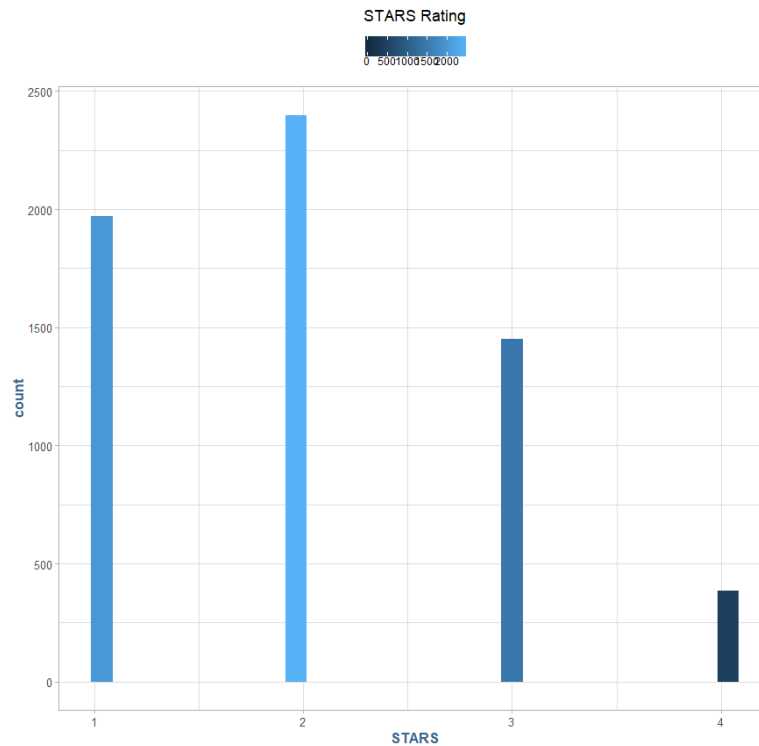


The distribution is skewed to the right, where the majority (approximately 70%) of our wines have a "bad" rating of one

or two. These star counts do not exactly follow a Poisson distribution, it is relatively close in general form:
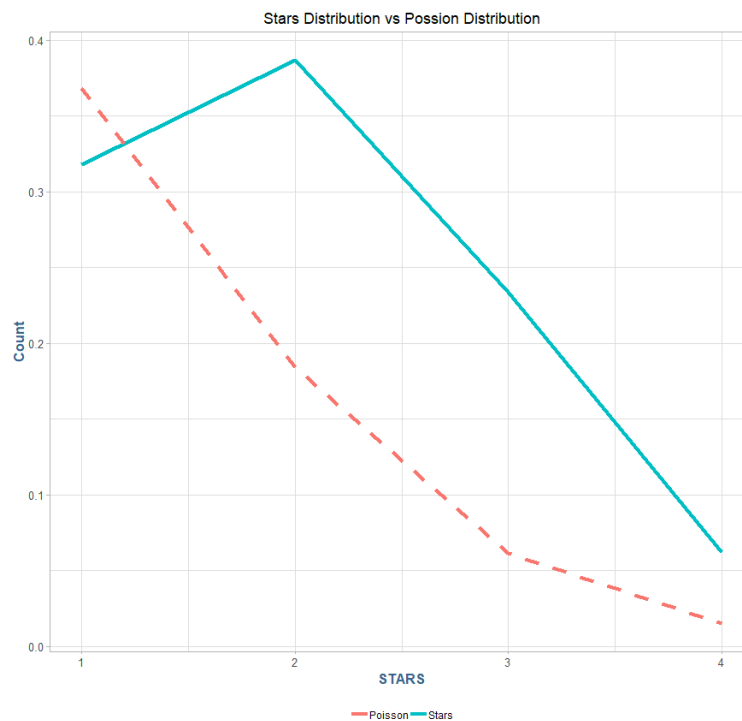
We will favor the Poisson distribution over the Gaussian distribution as we are predicting discrete scale data instead of continuous. We will again run a correlation matrix on the stars data given that we have pulled a smaller subset of the wines and removed any wines without a star rating, so this could yield some additional relationships:



Correlogram of Wine Metrics ~ Stars
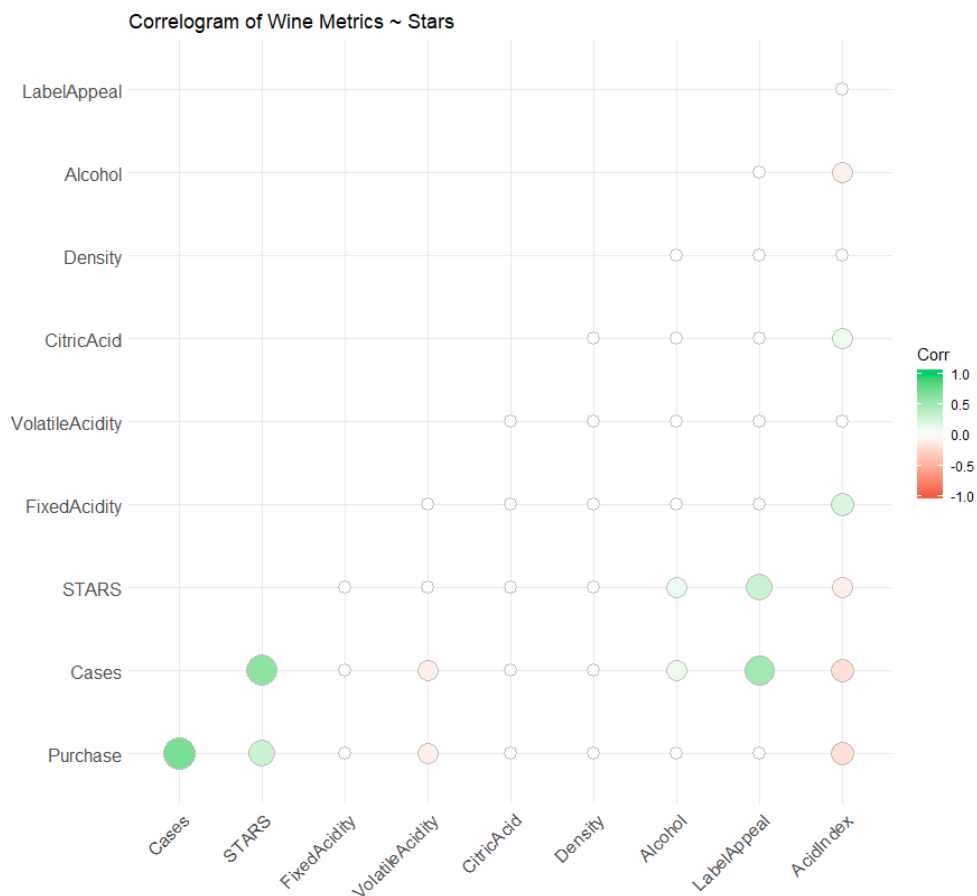
Looking across the STARS row, we see slightly negative relationships to volatile acidity and acidity index, and a positive correlation to cases and label appeal, as we would have expected from our prior analysis. Given that "cases" is the standout variable here in terms of being associated with the stars 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.

**Available Columns**
Cases
FixedAcidity
VolatileAcidity
CitricAcid
ResidualSugar
Chlorides
FreeSulfurDioxide
TotalSulfurDioxide
Density
pH
Sulphates
Alcohol
LabelAppeal
AcidIndex
Norm_Chlorides
Norm_CitricAcid
Norm_FixedAcidity
Norm_FreeSulfurDioxide
Norm_ResidualSugar
Norm_TotalSulfurDioxide
Norm_Sulphates
Norm_VolAcidity

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 $\beta_1$ 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, we can look at the deviation vs fitted and a QQ-plot of the deviations from the model on the training data:
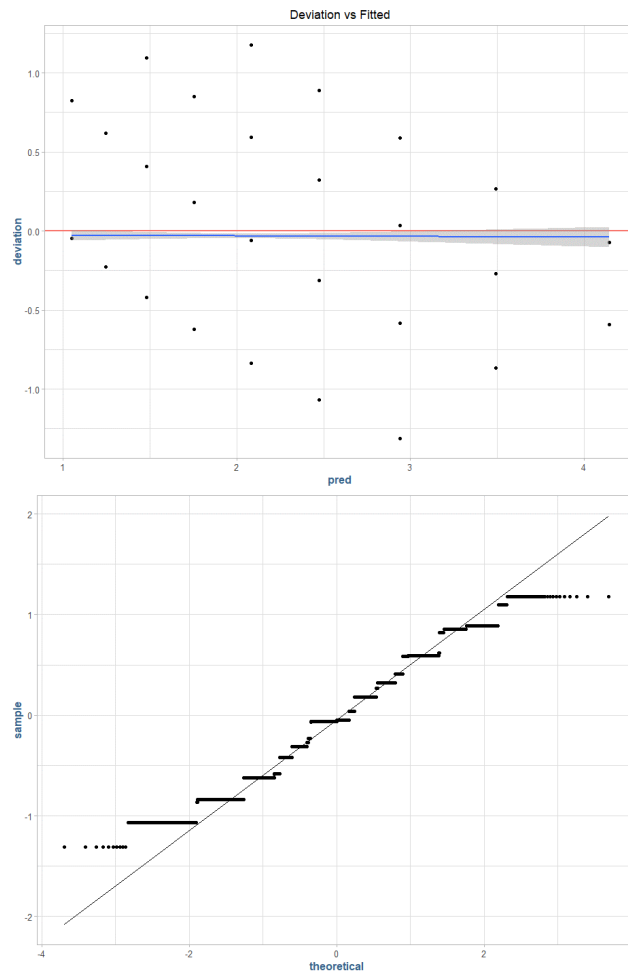
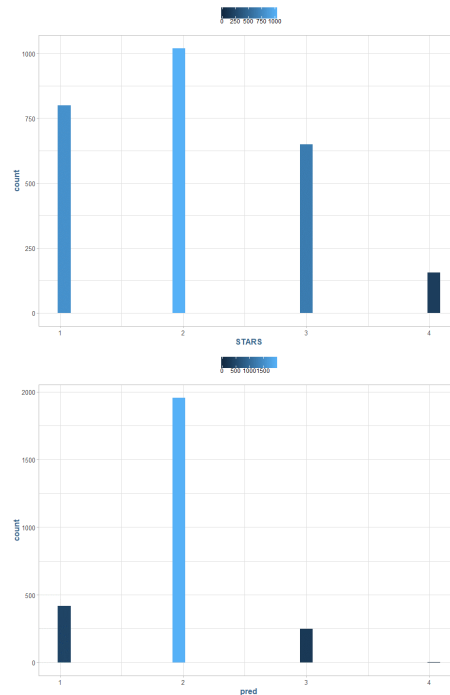As for the out of sample predictions, the model correctly predicted the stars rating for approximately 49.3% of the bottles of wine we kept out of sample in the hold-out set. We can look at the actual ratings distribution in the test set vs the predictions made on the test set:



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:

| STARS | PctAccurate |
|---|---|
| 1 | 39.682540 |
| 2 | 84.289746 |
| 3 | 16.747967 |
| 4 | 1.025641 |

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

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

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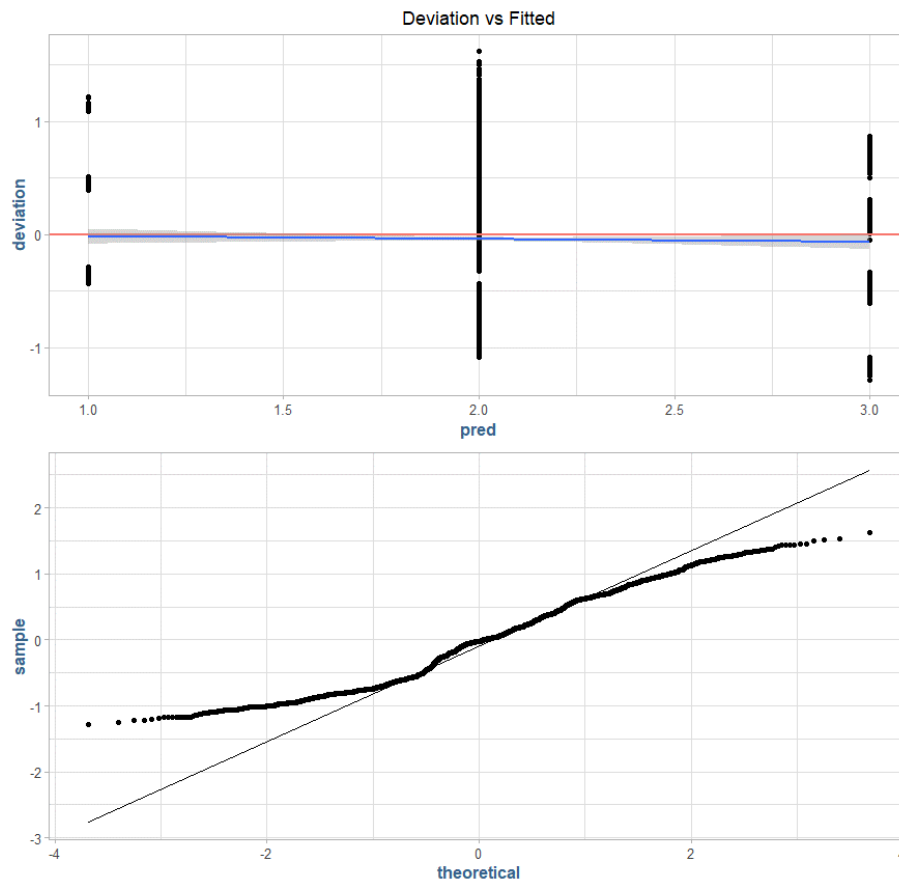
The second model correctly predicted the rating 41.7% for all the bottles in the out of sample test set. By category, we see that there is a significant breakdown in the model when it comes to four start bottles of wine:

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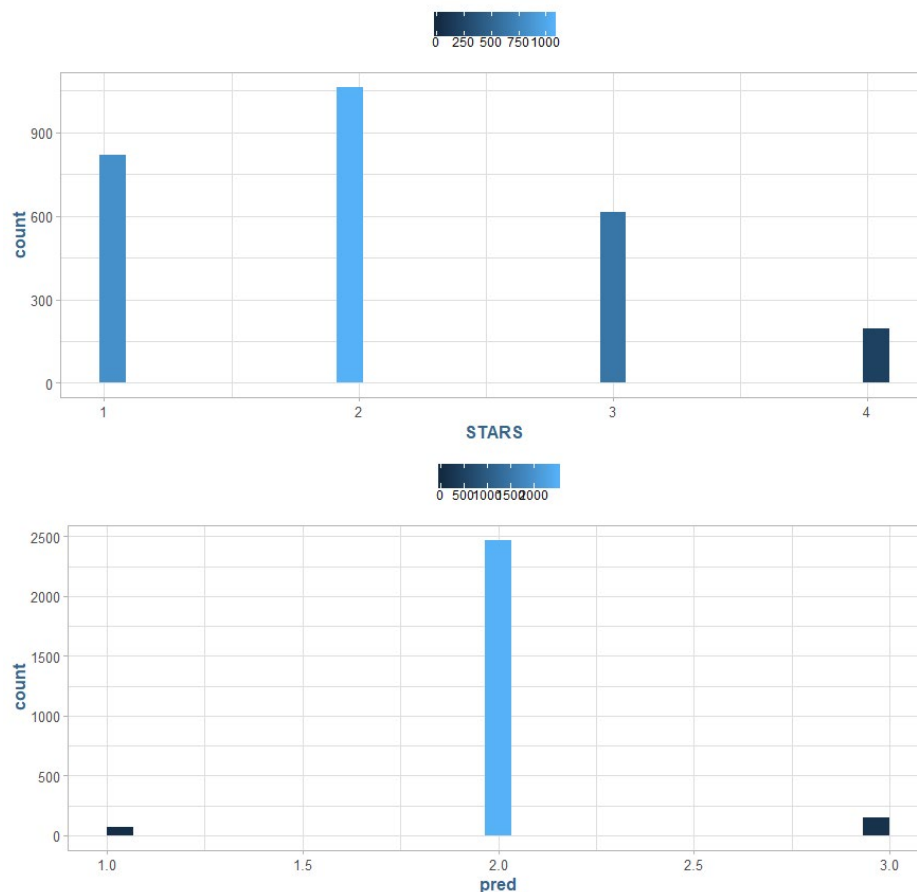
Our final attempt at the stars rating will attempt to combine the previous two models in order to get a more robust modeling procedure.

## STARS Model 3

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

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 we can see the diagnostics of the model is a slight improvement as well:
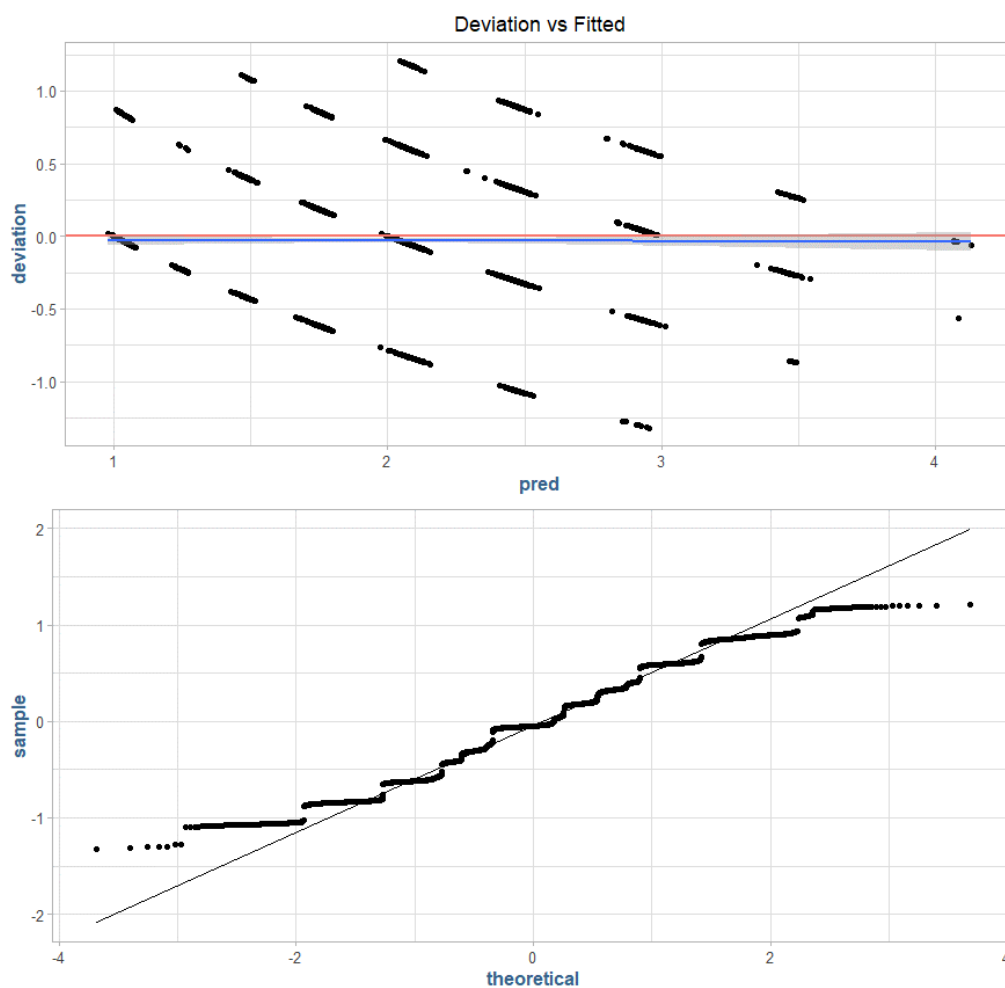
However, more importantly our final model has an overall prediction accuracy of 48.7% and we see substantial

increases in prediction 3 of the 4 areas in the stars rating categories:

| STARS | PctAccurate |
|---|---|
| 1 | 36.874237 |
| 2 | 82.878645 |
| 3 | 19.512195 |
| 4 | 4.102564 |

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:
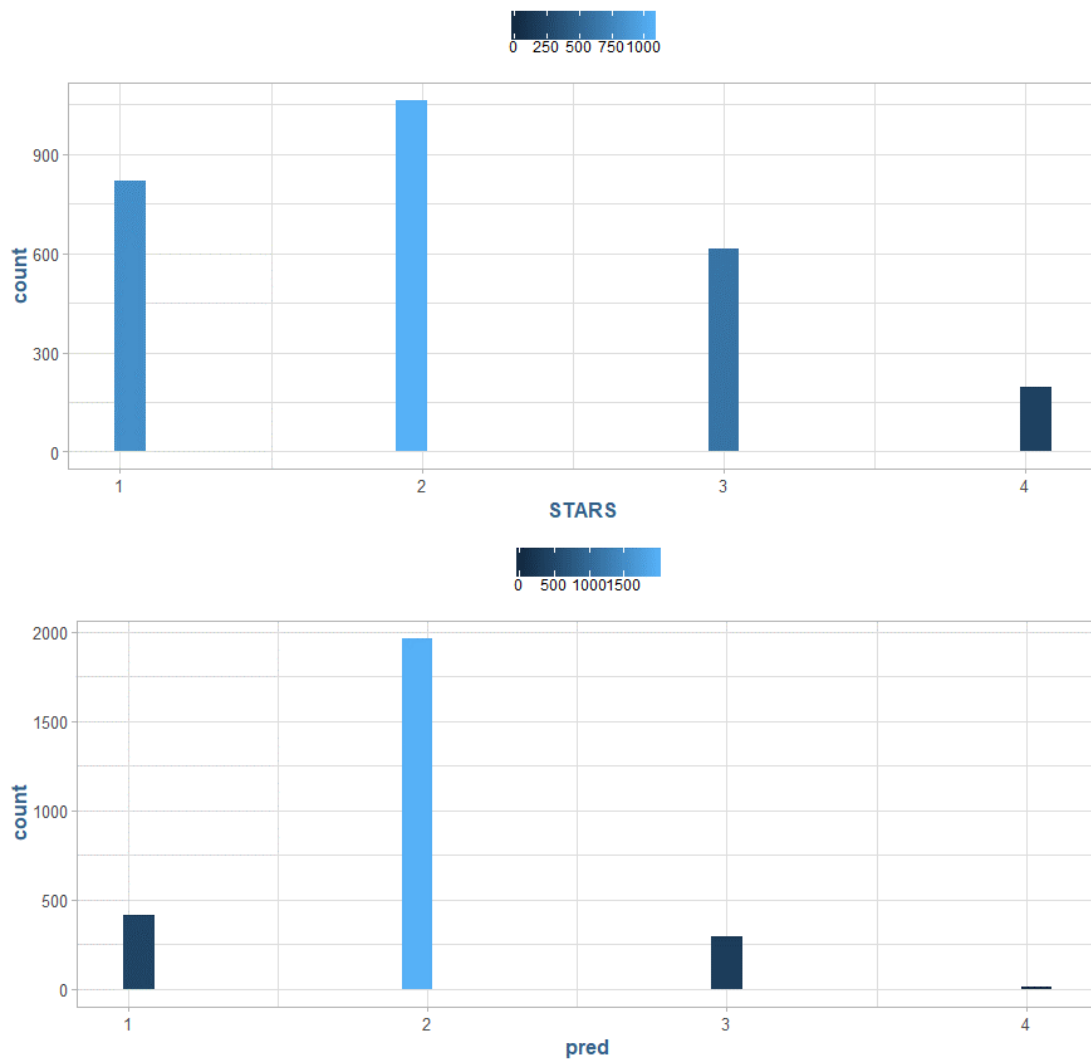
## PURCHASE MODEL

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%:
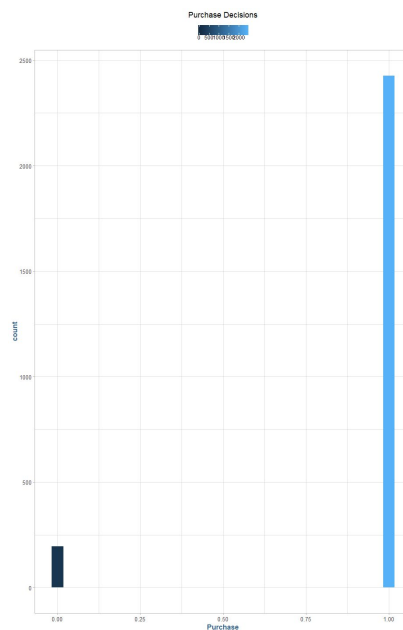


We also note that any wine we observe with over 2 stars has a 100% likelihood of getting purchased:

| 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:
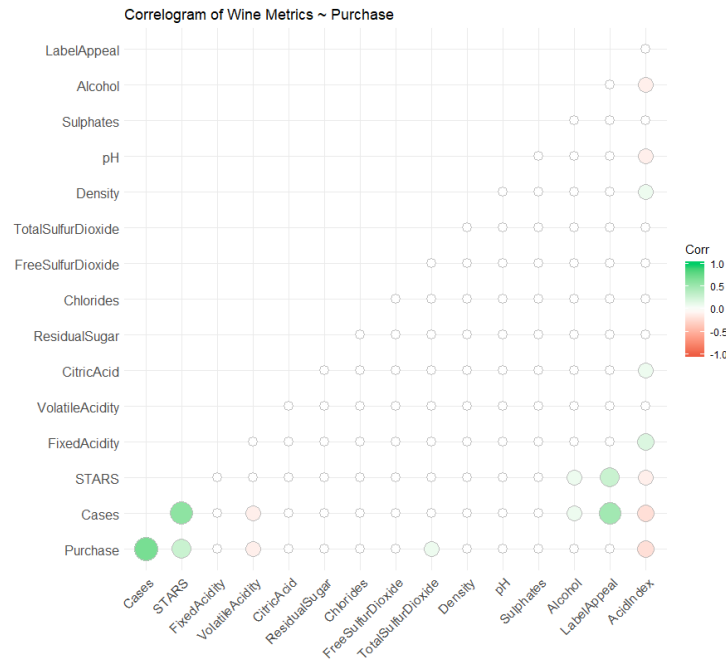


Above, we can see that there is an obviously strong relationship to cases, which implies a purchase, so we will discard that variable in this analysis. However, we also see positive correlations with stars rating, residual sugar, total sulfur dioxide, and negative correlations to volatile and total acidity.

For the purchase model we will employ similar variable selection techniques that we did in the stars rating model, namely forward, backward and stepwise variable selection. We 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 our 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.



Probability of Purchase Decision by Stars Rating
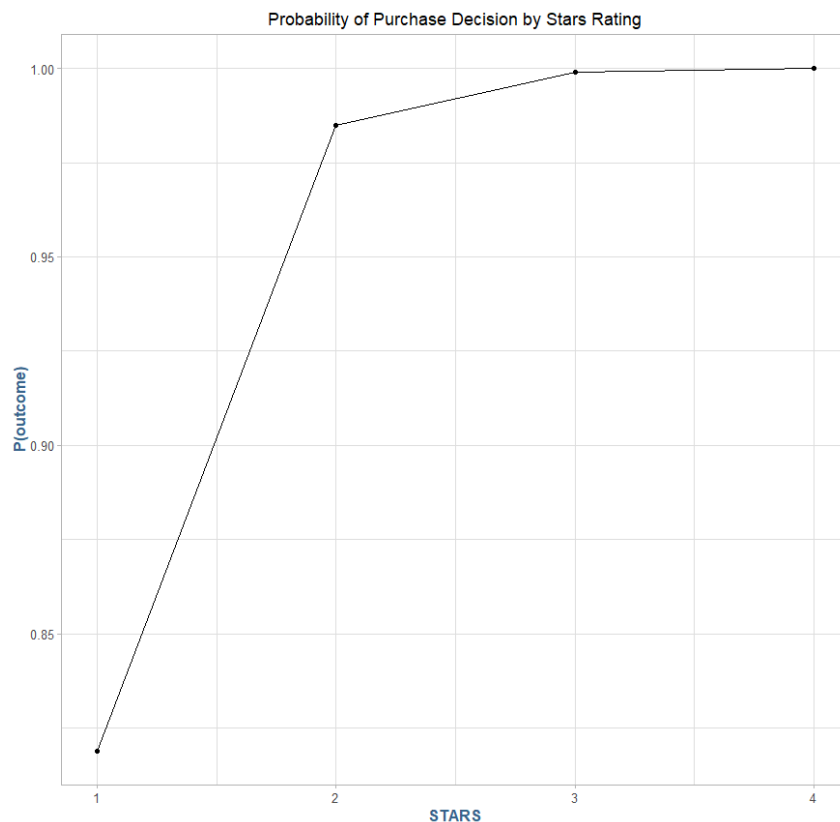
In the above diagram we see the probability of purchase decision based upon the stars rating of the wine, holding all other variables constant with their mean values.

For out of sample performance, we look at the probability of getting an accurate prediction for the purchase decision variable. We see that true purchases are predicted correctly 99.6% of the time on the test set, where as the do-not purchase value is only correctly predicted 8.9% of the time:

| Purchase | PctAccurate |
|---|---|
| 0 | 8.889 |
| 1 | 99.603 |

This is 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, would intuitively be highly correlated to the rating of the win. On the out of sample test data, the model performed extremely well for predicting purchases of three- and four-star wines, which would be our 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 we will examine in this report is the cases variable 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:
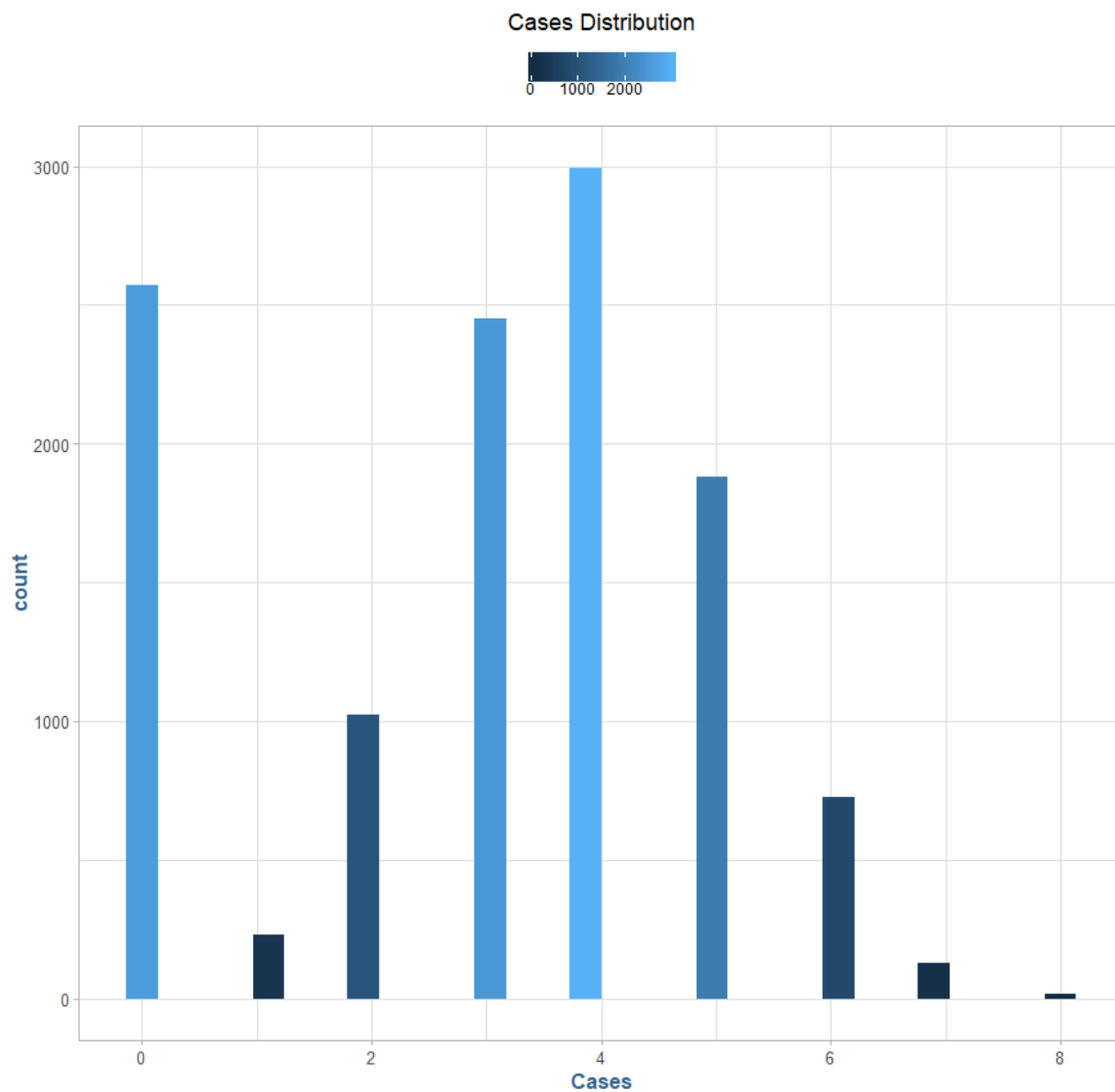


We can clearly see the dominate presence of zero values in this data set, with the rest of the values being relatively evenly distributed around the mean, mu = 4. Looking at the variables that have a high correlation to number of cases sold, we can see a clear pattern in the alcohol content of the wine to the number of cases sold:
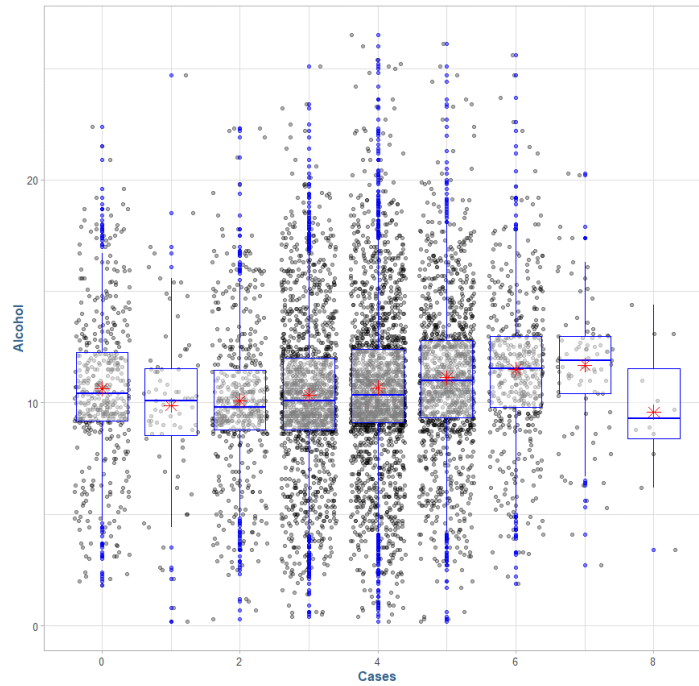
We note that there is not a high degree of any correlation between any of the chemical composition properties of the wines to the number of cases sold. For this modeling process, we declined to use the automated variable selection procedures that we employed in the previous two models, and instead start 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)*

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

*Is Zero (Binomial)?*

$$-5.011 - 3.830\beta_6 - 0.065\beta_7 - 0.534\beta_8$$

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%.

The interpretation of the coefficients in this model can be broken down by model part, for the binomial:

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, we must interpret all the coefficients together to get our estimated count. Here, we can interpret the stars rating 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, we can see the cases vs the predicted cases in the following diagram:



23
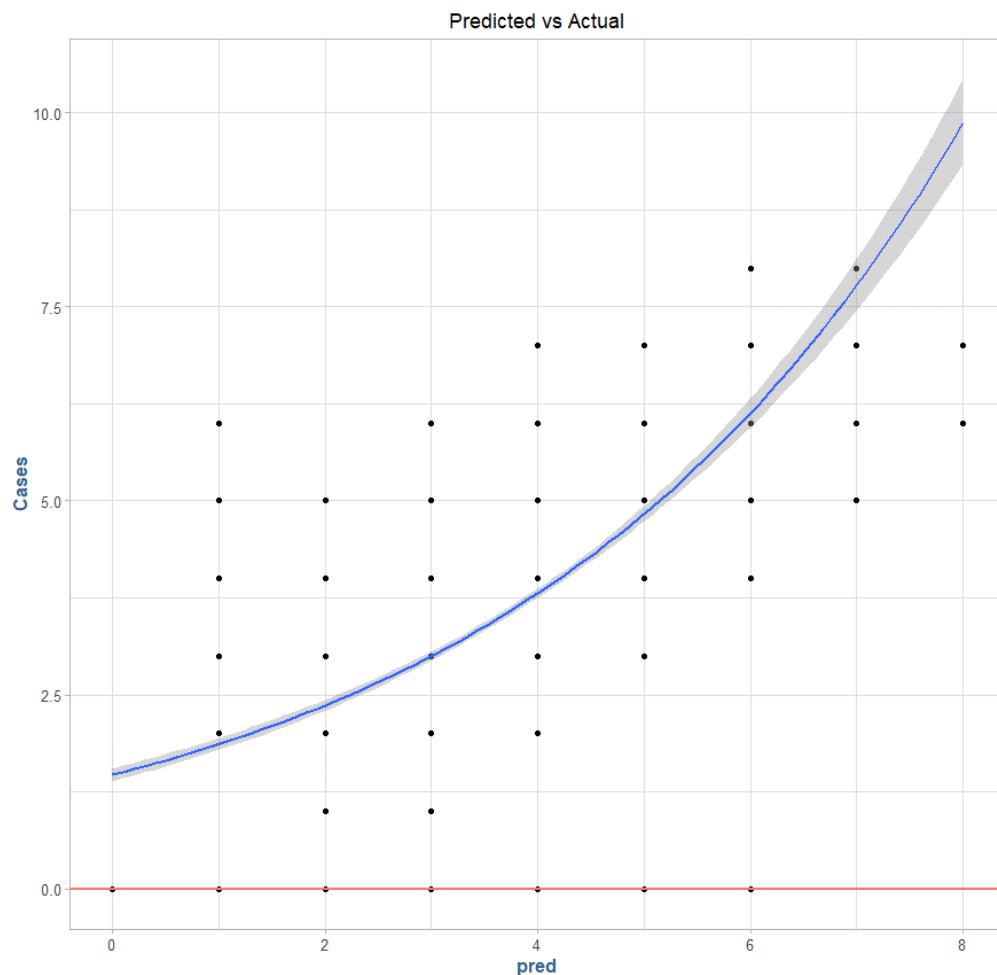
And the output from our residuals, which follow the Poisson as we would hope:
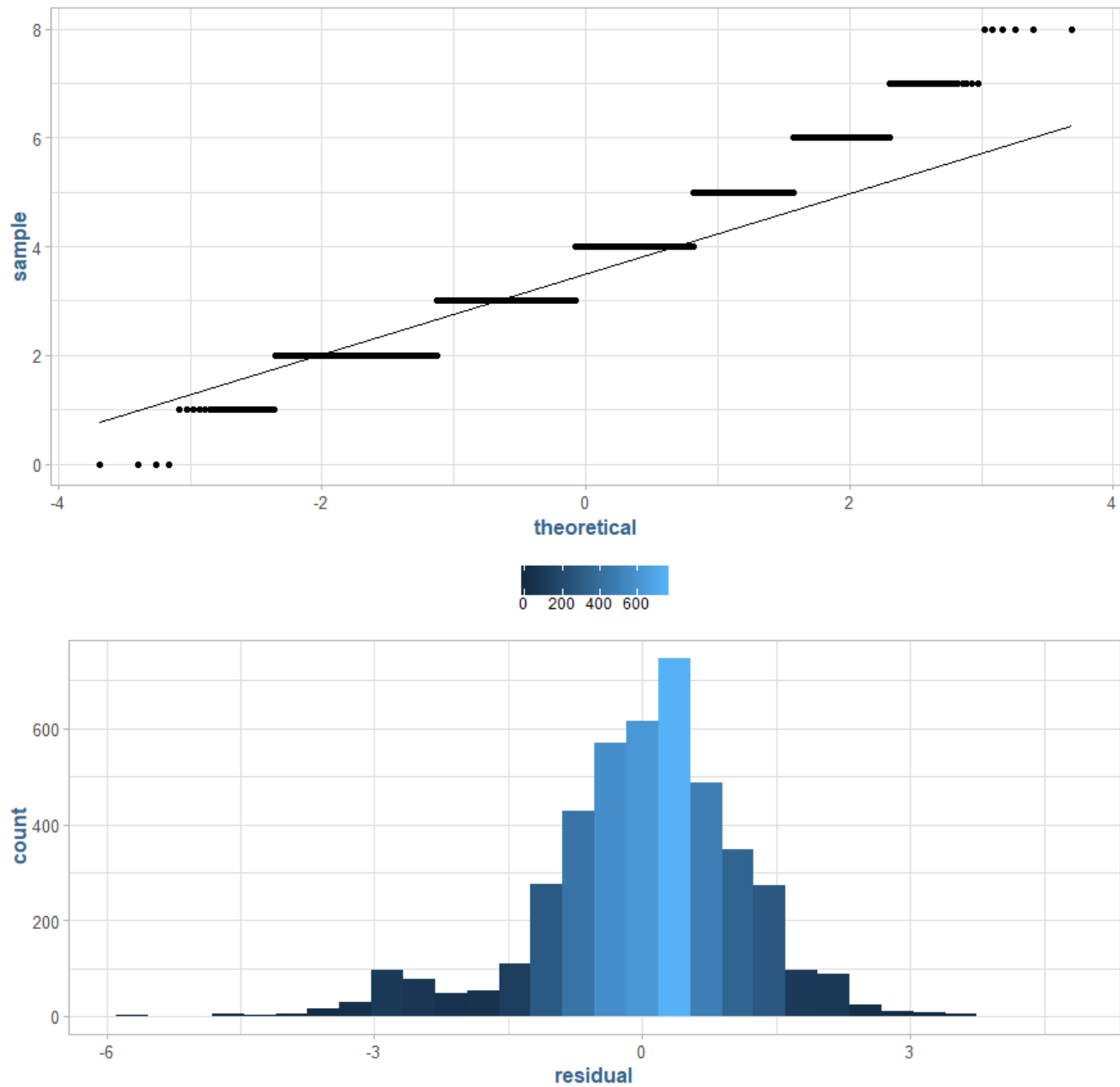


Additionally, the model performance on the test data is 40% of the out of sample cases were correctly predicted.

In this lab we explored a data set containing various attributes of bottles of wine with the goal of producing three distinct models: one for predicting an ordinal discrete variable, the stars rating, one for predicting a dichotomous variable that represents a binary yes/no decision on purchase, and finally a variable that is a heavily skewed toward zero Poisson, which represents 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 seem 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 surprisingly difficult modeling task. Starting with the exploratory data analysis which was one of the more difficult ones we have performed due to the hard to explain variable values, difficult to detect outliers and generally low correlations. However, we were able to press forward and produce three reasonable models for the task at hand, although further work on this data set would still be highly recommended.

## 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 |

Correlogram of Wine Metrics ~ Cases