Data 621 Homework 5

Critical Thinking Group 3: Vyannna Hill, Jose Rodriguez, and Christian Uriostegui

2023-12-09

Data Exploration

For wine evaluation, the team will review the data set for predicting the number of cases sold. The team included the lsit of features below for review.

Variable Name	Definition	Value
TARGET	Number of Cases Purchased	Response
AcidIndex	testing total acidity of wine by its weighted average	Predictor
Alcohol	Alcohol Content	Predictor
Chlorides	Chloride content of wine	Predictor
CitricAcid	Citric Acid Content	Predictor
Density	Density of Wine	Predictor
FixedAcidity	Fixed Acidity of Wine	Predictor
FreeSulfurDioxide	Sulfur Dioxide content of wine	Predictor
LabelAppeal	sentiment rating of the label	Predictor
STARS Wine	rating by a team of experts	Predictor
Sulphates	Sulfate conten of wine	Predictor
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	Predictor
VolatileAcidity	Volatile Acid content of wine	Predictor
pH	pH of wine	Predictor

From the predictor table, there are a few features that stand out. The features Label Appeal, STARS Wine, alcohol, and pH coefficients may have significance in the model as those items are most talk about in wine reviews. The team can keep those features in mind in the model building process later on.

The training data set has 12,795 observed wines and their ratings. From the summary, the team noticed that the features residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxiode, pH, sulphates, alcohol, and STARS have NA values. We suspect that these features are new to the evaluation process of wine sales. The largest NAs come from STARS, which could have a big influence in the regression models.

The team will need to impute for the missing values later on in the preparation process.

```
##     ï..INDEX         TARGET       FixedAcidity     VolatileAcidity
##  Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900
##  1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300
##  Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800
##  Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241
##  3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400
##  Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800
##
```

```
##    CitricAcid      ResidualSugar        Chlorides       FreeSulfurDioxide
##  Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00
##  1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00
##  Median : 0.3100   Median :   3.900   Median : 0.0460   Median :  30.00
##  Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85
##  3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00
##  Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00
##                    NA's   :616        NA's   :638       NA's   :647
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
##  1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800
##  Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000
##  Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
##  3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600
##  Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
##  NA's   :682                         NA's   :395     NA's   :1210
##     Alcohol       LabelAppeal          AcidIndex          STARS
##  Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median :10.40   Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##  NA's   :653                                          NA's   :3359
```

```
##           ï..INDEX             TARGET       FixedAcidity    VolatileAcidity
##                  0                  0                  0                  0
##         CitricAcid      ResidualSugar          Chlorides  FreeSulfurDioxide
##                  0                616                638                647
## TotalSulfurDioxide            Density                 pH          Sulphates
##                682                  0                395               1210
##            Alcohol        LabelAppeal          AcidIndex              STARS
##                653                  0                  0               3359
```

Predictor variables distributions Looking at the plots below, the team noticed that majority of features follow a near normal distribution. The only feature with a slight skewness is AcidIndex, which could benefit from a log transformation!

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 616 rows containing non-finite values (`stat_bin()`).
```

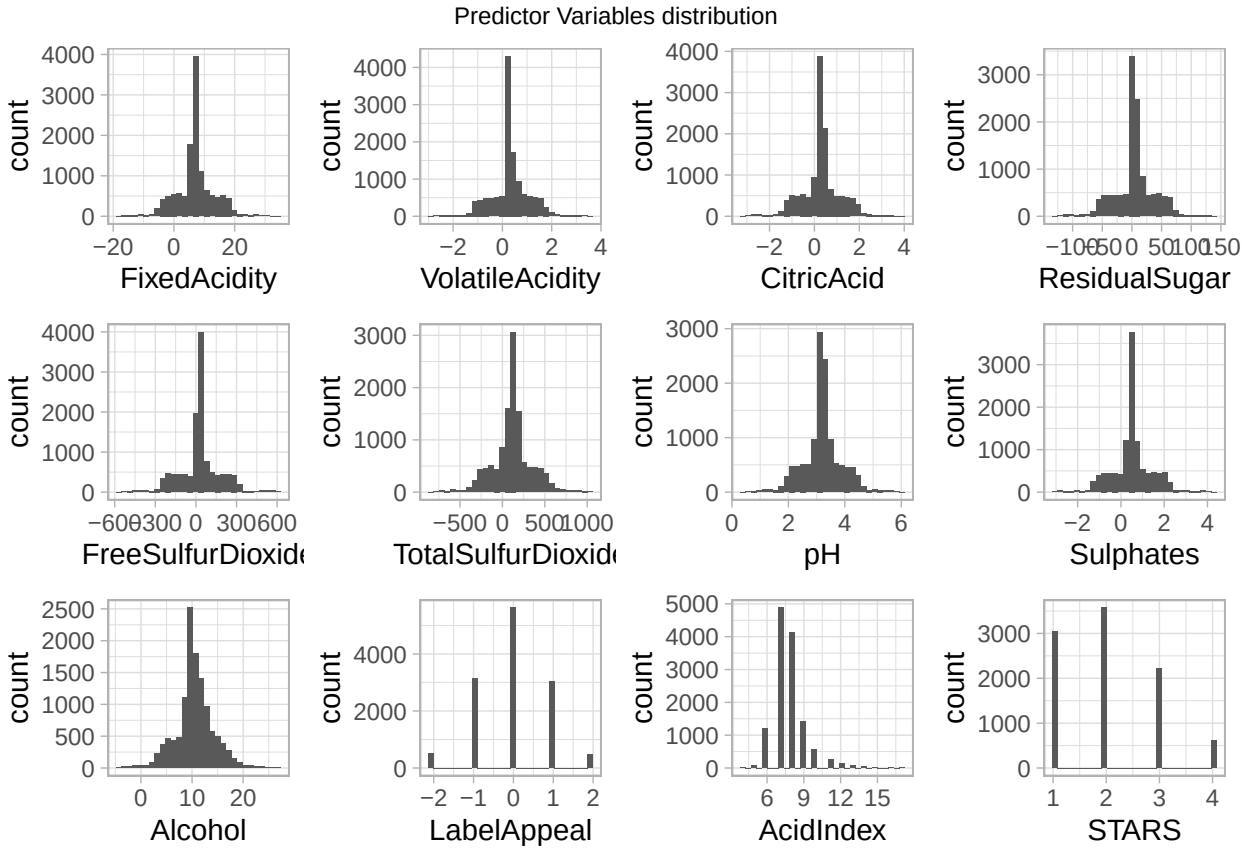
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 647 rows containing non-finite values (`stat_bin()`).
```

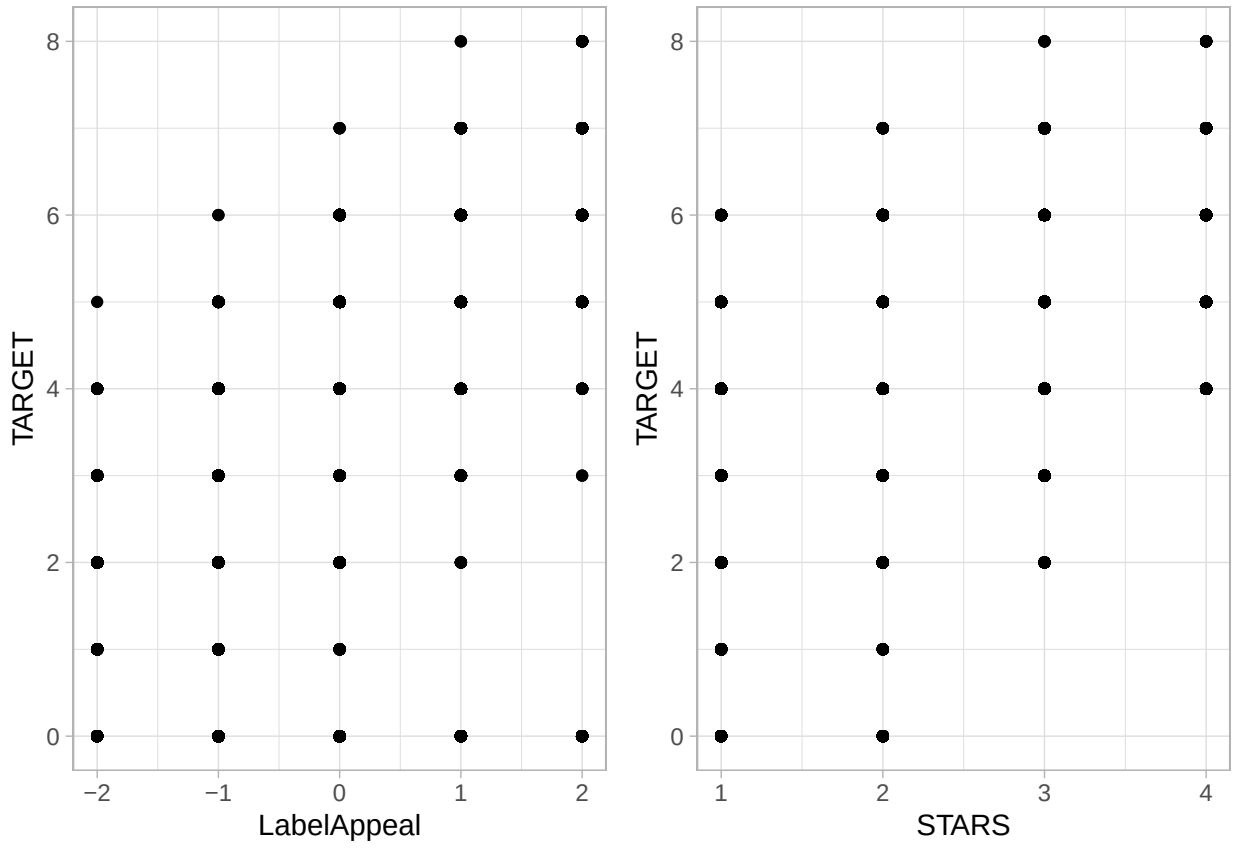
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 682 rows containing non-finite values (`stat_bin()`).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 395 rows containing non-finite values (`stat_bin()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 1210 rows containing non-finite values (`stat_bin()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 653 rows containing non-finite values (`stat_bin()`).

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3359 rows containing non-finite values (`stat_bin()`).
```

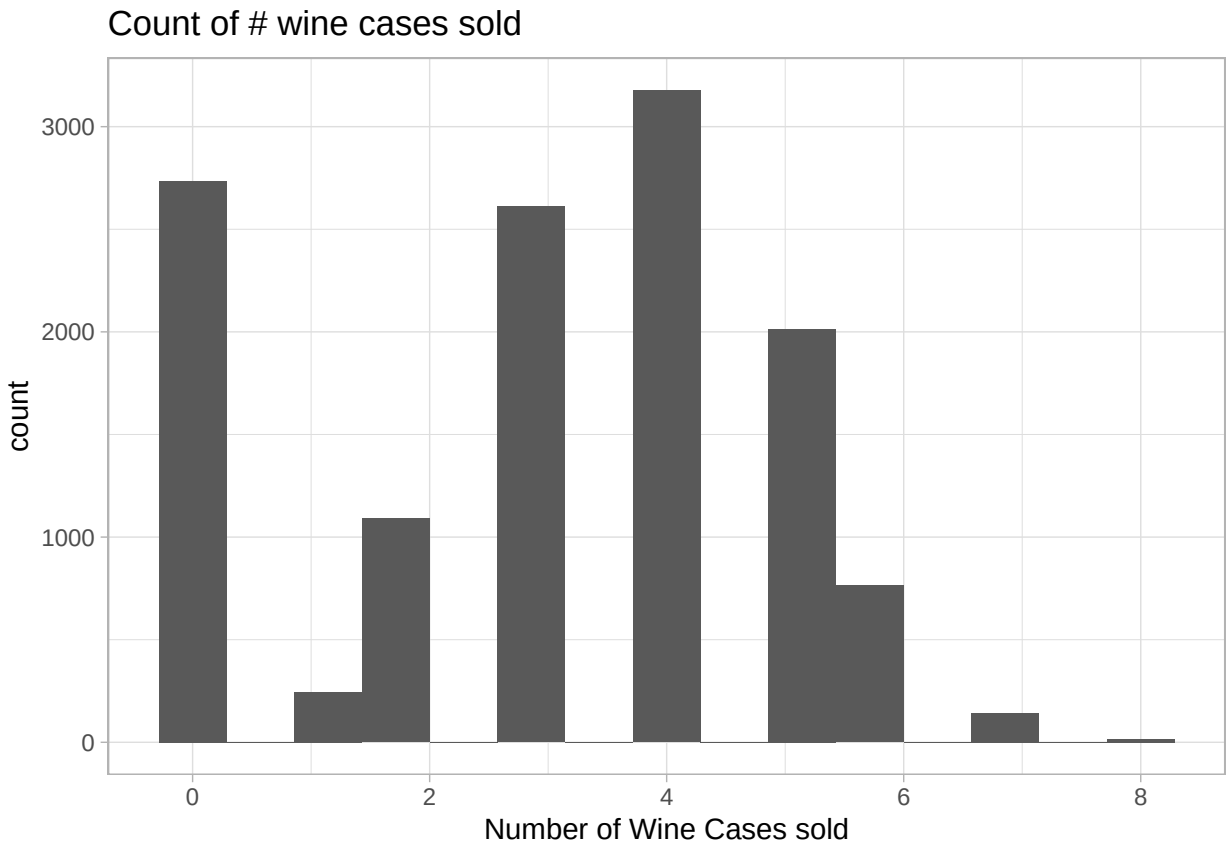



```
## Warning: Removed 3359 rows containing missing values (`geom_point()`).
```



Response distribution and Dispersion For the model building process, there is a assumption for Poisson regression that the response features do not included a majority of zeros. Checking for this assumption, the team noticed that the target response have a large share of zero values in the data set. We can assume that the data set is zero-inflated from this check. This can mean the Poisson regression model's fit will not be closest fit.

When checking the variance and mean of the response, we did see that the variance is a bit more than the mean of the response. This could mean the models could have over-dispersion, but a formal test will be used in this theory.




```
## # A tibble: 9 x 3
##   TARGET total    freq
##    <int> <int>   <dbl>
## 1      0  2734 0.214
## 2      1   244 0.0191
## 3      2  1091 0.0853
## 4      3  2611 0.204
## 5      4  3177 0.248
## 6      5  2014 0.157
## 7      6   765 0.0598
## 8      7   142 0.0111
## 9      8    17 0.00133
```

```
## [1] 3.710895
```

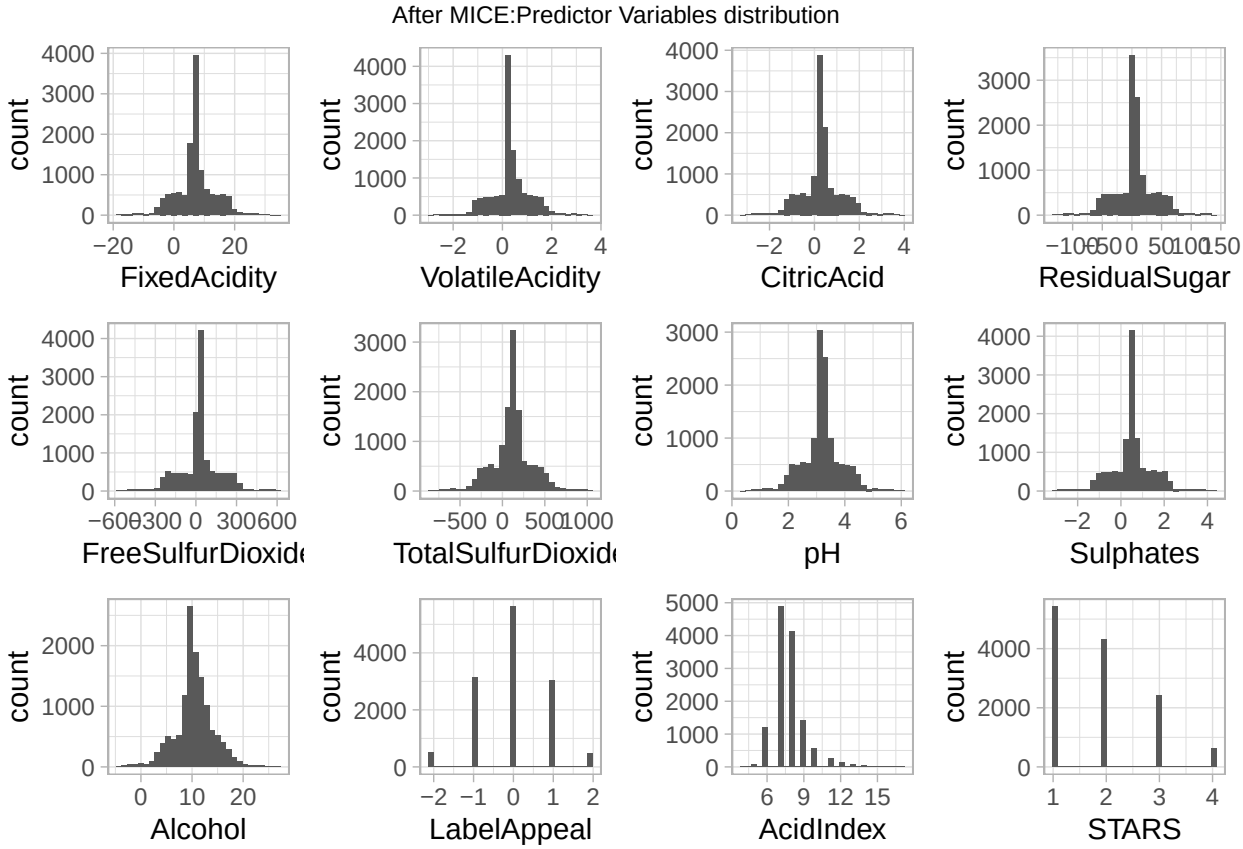
```
## [1] 3.029074
```

Data Preparation

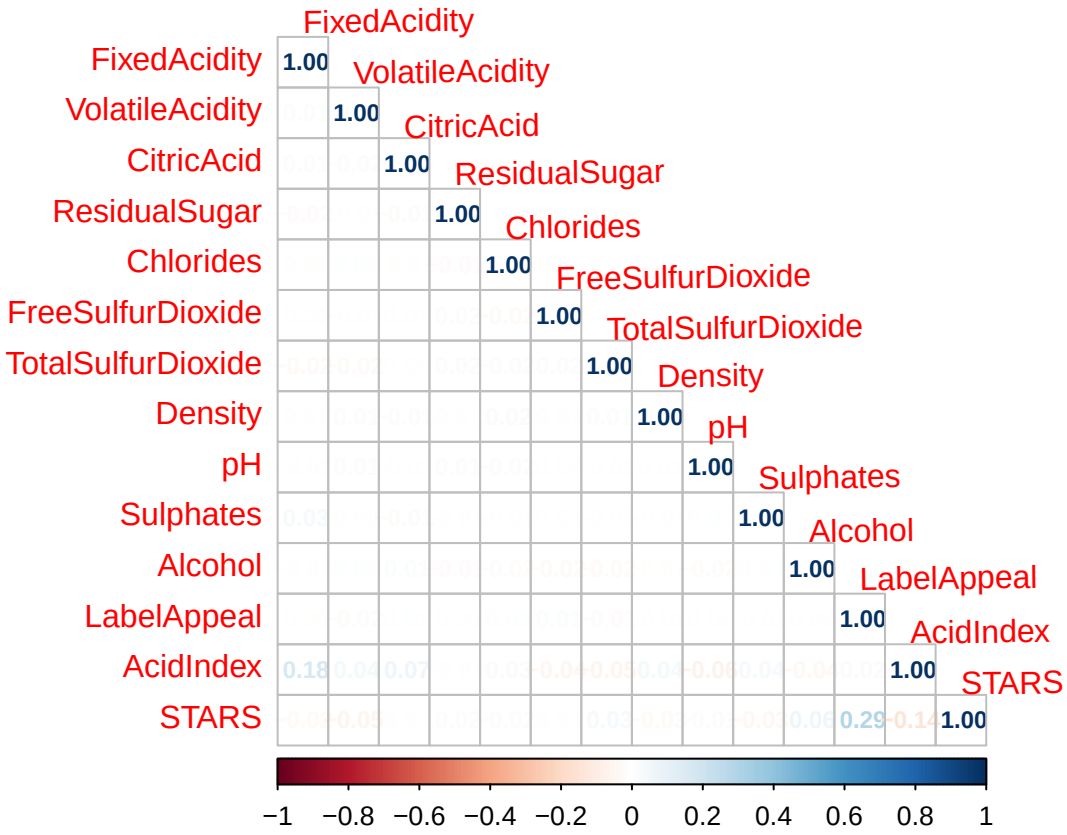
From the exploration, the only major issue of the data set is the missing values. As removal will reduce the model's performance, the team decided imputation is the best route.

Imputation via MICE The team imputed the missing values with MICE. The mice package will predict the missing value of the observation based on random complete cases in the data set.

```
##
##  iter imp variable
##    1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    1   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    1   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    1   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    1   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    2   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    2   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    2   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    2   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    3   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    3   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    3   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    3   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    4   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    4   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    4   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    4   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    5   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    5   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    5   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##    5   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
```



Feature Correlation The team reviewed correlation between features for any high correlation! There is no high correlation between the features.



Build Models

From the previous parts, the team is aware of that the response distribution has many zeros in the data set. This can influence how the Poisson regression models fit against the data set. After the review of models of poisson and negative binomial regressions, the team will also fit the training for the zero-inflated data set.

Poisson Model 1 For Poisson model one, the team included all features from the data set in the formula. The summary shows multiple features with coefficients that are statistically significant. Volatile Acidity, Chlorides, FreeSulfurDioxide, totalsulfurdioxide, ph, sulphates, labelAppeal, acidIndex, and stars are statistical significant features.

From earlier, the team appear correct in its assumptions on Label Appeal, STARS, and pH. Looking at its coefficients, labelappeal has a positive coefficient of (1.43e-01). This means the wine sees a additive effect of 1.43 on its cases sold by its label appeal. The additive effect appears in the alcohol as cases of wine increase its sales by (6.36e-01) if the alcohol percentage is higher.

Calling back to the influx of zero counts, let's check on the model's dispersion. The dispersion is 0.89, which was expected with the high zero count. However, the dispersion is not greater than one and p-value is greater than 0.05 so the model is technically dispersed. The team can see if model two's dispersion score lessens with a optimal subset of features, but it is not guaranteed.

```
##
## Call:
## glm(formula = TARGET ~ ., family = "poisson", data = train.clean)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9696  -0.6807   0.1261   0.6289   2.6448
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.516e+00  1.959e-01   7.741 9.83e-15 ***
## FixedAcidity        -6.194e-04  8.196e-04  -0.756  0.44981
## VolatileAcidity     -3.996e-02  6.521e-03  -6.128 8.87e-10 ***
## CitricAcid           1.014e-02  5.891e-03   1.721  0.08524 .
## ResidualSugar       -7.370e-07  1.499e-04  -0.005  0.99608
## Chlorides           -4.854e-02  1.606e-02  -3.022  0.00251 **
## FreeSulfurDioxide    1.436e-04  3.411e-05   4.211 2.54e-05 ***
## TotalSulfurDioxide   9.031e-05  2.209e-05   4.088 4.35e-05 ***
## Density             -3.227e-01  1.922e-01  -1.679  0.09317 .
## pH                  -1.799e-02  7.518e-03  -2.393  0.01670 *
## Sulphates           -1.306e-02  5.476e-03  -2.386  0.01704 *
## Alcohol              2.530e-03  1.373e-03   1.842  0.06540 .
## LabelAppeal          1.439e-01  6.082e-03  23.664  < 2e-16 ***
## AcidIndex           -9.797e-02  4.516e-03 -21.694  < 2e-16 ***
## STARS                3.375e-01  5.619e-03  60.066  < 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: 22861  on 12794  degrees of freedom
## Residual deviance: 16018  on 12780  degrees of freedom
## AIC: 47990
##
## Number of Fisher Scoring iterations: 5


##
##  Overdispersion test
##
## data:  pfit1
## z = -9.3457, p-value = 1
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##  0.8955757
```

Poisson Model 2 For this model, the team used the subset of the features that found to be statistically significant from the base model. There were a slight improvement to the AIC score from the previous iteration. There weren't any significant changes to the null deviance or the degrees of freedom. The coefficients saw slight changes in its p values with the feature subset.

Poisson model two saw a slight reduction in its dispersion score, but poisson regression may not be the best fit for the model

```
##
## Call:
## glm(formula = TARGET ~ STARS + LabelAppeal + pH + TotalSulfurDioxide +
```

```
##     FreeSulfurDioxide + Sulphates + Chlorides + AcidIndex + VolatileAcidity,
##     family = "poisson", data = train.clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9646  -0.6892   0.1233   0.6297   2.6477
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.226e+00  4.627e-02  26.486  < 2e-16 ***
## STARS                3.385e-01  5.602e-03  60.435  < 2e-16 ***
## LabelAppeal          1.438e-01  6.082e-03  23.649  < 2e-16 ***
## pH                  -1.823e-02  7.515e-03  -2.426  0.01528 *
## TotalSulfurDioxide   8.906e-05  2.207e-05   4.036 5.44e-05 ***
## FreeSulfurDioxide    1.420e-04  3.409e-05   4.164 3.13e-05 ***
## Sulphates           -1.311e-02  5.473e-03  -2.395  0.01662 *
## Chlorides           -5.004e-02  1.605e-02  -3.117  0.00182 **
## AcidIndex           -9.866e-02  4.452e-03 -22.160  < 2e-16 ***
## VolatileAcidity     -4.021e-02  6.520e-03  -6.167 6.94e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 16028  on 12785  degrees of freedom
## AIC: 47990
##
## Number of Fisher Scoring iterations: 5


##
##  Overdispersion test
##
## data:  pfit2
## z = -9.2491, p-value = 1
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##  0.8966245
```

Negative Binomial Model 1 Bouncing from the poisson models, negative binomial regression has a slight advantage. Negative binomial regression supposedly can handle count models with over-dispersion with correction based on the parameter. Although the previous poisson models were not statistically proven with over-dispersion, let's see if the built in correction helps the model fit.

For Negative Binomial Model one, let's use the optimal subset from poisson model two for features. From this run, there wasn't much of a change in the null deviance compared to poisson model two's results. The coefficients and the p-values are almost the same compared. When observing the likelihood score, the figure is very large. This points to the model not having the best fit with the data set provided as the dataset is suspected zero-inflated.

Checking on dispersion, the model is not over dispersed as the p-value does not break the null hypothesis.

```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
```

```
## control$trace > : iteration limit reached

## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached

##
## Call:
## glm.nb(formula = TARGET ~ STARS + LabelAppeal + pH + TotalSulfurDioxide +
##     FreeSulfurDioxide + Sulphates + Chlorides + AcidIndex + VolatileAcidity,
##     data = train.clean, init.theta = 48491.92994, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9645  -0.6892   0.1233   0.6297   2.6476
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.226e+00  4.627e-02  26.485  < 2e-16 ***
## STARS                3.385e-01  5.602e-03  60.433  < 2e-16 ***
## LabelAppeal          1.438e-01  6.082e-03  23.648  < 2e-16 ***
## pH                  -1.823e-02  7.516e-03  -2.426  0.01528 *
## TotalSulfurDioxide   8.907e-05  2.207e-05   4.036 5.44e-05 ***
## FreeSulfurDioxide    1.420e-04  3.409e-05   4.164 3.13e-05 ***
## Sulphates           -1.311e-02  5.473e-03  -2.395  0.01663 *
## Chlorides           -5.005e-02  1.605e-02  -3.117  0.00182 **
## AcidIndex           -9.866e-02  4.452e-03 -22.160  < 2e-16 ***
## VolatileAcidity     -4.021e-02  6.521e-03  -6.167 6.94e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48491.93) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 16027  on 12785  degrees of freedom
## AIC: 47992
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48492
##           Std. Err.:  55919
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -47970.08

## Likelihood ratio test of H0: Poisson, as restricted NB model:
## n.b., the distribution of the test-statistic under H0 is non-standard
## e.g., see help(odTest) for details/references
##
## Critical value of test statistic at the alpha= 0.05 level: 2.7055
## Chi-Square Test Statistic =  -0.2228 p-value = 0.5
```

Negative Binomial Model 2 Building off the last model, the team could see if the features STARS and AcidIndex have a effect on the deviance. From the distribution plots, those features distribution are slightly

right-skewed. This model will use the same subset of features but log transform STARS and AcidIndex. This change lowered the null deviance by 496 pts. This improvement in the model caused the STARS's coefficient to increased to (7.36e-01), which points out the feature has a stronger influence in more wine cases sold like the wine's sulfur dioxide content.

```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached

## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached


##
## Call:
## glm.nb(formula = TARGET ~ log(STARS) + LabelAppeal + pH + TotalSulfurDioxide +
##     FreeSulfurDioxide + Sulphates + Chlorides + log(AcidIndex) +
##     VolatileAcidity, data = train.clean, init.theta = 48189.24273,
##     link = log)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.05403  -0.65059   0.08671   0.60434   2.52504
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          2.128e+00  7.836e-02  27.160  < 2e-16 ***
## log(STARS)           7.360e-01  1.159e-02  63.481  < 2e-16 ***
## LabelAppeal          1.408e-01  6.049e-03  23.271  < 2e-16 ***
## pH                  -1.518e-02  7.524e-03  -2.018 0.043594 *
## TotalSulfurDioxide   8.505e-05  2.205e-05   3.858 0.000114 ***
## FreeSulfurDioxide    1.367e-04  3.406e-05   4.014 5.97e-05 ***
## Sulphates           -1.282e-02  5.469e-03  -2.345 0.019021 *
## Chlorides           -5.067e-02  1.604e-02  -3.159 0.001585 **
## log(AcidIndex)      -7.038e-01  3.523e-02 -19.980  < 2e-16 ***
## VolatileAcidity     -3.896e-02  6.531e-03  -5.966 2.43e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48189.24) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 15531  on 12785  degrees of freedom
## AIC: 47496
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48189
##           Std. Err.:  51121
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -47473.63

## Likelihood ratio test of H0: Poisson, as restricted NB model:
```

```
## n.b., the distribution of the test-statistic under H0 is non-standard
## e.g., see help(odTest) for details/references
##
## Critical value of test statistic at the alpha= 0.05 level: 2.7055
## Chi-Square Test Statistic =  -0.2633 p-value = 0.5
```

Multiple Linear Model 1 The team revisited linear regression with the current optimal subset of features. Surprisingly, the significance change with this regression model. STARS and AcidIndex coefficients also doubled in the this regression model. Although this model's explains 46% of the total variance, it might be in the best interest to re-examine the features any shifts towards significance in the next model.

```
##
## Call:
## lm(formula = TARGET ~ log(STARS) + LabelAppeal + pH + TotalSulfurDioxide +
##     FreeSulfurDioxide + Sulphates + Chlorides + log(AcidIndex) +
##     VolatileAcidity, data = train.clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6360 -0.9979  0.1205  1.0061  4.3061
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.948e+00  1.811e-01  32.843  < 2e-16 ***
## log(STARS)           2.232e+00  2.764e-02  80.775  < 2e-16 ***
## LabelAppeal          4.434e-01  1.454e-02  30.488  < 2e-16 ***
## pH                  -3.464e-02  1.832e-02  -1.891  0.05865 .
## TotalSulfurDioxide   2.394e-04  5.355e-05   4.470 7.88e-06 ***
## FreeSulfurDioxide    3.875e-04  8.331e-05   4.651 3.33e-06 ***
## Sulphates           -3.537e-02  1.334e-02  -2.652  0.00802 **
## Chlorides           -1.599e-01  3.912e-02  -4.087 4.41e-05 ***
## log(AcidIndex)      -1.924e+00  8.034e-02 -23.948  < 2e-16 ***
## VolatileAcidity     -1.169e-01  1.587e-02  -7.367 1.86e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.404 on 12785 degrees of freedom
## Multiple R-squared:  0.4695, Adjusted R-squared:  0.4691
## F-statistic:  1257 on 9 and 12785 DF,  p-value: < 2.2e-16
```

Multiple Linear Model 2 For linear model 2, the team revisited the full list of features (+ the log transformed STARS & AcidIndex) and used AICstep for the adjusted list of features with the lowest AIC possible. In this feature selection, the AIC is reduced to the lowest score thus far. Alcohol is now a feature that's statistically significant in the model.

```
## Start:  AIC=8677.89
## TARGET ~ FixedAcidity + FixedAcidity + VolatileAcidity + CitricAcid +
##     ResidualSugar + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     Density + pH + Sulphates + Alcohol + LabelAppeal + log(AcidIndex) +
##     log(STARS)
##
##                       Df Sum of Sq   RSS     AIC
```

```
## - ResidualSugar       1       0.0 25152  8675.9
## - FixedAcidity        1       1.9 25154  8676.9
## <none>                            25152  8677.9
## - CitricAcid          1       6.4 25158  8679.1
## - pH                  1       6.5 25158  8679.2
## - Density             1       6.7 25159  8679.3
## - Sulphates           1      13.8 25166  8682.9
## - Alcohol             1      21.4 25173  8686.8
## - Chlorides           1      31.1 25183  8691.7
## - TotalSulfurDioxide  1      40.8 25193  8696.6
## - FreeSulfurDioxide   1      44.1 25196  8698.3
## - VolatileAcidity     1     105.8 25258  8729.6
## - log(AcidIndex)      1    1067.2 26219  9207.6
## - LabelAppeal         1    1834.1 26986  9576.5
## - log(STARS)          1   12727.2 37879 13915.0
##
## Step:  AIC=8675.89
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + log(AcidIndex) + log(STARS)
##
##                      Df Sum of Sq   RSS     AIC
## - FixedAcidity        1       1.9 25154  8674.9
## <none>                            25152  8675.9
## - CitricAcid          1       6.4 25158  8677.1
## - pH                  1       6.5 25158  8677.2
## - Density             1       6.7 25159  8677.3
## + ResidualSugar       1       0.0 25152  8677.9
## - Sulphates           1      13.8 25166  8680.9
## - Alcohol             1      21.4 25173  8684.8
## - Chlorides           1      31.1 25183  8689.7
## - TotalSulfurDioxide  1      40.8 25193  8694.6
## - FreeSulfurDioxide   1      44.1 25196  8696.3
## - VolatileAcidity     1     105.8 25258  8727.6
## - log(AcidIndex)      1    1067.2 26219  9205.6
## - LabelAppeal         1    1834.1 26986  9574.5
## - log(STARS)          1   12731.4 37883 13914.4
##
## Step:  AIC=8674.86
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + log(AcidIndex) + log(STARS)
##
##                      Df Sum of Sq   RSS     AIC
## <none>                            25154  8674.9
## + FixedAcidity        1       1.9 25152  8675.9
## - CitricAcid          1       6.4 25160  8676.1
## - pH                  1       6.5 25160  8676.2
## - Density             1       6.7 25160  8676.2
## + ResidualSugar       1       0.0 25154  8676.9
## - Sulphates           1      14.1 25168  8680.0
## - Alcohol             1      21.4 25175  8683.7
## - Chlorides           1      31.0 25185  8688.6
## - TotalSulfurDioxide  1      41.2 25195  8693.8
```

```
## - FreeSulfurDioxide   1      43.9 25198  8695.2
## - VolatileAcidity     1     105.9 25260  8726.6
## - log(AcidIndex)      1    1115.3 26269  9228.0
## - LabelAppeal         1    1835.3 26989  9573.9
## - log(STARS)          1   12730.0 37884 13912.6

##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + log(AcidIndex) + log(STARS), data = train.clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6547 -0.9946  0.1202  1.0032  4.3592
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.668e+00  4.949e-01  13.474  < 2e-16 ***
## VolatileAcidity    -1.164e-01  1.586e-02  -7.337 2.31e-13 ***
## CitricAcid          2.598e-02  1.443e-02   1.800 0.071829 .
## Chlorides          -1.553e-01  3.912e-02  -3.969 7.26e-05 ***
## FreeSulfurDioxide   3.935e-04  8.329e-05   4.724 2.33e-06 ***
## TotalSulfurDioxide  2.449e-04  5.355e-05   4.573 4.85e-06 ***
## Density            -8.609e-01  4.682e-01  -1.839 0.065942 .
## pH                 -3.331e-02  1.831e-02  -1.819 0.068932 .
## Sulphates          -3.565e-02  1.333e-02  -2.674 0.007511 **
## Alcohol             1.099e-02  3.332e-03   3.298 0.000976 ***
## LabelAppeal         4.440e-01  1.454e-02  30.539  < 2e-16 ***
## log(AcidIndex)     -1.919e+00  8.059e-02 -23.807  < 2e-16 ***
## log(STARS)          2.226e+00  2.767e-02  80.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.403 on 12782 degrees of freedom
## Multiple R-squared:  0.4702, Adjusted R-squared:  0.4697
## F-statistic: 945.3 on 12 and 12782 DF,  p-value: < 2.2e-16
```

Bonus| zero-inflated Model Zero Inflated regression deals with two tasks, which finds the distribution of the non zero distribution and one that's the excess in zeros. For this regression model, the team will use the pscl package for the zero inflated regression function. For the feature selection, let's use the features from negative binomial model 2.

This model achieved the highest log-likelihood out of all the current models. This means the zero-inflated model has the closet fit to the data set. It might be too early for that call.

Let's also make another zero inflated regression model with the linear model's subset of features. This model will be to compare the coefficients that best match the linear model and if these features work better with zero inflation.

```
##
## Call:
## zeroinfl(formula = TARGET ~ log(STARS) + LabelAppeal + pH + TotalSulfurDioxide +
##     FreeSulfurDioxide + Sulphates + Chlorides + log(AcidIndex) + VolatileAcidity,
```

```
##     data = train.clean)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.0753 -0.4728  0.0202  0.4430  4.7788
##
## Count model coefficients (poisson with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.439e+00  8.480e-02  16.975  < 2e-16 ***
## log(STARS)          2.468e-01  1.280e-02  19.277  < 2e-16 ***
## LabelAppeal         2.374e-01  6.330e-03  37.501  < 2e-16 ***
## pH                  6.047e-03  7.813e-03   0.774   0.4390
## TotalSulfurDioxide -1.801e-05  2.208e-05  -0.816   0.4146
## FreeSulfurDioxide   3.018e-05  3.452e-05   0.874   0.3820
## Sulphates           3.538e-04  5.669e-03   0.062   0.9502
## Chlorides          -2.161e-02  1.662e-02  -1.301   0.1934
## log(AcidIndex)     -1.590e-01  3.872e-02  -4.107 4.01e-05 ***
## VolatileAcidity    -1.339e-02  6.770e-03  -1.978   0.0479 *
##
## Zero-inflation model coefficients (binomial with logit link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -8.6350916  0.4565508 -18.914  < 2e-16 ***
## log(STARS)         -4.5638604  0.1380847 -33.051  < 2e-16 ***
## LabelAppeal         0.7159722  0.0402796  17.775  < 2e-16 ***
## pH                  0.1854030  0.0471218   3.935 8.34e-05 ***
## TotalSulfurDioxide -0.0008189  0.0001366  -5.996 2.02e-09 ***
## FreeSulfurDioxide  -0.0008467  0.0002139  -3.958 7.54e-05 ***
## Sulphates           0.1223158  0.0342603   3.570 0.000357 ***
## Chlorides           0.2703918  0.1011278   2.674 0.007501 **
## log(AcidIndex)      3.7390415  0.2002230  18.674  < 2e-16 ***
## VolatileAcidity     0.2067694  0.0398681   5.186 2.14e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 27
## Log-likelihood: -2.107e+04 on 20 Df


##
## Call:
## zeroinfl(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + log(AcidIndex) + log(STARS), data = train.clean)
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.05952 -0.46541  0.02248  0.43900  4.79026
##
## Count model coefficients (poisson with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.655e+00  2.131e-01   7.764 8.23e-15 ***
## VolatileAcidity    -1.361e-02  6.770e-03  -2.011 0.044370 *
## CitricAcid          1.002e-03  6.061e-03   0.165 0.868752
## Chlorides          -1.916e-02  1.664e-02  -1.152 0.249308
## FreeSulfurDioxide   3.351e-05  3.454e-05   0.970 0.331945
```

```
## TotalSulfurDioxide -1.456e-05  2.210e-05  -0.659 0.509921
## Density            -3.133e-01  1.993e-01  -1.572 0.115966
## pH                  6.705e-03  7.814e-03   0.858 0.390868
## Sulphates           7.624e-05  5.671e-03   0.013 0.989273
## Alcohol             6.837e-03  1.406e-03   4.863 1.16e-06 ***
## LabelAppeal         2.379e-01  6.330e-03  37.590  < 2e-16 ***
## log(AcidIndex)     -1.469e-01  3.886e-02  -3.779 0.000157 ***
## log(STARS)          2.417e-01  1.283e-02  18.833  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -9.2315103  1.2577084  -7.340 2.14e-13 ***
## VolatileAcidity     0.2030009  0.0398720   5.091 3.56e-07 ***
## CitricAcid         -0.0594585  0.0369446  -1.609 0.107530
## Chlorides           0.2710310  0.1012449   2.677 0.007429 **
## FreeSulfurDioxide  -0.0008310  0.0002141  -3.881 0.000104 ***
## TotalSulfurDioxide -0.0007980  0.0001369  -5.830 5.55e-09 ***
## Density             0.2548504  1.1931137   0.214 0.830858
## pH                  0.1871599  0.0472220   3.963 7.39e-05 ***
## Sulphates           0.1196589  0.0343037   3.488 0.000486 ***
## Alcohol             0.0271805  0.0083981   3.237 0.001210 **
## LabelAppeal         0.7204506  0.0403300  17.864  < 2e-16 ***
## log(AcidIndex)      3.7760172  0.2011864  18.769  < 2e-16 ***
## log(STARS)         -4.5774277  0.1382757 -33.104  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 33
## Log-likelihood: -2.105e+04 on 26 Df
```

Select Models

There are a total of eight models for selection. The team will compared the performances of each model and its measure of fitness with the count data.

For starters, the team can put aside the poisson models from the model comparison as the data set's response counts are zeros. This lessens the amount of models for comparison as the previous performances improved from the initial models.

####Battle of the models

Looking at the first round of stats, the highest likelihood goes to zero inflated model two with a score of -21,048.24 with model one following behind. When it comes to the mc-fadden R^2, the linear model two has the highest R^2 of 0.15. For this round elimnation, the negative binomial models can be eliminated from the decision process.

```
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
## fitting null model for pseudo-r2
```

Likelihoods, AICs, and BICs Zero inflated Model two appears to the the winner in this chart. It holds the highest log likelihood and the lowest AIC and BIC. The team also check that zero inflated model two is statistically different than model one as a safety measure.

```
## Likelihood ratio test
##
## Model 1: TARGET ~ log(STARS) + LabelAppeal + pH + TotalSulfurDioxide +
##     FreeSulfurDioxide + Sulphates + Chlorides + log(AcidIndex) +
##     VolatileAcidity
## Model 2: TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + log(AcidIndex) + log(STARS)
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1  20 -21068
## 2  26 -21049  6 36.931  1.816e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Predicting the number of cases Using the winning model; zero inflated model two, let's see some of the predicted wine cases sold!

```
##
##  iter imp variable
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA

## Warning: Number of logged events: 1

##        1        2        3        4        5        6
## 0.000000 3.460069 0.000000 0.000000 3.586713 5.151381
```

Appendix

```
library(tidyverse)
library(MASS)
library(ggpubr)
library(caret)
library(AER)
library(GGally)
library(pscl)
library(ggpubr)
library(mice)
library(corrplot)
training_set<-read.csv("https://raw.githubusercontent.com/Vy4thewin/criticalthinking3/main/wine-training
testing_set<-read.csv("https://raw.githubusercontent.com/Vy4thewin/criticalthinking3/main/wine-evaluati

#summary statistics
summary(training_set)
colSums(is.na(training_set))

#reviewing the distributions of the predictors
g1<-ggplot(aes(x=FixedAcidity),data=training_set)+geom_histogram()+theme_light()
g2<-ggplot(aes(x=VolatileAcidity),data=training_set)+geom_histogram()+theme_light()
g3<-ggplot(aes(x=CitricAcid),data=training_set)+geom_histogram()+theme_light()
g4<-ggplot(aes(x=ResidualSugar),data=training_set)+geom_histogram()+theme_light()
g5<-ggplot(aes(x=Chlorides),data=training_set)+geom_histogram()+theme_light()
g5<-ggplot(aes(x=FreeSulfurDioxide),data=training_set)+geom_histogram()+theme_light()
g6<-ggplot(aes(x=TotalSulfurDioxide),data=training_set)+geom_histogram()+theme_light()
g7<-ggplot(aes(x=Density),data=training_set)+geom_histogram()+theme_light()
g7<-ggplot(aes(x=pH),data=training_set)+geom_histogram()+theme_light()
g8<-ggplot(aes(x=Sulphates),data=training_set)+geom_histogram()+theme_light()
g9<-ggplot(aes(x=Alcohol),data=training_set)+geom_histogram()+theme_light()
g10<-ggplot(aes(x=LabelAppeal),data=training_set)+geom_histogram()+theme_light()
g11<-ggplot(aes(x=AcidIndex),data=training_set)+geom_histogram()+theme_light()
g12<-ggplot(aes(x=STARS),data=training_set)+geom_histogram()+theme_light()
plt<-ggarrange(g1,g2,g3,g4,g5,g6,g7,g8,g9,g10,g11,g12,ncol = 4,nrow = 3)
annotate_figure(plt,top = text_grob("Predictor Variables distribution",size=9))

### Interaction of selected features and the response
g1<-training_set%>%ggplot(aes(x=LabelAppeal,y=TARGET))+geom_point()+theme_light()
g2<-training_set%>%ggplot(aes(x=STARS,y=TARGET))+geom_point()+theme_light()
ggarrange(g1,g2,ncol=2)

#ggpairs(train.clean%>%dplyr::select(-c(TARGET)))+theme_light()
corrplot(cor(train.clean[,2:15]),method = "number",type="lower", tl.srt = .71,number.cex=0.75

#review distribution of response
#too many zeros-> zero inflated
training_set%>%ggplot(aes(x=TARGET))+geom_histogram(bins=15)+theme_light()+labs(title="Count of # wine

#table view of response counts
training_set%>%group_by(TARGET)%>%summarise(total=n())%>%mutate(freq=total/sum(total))
```

```r
#check for over-dispersion. Var>Mean~ possible over-dispersion
var(training_set$TARGET)
mean(training_set$TARGET)

#Impute via mice with lasso norm
train.clean<-complete(mice(training_set,method = "pmm",seed = 333))
train.clean<-train.clean%>%dplyr::select(-c("ï..INDEX"))

#reviewing the distributions of the predictors after mice
g1<-ggplot(aes(x=FixedAcidity),data=train.clean)+geom_histogram()+theme_light()
g2<-ggplot(aes(x=VolatileAcidity),data=train.clean)+geom_histogram()+theme_light()
g3<-ggplot(aes(x=CitricAcid),data=train.clean)+geom_histogram()+theme_light()
g4<-ggplot(aes(x=ResidualSugar),data=train.clean)+geom_histogram()+theme_light()
g5<-ggplot(aes(x=Chlorides),data=train.clean)+geom_histogram()+theme_light()
g5<-ggplot(aes(x=FreeSulfurDioxide),data=train.clean)+geom_histogram()+theme_light()
g6<-ggplot(aes(x=TotalSulfurDioxide),data=train.clean)+geom_histogram()+theme_light()
g7<-ggplot(aes(x=Density),data=train.clean)+geom_histogram()+theme_light()
g7<-ggplot(aes(x=pH),data=train.clean)+geom_histogram()+theme_light()
g8<-ggplot(aes(x=Sulphates),data=train.clean)+geom_histogram()+theme_light()
g9<-ggplot(aes(x=Alcohol),data=train.clean)+geom_histogram()+theme_light()
g10<-ggplot(aes(x=LabelAppeal),data=train.clean)+geom_histogram()+theme_light()
g11<-ggplot(aes(x=AcidIndex),data=train.clean)+geom_histogram()+theme_light()
g12<-ggplot(aes(x=STARS),data=train.clean)+geom_histogram()+theme_light()
plt<-ggarrange(g1,g2,g3,g4,g5,g6,g7,g8,g9,g10,g11,g12,ncol = 4,nrow = 3)
annotate_figure(plt,top = text_grob("After MICE:Predictor Variables distribution",size=9))

#ggpairs(train.clean%>%dplyr::select(-c(TARGET)))+theme_light()
corrplot(cor(train.clean[,2:15]),method = "number",type="lower", tl.srt = .71,number.cex=0.75)

pfit1<-glm(TARGET~.,data=train.clean,family="poisson")
summary(pfit1)

#Dispersion check
dispersiontest(pfit1)

#subset of statistically significant features
pfit2<-glm(TARGET~ STARS+ LabelAppeal  + pH +TotalSulfurDioxide +  FreeSulfurDioxide+ Sulphates + Chlor
summary(pfit2)

#dispersion check
dispersiontest(pfit2)

nbfit1<-glm.nb(TARGET~ STARS+ LabelAppeal  + pH +TotalSulfurDioxide +  FreeSulfurDioxide+ Sulphates + Cl
summary(nbfit1)

#dispersion check
odTest(nbfit1)


nbfit2<-glm.nb(TARGET ~ log(STARS) + LabelAppeal + pH + TotalSulfurDioxide +FreeSulfurDioxide + Sulphat
summary(nbfit2)

#dispersion check
```

```
odTest(nbfit2)

lm1<-lm(TARGET ~ log(STARS) + LabelAppeal + pH + TotalSulfurDioxide +FreeSulfurDioxide + Sulphates + Ch
summary(lm1)

#use stepAIC to achieve lowest AIC in the linear model
lm2<-lm(TARGET~FixedAcidity+FixedAcidity+VolatileAcidity+CitricAcid+ResidualSugar+Chlorides+FreeSulfurD
lm2<-lm2%>%stepAIC(direction="both")
summary(lm2)

#using  zeroinfl() with nb's best features
zerop1<-zeroinfl(TARGET ~ log(STARS) + LabelAppeal + pH + TotalSulfurDioxide +FreeSulfurDioxide + Sulpha

summary(zerop1)

#second zero inflated model with lm's best features
zerop2<-zeroinfl(TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
    FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
    Alcohol + LabelAppeal + log(AcidIndex) + log(STARS), data = train.clean)
summary(zerop2)

#pull model statistics from the competing models
model.stats<-cbind(nbfit1=pR2(nbfit1,method="mcfadden"),nbfit2=pR2(nbfit2,method="mcfadden"),lm1=pR2(lm

#viewing new stats
model.stats<-cbind(zeroinfl1=c(logLik(zerop1),AIC(zerop1),BIC(zerop1)),zeroinfl2=c(logLik(zerop2),AIC(z

colnames(model.stats) <- c("Log Likelihood","AIC","BIC")

#checking stats difference between zero infalted models
lrtest(zerop1,zerop2)

#Reflecting transformation done on the train set
test.clean<-complete(mice(testing_set,method = "pmm",seed = 333))
test.clean<-test.clean%>%mutate(AcidIndex=log(AcidIndex),STARS=log(STARS))

#predicting based on the count version of the zero inflated model
wine.sold<-predict(zerop2,newdata =test.clean)
head(wine.sold)

#Reflecting transformation done on the train set
test.clean<-complete(mice(testing_set,method = "pmm",seed = 333))
test.clean<-test.clean%>%mutate(AcidIndex=log(AcidIndex),STARS=log(STARS))

#predicting based on the count version of the zero inflated model
wine.sold<-predict(zerop2,newdata =test.clean)
head(wine.sold)
```