DATA-621 Project 5

Harpreet Shoker

19-Jul-2018

COVER PAGE

DATA621-Assignment-5

By - Harpreet Shoker

Date - 18-Jul-2018

University - City university of New York

Professor - Marcus Ellis

Abstract

In this homework assignment, we will explore, analyze and model a data set containing information on approximately 12795 commercially available wines using 16 variables. The variables are mostly related to the chemical properties of the wine being sold. The response variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales.

Our objective is to build a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine. To attain our objective, we will be following the below best practice steps and guidelines:

1 -Data Exploration

2 -Data Preparation

3 -Build Models

4 -Select Models

DATA EXPLORATION

Reading the wine training data set from github

```
train <- read.csv('https://raw.githubusercontent.com/Harpreet1984/DATA621/master/HW5/wine-training-data
train$ï..INDEX = NULL
head(train)
```

```
##   TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
## 1      3          3.2           1.160      -0.98          54.2    -0.567
## 2      3          4.5           0.160      -0.81          26.1    -0.425
## 3      5          7.1           2.640      -0.88          14.8     0.037
## 4      3          5.7           0.385       0.04          18.8    -0.425
## 5      4          8.0           0.330      -1.26           9.4        NA
## 6      0         11.3           0.320       0.59           2.2     0.556
##   FreeSulfurDioxide TotalSulfurDioxide Density   pH Sulphates Alcohol
## 1                NA                268 0.99280 3.33     -0.59     9.9
## 2                15               -327 1.02792 3.38      0.70      NA
## 3               214                142 0.99518 3.12      0.48    22.0
## 4                22                115 0.99640 2.24      1.83     6.2
## 5              -167                108 0.99457 3.12      1.77    13.7
## 6               -37                 15 0.99940 3.20      1.29    15.4
##   LabelAppeal AcidIndex STARS
## 1           0         8     2
## 2          -1         7     3
## 3          -1         8     3
## 4          -1         6     1
## 5           0         9     2
## 6           0        11    NA
```

```
test <- read.csv('https://raw.githubusercontent.com/Harpreet1984/DATA621/master/HW5/wine-evaluation-data
```

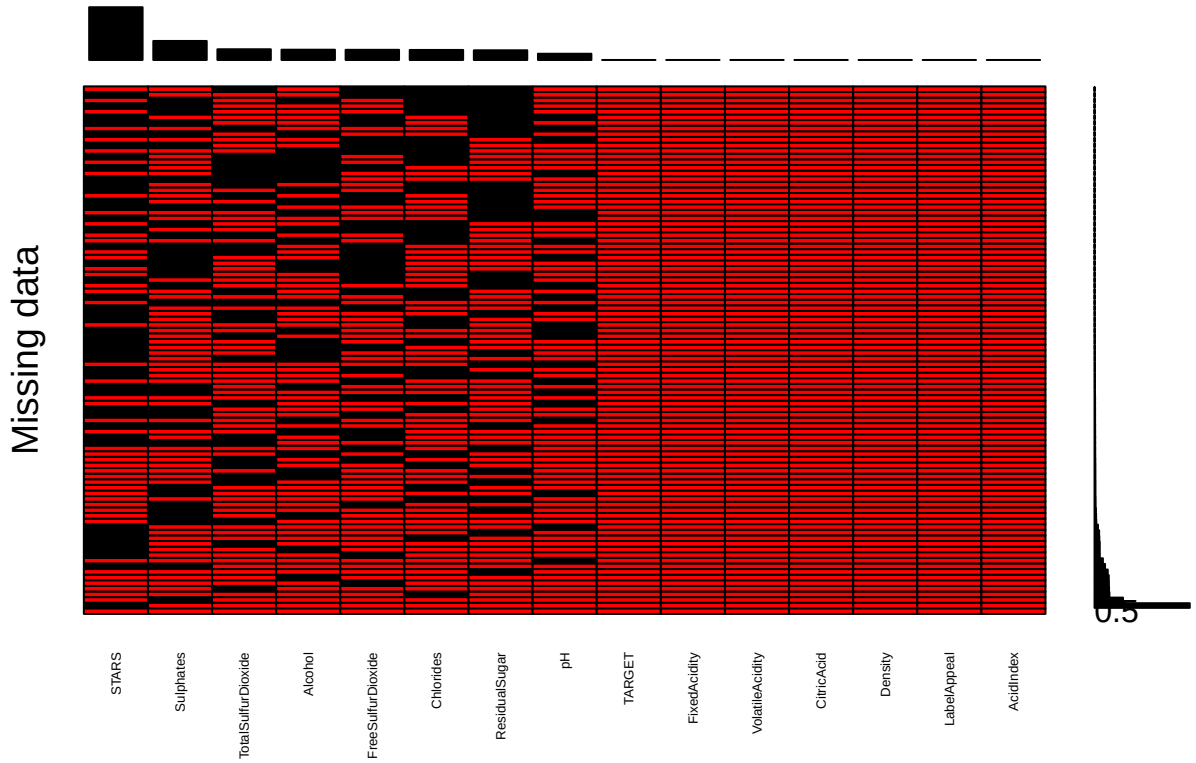
Only data from the training_set may be used The training_set contains| 12795 raw observations The training_set contains| 16 features (including index and targets)

```
summary(train)
```

```
##      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
```

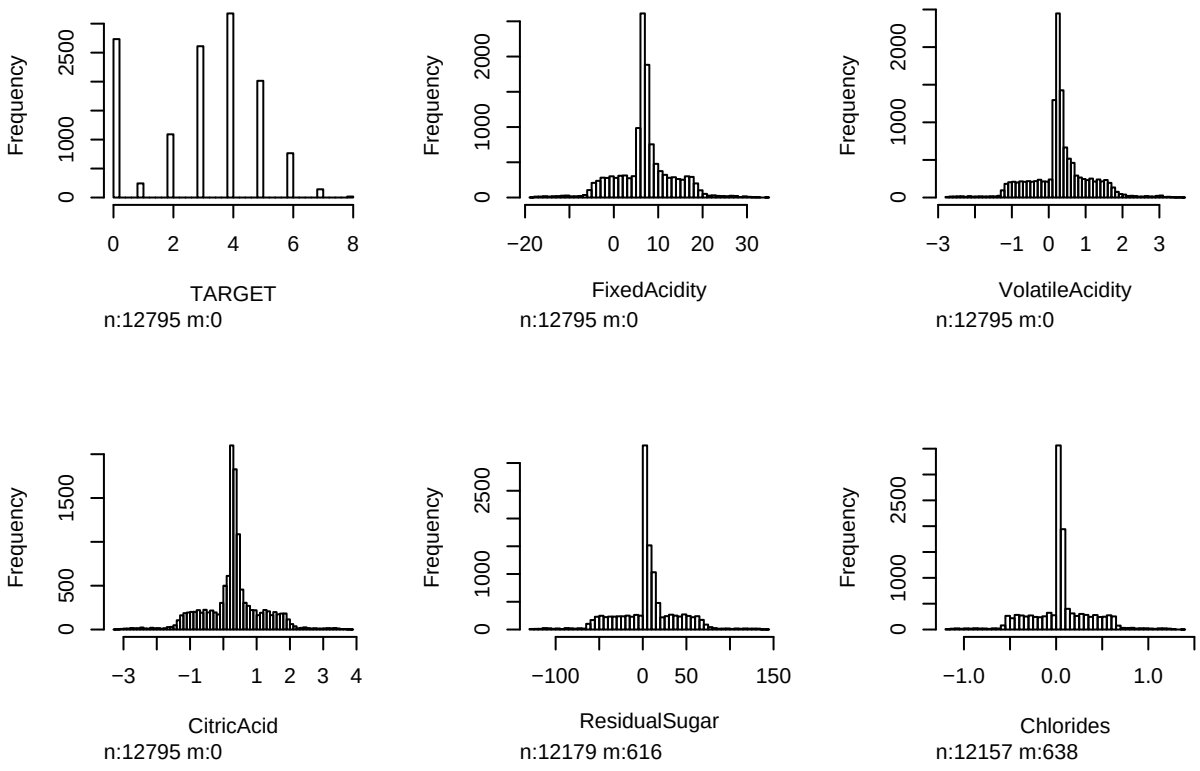
our dataset has almost all the means/medians match up. Our observation is that all 15 features other than AcidIndex have virtualy no skew, kurtosis, or standard error.Here we are not transforming any data, but before we move on to making models we should see if we can impute some data using the means of otherwise normalized features. The last feature of this dataset is that we're doing 'count' data, with a high zero count. (Meaning many people buy 0 crates of wine) This is problematic for both all three models we'll be making. I'll consider the use of zero-inflated models.

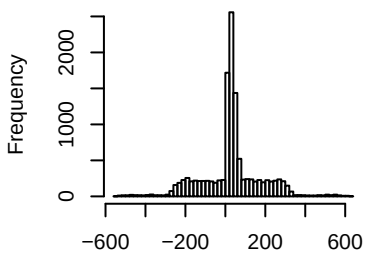



```
##
##  Variables sorted by number of missings:
##            Variable Count
##               STARS  3359
##           Sulphates  1210
##  TotalSulfurDioxide   682
##             Alcohol   653
##   FreeSulfurDioxide   647
##           Chlorides   638
##       ResidualSugar   616
##                  pH   395
##              TARGET     0
##        FixedAcidity     0
```

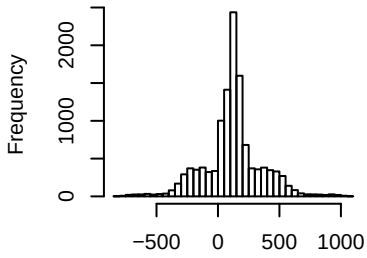
```
##     VolatileAcidity     0
##          CitricAcid     0
##             Density     0
##         LabelAppeal     0
##           AcidIndex     0
```

We see here in the above figure some missing data with almost a fourth of STARS missing.
Since our data has already been normalized, during imputation we will be using means to impute objective features and predicting STARS (being subjective, but correlated to objective qualities) through OLS of the other features.

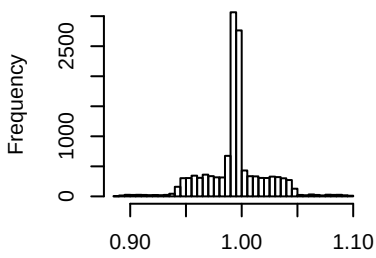




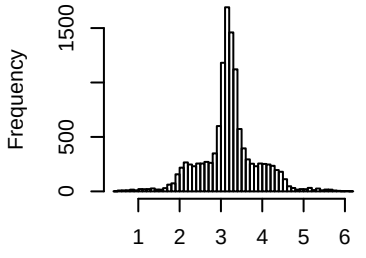
FreeSulfurDioxide
n:12148 m:647



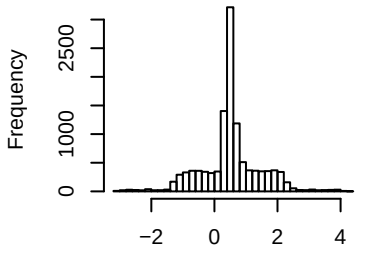
TotalSulfurDioxide
n:12113 m:682



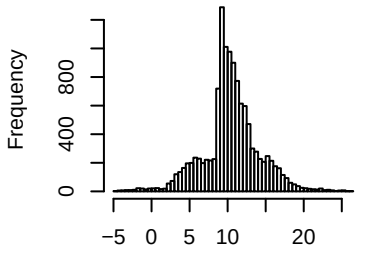
Density
n:12795 m:0



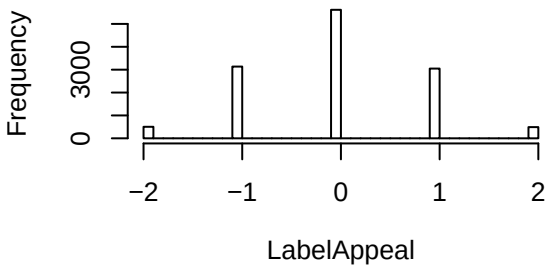
pH
n:12400 m:395



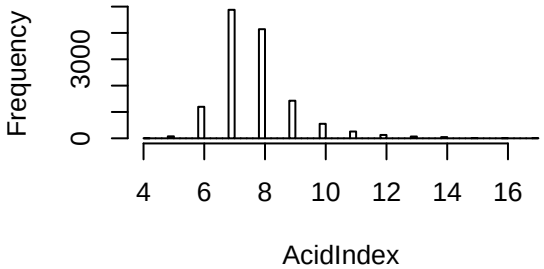
Sulphates
n:11585 m:1210



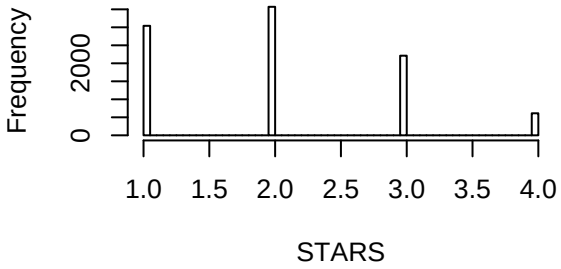
Alcohol
n:12142 m:653



n:12795 m:0

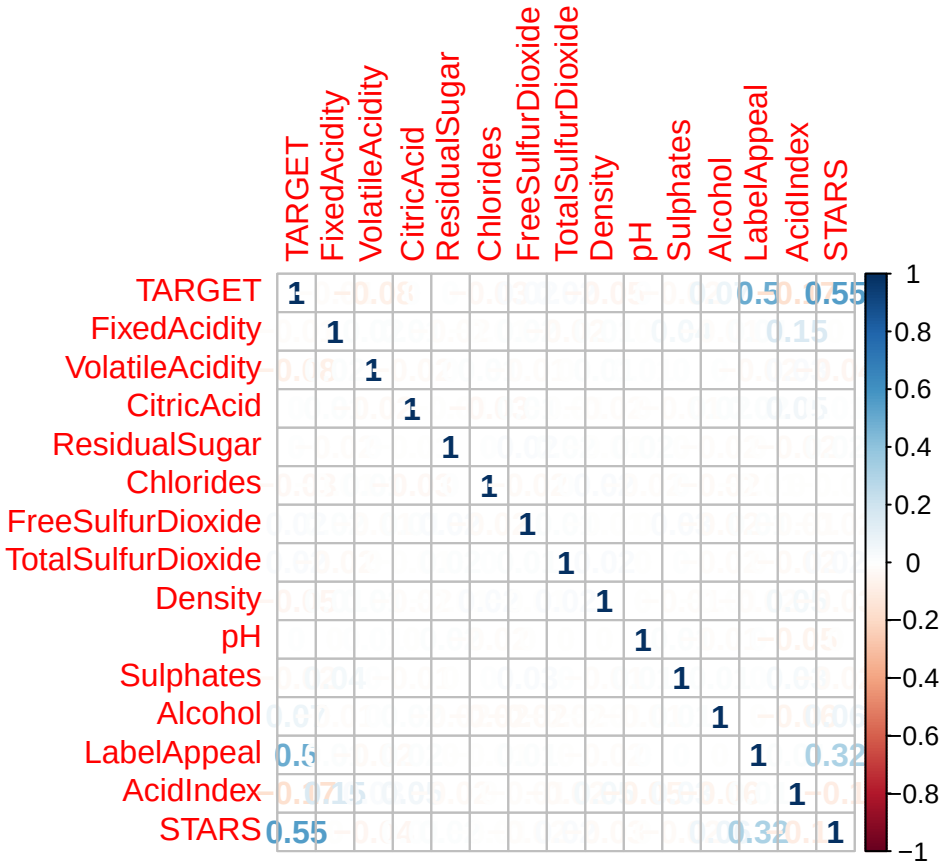


n:12795 m:0



n:9436 m:3359

.



Looking at correlation matrix to see what variables might make good predictive, parsimonious model.Wine sells if critics love it, the label is appealing, and its not too acidic.

DATA PREPARATION

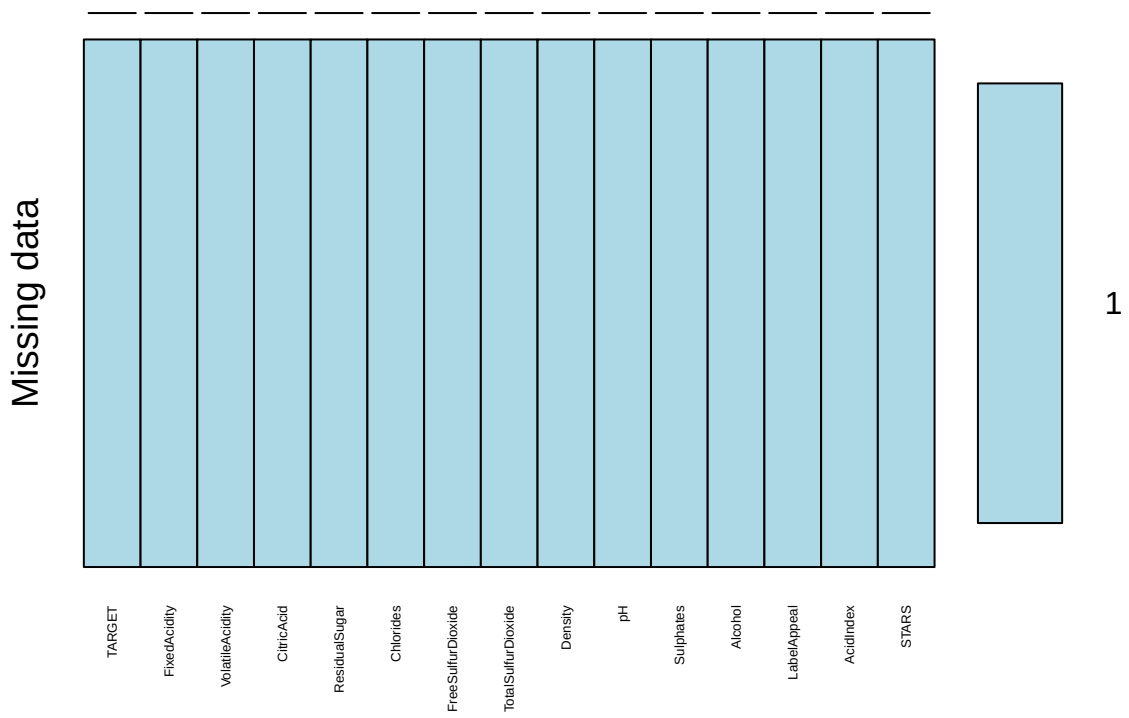
Since our data has normalised values, we are not doing transformations since the distributions are more or less normally distributed. Transforming our features into buckets or some type of ordinal data seems like needless data loss.

The results of data imputation + throwing out observations with missing values, we lost 2 out of almost 13k observations. We'll seperate some data before we train models, for cross validation later. Lets start making models

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

##
```

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




```
##
##  Variables sorted by number of missings:
##             Variable Count
##               TARGET     0
##        FixedAcidity     0
##     VolatileAcidity     0
##          CitricAcid     0
##       ResidualSugar     0
##           Chlorides     0
##   FreeSulfurDioxide     0
##  TotalSulfurDioxide     0
##             Density     0
##                  pH     0
##           Sulphates     0
```

```
##             Alcohol     0
##         LabelAppeal     0
##           AcidIndex     0
##               STARS     0
```

BUILD MODELS

Building these models here

1. Standara poisson Model
2. Zero - inflated poisson model - (target variable has a very high 0 occurence)
3. Zero - inflated negative Binomial
4. Linear Model

Let us first build a base line model with all variables to get better analyzing of the results

```
summary(pBase)
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = "poisson", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3455  -0.7024   0.1221   0.6250   2.3598
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.505e+00  2.522e-01   5.968 2.40e-09 ***
## FixedAcidity        -1.049e-03  1.056e-03  -0.994  0.32019
## VolatileAcidity     -4.347e-02  8.441e-03  -5.150 2.61e-07 ***
## CitricAcid           9.852e-03  7.655e-03   1.287  0.19810
## ResidualSugar       -1.178e-04  1.947e-04  -0.605  0.54513
## Chlorides           -5.986e-02  2.037e-02  -2.939  0.00329 **
## FreeSulfurDioxide    9.571e-05  4.391e-05   2.180  0.02927 *
## TotalSulfurDioxide   9.032e-05  2.856e-05   3.162  0.00156 **
## Density             -3.020e-01  2.474e-01  -1.221  0.22216
## pH                  -1.768e-02  9.658e-03  -1.830  0.06722 .
## Sulphates           -1.930e-02  7.161e-03  -2.695  0.00703 **
## Alcohol              9.895e-04  1.771e-03   0.559  0.57628
## LabelAppeal          1.442e-01  7.816e-03  18.448  < 2e-16 ***
## AcidIndex           -9.326e-02  5.777e-03 -16.142  < 2e-16 ***
## STARS                3.307e-01  6.941e-03  47.644  < 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: 13689.4  on 7678  degrees of freedom
## Residual deviance:  9523.1  on 7664  degrees of freedom
## AIC: 28724
##
## Number of Fisher Scoring iterations: 5
```

From the above results we see almost no correlation to our target variables in most variables

Model 1 - Standard poisson Model

parsimonious model of just STARS, LabelAppeal,and AcidIndex. . . Take out any variable deemed unfit and then run it through a zero-inflated poisson model for comparison.

```
summary(m1P)
```

```
##
## Call:
## glm(formula = TARGET ~ STARS + LabelAppeal + AcidIndex, family = "poisson",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3936  -0.7145   0.1540   0.6244   2.4888
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.155042   0.047838   24.14   <2e-16 ***
## STARS        0.333277   0.006909   48.24   <2e-16 ***
## LabelAppeal  0.144164   0.007811   18.46   <2e-16 ***
## AcidIndex   -0.095581   0.005665  -16.87   <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: 13689.4  on 7678  degrees of freedom
## Residual deviance:  9589.1  on 7675  degrees of freedom
## AIC: 28768
##
## Number of Fisher Scoring iterations: 5
```

Model - 2 Zero inflated Model

We'll make two zero-inflated models, with and without our lowly correlated VolatileAcidity variable

With highly correlated variabls

```
summary(m1ZIP)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity,
##     data = train, dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.30020 -0.41519  0.02999  0.44154  3.92847
##
## Count model coefficients (poisson with log link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.229570   0.051347  23.946  < 2e-16 ***
## STARS            0.111007   0.007898  14.055  < 2e-16 ***
## LabelAppeal      0.241204   0.008155  29.579  < 2e-16 ***
## AcidIndex       -0.022299   0.006201  -3.596 0.000323 ***
```

```
## VolatileAcidity -0.015789   0.008800  -1.794 0.072788 .
##
## Zero-inflation model coefficients (binomial with logit link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.81415    0.25413  -7.139 9.43e-13 ***
## STARS           -2.42199    0.08895 -27.227  < 2e-16 ***
## LabelAppeal      0.72321    0.05391  13.416  < 2e-16 ***
## AcidIndex        0.44698    0.03034  14.734  < 2e-16 ***
## VolatileAcidity  0.24800    0.05346   4.639 3.50e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 17
## Log-likelihood: -1.265e+04 on 10 Df
```

Without highly correlated variables

```
summary(m2ZIP)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ STARS + LabelAppeal + AcidIndex, data = train,
##     dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.28811 -0.42556  0.02245  0.46292  4.21900
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.226606   0.051292  23.914  < 2e-16 ***
## STARS        0.111199   0.007898  14.080  < 2e-16 ***
## LabelAppeal  0.241871   0.008154  29.664  < 2e-16 ***
## AcidIndex   -0.022607   0.006203  -3.645 0.000268 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.72692    0.25328  -6.818 9.21e-12 ***
## STARS       -2.42934    0.08922 -27.228  < 2e-16 ***
## LabelAppeal  0.72701    0.05402  13.458  < 2e-16 ***
## AcidIndex    0.44804    0.03038  14.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 15
## Log-likelihood: -1.266e+04 on 8 Df
```

From the coefficients we see its good to include VolatileAcidity in any Zero inflated model, as it definitely helps predict 0 counts. However the trade off is that it becomes a confounding variable while predicting the actual count above 0.

Lets build our negative binomial regression, then again with zero-inflation adjustment.

```
summary(nbBase)
```

```
##
## Call:
```

```
## MASS::glm.nb(formula = TARGET ~ ., data = train, init.theta = 48469.64451,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3454  -0.7024   0.1221   0.6250   2.3598
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.505e+00  2.522e-01   5.968 2.40e-09 ***
## FixedAcidity        -1.049e-03  1.056e-03  -0.994  0.32020
## VolatileAcidity     -4.347e-02  8.441e-03  -5.150 2.61e-07 ***
## CitricAcid           9.852e-03  7.655e-03   1.287  0.19811
## ResidualSugar       -1.178e-04  1.947e-04  -0.605  0.54515
## Chlorides           -5.986e-02  2.037e-02  -2.939  0.00329 **
## FreeSulfurDioxide    9.572e-05  4.391e-05   2.180  0.02928 *
## TotalSulfurDioxide   9.033e-05  2.856e-05   3.162  0.00156 **
## Density             -3.020e-01  2.474e-01  -1.221  0.22217
## pH                  -1.768e-02  9.658e-03  -1.830  0.06722 .
## Sulphates           -1.930e-02  7.162e-03  -2.695  0.00703 **
## Alcohol              9.896e-04  1.771e-03   0.559  0.57630
## LabelAppeal          1.442e-01  7.817e-03  18.447  < 2e-16 ***
## AcidIndex           -9.326e-02  5.777e-03 -16.142  < 2e-16 ***
## STARS                3.307e-01  6.941e-03  47.643  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48469.64) family taken to be 1)
##
##     Null deviance: 13688.8  on 7678  degrees of freedom
## Residual deviance:  9522.7  on 7664  degrees of freedom
## AIC: 28726
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48470
##           Std. Err.:  72613
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -28694.12
```

Same stats as the baseline poisson model, lets compare them using vuong test.

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A    p-value
## Raw                   18.98088 model1 > model2 < 2.22e-16
## AIC-corrected         18.98088 model1 > model2 < 2.22e-16
## BIC-corrected         18.98088 model1 > model2 < 2.22e-16
```

poisson model being better in this case; because I'm betting the data is overdispersed. Lets look.

```
## [1] 3.027347
```

```
## [1] 3
```

```
AER::dispersiontest(pBase,trafo =2)
```

```
##
##  Overdispersion test
##
## data:  pBase
## z = -17.744, p-value = 1
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##       alpha
## -0.07597519
```

the change may be a result of imputation methods we used.

Model 3 Zero inflated negative Binomial

negative binomial, zero inflated model using everything statistically significant and one with our most salient variables.

Lastly, if I assume that wholesalers buy wine according to customer preference. I imagine most customers like fancy looking, sweet tasting wines. So I'll make a minimal feature model including an interaction between alcohol (sweetness) and label appeal.

```
summary(m3ZInb)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ STARS + LabelAppeal + Alcohol:LabelAppeal,
##     data = train, dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.28333 -0.45537  0.00367  0.45516  5.91477
##
## Count model coefficients (negbin with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.051e+00  1.866e-02  56.331   <2e-16 ***
## STARS               1.125e-01  7.901e-03  14.244   <2e-16 ***
## LabelAppeal         2.372e-01  2.285e-02  10.384   <2e-16 ***
## LabelAppeal:Alcohol 4.724e-04  2.019e-03   0.234    0.815
## Log(theta)          1.550e+01  9.498e+00   1.632    0.103
##
## Zero-inflation model coefficients (binomial with logit link):
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.87990    0.10092  18.628  < 2e-16 ***
## STARS               -2.44895    0.08693 -28.172  < 2e-16 ***
## LabelAppeal          0.63498    0.14240   4.459 8.23e-06 ***
## LabelAppeal:Alcohol  0.01337    0.01290   1.037      0.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 5407745.3594
## Number of iterations in BFGS optimization: 39
## Log-likelihood: -1.28e+04 on 9 Df
```

```
summary(m1ZInb)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity +
##     Alcohol, data = train, dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.29938 -0.40787  0.03028  0.43743  3.92667
##
## Count model coefficients (negbin with log link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.175102   0.055361  21.226  < 2e-16 ***
## STARS            0.109510   0.007916  13.834  < 2e-16 ***
## LabelAppeal      0.241448   0.008158  29.598  < 2e-16 ***
## AcidIndex       -0.021411   0.006207  -3.450 0.000562 ***
## VolatileAcidity -0.016092   0.008799  -1.829 0.067422 .
## Alcohol          0.004824   0.001826   2.642 0.008243 **
## Log(theta)      17.738514         NA      NA       NA
##
## Zero-inflation model coefficients (binomial with logit link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.97405    0.28229  -6.993 2.69e-12 ***
## STARS           -2.41664    0.08866 -27.256  < 2e-16 ***
## LabelAppeal      0.72362    0.05389  13.427  < 2e-16 ***
## AcidIndex        0.44751    0.03031  14.766  < 2e-16 ***
## VolatileAcidity  0.24637    0.05337   4.616 3.90e-06 ***
## Alcohol          0.01463    0.01119   1.308    0.191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 50552037.539
## Number of iterations in BFGS optimization: 51
## Log-likelihood: -1.264e+04 on 13 Df
```

```
summary(m2ZInb)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ STARS + LabelAppeal + AcidIndex, data = train,
##     dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.28811 -0.42556  0.02245  0.46292  4.21901
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.226602   0.051292  23.914  < 2e-16 ***
## STARS        0.111199   0.007898  14.080  < 2e-16 ***
## LabelAppeal  0.241871   0.008154  29.664  < 2e-16 ***
## AcidIndex   -0.022607   0.006203  -3.645 0.000268 ***
## Log(theta)  16.199059         NA      NA       NA
```

```
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.72691    0.25328  -6.818 9.21e-12 ***
## STARS       -2.42934    0.08922 -27.228  < 2e-16 ***
## LabelAppeal  0.72701    0.05402  13.458  < 2e-16 ***
## AcidIndex    0.44804    0.03038  14.748  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 10843314.0342
## Number of iterations in BFGS optimization: 39
## Log-likelihood: -1.266e+04 on 9 Df
```

Model 4 Linear model

```
## Start:  AIC=5425.05
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +
##     Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Sum of Sq   RSS    AIC
## - ResidualSugar        1       1.1 15504 5423.6
## - FixedAcidity         1       2.0 15505 5424.0
## <none>                             15503 5425.0
## - Density              1       4.1 15508 5425.1
## - CitricAcid           1       4.7 15508 5425.4
## - pH                   1       5.6 15509 5425.8
## - Alcohol              1       8.1 15512 5427.1
## - FreeSulfurDioxide    1      11.3 15515 5428.6
## - Sulphates            1      17.9 15521 5431.9
## - TotalSulfurDioxide   1      21.4 15525 5433.6
## - Chlorides            1      25.3 15529 5435.5
## - VolatileAcidity      1      84.0 15588 5464.6
## - AcidIndex            1     678.6 16182 5752.0
## - LabelAppeal          1    1165.9 16669 5979.8
## - STARS                1    7309.8 22813 8389.3
##
## Step:  AIC=5423.57
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS
##
##                       Df Sum of Sq   RSS    AIC
## - FixedAcidity         1       1.9 15506 5422.5
## <none>                             15504 5423.6
## - Density              1       4.1 15509 5423.6
## - CitricAcid           1       4.7 15509 5423.9
## - pH                   1       5.6 15510 5424.3
## - Alcohol              1       8.2 15513 5425.6
## - FreeSulfurDioxide    1      11.1 15516 5427.1
## - Sulphates            1      17.8 15522 5430.4
## - TotalSulfurDioxide   1      21.2 15526 5432.1
## - Chlorides            1      25.2 15530 5434.0
## - VolatileAcidity      1      84.0 15588 5463.1
```

```
## - AcidIndex           1     678.6 16183 5750.5
## - LabelAppeal         1    1166.1 16671 5978.4
## - STARS               1    7309.2 22814 8387.4
##
## Step:  AIC=5422.52
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS
##
##                       Df Sum of Sq   RSS    AIC
## <none>                             15506 5422.5
## - Density              1       4.1 15510 5422.5
## - CitricAcid           1       4.6 15511 5422.8
## - pH                   1       5.6 15512 5423.3
## - Alcohol              1       8.3 15515 5424.6
## - FreeSulfurDioxide    1      11.0 15517 5426.0
## - Sulphates            1      18.1 15524 5429.5
## - TotalSulfurDioxide   1      21.4 15528 5431.1
## - Chlorides            1      25.0 15531 5432.9
## - VolatileAcidity      1      84.2 15591 5462.1
## - AcidIndex            1     713.7 16220 5766.0
## - LabelAppeal          1    1167.2 16674 5977.8
## - STARS                1    7309.0 22815 8386.0
```

```
summary(Linear_model)
```

```
##
## Call:
## lm(formula = TARGET ~ LabelAppeal + STARS, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0635 -0.8934  0.2322  0.9277  3.9042
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.97325    0.03669   26.52   <2e-16 ***
## LabelAppeal  0.42699    0.01945   21.96   <2e-16 ***
## STARS        1.12253    0.01783   62.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.464 on 7676 degrees of freedom
## Multiple R-squared:  0.4202, Adjusted R-squared:  0.4201
## F-statistic:  2782 on 2 and 7676 DF,  p-value: < 2.2e-16
```

```
summary(mLin1)
```

```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS, data = train)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -5.1659 -0.8790  0.1836  1.0060  4.1041
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.841e+00  6.197e-01   6.198 6.00e-10 ***
## VolatileAcidity     -1.342e-01  2.081e-02  -6.451 1.18e-10 ***
## CitricAcid           2.867e-02  1.895e-02   1.513 0.130275
## Chlorides           -1.777e-01  5.052e-02  -3.518 0.000437 ***
## FreeSulfurDioxide    2.539e-04  1.089e-04   2.332 0.019735 *
## TotalSulfurDioxide   2.284e-04  7.016e-05   3.255 0.001138 **
## Density             -8.661e-01  6.090e-01  -1.422 0.155010
## pH                  -3.952e-02  2.378e-02  -1.662 0.096617 .
## Sulphates           -5.241e-02  1.754e-02  -2.988 0.002819 **
## Alcohol              8.849e-03  4.374e-03   2.023 0.043086 *
## LabelAppeal          4.553e-01  1.895e-02  24.022  < 2e-16 ***
## AcidIndex           -2.336e-01  1.244e-02 -18.783  < 2e-16 ***
## STARS                1.057e+00  1.759e-02  60.112  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.422 on 7666 degrees of freedom
## Multiple R-squared:  0.4537, Adjusted R-squared:  0.4528
## F-statistic: 530.5 on 12 and 7666 DF,  p-value: < 2.2e-16
```

Only using STARS and LabelAppeal can compete with a stepwise backward regression of every variable within .3 adjusted R2. While the stepwise regression has a better F-statistic in this case,.

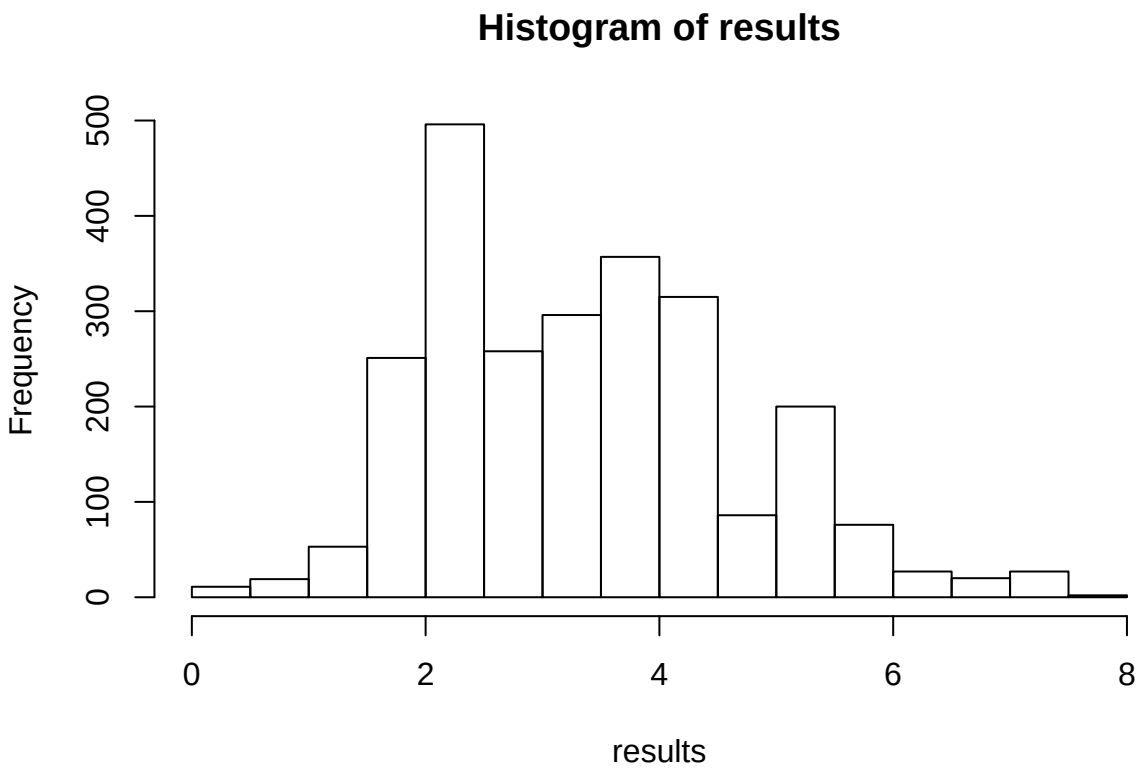
SELECT MODELS

Linear model with two variables LabelAppeal and STARS got a R2 of .42 However we would like a model that can delineate between will and won't sell. ZIP and ZINB were able to ascertain that we can best provide that information using acid index as well.

Testing the RMSE of our best ZIP model against the RMSE (while only using real predictions numbers (no decimals)) or our best (simplest) linear model we see that the simple linear model has a high RMSE at 1.44 to the ZIPS 1.35. I see no reason not to use AcidIndex, for a three variable ZIP model. Lets send in our predictions.

```
## [1] 1.441672 1.355328
```

```
results = predict(m2ZIP,test)
hist(results)
```




```
write.csv(results,'WinePredictions.csv')
```