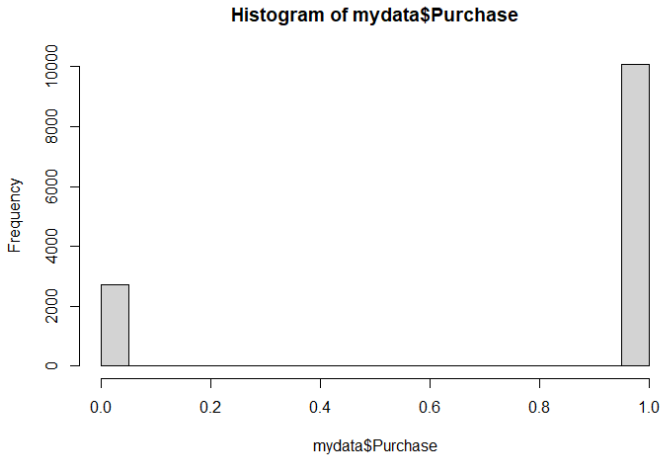
Model_Assignment_4

Jonah Muniz

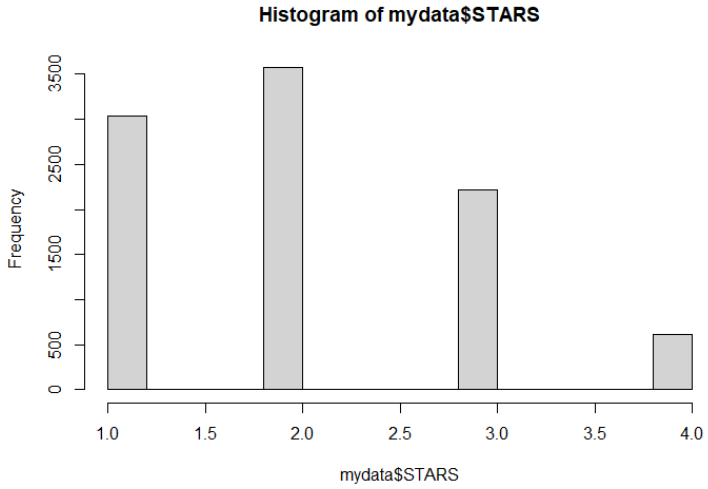
Task 1: Exploratory Data Analysis and Data Prep

The wine.csv data set is being used to perform this modeling assignment. The wine.csv dataset contains 17 variables and 12795 observations. Each row can be considered a bottle of wine. Some sample variables are, Purchase, Cases, Stars, FixedAcidity, VolatileAcidity, CholoricAcid, etc. Once the dataset was uploaded into R, some EDA was performed to gain a better understanding of the variables and their relationship with one another. At first a histogram was created for all continuous variables. Although each histogram will not be included in this report, a sample of the histogram of Purchases, Stars, and AcidIndex can be seen below.



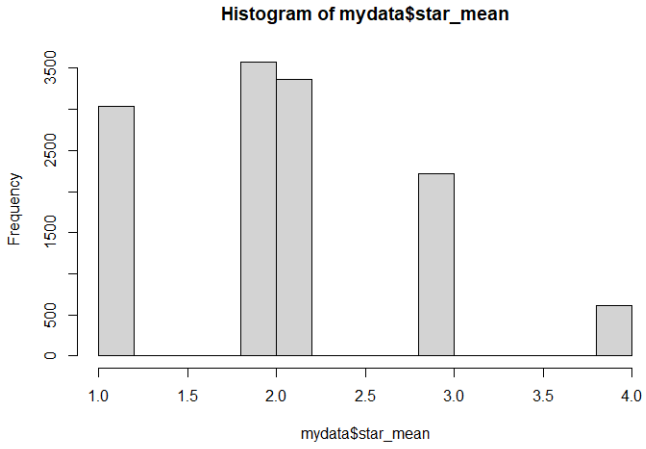
As can be seen in the histogram above, Purchase is either 0 or 1. The Purchase variable is a flag indicating whether this wine has been purchased or not. It can be seen from the histogram that majority of the wines in the wine dataset have been purchased. This histogram also points the modeler to use a logistic regression if they chose to predict whether a wine will be purchased or not due to the binomial behavior seen in the histogram.

The histogram for Stars can be seen below.



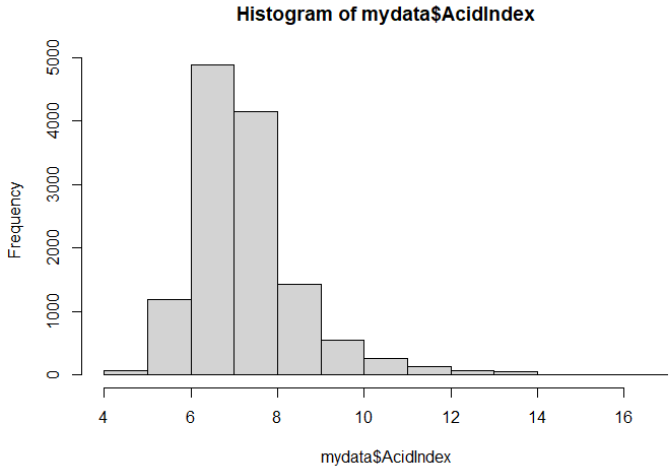
As can be seen above, all Stars values are integers in the range of 1-4. One thing that you do not see is that there are approximately 3359 Stars values that are NA. This can be handled in multiple ways, deleting all rows with NA, imputing the Stars values by using the mean or median. Based off of the summary statistic of Stars, seen below, the mean of Stars was used to impute the value of NA. The new Stars variable with NA replaced with mean(Stars) can be seen below as well.

```
summary(mydata$STARS)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
1.000   1.000   2.000   2.042   3.000   4.000    3359
```



You can now see the additional bar that has been added to the histogram at approximately 2.042.

Lastly, a histogram of AcidIndex can be seen below.



As can be seen above, the histogram for AcidIndex is pretty normal with a slight right tail.

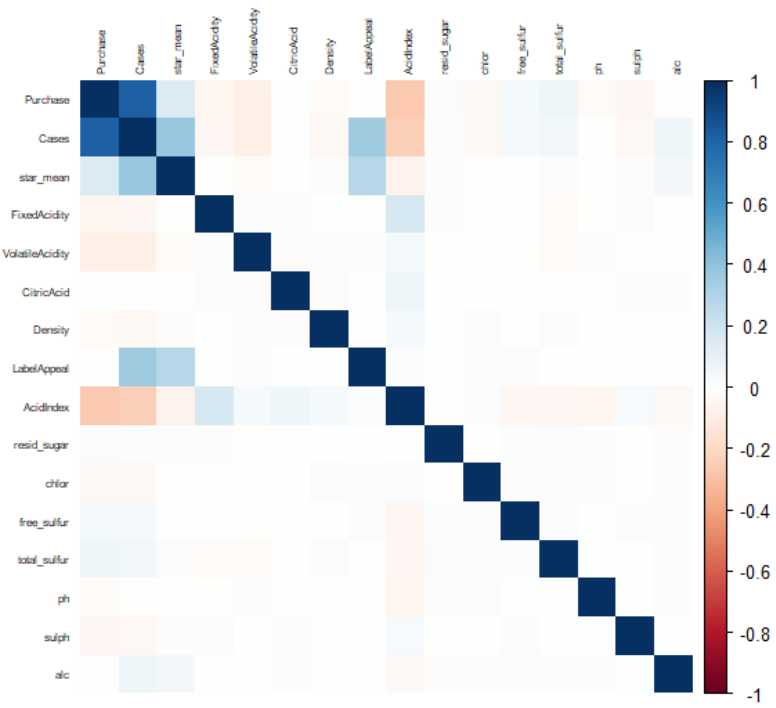
Stars is not the only variable in the wine dataset that contains NA values. The summary statistics of all 17 variables can be seen below. More importantly you can see the 8 variables that contain NA values.

```
      ï..INDEX        Purchase          Cases           STARS        FixedAcidity     VolatileAcidity     CitricAcid      ResidualSugar
 Min.   :    1   Min.   :0.0000   Min.   :0.000   Min.   :1.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400   Min.   :-127.800
 1st Qu.: 4038   1st Qu.:1.0000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300   1st Qu.:  -2.000
 Median : 8110   Median :1.0000   Median :3.000   Median :2.000   Median :  6.900   Median : 0.2800   Median : 0.3100   Median :   3.900
 Mean   : 8070   Mean   :0.7863   Mean   :3.029   Mean   :2.042   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084   Mean   :   5.419
 3rd Qu.:12106   3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800   3rd Qu.:  15.900
 Max.   :16129   Max.   :1.0000   Max.   :8.000   Max.   :4.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600   Max.   : 141.150
                                                  NA's   :3359                                                          NA's   :616
   Chlorides       FreeSulfurDioxide TotalSulfurDioxide    Density             pH          Sulphates          Alcohol
 Min.   :-1.1710   Min.   :-555.00   Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300   Min.   :-4.70
 1st Qu.:-0.0310   1st Qu.:   0.00   1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800   1st Qu.: 9.00
 Median : 0.0460   Median :  30.00   Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000   Median :10.40
 Mean   : 0.0548   Mean   :  30.85   Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271   Mean   :10.49
 3rd Qu.: 0.1530   3rd Qu.:  70.00   3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600   3rd Qu.:12.40
 Max.   : 1.3510   Max.   : 623.00   Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400   Max.   :26.50
 NA's   :638       NA's   :647       NA's   :682                         NA's   :395     NA's   :1210      NA's   :653
  LabelAppeal         AcidIndex
 Min.   :-2.000000   Min.   : 4.000
 1st Qu.:-1.000000   1st Qu.: 7.000
 Median : 0.000000   Median : 8.000
 Mean   :-0.009066   Mean   : 7.773
 3rd Qu.: 1.000000   3rd Qu.: 8.000
 Max.   : 2.000000   Max.   :17.000
```

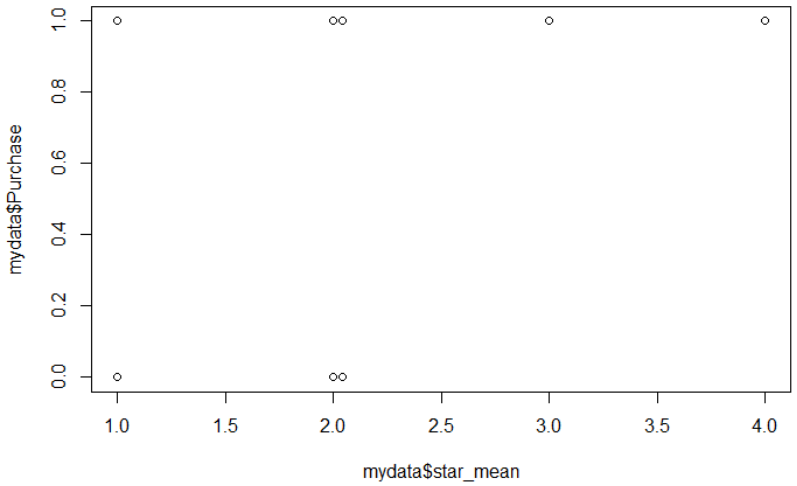
The 8 variables that contained NA values had the NA values replaced with the mean value of the variable. This was done to ensure we keep as much data as possible available to us for modeling. The new summary statistics for the 8 variables can be seen below.

```
   star_mean      resid_sugar          chlor           free_sulfur       total_sulfur          ph            sulph
 Min.   :1.000   Min.   :-127.800   Min.   :-1.17100   Min.   :-555.00   Min.   :-823.0   Min.   :0.480   Min.   :-3.1300
 1st Qu.:2.000   1st Qu.:   0.900   1st Qu.: 0.00000   1st Qu.:   5.00   1st Qu.:  34.0   1st Qu.:2.970   1st Qu.: 0.3400
 Median :2.000   Median :   4.900   Median : 0.04800   Median :  30.85   Median : 120.7   Median :3.208   Median : 0.5271
 Mean   :2.042   Mean   :   5.419   Mean   : 0.05482   Mean   :  30.85   Mean   : 120.7   Mean   :3.208   Mean   : 0.5271
 3rd Qu.:2.042   3rd Qu.:  14.900   3rd Qu.: 0.12800   3rd Qu.:  64.00   3rd Qu.: 198.0   3rd Qu.:3.450   3rd Qu.: 0.7700
 Max.   :4.000   Max.   : 141.150   Max.   : 1.35100   Max.   : 623.00   Max.   :1057.0   Max.   :6.130   Max.   : 4.2400
      alc
 Min.   :-4.70
 1st Qu.: 9.10
 Median :10.49
 Mean   :10.49
 3rd Qu.:12.20
 Max.   :26.50
```

Now that all NA values in the dataset have been addressed, the correlation matrix for the quantitative variables can be conducted to understand which variables have a strong relationship with one another. The Pearson correlation matrix can be seen below.



As can be seen above, majority of the variables are not correlated with one another. You can see that there is a strong correlation with Cases and Purchase which makes a lot of sense. You cannot have a Case unless there was a Purchase. Star_mean, the imputed Stars variable, has a strong correlation with both Purchase and Cases which also makes sense. Higher rated wines usually get purchased and when they do multiple cases are purchased. Star_mean also has a correlation with LabelAppeal. Purchase has a correlation with star_mean and AcidIndex. Lastly, Cases has a correlation with LabelAppeal and AcidIndex. The rest of the correlations seem insignificant. The scatter plots were then generated to evaluate the relationship these variables had with one another. A sample scatter plot can be seen below of star_mean and purchase.



As can be seen above, there seems to be a cutoff value for star_mean where once star_mean passes the value, it is purchased. This is indicative of an S-curve, something very familiar and desirable when modeling a logistic regression.

Task 2: Modeling

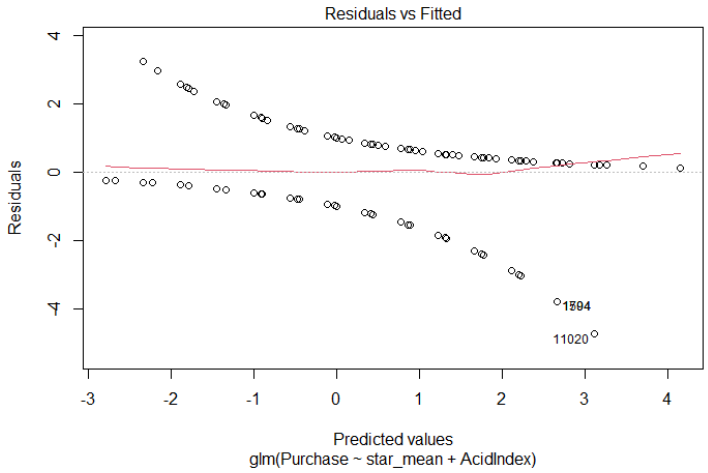
There are multiple models that can be fitted to predict a variable in the wine dataset. For example, the Stars value can be predicted, as well as whether a wine will be purchased, and if it is purchased, how many cases will be purchased. Based off the variable correlations and the behaviors witnessed when looking at scatterplots, a logistic regression model was selected to predict whether a wine will be purchased or not. This type of modeling technique was used for this use case because logistic regression models are binomial, which means they are great at predicting whether an event occurs, like a purchase.

Now that the model technique and the objective have been defined, the process of variable selection was needed. Based on the Pearson correlation matrix produced during the EDA process, I chose to select the variables for the model by hand. There were a few variables that showed any correlation, negative or positive, to the Purchase variables. These variables were Cases, star_mean, and AcidIndex. Upon further investigation, Cases was no used in the model due to its strong correlation with Purchase and leading to 100% accuracy due to its dependency of one another. A regression model was then fitted using star_mean and AcidIndex as the explanatory variables for Purchase. A summary of the model can be seen below, as well as the interpretation of the coefficients and checks on assumptions via diagnostic.

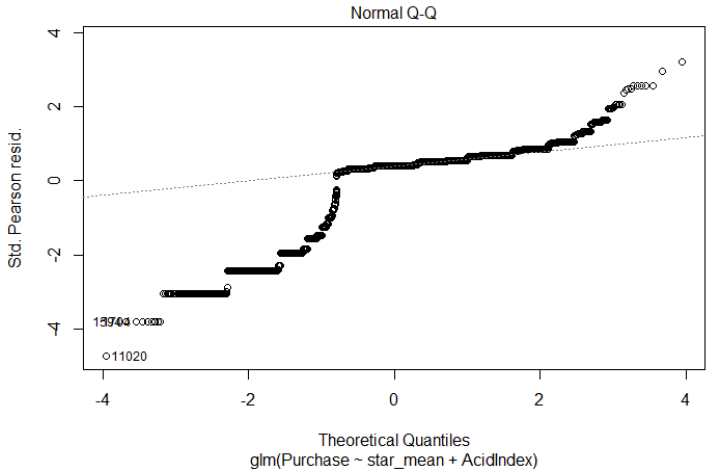
```
Call:
glm(formula = Purchase ~ star_mean + AcidIndex, family = binomial,
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5136   0.3549   0.5642   0.6907   2.2043

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.81368    0.14742   25.87   <2e-16 ***
star_mean    0.53083    0.03234   16.41   <2e-16 ***
AcidIndex   -0.44545    0.01661  -26.82   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13276  on 12794  degrees of freedom
Residual deviance: 12150  on 12792  degrees of freedom
AIC: 12156

Number of Fisher Scoring iterations: 4
```

As can be seen in the summary above, the equation for Purchase = 3.8137 + 0.5308*star_mean – 0.4455*AcidIndex. A hypothesis test can be conducted for the coefficients. The null hypothesis is that beta1 = 0. The alternate hypothesis is that beta1 does not equal 0. Based on the p-value < 0.05 the null hypothesis can be rejected. When star_mean increases by one unit Purchase increases by 70.0346% when all other coefficients are held constant. The null hypothesis is that beta2 = 0. The alternate hypothesis is that beta2 does not equal 0. Based on the p-value < 0.05 the null hypothesis can be rejected. When AcidIndex increases by one unit Purchase decreases by 35.9463% when all other coefficients are held constant. The AIC for this mode is 12156 and the goodness of fit is 0.9498 which is

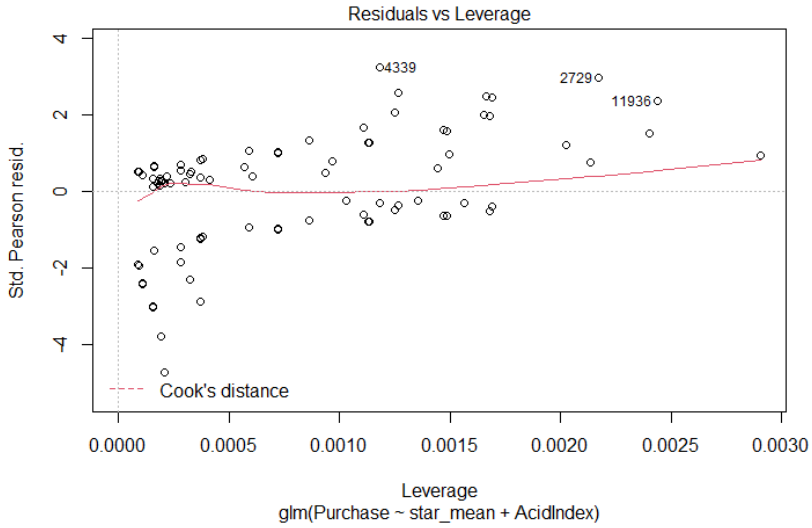
close to the ideal goodness of fit of 1. The diagnostic plots can be analyzed as well to see if the assumptions hold for the model.



As can be seen above, although the red line is very close to the ideal line at 0, there is a distinct pattern with the residuals when plotted against the fitted values. This shows that the residuals are not random.



As can be seen in the q-q plot the model is not normal or linear which is not surprising due to the fact that the logistic regression model was used due to the nonlinear behavior in Purchase and the explanatory variables.



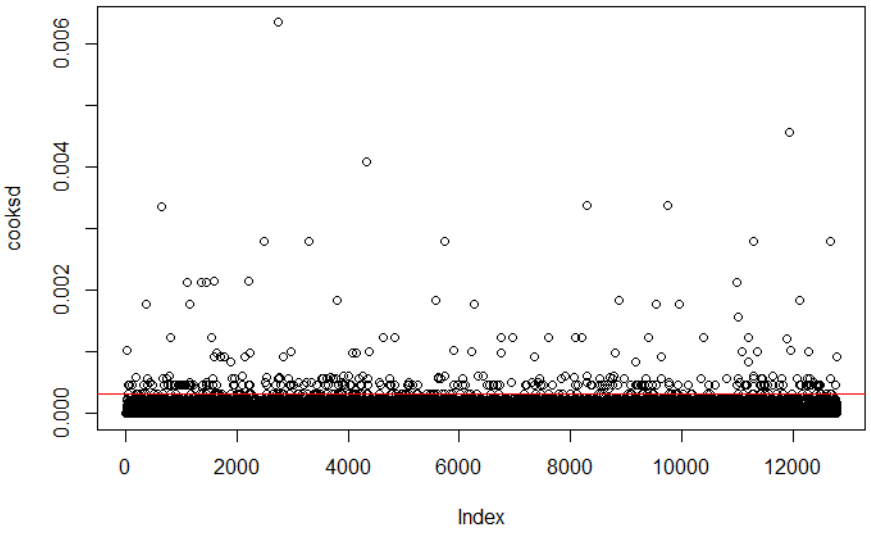
Lastly, when reviewing the residuals vs leverage plot potentially influential data points can be seen in the plot as 4339, 2729, and 11936.

A confusion matrix can then be created to measure the accuracy of the model to predict whether a purchase occurs. Below the confusion matric for this model can be found.

target	prediction 0	1
0	394	2340
1	214	9847

The model's accuracy can then be calculated, # of predictions correct/Total predictions = (9847+394)/12795*100 = 80.0391%.

Due to these influential data points the cook's distance for each point was determined in the model. A point was considered influential if the cook's distance exceeded 4/# of rows in the dataset. A plot can be seen below of the cook's distance for each point and the red cutoff line for an allowable cook's distance.

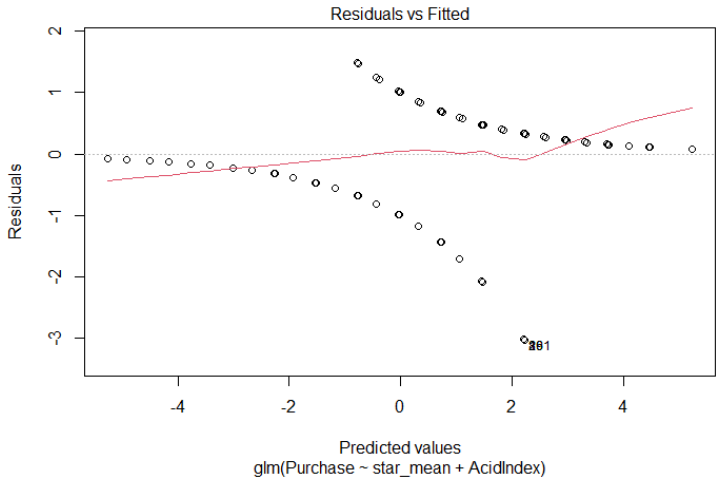


All values above the red line were analyzed and determined to be removed from the dataset. Total data points in the dataset were reduced from 12795 to 12358 which is a -3.4154% decrease in total observations which was a comfortable number of points removed to still feel confident in any conclusions from this dataset limiting any bias. A regression model was then refitted with these influential datapoints removed. A summary of the model can be seen below, as well as the interpretation of the coefficients and checks on assumptions via diagnostic.

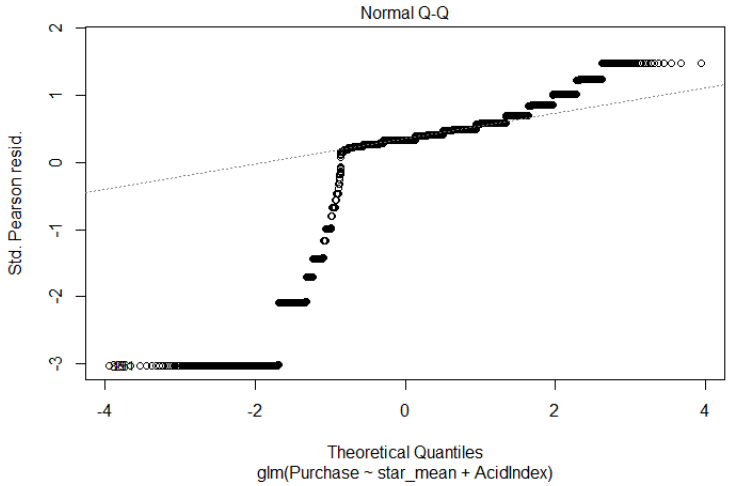
```
Call:
glm(formula = Purchase ~ star_mean + AcidIndex, family = binomial,
    data = mydata_scrub)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1561   0.3165   0.4573   0.6470   1.5238

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.65157    0.18747   35.48   <2e-16 ***
star_mean    0.39328    0.03499   11.24   <2e-16 ***
AcidIndex   -0.74759    0.02119  -35.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12161  on 12357  degrees of freedom
Residual deviance: 10372  on 12355  degrees of freedom
AIC: 10378

Number of Fisher Scoring iterations: 5
```

As can be seen in the summary above, the equation for Purchase = 6.6516 + 0.3933*star_mean – 0.7476 *AcidIndex. A hypothesis test can be conducted for the coefficients. The null hypothesis is that beta1 = 0. The alternate hypothesis is that beta1 does not equal 0. Based on the p-value < 0.05 the null hypothesis can be rejected. When star_mean increases by one unit Purchase increases by 48.1832% when all other coefficients are held constant. The null hypothesis is that beta2 = 0. The alternate hypothesis is that beta2 does not equal 0. Based on the p-value < 0.05 the null hypothesis can be rejected. When AcidIndex increases by one unit Purchase decreases by 52.6494% when all other

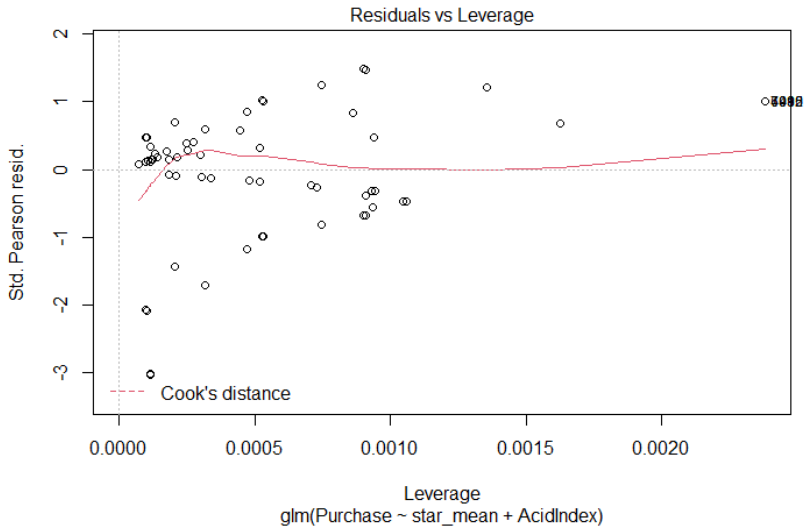
coefficients are held constant. The AIC for this mode is 10378 and the goodness of fit is 0.8395 which is close to the ideal goodness of fit of 1. The diagnostic plots can be analyzed as well to see if the assumptions hold for the model.



As can be seen above, although the red line is very close to the ideal line at 0, there is a distinct pattern with the residuals when plotted against the fitted values. This shows that the residuals are not random.



As can be seen in the q-q plot the model is not normal or linear which is not surprising due to the fact that the logistic regression model was used due to the nonlinear behavior in Purchase and the explanatory variables.



Lastly, when reviewing the residuals vs leverage plot, the previous potentially influential points were removed. There seems to be additional points that may need to be reviewed but all remaining data points are not considered influential.

A confusion matrix can then be created to measure the accuracy of the model to predict whether a purchase occurs. Below the confusion matric for this model can be found.

```
       prediction
target    0    1
     0  606 1792
     1  304 9656
```

The model's accuracy can then be calculated, # of predictions correct/Total predictions = (9656+606)/12358*100 = 83.0393%.

Both models can now be compared to see which is the best at accurately predicting is a bottle of wine will be purchased. Below is a table comparing the two models AIC and accuracy metrics.

Model	AIC	Accuracy
Logistic Regression (All Observations)	12156	80.0391
Logistic Regression (Influential Removed)	10378	83.0393

As can be seen in the table above the model with the influential datapoints removed performs better when comparing across AIC and Accuracy. Due to this, the model using stars_mean and AcidIndex as explanatory variables of Purchase with influential points removed is the best model to predict if a wine will be purchased in the future.

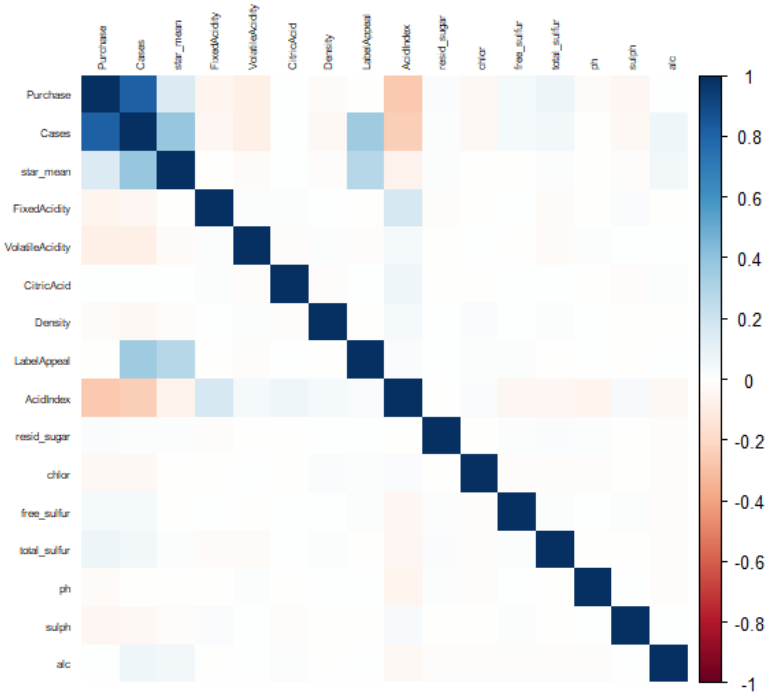
Task 3: Conclusions and Reflections

After performing the above analysis with the wine data set, I have learned that only two variables are necessary to predict if a bottle of wine will be purchased, total stars and AcidIndex. It was discovered that when you increase the stars of a wine by one the chances of that wine being purchased increases

by ~48% when everything else is kept the same. When AcidIndex increases by one, the chances that the wine will be purchase decreases by ~53%. The discovery that the probability of purchase going up when stars increase was not a novel find but learning how much the probability decreases when AcidIndex increases was novel. I am not a big wine drinker, but I did not know how important acidity level is. I am curious how someone would know the acidity level of a wine without tasting it. It was also surprising that some of the other variables had no significant effect on the probability of purchasing a bottle of wine. Based off this model and analysis I would advice someone in this field that if they were trying to sell more bottles, I would lower the AcidIndex of their wine. I assume it would be easier to change the AcidIndex of a wine then increasing the stars of a wine.

Extra Credit:

Now that a logistic regression model has been fitted to predict whether a wine was purchases or not, a Poisson regression model can be fitted to understand if a wine is purchased, how many cases was purchased. Now that the model of choice has been decided, the same correlation matrix produced to understand which variables are correlated to Purchase can be used to understand which variables are correlated with Cases. Please see below for the correlation matrix.



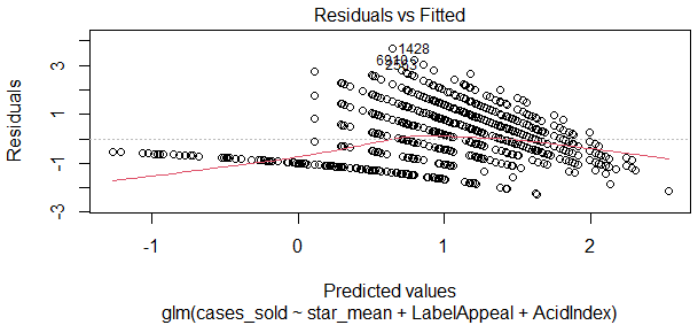
As can be seen in the matrix above, there is a strong correlation with Purchase which we discuss in the assignment above. There is also a correlation with star_mean, LabelAppeal and AcidIndex. Cases is correlated with the same variables as Purchase, which makes sense due to their innate correlation, but there is a new variable that is correlated with Cases that is not with Purchase, LabelAppeal.

The logistic regression models predicted values are leveraged to predict the number of cases solid for the bottles that were predicted to be sold. A new variable of cases_sold was created to ensure that only the wine bottles that were predicted to be purchased leveraging the logistic regression model are leveraged to train the Poisson regression model. If the predicted value for Purchase is 1, cases_sold is

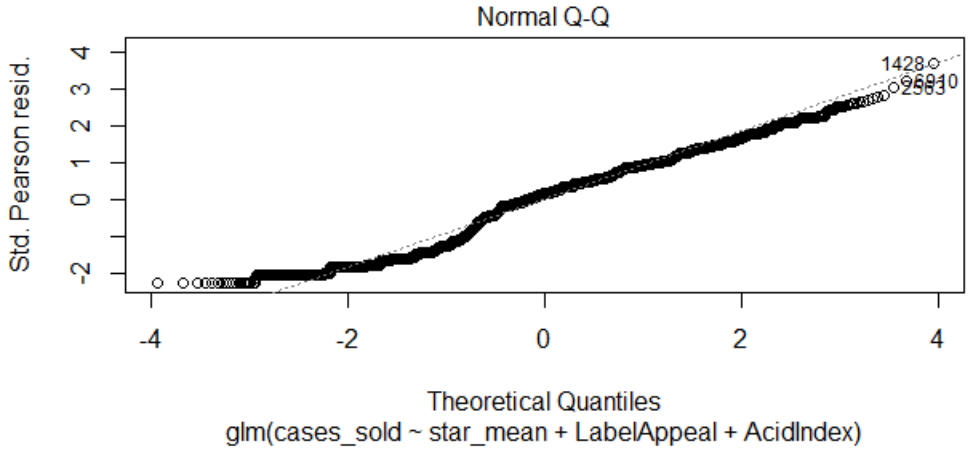
replaced with the original cases data, if predicted value for Purchase is 0, cases_sold is set to 0. A summary of the Poisson model can be found below.

```
Call:
glm(formula = cases_sold ~ star_mean + LabelAppeal + AcidIndex,
    family = poisson, data = mydata_scrub)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1993  -0.6724   0.1578   0.6024   2.8299

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.576854   0.041760   61.71   <2e-16 ***
star_mean    0.184198   0.006630   27.79   <2e-16 ***
LabelAppeal  0.208863   0.006142   34.01   <2e-16 ***
AcidIndex   -0.248282   0.005021  -49.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22469  on 12357  degrees of freedom
Residual deviance: 16692  on 12354  degrees of freedom
AIC: 47403

Number of Fisher Scoring iterations: 5
```

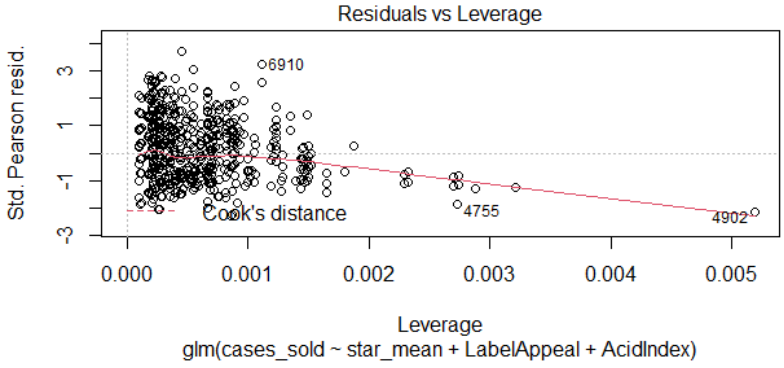
As can be seen in the summary above, the equation for cases_sold = 2.5769 + 0.1842*star_mean + 0.2089*LabelAppeal – 0.2483 *AcidIndex. A hypothesis test can be conducted for the coefficients. The null hypothesis is that beta1 = 0. The alternate hypothesis is that beta1 does not equal 0. Based on the p-value < 0.05 the null hypothesis can be rejected. When star_mean increases by one unit, cases_sold increases by 20.2257% when all other coefficients are held constant. The null hypothesis is that beta2 = 0. The alternate hypothesis is that beta2 does not equal 0. Based on the p-value < 0.05 the null hypothesis can be rejected. When LabelAppeal increases by one unit, cases_sold increases by 23.2276% when all other coefficients are held constant. The null hypothesis is that beta3 = 0. The alternate hypothesis is that beta3 does not equal 0. Based on the p-value < 0.05 the null hypothesis can be rejected. When AcidIndex increases by one unit, cases_sold decreases by 21.9761% when all other coefficients are held constant. The AIC for this mode is 47403 and the goodness of fit is 1.3511 which is close to the ideal goodness of fit of 1. The diagnostic plots can be analyzed as well to see if the assumptions hold for the model.



As can be seen above, the red trendline doe not follow the ideal straight random line. You can see that there is a distinct pattern, and the residuals are not random.



As can be seen in the q-q plot the model is slightly normal or linear. The points seem to follow the ideal straight line on the q-q plot except for a long tail found on the bottom left of the plot.

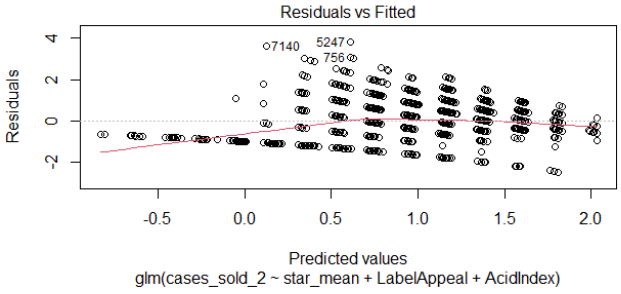


Lastly, when reviewing the residuals vs leverage plot, you can see some potentially high leverage data points that will be investigated and potentially removed if cooks distance exceeds the limit.

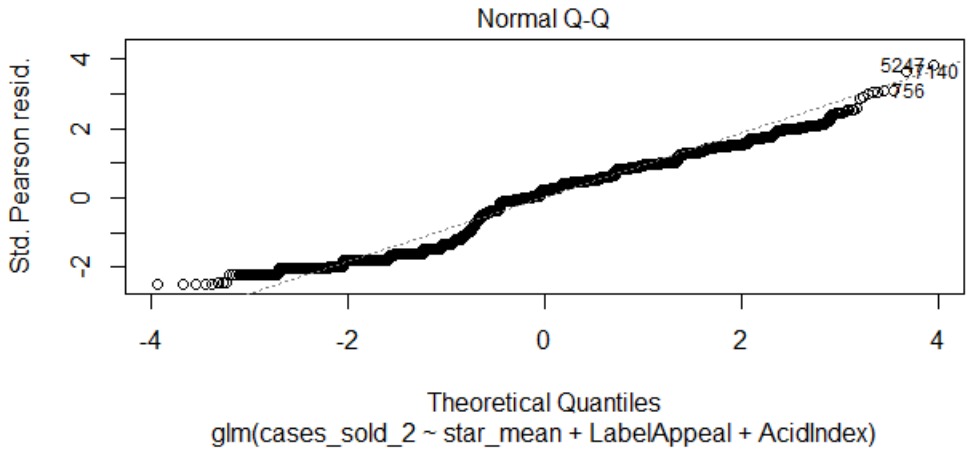
After analyzing the data points and calculating the cook's distance, 286 data points were found to have a cook's distance greater than 0.000313. The above model is rerun now that these influential points are removed to review if the performance improves. A summary of the Poisson model with the influential points removed can be found below.

```
Call:
glm(formula = cases_sold_2 ~ star_mean + LabelAppeal + AcidIndex,
    family = poisson, data = mydata_scrub_2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4954  -0.6383   0.2278   0.6086   2.8980

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.115343   0.040202   52.62   <2e-16 ***
star_mean    0.215142   0.006688   32.17   <2e-16 ***
LabelAppeal  0.220838   0.006332   34.88   <2e-16 ***
AcidIndex   -0.197754   0.004837  -40.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 23097  on 12508  degrees of freedom
Residual deviance: 17764  on 12505  degrees of freedom
AIC: 48393

Number of Fisher Scoring iterations: 5
```

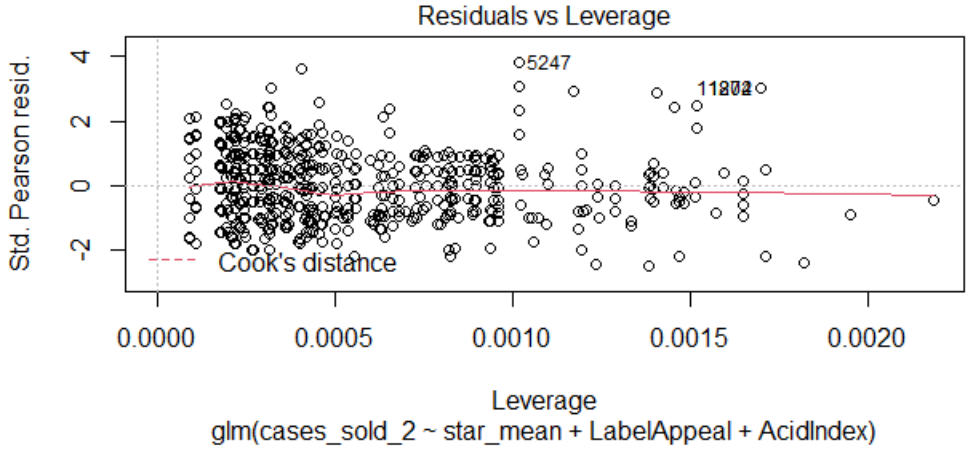
As can be seen in the summary above, the equation for cases_sold = 2.1153 + 0.2151*star_mean + 0.2208*LabelAppeal – 0.1978 *AcidIndex. A hypothesis test can be conducted for the coefficients. The null hypothesis is that beta1 = 0. The alternate hypothesis is that beta1 does not equal 0. Based on the p-value < 0.05 the null hypothesis can be rejected. When star_mean increases by one unit, cases_sold increases by 24.0038% when all other coefficients are held constant. The null hypothesis is that beta2 = 0. The alternate hypothesis is that beta2 does not equal 0. Based on the p-value < 0.05 the null hypothesis can be rejected. When LabelAppeal increases by one unit, cases_sold increases by 24.7121% when all other coefficients are held constant. The null hypothesis is that beta3 = 0. The alternate hypothesis is that beta3 does not equal 0. Based on the p-value < 0.05 the null hypothesis can be rejected. When AcidIndex increases by one unit, cases_sold decreases by 17.9428% when all other coefficients are held constant. The AIC for this mode is 48393 and the goodness of fit is 1.420552 which is close to the ideal goodness of fit of 1. The diagnostic plots can be analyzed as well to see if the assumptions hold for the model.



As can be seen above, the red trendline does not follow the ideal straight random line. You can see that there is a distinct pattern, and the residuals are not random.



As can be seen in the q-q plot the model is slightly normal or linear. The points seem to follow the ideal straight line on the q-q plot except for a long tail found on the bottom left of the plot.



Lastly, when reviewing the residuals vs leverage plot, you can see that the trendline is closer to the ideal straight line.

Now that we have two Poisson regression models to predict the number of cases sold, the mean squared error for both models can be calculated to see which model is better at predicting the number of cases sold. Below is a table comparing the two models and their mean squared error.

Model	Mean Squared Error
Poisson #1	0.3984
Poisson #2	0.4047

As can be seen in the table above, the first Poisson model is better at predicting the number of cases sold.

After completing this additional model fitting it can be said that star_mean, LabelAppeal, and AcidIndex are important variables when predicting the number of cases sold. If a wine maker is trying to increase their total cases sold, I would suggest that they work on their label appeal and try to increase the overall appeal. If they are able to increase their label appeal by one, they will see an increase of ~23% in the number of cases sold.