Critical Thinking Group 4 - HW5 - Wine

Sreejaya, Suman, Vuthy

November 15, 2016

Overview

The objective of this assignment is to predict the number of cases of wine that will be sold based on the properties of the wine. A count regression model will be used to predict wine sales.

Dataset
Wine - Training data
Wine - Evaluation Data

Data Exploration

```
##     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
```

```
## Observations: 12,795
## Variables: 16
```

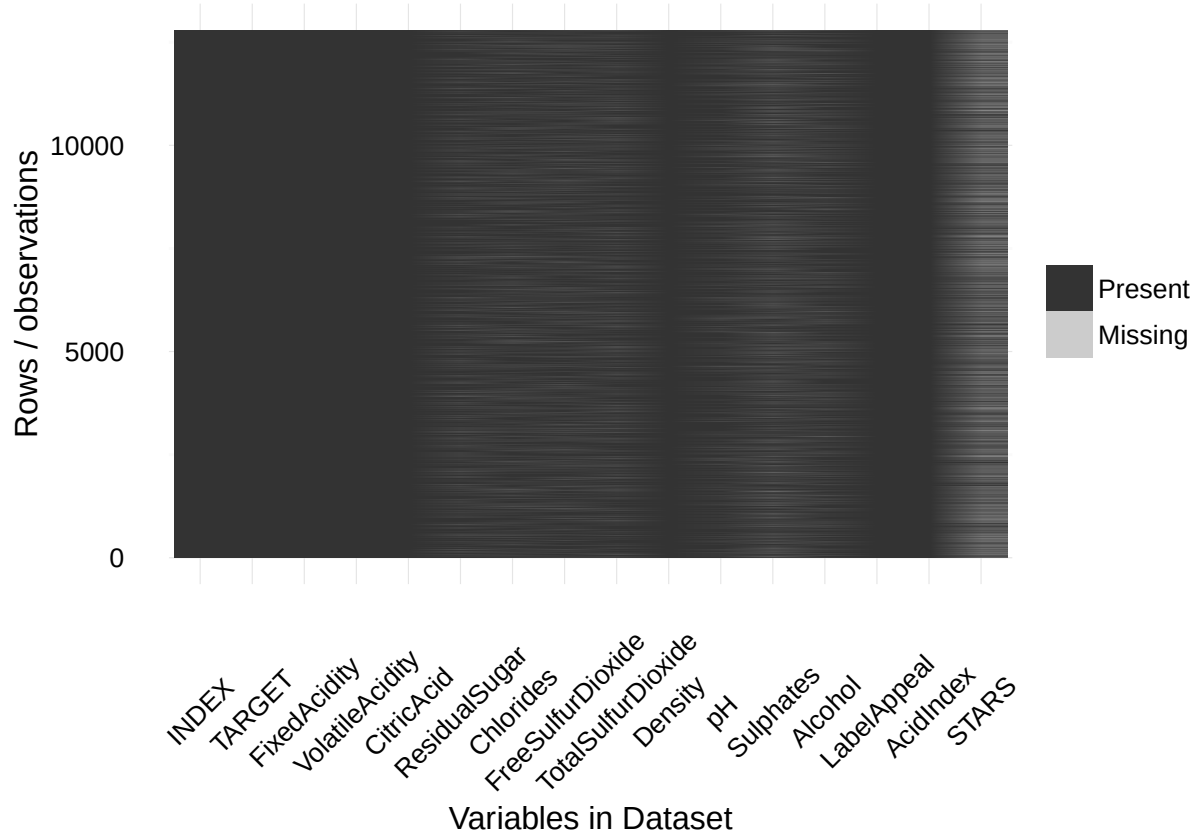
```
## $ ï..INDEX           <int> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16...
## $ TARGET             <int> 3, 3, 5, 3, 4, 0, 0, 4, 3, 6, 0, 4, 3, 7, 4...
## $ FixedAcidity       <dbl> 3.2, 4.5, 7.1, 5.7, 8.0, 11.3, 7.7, 6.5, 14...
## $ VolatileAcidity    <dbl> 1.160, 0.160, 2.640, 0.385, 0.330, 0.320, 0...
## $ CitricAcid         <dbl> -0.98, -0.81, -0.88, 0.04, -1.26, 0.59, -0....
## $ ResidualSugar      <dbl> 54.20, 26.10, 14.80, 18.80, 9.40, 2.20, 21....
## $ Chlorides          <dbl> -0.567, -0.425, 0.037, -0.425, NA, 0.556, 0...
## $ FreeSulfurDioxide  <dbl> NA, 15, 214, 22, -167, -37, 287, 523, -213,...
## $ TotalSulfurDioxide <dbl> 268, -327, 142, 115, 108, 15, 156, 551, NA,...
## $ Density            <dbl> 0.99280, 1.02792, 0.99518, 0.99640, 0.99457...
## $ pH                 <dbl> 3.33, 3.38, 3.12, 2.24, 3.12, 3.20, 3.49, 3...
## $ Sulphates          <dbl> -0.59, 0.70, 0.48, 1.83, 1.77, 1.29, 1.21, ...
## $ Alcohol            <dbl> 9.9, NA, 22.0, 6.2, 13.7, 15.4, 10.3, 11.6,...
## $ LabelAppeal        <int> 0, -1, -1, -1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 2...
## $ AcidIndex          <int> 8, 7, 8, 6, 9, 11, 8, 7, 6, 8, 5, 10, 7, 8,...
## $ STARS              <int> 2, 3, 3, 1, 2, NA, NA, 3, NA, 4, 1, 2, 2, 3...
```

Looks like the INDEX column name need to be corrected.

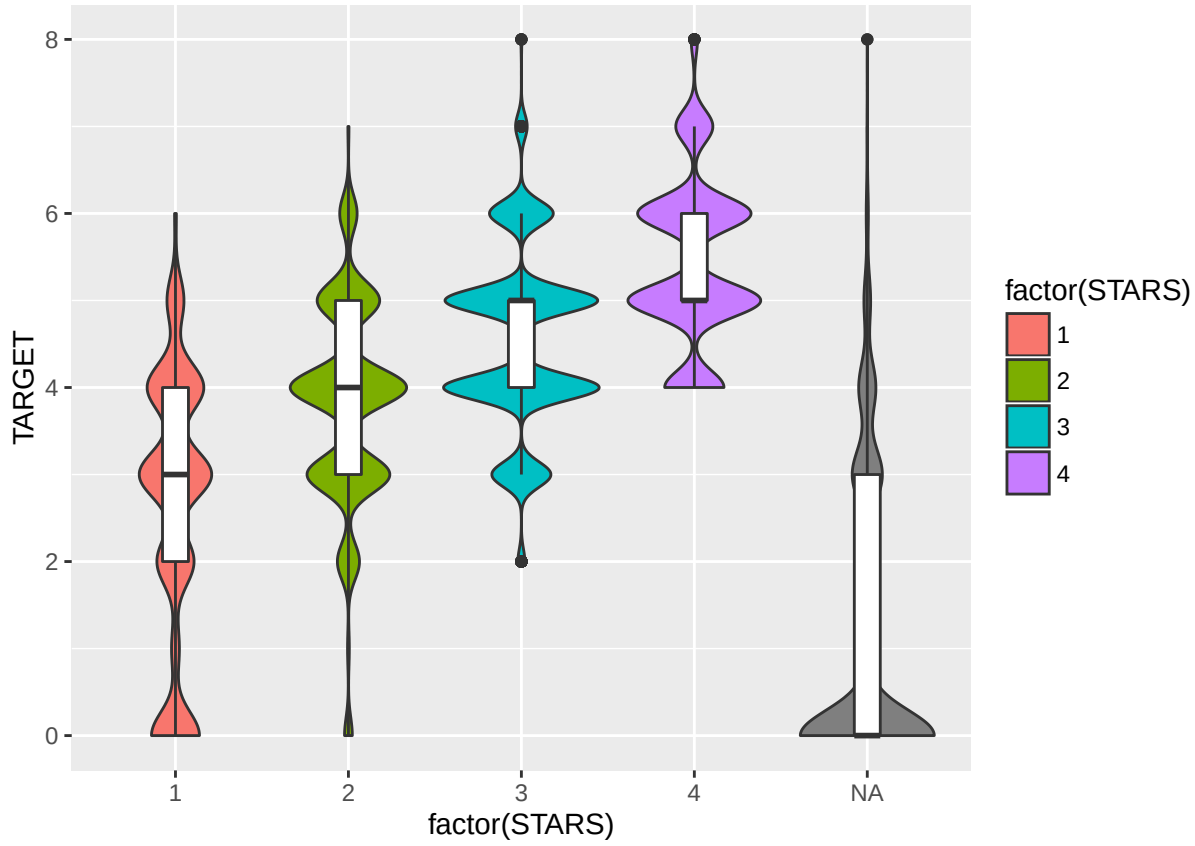
Missing Data

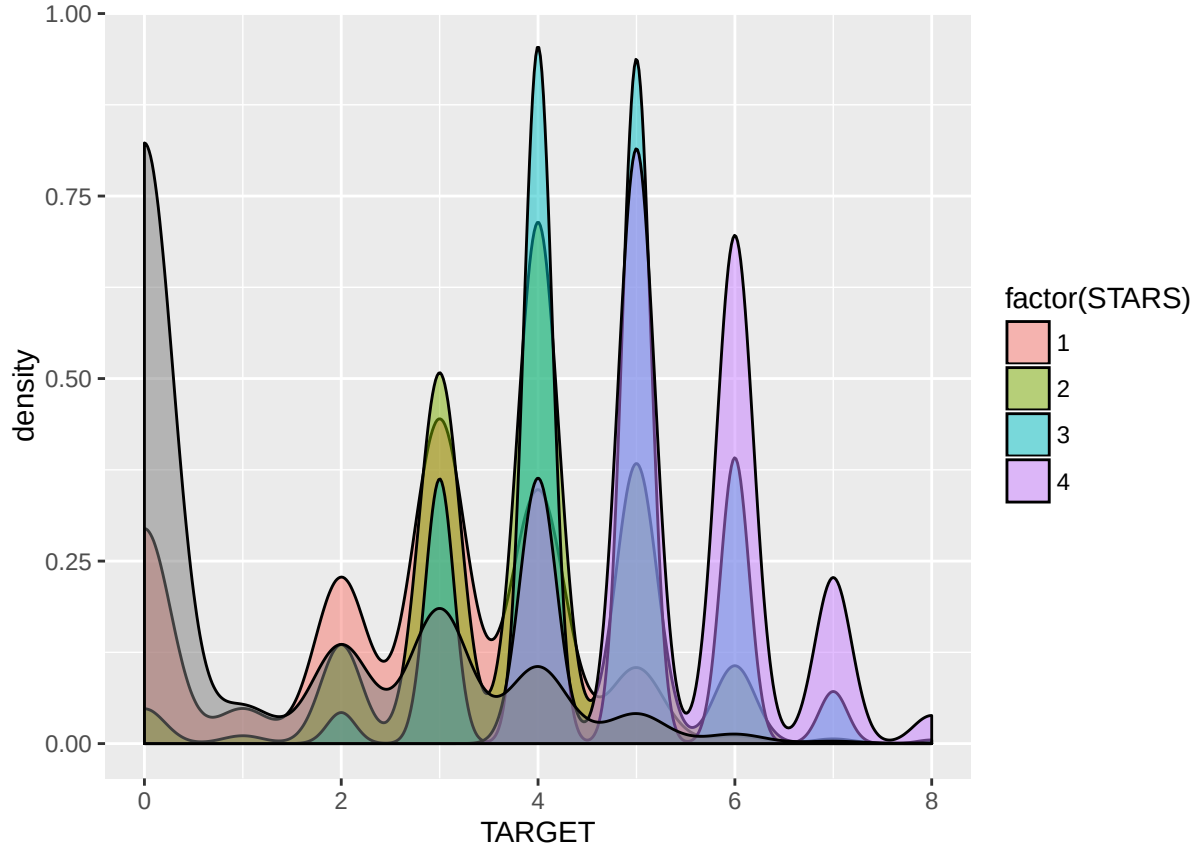
Eight of the variables have missing data.

Lets explore more on the missing values here:



Though there are lot of missing values, we could not see a definite pattern here, but we difinitely notice that there are highest number of missing values for *STARS* variable.

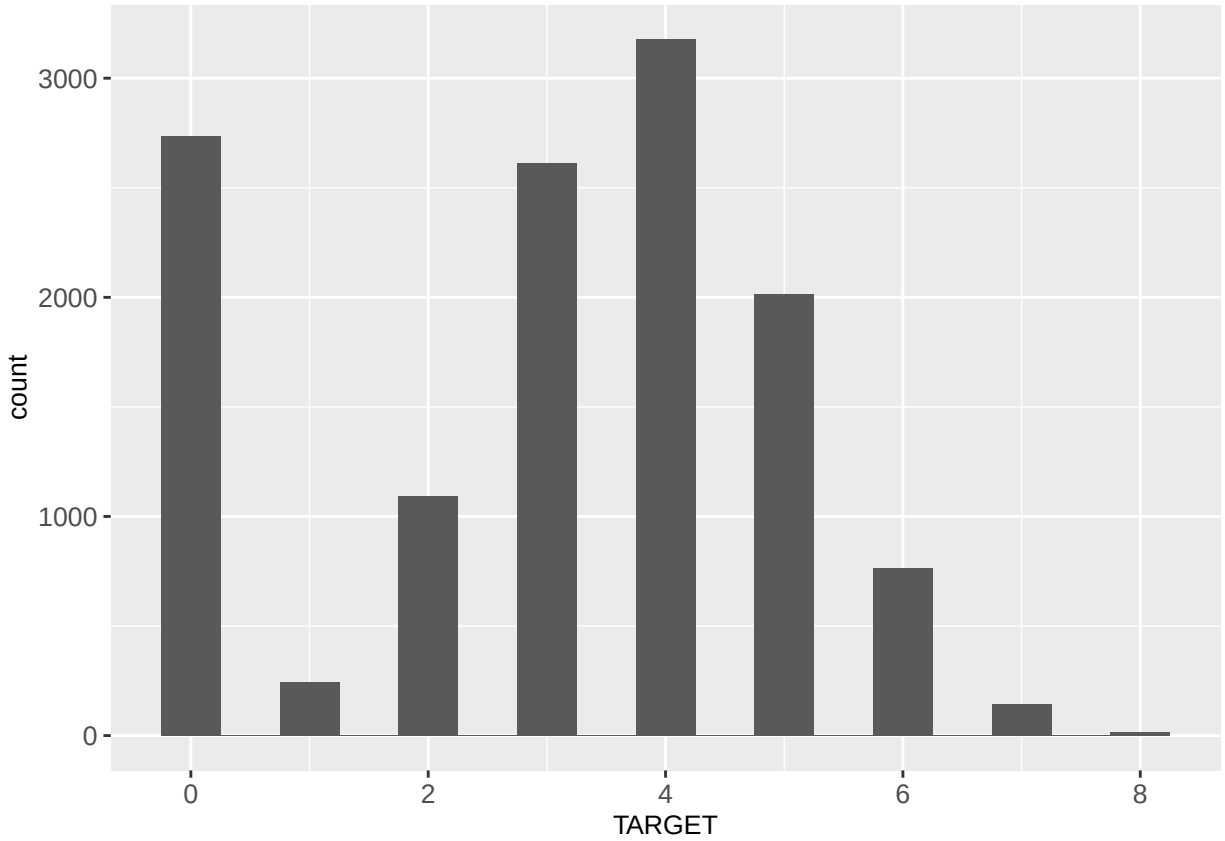

8
6
4
2
0
TARGET
1
2
3
4
NA
factor(STARS)
factor(STARS)
1
2
3
4



From the above diagrams, we notice that the NAs for STARS showing us a different distribution. So, we have to take care of this in the data preparation.

Data Distributions

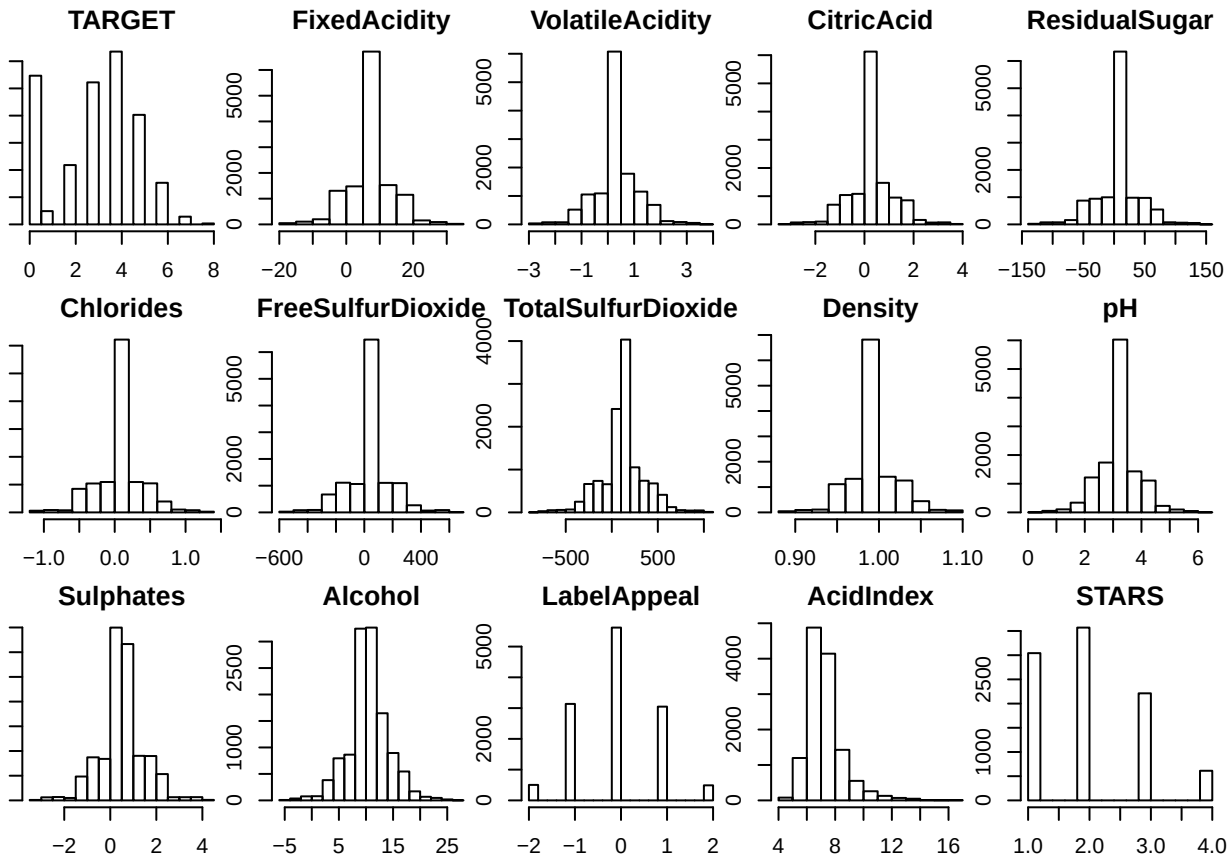
Lets check the overall distribution of the TAGET variable (which is a *count variable* indicating the number of sample cases):

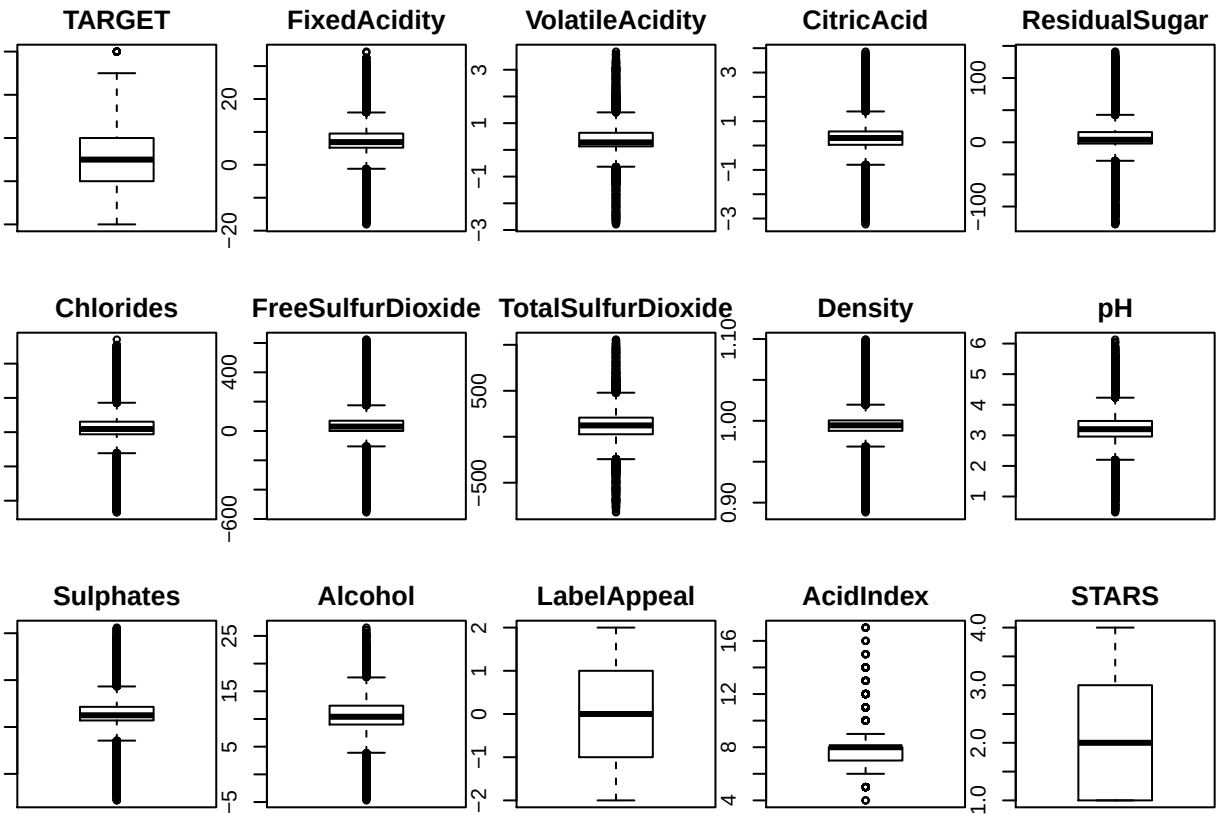


The above *TARGET* distribution has lot of *ZERO* values, which would indicate the *no sample cases purchased*, which could be due to NA values presence Or, some business reasons. But overall this appears close to *Poisson* distribution.

Lets check other variables distributions:

```
##  [1] "INDEX"             "TARGET"             "FixedAcidity"
##  [4] "VolatileAcidity"   "CitricAcid"         "ResidualSugar"
##  [7] "Chlorides"         "FreeSulfurDioxide"  "TotalSulfurDioxide"
## [10] "Density"           "pH"                 "Sulphates"
## [13] "Alcohol"           "LabelAppeal"        "AcidIndex"
## [16] "STARS"
```

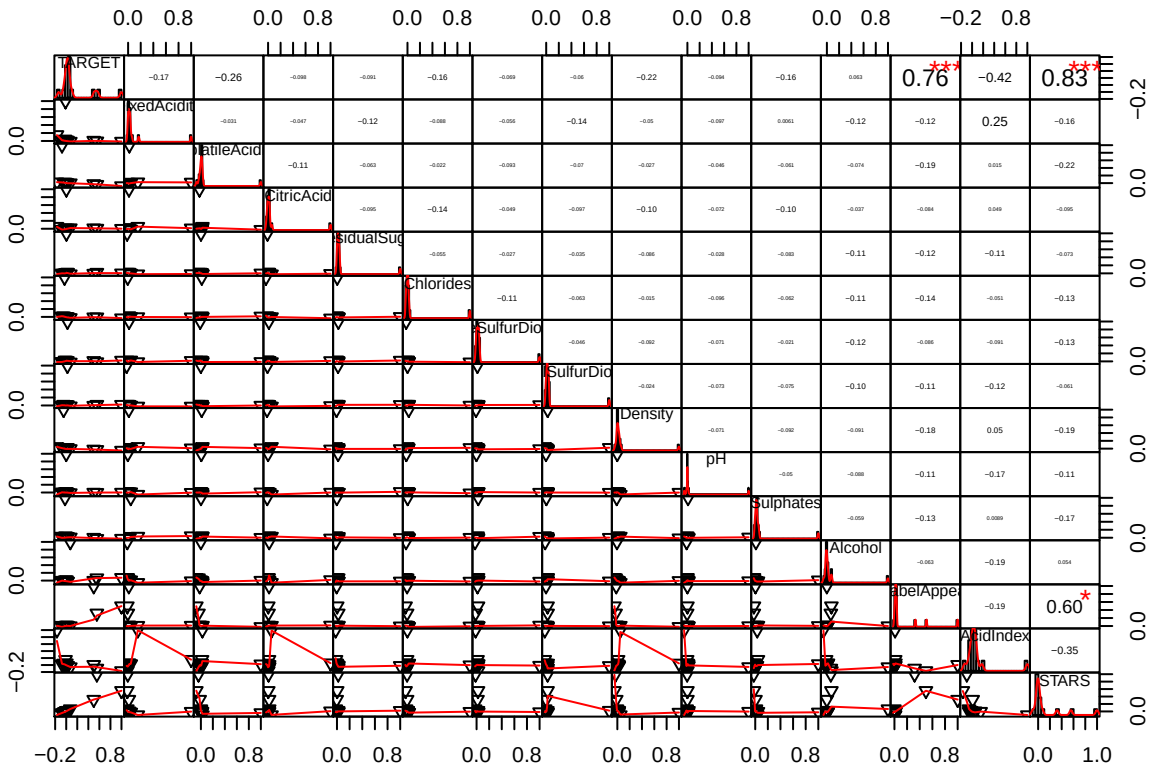

TARGET
FixedAcidity
VolatileAcidity
CitricAcid
ResidualSugar
Chlorides
FreeSulfurDioxide
TotalSulfurDioxide
Density
pH
Sulphates
Alcohol
LabelAppeal
AcidIndex
STARS



Most of the variables appears to be normally distributed.

Correlations

Lets visualize the correlation graph:



The above indicates the *STARS* and *LabelAppeal* are significant variables from correlation perspective.

Data Preparation

```
##      TARGET       FixedAcidity     VolatileAcidity     CitricAcid
##  Min.   :0.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400
##  1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300
##  Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100
##  Mean   :3.029   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084
##  3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800
##  Max.   :8.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600
##
##  ResidualSugar        Chlorides       FreeSulfurDioxide TotalSulfurDioxide
##  Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00   Min.   :-823.0
##  1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00   1st Qu.:  27.0
##  Median :   3.900   Median : 0.0460   Median :  30.00   Median : 123.0
##  Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85   Mean   : 120.7
##  3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00   3rd Qu.: 208.0
##  Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00   Max.   :1057.0
##  NA's   :616        NA's   :638       NA's   :647       NA's   :682
##     Density             pH          Sulphates          Alcohol
##  Min.   :0.8881   Min.   :0.480   Min.   :-3.1300   Min.   :-4.700
##  1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800   1st Qu.: 8.700
##  Median :0.9945   Median :3.200   Median : 0.5000   Median :10.200
##  Mean   :0.9942   Mean   :3.208   Mean   : 0.5271   Mean   : 9.954
```

```
##  3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600   3rd Qu.:12.200
##  Max.   :1.0992   Max.   :6.130   Max.   : 4.2400   Max.   :26.500
##                   NA's   :395     NA's   :1210
##   LabelAppeal          AcidIndex
##  Min.   :-2.000000   Min.   : 4.000
##  1st Qu.:-1.000000   1st Qu.: 7.000
##  Median : 0.000000   Median : 8.000
##  Mean   :-0.009066   Mean   : 7.773
##  3rd Qu.: 1.000000   3rd Qu.: 8.000
##  Max.   : 2.000000   Max.   :17.000
##
##  C:/Users/administrator.BARNETTASSOCIAT/Documents/GitHub/DATA621-HW/HW5/CT4-HW5.Rmd_1
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2377
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
##  C:/Users/administrator.BARNETTASSOCIAT/Documents/GitHub/DATA621-HW/HW5/CT4-HW5.Rmd_2
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.279
##  3rd Qu.:1.000
##  Max.   :1.000
##
##  C:/Users/administrator.BARNETTASSOCIAT/Documents/GitHub/DATA621-HW/HW5/CT4-HW5.Rmd_3
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1729
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
##  C:/Users/administrator.BARNETTASSOCIAT/Documents/GitHub/DATA621-HW/HW5/CT4-HW5.Rmd_4
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.04783
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
##  C:/Users/administrator.BARNETTASSOCIAT/Documents/GitHub/DATA621-HW/HW5/CT4-HW5.Rmd_NA
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2625
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
```

Build Models

Model Selection

Appendix