Assignment 10

June 2021

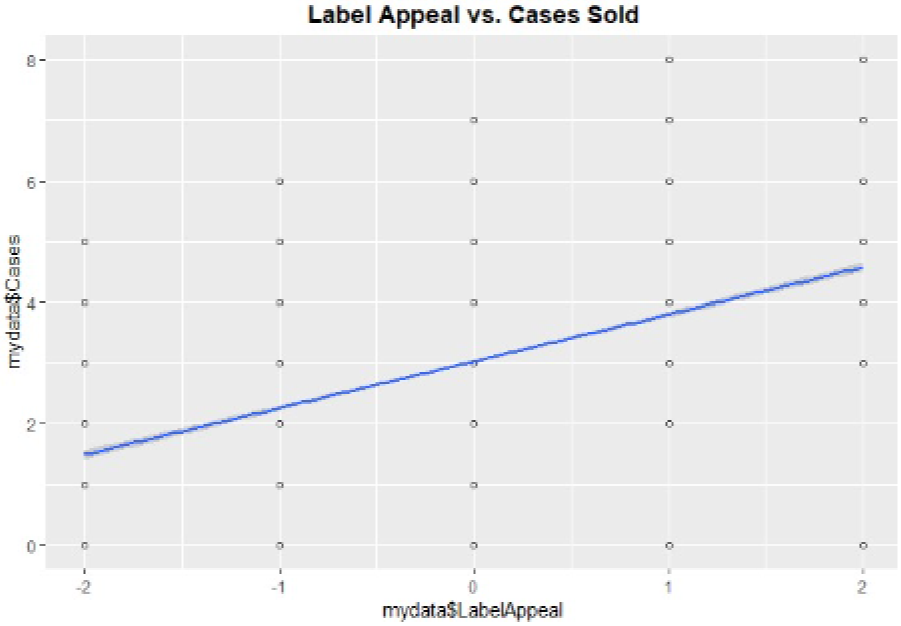
Task 1)

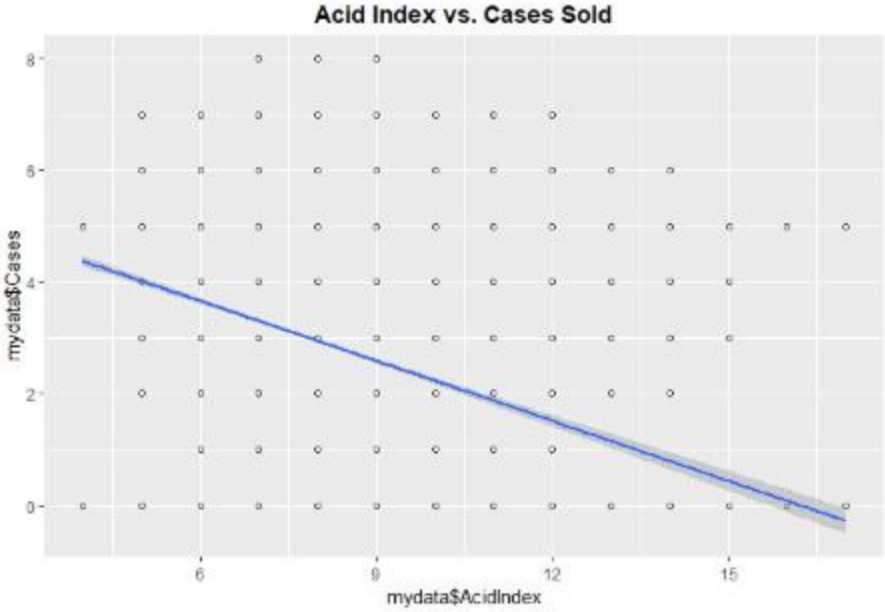
After viewing the wine data set, I decided to choose the number of cases sold as the response or target variable to build the predictive model around. I didn't want to use the variable 'Stars' as I though this may be a potential predictor variable instead for the number of cases sold. Due to the large size of the dataset which has over 12,000 rows, I will split the data into a train-test split with 70% in train and 30% in test to asses any of our modeling decisions. These will be called train.df, and test.df respectively for reference.

To perform exploratory data analysis, I will make histograms, get the mean, median, standard deviations, minimum, and maximum, and null values for all of the continuous variables which each have normal distributions with no major outliers. There are quite a bit of nulls though, including 653 for alcohol, 1210 for sulphates, 682 for sulfur dioxide, and 647 for free sulfur dioxide. There are less than 5% of the total data records with no correlation to the number of cases sol so I have decide to exclude these variables with he exception of alcohol. Since the null alcohol value indicates no alcohol present in the wine, I will replace these values with zeroes rather than removing the records.

Another variable I am interested in testing the correlation with the rating of the wine is adding together fixed acidity and volatile acidity to form a new variable of total acidity. After having done this initial exploration and created a new variable, I will do more exploration and look at the correlations between each individual variable and number of cases sold. This will include testing correlation between cases sold and the remaining non-continuous variables.

After viewing the correlation matrices and scatterplots for the predictor variables and cases sold, I can see that the strongest correlations came from label and the acid variable. Additionally the scatterplots and correlation metrics showed some correlation between cases dold and fixed acididfty, volatile acidity, acidity total, and density. I wil start the modeling process with all six of the predictor variables and compare the summary statistics between each model to determine which fits best and analyze them fully.





2)

As discussed previously, I decided to include label appeal, acid index, fixed acidity, volatile acidy, acidity and density in my model. To determine the best fitting model, I will remove each variable with the lowest correlation to from six different models. To pick the model that fits best, I will compare adjusted r-squared, f-statistics, and then check the underlying assumptions via diagnostics. I will create the models based off the training data and validate its accuracy with the test data.

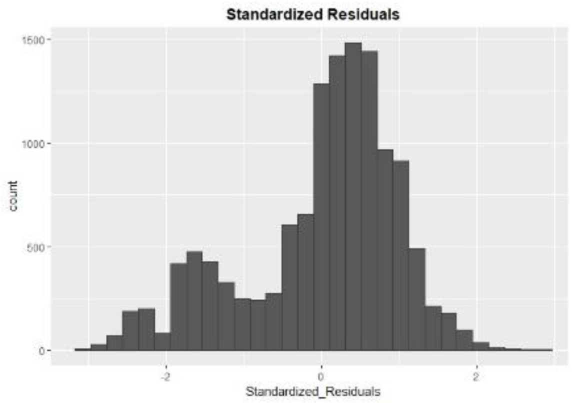
After selecting the best fitting model, I noted that a model including label and acid index is the best fit for predating the number of cases sold. When computing the summary statistics for this model, we get this model:

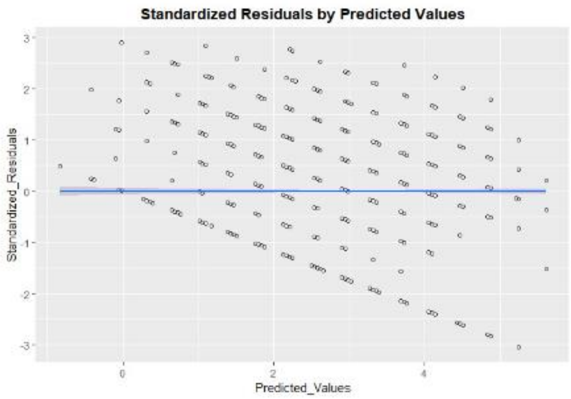
Predicted cases sold = 5.9063 + .78X1 - .37X2

I got to this model because 5.9063 is the starting point on the y axis when x = 0. This is before label appeal and the acid index is factored into the equation. .78 – for every 1 unit increase in label appeal, the number of cases sold increase by .78%. .37 – for every 1 unit increase in acid index, the number of cases dol decrease by .37%.

These coefficients seem correct as increasing the label appeal should logically increase the number of cases sold. This would also be inverse for increasing the acid index levels.

After viewing the results, I put the model to test with predicated values from our test dataset and I got the following plots that summarize the goodness of fit:





3)

As discussed in the previous section, there are some issues with our best fitting model. The standardized residuals of the predicate values skewed to the left. When gathering summary statistics for the final model, we noted the r-squared vale for our best fitting model was only 19.2% which means that only 19.2% of the variability is explained within the model.

I believe the model did not fit as well as we wanted because the dataset consisted of variables that are mostly related to the chemical properties of the wine and it did not include variables such as manufacturer, bottle shape, bottle size, price, wine type, location of grapes, and other related variables. If I couple the potential predictor variables with the label appeal and acid index we might have a better fitting model. In sum, I would not recommend any action items from our final model and instead believe that we should look at different data sources prior to building a complete model to predict the number of cases sold.