Count Regression - Wine Dataset

Shyam BV

April 27, 2018

Contents

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET	Number of Cases Purchased	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.	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	A high number of stars suggests high sales
Sulphates	Sulfate conten of wine	
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	
VolatileAcidity	Volatile Acid content of wine	
pH	pH of wine	

Figure 1: Data Definition.

1 Count Regression Model : Predicting the cases of wine that will be sold using the chemical properties of wine

---

1.1 Overview

Purpose of this assignment is to explore, analyze and model a dataset containing information on wine chemical information and wine cases sold. Each record has a response variable with the number of cases sold on that particular wine. The main objective is to build a Count Regression model on the training data set to predict the wine cases which will be sold depending on chemical properties.

1.2 Data Exploration

Since the data is observational, we will analyze all the missing values and explore the summary statistics of each predictor variable. Also plot charts and see how individual variable affect the wine cases sold.

1.2.1 Summary Statistics

Below are the inference from the summary:

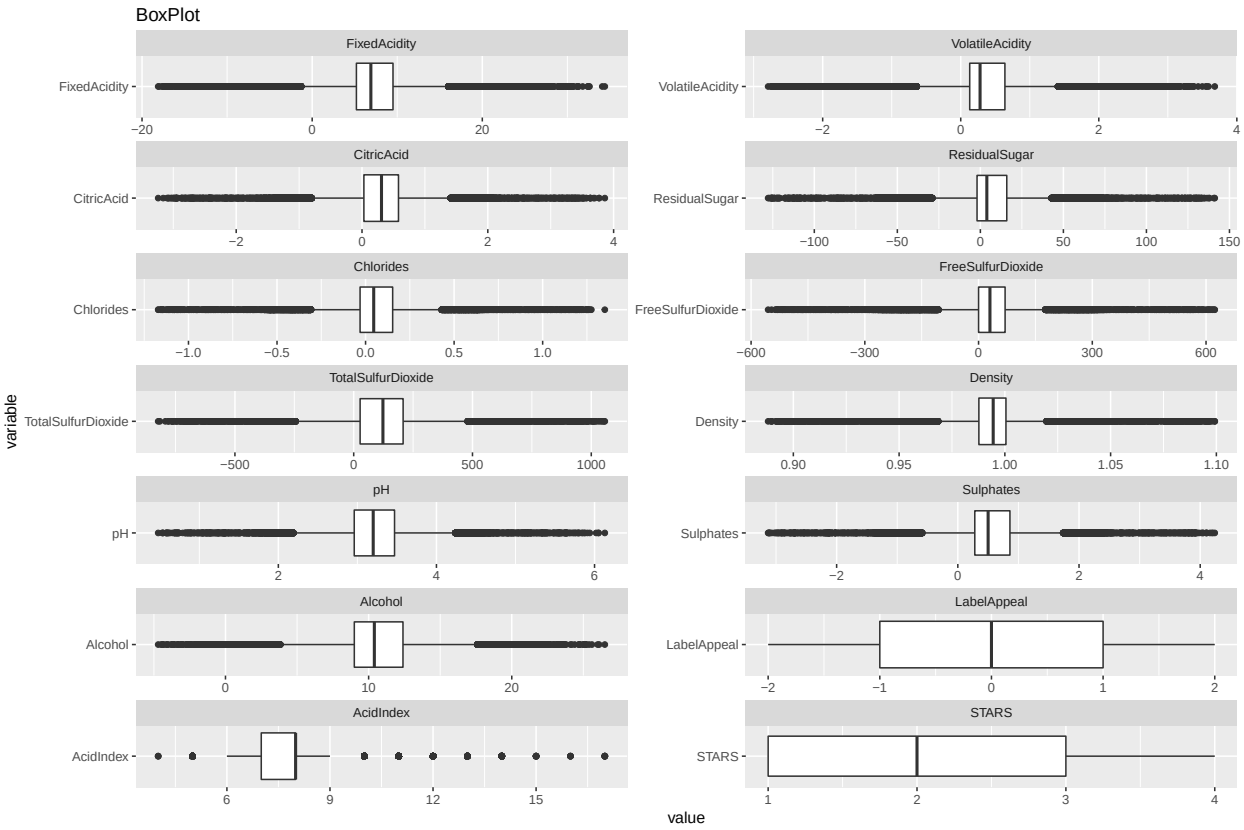
1. Summary data shows that there are some variables(`ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol, STARS`) which are having missing values. We need to perform some form of imputation for those.
2. `TARGET` is in the number of cases sold.
3. Variables are in different scales. Some observations starts from negative value and some from 0 or positive value.
4. There is an ordinal categorical predictors called `STARS`. This is the rating of the wine.
5. `LabelAppeal` is a label design score of the customers. It can be treated as an categorical variable.

1.2.2 Predictor variable Plots

As the `Index` column does not add any value to the dataset, we will remove it.

1.2.2.1 Box plot

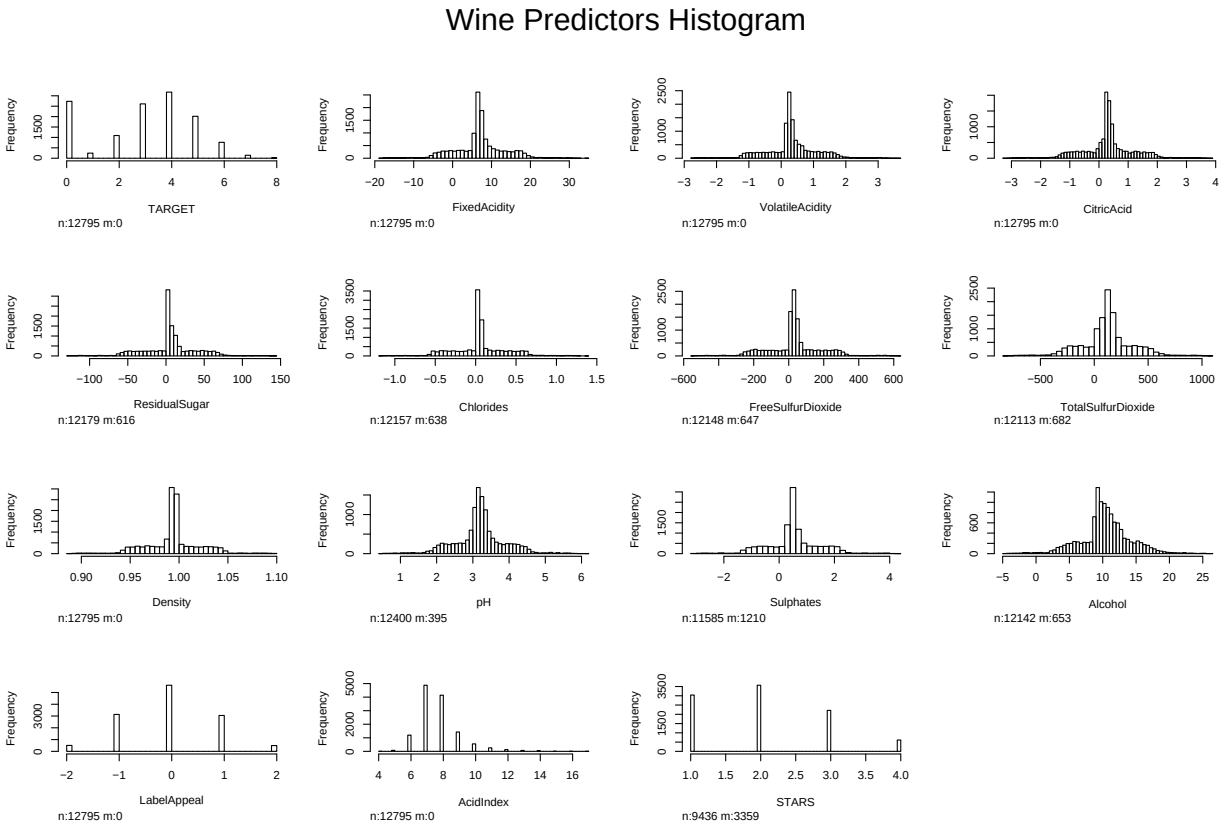
Below is the box plot of all the predictor variables.



Above plots show that different variables have many observations outside the mean/median range.

1.2.3 Histogram of Variables

Below plots show the histogram plot of the predictor variables.

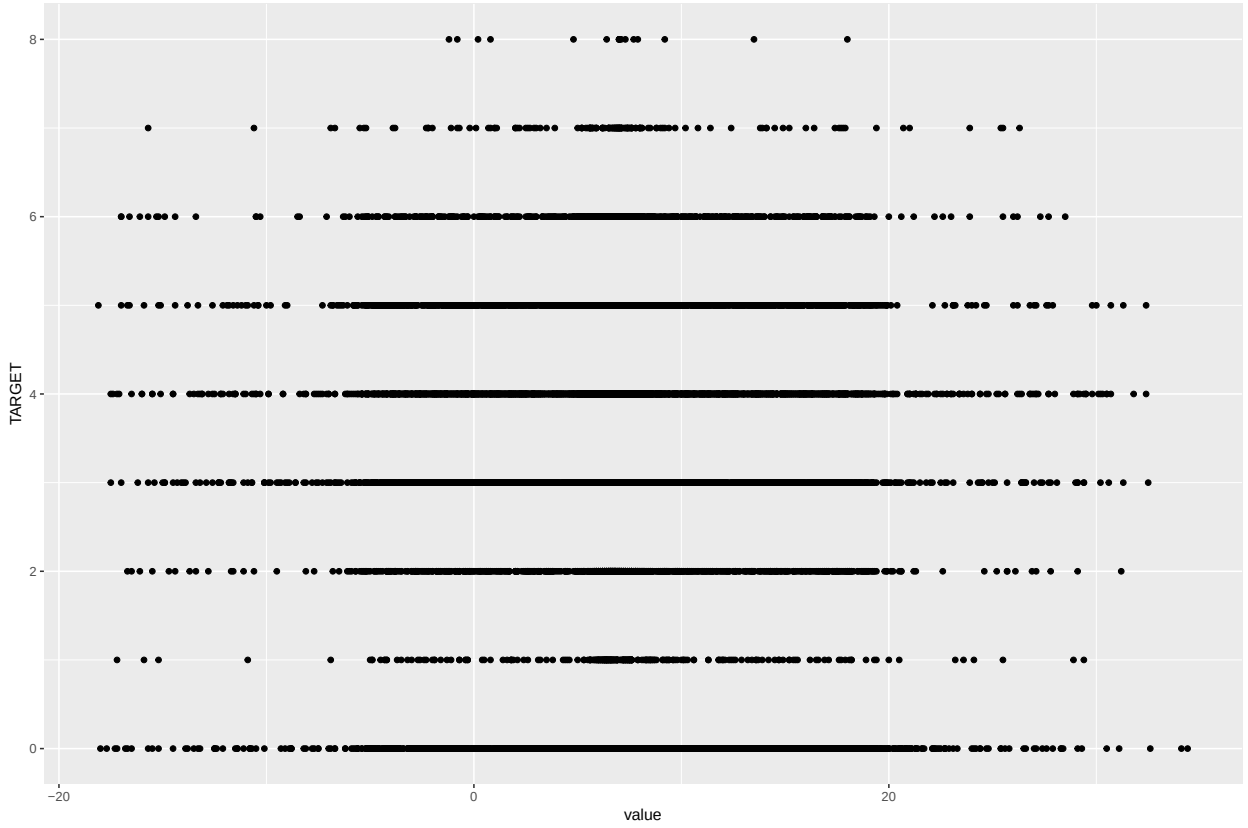


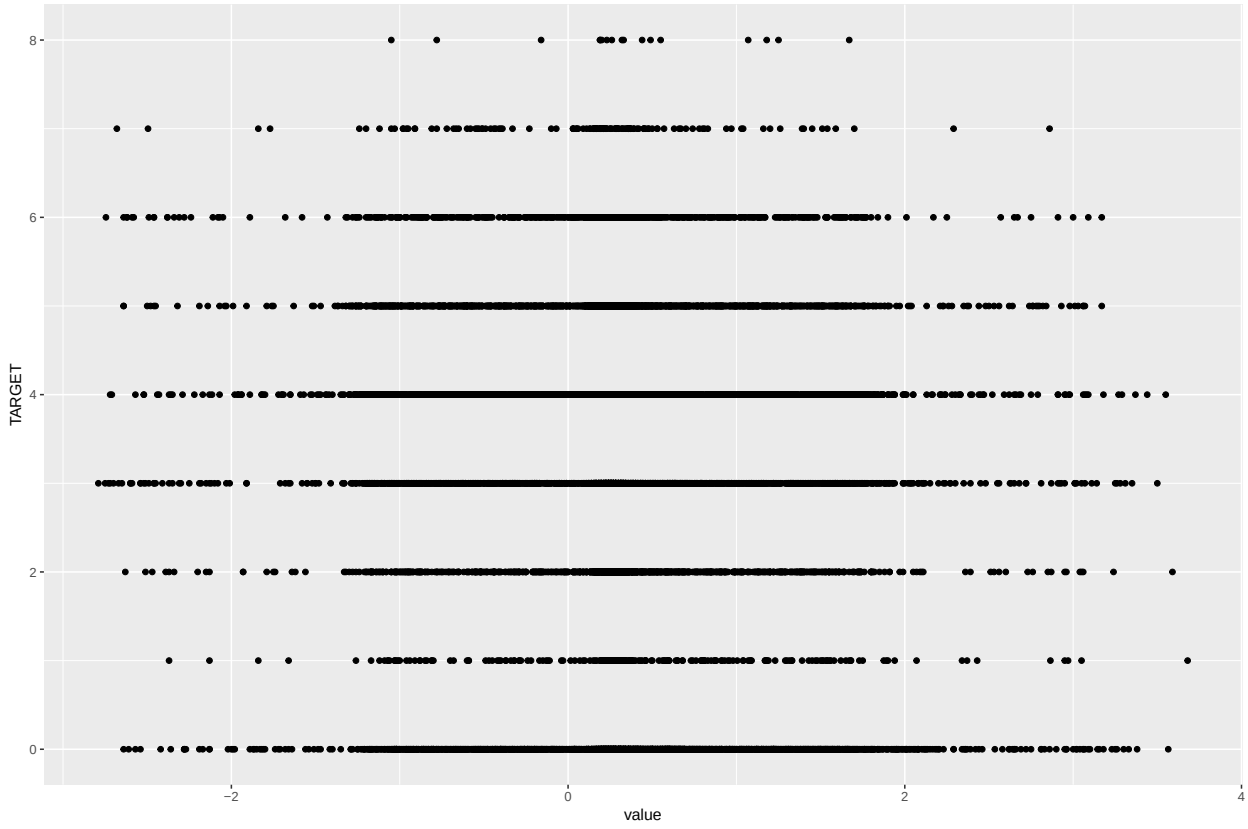
Histogram plots show that the varibles `FixedAcidity, VolatileAcidity, CitricAcid, ResidualSugar, Chlorides, FreeSulfurDioxide, Density, pH, Sulphates,Alcohol, AcidIndex` are heavily centered around the mean.

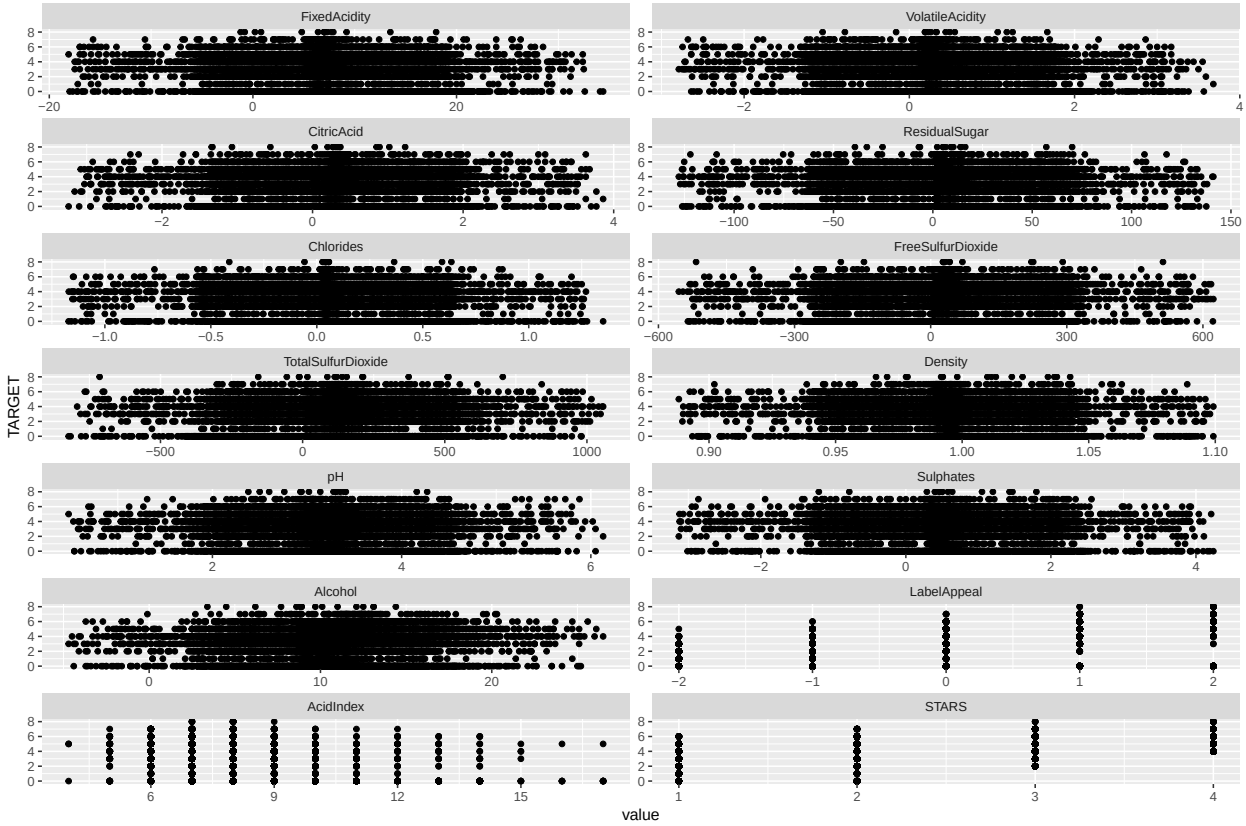
`LabelAppeal and STARS` are ordinal categorical variable.

1.2.3.1 Target vs Other Variables

Below plots show how each predictor variables impact the target variable.


TARGET
8
6
4
2
0
−20
0
20
value

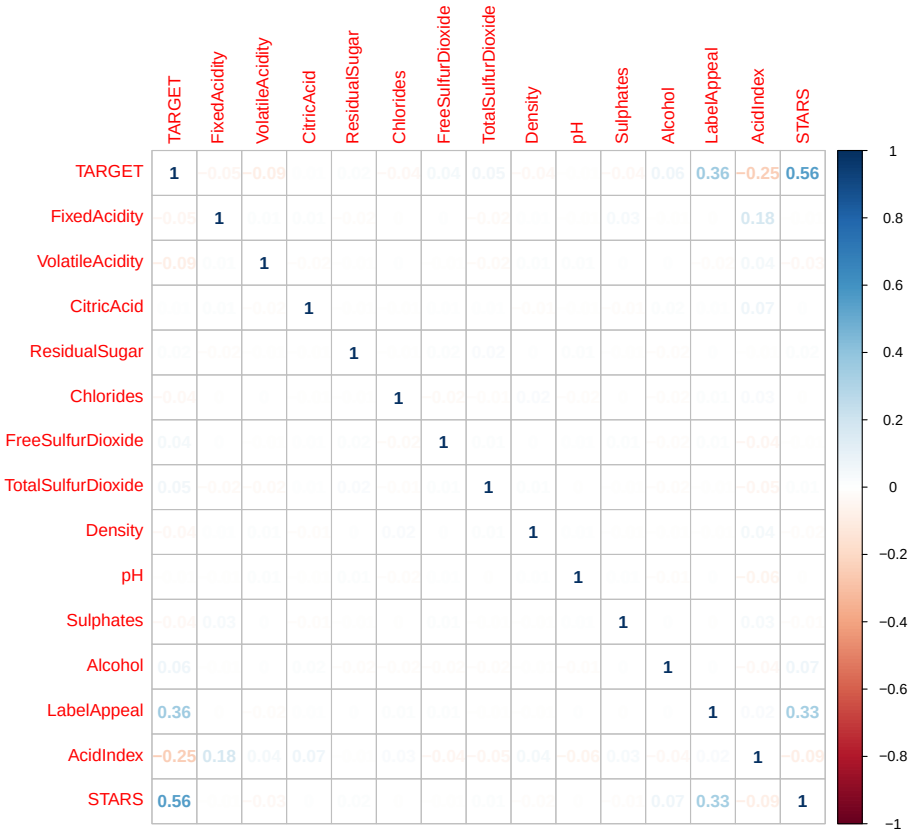

TARGET
8
6
4
2
0
−2
0
2
4
value



Above plot does not explain the individual relationship between the variables. TARGET variable is evenly distributed.

1.2.3.2 Correlation matrix

Lets check the relation between the variables.This will show how each variables are correlated.




```
##          Var1        Var2      Freq
## 1       STARS      TARGET 0.5546857
## 2      TARGET       STARS 0.5546857
## 3 LabelAppeal      TARGET 0.4979465
## 4      TARGET LabelAppeal 0.4979465
## 5       STARS LabelAppeal 0.3188970
## 6 LabelAppeal       STARS 0.3188970
```

It looks like the highest correlation number is around 0.55. So we can conclude that the variables are not highly correlated and does not suffer from multi-collinearity.

1.2.4 Target Variable analysis

Lets dig deep into the target variable and analyze the values in it.

```
## [1] "Unique TARGET Counts:"
```

```
## [1] 3 5 4 0 6 7 2 1 8
```

```
## # A tibble: 3 x 2
##   STARS `n()`
##   <int> <int>
## 1     1   607
## 2     2    89
## 3    NA  2038
```

Above table shows that the most of the TARGET values with 0 has zero stars. This seems to be some sort of pattern between those variables. So we will assign 0 for all the missing values in STARS variable.

1.3 Data Preparation

In the data analysis we found out that there are some variables with missing values. So in this section, we will address it and we will also the transformation of variables.

1.3.1 Dropping NA rows

We need to drop the observations which has multiple `NA` values.

1.3.2 Imputation

As a next step we need to create a statergy and apply impuation for the missing values variables `ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol`.

Now we will impute 0 for all the missing values in STARS variable.

1.3.3 Ordered variables

As a next step, we need to convert the categorial variables to the dummy variables. Dummy variables will create a unique variable for each category.

Here the `LabelAppeal and STARS` are considered as ordinal categorical variables. Because LabelAppeal rating depends on the customer who likes the label. STARS is the rating of the wine. So it cannot be treated as regular categorical variable.

1.3.4 Training and Test dataset

Now we will split the dataset into training and testing dataset. This will allow us to validate the model and tune the model.

1.4 Build Models

In this section, we will build different models to predict the wine cases. As the TARGET variable is a count of cases, we need to use count regression techniques. Also the TARGET variable can be transformed into different categories and create a multinomial model. Lets develop different models one by one.

1.4.1 Model 1 - Poisson Regression models

A random variable Y is said to have a poisson distribution with parameter μ if it takes the integer values y=0, 2, 4, 6 with probability

$Pr(Y = y) = (e^{-}\mu\mu^{y})/y!$

Lets verify if the mean and the variance of the response variable are same.

```
## [1] 3.042099
```

```
## [1] 3.747278
```

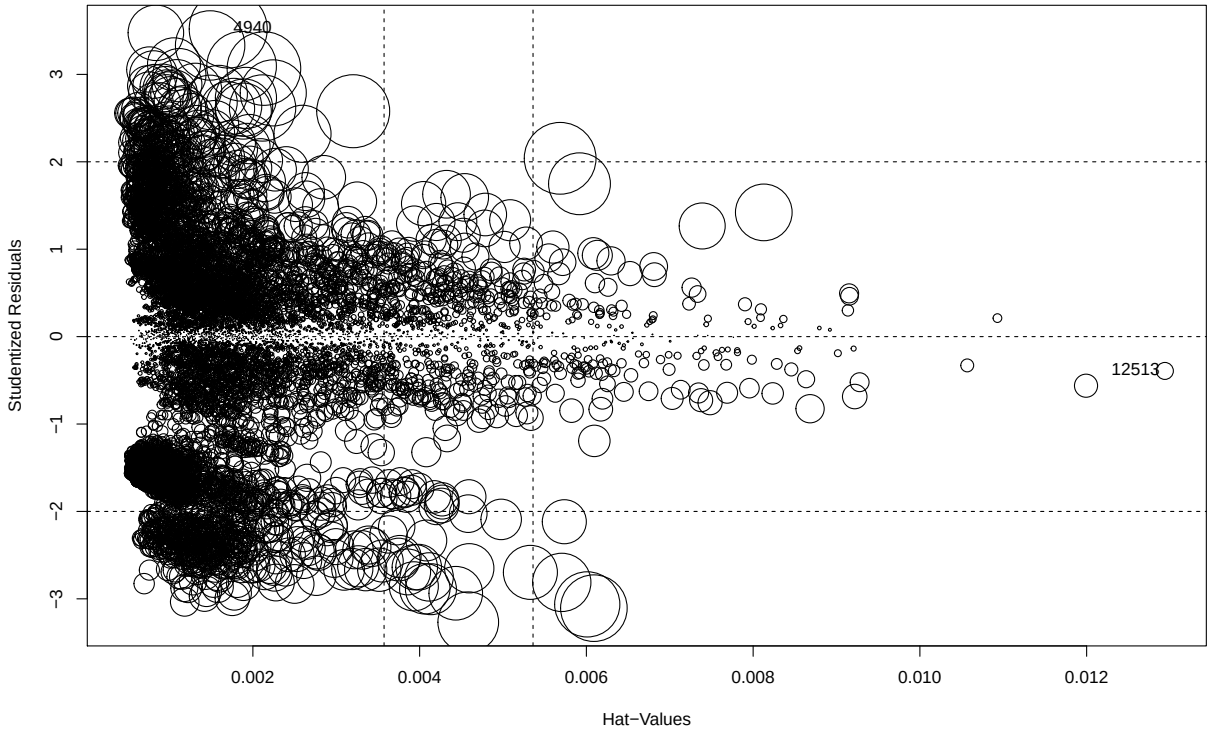
It seems there is a difference between the mean and variance. It suffers from overdispersion. However, it is a minor difference. We will assume the dispersion as 1.

```
##
## Call:
## glm(formula = TARGET ~ ., family = "poisson", data = train.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2603  -0.6480  -0.0029   0.4518   3.4998
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.914e+00  2.321e-01   8.247  < 2e-16 ***
## FixedAcidity         5.681e-04  9.741e-04   0.583  0.55978
## VolatileAcidity     -3.157e-02  7.770e-03  -4.063 4.85e-05 ***
## CitricAcid           6.054e-03  7.069e-03   0.856  0.39173
## ResidualSugar        7.444e-05  1.830e-04   0.407  0.68407
## Chlorides           -4.900e-02  1.951e-02  -2.512  0.01202 *
## FreeSulfurDioxide    9.298e-05  4.206e-05   2.211  0.02705 *
## TotalSulfurDioxide   8.559e-05  2.699e-05   3.171  0.00152 **
## Density             -2.169e-01  2.286e-01  -0.948  0.34291
## pH                  -6.037e-03  9.086e-03  -0.664  0.50643
## Sulphates           -1.390e-02  6.827e-03  -2.037  0.04169 *
## Alcohol              3.673e-03  1.681e-03   2.185  0.02886 *
## LabelAppeal.L        5.871e-01  3.343e-02  17.564  < 2e-16 ***
## LabelAppeal.Q       -8.478e-02  2.807e-02  -3.020  0.00253 **
## LabelAppeal.C        2.323e-02  1.972e-02   1.178  0.23888
## LabelAppeal^4        8.545e-03  1.245e-02   0.687  0.49238
## AcidIndex           -8.481e-02  5.466e-03 -15.514  < 2e-16 ***
## STARS.L              9.613e-01  1.964e-02  48.937  < 2e-16 ***
## STARS.Q             -3.994e-01  1.689e-02 -23.638  < 2e-16 ***
## STARS.C              1.379e-01  1.447e-02   9.531  < 2e-16 ***
## STARS^4             -1.362e-02  1.182e-02  -1.152  0.24912
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16086.3  on 8954  degrees of freedom
## Residual deviance:  9563.2  on 8934  degrees of freedom
## AIC: 31986
##
## Number of Fisher Scoring iterations: 6
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Sulphates + Alcohol + LabelAppeal +
##     AcidIndex + STARS, family = "poisson", data = train.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2635  -0.6462  -0.0035   0.4480   3.5135
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)         1.679e+00  4.709e-02  35.650  < 2e-16 ***
## VolatileAcidity    -3.176e-02  7.770e-03  -4.088 4.35e-05 ***
## Chlorides          -4.931e-02  1.950e-02  -2.529  0.01144 *
## FreeSulfurDioxide   9.355e-05  4.203e-05   2.226  0.02602 *
## TotalSulfurDioxide  8.531e-05  2.698e-05   3.162  0.00157 **
## Sulphates          -1.375e-02  6.822e-03  -2.015  0.04390 *
## Alcohol             3.728e-03  1.680e-03   2.220  0.02644 *
## LabelAppeal.L       5.864e-01  3.342e-02  17.548  < 2e-16 ***
## LabelAppeal.Q      -8.497e-02  2.807e-02  -3.027  0.00247 **
## LabelAppeal.C       2.294e-02  1.972e-02   1.163  0.24475
## LabelAppeal^4       8.307e-03  1.244e-02   0.668  0.50435
## AcidIndex          -8.408e-02  5.379e-03 -15.632  < 2e-16 ***
## STARS.L             9.617e-01  1.964e-02  48.971  < 2e-16 ***
## STARS.Q            -3.997e-01  1.689e-02 -23.664  < 2e-16 ***
## STARS.C             1.377e-01  1.446e-02   9.523  < 2e-16 ***
## STARS^4            -1.337e-02  1.182e-02  -1.131  0.25809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16086.3  on 8954  degrees of freedom
## Residual deviance:  9565.8  on 8939  degrees of freedom
## AIC: 31978
##
## Number of Fisher Scoring iterations: 6
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: TARGET
##
## Terms added sequentially (first to last)
##
##
##                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                8954    16086.3
## VolatileAcidity     1     93.6      8953    15992.8 < 2.2e-16 ***
## Chlorides           1     18.4      8952    15974.3 1.760e-05 ***
## FreeSulfurDioxide   1     21.6      8951    15952.8 3.424e-06 ***
## TotalSulfurDioxide  1     23.1      8950    15929.7 1.575e-06 ***
## Sulphates           1     20.0      8949    15909.7 7.659e-06 ***
## Alcohol             1     31.4      8948    15878.3 2.061e-08 ***
## LabelAppeal         4   1544.7      8944    14333.6 < 2.2e-16 ***
## AcidIndex           1    749.8      8943    13583.8 < 2.2e-16 ***
## STARS               4   4018.0      8939     9565.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
##          StudRes         Hat        CookD
## 4940   3.5192512 0.001704163 0.0025453042
## 12513 -0.3926132 0.012938389 0.0001205769
```

Removing the outliers and refitting the model

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Sulphates + Alcohol + LabelAppeal +
##     AcidIndex + STARS, family = "poisson", data = train.df[-c(4940,
##     12513)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2635  -0.6462  -0.0035   0.4480   3.5135
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.679e+00  4.709e-02  35.650  < 2e-16 ***
## VolatileAcidity     -3.176e-02  7.770e-03  -4.088 4.35e-05 ***
## Chlorides           -4.931e-02  1.950e-02  -2.529  0.01144 *
## FreeSulfurDioxide    9.355e-05  4.203e-05   2.226  0.02602 *
## TotalSulfurDioxide   8.531e-05  2.698e-05   3.162  0.00157 **
## Sulphates           -1.375e-02  6.822e-03  -2.015  0.04390 *
## Alcohol              3.728e-03  1.680e-03   2.220  0.02644 *
## LabelAppeal.L        5.864e-01  3.342e-02  17.548  < 2e-16 ***
## LabelAppeal.Q       -8.497e-02  2.807e-02  -3.027  0.00247 **
```

```
## LabelAppeal.C       2.294e-02  1.972e-02   1.163  0.24475
## LabelAppeal^4       8.307e-03  1.244e-02   0.668  0.50435
## AcidIndex          -8.408e-02  5.379e-03 -15.632  < 2e-16 ***
## STARS.L             9.617e-01  1.964e-02  48.971  < 2e-16 ***
## STARS.Q            -3.997e-01  1.689e-02 -23.664  < 2e-16 ***
## STARS.C             1.377e-01  1.446e-02   9.523  < 2e-16 ***
## STARS^4            -1.337e-02  1.182e-02  -1.131  0.25809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16086.3  on 8954  degrees of freedom
## Residual deviance:  9565.8  on 8939  degrees of freedom
## AIC: 31978
##
## Number of Fisher Scoring iterations: 6
```

Removing outliers did not make any difference in the model. Lets perform final prediction of the model.

1.4.2 Zero Inflated Poisson

As the TARGET variable has many 0 counts, we can look for alternative models. Zero inflated poisson model is an another type of model which can be created for this type of dataset.

1.4.2.1 Model 2.1 - Zero Inflated Poisson with New Var

In this model, we will build a fresh model and reduce the varaibles which are not statistically significant.

```
##
## Call:
## zeroinfl(formula = TARGET ~ ., data = train.df, dist = "poisson")
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.282909 -0.420265 -0.003459  0.375103  4.836143
##
## Count model coefficients (poisson with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.492e+00  2.387e-01   6.251 4.07e-10 ***
## FixedAcidity        7.780e-04  9.974e-04   0.780 0.435402
## VolatileAcidity    -1.247e-02  7.994e-03  -1.560 0.118732
## CitricAcid          6.987e-04  7.194e-03   0.097 0.922633
## ResidualSugar      -3.077e-05  1.876e-04  -0.164 0.869746
## Chlorides          -2.394e-02  2.005e-02  -1.194 0.232397
## FreeSulfurDioxide   1.118e-05  4.252e-05   0.263 0.792668
## TotalSulfurDioxide -1.056e-05  2.674e-05  -0.395 0.692736
## Density            -1.889e-01  2.353e-01  -0.803 0.422020
## pH                  4.914e-03  9.278e-03   0.530 0.596405
## Sulphates          -3.838e-04  7.006e-03  -0.055 0.956306
## Alcohol             7.197e-03  1.712e-03   4.204 2.62e-05 ***
## LabelAppeal.L       8.667e-01  3.626e-02  23.905  < 2e-16 ***
## LabelAppeal.Q      -1.975e-01  3.029e-02  -6.520 7.05e-11 ***
## LabelAppeal.C       4.276e-02  2.103e-02   2.033 0.042035 *
```

```
## LabelAppeal^4      -1.880e-03  1.299e-02  -0.145 0.884890
## AcidIndex          -2.038e-02  5.842e-03  -3.489 0.000485 ***
## STARS.L             2.954e-01  2.071e-02  14.261  < 2e-16 ***
## STARS.Q             1.058e-02  1.753e-02   0.603 0.546351
## STARS.C            -2.044e-02  1.499e-02  -1.364 0.172668
## STARS^4             6.716e-03  1.214e-02   0.553 0.580049
##
## Zero-inflation model coefficients (binomial with logit link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.513e+01  1.779e+02  -0.085 0.932247
## FixedAcidity       -2.633e-04  6.565e-03  -0.040 0.968008
## VolatileAcidity     1.964e-01  5.242e-02   3.746 0.000179 ***
## CitricAcid         -3.482e-02  4.801e-02  -0.725 0.468313
## ResidualSugar      -9.920e-04  1.241e-03  -0.800 0.423929
## Chlorides           2.270e-01  1.326e-01   1.712 0.086919 .
## FreeSulfurDioxide  -8.665e-04  2.911e-04  -2.977 0.002912 **
## TotalSulfurDioxide -1.050e-03  1.831e-04  -5.737 9.65e-09 ***
## Density             1.227e+00  1.571e+00   0.781 0.434847
## pH                  1.568e-01  6.049e-02   2.592 0.009540 **
## Sulphates           1.416e-01  4.681e-02   3.025 0.002483 **
## Alcohol             3.280e-02  1.147e-02   2.859 0.004252 **
## LabelAppeal.L       2.441e+00  2.962e-01   8.244  < 2e-16 ***
## LabelAppeal.Q      -6.670e-01  2.498e-01  -2.670 0.007586 **
## LabelAppeal.C       7.758e-02  1.634e-01   0.475 0.634994
## LabelAppeal^4      -1.643e-01  8.991e-02  -1.827 0.067693 .
## AcidIndex           4.529e-01  3.158e-02  14.340  < 2e-16 ***
## STARS.L            -1.869e+01  5.159e+02  -0.036 0.971093
## STARS.Q            -1.989e+00  4.360e+02  -0.005 0.996360
## STARS.C             5.091e+00  3.600e+02   0.014 0.988715
## STARS^4             4.336e+00  2.175e+02   0.020 0.984098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 50
## Log-likelihood: -1.425e+04 on 42 Df

## [1] 31891.31

##                       Name      AIC      MSE
## 1          Poisson Model 31978.11 1.731961
## 2 Zero Inflation Poisson 31891.31 3.081792
```

Above results show that the AIC is less, but the MSE gets increased in Zero inflaction possion model.

1.4.2.2 Model 2.2 - Zero Inflated Poisson - Old var

In this model, we will used the variables which were already selected in the poisson model.

```
##
## Call:
## zeroinfl(formula = TARGET ~ . - FixedAcidity - ResidualSugar - CitricAcid -
##     Density - pH, data = train.df, dist = "poisson")
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.282581 -0.421480 -0.003867  0.376163  4.763440
```

```
##
## Count model coefficients (poisson with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.325e+00  4.979e-02  26.604  < 2e-16 ***
## VolatileAcidity     -1.246e-02  7.992e-03  -1.559 0.118886
## Chlorides           -2.427e-02  2.004e-02  -1.211 0.225974
## FreeSulfurDioxide    1.129e-05  4.249e-05   0.266 0.790365
## TotalSulfurDioxide -1.085e-05  2.673e-05  -0.406 0.684887
## Sulphates           -9.917e-05  7.002e-03  -0.014 0.988700
## Alcohol              7.188e-03  1.711e-03   4.201 2.66e-05 ***
## LabelAppeal.L        8.674e-01  3.622e-02  23.949  < 2e-16 ***
## LabelAppeal.Q       -1.983e-01  3.026e-02  -6.554 5.61e-11 ***
## LabelAppeal.C        4.304e-02  2.102e-02   2.048 0.040608 *
## LabelAppeal^4       -1.891e-03  1.298e-02  -0.146 0.884200
## AcidIndex           -2.032e-02  5.761e-03  -3.528 0.000419 ***
## STARS.L              2.954e-01  2.071e-02  14.263  < 2e-16 ***
## STARS.Q              1.057e-02  1.754e-02   0.603 0.546781
## STARS.C             -2.046e-02  1.499e-02  -1.365 0.172117
## STARS^4              6.679e-03  1.214e-02   0.550 0.582099
##
## Zero-inflation model coefficients (binomial with logit link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.338e+01  1.882e+02  -0.071 0.943297
## VolatileAcidity      1.996e-01  5.240e-02   3.810 0.000139 ***
## Chlorides            2.327e-01  1.324e-01   1.757 0.078877 .
## FreeSulfurDioxide   -8.576e-04  2.907e-04  -2.950 0.003174 **
## TotalSulfurDioxide -1.058e-03  1.831e-04  -5.778 7.55e-09 ***
## Sulphates            1.412e-01  4.679e-02   3.019 0.002536 **
## Alcohol              3.201e-02  1.144e-02   2.799 0.005132 **
## LabelAppeal.L        2.437e+00  2.959e-01   8.235  < 2e-16 ***
## LabelAppeal.Q       -6.635e-01  2.496e-01  -2.658 0.007862 **
## LabelAppeal.C        7.737e-02  1.635e-01   0.473 0.636104
## LabelAppeal^4       -1.625e-01  8.981e-02  -1.809 0.070407 .
## AcidIndex            4.445e-01  3.063e-02  14.512  < 2e-16 ***
## STARS.L             -1.879e+01  5.454e+02  -0.034 0.972514
## STARS.Q             -2.017e+00  4.609e+02  -0.004 0.996509
## STARS.C              5.124e+00  3.811e+02   0.013 0.989272
## STARS^4              4.366e+00  2.305e+02   0.019 0.984888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 39
## Log-likelihood: -1.425e+04 on 32 Df
```

```
##                             Name      AIC      MSE
## 1                 Poisson Model 31978.11 1.731961
## 2        Zero Inflation Poisson 31891.31 3.081792
## 3 Zero-Inflated Poisson old var 28565.70 1.651993
```

1.4.2.3 Vuong Test

In this section, we will compare poisson model and zero inflated model.

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
```

```
##  null that the models are indistinguishible)
## ------------------------------------------------------------
##                Vuong z-statistic             H_A    p-value
## Raw                   -37.21389 model2 > model1 < 2.22e-16
## AIC-corrected         -36.97602 model2 > model1 < 2.22e-16
## BIC-corrected         -36.13158 model2 > model1 < 2.22e-16
```

Vuong test compares the zero-inflated model with an ordinary poisson regression model. In this dataset, we can see that our test statistic is significant, indicating that the zero-inflated model is superior to the standard poisson model.

1.4.3 Model 3 - Negative Binomial model

Negative Binomial regession model is used when we have an overdispersion. It means the variance is higher than the mean. And there is a large number of `0` counts.

As we have seen before, that the variance of this dataset is higher than the mean and the histogram plot shows that there are higher number of zeros, we will model negative binomial model.

1.4.3.1 Model 3.1 - Normal negative binomial model

In this model, we will use a general negative binomial model and see how the model performs in the test dataset.

```
##
## Call:
## glm.nb(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Sulphates + Alcohol + LabelAppeal +
##     AcidIndex + STARS, data = train.df, init.theta = 40911.51538,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2633  -0.6461  -0.0035   0.4479   3.5133
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.679e+00  4.710e-02  35.649  < 2e-16 ***
## VolatileAcidity     -3.176e-02  7.770e-03  -4.088 4.35e-05 ***
## Chlorides           -4.931e-02  1.950e-02  -2.529  0.01144 *
## FreeSulfurDioxide    9.356e-05  4.203e-05   2.226  0.02603 *
## TotalSulfurDioxide   8.531e-05  2.698e-05   3.162  0.00157 **
## Sulphates           -1.375e-02  6.823e-03  -2.015  0.04390 *
## Alcohol              3.728e-03  1.680e-03   2.220  0.02645 *
## LabelAppeal.L        5.864e-01  3.342e-02  17.547  < 2e-16 ***
## LabelAppeal.Q       -8.497e-02  2.807e-02  -3.027  0.00247 **
## LabelAppeal.C        2.294e-02  1.972e-02   1.163  0.24474
## LabelAppeal^4        8.308e-03  1.244e-02   0.668  0.50434
## AcidIndex           -8.408e-02  5.379e-03 -15.632  < 2e-16 ***
## STARS.L              9.617e-01  1.964e-02  48.969  < 2e-16 ***
## STARS.Q             -3.997e-01  1.689e-02 -23.663  < 2e-16 ***
## STARS.C              1.377e-01  1.446e-02   9.522  < 2e-16 ***
## STARS^4             -1.337e-02  1.182e-02  -1.131  0.25811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for Negative Binomial(40911.52) family taken to be 1)
##
##     Null deviance: 16085.5  on 8954  degrees of freedom
## Residual deviance:  9565.4  on 8939  degrees of freedom
## AIC: 31980
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  40912
##           Std. Err.:  41229
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -31946.4
```

Model evaluation shows that the AIC and MSE is almost similar to Poisson distribution.

1.4.3.2 Model 3.2 - Negative Binomial with zero inflated

In this model, we will have a combination of negative binomial model which is zero-inflated.

```
##
## Call:
## zeroinfl(formula = TARGET ~ . - STARS - Density - FixedAcidity -
##     ResidualSugar - Alcohol, data = train.df, dist = c("negbin"))
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.0635 -0.3285  0.2117  0.4633  5.4630
##
## Count model coefficients (negbin with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.383e+00  5.667e-02  24.407  < 2e-16 ***
## VolatileAcidity    -1.408e-02  8.151e-03  -1.727 0.084118 .
## CitricAcid          1.192e-03  7.309e-03   0.163 0.870429
## Chlorides          -2.078e-02  2.048e-02  -1.015 0.310237
## FreeSulfurDioxide  -1.110e-05  4.305e-05  -0.258 0.796607
## TotalSulfurDioxide -2.386e-05  2.706e-05  -0.882 0.377739
## pH                  2.573e-03  9.477e-03   0.271 0.786024
## Sulphates           1.289e-03  7.132e-03   0.181 0.856562
## LabelAppeal.L       1.012e+00  3.585e-02  28.232  < 2e-16 ***
## LabelAppeal.Q      -1.990e-01  3.069e-02  -6.484 8.96e-11 ***
## LabelAppeal.C       3.089e-02  2.150e-02   1.436 0.150901
## LabelAppeal^4      -6.154e-04  1.334e-02  -0.046 0.963193
## AcidIndex          -2.135e-02  5.889e-03  -3.626 0.000288 ***
## Log(theta)          1.696e+01         NA      NA       NA
##
## Zero-inflation model coefficients (binomial with logit link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -6.2731053  0.2666419 -23.526  < 2e-16 ***
## VolatileAcidity     0.2778860  0.0410485   6.770 1.29e-11 ***
## CitricAcid         -0.0892818  0.0367841  -2.427 0.015216 *
## Chlorides           0.3349931  0.1016804   3.295 0.000986 ***
## FreeSulfurDioxide  -0.0007212  0.0002186  -3.299 0.000969 ***
```

```
## TotalSulfurDioxide -0.0008691  0.0001392  -6.241 4.34e-10 ***
## pH                  0.1321256  0.0472831   2.794 0.005200 **
## Sulphates           0.1607558  0.0357984   4.491 7.10e-06 ***
## LabelAppeal.L       0.7950484  0.2598146   3.060 0.002213 **
## LabelAppeal.Q      -0.5569771  0.2207179  -2.523 0.011620 *
## LabelAppeal.C       0.1612794  0.1426644   1.130 0.258274
## LabelAppeal^4      -0.0822510  0.0747836  -1.100 0.271396
## AcidIndex           0.5106459  0.0230413  22.162  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 23124109.6249
## Number of iterations in BFGS optimization: 56
## Log-likelihood: -1.593e+04 on 27 Df
```

This results to a similar model like zero-inflated model with Poisson.

1.4.4 Model 4 - Multiple linear regression

As the TARGET variable involves the counts, this might not be the suitable model for wine dataset. But to just get an overall idea, we will create a linear model and see.

1.4.4.1 Model 4.1 - Multiple Linear regression without Transformation

```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Sulphates + Alcohol + LabelAppeal +
##     AcidIndex + STARS, data = train.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0407 -0.8568  0.0238  0.8384  5.5428
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.976e+00  9.644e-02  51.603  < 2e-16 ***
## VolatileAcidity    -9.799e-02  1.773e-02  -5.527 3.35e-08 ***
## Chlorides          -1.582e-01  4.457e-02  -3.551 0.000386 ***
## FreeSulfurDioxide   2.829e-04  9.623e-05   2.940 0.003292 **
## TotalSulfurDioxide  2.321e-04  6.134e-05   3.785 0.000155 ***
## Sulphates          -3.618e-02  1.560e-02  -2.319 0.020399 *
## Alcohol             1.284e-02  3.827e-03   3.355 0.000797 ***
## LabelAppeal.L       1.586e+00  6.570e-02  24.140  < 2e-16 ***
## LabelAppeal.Q       1.132e-01  5.535e-02   2.046 0.040825 *
## LabelAppeal.C       1.489e-02  4.020e-02   0.370 0.711086
## LabelAppeal^4       3.093e-02  2.701e-02   1.145 0.252175
## AcidIndex          -2.109e-01  1.065e-02 -19.808  < 2e-16 ***
## STARS.L             2.764e+00  4.778e-02  57.839  < 2e-16 ***
## STARS.Q            -5.028e-01  4.138e-02 -12.150  < 2e-16 ***
## STARS.C             1.224e-01  3.516e-02   3.480 0.000504 ***
## STARS^4             5.781e-02  2.927e-02   1.975 0.048260 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.311 on 8939 degrees of freedom
## Multiple R-squared:  0.5422, Adjusted R-squared:  0.5414
## F-statistic: 705.8 on 15 and 8939 DF,  p-value: < 2.2e-16
```

On a inital look, the model is not doing a bad job at prediction. However, this is not a good model on counts TARGET variable.

1.4.4.2 Model 4.2 - Multiple linear regression with Transformation

In the previous multiple linear regression model, we have have a limitation on the dependent variable. Limitation is count cannot be a negative value. However the limitation is not there on the predictor variables.

To overcome this limitation we will transform the dependent variable using log transformation. It will overcome this limitation.

```
##
## Call:
## lm(formula = log(train.df$TARGET + 1e-10) ~ ., data = train.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.6783  -2.0778   0.5565   3.6521  22.7738
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.9792688  3.0931214   3.226 0.001259 **
## FixedAcidity         0.0023789  0.0129412   0.184 0.854153
## VolatileAcidity     -0.5072853  0.1034896  -4.902 9.66e-07 ***
## CitricAcid           0.1151489  0.0942348   1.222 0.221764
## ResidualSugar        0.0030999  0.0024369   1.272 0.203373
## Chlorides           -0.7172095  0.2602238  -2.756 0.005861 **
## FreeSulfurDioxide    0.0015740  0.0005618   2.802 0.005097 **
## TotalSulfurDioxide   0.0022324  0.0003581   6.235 4.74e-10 ***
## Density             -1.2815272  3.0590344  -0.419 0.675276
## pH                  -0.2902889  0.1206907  -2.405 0.016183 *
## Sulphates           -0.3048603  0.0911065  -3.346 0.000823 ***
## Alcohol             -0.0522131  0.0223443  -2.337 0.019474 *
## LabelAppeal.L       -2.6801185  0.3835052  -6.988 2.98e-12 ***
## LabelAppeal.Q        0.6844829  0.3231310   2.118 0.034179 *
## LabelAppeal.C        0.1397614  0.2346212   0.596 0.551398
## LabelAppeal^4        0.1825551  0.1576390   1.158 0.246871
## AcidIndex           -1.2520312  0.0635057 -19.715  < 2e-16 ***
## STARS.L             11.2845870  0.2788792  40.464  < 2e-16 ***
## STARS.Q             -5.5507992  0.2415113 -22.984  < 2e-16 ***
## STARS.C              1.3170871  0.2052662   6.416 1.47e-10 ***
## STARS^4              0.1773843  0.1708162   1.038 0.299088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.65 on 8934 degrees of freedom
## Multiple R-squared:  0.4134, Adjusted R-squared:  0.4121
## F-statistic: 314.8 on 20 and 8934 DF,  p-value: < 2.2e-16
```

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

```
## lm(formula = log(train.df$TARGET + 1e-10) ~ VolatileAcidity +
##     Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + pH +
##     Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS, data = train.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.8141  -2.0751   0.5468   3.6325  23.0554
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          8.7243040  0.7056711  12.363  < 2e-16 ***
## VolatileAcidity     -0.5099251  0.1034702  -4.928 8.45e-07 ***
## Chlorides           -0.7236461  0.2601272  -2.782 0.005416 **
## FreeSulfurDioxide    0.0015884  0.0005616   2.828 0.004690 **
## TotalSulfurDioxide   0.0022435  0.0003579   6.268 3.83e-10 ***
## pH                  -0.2892936  0.1206652  -2.397 0.016528 *
## Sulphates           -0.3047400  0.0910371  -3.347 0.000819 ***
## Alcohol             -0.0517461  0.0223362  -2.317 0.020543 *
## LabelAppeal.L       -2.6796694  0.3834530  -6.988 2.98e-12 ***
## LabelAppeal.Q        0.6923713  0.3230199   2.143 0.032105 *
## LabelAppeal.C        0.1427172  0.2345880   0.608 0.542955
## LabelAppeal^4        0.1820737  0.1576252   1.155 0.248078
## AcidIndex           -1.2464237  0.0622954 -20.008  < 2e-16 ***
## STARS.L             11.2911547  0.2788434  40.493  < 2e-16 ***
## STARS.Q             -5.5548162  0.2414740 -23.004  < 2e-16 ***
## STARS.C              1.3172463  0.2052086   6.419 1.44e-10 ***
## STARS^4              0.1823697  0.1707824   1.068 0.285618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.65 on 8938 degrees of freedom
## Multiple R-squared:  0.4132, Adjusted R-squared:  0.4121
## F-statistic: 393.3 on 16 and 8938 DF,  p-value: < 2.2e-16
```

Log transformed model is not performing well in this type of dataset. Adjusted R-squared is low compared to the previous multiple linear regression model.

1.4.5 Model Interpreation

As we have build different models, we will compare the models and interprent the poisson model.

Comparing all the model variables, it seems some of the variables are statistically significant in all the models. They are `LABELAPPEAL, TotalSulfurDioxide, FreeSulfurDioxide, Sulphates, VolatileAcidity`.

Model interpreation of poission model is as follows

For per unit increase, given the other variables are held constant in the model,

1. VolatileAcidity - Difference in the logs of expected count would decrease by 0.03.
2. Chlorides - Difference in the logs of expected count would decrease by 0.04.
3. FreeSulfurDioxide - Difference in the logs of expected count would increase by 0.0009.
4. TotalSulfurDioxide - Difference in the logs of expected count would increase by 0.0008.
5. Sulphates - Difference in the logs of expected count would decrease by 0.01.

6. Alcohol - Difference in the logs of expected count would increase by 0.03.
7. LabelAppeal.L - Estimated poisson regression cefficient comparing LabelAppeal -1 to -2, given the other variables are held constant. The difference in the logs of expected counts is expected to be 0.5 unit higher for LabelAppeal -1 compared to -2.
8. LabelAppeal.Q - Estimated poisson regression cefficient comparing LabelAppeal 0 to -2, given the other variables are held constant. The difference in the logs of expected counts is expected to be 0.08 unit lower for LabelAppeal 0 compared to -2.
9. LabelAppeal.L - Estimated poisson regression cefficient comparing LabelAppeal 1 to -2, given the other variables are held constant. The difference in the logs of expected counts is expected to be 0.02 unit higher for LabelAppeal 1 compared to -2.
10. LabelAppeal.L - Estimated poisson regression cefficient comparing LabelAppeal 2 to -2, given the other variables are held constant. The difference in the logs of expected counts is expected to be 0.008 unit higher for LabelAppeal 2 compared to -2.
11. Alcohol - Difference in the logs of expected count would decrease by 0.08.
12. STARS.L - Estimated poisson regression cefficient comparing STARS 1 to 0, given the other variables are held constant. The difference in the logs of expected counts is expected to be 0.9 unit higher for STARS 1 compared to 0.
13. STARS.Q - Estimated poisson regression cefficient comparing STARS 2 to 0, given the other variables are held constant. The difference in the logs of expected counts is expected to be 0.3 unit lower for STARS 2 compared to 0.
14. STARS.C - Estimated poisson regression cefficient comparing STARS 3 to 0, given the other variables are held constant. The difference in the logs of expected counts is expected to be 0.1 unit higher for STARS 3 compared to 0.
15. STARS^4 - Estimated poisson regression cefficient comparing STARS 4 to 0, given the other variables are held constant. The difference in the logs of expected counts is expected to be 0.01 unit lower for STARS 4 compared to 0.

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Sulphates + Alcohol + LabelAppeal +
##     AcidIndex + STARS, family = "poisson", data = train.df[-c(4940,
##     12513)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2635  -0.6462  -0.0035   0.4480   3.5135
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.679e+00  4.709e-02  35.650  < 2e-16 ***
## VolatileAcidity     -3.176e-02  7.770e-03  -4.088 4.35e-05 ***
## Chlorides           -4.931e-02  1.950e-02  -2.529  0.01144 *
## FreeSulfurDioxide    9.355e-05  4.203e-05   2.226  0.02602 *
## TotalSulfurDioxide   8.531e-05  2.698e-05   3.162  0.00157 **
## Sulphates           -1.375e-02  6.822e-03  -2.015  0.04390 *
## Alcohol              3.728e-03  1.680e-03   2.220  0.02644 *
## LabelAppeal.L        5.864e-01  3.342e-02  17.548  < 2e-16 ***
## LabelAppeal.Q       -8.497e-02  2.807e-02  -3.027  0.00247 **
## LabelAppeal.C        2.294e-02  1.972e-02   1.163  0.24475
```

```
## LabelAppeal^4       8.307e-03  1.244e-02   0.668  0.50435
## AcidIndex          -8.408e-02  5.379e-03 -15.632  < 2e-16 ***
## STARS.L             9.617e-01  1.964e-02  48.971  < 2e-16 ***
## STARS.Q            -3.997e-01  1.689e-02 -23.664  < 2e-16 ***
## STARS.C             1.377e-01  1.446e-02   9.523  < 2e-16 ***
## STARS^4            -1.337e-02  1.182e-02  -1.131  0.25809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16086.3  on 8954  degrees of freedom
## Residual deviance:  9565.8  on 8939  degrees of freedom
## AIC: 31978
##
## Number of Fisher Scoring iterations: 6
```

1.5 Select Model

We have created various models and interpreted the them. Finally we have to select the best performing model. We have calculated the AIC and MSE for all the models.

```
##                                                  Name      AIC      MSE
## 1                                     Poisson Model 31978.11 1.731961
## 2                           Zero Inflation Poisson 31891.31 3.081792
## 3                     Zero-Inflated Poisson old var 28565.70 1.651993
## 4                                 Negative Binomial 31980.40 1.731961
## 5 Zero-inflated Negative Binomial Stepwise selection 31918.28 3.118260
## 6     Zero-inflated Negative Binomial same Variables 28567.70 1.652253
## 7                    Multiple Linear model stepwise 30279.38 1.779109
```

From above results, we see that zero inflated Poisson model with same variables used by Poisson model is performing better. However, Poisson model is more interpretable. So we will select that model as our final model compared to other models.

1.5.1 Prediction on evaluation dataset

Finally lets predict on the evaluation dataset.

```
## [1] 3.034783
```

```
## [1] 2.226744
```

```
##   TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
## 1      1          5.4          -0.860       0.27         -10.7     0.092
## 2      4         12.4           0.385      -0.76         -19.7     1.169
## 3      3          7.2           1.750       0.17         -33.0     0.065
## 4      2          6.2           0.100       1.80           1.0    -0.179
## 5      1         11.4           0.210       0.28           1.2     0.038
## 6      5         17.6           0.040      -1.15           1.4     0.535
##   FreeSulfurDioxide TotalSulfurDioxide Density   pH Sulphates Alcohol
## 1                23                398 0.98527 5.02      0.64   12.30
## 2               -37                 68 0.99048 3.37      1.09   16.00
## 3                 9                 76 1.04641 4.61      0.68    8.55
## 4               104                 89 0.98877 3.20      2.11   12.30
```

```
## 5                70                53 1.02899 2.54     -0.07    4.80
## 6              -250               140 0.95028 3.06     -0.02   11.40
##   LabelAppeal AcidIndex STARS
## 1          -1         6     0
## 2           0         6     2
## 3           0         8     1
## 4          -1         8     1
## 5           0        10     0
## 6           1         8     4
```

1.6 References

1. Pavan HW3 Markdown
2. stats.idre.ucla.edu/stata/output/poisson-regression/