Western University

Using Generalized Linear Models to predict Wine Sales

STATS 9155B

Submitted By:
Debduti Sengupta- 251253766
Parth Sawhney- 251247405

1. Introduction

In this project, our aim is to analyze the Wine dataset and build a model to predict the number of wine cases sold, based on its features, most of which are chemical properties. This dataset comes from a Kaggle competition for wine sales prediction and the dataset can be found on Github. We will cover the following topics in this project:

a. We will do an initial analysis to find out nulls in the dataset and process them
b. We will do exploratory analysis on all the variables and produce relevant boxplots, scatter plots and histograms
c. We will then do some feature engineering that will help with building our models
d. We will build four different models- Poisson Regression Model, Negative Binomial Model, Zero Inflated Poisson Model and Zero Inflated Negative Binomial Model. We will compare their performances based on metrics such as Root Mean Squared Error, Mean Absolute Error, AIC and BIC and select a final best model

2. Dataset

The dataset **Wine.csv** contains 12795 records and 14 predictors (excluding INDEX). Our response variable is **TARGET.** The dataset is obtained from a Kaggle competition for prediction of wine sales based on its attributes.

```
'data.frame':	12795 obs. of  16 variables:
 $ 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 ...
```

Figure 1

2.1 Data Description

Attached below is the data description of the Wine.csv dataset.

VARIABLE NAME	DEFINITION
INDEX	Identification Variable (do not use)
TARGET	Number of Cases Purchased
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.
ResidualSugar	Residual Sugar of wine
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor
Sulphates	Sulfate conten of wine
TotalSulfurDioxide	Total Sulfur Dioxide of Wine
VolatileAcidity	Volatile Acid content of wine
pH	pH of wine

Figure 2

The following is a summary of the dataset.

```
> summary(wine)
     INDEX           TARGET       FixedAcidity     VolatileAcidity     CitricAcid     
 Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400  
 1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300  
 Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100  
 Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084  
 3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800  
 Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600  

 ResidualSugar        Chlorides       FreeSulfurDioxide TotalSulfurDioxide    Density      
 Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00   Min.   :-823.0     Min.   :0.8881  
 1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00   1st Qu.:  27.0     1st Qu.:0.9877  
 Median :   3.900   Median : 0.0460   Median :  30.00   Median : 123.0     Median :0.9945  
 Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85   Mean   : 120.7     Mean   :0.9942  
 3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00   3rd Qu.: 208.0     3rd Qu.:1.0005  
 Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00   Max.   :1057.0     Max.   :1.0992  
 NA's   :616        NA's   :638       NA's   :647       NA's   :682                        
       pH          Sulphates          Alcohol       LabelAppeal          AcidIndex     
 Min.   :0.480   Min.   :-3.1300   Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000  
 1st Qu.:2.960   1st Qu.: 0.2800   1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000  
 Median :3.200   Median : 0.5000   Median :10.40   Median : 0.000000   Median : 8.000  
 Mean   :3.208   Mean   : 0.5271   Mean   :10.49   Mean   :-0.009066   Mean   : 7.773  
 3rd Qu.:3.470   3rd Qu.: 0.8600   3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000  
 Max.   :6.130   Max.   : 4.2400   Max.   :26.50   Max.   : 2.000000   Max.   :17.000  
 NA's   :395     NA's   :1210      NA's   :653                                          
     STARS      
 Min.   :1.000  
 1st Qu.:1.000  
 Median :2.000  
 Mean   :2.042  
 3rd Qu.:3.000  
 Max.   :4.000  
 NA's   :3359   
```

Figure 3

The above summary statistics reveals that there are NULLS ResidualSugar, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, Sulphates, Alcohol, and Stars. We also seem to have outliers in some of the columns. We will take a closer look into these in our exploratory data analysis and clean our data accordingly.

3. Methods

We conducted the following steps for our project:

Exploratory Data Analysis: We checked the distribution of all the variables and created histograms and box plots for them to visualize them. We also plotted the distribution of percentage of missing values in the dataset. We also created a correlation matrix and scatter plots to understand the correlation of the features with the TARGET.

Data Preparation: Outliers, as observed from the histogram of predictor variables are removed from the dataset. Binary flags are added to each variable to indicate whether they have NULLS. The missing values are re-imputed using predictive mean matching.

Data Analysis: We fitted four different models- Poisson Regression , Negative Binomial, Zero Inflated Poisson Regression and Zero Inflated Negative Binomial Regression models. We ran a full model with all the variables, and then we did a stepwise backward selection using AIC to select the significant predictors. This process was done for both Poisson and Negative Binomial Models. The goodness of fit was checked using Pearson Chi-Square test for the models. Since there was slight overdispersion and there was a large number of zeros in our response variable, we proceeded to run zero-inflated regression models for both Poisson and Negative Binomial Regression Models. We ran the Vuong test between our base model and zero inflated model to determine whether the zero inflated model had any difference with the base model.

Finally, we compared all four models based on different metrics like their Root Mean Squared Error, Mean Absolute Error, AIC, and BIC and chose the best performing model.

4. Results

Section 1: Exploratory Data Analysis

We ran a univariate and multivariate analysis for all the variables as follows:

Target
For the response variable TARGET, see that there are quite a number of zeros and there is an outlier at around 8 cases. The majority of the number of wine cases falls around the mean of 3.

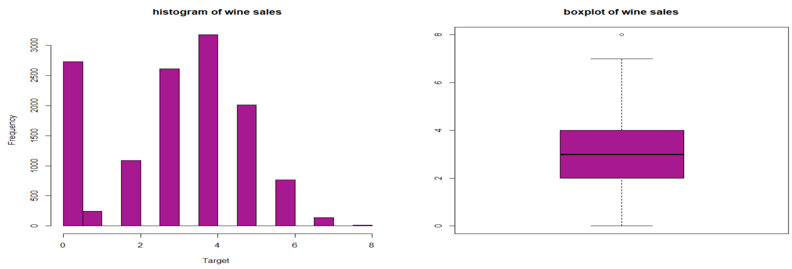


Figure 4

FixedAcidity and VolatileAcidity
Fixed Acidity has a symmetrical bell shape, and both fixed and volatile acidity fields have some outliers on both extreme ends. For the former, the outliers are less than - 5 and greater than 20, for the latter they are less than -1.5 and greater than 2.

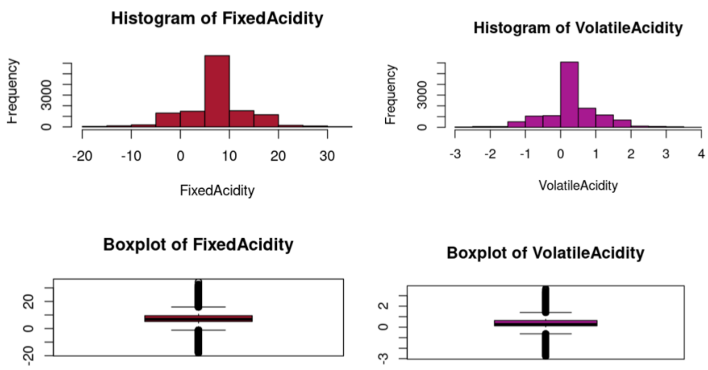


Figure 5

CitricAcid and ResidualSugar
The histogram of CitricAcid shows a symmetric bell shape with noticeable outliers less than -1.5 and greater than 2. Residual Sugars have a symmetric bell shaped distribution with outliers less than -65 and greater than 65.

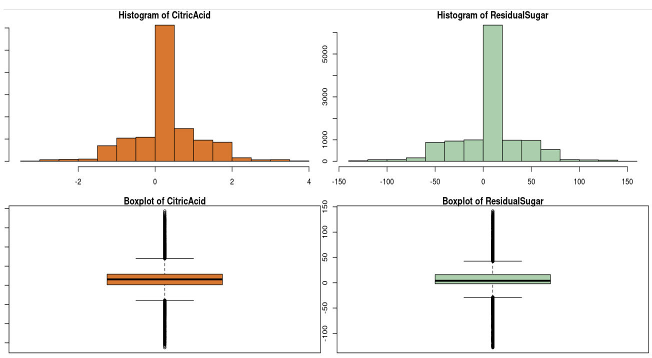


Figure 6

Chlorides and FreeSulfur Dioxide
The histogram of Chlorides shows a symmetric bell shape with noticeable outliers around less than -0.6 and greater than 0.7.The histogram of FreeSulfurDioxide shows a symmetric bell shape with noticeable outliers around less than -275 and greater than 350.

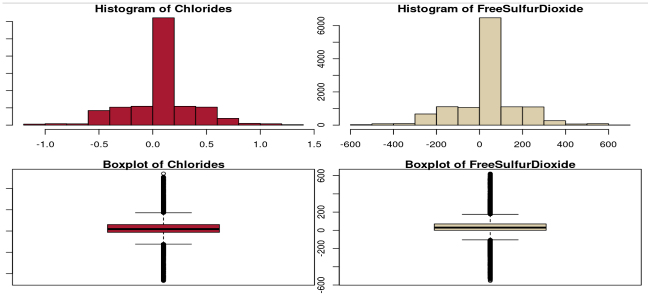


Figure 7

Ph, Sulphates, TotalSulfurDioxide and Density
The pH histogram has a symmetric bell shape with prominent outliers, with the majority of the values hovering around 3.208. Sulphates also has a symmetric bell shape with significant outliers around -1.5 and larger than 2.5. Density and TotalSulphurDioxide histograms both have a symmetric bell shape with some outliers. Most of the variables have a similar kind of bell shaped distribution.

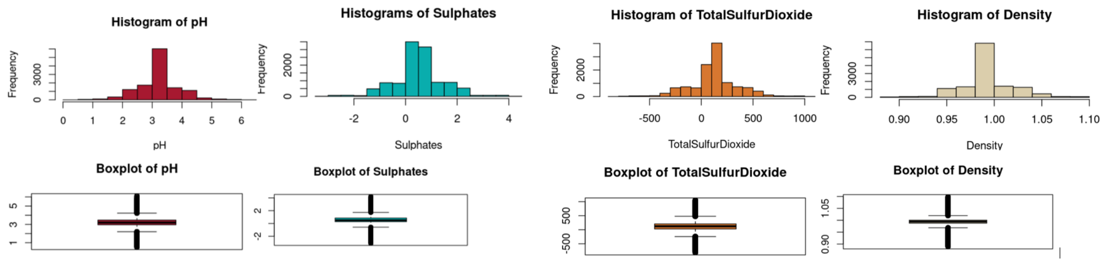


Figure 8

Alcohol and Acid Index
A histogram and boxplot of Alcohol & AcidIndex are shown in the figure above. Alcohol's histogram has a symmetric bell shape with prominent outliers in the range of less than 2 to higher than 20. AcidIndex's histogram has a small right skew with a few outliers. AcidIndex's box plot likewise shows that the median number is 8.

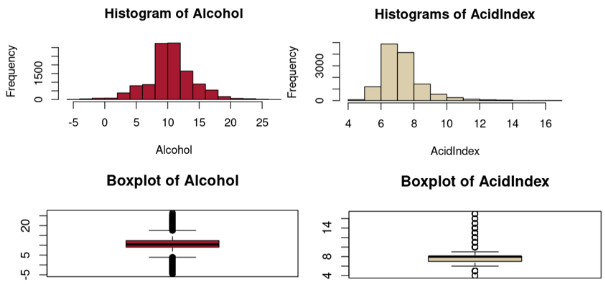


Figure 9

Label Appeal and STARS
LabelAppeal's histogram has a symmetric bell shape. The majority of the results are around -0.009066, which is the mean. LabelAppeal is mostly in the range of -1.0 to 1.0. The STARS histogram has a small right skew. It's also important to note that the STARS data set has the most missing values.

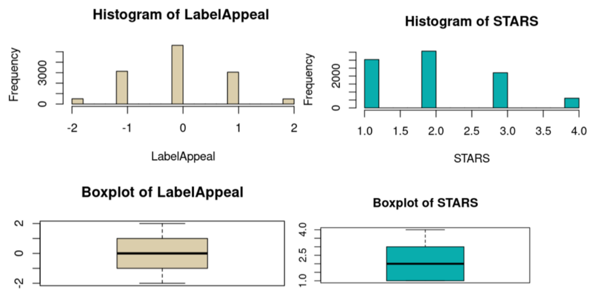


Figure 10

Section 2: Data Preparation
Missing Values for Variables:

```
           index           target       fixedacidity  volatileacidity        citricacid
        0.000000         0.000000           0.000000         0.000000          0.000000
   residualsugar        chlorides freesulfurdioxide totalsulfurdioxide          density
        4.814381         4.986323           5.056663         5.330207          0.000000
              ph        sulphates            alcohol       labelappeal        acidindex
        3.087143         9.456819           5.103556         0.000000          0.000000
           stars
       26.252442
```

Figure 3: Percentage of missing values in data

The variables in the wine data set have missing data, as shown in the figure above. To impute the missing data, we'll utilize the MICE package with pmm (predictive mean matching). The MICE package basically employs an algorithm that predicts and imputes missing values based on information from other variables in the dataset. We must deal with missing values because Poisson, Binomial type regression models cannot handle them and must be dealt with before using these modeling techniques. Stars had the highest percentage of missing data at 26.25%.

Outliers treatment and Flag Variables:
Outliers, as observed from the boxplots, are removed from the file, and flags are added for predictors that have NULL values in them and they are set to 1.
Because the data set had a number of variables with missing data, we constructed these flag variables. Furthermore, there's a good probability that a missing variable is really predictive of the target variable, which would improve the model's accuracy.

```
1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300
Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100
Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084
3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800
Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600

residualsugar        chlorides         freesulfurdioxide totalsulfurdioxide    density
Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00   Min.   :-823.0     Min.   :0.8881
1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00   1st Qu.:  27.0     1st Qu.:0.9877
Median :   3.900   Median : 0.0460   Median :  30.00   Median : 123.0     Median :0.9945
Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85   Mean   : 120.7     Mean   :0.9942
3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00   3rd Qu.: 208.0     3rd Qu.:1.0005
Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00   Max.   :1057.0     Max.   :1.0992
NA's   :616        NA's   :638       NA's   :647       NA's   :682
      ph           sulphates          alcohol        labelappeal          acidindex
Min.   :0.480   Min.   :-3.1300   Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000
1st Qu.:2.960   1st Qu.: 0.2800   1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000
Median :3.200   Median : 0.5000   Median :10.40   Median : 0.000000   Median : 8.000
Mean   :3.208   Mean   : 0.5271   Mean   :10.49   Mean   :-0.009066   Mean   : 7.773
3rd Qu.:3.470   3rd Qu.: 0.8600   3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000
Max.   :6.130   Max.   : 4.2400   Max.   :26.50   Max.   : 2.000000   Max.   :17.000
NA's   :395     NA's   :1210      NA's   :653
    stars       noresidualsugar    nochlorides       nofreesulfurdioxide nototalsulfurdioxide
Min.   :1.000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000     Min.   :0.0000
1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000     1st Qu.:0.0000
Median :2.000   Median :0.00000   Median :0.00000   Median :0.00000     Median :0.0000
Mean   :2.042   Mean   :0.04814   Mean   :0.04986   Mean   :0.05057     Mean   :0.0533
3rd Qu.:3.000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000     3rd Qu.:0.0000
Max.   :4.000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000     Max.   :1.0000
NA's   :3359
     noph            nosulphates        noalcohol           nostars
Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
Median :0.00000   Median :0.00000   Median :0.00000   Median :0.0000
Mean   :0.03087   Mean   :0.09457   Mean   :0.05104   Mean   :0.2625
3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000
Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
```

Figure 11

To get an overall idea about which variables might be correlated with the TARGET, we created a correlation matrix.

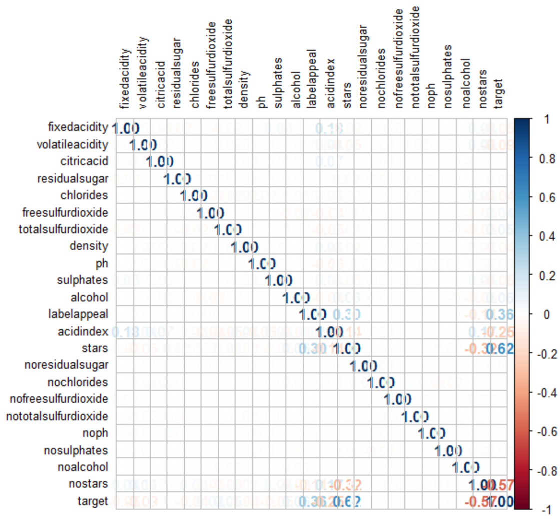
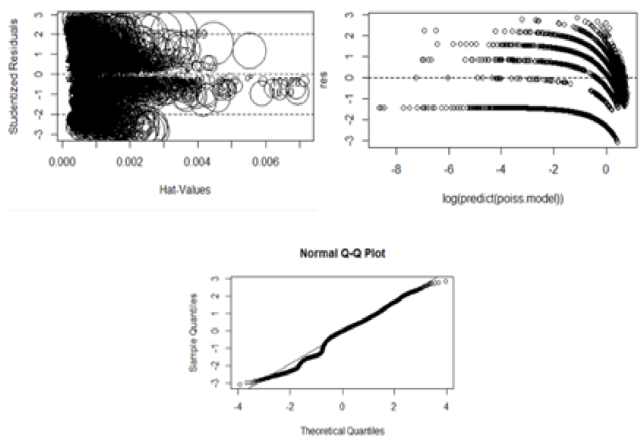


Figure: 12

Section 3: Build Model
Model 1: Poisson Regression

```
> summary(poisson.back)

Call:
glm(formula = target ~ volatileacidity + chlorides + freesulfurdioxide +
    totalsulfurdioxide + ph + sulphates + alcohol + labelappeal +
    acidindex + stars + nostars, family = "poisson", data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.07969  -0.71550   0.00844   0.48123   2.87818

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.326e+00  4.895e-02  27.087  < 2e-16 ***
volatileacidity    -2.896e-02  6.522e-03  -4.440 9.01e-06 ***
chlorides          -4.058e-02  1.599e-02  -2.539  0.01113 *
freesulfurdioxide   9.048e-05  3.407e-05   2.656  0.00792 **
totalsulfurdioxide  7.194e-05  2.220e-05   3.240  0.00119 **
ph                 -1.299e-02  7.522e-03  -1.726  0.08430 .
sulphates          -1.155e-02  5.462e-03  -2.114  0.03451 *
alcohol             2.691e-03  1.371e-03   1.963  0.04965 *
labelappeal         1.364e-01  6.137e-03  22.227  < 2e-16 ***
acidindex          -7.414e-02  4.519e-03 -16.407  < 2e-16 ***
stars               2.485e-01  5.857e-03  42.419  < 2e-16 ***
nostars            -8.826e-01  1.735e-02 -50.858  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22861  on 12794  degrees of freedom
Residual deviance: 12935  on 12783  degrees of freedom
AIC: 44901

Number of Fisher Scoring iterations: 6
```



```
> anova(poiss.model, test="Chisq")
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                               12794      22861
volatileacidity     1    123.6     12793      22737 < 2.2e-16 ***
freesulfurdioxide   1     27.1     12792      22710 1.895e-07 ***
totalsulfurdioxide  1     38.9     12791      22671 4.572e-10 ***
sulphates           1     33.8     12790      22638 6.095e-09 ***
chlorides           1     24.7     12789      22613 6.820e-07 ***
alcohol             1     64.1     12788      22549 1.184e-15 ***
ph                  1      1.8     12787      22547    0.1834
labelappeal         1   1981.3     12786      20566 < 2.2e-16 ***
acidindex           1   1016.1     12785      19550 < 2.2e-16 ***
stars               1   3474.6     12784      16075 < 2.2e-16 ***
nostars             1   3140.4     12783      12934 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> with(poiss.model, cbind(res.deviance = deviance, df = df.residual,
+                          p = pchisq(deviance, df.residual, lower.tail=FALSE)))
     res.deviance    df         p
[1,]     12934.55 12783 0.1715032
```

```
> dispersiontest(poiss.model)

        Overdispersion test

data:  poiss.model
z = -16.767, p-value = 1
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
 0.8199568
```

Figure: 13

Observation: Poisson regression models use the log link function to approximate regression processes for a count variable distributed such that the variance is equal to the mean. The backward stepwise AIC feature selection algorithm returned a Poisson model with ten predictors. The deviance of residuals, which is a measure of model fit of a generalized linear model, shows that the null deviance is 22861 and the residual deviance is 12935. As a result, since the goodness-of-fit chi-squared test is not statistically significant as we got p-values of 0.17. So, we have no evidence to reject the Null hypothesis that the Poisson Model fits well. Also, the Analysis of Deviance table shows the difference between the null deviance and the residual deviance i.e wider the gap, the better the predictor. The table shows that stars, labelappeal, acidindex, and no_stars significantly reduce the residual deviance and have very small p-values. Furthermore, the results show that the data is not overdispersed as indicated by the dispersion test. As per the QQ plot, the data is relatively normal. The results also show an AIC of 44901.

In terms of the coefficients, the model's coefficients make intuitive sense. Stars and labelappeal, for example, are both positive. This implies that when label attractiveness and star ratings rise, the number of sample cases of wine purchased rises as well, which makes intuitive sense in terms of wine sales. Furthermore, the fact that no_stars is negative shows that the number of sample cases of wine purchased drops as the number of wines with no STARS (e.g., N/A) increases, which makes logical wine sales sense.

Model 2: Negative Binomial Regression

```
hi-Square Test Statistic =  -0.3835 p-value = 0.5
 summary(negbinomial.mod)

all:
lm.nb(formula = target ~ volatileacidity + +freesulfurdioxide +
    totalsulfurdioxide + sulphates + chlorides + alcohol + ph +
    labelappeal + acidindex + stars + nostars, data = train,
    init.theta = 45272.9192, link = log)

eviance Residuals:
     Min        1Q    Median        3Q       Max
3.07960  -0.71548   0.00841   0.48121   2.87811

oefficients:
                    Estimate Std. Error z value Pr(>|z|)
Intercept)         1.326e+00  4.895e-02  27.086  < 2e-16 ***
olatileacidity    -2.896e-02  6.523e-03  -4.440 9.01e-06 ***
reesulfurdioxide   9.048e-05  3.407e-05   2.655  0.00792 **
otalsulfurdioxide  7.194e-05  2.220e-05   3.240  0.00119 **
ulphates          -1.155e-02  5.462e-03  -2.114  0.03451 *
hlorides          -4.058e-02  1.599e-02  -2.539  0.01113 *
lcohol             2.691e-03  1.371e-03   1.963  0.04967 *
h                 -1.299e-02  7.523e-03  -1.726  0.08430 .
abelappeal         1.364e-01  6.137e-03  22.226  < 2e-16 ***
cidindex          -7.414e-02  4.519e-03 -16.406  < 2e-16 ***
tars               2.485e-01  5.858e-03  42.418  < 2e-16 ***
ostars            -8.826e-01  1.736e-02 -50.857  < 2e-16 ***
--
ignif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Dispersion parameter for Negative Binomial(45272.92) family taken to be 1)

    Null deviance: 22860  on 12794  degrees of freedom
esidual deviance: 12934  on 12783  degrees of freedom
IC: 44903
```

```
     res.deviance    df         p
[1,]     12934.07 12783 0.1722504
> odTest(negbinomial.mod)
Likelihood ratio test of H0: Poisson, as restricted NB model:
n.b., the distribution of the test-statistic under H0 is non-standard
e.g., see help(odTest) for details/references

Critical value of test statistic at the alpha= 0.05 level: 2.7055
Chi-Square Test Statistic =  -0.3835 p-value = 0.5
> summary(negbinomial.mod)
```

Figure: 14

Observation: Our final Negative Binomial model was a parsimonious model created with only those variables that were deemed significant by the AIC test. We ran an odTest to compare the log-likelihood ratios of a Negative Binomial regression to the restriction of a Poisson regression mean=variance.

The results show that we should accept the Poisson regression model because the test statistic of -0.3835 is less than 2.7055 with a p-value of 0.5. The deviance of residuals, which is a measure of model fit of a generalized linear model, shows that the null deviance is 22860 and the residual deviance is 12934. The results also show an AIC of 44903, 2*log likelihood of -44876.95, and Theta of 45273. One common cause of overdispersion is the presence of excess zeros.

Model 3: Zero Inflated Poisson Regression

```
Count model coefficients (poisson with log link):
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)        1.132e+00  5.117e-02  22.128  < 2e-16 ***
volatileacidity   -1.240e-02  6.713e-03  -1.847 0.064691 .  
freesulfurdioxide  2.338e-05  3.443e-05   0.679 0.497047    
totalsulfurdioxide -1.604e-05  2.211e-05  -0.726 0.468127    
sulphates          3.260e-04  5.620e-03   0.058 0.953733    
chlorides         -2.596e-02  1.639e-02  -1.583 0.113322    
alcohol            6.316e-03  1.399e-03   4.515 6.34e-06 ***
ph                 4.574e-03  7.734e-03   0.591 0.554244    
labelappeal        2.257e-01  6.376e-03  35.400  < 2e-16 ***
acidindex         -1.853e-02  4.845e-03  -3.825 0.000131 ***
stars              1.182e-01  6.201e-03  19.068  < 2e-16 ***
nostars           -1.693e-01  1.846e-02  -9.174  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -3.9014785  0.3431557 -11.369  < 2e-16 ***
volatileacidity     0.2135367  0.0498425   4.284 1.83e-05 ***
freesulfurdioxide  -0.0005924  0.0002678  -2.212   0.0270 *  
totalsulfurdioxide -0.0010107  0.0001666  -6.067 1.30e-09 ***
sulphates           0.1634195  0.0416523   3.923 8.73e-05 ***
chlorides           0.0921370  0.1204723   0.765   0.4444    
alcohol             0.0258440  0.0104459   2.474   0.0134 *  
ph                  0.2493076  0.0570175   4.372 1.23e-05 ***
labelappeal         1.0010760  0.0525739  19.041  < 2e-16 ***
acidindex           0.4391702  0.0278681  15.759  < 2e-16 ***
stars              -2.5681640  0.0924650 -27.774  < 2e-16 ***
nostars             3.2614979  0.0910217  35.832  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> vuong(poiss.model, zinp.mod)
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                   -44.77167 model2 > model1 < 2.22e-16
AIC-corrected         -44.55430 model2 > model1 < 2.22e-16
BIC-corrected         -43.74385 model2 > model1 < 2.22e-16
> 
```

Figure : 15

Observation: An oversupply of zero data can skew the Poisson and negative binomial models. Zero-inflated models presume there are two types of values in the distribution: true zero measurements and another set of values that follow a more usual distribution. These models categorize values into their appropriate categories before predicting their outcomes using distinct sets of coefficients for each. Because almost 2,500 wines sold zero cases in this situation, zero-inflated models may be more accurate. The Vuong test compares the zero-inflated model with a standard Poisson regression model. The Vuong test shows that our test statistic is significant, indicating that the zero-inflated model is an improvement over the standard Poisson model.

Model 4: Zero Inflated Negative Binomial Regression

```
Count model coefficients (negbin with log link):
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)         1.132e+00  5.117e-02  22.128  < 2e-16 ***
volatileacidity    -1.240e-02  6.713e-03  -1.847 0.064689 .  
freesulfurdioxide   2.338e-05  3.443e-05   0.679 0.497058    
totalsulfurdioxide -1.604e-05  2.211e-05  -0.726 0.468059    
sulphates           3.259e-04  5.620e-03   0.058 0.953760    
chlorides          -2.596e-02  1.639e-02  -1.584 0.113304    
alcohol             6.316e-03  1.399e-03   4.515 6.34e-06 ***
ph                  4.573e-03  7.734e-03   0.591 0.554358    
labelappeal         2.257e-01  6.376e-03  35.400  < 2e-16 ***
acidindex          -1.853e-02  4.845e-03  -3.825 0.000131 ***
stars               1.182e-01  6.201e-03  19.068  < 2e-16 ***
nostars            -1.693e-01  1.846e-02  -9.174  < 2e-16 ***
Log(theta)          1.838e+01  1.964e+00   9.362  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -3.9014673  0.3431597 -11.369  < 2e-16 ***
volatileacidity     0.2135388  0.0498431   4.284 1.83e-05 ***
freesulfurdioxide  -0.0005924  0.0002678  -2.212   0.0270 *  
totalsulfurdioxide -0.0010107  0.0001666  -6.067 1.30e-09 ***
sulphates           0.1634235  0.0416529   3.923 8.73e-05 ***
chlorides           0.0921160  0.1204738   0.765   0.4445    
alcohol             0.0258436  0.0104461   2.474   0.0134 *  
ph                  0.2493023  0.0570182   4.372 1.23e-05 ***
labelappeal         1.0011030  0.0525751  19.041  < 2e-16 ***
acidindex           0.4391693  0.0278684  15.759  < 2e-16 ***
stars              -2.5681947  0.0924674 -27.774  < 2e-16 ***
nostars             3.2615363  0.0910239  35.832  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 96349759.1064 
Number of iterations in BFGS optimization: 39 
Log-likelihood: -1.997e+04 on 25 Df
> 
```

```
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                   -44.77786 model2 > model1 < 2.22e-16
AIC-corrected         -44.56047 model2 > model1 < 2.22e-16
BIC-corrected         -43.74998 model2 > model1 < 2.22e-16
> 
```

Figure: 16

Observation: Since almost 2,500 wines sold zero cases in this situation, zero-inflated models may be more accurate. The Vuong test compares the zero-inflated model with a standard Negative Binomial regression model. The Vuong test shows that our test statistic is significant, indicating that the zero-inflated model is an improvement over the standard Negative Binomial model.

5. Discussions:

We have the following points for the final interpretation and discussion of our models

a. Regression Coefficients of the Models

In Figure 17 we see the plot depicting the regression coefficients of the **Poisson** vs **Negative Binomial Models**. Both models assign almost similar coefficients to the predictors. Missing values for stars and acidindex values significantly harm sales while labelappeal and present higher values for stars lead to higher sales.

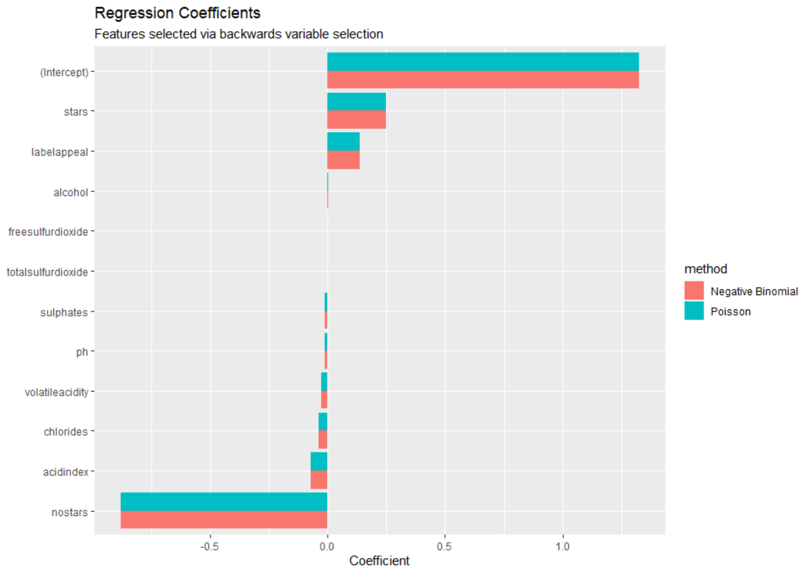


Figure: 17

In Figure 18, we see the regression coefficients from the Zero Inflated Models.

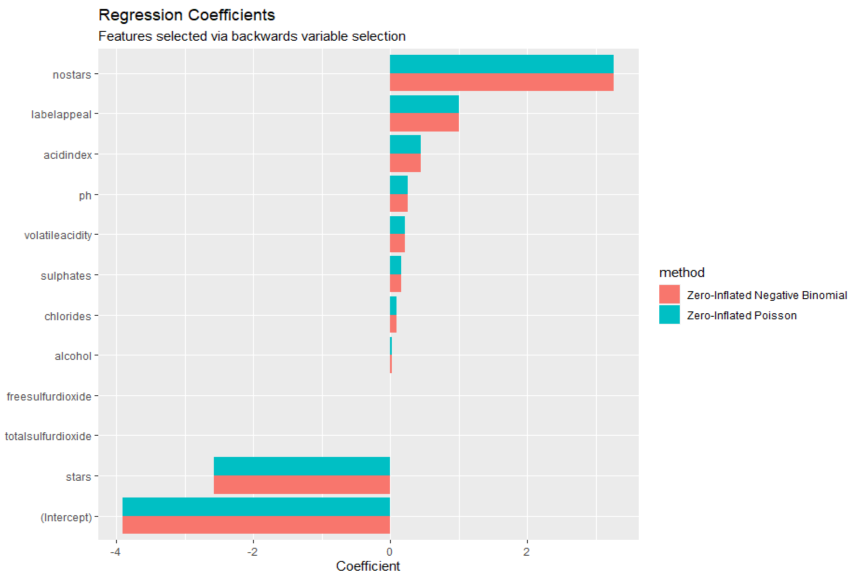


Figure: 18

Both sets of models use the same predictors. Some coefficients have been flipped in the zero-inflated models, Missing Stars and acidity values are strong positive influences on sales. Label appeal is similarly positive as in previous models, but actual star ratings are negative. Our assumption is that higher star ratings also come with a higher price tag, which can explain the negative coefficient for STARS.

b. Model Evaluation

The four models have been evaluated based on Root Mean Square Error(rmse), Mean Absolute Error(MAE), AIC, and BIC.

Model	rmse	mae	AIC	BIC
Poisson	2.578925	2.209098	44900.57	44990.05
Negative Binomial	2.578925	2.209097	44902.95	44999.89
Zero-Inflated Poisson	1.163609	0.869497	39981.29	40160.25
Zero-Inflated Negative Binomial	1.163609	0.869497	39983.29	40169.71

The error rates for the Zero Inflated models are lower than the regular models. Although the AIC and BIC values are slightly better for the Zero Inflated Poisson Models, our inference is to choose the **Zero Inflated Negative Binomial** Model because of the slight overdispersion present in the response variable and large number of zeros in the response variable.

6. References:

[1] Cameron A. C. and Trivedi P. K., (2013). Regression Analysis of Count Data, Second Edition, Econometric Society Monograph No. 53, Cambridge University Press, Cambridge.

[2] Dunn, P.K.; Smyth, G.K. (2018). Generalized Linear Models with Examples in R. New York: Springer. doi:10.1007/978-1-4419-0118-7.

[3] Kida, Y. (2019). Generalized Linear Models - Introduction to advanced statistical modeling. https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab.

7. Appendix:

7.1 Code Sections

Part 0	Import Libraries and load the data. Check Summary Statistics
Part 1	Exploratory Data Analysis.
Part 2	Data Preparation including outlier removal, imputing missing values and including flags
Part 3	Model Building
Part 4	Model Evaluation

7.2 R Source Code:

```
#Part 0: Load & Prepare Data. In  part 0, we import necessary
libraries, load
#and check the summary statistics of the data.


library(readr)
library(dplyr)
library(zoo)
library(psych)
library(ROCR)
library(corrplot)
library(car)
library(InformationValue)
library(rJava)
library(pbkrtest)
library(car)
library(leaps)
library(MASS)
library(corrplot)
library(glm2)
library(aod)
library(mice)
library(Hmisc)
library(xlsxjars)
library(xlsx)

library(VIM)
library(pROC)
library(pscl) # For "counting" models (e.g., Poisson and Negative
Binomial)
```

```
library(ggplot2) # For graphical tools
library(readr)
library(corrplot)


#setwd

# Read the wine dataset

wine=read.csv("Wine_Training.csv",header=T)

head(wine,1)
```

#Part 1: We do our exploratory data analysis.We check the distribution of each predictor

```
#Data Quality Check
str(wine)
summary(wine)

library(Hmisc)
describe(wine)

nulls <- data.frame(col = as.character(colnames(wine)),
                                                             pct_null =
colSums(is.na(wine))*100/(colSums(is.na(wine))+colSums(!is.na(wine
))))%>%
  filter(col != 'INDEX')
ggplot(nulls, aes(x = col, y = pct_null))+
  geom_bar(stat = 'identity')+
  coord_flip()+
  labs(title = 'Distribution of Missing Data',
       x = element_blank(), y = 'Percent of Information Missing')+
  ylim(0,100)

#TARGET
par(mfrow=c(1,2))
hist(wine$TARGET, col = "#A71990", xlab = "Target ", main =
"histogram of wine sales")
boxplot(wine$TARGET, col = "#A71990", main = "boxplot of wine
sales")
par(mfrow = c(1,1))
#Chemistry

# FixedAcidity and VolatileAcidity
dev.off()
par("mar")
par(mar=c(3,1,1,1))
par(mfrow=c(2,2))
hist(wine$FixedAcidity, col = "#A71930", xlab ="FixedAcidity",
main = "Histogram of FixedAcidity")
hist(wine$VolatileAcidity, col = "#A71990", xlab =
"VolatileAcidity", main = "Histogram of VolatileAcidity")
boxplot(wine$FixedAcidity, col = "#A71930", main = "Boxplot of
FixedAcidity")
```

```
boxplot(wine$VolatileAcidity, col = "#A71990", main = "Boxplot of
VolatileAcidity")
par(mfrow=c(1,1))

# CitricAcid and ResidualSugar
par(mfrow=c(2,2))
hist(wine$CitricAcid, col = "#D77730", xlab = "CitricAcid", main =
"Histogram of CitricAcid")
hist(wine$ResidualSugar, col = "#ABCEAC", xlab = "ResidualSugar ",
main = "Histogram of ResidualSugar")
boxplot(wine$CitricAcid,  col  =  "#D77730",  main  =  "Boxplot  of
CitricAcid")
boxplot(wine$ResidualSugar,  col  =  "#ABCEAC",  main  =  "Boxplot of
ResidualSugar")
par(mfrow=c(1,1))

#Chlorides and FreeSulfur Dioxide
par(mfrow=c(2,2))
hist(wine$Chlorides, col = "#A71930", xlab = "Chlorides", main =
"Histogram of Chlorides")
hist(wine$FreeSulfurDioxide,    col    =    "#DBCEAC",    xlab    =
"FreeSulfurDioxide ", main = "Histogram of FreeSulfurDioxide")
boxplot(wine$Chlorides,  col  =  "#A71930",  main  =  "Boxplot  of
Chlorides")
boxplot(wine$FreeSulfurDioxide, col = "#DBCEAC", main = "Boxplot
of FreeSulfurDioxide")
par(mfrow=c(1,1))

#TotalSulfurDioxide and Density
par(mfrow=c(1,1))
hist(wine$TotalSulfurDioxide,    col    =    "#D77730",    xlab    =
"TotalSulfurDioxide", main = "Histogram of TotalSulfurDioxide")
hist(wine$Density,  col  =  "#DBCEAC",  xlab  =  "Density",  main  =
"Histogram of Density")
boxplot(wine$TotalSulfurDioxide, col = "#D77730", main = "Boxplot
of TotalSulfurDioxide")
boxplot(wine$Density,   col   =   "#DBCEAC",   main   =   "Boxplot   of
Density")
par(mfrow=c(1,1))

#pH and Sulphates
par(mfrow=c(2,2))
hist(wine$pH, col = "#A71930", xlab = "pH", main = "Histogram of
pH")
hist(wine$Sulphates, col = "#09ADAD", xlab = "Sulphates", main =
"Histograms of Sulphates")
boxplot(wine$pH, col = "#A71930", main = "Boxplot of pH")
boxplot(wine$Sulphates,  col  =  "#09ADAD",  main  =  "Boxplot  of
Sulphates")
par(mfrow=c(1,1))


#Alcohol and Acid Index
par(mfrow=c(2,2))
hist(wine$Alcohol,  col  =  "#A71930",  xlab  =  "Alcohol",  main  =
"Histogram of Alcohol")
```

```
hist(wine$AcidIndex, col = "#DBCEAC", xlab = "AcidIndex", main =
"Histograms of AcidIndex")
boxplot(wine$Alcohol, col = "#A71930", main = "Boxplot of
Alcohol")
boxplot(wine$AcidIndex, col = "#DBCEAC", main = "Boxplot of
AcidIndex")
par(mfrow=c(1,1))

#Label Appeal and STARS
par(mfrow=c(2,2))
hist(wine$LabelAppeal, col = "#DBCEAC", xlab = "LabelAppeal", main
= "Histogram of LabelAppeal ")
hist(wine$STARS, col = "#09ADAD", xlab = "STARS", main =
"Histogram of STARS")
boxplot(wine$LabelAppeal, col = "#DBCEAC", main = "Boxplot of
LabelAppeal")
boxplot(wine$STARS, col = "#09ADAD", main = "Boxplot of STARS")
par(mfrow=c(1,1))



##################################################################
###########################
```

##Part 2: Data Preparation

```
##In Part 2, we do data preparation to create our models. We do
NULL #handling, create flags to indicate which predictors had
NULLS or #significant amount of outliers.

#Check missing data percentage
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(wine,2,pMiss)

# Outliers, as observed from predictor histograms and boxplots are
#removed and nulls are imputed
# using predictive mean matching.

wine$NoResidualSugar <- 0
wine$NoResidualSugar [is.na(wine$ResidualSugar)] <- 1

wine$NoChlorides  <- 0
wine$NoChlorides [is.na(wine$Chlorides)] <- 1

wine$NoFreeSulfurDioxide <- 0
wine$NoFreeSulfurDioxide[is.na(wine$FreeSulfurDioxide)] <- 1

wine$NoTotalSulfurDioxide <- 0
wine$NoTotalSulfurDioxide[is.na(wine$TotalSulfurDioxide)] <- 1

wine$NopH <- 0
wine$NopH[is.na(wine$pH)] <- 1

wine$NoSulphates <- 0
wine$NoSulphates [is.na(wine$Sulphates)] <- 1

wine$NoResidualSugar <- 0
wine$NoResidualSugar [is.na(wine$ResidualSugar)] <- 1
```

```
wine$NoAlcohol <- 0
wine$NoAlcohol [is.na(wine$Alcohol)] <- 1

wine$NoSTARS<- 0
wine$NoSTARS [is.na(wine$STARS)] <- 1

str(wine)

colnames(wine) <- tolower(colnames(wine))



library(mice)
tempData <- mice(wine,m=5,maxit=50,meth='pmm',seed=500)
summary(tempData)


train <- complete(tempData,1)
apply(train,2,pMiss)
summary(train)

colnames(train) <- tolower(colnames(train))
names(train)

###################################################################
##############################


#Correlation Matrix
correl <- subset(train, select=c(
  'fixedacidity',
  'volatileacidity',
  'citricacid',
  'residualsugar',
  'chlorides',
  'freesulfurdioxide',
  'totalsulfurdioxide',
  'density',
  'ph',
  'sulphates',
  'alcohol',
  'labelappeal',
  'acidindex',
  'stars',
  'noresidualsugar',
  'nochlorides',
  'nofreesulfurdioxide',
  'nototalsulfurdioxide',
  'noph',
  'nosulphates',
  'noalcohol',
  'nostars',
  'target'))

require(corrplot)
```

```
mcor <- cor(correl)
corrplot(mcor,           method="number",           shade.col=NA,
tl.col="black",tl.cex=0.8)
par(mfrow=c(1,1))

#################################################################
#####################################################
```

 # Part3: Model Building

```
#Model 1:Poisson

library(MASS)

base_poisson <- glm(target ~ ., family="poisson", data=train)
summary(base_poisson)
#Using AIC Stepwise for variable selection

poisson.back     <-     stepAIC(base_poisson,     direction     =
'backward',trace=0)
summary(poisson.back)
poiss.model   <-   glm(formula   =   target   ~   volatileacidity   +
+freesulfurdioxide+ totalsulfurdioxide +
                            sulphates + chlorides+ alcohol + ph+
labelappeal + acidindex +
                    stars + nostars,
                  family = "poisson", data = train)

poisson.coeffs          <-          data.frame(var          =
names(poiss.model$coefficients),
                                                  coefficient  =
poiss.model$coefficients)%>%
  mutate(method = 'Poisson')

anova(poiss.model, test="Chisq")

with(poiss.model, cbind(res.deviance = deviance, df = df.residual,
                                p  =  pchisq(deviance,  df.residual,
lower.tail=FALSE)))

library(AER)
deviance(poiss.model)/poiss.model$df.residual
dispersiontest(poiss.model)

#what type of dispersion does sample have?
mean(train$target)
var(train$target)

library(car)
influencePlot(poiss.model)
res <- residuals(poiss.model, type="deviance")
plot(log(predict(poiss.model)), res)
abline(h=0, lty=2)
qqnorm(res)
qqline(res)
```

```
#####################################################################
########################################

#Model2: Negative Binomial

base_nb <- glm.nb(target ~ ., data=train)
#Using AIC Stepwise for variable selection
nb.back <- stepAIC(base_nb, direction = 'backward',trace=0)

summary(nb.back)

negbinomial.mod <- glm.nb(formula = target ~ volatileacidity +
+freesulfurdioxide+ totalsulfurdioxide +
                              sulphates + chlorides+ alcohol + ph+
labelappeal + acidindex +
                           stars + nostars,data = train)

negbinomial.coeffs <- data.frame(var =
names(negbinomial.mod$coefficients),
                                                   coefficient =
negbinomial.mod$coefficients)%>%
  mutate(method = 'Negative Binomial')
odTest(negbinomial.mod)

summary(negbinomial.mod)

with(negbinomial.mod, cbind(res.deviance = deviance, df =
df.residual,
                                p = pchisq(deviance, df.residual,
lower.tail=FALSE)))

library(ggplot2)

ggplot(bind_rows(negbinomial.coeffs, poisson.coeffs),
        aes(x = reorder(var, coefficient), y = coefficient, fill =
method))+
  geom_col(position = 'dodge')+
  coord_flip()+
  labs(y = 'Coefficient',
       x = element_blank(),
       title = 'Regression Coefficients',
           subtitle = 'Features selected via backwards variable
selection')



#theme_gray()

###############################################

# Zero Inflated Regression

# Zero Inflated Poisson

zinp.mod <- pscl::zeroinfl(formula = target ~ volatileacidity
+freesulfurdioxide+ totalsulfurdioxide +
```

```
                                  sulphates + chlorides+ alcohol +
ph+labelappeal + acidindex +
                             stars + nostars,
                           data = train)


zinp.coeffs <- data.frame(var = names(zinp.mod$coefficients$zero),
                                                      coefficient  =
zinp.mod$coefficients$zero)%>%
  mutate(method = 'Zero-Inflated Poisson')

summary(zinp.mod)

vuong(poiss.model, zinp.mod)

#Zero Inflated Neg Binom

zinng.mod  <-  pscl::zeroinfl(formula  =  target  ~  volatileacidity
+freesulfurdioxide+ totalsulfurdioxide +
                              sulphates + chlorides+ alcohol + ph+
labelappeal + acidindex +
                              stars + nostars,
                            data = train, dist = "negbin")
zinng.coeffs            <-            data.frame(var             =
names(zinng.mod$coefficients$zero),
                                                      coefficient  =
zinng.mod$coefficients$zero)%>%
  mutate(method = 'Zero-Inflated Negative Binomial')

summary(zinng.mod)

vuong(negbinomial.mod, zinng.mod)


#Part  4,  we  compare  the  model  regression  coefficients.  We  also
#compare the models based on
#Root Mean Squared Error(rmse), Mean Absolute Error(MAE) and their
#AIC and BIC scores.
#Based on everything, we choose the final model

# Comparing Coefficients of Zero Inflated Models

ggplot(bind_rows(zinp.coeffs, zinng.coeffs),
        aes(x = reorder(var, coefficient), y = coefficient, fill =
method))+
  geom_col(position = 'dodge')+
  coord_flip()+
  labs(y = 'Coefficient',
       x = element_blank(),
       title = 'Regression Coefficients',
           subtitle  =  'Features  selected  via  backwards  variable
selection')+

  theme_gray()
```

```
##################################################################
##################
# Model Evaluation
# Calculating mae and rmse on full train data.We also calculate
AIC #and BIC score s of all
# the models

library(ModelMetrics)

columns <- c('Poisson', 'Negative Binomial','Zero-Inflated
Poisson','Zero-Inflated Negative Binomial')

poiss.mae <- mae(train$target, predict(poiss.model))
poiss.rmse <- rmse(train$target, predict(poiss.model))
AIC.poiss <- AIC(poiss.model)
BIC.poiss <- BIC(poiss.model)

negbin.mae <- mae(train$target, predict(negbinomial.mod))
negbin.rmse <- rmse(train$target, predict(negbinomial.mod))
AIC.nbr <- AIC(negbinomial.mod)
BIC.nbr <- BIC(negbinomial.mod)

zinp.mae <- mae(train$target, predict(zinp.mod))
zinp.rmse <- rmse(train$target, predict(zinp.mod))
AIC.zinp <- AIC(zinp.mod)
BIC.zinp <- BIC(zinp.mod)

zinng.mae <- mae(train$target, predict(zinng.mod))
zinng.rmse <- rmse(train$target, predict(zinng.mod))
AIC.zinng <- AIC(zinng.mod)
BIC.zinng <- BIC(zinng.mod)

data.frame(
  columns,
  rmse = c(poiss.rmse, negbin.rmse, zinp.rmse, zinng.rmse),
  mae = c(poiss.mae, negbin.mae, zinp.mae, zinng.mae),
  AIC = c(AIC.poiss,AIC.nbr,AIC.zinp,AIC.zinng),
  BIC = c(BIC.poiss,BIC.nbr,BIC.zinp,BIC.zinng)
)
```