DS621_Group2_HW5_JC

Jagdish Chhabria

5/10/2021

Group 2 members: *Diego Correa, Jagdish Chhabria, Orli Khaimova, Richard Zheng, Stephen Haslett.*

Assignment 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: VARIABLE

Task 1: Data Exploration

Describe the size and the variables in the wine training data set.

```
## 'data.frame':    12795 obs. of  16 variables:
##  $ ï..INDEX          : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET            : int  3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity      : num  3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity   : num  1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
##  $ CitricAcid        : num  -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
##  $ ResidualSugar     : num  54.2 26.1 14.8 18.8 9.4 ...
##  $ Chlorides         : num  -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...
##  $ FreeSulfurDioxide : num  NA 15 214 22 -167 -37 287 523 -213 62 ...
##  $ TotalSulfurDioxide: num  268 -327 142 115 108 15 156 551 NA 180 ...
##  $ Density           : num  0.993 1.028 0.995 0.996 0.995 ...
##  $ pH                : num  3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
##  $ Sulphates         : num  -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...
##  $ Alcohol           : num  9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
##  $ LabelAppeal       : int  0 -1 -1 -1 0 0 0 1 0 0 ...
##  $ AcidIndex         : int  8 7 8 6 9 11 8 7 6 8 ...
##  $ STARS             : int  2 3 3 1 2 NA NA 3 NA 4 ...
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

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

```
# Remove the index variable
training_set<-training_set%>%dplyr::select(-INDEX)
#%>%mutate(TARGET=as.factor(TARGET))
#training_set<-training_set%>%dplyr::mutate(TARGET=as.factor(TARGET))
```

```
# Check the structure of the training dataset.
str(training_set)
```

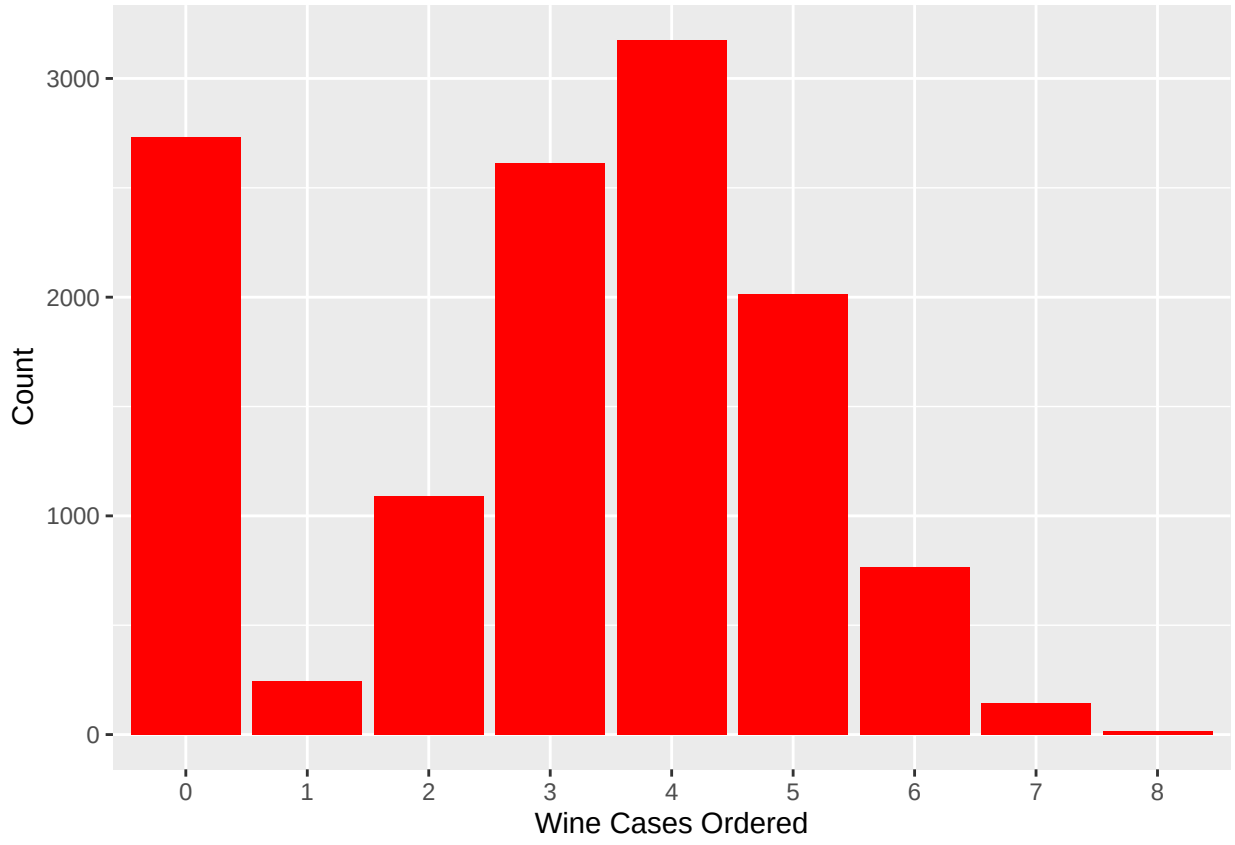
```
## 'data.frame':    12795 obs. of  15 variables:
##  $ TARGET            : int  3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity      : num  3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity   : num  1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
##  $ CitricAcid        : num  -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
##  $ ResidualSugar     : num  54.2 26.1 14.8 18.8 9.4 ...
##  $ Chlorides         : num  -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...
##  $ FreeSulfurDioxide : num  NA 15 214 22 -167 -37 287 523 -213 62 ...
##  $ TotalSulfurDioxide: num  268 -327 142 115 108 15 156 551 NA 180 ...
##  $ Density           : num  0.993 1.028 0.995 0.996 0.995 ...
##  $ pH                : num  3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
##  $ Sulphates         : num  -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...
##  $ Alcohol           : num  9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
##  $ LabelAppeal       : int  0 -1 -1 -1 0 0 0 1 0 0 ...
##  $ AcidIndex         : int  8 7 8 6 9 11 8 7 6 8 ...
##  $ STARS             : int  2 3 3 1 2 NA NA 3 NA 4 ...
```

```
describe(training_set)
```

```
##                      vars     n   mean     sd median trimmed    mad     min
## TARGET                  1 12795   3.03   1.93   3.00    3.05   1.48    0.00
## FixedAcidity            2 12795   7.08   6.32   6.90    7.07   3.26  -18.10
## VolatileAcidity         3 12795   0.32   0.78   0.28    0.32   0.43   -2.79
## CitricAcid              4 12795   0.31   0.86   0.31    0.31   0.42   -3.24
## ResidualSugar           5 12179   5.42  33.75   3.90    5.58  15.72 -127.80
## Chlorides               6 12157   0.05   0.32   0.05    0.05   0.13   -1.17
## FreeSulfurDioxide       7 12148  30.85 148.71  30.00   30.93  56.34 -555.00
## TotalSulfurDioxide      8 12113 120.71 231.91 123.00  120.89 134.92 -823.00
## Density                 9 12795   0.99   0.03   0.99    0.99   0.01    0.89
## pH                     10 12400   3.21   0.68   3.20    3.21   0.39    0.48
## Sulphates              11 11585   0.53   0.93   0.50    0.53   0.44   -3.13
## Alcohol                12 12142  10.49   3.73  10.40   10.50   2.37   -4.70
## LabelAppeal            13 12795  -0.01   0.89   0.00   -0.01   1.48   -2.00
## AcidIndex              14 12795   7.77   1.32   8.00    7.64   1.48    4.00
```

```
## STARS                15  9436   2.04   0.90   2.00    1.97   1.48    1.00
##                        max   range  skew kurtosis   se
## TARGET                8.00    8.00 -0.33    -0.88 0.02
## FixedAcidity         34.40   52.50 -0.02     1.67 0.06
## VolatileAcidity       3.68    6.47  0.02     1.83 0.01
## CitricAcid            3.86    7.10 -0.05     1.84 0.01
## ResidualSugar       141.15  268.95 -0.05     1.88 0.31
## Chlorides             1.35    2.52  0.03     1.79 0.00
## FreeSulfurDioxide   623.00 1178.00  0.01     1.84 1.35
## TotalSulfurDioxide 1057.00 1880.00 -0.01     1.67 2.11
## Density               1.10    0.21 -0.02     1.90 0.00
## pH                    6.13    5.65  0.04     1.65 0.01
## Sulphates             4.24    7.37  0.01     1.75 0.01
## Alcohol              26.50   31.20 -0.03     1.54 0.03
## LabelAppeal           2.00    4.00  0.01    -0.26 0.01
## AcidIndex            17.00   13.00  1.65     5.19 0.01
## STARS                 4.00    3.00  0.45    -0.69 0.01
```

```
# Plot the distribution of the TARGET variable
#ggplot(training_set,aes(x=training_set$TARGET))+geom_histogram()
wine.cases<-table(training_set$TARGET)%>%data.frame()
wine.cases%>%ggplot(aes(x=Var1,y=Freq))+geom_bar(stat="identity", fill="red")+ labs(x = "Wine Cases Ord
```



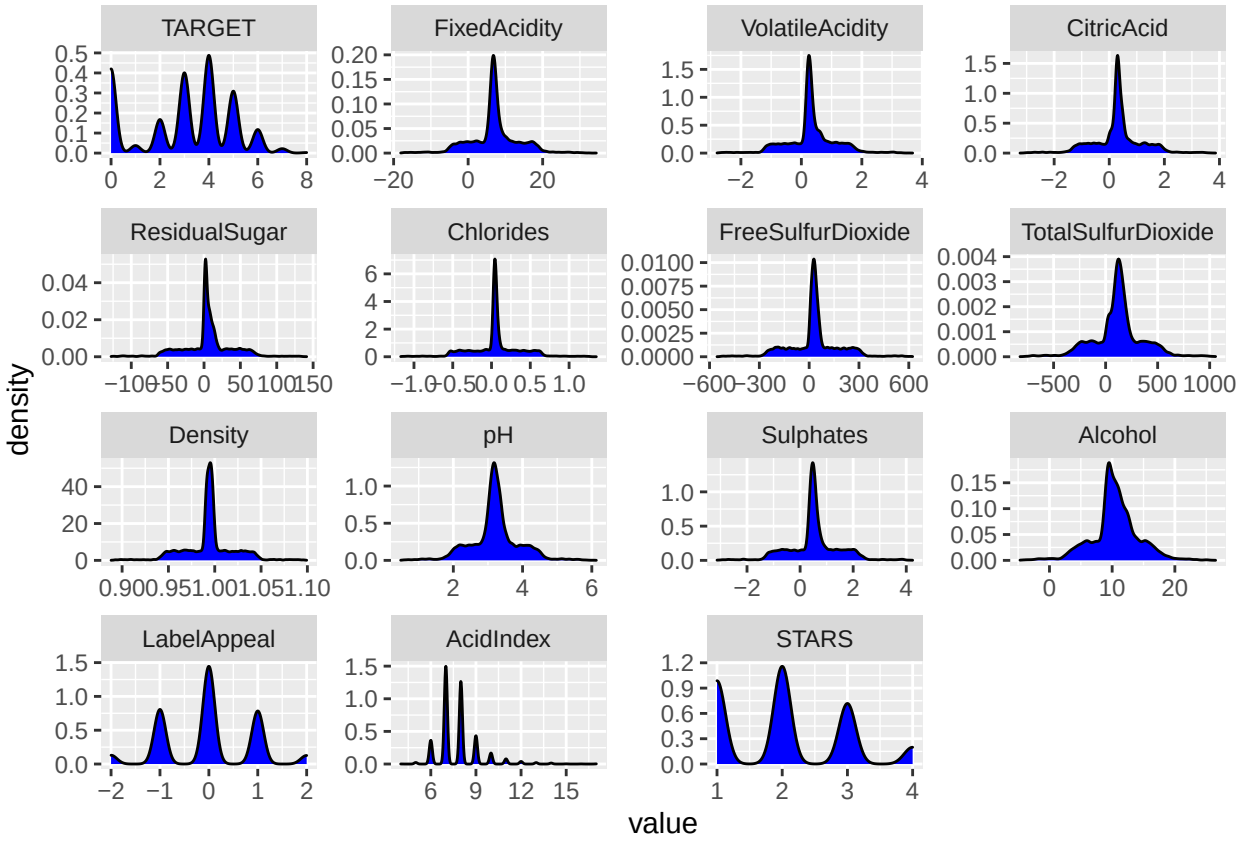

```
# Check count and proportion of 0 values for the TARGET variable
training_set%>%filter(TARGET==0)%>%summarise(n=n())%>%mutate(freq=round(n/nrow(training_set),4))
```

```
##      n   freq
## 1 2734 0.2137
```

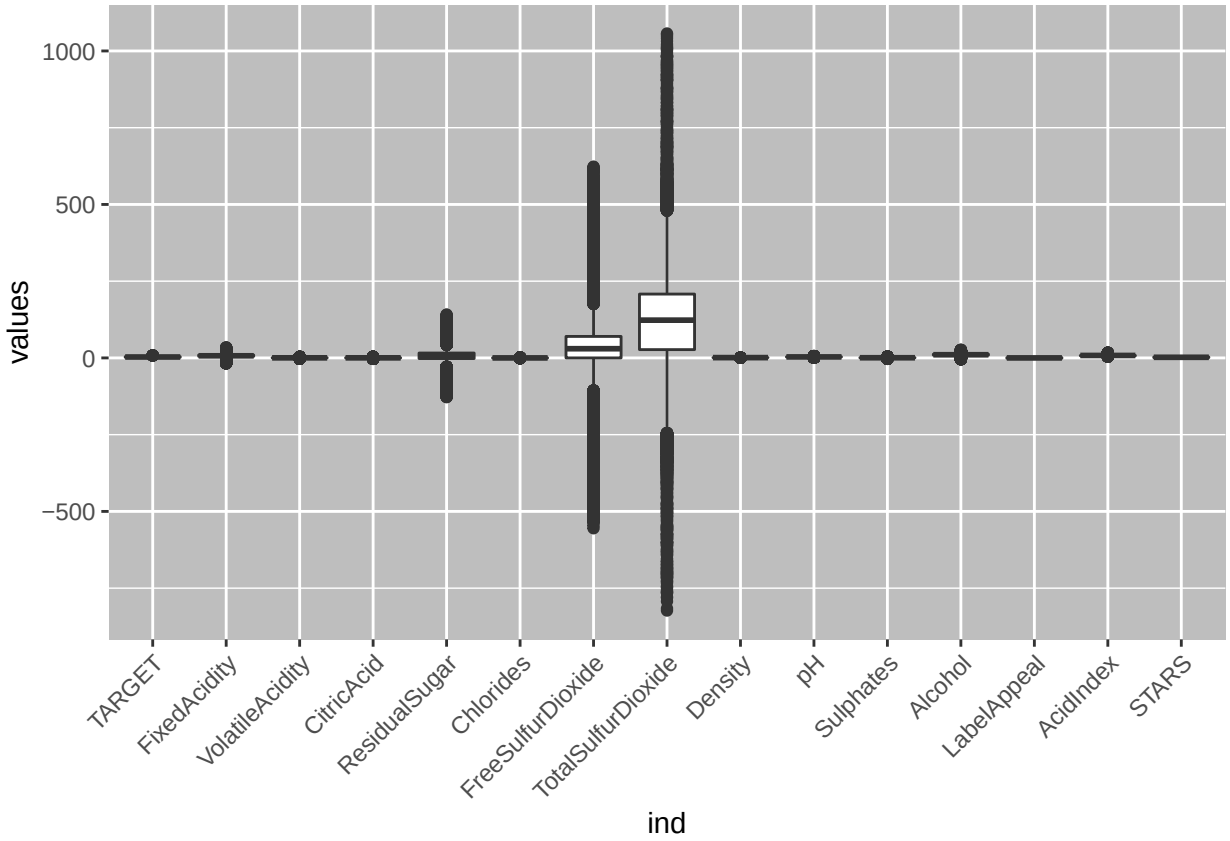
From the above, we can see that about 21% of the records have a count = 0. Given that more than a fifth of the target variable values are 0, this could be considered as a "zero-inflated" dataset.

```
# Check distributions for all the variables
melt(training_set)%>%ggplot(aes(x=value))+geom_density(fill='blue')+facet_wrap(~variable,scales='free')
```

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



From the above, we can see that 4 of the variables including the target variable are multi-modal, while the rest look leptokurtic. Given that we don't intend to use linear regression as the model, we will not attempt to transform the varaibles to make them more normally distributed.



From the above, we can see that these variables have significant outliers: TotalSulfurDioxide, FreeSulfurDioxide and ResidualSugar.

```
# Function to remove outliers
remove_outliers<-function(x) {
quant <- quantile(x, probs=c(.25, .75), na.rm = T)
cap <- quantile(x, probs=c(.05, .95), na.rm = T)
H <- 1.5 * IQR(x, na.rm = T)
x[x < (quant[1] - H)] <- cap[1]
x[x > (quant[2] + H)] <- cap[2]

return(x)
}
```

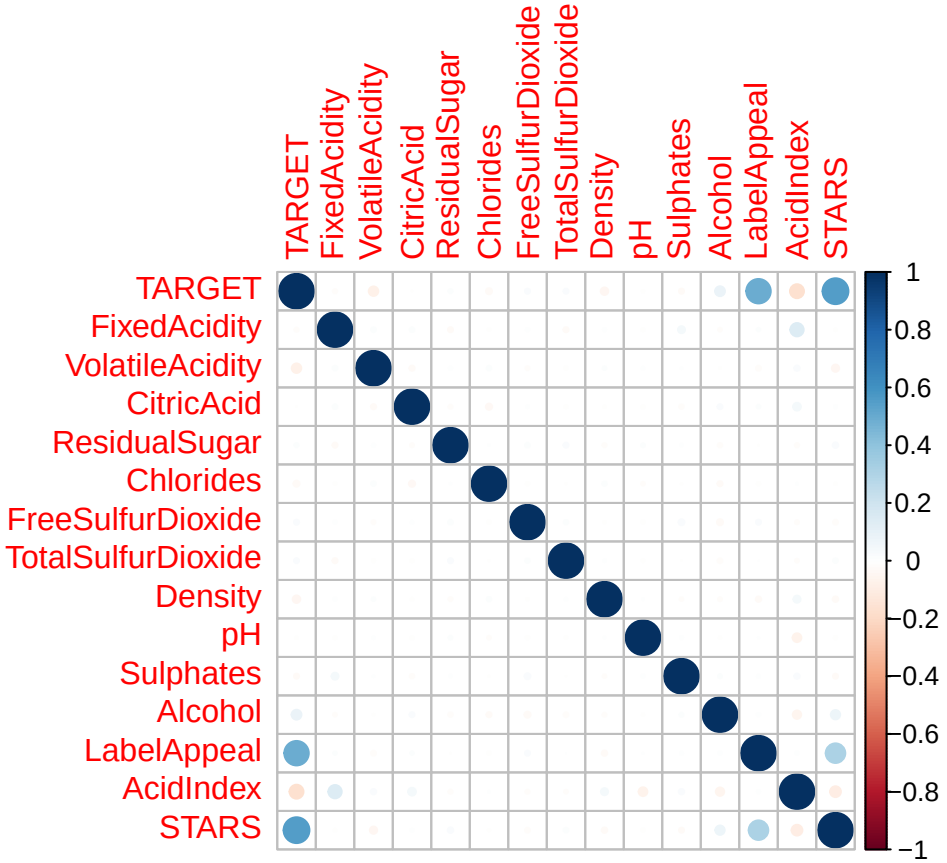
```
# Remove outliers from training data
training_set$FixedAcidity <- remove_outliers(training_set$FixedAcidity)
training_set$VolatileAcidity <- remove_outliers(training_set$VolatileAcidity)
training_set$CitricAcid <- remove_outliers(training_set$CitricAcid)
training_set$ResidualSugar <- remove_outliers(training_set$ResidualSugar)
training_set$Chlorides <- remove_outliers(training_set$Chlorides)
training_set$FreeSulfurDioxide <- remove_outliers(training_set$FreeSulfurDioxide)
training_set$TotalSulfurDioxide <- remove_outliers(training_set$TotalSulfurDioxide)
training_set$Density <- remove_outliers(training_set$Density)
training_set$pH <- remove_outliers(training_set$pH)
training_set$Sulphates <- remove_outliers(training_set$Sulphates)
training_set$Alcohol <- remove_outliers(training_set$Alcohol)
training_set$AcidIndex <- remove_outliers(training_set$AcidIndex)
```

	x
TARGET	1.0000000
STARS	0.5587938
LabelAppeal	0.3565005
Alcohol	0.0650611
TotalSulfurDioxide	0.0530651
FreeSulfurDioxide	0.0413681
ResidualSugar	0.0198635
CitricAcid	0.0120351
pH	-0.0107792
Density	-0.0315375
Chlorides	-0.0339048
Sulphates	-0.0394213
FixedAcidity	-0.0510757
VolatileAcidity	-0.0891214
AcidIndex	-0.2353997

```
# Remove outliers from evaluation data
evaluation_set$FixedAcidity <- remove_outliers(evaluation_set$FixedAcidity)
evaluation_set$VolatileAcidity <- remove_outliers(evaluation_set$VolatileAcidity)
evaluation_set$CitricAcid <- remove_outliers(evaluation_set$CitricAcid)
evaluation_set$ResidualSugar <- remove_outliers(evaluation_set$ResidualSugar)
evaluation_set$Chlorides <- remove_outliers(evaluation_set$Chlorides)
evaluation_set$FreeSulfurDioxide <- remove_outliers(evaluation_set$FreeSulfurDioxide)
evaluation_set$TotalSulfurDioxide <- remove_outliers(evaluation_set$TotalSulfurDioxide)
evaluation_set$Density <- remove_outliers(evaluation_set$Density)
evaluation_set$pH <- remove_outliers(evaluation_set$pH)
evaluation_set$Sulphates <- remove_outliers(evaluation_set$Sulphates)
evaluation_set$Alcohol <- remove_outliers(evaluation_set$Alcohol)
evaluation_set$AcidIndex <- remove_outliers(evaluation_set$AcidIndex)
```

From the above, we can see that the STARS, LabelAppeal and AcidIndex variables are strogly correlated with the TARGET variable.

	x
STARS	3359
Sulphates	1210
TotalSulfurDioxide	682
Alcohol	653
FreeSulfurDioxide	647
Chlorides	638
ResidualSugar	616
pH	395
TARGET	0
FixedAcidity	0
VolatileAcidity	0
CitricAcid	0
Density	0
LabelAppeal	0
AcidIndex	0

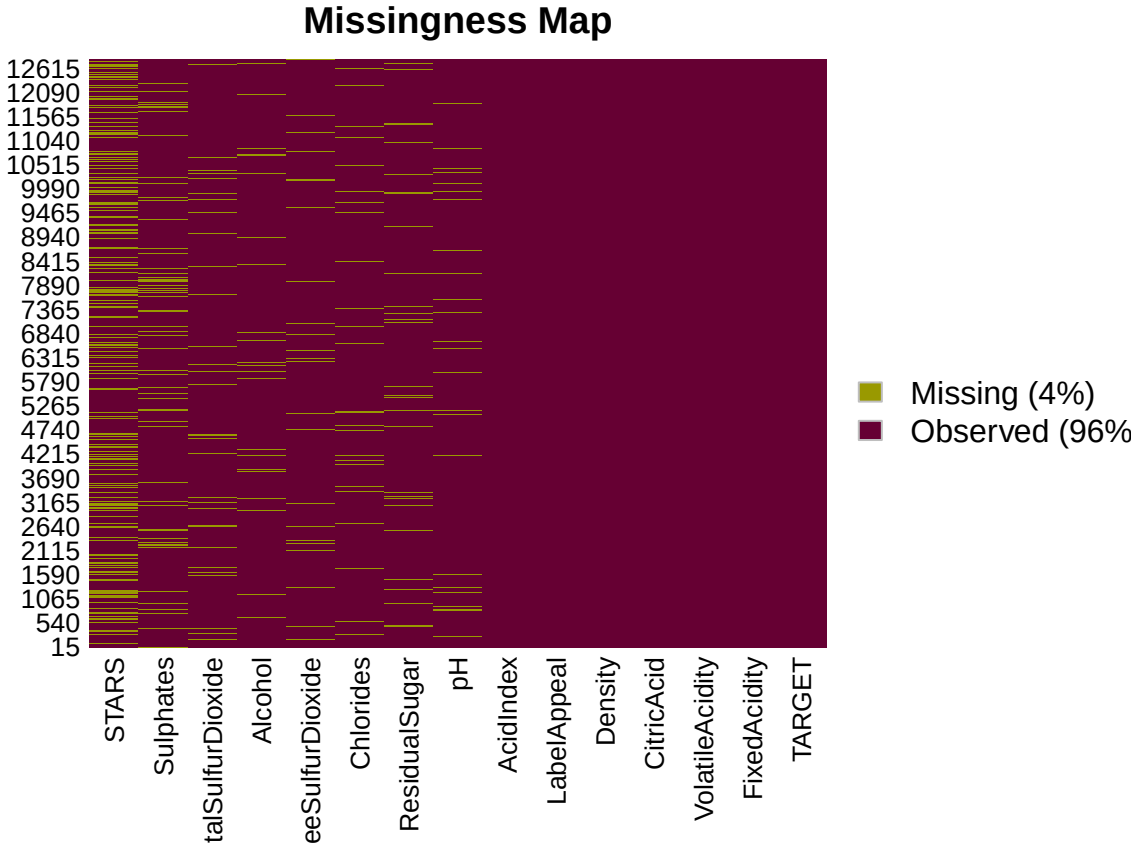


Missing Values

We check for missing values for each of the variables

We plot the % of missing values below.

	x
TARGET	3335
STARS	841
Sulphates	310
Alcohol	185
ResidualSugar	168
TotalSulfurDioxide	157
FreeSulfurDioxide	152
Chlorides	138
pH	104
IN	0
FixedAcidity	0
VolatileAcidity	0
CitricAcid	0
Density	0
LabelAppeal	0
AcidIndex	0



We impute the missing values using the predictive mean matching algorithm from the mice library.

We now use re-check for missing values.

We impute the missing values using the predictive mean matching algorithm from the mice library.

We now use re-check for missing values.

	x
TARGET	0
FixedAcidity	0
VolatileAcidity	0
CitricAcid	0
ResidualSugar	0
Chlorides	0
FreeSulfurDioxide	0
TotalSulfurDioxide	0
Density	0
pH	0
Sulphates	0
Alcohol	0
LabelAppeal	0
AcidIndex	0
STARS	0

	x
TARGET	3335
IN	0
FixedAcidity	0
VolatileAcidity	0
CitricAcid	0
ResidualSugar	0
Chlorides	0
FreeSulfurDioxide	0
TotalSulfurDioxide	0
Density	0
pH	0
Sulphates	0
Alcohol	0
LabelAppeal	0
AcidIndex	0
STARS	0

Model Building

For modeling count variables, the typical choices of model are:

1) Poisson regression: This is often used for modeling count data because it fits the framework.
2) Negative binomial regression: This can be used for over-dispersed count data i.e. when the conditional variance exceeds the conditional mean. It can be considered as a generalization of Poisson regression since it has the same mean structure as Poisson regression and it has an extra parameter to model the over-dispersion.
3) Zero-inflated regression model: This model attempts to handle the excess zeros problem. Two kinds of zeros can exist in the data: "true zeros" and "excess zeros". Zero-inflated models estimate two equations simultaneously, one for the count model and one for the excess zeros.

When it comes to count variables, the Poisson regression model (or one of its variants) have a number of advantages over an ordinary linear regression model, including a skew, discrete distribution, and the restriction of predicted values to non-negative numbers. A Poisson model is similar to an ordinary linear regression, with two exceptions. First, it assumes that the errors follow a Poisson, not a normal, distribution. Second, rather than modeling Y as a linear function of the regression coefficients, it models the natural log of the response variable, ln(Y), as a linear function of the coefficients. The Poisson model assumes that the mean and variance of the errors are equal. But usually in practice the variance of the errors is larger than the mean (although it can also be smaller). When the variance is larger than the mean, there are two extensions of the Poisson model that work well. In the over-dispersed Poisson model, an extra parameter is included which estimates how much larger the variance is than the mean. This parameter estimate is then used to correct for the effects of the larger variance on the p-values. An alternative is a negative binomial model.

```
# Fit the Poisson model for the count variable, with all predictors included
summary(model1 <- glm(TARGET~., family="poisson", data=training_set_imputed))
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = "poisson", data = training_set_imputed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9269  -0.6750   0.1287   0.6347   2.4535
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.586e+00  2.001e-01   7.925 2.28e-15 ***
## FixedAcidity        -9.671e-04  9.191e-04  -1.052 0.292666
## VolatileAcidity     -4.308e-02  7.248e-03  -5.944 2.78e-09 ***
## CitricAcid           1.424e-02  6.529e-03   2.181 0.029211 *
## ResidualSugar        1.130e-04  1.585e-04   0.713 0.475715
## Chlorides           -4.102e-02  1.716e-02  -2.391 0.016811 *
## FreeSulfurDioxide    1.437e-04  3.469e-05   4.144 3.41e-05 ***
## TotalSulfurDioxide   9.889e-05  2.546e-05   3.884 0.000103 ***
## Density             -3.526e-01  1.950e-01  -1.808 0.070642 .
## pH                  -2.181e-02  8.601e-03  -2.536 0.011228 *
## Sulphates           -1.608e-02  5.981e-03  -2.688 0.007182 **
## Alcohol              2.807e-03  1.585e-03   1.772 0.076462 .
## LabelAppeal          1.438e-01  6.073e-03  23.686  < 2e-16 ***
## AcidIndex           -1.035e-01  5.203e-03 -19.900  < 2e-16 ***
```

```
## STARS               3.415e-01  5.602e-03  60.965  < 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: 16067  on 12780  degrees of freedom
## AIC: 48039
##
## Number of Fisher Scoring iterations: 5
```

Deviance residuals are approximately normally distributed if the model is specified correctly. From the results above, we can see that there is some skeweness in the deviance residuals since median is not quite zero (it is 0.06). Next we examine the Poisson regression coefficients for each of the variables along with the standard errors, z-scores, p-values and 95% confidence intervals for the coefficients. The coefficient for the Alcohol variable is 0.0028. This means that the expected log count for a one-unit increase in Alcohol is .0028.

Based on the p-values above, it looks like the following predictors have a significant impact on the number of wine cases ordered: VolatileAcidity, Alcohol, LabelAppeal, AcidIndex, STARS. All the co-efficients are very small though.

```
# Forward step through this model to find the best predictors
model1.5 <- stepAIC(model1, trace = F)
summary(model1.5)
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS, family = "poisson",
##     data = training_set_imputed)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9270  -0.6783   0.1282   0.6317   2.4737
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.587e+00  2.001e-01   7.930 2.19e-15 ***
## VolatileAcidity     -4.318e-02  7.247e-03  -5.959 2.54e-09 ***
## CitricAcid           1.414e-02  6.529e-03   2.165  0.03035 *
## Chlorides           -4.106e-02  1.716e-02  -2.393  0.01670 *
## FreeSulfurDioxide    1.437e-04  3.468e-05   4.145 3.40e-05 ***
## TotalSulfurDioxide   9.969e-05  2.545e-05   3.916 8.99e-05 ***
## Density             -3.539e-01  1.950e-01  -1.815  0.06956 .
## pH                  -2.170e-02  8.600e-03  -2.524  0.01162 *
## Sulphates           -1.620e-02  5.980e-03  -2.709  0.00676 **
## Alcohol              2.789e-03  1.584e-03   1.760  0.07835 .
## LabelAppeal          1.439e-01  6.073e-03  23.698  < 2e-16 ***
## AcidIndex           -1.043e-01  5.145e-03 -20.279  < 2e-16 ***
## STARS                3.416e-01  5.602e-03  60.984  < 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: 16068  on 12782  degrees of freedom
## AIC: 48036
##
## Number of Fisher Scoring iterations: 5
```

The forward step algorithm shows that the best fit model results in 2 of the predictors being discarded: FixedAcidity and Residualsugar.

```
# Check for dispersion with this model
dispersiontest(model1.5,trafo=1)
```

```
##
##  Overdispersion test
##
## data:  model1.5
## z = -9.4259, p-value = 1
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##      alpha
## -0.1047475
```

From the above results, it seems that there is underdispersion in the data, since $c < 0$.

```
# Fit the Negative Binomial model for the count variable, with all predictors included
summary(model2<-glm.nb(TARGET~.,data=training_set_imputed))
```

```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = training_set_imputed, init.theta = 48882.01821,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9269  -0.6749   0.1287   0.6347   2.4534
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.586e+00  2.001e-01   7.925 2.29e-15 ***
## FixedAcidity        -9.672e-04  9.191e-04  -1.052 0.292675
## VolatileAcidity     -4.308e-02  7.248e-03  -5.944 2.78e-09 ***
## CitricAcid           1.424e-02  6.529e-03   2.181 0.029216 *
## ResidualSugar        1.130e-04  1.585e-04   0.713 0.475708
## Chlorides           -4.102e-02  1.716e-02  -2.391 0.016812 *
## FreeSulfurDioxide    1.437e-04  3.469e-05   4.144 3.41e-05 ***
## TotalSulfurDioxide   9.889e-05  2.546e-05   3.884 0.000103 ***
## Density             -3.526e-01  1.950e-01  -1.808 0.070647 .
## pH                  -2.181e-02  8.602e-03  -2.536 0.011228 *
```

```
## Sulphates          -1.608e-02  5.981e-03  -2.688 0.007182 **
## Alcohol             2.807e-03  1.585e-03   1.772 0.076477 .
## LabelAppeal         1.438e-01  6.073e-03  23.685  < 2e-16 ***
## AcidIndex          -1.035e-01  5.203e-03 -19.900  < 2e-16 ***
## STARS               3.415e-01  5.603e-03  60.963  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48882.02) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 16066  on 12780  degrees of freedom
## AIC: 48041
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48882
##           Std. Err.:  56641
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -48008.87
```

```
# Forward step through this model to find the best predictors
model2.5 <- stepAIC(model2, trace = F)
summary(model2.5)
```

```
##
## Call:
## glm.nb(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS, data = training_set_imputed,
##     init.theta = 48891.91523, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9269  -0.6782   0.1282   0.6317   2.4736
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.587e+00  2.001e-01   7.930 2.20e-15 ***
## VolatileAcidity    -4.318e-02  7.247e-03  -5.959 2.54e-09 ***
## CitricAcid          1.414e-02  6.529e-03   2.165  0.03036 *
## Chlorides          -4.106e-02  1.716e-02  -2.393  0.01670 *
## FreeSulfurDioxide   1.438e-04  3.468e-05   4.145 3.40e-05 ***
## TotalSulfurDioxide  9.969e-05  2.546e-05   3.916 8.99e-05 ***
## Density            -3.540e-01  1.950e-01  -1.815  0.06956 .
## pH                 -2.170e-02  8.600e-03  -2.524  0.01162 *
## Sulphates          -1.620e-02  5.980e-03  -2.709  0.00676 **
## Alcohol             2.789e-03  1.584e-03   1.760  0.07836 .
## LabelAppeal         1.439e-01  6.073e-03  23.697  < 2e-16 ***
## AcidIndex          -1.043e-01  5.145e-03 -20.279  < 2e-16 ***
## STARS               3.416e-01  5.602e-03  60.982  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(48891.92) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 16068  on 12782  degrees of freedom
## AIC: 48039
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  48892
##           Std. Err.:  56661
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -48010.51
```

```
# Check for overdispersion with this model
odTest(model2.5)
```

```
## Likelihood ratio test of H0: Poisson, as restricted NB model:
## n.b., the distribution of the test-statistic under H0 is non-standard
## e.g., see help(odTest) for details/references
##
## Critical value of test statistic at the alpha= 0.05 level: 2.7055
## Chi-Square Test Statistic =  -0.2206 p-value = 0.5
```

Based on the above test statistic value, we fail to reject the Null Hypothesis which states that the Poisson model is better suited for this dataset. So we stick with the Poisson model instead of the Negative Binomial model.

```
# Fit a zero-inflation poisson model with all predictors included
summary(model3<-zeroinfl(TARGET~.,data=training_set,dist="poisson"))
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ ., data = training_set, dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.28979 -0.29061  0.04278  0.34137  2.79555
##
## Count model coefficients (poisson with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.463e+00  2.597e-01   5.632 1.78e-08 ***
## FixedAcidity         3.283e-04  1.197e-03   0.274 0.783841
## VolatileAcidity     -1.208e-02  9.451e-03  -1.278 0.201169
## CitricAcid          -1.510e-03  8.551e-03  -0.177 0.859831
## ResidualSugar       -1.489e-04  2.078e-04  -0.717 0.473574
## Chlorides           -2.618e-02  2.260e-02  -1.158 0.246677
## FreeSulfurDioxide    2.466e-05  4.520e-05   0.546 0.585331
## TotalSulfurDioxide -2.499e-05  3.277e-05  -0.763 0.445693
## Density             -3.019e-01  2.530e-01  -1.193 0.232809
```

```
## pH                   4.326e-03  1.124e-02   0.385 0.700343
## Sulphates            1.453e-03  7.852e-03   0.185 0.853198
## Alcohol              7.246e-03  2.077e-03   3.488 0.000486 ***
## LabelAppeal          2.104e-01  8.161e-03  25.779  < 2e-16 ***
## AcidIndex           -2.169e-02  7.079e-03  -3.063 0.002189 **
## STARS                1.124e-01  7.923e-03  14.186  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -7.0149200  2.9041332  -2.415  0.01571 *
## FixedAcidity       -0.0030779  0.0130603  -0.236  0.81369
## VolatileAcidity     0.2940584  0.1027079   2.863  0.00420 **
## CitricAcid          0.0187661  0.0933737   0.201  0.84072
## ResidualSugar      -0.0054546  0.0022460  -2.429  0.01516 *
## Chlorides           0.1414853  0.2547292   0.555  0.57860
## FreeSulfurDioxide  -0.0010174  0.0005165  -1.970  0.04885 *
## TotalSulfurDioxide -0.0011826  0.0003636  -3.253  0.00114 **
## Density             2.7260489  2.7846957   0.979  0.32761
## pH                  0.2963530  0.1272129   2.330  0.01983 *
## Sulphates           0.2401191  0.0890932   2.695  0.00704 **
## Alcohol             0.0471531  0.0227282   2.075  0.03802 *
## LabelAppeal         0.7465259  0.0911853   8.187 2.68e-16 ***
## AcidIndex           0.6268839  0.0707353   8.862  < 2e-16 ***
## STARS              -3.7833212  0.3773277 -10.027  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 40
## Log-likelihood: -1.114e+04 on 30 Df
```

```
# Fit the Poisson model for the count variable, with selected predictors only
summary(model4<-glm(TARGET~VolatileAcidity+Alcohol+LabelAppeal+AcidIndex+STARS, family="poisson", data=
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + Alcohol + LabelAppeal +
##     AcidIndex + STARS, family = "poisson", data = training_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2229  -0.2696   0.0689   0.3729   1.6675
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.231366   0.050512  24.378  < 2e-16 ***
## VolatileAcidity -0.026470   0.007862  -3.367 0.000761 ***
## Alcohol          0.005454   0.001724   3.163 0.001562 **
## LabelAppeal      0.181030   0.006714  26.964  < 2e-16 ***
## AcidIndex       -0.053081   0.005700  -9.312  < 2e-16 ***
## STARS            0.185238   0.006311  29.352  < 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: 8176.2  on 8962  degrees of freedom
## Residual deviance: 5579.8  on 8957  degrees of freedom
##   (3832 observations deleted due to missingness)
## AIC: 32260
##
## Number of Fisher Scoring iterations: 5
```

```
# Fit a zero-inflation poisson model with selected predictors only
summary(model5<-zeroinfl(TARGET~VolatileAcidity+Alcohol+LabelAppeal+AcidIndex+STARS,data=training_set,di
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ VolatileAcidity + Alcohol + LabelAppeal +
##     AcidIndex + STARS, data = training_set, dist = "poisson")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.32380 -0.28908  0.04665  0.34499  2.21850
##
## Count model coefficients (poisson with log link):
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.178035   0.051868  22.712  < 2e-16 ***
## VolatileAcidity -0.010714   0.008013  -1.337    0.181
## Alcohol           0.007937   0.001755   4.522 6.13e-06 ***
## LabelAppeal       0.212394   0.006872  30.906  < 2e-16 ***
## AcidIndex        -0.023077   0.005919  -3.899 9.67e-05 ***
## STARS             0.112028   0.006669  16.798  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -3.23899    0.63776  -5.079  3.8e-07 ***
## VolatileAcidity  0.33099    0.08610   3.844 0.000121 ***
## Alcohol          0.04803    0.01870   2.568 0.010219 *
## LabelAppeal      0.70979    0.07759   9.148  < 2e-16 ***
## AcidIndex        0.61347    0.05786  10.603  < 2e-16 ***
## STARS           -3.84393    0.36139 -10.637  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 20
## Log-likelihood: -1.556e+04 on 12 Df
```

```
# Perform a vuong test to compare model 4 and model 5
vuong(model4, model5)
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A    p-value
## Raw                  -14.49454 model2 > model1 < 2.22e-16
## AIC-corrected        -14.34069 model2 > model1 < 2.22e-16
## BIC-corrected        -13.79448 model2 > model1 < 2.22e-16
```

The Vuong test compares the zero-inflated model (model 5) with the ordinary Poisson regression model (model 4). In this case, we can see that our test statistic is significant, indicating that the zero-inflated model is superior to the standard Poisson model.

```
# Predict the target count for the evaluation dataset
evaluation_set$TARGET<-predict(model5,type = 'response',newdata = evaluation_set)
```

```
str(evaluation_set)
```

```
## 'data.frame':    3335 obs. of  16 variables:
##  $ IN                : int  3 9 10 18 21 30 31 37 39 47 ...
##  $ TARGET            : num  NA 3.99 2.45 2.44 NA ...
##  $ FixedAcidity      : num  5.4 12.4 7.2 6.2 11.4 17.5 17.5 17.5 11.6 3.8 ...
##  $ VolatileAcidity   : num  -1.046 0.385 1.64 0.1 0.21 ...
##  $ CitricAcid        : num  0.27 -0.76 0.17 1.81 0.28 ...
##  $ ResidualSugar     : num  -10.7 -19.7 -52.8 1 1.2 1.4 4.6 31.9 -52.8 -7.7 ...
##  $ Chlorides         : num  0.092 0.593 0.065 -0.179 0.038 ...
##  $ FreeSulfurDioxide : num  23 -37 9 104 70 -218 10 115 35 40 ...
##  $ TotalSulfurDioxide: num  398 68 76 89 53 140 17 381 83 129 ...
##  $ Density           : num  0.985 0.99 1.04 0.989 1.04 ...
##  $ pH                : num  4.39 3.37 4.39 3.2 2.54 3.06 3.07 2.99 3.32 4.39 ...
##  $ Sulphates         : num  0.64 1.09 0.68 2.08 -0.07 ...
##  $ Alcohol           : num  12.3 16 8.55 12.3 4.8 11.4 8.5 11.4 4.2 10.9 ...
##  $ LabelAppeal       : int  -1 0 0 -1 0 1 0 1 0 0 ...
##  $ AcidIndex         : num  6 6 8 8 10 8 10 7 10 7 ...
##  $ STARS             : int  NA 2 1 1 NA 4 3 NA NA NA ...
```