Count Regression Model

Abdellah AitElmouden | Gabriel Abreu | Jered Ataky | Patrick Maloney

5/22/2021

Introduction

The goal of this assignment is to explore, analyze and model a dataset containing information on approximately 12,000 commercially available wines. The dataset 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. The target variable, cases of wine sold, is count data and therefore will be modeled using appropriate techniques such as Poisson and Negative Binomial regressions.

Data Exploration

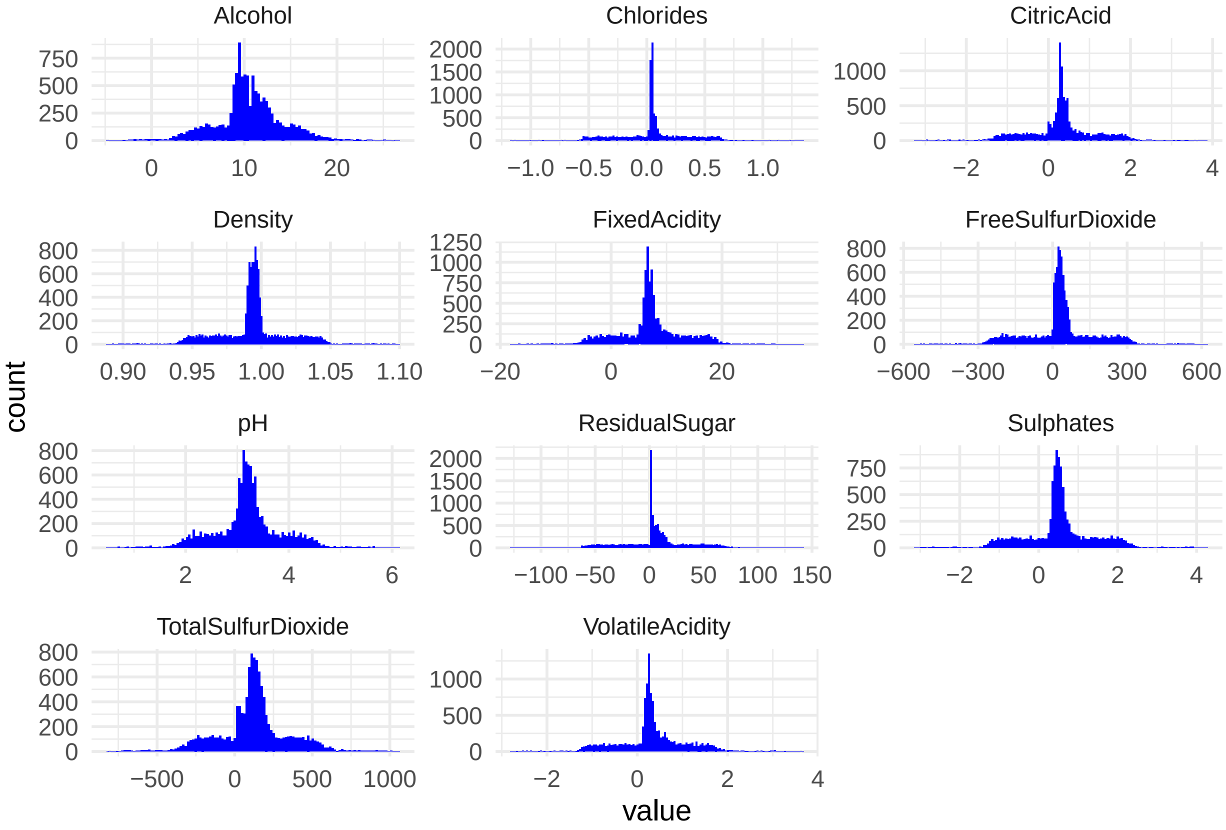
```
## Rows: 12,795
## Columns: 16
## $ ï..INDEX           <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19~
## $ TARGET             <int> 3, 3, 5, 3, 4, 0, 0, 4, 3, 6, 0, 4, 3, 7, 4, 0, 0, ~
## $ FixedAcidity       <dbl> 3.2, 4.5, 7.1, 5.7, 8.0, 11.3, 7.7, 6.5, 14.8, 5.5,~
## $ VolatileAcidity    <dbl> 1.160, 0.160, 2.640, 0.385, 0.330, 0.320, 0.290, -1~
## $ CitricAcid         <dbl> -0.98, -0.81, -0.88, 0.04, -1.26, 0.59, -0.40, 0.34~
## $ ResidualSugar      <dbl> 54.20, 26.10, 14.80, 18.80, 9.40, 2.20, 21.50, 1.40~
## $ Chlorides          <dbl> -0.567, -0.425, 0.037, -0.425, NA, 0.556, 0.060, 0.~
## $ FreeSulfurDioxide  <dbl> NA, 15, 214, 22, -167, -37, 287, 523, -213, 62, 551~
## $ TotalSulfurDioxide <dbl> 268, -327, 142, 115, 108, 15, 156, 551, NA, 180, 65~
## $ Density            <dbl> 0.99280, 1.02792, 0.99518, 0.99640, 0.99457, 0.9994~
## $ pH                 <dbl> 3.33, 3.38, 3.12, 2.24, 3.12, 3.20, 3.49, 3.20, 4.9~
## $ Sulphates          <dbl> -0.59, 0.70, 0.48, 1.83, 1.77, 1.29, 1.21, NA, 0.26~
## $ Alcohol            <dbl> 9.9, NA, 22.0, 6.2, 13.7, 15.4, 10.3, 11.6, 15.0, 1~
## $ LabelAppeal        <int> 0, -1, -1, -1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 2, 0, 0, ~
## $ AcidIndex          <int> 8, 7, 8, 6, 9, 11, 8, 7, 6, 8, 5, 10, 7, 8, 9, 8, 9~
## $ STARS              <int> 2, 3, 3, 1, 2, NA, NA, 3, NA, 4, 1, 2, 2, 3, NA, NA~
```

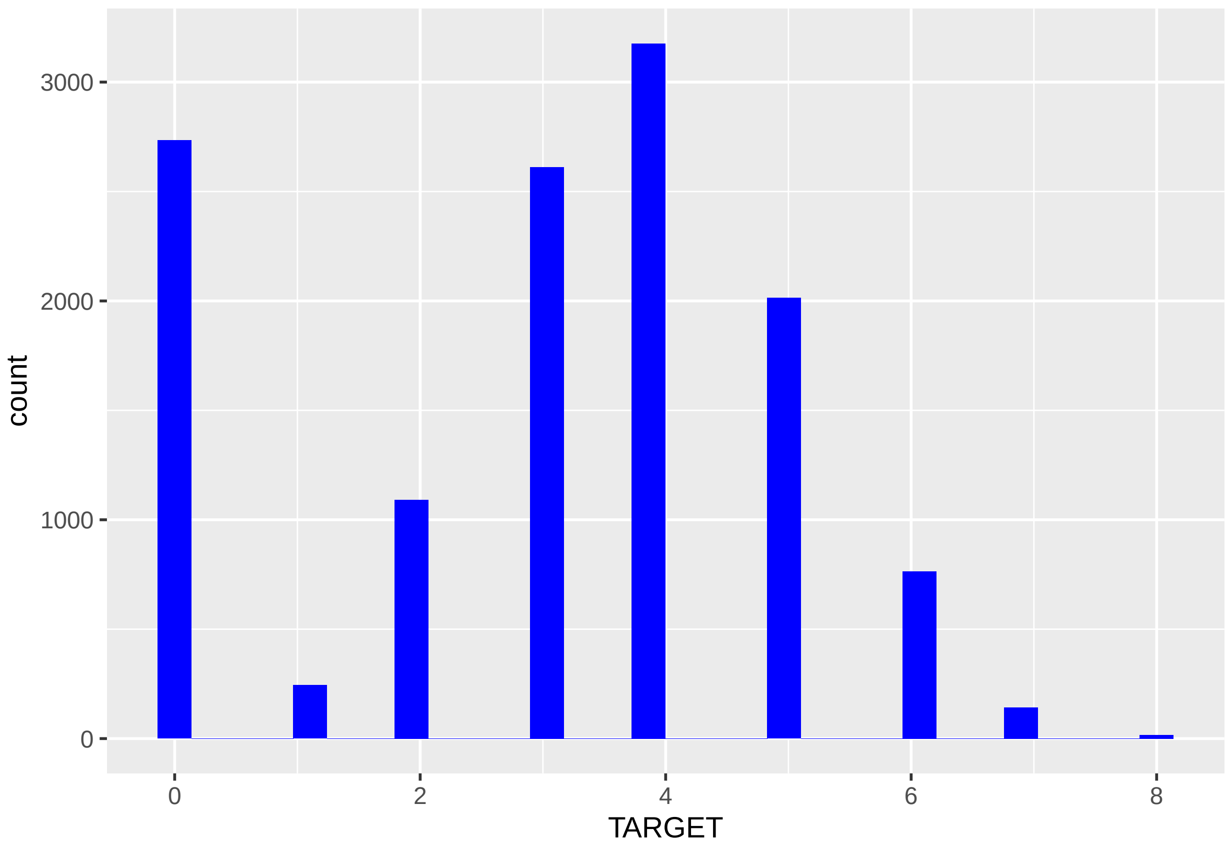
All the variable in this dataset are numeric and continuous except for AcidIndex, STARS and LabelAppeal which are discrete. The target variable TARGET is also discrete. There are a number of missing observations for certain chemical composition variables as well as a large number of wines with no STARS rating. The distribution of the continuous variables appear well centered.

```
##     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
```

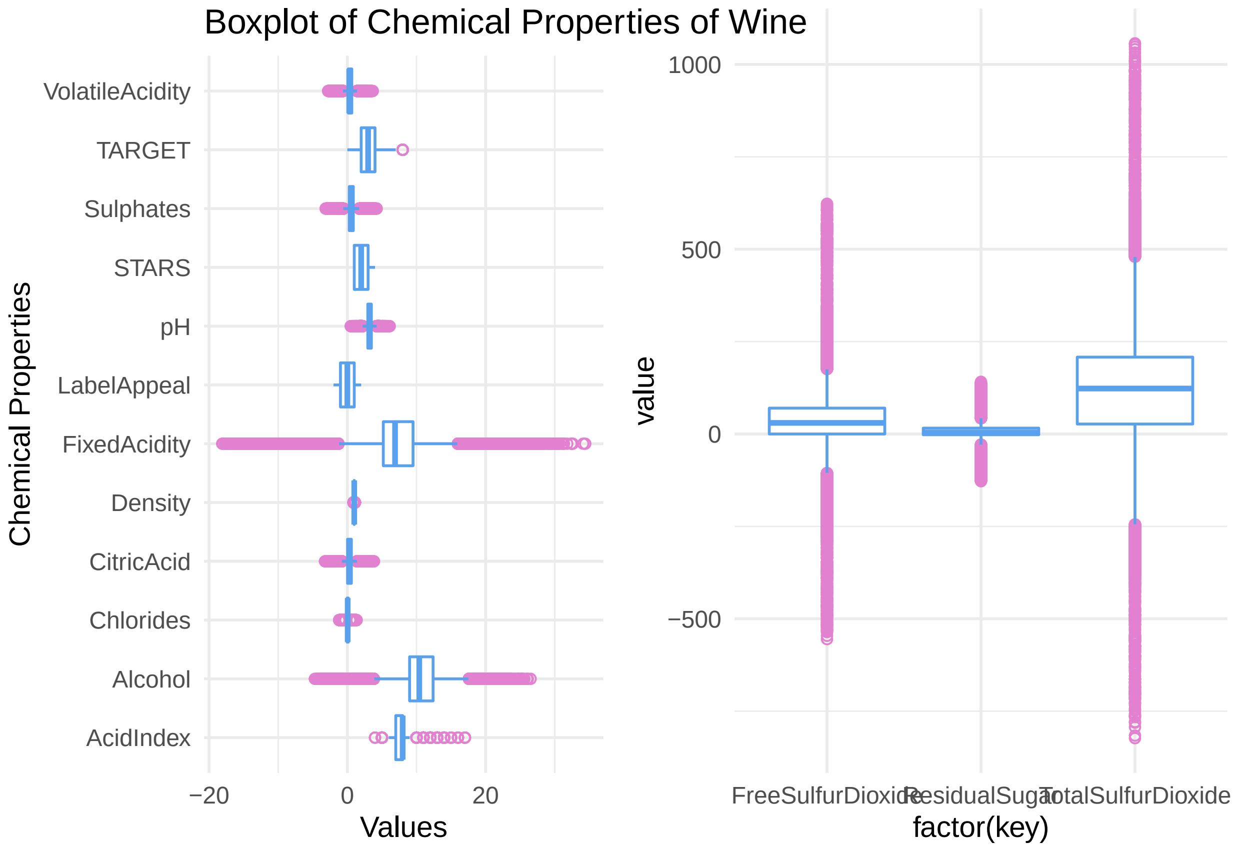
Visualization

To take a look at the distributions of the dataset, we plot some histograms and can see that all continuous variables are centered and close to normally distributed. We also note that some variables are centered around zero and take negative values which is unexpected and will be investigated further and transformed for our analysis. The distribution of the TARGET variable looks like it could be well described by the poisson distribution which has equal mean and variance but the high number of zero values justifies the use of a zero-inflated model as well. The mean and variance of TARGET are 3.02 and 3.71 respectively, which is close enough to satisfy the equal mean-variance assumption of the poisson distribution.


Alcohol
Chlorides
CitricAcid
Density
FixedAcidity
FreeSulfurDioxide
pH
ResidualSugar
Sulphates
TotalSulfurDioxide
VolatileAcidity
count
value
750
500
250
0
0
10
20
2000
1500
1000
500
0
−1.0
−0.5
0.0
0.5
1.0
1000
500
0
−2
0
2
4
800
600
400
200
0
0.90
0.95
1.00
1.05
1.10
1250
1000
750
500
250
0
−20
0
20
800
600
400
200
0
−600
−300
0
300
600
800
600
400
200
0
2
4
6
2000
1500
1000
500
0
−100
−50
0
50
100
150
750
500
250
0
−2
0
2
4
800
600
400
200
0
−500
0
500
1000
1000
500
0
−2
0
2
4



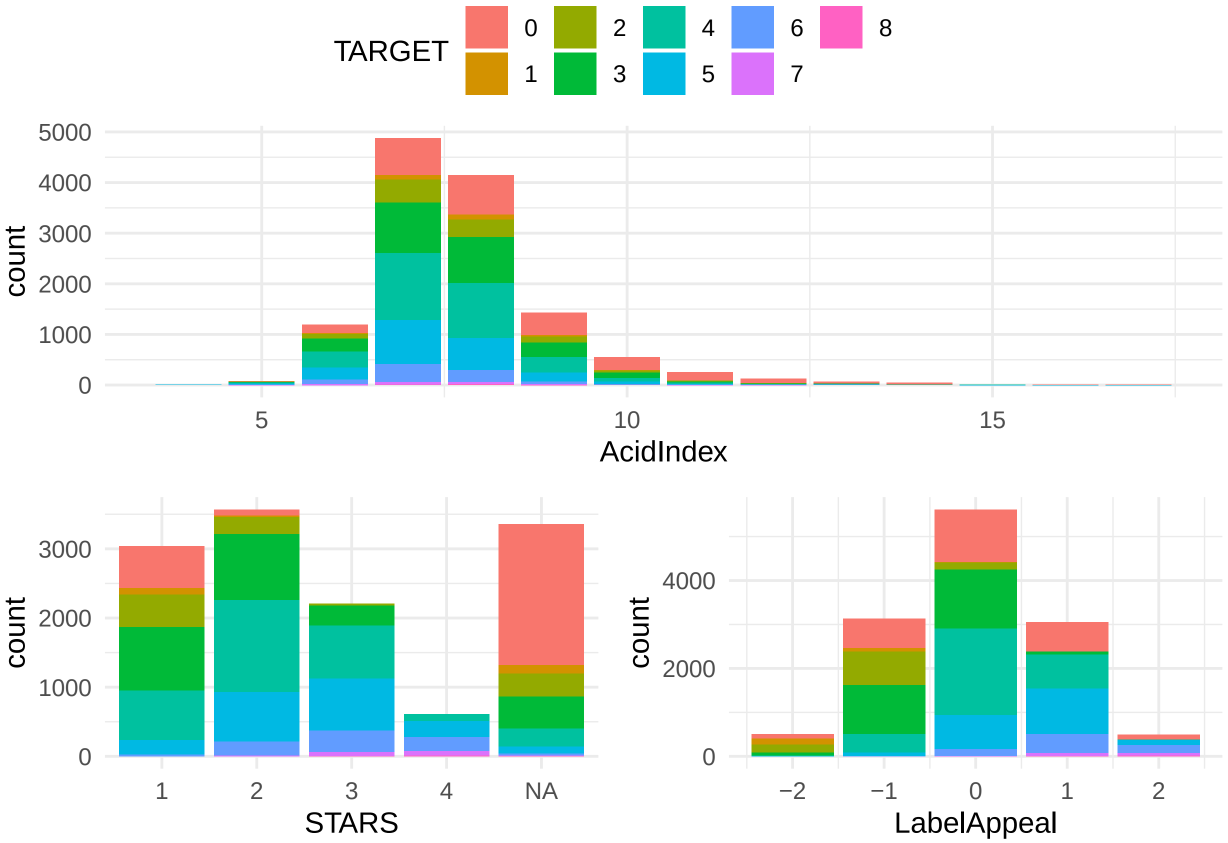
below box plot shows several variables in to two panels. Some variables such as TotalSulfurDioxide, FreeSulfurDioxide, and ResidualSugar have large ranges compared to other variables. Therefore, we separated those variables in to a different panel to view their distribution. From both panel, we can tell a high number of variables have numerous outliers.



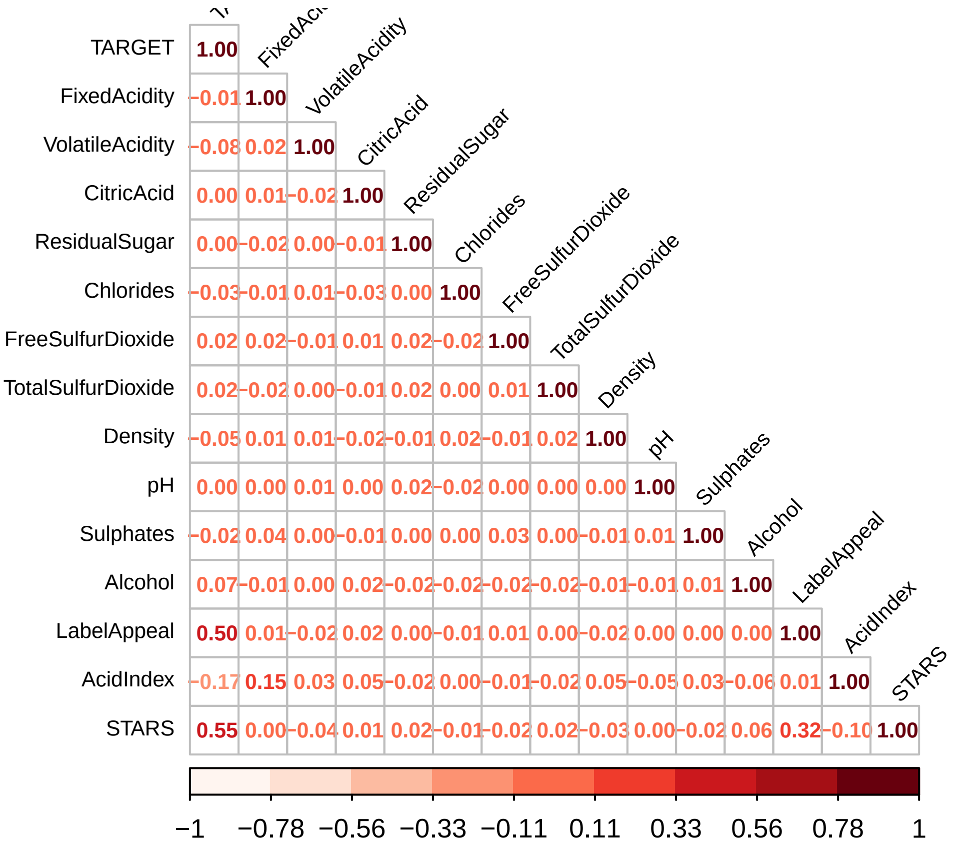
by taking a look at the bar char below, we can conclude the following point:

- AcidIndex tells us that large quantity of wine were sold with the index number 7 and 8.
- LabelAppeal tells us generic labeled wine sells the most;
- However, better label does yield higher number of wine samples per order.
- STARS tells us excellent quality does not result in high wine orders. It could be due to high star wine bottle's high price tag.

	Correlation
TARGET	1.0000000
STARS	0.5546857
LabelAppeal	0.4979465
Alcohol	0.0737771
FreeSulfurDioxide	0.0226398
TotalSulfurDioxide	0.0216021
ResidualSugar	0.0035196
CitricAcid	0.0023450
pH	0.0002199
FixedAcidity	-0.0125381
Sulphates	-0.0212204
Chlorides	-0.0304301
Density	-0.0475989
VolatileAcidity	-0.0759979
AcidIndex	-0.1676431



The plot below represent the correlation table and plot, and we can see that STARS and LabelAppeal are most positively correlated variables with the response variable. We expected this because our variable description mentions these variable's theoretical affect are higher than other variables. Also, we some mild negative correlation between the response variable and AcidIndex variable.



	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol	LabelAppeal	AcidIndex	STARS
TARGET	1.00														
FixedAcidity	-0.01	1.00													
VolatileAcidity	-0.08	0.02	1.00												
CitricAcid	0.00	0.01	-0.02	1.00											
ResidualSugar	0.00	-0.02	0.00	-0.01	1.00										
Chlorides	-0.03	-0.01	0.01	-0.03	0.00	1.00									
FreeSulfurDioxide	0.02	0.02	-0.01	0.01	0.02	-0.02	1.00								
TotalSulfurDioxide	0.02	-0.02	0.00	-0.01	0.02	0.00	0.01	1.00							
Density	-0.05	0.01	0.01	-0.02	-0.01	0.02	-0.01	0.02	1.00						
pH	0.00	0.00	0.01	0.00	0.02	-0.02	0.00	0.00	0.00	1.00					
Sulphates	-0.02	0.04	0.00	-0.01	0.00	0.00	0.03	0.00	-0.01	0.01	1.00				
Alcohol	0.07	-0.01	0.00	0.02	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01	0.01	1.00			
LabelAppeal	0.50	0.01	-0.02	0.02	0.00	-0.01	0.01	0.00	-0.02	0.00	0.00	0.00	1.00		
AcidIndex	-0.17	0.15	0.03	0.05	-0.02	0.00	-0.01	-0.02	0.05	-0.05	0.03	-0.06	0.01	1.00	
STARS	0.55	0.00	-0.04	0.01	0.02	-0.01	-0.02	0.02	-0.03	0.00	-0.02	0.06	0.32	-0.10	1.00

Data Preparation

To explore the missing variables we Used the aggr function from VIM package, from the plot we see several variables have missing values. According to UCI Machine Learning, who published this dataset, all wine contain some natural sulfites. Therefore, to avoid creating problems while analyzing our data we will impute the missing values for sulfite chemical properties. Also We will impute values of wines with less than 1 gram/liter of sugar. Matter of fact, since all missing values are missing at random, we will impute all the missing values using the mice package and random forest method. Mice package uses multivariate imputations to estimate the missing values. Using multiple imputations helps in resolving the uncertainty for the missing values. Our target variable will be removed as a predictor variable but still will be imputed. Our response variables will be removed as predictor variables but still will be imputed.

```
##
##  iter imp variable
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
```

```
##
##  iter imp variable
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
```

```
## [1] "Missing value after imputation: 0"
```

Model Building

Model 1: Poisson (Raw data)

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = train_data)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2158  -0.2734   0.0616   0.3732   1.6830
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.593e+00  2.506e-01   6.359 2.03e-10 ***
## FixedAcidity        3.293e-04  1.053e-03   0.313  0.75447
## VolatileAcidity    -2.560e-02  8.353e-03  -3.065  0.00218 **
## CitricAcid         -7.259e-04  7.575e-03  -0.096  0.92365
## ResidualSugar      -6.141e-05  1.941e-04  -0.316  0.75165
## Chlorides          -3.007e-02  2.056e-02  -1.463  0.14346
## FreeSulfurDioxide   6.734e-05  4.404e-05   1.529  0.12620
## TotalSulfurDioxide  2.081e-05  2.855e-05   0.729  0.46618
## Density            -3.725e-01  2.462e-01  -1.513  0.13026
## pH                 -4.661e-03  9.598e-03  -0.486  0.62722
## Sulphates          -5.164e-03  7.051e-03  -0.732  0.46398
## Alcohol             3.948e-03  1.771e-03   2.229  0.02579 *
## LabelAppeal         1.771e-01  7.954e-03  22.271  < 2e-16 ***
## AcidIndex          -4.870e-02  5.903e-03  -8.251  < 2e-16 ***
## STARS               1.871e-01  7.487e-03  24.993  < 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: 5844.1  on 6435  degrees of freedom
## Residual deviance: 4009.1  on 6421  degrees of freedom
##   (6359 observations deleted due to missingness)
## AIC: 23172
##
## Number of Fisher Scoring iterations: 5
```

```
## [1] "Goodness of Fit Test:"
```

```
##      res.deviance   df p
## [1,]     4009.142 6421 1
```

From the output we can say that the deviance residuals is quite symmetrical. This means that the predicted points are close to actual observed points. As can be seen in our correlation table, this is as predicted. STARS, LabelAppeal and AcidIndex are significant variables. And the variation in standard error is low. The goodness of fit test has a high p value which indicates that the model fits the data well.

Model 2: Poisson (Imputed Data)

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = train_data_imputed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8387  -0.5031   0.2162   0.6282   2.6384
##
```

```
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)    
## (Intercept)          2.131e+00  1.956e-01  10.898  < 2e-16 ***
## FixedAcidity        -2.187e-04  8.196e-04  -0.267 0.789613    
## VolatileAcidity     -5.200e-02  6.480e-03  -8.024 1.03e-15 ***
## CitricAcid           1.371e-02  5.893e-03   2.326 0.020010 *  
## ResidualSugar        1.253e-04  1.504e-04   0.833 0.404695    
## Chlorides           -5.865e-02  1.601e-02  -3.663 0.000250 ***
## FreeSulfurDioxide    1.510e-04  3.425e-05   4.407 1.05e-05 ***
## TotalSulfurDioxide   1.178e-04  2.206e-05   5.337 9.46e-08 ***
## Density             -4.318e-01  1.921e-01  -2.247 0.024623 *  
## pH                  -2.156e-02  7.524e-03  -2.865 0.004169 ** 
## Sulphates           -1.920e-02  5.455e-03  -3.519 0.000433 ***
## Alcohol              5.684e-03  1.376e-03   4.131 3.62e-05 ***
## LabelAppeal          2.001e-01  6.087e-03  32.868  < 2e-16 ***
## AcidIndex           -1.240e-01  4.453e-03 -27.851  < 2e-16 ***
## STARS                1.665e-01  5.822e-03  28.598  < 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: 18736  on 12780  degrees of freedom
## AIC: 50708
## 
## Number of Fisher Scoring iterations: 5
```

```
## [1] "Goodness of Fit Test:"
```

```
##      res.deviance    df             p
## [1,]     18735.63 12780 2.506431e-234
```

The deviance residuals are the same as before. Imputation, on the other hand, introduces more important variables into the model. Furthermore, the AIC score increased significantly from 23172 to 50384. In addition, the deviance residuals fell from 40,000 to 18,412. The goodness of fit measure, on the other hand, has a very low p value, indicating that this model does not fit the data well. Since the residual deviance is greater than the degrees of freedom, then some over-dispersion exists.

Given what we observed when looking at the distribution of TARGET, we should expect the inflated count of zeros to affect the model and bias results. For this reason, we move to a zero-inflated model to reflect this.

Model 3: Quasipoisson Model

We try a quasipoisson model to account for any overdispersion and to see if the results change significantly. As seen in the summary below, the models are nearly identical.

```
## 
## Call:
## glm(formula = TARGET ~ ., family = quasipoisson(link = "log"), 
##     data = train_data_imputed)
## 
## Deviance Residuals: 
```

```
##     Min       1Q   Median       3Q      Max
## -3.8387  -0.5031   0.2162   0.6282   2.6384
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.131e+00  1.923e-01  11.081  < 2e-16 ***
## FixedAcidity       -2.187e-04  8.061e-04  -0.271 0.786171
## VolatileAcidity    -5.200e-02  6.373e-03  -8.159 3.71e-16 ***
## CitricAcid          1.371e-02  5.796e-03   2.365 0.018034 *
## ResidualSugar       1.253e-04  1.479e-04   0.847 0.396869
## Chlorides          -5.865e-02  1.575e-02  -3.724 0.000197 ***
## FreeSulfurDioxide   1.510e-04  3.369e-05   4.481 7.48e-06 ***
## TotalSulfurDioxide  1.178e-04  2.170e-05   5.426 5.86e-08 ***
## Density            -4.318e-01  1.890e-01  -2.285 0.022330 *
## pH                 -2.156e-02  7.400e-03  -2.913 0.003584 **
## Sulphates          -1.920e-02  5.365e-03  -3.578 0.000347 ***
## Alcohol             5.684e-03  1.353e-03   4.200 2.69e-05 ***
## LabelAppeal         2.001e-01  5.986e-03  33.419  < 2e-16 ***
## AcidIndex          -1.240e-01  4.379e-03 -28.318  < 2e-16 ***
## STARS               1.665e-01  5.726e-03  29.078  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.9672522)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 18736  on 12780  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
## [1] "Goodness of Fit Test:"
```

```
##      res.deviance    df             p
## [1,]     18735.63 12780 2.506431e-234
```

Model 4: Zero Inflated

We saw earlier that the dependent variable had an excess number of zeros which skewed the distribution from a typical poisson. The zero inflated model generates coefficients for the zero count part of the model as well as for the count part.

```
##
## Call:
## zeroinfl(formula = TARGET ~ ., data = train_data_imputed, dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.2647 -0.3637  0.1660  0.5024  4.5568
##
## Count model coefficients (poisson with log link):
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.436e+00  2.064e-01   6.957 3.48e-12 ***
## FixedAcidity        2.761e-04  8.583e-04   0.322  0.74772    
## VolatileAcidity    -1.340e-02  6.869e-03  -1.950  0.05113 .  
## CitricAcid          1.665e-06  6.141e-03   0.000  0.99978    
## ResidualSugar      -8.537e-05  1.574e-04  -0.542  0.58768    
## Chlorides          -1.864e-02  1.689e-02  -1.103  0.26983    
## FreeSulfurDioxide   3.421e-05  3.517e-05   0.973  0.33071    
## TotalSulfurDioxide -2.484e-05  2.242e-05  -1.108  0.26781    
## Density            -3.173e-01  2.025e-01  -1.567  0.11715    
## pH                  7.552e-03  7.929e-03   0.952  0.34089    
## Sulphates          -4.784e-04  5.758e-03  -0.083  0.93378    
## Alcohol             7.454e-03  1.431e-03   5.208 1.90e-07 ***
## LabelAppeal         2.490e-01  6.413e-03  38.830  < 2e-16 ***
## AcidIndex          -1.632e-02  4.985e-03  -3.274  0.00106 ** 
## STARS               9.228e-02  6.282e-03  14.690  < 2e-16 ***
## 
## Zero-inflation model coefficients (binomial with logit link):
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)        -6.0101503  1.0351369  -5.806 6.39e-09 ***
## FixedAcidity        0.0022671  0.0043130   0.526  0.59913    
## VolatileAcidity     0.2410929  0.0346521   6.958 3.46e-12 ***
## CitricAcid         -0.0808575  0.0313726  -2.577  0.00996 ** 
## ResidualSugar      -0.0012063  0.0007978  -1.512  0.13049    
## Chlorides           0.2513384  0.0847036   2.967  0.00300 ** 
## FreeSulfurDioxide  -0.0007432  0.0001806  -4.115 3.86e-05 ***
## TotalSulfurDioxide -0.0008516  0.0001157  -7.359 1.85e-13 ***
## Density             0.8678799  1.0174164   0.853  0.39365    
## pH                  0.1862523  0.0400600   4.649 3.33e-06 ***
## Sulphates           0.1201562  0.0291157   4.127 3.68e-05 ***
## Alcohol             0.0086516  0.0072511   1.193  0.23282    
## LabelAppeal         0.3177151  0.0331974   9.570  < 2e-16 ***
## AcidIndex           0.4798872  0.0196945  24.366  < 2e-16 ***
## STARS              -0.4876110  0.0340780 -14.309  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Number of iterations in BFGS optimization: 36 
## Log-likelihood: -2.257e+04 on 30 Df
```

```
## 
## Call:
## zeroinfl(formula = TARGET ~ . - FixedAcidity - Density, data = train_data_imputed, 
##     dist = "poisson")
## 
## Pearson residuals:
##     Min      1Q  Median      3Q     Max 
## -2.2637 -0.3611  0.1657  0.5028  4.2429 
## 
## Count model coefficients (poisson with log link):
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         1.123e+00  5.218e-02  21.528  < 2e-16 ***
## VolatileAcidity    -1.348e-02  6.869e-03  -1.963  0.04966 *  
## CitricAcid          8.106e-05  6.141e-03   0.013  0.98947    
## ResidualSugar      -8.645e-05  1.574e-04  -0.549  0.58294    
## Chlorides          -1.934e-02  1.688e-02  -1.145  0.25207    
## FreeSulfurDioxide   3.317e-05  3.515e-05   0.944  0.34541    
## TotalSulfurDioxide -2.555e-05  2.241e-05  -1.140  0.25422    
## pH                  7.531e-03  7.930e-03   0.950  0.34229    
## Sulphates          -3.744e-04  5.756e-03  -0.065  0.94814    
## Alcohol             7.471e-03  1.431e-03   5.221 1.78e-07 ***
## LabelAppeal         2.490e-01  6.412e-03  38.831  < 2e-16 ***
## AcidIndex          -1.647e-02  4.939e-03  -3.336  0.00085 ***
## STARS               9.240e-02  6.282e-03  14.709  < 2e-16 ***
## 
## Zero-inflation model coefficients (binomial with logit link):
##                      Estimate Std. Error z value Pr(>|z|)    
## (Intercept)        -5.1500574  0.2384962 -21.594  < 2e-16 ***
## VolatileAcidity     0.2418331  0.0346319   6.983 2.89e-12 ***
## CitricAcid         -0.0813522  0.0313463  -2.595  0.00945 ** 
## ResidualSugar      -0.0012059  0.0007977  -1.512  0.13057    
## Chlorides           0.2517730  0.0846799   2.973  0.00295 ** 
## FreeSulfurDioxide  -0.0007424  0.0001806  -4.111 3.94e-05 ***
## TotalSulfurDioxide -0.0008507  0.0001157  -7.353 1.93e-13 ***
## pH                  0.1864564  0.0400542   4.655 3.24e-06 ***
## Sulphates           0.1203503  0.0291009   4.136 3.54e-05 ***
## Alcohol             0.0086686  0.0072495   1.196  0.23180    
## LabelAppeal         0.3170975  0.0331851   9.555  < 2e-16 ***
## AcidIndex           0.4822325  0.0193585  24.911  < 2e-16 ***
## STARS              -0.4876789  0.0340664 -14.316  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Number of iterations in BFGS optimization: 31 
## Log-likelihood: -2.257e+04 on 26 Df
```

Model 5: Linear Model

Let build a multiple linear regression

```
## 
## Call:
## lm(formula = TARGET ~ ., data = train_data_imputed)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.8273 -0.6997  0.3798  1.1196  4.3471 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         5.755e+00  5.610e-01  10.259  < 2e-16 ***
## FixedAcidity       -3.077e-04  2.356e-03  -0.131 0.896082    
## VolatileAcidity    -1.571e-01  1.871e-02  -8.399  < 2e-16 ***
## CitricAcid          3.931e-02  1.703e-02   2.308 0.020999 *  
## ResidualSugar       4.399e-04  4.331e-04   1.016 0.309780    
## Chlorides          -1.893e-01  4.598e-02  -4.118 3.85e-05 ***
## FreeSulfurDioxide   4.574e-04  9.873e-05   4.633 3.64e-06 ***
## TotalSulfurDioxide  3.422e-04  6.322e-05   5.413 6.32e-08 ***
## Density            -1.262e+00  5.524e-01  -2.285 0.022323 *  
## pH                 -5.770e-02  2.161e-02  -2.671 0.007575 **
## Sulphates          -5.632e-02  1.571e-02  -3.585 0.000339 ***
## Alcohol             1.951e-02  3.948e-03   4.941 7.86e-07 ***
## LabelAppeal         6.035e-01  1.740e-02  34.687  < 2e-16 ***
## AcidIndex          -3.285e-01  1.138e-02 -28.858  < 2e-16 ***
## STARS               5.410e-01  1.736e-02  31.171  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.655 on 12780 degrees of freedom
## Multiple R-squared:  0.2623, Adjusted R-squared:  0.2615
## F-statistic: 324.5 on 14 and 12780 DF,  p-value: < 2.2e-16
```

Model 6: Negative Binomial

```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = train_data_imputed, link = "log",
##     init.theta = 33879.1139)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8385  -0.5031   0.2162   0.6281   2.6383
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         2.131e+00  1.956e-01  10.898  < 2e-16 ***
## FixedAcidity       -2.187e-04  8.197e-04  -0.267 0.789597
## VolatileAcidity    -5.200e-02  6.481e-03  -8.023 1.03e-15 ***
## CitricAcid          1.371e-02  5.893e-03   2.326 0.020017 *
## ResidualSugar       1.253e-04  1.504e-04   0.833 0.404699
## Chlorides          -5.865e-02  1.601e-02  -3.663 0.000250 ***
## FreeSulfurDioxide   1.510e-04  3.426e-05   4.407 1.05e-05 ***
## TotalSulfurDioxide  1.178e-04  2.206e-05   5.337 9.47e-08 ***
## Density            -4.318e-01  1.921e-01  -2.247 0.024628 *
## pH                 -2.156e-02  7.524e-03  -2.865 0.004170 **
## Sulphates          -1.920e-02  5.456e-03  -3.519 0.000433 ***
## Alcohol             5.684e-03  1.376e-03   4.130 3.62e-05 ***
## LabelAppeal         2.001e-01  6.087e-03  32.866  < 2e-16 ***
## AcidIndex          -1.240e-01  4.453e-03 -27.850  < 2e-16 ***
## STARS               1.665e-01  5.823e-03  28.596  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(33879.11) family taken to be 1)
##
##     Null deviance: 22859  on 12794  degrees of freedom
## Residual deviance: 18735  on 12780  degrees of freedom
## AIC: 50710
##
## Number of Fisher Scoring iterations: 1
##
##
```

```
##              Theta:  33879
##          Std. Err.:  59308
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -50677.75
```

Model Selection

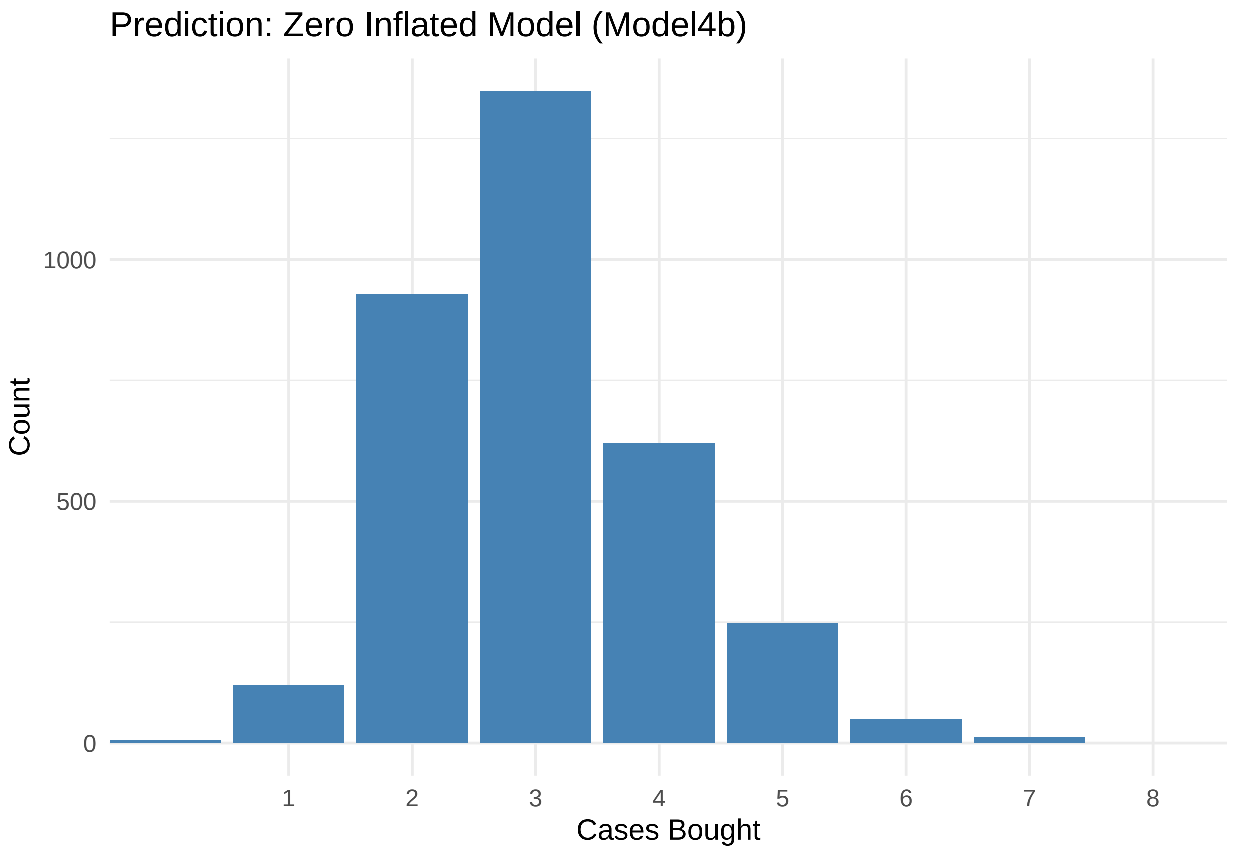
Based on the models tested, we select the Zero Inflated model due to the highest accuracy compared to other models. Also, Zero Inflation model corrects many zeros which are dominated in poisson distributions especially in this case where there are many zeros normally distributed.

```
##                Poisson (Imputed) Quasipoisson Model Zero Inflated
## Accuracy                    0.24               0.24          0.26
## Kappa                       0.10               0.10          0.12
## AccuracyLower               0.24               0.24          0.26
## AccuracyUpper               0.25               0.25          0.27
## AccuracyNull                0.25               0.25          0.25
## AccuracyPValue              0.83               0.83          0.00
## McnemarPValue                NaN                NaN           NaN
```

```
##
##  iter imp variable
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   2   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   3   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   4   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   5   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
```

```
##
##  iter imp variable
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  STA
##   1   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   1   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   1   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   1   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   2   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   2   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   2   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   2   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   3   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   3   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   3   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   3   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   4   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   4   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   4   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   4   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   5   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   5   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   5   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
##   5   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol  ST
```

```
## [1] "Missing value after imputation: 3335"
```



Appendix

```
library(corrplot)
library(tidyverse)
library(Hmisc)
library(PerformanceAnalytics)
library(mice)
library(gt)
library(caret)
library(bnstruct)
library(VIM)
library(corrr)
library(kableExtra)
library(rpart)
library(gtsummary)
library(reshape)
library(pROC)
library(randomForest)
library(pscl)
library(skimr)
## Data Exploration
train_data <- read.csv("./data/wine-training-data.csv", header = TRUE)
test_data <- read.csv("./data/wine-evaluation-data.csv", header = TRUE)
glimpse(train_data)
skim(train_data)
# remove index column as it is not needed
train_data <- train_data %>%
  dplyr::select(-"ï..INDEX")
data_test <- test_data %>%
  dplyr::select(-"IN")
## Visualization
# histogram
train_data %>%
  dplyr::select(-c("AcidIndex", "STARS", "TARGET", "LabelAppeal")) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~key, scale = "free",  ncol = 3) +
  geom_histogram(binwidth = function(x) 2 * IQR(x) / (length(x)^(1/3)), fill="blue") +
  theme_minimal()
train_data %>%
  ggplot(aes(x=TARGET)) +
  geom_histogram(fill='blue')
# boxplot
p1 <- train_data %>%
  dplyr::select(-c("TotalSulfurDioxide", "FreeSulfurDioxide", "ResidualSugar")) %>%
  gather(na.rm = TRUE) %>%
  ggplot(aes(factor(key), value)) +
  geom_boxplot(outlier.colour = "#e281cf", outlier.shape = 1,  color = "#5aa1ed") +
  coord_flip() +
  labs(title = "Boxplot of Chemical Properties of Wine", x = "Chemical Properties", y = "Values") +
  theme_minimal()
p2 <- train_data %>%
  dplyr::select(c("TotalSulfurDioxide", "FreeSulfurDioxide", "ResidualSugar")) %>%
  gather(na.rm = TRUE) %>%
  ggplot(aes(factor(key), value)) +
  geom_boxplot(outlier.colour = "#e281cf", outlier.shape = 1, color = "#5aa1ed") +
  #labs(title = "Boxplot of Chemical Properties of Wine", x = "Chemical Properties", y = "Values") +
  theme_minimal()
ggarrange(p1, p2)
# barchart
p3 <- train_data %>%
  dplyr::select(TARGET, STARS) %>%
  mutate(STARS = as.factor(STARS),
         TARGET = as.factor(TARGET)) %>%
  ggplot(aes(STARS)) +
  geom_bar(aes(fill = TARGET)) +
  theme_minimal()
p4 <- train_data %>%
  dplyr::select(TARGET, LabelAppeal) %>%
  mutate(STARS = as.factor(LabelAppeal),
         TARGET = as.factor(TARGET)) %>%
  ggplot(aes(LabelAppeal)) +
  geom_bar(aes(fill = TARGET)) +
  theme_minimal()
p5 <- train_data %>%
  dplyr::select(TARGET, AcidIndex) %>%
  mutate(STARS = as.factor(AcidIndex),
         TARGET = as.factor(TARGET)) %>%
  ggplot(aes(AcidIndex)) +
  geom_bar(aes(fill = TARGET)) +
  theme_minimal()
ggarrange(p5, ggarrange(p3, p4, ncol = 2, nrow = 1, legend = "none"), nrow = 2, common.legend = TRUE)
# top correlation
wine_train_corr <- train_data %>%
  drop_na() %>%
  cor()
kable(sort(wine_train_corr[,1], decreasing = T), col.names = c("Correlation")) %>%
  kable_styling(full_width = F)
# correlation plot
corrplot(wine_train_corr,
         method = "number",
         type = "lower",
         col = brewer.pal(n = 15, name = "Reds"),
         number.cex = .7, tl.cex = .7,
         tl.col = "black", tl.srt = 45)
# missing value columns
aggr(train_data,
     sortVars=TRUE,
     labels=names(train_data),
     cex.axis=.5,
     bars = FALSE,
     col = c("white", "#E46726"),
     combined = TRUE,
     #border = NA,
     ylab = "Missing Values")
# imputating train data
init <- mice(train_data)
meth <- init$method
```

```
predM <- init$predictorMatrix
predM[, c("TARGET")] <- 0 #this code will remove the variable as a predictor but still will be imputed
train_data_impute <- mice(train_data, method = 'rf', predictorMatrix=predM)
train_data_imputed <-mice:: complete(train_data_impute)
print(paste0("Missing value after imputation: ", sum(is.na(train_data_imputed))))
## Model Building
### Model 1: Poisson (Raw data)
# poisson model with the missing values
model1 <- glm(TARGET ~ ., family = poisson, train_data)
summary(model1)
print('Goodness of Fit Test:')
with(model1, cbind(res.deviance = deviance, df = df.residual,  p = pchisq(deviance, df.residual, lower.t
### Model 2: Poisson (Imputed Data)
# poisson model with the imputed values
model2 <- glm(TARGET ~ ., family = poisson, train_data_imputed)
summary(model2)
print('Goodness of Fit Test:')
with(model2, cbind(res.deviance = deviance, df = df.residual, p = pchisq(deviance, df.residual, lower.ta
### Model 3: Quasipoisson Model
# poisson model with the imputed values
model3 <- glm(TARGET ~ ., family = quasipoisson(link='log'), train_data_imputed)
summary(model3)
print('Goodness of Fit Test:')
with(model3, cbind(res.deviance = deviance, df = df.residual, p = pchisq(deviance, df.residual, lower.ta
### Model 4: Zero Inflated
model4 <- zeroinfl(TARGET ~., train_data_imputed, dist = 'poisson')
summary(model4)
model4b <- zeroinfl(TARGET ~ . - FixedAcidity - Density, train_data_imputed, dist = 'poisson')
summary(model4b)
## Model Selection
pred_train <- data.frame(TARGET=train_data_imputed$TARGET,
                         model2=model2$fitted.values,
                         model3=model3$fitted.values,
                         model4b=model4b$fitted.values)
pred_train <- round(pred_train, 0)
colnames(pred_train) <- c("TARGET","Poisson (Imputed)" ,"Quasipoisson Model", "Zero Inflated")
pred_train %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~key, scale = "free",  ncol = 4) +
  geom_bar(fill="blue") +
  theme_minimal() + labs(x="Cases Bought", y = "Count", title = "Prediction Histogram")
model2_fitted.values <- factor(round(model2$fitted.values),levels=rev(0:9))
model3_fitted.values <- factor(round(model3$fitted.values),levels=rev(0:9))
model4b_fitted.values <- factor(round(model4b$fitted.values),levels=rev(0:9))
m2_cfm <- confusionMatrix(model2_fitted.values, factor(train_data_imputed$TARGET,levels=rev(0:9)))
m3_cfm <- confusionMatrix(model3_fitted.values, factor(train_data_imputed$TARGET,levels=rev(0:9)))
m4_cfm <- confusionMatrix(model4b_fitted.values, factor(train_data_imputed$TARGET,levels=rev(0:9)))
models_sum <- data.frame(m2_cfm$overall, m3_cfm$overall, m4_cfm$overall)
colnames(models_sum) <- c("Poisson (Imputed)" ,"Quasipoisson Model", "Zero Inflated")
round(models_sum, 2)
init <- mice(data_test)
meth <- init$method
predM <- init$predictorMatrix
```

```
predM[, c("TARGET")] <- 0 #this code will remove the variable as a predictor but still will be imputed
data_test_impute <- mice(data_test, method = 'rf', predictorMatrix=predM)
data_test_imputed <- mice::complete(data_test_impute)
print(paste0("Missing value after imputation: ", sum(is.na(data_test_imputed))))
test_predict <- predict(model4b, newdata=data_test_imputed)
test_predict <- round(test_predict,0)
data_pred <- data.frame(TARGET=test_predict)
ggplot(data_pred, aes(x=TARGET)) + geom_bar(fill="steelblue") + theme_minimal() +
  labs(y="Count", title = "Prediction: Zero Inflated Model (Model4b)") +
    scale_x_discrete(name = "Cases Bought", limits=c("1","2","3","4", "5", "6", "7", "8"))
write.csv(test_predict, "WinePredictions.csv")
```