

DATA621__Homework5

Jenny

December 11, 2018

```
install.packages("DataExplorer") install.packages("corrplot") install.packages("leaps") install.packages("mlogit")
install.packages("lattice")
```

```
library(ggplot2)
library(stringr)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(DataExplorer)
library(leaps)
library(mlogit)
```

```
## Loading required package: Formula
```

```
## Loading required package: maxLik
```

```
## Loading required package: miscTools
```

```
##
```

```
## Please cite the 'maxLik' package as:
```

```
## Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maximum likelihood estimation in R. C
```

```
##
```

```
## If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a forum o
```

```
## https://r-forge.r-project.org/projects/maxlik/
```

```
library(reshape2)
```

Introduction

In this homework, I will explore, analyze and model a data set containing 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.

My 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.

Read the data set

```
train_data <- read.csv("https://raw.githubusercontent.com/JennierJ/DATA621/master/Homework5/wine-training")
eval_data <- read.csv("https://raw.githubusercontent.com/JennierJ/DATA621/master/Homework5/wine-evaluation")
```

Data Exploration and Cleaning

```
head(train_data)
```

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

```
summary(train_data)
```

```
##      i..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
```

```
nrow(train_data)
```

```
## [1] 12795
```

```
train_data <- train_data[, -1]
```

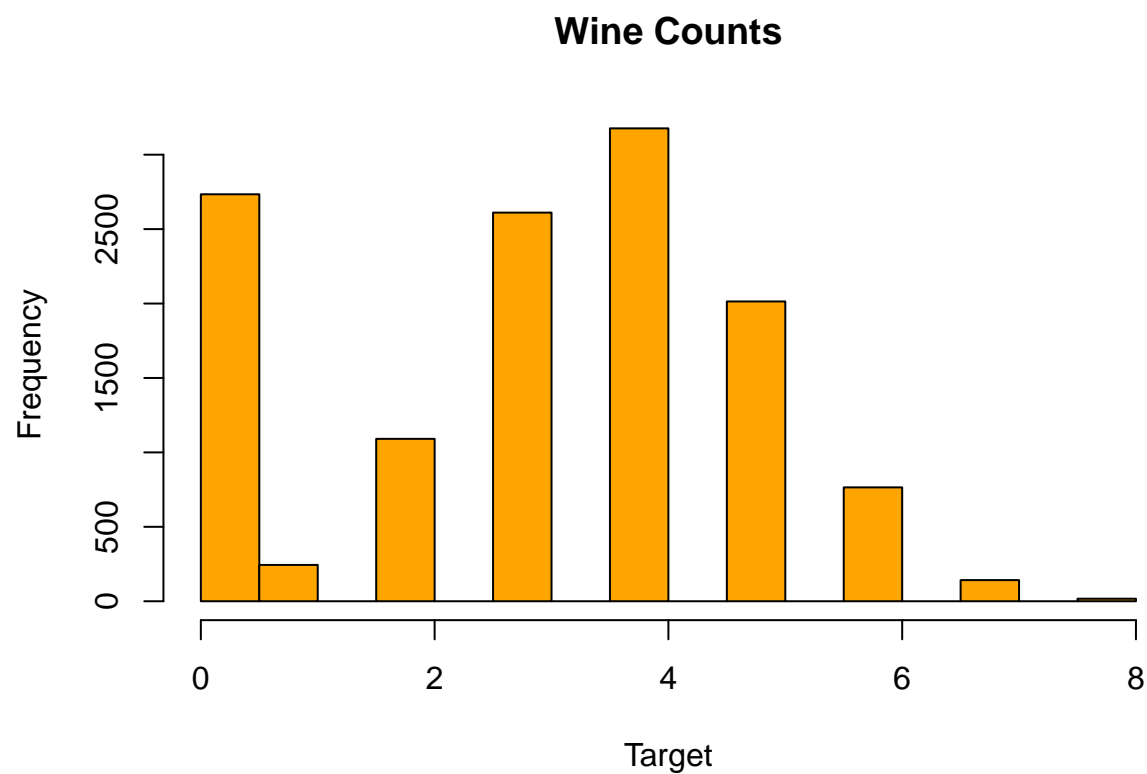
```
eval_data <- eval_data[, -1]
```

```
summary(eval_data)
```

```
##   TARGET      FixedAcidity  VolatileAcidity  CitricAcid
## Mode:logical  Min.      :-18.200    Min.      :-2.8300    Min.      :-3.1200
## NA's:3335     1st Qu.:  5.200    1st Qu.: 0.0800    1st Qu.: 0.0000
##              Median :  6.900    Median : 0.2800    Median : 0.3100
##              Mean   :  6.864    Mean   : 0.3103    Mean   : 0.3124
##              3rd Qu.:  9.000    3rd Qu.: 0.6300    3rd Qu.: 0.6050
##              Max.   : 33.500    Max.   : 3.6100    Max.   : 3.7600
##
## ResidualSugar      Chlorides      FreeSulfurDioxide
## Min.      :-128.300    Min.      :-1.15000    Min.      :-563.00
## 1st Qu.:  -2.600    1st Qu.: 0.01600    1st Qu.:  3.00
## Median :   3.600    Median : 0.04700    Median : 30.00
## Mean   :   5.319    Mean   : 0.06143    Mean   : 34.95
## 3rd Qu.:  17.200    3rd Qu.: 0.17100    3rd Qu.: 79.25
## 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
```

Histograms

```
hist(train_data$TARGET, col = "orange", xlab = " Target ", main = "Wine Counts")
```

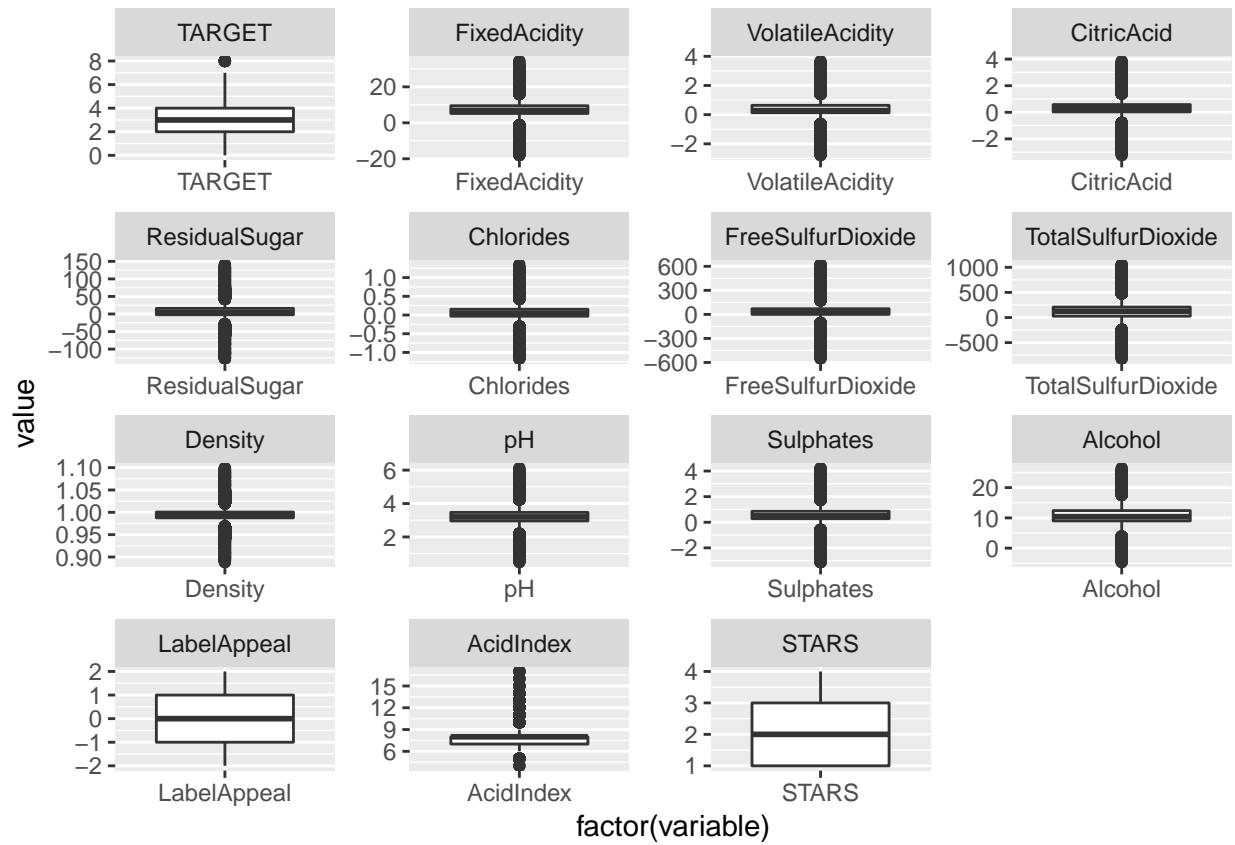


Boxplot

```
ggplot(melt(train_data), aes(x=factor(variable), y=value)) + facet_wrap(~variable, scale="free") + geom_boxplot()
```

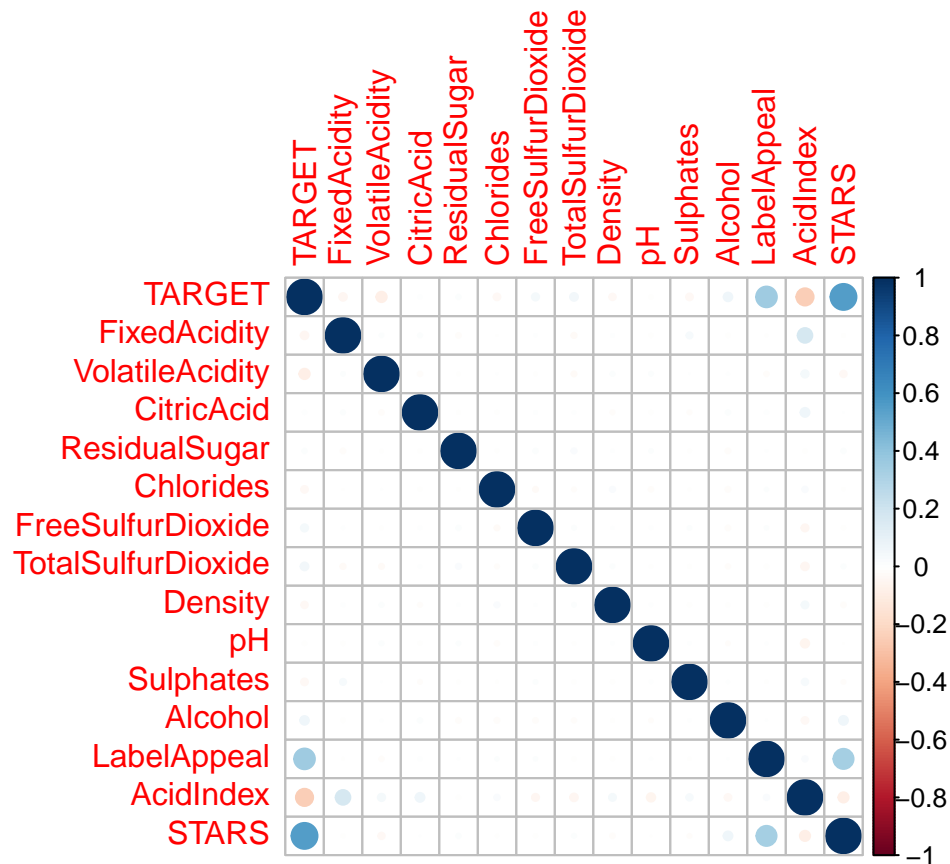
```
## No id variables; using all as measure variables
```

```
## Warning: Removed 8200 rows containing non-finite values (stat_boxplot).
```



Correlation

```
corrplot(as.matrix(cor(train_data, use = "pairwise.complete")),method = "circle")
```



Data Preparation

Replacing N/A with mean for calculation

```
train_data$ResidualSugar[is.na(train_data$ResidualSugar)] <- mean(train_data$ResidualSugar, na.rm=TRUE)
train_data$Chlorides[is.na(train_data$Chlorides)] <- mean(train_data$Chlorides, na.rm=TRUE)
train_data$FreeSulfurDioxide[is.na(train_data$FreeSulfurDioxide)] <- mean(train_data$FreeSulfurDioxide, na.rm=TRUE)
train_data$TotalSulfurDioxide[is.na(train_data$TotalSulfurDioxide)] <- mean(train_data$TotalSulfurDioxide, na.rm=TRUE)
train_data$pH[is.na(train_data$pH)] <- mean(train_data$pH, na.rm=TRUE)
train_data$Sulphates[is.na(train_data$Sulphates)] <- mean(train_data$Sulphates, na.rm=TRUE)
train_data$Sulphates[is.na(train_data$Sulphates)] <- mean(train_data$Sulphates, na.rm=TRUE)
train_data$Alcohol[is.na(train_data$Alcohol)] <- mean(train_data$Alcohol, na.rm=TRUE)
train_data$STARS[is.na(train_data$STARS)] <- mean(train_data$STARS, na.rm=TRUE)
```

```
eval_data$ResidualSugar[is.na(eval_data$ResidualSugar)] <- mean(eval_data$ResidualSugar, na.rm=TRUE)
eval_data$Chlorides[is.na(eval_data$Chlorides)] <- mean(eval_data$Chlorides, na.rm=TRUE)
eval_data$FreeSulfurDioxide[is.na(eval_data$FreeSulfurDioxide)] <- mean(eval_data$FreeSulfurDioxide, na.rm=TRUE)
eval_data$TotalSulfurDioxide[is.na(eval_data$TotalSulfurDioxide)] <- mean(eval_data$TotalSulfurDioxide, na.rm=TRUE)
eval_data$pH[is.na(eval_data$pH)] <- mean(eval_data$pH, na.rm=TRUE)
eval_data$Sulphates[is.na(eval_data$Sulphates)] <- mean(eval_data$Sulphates, na.rm=TRUE)
eval_data$Alcohol[is.na(eval_data$Alcohol)] <- mean(eval_data$Alcohol, na.rm=TRUE)
eval_data$STARS[is.na(eval_data$STARS)] <- mean(eval_data$STARS, na.rm=TRUE)
```

Build Model

Poission Distribution

```
model1 = glm(TARGET ~., data = train_data, family = poisson)
summary(model1)
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = train_data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.5118  -0.5144   0.2080   0.6344   2.5664
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.045e+00  1.957e-01  10.448  < 2e-16 ***
## FixedAcidity   -4.444e-04  8.194e-04  -0.542  0.587585
## VolatileAcidity -5.097e-02  6.492e-03  -7.851  4.12e-15 ***
## CitricAcid      1.343e-02  5.892e-03   2.280  0.022631 *
## ResidualSugar   1.489e-04  1.545e-04   0.964  0.335073
## Chlorides       -6.058e-02  1.645e-02  -3.683  0.000231 ***
## FreeSulfurDioxide 1.420e-04  3.513e-05   4.041  5.31e-05 ***
## TotalSulfurDioxide 1.072e-04  2.268e-05   4.727  2.28e-06 ***
## Density        -4.364e-01  1.921e-01  -2.272  0.023074 *
## pH              -2.411e-02  7.639e-03  -3.156  0.001599 **
## Sulphates       -1.901e-02  5.738e-03  -3.313  0.000924 ***
## Alcohol         5.528e-03  1.410e-03   3.920  8.85e-05 ***
## LabelAppeal     1.996e-01  6.014e-03  33.180  < 2e-16 ***
## AcidIndex       -1.232e-01  4.461e-03 -27.616  < 2e-16 ***
## STARS           2.113e-01  6.491e-03  32.543  < 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: 18511  on 12780  degrees of freedom
## AIC: 50483
##
## Number of Fisher Scoring iterations: 5
```

Exclude the predictor that are not significant

```
model2 = glm(TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide
              + TotalSulfurDioxide + Density + pH + Sulphates + Alcohol + LabelAppeal
              + AcidIndex + STARS, data = train_data,
              family = poisson)
summary(model2)
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##      FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##      Alcohol + LabelAppeal + AcidIndex + STARS, family = poisson,
```

```

##      data = train_data)
##
## Deviance Residuals:
##      Min        1Q      Median        3Q        Max
## -3.5135  -0.5169   0.2084   0.6350   2.5697
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.045e+00  1.957e-01  10.451 < 2e-16 ***
## VolatileAcidity -5.106e-02  6.491e-03  -7.866 3.68e-15 ***
## CitricAcid      1.336e-02  5.892e-03   2.267 0.023405 *
## Chlorides      -6.056e-02  1.645e-02  -3.681 0.000232 ***
## FreeSulfurDioxide 1.422e-04  3.513e-05   4.047 5.19e-05 ***
## TotalSulfurDioxide 1.078e-04  2.268e-05   4.753 2.00e-06 ***
## Density       -4.366e-01  1.921e-01  -2.273 0.023025 *
## pH            -2.401e-02  7.637e-03  -3.143 0.001671 **
## Sulphates     -1.910e-02  5.737e-03  -3.328 0.000873 ***
## Alcohol        5.503e-03  1.410e-03   3.904 9.47e-05 ***
## LabelAppeal    1.996e-01  6.014e-03  33.190 < 2e-16 ***
## AcidIndex     -1.236e-01  4.409e-03 -28.031 < 2e-16 ***
## STARS          2.113e-01  6.491e-03  32.554 < 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: 18513  on 12782  degrees of freedom
## AIC: 50481
##
## Number of Fisher Scoring iterations: 5

```

```

# Linera Model
model3 = lm(TARGET ~., data = train_data)
summary(model3)

```

```

##
## Call:
## lm(formula = TARGET ~ ., data = train_data)
##
## Residuals:
##      Min        1Q      Median        3Q        Max
## -5.2735  -0.7440   0.3694   1.1250   4.3210
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.451e+00  5.543e-01   9.833 < 2e-16 ***
## FixedAcidity   -1.152e-03  2.326e-03  -0.495 0.620510
## VolatileAcidity -1.564e-01  1.847e-02  -8.470 < 2e-16 ***
## CitricAcid      4.009e-02  1.681e-02   2.385 0.017085 *
## ResidualSugar    4.892e-04  4.392e-04   1.114 0.265349
## Chlorides      -1.946e-01  4.660e-02  -4.175 2.99e-05 ***
## FreeSulfurDioxide 4.284e-04  9.988e-05   4.289 1.81e-05 ***
## TotalSulfurDioxide 3.122e-04  6.417e-05   4.865 1.16e-06 ***
## Density       -1.289e+00  5.453e-01  -2.364 0.018098 *

```



```
## pH -6.458e-02 2.164e-02 -2.984 0.002850 **
## Sulphates -5.560e-02 1.631e-02 -3.409 0.000654 ***
## Alcohol 1.929e-02 3.991e-03 4.832 1.37e-06 ***
## LabelAppeal 6.042e-01 1.693e-02 35.684 < 2e-16 ***
## AcidIndex -3.290e-01 1.122e-02 -29.313 < 2e-16 ***
## STARS 7.153e-01 1.953e-02 36.617 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.634 on 12780 degrees of freedom
## Multiple R-squared: 0.2811, Adjusted R-squared: 0.2803
## F-statistic: 356.9 on 14 and 12780 DF, p-value: < 2.2e-16
```

Model Selection

```
predict1 <- predict(model2, newdata=eval_data, type="response")
summary(predict1)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.8397  2.3207  2.8487  3.0524  3.5567  9.5182
```

```
predict2 <- predict(model3, newdata=eval_data, type="response")
summary(predict2)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.3319  2.3733  2.9959  3.0525  3.6876  6.6712
```