DATA 621 HW5

Ahsanul Choudhury

May 6, 2018

Contents

Introduction

The purpose of this excercise is to 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.

Our 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.

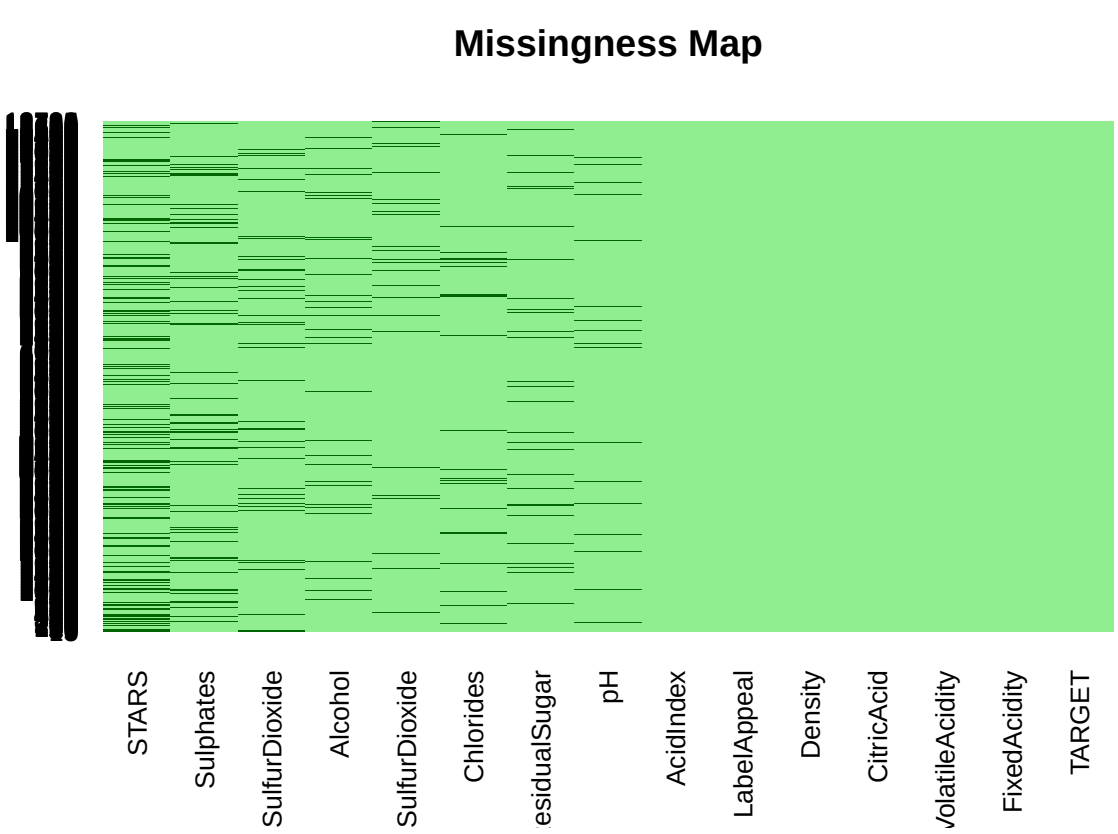
Data Exploration

Let's look at the data first; there are 12795 observations and 16 variables, following table contains the names of the variable and a brief description of each variable:

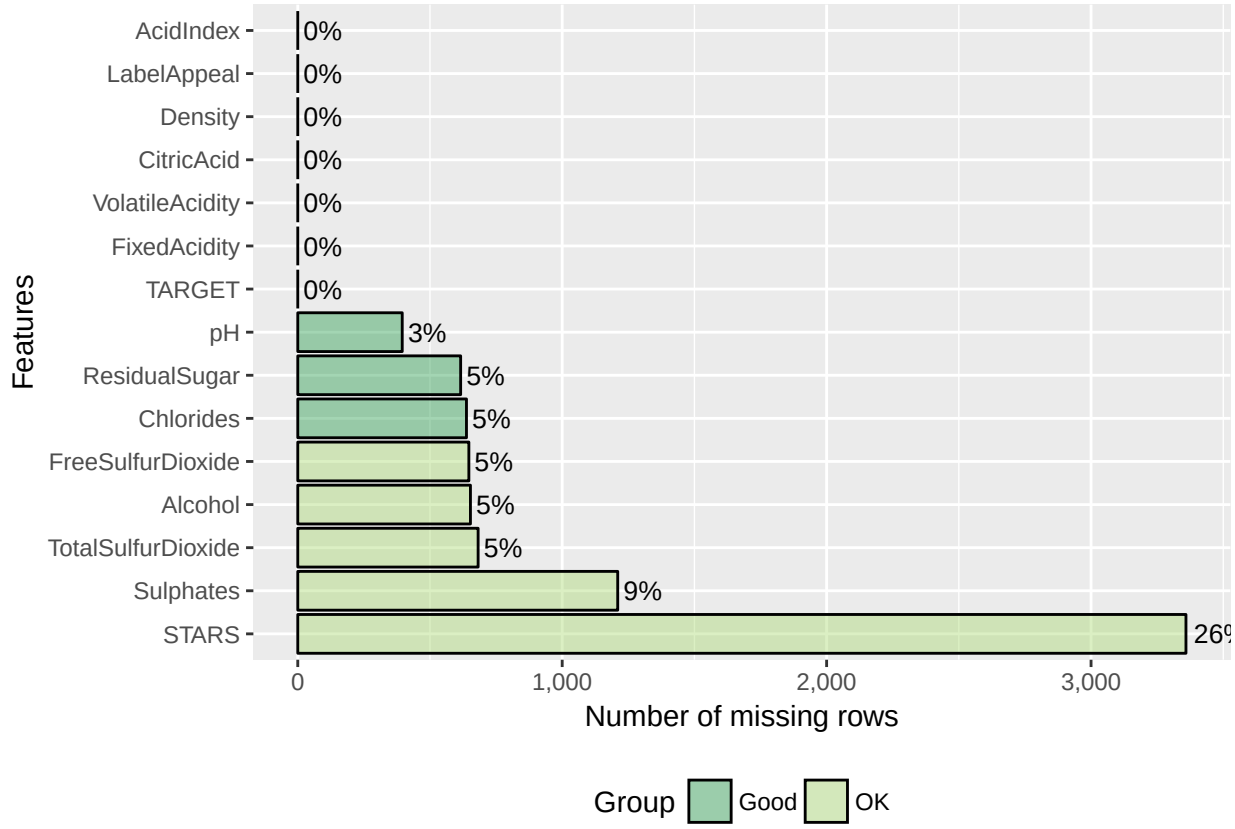
Variables	Description
ï..INDEX	Identification Variable
TARGET	Number of Cases Purchased
FixedAcidity	Fixed Acidity of Wine
VolatileAcidity	Volatile Acid content of wine
CitricAcid	Citric Acid Content
ResidualSugar	Residual Sugar of wine
Chlorides	Chloride content of wine
FreeSulfurDioxide	Sulfur Dioxide content of wine
TotalSulfurDioxide	Total Sulfur Dioxide of Wine
Density	Density of Wine
pH	pH of wine
Sulphates	Sulfate content of wine
Alcohol	Alcohol Content
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.
AcidIndex	Proprietary method of testing total acidity of wine by using a weighted average
STARS	Wine rating by a team of experts. 4 Stars = Excellent, 1 Star = Poor

There are a large numbers of NA values in the data set, the following table shows the the NA values in each of the variables.

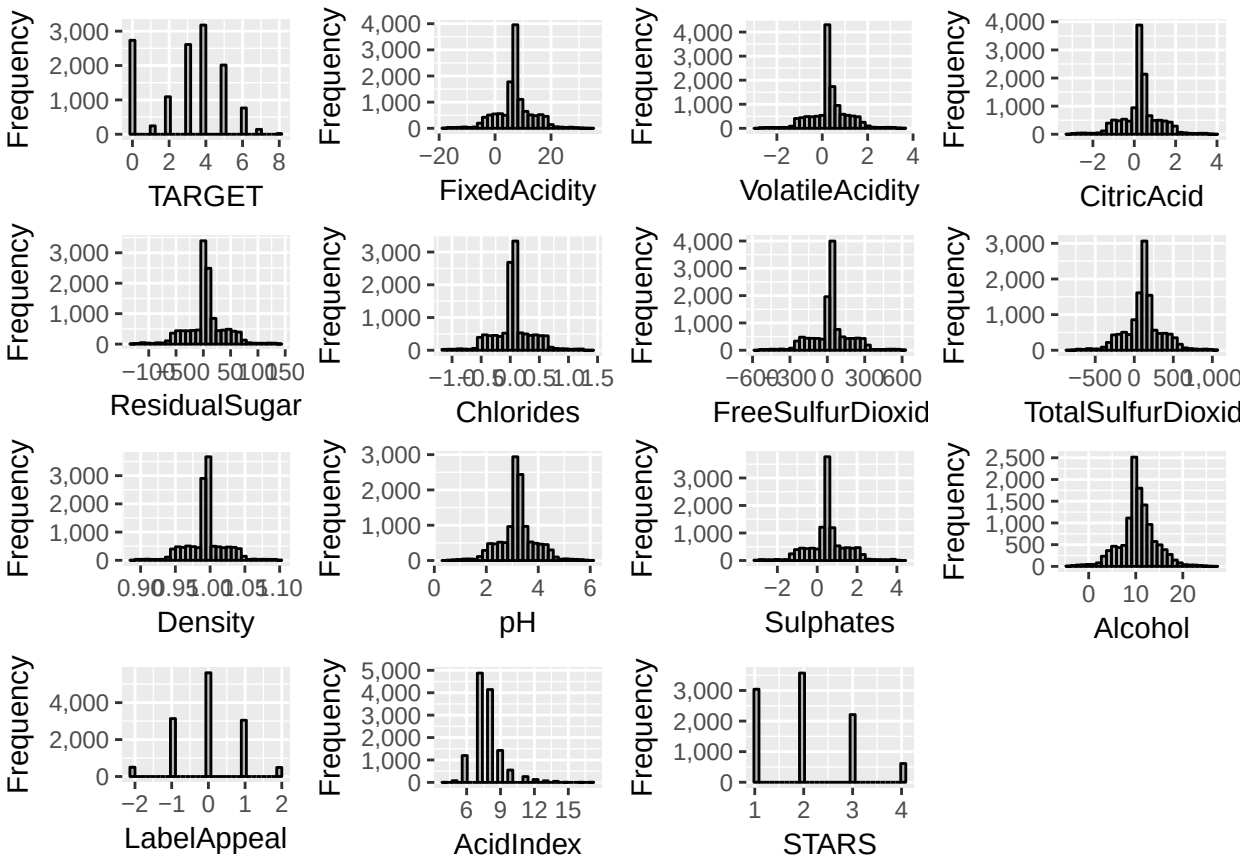
	Missing Data
TARGET	0
FixedAcidity	0
VolatileAcidity	0
CitricAcid	0
ResidualSugar	616
Chlorides	638
FreeSulfurDioxide	647
TotalSulfurDioxide	682
Density	0
pH	395
Sulphates	1210
Alcohol	653
LabelAppeal	0
AcidIndex	0
STARS	3359

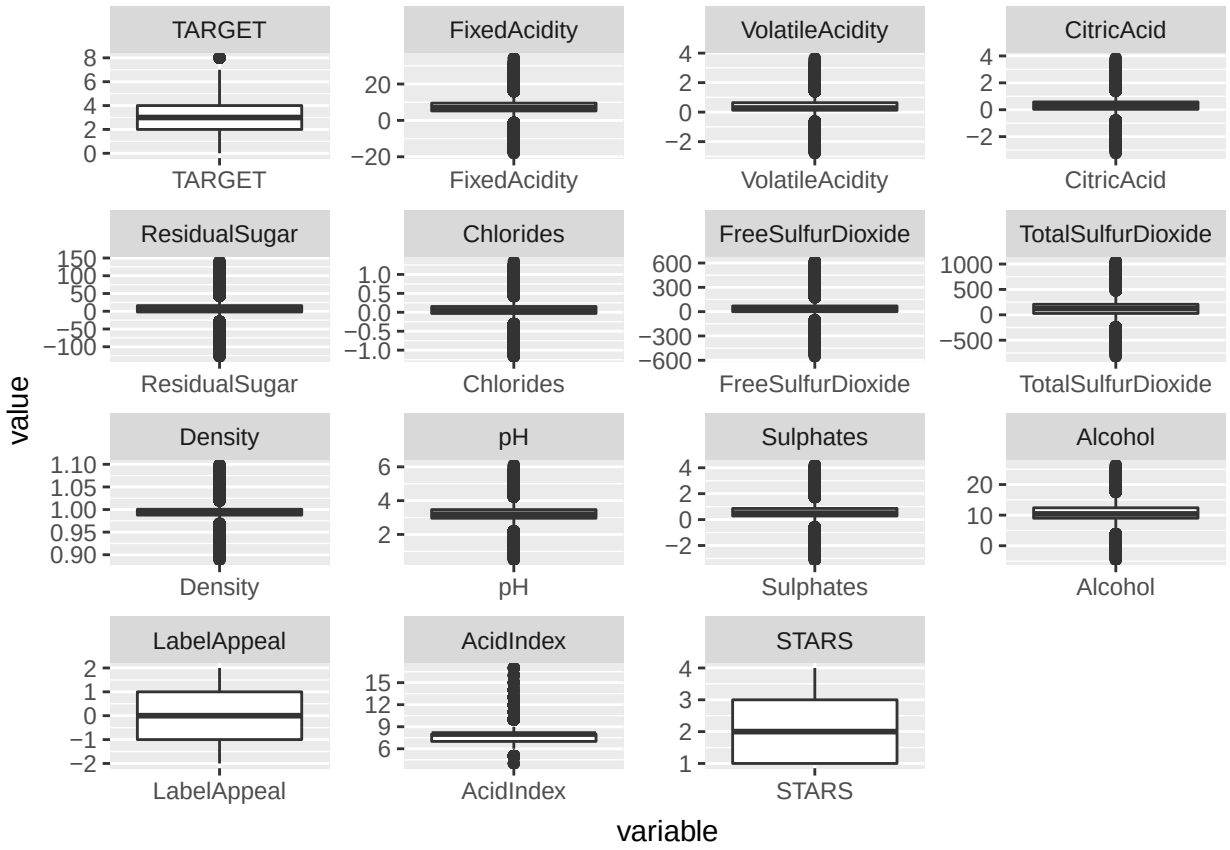


In the *Missingness Map* we do not see any pattern in the dataset for the missing values, **STARS** has the highest number of missing values with 3359 followed by **Sulphates** with 1210 missing values. The following graph shows the percentage of missing data in each variables.



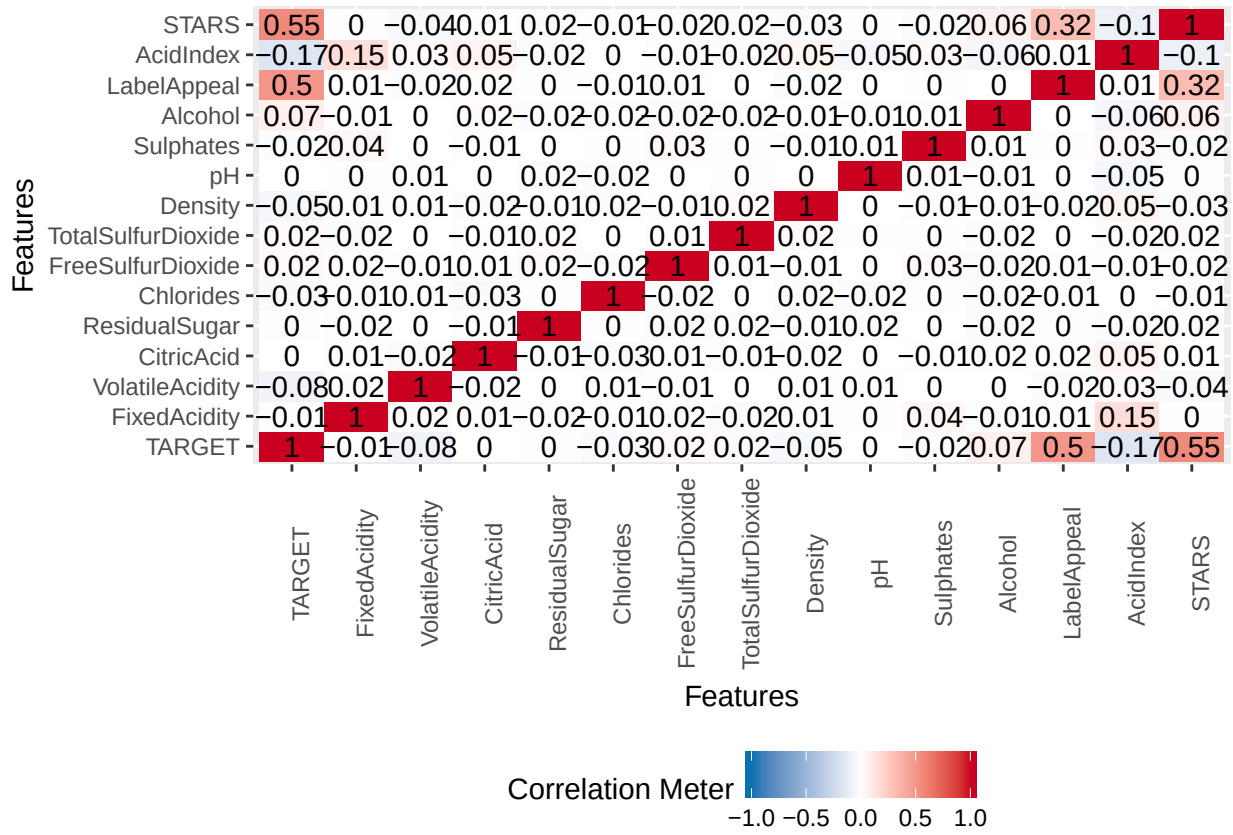
In order to see the distribution and outliers of the variables we next we will plot histograms and boxplot for each of the variables.


Frequency
TARGET
FixedAcidity
VolatileAcidity
CitricAcid
ResidualSugar
Chlorides
FreeSulfurDioxid
TotalSulfurDioxid
Density
pH
Sulphates
Alcohol
LabelAppeal
AcidIndex
STARS



From the histograms we can see most of the variables are fairly normally distributed and the boxplots indicates presence of outliers.

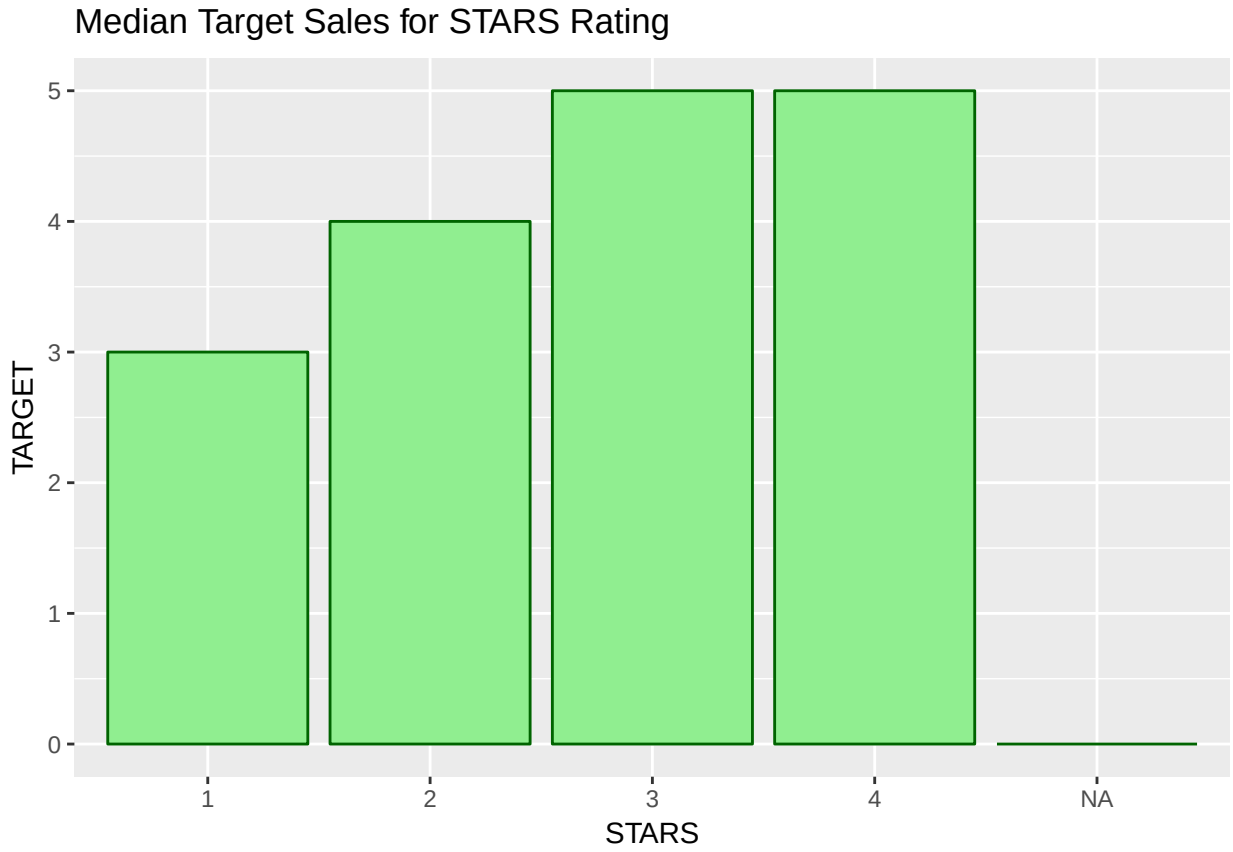
Correlation Plot of Variables



The correlation plot shows only weak to moderate correlation between the variables.

Data Preparation

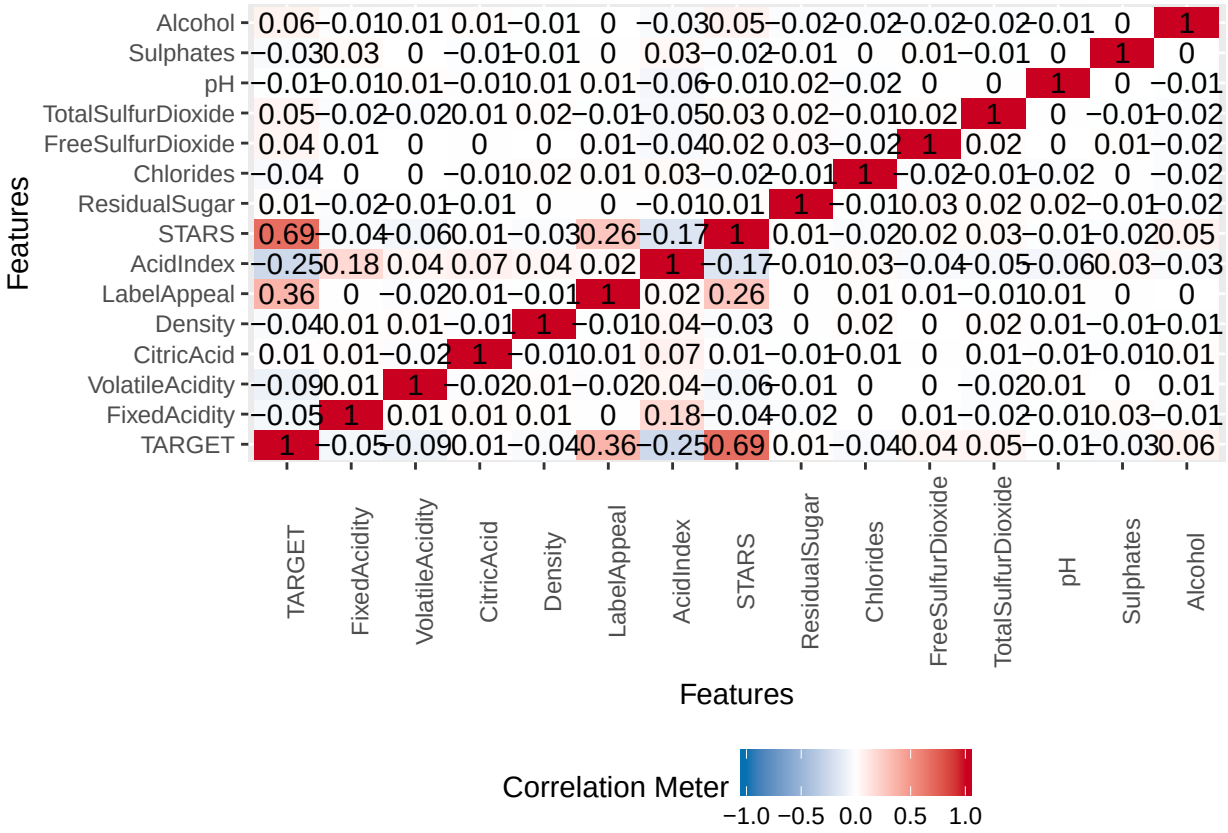
In order to prepare the data for building models we will handle the missing values first. **STARS** has 26% of data missing and before handling the missing values in **STARS** let's look at the distribution of **TARGET** based on **STARS**:



We can see if a wine in not rated by the team of experts sells poorly, which indicates missing **STAR** rating is, in fact, predictive of **TARGET**. So, we can conclude missing values in **STARS** do not require any replacement and we will simply replace the missing values with zero.

For the missing values in all the other variables we imputed data using predictive mean matching approach from *mice* package in R.

Correlation Plot After Missing Data Handling



Build Models

For our exercise we will build following six models:

- Model 1: Poisson
- Model 2: Poisson Reduced
- Model 3: Negative Binomial
- Model 4: Negative Binomial Reduced
- Model 5: Backward Stepwise Multiple Linear Regression
- Model 6: Zero Dispersion Counts

Please follow the link provided in Appendix for R code for the models

Select Models

Model	AIC	BIC
Poisson	46699.87	46811.72
Poisson Reduced	46709.05	46776.17
Negative Binomial	46702.15	46821.46
Negative Binomial Reduced	46711.33	46785.90
Backward Stepwise Multiple Linear Regression	43509.04	43613.44
Zero Dispersion Counts	41695.15	NA

Based on the numbers on the above table we will select model 6 or the *Zero Dispersion Counts* model to use for our model prediction.

Summary of *Zero Dispersion Counts* model given below:

```
##
## Call:
## zeroinfl(formula = TARGET ~ . | STARS, data = train, dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.31560 -0.53726  0.01947  0.40923  2.89564
##
## Count model coefficients (negbin with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.551e+00  2.015e-01   7.699 1.37e-14 ***
## FixedAcidity        3.029e-04  8.395e-04   0.361   0.7183
## VolatileAcidity    -1.512e-02  6.713e-03  -2.253   0.0243 *
## CitricAcid          1.465e-03  6.014e-03   0.244   0.8075
## Density            -2.729e-01  1.978e-01  -1.380   0.1676
## LabelAppeal         2.240e-01  6.324e-03  35.416  < 2e-16 ***
## AcidIndex          -3.147e-02  5.038e-03  -6.245 4.23e-10 ***
## STARS               1.007e-01  5.224e-03  19.281  < 2e-16 ***
## ResidualSugar      -4.896e-05  1.541e-04  -0.318   0.7507
## Chlorides          -2.108e-02  1.649e-02  -1.278   0.2011
## FreeSulfurDioxide   3.454e-05  3.432e-05   1.006   0.3142
## TotalSulfurDioxide -1.837e-06  2.199e-05  -0.084   0.9334
## pH                  1.875e-03  7.747e-03   0.242   0.8087
## Sulphates          -1.924e-03  5.616e-03  -0.343   0.7320
## Alcohol             6.335e-03  1.397e-03   4.534 5.78e-06 ***
## Log(theta)          2.019e+01  1.299e+00  15.540  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.38032    0.03624   10.50   <2e-16 ***
## STARS       -2.22396    0.05419  -41.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 588048265.9193
## Number of iterations in BFGS optimization: 65
## Log-likelihood: -2.083e+04 on 18 Df
```

Make Prediction

My prediction .csv file can be found on the following link:
https://github.com/choudhury1023/DATA-621/blob/master/HW%205/HW5preds.csv

Appendix

For R code please follow the link below:
https://github.com/choudhury1023/DATA-621/blob/master/HW%205/Ahsanul_Choudhury_HW5.Rmd