Project 5

Michael Muller

April 30, 2018

Overview

The task

Create three pairs of models out of the wine-crates-sold dataset.

|Poisson Regression models| *Predicting the count of wine-crates sold for a given year assuming the results follow a poisson distribution*

|Negative binomial models| *Predicting the count of wine-crates sold for a given year assuming the results follow a negative binomal distribution*

|Multiple linear regression model | *Predicting the number of wine-crates sold for a given year assuming the results follow some degree of linear correlation*

The restraints

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)*

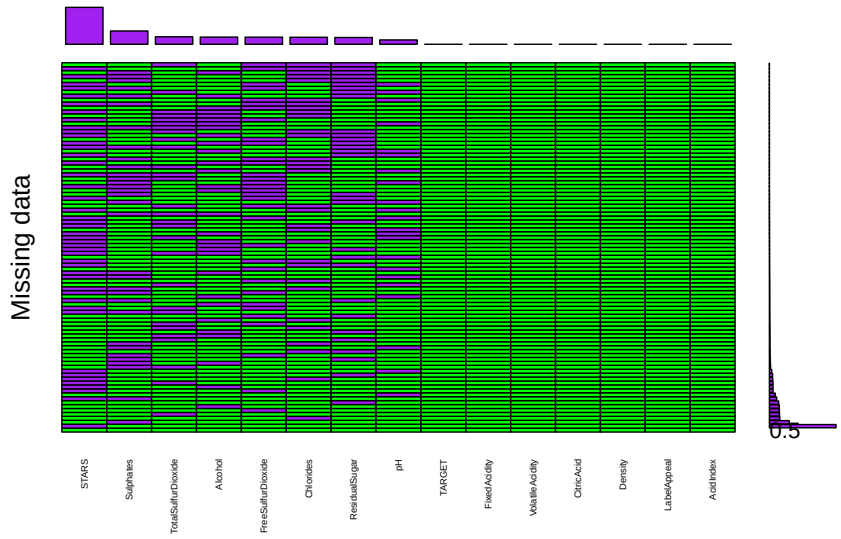
1. DATA EXPLORATION

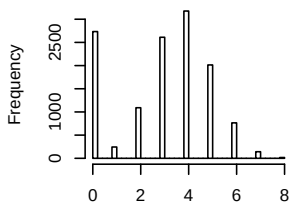
Taking a look into our dataset we see almost all the means/medians match up. We have some tough critics, with an average of 2 stars given, but the most impressive observation is that all 15 features other than AcidIndex have virtualy no skew, kurtosis, or standard error. I believe this data has been tampered with beforehand (looks to be normalized) so I won't transform any of the 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 salient 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-augmented and zero-inflated models.

Lets check out the data's sparsity using the MICE package in R.
Interpretting the figure below; we have quite a bit of missing data with almost a fourth of STARS missing. Since our data has already been normalized, during imputation I'll be using means to impute objective features and predicting STARS (being subjective, but correlated to objective qualities) through OLS of the other features.

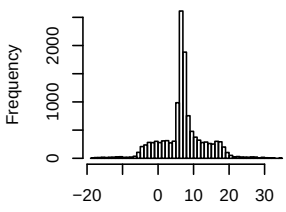
```
> library(psych)
> describe(train)
                   vars     n    mean      sd  median trimmed     mad     min      max    range  skew kurtosis    se
ï..INDEX              1 12795 8069.98 4656.91 8110.00 8071.03 5977.84    1.00 16129.00 16128.00  0.00    -1.20 41.17
TARGET                2 12795    3.03    1.93    3.00    3.05    1.48    0.00     8.00     8.00 -0.33    -0.88  0.02
FixedAcidity          3 12795    7.08    6.32    6.90    7.07    3.26  -18.10    34.40    52.50 -0.02     1.67  0.06
VolatileAcidity       4 12795    0.32    0.78    0.28    0.32    0.43   -2.79     3.68     6.47  0.02     1.83  0.01
CitricAcid            5 12795    0.31    0.86    0.31    0.31    0.42   -3.24     3.86     7.10 -0.05     1.84  0.01
ResidualSugar         6 12179    5.42   33.75    3.90    5.58   15.72 -127.80   141.15   268.95 -0.05     1.88  0.31
Chlorides             7 12157    0.05    0.32    0.05    0.05    0.13   -1.17     1.35     2.52  0.03     1.79  0.00
FreeSulfurDioxide     8 12148   30.85  148.71   30.00   30.93   56.34 -555.00   623.00  1178.00  0.01     1.84  1.35
TotalSulfurDioxide    9 12113  120.71  231.91  123.00  120.89  134.92 -823.00  1057.00  1880.00 -0.01     1.67  2.11
Density              10 12795    0.99    0.03    0.99    0.99    0.01    0.89     1.10     0.21 -0.02     1.90  0.00
pH                   11 12400    3.21    0.68    3.20    3.21    0.39    0.48     6.13     5.65  0.04     1.65  0.01
Sulphates            12 11585    0.53    0.93    0.50    0.53    0.44   -3.13     4.24     7.37  0.01     1.75  0.01
Alcohol              13 12142   10.49    3.73   10.40   10.50    2.37   -4.70    26.50    31.20 -0.03     1.54  0.03
LabelAppeal          14 12795   -0.01    0.89    0.00   -0.01    1.48   -2.00     2.00     4.00  0.01    -0.26  0.01
AcidIndex            15 12795    7.77    1.32    8.00    7.64    1.48    4.00    17.00    13.00  1.65     5.19  0.01
STARS                16  9436    2.04    0.90    2.00    1.97    1.48    1.00     4.00     3.00  0.45    -0.69  0.01
>
```

Figure 1: Data Description

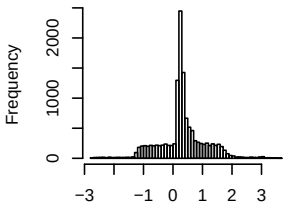




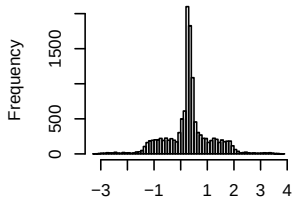
n:12795 m:0



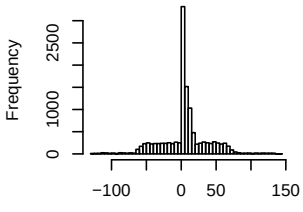
n:12795 m:0



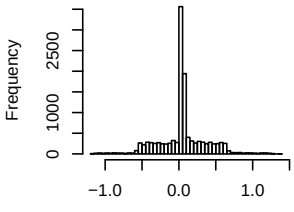
n:12795 m:0



n:12795 m:0

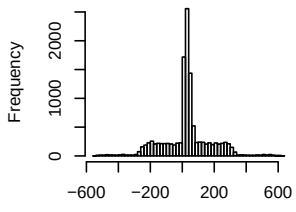


n:12179 m:616

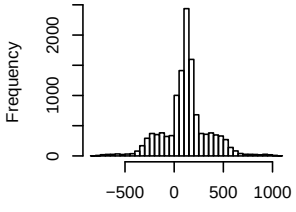


n:12157 m:638

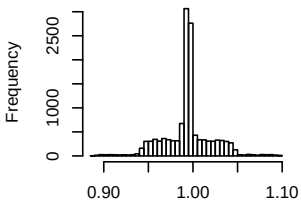
Continued



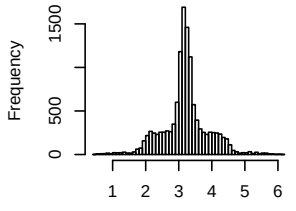
n:12148 m:647



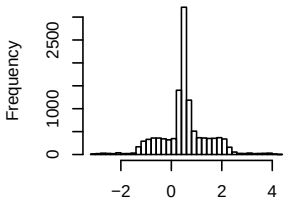
n:12113 m:682



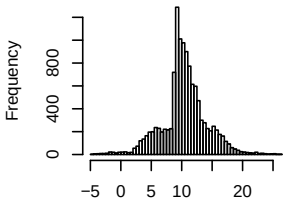
n:12795 m:0



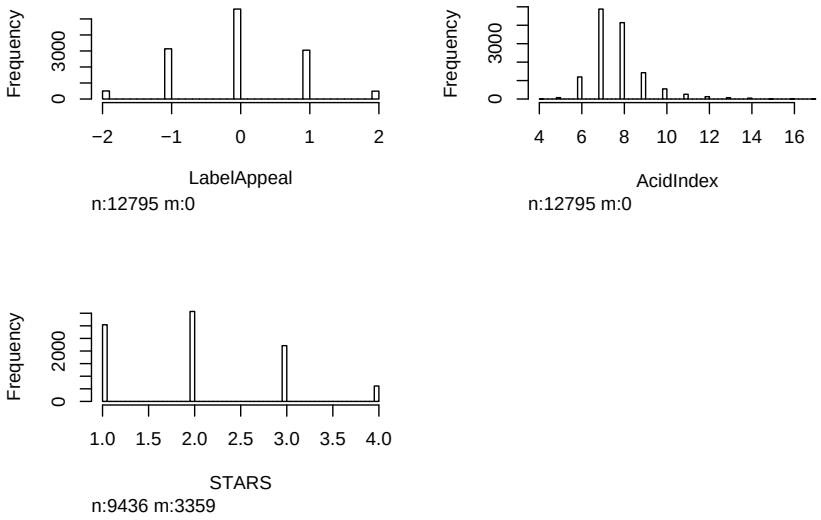
n:12400 m:395



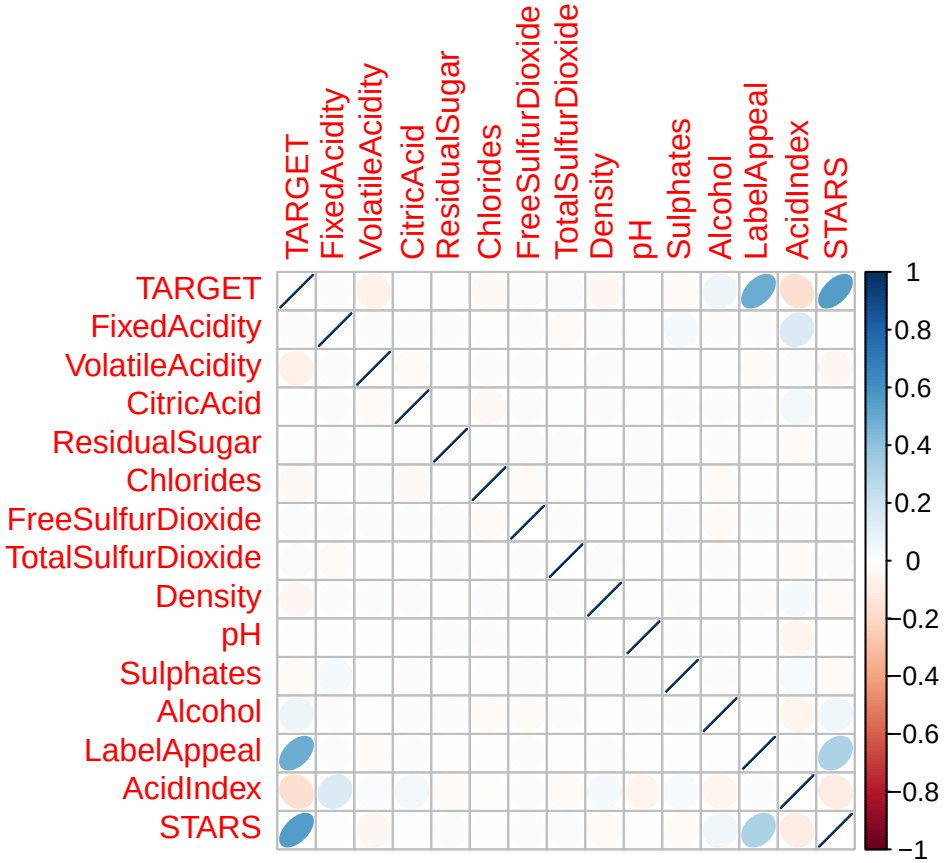
n:11585 m:1210



n:12142 m:653



Before we move on, lets check our correlation matrix to see what variables might make good predictive, parsimonious model. Not surprisingly, wine sells if critics love it, the label is appealing, and its not too acidic.

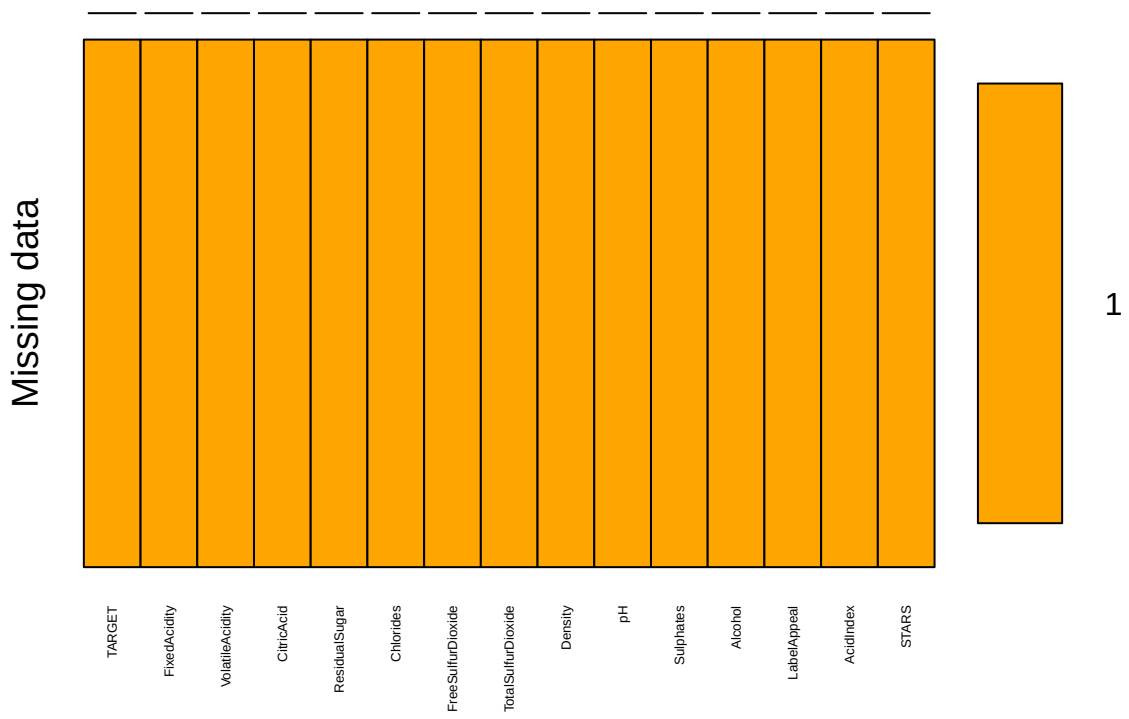


2. DATA PREPERATION

Again, since our data has been pre-tampered with, values normalized. I don't see it fit or wise to make anymore 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
```

3. BUILD MODELS

We'll start by building a standard poisson model, and build another zero-inflated poisson model because our target variable has a very high 0 occurence. ZIP regression is used to model count data that has an excess of zeroes. What ZIP does to account for the zeroes is it attempts to filter them through a logit model; and build answers only on those who pass the threshhold for 'above zero.'

First we'll build a base line model as always, with all the variables and analyze the results.

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Call:
## glm(formula = TARGET ~ ., family = "poisson", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3454  -0.7024   0.1222   0.6250   2.3599
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.506e+00  2.522e-01   5.970 2.38e-09 ***
## FixedAcidity        -1.049e-03  1.056e-03  -0.994  0.32018
## VolatileAcidity     -4.347e-02  8.441e-03  -5.150 2.61e-07 ***
## CitricAcid           9.851e-03  7.655e-03   1.287  0.19810
## ResidualSugar       -1.178e-04  1.947e-04  -0.605  0.54508
## Chlorides           -5.986e-02  2.037e-02  -2.939  0.00329 **
## FreeSulfurDioxide    9.572e-05  4.391e-05   2.180  0.02927 *
## TotalSulfurDioxide   9.033e-05  2.856e-05   3.163  0.00156 **
## Density             -3.023e-01  2.474e-01  -1.222  0.22166
## pH                  -1.768e-02  9.658e-03  -1.830  0.06719 .
## Sulphates           -1.930e-02  7.161e-03  -2.695  0.00703 **
## Alcohol              9.898e-04  1.771e-03   0.559  0.57619
## LabelAppeal          1.442e-01  7.816e-03  18.448  < 2e-16 ***
## AcidIndex           -9.326e-02  5.777e-03 -16.143  < 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
```

Seeing as we have almost no correlation to our TARGET in most variables, we can make our next model with a very 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.

```
##
## Call:
## glm(formula = TARGET ~ STARS + LabelAppeal + AcidIndex, family = "poisson",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -3.3935  -0.7145   0.1540   0.6244   2.4888
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.155090   0.047837   24.15   <2e-16 ***
## STARS        0.333274   0.006909   48.24   <2e-16 ***
## LabelAppeal  0.144165   0.007811   18.46   <2e-16 ***
## AcidIndex   -0.095587   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:  9589.2  on 7675  degrees of freedom
## AIC: 28768
##
## Number of Fisher Scoring iterations: 5
```

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

```
## Warning: package 'pscl' was built under R version 3.4.4
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity,
##     data = train, dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.3002 -0.4151  0.0300  0.4415  3.9283
##
## Count model coefficients (poisson with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.229588   0.051347  23.947  < 2e-16 ***
## STARS             0.111007   0.007898  14.055  < 2e-16 ***
## LabelAppeal       0.241204   0.008155  29.579  < 2e-16 ***
## AcidIndex        -0.022301   0.006201  -3.596 0.000323 ***
## VolatileAcidity -0.015789   0.008800  -1.794 0.072787 .
##
## Zero-inflation model coefficients (binomial with logit link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.81449    0.25413  -7.140 9.32e-13 ***
## STARS           -2.42194    0.08895 -27.227  < 2e-16 ***
## LabelAppeal      0.72319    0.05391  13.416  < 2e-16 ***
## AcidIndex        0.44702    0.03034  14.735  < 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
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ STARS + LabelAppeal + AcidIndex, data = train,
##     dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.28809 -0.42555  0.02245  0.46293  4.21877
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.226625   0.051292  23.914  < 2e-16 ***
## STARS        0.111198   0.007898  14.080  < 2e-16 ***
## LabelAppeal  0.241871   0.008154  29.664  < 2e-16 ***
## AcidIndex   -0.022609   0.006203  -3.645 0.000267 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.72726    0.25327   -6.82 9.12e-12 ***
## STARS       -2.42929    0.08922  -27.23  < 2e-16 ***
## LabelAppeal  0.72699    0.05402   13.46  < 2e-16 ***
## AcidIndex    0.44808    0.03038   14.75  < 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
```

Looking at our logit coefficients, it seems wise to include VolatileAcidity in any ZI 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.

```
## 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:
## MASS::glm.nb(formula = TARGET ~ ., data = train, init.theta = 48469.4802,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3453  -0.7024   0.1222   0.6250   2.3598
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.506e+00  2.522e-01   5.969 2.38e-09 ***
```

```
## FixedAcidity       -1.050e-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  0.19811
## ResidualSugar      -1.178e-04  1.947e-04  -0.605  0.54510
## Chlorides          -5.987e-02  2.037e-02  -2.939  0.00329 **
## FreeSulfurDioxide   9.572e-05  4.391e-05   2.180  0.02927 *
## TotalSulfurDioxide  9.033e-05  2.856e-05   3.162  0.00156 **
## Density            -3.023e-01  2.474e-01  -1.222  0.22168
## pH                 -1.768e-02  9.658e-03  -1.830  0.06719 .
## Sulphates          -1.930e-02  7.162e-03  -2.695  0.00703 **
## Alcohol             9.898e-04  1.771e-03   0.559  0.57621
## LabelAppeal         1.442e-01  7.817e-03  18.447  < 2e-16 ***
## AcidIndex          -9.326e-02  5.777e-03 -16.143  < 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(48469.48) 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:  48469
##           Std. Err.:  72613
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -28694.16
```

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                   18.98077 model1 > model2 < 2.22e-16
## AIC-corrected         18.98077 model1 > model2 < 2.22e-16
## BIC-corrected         18.98077 model1 > model2 < 2.22e-16
```

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

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

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

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

```
##       alpha
## -0.07597474
```

Guess I was wrong; this may be a result of my imputation methods. Anyway lets make a 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.

```
## Warning in sqrt(diag(vc)[np]): NaNs produced

## Warning in sqrt(diag(vc)[np]): NaNs produced
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ STARS + LabelAppeal + Alcohol:LabelAppeal,
##     data = train, dist = "negbin")
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.283221 -0.455356  0.003673  0.455159  5.915056
##
## Count model coefficients (negbin with log link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.051e+00  1.866e-02  56.332   <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 4.723e-04  2.019e-03   0.234    0.815
## Log(theta)          1.550e+01  1.069e+01   1.450    0.147
##
## Zero-inflation model coefficients (binomial with logit link):
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.87983    0.10091  18.628  < 2e-16 ***
## STARS               -2.44887    0.08693 -28.172  < 2e-16 ***
## LabelAppeal          0.63497    0.14239   4.459 8.22e-06 ***
## LabelAppeal:Alcohol  0.01337    0.01290   1.037      0.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 5404419.7178
## Number of iterations in BFGS optimization: 39
## Log-likelihood: -1.28e+04 on 9 Df
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ STARS + LabelAppeal + AcidIndex + VolatileAcidity +
##     Alcohol, data = train, dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.29937 -0.40781  0.03028  0.43744  3.92645
##
## Count model coefficients (negbin with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.175121   0.055361  21.226  < 2e-16 ***
## STARS            0.109509   0.007916  13.834  < 2e-16 ***
```

```
## LabelAppeal      0.241448   0.008158  29.598  < 2e-16 ***
## AcidIndex       -0.021413   0.006207  -3.450 0.000561 ***
## VolatileAcidity -0.016092   0.008799  -1.829 0.067425 .
## Alcohol          0.004824   0.001826   2.642 0.008243 **
## Log(theta)      18.922364         NA      NA       NA
##
## Zero-inflation model coefficients (binomial with logit link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.97442    0.28229  -6.994 2.66e-12 ***
## STARS           -2.41658    0.08866 -27.256  < 2e-16 ***
## LabelAppeal      0.72360    0.05389  13.427  < 2e-16 ***
## AcidIndex        0.44755    0.03031  14.767  < 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 = 165149892.8501
## Number of iterations in BFGS optimization: 51
## Log-likelihood: -1.264e+04 on 13 Df
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ STARS + LabelAppeal + AcidIndex, data = train,
##     dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.28809 -0.42555  0.02245  0.46293  4.21879
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.226621   0.051292  23.914  < 2e-16 ***
## STARS        0.111198   0.007898  14.080  < 2e-16 ***
## LabelAppeal  0.241871   0.008154  29.664  < 2e-16 ***
## AcidIndex   -0.022609   0.006203  -3.645 0.000267 ***
## Log(theta)  16.212402         NA      NA       NA
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.72725    0.25327   -6.82 9.12e-12 ***
## STARS       -2.42929    0.08922  -27.23  < 2e-16 ***
## LabelAppeal  0.72699    0.05402   13.46  < 2e-16 ***
## AcidIndex    0.44808    0.03038   14.75  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 10988965.3482
## Number of iterations in BFGS optimization: 39
## Log-likelihood: -1.266e+04 on 9 Df
```

I was completely wrong on my assumption of interaction. Lets move on to linear models.

```
## Start:  AIC=5425.08
## 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 15506 5424.1
## <none>                             15504 5425.1
## - 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 5428.7
## - Sulphates            1      17.9 15521 5431.9
## - TotalSulfurDioxide   1      21.4 15525 5433.7
## - Chlorides            1      25.3 15529 5435.6
## - VolatileAcidity      1      84.0 15588 5464.6
## - AcidIndex            1     678.7 16182 5752.1
## - LabelAppeal          1    1165.9 16669 5979.9
## - STARS                1    7309.7 22813 8389.3
##
## Step:  AIC=5423.6
## 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 5422.6
## <none>                             15504 5423.6
## - Density              1       4.1 15509 5423.6
## - CitricAcid           1       4.7 15509 5423.9
## - pH                   1       5.6 15510 5424.4
## - Alcohol              1       8.2 15513 5425.7
## - 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 5434.1
## - VolatileAcidity      1      84.0 15589 5463.1
## - AcidIndex            1     678.8 16183 5750.6
## - LabelAppeal          1    1166.1 16671 5978.4
## - STARS                1    7309.1 22814 8387.4
##
## Step:  AIC=5422.55
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS
##
##                       Df Sum of Sq   RSS    AIC
## <none>                             15506 5422.6
## - Density              1       4.1 15511 5422.6
## - CitricAcid           1       4.6 15511 5422.8
## - pH                   1       5.6 15512 5423.3
## - Alcohol              1       8.3 15515 5424.7
## - FreeSulfurDioxide    1      11.0 15518 5426.0
## - Sulphates            1      18.1 15524 5429.5
```

```
## - TotalSulfurDioxide  1      21.4 15528 5431.2
## - Chlorides           1      25.0 15532 5432.9
## - VolatileAcidity     1      84.2 15591 5462.1
## - AcidIndex           1     713.8 16220 5766.1
## - LabelAppeal         1    1167.2 16674 5977.9
## - STARS               1    7309.0 22815 8386.0
```

```
##
## Call:
## lm(formula = TARGET ~ LabelAppeal + STARS, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0636 -0.8934  0.2322  0.9277  3.9042
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.97326    0.03669   26.52   <2e-16 ***
## LabelAppeal  0.42699    0.01945   21.96   <2e-16 ***
## STARS        1.12252    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
```

```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1658 -0.8791  0.1835  1.0061  4.1042
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.843e+00  6.197e-01   6.201 5.89e-10 ***
## VolatileAcidity     -1.342e-01  2.081e-02  -6.451 1.18e-10 ***
## CitricAcid           2.867e-02  1.895e-02   1.513 0.130291
## Chlorides           -1.777e-01  5.052e-02  -3.519 0.000436 ***
## FreeSulfurDioxide    2.539e-04  1.089e-04   2.332 0.019729 *
## TotalSulfurDioxide   2.284e-04  7.016e-05   3.255 0.001138 **
## Density             -8.676e-01  6.090e-01  -1.425 0.154294
## pH                  -3.953e-02  2.378e-02  -1.662 0.096485 .
## Sulphates           -5.241e-02  1.754e-02  -2.988 0.002820 **
## Alcohol              8.849e-03  4.374e-03   2.023 0.043083 *
## LabelAppeal          4.553e-01  1.895e-02  24.022  < 2e-16 ***
## AcidIndex           -2.336e-01  1.244e-02 -18.785  < 2e-16 ***
## STARS                1.057e+00  1.759e-02  60.111  < 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
```

Interesting to see just 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, I want to make sure I'm not overfitting a model with too many variables.

4. SELECT MODELS

I like that a two variable linear model can get a R2 of .42, however I would like a model that can delineate between will and won't sell. Previously our 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 1.355353
```