Eric_Hirsch_621_Assignment_3

Predicting Wine Cases Bought

Eric Hirsch

4/7/2022

Contents

1. Data Exploration

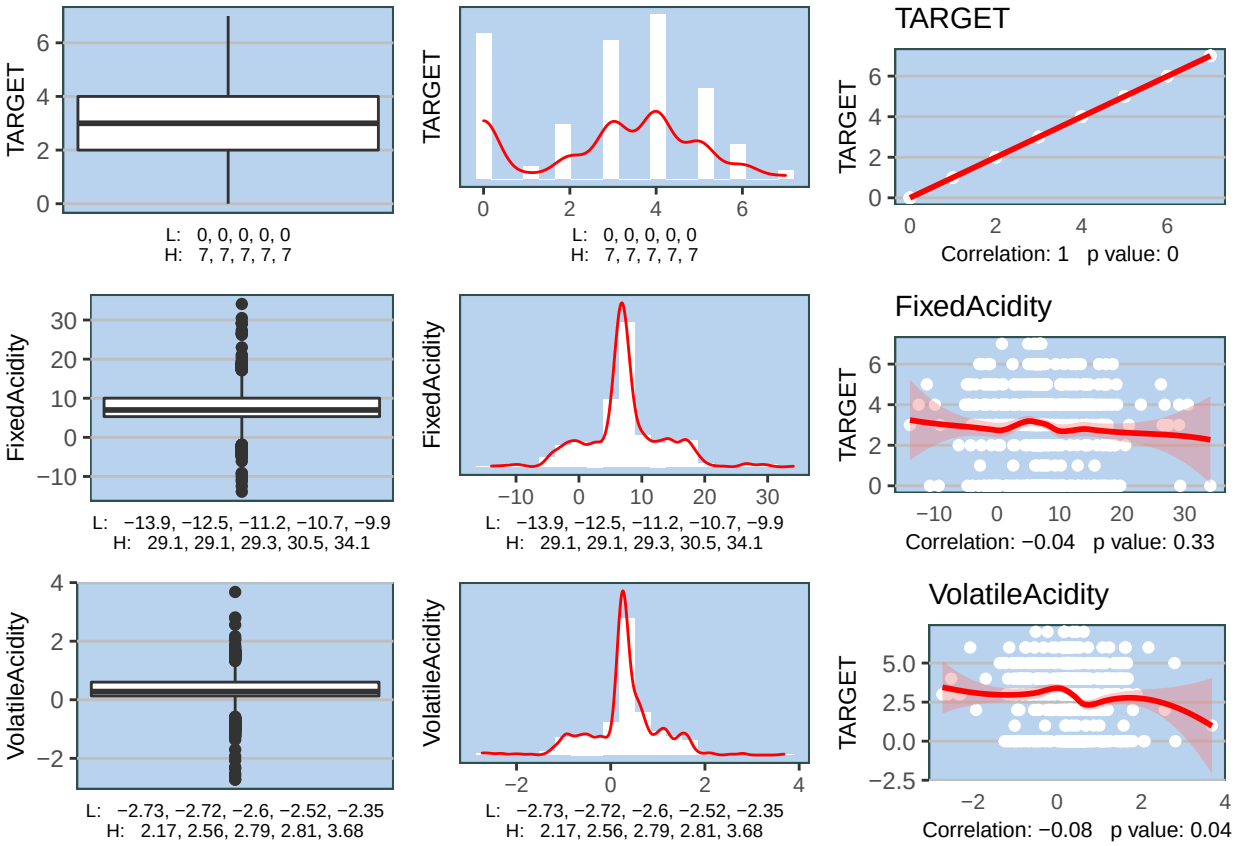
A. Summary Statistics We are modeling a data set containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable, "TARGET", is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine.

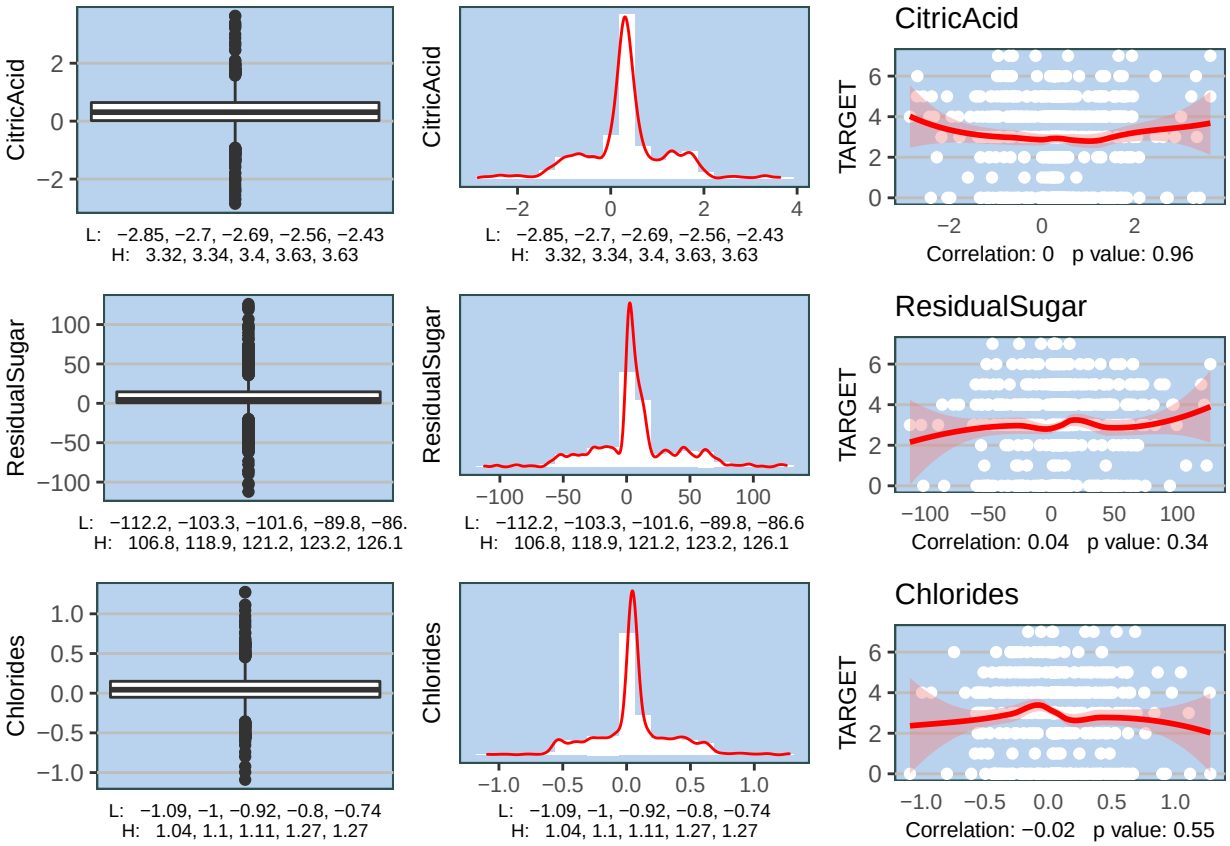
We first examine the data. The dataset consists of 12795 observations and 15 variables, all numeric. Some of the variables are count data, including the dependent varable "TARGET." The target variable has a minimum of 0, a maximum of 8 and a median of 3. Because of the small number of counts, this dataset is likely to be best modeled with a poisson or negative binomial count model.

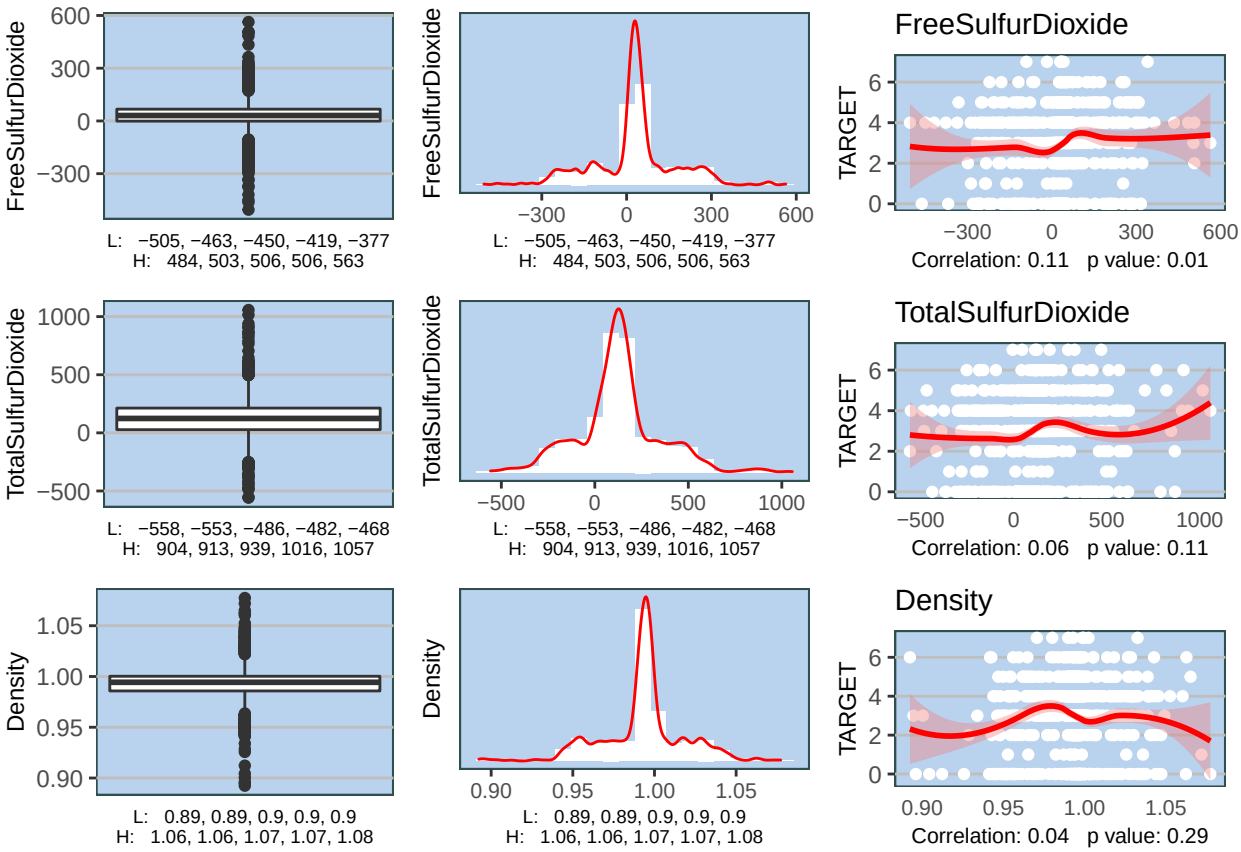
The mssing values should also be noted. There are a large number of missing values. In the case of STARS, e.g., 26% of the values (3359) are missing. We will need to make some decisions about these missing values.

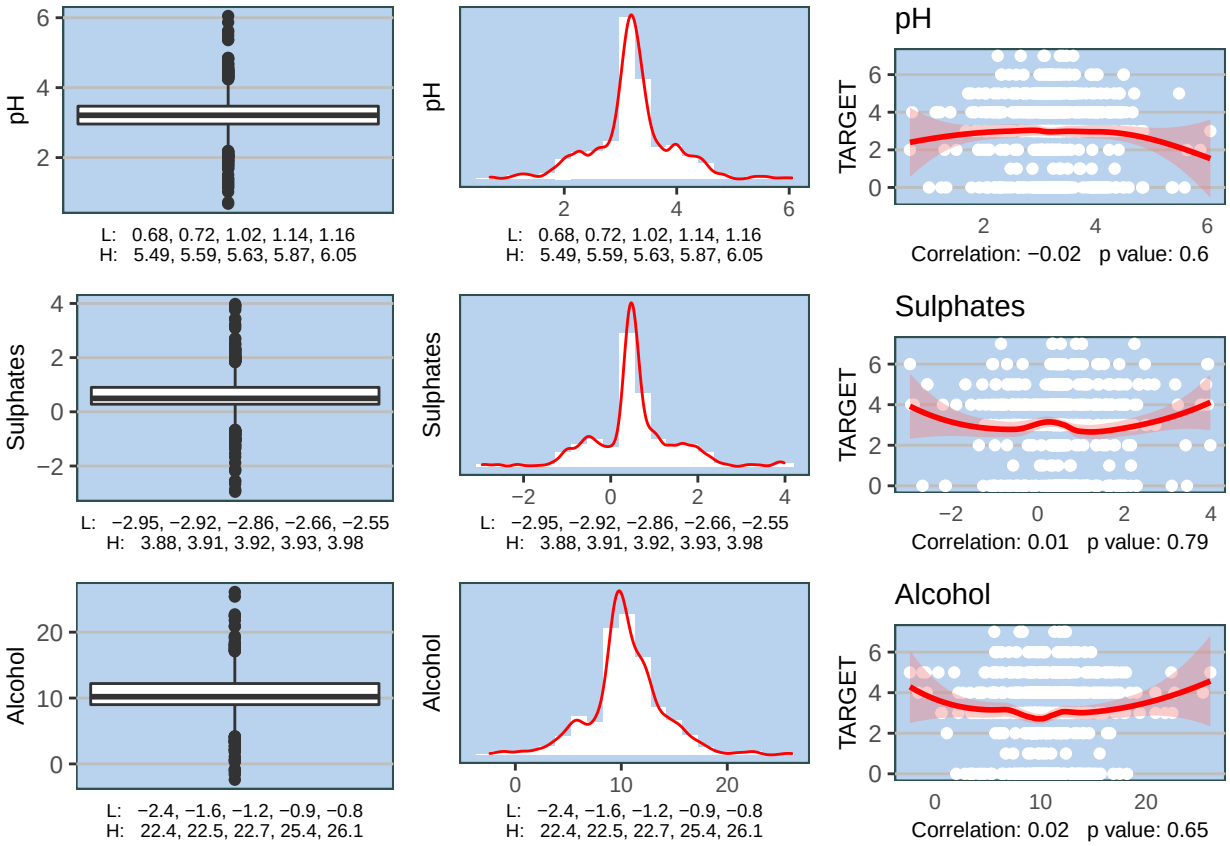
```
##      TARGET       FixedAcidity     VolatileAcidity     CitricAcid
##  Min.   :0.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400
##  1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300
##  Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100
##  Mean   :3.029   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084
##  3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800
##  Max.   :8.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600
##
##  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
##  Min.   :0.8881   Min.   :0.480   Min.   :-3.1300   Min.   :-4.70
##  1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800   1st Qu.: 9.00
##  Median :0.9945   Median :3.200   Median : 0.5000   Median :10.40
##  Mean   :0.9942   Mean   :3.208   Mean   : 0.5271   Mean   :10.49
##  3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600   3rd Qu.:12.40
##  Max.   :1.0992   Max.   :6.130   Max.   : 4.2400   Max.   :26.50
##                   NA's   :395     NA's   :1210      NA's   :653
##   LabelAppeal          AcidIndex          STARS
##  Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##                                       NA's   :3359

## 'data.frame':    12795 obs. of  15 variables:
##  $ TARGET            : int  3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity      : num  3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity   : num  1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
##  $ CitricAcid        : num  -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
##  $ ResidualSugar     : num  54.2 26.1 14.8 18.8 9.4 ...
##  $ Chlorides         : num  -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...
##  $ FreeSulfurDioxide : num  NA 15 214 22 -167 -37 287 523 -213 62 ...
##  $ TotalSulfurDioxide: num  268 -327 142 115 108 15 156 551 NA 180 ...
##  $ Density           : num  0.993 1.028 0.995 0.996 0.995 ...
##  $ pH                : num  3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
##  $ Sulphates         : num  -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...
##  $ Alcohol           : num  9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
##  $ LabelAppeal       : int  0 -1 -1 -1 0 0 0 1 0 0 ...
##  $ AcidIndex         : int  8 7 8 6 9 11 8 7 6 8 ...
##  $ STARS             : int  2 3 3 1 2 NA NA 3 NA 4 ...
```

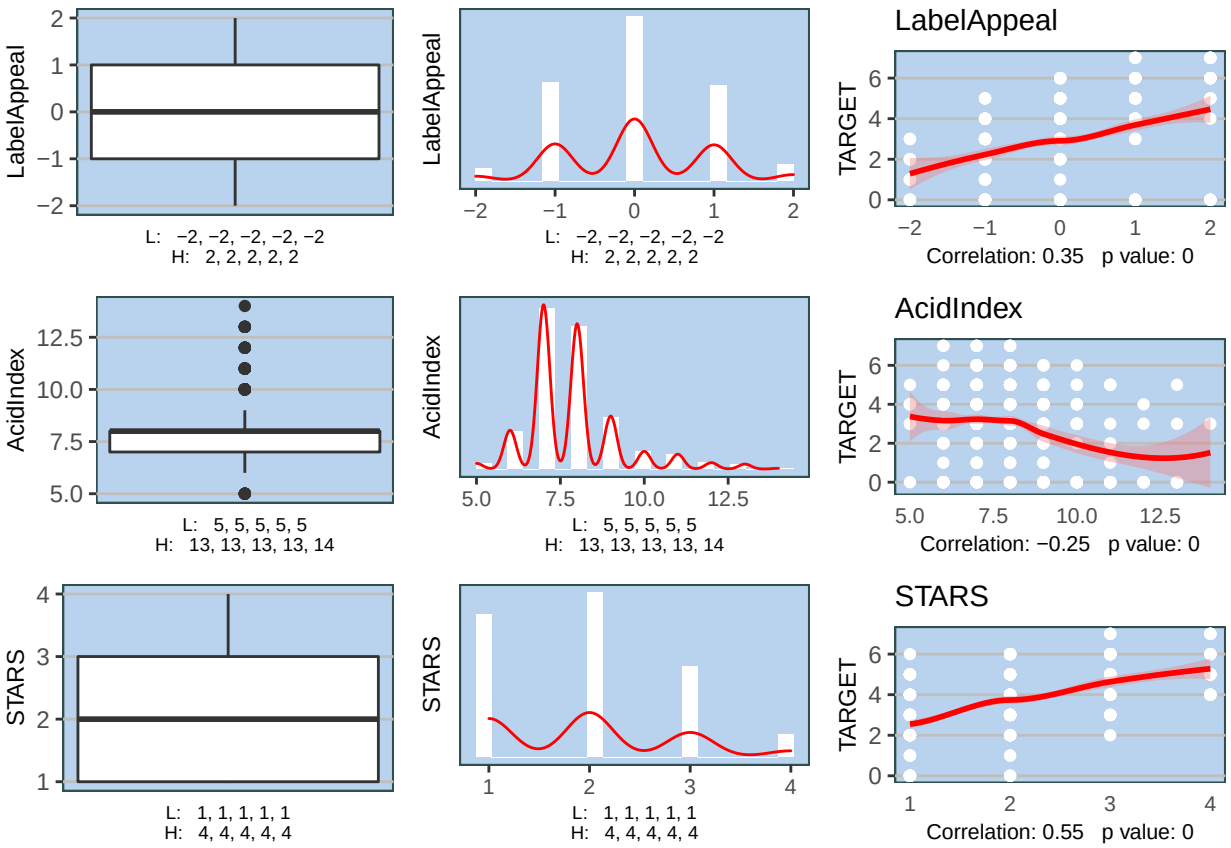
Now we examine distributions:


TARGET
L: 0, 0, 0, 0, 0
H: 7, 7, 7, 7, 7
TARGET
L: 0, 0, 0, 0, 0
H: 7, 7, 7, 7, 7
TARGET
Correlation: 1 p value: 0
FixedAcidity
L: −13.9, −12.5, −11.2, −10.7, −9.9
H: 29.1, 29.1, 29.3, 30.5, 34.1
FixedAcidity
L: −13.9, −12.5, −11.2, −10.7, −9.9
H: 29.1, 29.1, 29.3, 30.5, 34.1
FixedAcidity
Correlation: −0.04 p value: 0.33
VolatileAcidity
L: −2.73, −2.72, −2.6, −2.52, −2.35
H: 2.17, 2.56, 2.79, 2.81, 3.68
VolatileAcidity
L: −2.73, −2.72, −2.6, −2.52, −2.35
H: 2.17, 2.56, 2.79, 2.81, 3.68
VolatileAcidity
Correlation: −0.08 p value: 0.04


CitricAcid
2
0
−2
L: −2.85, −2.7, −2.69, −2.56, −2.43
H: 3.32, 3.34, 3.4, 3.63, 3.63
CitricAcid
−2
0
2
4
L: −2.85, −2.7, −2.69, −2.56, −2.43
H: 3.32, 3.34, 3.4, 3.63, 3.63
CitricAcid
TARGET
6
4
2
0
−2
0
2
Correlation: 0 p value: 0.96
ResidualSugar
100
50
0
−50
−100
L: −112.2, −103.3, −101.6, −89.8, −86.
H: 106.8, 118.9, 121.2, 123.2, 126.1
ResidualSugar
−100
−50
0
50
100
L: −112.2, −103.3, −101.6, −89.8, −86.6
H: 106.8, 118.9, 121.2, 123.2, 126.1
ResidualSugar
TARGET
6
4
2
0
−100
−50
0
50
100
Correlation: 0.04 p value: 0.34
Chlorides
1.0
0.5
0.0
−0.5
−1.0
L: −1.09, −1, −0.92, −0.8, −0.74
H: 1.04, 1.1, 1.11, 1.27, 1.27
Chlorides
−1.0
−0.5
0.0
0.5
1.0
L: −1.09, −1, −0.92, −0.8, −0.74
H: 1.04, 1.1, 1.11, 1.27, 1.27
Chlorides
TARGET
6
4
2
0
−1.0
−0.5
0.0
0.5
1.0
Correlation: −0.02 p value: 0.55

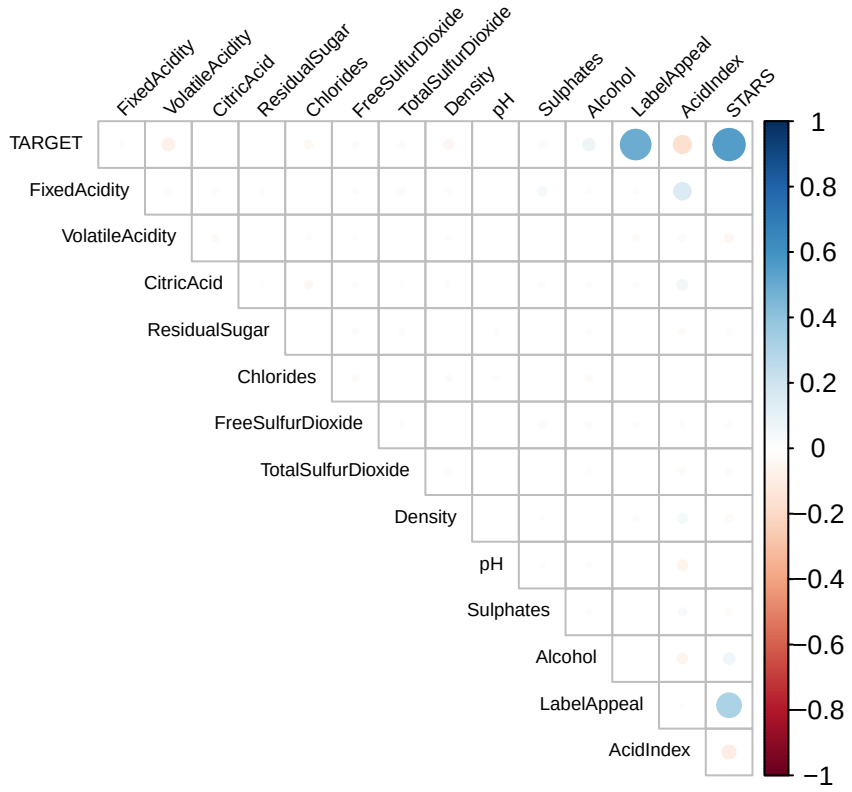

FreeSulfurDioxide
600
300
0
−300
L: −505, −463, −450, −419, −377
H: 484, 503, 506, 506, 563
FreeSulfurDioxide
−300
0
300
600
L: −505, −463, −450, −419, −377
H: 484, 503, 506, 506, 563
FreeSulfurDioxide
TARGET
6
4
2
0
−300
0
300
600
Correlation: 0.11 p value: 0.01
TotalSulfurDioxide
1000
500
0
−500
L: −558, −553, −486, −482, −468
H: 904, 913, 939, 1016, 1057
TotalSulfurDioxide
−500
0
500
1000
L: −558, −553, −486, −482, −468
H: 904, 913, 939, 1016, 1057
TotalSulfurDioxide
TARGET
6
4
2
0
−500
0
500
1000
Correlation: 0.06 p value: 0.11
Density
1.05
1.00
0.95
0.90
L: 0.89, 0.89, 0.9, 0.9, 0.9
H: 1.06, 1.06, 1.07, 1.07, 1.08
Density
0.90
0.95
1.00
1.05
L: 0.89, 0.89, 0.9, 0.9, 0.9
H: 1.06, 1.06, 1.07, 1.07, 1.08
Density
TARGET
6
4
2
0
0.90
0.95
1.00
1.05
Correlation: 0.04 p value: 0.29


pH
L: 0.68, 0.72, 1.02, 1.14, 1.16
H: 5.49, 5.59, 5.63, 5.87, 6.05
pH
L: 0.68, 0.72, 1.02, 1.14, 1.16
H: 5.49, 5.59, 5.63, 5.87, 6.05
pH
TARGET
Correlation: −0.02 p value: 0.6
Sulphates
L: −2.95, −2.92, −2.86, −2.66, −2.55
H: 3.88, 3.91, 3.92, 3.93, 3.98
Sulphates
L: −2.95, −2.92, −2.86, −2.66, −2.55
H: 3.88, 3.91, 3.92, 3.93, 3.98
Sulphates
TARGET
Correlation: 0.01 p value: 0.79
Alcohol
L: −2.4, −1.6, −1.2, −0.9, −0.8
H: 22.4, 22.5, 22.7, 25.4, 26.1
Alcohol
L: −2.4, −1.6, −1.2, −0.9, −0.8
H: 22.4, 22.5, 22.7, 25.4, 26.1
Alcohol
TARGET
Correlation: 0.02 p value: 0.65



Looking at boxplots, histograms, and scatterplots against the target variable for each variable in the dataset, we see some areas of interest. First, the target variable appears to be normally distributed for wines that had at least one purchase - but a large percentage of observations have zero purchases. This suggests the need for a zero-inflated model. Second, most of the variables exhibit the same pattern - a high peak at the mean of the distribution with long tails on either end. Finally, AcidIndex, LabelAppeal and Stars have, by far, the highest direct correlations with TARGET.

B. Multicollinearity Looked at from a strictly pairwise view, there is very little mutlicollinearity in this database. This is, frankly, very surprising given pairings like “FreeSulfurDixide” and “TotalSulfurDioxide”, and “FixedAcidity”, “VolatileAcidity” and “Ph”. If it were possible, I would definitely want to explore these findings with a chemist to understand them better and verify them.



C. A preliminary exploratory model Whether we choose a poisson or negative binomial model will depend on the residuals of our poisson model. Once we account for missing values, add columns and trasnform columns, our model will change. It is therefore a good exercise to run a poisson model at this point to get some insight into the data.

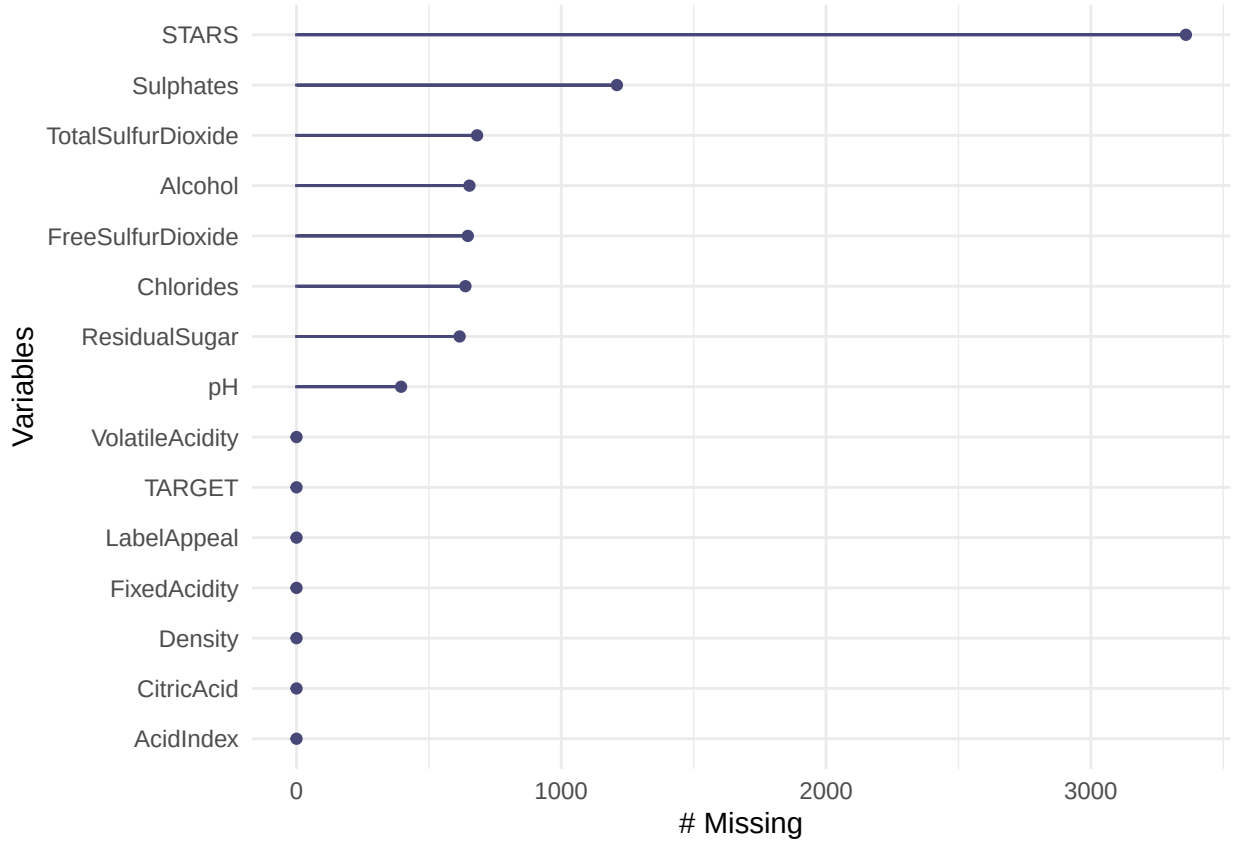
```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2158  -0.2734   0.0616   0.3732   1.6830
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.593e+00  2.506e-01   6.359 2.03e-10 ***
## FixedAcidity         3.293e-04  1.053e-03   0.313  0.75447
## VolatileAcidity     -2.560e-02  8.353e-03  -3.065  0.00218 **
## CitricAcid          -7.259e-04  7.575e-03  -0.096  0.92365
## ResidualSugar       -6.141e-05  1.941e-04  -0.316  0.75165
## Chlorides           -3.007e-02  2.056e-02  -1.463  0.14346
## FreeSulfurDioxide    6.734e-05  4.404e-05   1.529  0.12620
## TotalSulfurDioxide   2.081e-05  2.855e-05   0.729  0.46618
## Density             -3.725e-01  2.462e-01  -1.513  0.13026
## pH                  -4.661e-03  9.598e-03  -0.486  0.62722
```

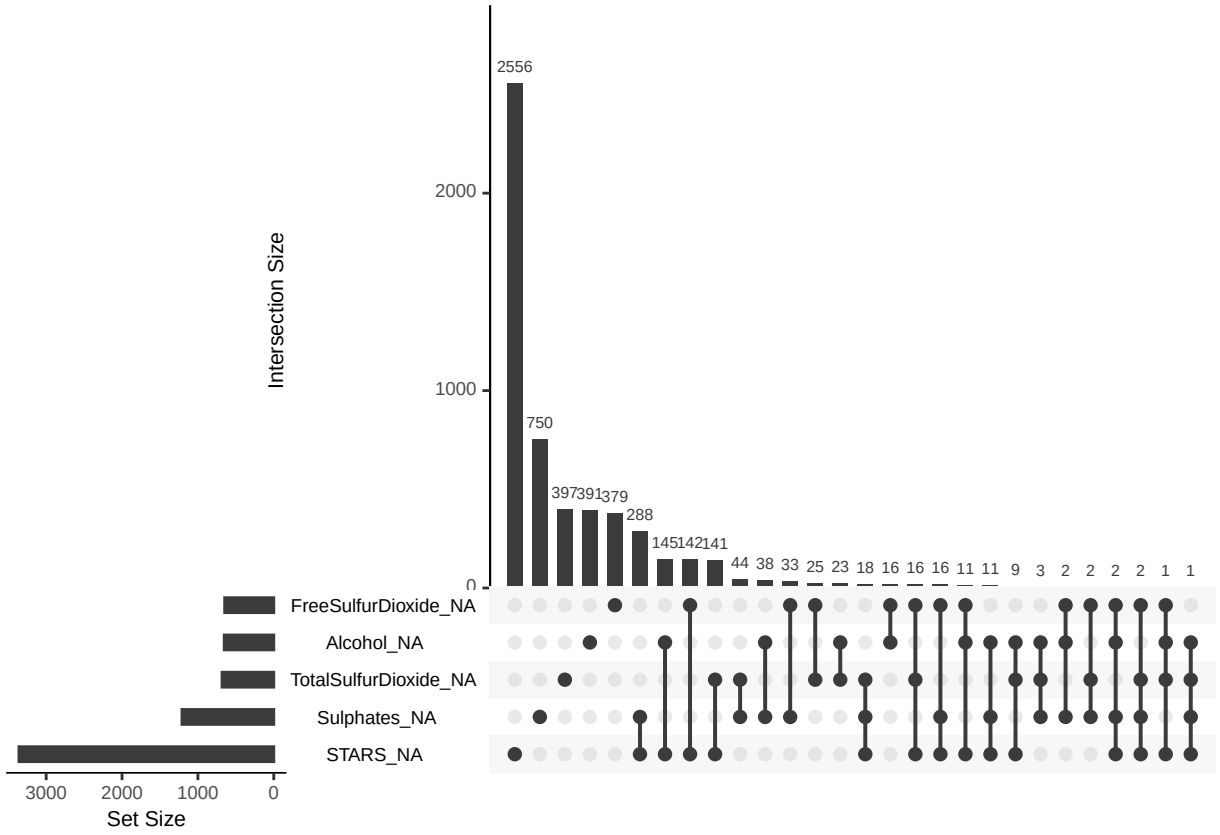
```
## Sulphates          -5.164e-03  7.051e-03  -0.732  0.46398
## Alcohol             3.948e-03  1.771e-03   2.229  0.02579 *
## LabelAppeal         1.771e-01  7.954e-03  22.271  < 2e-16 ***
## AcidIndex          -4.870e-02  5.903e-03  -8.251  < 2e-16 ***
## STARS               1.871e-01  7.487e-03  24.993  < 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: 5844.1  on 6435  degrees of freedom
## Residual deviance: 4009.1  on 6421  degrees of freedom
##   (6359 observations deleted due to missingness)
## AIC: 23172
##
## Number of Fisher Scoring iterations: 5
```

We can see that The residual deviance falls far short of the degrees of freedom, suggesting signifcant overdispersion. However, improvement in the model may help to account for more of the variance. Also, we are missing almost half the data due to missing values.

2. Data Preparation

A. Address Missing Values We consider the missing values.

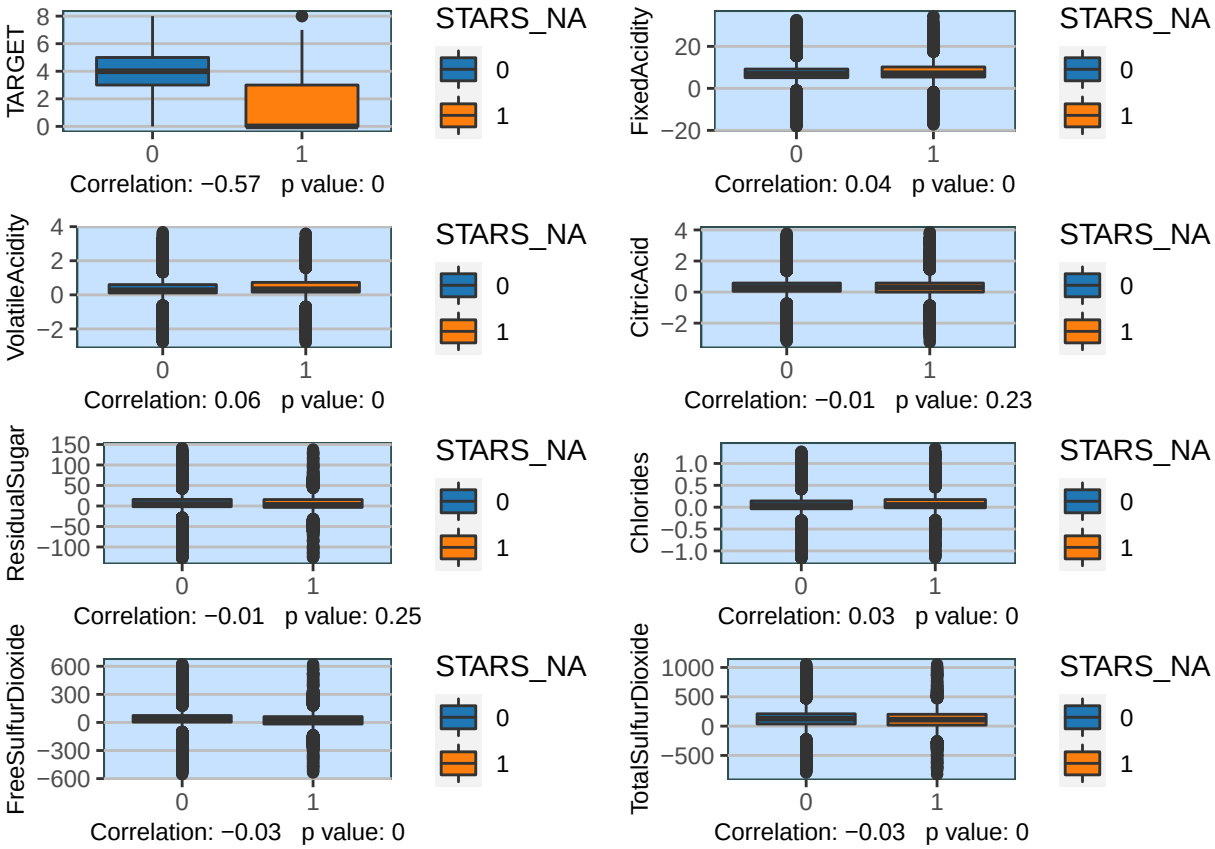


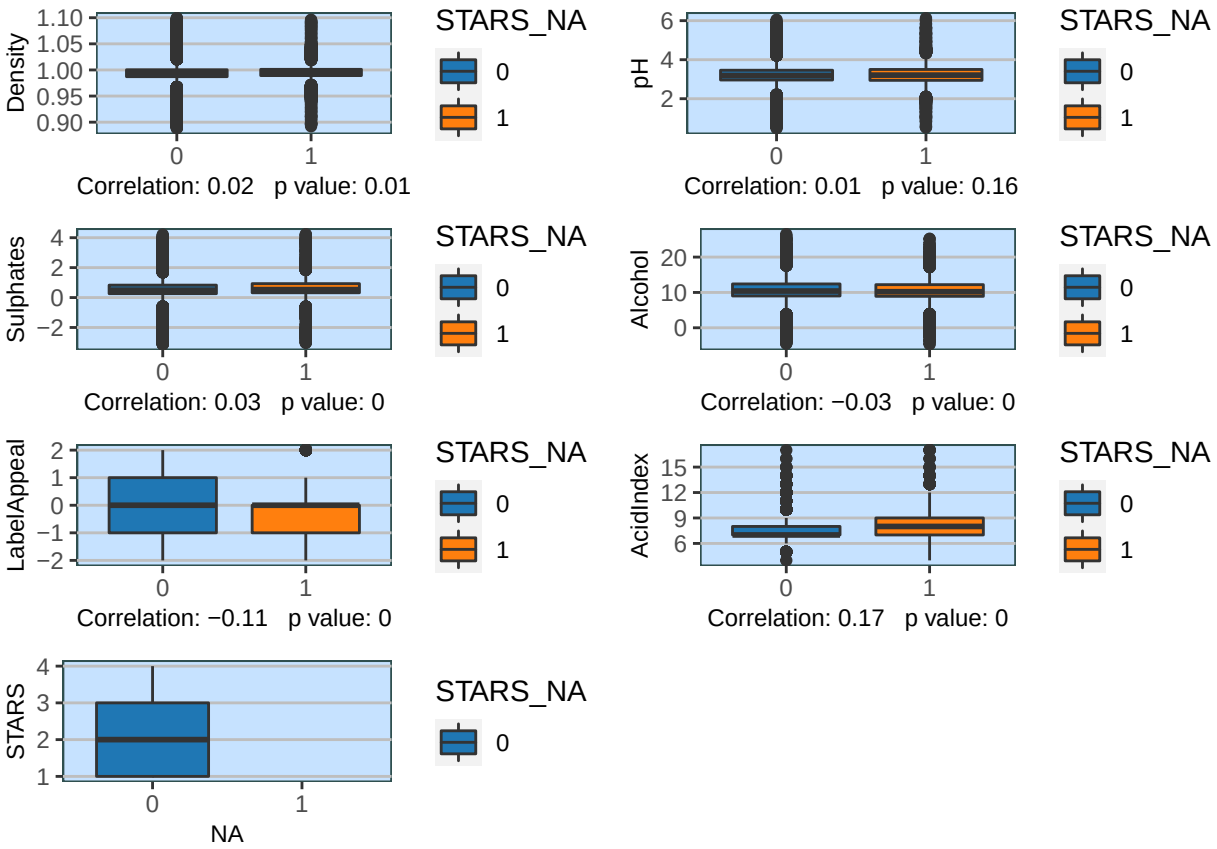


Over half of the records have missing values. Missings are confined to 8 variables. Interestingly, there is not a lot of overlap among the missings so we add these to the dataset.

With so many missing values, it's prudent to create flags to track what is missing.

We investigate the STARS missing flag, STARS_NA, further:


TARGET
8
6
4
2
0
0
1
Correlation: −0.57 p value: 0
STARS_NA
0
1
FixedAcidity
20
0
−20
Correlation: 0.04 p value: 0
VolatileAcidity
4
2
0
−2
Correlation: 0.06 p value: 0
CitricAcid
Correlation: −0.01 p value: 0.23
ResidualSugar
150
100
50
0
−50
−100
Correlation: −0.01 p value: 0.25
Chlorides
1.0
0.5
0.0
−0.5
−1.0
Correlation: 0.03 p value: 0
FreeSulfurDioxide
600
300
0
−300
−600
Correlation: −0.03 p value: 0
TotalSulfurDioxide
1000
500
0
−500
Correlation: −0.03 p value: 0



We can see how strongly Stars_na correlates with TARGET. Wines that have no stars tend to get bought at a far lower rate than those that do. The variance is also much higher.

We can see outliers at the top of the first boxplot - this represents two wines that were bought at the highest quantity despite having missing stars. We remove them from the analysis since they are clearly idiosyncratic anomalies.

We impute values for the NAs. Because the most significant of them are in the STARS category, a count category from 1 to 5, we will simply impute the median.

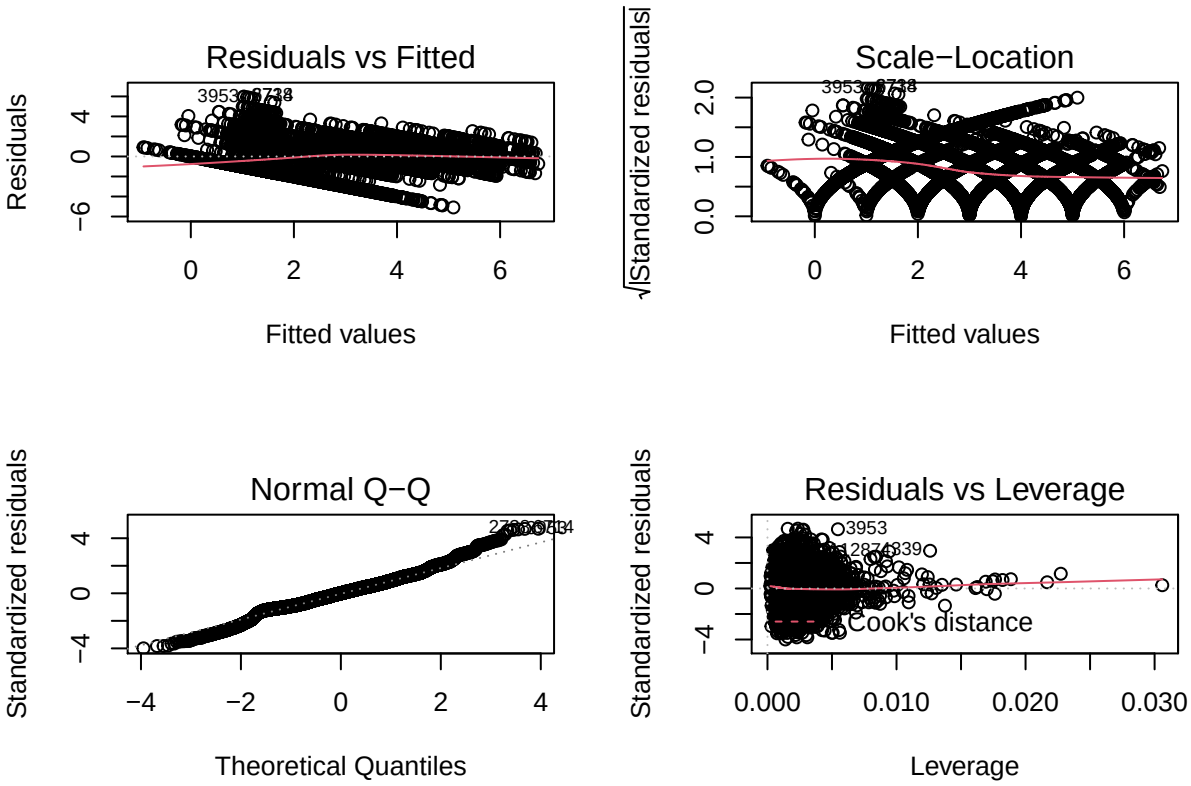
B. Transformations and interaction terms We do a log transformation of AcidIndex (logAcidIndex) to account for the skew. We also add two interaction terms that through analysis were discovered to be possible candidates - inter_Starsna_labelappeal and inter_stars_acidindex.

3. Build Models

A. Base Model We begin with a base model - OLS on all of the variables. We create two versions - one with all variables retained, and one with backward elimination to minimize the AIC.

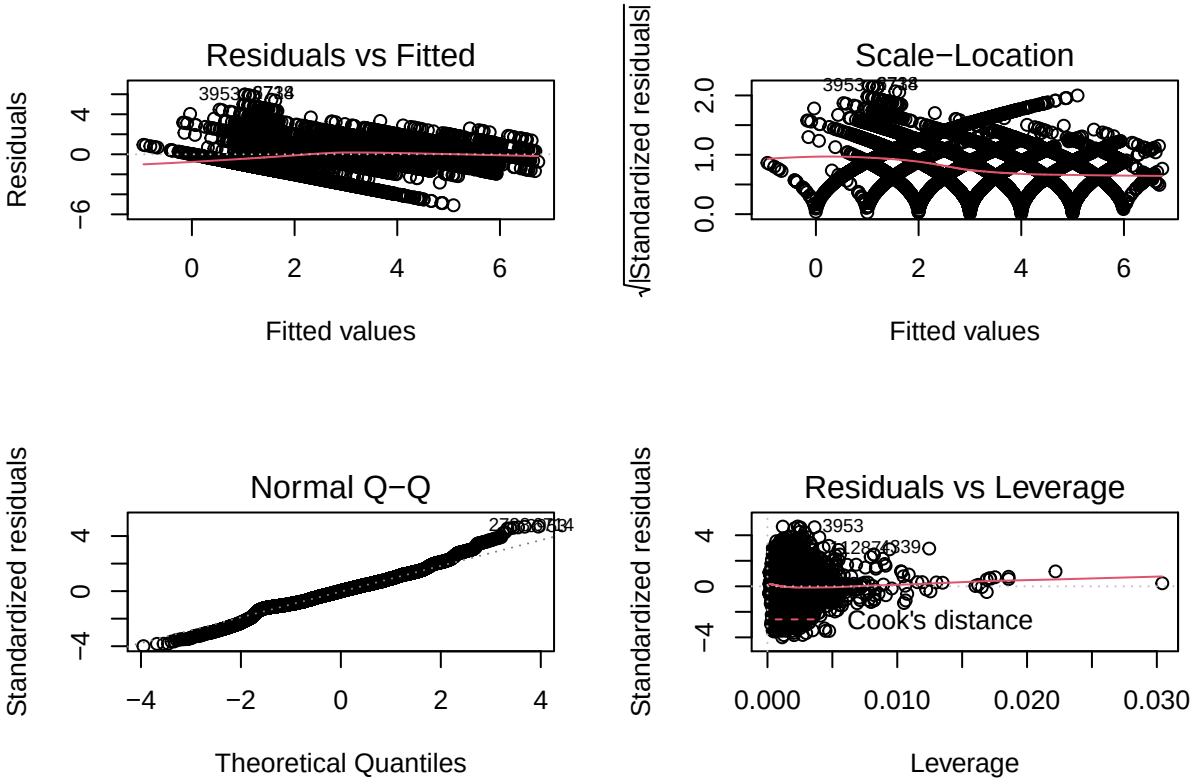
```
##
## Call:
## lm(formula = fla, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0938 -0.8494  0.0003  0.7571  5.9744
```

```
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.0282967  0.0112842 268.367  < 2e-16 ***
## FixedAcidity                 0.0046003  0.0114799   0.401 0.688624
## VolatileAcidity             -0.0721016  0.0113295  -6.364 2.03e-10 ***
## CitricAcid                   0.0125233  0.0113229   1.106 0.268743
## ResidualSugar                0.0064658  0.0113017   0.572 0.567254
## Chlorides                   -0.0338565  0.0113057  -2.995 0.002753 **
## FreeSulfurDioxide            0.0389938  0.0113082   3.448 0.000566 ***
## TotalSulfurDioxide           0.0492000  0.0113195   4.346 1.39e-05 ***
## Density                     -0.0233799  0.0113073  -2.068 0.038689 *
## pH                          -0.0207665  0.0113211  -1.834 0.066630 .
## Sulphates                   -0.0275255  0.0113056  -2.435 0.014918 *
## Alcohol                      0.0497245  0.0113348   4.387 1.16e-05 ***
## LabelAppeal                  0.5959203  0.0141790  42.028  < 2e-16 ***
## AcidIndex                   -0.6963492  0.0817212  -8.521  < 2e-16 ***
## STARS                       -0.1836521  0.0749295  -2.451 0.014259 *
## STARS_NA                    -1.0280170  0.0116853 -87.975  < 2e-16 ***
## Sulphates_NA                -0.0091637  0.0112937  -0.811 0.417153
## TotalSulpherDioxide_NA       0.0087213  0.0112944   0.772 0.440018
## Alcohol_NA                   0.0095086  0.0112904   0.842 0.399697
## FreeSulfurDioxide_NA         0.0111519  0.0112928   0.988 0.323403
## Chlorides_NA                -0.0003343  0.0112996  -0.030 0.976399
## ResidualSugar_NA             0.0160588  0.0112941   1.422 0.155086
## pH_NA                       -0.0171818  0.0112913  -1.522 0.128113
## inter_Starsna_labelappeal -0.3176037  0.0136943 -23.192  < 2e-16 ***
## inter_stars_acidindex        0.7917448  0.0793597   9.977  < 2e-16 ***
## logAcidIndex                 0.1187173  0.0781122   1.520 0.128578
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.276 on 12767 degrees of freedom
## Multiple R-squared:  0.5615, Adjusted R-squared:  0.5606
## F-statistic: 653.9 on 25 and 12767 DF,  p-value: < 2.2e-16
```




```
## NULL
## [1] "AIC:  42575.1862507901"

##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS + STARS_NA + ResidualSugar_NA +
##     pH_NA + inter_Starsna_labelappeal + inter_stars_acidindex +
##     logAcidIndex, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0955 -0.8473  0.0014  0.7604  5.9725
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.02830    0.01128 268.402  < 2e-16 ***
## VolatileAcidity            -0.07264    0.01132  -6.416 1.45e-10 ***
## Chlorides                  -0.03405    0.01130  -3.012 0.002597 **
## FreeSulfurDioxide           0.03929    0.01130   3.476 0.000511 ***
## TotalSulfurDioxide          0.04937    0.01131   4.364 1.29e-05 ***
## Density                    -0.02337    0.01130  -2.068 0.038691 *
## pH                         -0.02071    0.01132  -1.830 0.067227 .
## Sulphates                  -0.02736    0.01130  -2.422 0.015457 *
```

```
## Alcohol                     0.04986    0.01133   4.402 1.08e-05 ***
## LabelAppeal                 0.59595    0.01417  42.043  < 2e-16 ***
## AcidIndex                  -0.69667    0.08164  -8.534  < 2e-16 ***
## STARS                      -0.18413    0.07490  -2.458 0.013975 *
## STARS_NA                   -1.02812    0.01168 -88.043  < 2e-16 ***
## ResidualSugar_NA            0.01611    0.01129   1.427 0.153507
## pH_NA                      -0.01702    0.01129  -1.508 0.131700
## inter_Starsna_labelappeal -0.31804    0.01369 -23.234  < 2e-16 ***
## inter_stars_acidindex       0.79267    0.07933   9.992  < 2e-16 ***
## logAcidIndex                0.12038    0.07809   1.542 0.123205
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.276 on 12775 degrees of freedom
## Multiple R-squared:  0.5613, Adjusted R-squared:  0.5608
## F-statistic: 961.6 on 17 and 12775 DF,  p-value: < 2.2e-16
```




```
## NULL
## [1] "AIC:  42563.8267085174"
```

Many of the varaibles are significant, including both our interaction terms. The AIC of the inclusive model is 42575 while the AIC maxed model is 42563.

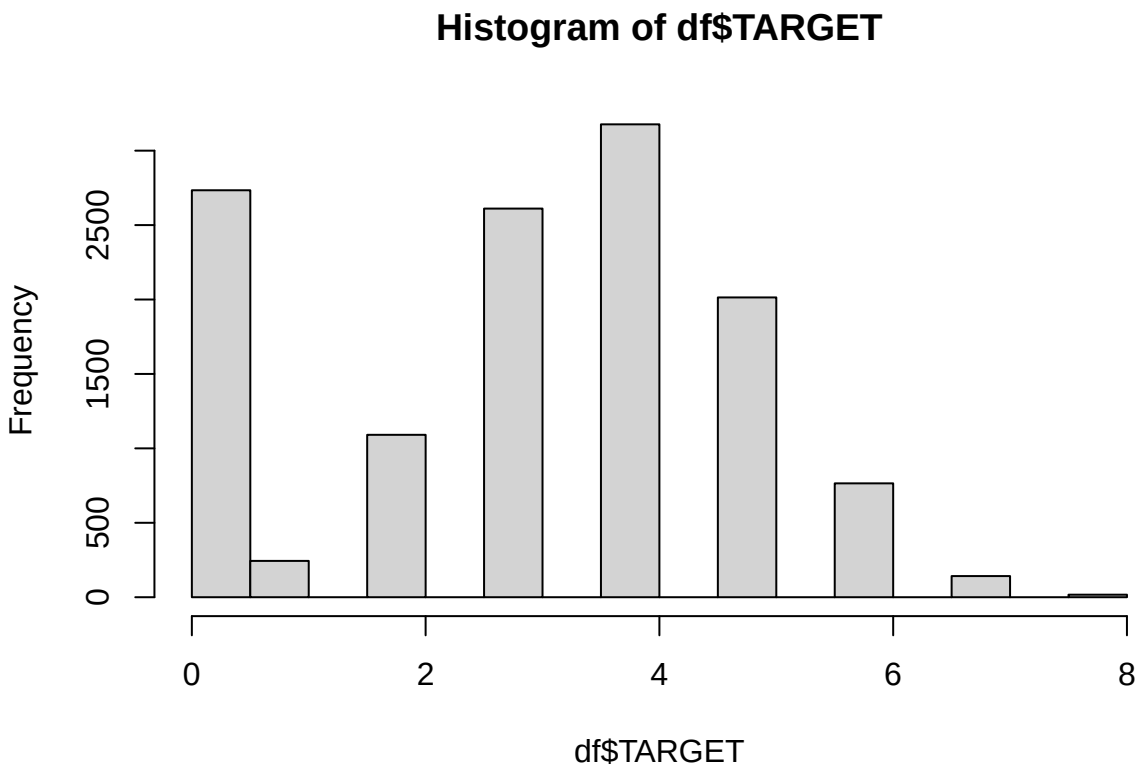
B. Poisson GLM Model

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = df7a)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2679  -0.6154  -0.0150   0.4266   3.8258
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 0.9679425  0.0059134 163.688  < 2e-16 ***
## FixedAcidity                0.0012265  0.0051745   0.237 0.812636
## VolatileAcidity            -0.0234855  0.0051237  -4.584 4.57e-06 ***
## CitricAcid                  0.0045104  0.0050798   0.888 0.374587
## ResidualSugar               0.0015635  0.0051037   0.306 0.759339
## Chlorides                  -0.0112118  0.0051193  -2.190 0.028518 *
## FreeSulfurDioxide           0.0135826  0.0050852   2.671 0.007562 **
## TotalSulfurDioxide          0.0171822  0.0051451   3.340 0.000839 ***
## Density                    -0.0086974  0.0050946  -1.707 0.087788 .
## pH                         -0.0078654  0.0051212  -1.536 0.124574
## Sulphates                  -0.0094033  0.0051037  -1.842 0.065410 .
## Alcohol                     0.0149330  0.0051230   2.915 0.003558 **
## LabelAppeal                 0.1643427  0.0058329  28.175  < 2e-16 ***
## AcidIndex                  -0.5120993  0.0481215 -10.642  < 2e-16 ***
## STARS                      -0.1700074  0.0316107  -5.378 7.53e-08 ***
## STARS_NA                   -0.4576224  0.0076481 -59.834  < 2e-16 ***
## Sulphates_NA               -0.0029582  0.0051421  -0.575 0.565102
## TotalSulpherDioxide_NA      0.0035830  0.0050445   0.710 0.477539
## Alcohol_NA                  0.0027557  0.0050748   0.543 0.587117
## FreeSulfurDioxide_NA        0.0034029  0.0050843   0.669 0.503310
## Chlorides_NA               -0.0008518  0.0050692  -0.168 0.866550
## ResidualSugar_NA            0.0050701  0.0050104   1.012 0.311576
## pH_NA                      -0.0067501  0.0051738  -1.305 0.191999
## inter_Starsna_labelappeal -0.0964410  0.0088294 -10.923  < 2e-16 ***
## inter_stars_acidindex       0.3354971  0.0338729   9.905  < 2e-16 ***
## logAcidIndex                0.2454432  0.0436868   5.618 1.93e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 22850  on 12792  degrees of freedom
## Residual deviance: 13469  on 12767  degrees of freedom
## AIC: 45455
##
## Number of Fisher Scoring iterations: 6
```

The deviance and residuals have come much closer together, close enough that we may use the poisson distribution rather than the negative binomial. In a work situation we might want to verify that the imputed missing values have not overly artificially reduced the variance, but we will not do that here.

The AIC has increased in this model. The poisson model by itself is not a better model than the base model.

C. Zero Inflated model Over 20% of the wines had zero cases bought, second only to 4 cases, with a 27% share of cases bought.



For this reason, a zero-inflated model may give us the best results.

```
##
## Call:
## zeroinfl(formula = TARGET ~ ., data = df7a, dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.30413 -0.41515 -0.01422  0.38094  6.38824
##
## Count model coefficients (poisson with log link):
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               1.2566690  0.0063997 196.364  < 2e-16 ***
## FixedAcidity              0.0018688  0.0053125   0.352    0.725
## VolatileAcidity          -0.0101436  0.0052641  -1.927    0.054 .
## CitricAcid                0.0013181  0.0051931   0.254    0.800
## ResidualSugar            -0.0017617  0.0052221  -0.337    0.736
## Chlorides                -0.0075648  0.0052524  -1.440    0.150
## FreeSulfurDioxide         0.0042884  0.0051799   0.828    0.408
## TotalSulfurDioxide       -0.0035331  0.0052453  -0.674    0.501
## Density                  -0.0075084  0.0052523  -1.430    0.153
## pH                        0.0029343  0.0052539   0.559    0.576
## Sulphates                 0.0001724  0.0052461   0.033    0.974
## Alcohol                   0.0251933  0.0052235   4.823 1.41e-06 ***
```

```
## LabelAppeal                0.1894484  0.0059677  31.746  < 2e-16 ***
## AcidIndex                 -0.0193548  0.0491256  -0.394    0.694
## STARS                      0.0332222  0.0330670   1.005    0.315
## STARS_NA                  -0.0697802  0.0081782  -8.532  < 2e-16 ***
## Sulphates_NA              -0.0011892  0.0052608  -0.226    0.821
## TotalSulpherDioxide_NA    -0.0005542  0.0051801  -0.107    0.915
## Alcohol_NA                 0.0006472  0.0051991   0.124    0.901
## FreeSulfurDioxide_NA       0.0018759  0.0051867   0.362    0.718
## Chlorides_NA               0.0006192  0.0051969   0.119    0.905
## ResidualSugar_NA           0.0049258  0.0051158   0.963    0.336
## pH_NA                     -0.0010609  0.0053083  -0.200    0.842
## inter_Starsna_labelappeal  0.0819719  0.0095532   8.581  < 2e-16 ***
## inter_stars_acidindex      0.0575474  0.0354591   1.623    0.105
## logAcidIndex              -0.0305925  0.0433916  -0.705    0.481
##
## Zero-inflation model coefficients (binomial with logit link):
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -4.153236   0.270076 -15.378  < 2e-16 ***
## FixedAcidity               0.001972   0.035014   0.056 0.955077
## VolatileAcidity            0.142489   0.034452   4.136 3.54e-05 ***
## CitricAcid                -0.020594   0.034525  -0.597 0.550840
## ResidualSugar             -0.037284   0.034267  -1.088 0.276576
## Chlorides                  0.032560   0.034239   0.951 0.341619
## FreeSulfurDioxide         -0.107751   0.035122  -3.068 0.002156 **
## TotalSulfurDioxide        -0.218712   0.034403  -6.357 2.05e-10 ***
## Density                    0.022489   0.035188   0.639 0.522738
## pH                         0.137646   0.034352   4.007 6.15e-05 ***
## Sulphates                  0.117461   0.034643   3.391 0.000697 ***
## Alcohol                    0.102863   0.034759   2.959 0.003083 **
## LabelAppeal                0.634393   0.069115   9.179  < 2e-16 ***
## AcidIndex                  1.459740   0.271441   5.378 7.54e-08 ***
## STARS                     -2.501154   0.448843  -5.572 2.51e-08 ***
## STARS_NA                   2.604195   0.155933  16.701  < 2e-16 ***
## Sulphates_NA               0.027979   0.033428   0.837 0.402598
## TotalSulpherDioxide_NA    -0.040783   0.035091  -1.162 0.245154
## Alcohol_NA                -0.029519   0.034148  -0.864 0.387341
## FreeSulfurDioxide_NA      -0.022604   0.033857  -0.668 0.504360
## Chlorides_NA               0.023712   0.035265   0.672 0.501329
## ResidualSugar_NA          -0.004684   0.034251  -0.137 0.891213
## pH_NA                      0.054680   0.033839   1.616 0.106119
## inter_Starsna_labelappeal  0.038714   0.044515   0.870 0.384470
## inter_stars_acidindex     -0.507472   0.346826  -1.463 0.143416
## logAcidIndex              -0.729104   0.268879  -2.712 0.006695 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 62
## Log-likelihood: -2.032e+04 on 52 Df

## [1] "AIC:  40747.8680869199"
```

In this model the AIC has fallen to 40747. This is the best model so far, even better than our optimized OLS model.

D. Optimized Zero Inflated model We can use backward elimination to see if we can decrease our AIC.

```
##
## Call:
## zeroinfl(formula = TARGET ~ ., data = df7b, dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.29768 -0.41650 -0.01853  0.38153  6.27907
##
## Count model coefficients (poisson with log link):
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.2567952  0.0063976 196.447  < 2e-16 ***
## VolatileAcidity           -0.0101709  0.0052585  -1.934   0.0531 .
## FreeSulfurDioxide          0.0042929  0.0051748   0.830   0.4068
## TotalSulfurDioxide        -0.0036457  0.0052417  -0.696   0.4867
## pH                         0.0031258  0.0052496   0.595   0.5516
## Sulphates                  0.0003422  0.0052435   0.065   0.9480
## Alcohol                    0.0254536  0.0052181   4.878 1.07e-06 ***
## LabelAppeal                0.1893531  0.0059657  31.740  < 2e-16 ***
## AcidIndex                 -0.0179753  0.0490830  -0.366   0.7142
## STARS                      0.0345365  0.0330526   1.045   0.2961
## STARS_NA                  -0.0699407  0.0081760  -8.554  < 2e-16 ***
## inter_Starsna_labelappeal  0.0819684  0.0095458   8.587  < 2e-16 ***
## inter_stars_acidindex      0.0563841  0.0354460   1.591   0.1117
## logAcidIndex              -0.0314879  0.0433379  -0.727   0.4675
##
## Zero-inflation model coefficients (binomial with logit link):
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -4.15456    0.26981 -15.398  < 2e-16 ***
## VolatileAcidity            0.14310    0.03433   4.168 3.08e-05 ***
## FreeSulfurDioxide         -0.11119    0.03509  -3.169 0.001529 **
## TotalSulfurDioxide        -0.22086    0.03436  -6.428 1.29e-10 ***
## pH                         0.13851    0.03428   4.040 5.34e-05 ***
## Sulphates                  0.11505    0.03453   3.331 0.000864 ***
## Alcohol                    0.10393    0.03474   2.992 0.002771 **
## LabelAppeal                0.63143    0.06897   9.155  < 2e-16 ***
## AcidIndex                  1.44750    0.27114   5.339 9.37e-08 ***
## STARS                     -2.52559    0.44858  -5.630 1.80e-08 ***
## STARS_NA                   2.60588    0.15573  16.734  < 2e-16 ***
## inter_Starsna_labelappeal  0.04176    0.04445   0.939 0.347525
## inter_stars_acidindex     -0.48548    0.34663  -1.401 0.161335
## logAcidIndex              -0.72284    0.26855  -2.692 0.007111 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 36
## Log-likelihood: -2.033e+04 on 28 Df

## [1] "AIC:  40715.2278588892"
```

Our final model has an AIC of 40715, an improvement over the unoptimized model.

Even though our interaction terms are not significant, the AIC climbs considerably when they are removed from the model. Therefore we retain them.

One issue is that some of our coefficients are no longer intuitive, particularly STARS and STARS_NA which are both the opposite of what they should be . However, with the presence of interaction terms involving these variables, their coefficients are not particularly stable. When the interaction terms are removed, the coefficents return to the direction we would expect.

4. Select Model

We choose our optimized zero-inflated poisson model, as it had the lowest AIC and makes the most common sense - it is better than OLS and standard Poisson for count data with a high proportion of zeroes.

The final step is to make predictions on the evaluation set:

```
##   predict(m, newdata = df7aEval, type = "response")
## 1                                          1.6314051
## 2                                          3.9492335
## 3                                          2.4109403
## 4                                          2.4147120
## 5                                          0.7775015
## 6                                          5.6363950
```

5. Conclusion

We examined 12795 records of wine purchases to create a predictive model of how many crates of wine would be bought based on several chemical chacteristics. Our best model was a zero-inflated poisson model, using backward elimination to minimize AIC. The analysis had first suggested a negative binomial modelwould be more appropriate, but dispersion was reduced considerable after feature engineering.

A few enhancements to the model increased accuracy and lowered AIC, including a log transformation and two interaction terms. The final model perfomred considerably better than the base model.