Wine

Daniel Brooks (daniel.brooks@spsmail.cuny.edu), Daniel Fanelli (daniel.fanelli@spsmail.cuny.edu), Christopher Fenton (christopher.fenton@spsmail.cuny.edu), James Hamski (james.hamski@spsmail.cuny.edu), Youqing Xiang (youqing.xiang@spsmail.cuny.edu)

7/11/2016

The purpose of this analysis is to develop models to predict the number of cases of wine samples a large wine manufacturer should offer to distributors to maximize wine sales.

Our data shows the chemical properties of commercially available wines as well as factors such as STARS ratings.

Our response variable is the number of sample cases purchased by distribution companies after sampling, a variable that has a direct correlation to overall wine sales. These cases are 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.

If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales.

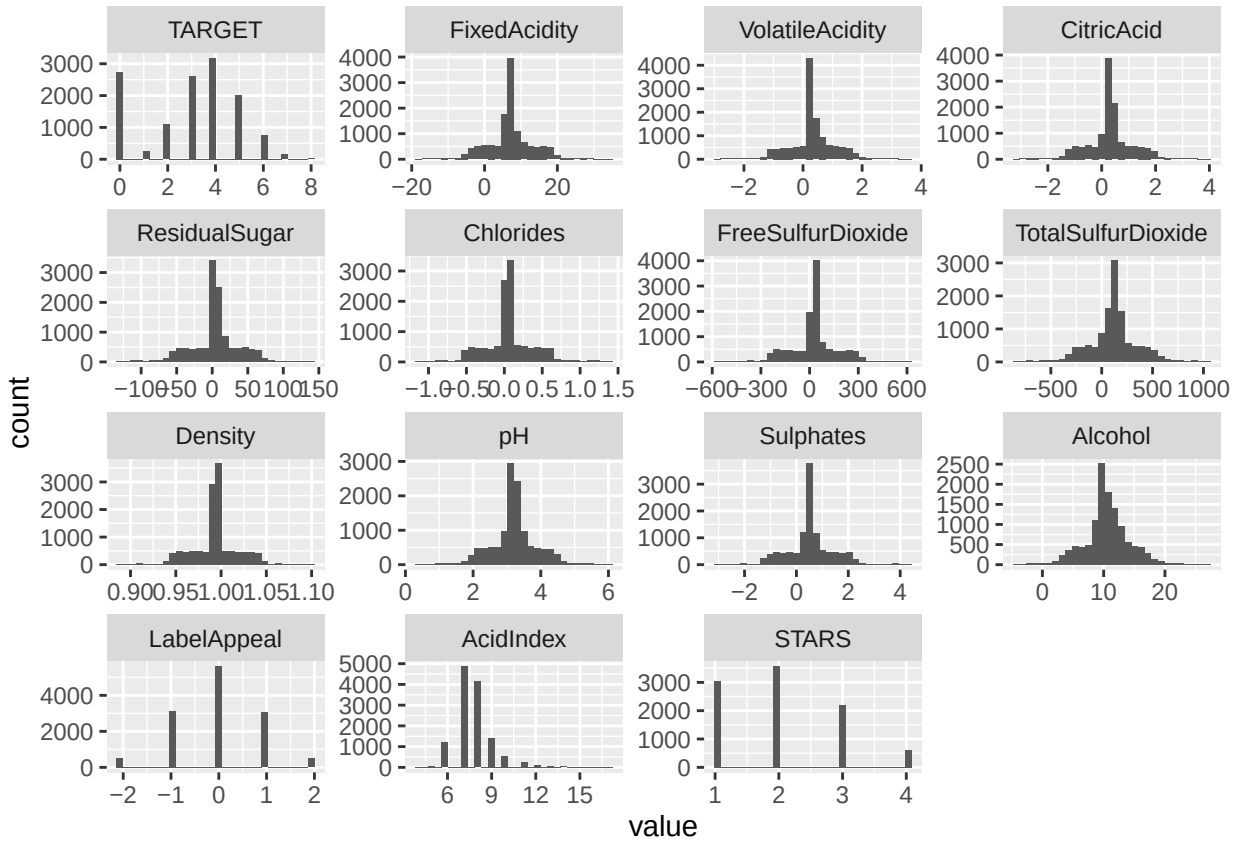
Our training dataset includes information on 12795 wines. Each wine has 14 potential predictor variables, and 1 response variable. The response variable is "TARGET", which is the number of cases purchased.

1) Data Exploration

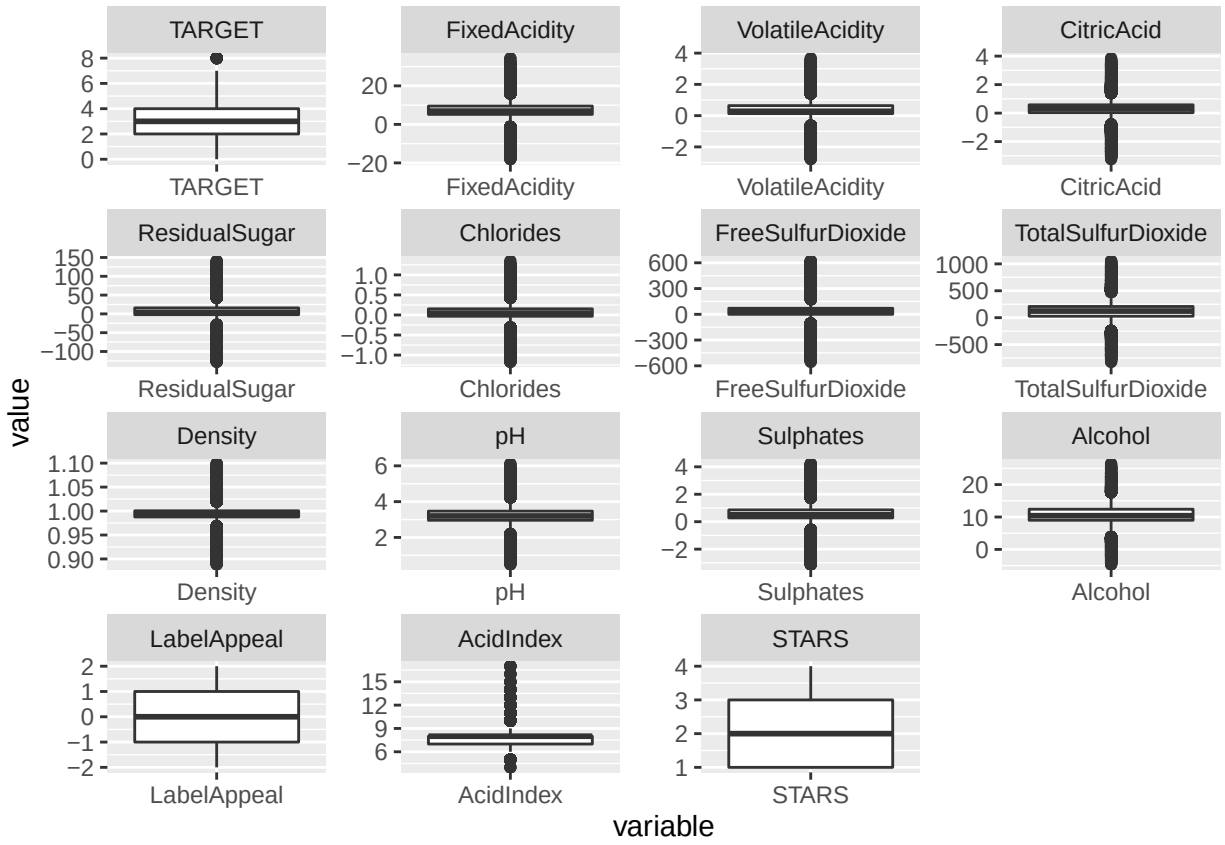
A basic analysis of the numerical variables is below:

- Our histograms show normally distributed variables
- Our Box-Plots seem to have large amounts of instances outside of the 2nd and 3rd quartiles. We will examine this further.

WINE Data Histograms

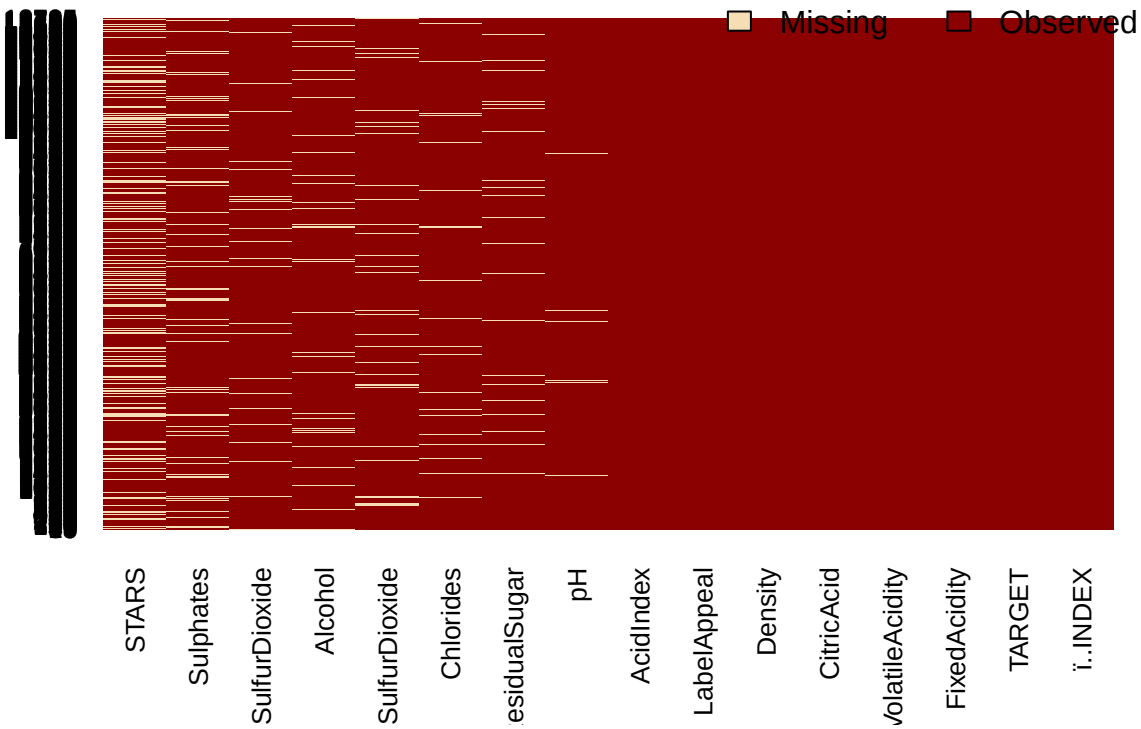


WINE Data BoxPlots

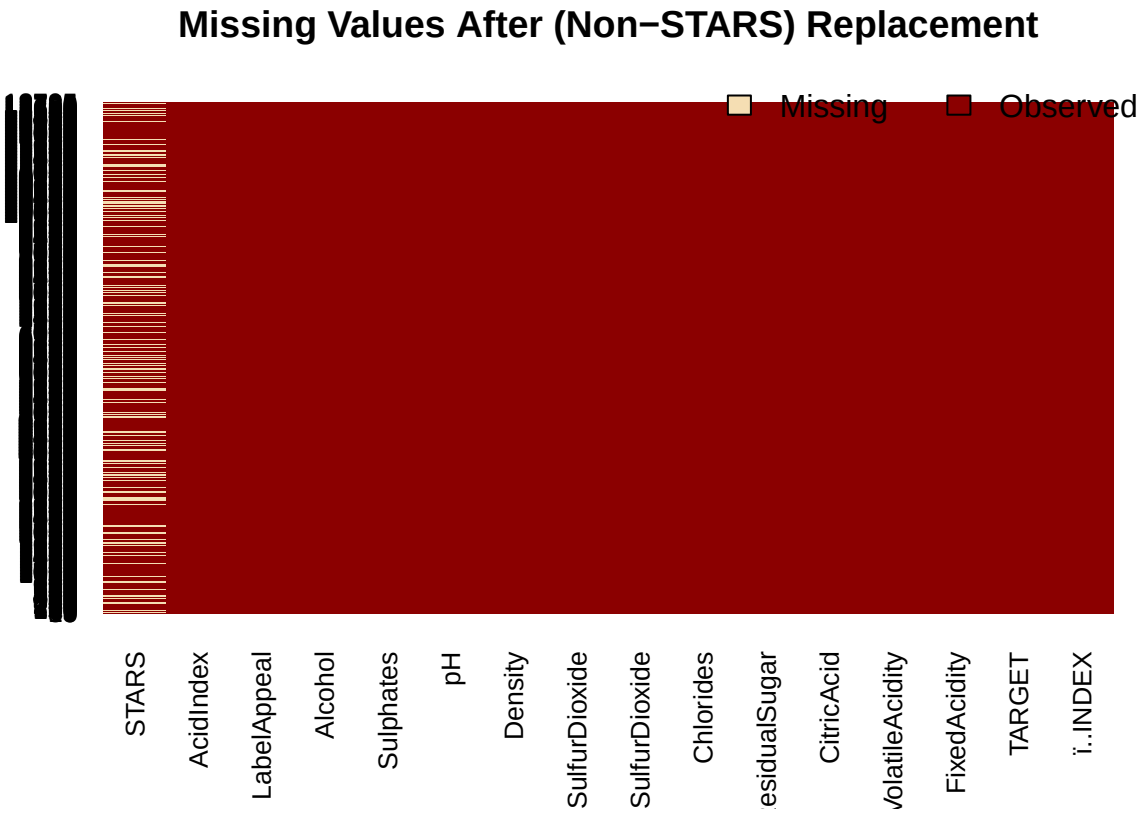


Explore NA's:

The below table shows a summary of the NA values in the data. Only STARS had an NA frequency higher than 10%, so this was a concern. All NA values were thus replaced with samples from their respective collections, except for STARS, which required further analysis.

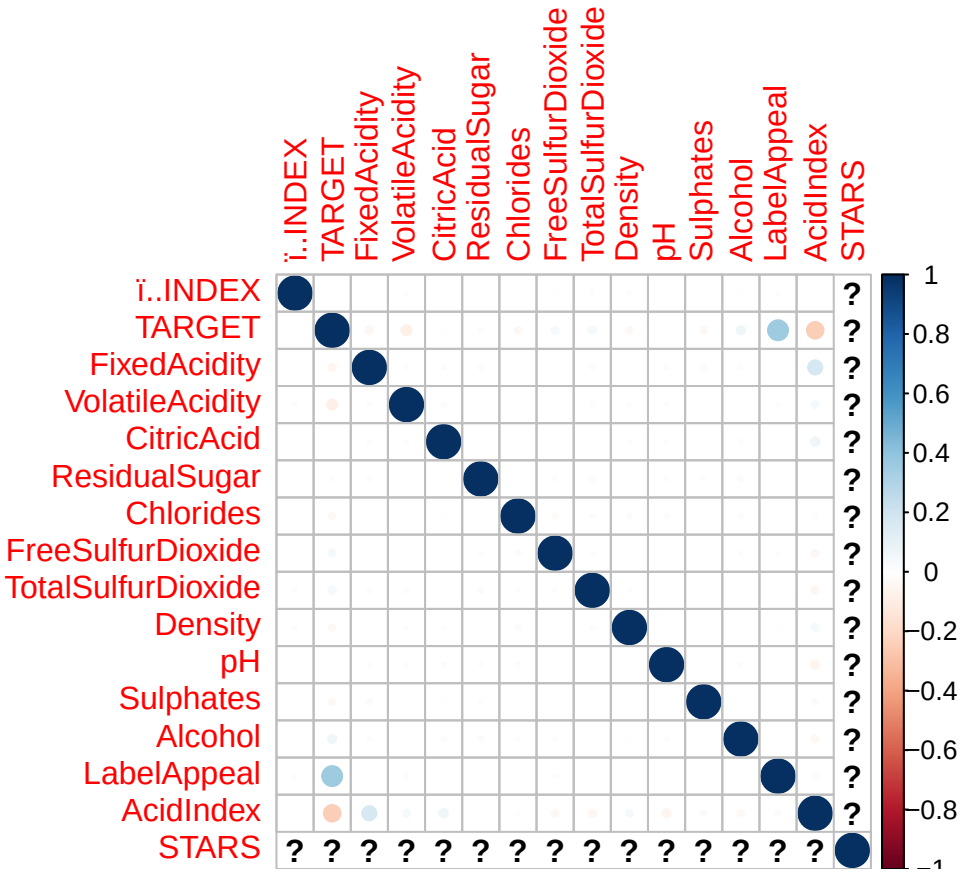


	not_na_count	na_count	na_pct
ï..INDEX	12795	0	0.0000000
TARGET	12795	0	0.0000000
FixedAcidity	12795	0	0.0000000
VolatileAcidity	12795	0	0.0000000
CitricAcid	12795	0	0.0000000
ResidualSugar	12179	616	0.0481438
Chlorides	12157	638	0.0498632
FreeSulfurDioxide	12148	647	0.0505666
TotalSulfurDioxide	12113	682	0.0533021
Density	12795	0	0.0000000
pH	12400	395	0.0308714
Sulphates	11585	1210	0.0945682
Alcohol	12142	653	0.0510356
LabelAppeal	12795	0	0.0000000
AcidIndex	12795	0	0.0000000
STARS	9436	3359	0.2625244



Corrlation and Covariance:

There does not seem to be much correlation, much less any multi-collinearity issues:



Summary of Wine Data Correlation:

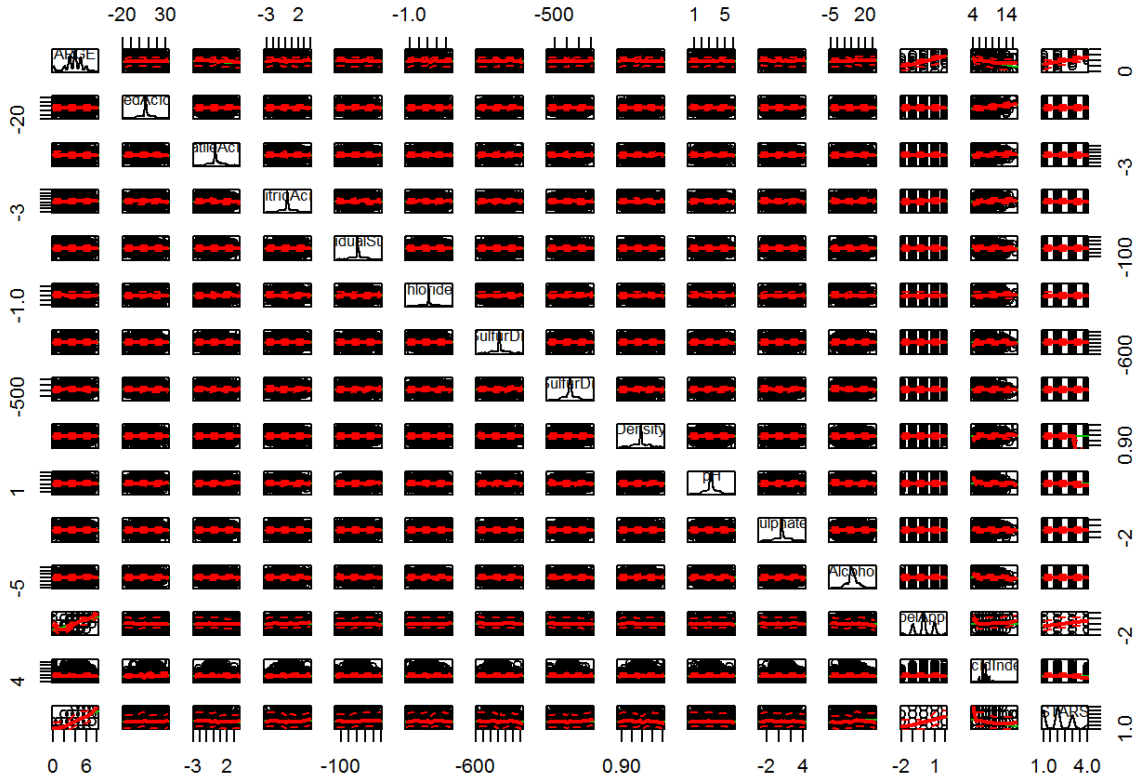
See the mean correlation for each column. Each is near zero, showing us that the columns themselves are independent.

```
##     ï..INDEX            TARGET           FixedAcidity
##  Min.   :-0.012618   Min.   :-0.161101   Min.   :-0.018734
##  1st Qu.:-0.003560   1st Qu.:-0.020687   1st Qu.:-0.008983
##  Median : 0.003618   Median : 0.007775   Median :-0.001030
##  Mean   : 0.065308   Mean   : 0.119635   Mean   : 0.071673
##  3rd Qu.: 0.009586   3rd Qu.: 0.038658   3rd Qu.: 0.012521
##  Max.   : 1.000000   Max.   : 1.000000   Max.   : 1.000000
##  VolatileAcidity       CitricAcid       ResidualSugar
##  Min.   :-0.070884   Min.   :-0.02178   Min.   :-0.0184162
##  1st Qu.:-0.014711   1st Qu.:-0.01281   1st Qu.:-0.0067129
##  Median :-0.002011   Median : 0.00337   Median : 0.0006643
##  Mean   : 0.056294   Mean   : 0.06480   Mean   : 0.0634631
##  3rd Qu.: 0.010878   3rd Qu.: 0.01241   3rd Qu.: 0.0107101
##  Max.   : 1.000000   Max.   : 1.00000   Max.   : 1.0000000
##    Chlorides          FreeSulfurDioxide   TotalSulfurDioxide
##  Min.   :-0.026547   Min.   :-0.021294   Min.   :-0.017161
##  1st Qu.:-0.014482   1st Qu.:-0.009228   1st Qu.:-0.008201
##  Median :-0.003756   Median : 0.008293   Median : 0.001479
##  Mean   : 0.057386   Mean   : 0.064848   Mean   : 0.063714
##  3rd Qu.: 0.002508   3rd Qu.: 0.011273   3rd Qu.: 0.016375
##  Max.   : 1.000000   Max.   : 1.000000   Max.   : 1.000000
##     Density                pH              Sulphates
```

```
##  Min.   :-0.032332   Min.   :-0.063540   Min.   :-0.0140547
##  1st Qu.:-0.009749   1st Qu.:-0.012286   1st Qu.:-0.0052085
##  Median :-0.003908   Median :-0.005150   Median : 0.0001468
##  Mean   : 0.062730   Mean   : 0.056203   Mean   : 0.0646953
##  3rd Qu.: 0.014786   3rd Qu.: 0.004502   3rd Qu.: 0.0102071
##  Max.   : 1.000000   Max.   : 1.000000   Max.   : 1.0000000
##     Alcohol            LabelAppeal           AcidIndex
##  Min.   :-0.054901   Min.   :-0.019551   Min.   :-0.161101
##  1st Qu.:-0.015143   1st Qu.:-0.002134   1st Qu.:-0.026981
##  Median :-0.003438   Median : 0.005229   Median :-0.002396
##  Mean   : 0.063010   Mean   : 0.117717   Mean   : 0.056447
##  3rd Qu.: 0.007351   3rd Qu.: 0.015651   3rd Qu.: 0.037894
##  Max.   : 1.000000   Max.   : 1.000000   Max.   : 1.000000
##      STARS
##  Min.   :-0.0862589
##  1st Qu.:-0.0094700
##  Median :-0.0001328
##  Mean   : 0.1135531
##  3rd Qu.: 0.0280086
##  Max.   : 1.0000000
```

Scatterplot of Wine Data:

This figure also serves to show the general horizontal (lack of) correlation between columns.



Summary of Wine Data Covariance:

Most of the covariances are low, though a few stick out as possibly high.

```
##     ï..INDEX            TARGET           FixedAcidity
##  Min.   :     -40   Min.   : -0.29481   Min.   :-24.73691
##  1st Qu.:     -14   1st Qu.: -0.01485   1st Qu.: -0.19551
##  Median :       6   Median :  0.23173   Median : -0.00699
##  Mean   : 1338286   Mean   : 11.08503   Mean   :  0.90763
##  3rd Qu.:     437   3rd Qu.:  1.18963   3rd Qu.:  0.08888
##  Max.   :21396100   Max.   :161.04070   Max.   : 39.46245
##  VolatileAcidity        CitricAcid       ResidualSugar
##  Min.   :-25.553108   Min.   :-0.42667   Min.   :  -2.3717
##  1st Qu.: -0.094280   1st Qu.:-0.00631   1st Qu.:  -0.1793
##  Median : -0.006559   Median : 0.00432   Median :   0.0826
##  Mean   : -1.781516   Mean   : 2.91247   Mean   : 161.9554
##  3rd Qu.:  0.005755   3rd Qu.: 0.06194   3rd Qu.:   9.9693
##  Max.   :  0.608814   Max.   :45.54073   Max.   :1266.0105
##    Chlorides         FreeSulfurDioxide   TotalSulfurDioxide
##  Min.   :-1.007353   Min.   :   -6.472   Min.   :  -24.74
##  1st Qu.:-0.015252   1st Qu.:   -1.022   1st Qu.:   -1.48
##  Median :-0.002520   Median :    1.200   Median :   -0.02
##  Mean   : 0.571868   Mean   : 1801.910   Mean   : 3874.66
##  3rd Qu.: 0.000199   3rd Qu.:   17.977   3rd Qu.:   44.38
##  Max.   :10.331595   Max.   :22242.541   Max.   :52650.13
##     Density                 pH              Sulphates
##  Min.   :-1.2940244   Min.   :-39.51430   Min.   :-10.412234
##  1st Qu.:-0.0004136   1st Qu.: -0.04437   1st Qu.: -0.006541
##  Median :-0.0000975   Median : -0.00168   Median :  0.000128
##  Mean   :-0.0712490   Mean   : -2.52035   Mean   : -0.352708
##  3rd Qu.: 0.0008329   3rd Qu.:  0.00305   3rd Qu.:  0.055571
##  Max.   : 0.1223139   Max.   :  0.55446   Max.   :  2.895748
##     Alcohol           LabelAppeal         AcidIndex
##  Min.   :-27.70921   Min.   :-1.19394   Min.   :-37.86421
##  1st Qu.: -0.76080   1st Qu.:-0.00245   1st Qu.: -0.25522
##  Median : -0.01548   Median : 0.00723   Median : -0.02517
##  Mean   : -2.18033   Mean   : 5.19064   Mean   : -2.73474
##  3rd Qu.:  0.02251   3rd Qu.: 0.37402   3rd Qu.:  0.02357
##  Max.   : 13.98369   Max.   :80.89284   Max.   :  1.38969
##      STARS
##  Min.   :-1.268509
##  1st Qu.:-0.012080
##  Median : 0.000118
##  Mean   : 0.325995
##  3rd Qu.: 0.524104
##  Max.   : 3.323740
```

2) Data Preparation

Looking for Patterns in the 'STARS' NA Values:

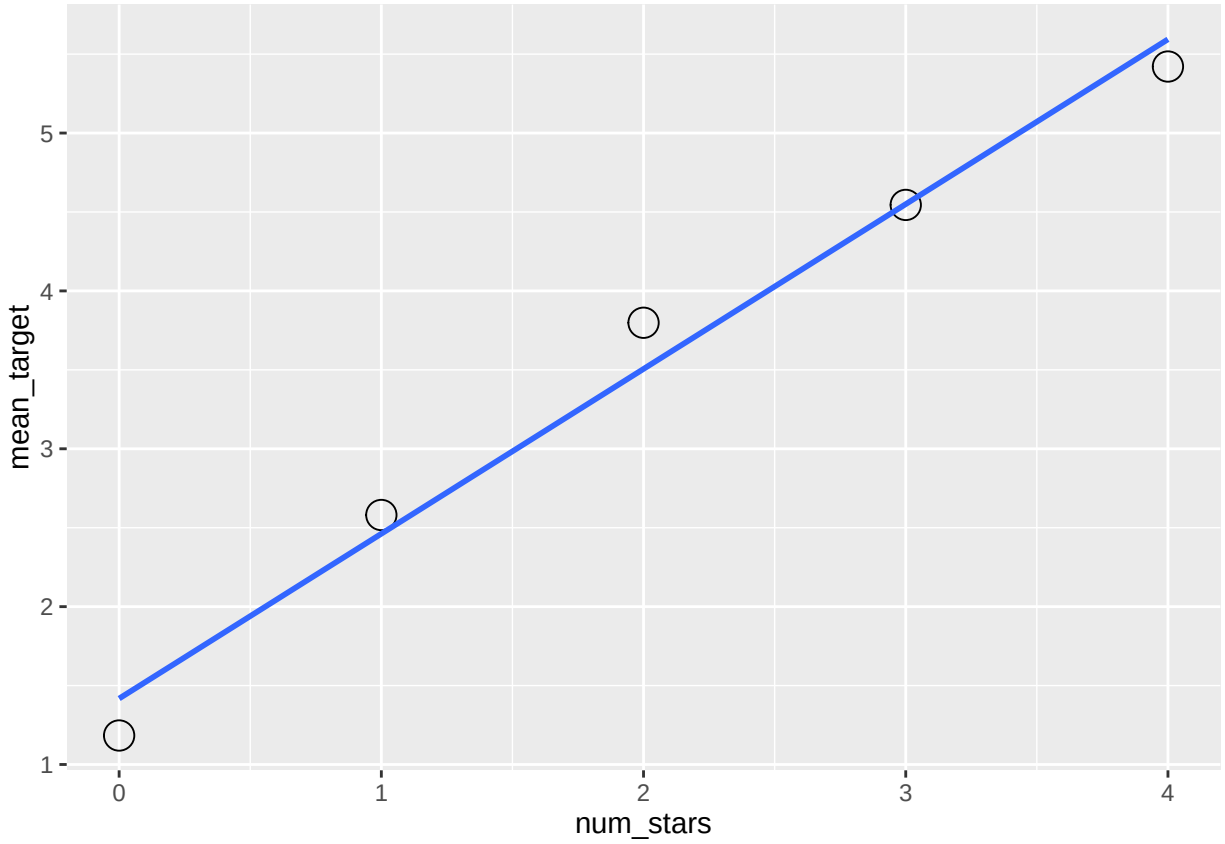
Next, due to such a high NA percent ***(26.25%)***

Note that there are no ZEROS in the STARS field, 1 is the min:

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

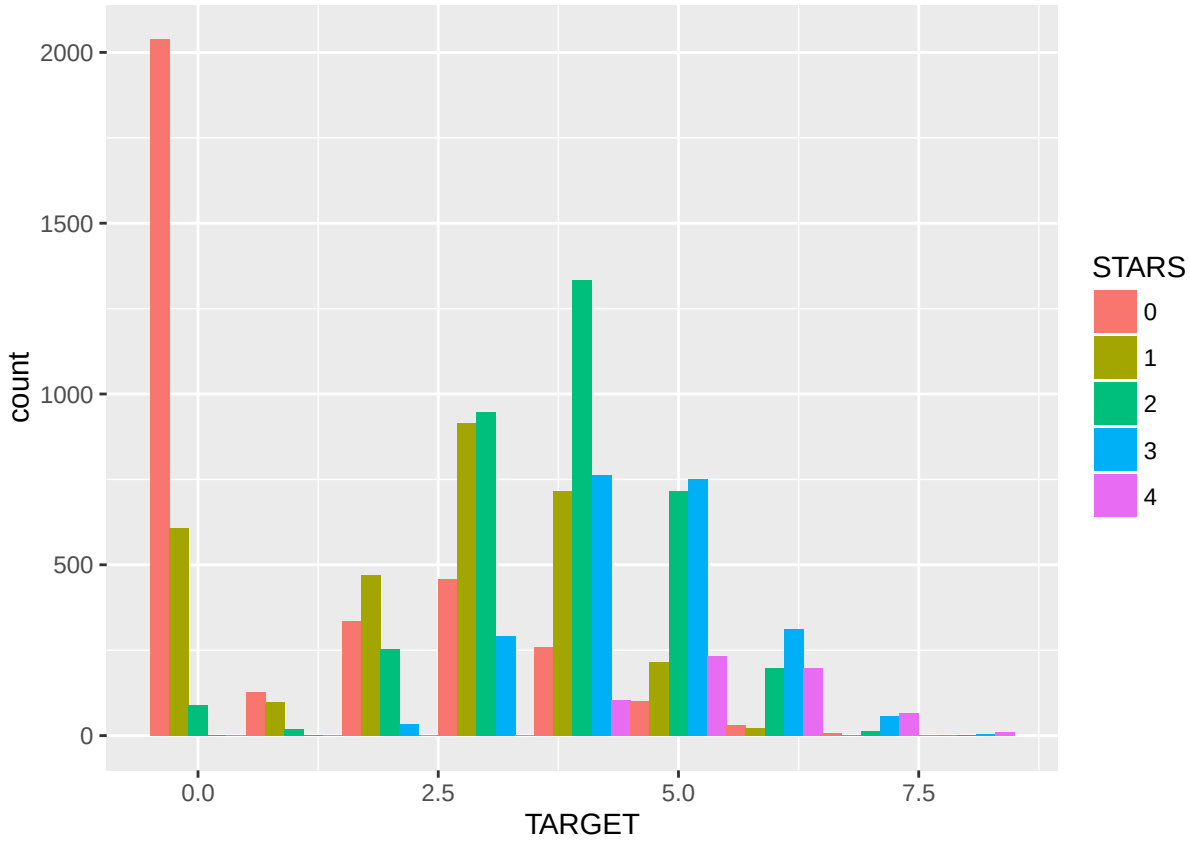
Graphically, it seems that a blank stars field is analagous to a ZERO (see chart below).

```
##   num_stars mean_target
## 1         0    1.183686
## 2         1    2.580539
## 3         2    3.798319
## 4         3    4.544756
## 5         4    5.421569
```



Though this calculation may be imperfect at the moment, we will show later that the calculation is to be discarded, and thus not worth figuring out a closer approximation for NA's replacement in the ***STARS*** field.

After the fill, we have what looks like a ***Zero-Inflated*** model on our hands.



3) Build Models

We will build 4 data sets from our training data:

- TRAINING set where NAs were replaced by ZEROs in the SCORE column.
- TRAINING set where NAs were ***NOT*** replaced by ZEROs in the SCORE column.
- TEST set where NAs were replaced by ZEROs in the SCORE column.
- TEST set where NAs were ***NOT*** replaced by ZEROs in the SCORE column.

Simple Step Selection will be used for attribute selection, and we will build 3 models for both sets, plus a "zerinfl" negative binomial, yielding a total of 7 models:

- ***A Poisson GLM with SCORE: "Zeros-for-NAs"***
- ***A Poisson GLM with SCORE: "NAs Removed"***
- ***A Negative Binomial GLM with SCORE: "Zeros-for-NAs"***
- ***A Negative Binomial GLM with SCORE: "NAs Removed"***
- ***A Multiple Linear Regression Model with SCORE: "Zeros-for-NAs"***
- ***A Multiple Linear Regression Model with SCORE: "NAs Removed"***
- ***A Negative Binomial Model via "zeroinfl()": "Zeros-for-NAs"***

Summaries for these Models are below:

(To show our point regarding coefficients, we will show summaries for the 2 Negative Binomial Distributions. Summaries of all 7 models, however, are located in the appendix).

```r
fit.nb.zeros <- step(glm.nb(TARGET ~ . , data = wine.zeros.train), trace = FALSE)
summary(fit.nb.zeros)
```

```
##
## Call:
## glm.nb(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide +
##     TotalSulfurDioxide + Sulphates + Alcohol + LabelAppeal +
##     AcidIndex + STARS, data = wine.zeros.train, init.theta = 41523.55951,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2390  -0.6311   0.0000   0.4408   3.6277
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.208e-01  3.206e-01   0.377  0.70639
## VolatileAcidity     -3.136e-02  7.557e-03  -4.150 3.33e-05 ***
## FreeSulfurDioxide    9.342e-05  3.959e-05   2.360  0.01827 *
## TotalSulfurDioxide   6.691e-05  2.561e-05   2.613  0.00898 **
## Sulphates           -1.068e-02  6.254e-03  -1.708  0.08769 .
## Alcohol              4.616e-03  1.596e-03   2.893  0.00382 **
## LabelAppeal-1        2.662e-01  4.400e-02   6.049 1.46e-09 ***
## LabelAppeal0         4.505e-01  4.295e-02  10.488  < 2e-16 ***
## LabelAppeal1         5.920e-01  4.367e-02  13.558  < 2e-16 ***
## LabelAppeal2         7.332e-01  4.930e-02  14.873  < 2e-16 ***
## AcidIndex5          -3.325e-01  3.252e-01  -1.023  0.30653
## AcidIndex6          -2.626e-01  3.173e-01  -0.828  0.40789
## AcidIndex7          -3.019e-01  3.169e-01  -0.953  0.34078
## AcidIndex8          -3.299e-01  3.170e-01  -1.041  0.29807
## AcidIndex9          -4.468e-01  3.174e-01  -1.408  0.15918
## AcidIndex10         -5.977e-01  3.189e-01  -1.874  0.06091 .
## AcidIndex11         -9.226e-01  3.231e-01  -2.856  0.00429 **
## AcidIndex12         -9.078e-01  3.300e-01  -2.751  0.00594 **
## AcidIndex13         -7.657e-01  3.332e-01  -2.298  0.02154 *
## AcidIndex14         -9.978e-01  3.533e-01  -2.824  0.00474 **
## AcidIndex15         -3.222e-01  4.290e-01  -0.751  0.45261
## AcidIndex16         -3.787e+01  3.355e+07   0.000  1.00000
## AcidIndex17         -1.272e+00  5.484e-01  -2.319  0.02039 *
## STARS1               7.290e-01  2.262e-02  32.227  < 2e-16 ***
## STARS2               1.058e+00  2.113e-02  50.048  < 2e-16 ***
## STARS3               1.170e+00  2.225e-02  52.577  < 2e-16 ***
## STARS4               1.300e+00  2.807e-02  46.292  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(41523.56) family taken to be 1)
##
##     Null deviance: 17041  on 9595  degrees of freedom
## Residual deviance: 10185  on 9569  degrees of freedom
## AIC: 34235
##
## Number of Fisher Scoring iterations: 1
```

```
##
##
##               Theta:  41524
##           Std. Err.:  40652
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -34179.03
```

```
fit.nb.nozeros <- step(glm.nb(TARGET ~ . , data = wine.nozeros.train), trace = FALSE)
summary(fit.nb.nozeros)
```

```
##
## Call:
## glm.nb(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide +
##     Alcohol + LabelAppeal + AcidIndex + STARS, data = wine.nozeros.train,
##     init.theta = 132309.9196, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2643  -0.2662   0.0492   0.3994   1.7733
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         3.202e-01  4.512e-01   0.710  0.47790
## VolatileAcidity    -2.543e-02  8.013e-03  -3.174  0.00150 **
## FreeSulfurDioxide   7.973e-05  4.191e-05   1.902  0.05711 .
## Alcohol             5.112e-03  1.693e-03   3.019  0.00253 **
## LabelAppeal-1       2.966e-01  5.113e-02   5.801 6.61e-09 ***
## LabelAppeal0        5.114e-01  4.998e-02  10.233  < 2e-16 ***
## LabelAppeal1        6.697e-01  5.064e-02  13.224  < 2e-16 ***
## LabelAppeal2        8.108e-01  5.589e-02  14.507  < 2e-16 ***
## AcidIndex5          1.559e-01  4.540e-01   0.343  0.73123
## AcidIndex6          1.821e-01  4.482e-01   0.406  0.68460
## AcidIndex7          1.523e-01  4.479e-01   0.340  0.73376
## AcidIndex8          1.264e-01  4.479e-01   0.282  0.77773
## AcidIndex9          6.587e-02  4.483e-01   0.147  0.88318
## AcidIndex10        -2.482e-02  4.497e-01  -0.055  0.95598
## AcidIndex11        -1.793e-01  4.527e-01  -0.396  0.69200
## AcidIndex12        -1.289e-01  4.591e-01  -0.281  0.77886
## AcidIndex13        -1.049e-01  4.599e-01  -0.228  0.81951
## AcidIndex14        -4.651e-01  4.816e-01  -0.966  0.33410
## AcidIndex15         8.730e-02  5.587e-01   0.156  0.87582
## AcidIndex16        -3.765e+01  5.917e+07   0.000  1.00000
## AcidIndex17         6.789e-03  6.336e-01   0.011  0.99145
## STARS2              3.279e-01  1.666e-02  19.688  < 2e-16 ***
## STARS3              4.387e-01  1.819e-02  24.123  < 2e-16 ***
## STARS4              5.644e-01  2.517e-02  22.424  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(132309.9) family taken to be 1)
##
##     Null deviance: 6582.8  on 7076  degrees of freedom
## Residual deviance: 4339.4  on 7053  degrees of freedom
```

```
## AIC: 25368
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  132310
##           Std. Err.:  202837
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -25317.94
```

Some Notes on the Models

- Though the 5th and 6th models are not GLMs, and thus should be used with normal distributions, we need to remember that the TARGET variable did have a normal distribution before the ZEROS replaced the NAs in the SCORE column.
- There seem to be MORE significant fields in the "with zeros" model than in the "no zeros" model. Will this mean higher accuracy?

The Output of these 6 Models:

Model	SE	SD	AIC	BIC	LogLik
Poisson no 0s	0.84	0.89	25365.82	25530.57	-12658.91
Poisson w/ 0s	0.84	0.89	25367.94	25539.56	-12658.97
Negative Binomial no 0s	0.84	0.87	22019.33	22197.81	-10983.67
Negative Binomial w/0s	1.02	1.25	34232.72	34426.29	-17089.36
Multiple Linear Regression no 0s	1.02	1.25	34235.03	34435.76	-17089.51
Multiple Linear Regression w/ 0s	1.02	1.23	32395.08	32617.32	-16166.54
Zero-Infl Negative Binomial	1.01	1.24	31261.41	NaN	-15597.71

Model Co-efficients

(Similar observations hold for the 4 models not shown also)

In the ***Negative Binomial Fit With Zeroes***, the following estimates are displayed for STARS*:

- STARS1 = 7.292e-01
- STARS2 = 1.058e+00
- STARS3 = 1.171e+00
- STARS4 = 1.301e+00

In the ***Negative Binomial Fit With NO Zeroes***, the following estimates are displayed for STARS*:

- STARS2 = 3.280e-01
- STARS3 = 4.389e-01
- STARS4 = 5.650e-01

Note that STARS1 has a huge increase in impact in the first fit, showing what an impact that NA to Zero translation had. (We'll see soon whether that impact was one of clarifying or confusing.)

4) Select Models

Before selecting a model, a quick explanation of why the "no ZEROs" models performed better:

One might think that either removing or keeping a value (such as the "Perfect-Graphical-Fit" Zeros-for-NAs) would possibly ***improve*** a model's accuracy, but at least ***maintain*** it. In this case, however, we saw a relatively large drop in performance of the models due to the inclusion of this attribute. Why would this be?

One probable explanation is that true customers that actually ***BOUGHT*** the product took the time to fill their surveys out accurately. Customers who didn't purchase the product (with less stake in the game and/or lack of knowledge of the product) simply did not contribute such useful data. Due to this, the Zero-Inflated model is basically gone, and a linear model with a normal distribution again seems reasonable.

We know that Poisson Regression is actually a special case of Negative Binomial Regression (where the mean and the variance are equal), but in this case, the Poisson Regression did not yield more accurate results. We know that if there is overdispersion in the Poisson, then the estimates from the Poisson regression model are consistent but inefficient. It seems from our Box Plots in Figure 1.2 that these overdispersions may have occurred.

Based on this and the above results table, it seems that the ***Negative Binomial no 0s*** is our winner.

"When comparing models fitted by maximum likelihood to the same data, the smaller the AIC or BIC, the better the fit.""

Our results have been written to ***DATA621-HW-5-RESULTS.csv***, and a sampling is shown below:

```
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##   1   4   3   2   1   6   2   2   1   1   0   1   4   0   1   2   2   1
##  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36
##   4   5   2   1   2   2   4   5   3   6   5   2   0   1   5   3   1   4
##  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
##   3   4   1   1   2   0   1   1   0   4   3   5   4   2   4   0   4   1
##  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72
##   2   0   0   4   1   1   0   0   0   2   3   2   4   5   1   4   0   0
##  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
##   2   5   2   3   0   4   5   0   2   2   2   2   3   4   3   5   0   5
##  91  92  93  94  95  96  97  98  99 100
##   2   1   1   3   2   0   4   3   3   1
```

Appendix

Summaries of all 7 Models:

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide +
##     Alcohol + LabelAppeal + AcidIndex + STARS, family = poisson,
##     data = wine.nozeros.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2643  -0.2663   0.0492   0.3994   1.7734
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)        3.202e-01  4.511e-01   0.710  0.47789
## VolatileAcidity   -2.543e-02  8.013e-03  -3.174  0.00150 **
## FreeSulfurDioxide  7.973e-05  4.191e-05   1.902  0.05711 .
## Alcohol            5.112e-03  1.693e-03   3.019  0.00253 **
## LabelAppeal-1      2.966e-01  5.113e-02   5.801 6.61e-09 ***
## LabelAppeal0       5.114e-01  4.998e-02  10.233  < 2e-16 ***
## LabelAppeal1       6.697e-01  5.064e-02  13.224  < 2e-16 ***
## LabelAppeal2       8.108e-01  5.589e-02  14.507  < 2e-16 ***
## AcidIndex5         1.559e-01  4.540e-01   0.343  0.73123
## AcidIndex6         1.821e-01  4.482e-01   0.406  0.68459
## AcidIndex7         1.523e-01  4.479e-01   0.340  0.73375
## AcidIndex8         1.264e-01  4.479e-01   0.282  0.77772
## AcidIndex9         6.587e-02  4.483e-01   0.147  0.88317
## AcidIndex10       -2.482e-02  4.496e-01  -0.055  0.95598
## AcidIndex11       -1.793e-01  4.527e-01  -0.396  0.69200
## AcidIndex12       -1.289e-01  4.591e-01  -0.281  0.77886
## AcidIndex13       -1.049e-01  4.599e-01  -0.228  0.81951
## AcidIndex14       -4.651e-01  4.816e-01  -0.966  0.33409
## AcidIndex15        8.730e-02  5.587e-01   0.156  0.87582
## AcidIndex16       -1.216e+01  1.727e+02  -0.070  0.94385
## AcidIndex17        6.789e-03  6.336e-01   0.011  0.99145
## STARS2             3.279e-01  1.666e-02  19.688  < 2e-16 ***
## STARS3             4.387e-01  1.819e-02  24.123  < 2e-16 ***
## STARS4             5.644e-01  2.517e-02  22.425  < 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: 6583.0  on 7076  degrees of freedom
## Residual deviance: 4339.4  on 7053  degrees of freedom
## AIC: 25366
##
## Number of Fisher Scoring iterations: 9


##
## Call:
## glm.nb(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide +
##     Alcohol + LabelAppeal + AcidIndex + STARS, data = wine.nozeros.train,
##     init.theta = 132309.9196, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2643  -0.2662   0.0492   0.3994   1.7733
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        3.202e-01  4.512e-01   0.710  0.47790
## VolatileAcidity   -2.543e-02  8.013e-03  -3.174  0.00150 **
## FreeSulfurDioxide  7.973e-05  4.191e-05   1.902  0.05711 .
## Alcohol            5.112e-03  1.693e-03   3.019  0.00253 **
## LabelAppeal-1      2.966e-01  5.113e-02   5.801 6.61e-09 ***
## LabelAppeal0       5.114e-01  4.998e-02  10.233  < 2e-16 ***
## LabelAppeal1       6.697e-01  5.064e-02  13.224  < 2e-16 ***
```

```
## LabelAppeal2       8.108e-01  5.589e-02  14.507  < 2e-16 ***
## AcidIndex5         1.559e-01  4.540e-01   0.343  0.73123
## AcidIndex6         1.821e-01  4.482e-01   0.406  0.68460
## AcidIndex7         1.523e-01  4.479e-01   0.340  0.73376
## AcidIndex8         1.264e-01  4.479e-01   0.282  0.77773
## AcidIndex9         6.587e-02  4.483e-01   0.147  0.88318
## AcidIndex10       -2.482e-02  4.497e-01  -0.055  0.95598
## AcidIndex11       -1.793e-01  4.527e-01  -0.396  0.69200
## AcidIndex12       -1.289e-01  4.591e-01  -0.281  0.77886
## AcidIndex13       -1.049e-01  4.599e-01  -0.228  0.81951
## AcidIndex14       -4.651e-01  4.816e-01  -0.966  0.33410
## AcidIndex15        8.730e-02  5.587e-01   0.156  0.87582
## AcidIndex16       -3.765e+01  5.917e+07   0.000  1.00000
## AcidIndex17        6.789e-03  6.336e-01   0.011  0.99145
## STARS2             3.279e-01  1.666e-02  19.688  < 2e-16 ***
## STARS3             4.387e-01  1.819e-02  24.123  < 2e-16 ***
## STARS4             5.644e-01  2.517e-02  22.424  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(132309.9) family taken to be 1)
##
##     Null deviance: 6582.8  on 7076  degrees of freedom
## Residual deviance: 4339.4  on 7053  degrees of freedom
## AIC: 25368
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  132310
##           Std. Err.:  202837
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -25317.94


##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide + Density +
##     Alcohol + LabelAppeal + AcidIndex + STARS, data = wine.nozeros.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3174 -0.5322  0.1050  0.7376  3.2163
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.760e+00  1.253e+00   1.405  0.16016
## VolatileAcidity   -9.350e-02  1.745e-02  -5.360 8.60e-08 ***
## FreeSulfurDioxide  3.004e-04  9.178e-05   3.273  0.00107 **
## Density           -1.085e+00  5.141e-01  -2.110  0.03487 *
## Alcohol            1.877e-02  3.669e-03   5.116 3.20e-07 ***
## LabelAppeal-1      6.108e-01  8.173e-02   7.474 8.71e-14 ***
## LabelAppeal0       1.260e+00  7.986e-02  15.773  < 2e-16 ***
## LabelAppeal1       1.916e+00  8.272e-02  23.165  < 2e-16 ***
```

```
## LabelAppeal2       2.663e+00  1.046e-01  25.459  < 2e-16 ***
## AcidIndex5         7.429e-01  1.159e+00   0.641  0.52148
## AcidIndex6         8.505e-01  1.147e+00   0.742  0.45828
## AcidIndex7         7.365e-01  1.146e+00   0.643  0.52052
## AcidIndex8         6.339e-01  1.146e+00   0.553  0.58028
## AcidIndex9         4.269e-01  1.147e+00   0.372  0.70972
## AcidIndex10        1.372e-01  1.149e+00   0.119  0.90494
## AcidIndex11       -2.496e-01  1.152e+00  -0.217  0.82850
## AcidIndex12       -1.734e-01  1.162e+00  -0.149  0.88136
## AcidIndex13       -9.105e-02  1.163e+00  -0.078  0.93761
## AcidIndex14       -8.056e-01  1.178e+00  -0.684  0.49395
## AcidIndex15        4.514e-01  1.324e+00   0.341  0.73305
## AcidIndex16       -2.047e+00  1.620e+00  -1.263  0.20650
## AcidIndex17        1.485e-01  1.623e+00   0.092  0.92710
## STARS2             1.010e+00  3.313e-02  30.480  < 2e-16 ***
## STARS3             1.497e+00  3.859e-02  38.789  < 2e-16 ***
## STARS4             2.173e+00  6.126e-02  35.463  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.144 on 7052 degrees of freedom
## Multiple R-squared:  0.465,  Adjusted R-squared:  0.4632
## F-statistic: 255.4 on 24 and 7052 DF,  p-value: < 2.2e-16


##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide +
##     TotalSulfurDioxide + Sulphates + Alcohol + LabelAppeal +
##     AcidIndex + STARS, family = poisson, data = wine.zeros.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2391  -0.6311  -0.0001   0.4408   3.6279
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.207e-01  3.206e-01   0.377  0.70647
## VolatileAcidity     -3.136e-02  7.556e-03  -4.150 3.33e-05 ***
## FreeSulfurDioxide    9.342e-05  3.958e-05   2.360  0.01827 *
## TotalSulfurDioxide   6.691e-05  2.561e-05   2.613  0.00898 **
## Sulphates           -1.068e-02  6.254e-03  -1.708  0.08769 .
## Alcohol              4.616e-03  1.596e-03   2.893  0.00382 **
## LabelAppeal-1        2.662e-01  4.400e-02   6.049 1.46e-09 ***
## LabelAppeal0         4.505e-01  4.295e-02  10.488  < 2e-16 ***
## LabelAppeal1         5.920e-01  4.367e-02  13.558  < 2e-16 ***
## LabelAppeal2         7.332e-01  4.930e-02  14.874  < 2e-16 ***
## AcidIndex5          -3.325e-01  3.252e-01  -1.022  0.30656
## AcidIndex6          -2.626e-01  3.173e-01  -0.828  0.40793
## AcidIndex7          -3.019e-01  3.169e-01  -0.953  0.34081
## AcidIndex8          -3.298e-01  3.170e-01  -1.041  0.29810
## AcidIndex9          -4.468e-01  3.174e-01  -1.408  0.15920
## AcidIndex10         -5.977e-01  3.189e-01  -1.874  0.06091 .
## AcidIndex11         -9.226e-01  3.231e-01  -2.856  0.00429 **
## AcidIndex12         -9.077e-01  3.300e-01  -2.751  0.00594 **
```

```
## AcidIndex13        -7.657e-01  3.331e-01  -2.298  0.02154 *
## AcidIndex14        -9.977e-01  3.533e-01  -2.824  0.00474 **
## AcidIndex15        -3.221e-01  4.289e-01  -0.751  0.45265
## AcidIndex16        -1.316e+01  1.410e+02  -0.093  0.92562
## AcidIndex17        -1.272e+00  5.484e-01  -2.319  0.02040 *
## STARS1              7.290e-01  2.262e-02  32.228  < 2e-16 ***
## STARS2              1.058e+00  2.113e-02  50.049  < 2e-16 ***
## STARS3              1.170e+00  2.225e-02  52.578  < 2e-16 ***
## STARS4              1.300e+00  2.807e-02  46.294  < 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: 17042  on 9595  degrees of freedom
## Residual deviance: 10185  on 9569  degrees of freedom
## AIC: 34233
##
## Number of Fisher Scoring iterations: 10


##
## Call:
## glm.nb(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide +
##     TotalSulfurDioxide + Sulphates + Alcohol + LabelAppeal +
##     AcidIndex + STARS, data = wine.zeros.train, init.theta = 41523.55951,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2390  -0.6311   0.0000   0.4408   3.6277
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.208e-01  3.206e-01   0.377  0.70639
## VolatileAcidity    -3.136e-02  7.557e-03  -4.150 3.33e-05 ***
## FreeSulfurDioxide   9.342e-05  3.959e-05   2.360  0.01827 *
## TotalSulfurDioxide  6.691e-05  2.561e-05   2.613  0.00898 **
## Sulphates          -1.068e-02  6.254e-03  -1.708  0.08769 .
## Alcohol             4.616e-03  1.596e-03   2.893  0.00382 **
## LabelAppeal-1       2.662e-01  4.400e-02   6.049 1.46e-09 ***
## LabelAppeal0        4.505e-01  4.295e-02  10.488  < 2e-16 ***
## LabelAppeal1        5.920e-01  4.367e-02  13.558  < 2e-16 ***
## LabelAppeal2        7.332e-01  4.930e-02  14.873  < 2e-16 ***
## AcidIndex5         -3.325e-01  3.252e-01  -1.023  0.30653
## AcidIndex6         -2.626e-01  3.173e-01  -0.828  0.40789
## AcidIndex7         -3.019e-01  3.169e-01  -0.953  0.34078
## AcidIndex8         -3.299e-01  3.170e-01  -1.041  0.29807
## AcidIndex9         -4.468e-01  3.174e-01  -1.408  0.15918
## AcidIndex10        -5.977e-01  3.189e-01  -1.874  0.06091 .
## AcidIndex11        -9.226e-01  3.231e-01  -2.856  0.00429 **
## AcidIndex12        -9.078e-01  3.300e-01  -2.751  0.00594 **
## AcidIndex13        -7.657e-01  3.332e-01  -2.298  0.02154 *
## AcidIndex14        -9.978e-01  3.533e-01  -2.824  0.00474 **
## AcidIndex15        -3.222e-01  4.290e-01  -0.751  0.45261
```

```
## AcidIndex16        -3.787e+01  3.355e+07   0.000  1.00000
## AcidIndex17        -1.272e+00  5.484e-01  -2.319  0.02039 *
## STARS1              7.290e-01  2.262e-02  32.227  < 2e-16 ***
## STARS2              1.058e+00  2.113e-02  50.048  < 2e-16 ***
## STARS3              1.170e+00  2.225e-02  52.577  < 2e-16 ***
## STARS4              1.300e+00  2.807e-02  46.292  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(41523.56) family taken to be 1)
##
##     Null deviance: 17041  on 9595  degrees of freedom
## Residual deviance: 10185  on 9569  degrees of freedom
## AIC: 34235
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  41524
##           Std. Err.:  40652
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -34179.03


##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS, data = wine.zeros.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0025 -0.8552  0.0462  0.8343  6.0278
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.712e+00  1.056e+00   2.569 0.010203 *
## VolatileAcidity    -9.805e-02  1.709e-02  -5.736 9.96e-09 ***
## Chlorides          -7.519e-02  4.204e-02  -1.789 0.073706 .
## FreeSulfurDioxide   2.669e-04  9.015e-05   2.960 0.003080 **
## TotalSulfurDioxide  1.920e-04  5.792e-05   3.315 0.000919 ***
## Density            -9.282e-01  5.026e-01  -1.847 0.064799 .
## pH                 -3.062e-02  1.984e-02  -1.544 0.122726
## Sulphates          -2.917e-02  1.423e-02  -2.050 0.040383 *
## Alcohol             1.425e-02  3.599e-03   3.960 7.56e-05 ***
## LabelAppeal-1       4.082e-01  7.229e-02   5.647 1.68e-08 ***
## LabelAppeal0        8.611e-01  7.051e-02  12.213  < 2e-16 ***
## LabelAppeal1        1.356e+00  7.366e-02  18.404  < 2e-16 ***
## LabelAppeal2        1.985e+00  9.824e-02  20.211  < 2e-16 ***
## AcidIndex5         -1.256e+00  9.426e-01  -1.332 0.182744
## AcidIndex6         -1.044e+00  9.262e-01  -1.127 0.259725
## AcidIndex7         -1.167e+00  9.255e-01  -1.261 0.207452
## AcidIndex8         -1.266e+00  9.256e-01  -1.368 0.171328
## AcidIndex9         -1.583e+00  9.261e-01  -1.709 0.087433 .
```

```
## AcidIndex10        -1.883e+00  9.276e-01  -2.030 0.042432 *
## AcidIndex11        -2.336e+00  9.300e-01  -2.512 0.012010 *
## AcidIndex12        -2.291e+00  9.346e-01  -2.452 0.014239 *
## AcidIndex13        -2.352e+00  9.424e-01  -2.496 0.012577 *
## AcidIndex14        -2.316e+00  9.506e-01  -2.436 0.014876 *
## AcidIndex15        -1.245e+00  1.095e+00  -1.138 0.255349
## AcidIndex16        -3.130e+00  1.133e+00  -2.762 0.005747 **
## AcidIndex17        -2.784e+00  1.069e+00  -2.605 0.009198 **
## STARS1              1.300e+00  3.817e-02  34.049  < 2e-16 ***
## STARS2              2.359e+00  3.727e-02  63.285  < 2e-16 ***
## STARS3              2.894e+00  4.306e-02  67.207  < 2e-16 ***
## STARS4              3.624e+00  6.851e-02  52.897  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.306 on 9566 degrees of freedom
## Multiple R-squared:  0.539,  Adjusted R-squared:  0.5376
## F-statistic: 385.6 on 29 and 9566 DF,  p-value: < 2.2e-16


##
## Call:
## zeroinfl(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     Density + Alcohol + LabelAppeal + AcidIndex + STARS | STARS,
##     data = wine.zeros.train, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.3497 -0.5268  0.0178  0.4471  2.7038
##
## Count model coefficients (negbin with log link):
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        7.213e-01  3.904e-01   1.848   0.0647 .
## VolatileAcidity   -1.465e-02  7.730e-03  -1.896   0.0580 .
## Chlorides         -1.703e-02  1.898e-02  -0.897   0.3697
## FreeSulfurDioxide  4.874e-05  3.991e-05   1.221   0.2219
## Density           -2.421e-01  2.273e-01  -1.065   0.2868
## Alcohol            6.660e-03  1.628e-03   4.092 4.28e-05 ***
## LabelAppeal-1      3.748e-01  4.540e-02   8.256  < 2e-16 ***
## LabelAppeal0       6.452e-01  4.443e-02  14.523  < 2e-16 ***
## LabelAppeal1       8.325e-01  4.525e-02  18.398  < 2e-16 ***
## LabelAppeal2       9.930e-01  5.094e-02  19.495  < 2e-16 ***
## AcidIndex5        -5.178e-03  3.252e-01  -0.016   0.9873
## AcidIndex6         3.377e-02  3.172e-01   0.106   0.9152
## AcidIndex7         5.686e-03  3.168e-01   0.018   0.9857
## AcidIndex8        -8.558e-03  3.168e-01  -0.027   0.9785
## AcidIndex9        -5.797e-02  3.173e-01  -0.183   0.8550
## AcidIndex10       -1.183e-01  3.191e-01  -0.371   0.7107
## AcidIndex11       -2.300e-01  3.253e-01  -0.707   0.4794
## AcidIndex12       -1.492e-01  3.346e-01  -0.446   0.6557
## AcidIndex13       -8.272e-02  3.377e-01  -0.245   0.8065
## AcidIndex14       -1.942e-01  3.800e-01  -0.511   0.6093
## AcidIndex15        1.276e-01  4.433e-01   0.288   0.7734
## AcidIndex16       -1.317e+01  2.716e+02  -0.048   0.9613
## AcidIndex17       -2.560e-01  6.208e-01  -0.412   0.6801
```

```
## STARS1             3.017e-02  2.439e-02   1.237   0.2161
## STARS2             1.579e-01  2.275e-02   6.943 3.84e-12 ***
## STARS3             2.551e-01  2.378e-02  10.726  < 2e-16 ***
## STARS4             3.700e-01  2.942e-02  12.577  < 2e-16 ***
## Log(theta)         1.754e+01  2.130e+00   8.235  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.33735    0.04319   7.811 5.67e-15 ***
## STARS1       -1.96541    0.07747 -25.369  < 2e-16 ***
## STARS2       -6.19471    0.89730  -6.904 5.07e-12 ***
## STARS3      -20.00694  456.23681  -0.044    0.965
## STARS4      -20.00141  869.14230  -0.023    0.982
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 41602484.5139
## Number of iterations in BFGS optimization: 95
## Log-likelihood: -1.56e+04 on 33 Df
```

All Code:

```
DO_PERFORM_STEPS <- TRUE
DO_SCATTERPLOT = FALSE

library(stringr)
library(knitr)
library(sandwich)
library(MASS)
library(ROCR)
library(pROC)
library(ggplot2)
library(pscl)
library(boot)
library(gridExtra)
library(Amelia)
library(car)
library(plyr)
library(psych)
library(reshape2)

wine <- read.csv('wine-training-data.csv', stringsAsFactors = FALSE)

do_factors <- function(wine_instance){
  wine_instance <- within(wine_instance, {
      LabelAppeal <- factor(LabelAppeal)
      AcidIndex <- factor(AcidIndex)
      STARS <- factor(STARS)
  })
  return (wine_instance)
}
wine_no_indexes <- wine[,-c(1)]
```

```r
d <- melt(wine_no_indexes[,sapply(wine_no_indexes, is.numeric)])

ggplot(d,aes(x = value)) +
    facet_wrap(~variable,scales = "free") +
    geom_histogram()
ggplot(d, aes(variable, value)) +
  facet_wrap(~variable,scales = "free") +
  geom_boxplot()
not_na_count <- sapply(wine, function(y) sum(length(which(!is.na(y)))))
na_count <- sapply(wine, function(y) sum(length(which(is.na(y)))))
na_pct <- na_count / (na_count + not_na_count)

na_summary_df <- data.frame(not_na_count,na_count,na_pct)

missmap(wine, main = "Missing Values Before (Non-STARS) Replacement")

kable(na_summary_df)

# doing this because whole data set would be gone otherwise, show that in numbers!!!!!!!!!!!!!
wine$ResidualSugar[is.na(wine$ResidualSugar)] <- sample(wine$ResidualSugar[!is.na(wine$ResidualSugar)])
wine$Chlorides[is.na(wine$Chlorides)] <- sample(wine$Chlorides[!is.na(wine$Chlorides)])
wine$FreeSulfurDioxide[is.na(wine$FreeSulfurDioxide)] <- sample(wine$FreeSulfurDioxide[!is.na(wine$Free
wine$TotalSulfurDioxide[is.na(wine$TotalSulfurDioxide)] <- sample(wine$TotalSulfurDioxide[!is.na(wine$T
wine$pH[is.na(wine$pH)] <- sample(wine$pH[!is.na(wine$pH)])
wine$Sulphates[is.na(wine$Sulphates)] <- sample(wine$Sulphates[!is.na(wine$Sulphates)])
wine$Alcohol[is.na(wine$Alcohol)] <- sample(wine$Alcohol[!is.na(wine$Alcohol)])

missmap(wine, main = "Missing Values After (Non-STARS) Replacement")

library(corrplot)
# remove index and stars (stars has nulls so shows up blank)
numeric_cols <- sapply(wine[,-c(1,16)], is.numeric)
M <- cor(wine[ , numeric_cols])
corrplot(M, method="circle")
summary(cor(wine[ , numeric_cols], use="complete.obs"))
if(DO_SCATTERPLOT){
  scatterplotMatrix(wine[,-c(1)])
}
summary(cov(wine[ , numeric_cols], use="complete.obs"))
min(wine$STARS[!is.na(wine$STARS)])
wine$STARS[is.na(wine$STARS)] <- 0

wine <- do_factors(wine)

m0 <- mean(wine$TARGET[wine$STARS == 0])
m1 <- mean(wine$TARGET[wine$STARS == 1])
m2 <- mean(wine$TARGET[wine$STARS == 2])
m3 <- mean(wine$TARGET[wine$STARS == 3])
m4 <- mean(wine$TARGET[wine$STARS == 4])

stars_summary_df <- data.frame(cbind(num_stars = c(0,1,2,3,4), mean_target = c(m0,m1,m2,m3,m4)))
stars_summary_df
```

```
ggplot(stars_summary_df, aes(num_stars, mean_target)) + geom_point(shape=1, size=5) + geom_smooth(metho
ggplot(wine, aes(TARGET, fill = STARS)) + geom_histogram(binwidth=1, bins = 8,  position="dodge")
wine.zeros <- wine
wine.nozeros <- wine[wine$STARS != 0,]

cutoff.zeros <- nrow(wine.zeros)*.75
wine.zeros.train <- wine.zeros[1:cutoff.zeros,]
wine.zeros.test <- wine.zeros[(cutoff.zeros+1):nrow(wine.zeros),]

cutoff.nozeros <- nrow(wine.nozeros)*.75
wine.nozeros.train <- wine.nozeros[1:cutoff.nozeros,]
wine.nozeros.test <- wine.nozeros[(cutoff.nozeros+1):nrow(wine.nozeros),]

wine.zeros <- do_factors(wine.zeros)
wine.nozeros <- do_factors(wine.nozeros)
wine.zeros.train <- do_factors(wine.zeros.train)
wine.zeros.test <- do_factors(wine.zeros.test)
wine.nozeros.train <- do_factors(wine.nozeros.train)
wine.nozeros.test <- do_factors(wine.nozeros.test)

fit.poisson.zeros <- step(glm(TARGET ~ . , family = poisson, data = wine.zeros.train), trace = FALSE)
#summary(fit.poisson.zeros)
fit.poisson.nozeros <- step(glm(TARGET ~ . , family = poisson, data = wine.nozeros.train), trace = FALS
#summary(fit.poisson.nozeros)
fit.nb.zeros <- step(glm.nb(TARGET ~ . , data = wine.zeros.train), trace = FALSE)
summary(fit.nb.zeros)
fit.nb.nozeros <- step(glm.nb(TARGET ~ . , data = wine.nozeros.train), trace = FALSE)
summary(fit.nb.nozeros)
fit.mlr.zeros <- step(lm(TARGET ~ ., data = wine.zeros.train), trace = FALSE)
#summary(fit.mlr.zeros)
fit.mlr.nozeros <- step(lm(TARGET ~ ., data = wine.nozeros.train), trace = FALSE)
#summary(fit.mlr.nozeros)
fit.nb.zeroinfl <- zeroinfl(TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide + Density + Alcohol
#summary(fit.nb.zeroinfl)
calc_sd <- function(fit, data){
  prediction <- predict(fit, newdata=data, type='response')
  difference <- (prediction - mean(data$TARGET))
  difference_squared <- difference * difference
  return (mean(sqrt(difference_squared)))
}

calc_se <- function(fit, data){
  prediction <- predict(fit, newdata=data, type='response')
  difference <- (prediction - data$TARGET)
  difference_squared <- difference * difference
  return (mean(sqrt(difference_squared)))
}
#####################################################################
# SD Calcs:
#####################################################################
sd.poisson.nozeros <- calc_sd(fit.poisson.nozeros, wine.nozeros.test)
sd.nb.nozeros <- calc_sd(fit.nb.nozeros, wine.nozeros.test)
sd.mlr.nozeros <- calc_sd(fit.mlr.nozeros, wine.nozeros.test)
```

```
sd.poisson.zeros <- calc_sd(fit.poisson.zeros, wine.zeros.test)
sd.nb.zeros <- calc_sd(fit.nb.zeros, wine.zeros.test)
sd.mlr.zeros <- calc_sd(fit.mlr.zeros, wine.zeros.test)

sd.nb.zeroinfl <- calc_sd(fit.nb.zeroinfl, wine.zeros.test)

SD <- format(c(sd.poisson.nozeros, sd.nb.nozeros, sd.mlr.nozeros, sd.poisson.zeros, sd.nb.zeros, sd.mlr

########################################################################
# SE Calcs:
########################################################################
se.poisson.nozeros <- calc_se(fit.poisson.nozeros, wine.nozeros.test)
se.nb.nozeros <- calc_se(fit.nb.nozeros, wine.nozeros.test)
se.mlr.nozeros <- calc_se(fit.mlr.nozeros, wine.nozeros.test)

se.poisson.zeros <- calc_se(fit.poisson.zeros, wine.zeros.test)
se.nb.zeros <- calc_se(fit.nb.zeros, wine.zeros.test)
se.mlr.zeros <- calc_se(fit.mlr.zeros, wine.zeros.test)

se.nb.zeroinfl <- calc_se(fit.nb.zeroinfl, wine.zeros.test)


SE <- format(c(se.poisson.nozeros, se.nb.nozeros, se.mlr.nozeros, se.poisson.zeros, se.nb.zeros, se.mlr

########################################################################
# AIC Calcs:
########################################################################
AIC <- format(c(AIC(fit.poisson.nozeros), AIC(fit.nb.nozeros), AIC(fit.mlr.nozeros), AIC(fit.poisson.ze

########################################################################
# BIC Calcs:
########################################################################
BIC <- format(c(BIC(fit.poisson.nozeros),BIC(fit.nb.nozeros),BIC(fit.mlr.nozeros),BIC(fit.poisson.zeros

########################################################################
# MDL Co-efficients:
########################################################################
all_fits <- c(fit.poisson.nozeros,fit.nb.nozeros,fit.mlr.nozeros,fit.poisson.zeros,fit.nb.zeros,fit.mlr

########################################################################
# LogLik Calcs:
########################################################################
LogLik <- format(c(logLik(fit.poisson.nozeros),logLik(fit.nb.nozeros),logLik(fit.mlr.nozeros),logLik(fi

Model <- c("Poisson no 0s", "Poisson w/ 0s", "Negative Binomial no 0s", "Negative Binomial w/0s", "Multi
kable(cbind(Model, SE, SD, AIC, BIC, LogLik))
file_name <- "DATA621-HW-5-RESULTS.csv"
##########3 ************************************
wine.final.test.data <- read.csv('wine-evaluation-data.csv', stringsAsFactors = FALSE)
colnames(wine.final.test.data) <- colnames(wine)
wine.final.test.data$STARS[is.na(wine.final.test.data$STARS)] <- 0
wine.final.test.data <- do_factors(wine.final.test.data)
final_prediction <- predict(fit.nb.zeros, newdata=wine.final.test.data, type='response')
```

```r
#final_prediction <- predict(fit.nb.nozeros, newdata=wine.final.test.data, type='response')
# not offering half bottles:
final_prediction <- round(final_prediction, digits = 0)
# if NA, then no bottles of wine:
final_prediction[is.na(final_prediction)] <- 0

head(final_prediction, n=100)
write.csv(final_prediction, file = file_name, fileEncoding = "UTF-8", na = "NA")
summary(fit.poisson.nozeros)
summary(fit.nb.nozeros)
summary(fit.mlr.nozeros)
summary(fit.poisson.zeros)
summary(fit.nb.zeros)
summary(fit.mlr.zeros)
summary(fit.nb.zeroinfl)
##
```