DATA 621 - HW #5

Angel Gallardo, Shamecca Marshall

2024-12-10

Contents

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

Variable Name	Definition	Theoretical Effect
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	
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
```

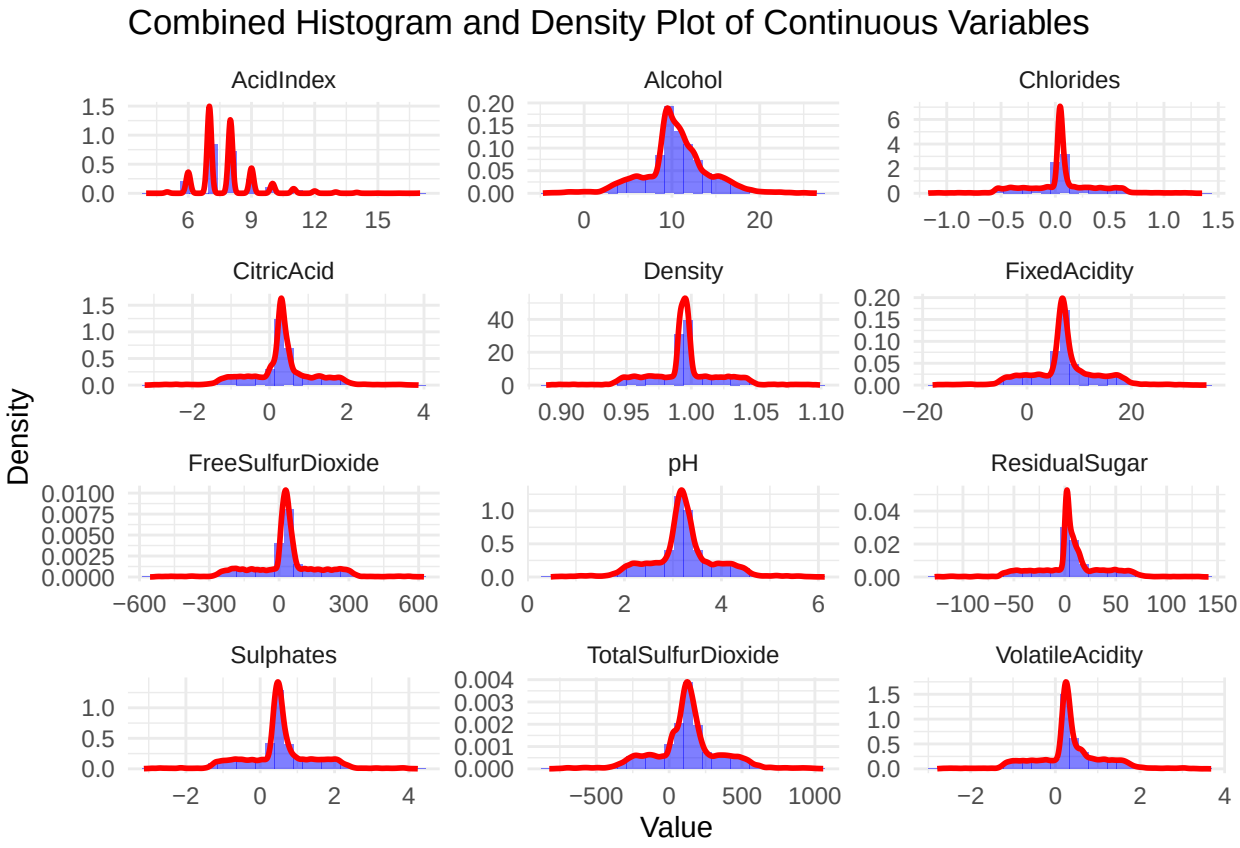

Figure 1: Histograms with Overlaid Density Plots for All Continuous Variables

Figure 1 presents the histograms and overlaid density plots for all continuous predictor variables in the dataset. While some variables, such as `Alcohol` and `Density`, exhibit relatively normal distributions, others, like `FreeSulfurDioxide`, `ResidualSugar`, and `TotalSulfurDioxide`, have extreme outliers and skewed distributions. The variability in distributions suggests that certain variables might benefit from transformations or adjustments to improve model performance. However, the overall spread of the data provides a good basis for analysis without immediate transformation in some cases.

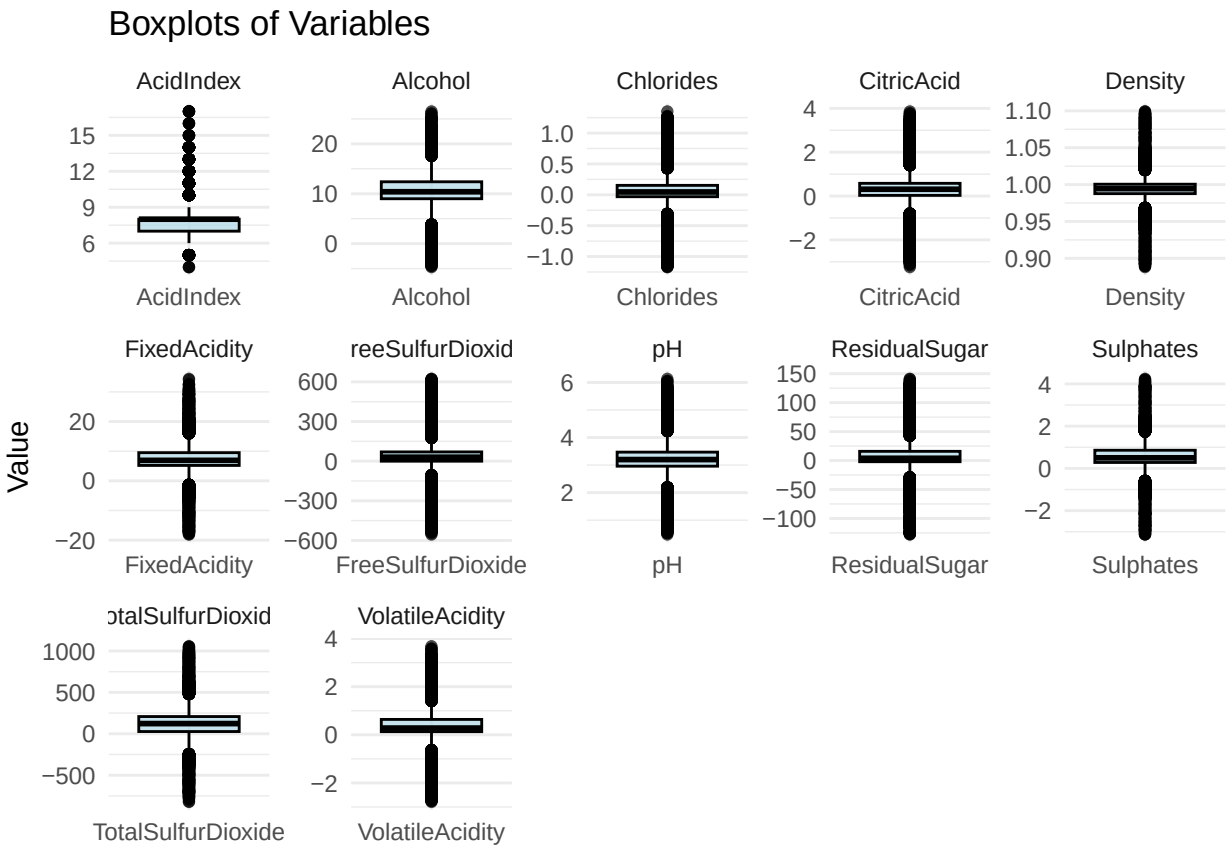

Figure 2: Boxplots for Continuous Variables

Figure 2 displays the boxplots for all continuous predictor variables, highlighting their spread, medians, and potential outliers. The boxplots reveal that certain variables, such as `FreeSulfurDioxide`, `ResidualSugar`, and `TotalSulfurDioxide`, exhibit a large number of extreme values (outliers), which suggests significant variability in the dataset. On the other hand, variables like `Density` and `pH` demonstrate much tighter distributions with fewer outliers.

Key observations include: - `Alcohol`, `FixedAcidity`, and `Sulphates` have a relatively uniform spread, with fewer extreme deviations compared to other variables. - Variables such as `Chlorides` and `CitricAcid` show distributions concentrated around the median, but the presence of outliers indicates some inconsistencies in data values. - The wide range in variables like `ResidualSugar` and `TotalSulfurDioxide` suggests potential skewness or extreme cases that might influence model performance.

These findings indicate that careful preprocessing, such as scaling or transforming specific variables, may be necessary to handle the observed outliers effectively in downstream analyses. However, the overall distribution of the variables offers a diverse dataset for building predictive models.

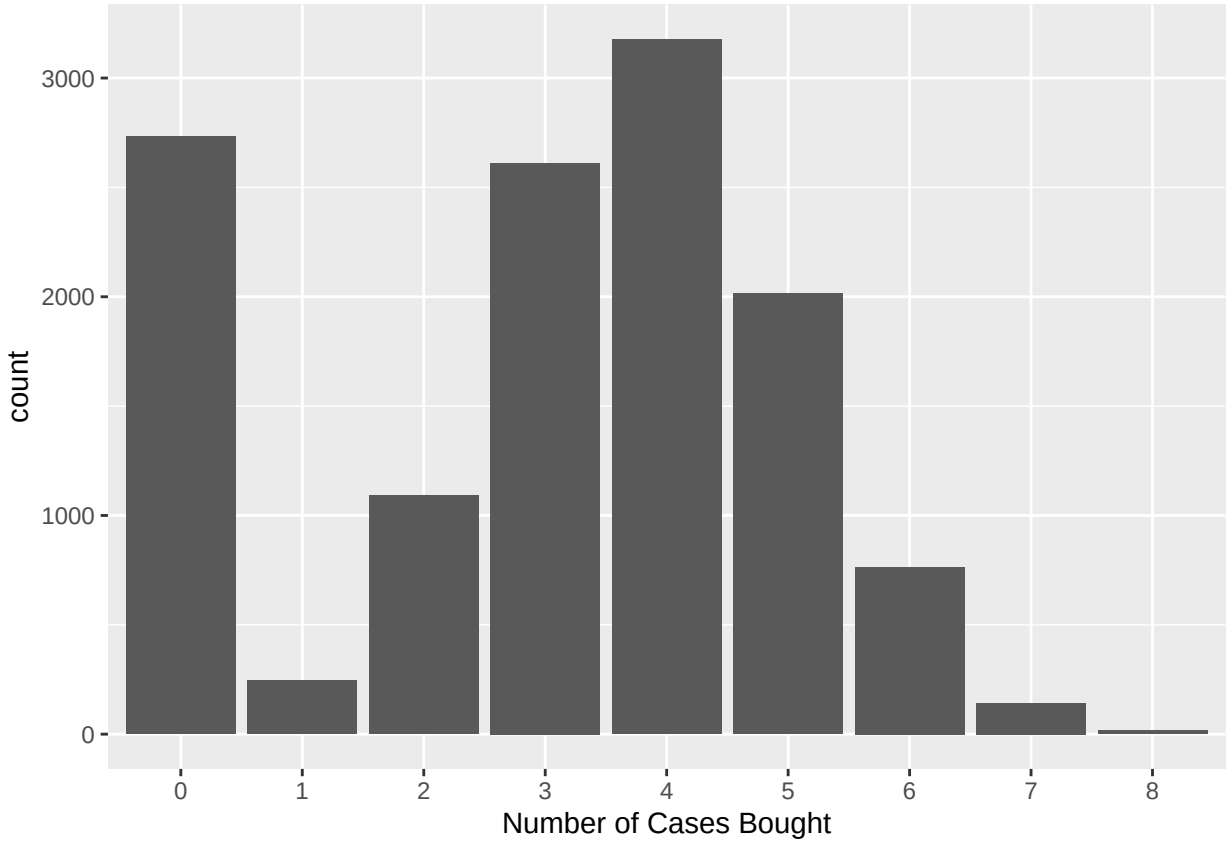

Figure 3: Bar chart of the number of cases bought.

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.

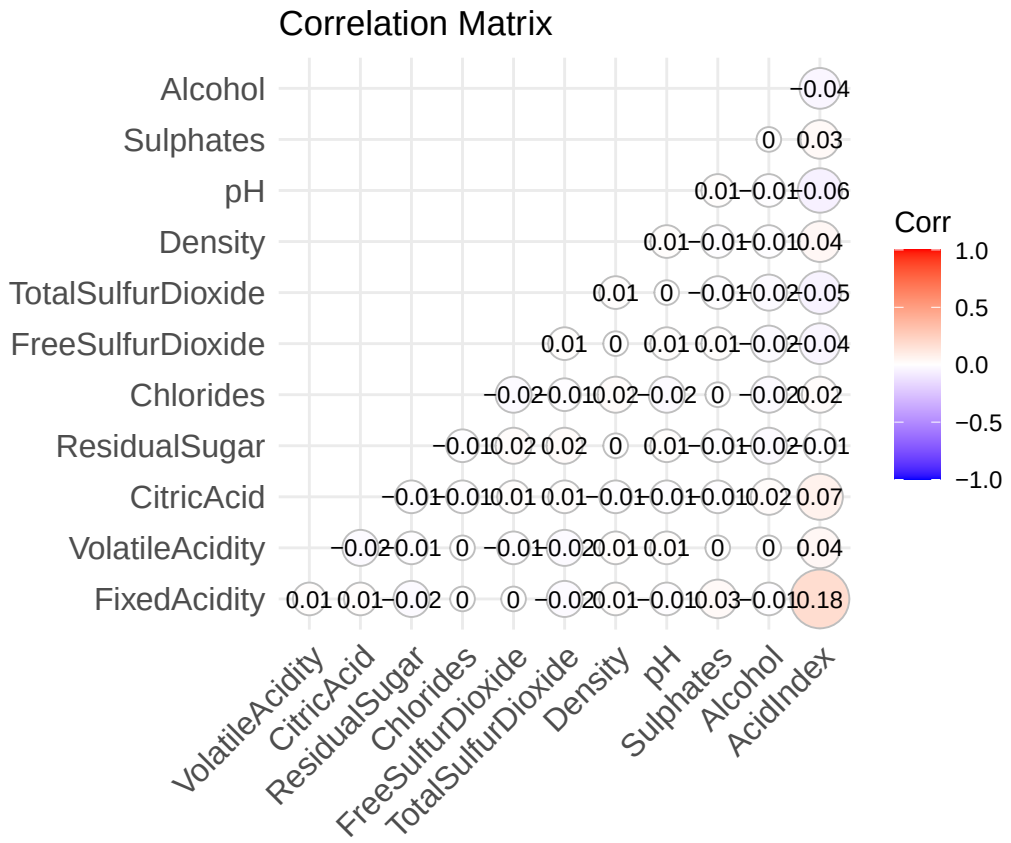

Figure 4: Correlation Matrix for Continuous Predictor Variables

Figure 4 visualizes the correlations between all continuous predictor variables using a correlation matrix. Most of the correlations are close to zero, indicating a general lack of strong multicollinearity among the predictors. This suggests that the continuous variables are largely independent and can contribute unique information to the regression models.

Key observations include: - **`AcidIndex` and `FixedAcidity`**: These variables exhibit a moderate positive correlation, indicating a potential relationship that should be considered when including both in regression models. - **`VolatileAcidity` and `CitricAcid`**: A weak negative correlation is present, suggesting that as one increases, the other slightly decreases. - Other variables, such as `Alcohol`, `Sulphates`, and `Density`, show minimal correlations with other predictors, further confirming the absence of significant multicollinearity.

In conclusion, Figure 4 confirms that multicollinearity is not a major issue in this dataset, allowing most continuous variables to be included in the regression models without significant adjustments. However, pairs with moderate correlations, such as `AcidIndex` and `FixedAcidity`, may require careful monitoring to avoid redundancy.

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.

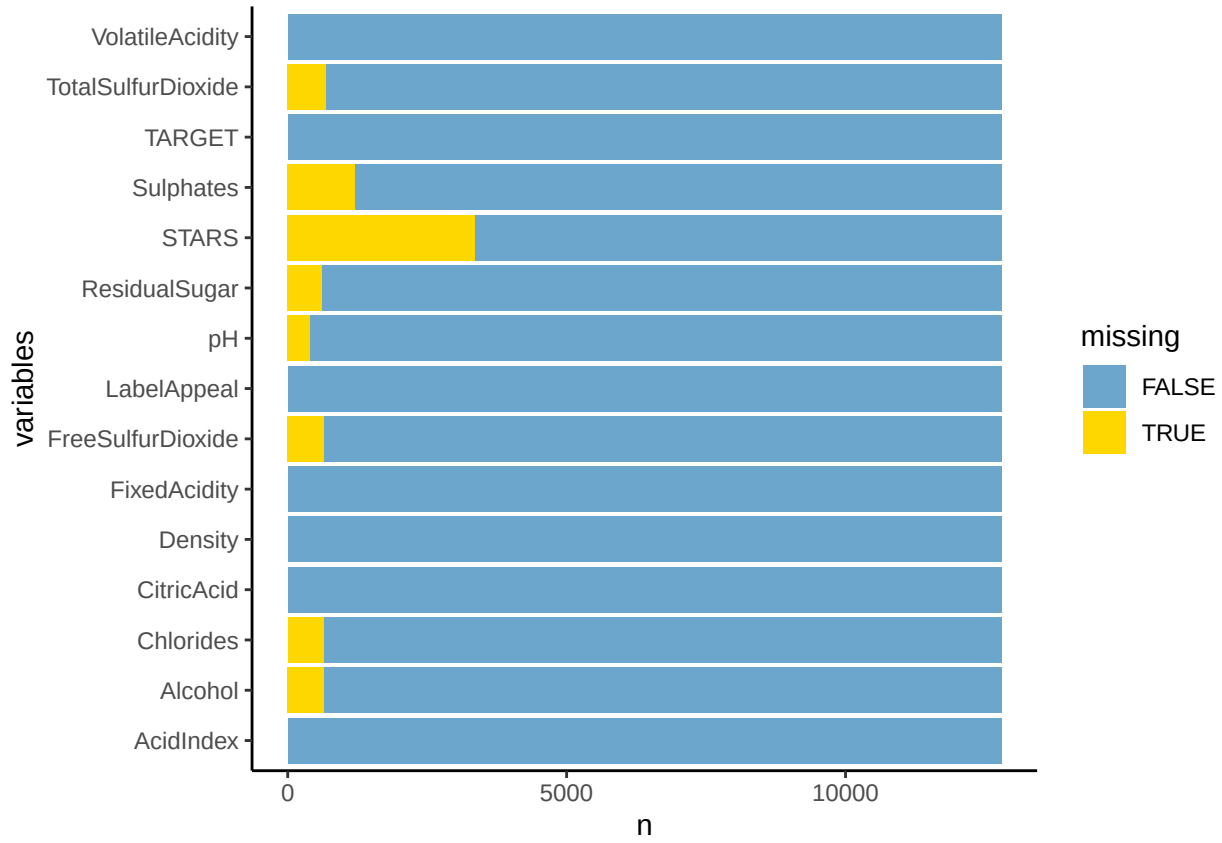

Figure 5: Barplot of number of missing values for each predictor. **Figure 5: Barplot of Number of Missing Values for Each Predictor**

Figure 5 illustrates the absence of missing values for all predictors in the dataset. The barplot shows that every variable has complete observations (indicated by all bars being fully labeled as `FALSE` for missing). This finding suggests the dataset is clean and does not require any imputation or handling of missing data during preprocessing.

Key Insights:

1. **Clean Dataset**: All variables (`VolatileAcidity`, `TotalSulfurDioxide`, `Alcohol`, etc.) are fully populated with no missing entries, which simplifies the data preparation process.
2. **Efficiency in Modeling**: Since there are no missing values, modeling efforts can focus on transformation and feature engineering without dedicating resources to missing value imputation.
3. **Data Quality**: The absence of missing data is a positive indicator of high-quality data collection and curation, providing a strong foundation for reliable analysis.

In conclusion, Figure 5 confirms that no predictors require imputation or deletion due to missing values, allowing for direct exploration, transformation, and modeling of all variables. This saves time and ensures consistency across the dataset.

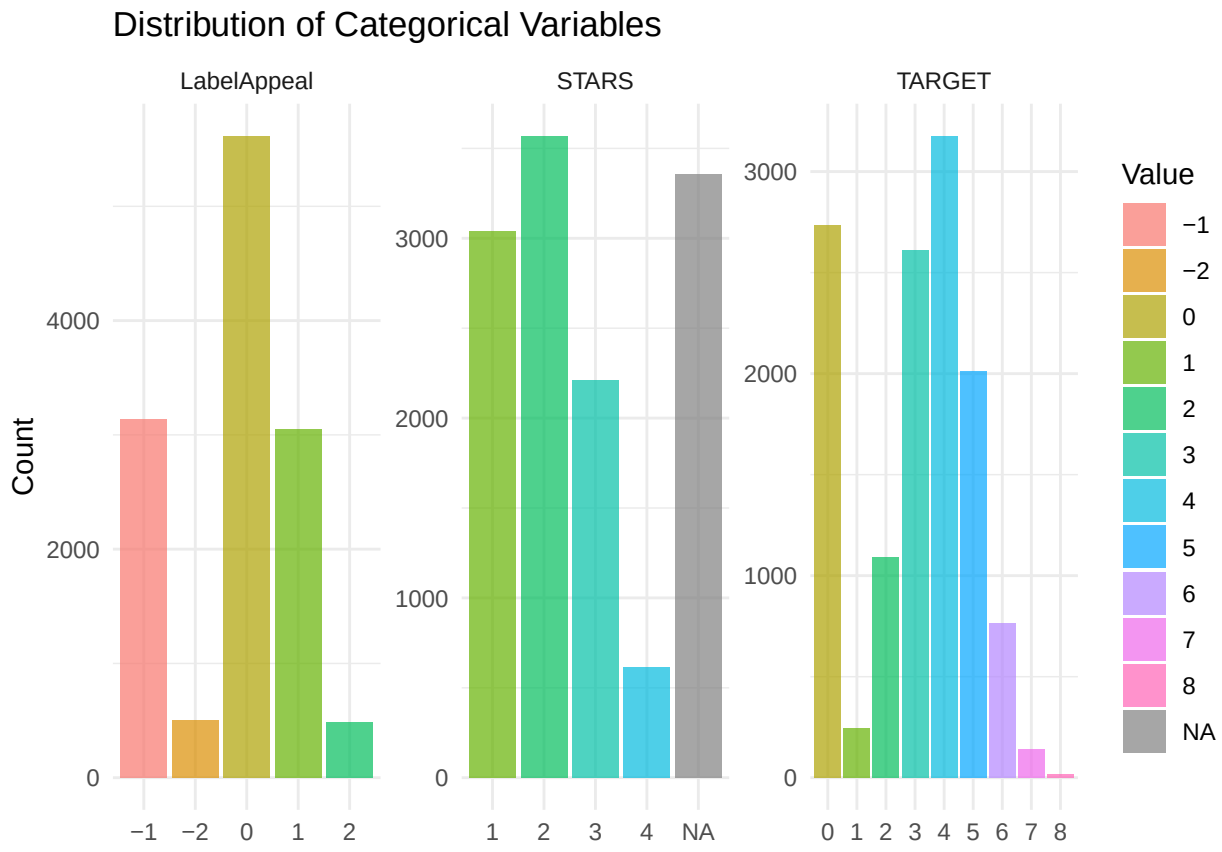

Figure 6: Distribution of Categorical Variables

Figure 6 displays the distribution of three categorical variables in the dataset: `LabelAppeal`, `STARS`, and `TARGET`. Key observations from this figure include:

- **`LabelAppeal`**: This variable ranges from -2 to 2, with the majority of observations centered at 0. Negative values are also frequent, indicating many wines may have less appealing labels, while higher positive values are relatively sparse.
- **`STARS`**: The distribution is skewed toward lower star ratings, with most wines rated as 1 or 2 stars. Very few wines achieve the highest rating of 4 stars.
- **`TARGET`**: The number of cases bought shows that 0 cases (no purchase) is the most common outcome, followed by a higher frequency of smaller purchases (1-3 cases). The frequency decreases substantially as the number of cases purchased increases, reflecting a trend where large purchases (e.g., 7-8 cases) are rare.

Examining Feature Multicollinearity

While Figure 6 provides insights into the categorical variables, understanding multicollinearity requires examining relationships between continuous variables. Multicollinearity can inflate variances in regression coefficients, making it crucial to identify and address correlations between features.

A correlation plot is typically used to assess relationships between continuous variables, as shown earlier in the report. This method helps determine whether any two continuous predictors are highly correlated,

which would necessitate removing or combining variables to avoid redundancy. For categorical variables like `LabelAppeal` and `STARS`, methodologies such as tetrachoric correlation may be applied, though in this dataset, categorical multicollinearity is less critical due to the limited scope of binary predictors.

In conclusion, while Figure 6 provides a detailed overview of categorical variable distributions, the analysis of multicollinearity primarily focuses on continuous predictors to ensure the integrity of the regression models.

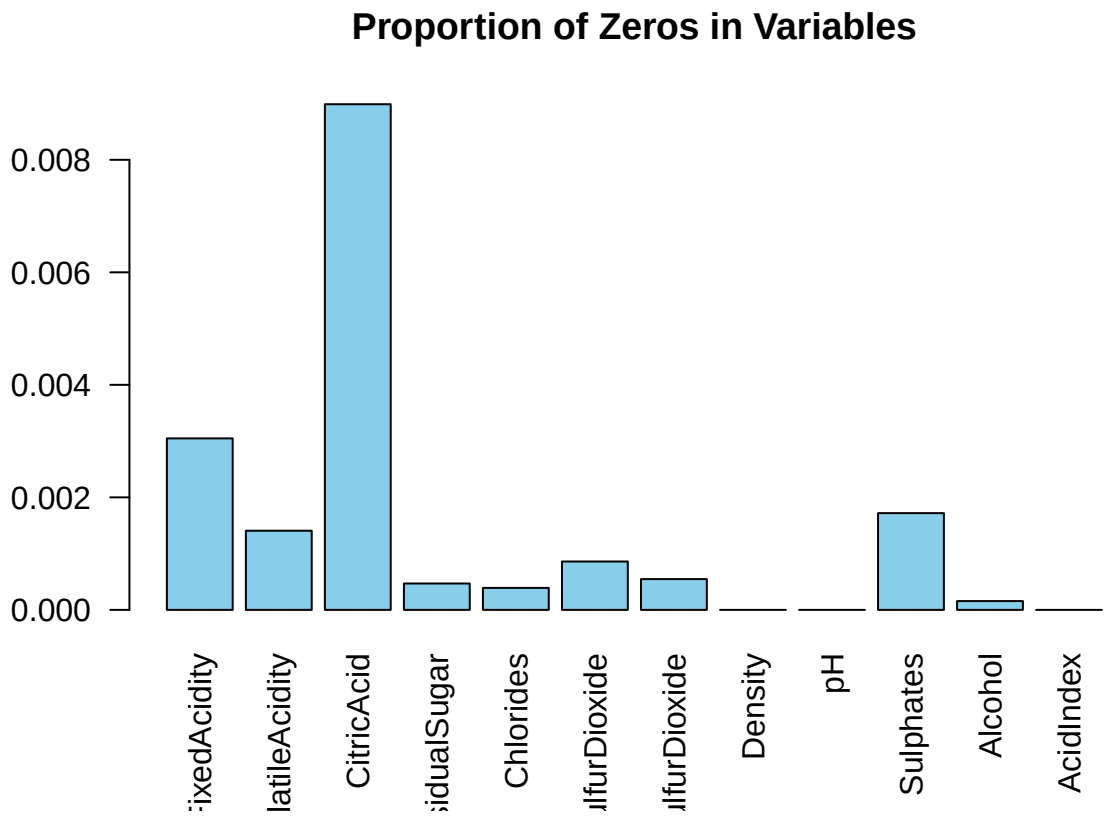

Figure 7: Proportion of Zeros in Variables

Figure 7 illustrates the proportion of zero values present in each continuous variable. The barplot reveals that certain variables, such as `CitricAcid`, have a significantly higher proportion of zero values compared to others. These zeros could represent meaningful characteristics of the data, such as absence or non-detection of specific chemical components, rather than missing or invalid entries.

Key Findings:

1. **High Proportion of Zeros**:
   - `CitricAcid` has the highest proportion of zeros among all variables, suggesting that many wine samples may lack measurable amounts of citric acid.
   - `Sulphates` and `VolatileAcidity` also exhibit smaller proportions of zeros, indicating their absence in some wines but to a lesser extent than `CitricAcid`.
2. **Variables with No Zeros**:
   - Several variables, such as `pH`, `Alcohol`, and `AcidIndex`, have no zeros in the dataset. This implies these features consistently have measurable, non-zero values across all samples.
3. **Potential Impact on Modeling**:

- The presence of zeros in variables like `CitricAcid` and `Sulphates` may introduce sparsity into the dataset, potentially influencing model performance. Special handling, such as flagging these cases with binary indicators or applying transformations, could be considered.
- Variables with no zeros, such as `pH` and `Alcohol`, can be modeled directly without additional preprocessing related to sparsity.

4. **Interpretation**:
    - Zeros in variables like `CitricAcid` and `Sulphates` could represent distinct characteristics of certain wine types. For example, wines with zero citric acid might belong to specific styles or production methods.
    - These zeros may carry predictive power for target outcomes (e.g., wine quality or sales) and should be evaluated carefully.

Conclusion:

The proportion of zeros in variables like `CitricAcid` and `Sulphates` highlights the need for targeted feature engineering. These variables may require additional attention during data preprocessing to ensure that their sparsity does not negatively impact the model's performance, while also leveraging the potential information these zeros might convey.

Data Preparation

Dealing with Missing Values

In general, imputing missing values using means or medians is considered acceptable if the missing data accounts for no more than 5% of the sample, as noted by Peng et al. (2006). However, when the proportion of missing values exceeds 20%, these simple imputation methods can artificially reduce variability, as they impute values centered around the variable's distribution, thereby failing to reflect the true spread of the data.

To address this, our team opted for a more robust approach: Multiple Imputation using Chained Equations (MICE) in R.

The MICE package implements a method where each incomplete variable is imputed using a model tailored specifically for that variable. As explained by Alice, plausible values are drawn from a distribution designed for the specific missing data points. Among the various imputation methods available within MICE, we selected Predictive Mean Matching (PMM), which is particularly suited for quantitative data.

Van Buuren describes PMM as a method that selects values from the observed data that are most likely to belong to the variable in the observation with the missing value. This approach ensures that only plausible values are chosen, avoiding issues such as imputing negative values where they would be inappropriate. Additionally, PMM avoids artificially reducing variability by using multiple regression models, which preserve the natural spread of errors. The method also accounts for uncertainty in imputation by generating multiple plausible values, leading to more reliable standard errors.

As noted by Marshall et al. (2010), a simulation study on skewed data concluded that predictive mean matching "may be the preferred approach provided that less than 50% of the cases have missing data." This reinforces the validity of using PMM for our dataset, ensuring that the imputation process reflects the true variability and distribution of the data while minimizing bias.

```
[1] 8200
```

```
[1] 2055
```

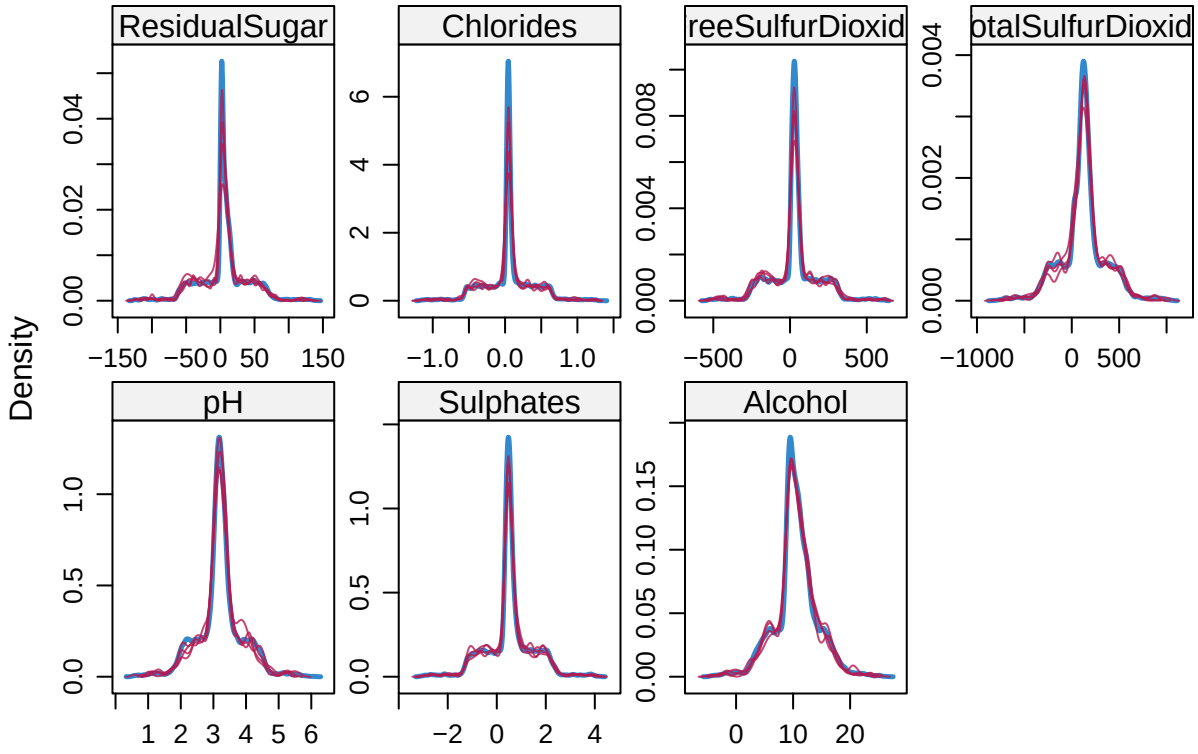

Figure 8: Density Plots for Variables with Missing Data
The density plots show the comparison of distributions between non-missing data (blue lines) and imputed data (red lines) for variables with missing values. The imputed data were generated using multiple imputations, with the number of imputations set to 4. The close alignment between the red and blue lines indicates that the distributions of the imputed data closely match those of the non-missing data, which is desirable. If significant discrepancies were observed, alternative imputation methods would need to be considered to improve the imputation quality.

Split Data Into Testing and Training

The dataset was divided into training and evaluation subsets, with 8200 observations allocated to the training subset (`wine_train`) and 5390 observations to the evaluation subset (`wine_eval`). These values reflect the original dataset. A similar division was applied to the dataset with imputed missing values, maintaining the same number of observations in `wine_train` and `wine_eval`.

```
  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

This section presents the coefficients and p-values for each of the models generated. For the stepAIC models, the selection direction was configured to `both`. The performance metrics for all models are detailed in the "Model Selection" section of this report.

Poisson Regression Models

This analysis involved constructing four distinct Poisson regression models using both the original and imputed/modified datasets. The models are as follows:

- A Poisson regression model based on the original dataset
- A Poisson regression model based on the modified dataset
- A Poisson regression model with significant features selected via stepAIC on the original dataset
- A Poisson regression model with significant features selected via stepAIC on the modified dataset

Poisson Regression Model Using Original Data The p-values for the coefficients in this model are presented below. At a 95% confidence level, `LabelAppeal`, `STARS`, `VolatileAcidity`, `AcidIndex`, and the `Intercept` are statistically significant. As previously discussed in the report, `STARS`, `LabelAppeal`, and `AcidIndex` are strongly correlated with the `TARGET` variable, which accounts for their low p-values.

```
                     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 Using Modified Data Similarly, the same highly correlated variables exhibit low p-values in this model. Notably, the p-values for these variables appear to be even lower compared to those observed in the Poisson regression model using the original dataset.

```
                      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 Regression Using Original Data Apart from `Chlorides` and `Alcohol`, all other variables are statistically significant in this model. As expected, the three variables—`STARS`, `LabelAppeal`, and `AcidIndex`—remain significant, consistent with previous findings.

```
                  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 Regression Using Modified Data This model reveals that with the imputed dataset, `FreeSulfurDioxide`, `TotalSulfurDioxide`, and `VolatileAcidity` are statistically significant vari-

ables. Sulfur dioxide plays a critical role in preserving wine by preventing oxidation and browning. Consequently, the levels of sulfur dioxide are significant factors influencing the number of wine cases purchased (refer to 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

This analysis included four distinct negative binomial models, constructed using both the original and imputed/modified datasets. The models are as follows:

- Negative binomial model using the original dataset
- Negative binomial model using the modified dataset
- Negative binomial model with significant features selected via stepAIC on the original dataset
- Negative binomial model with significant features selected via stepAIC on the modified dataset

Negative Binomial Model Using Original Data The p-values for the coefficients in this model are presented below. At a 95% confidence level, `LabelAppeal`, `STARS`, `VolatileAcidity`, `AcidIndex`, and the `Intercept` are statistically significant. As highlighted earlier in the report, `STARS`, `LabelAppeal`, and `AcidIndex` are strongly correlated with the `TARGET` variable, which explains their low p-values. Additionally, the selected variables and their p-values are very similar to those observed in the Poisson regression model using the original dataset.

```
Call:
glm.nb(formula = TARGET ~ ., data = original_train %>% dplyr::mutate(TARGET = as.numeric(TARGET)),
    init.theta = 241044.7343, link = log)

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(241044.7) 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.:  522595
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -16672.03
```

Negative Binomial Model Using Modified Data In this model, the same highly correlated variables exhibit low p-values, along with `FreeSulfurDioxide` and `TotalSulfurDioxide`, which were not statistically significant in the model using the original dataset. Additionally, `Chlorides` shows borderline statistical significance. Notably, the p-values for these variables are lower than those observed in the negative binomial model with original data. Furthermore, the selected variables and their p-values in this model closely align with those in the Poisson regression model using the modified dataset.

```
Call:
glm.nb(formula = TARGET ~ ., data = modified_train %>% dplyr::mutate(TARGET = as.numeric(TARGET)),
    init.theta = 103966.2431, link = log)

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(103966.2) 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 Using Original Data In this model, all variables except `Chlorides` and `Alcohol` are statistically significant. The three variables previously tested against `TARGET` using the Chi-square test—`STARS`, `LabelAppeal`, and `AcidIndex`—are included in this model, as expected. Moreover, the selected variables and their p-values are very similar to those observed in the Step AIC Poisson regression model using the original dataset.

```
Call:
glm.nb(formula = TARGET ~ VolatileAcidity + Chlorides + Alcohol +
    LabelAppeal + AcidIndex + STARS, data = original_train %>%
    dplyr::mutate(TARGET = as.numeric(TARGET)), init.theta = 240805.0189,
    link = log)

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(240805) 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:  240805
          Std. Err.:  521943
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -16675.07
```

Step AIC for Negative Binomial Model Using Modified Data Similar to the Step AIC Poisson regression model with modified data, the selected variables and their p-values in this model are largely consistent. This alignment reinforces the stability of the variable selection process and the significance of the chosen predictors in both models.

```
Call:
glm.nb(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
    TotalSulfurDioxide + Alcohol + LabelAppeal + AcidIndex +
    STARS, data = modified_train %>% dplyr::mutate(TARGET = as.numeric(TARGET)),
    init.theta = 103836.0548, link = log)

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(103836.1) 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

This analysis involved constructing four multiple linear regression models using both the original and imputed/modified datasets. The models are as follows:

- Multiple linear regression model using the original dataset
- Multiple linear regression model using the modified dataset
- Multiple linear regression model with significant features selected via stepAIC on the original dataset
- Multiple linear regression model with significant features selected via stepAIC on the modified dataset

Multiple Linear Regression Model Using Original Data The p-values for the coefficients in this model are presented below. At a 95% confidence level, `LabelAppeal`, `STARS`, `VolatileAcidity`, `Chlorides`, `Alcohol`, `AcidIndex`, and the `Intercept` are statistically significant. As discussed earlier in the report, `STARS`, `LabelAppeal`, and `AcidIndex` are strongly correlated with the `TARGET` variable, making their low p-values 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 Using Modified Data In this model, the same highly correlated variables continue to exhibit low p-values, along with `FreeSulfurDioxide` and `TotalSulfurDioxide`, which were not statistically significant in the model using the original dataset. Additionally, the p-value for `VolatileAcidity` has decreased further, indicating stronger statistical significance, while the p-value for `Alcohol` has increased slightly but remains 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 Using Original Data In this model, all variables except `FreeSulfurDioxide` and `TotalSulfurDioxide` are statistically significant. The three variables previously tested against `TARGET` using the Chi-square test—`STARS`, `LabelAppeal`, and `AcidIndex`—are included, as expected. Essentially, all variables that were statistically significant in the multiple linear regression model using the original dataset are retained in this model.

```
                      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 Using Modified Data In this model, all variables except `CitricAcid` and `Sulphates` are statistically significant. As expected, the three variables previously tested against `TARGET` using the Chi-square test—`STARS`, `LabelAppeal`, and `AcidIndex`—are included in this model. Essentially, all variables that were statistically significant in the multiple linear regression model using the modified dataset are retained 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
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

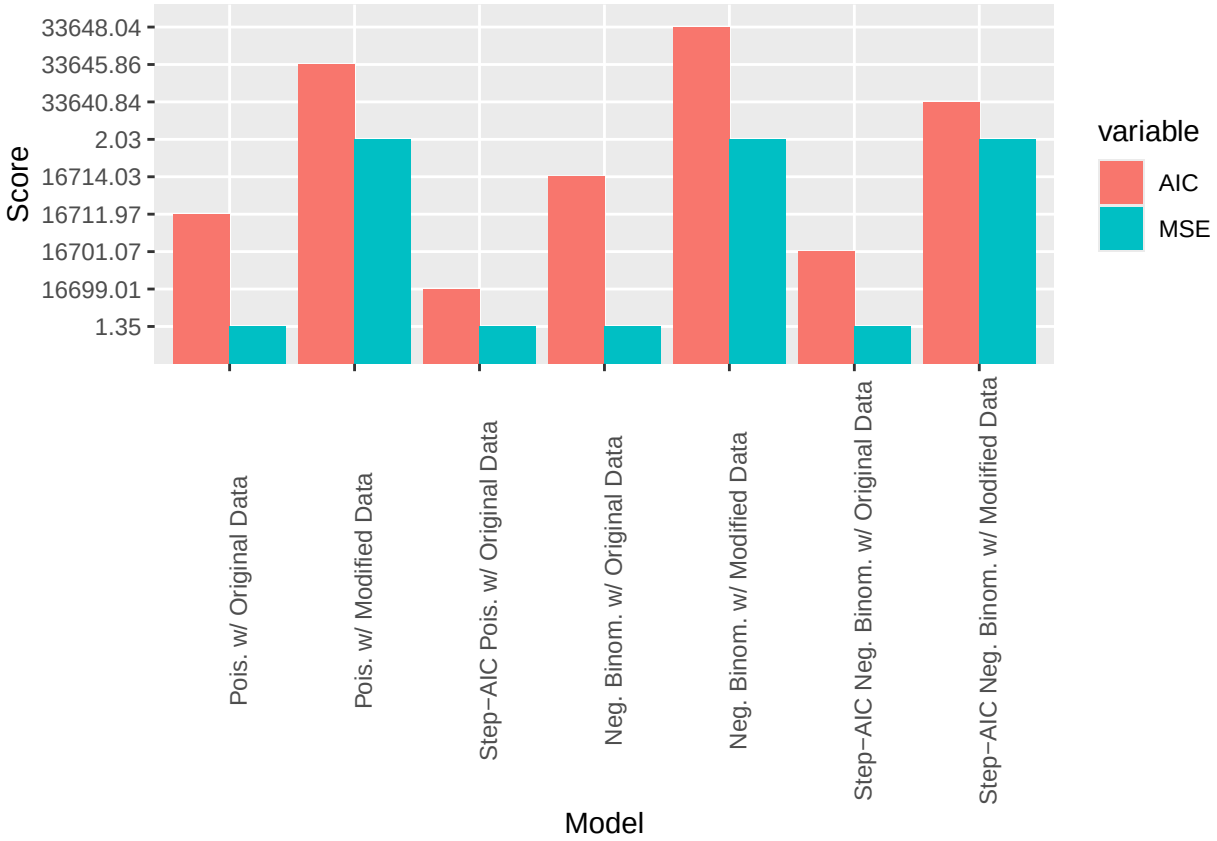

Figure 8: Bar Chart of Metrics for Binary Logistic Regression Models

Figure 8 illustrates that the Step-AIC Poisson model using the original data outperforms all other models. While the Mean Squared Error (MSE) remains consistent across all count regression models when using the original data, the Akaike Information Criterion (AIC) differs. Among these, the Step-AIC Poisson model with original data achieves the lowest AIC, indicating it is the most efficient and well-fitted model.

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

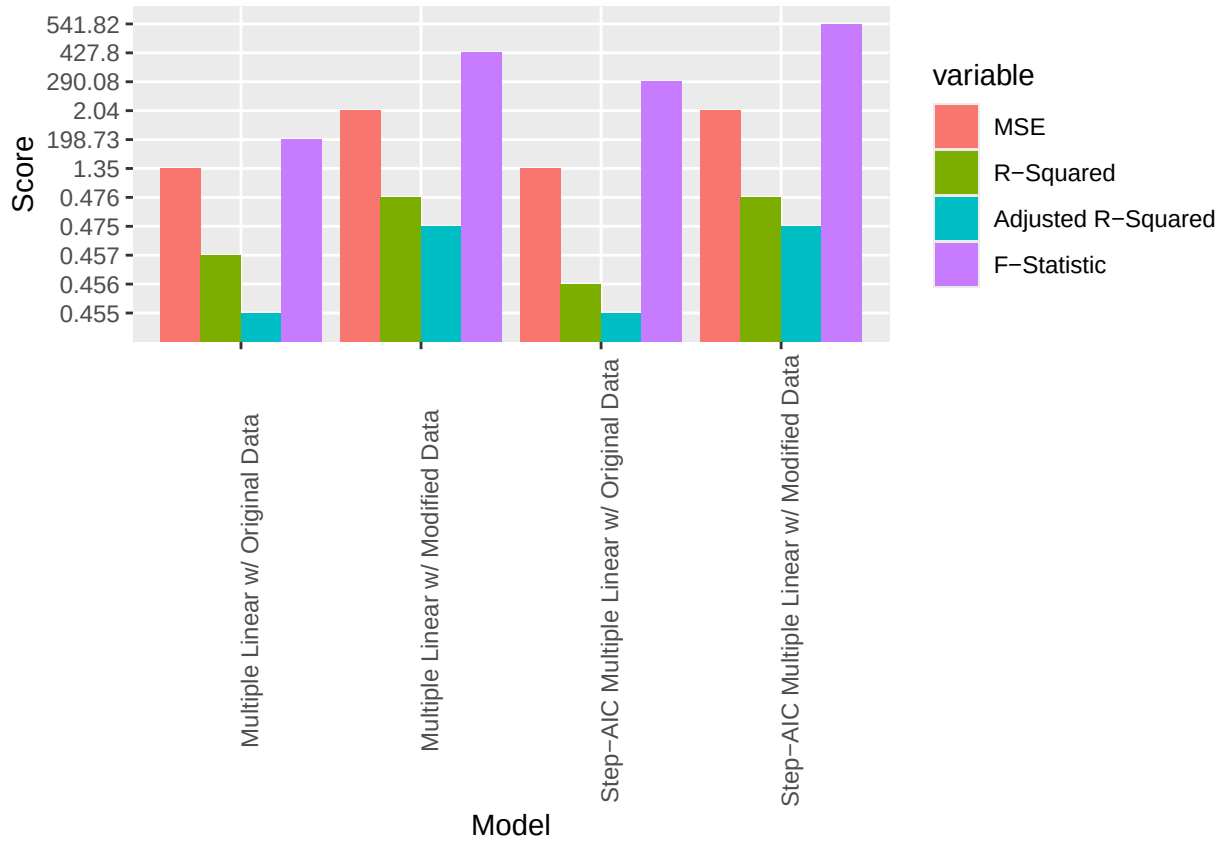

Figure 8: Metrics Bar Chart for Multiple Linear Regression Models

Among the linear regression models, the Step-AIC multiple linear regression model using modified data outperforms the rest. When compared to the multiple linear regression and Step-AIC models using the original dataset, it demonstrates higher R-squared and adjusted R-squared values. Additionally, the Step-AIC multiple linear regression model with modified data achieves a slightly higher F-statistic compared to the standard multiple linear regression model with modified data. These metrics indicate that the Step-AIC multiple linear regression model with modified data is the most effective, as it outperforms the other models in 3 out of the 4 evaluation criteria.

Given that the distribution of the imputed data closely aligns with the original dataset, it is reasonable to conclude that the Step-AIC multiple linear regression model with modified data will generalize well when applied to new data.

However, when considering Figure 8, Figure 9, and the model summaries provided in the "Build Models" section, the Step-AIC Poisson regression model using the original data emerges as the best overall model. It is more parsimonious and simpler than the Step-AIC multiple linear regression model with modified data while maintaining strong performance. This model allows for reliable predictions of the number of wine cases ordered based on the wine characteristics outlined in the "Step AIC for Poisson with Original Data" section.