DATA621 Homework 5

Javern Wilson, Joseph Simone, Jack Russo

4/27/2020

Contents

Overview

In this homework assignment, we will explore, analyze and model a data set containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable 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. HINT: Sometimes, the fact that a variable is missing is actually predictive of the target. We will only use the variables given to us (or variables that we derive from the variables provided). Below is a short description of the variables of interest in the data set:

VARIABLE NAME DEFINITION THEORETICAL EFFECT * `INDEX`: Identification Variable (do not use) - **EFFECT:** None * `TARGET` Number of Cases Purchased - **EFFECT:** None * `AcidIndex`: Proprietary method of testing total acidity of wine by using a weighted average * `Alcohol`: Alcohol Content * `Chlorides`: Chloride content of wine * `CitricAcid`: Citric Acid Content * `Density`: Density of Wine * `FixedAcidity`: Fixed Acidity of Wine * `FreeSulfurDioxide`: Sulfur Dioxide content of wine * `LabelAppeal`: Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest customes don't like the design. - **EFFECT**: Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales. * `ResidualSugar`: Residual Sugar of wine STARS Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor - **EFFECT:** A high number of stars suggests high sales * `Sulphates`: Sulfate conten of winev * `TotalSulfurDioxide`: Total Sulfur Dioxide of Wine * `VolatileAcidity`: Volatile Acid content of wine * `pH`: pH of wine

DATA EXPLORATION

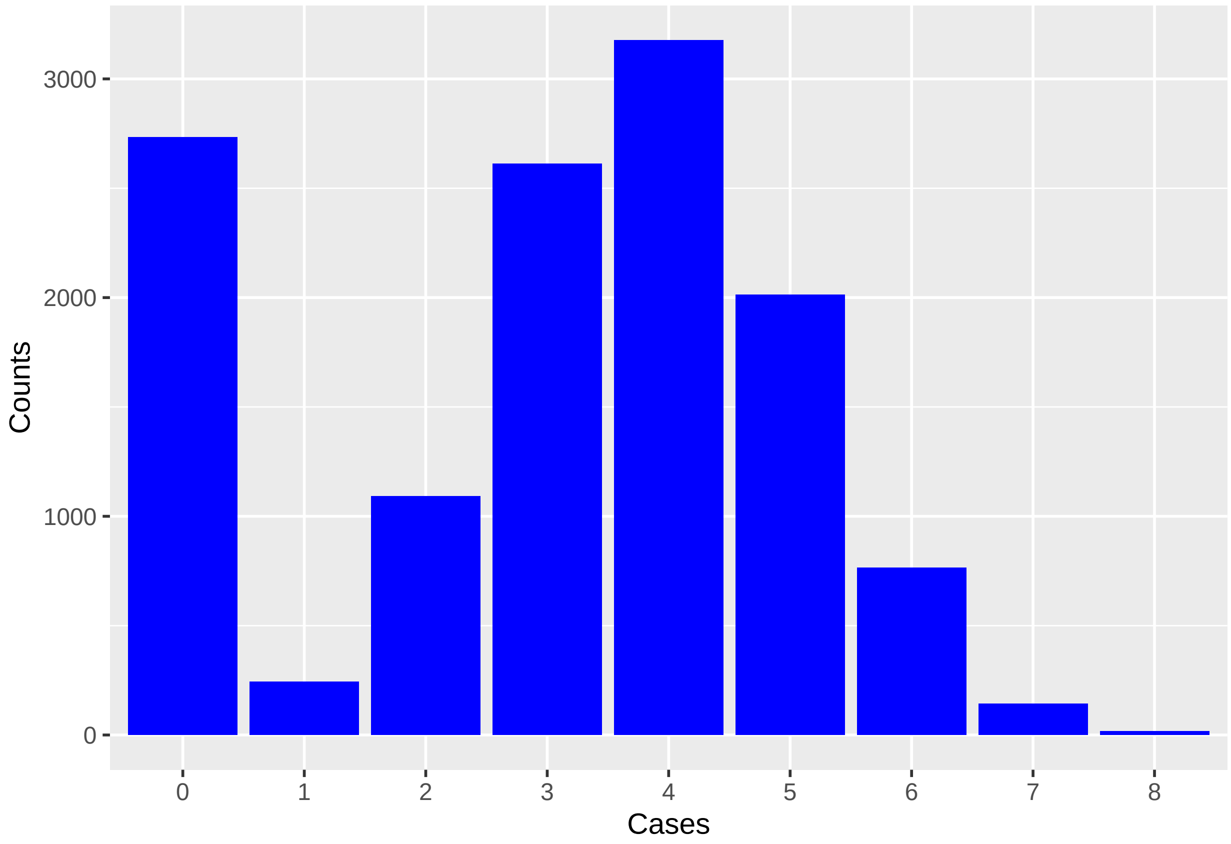
Preview

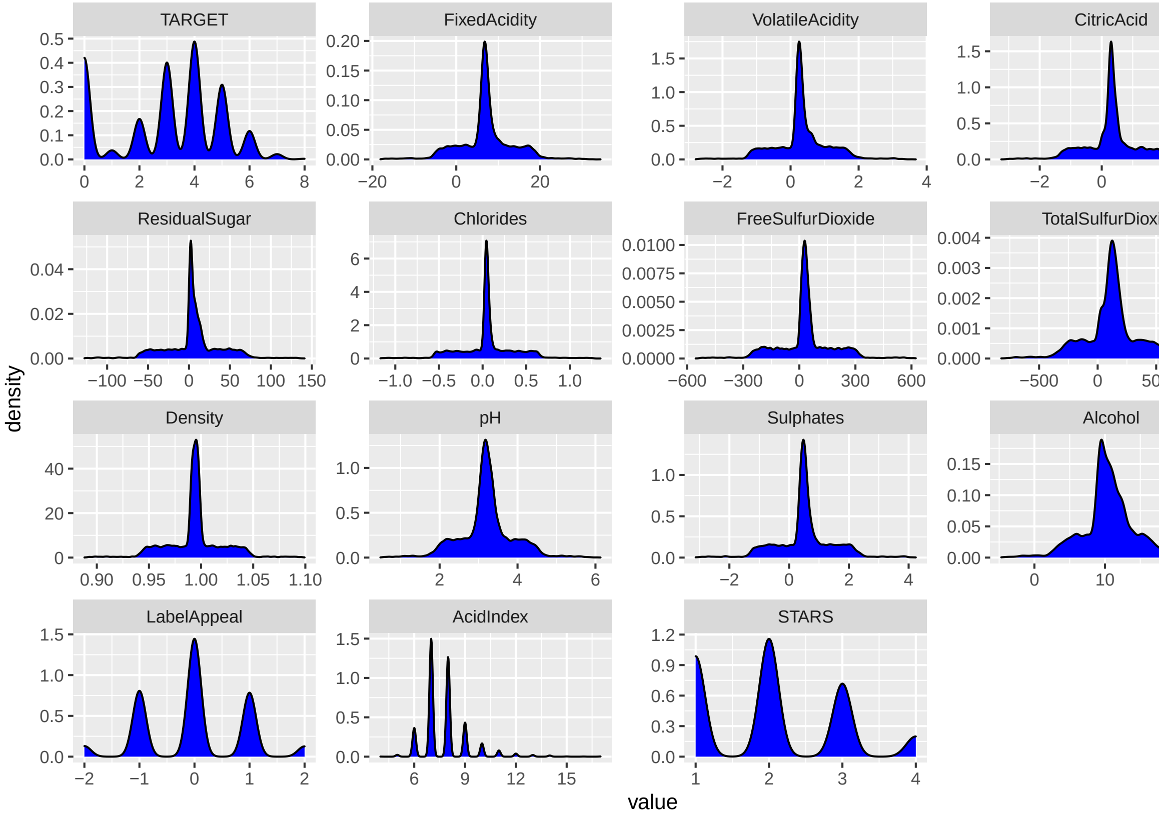
```
## # A tibble: 6 x 16
##   ï..INDEX TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar
##      <int>  <int>        <dbl>           <dbl>      <dbl>         <dbl>
## 1        1      3          3.2           1.16       -0.98          54.2
## 2        2      3          4.5           0.16       -0.81          26.1
## 3        4      5          7.1           2.64       -0.88          14.8
## 4        5      3          5.7           0.385       0.04          18.8
## 5        6      4          8             0.33       -1.26           9.4
## 6        7      0         11.3           0.32        0.59           2.2
## # ... with 10 more variables: Chlorides <dbl>, FreeSulfurDioxide <dbl>,
## #   TotalSulfurDioxide <dbl>, Density <dbl>, pH <dbl>, Sulphates <dbl>,
## #   Alcohol <dbl>, LabelAppeal <int>, AcidIndex <int>, STARS <int>
```

```
## 'data.frame':    12795 obs. of  16 variables:
##  $ ï..INDEX          : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET            : int  3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity      : num  3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity   : num  1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
##  $ CitricAcid        : num  -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
##  $ ResidualSugar     : num  54.2 26.1 14.8 18.8 9.4 ...
##  $ Chlorides         : num  -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...
##  $ FreeSulfurDioxide : num  NA 15 214 22 -167 -37 287 523 -213 62 ...
##  $ TotalSulfurDioxide: num  268 -327 142 115 108 15 156 551 NA 180 ...
##  $ Density           : num  0.993 1.028 0.995 0.996 0.995 ...
##  $ pH                : num  3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
##  $ Sulphates         : num  -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...
##  $ Alcohol           : num  9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
##  $ LabelAppeal       : int  0 -1 -1 -1 0 0 0 1 0 0 ...
##  $ AcidIndex         : int  8 7 8 6 9 11 8 7 6 8 ...
##  $ STARS             : int  2 3 3 1 2 NA NA 3 NA 4 ...
```

```
##     ï..INDEX         TARGET       FixedAcidity     VolatileAcidity
##  Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900
##  1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300
##  Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800
##  Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241
##  3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400
##  Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800
##
##    CitricAcid      ResidualSugar       Chlorides       FreeSulfurDioxide
##  Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00
##  1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00
##  Median : 0.3100   Median :   3.900   Median : 0.0460   Median :  30.00
##  Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85
##  3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00
##  Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00
##                    NA's   :616        NA's   :638       NA's   :647
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
##  1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800
##  Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000
##  Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
##  3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600
##  Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
##  NA's   :682                         NA's   :395     NA's   :1210
##     Alcohol       LabelAppeal         AcidIndex         STARS
##  Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median :10.40   Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##  NA's   :653                                          NA's   :3359
```

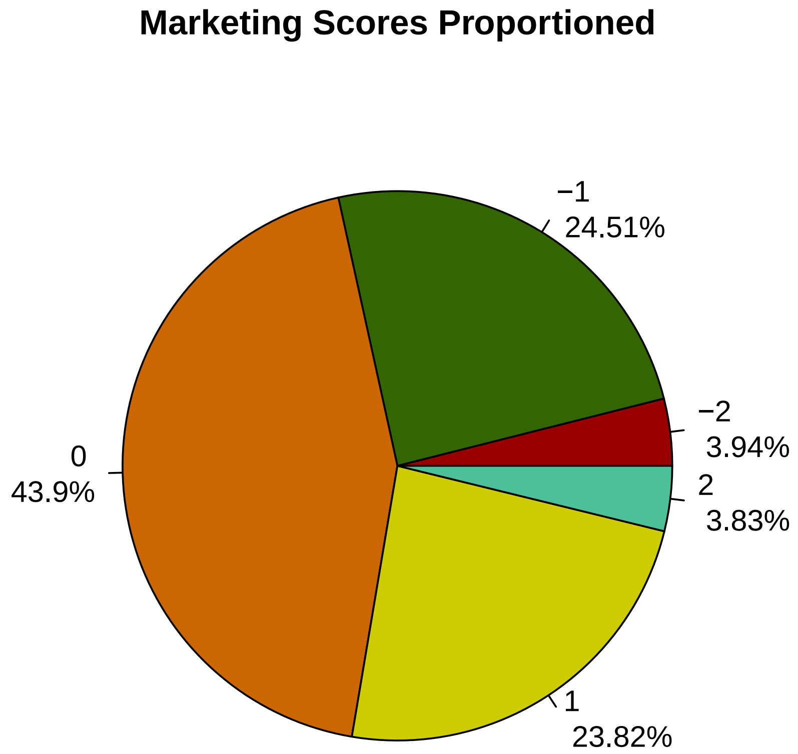
Top Amount of cases purchased





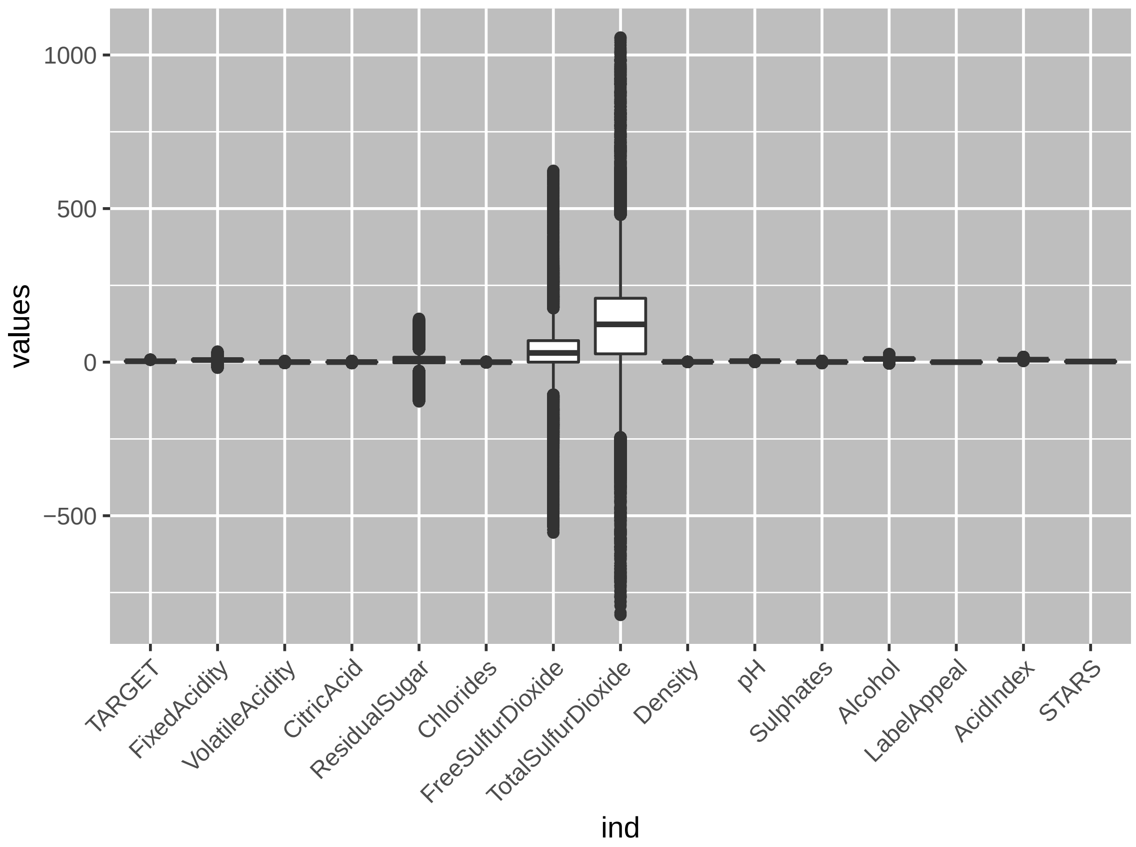
### Skewness in Data

A few of the variables have multimodal distribution (`TARGET`, `LabelAppeal`, `STARS`) while the others seem to be normally distrbuted due to bell curve they display. ### Marketing Scores

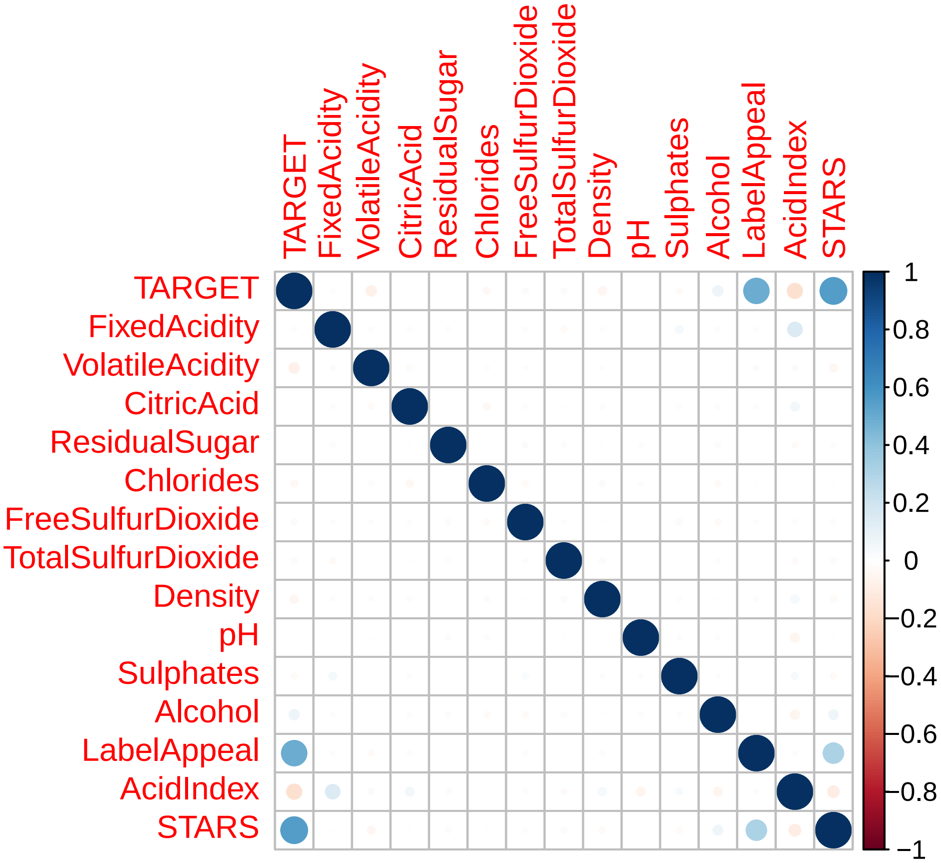


About 28% of the wine are not favored by customers based on their label designs ### Boxplot: Exploring Outliers

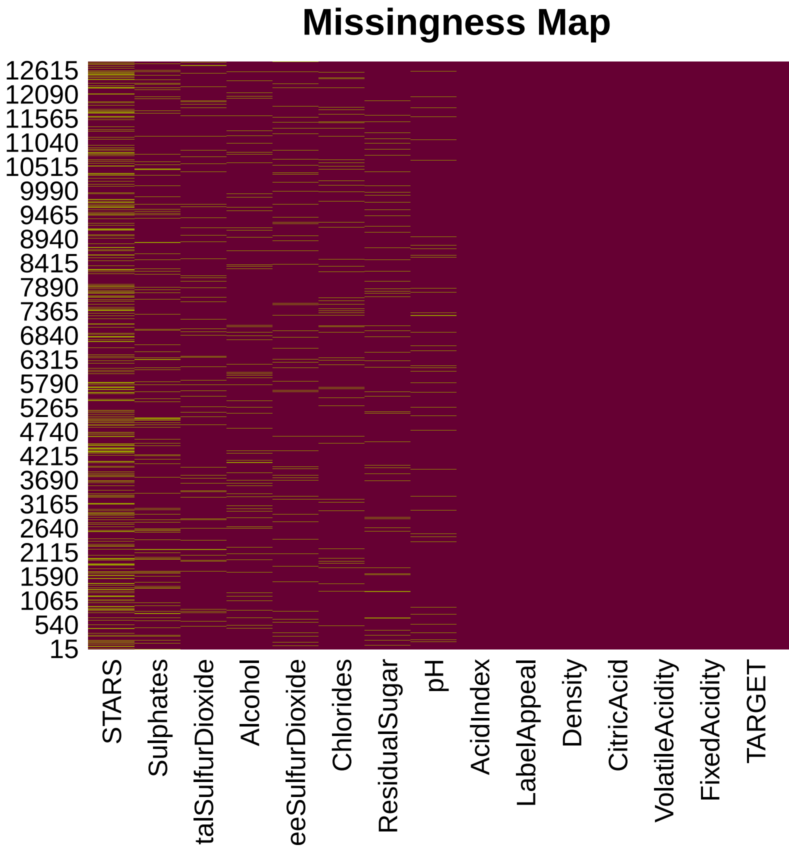
```
## Warning: Removed 8200 rows containing non-finite values (stat_boxplot).
```


values
1000
500
0
−500
ind
TARGET
FixedAcidity
VolatileAcidity
CitricAcid
ResidualSugar
Chlorides
FreeSulfurDioxide
TotalSulfurDioxide
Density
pH
Sulphates
Alcohol
LabelAppeal
AcidIndex
STARS

Correlation



We can see that there is come moderate but postive corrleation among the target variable and predictors



`STARS` and `LabelAppeal`. ### Missing Values
4% of the data is missing which we will later handle as we move forward ## DATA PREPARATION ### Handing Negative values ### Creates summary metrics table

Var	negative_value
Chlorides	0.26
ResidualSugar	0.26
FreeSulfurDioxide	0.25
CitricAcid	0.23
VolatileAcidity	0.22
TotalSulfurDioxide	0.21
Sulphates	0.20
FixedAcidity	0.13
Alcohol	0.01

New Variable variables

Conversion of negative values to absolute

```
##                    FixedAcidity VolatileAcidity CitricAcid ResidualSugar
## FixedAcidity               1.00            0.01       0.00          0.01
## VolatileAcidity            0.01            1.00       0.00          0.00
## CitricAcid                 0.00            0.00       1.00         -0.01
## ResidualSugar              0.01            0.00      -0.01          1.00
## Chlorides                  0.00            0.00       0.00          0.00
## FreeSulfurDioxide          0.00           -0.01       0.01          0.00
## TotalSulfurDioxide        -0.01           -0.03       0.01          0.01
## Density                    0.00            0.00      -0.01          0.00
## pH                         0.00            0.01       0.00          0.00
## Sulphates                  0.02            0.00       0.02          0.00
## Alcohol                   -0.01            0.01      -0.01         -0.01
## LabelAppeal                0.00           -0.02       0.02          0.00
## AcidIndex                  0.18            0.04       0.04         -0.01
## STARS                     -0.02           -0.03       0.00          0.01
## TARGET                    -0.05           -0.07       0.01          0.01
## BoundSulfurDioxide         0.00           -0.03       0.02          0.01
## PerVolume                 -0.49            0.47       0.00          0.00
##                    Chlorides FreeSulfurDioxide TotalSulfurDioxide Density    pH
## FixedAcidity            0.00              0.00              -0.01    0.00  0.00
## VolatileAcidity         0.00             -0.01              -0.03    0.00  0.01
## CitricAcid              0.00              0.01               0.01   -0.01  0.00
## ResidualSugar           0.00              0.00               0.01    0.00  0.00
## Chlorides               1.00              0.00              -0.01    0.02  0.01
## FreeSulfurDioxide       0.00              1.00               0.02    0.01  0.00
## TotalSulfurDioxide     -0.01              0.02               1.00    0.02  0.01
## Density                 0.02              0.01               0.02    1.00  0.01
## pH                      0.01              0.00               0.01    0.01  1.00
## Sulphates               0.02             -0.01              -0.01    0.01  0.01
## Alcohol                 0.00             -0.01              -0.03   -0.01 -0.01
## LabelAppeal            -0.01              0.01              -0.01   -0.01  0.00
## AcidIndex               0.03             -0.02              -0.04    0.04 -0.06
## STARS                  -0.01              0.00               0.01   -0.02  0.00
## TARGET                 -0.02              0.02               0.03   -0.04 -0.01
## BoundSulfurDioxide     -0.01              0.27               0.75    0.01  0.01
## PerVolume               0.01             -0.01              -0.02    0.00  0.02
##                    Sulphates Alcohol LabelAppeal AcidIndex STARS TARGET
## FixedAcidity            0.02   -0.01        0.00      0.18 -0.02  -0.05
## VolatileAcidity         0.00    0.01       -0.02      0.04 -0.03  -0.07
## CitricAcid              0.02   -0.01        0.02      0.04  0.00   0.01
## ResidualSugar           0.00   -0.01        0.00     -0.01  0.01   0.01
## Chlorides               0.02    0.00       -0.01      0.03 -0.01  -0.02
## FreeSulfurDioxide      -0.01   -0.01        0.01     -0.02  0.00   0.02
## TotalSulfurDioxide     -0.01   -0.03       -0.01     -0.04  0.01   0.03
## Density                 0.01   -0.01       -0.01      0.04 -0.02  -0.04
## pH                      0.01   -0.01        0.00     -0.06  0.00  -0.01
## Sulphates               1.00    0.00        0.00      0.03  0.00  -0.03
## Alcohol                 0.00    1.00        0.00     -0.04  0.07   0.06
## LabelAppeal             0.00    0.00        1.00      0.02  0.34   0.36
## AcidIndex               0.03   -0.04        0.02      1.00 -0.09  -0.25
## STARS                   0.00    0.07        0.34     -0.09  1.00   0.36
## TARGET                 -0.03    0.06        0.36     -0.25  0.36   1.00
## BoundSulfurDioxide     -0.01   -0.02       -0.01      0.00  0.00   0.01
## PerVolume               0.00    0.02       -0.01     -0.03 -0.01  -0.03
##                    BoundSulfurDioxide PerVolume
## FixedAcidity                     0.00     -0.49
## VolatileAcidity                 -0.03      0.47
## CitricAcid                       0.02      0.00
## ResidualSugar                    0.01      0.00
## Chlorides                       -0.01      0.01
## FreeSulfurDioxide                0.27     -0.01
## TotalSulfurDioxide               0.75     -0.02
## Density                          0.01      0.00
## pH                               0.01      0.02
## Sulphates                       -0.01      0.00
## Alcohol                         -0.02      0.02
## LabelAppeal                     -0.01     -0.01
## AcidIndex                        0.00     -0.03
## STARS                            0.00     -0.01
## TARGET                           0.01     -0.03
## BoundSulfurDioxide               1.00     -0.02
## PerVolume                       -0.02      1.00
##
## n= 12795
##
##
## P
##                    FixedAcidity VolatileAcidity CitricAcid ResidualSugar
## FixedAcidity                    0.2489          0.6205     0.4985
## VolatileAcidity    0.2489                       0.7764     0.9118
## CitricAcid         0.6205       0.7764                     0.1087
## ResidualSugar      0.4985       0.9118          0.1087
## Chlorides          0.6955       0.7050          0.6649     0.9794
## FreeSulfurDioxide  0.5905       0.1836          0.4856     0.6594
## TotalSulfurDioxide 0.1810       0.0021          0.5315     0.1422
## Density            0.9949       0.6341          0.2196     0.8290
## pH                 0.9041       0.1369          0.7575     0.7251
## Sulphates          0.0180       0.8996          0.0605     0.6276
## Alcohol            0.1504       0.0924          0.4265     0.4327
## LabelAppeal        0.8000       0.0825          0.0501     0.8457
## AcidIndex          0.0000       0.0000          0.0000     0.1534
## STARS              0.0048       0.0010          0.7936     0.2112
## TARGET             0.0000       0.0000          0.1145     0.5294
## BoundSulfurDioxide 0.9206       0.0015          0.0555     0.4094
## PerVolume          0.0000       0.0000          0.7383     0.8696
##                    Chlorides FreeSulfurDioxide TotalSulfurDioxide Density
## FixedAcidity       0.6955    0.5905            0.1810             0.9949
## VolatileAcidity    0.7050    0.1836            0.0021             0.6341
## CitricAcid         0.6649    0.4856            0.5315             0.2196
## ResidualSugar      0.9794    0.6594            0.1422             0.8290
## Chlorides                    0.6772            0.3179             0.0473
## FreeSulfurDioxide  0.6772                      0.0880             0.5348
## TotalSulfurDioxide 0.3179    0.0880                               0.0350
## Density            0.0473    0.5348            0.0350
## pH                 0.4103    0.7903            0.0913             0.3608
## Sulphates          0.0274    0.5023            0.1872             0.2285
## Alcohol            0.6878    0.3126            0.0006             0.5169
## LabelAppeal        0.4860    0.2027            0.1899             0.2892
## AcidIndex          0.0013    0.0125            0.0000             0.0000
## STARS              0.5135    0.6221            0.5716             0.0274
## TARGET             0.0089    0.0096            0.0002             0.0000
## BoundSulfurDioxide 0.3971    0.0000            0.0000             0.2230
## PerVolume          0.1932    0.3111            0.0879             0.8528
##                    pH     Sulphates Alcohol LabelAppeal AcidIndex STARS  TARGET
## FixedAcidity       0.9041 0.0180    0.1504  0.8000      0.0000    0.0048 0.0000
## VolatileAcidity    0.1369 0.8996    0.0924  0.0825      0.0000    0.0010 0.0000
## CitricAcid         0.7575 0.0605    0.4265  0.0501      0.0000    0.7936 0.1145
## ResidualSugar      0.7251 0.6276    0.4327  0.8457      0.1534    0.2112 0.5294
## Chlorides          0.4103 0.0274    0.6878  0.4860      0.0013    0.5135 0.0089
## FreeSulfurDioxide  0.7903 0.5023    0.3126  0.2027      0.0125    0.6221 0.0096
## TotalSulfurDioxide 0.0913 0.1872    0.0006  0.1899      0.0000    0.5716 0.0002
## Density            0.3608 0.2285    0.5169  0.2892      0.0000    0.0274 0.0000
## pH                        0.1518    0.3289  0.7202      0.0000    0.9060 0.3373
## Sulphates          0.1518           0.9905  0.6856      0.0001    0.7025 0.0025
## Alcohol            0.3289 0.9905            0.6871      0.0000    0.0000 0.0000
## LabelAppeal        0.7202 0.6856    0.6871              0.0051    0.0000 0.0000
## AcidIndex          0.0000 0.0001    0.0000  0.0051                0.0000 0.0000
## STARS              0.9060 0.7025    0.0000  0.0000      0.0000           0.0000
## TARGET             0.3373 0.0025    0.0000  0.0000      0.0000    0.0000
## BoundSulfurDioxide 0.1530 0.1575    0.0774  0.4717      0.6189    0.7471 0.4989
## PerVolume          0.0351 0.8772    0.0300  0.2185      0.0013    0.3883 0.0039
##                    BoundSulfurDioxide PerVolume
## FixedAcidity       0.9206             0.0000
## VolatileAcidity    0.0015             0.0000
## CitricAcid         0.0555             0.7383
## ResidualSugar      0.4094             0.8696
## Chlorides          0.3971             0.1932
## FreeSulfurDioxide  0.0000             0.3111
## TotalSulfurDioxide 0.0000             0.0879
## Density            0.2230             0.8528
## pH                 0.1530             0.0351
## Sulphates          0.1575             0.8772
## Alcohol            0.0774             0.0300
## LabelAppeal        0.4717             0.2185
## AcidIndex          0.6189             0.0013
## STARS              0.7471             0.3883
## TARGET             0.4989             0.0039
## BoundSulfurDioxide                    0.0085
## PerVolume          0.0085
```

BUILD MODELS

(at least two for each) ### Poisson Models

```
## 
## Call:
## glm(formula = TARGET ~ ., family = "poisson", data = wine_train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.9147  -0.4943   0.2180   0.6309   2.6165  
## 
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         1.810e+00  1.959e-01   9.242  < 2e-16 ***
## FixedAcidity       -1.047e-03  1.261e-03  -0.830 0.406545    
## VolatileAcidity    -5.792e-02  1.128e-02  -5.137 2.80e-07 ***
## CitricAcid          1.857e-02  8.290e-03   2.240 0.025084 *  
## ResidualSugar       6.505e-05  2.032e-04   0.320 0.748833    
## Chlorides          -3.047e-02  2.170e-02  -1.404 0.160216    
## FreeSulfurDioxide   1.630e-04  5.040e-05   3.233 0.001224 ** 
## TotalSulfurDioxide  2.449e-04  4.839e-05   5.060 4.18e-07 ***
## Density            -4.809e-01  1.921e-01  -2.504 0.012273 *  
## pH                 -2.344e-02  7.523e-03  -3.116 0.001834 ** 
## Sulphates          -1.665e-02  7.869e-03  -2.116 0.034350 *  
## Alcohol             6.097e-03  1.408e-03   4.331 1.48e-05 ***
## LabelAppeal         1.996e-01  6.116e-03  32.641  < 2e-16 ***
## AcidIndex          -1.239e-01  4.465e-03 -27.761  < 2e-16 ***
## STARS               1.617e-01  5.832e-03  27.724  < 2e-16 ***
## BoundSulfurDioxide -1.662e-04  4.449e-05  -3.736 0.000187 ***
## PerVolume          -3.281e-02  5.229e-02  -0.627 0.530385    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 18855  on 12778  degrees of freedom
## AIC: 50832
## 
## Number of Fisher Scoring iterations: 5
```

```
## 
## Call:
## glm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides + 
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates + 
##     Alcohol + LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide, 
##     family = "poisson", data = wine_train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.9108  -0.4940   0.2173   0.6300   2.6143  
## 
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)          1.807e+00  1.957e-01   9.232  < 2e-16 ***
## VolatileAcidity     -6.185e-02  9.416e-03  -6.569 5.07e-11 ***
## CitricAcid           1.860e-02  8.289e-03   2.244 0.024857 *  
## Chlorides           -3.070e-02  2.170e-02  -1.415 0.157008    
## FreeSulfurDioxide    1.632e-04  5.039e-05   3.239 0.001199 ** 
## TotalSulfurDioxide   2.453e-04  4.839e-05   5.068 4.01e-07 ***
## Density             -4.801e-01  1.920e-01  -2.500 0.012419 *  
## pH                  -2.361e-02  7.520e-03  -3.140 0.001692 ** 
## Sulphates           -1.681e-02  7.867e-03  -2.137 0.032596 *  
## Alcohol              6.091e-03  1.408e-03   4.327 1.51e-05 ***
## LabelAppeal          1.997e-01  6.115e-03  32.649  < 2e-16 ***
## AcidIndex           -1.245e-01  4.404e-03 -28.276  < 2e-16 ***
## STARS                1.617e-01  5.832e-03  27.733  < 2e-16 ***
## BoundSulfurDioxide  -1.663e-04  4.449e-05  -3.739 0.000185 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 18856  on 12781  degrees of freedom
## AIC: 50826
##
## Number of Fisher Scoring iterations: 5
```

Negative Binomial Models

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

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

```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = wine_train, init.theta = 32573.82814,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9145  -0.4943   0.2180   0.6308   2.6164
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.810e+00  1.959e-01   9.242  < 2e-16 ***
## FixedAcidity       -1.047e-03  1.262e-03  -0.830 0.406549
## VolatileAcidity    -5.792e-02  1.128e-02  -5.136 2.80e-07 ***
## CitricAcid          1.857e-02  8.291e-03   2.240 0.025092 *
## ResidualSugar       6.506e-05  2.032e-04   0.320 0.748812
## Chlorides          -3.047e-02  2.170e-02  -1.404 0.160226
## FreeSulfurDioxide   1.630e-04  5.040e-05   3.233 0.001225 **
## TotalSulfurDioxide  2.449e-04  4.839e-05   5.060 4.19e-07 ***
## Density            -4.809e-01  1.921e-01  -2.504 0.012276 *
## pH                 -2.344e-02  7.524e-03  -3.116 0.001835 **
## Sulphates          -1.665e-02  7.869e-03  -2.116 0.034356 *
## Alcohol             6.097e-03  1.408e-03   4.331 1.48e-05 ***
## LabelAppeal         1.996e-01  6.117e-03  32.639  < 2e-16 ***
## AcidIndex          -1.239e-01  4.465e-03 -27.760  < 2e-16 ***
## STARS               1.617e-01  5.833e-03  27.723  < 2e-16 ***
## BoundSulfurDioxide -1.662e-04  4.449e-05  -3.735 0.000187 ***
## PerVolume          -3.281e-02  5.229e-02  -0.627 0.530415
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(32573.83) family taken to be 1)
##
##     Null deviance: 22859  on 12794  degrees of freedom
## Residual deviance: 18854  on 12778  degrees of freedom
## AIC: 50834
##
## Number of Fisher Scoring iterations: 1
```

```
##
##
##               Theta:  32574
##           Std. Err.:  59283
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -50797.6
```

```
##
## Call:
## glm.nb(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide,
##     data = wine_train, init.theta = 32570.2802, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9106  -0.4940   0.2173   0.6300   2.6142
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.807e+00  1.957e-01   9.231  < 2e-16 ***
## VolatileAcidity    -6.186e-02  9.417e-03  -6.569 5.08e-11 ***
## CitricAcid          1.860e-02  8.290e-03   2.244 0.024865 *
## Chlorides          -3.070e-02  2.170e-02  -1.415 0.157018
## FreeSulfurDioxide   1.632e-04  5.040e-05   3.239 0.001199 **
## TotalSulfurDioxide  2.453e-04  4.839e-05   5.068 4.02e-07 ***
## Density            -4.801e-01  1.921e-01  -2.500 0.012422 *
## pH                 -2.361e-02  7.520e-03  -3.139 0.001692 **
## Sulphates          -1.681e-02  7.867e-03  -2.137 0.032601 *
## Alcohol             6.091e-03  1.408e-03   4.327 1.51e-05 ***
## LabelAppeal         1.997e-01  6.116e-03  32.648  < 2e-16 ***
## AcidIndex          -1.245e-01  4.404e-03 -28.275  < 2e-16 ***
## STARS               1.617e-01  5.832e-03  27.732  < 2e-16 ***
## BoundSulfurDioxide -1.663e-04  4.449e-05  -3.738 0.000185 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(32570.28) family taken to be 1)
##
##     Null deviance: 22859  on 12794  degrees of freedom
## Residual deviance: 18855  on 12781  degrees of freedom
## AIC: 50828
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  32570
##           Std. Err.:  59277
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -50798.43
```

Multiple Linear Regression Models

```
##
## Call:
## lm(formula = TARGET ~ ., data = wine_train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8909 -0.7215  0.3896  1.1253  4.4525
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.174e+00  5.642e-01   9.170  < 2e-16 ***
## FixedAcidity       -2.899e-03  3.624e-03  -0.800  0.42381
## VolatileAcidity    -1.567e-01  3.170e-02  -4.943 7.77e-07 ***
## CitricAcid          5.901e-02  2.429e-02   2.429  0.01514 *
## ResidualSugar       5.614e-05  5.893e-04   0.095  0.92409
## Chlorides          -1.058e-01  6.242e-02  -1.696  0.09000 .
## FreeSulfurDioxide   4.823e-04  1.482e-04   3.253  0.00114 **
## TotalSulfurDioxide  7.554e-04  1.425e-04   5.300 1.17e-07 ***
## Density            -1.371e+00  5.548e-01  -2.472  0.01346 *
## pH                 -5.957e-02  2.168e-02  -2.747  0.00602 **
## Sulphates          -4.886e-02  2.248e-02  -2.174  0.02973 *
## Alcohol             2.099e-02  4.065e-03   5.164 2.45e-07 ***
## LabelAppeal         6.000e-01  1.758e-02  34.131  < 2e-16 ***
## AcidIndex          -3.264e-01  1.145e-02 -28.501  < 2e-16 ***
## STARS2              7.165e-01  3.550e-02  20.186  < 2e-16 ***
## STARS3              1.063e+00  4.176e-02  25.447  < 2e-16 ***
## STARS4              1.562e+00  6.742e-02  23.167  < 2e-16 ***
## BoundSulfurDioxide -5.427e-04  1.314e-04  -4.131 3.63e-05 ***
## PerVolume          -1.362e-01  1.494e-01  -0.912  0.36187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.663 on 12776 degrees of freedom
## Multiple R-squared:  0.2562, Adjusted R-squared:  0.2551
## F-statistic: 244.5 on 18 and 12776 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide,
##     data = wine_train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8862 -0.7213  0.3906  1.1225  4.4558
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.1569379  0.5635980   9.150  < 2e-16 ***
## VolatileAcidity    -0.1725773  0.0265013  -6.512 7.69e-11 ***
## CitricAcid          0.0590612  0.0242842   2.432  0.01503 *
## Chlorides          -0.1065731  0.0624058  -1.708  0.08771 .
## FreeSulfurDioxide   0.0004826  0.0001482   3.256  0.00113 **
## TotalSulfurDioxide  0.0007556  0.0001425   5.303 1.16e-07 ***
## Density            -1.3684490  0.5546977  -2.467  0.01364 *
## pH                 -0.0600242  0.0216759  -2.769  0.00563 **
## Sulphates          -0.0492813  0.0224708  -2.193  0.02832 *
## Alcohol             0.0209603  0.0040646   5.157 2.55e-07 ***
## LabelAppeal         0.6001110  0.0175765  34.143  < 2e-16 ***
## AcidIndex          -0.3277181  0.0112473 -29.138  < 2e-16 ***
## STARS2              0.7168022  0.0354876  20.199  < 2e-16 ***
## STARS3              1.0628894  0.0417553  25.455  < 2e-16 ***
## STARS4              1.5621130  0.0674103  23.173  < 2e-16 ***
## BoundSulfurDioxide -0.0005419  0.0001313  -4.126 3.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.662 on 12779 degrees of freedom
## Multiple R-squared:  0.2561, Adjusted R-squared:  0.2552
## F-statistic: 293.3 on 15 and 12779 DF,  p-value: < 2.2e-16
```

SELECT MODELS

To select the models, we'll use AIC and MSE to measure accuracy of the predicted values. Below, the Poisson, Negative Binomial, and Multiple Linear Regression have been compared to select the model with the lowest AIC.

Comparison of Poisson Models

We'll need to compare the AIC's of each Possion Model.

```
## [1] 50831.51
```

```
## [1] 50826.34
```

Poisson Model 2 proves to have the lower AIC of the two, with a 50826.34 AIC. Below is the formula for Possion Model 2.

```
## [[1]]
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide
```

Comparison of Negative Binomial Models

We'll need to compare the AIC's of each Negative Binomial Model.

```
## [1] 50833.6
```

```
## [1] 50828.43
```

Negative Binomial Model 2 proves to have the lower AIC of the two, with a 50828.43 AIC. Below is the formula for Negative Binomial Model 2.

```
## [[1]]
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide
```

Comparsion of Multiple Linar Models

We'll need to compare the Adjusted R Squares of each Linear Model.

```
## [1] 0.2551296
```

```
## [1] 0.2552485
```

Linear Model 2 proves to have the higher Adjusted R Squares, with a value of 0.2552485. Below is the formula for Linear Model 2.

```
## [[1]]
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide
```

Mean Square Error The Mean Square Error measures the averaged square different between the etsimated values and the actual value. The lower the value of the MSE, the more accurately the model is able to predict the values.

$$\text{MSE} = \frac{1}{n}\sum(y - \hat{y})^2$$

Comparison of Possion and Negative Binomial Model's By evaluating the AIC's and MSE's of each model, we can choose the best one be looking at the lowest AIC and lowest MSE.

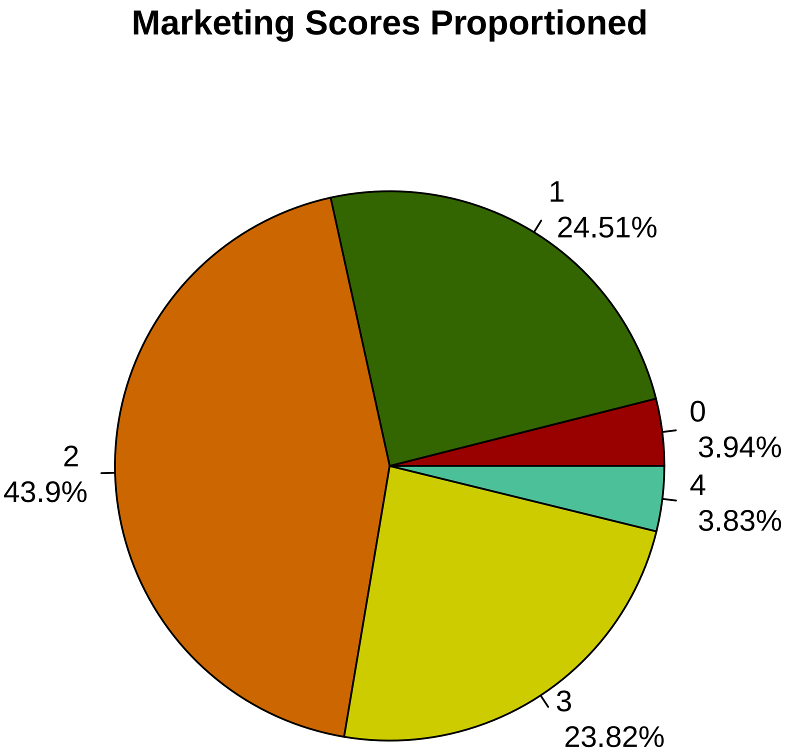
	Possion Model 1	Possion Model 2	Negative Binomial Model 1	Negative Binomial Model 2
MSE	7.07970144711237	7.07976751621997	7.07969989096655	7.07976596263758
AIC	50831.5145571202	50826.3420675487	50833.6039683312	50828.4314772116

Though Poisson Model 2 has a slightly higher MSE than Negative Binomial Model 2, it does have a lower AIC.

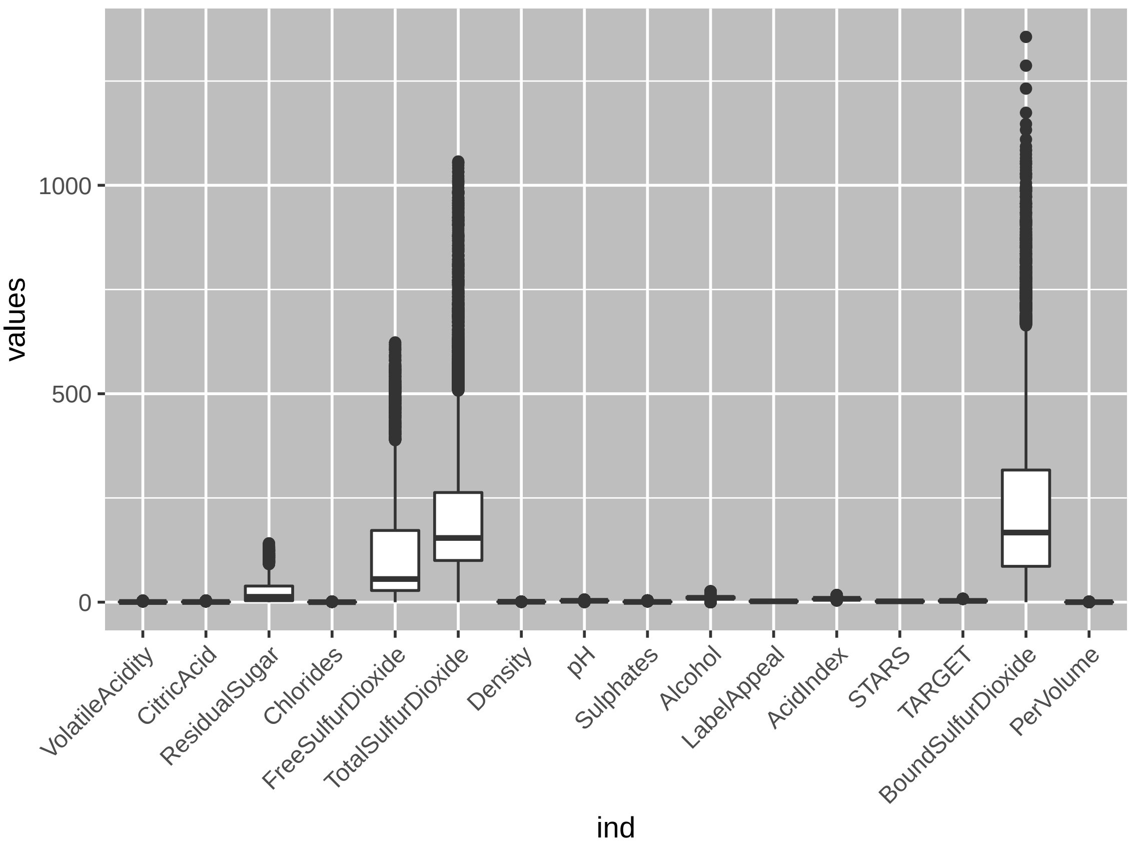
Transform Evaluation Data Set

```
## # A tibble: 10 x 18
##       IN TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
##    <int> <lgl>         <dbl>           <dbl>      <dbl>         <dbl>     <dbl>
##  1     3 NA              5.4           0.86        0.27          10.7     0.092
##  2     9 NA             12.4           0.385       0.76          19.7     1.17
##  3    10 NA              7.2           1.75        0.17          33       0.065
##  4    18 NA              6.2           0.1         1.8            1       0.179
##  5    21 NA             11.4           0.21        0.28           1.2     0.038
##  6    30 NA              17.6           0.04        1.15           1.4     0.535
##  7    31 NA              15.5           0.53        0.53           4.6     1.26
##  8    37 NA              15.9           1.19        1.14          31.9     0.299
##  9    39 NA              11.6           0.32        0.55          50.9     0.076
## 10    47 NA               3.8           0.22        0.31           7.7     0.039
## # ... with 11 more variables: FreeSulfurDioxide <dbl>,
## #   TotalSulfurDioxide <dbl>, Density <dbl>, pH <dbl>, Sulphates <dbl>,
## #   Alcohol <dbl>, LabelAppeal <int>, AcidIndex <int>, STARS <int>,
## #   BoundSulfurDioxide <dbl>, TARGET_FLAG <dbl>
```

A few of the variables have multimodal distribution (`TARGET`, `LabelAppeal`, `STARS`) while the others seem to be normally distrbuted due to bell curve they display. ### Marketing Scores



About 28% of the wine are not favored by customers based on their label designs ### Boxplot: Exploring Outliers



4% of the data is missing which we will later handle as we move forward

BUILD MODELS

(at least two for each) ### Poisson Models

```
##
## Call:
## glm(formula = TARGET ~ ., family = "poisson", data = wine_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9147  -0.4943   0.2180   0.6309   2.6165
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.810e+00  1.959e-01   9.242  < 2e-16 ***
## FixedAcidity       -1.047e-03  1.261e-03  -0.830 0.406545
## VolatileAcidity    -5.792e-02  1.128e-02  -5.137 2.80e-07 ***
## CitricAcid          1.857e-02  8.290e-03   2.240 0.025084 *
## ResidualSugar       6.505e-05  2.032e-04   0.320 0.748833
## Chlorides          -3.047e-02  2.170e-02  -1.404 0.160216
## FreeSulfurDioxide   1.630e-04  5.040e-05   3.233 0.001224 **
## TotalSulfurDioxide  2.449e-04  4.839e-05   5.060 4.18e-07 ***
## Density            -4.809e-01  1.921e-01  -2.504 0.012273 *
## pH                 -2.344e-02  7.523e-03  -3.116 0.001834 **
## Sulphates          -1.665e-02  7.869e-03  -2.116 0.034350 *
## Alcohol             6.097e-03  1.408e-03   4.331 1.48e-05 ***
## LabelAppeal         1.996e-01  6.116e-03  32.641  < 2e-16 ***
## AcidIndex          -1.239e-01  4.465e-03 -27.761  < 2e-16 ***
## STARS               1.617e-01  5.832e-03  27.724  < 2e-16 ***
## BoundSulfurDioxide -1.662e-04  4.449e-05  -3.736 0.000187 ***
## PerVolume          -3.281e-02  5.229e-02  -0.627 0.530385
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 18855  on 12778  degrees of freedom
## AIC: 50832
##
## Number of Fisher Scoring iterations: 5
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide,
##     family = "poisson", data = wine_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9108  -0.4940   0.2173   0.6300   2.6143
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.807e+00  1.957e-01   9.232  < 2e-16 ***
## VolatileAcidity    -6.185e-02  9.416e-03  -6.569 5.07e-11 ***
## CitricAcid          1.860e-02  8.289e-03   2.244 0.024857 *
## Chlorides          -3.070e-02  2.170e-02  -1.415 0.157008
## FreeSulfurDioxide   1.632e-04  5.039e-05   3.239 0.001199 **
## TotalSulfurDioxide  2.453e-04  4.839e-05   5.068 4.01e-07 ***
## Density            -4.801e-01  1.920e-01  -2.500 0.012419 *
## pH                 -2.361e-02  7.520e-03  -3.140 0.001692 **
## Sulphates          -1.681e-02  7.867e-03  -2.137 0.032596 *
## Alcohol             6.091e-03  1.408e-03   4.327 1.51e-05 ***
## LabelAppeal         1.997e-01  6.115e-03  32.649  < 2e-16 ***
## AcidIndex          -1.245e-01  4.404e-03 -28.276  < 2e-16 ***
## STARS               1.617e-01  5.832e-03  27.733  < 2e-16 ***
## BoundSulfurDioxide -1.663e-04  4.449e-05  -3.739 0.000185 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 18856  on 12781  degrees of freedom
## AIC: 50826
```

```
##
## Number of Fisher Scoring iterations: 5
```

Negative Binomial Models

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

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

```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = wine_train, init.theta = 32573.82814,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9145  -0.4943   0.2180   0.6308   2.6164
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.810e+00  1.959e-01   9.242  < 2e-16 ***
## FixedAcidity        -1.047e-03  1.262e-03  -0.830 0.406549
## VolatileAcidity     -5.792e-02  1.128e-02  -5.136 2.80e-07 ***
## CitricAcid           1.857e-02  8.291e-03   2.240 0.025092 *
## ResidualSugar        6.506e-05  2.032e-04   0.320 0.748812
## Chlorides           -3.047e-02  2.170e-02  -1.404 0.160226
## FreeSulfurDioxide    1.630e-04  5.040e-05   3.233 0.001225 **
## TotalSulfurDioxide  2.449e-04  4.839e-05   5.060 4.19e-07 ***
## Density             -4.809e-01  1.921e-01  -2.504 0.012276 *
## pH                  -2.344e-02  7.524e-03  -3.116 0.001835 **
## Sulphates           -1.665e-02  7.869e-03  -2.116 0.034356 *
## Alcohol              6.097e-03  1.408e-03   4.331 1.48e-05 ***
## LabelAppeal          1.996e-01  6.117e-03  32.639  < 2e-16 ***
## AcidIndex           -1.239e-01  4.465e-03 -27.760  < 2e-16 ***
## STARS                1.617e-01  5.833e-03  27.723  < 2e-16 ***
## BoundSulfurDioxide -1.662e-04  4.449e-05  -3.735 0.000187 ***
## PerVolume           -3.281e-02  5.229e-02  -0.627 0.530415
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(32573.83) family taken to be 1)
##
##     Null deviance: 22859  on 12794  degrees of freedom
## Residual deviance: 18854  on 12778  degrees of freedom
## AIC: 50834
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  32574
##           Std. Err.:  59283
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -50797.6
```

```
##
## Call:
## glm.nb(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide,
##     data = wine_train, init.theta = 32570.2802, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9106  -0.4940   0.2173   0.6300   2.6142
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.807e+00  1.957e-01   9.231  < 2e-16 ***
## VolatileAcidity    -6.186e-02  9.417e-03  -6.569 5.08e-11 ***
## CitricAcid          1.860e-02  8.290e-03   2.244 0.024865 *
## Chlorides          -3.070e-02  2.170e-02  -1.415 0.157018
## FreeSulfurDioxide   1.632e-04  5.040e-05   3.239 0.001199 **
## TotalSulfurDioxide  2.453e-04  4.839e-05   5.068 4.02e-07 ***
## Density            -4.801e-01  1.921e-01  -2.500 0.012422 *
## pH                 -2.361e-02  7.520e-03  -3.139 0.001692 **
## Sulphates          -1.681e-02  7.867e-03  -2.137 0.032601 *
## Alcohol             6.091e-03  1.408e-03   4.327 1.51e-05 ***
## LabelAppeal         1.997e-01  6.116e-03  32.648  < 2e-16 ***
## AcidIndex          -1.245e-01  4.404e-03 -28.275  < 2e-16 ***
## STARS               1.617e-01  5.832e-03  27.732  < 2e-16 ***
## BoundSulfurDioxide -1.663e-04  4.449e-05  -3.738 0.000185 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(32570.28) family taken to be 1)
##
##     Null deviance: 22859  on 12794  degrees of freedom
## Residual deviance: 18855  on 12781  degrees of freedom
## AIC: 50828
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  32570
##           Std. Err.:  59277
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -50798.43
```

Multiple Linear Regression Models

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

```
## lm(formula = TARGET ~ ., data = wine_train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8909 -0.7215  0.3896  1.1253  4.4525
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.174e+00  5.642e-01   9.170  < 2e-16 ***
## FixedAcidity       -2.899e-03  3.624e-03  -0.800  0.42381
## VolatileAcidity    -1.567e-01  3.170e-02  -4.943 7.77e-07 ***
## CitricAcid          5.901e-02  2.429e-02   2.429  0.01514 *
## ResidualSugar       5.614e-05  5.893e-04   0.095  0.92409
## Chlorides          -1.058e-01  6.242e-02  -1.696  0.09000 .
## FreeSulfurDioxide   4.823e-04  1.482e-04   3.253  0.00114 **
## TotalSulfurDioxide  7.554e-04  1.425e-04   5.300 1.17e-07 ***
## Density            -1.371e+00  5.548e-01  -2.472  0.01346 *
## pH                 -5.957e-02  2.168e-02  -2.747  0.00602 **
## Sulphates          -4.886e-02  2.248e-02  -2.174  0.02973 *
## Alcohol             2.099e-02  4.065e-03   5.164 2.45e-07 ***
## LabelAppeal         6.000e-01  1.758e-02  34.131  < 2e-16 ***
## AcidIndex          -3.264e-01  1.145e-02 -28.501  < 2e-16 ***
## STARS2              7.165e-01  3.550e-02  20.186  < 2e-16 ***
## STARS3              1.063e+00  4.176e-02  25.447  < 2e-16 ***
## STARS4              1.562e+00  6.742e-02  23.167  < 2e-16 ***
## BoundSulfurDioxide -5.427e-04  1.314e-04  -4.131 3.63e-05 ***
## PerVolume          -1.362e-01  1.494e-01  -0.912  0.36187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.663 on 12776 degrees of freedom
## Multiple R-squared:  0.2562, Adjusted R-squared:  0.2551
## F-statistic: 244.5 on 18 and 12776 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide,
##     data = wine_train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8862 -0.7213  0.3906  1.1225  4.4558
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.1569379  0.5635980   9.150  < 2e-16 ***
## VolatileAcidity    -0.1725773  0.0265013  -6.512 7.69e-11 ***
## CitricAcid          0.0590612  0.0242842   2.432  0.01503 *
## Chlorides          -0.1065731  0.0624058  -1.708  0.08771 .
## FreeSulfurDioxide   0.0004826  0.0001482   3.256  0.00113 **
## TotalSulfurDioxide  0.0007556  0.0001425   5.303 1.16e-07 ***
## Density            -1.3684490  0.5546977  -2.467  0.01364 *
## pH                 -0.0600242  0.0216759  -2.769  0.00563 **
## Sulphates          -0.0492813  0.0224708  -2.193  0.02832 *
## Alcohol             0.0209603  0.0040646   5.157 2.55e-07 ***
## LabelAppeal         0.6001110  0.0175765  34.143  < 2e-16 ***
## AcidIndex          -0.3277181  0.0112473 -29.138  < 2e-16 ***
## STARS2              0.7168022  0.0354876  20.199  < 2e-16 ***
## STARS3              1.0628894  0.0417553  25.455  < 2e-16 ***
## STARS4              1.5621130  0.0674103  23.173  < 2e-16 ***
## BoundSulfurDioxide -0.0005419  0.0001313  -4.126 3.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.662 on 12779 degrees of freedom
## Multiple R-squared:  0.2561, Adjusted R-squared:  0.2552
## F-statistic: 293.3 on 15 and 12779 DF,  p-value: < 2.2e-16
```

SELECT MODELS

To select the models, we'll use AIC and MSE to measure accuracy of the predicted values. Below, the Poisson, Negative Binomial, and Multiple Linear Regression have been compared to select the model with the lowest AIC.

Comparison of Poisson Models

SELECT MODELS

```
## [[1]]
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide
```

```
## [[1]]
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide
```

```
## [[1]]
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide
```

We'll need to compare the AIC's of each Possion Model.

```
## [1] 50831.51
```

```
## [1] 50826.34
```

Poisson Model 2 proves to have the lower AIC of the two, with a 50826.34 AIC. Below is the formula for Possion Model 2.

```
## [[1]]
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide
```

Comparison of Negative Binomial Models

We'll need to compare the AIC's of each Negative Binomial Model.

```
## [1] 50833.6
```

```
## [1] 50828.43
```

Negative Binomial Model 2 proves to have the lower AIC of the two, with a 50828.43 AIC. Below is the formula for Negative Binomial Model 2.

```
## [[1]]
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide
```

Comparsion of Multiple Linar Models

We'll need to compare the Adjusted R Squares of each Linear Model.

```
## [1] 0.2551296
```

```
## [1] 0.2552485
```

Linear Model 2 proves to have the higher Adjusted R Squares, with a value of 0.2552485. Below is the formula for Linear Model 2.

```
## [[1]]
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS + BoundSulfurDioxide
```

Mean Square Error The Mean Square Error measures the averaged square different between the etsimated values and the actual value. The lower the value of the MSE, the more accurately the model is able to predict the values.

$$\mathrm{MSE} = \frac{1}{n}\sum(y - \hat{y})^2$$

Comparison of Possion and Negative Binomial Model's By evaluating the AIC's and MSE's of each model, we can choose the best one be looking at the lowest AIC and lowest MSE.

	Possion Model 1	Possion Model 2	Negative Binomial Model 1	Negative Binomial Model 2
MSE	7.07970144711237	7.07976751621997	7.07969989096655	7.07976596263758
AIC	50831.5145571202	50826.3420675487	50833.6039683312	50828.4314772116

Though Poisson Model 2 has a slightly higher MSE than Negative Binomial Model 2, it does have a lower AIC.

Transform Evaluation Data Set

```
## # A tibble: 10 x 18
##       IN TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
##    <int> <lgl>         <dbl>           <dbl>      <dbl>         <dbl>     <dbl>
##  1     3 NA              5.4           0.86        0.27          10.7     0.092
##  2     9 NA             12.4           0.385       0.76          19.7     1.17
##  3    10 NA              7.2           1.75        0.17          33       0.065
##  4    18 NA              6.2           0.1         1.8            1       0.179
##  5    21 NA             11.4           0.21        0.28           1.2     0.038
##  6    30 NA             17.6           0.04        1.15           1.4     0.535
##  7    31 NA             15.5           0.53        0.53           4.6     1.26
##  8    37 NA             15.9           1.19        1.14          31.9     0.299
##  9    39 NA             11.6           0.32        0.55          50.9     0.076
## 10    47 NA              3.8           0.22        0.31           7.7     0.039
## # ... with 11 more variables: FreeSulfurDioxide <dbl>,
## #   TotalSulfurDioxide <dbl>, Density <dbl>, pH <dbl>, Sulphates <dbl>,
## #   Alcohol <dbl>, LabelAppeal <int>, AcidIndex <int>, STARS <int>,
## #   BoundSulfurDioxide <dbl>, TARGET_FLAG <dbl>
```

Appendix

Project Source Code Evaluation CSV File