DATA 621: BUSINESS ANALYTICS AND DATA MINING HOMEWORK#5 Assignment Requirements

Group 2 - Gabriel Campos, Melissa Bowman, Alexander Khaykin, & Jennifer Abinette

Last edited December 05, 2023

Contents

1 Overview

In this homework assignment, you will explore, analyze and model a data set containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales.

Your objective is to build a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine. HINT: Sometimes, the fact that a variable is missing is actually predictive of the target. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

```
## Warning: package 'pscl' was built under R version 4.3.2
```

```
## Warning: package 'Metrics' was built under R version 4.3.2
```

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	

1.1 Deliverables

- A write-up submitted in PDF format. Your write-up should have four sections. Each one is described below. You may assume you are addressing me as a fellow data scientist, so do not need to shy away from technical details.
- Assigned predictions (number of cases of wine sold) for the evaluation data set.
- Include your R statistical programming code in an Appendix.

1.2 Write Up:

1.2.1 1. DATA EXPLORATION (25 Points)

Describe the size and the variables in the wine training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas.

a. Mean / Standard Deviation / Median

b. Bar Chart or Box Plot of the data
c. Is the data correlated to the target variable (or to other variables?)
d. Are any of the variables missing and need to be imputed “fixed”?

1.2.2 2. DATA PREPARATION (25 Points)

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations.

a. Fix missing values (maybe with a Mean or Median value)
b. Create flags to suggest if a variable was missing
c. Transform data by putting it into buckets
d. Mathematical transforms such as log or square root (or use Box-Cox)
e. Combine variables (such as ratios or adding or multiplying) to create new variables

1.2.3 3. BUILD MODELS (25 Points)

Using the training data set, build at least two different poisson regression models, at least two different negative binomial regression models, and at least two multiple linear regression models, using different variables (or the same variables with different transformations). Sometimes poisson and negative binomial regression models give the same results. If that is the case, comment on that. Consider changing the input variables if that occurs so that you get different models. Although not covered in class, you may also want to consider building zero-inflated poisson and negative binomial regression models. You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done

Discuss the coefficients in the models, do they make sense? In this case, about the only thing you can comment on is the number of stars and the wine label appeal. However, you might comment on the coefficient and magnitude of variables and how they are similar or different from model to model. For example, you might say “pH seems to have a major positive impact in my poisson regression model, but a negative effect in my multiple linear regression model”. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

1.2.4 4. SELECT MODELS (25 Points)

Decide on the criteria for selecting the best count regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your models.

For the count regression model, will you use a metric such as AIC, average squared error, etc.? Be sure to explain how you can make inferences from the model, and discuss other relevant model output. If you like the multiple linear regression model the best, please say why. However, you must select a count regression model for model deployment. Using the training data set, evaluate the performance of the count regression model. Make predictions using the evaluation data set.

2 Import Data

```r
df_wine_eval <-
  read.csv(paste0(url_git,"wine-evaluation-data.csv"))

head(df_wine_eval)
```

```
##   IN TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
## 1  3     NA          5.4          -0.860       0.27         -10.7     0.092
## 2  9     NA         12.4           0.385      -0.76         -19.7     1.169
## 3 10     NA          7.2           1.750       0.17         -33.0     0.065
## 4 18     NA          6.2           0.100       1.80           1.0    -0.179
## 5 21     NA         11.4           0.210       0.28           1.2     0.038
## 6 30     NA         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    NA
## 2           0         6     2
## 3           0         8     1
## 4          -1         8     1
## 5           0        10    NA
## 6           1         8     4
```

```r
df_wine_train <-
  read.csv(paste0(url_git,"wine-training-data.csv"))
head(df_wine_train)
```

```
##   INDEX TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
## 1     1      3          3.2           1.160      -0.98          54.2    -0.567
## 2     2      3          4.5           0.160      -0.81          26.1    -0.425
## 3     4      5          7.1           2.640      -0.88          14.8     0.037
## 4     5      3          5.7           0.385       0.04          18.8    -0.425
## 5     6      4          8.0           0.330      -1.26           9.4        NA
## 6     7      0         11.3           0.320       0.59           2.2     0.556
##   FreeSulfurDioxide TotalSulfurDioxide Density   pH Sulphates Alcohol
## 1                NA                268 0.99280 3.33     -0.59     9.9
## 2                15               -327 1.02792 3.38      0.70      NA
## 3               214                142 0.99518 3.12      0.48    22.0
## 4                22                115 0.99640 2.24      1.83     6.2
## 5              -167                108 0.99457 3.12      1.77    13.7
## 6               -37                 15 0.99940 3.20      1.29    15.4
##   LabelAppeal AcidIndex STARS
## 1           0         8     2
## 2          -1         7     3
## 3          -1         8     3
## 4          -1         6     1
## 5           0         9     2
## 6           0        11    NA
```

Of training variable:

```r
print(skim(df_wine_train))
```

```
## -- Data Summary ------------------------
##                            Values
## Name                       df_wine_train
## Number of rows             12795
## Number of columns          16
## _______________________
## Column type frequency:
##   numeric                  16
## ________________________
## Group variables            None
##
## -- Variable type: numeric ------------------------------------------------------
##    skim_variable      n_missing complete_rate       mean        sd       p0
##  1 INDEX                      0         1     8070.      4657.         1
##  2 TARGET                     0         1        3.03       1.93      0
##  3 FixedAcidity               0         1        7.08       6.32    -18.1
##  4 VolatileAcidity            0         1        0.324      0.784    -2.79
##  5 CitricAcid                 0         1        0.308      0.862    -3.24
##  6 ResidualSugar            616         0.952    5.42      33.7    -128.
##  7 Chlorides                638         0.950    0.0548     0.318    -1.17
##  8 FreeSulfurDioxide        647         0.949   30.8      149.     -555
##  9 TotalSulfurDioxide       682         0.947  121.       232.     -823
## 10 Density                    0         1        0.994      0.0265    0.888
## 11 pH                       395         0.969    3.21       0.680     0.48
## 12 Sulphates               1210         0.905    0.527      0.932    -3.13
## 13 Alcohol                  653         0.949   10.5        3.73     -4.7
## 14 LabelAppeal                0         1       -0.00907    0.891    -2
## 15 AcidIndex                  0         1        7.77       1.32      4
## 16 STARS                   3359         0.737    2.04       0.903     1
##          p25      p50       p75     p100 hist
##  1 4038.    8110     12106.    16129
##  2    2        3         4        8
##  3    5.2      6.9       9.5     34.4
##  4    0.13     0.28      0.64     3.68
##  5    0.03     0.31      0.58     3.86
##  6   -2        3.9      15.9    141.
##  7   -0.031    0.046     0.153    1.35
##  8    0       30        70      623
##  9   27      123       208     1057
## 10    0.988    0.994     1.00     1.10
## 11    2.96     3.2       3.47     6.13
## 12    0.28     0.5       0.86     4.24
## 13    9       10.4      12.4     26.5
## 14   -1        0         1        2
## 15    7        8         8       17
## 16    1        2         3        4
```

```r
summary(df_wine_train)
```

```
##       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
```

Of evaluated variable:

```
print(skim(df_wine_eval))
```

```
## -- Data Summary ------------------------
##                            Values
## Name                       df_wine_eval
## Number of rows             3335
## Number of columns          16
## _______________________
## Column type frequency:
##   logical                  1
##   numeric                  15
## ________________________
## Group variables            None
##
## -- Variable type: logical ------------------------------------------------------
##   skim_variable n_missing complete_rate mean count
## 1 TARGET             3335             0  NaN ": "
##
## -- Variable type: numeric ------------------------------------------------------
```

```
##    skim_variable      n_missing complete_rate      mean        sd       p0
##  1 IN                         0         1     8048.     4655.        3
##  2 FixedAcidity               0         1        6.86      6.32    -18.2
##  3 VolatileAcidity            0         1        0.310     0.807    -2.83
##  4 CitricAcid                 0         1        0.312     0.871    -3.12
##  5 ResidualSugar            168         0.950    5.32     34.4    -128.
##  6 Chlorides                138         0.959    0.0614    0.314    -1.15
##  7 FreeSulfurDioxide        152         0.954   34.9     150.     -563
##  8 TotalSulfurDioxide       157         0.953  123.      226.     -769
##  9 Density                    0         1        0.995     0.0262    0.890
## 10 pH                       104         0.969    3.24      0.676     0.6
## 11 Sulphates                310         0.907    0.535     0.905    -3.07
## 12 Alcohol                  185         0.945   10.6       3.76     -4.2
## 13 LabelAppeal                0         1        0.0135    0.889    -2
## 14 AcidIndex                  0         1        7.75      1.32      5
## 15 STARS                    841         0.748    2.04      0.913     1
##         p25      p50       p75     p100 hist
##  1 4018.    7906     12061     16130
##  2    5.2      6.9       9        33.5
##  3    0.08     0.28      0.63      3.61
##  4    0        0.31      0.605     3.76
##  5   -2.6      3.6      17.2     145.
##  6    0.016    0.047     0.171     1.26
##  7    3       30        79.2     617
##  8   27.2    124       210      1004
##  9    0.988    0.995     1.00      1.10
## 10    2.98     3.21      3.49      6.21
## 11    0.33     0.5       0.82      4.18
## 12    9       10.4      12.5      25.6
## 13   -1        0         1         2
## 14    7        8         8        17
## 15    1        2         3         4
```

```r
summary(df_wine_eval)
```

```
##        IN          TARGET         FixedAcidity     VolatileAcidity
##  Min.   :    3   Mode:logical   Min.   :-18.200   Min.   :-2.8300
##  1st Qu.: 4018   NA's:3335      1st Qu.:  5.200   1st Qu.: 0.0800
##  Median : 7906                  Median :  6.900   Median : 0.2800
##  Mean   : 8048                  Mean   :  6.864   Mean   : 0.3103
##  3rd Qu.:12061                  3rd Qu.:  9.000   3rd Qu.: 0.6300
##  Max.   :16130                  Max.   : 33.500   Max.   : 3.6100
##
##    CitricAcid      ResidualSugar        Chlorides        FreeSulfurDioxide
##  Min.   :-3.1200   Min.   :-128.300   Min.   :-1.15000   Min.   :-563.00
##  1st Qu.: 0.0000   1st Qu.:  -2.600   1st Qu.: 0.01600   1st Qu.:   3.00
##  Median : 0.3100   Median :   3.600   Median : 0.04700   Median :  30.00
##  Mean   : 0.3124   Mean   :   5.319   Mean   : 0.06143   Mean   :  34.95
##  3rd Qu.: 0.6050   3rd Qu.:  17.200   3rd Qu.: 0.17100   3rd Qu.:  79.25
##  Max.   : 3.7600   Max.   : 145.400   Max.   : 1.26300   Max.   : 617.00
##                    NA's   :168        NA's   :138        NA's   :152
##  TotalSulfurDioxide    Density              pH           Sulphates
##  Min.   :-769.00    Min.   :0.8898   Min.   :0.600   Min.   :-3.0700
##  1st Qu.:  27.25    1st Qu.:0.9883   1st Qu.:2.980   1st Qu.: 0.3300
```

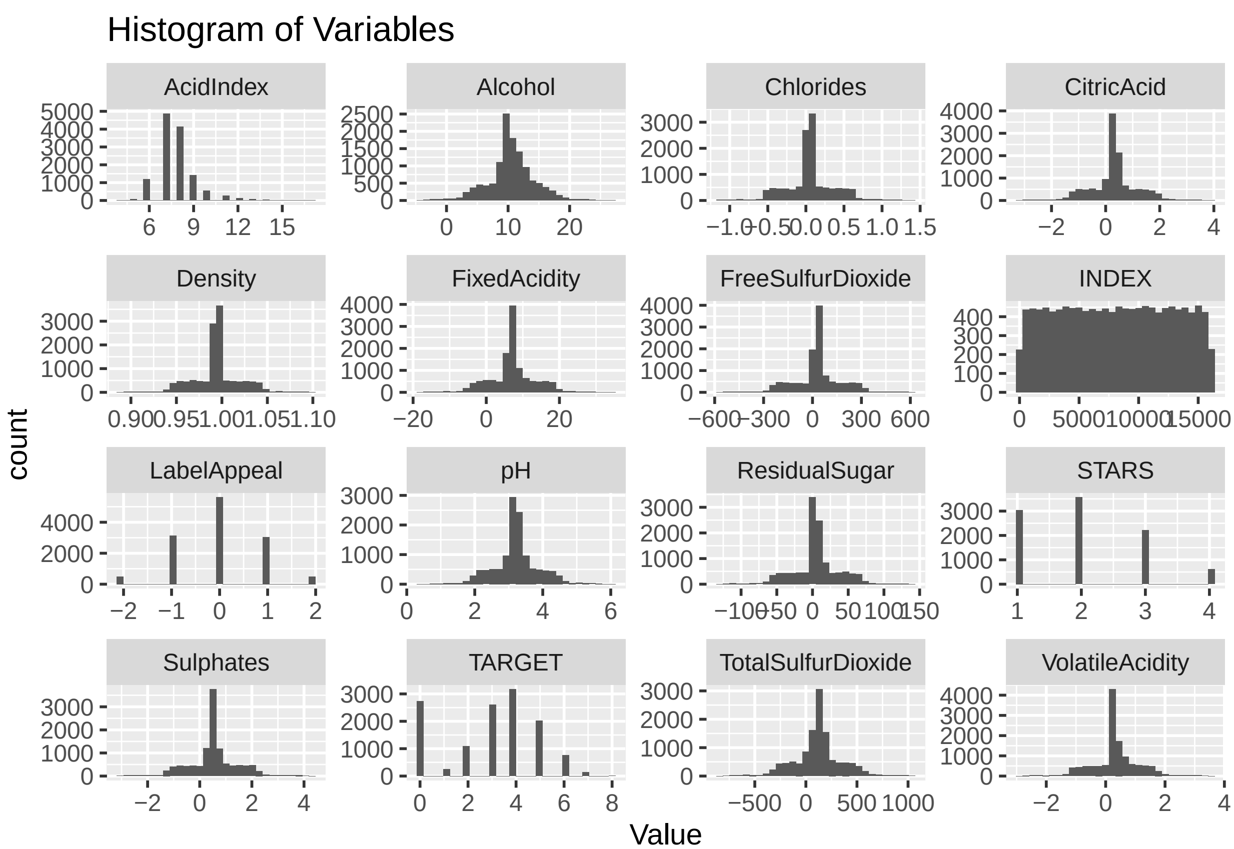
```
##  Median : 124.00    Median :0.9946   Median :3.210   Median : 0.5000
##  Mean   : 123.41    Mean   :0.9947   Mean   :3.237   Mean   : 0.5346
##  3rd Qu.: 210.00    3rd Qu.:1.0005   3rd Qu.:3.490   3rd Qu.: 0.8200
##  Max.   :1004.00    Max.   :1.0998   Max.   :6.210   Max.   : 4.1800
##  NA's   :157                         NA's   :104     NA's   :310
##     Alcohol        LabelAppeal          AcidIndex          STARS
##  Min.   :-4.20   Min.   :-2.00000   Min.   : 5.000   Min.   :1.00
##  1st Qu.: 9.00   1st Qu.:-1.00000   1st Qu.: 7.000   1st Qu.:1.00
##  Median :10.40   Median : 0.00000   Median : 8.000   Median :2.00
##  Mean   :10.58   Mean   : 0.01349   Mean   : 7.748   Mean   :2.04
##  3rd Qu.:12.50   3rd Qu.: 1.00000   3rd Qu.: 8.000   3rd Qu.:3.00
##  Max.   :25.60   Max.   : 2.00000   Max.   :17.000   Max.   :4.00
##  NA's   :185                                         NA's   :841
```

Looking at histogram

```r
# Gather the data into a long format
data_long <- gather(df_wine_train, key = "Variable", value = "Value")

ggplot(data_long, aes(x = Value)) +
  geom_histogram() +
  facet_wrap(~Variable, scales = "free") +
  labs(title = "Histogram of Variables")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Relatively normal data. We do not have to correct any variables

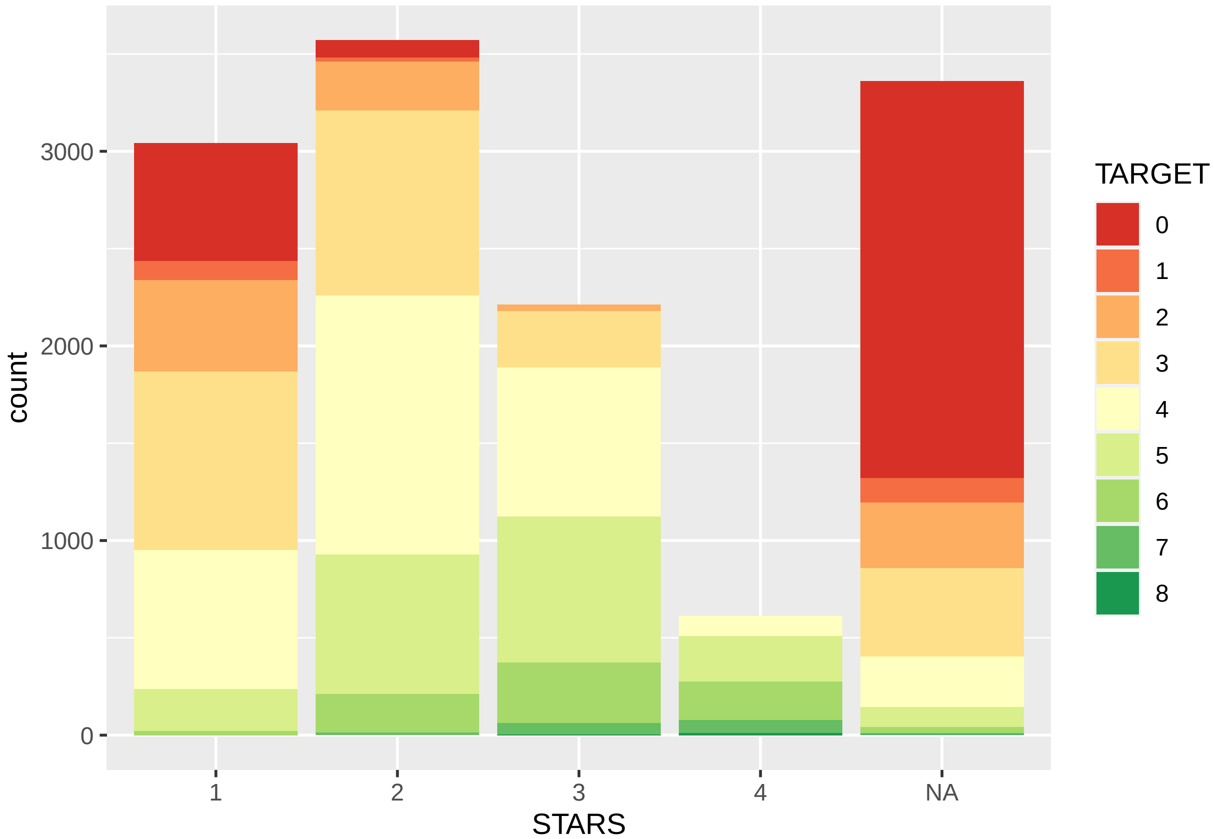
```
# Create a correlation matrix for all variables
(cor_matrix <- cor(df_wine_train, use='complete.obs'))
```

```
##                           INDEX        TARGET FixedAcidity VolatileAcidity
## INDEX               1.0000000000  0.0236764338 -0.002831415   -0.0008743296
## TARGET              0.0236764338  1.0000000000 -0.012538100   -0.0759978765
## FixedAcidity       -0.0028314152 -0.0125380998  1.000000000    0.0190109733
## VolatileAcidity    -0.0008743296 -0.0759978765  0.019010973    1.0000000000
## CitricAcid          0.0278869710  0.0023450490  0.014000376   -0.0234315631
## ResidualSugar       0.0208952098  0.0035195999 -0.015429391    0.0015279517
## Chlorides           0.0026827829 -0.0304301331 -0.006104447    0.0148489225
## FreeSulfurDioxide   0.0046416504  0.0226398054  0.015438463   -0.0114408079
## TotalSulfurDioxide  0.0064949038  0.0216020726 -0.023323485   -0.0007434083
## Density            -0.0034840089 -0.0475989086  0.011574241    0.0130977690
## pH                 -0.0274556333  0.0002198557 -0.004553886    0.0072030364
## Sulphates          -0.0053946247 -0.0212203783  0.042229181    0.0015161001
## Alcohol            -0.0024453460  0.0737771084 -0.013085026    0.0002603082
## LabelAppeal         0.0314911460  0.4979464796  0.011375965   -0.0202419713
## AcidIndex           0.0055244862 -0.1676430648  0.154167846    0.0250529742
## STARS              -0.0057807296  0.5546857223 -0.004937345   -0.0402432388
##                       CitricAcid ResidualSugar     Chlorides FreeSulfurDioxide
## INDEX               0.0278869710   0.020895210  0.0026827829       0.004641650
## TARGET              0.0023450490   0.003519600 -0.0304301331       0.022639805
## FixedAcidity        0.0140003760  -0.015429391 -0.0061044471       0.015438463
## VolatileAcidity    -0.0234315631   0.001527952  0.0148489225      -0.011440808
## CitricAcid          1.0000000000  -0.009843146 -0.0335608661       0.012113248
## ResidualSugar      -0.0098431456   1.000000000  0.0041215692       0.021959113
## Chlorides          -0.0335608661   0.004121569  1.0000000000      -0.020492488
## FreeSulfurDioxide   0.0121132485   0.021959113 -0.0204924876       1.000000000
## TotalSulfurDioxide -0.0099174506   0.017030939  0.0004188605       0.013461673
## Density            -0.0169919691  -0.007120841  0.0206724860      -0.008663509
## pH                 -0.0007581304   0.017563769 -0.0179702278      -0.002008516
## Sulphates          -0.0144237270  -0.002705775  0.0026187777       0.026829029
## Alcohol             0.0169864284  -0.018943324 -0.0228849573      -0.023867458
## LabelAppeal         0.0153315666  -0.004579308 -0.0063870237       0.014960087
## AcidIndex           0.0545838104  -0.020301890 -0.0017134096      -0.014733717
## STARS               0.0071401699   0.019665541 -0.0063242568      -0.015390398
##                    TotalSulfurDioxide      Density            pH    Sulphates
## INDEX                    0.0064949038 -0.003484009 -0.0274556333 -0.005394625
## TARGET                   0.0216020726 -0.047598909  0.0002198557 -0.021220378
## FixedAcidity            -0.0233234848  0.011574241 -0.0045538857  0.042229181
## VolatileAcidity         -0.0007434083  0.013097769  0.0072030364  0.001516100
## CitricAcid              -0.0099174506 -0.016991969 -0.0007581304 -0.014423727
## ResidualSugar            0.0170309394 -0.007120841  0.0175637691 -0.002705775
## Chlorides                0.0004188605  0.020672486 -0.0179702278  0.002618778
## FreeSulfurDioxide        0.0134616726 -0.008663509 -0.0020085157  0.026829029
## TotalSulfurDioxide       1.0000000000  0.023167955 -0.0034227601  0.002504051
## Density                  0.0231679548  1.000000000 -0.0020192285 -0.010609294
## pH                      -0.0034227601 -0.002019229  1.0000000000  0.010449255
## Sulphates                0.0025040509 -0.010609294  0.0104492547  1.000000000
## Alcohol                 -0.0168515467 -0.006128355 -0.0122034469  0.010844330
## LabelAppeal             -0.0027237419 -0.018094403  0.0002181758  0.003768700
```

```
## AcidIndex              -0.0221292631  0.047778830 -0.0537128921  0.031071782
## STARS                   0.0220949002 -0.028492455 -0.0044002985 -0.023135130
##                         Alcohol   LabelAppeal    AcidIndex        STARS
## INDEX              -0.0024453460  0.0314911460  0.005524486 -0.005780730
## TARGET              0.0737771084  0.4979464796 -0.167643065  0.554685722
## FixedAcidity       -0.0130850260  0.0113759650  0.154167846 -0.004937345
## VolatileAcidity     0.0002603082 -0.0202419713  0.025052974 -0.040243239
## CitricAcid          0.0169864284  0.0153315666  0.054583810  0.007140170
## ResidualSugar      -0.0189433242 -0.0045793083 -0.020301890  0.019665541
## Chlorides          -0.0228849573 -0.0063870237 -0.001713410 -0.006324257
## FreeSulfurDioxide  -0.0238674577  0.0149600871 -0.014733717 -0.015390398
## TotalSulfurDioxide -0.0168515467 -0.0027237419 -0.022129263  0.022094900
## Density            -0.0061283546 -0.0180944026  0.047778830 -0.028492455
## pH                 -0.0122034469  0.0002181758 -0.053712892 -0.004400299
## Sulphates           0.0108443299  0.0037686996  0.031071782 -0.023135130
## Alcohol             1.0000000000 -0.0006449123 -0.055891906  0.064854486
## LabelAppeal        -0.0006449123  1.0000000000  0.010300984  0.318897022
## AcidIndex          -0.0558919056  0.0103009840  1.000000000 -0.095482582
## STARS               0.0648544864  0.3188970216 -0.095482582  1.000000000
```

Only 3 real variable that relate to TARGET which are LabelAppeal, AcidIndex, STARS. STARS though has a lot of NA values

```r
df_wine_train %>%
  mutate(STARS = as.factor(STARS),
         TARGET = as.factor(TARGET)) %>%
  ggplot(aes(STARS)) +
  geom_bar(aes(fill = TARGET)) +
  scale_fill_brewer(palette = "RdYlGn")
```



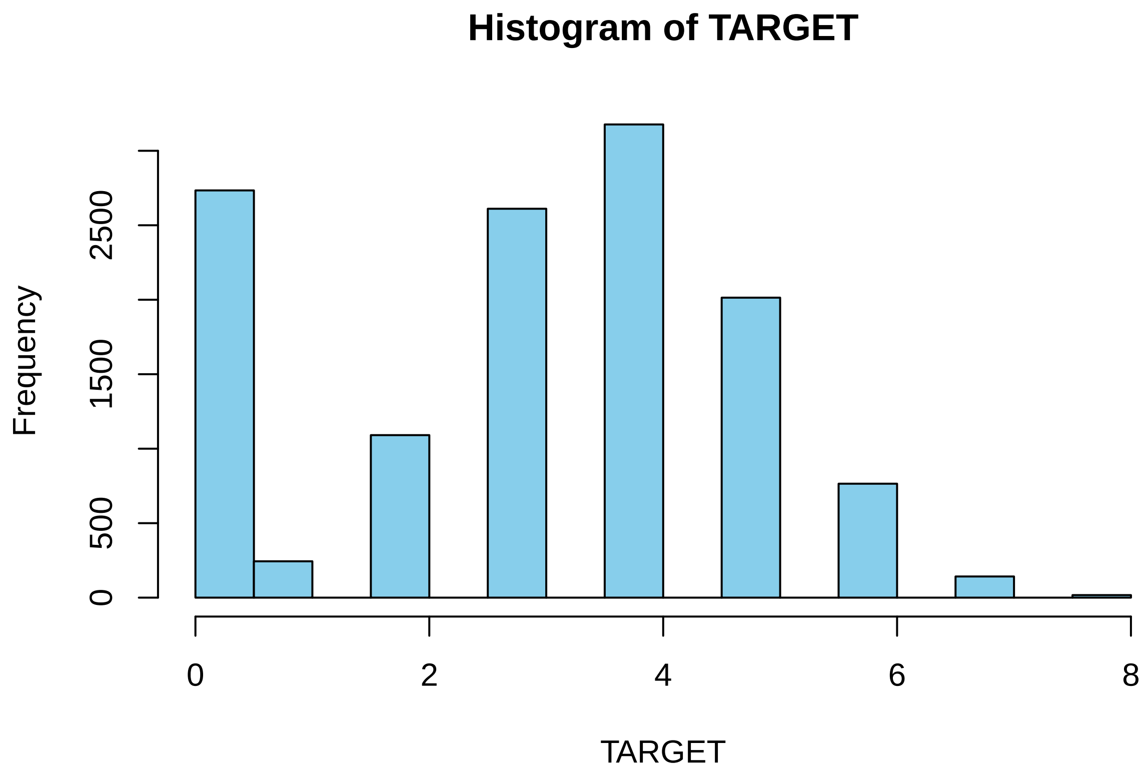
Because STARS has a lot of NA values that relate to a TARGET value of 0 we should make STARS NA zero instead of eliminating NA values.

```
df_wine_train_transformed <- df_wine_train %>%
    mutate(STARS = replace(STARS, is.na(STARS) , 0))

df_wine_eval_transformed  <- df_wine_eval %>%
    mutate(STARS = replace(STARS, is.na(STARS) , 0))
```

(Might not need this histogram)

```
# Plot a histogram
hist(df_wine_train$TARGET, main = "Histogram of TARGET", xlab = "TARGET",
     col = "skyblue", border = "black")
```




```r
# Calculate the percentage of unique values in the TARGET variable
target_table <- table(df_wine_train$TARGET)
target_percentage <- prop.table(target_table) * 100


rounded_percentage <- round(target_percentage, 2)


print(rounded_percentage)
```

```
##
##     0     1     2     3     4     5     6     7     8
## 21.37  1.91  8.53 20.41 24.83 15.74  5.98  1.11  0.13
```

Since there are an excess of zero values in the data set, the Poisson and Negative Binomial Regression may not be able to give the best model outcome. Therefore, we will also test Hurdle Poisson and Zero-Inflated Poisson Regression models to see if these models work best. To compare these models, we will be using the The Root Mean Squared Error (RMSE). The lowest number will tell us which model works best.

Train-test split

```r
set.seed(100)
n <- nrow(df_wine_train_transformed)
train_index <- sample(1:n, 0.8 * n)  # 80% for training, 20% for testing
df_train <- df_wine_train_transformed[train_index, ]
df_test <- df_wine_train_transformed[-train_index, ]
```

#Poisson Regression Model

```
poisson_model <- glm(TARGET ~ LabelAppeal + AcidIndex + STARS, data = df_train,
                     family = poisson)
#summary(poisson_model)
```

Prediction of test-split data (will need to be rounded to full numbers?)

```
poisson_preds <- predict(poisson_model, newdata = df_test, type = "response")
```

RMSE

```
poisson_rmse <- sqrt(mean((poisson_preds - df_test$TARGET)^2))
```

3 Negative Binomial Regression

Model

```
neg_binom_model <- glm.nb(TARGET ~ LabelAppeal + AcidIndex + STARS,
                          data = df_train)
```

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

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

```
#summary(neg_binom_model)
```

Prediction of test-split data (will need to be rounded to full numbers?)

```
neg_binom_preds <- predict(neg_binom_model, newdata = df_test,
                           type = "response")
```

RMSE

```
neg_binom_rmse <- sqrt(mean((neg_binom_preds - df_test$TARGET)^2))
```

4 Hurdle Poisson Regression

Model

```
hurdle_poisson_model <- hurdle(TARGET ~ LabelAppeal + AcidIndex + STARS,
                               data = df_train, dist = "poisson")
#summary(hurdle_poisson_model)
```

Prediction of test-split data (will need to be rounded to full numbers?)

```r
hurdle_preds <- predict(hurdle_poisson_model, newdata = df_test,
                        type = "response")
```

RMSE

```r
hurdle_rmse <- sqrt(mean((hurdle_preds - df_test$TARGET)^2))
```

5 Zero-Inflated Poisson Regression

Model

```r
zip_model <- zeroinfl(TARGET ~ LabelAppeal + AcidIndex + STARS | 1,
                      data = df_train, dist = "poisson")
summary(zip_model)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ LabelAppeal + AcidIndex + STARS | 1, data = df_train,
##     dist = "poisson")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.6328 -0.3246  0.1745  0.4957  2.8957
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.378339   0.045278   30.44   <2e-16 ***
## LabelAppeal  0.193934   0.007571   25.62   <2e-16 ***
## AcidIndex   -0.061714   0.005828  -10.59   <2e-16 ***
## STARS        0.182323   0.007230   25.22   <2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.81723    0.04322  -42.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 11
## Log-likelihood: -1.834e+04 on 5 Df
```

Prediction of test-split data (will need to be rounded to full numbers?)

```r
zip_preds <- predict(zip_model, newdata = df_test, type = "response")
```

RMSE

```r
zip_rmse <- sqrt(mean((zip_preds - df_test$TARGET)^2))
```

6 Compare RMSE

```
comparison <- data.frame(
  Model = c("Poisson", "Negative Binomial", "Hurdle Poisson",
            "Zero-Inflated Poisson"),
  RMSE = c(poisson_rmse, neg_binom_rmse, hurdle_rmse, zip_rmse)
)

print(comparison)
```

```
##                   Model     RMSE
## 1               Poisson 1.437424
## 2     Negative Binomial 1.437428
## 3        Hurdle Poisson 1.318200
## 4 Zero-Inflated Poisson 1.451817
```

7 Predict using the hurdle_poisson_model

```
eval_preds <- predict(hurdle_poisson_model,
                      newdata = df_wine_eval_transformed, type = "response")
```