DATA 621 - Homework 5

Coffy Andrews-Guo, Krutika Patel, Alec McCabe, Ahmed Elsaeyed, Peter Phung

2022-11-29

Problem Statement and Goals

In this report, we generate a count regression model that is able to predict the number of cases of wine that will be sold given certain properties of the wine. The independent and dependent variables that are used in order to generate this model use data from 12,000 commercially available wines. The analysis detailed in this report shows the testing of several models:

- Four different poisson regression models
- Four different negative binomial regression models
- Four different multiple linear regression models

From these models, a best model was selected based on model performance and various metrics. Note that the multiple linear regression models were provided in this analysis for comparison purposes and ultimately a count regression model was selected for model deployment.

Data Exploration

The following is a summary of the variables provided within the data to generate the count regression model.

Variable Name	Definition	Theoretical Effect
INDEX	Identification Variable (do not use)	None
TARGET	Number of Cases Purchased	None
AcidIndex	Proprietary method of testing total acidity of wine by using a weighted average	
Alcohol	Alcohol Content	
Chlorides	Chloride content of wine	
CitricAcid	Citric Acid Content	
Density	Density of Wine	
FixedAcidity	Fixed Acidity of Wine	
FreeSulfurDioxide	Sulfur Dioxide content of wine	
LabelAppeal	Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customes don't like the design.	Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales.
ResidualSugar	Residual Sugar of wine	

Variable Name	Definition	Theoretical Effect
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor	A high number of stars suggests high sales
Sulphates	Sulfate content of wine	
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	
VolatileAcidity	Volatile Acid content of wine	
pH	pH of wine	

Table 1: Variables in the dataset

A summary of the variables is shown below. The summary itself reveals some interesting characteristics about the data. `Density`, `pH`, `AcidIndex`, `STARS`, and `LabelAppeal` are the only variables where their minimums are not negative, while the rest of the predictor variables are negative. It would also seem that `TARGET`, `LabelAppeal` and `STARS` are discrete variables and were therefore treated as such throughout this report. Note that the summary below shows the `INDEX` variable which was ignored throughout this analysis.

```
     TARGET      FixedAcidity     VolatileAcidity     CitricAcid
 4      :3177   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400
 0      :2734   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300
 3      :2611   Median :  6.900   Median : 0.2800   Median : 0.3100
 5      :2014   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084
 2      :1091   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800
 6      : 765   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600
 (Other): 403
 ResidualSugar        Chlorides       FreeSulfurDioxide TotalSulfurDioxide
 Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00   Min.   :-823.0
 1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00   1st Qu.:  27.0
 Median :   3.900   Median : 0.0460   Median :  30.00   Median : 123.0
 Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85   Mean   : 120.7
 3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00   3rd Qu.: 208.0
 Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00   Max.   :1057.0
 NA's   :616        NA's   :638       NA's   :647       NA's   :682
    Density              pH          Sulphates          Alcohol       LabelAppeal
 Min.   :0.8881   Min.   :0.480   Min.   :-3.1300   Min.   :-4.70   -2: 504
 1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800   1st Qu.: 9.00   -1:3136
 Median :0.9945   Median :3.200   Median : 0.5000   Median :10.40   0 :5617
 Mean   :0.9942   Mean   :3.208   Mean   : 0.5271   Mean   :10.49   1 :3048
 3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600   3rd Qu.:12.40   2 : 490
 Max.   :1.0992   Max.   :6.130   Max.   : 4.2400   Max.   :26.50
                  NA's   :395     NA's   :1210      NA's   :653
   AcidIndex        STARS
 Min.   : 4.000   1   :3042
 1st Qu.: 7.000   2   :3570
 Median : 8.000   3   :2212
 Mean   : 7.773   4   : 612
 3rd Qu.: 8.000   NA's:3359
 Max.   :17.000
```

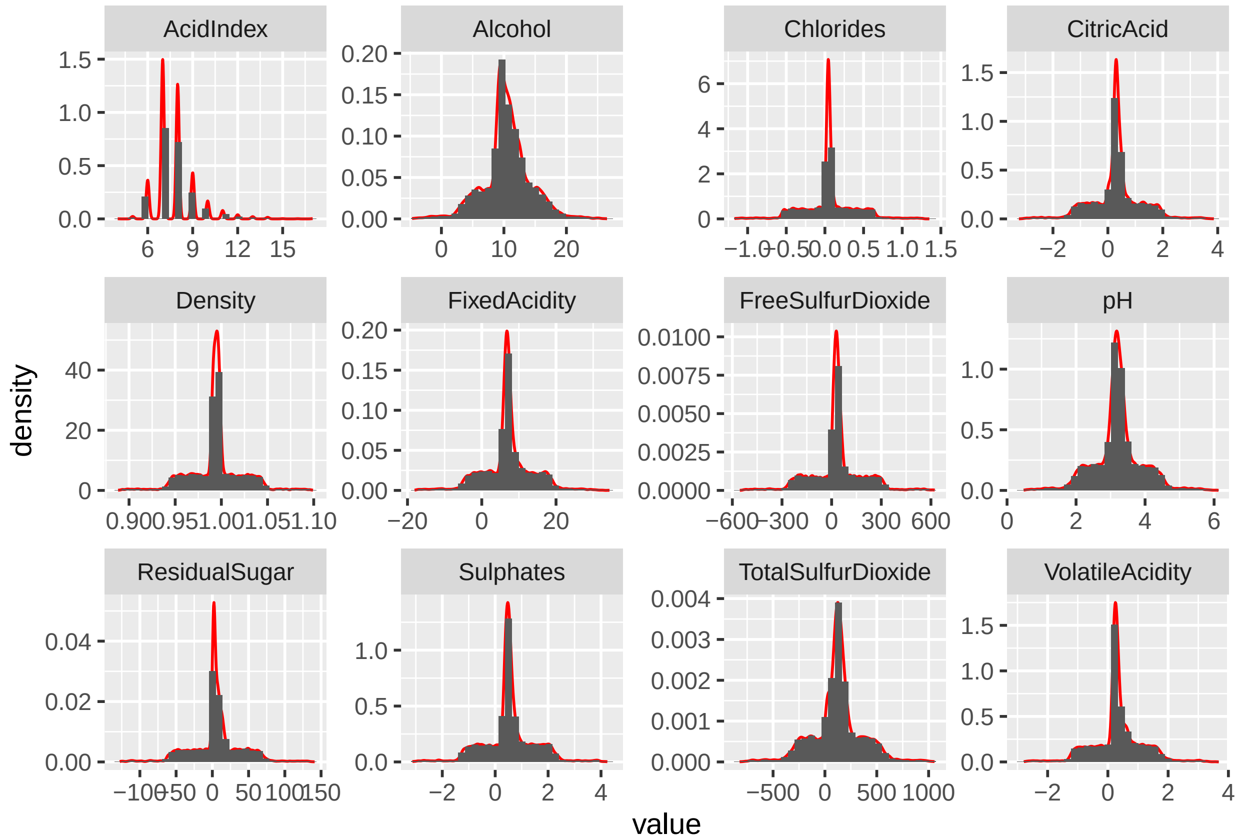

Figure 1: Histograms for all of the variables.

Figure 1 shows us that the histograms for the continuous predictor variables assume somewhat of a normal distirbution. Therefore, the team reasoned that these variables did not require any transformation.

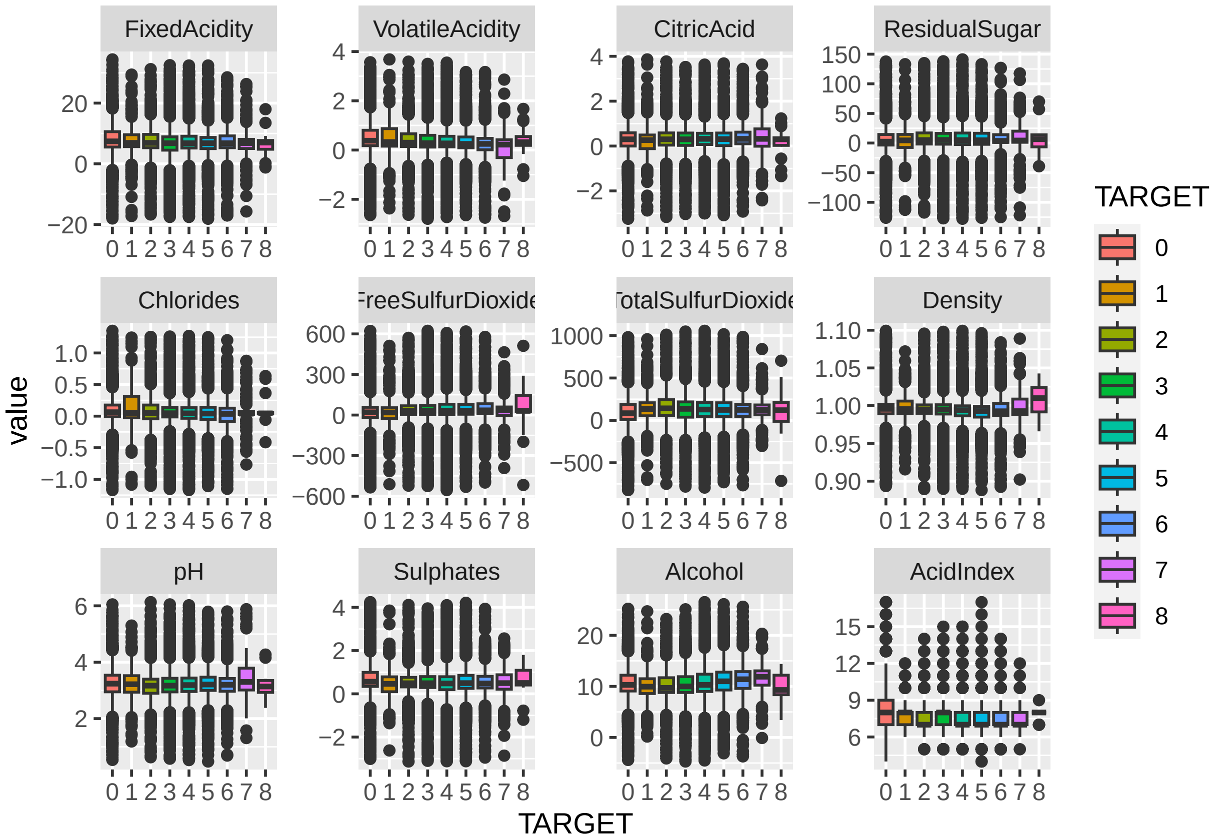

Figure 2: Boxplots for the dataset

Figure 2 points out that there are way less outliers when 8 cases are purchased compared to the under 8 cases. Figure 2 also shows that the number of outliers decreases as the number of cases increases. It would seem that people tend to buy higher amounts of wine with the following characteristics:

- Fixed acidity is 0
- Volatile acidity is 0
- Residual sugar is 0
- Chlorides is 0
- Sulfur dioxide content is 0
- Total sulfur dioxide is 0
- Density is 1
- pH is 3 (The optimal pH for wine is about 3.0 to 3.4 (source))
- Sulphates is 0
- Alcohol content is 9%
- The weighted average of the acidity of the wine is ~8

This indicates that the more higher quality the wine, the more amounts of it that people will purchase. Also, if we look at Figure 3, we can assume that affluent people buy more cases, which is why there is so few purchases of 8 cases of wine. Figure 3 shows us that, many people tend to generally buy a bottle, which is why the count for `0` is significantly high. Ignoring this `0`, we can see that the rest of the graph takes on a normal distribution.

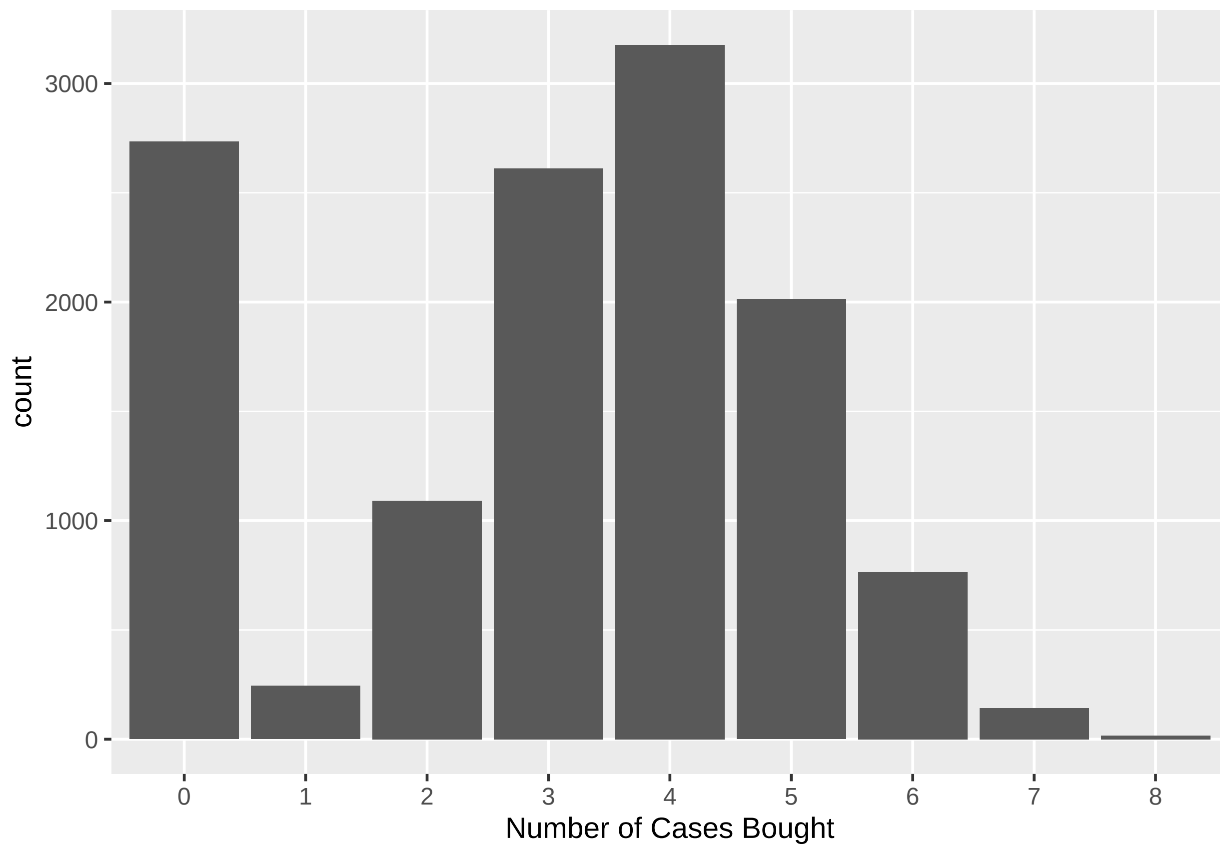

Figure 3: Bar chart of the number of cases bought.

Examining Feature Multicollinearity

Finally, it is imperative to understand which features are correlated with each other in order to address and avoid multicollinearity within our models. By using a correlation plot, we can visualize the relationships between certain features. The correlation plot is only able to determine the correlation for continuous variables. There are methodologies to determine correlations for categorical variables (tetrachoric correlation). However there is only one binary predictor variable which is why the multicollinearity will only be considered for the continuous variables.

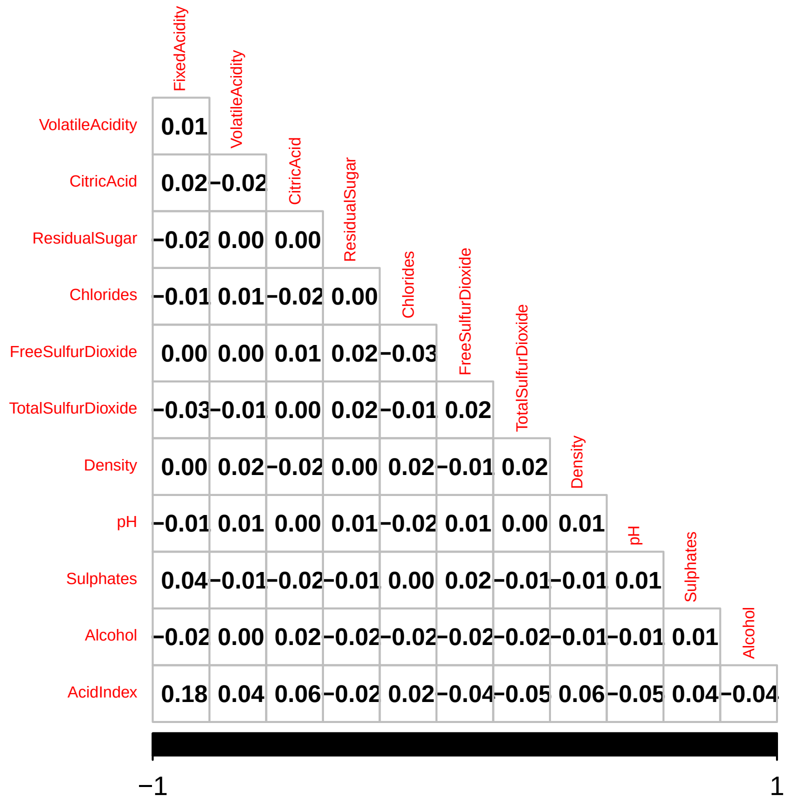

Figure 4: Multicollinearity plot for continuous predictor variables

Figure 4 shows that there isn't much multicollinearity between the continuous variables. In fact the correlations themselves are near 0 for all of the continuous predictor variables. `AcidIndex` has a weak positive correlation with `FixedAcidity` and will therefore be ignored.

Variable	P-Value
STARS	0
AcidIndex	2.82264623433189e-189
LabelAppeal	0

Table 2: Chi-Square test p-values for categorical variables against `TARGET` variable.

We decided to perform Chi-Square tests to determine the correlations between the categorical predictor variables and the `TARGET` variable to see if we can reject the null (they are independent). Table 2 above reveals that all of these variables have a p-value of less than 0.05, which indicates that these variables are correlated with the `TARGET` variable. For `STARS` and `LabelAppeal`, this is to be expected based on the theoretical effects for these variables. We decided to not omit any variables based on these results.

NA exploration

As can be seen in Figure 5, some of the columns have missing values. These missing values were imputed using the MICE algorithm. The methodology that was used is explained in the "Dealing with Missing Values" section.

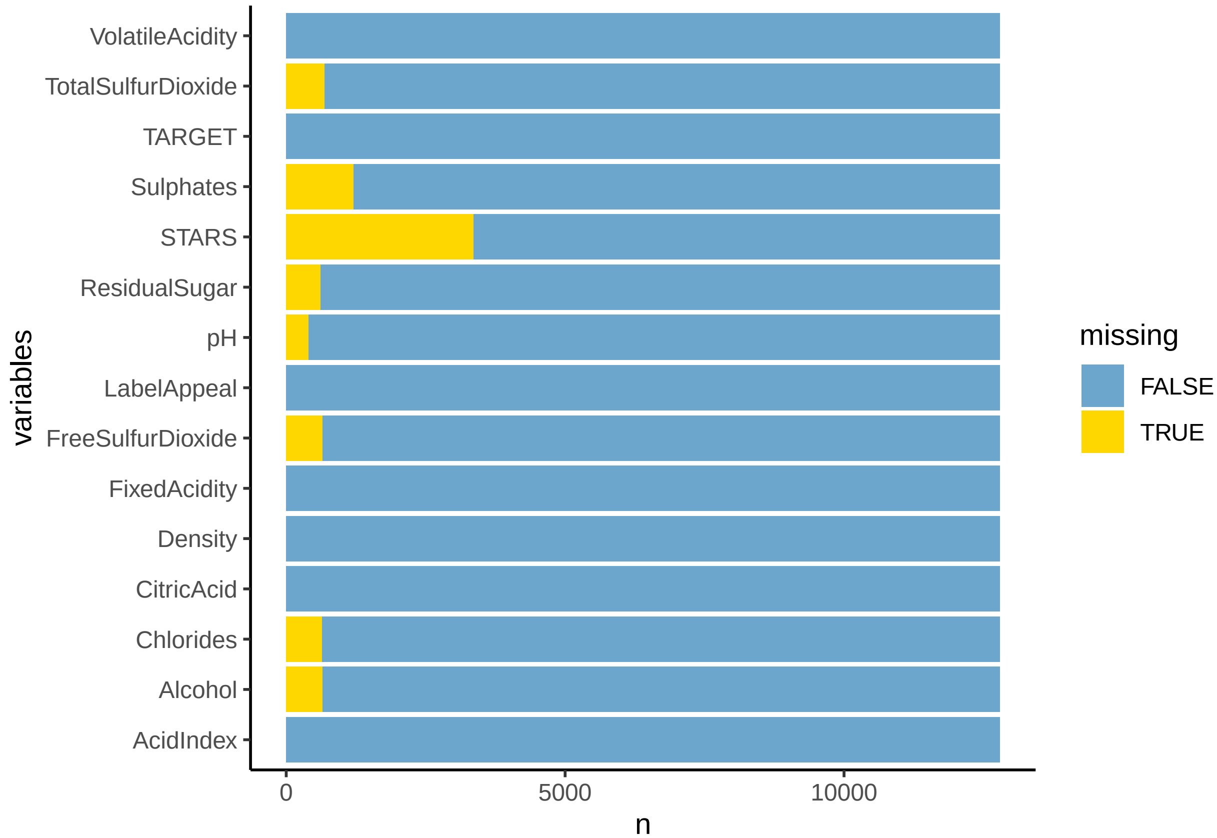

Figure 5: Barplot of number of missing values for each predictor.

Data Preparation

Dealing with Missing Values

In general, imputations by the means/medians is acceptable if the missing values only account for 5% of the sample. Peng et al.(2006) However, should the degree of missing values exceed 20% then using these simple imputation approaches will result in an artificial reduction in variability due to the fact that values are being imputed at the center of the variable's distribution.

Our team decided to employ another technique to handle the missing values: Multiple Regression Imputation using the MICE package.

The MICE package in R implements a methodology where each incomplete variable is imputed by a separate model. Alice points out that plausible values are drawn from a distribution specifically designed for each missing datapoint. Many imputation methods can be used within the package. The one that was selected for the data being analyzed in this report is PMM (Predictive Mean Matching), which is used for quantitative data.

Van Buuren explains that PMM works by selecting values from the observed/already existing data that would most likely belong to the variable in the observation with the missing value. The advantage of this is that it selects values that must exist from the observed data, so no negative values will be used to impute missing data.Not only that, it circumvents the shrinking of errors by using multiple regression models. The variability between the different imputed values gives a wider, but more correct standard error. Uncertainty is inherent in imputation which is why having multiple imputed values is important. Not only that. Marshall et al. 2010 points out that:

"Another simulation study that addressed skewed data concluded that predictive mean matching 'may be the preferred approach provided that less than 50% of the cases have missing data. . . '

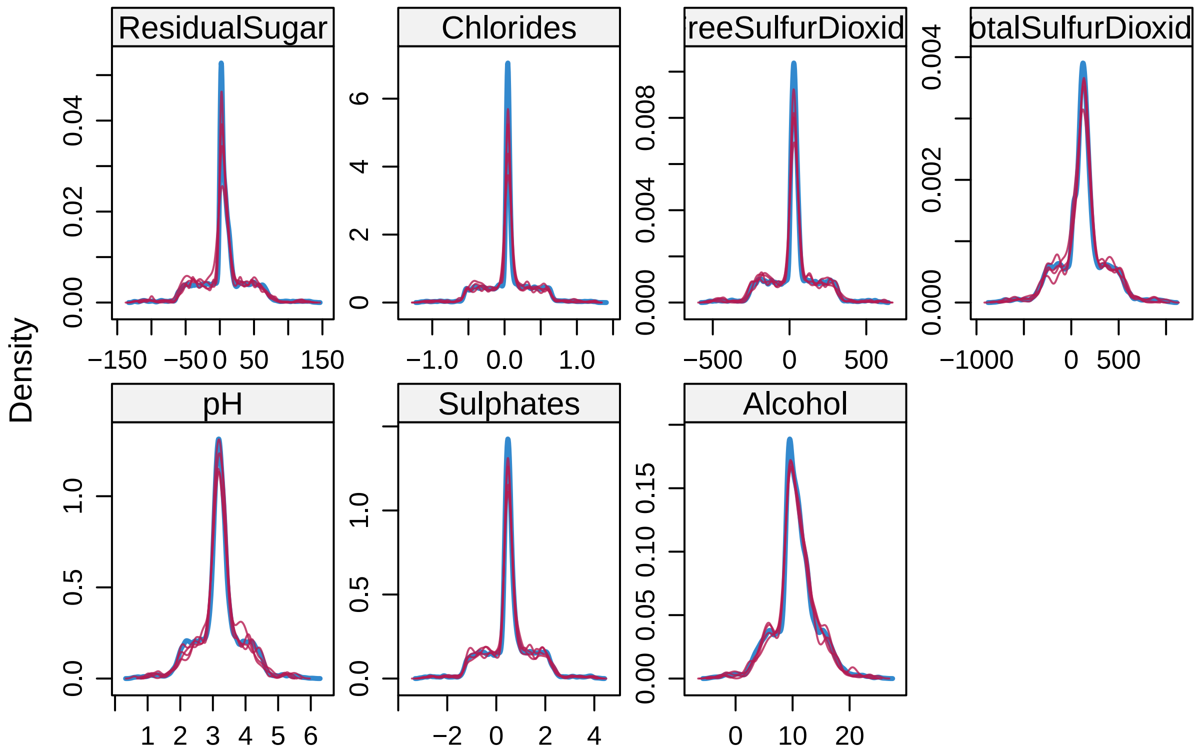

Figure 6: Density plots for variables containing missing data. The number of multiple imputations was set to 4. Each of the red lines represents the distribution for each imputation.

The blue lines for each of the graphs in Figure 6 represent the distributions the non-missing data for each of the variables while the red lines represent the distributions for the imputed data. Note that the distributions for the imputed data for each of the iterations closely matches the distributions for the non-missing data, which is ideal. If the distributions did not match so well, than another imputing method would have had to have been used.

Split Data Into Testing and Training

The data was into testing and training subsets such that 70% of it will be used to train, and 30% to test. The first row shows the split for the testing data while the second row shows the split for the training data. The first two rows are for the original data set, while the last two rows are for the data set with imputed NA values.

```
  0   1   2   3   4   5   6   7   8
143  25 159 441 603 387 145  26   3
```

```
   0    1    2    3    4    5    6    7    8
 332   58  372 1028 1407  903  337   61    6
```

```
  0   1   2   3   4   5   6   7   8
820  73 327 783 953 604 229  43   5
```

```
   0    1    2    3    4    5    6    7    8 
1914  171  764 1828 2224 1410  536   99   12 
```

Build Models

In this section, the coefficients and p-values for each of the models generated are shown. Note that for the stepAIC models, the selection direction was set to `both`. The metrics for each of the models are shown in the "Model Selection" section in this report.

Poisson Regression Models

There were 4 different poisson regression models that were constructed in this analysis using imputed/modified and original data. They are:

- Poisson regression model using original data
- Poisson regression model using modified data
- Poisson regression model with significant features selected using stepAIC using original data.
- Poisson regression model with significant features selected using stepAIC using modified data.

Poission Regression Model with Original Data The p-values for the coefficients for this model are shown below. The `LabelAppeal`, `STARS`, `VolatileAcidity`, `AcidIndex`, and `Intercept` are statistically significant when using a 95% confidence interval. It was shown earlier in the report that `STARS`, `LabelAppeal` and `AcidIndex` were highly correlated with the `TARGET` variable, so these low p-values are to be expected.

```
                     Estimate  Std. Error z value  Pr(>|z|)
(Intercept)         1.4416e+00  2.6779e-01  5.3832 7.317e-08
FixedAcidity        5.2608e-04  1.1162e-03  0.4713  0.637431
VolatileAcidity    -1.9124e-02  8.8193e-03 -2.1684  0.030128
CitricAcid          2.1219e-04  8.1845e-03  0.0259  0.979317
ResidualSugar       1.3206e-05  2.0418e-04  0.0647  0.948433
Chlorides          -3.2633e-02  2.1665e-02 -1.5063  0.131997
FreeSulfurDioxide   4.1218e-05  4.6534e-05  0.8858  0.375747
TotalSulfurDioxide  2.3185e-05  2.9962e-05  0.7738  0.439040
Density            -1.9199e-01  2.5814e-01 -0.7438  0.457027
pH                 -5.2934e-03  1.0212e-02 -0.5183  0.604232
Sulphates          -6.2419e-03  7.4458e-03 -0.8383  0.401859
Alcohol             3.3917e-03  1.8926e-03  1.7921  0.073115
LabelAppeal-1       1.7693e-01  5.2083e-02  3.3971  0.000681
LabelAppeal0        3.4321e-01  5.0834e-02  6.7517 1.462e-11
LabelAppeal1        4.6510e-01  5.1713e-02  8.9938 < 2.2e-16
LabelAppeal2        5.6419e-01  5.8362e-02  9.6671 < 2.2e-16
AcidIndex          -3.7300e-02  6.1878e-03 -6.0280 1.660e-09
STARS2              2.4724e-01  1.7964e-02 13.7634 < 2.2e-16
STARS3              3.3747e-01  1.9880e-02 16.9755 < 2.2e-16
STARS4              4.3899e-01  2.9142e-02 15.0638 < 2.2e-16

n = 4504 p = 20
Deviance = 1638.67537 Null Deviance = 2707.94272 (Difference = 1069.26735)
```

Poisson Regression Model with Modified Data Once again, the same highly correlated variables have low p-values. With that being said, it would appear that the p-values for these variables is lower than the p-values shown in the poisson regression model with original data.

```
                     Estimate  Std. Error  z value  Pr(>|z|)
```

```
(Intercept)         1.4420e+00  2.0372e-01   7.0786 1.456e-12
FixedAcidity        1.7602e-05  8.5235e-04   0.0207 0.9835238
VolatileAcidity    -2.9866e-02  6.8283e-03  -4.3738 1.221e-05
CitricAcid          6.8766e-03  6.0917e-03   1.1288 0.2589671
ResidualSugar      -6.1602e-05  1.5835e-04  -0.3890 0.6972537
Chlorides          -3.2173e-02  1.6488e-02  -1.9513 0.0510235
FreeSulfurDioxide   9.9303e-05  3.5323e-05   2.8113 0.0049346
TotalSulfurDioxide  6.2451e-05  2.2812e-05   2.7376 0.0061886
Density            -1.4439e-01  1.9848e-01  -0.7275 0.4669093
pH                 -7.4129e-03  7.8049e-03  -0.9498 0.3422243
Sulphates          -7.8606e-03  5.6926e-03  -1.3808 0.1673256
Alcohol             2.5947e-03  1.4193e-03   1.8282 0.0675219
LabelAppeal-1       1.3673e-01  3.5501e-02   3.8514 0.0001174
LabelAppeal0        2.6813e-01  3.4557e-02   7.7593 8.543e-15
LabelAppeal1        3.6710e-01  3.5326e-02  10.3919 < 2.2e-16
LabelAppeal2        4.8019e-01  4.1133e-02  11.6739 < 2.2e-16
AcidIndex          -6.7667e-02  4.5626e-03 -14.8307 < 2.2e-16
STARS2              4.5822e-01  1.3261e-02  34.5545 < 2.2e-16
STARS3              6.0932e-01  1.4965e-02  40.7170 < 2.2e-16
STARS4              7.1565e-01  2.2023e-02  32.4956 < 2.2e-16

n = 8958 p = 20
Deviance = 5698.18177 Null Deviance = 9674.18100 (Difference = 3975.99923)
```

Step AIC for Poisson with Original Data With the exception of `Chlorides` and `Alcohol`, the rest of the variables are statistically significant and those same 3 variables (`STARS`, `LabelAppeal` and `AcidIndex`) are present in this model which is to be expected.

```
                  Estimate Std. Error z value  Pr(>|z|)
(Intercept)      1.2374839  0.0727173 17.0177 < 2.2e-16
VolatileAcidity -0.0192110  0.0088121 -2.1801  0.029252
Chlorides       -0.0332443  0.0216266 -1.5372  0.124246
Alcohol          0.0033364  0.0018910  1.7643  0.077681
LabelAppeal-1    0.1772110  0.0520719  3.4032  0.000666
LabelAppeal0     0.3431725  0.0508213  6.7525 1.453e-11
LabelAppeal1     0.4655570  0.0516916  9.0064 < 2.2e-16
LabelAppeal2     0.5640479  0.0583399  9.6683 < 2.2e-16
AcidIndex       -0.0371248  0.0060890 -6.0970 1.081e-09
STARS2           0.2474267  0.0179514 13.7831 < 2.2e-16
STARS3           0.3387887  0.0198449 17.0719 < 2.2e-16
STARS4           0.4390154  0.0291173 15.0775 < 2.2e-16

n = 4504 p = 12
Deviance = 1641.71807 Null Deviance = 2707.94272 (Difference = 1066.22465)
```

Step AIC for Poisson with Modified Data This model indicates that when using the imputed data, the `FreeSulfurDioxide`, `TotalSulfurDioxide`, and `VolatileAcidity` variables are statistically significant. Grogan indicates that "sulfur dioxide preserves wine, preventing oxidation and browning", so the amount of it is important in how many cases are bought (see Figure 2 boxplot for these variables).

```
                      Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)         1.2692e+00  5.1302e-02  24.7403 < 2.2e-16
VolatileAcidity    -2.9978e-02  6.8270e-03  -4.3911 1.128e-05
Chlorides          -3.2486e-02  1.6466e-02  -1.9729 0.0485114
FreeSulfurDioxide   9.8728e-05  3.5308e-05   2.7962 0.0051703
```

```
TotalSulfurDioxide  6.2166e-05  2.2792e-05   2.7276 0.0063800
Alcohol             2.6342e-03  1.4183e-03   1.8573 0.0632701
LabelAppeal-1       1.3664e-01  3.5497e-02   3.8494 0.0001184
LabelAppeal0        2.6803e-01  3.4551e-02   7.7576 8.655e-15
LabelAppeal1        3.6691e-01  3.5321e-02  10.3880 < 2.2e-16
LabelAppeal2        4.7937e-01  4.1120e-02  11.6578 < 2.2e-16
AcidIndex          -6.7322e-02  4.4877e-03 -15.0014 < 2.2e-16
STARS2              4.5901e-01  1.3250e-02  34.6417 < 2.2e-16
STARS3              6.1049e-01  1.4947e-02  40.8434 < 2.2e-16
STARS4              7.1645e-01  2.2012e-02  32.5481 < 2.2e-16

n = 8958 p = 14
Deviance = 5702.98487 Null Deviance = 9674.18100 (Difference = 3971.19613)
```

Negative Binomial Models

There were 4 different negative binomial models that were constructed in this analysis using imputed/modified and original data. They are:

- Negative binomial model using original data
- Negative binomial model using modified data
- Negative binomial model with significant features selected using stepAIC using original data.
- Negative binomial model with significant features selected using stepAIC using modified data.

Negative Binomial Model with Original Data The p-values for the coefficients for this model are shown below. The `LabelAppeal`, `STARS`, `VolatileAcidity`, `AcidIndex`, and `Intercept` are statistically significant when using a 95% confidence interval. It was shown earlier in the report that `STARS`, `LabelAppeal` and `AcidIndex` were highly correlated with the `TARGET` variable, so these low p-values are to be expected. In fact, the selected variables and the p-values for this model and the poisson regression model with original data are more or less the same.

```
Call:
glm.nb(formula = TARGET ~ ., data = original_train %>% dplyr::mutate(TARGET = as.numeric(TARGET)),
    init.theta = 241045.1812, link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.59914  -0.24871   0.04379   0.34233   1.51828

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.442e+00  2.678e-01   5.383 7.32e-08 ***
FixedAcidity        5.261e-04  1.116e-03   0.471 0.637433
VolatileAcidity    -1.912e-02  8.819e-03  -2.168 0.030129 *
CitricAcid          2.122e-04  8.185e-03   0.026 0.979319
ResidualSugar       1.321e-05  2.042e-04   0.065 0.948431
Chlorides          -3.263e-02  2.166e-02  -1.506 0.132000
FreeSulfurDioxide   4.122e-05  4.654e-05   0.886 0.375751
TotalSulfurDioxide  2.319e-05  2.996e-05   0.774 0.439043
Density            -1.920e-01  2.581e-01  -0.744 0.457031
pH                 -5.293e-03  1.021e-02  -0.518 0.604233
Sulphates          -6.242e-03  7.446e-03  -0.838 0.401862
Alcohol             3.392e-03  1.893e-03   1.792 0.073119 .
LabelAppeal-1       1.769e-01  5.208e-02   3.397 0.000681 ***
```

```
LabelAppeal0        3.432e-01  5.083e-02   6.752 1.46e-11 ***
LabelAppeal1        4.651e-01  5.171e-02   8.994  < 2e-16 ***
LabelAppeal2        5.642e-01  5.836e-02   9.667  < 2e-16 ***
AcidIndex          -3.730e-02  6.188e-03  -6.028 1.66e-09 ***
STARS2              2.472e-01  1.796e-02  13.763  < 2e-16 ***
STARS3              3.375e-01  1.988e-02  16.975  < 2e-16 ***
STARS4              4.390e-01  2.914e-02  15.064  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(241045.2) family taken to be 1)

    Null deviance: 2707.9  on 4503  degrees of freedom
Residual deviance: 1638.7  on 4484  degrees of freedom
AIC: 16714

Number of Fisher Scoring iterations: 1

              Theta:  241045
          Std. Err.:  522593
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -16672.03
```

Negative Binomial Model with Modified Data Once again, the same highly correlated variables have low p-values along with the `FreeSulfurDioxide` and `TotalSulfurDioxide`, and almost but not quite, `Chlorides`, which were not statistically significant when the original data was used. With that being said, it would appear that the p-values for these variables is lower than the p-values shown in the negative binomial model with original data. In fact, the selected variables and the p-values for this model and the poisson regression model with modified data are more or less the same.

```
Call:
glm.nb(formula = TARGET ~ ., data = modified_train %>% dplyr::mutate(TARGET = as.numeric(TARGET)),
    init.theta = 103966.2608, link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.61209  -0.54335   0.04884   0.47504   2.40421

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.442e+00  2.037e-01   7.079 1.46e-12 ***
FixedAcidity        1.760e-05  8.524e-04   0.021 0.983524
VolatileAcidity    -2.987e-02  6.828e-03  -4.374 1.22e-05 ***
CitricAcid          6.877e-03  6.092e-03   1.129 0.258976
ResidualSugar      -6.160e-05  1.584e-04  -0.389 0.697261
Chlorides          -3.217e-02  1.649e-02  -1.951 0.051025 .
FreeSulfurDioxide   9.930e-05  3.532e-05   2.811 0.004935 **
TotalSulfurDioxide  6.245e-05  2.281e-05   2.738 0.006189 **
Density            -1.444e-01  1.985e-01  -0.728 0.466919
pH                 -7.413e-03  7.805e-03  -0.950 0.342220
Sulphates          -7.861e-03  5.693e-03  -1.381 0.167321
Alcohol             2.595e-03  1.419e-03   1.828 0.067532 .
```

```
LabelAppeal-1       1.367e-01  3.550e-02   3.851 0.000117 ***
LabelAppeal0        2.681e-01  3.456e-02   7.759 8.55e-15 ***
LabelAppeal1        3.671e-01  3.533e-02  10.392  < 2e-16 ***
LabelAppeal2        4.802e-01  4.113e-02  11.674  < 2e-16 ***
AcidIndex          -6.767e-02  4.563e-03 -14.831  < 2e-16 ***
STARS2              4.582e-01  1.326e-02  34.554  < 2e-16 ***
STARS3              6.093e-01  1.497e-02  40.716  < 2e-16 ***
STARS4              7.156e-01  2.202e-02  32.495  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(103966.3) family taken to be 1)

    Null deviance: 9673.9  on 8957  degrees of freedom
Residual deviance: 5698.0  on 8938  degrees of freedom
AIC: 33648

Number of Fisher Scoring iterations: 1

              Theta:  103966
          Std. Err.:  133381
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -33606.04
```

Step AIC for Negative Binomial Model with Original Data With the exception of `Chlorides` and `Alcohol`, the rest of the variables are statistically significant and those 3 variables that were tested against `TARGET` using the Chi-square test (`STARS`, `LabelAppeal` and `AcidIndex`) are present in this model which is to be expected. In fact, the selected variables and the p-values for this model and the Step AIC for poisson regression model with original data are more or less the same.

```
Call:
glm.nb(formula = TARGET ~ VolatileAcidity + Chlorides + Alcohol +
    LabelAppeal + AcidIndex + STARS, data = original_train %>%
    dplyr::mutate(TARGET = as.numeric(TARGET)), init.theta = 240803.8125,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6060  -0.2454   0.0456   0.3438   1.5118

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.237485   0.072718  17.018  < 2e-16 ***
VolatileAcidity -0.019211   0.008812  -2.180 0.029254 *
Chlorides       -0.033244   0.021627  -1.537 0.124249
Alcohol          0.003336   0.001891   1.764 0.077685 .
LabelAppeal-1    0.177211   0.052072   3.403 0.000666 ***
LabelAppeal0     0.343172   0.050822   6.752 1.45e-11 ***
LabelAppeal1     0.465557   0.051692   9.006  < 2e-16 ***
LabelAppeal2     0.564048   0.058341   9.668  < 2e-16 ***
AcidIndex       -0.037125   0.006089  -6.097 1.08e-09 ***
STARS2           0.247427   0.017952  13.783  < 2e-16 ***
```

```
STARS3           0.338789   0.019845  17.072  < 2e-16 ***
STARS4           0.439015   0.029118  15.077  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(240803.8) family taken to be 1)

    Null deviance: 2707.9  on 4503  degrees of freedom
Residual deviance: 1641.7  on 4492  degrees of freedom
AIC: 16701

Number of Fisher Scoring iterations: 1

              Theta:  240804
          Std. Err.:  521942
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -16675.07
```

Step AIC for Negative Binomial Model with Modified Data Once again, the selected variables and the p-values for this model and the Step AIC for poisson regression model with modified data are more or less the same.

```
Call:
glm.nb(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
    TotalSulfurDioxide + Alcohol + LabelAppeal + AcidIndex +
    STARS, data = modified_train %>% dplyr::mutate(TARGET = as.numeric(TARGET)),
    init.theta = 103835.9693, link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.59681  -0.54206   0.05122   0.47267   2.42628

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.269e+00  5.130e-02  24.740  < 2e-16 ***
VolatileAcidity    -2.998e-02  6.827e-03  -4.391 1.13e-05 ***
Chlorides          -3.249e-02  1.647e-02  -1.973 0.048513 *
FreeSulfurDioxide   9.873e-05  3.531e-05   2.796 0.005171 **
TotalSulfurDioxide  6.217e-05  2.279e-05   2.728 0.006380 **
Alcohol             2.634e-03  1.418e-03   1.857 0.063280 .
LabelAppeal-1       1.366e-01  3.550e-02   3.849 0.000118 ***
LabelAppeal0        2.680e-01  3.455e-02   7.757 8.66e-15 ***
LabelAppeal1        3.669e-01  3.532e-02  10.388  < 2e-16 ***
LabelAppeal2        4.794e-01  4.112e-02  11.658  < 2e-16 ***
AcidIndex          -6.732e-02  4.488e-03 -15.001  < 2e-16 ***
STARS2              4.590e-01  1.325e-02  34.641  < 2e-16 ***
STARS3              6.105e-01  1.495e-02  40.843  < 2e-16 ***
STARS4              7.164e-01  2.201e-02  32.547  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(103836) family taken to be 1)
```

```
    Null deviance: 9673.9  on 8957  degrees of freedom
Residual deviance: 5702.8  on 8944  degrees of freedom
AIC: 33641

Number of Fisher Scoring iterations: 1

              Theta:  103836
          Std. Err.:  133139
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -33610.84
```

Multiple Linear Regression Models

There were 4 different multiple linear regression models that were constructed in this analysis using imputed/modified and original data. They are:

- Multiple linear regression model using original data
- Multiple linear regression model using modified data
- Multiple linear regression model with significant features selected using stepAIC using original data.
- Multiple linear regression model with significant features selected using stepAIC using modified data.

Multiple Linear Regression Model with Original Data The p-values for the coefficients for this model are shown below. The `LabelAppeal`, `STARS`, `VolatileAcidity`, `Chlorides`, `Alcohol`, `AcidIndex`, and `Intercept` are statistically significant when using a 95% confidence interval. It was shown earlier in the report that `STARS`, `LabelAppeal` and `AcidIndex` were highly correlated with the `TARGET` variable, so these low p-values are to be expected.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.6537e+00	6.5604e-01	7.0935	1.512e-12
FixedAcidity	2.7365e-03	2.7456e-03	0.9967	0.3189695
VolatileAcidity	-9.0222e-02	2.1728e-02	-4.1523	3.353e-05
CitricAcid	1.8870e-03	2.0249e-02	0.0932	0.9257579
ResidualSugar	-1.6880e-05	5.0335e-04	-0.0335	0.9732499
Chlorides	-1.5315e-01	5.3509e-02	-2.8622	0.0042271
FreeSulfurDioxide	1.9276e-04	1.1449e-04	1.6836	0.0923289
TotalSulfurDioxide	1.1301e-04	7.3585e-05	1.5358	0.1246541
Density	-8.9159e-01	6.3746e-01	-1.3986	0.1619872
pH	-2.1876e-02	2.5220e-02	-0.8674	0.3857608
Sulphates	-2.4703e-02	1.8332e-02	-1.3475	0.1778813
Alcohol	1.6244e-02	4.6425e-03	3.4991	0.0004715
LabelAppeal-1	5.1536e-01	1.0241e-01	5.0324	5.032e-07
LabelAppeal0	1.1871e+00	1.0004e-01	11.8662	< 2.2e-16
LabelAppeal1	1.8164e+00	1.0357e-01	17.5368	< 2.2e-16
LabelAppeal2	2.4244e+00	1.2977e-01	18.6829	< 2.2e-16
AcidIndex	-1.6520e-01	1.4597e-02	-11.3173	< 2.2e-16
STARS2	1.0171e+00	4.1367e-02	24.5866	< 2.2e-16
STARS3	1.5039e+00	4.8201e-02	31.2012	< 2.2e-16
STARS4	2.1549e+00	7.9087e-02	27.2468	< 2.2e-16

```
n = 4504, p = 20, Residual SE = 1.14099, R-Squared = 0.46
```

Multiple Linear Regression Model with Modified Data Once again, the same highly correlated variables have low p-values along with the `FreeSulfurDioxide` and `TotalSulfurDioxide` variables, which were not statistically significant when the original data was used. With that being said, it would appear that the p-value for `VolatileAcidity` has decreased further while the p-value for `Alcohol` has increased but is still statistically significant.

```
                     Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         4.4322e+00  5.6661e-01   7.8223 5.777e-15
FixedAcidity        6.2514e-04  2.3869e-03   0.2619  0.793403
VolatileAcidity    -1.2096e-01  1.9048e-02  -6.3501 2.256e-10
CitricAcid          2.5663e-02  1.7058e-02   1.5044  0.132503
ResidualSugar      -2.3788e-04  4.4219e-04  -0.5380  0.590617
Chlorides          -1.3414e-01  4.6288e-02  -2.8980  0.003765
FreeSulfurDioxide   4.0576e-04  9.9291e-05   4.0865 4.417e-05
TotalSulfurDioxide  2.4930e-04  6.3518e-05   3.9249 8.742e-05
Density            -5.4382e-01  5.5567e-01  -0.9787  0.327764
pH                 -2.2353e-02  2.1837e-02  -1.0236  0.306036
Sulphates          -2.9184e-02  1.5895e-02  -1.8360  0.066389
Alcohol             1.1460e-02  3.9620e-03   2.8926  0.003830
LabelAppeal-1       3.4631e-01  8.0536e-02   4.3001 1.725e-05
LabelAppeal0        8.0456e-01  7.8444e-02  10.2564 < 2.2e-16
LabelAppeal1        1.2578e+00  8.1916e-02  15.3548 < 2.2e-16
LabelAppeal2        1.8727e+00  1.0825e-01  17.3008 < 2.2e-16
AcidIndex          -2.4019e-01  1.1595e-02 -20.7160 < 2.2e-16
STARS2              1.6065e+00  3.4564e-02  46.4797 < 2.2e-16
STARS3              2.4045e+00  4.2611e-02  56.4305 < 2.2e-16
STARS4              3.1019e+00  7.1846e-02  43.1742 < 2.2e-16

n = 8958, p = 20, Residual SE = 1.39557, R-Squared = 0.48
```

Step AIC for Multiple Linear Regression Model with Original Data With the exception of `FreeSulfurDioxide` and `TotalSulfurDioxide`, the rest of the variables are statistically significant and those 3 variables that were tested against `TARGET` using the Chi-square test (`STARS`, `LabelAppeal` and `AcidIndex`) are present in this model which is to be expected. Basically all of the statistically significant from the multiple linear regression model with original data are used here.

```
                     Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         3.6932e+00  1.5945e-01  23.1624 < 2.2e-16
VolatileAcidity    -9.0231e-02  2.1716e-02  -4.1550 3.314e-05
Chlorides          -1.5425e-01  5.3464e-02  -2.8851 0.0039313
FreeSulfurDioxide   1.9168e-04  1.1440e-04   1.6754 0.0939159
TotalSulfurDioxide  1.0845e-04  7.3521e-05   1.4751 0.1402580
Alcohol             1.6233e-02  4.6384e-03   3.4998 0.0004702
LabelAppeal-1       5.1333e-01  1.0239e-01   5.0136 5.549e-07
LabelAppeal0        1.1860e+00  1.0002e-01  11.8572 < 2.2e-16
LabelAppeal1        1.8163e+00  1.0355e-01  17.5402 < 2.2e-16
LabelAppeal2        2.4205e+00  1.2974e-01  18.6570 < 2.2e-16
AcidIndex          -1.6365e-01  1.4346e-02 -11.4077 < 2.2e-16
STARS2              1.0173e+00  4.1333e-02  24.6132 < 2.2e-16
STARS3              1.5076e+00  4.8148e-02  31.3121 < 2.2e-16
STARS4              2.1563e+00  7.9059e-02  27.2746 < 2.2e-16

n = 4504, p = 14, Residual SE = 1.14091, R-Squared = 0.46
```

Step AIC for Multiple Linear Regression Model with Modified Data With the exception of `CitricAcid` and `Sulphates`, the rest of the variables are statistically significant and those 3 variables that were tested against `TARGET` using the Chi-square test (`STARS`, `LabelAppeal` and `AcidIndex`) are present in this model which is to be expected. Basically all of the statistically significant from the multiple linear regression model with modified data are used here.

```
                      Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)          3.8176e+00  1.2556e-01  30.4034 < 2.2e-16
VolatileAcidity     -1.2140e-01  1.9043e-02  -6.3750 1.921e-10
CitricAcid           2.6097e-02  1.7053e-02   1.5304  0.125957
Chlorides           -1.3460e-01  4.6244e-02  -2.9107  0.003616
FreeSulfurDioxide    4.0343e-04  9.9244e-05   4.0650 4.843e-05
TotalSulfurDioxide   2.4722e-04  6.3475e-05   3.8947 9.905e-05
Sulphates           -2.9014e-02  1.5887e-02  -1.8262  0.067849
Alcohol              1.1502e-02  3.9601e-03   2.9046  0.003686
LabelAppeal-1        3.4539e-01  8.0524e-02   4.2893 1.811e-05
LabelAppeal0         8.0378e-01  7.8432e-02  10.2481 < 2.2e-16
LabelAppeal1         1.2570e+00  8.1894e-02  15.3492 < 2.2e-16
LabelAppeal2         1.8695e+00  1.0821e-01  17.2765 < 2.2e-16
AcidIndex           -2.3953e-01  1.1401e-02 -21.0100 < 2.2e-16
STARS2               1.6077e+00  3.4535e-02  46.5525 < 2.2e-16
STARS3               2.4072e+00  4.2565e-02  56.5525 < 2.2e-16
STARS4               3.1027e+00  7.1831e-02  43.1945 < 2.2e-16

n = 8958, p = 16, Residual SE = 1.39544, R-Squared = 0.48
```

Model Selection

Binary Logistic Regression Models

Model	AIC	MSE
Pois. w/ Original Data	16711.97	1.35
Pois. w/ Modified Data	33645.86	2.03
Step-AIC Pois. w/ Original Data	16699.01	1.35
Step-AIC Pois. w/ Modified Data	33638.66	2.03
Neg. Binom. w/ Original Data	16714.03	1.35
Neg. Binom. w/ Modified Data	33648.04	2.03
Step-AIC Neg. Binom. w/ Original Data	16701.07	1.35
Step-AIC Neg. Binom. w/ Modified Data	33640.84	2.03

Table 3: Model metrics for binary logistic regression models

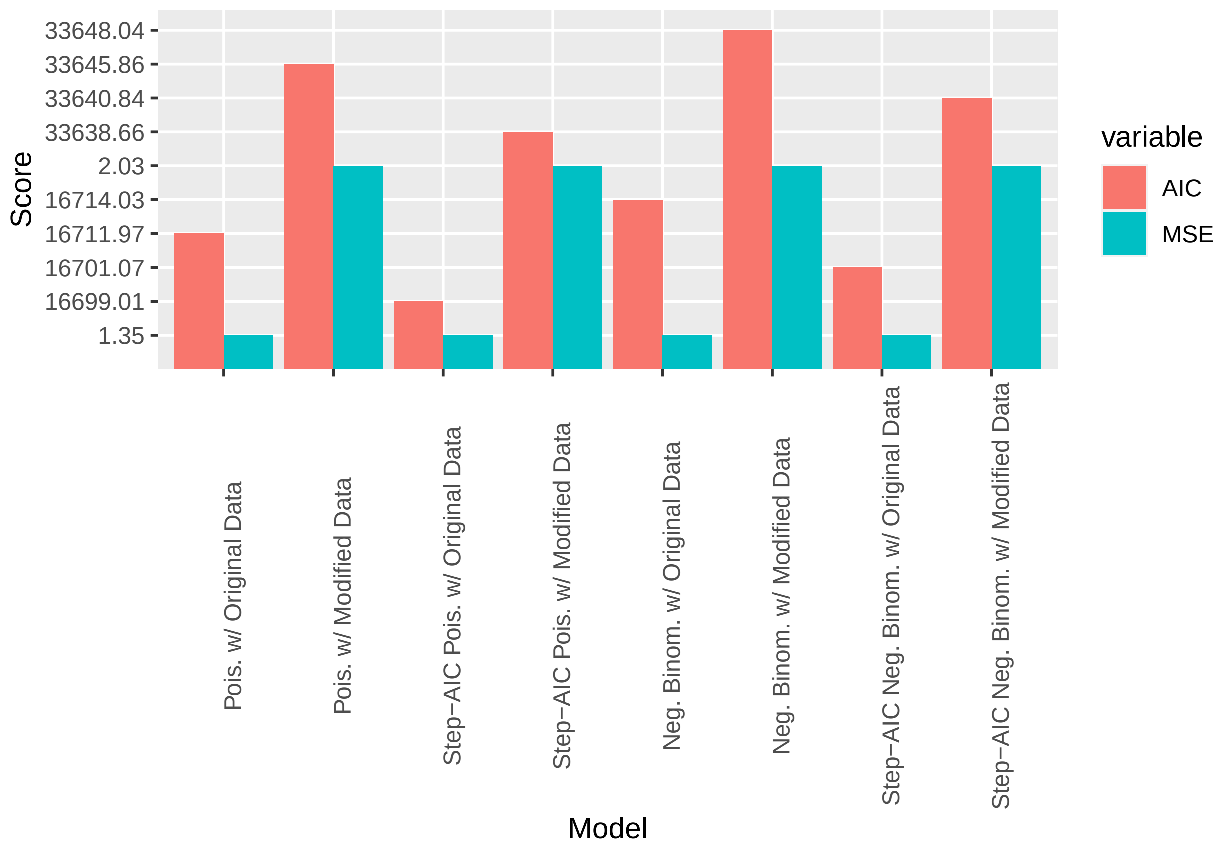

Figure 7: Bar chart of metrics for binary logistic regression models

Figure 7 shows us that the Step-AIC poisson model with original data performs best out of all of the models. Even though the MSE is the same for all of the count regression models when using the original data, the AIC varies between each of them, and the Step-AIC poisson model with original data has the lowest AIC.

Multiple Linear Regression Models

Model	MSE	R-Squared	Adjusted R-Squared	F-Statistic
Multiple Linear w/ Original Data	1.35	0.457	0.455	198.73
Multiple Linear w/ Modified Data	2.04	0.476	0.475	427.8
Step-AIC Multiple Linear w/ Original Data	1.35	0.456	0.455	290.08
Step-AIC Multiple Linear w/ Modified Data	2.04	0.476	0.475	541.82

Table 4: Model metrics for multiple linear regression models

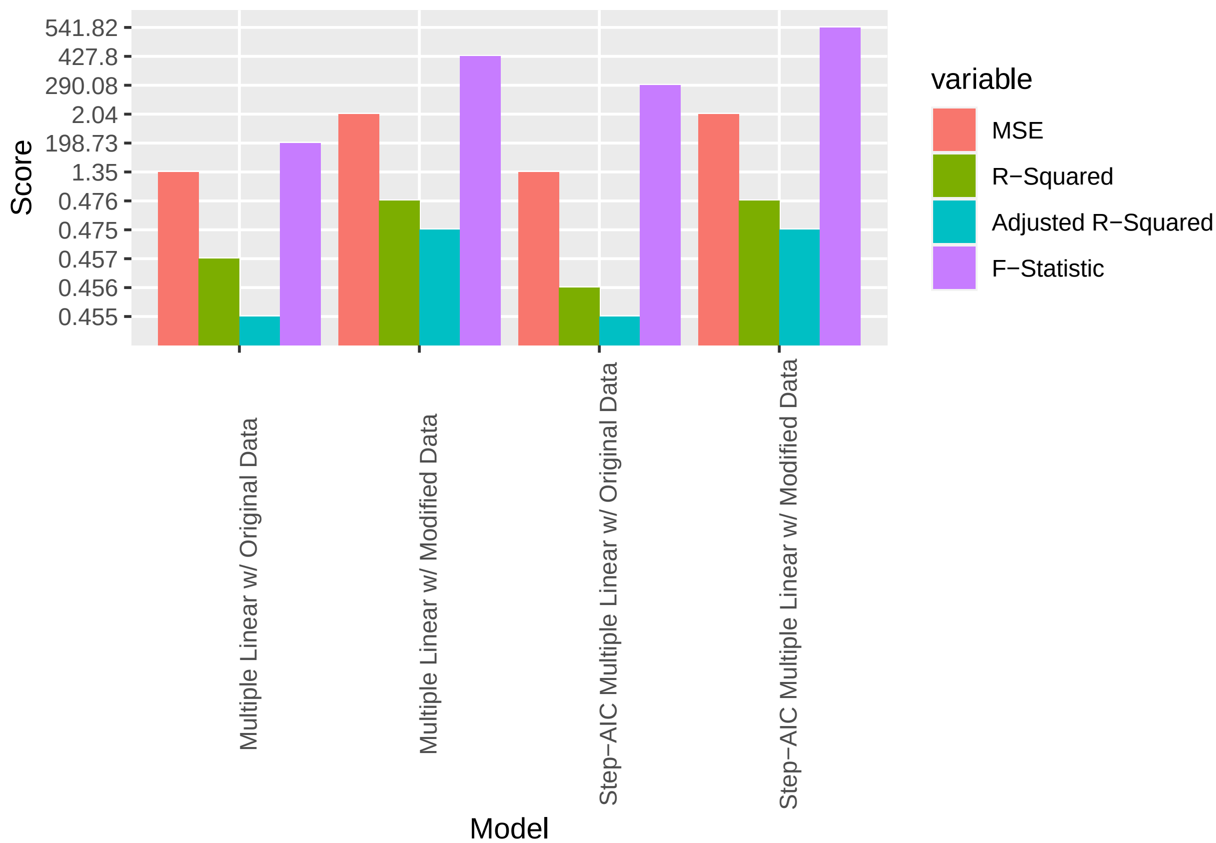

Figure 8: Metrics bar chart for multiple linear regression models

Among the linear regression models, the Step-AIC multiple linear regression model with modified data performs the best. When compared to the multiple linear and step-AIC models using original data, the R-squared and adjusted R-squareds are higher. Also the Step-AIC multiple linear regression model with modified data has a slightly higher F-statistic score than the multiple linear regression model with modified data, making this model the best model since 3 out of the 4 metrics for this model beat out the rest of the models. Since the distribution for the imputed data is roughly the same as the distribution as the original data, we can conclude that the Step-AIC multiple linear regression model with modified data will perform well when presented with new data.

Based on the results shown in Figure 7 and Figure 8 and the model summaries in the “Build Models” section, the Step AIC poisson regression model with original data is the best model out of all of these models. It is more parsimonious than the Step-AIC multiple linear regression model with modified data, making it the best overall model. With this model, we are able to generate predictions for an approximate number of wine cases that could be ordered based on the wine characteristics (predictor variables) shown in the “Step AIC for Poisson with Original Data” section.