Homework #5: Count Regression Models

Douglas Barley, Ethan Haley, Isabel Magnus, John Mazon, Vinayak Kamath, Arushi

11/28/2021

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:

-**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 content of wine

-**TotalSulfurDioxide**: Total Sulfur Dioxide of Wine

-**VolatileAcidity**: Volatile Acid content of wine

-**pH**: pH of wine

```
#importing the train an eval data
wine_train_df<- read.csv("https://raw.githubusercontent.com/johnm1990/msds-621/main/wine-training-data.
wine_train_df <- wine_train_df[,2:16]
wine_eval_df<- read.csv("https://raw.githubusercontent.com/johnm1990/msds-621/main/wine-evaluation-data
wine_eval_df <- wine_eval_df[,2:16]
#per assignment instructions, we don't use first column 'ID', so we remove it, we performed in above ma
```

DATA EXPLORATION

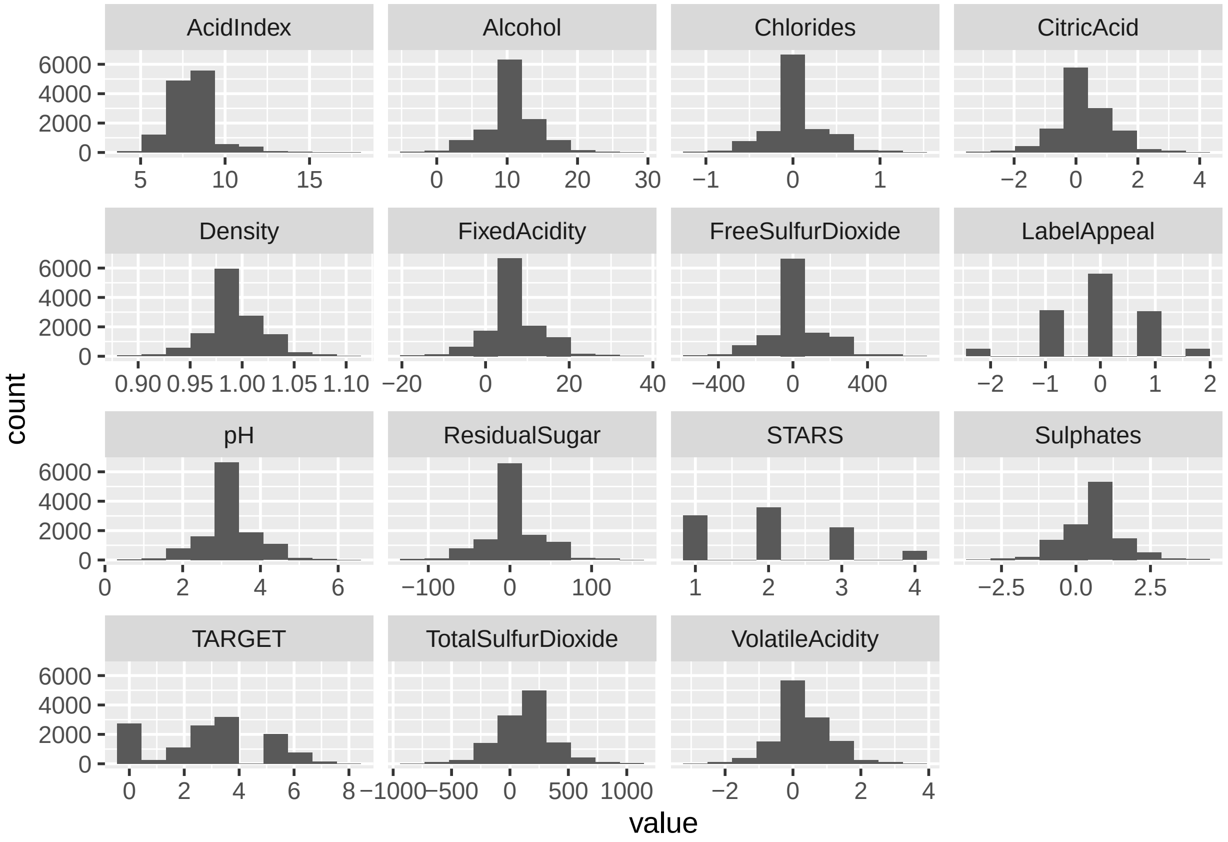
```
summary(wine_train_df)
```

```
##      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.   :-4.70
##  1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800   1st Qu.: 9.00
##  Median :0.9945   Median :3.200   Median : 0.5000   Median :10.40
##  Mean   :0.9942   Mean   :3.208   Mean   : 0.5271   Mean   :10.49
##  3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600   3rd Qu.:12.40
##  Max.   :1.0992   Max.   :6.130   Max.   : 4.2400   Max.   :26.50
##                   NA's   :395     NA's   :1210      NA's   :653
##   LabelAppeal          AcidIndex          STARS
##  Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##                                       NA's   :3359
```

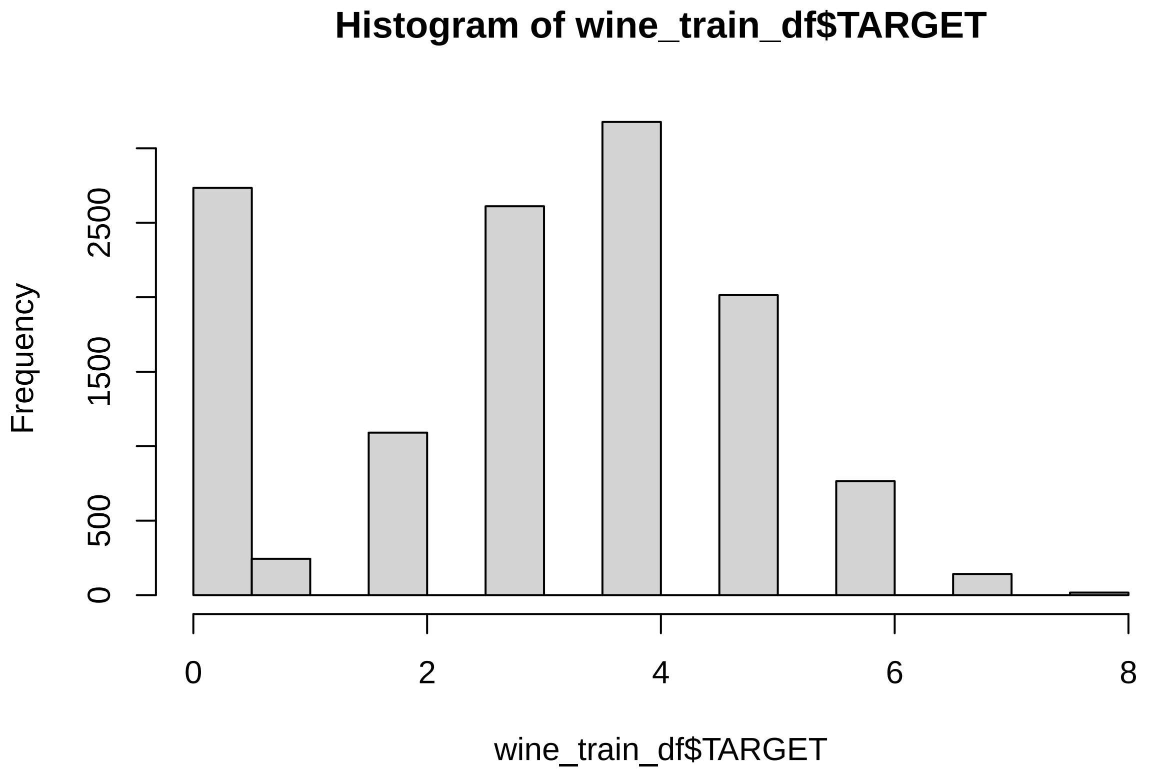
```
kable(format(sapply(wine_train_df, function(wine_train_df) c( "Stand dev" = round(sd(wine_train_df, na.
                         "Mean"= mean(wine_train_df,na.rm=TRUE),
                         "n" = length(wine_train_df),
                         "Median" = median(wine_train_df,na.rm = TRUE),
                         "CoeffofVariation" = sd(wine_train_df)/mean(wine_train_df,na.rm=TRUE),
                         "Minimum" = min(wine_train_df),
                         "Maximum" = max(wine_train_df),
                         "Upper Quantile" = quantile(wine_train_df,1,na.rm = TRUE),
                         "LowerQuartile" = quantile(wine_train_df,0,na.rm = TRUE)
                    )
), scientific = FALSE)
)
```

	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	C
Stand dev	1.930000000	6.320000000	0.780000000	0.860000000	33.750000000	0
Mean	3.029073857	7.075717077	0.324103947	0.308412661	5.418733065	0
n	12795.000000000	12795.000000000	12795.000000000	12795.000000000	12795.000000000	1
Median	3.000000000	6.900000000	0.280000000	0.310000000	3.900000000	0
CoeffofVariation	0.635959475	0.892862644	2.419020945	2.795215292	NA	N
Minimum	0.000000000	-18.100000000	-2.790000000	-3.240000000	NA	N
Maximum	8.000000000	34.400000000	3.680000000	3.860000000	NA	N
Upper Quantile.100%	8.000000000	34.400000000	3.680000000	3.860000000	141.150000000	1
LowerQuartile.0%	0.000000000	-18.100000000	-2.790000000	-3.240000000	-127.800000000	-

```
ggplot(gather(wine_train_df), aes(value)) +
    geom_histogram(bins = 10) +
    facet_wrap(~key, scales = 'free_x')
```



```
hist(wine_train_df$TARGET)
```



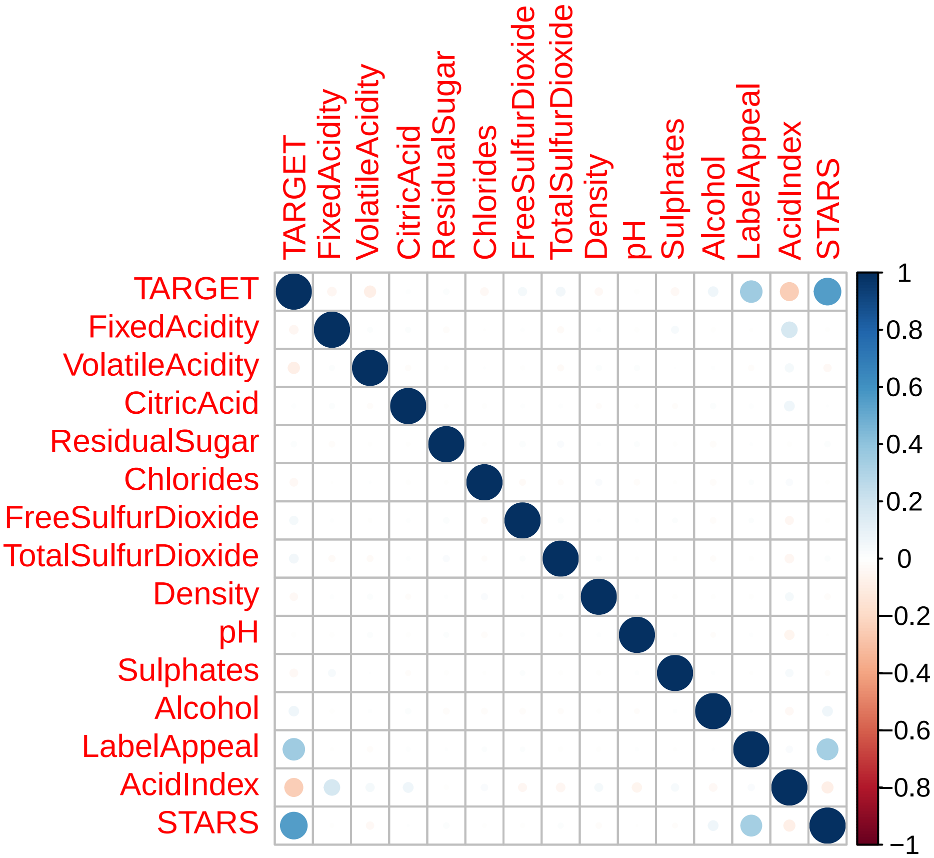
```
table(wine_train_df$TARGET)
```

```
##
##    0    1    2    3    4    5    6    7    8
## 2734  244 1091 2611 3177 2014  765  142   17
```

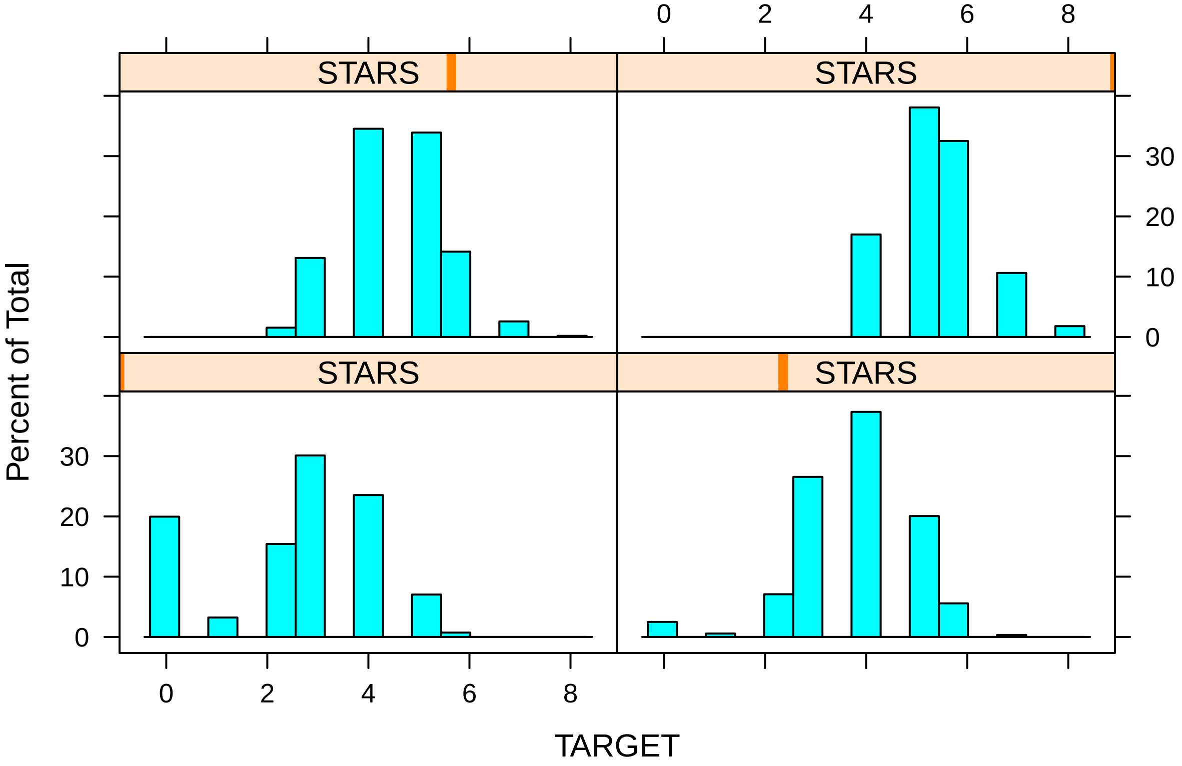
```
#Corr matrix and the scatterplot matrix
##correlation matrix
wine_train_df.rcorr = rcorr(as.matrix(wine_train_df))
wine_train_df.rcorr
```

```
##                    TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar
## TARGET               1.00        -0.05           -0.09       0.01          0.02
## FixedAcidity        -0.05         1.00            0.01       0.01         -0.02
## VolatileAcidity     -0.09         0.01            1.00      -0.02         -0.01
## CitricAcid           0.01         0.01           -0.02       1.00         -0.01
## ResidualSugar        0.02        -0.02           -0.01      -0.01          1.00
## Chlorides           -0.04         0.00            0.00      -0.01         -0.01
## FreeSulfurDioxide    0.04         0.00           -0.01       0.01          0.02
## TotalSulfurDioxide   0.05        -0.02           -0.02       0.01          0.02
## Density             -0.04         0.01            0.01      -0.01          0.00
## pH                  -0.01        -0.01            0.01      -0.01          0.01
## Sulphates           -0.04         0.03            0.00      -0.01         -0.01
## Alcohol              0.06        -0.01            0.00       0.02         -0.02
## LabelAppeal          0.36         0.00           -0.02       0.01          0.00
## AcidIndex           -0.25         0.18            0.04       0.07         -0.01
## STARS                0.56        -0.01           -0.03       0.00          0.02
##                    Chlorides FreeSulfurDioxide TotalSulfurDioxide Density    pH
## TARGET                 -0.04              0.04               0.05   -0.04 -0.01
## FixedAcidity            0.00              0.00              -0.02    0.01 -0.01
## VolatileAcidity         0.00             -0.01              -0.02    0.01  0.01
## CitricAcid             -0.01              0.01               0.01   -0.01 -0.01
## ResidualSugar          -0.01              0.02               0.02    0.00  0.01
## Chlorides               1.00             -0.02              -0.01    0.02 -0.02
## FreeSulfurDioxide      -0.02              1.00               0.01    0.00  0.01
## TotalSulfurDioxide     -0.01              0.01               1.00    0.01  0.00
## Density                 0.02              0.00               0.01    1.00  0.01
## pH                     -0.02              0.01               0.00    0.01  1.00
## Sulphates               0.00              0.01              -0.01   -0.01  0.01
## Alcohol                -0.02             -0.02              -0.02   -0.01 -0.01
## LabelAppeal             0.01              0.01              -0.01   -0.01  0.00
## AcidIndex               0.03             -0.04              -0.05    0.04 -0.06
## STARS                   0.00             -0.01               0.01   -0.02  0.00
##                    Sulphates Alcohol LabelAppeal AcidIndex STARS
## TARGET                 -0.04    0.06        0.36     -0.25  0.56
## FixedAcidity            0.03   -0.01        0.00      0.18 -0.01
## VolatileAcidity         0.00    0.00       -0.02      0.04 -0.03
## CitricAcid             -0.01    0.02        0.01      0.07  0.00
## ResidualSugar          -0.01   -0.02        0.00     -0.01  0.02
## Chlorides               0.00   -0.02        0.01      0.03  0.00
## FreeSulfurDioxide       0.01   -0.02        0.01     -0.04 -0.01
## TotalSulfurDioxide     -0.01   -0.02       -0.01     -0.05  0.01
## Density                -0.01   -0.01       -0.01      0.04 -0.02
## pH                      0.01   -0.01        0.00     -0.06  0.00
## Sulphates               1.00    0.00        0.00      0.03 -0.01
## Alcohol                 0.00    1.00        0.00     -0.04  0.07
## LabelAppeal             0.00    0.00        1.00      0.02  0.33
## AcidIndex               0.03   -0.04        0.02      1.00 -0.09
## STARS                  -0.01    0.07        0.33     -0.09  1.00
##
## n
##                    TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar
## TARGET              12795        12795           12795      12795         12179
## FixedAcidity        12795        12795           12795      12795         12179
## VolatileAcidity     12795        12795           12795      12795         12179
## CitricAcid          12795        12795           12795      12795         12179
## ResidualSugar       12179        12179           12179      12179         12179
## Chlorides           12157        12157           12157      12157         11585
## FreeSulfurDioxide   12148        12148           12148      12148         11563
## TotalSulfurDioxide  12113        12113           12113      12113         11532
## Density             12795        12795           12795      12795         12179
## pH                  12400        12400           12400      12400         11802
## Sulphates           11585        11585           11585      11585         11030
## Alcohol             12142        12142           12142      12142         11563
## LabelAppeal         12795        12795           12795      12795         12179
## AcidIndex           12795        12795           12795      12795         12179
## STARS                9436         9436            9436       9436          8984
##                    Chlorides FreeSulfurDioxide TotalSulfurDioxide Density    pH
## TARGET                 12157             12148              12113   12795 12400
## FixedAcidity           12157             12148              12113   12795 12400
## VolatileAcidity        12157             12148              12113   12795 12400
## CitricAcid             12157             12148              12113   12795 12400
## ResidualSugar          11585             11563              11532   12179 11802
## Chlorides              12157             11544              11510   12157 11773
## FreeSulfurDioxide      11544             12148              11512   12148 11771
## TotalSulfurDioxide     11510             11512              12113   12113 11739
## Density                12157             12148              12113   12795 12400
## pH                     11773             11771              11739   12400 12400
## Sulphates              10991             10995              10973   11585 11228
## Alcohol                11538             11527              11497   12142 11771
## LabelAppeal            12157             12148              12113   12795 12400
## AcidIndex              12157             12148              12113   12795 12400
## STARS                   8969              8979               8942    9436  9154
##                    Sulphates Alcohol LabelAppeal AcidIndex STARS
## TARGET                 11585   12142       12795     12795  9436
## FixedAcidity           11585   12142       12795     12795  9436
## VolatileAcidity        11585   12142       12795     12795  9436
## CitricAcid             11585   12142       12795     12795  9436
## ResidualSugar          11030   11563       12179     12179  8984
## Chlorides              10991   11538       12157     12157  8969
## FreeSulfurDioxide      10995   11527       12148     12148  8979
## TotalSulfurDioxide     10973   11497       12113     12113  8942
## Density                11585   12142       12795     12795  9436
## pH                     11228   11771       12400     12400  9154
## Sulphates              11585   10989       11585     11585  8564
## Alcohol                10989   12142       12142     12142  8963
## LabelAppeal            11585   12142       12795     12795  9436
## AcidIndex              11585   12142       12795     12795  9436
## STARS                   8564    8963        9436      9436  9436
##
## P
##                    TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar
## TARGET                    0.0000       0.0000          0.3260     0.0688
## FixedAcidity       0.0000              0.1616          0.1072     0.0375
## VolatileAcidity    0.0000 0.1616                       0.0552     0.4744
## CitricAcid         0.3260 0.1072       0.0552                     0.4438
## ResidualSugar      0.0688 0.0375       0.4744          0.4438
## Chlorides          0.0000 0.9598       0.9134          0.3449     0.5471
## FreeSulfurDioxide  0.0000 0.5837       0.4354          0.4787     0.0600
## TotalSulfurDioxide 0.0000 0.0133       0.0203          0.4867     0.0158
## Density            0.0000 0.4638       0.0956          0.1145     0.6509
## pH                 0.2930 0.3172       0.1302          0.3322     0.1880
## Sulphates          0.0000 0.0009       0.9889          0.1621     0.4173
## Alcohol            0.0000 0.3018       0.6536          0.0603     0.0315
## LabelAppeal        0.0000 0.7034       0.0547          0.3279     0.7979
## AcidIndex          0.0000 0.0000       0.0000          0.0000     0.2989
## STARS              0.0000 0.5197       0.0008          0.9485     0.1126
##                    Chlorides FreeSulfurDioxide TotalSulfurDioxide Density
## TARGET             0.0000    0.0000            0.0000             0.0000
## FixedAcidity       0.9598    0.5837            0.0133             0.4638
## VolatileAcidity    0.9134    0.4354            0.0203             0.0956
## CitricAcid         0.3449    0.4787            0.4867             0.1145
## ResidualSugar      0.5471    0.0600            0.0158             0.6509
## Chlorides                    0.0264            0.1333             0.0125
## FreeSulfurDioxide  0.0264                      0.1410             0.7263
## TotalSulfurDioxide 0.1333    0.1410                               0.1584
## Density            0.0125    0.7263            0.1584
## pH                 0.0561    0.5117            0.6380             0.5207
## Sulphates          0.7302    0.2242            0.4550             0.3296
## Alcohol            0.0344    0.0460            0.0871             0.4267
## LabelAppeal        0.2466    0.2566            0.2834             0.2892
## AcidIndex          0.0054    0.0000            0.0000             0.0000
## STARS              0.6405    0.3895            0.1878             0.0757
##                    pH     Sulphates Alcohol LabelAppeal AcidIndex STARS
## TARGET             0.2930 0.0000    0.0000  0.0000      0.0000    0.0000
## FixedAcidity       0.3172 0.0009    0.3018  0.7034      0.0000    0.5197
## VolatileAcidity    0.1302 0.9889    0.6536  0.0547      0.0000    0.0008
## CitricAcid         0.3322 0.1621    0.0603  0.3279      0.0000    0.9485
## ResidualSugar      0.1880 0.4173    0.0315  0.7979      0.2989    0.1126
## Chlorides          0.0561 0.7302    0.0344  0.2466      0.0054    0.6405
## FreeSulfurDioxide  0.5117 0.2242    0.0460  0.2566      0.0000    0.3895
## TotalSulfurDioxide 0.6380 0.4550    0.0871  0.2834      0.0000    0.1878
## Density            0.5207 0.3296    0.4267  0.2892      0.0000    0.0757
## pH                        0.5618    0.2103  0.6450      0.0000    0.9627
## Sulphates          0.5618           0.6192  0.6757      0.0002    0.2548
## Alcohol            0.2103 0.6192            0.9099      0.0000    0.0000
## LabelAppeal        0.6450 0.6757    0.9099              0.0051    0.0000
## AcidIndex          0.0000 0.0002    0.0000  0.0051                0.0000
## STARS              0.9627 0.2548    0.0000  0.0000      0.0000
```

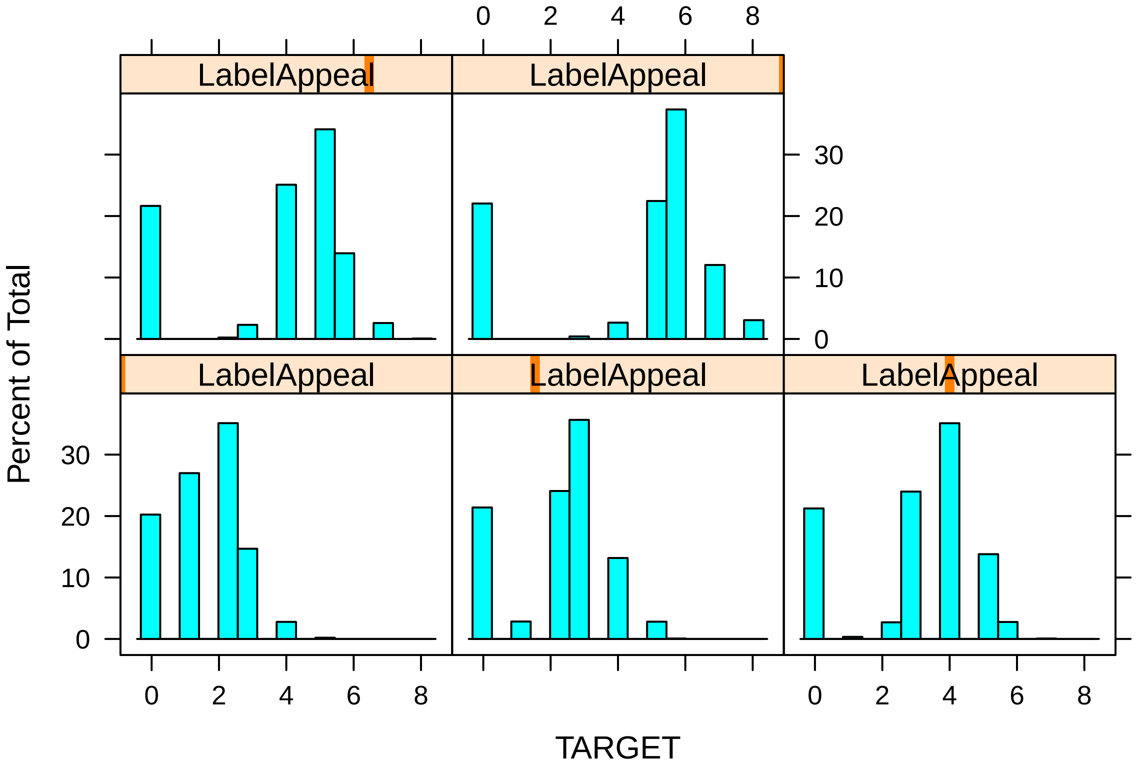
```
wine_train_df.cor = cor(wine_train_df, use = "pairwise.complete.obs")
corrplot(wine_train_df.cor)
```



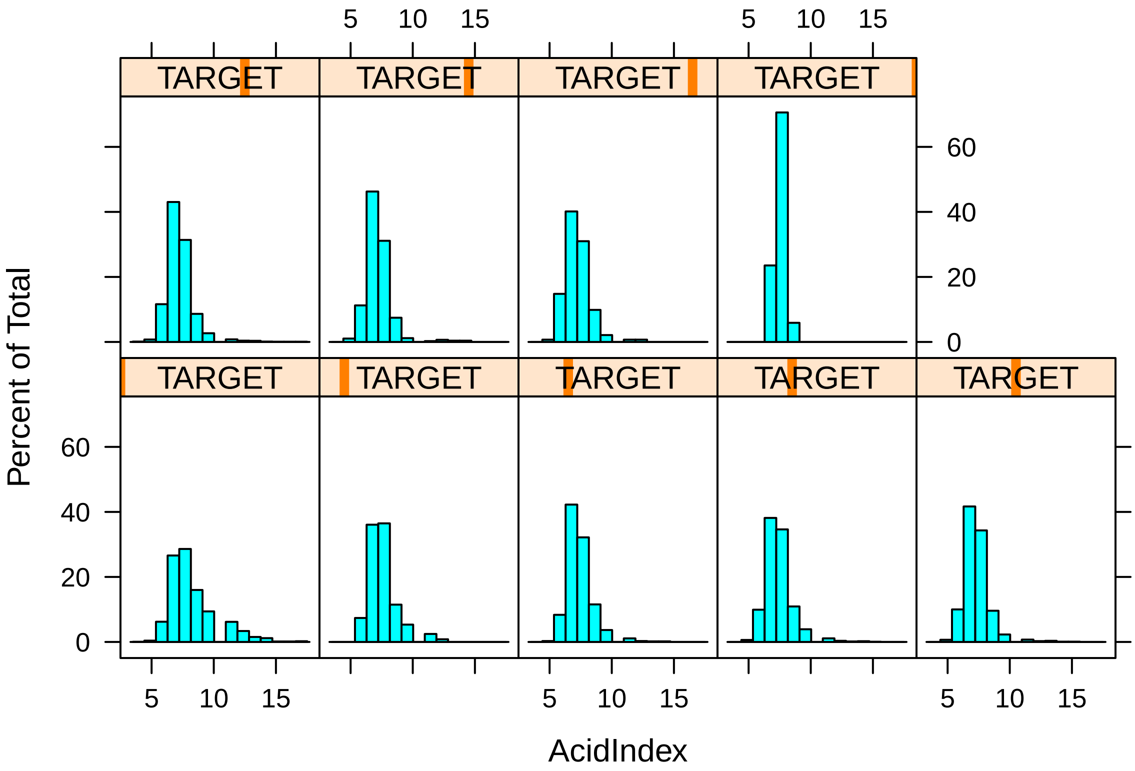
```
histogram(~ TARGET | STARS, data = wine_train_df)
```



```
histogram(~ TARGET | LabelAppeal, data = wine_train_df)
```



```
histogram(~ AcidIndex | TARGET, data = wine_train_df)
```



```
cor_stars_tgt <- cor.test(wine_train_df$STARS, wine_train_df$TARGET)
cor_stars_tgt
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine_train_df$STARS and wine_train_df$TARGET
## t = 65.446, df = 9434, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5447586 0.5725160
## sample estimates:
##       cor
## 0.5587938
```

```
cor_lbl_tgr <- cor.test(wine_train_df$LabelAppeal, wine_train_df$TARGET)
cor_lbl_tgr
```

```
##
##  Pearson's product-moment correlation
##
## data:  wine_train_df$LabelAppeal and wine_train_df$TARGET
## t = 43.158, df = 12793, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3412812 0.3715329
## sample estimates:
##       cor
## 0.3565005
```

```
cor_acid_tgt <- cor.test(wine_train_df$AcidIndex, wine_train_df$TARGET)
cor_acid_tgt
```

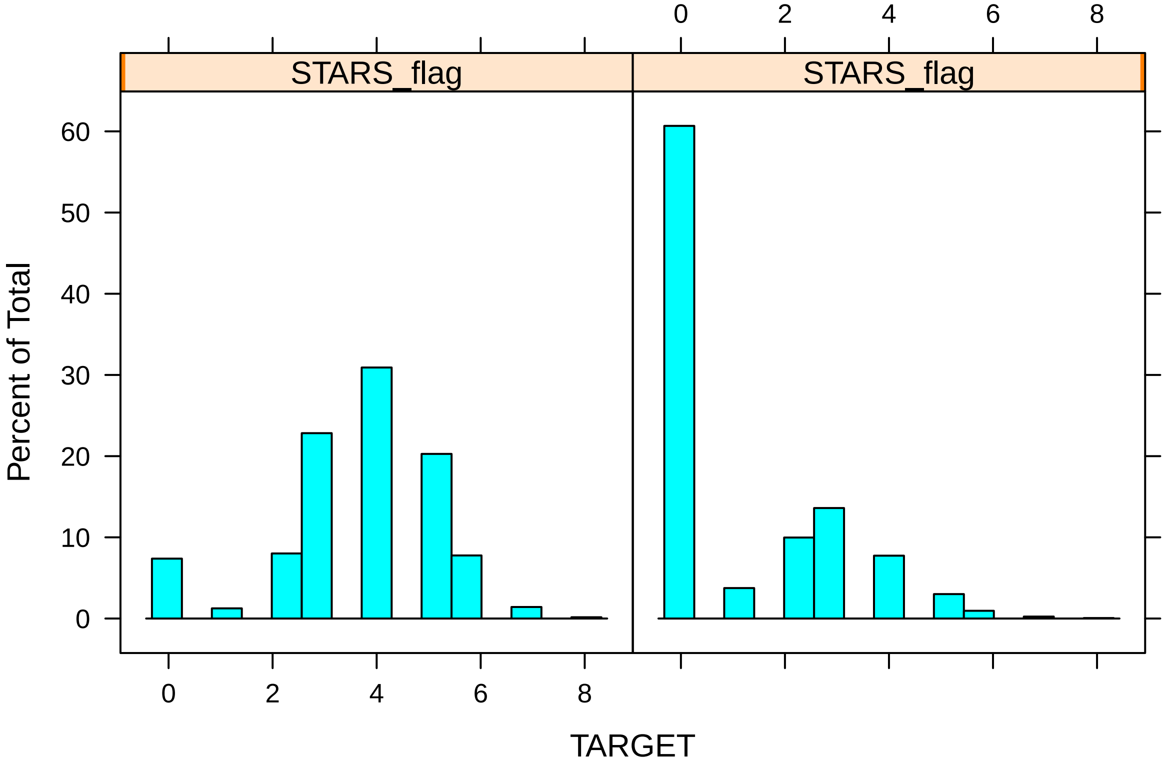
```
##
##  Pearson's product-moment correlation
##
## data:  wine_train_df$AcidIndex and wine_train_df$TARGET
## t = -28.712, df = 12793, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2622588 -0.2297013
## sample estimates:
##        cor
## -0.2460494
```

```
# Compute the analysis of variance, when has more than two groups perform ANOVA
res.aov <- aov(AcidIndex ~ TARGET, data = wine_train_df)
# Summary of the analysis
summary(res.aov)
```

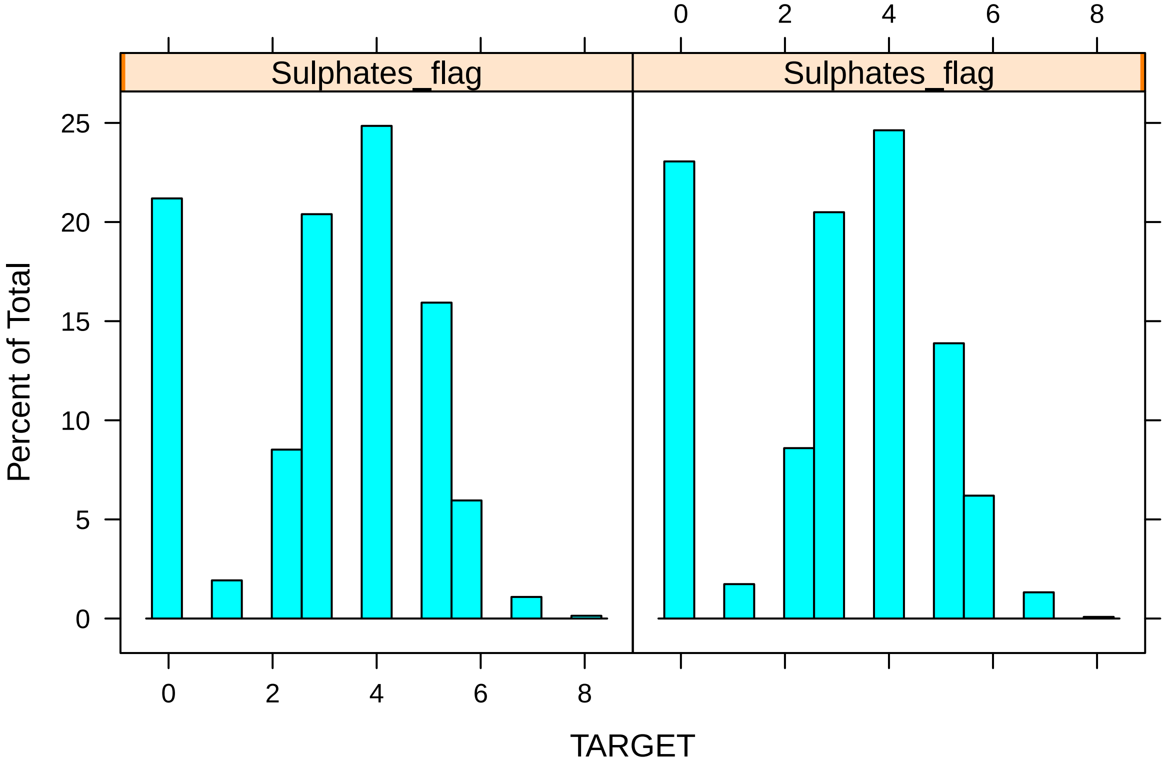
```
##                Df Sum Sq Mean Sq F value Pr(>F)
## TARGET          1   1358  1357.6   824.4 <2e-16 ***
## Residuals   12793  21067     1.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

DATA PREPARATION

```
wine_train_df <- wine_train_df %>%
            mutate(ResidualSugar= ifelse(is.na(ResidualSugar),
                                         mean(ResidualSugar, na.rm=TRUE),ResidualSugar),
                   Chlorides= ifelse(is.na(Chlorides),
                                         mean(Chlorides, na.rm=TRUE),Chlorides),
                   FreeSulfurDioxide= ifelse(is.na(FreeSulfurDioxide),
                                         mean(FreeSulfurDioxide, na.rm=TRUE),FreeSulfurDioxide),
                   TotalSulfurDioxide= ifelse(is.na(TotalSulfurDioxide),
                                         mean(TotalSulfurDioxide, na.rm=TRUE),TotalSulfurDioxide),
                   pH= ifelse(is.na(pH),
                                         mean(pH, na.rm=TRUE),pH),
                   Alcohol= ifelse(is.na(Alcohol),
                                         mean(Alcohol, na.rm=TRUE),Alcohol),
                   )

wine_eval_df <- wine_eval_df %>%
            mutate(ResidualSugar= ifelse(is.na(ResidualSugar),
                                         mean(ResidualSugar, na.rm=TRUE),ResidualSugar),
                   Chlorides= ifelse(is.na(Chlorides),
                                         mean(Chlorides, na.rm=TRUE),Chlorides),
                   FreeSulfurDioxide= ifelse(is.na(FreeSulfurDioxide),
                                         mean(FreeSulfurDioxide, na.rm=TRUE),FreeSulfurDioxide),
                   TotalSulfurDioxide= ifelse(is.na(TotalSulfurDioxide),
                                         mean(TotalSulfurDioxide, na.rm=TRUE),TotalSulfurDioxide),
                   pH= ifelse(is.na(pH),
                                         mean(pH, na.rm=TRUE),pH),
                   Alcohol= ifelse(is.na(Alcohol),
                                         mean(Alcohol, na.rm=TRUE),Alcohol),
                   )




#LOG TRANSFORMATION

wine_train_df$FreeSulfurDioxide_log <- log(wine_train_df$FreeSulfurDioxide + 1 - min(wine_train_df$FreeS
wine_train_df$TotalSulfurDioxide_log <- log(wine_train_df$TotalSulfurDioxide + 1 - min(wine_train_df$To

wine_eval_df$FreeSulfurDioxide_log <- log(wine_eval_df$FreeSulfurDioxide + 1 - min(wine_eval_df$FreeSul
wine_eval_df$TotalSulfurDioxide_log <- log(wine_eval_df$TotalSulfurDioxide + 1 - min(wine_eval_df$TotalS


# #Flags for N/A's:

wine_train_df <- wine_train_df %>%
            mutate(Sulphates_flag= ifelse(is.na(Sulphates),1,0),
                   STARS_flag= ifelse(is.na(STARS),1,0)
                   )
#flags = will create 1 if NA
histogram(~ TARGET | STARS_flag, data = wine_train_df)
```



```r
histogram(~ TARGET | Sulphates_flag, data = wine_train_df)
```



```
#corrective actions
wine_train_df <- wine_train_df %>%
            mutate(Sulphates= ifelse(is.na(Sulphates),
                                         mean(Sulphates, na.rm=TRUE),Sulphates),
                   STARS_merged=ifelse(is.na(STARS),0,STARS))

wine_eval_df <- wine_eval_df %>%
            mutate(Sulphates= ifelse(is.na(Sulphates),
                                         mean(Sulphates, na.rm=TRUE),Sulphates),
                   STARS_merged=ifelse(is.na(STARS),0,STARS))

table(wine_train_df$STARS)
```

```
##
##    1    2    3    4
## 3042 3570 2212  612
```

```
table(wine_train_df$STARS_merged)
```

```
##
##    0    1    2    3    4
## 3359 3042 3570 2212  612
```

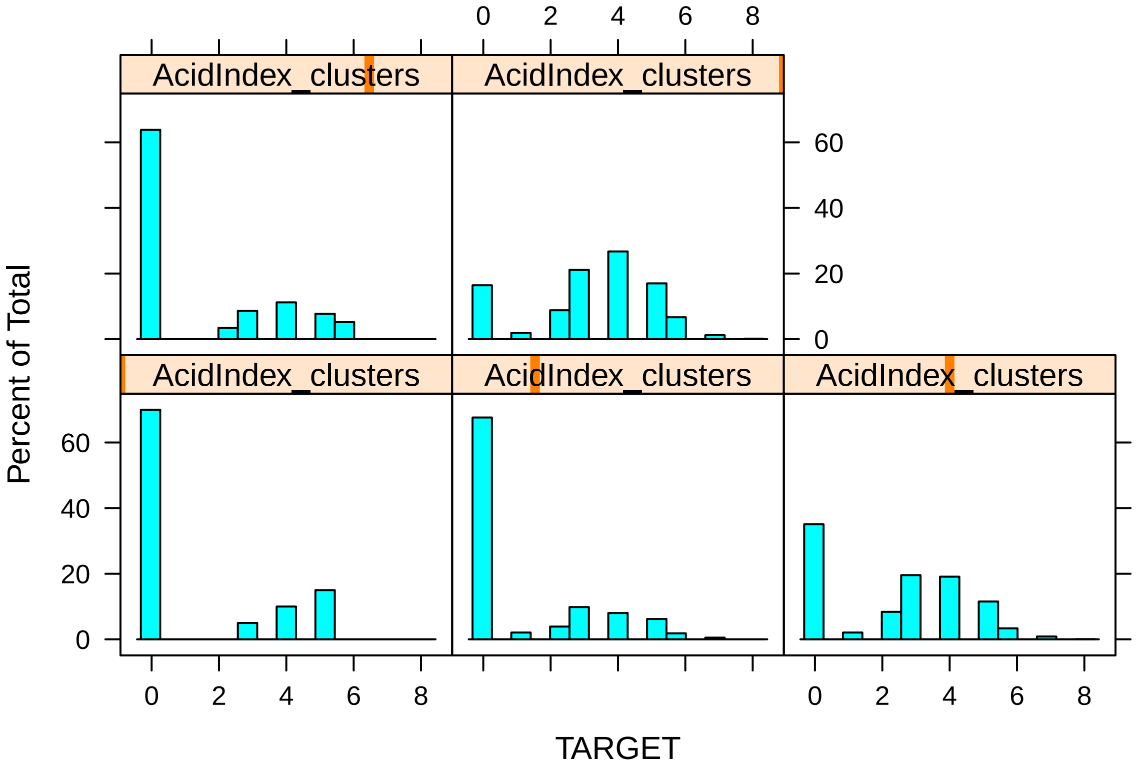
```
##you will see includes no 0 columns
table(wine_train_df$STARS)
```

```
##
##    1    2    3    4
## 3042 3570 2212  612
```

```
#creating clusters for acid index
kmeans.re <- kmeans(wine_train_df$AcidIndex, centers = 5)
table(kmeans.re$cluster)
```

```
##
##     1     2     3     4     5
##    20   386  1978   116 10295
```

```
wine_train_df$AcidIndex_clusters <- kmeans.re$cluster
histogram(~ TARGET | AcidIndex_clusters, data = wine_train_df)
```



BUILD THE MODELS

```r
#multiple reg
model.manual.mr <- lm(TARGET ~ STARS_merged+LabelAppeal+AcidIndex, data = wine_train_df)
summary(model.manual.mr)
```

```
## 
## Call:
## lm(formula = TARGET ~ STARS_merged + LabelAppeal + AcidIndex, 
##     data = wine_train_df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.5478 -0.9207  0.0973  0.9289  6.0697 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   3.212216   0.075692   42.44   <2e-16 ***
## STARS_merged  0.986226   0.010453   94.35   <2e-16 ***
## LabelAppeal   0.430953   0.013718   31.41   <2e-16 ***
## AcidIndex    -0.214113   0.009037  -23.69   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 1.33 on 12791 degrees of freedom
## Multiple R-squared:  0.5236, Adjusted R-squared:  0.5235 
## F-statistic:  4686 on 3 and 12791 DF,  p-value: < 2.2e-16
```

```r
#
fullmod_regressiondata <- wine_train_df %>%
  dplyr::select(TARGET,FixedAcidity,VolatileAcidity,CitricAcid,
    ResidualSugar,Chlorides,Density,pH,Sulphates,Alcohol,LabelAppeal,AcidIndex,
   FreeSulfurDioxide_log,TotalSulfurDioxide_log,
    STARS_merged)
#
model.full.mr  <- lm(TARGET ~ . , data = fullmod_regressiondata)
summary(model.full.mr)
```

```
## 
## Call:
## lm(formula = TARGET ~ ., data = fullmod_regressiondata)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.5451 -0.9491  0.0673  0.9066  5.9806 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             2.161e+00  5.705e-01   3.787 0.000153 ***
## FixedAcidity            2.723e-05  1.885e-03   0.014 0.988477    
## VolatileAcidity        -9.943e-02  1.498e-02  -6.637 3.34e-11 ***
## CitricAcid              2.088e-02  1.363e-02   1.532 0.125505    
## ResidualSugar           2.123e-04  3.560e-04   0.596 0.550990    
## Chlorides              -1.250e-01  3.778e-02  -3.308 0.000942 ***
## Density               -7.827e-01  4.420e-01  -1.771 0.076595 .  
## pH                    -3.447e-02  1.754e-02  -1.965 0.049465 *  
## Sulphates             -3.278e-02  1.322e-02  -2.480 0.013161 *  
## Alcohol                1.075e-02  3.234e-03   3.325 0.000886 ***
## LabelAppeal            4.330e-01  1.367e-02  31.675  < 2e-16 ***
## AcidIndex             -2.088e-01  9.213e-03 -22.668  < 2e-16 ***
## FreeSulfurDioxide_log  1.223e-01  3.669e-02   3.334 0.000860 ***
## TotalSulfurDioxide_log 1.576e-01  3.896e-02   4.046 5.24e-05 ***
## STARS_merged           9.769e-01  1.046e-02  93.426  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 1.325 on 12780 degrees of freedom
## Multiple R-squared:  0.5278, Adjusted R-squared:  0.5273 
## F-statistic:  1020 on 14 and 12780 DF,  p-value: < 2.2e-16
```

```
model.forward.mr <- model.full.mr %>% stepAIC(direction = "forward", trace = FALSE)
summary(model.forward.mr)
```

```
## 
## Call:
## lm(formula = TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + 
##     ResidualSugar + Chlorides + Density + pH + Sulphates + Alcohol + 
##     LabelAppeal + AcidIndex + FreeSulfurDioxide_log + TotalSulfurDioxide_log + 
##     STARS_merged, data = fullmod_regressiondata)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.5451 -0.9491  0.0673  0.9066  5.9806 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             2.161e+00  5.705e-01   3.787 0.000153 ***
## FixedAcidity            2.723e-05  1.885e-03   0.014 0.988477    
## VolatileAcidity        -9.943e-02  1.498e-02  -6.637 3.34e-11 ***
## CitricAcid              2.088e-02  1.363e-02   1.532 0.125505    
## ResidualSugar           2.123e-04  3.560e-04   0.596 0.550990    
## Chlorides              -1.250e-01  3.778e-02  -3.308 0.000942 ***
## Density                -7.827e-01  4.420e-01  -1.771 0.076595 .  
## pH                     -3.447e-02  1.754e-02  -1.965 0.049465 *  
## Sulphates              -3.278e-02  1.322e-02  -2.480 0.013161 *  
## Alcohol                 1.075e-02  3.234e-03   3.325 0.000886 ***
## LabelAppeal             4.330e-01  1.367e-02  31.675  < 2e-16 ***
## AcidIndex              -2.088e-01  9.213e-03 -22.668  < 2e-16 ***
## FreeSulfurDioxide_log   1.223e-01  3.669e-02   3.334 0.000860 ***
## TotalSulfurDioxide_log  1.576e-01  3.896e-02   4.046 5.24e-05 ***
## STARS_merged            9.769e-01  1.046e-02  93.426  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 1.325 on 12780 degrees of freedom
## Multiple R-squared:  0.5278, Adjusted R-squared:  0.5273 
## F-statistic:  1020 on 14 and 12780 DF,  p-value: < 2.2e-16
```

```
#Getting formula for the model
formula(model.forward.mr)
```

```
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar + 
##     Chlorides + Density + pH + Sulphates + Alcohol + LabelAppeal + 
##     AcidIndex + FreeSulfurDioxide_log + TotalSulfurDioxide_log + 
##     STARS_merged
```

```
model.backward.mr <- model.full.mr %>% stepAIC(direction = "backward", trace = FALSE)
summary(model.backward.mr)
```

```
## 
## Call:
## lm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides + 
##     Density + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + 
##     FreeSulfurDioxide_log + TotalSulfurDioxide_log + STARS_merged, 
##     data = fullmod_regressiondata)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.5457 -0.9467  0.0673  0.9064  5.9814 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             2.156070   0.570403   3.780 0.000158 ***
## VolatileAcidity        -0.099472   0.014981  -6.640 3.26e-11 ***
## CitricAcid              0.020824   0.013624   1.528 0.126428    
## Chlorides              -0.125077   0.037773  -3.311 0.000931 ***
## Density                -0.781561   0.441939  -1.768 0.077004 .  
## pH                     -0.034351   0.017541  -1.958 0.050220 .  
## Sulphates              -0.032832   0.013216  -2.484 0.012993 *  
## Alcohol                 0.010716   0.003233   3.314 0.000921 ***
## LabelAppeal             0.432995   0.013669  31.677  < 2e-16 ***
## AcidIndex              -0.208845   0.009074 -23.015  < 2e-16 ***
## FreeSulfurDioxide_log   0.122712   0.036681   3.345 0.000824 ***
## TotalSulfurDioxide_log  0.157958   0.038950   4.055 5.03e-05 ***
## STARS_merged            0.977012   0.010455  93.453  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 1.324 on 12782 degrees of freedom
## Multiple R-squared:  0.5278, Adjusted R-squared:  0.5273 
## F-statistic:  1190 on 12 and 12782 DF,  p-value: < 2.2e-16
```

```
AIC(model.backward.mr)
```

```
## [1] 43515.75
```

```
#Getting formula for the model
formula(model.backward.mr)
```

```
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + FreeSulfurDioxide_log +
##     TotalSulfurDioxide_log + STARS_merged
```

```
#manual poisson
model.manual.poisson <- glm(TARGET ~ STARS_merged+LabelAppeal+AcidIndex, data = wine_train_df,family =
summary(model.manual.poisson)
```

```
##
## Call:
## glm(formula = TARGET ~ STARS_merged + LabelAppeal + AcidIndex,
##     family = poisson, data = wine_train_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9872  -0.7168   0.0485   0.5527   3.2791
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.223551   0.036514   33.51   <2e-16 ***
## STARS_merged  0.313946   0.004507   69.65   <2e-16 ***
## LabelAppeal   0.132978   0.006060   21.95   <2e-16 ***
## AcidIndex    -0.088835   0.004462  -19.91   <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: 14804  on 12791  degrees of freedom
## AIC: 46754
##
## Number of Fisher Scoring iterations: 5
```

```
model.full.poisson  <- glm(TARGET ~ . , data = fullmod_regressiondata,family=poisson)
summary(model.full.poisson)
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = fullmod_regressiondata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9717  -0.7206   0.0689   0.5772   3.2241
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             8.244e-01  2.512e-01   3.282 0.001031 **
## FixedAcidity           -2.882e-04  8.205e-04  -0.351 0.725409
## VolatileAcidity        -3.344e-02  6.515e-03  -5.134 2.84e-07 ***
## CitricAcid              7.770e-03  5.892e-03   1.319 0.187282
## ResidualSugar           5.764e-05  1.546e-04   0.373 0.709370
## Chlorides              -4.156e-02  1.645e-02  -2.526 0.011527 *
## Density               -2.737e-01  1.920e-01  -1.426 0.153931    
## pH                    -1.571e-02  7.637e-03  -2.057 0.039639 *  
## Sulphates             -1.264e-02  5.749e-03  -2.198 0.027925 *  
## Alcohol                2.148e-03  1.410e-03   1.523 0.127676    
## LabelAppeal            1.333e-01  6.063e-03  21.993  < 2e-16 ***
## AcidIndex             -8.721e-02  4.547e-03 -19.179  < 2e-16 ***
## FreeSulfurDioxide_log  4.710e-02  1.617e-02   2.913 0.003582 ** 
## TotalSulfurDioxide_log  6.020e-02  1.779e-02   3.384 0.000715 ***
## STARS_merged           3.112e-01  4.531e-03  68.698  < 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: 14734  on 12780  degrees of freedom
## AIC: 46706
## 
## Number of Fisher Scoring iterations: 5
```

```
model.forward.poisson <- model.full.poisson %>% stepAIC(direction = "forward", trace = FALSE)
summary(model.forward.poisson)
```

```
## 
## Call:
## glm(formula = TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + 
##     ResidualSugar + Chlorides + Density + pH + Sulphates + Alcohol + 
##     LabelAppeal + AcidIndex + FreeSulfurDioxide_log + TotalSulfurDioxide_log + 
##     STARS_merged, family = poisson, data = fullmod_regressiondata)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.9717  -0.7206   0.0689   0.5772   3.2241  
## 
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)    
## (Intercept)             8.244e-01  2.512e-01   3.282 0.001031 ** 
## FixedAcidity           -2.882e-04  8.205e-04  -0.351 0.725409    
## VolatileAcidity        -3.344e-02  6.515e-03  -5.134 2.84e-07 ***
## CitricAcid              7.770e-03  5.892e-03   1.319 0.187282    
## ResidualSugar           5.764e-05  1.546e-04   0.373 0.709370    
## Chlorides              -4.156e-02  1.645e-02  -2.526 0.011527 *  
## Density                -2.737e-01  1.920e-01  -1.426 0.153931    
## pH                     -1.571e-02  7.637e-03  -2.057 0.039639 *  
## Sulphates              -1.264e-02  5.749e-03  -2.198 0.027925 *  
## Alcohol                 2.148e-03  1.410e-03   1.523 0.127676    
## LabelAppeal             1.333e-01  6.063e-03  21.993  < 2e-16 ***
## AcidIndex              -8.721e-02  4.547e-03 -19.179  < 2e-16 ***
## FreeSulfurDioxide_log   4.710e-02  1.617e-02   2.913 0.003582 ** 
## TotalSulfurDioxide_log  6.020e-02  1.779e-02   3.384 0.000715 ***
## STARS_merged            3.112e-01  4.531e-03  68.698  < 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: 14734  on 12780  degrees of freedom
## AIC: 46706
##
## Number of Fisher Scoring iterations: 5
```

```
#Getting formula for the model
formula(model.forward.poisson)
```

```
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +
##     Chlorides + Density + pH + Sulphates + Alcohol + LabelAppeal +
##     AcidIndex + FreeSulfurDioxide_log + TotalSulfurDioxide_log +
##     STARS_merged
```

```
model.backward.poisson<-model.full.poisson %>% stepAIC(direction = "backward", trace = FALSE)
summary(model.backward.poisson)
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + Chlorides + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + FreeSulfurDioxide_log +
##     TotalSulfurDioxide_log + STARS_merged, family = poisson,
##     data = fullmod_regressiondata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9799  -0.7206   0.0697   0.5792   3.2270
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             0.824164   0.251171   3.281 0.001033 **
## VolatileAcidity        -0.033647   0.006514  -5.166  2.4e-07 ***
## Chlorides              -0.041713   0.016450  -2.536 0.011223 *
## Density                -0.277539   0.191947  -1.446 0.148200
## pH                     -0.015612   0.007635  -2.045 0.040873 *
## Sulphates              -0.012782   0.005747  -2.224 0.026143 *
## Alcohol                 0.002182   0.001409   1.548 0.121515
## LabelAppeal             0.133393   0.006063  22.002  < 2e-16 ***
## AcidIndex              -0.087077   0.004491 -19.391  < 2e-16 ***
## FreeSulfurDioxide_log   0.047248   0.016169   2.922 0.003476 **
## TotalSulfurDioxide_log  0.060516   0.017784   3.403 0.000667 ***
## STARS_merged            0.311332   0.004530  68.734  < 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: 14736  on 12783  degrees of freedom
## AIC: 46702
##
## Number of Fisher Scoring iterations: 5
```

```
#Getting formula for the model
formula(model.backward.poisson)
```

```
## TARGET ~ VolatileAcidity + Chlorides + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + FreeSulfurDioxide_log +
##     TotalSulfurDioxide_log + STARS_merged
```

Backward consistently provided better results.

```
#negative binomial
model.manual.negbin <- glm.nb(TARGET ~ STARS_merged+LabelAppeal+AcidIndex, data = wine_train_df)
summary(model.manual.negbin)
```

```
##
## Call:
## glm.nb(formula = TARGET ~ STARS_merged + LabelAppeal + AcidIndex,
##     data = wine_train_df, init.theta = 48842.02805, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9872  -0.7168   0.0485   0.5527   3.2790
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.223558   0.036516   33.51   <2e-16 ***
## STARS_merged  0.313950   0.004508   69.65   <2e-16 ***
## LabelAppeal   0.132977   0.006060   21.94   <2e-16 ***
## AcidIndex    -0.088837   0.004463  -19.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48842.03) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 14804  on 12791  degrees of freedom
## AIC: 46757
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48842
##           Std. Err.:  50670
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -46746.7
```

```
#Step 1: Create a full model
model.full.negbin  <- glm.nb(TARGET ~ . , data = fullmod_regressiondata)
summary(model.full.negbin )
```

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

```
## glm.nb(formula = TARGET ~ ., data = fullmod_regressiondata, init.theta = 48988.32099,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9717  -0.7205   0.0689   0.5772   3.2239
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              8.244e-01  2.512e-01   3.282 0.001032 **
## FixedAcidity            -2.882e-04  8.205e-04  -0.351 0.725410
## VolatileAcidity         -3.345e-02  6.515e-03  -5.134 2.84e-07 ***
## CitricAcid               7.770e-03  5.892e-03   1.319 0.187294
## ResidualSugar            5.765e-05  1.547e-04   0.373 0.709355
## Chlorides               -4.156e-02  1.645e-02  -2.526 0.011528 *
## Density                 -2.737e-01  1.920e-01  -1.426 0.153939
## pH                      -1.571e-02  7.637e-03  -2.058 0.039638 *
## Sulphates               -1.264e-02  5.749e-03  -2.198 0.027925 *
## Alcohol                  2.148e-03  1.410e-03   1.523 0.127700
## LabelAppeal              1.333e-01  6.063e-03  21.992  < 2e-16 ***
## AcidIndex               -8.721e-02  4.547e-03 -19.178  < 2e-16 ***
## FreeSulfurDioxide_log    4.710e-02  1.617e-02   2.913 0.003583 **
## TotalSulfurDioxide_log   6.020e-02  1.779e-02   3.384 0.000715 ***
## STARS_merged             3.112e-01  4.531e-03  68.696  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for Negative Binomial(48988.32) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 14734  on 12780  degrees of freedom
## AIC: 46708
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48988
##           Std. Err.:  50753
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -46676.38
```

```r
model.forward.negbin <- model.full.negbin %>% stepAIC(direction = "forward", trace = FALSE)
summary(model.forward.negbin)
```

```
##
## Call:
## glm.nb(formula = TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid +
##     ResidualSugar + Chlorides + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + FreeSulfurDioxide_log + TotalSulfurDioxide_log +
##     STARS_merged, data = fullmod_regressiondata, init.theta = 48988.32099,
##     link = log)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.9717  -0.7205   0.0689   0.5772   3.2239
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             8.244e-01  2.512e-01   3.282 0.001032 **
## FixedAcidity           -2.882e-04  8.205e-04  -0.351 0.725410
## VolatileAcidity        -3.345e-02  6.515e-03  -5.134 2.84e-07 ***
## CitricAcid              7.770e-03  5.892e-03   1.319 0.187294
## ResidualSugar           5.765e-05  1.547e-04   0.373 0.709355
## Chlorides              -4.156e-02  1.645e-02  -2.526 0.011528 *
## Density                -2.737e-01  1.920e-01  -1.426 0.153939
## pH                     -1.571e-02  7.637e-03  -2.058 0.039638 *
## Sulphates              -1.264e-02  5.749e-03  -2.198 0.027925 *
## Alcohol                 2.148e-03  1.410e-03   1.523 0.127700
## LabelAppeal             1.333e-01  6.063e-03  21.992  < 2e-16 ***
## AcidIndex              -8.721e-02  4.547e-03 -19.178  < 2e-16 ***
## FreeSulfurDioxide_log   4.710e-02  1.617e-02   2.913 0.003583 **
## TotalSulfurDioxide_log  6.020e-02  1.779e-02   3.384 0.000715 ***
## STARS_merged            3.112e-01  4.531e-03  68.696  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48988.32) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 14734  on 12780  degrees of freedom
## AIC: 46708
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48988
##           Std. Err.:  50753
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -46676.38
```

```
#Getting formula for the model
formula(model.forward.negbin)
```

```
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +
##     Chlorides + Density + pH + Sulphates + Alcohol + LabelAppeal +
##     AcidIndex + FreeSulfurDioxide_log + TotalSulfurDioxide_log +
##     STARS_merged
```

```
model.backward.negbin <-model.full.negbin %>% stepAIC(direction = "backward", trace = FALSE)
summary(model.backward.negbin)
```

```
##
## Call:
## glm.nb(formula = TARGET ~ VolatileAcidity + Chlorides + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + FreeSulfurDioxide_log +
##     TotalSulfurDioxide_log + STARS_merged, data = fullmod_regressiondata, 
##     init.theta = 48991.45877, link = log)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.9798  -0.7206   0.0697   0.5791   3.2269  
## 
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)    
## (Intercept)             0.824163   0.251180   3.281 0.001034 ** 
## VolatileAcidity        -0.033648   0.006514  -5.166  2.4e-07 ***
## Chlorides              -0.041714   0.016451  -2.536 0.011224 *  
## Density                -0.277545   0.191954  -1.446 0.148208    
## pH                     -0.015613   0.007635  -2.045 0.040871 *  
## Sulphates              -0.012783   0.005747  -2.224 0.026143 *  
## Alcohol                 0.002182   0.001409   1.548 0.121537    
## LabelAppeal             0.133392   0.006063  22.001  < 2e-16 ***
## AcidIndex              -0.087078   0.004491 -19.391  < 2e-16 ***
## FreeSulfurDioxide_log   0.047248   0.016169   2.922 0.003477 ** 
## TotalSulfurDioxide_log  0.060518   0.017784   3.403 0.000667 ***
## STARS_merged            0.311336   0.004530  68.732  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for Negative Binomial(48991.46) family taken to be 1)
## 
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 14736  on 12783  degrees of freedom
## AIC: 46704
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##               Theta:  48991 
##           Std. Err.:  50765 
## Warning while fitting theta: iteration limit reached 
## 
##  2 x log-likelihood:  -46678.37
```

```
#Getting formula for the model
formula(model.backward.negbin)
```

```
## TARGET ~ VolatileAcidity + Chlorides + Density + pH + Sulphates + 
##     Alcohol + LabelAppeal + AcidIndex + FreeSulfurDioxide_log + 
##     TotalSulfurDioxide_log + STARS_merged
```

SELECT THE MODELS

```
stargazer(model.full.mr, model.forward.poisson, model.forward.negbin, title="Results", align=TRUE)#, he
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Sun, Dec 12, 2021 - 5:23:31 PM % Requires LaTeX packages: dcolumn

Table 1: Results

	Dependent variable:		
	TARGET		
	OLS	*Poisson*	*negative binomial*
	(1)	(2)	(3)
FixedAcidity	0.00003	−0.0003	−0.0003
	(0.002)	(0.001)	(0.001)
VolatileAcidity	−0.099***	−0.033***	−0.033***
	(0.015)	(0.007)	(0.007)
CitricAcid	0.021	0.008	0.008
	(0.014)	(0.006)	(0.006)
ResidualSugar	0.0002	0.0001	0.0001
	(0.0004)	(0.0002)	(0.0002)
Chlorides	−0.125***	−0.042**	−0.042**
	(0.038)	(0.016)	(0.016)
Density	−0.783*	−0.274	−0.274
	(0.442)	(0.192)	(0.192)
pH	−0.034**	−0.016**	−0.016**
	(0.018)	(0.008)	(0.008)
Sulphates	−0.033**	−0.013**	−0.013**
	(0.013)	(0.006)	(0.006)
Alcohol	0.011***	0.002	0.002
	(0.003)	(0.001)	(0.001)
LabelAppeal	0.433***	0.133***	0.133***
	(0.014)	(0.006)	(0.006)
AcidIndex	−0.209***	−0.087***	−0.087***
	(0.009)	(0.005)	(0.005)
FreeSulfurDioxide_log	0.122***	0.047***	0.047***
	(0.037)	(0.016)	(0.016)
TotalSulfurDioxide_log	0.158***	0.060***	0.060***
	(0.039)	(0.018)	(0.018)
STARS_merged	0.977***	0.311***	0.311***
	(0.010)	(0.005)	(0.005)
Constant	2.161***	0.824***	0.824***
	(0.570)	(0.251)	(0.251)
Observations	12,795	12,795	12,795
R^2	0.528		
Adjusted R^2	0.527		
Log Likelihood		-23,338.050	-23,339.190
θ			48,988.320 (50,752.710)
Akaike Inf. Crit.		46,706.100	46,708.380

Predictions

```
predict1 <- predict(model.forward.mr, newdata=wine_eval_df, type="response")
summary(predict1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.6415  1.9324  2.9853  3.0564  4.0508  6.9920
```

```
write.csv(predict1, 'predict1.csv', row.names = FALSE)
```

```
predict2 <- predict(model.forward.poisson, newdata=wine_eval_df, type="response")
summary(predict2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7213  1.9224  2.6847  3.0491  3.7932 10.0915
```

```
write.csv(predict2, 'predict2.csv', row.names = FALSE)
```

```
predict3 <- predict(model.forward.negbin, newdata=wine_eval_df, type="response")
summary(predict3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7213  1.9224  2.6847  3.0491  3.7932 10.0916
```

```
write.csv(predict3, 'predict3.csv', row.names = FALSE)
```