HW 5

Team 2

May 15, 2019

Contents

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.

Objective

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).

Dataset

Below is a short description of the variables of interest in the data set:

- `INDEX`: Identification Variable (do not use).
- `TARGET`: Number of Cases Purchased.
- `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.
- `ResidualSugar`: Residual Sugar of wine.
- `STARS`: Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor.
- `Sulphates`: Sulfate conten of wine.
- `TotalSulfurDioxide`: Total Sulfur Dioxide of Wine.
- `VolatileAcidity`: Volatile Acid content of wine.
- `pH`: pH of wine.

There is a theoretical effect for `LabelAppeal`, which suggests many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales. Additionally, a high number captured by the `STARS` variable is theorized to suggest high sales.

Dependencies

Replication of our work requires the following packages in Rstudio:

```
library(psych)
library(randomForest)
library(corrplot)
library(caret)
library(MASS)
library(dplyr)
library(tidyr)
library(AER)
library(ggplot2)
library(reshape2)
library(pROC)
library(Metrics)
```

PART 1: DATA EXPLORATION

First, we read the data as a csv and then examined the below variable from the `training` dataset.

```
training <- as.data.frame(read.csv("/Users/Olga/Desktop/DataMining/HW5/wine-training-data.csv"))
test <- as.data.frame(read.csv("/Users/Olga/Desktop/DataMining/HW5/wine-evaluation-data.csv"))
dim(training)
```

```
FALSE [1] 12795    16
```

The data set contains 12,795 cases, 13 predictors, 1 response variable and INDEX column. Each case is a commerically available wine, with the response variable being the number of cases purchased by restaurants and wine shops. 12 predictors are related to chemical properties of wine and 2 related to rating and design.

Summary Statistics

We look at summary of the data below.

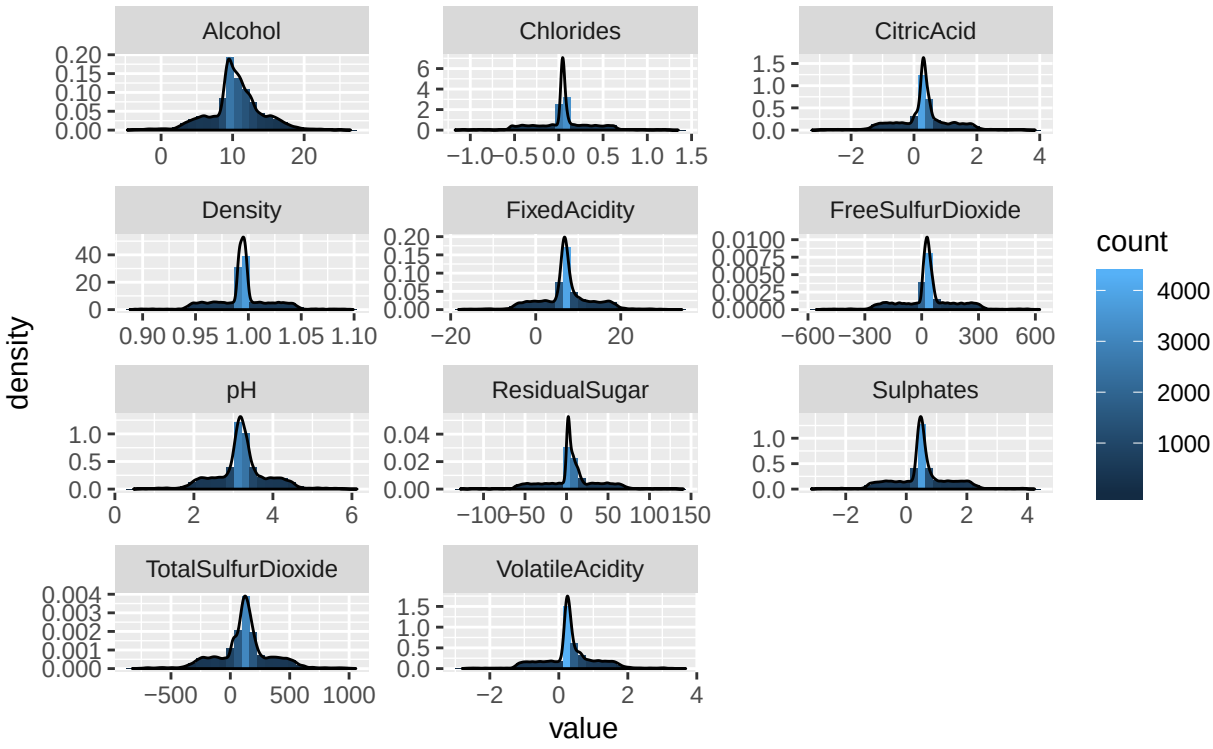
	vars	n	mean	sd	min	max	range	se
TARGET	1	12795	3.03	1.93	0.00	8.00	8.00	0.02
FixedAcidity	2	12795	7.08	6.32	-18.10	34.40	52.50	0.06
VolatileAcidity	3	12795	0.32	0.78	-2.79	3.68	6.47	0.01
CitricAcid	4	12795	0.31	0.86	-3.24	3.86	7.10	0.01
ResidualSugar	5	12179	5.42	33.75	-127.80	141.15	268.95	0.31
Chlorides	6	12157	0.05	0.32	-1.17	1.35	2.52	0.00
FreeSulfurDioxide	7	12148	30.85	148.71	-555.00	623.00	1178.00	1.35
TotalSulfurDioxide	8	12113	120.71	231.91	-823.00	1057.00	1880.00	2.11
Density	9	12795	0.99	0.03	0.89	1.10	0.21	0.00
pH	10	12400	3.21	0.68	0.48	6.13	5.65	0.01
Sulphates	11	11585	0.53	0.93	-3.13	4.24	7.37	0.01
Alcohol	12	12142	10.49	3.73	-4.70	26.50	31.20	0.03
LabelAppeal	13	12795	-0.01	0.89	-2.00	2.00	4.00	0.01
AcidIndex	14	12795	7.77	1.32	4.00	17.00	13.00	0.01
STARS	15	9436	2.04	0.90	1.00	4.00	3.00	0.01

```
FALSE             TARGET       FixedAcidity    VolatileAcidity
FALSE           0.000000           0.000000           0.000000
FALSE         CitricAcid      ResidualSugar          Chlorides
FALSE           0.000000           4.814381           4.986323
FALSE  FreeSulfurDioxide TotalSulfurDioxide            Density
FALSE           5.056663           5.330207           0.000000
FALSE                 pH          Sulphates            Alcohol
FALSE           3.087143           9.456819           5.103556
FALSE        LabelAppeal          AcidIndex              STARS
FALSE           0.000000           0.000000          26.252442
```

As we see 8 variables have missing values. The % of missing values vary from 3.08%(pH) to 26.25%(STARS). These values require imputation ot exclution to conduct further analysis, except variable STARS as we think that these variables should be equal to 0 star rating.

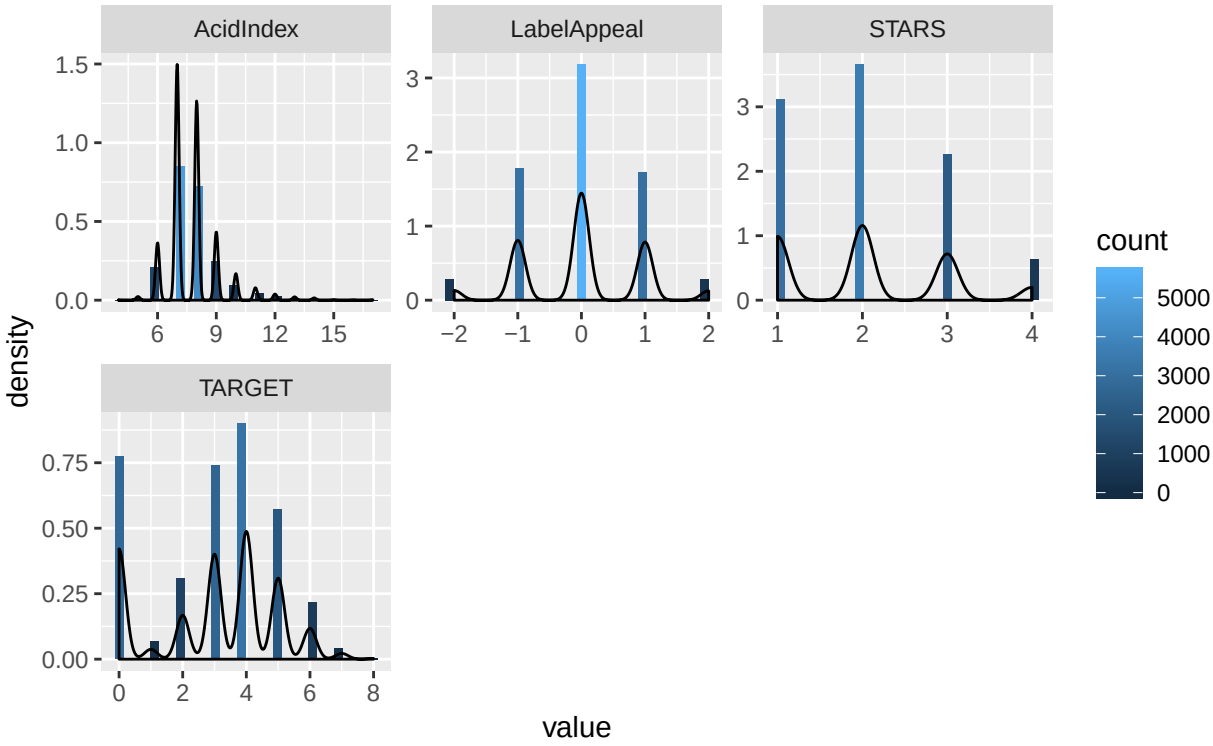
Distribution Of Continious Variables

Below, we examine the distribution of continious variables using histograms and density plots for each variable.



Most variables appear to be fairly normally distributed with a small spread. There is very little skew in all of these predictors.

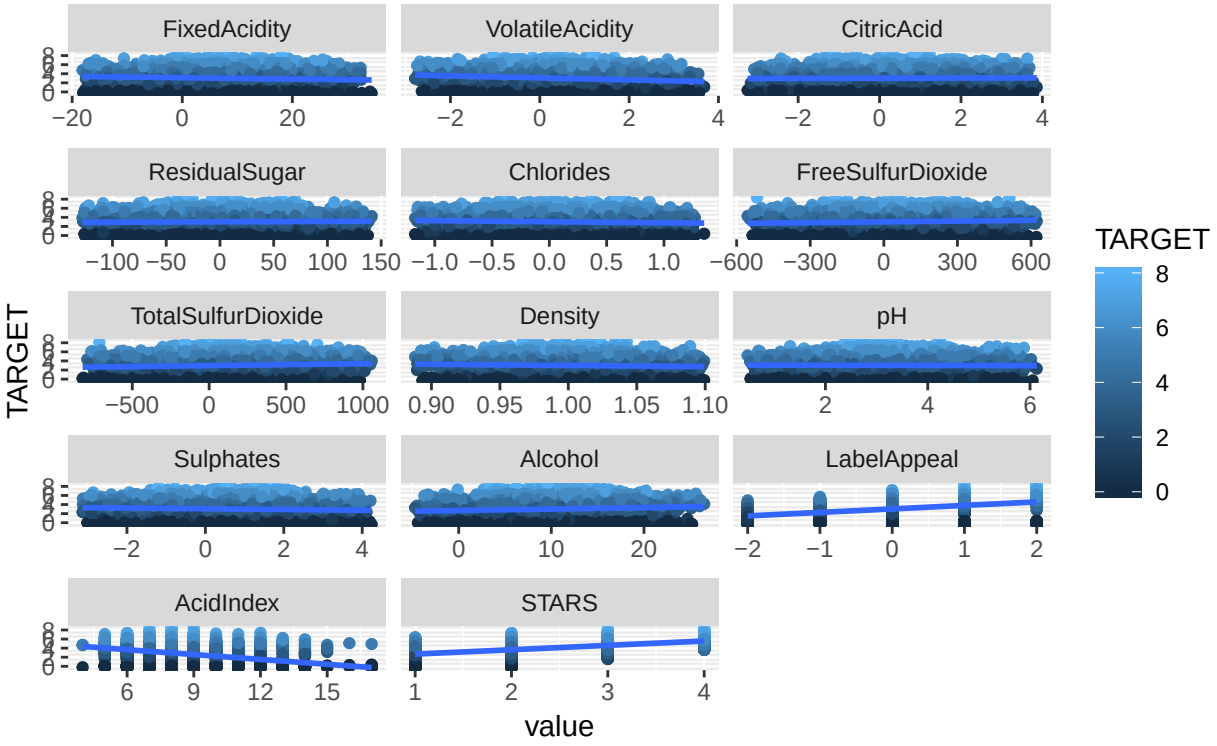
Poisson Distributions (binomial) for a discrete variables.



Correlation Plot Matrix

Scatter plot matrix with the `pairs` function does not seem useful given the amount of variables. Recommend using the ggplot correlation plot matrix, which shows linear relationship between the predictor and response variables.

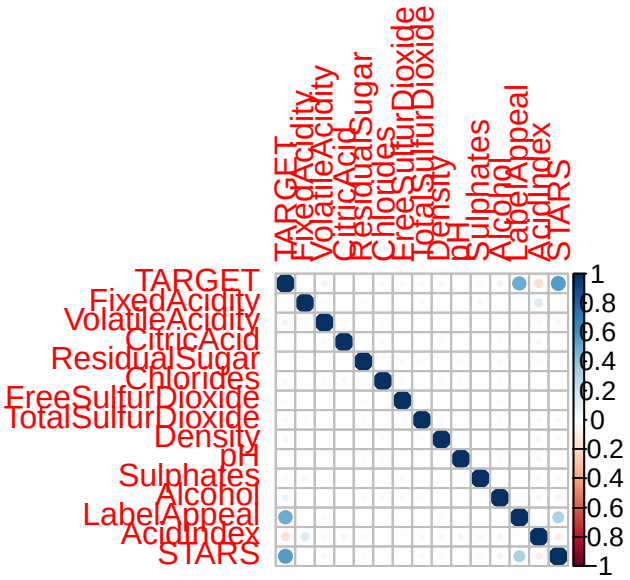
The correlation plot matrix below shows linear relationship between the predictor and response variables.



"LabelAppeal", 'AcidIndex' and "STARS show some correlation with target variable. It seems that ratings given by by experts and bottle aesthetics of the wine have a greater effect on the decision to purchase or not rather than any of the chemical properties (except 'AcidIndex')

Correlation

We can see our correlation matrix below. A dark blue circle represents a strong positive relationship and a dark red circle represents a strong negative relationship between two variables.



There is no strong collinearity in a data set.

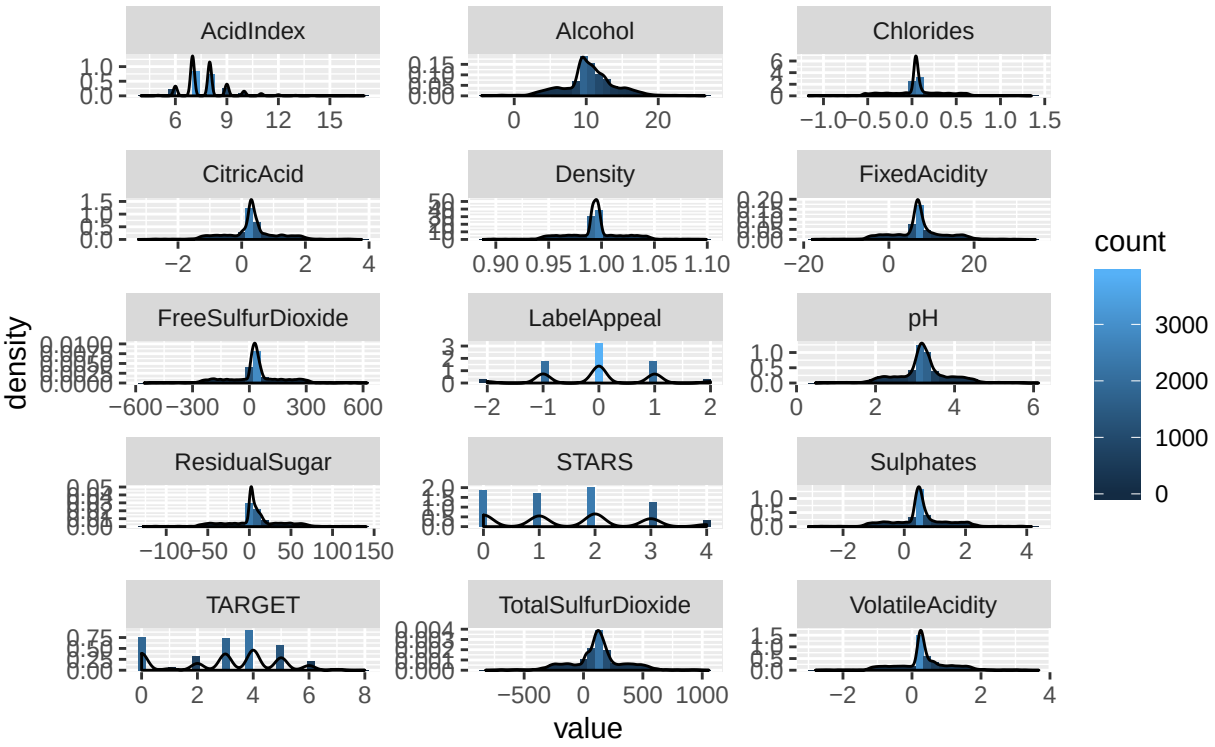
PART 2: DATA TRANSFORMATION

Handling missing vales

First of all we need to convert NAs in STARS variables to zero as we believe that NAs in STARS variable represent stars wine rating.

The nature of the training data set is that the predictors have very little skew, and the majority of values being centered around the mean. Taking into account that condition we can remove NAs or impute mean/median. Selecting the impute of mean/median will allow us just to add more values which are going to be centered around the mean. That's why we think that we can remove the rest of NAs.

The distribution of variables plot after NAs correction and deletion.



"training" data set does not contain meaninless values (negative ones when only positive ones are possible). Log or square root can help with transformation of skewed predictors. In our case it is "AcidIndex"

PART 3: BUILD MODELS

Poisson Regression

Target variable is a discrete variable, in this case, a simple transformation cannot produce normally distributed errors. The alternative is to use a Poisson model or one of its variants (negative binomial model).

Poisson model assumptions:

- the errors follow a Poisson, not a normal, distribution;
- it models the natural log of the response variable, ln(Y), as a linear function of the coefficients;
- the mean and variance of the errors are equal

Model 0

Building a poisson model based on data set where rows containing NAs were deleted.

```
FALSE
FALSE Call:
FALSE glm(formula = TARGET ~ ., family = "poisson", data = training)
FALSE
FALSE Deviance Residuals:
FALSE     Min       1Q   Median       3Q      Max
FALSE -2.9803  -0.7083   0.0639   0.5756   3.2351
FALSE
```

```
FALSE Coefficients:
FALSE                       Estimate Std. Error z value Pr(>|z|)
FALSE (Intercept)          1.618e+00  2.368e-01   6.830 8.49e-12 ***
FALSE FixedAcidity        -1.785e-04  1.001e-03  -0.178 0.858472
FALSE VolatileAcidity     -3.296e-02  7.888e-03  -4.178 2.94e-05 ***
FALSE CitricAcid           4.358e-03  7.178e-03   0.607 0.543785
FALSE ResidualSugar       -5.403e-05  1.831e-04  -0.295 0.767882
FALSE Chlorides           -4.827e-02  1.939e-02  -2.489 0.012815 *
FALSE FreeSulfurDioxide    1.275e-04  4.173e-05   3.057 0.002239 **
FALSE TotalSulfurDioxide   9.401e-05  2.698e-05   3.484 0.000493 ***
FALSE Density             -3.618e-01  2.332e-01  -1.552 0.120766
FALSE pH                  -1.708e-02  9.073e-03  -1.883 0.059759 .
FALSE Sulphates           -1.092e-02  6.657e-03  -1.640 0.101005
FALSE Alcohol              1.492e-03  1.677e-03   0.890 0.373490
FALSE LabelAppeal          1.324e-01  7.369e-03  17.963  < 2e-16 ***
FALSE AcidIndex           -8.671e-02  5.479e-03 -15.824  < 2e-16 ***
FALSE STARS                3.094e-01  5.532e-03  55.936  < 2e-16 ***
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
FALSE
FALSE (Dispersion parameter for poisson family taken to be 1)
FALSE
FALSE     Null deviance: 15334.3  on 8674  degrees of freedom
FALSE Residual deviance:  9962.1  on 8660  degrees of freedom
FALSE AIC: 31705
FALSE
FALSE Number of Fisher Scoring iterations: 5
```

```
FALSE       FixedAcidity    VolatileAcidity         CitricAcid
FALSE           1.025510           1.004797           1.005678
FALSE      ResidualSugar          Chlorides  FreeSulfurDioxide
FALSE           1.001886           1.003945           1.002879
FALSE TotalSulfurDioxide            Density                 pH
FALSE           1.003581           1.005744           1.004419
FALSE          Sulphates            Alcohol        LabelAppeal
FALSE           1.003035           1.011926           1.102394
FALSE          AcidIndex              STARS
FALSE           1.062624           1.141388
```

model_0 has AIC - 31705. There is no significant multicollinearity.

Model 1

Building a poisson model based on data set where rows containing NAs were replaced with mean.

```
FALSE
FALSE Call:
FALSE glm(formula = TARGET ~ ., family = "poisson", data = training_mean)
FALSE
FALSE Deviance Residuals:
FALSE     Min       1Q   Median       3Q      Max
FALSE -2.9803  -0.7083   0.0639   0.5756   3.2351
FALSE
FALSE Coefficients:
FALSE                       Estimate Std. Error z value Pr(>|z|)
```

```
FALSE (Intercept)         1.618e+00  2.368e-01   6.830 8.49e-12 ***
FALSE FixedAcidity       -1.785e-04  1.001e-03  -0.178 0.858472
FALSE VolatileAcidity    -3.296e-02  7.888e-03  -4.178 2.94e-05 ***
FALSE CitricAcid          4.358e-03  7.178e-03   0.607 0.543785
FALSE ResidualSugar      -5.403e-05  1.831e-04  -0.295 0.767882
FALSE Chlorides          -4.827e-02  1.939e-02  -2.489 0.012815 *
FALSE FreeSulfurDioxide   1.275e-04  4.173e-05   3.057 0.002239 **
FALSE TotalSulfurDioxide  9.401e-05  2.698e-05   3.484 0.000493 ***
FALSE Density            -3.618e-01  2.332e-01  -1.552 0.120766
FALSE pH                 -1.708e-02  9.073e-03  -1.883 0.059759 .
FALSE Sulphates          -1.092e-02  6.657e-03  -1.640 0.101005
FALSE Alcohol             1.492e-03  1.677e-03   0.890 0.373490
FALSE LabelAppeal         1.324e-01  7.369e-03  17.963  < 2e-16 ***
FALSE AcidIndex          -8.671e-02  5.479e-03 -15.824  < 2e-16 ***
FALSE STARS               3.094e-01  5.532e-03  55.936  < 2e-16 ***
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
FALSE
FALSE (Dispersion parameter for poisson family taken to be 1)
FALSE
FALSE     Null deviance: 15334.3  on 8674  degrees of freedom
FALSE Residual deviance:  9962.1  on 8660  degrees of freedom
FALSE AIC: 31705
FALSE
FALSE Number of Fisher Scoring iterations: 5
```

```
FALSE       FixedAcidity    VolatileAcidity         CitricAcid
FALSE           1.025510           1.004797           1.005678
FALSE      ResidualSugar          Chlorides  FreeSulfurDioxide
FALSE           1.001886           1.003945           1.002879
FALSE TotalSulfurDioxide            Density                 pH
FALSE           1.003581           1.005744           1.004419
FALSE          Sulphates            Alcohol        LabelAppeal
FALSE           1.003035           1.011926           1.102394
FALSE          AcidIndex              STARS
FALSE           1.062624           1.141388
```

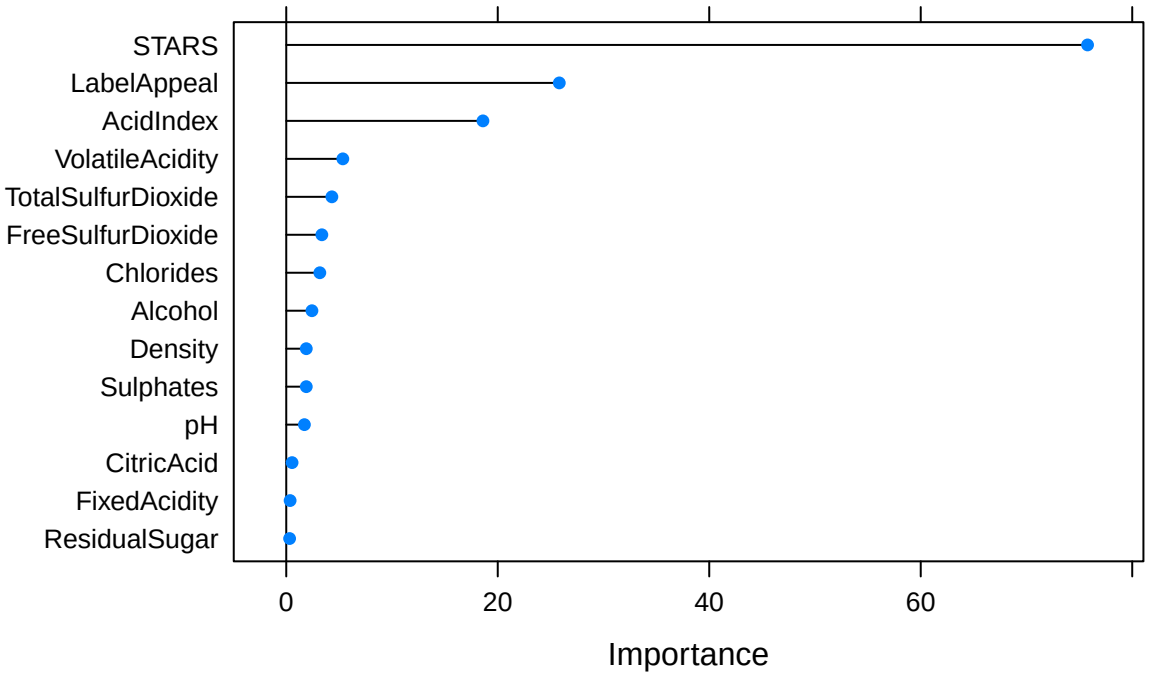
As we see both models shows same AIC - 31705. We belive that happens for the reason discussed in the part "data transformation" - handling missing values. There is no significant multicollinearity.

Model 2

Building a model based on the selected important variables using varImp() from caret package.

```
FALSE glm variable importance
FALSE
FALSE                    Overall
FALSE STARS              75.7764
FALSE LabelAppeal        25.8206
FALSE AcidIndex          18.6019
FALSE VolatileAcidity     5.3512
FALSE TotalSulfurDioxide  4.3154
FALSE FreeSulfurDioxide   3.3676
FALSE Chlorides           3.1753
FALSE Alcohol             2.4283
```

```
FALSE Density             1.8946
FALSE Sulphates           1.8936
FALSE pH                  1.7189
FALSE CitricAcid          0.5494
FALSE FixedAcidity        0.3652
FALSE ResidualSugar       0.3172
```



The most important variables is STARS which has 76 out of 100 scores, LbelAppeal - 26 and AcidIndex- 19. The rest variables are significantly leass important.

Model 2 will be built based on the most important variables only: STARS, LabelAppeal, AcidIndex

```
FALSE
FALSE Call:
FALSE glm(formula = TARGET ~ STARS + LabelAppeal + AcidIndex, family = "poisson",
FALSE     data = training_mean)
FALSE
FALSE Deviance Residuals:
FALSE     Min       1Q   Median       3Q      Max
FALSE -2.9867  -0.6865   0.0499   0.5659   3.2714
FALSE
FALSE Coefficients:
FALSE              Estimate Std. Error z value Pr(>|z|)
FALSE (Intercept)  1.222705   0.044152   27.69   <2e-16 ***
FALSE STARS        0.312331   0.005493   56.86   <2e-16 ***
FALSE LabelAppeal  0.132377   0.007361   17.98   <2e-16 ***
FALSE AcidIndex   -0.088132   0.005374  -16.40   <2e-16 ***
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
FALSE
FALSE (Dispersion parameter for poisson family taken to be 1)
FALSE
FALSE     Null deviance: 15334  on 8674  degrees of freedom
FALSE Residual deviance: 10018  on 8671  degrees of freedom
```

```
FALSE AIC: 31739
FALSE
FALSE Number of Fisher Scoring iterations: 5
```

```
FALSE       STARS LabelAppeal   AcidIndex
FALSE    1.127055    1.101034    1.026773
```

Model slighly deteriorated (based on AIC value), but not significanly and we used only 3 variables instead of 12. There is no significant multicollinearity.

Negative Binomial Regression

Negative binomial regression is for modeling count variables, usually for over-dispersed count outcome variables, that is when the conditional variance exceeds the conditional mean.

Before building Negative Binomial Regression we need to check if there is any evidence of overdispertion.

```
FALSE
FALSE   Overdispersion test
FALSE
FALSE data:  model_0
FALSE z = -9.8815, p-value = 1
FALSE alternative hypothesis: true dispersion is greater than 1
FALSE sample estimates:
FALSE dispersion
FALSE  0.8686586
```

The test shows that overdispersion takes place. In case with overdispertion Negative Binomial Regression may produce better model.

Model 3

Negative Binomial Regression model will be built using all variables.

```
FALSE
FALSE Call:
FALSE glm.nb(formula = TARGET ~ ., data = training, init.theta = 49024.77017,
FALSE     link = log)
FALSE
FALSE Deviance Residuals:
FALSE     Min       1Q   Median       3Q      Max
FALSE -2.9802  -0.7083   0.0639   0.5756   3.2350
FALSE
FALSE Coefficients:
FALSE                       Estimate Std. Error z value Pr(>|z|)
FALSE (Intercept)          1.618e+00  2.368e-01   6.830 8.50e-12 ***
FALSE FixedAcidity        -1.785e-04  1.001e-03  -0.178 0.858492
FALSE VolatileAcidity     -3.296e-02  7.888e-03  -4.178 2.94e-05 ***
FALSE CitricAcid           4.358e-03  7.178e-03   0.607 0.543793
FALSE ResidualSugar       -5.402e-05  1.831e-04  -0.295 0.767908
FALSE Chlorides           -4.827e-02  1.939e-02  -2.489 0.012816 *
FALSE FreeSulfurDioxide    1.276e-04  4.173e-05   3.056 0.002240 **
FALSE TotalSulfurDioxide   9.401e-05  2.698e-05   3.484 0.000493 ***
FALSE Density             -3.618e-01  2.332e-01  -1.552 0.120773
FALSE pH                  -1.708e-02  9.074e-03  -1.883 0.059760 .
```

```
FALSE Sulphates          -1.092e-02  6.657e-03  -1.640 0.101004
FALSE Alcohol             1.492e-03  1.677e-03   0.890 0.373527
FALSE LabelAppeal         1.324e-01  7.369e-03  17.963  < 2e-16 ***
FALSE AcidIndex          -8.671e-02  5.479e-03 -15.824  < 2e-16 ***
FALSE STARS               3.094e-01  5.532e-03  55.935  < 2e-16 ***
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
FALSE
FALSE (Dispersion parameter for Negative Binomial(49024.77) family taken to be 1)
FALSE
FALSE     Null deviance: 15333.7  on 8674  degrees of freedom
FALSE Residual deviance:  9961.8  on 8660  degrees of freedom
FALSE AIC: 31708
FALSE
FALSE Number of Fisher Scoring iterations: 1
FALSE
FALSE
FALSE               Theta:  49025
FALSE           Std. Err.:  61688
FALSE Warning while fitting theta: iteration limit reached
FALSE
FALSE  2 x log-likelihood:  -31675.54
```

The AIC is 31708 and it is similar that we have achived with Poisson model (AIC - 31705).

Model 4

Negative Binomial Regression model will be built using only the most important variables and taking log of AcidIndex (as AcidIndex is lightly skewed).

```
FALSE
FALSE Call:
FALSE glm.nb(formula = TARGET ~ STARS + LabelAppeal + log(AcidIndex),
FALSE     data = training, init.theta = 48974.32239, link = log)
FALSE
FALSE Deviance Residuals:
FALSE     Min       1Q   Median       3Q      Max
FALSE -2.9759  -0.6796   0.0612   0.5584   3.2900
FALSE
FALSE Coefficients:
FALSE                 Estimate Std. Error z value Pr(>|z|)
FALSE (Intercept)     1.887260   0.088335   21.36   <2e-16 ***
FALSE STARS           0.314005   0.005486   57.24   <2e-16 ***
FALSE LabelAppeal     0.131510   0.007360   17.87   <2e-16 ***
FALSE log(AcidIndex) -0.663076   0.042593  -15.57   <2e-16 ***
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
FALSE
FALSE (Dispersion parameter for Negative Binomial(48974.32) family taken to be 1)
FALSE
FALSE     Null deviance: 15334  on 8674  degrees of freedom
FALSE Residual deviance: 10054  on 8671  degrees of freedom
FALSE AIC: 31777
FALSE
```

```
FALSE Number of Fisher Scoring iterations: 1
FALSE
FALSE
FALSE               Theta:  48974
FALSE           Std. Err.:  62029
FALSE Warning while fitting theta: iteration limit reached
FALSE
FALSE  2 x log-likelihood:  -31767.46
```

The AIC of the model is 31777 which is slightly higher than the result achived with Poisson model (31739).

Multiple Linear Regression

Model 5

Multiple Linear Regression model will be built using all variables.

```
FALSE
FALSE Call:
FALSE glm(formula = TARGET ~ ., family = "gaussian", data = training)
FALSE
FALSE Deviance Residuals:
FALSE     Min       1Q   Median       3Q      Max
FALSE -4.5582  -0.9421   0.0522   0.9042   6.0007
FALSE
FALSE Coefficients:
FALSE                       Estimate Std. Error t value Pr(>|t|)
FALSE (Intercept)         4.212e+00  5.446e-01   7.734 1.16e-14 ***
FALSE FixedAcidity        8.444e-04  2.312e-03   0.365 0.714956
FALSE VolatileAcidity    -9.751e-02  1.822e-02  -5.351 8.96e-08 ***
FALSE CitricAcid          9.147e-03  1.665e-02   0.549 0.582723
FALSE ResidualSugar      -1.337e-04  4.215e-04  -0.317 0.751066
FALSE Chlorides          -1.417e-01  4.464e-02  -3.175 0.001502 **
FALSE FreeSulfurDioxide   3.240e-04  9.622e-05   3.368 0.000762 ***
FALSE TotalSulfurDioxide  2.682e-04  6.215e-05   4.315 1.61e-05 ***
FALSE Density            -1.019e+00  5.380e-01  -1.895 0.058182 .
FALSE pH                 -3.611e-02  2.101e-02  -1.719 0.085660 .
FALSE Sulphates          -2.907e-02  1.535e-02  -1.894 0.058317 .
FALSE Alcohol             9.400e-03  3.871e-03   2.428 0.015190 *
FALSE LabelAppeal         4.309e-01  1.669e-02  25.821  < 2e-16 ***
FALSE AcidIndex          -2.066e-01  1.111e-02 -18.602  < 2e-16 ***
FALSE STARS               9.691e-01  1.279e-02  75.776  < 2e-16 ***
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
FALSE
FALSE (Dispersion parameter for gaussian family taken to be 1.766497)
FALSE
FALSE     Null deviance: 31832  on 8674  degrees of freedom
FALSE Residual deviance: 15298  on 8660  degrees of freedom
FALSE AIC: 29572
FALSE
FALSE Number of Fisher Scoring iterations: 2
```

The AIC of the model is 29572 which is significantly lower than the result achived with Poisson or Negative Binomial Regression models.

Model 6

Multiple Linear Regression model will be built using only the most important variables.

```
FALSE
FALSE Call:
FALSE glm(formula = TARGET ~ STARS + LabelAppeal + AcidIndex, family = "gaussian",
FALSE     data = training)
FALSE
FALSE Deviance Residuals:
FALSE     Min       1Q   Median       3Q      Max
FALSE -4.5375  -0.9105   0.1041   0.9225   6.0630
FALSE
FALSE Coefficients:
FALSE             Estimate Std. Error t value Pr(>|t|)
FALSE (Intercept)  3.20236    0.09167   34.94   <2e-16 ***
FALSE STARS        0.97949    0.01277   76.68   <2e-16 ***
FALSE LabelAppeal  0.42972    0.01675   25.65   <2e-16 ***
FALSE AcidIndex   -0.21189    0.01089  -19.45   <2e-16 ***
FALSE ---
FALSE Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
FALSE
FALSE (Dispersion parameter for gaussian family taken to be 1.782014)
FALSE
FALSE     Null deviance: 31832  on 8674  degrees of freedom
FALSE Residual deviance: 15452  on 8671  degrees of freedom
FALSE AIC: 29637
FALSE
FALSE Number of Fisher Scoring iterations: 2
```

The AIC of the model is 29637 which is significantly lower than the result achived with Poisson or Negative Binomial Regression models.

PART 4: SELECT MODELS

Model Evaluation

We are going to evaluate models based on the following criteria: AIC, BIC and Average Squared Error

Splitting "training" data set on "train.data"" and "test.data"" in order to assess models using average squared error. Here is the piece of the randomly selected "test.data" from the "training" data set.

```
FALSE    TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
FALSE 3       5          7.1           2.640      -0.88         14.80     0.037
FALSE 4       3          5.7           0.385       0.04         18.80    -0.425
FALSE 6       0         11.3           0.320       0.59          2.20     0.556
FALSE 13      3          6.0           0.330      -1.06          3.00     0.518
FALSE 20      4          6.8           0.475      -0.20        -50.75     0.047
FALSE 21      3          5.8           0.760       0.33        -26.00    -0.195
FALSE    FreeSulfurDioxide TotalSulfurDioxide Density   pH Sulphates Alcohol
FALSE 3                214                142 0.99518 3.12      0.48    22.0
FALSE 4                 22                115 0.99640 2.24      1.83     6.2
FALSE 6                -37                 15 0.99940 3.20      1.29    15.4
FALSE 13                 5                378 0.96643 3.55     -0.86     3.9
```

```
FALSE 20               -88               508 0.99403 3.23      0.35    18.3
FALSE 21                87              -283 0.98850 3.07      0.56    11.4
FALSE    LabelAppeal AcidIndex STARS
FALSE 3           -1         8     3
FALSE 4           -1         6     1
FALSE 6            0        11     0
FALSE 13           1         7     2
FALSE 20          -1         8     2
FALSE 21          -1         6     1
```

```
FALSE [1] 3468   15
```

The results of the models evaluations are presented in the following table:

```
FALSE              AIC      BIC Sq.Error
FALSE model_0 31705.35 31811.38 6.761871
FALSE model_1 31705.35 31811.38 6.761871
FALSE model_2 31739.42 31767.70 6.766863
FALSE model_3 31707.54 31820.63 6.761866
FALSE model_4 31777.46 31812.81 6.769644
FALSE model_5 29571.63 29684.72 1.746332
FALSE model_6 29636.51 29671.86 1.767922
```

model_5 and model_6 have the best performance based on the selected criteria. It is unexpected taking into account the nature of the TARGET variable and detected overdispertion.

We select model_5 as the best model.

Forecasting

Preparing test data set for the prediction - applying the same transformations as for the training data set: handling NAs.

Making prediction using the best model (overall): model_5

```
FALSE   IN TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar
FALSE 1  3      2          5.4          -0.860       0.27         -10.7
FALSE 2  9      4         12.4           0.385      -0.76         -19.7
FALSE 3 10      2          7.2           1.750       0.17         -33.0
FALSE 4 18      2          6.2           0.100       1.80           1.0
FALSE 5 21      1         11.4           0.210       0.28           1.2
FALSE 6 30      6         17.6           0.040      -1.15           1.4
FALSE   Chlorides FreeSulfurDioxide TotalSulfurDioxide Density   pH Sulphates
FALSE 1     0.092               23                398 0.98527 5.02      0.64
FALSE 2     1.169              -37                 68 0.99048 3.37      1.09
FALSE 3     0.065                9                 76 1.04641 4.61      0.68
FALSE 4    -0.179              104                 89 0.98877 3.20      2.11
FALSE 5     0.038               70                 53 1.02899 2.54     -0.07
FALSE 6     0.535             -250                140 0.95028 3.06     -0.02
FALSE   Alcohol LabelAppeal AcidIndex STARS
FALSE 1   12.30          -1         6     0
FALSE 2   16.00           0         6     2
FALSE 3    8.55           0         8     1
FALSE 4   12.30          -1         8     1
FALSE 5    4.80           0        10     0
FALSE 6   11.40           1         8     4
```

Making prediction using the best model among count regression models (as requested in the HW): model_0

```
FALSE   IN TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar
FALSE 1  3      2          5.4          -0.860       0.27         -10.7
FALSE 2  9      4         12.4           0.385      -0.76         -19.7
FALSE 3 10      2          7.2           1.750       0.17         -33.0
FALSE 4 18      2          6.2           0.100       1.80           1.0
FALSE 5 21      1         11.4           0.210       0.28           1.2
FALSE 6 30      6         17.6           0.040      -1.15           1.4
FALSE   Chlorides FreeSulfurDioxide TotalSulfurDioxide Density   pH Sulphates
FALSE 1     0.092               23                398 0.98527 5.02      0.64
FALSE 2     1.169              -37                 68 0.99048 3.37      1.09
FALSE 3     0.065                9                 76 1.04641 4.61      0.68
FALSE 4    -0.179              104                 89 0.98877 3.20      2.11
FALSE 5     0.038               70                 53 1.02899 2.54     -0.07
FALSE 6     0.535             -250                140 0.95028 3.06     -0.02
FALSE   Alcohol LabelAppeal AcidIndex STARS TARGET2
FALSE 1   12.30          -1         6     0       2
FALSE 2   16.00           0         6     2       3
FALSE 3    8.55           0         8     1       2
FALSE 4   12.30          -1         8     1       2
FALSE 5    4.80           0        10     0       1
FALSE 6   11.40           1         8     4       6
```