DATA 621 Homework 5

Critical Thinking Group 1

December 05, 2021

Contents

Prepared for:

Prof. Dr. Nasrin Khansari

City University of New York, School of Professional Studies - Data 621

DATA 621 – Business Analytics and Data Mining

Home Work 5

Prepared by:

Critical Thinking Group 1

Vic Chan

Gehad Gad

Evan McLaughlin

Bruno de Melo

Anjal Hussan

Zhouxin Shi

Sie Siong Wong

Introduction

Problem

Our goal is to explore, analyze and model a dataset 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.

The 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

Below we'll display a few basic EDA techniques to gain insight into our wine dataset.

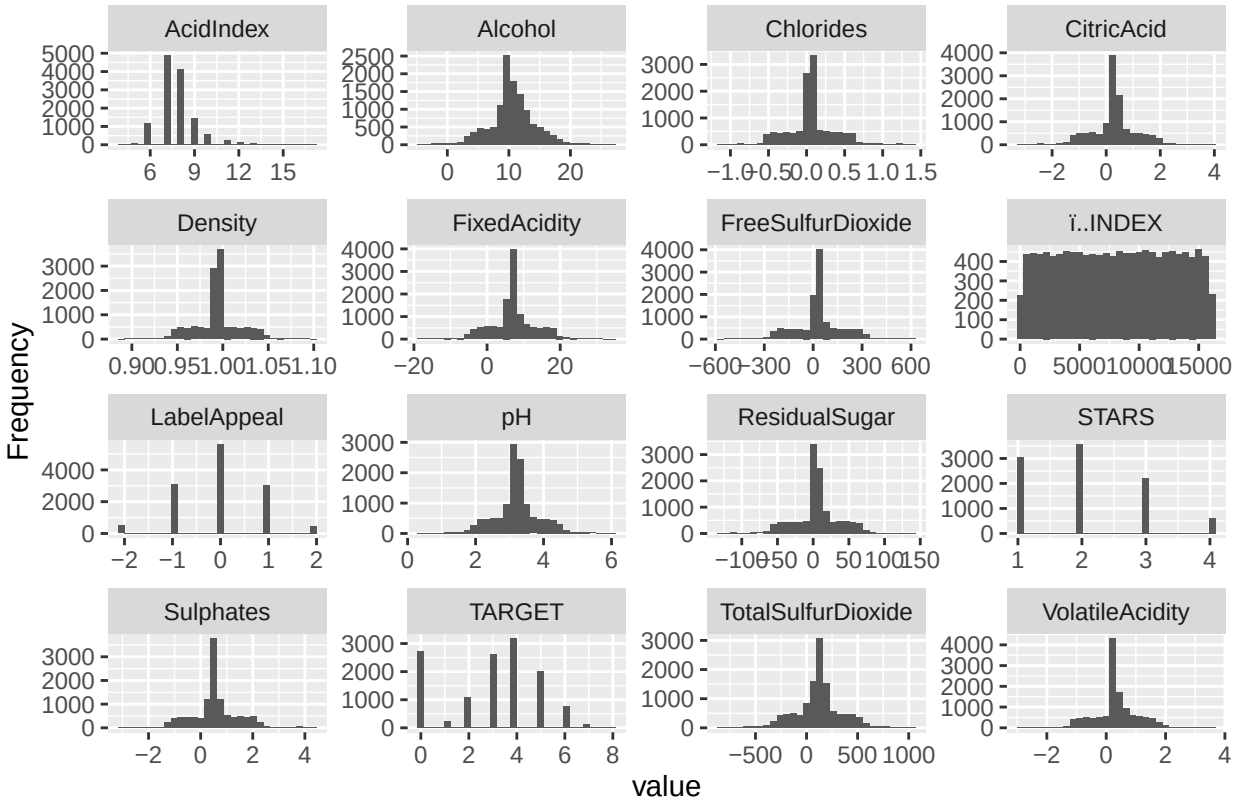
Basic Statistics

The data is 1.3 Mb in size. There are 12,795 rows and 15 columns (features). Of all 15 columns, 0 are discrete, 15 are continuous, and 0 are all missing. There are 8,200 missing values out of 191,925 data points.

```
##                        n    mean      sd  median     min      max  skew
## ï..INDEX           12795 8069.98 4656.91 8110.00    1.00 16129.00  0.00
## TARGET             12795    3.03    1.93    3.00    0.00     8.00 -0.33
## FixedAcidity       12795    7.08    6.32    6.90  -18.10    34.40 -0.02
## VolatileAcidity    12795    0.32    0.78    0.28   -2.79     3.68  0.02
## CitricAcid         12795    0.31    0.86    0.31   -3.24     3.86 -0.05
## ResidualSugar      12179    5.42   33.75    3.90 -127.80   141.15 -0.05
## Chlorides          12157    0.05    0.32    0.05   -1.17     1.35  0.03
## FreeSulfurDioxide  12148   30.85  148.71   30.00 -555.00   623.00  0.01
## TotalSulfurDioxide 12113  120.71  231.91  123.00 -823.00  1057.00 -0.01
## Density            12795    0.99    0.03    0.99    0.89     1.10 -0.02
## pH                 12400    3.21    0.68    3.20    0.48     6.13  0.04
## Sulphates          11585    0.53    0.93    0.50   -3.13     4.24  0.01
## Alcohol            12142   10.49    3.73   10.40   -4.70    26.50 -0.03
## LabelAppeal        12795   -0.01    0.89    0.00   -2.00     2.00  0.01
## AcidIndex          12795    7.77    1.32    8.00    4.00    17.00  1.65
##                    kurtosis
## ï..INDEX              -1.20
## TARGET                -0.88
## FixedAcidity           1.67
## VolatileAcidity        1.83
## CitricAcid             1.84
## ResidualSugar          1.88
## Chlorides              1.79
## FreeSulfurDioxide      1.84
## TotalSulfurDioxide     1.67
## Density                1.90
## pH                     1.65
## Sulphates              1.75
## Alcohol                1.54
## LabelAppeal           -0.26
## AcidIndex              5.19
```

It's useful to note a couple of things right off the bat with regard to our dataset: - There are several variables that have negative values. - ResidualSugar, Chlorides, FreeSulfurDioxide, and TotalSulfurDioxide all have quite a few missing values that we are going to need to deal with in order to assess the variables. - The Index column is useless and can be ignored.

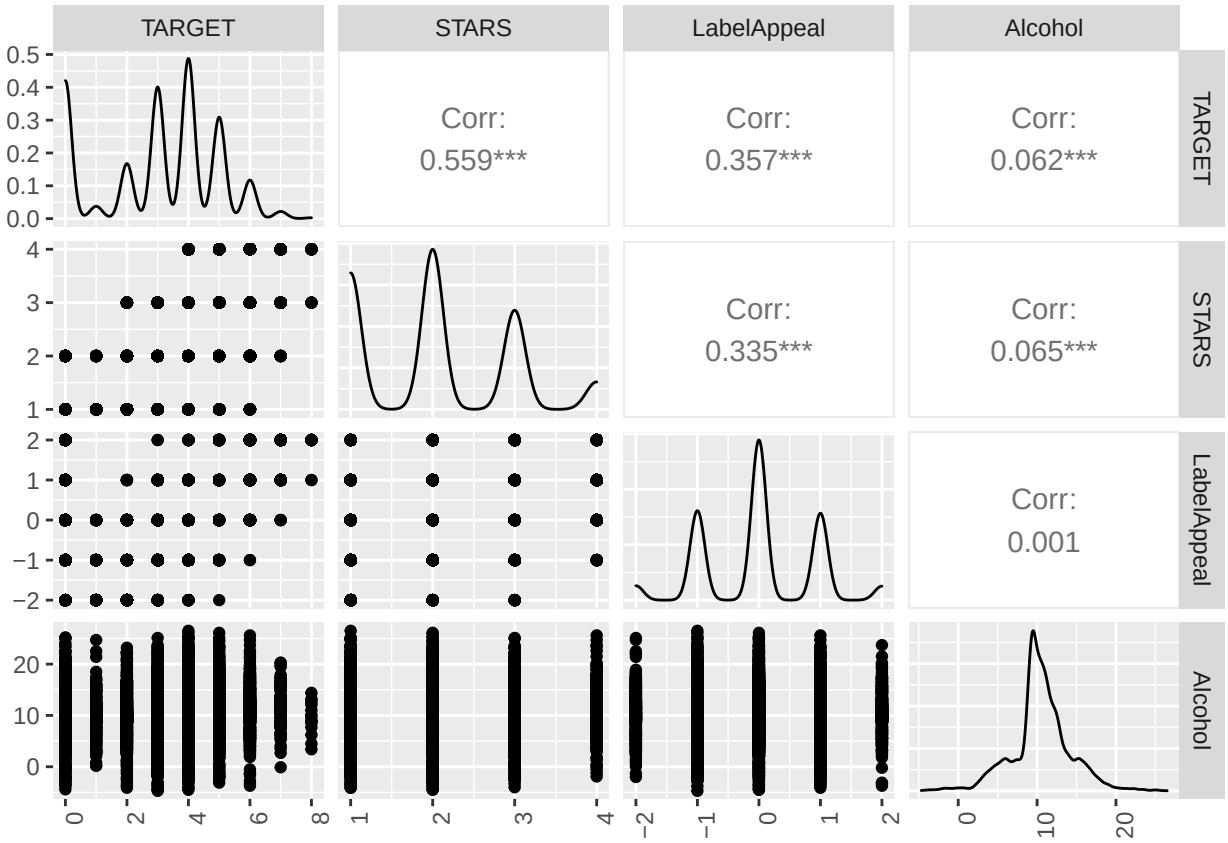
Histogram of Variables



Based on the histograms we can see that a lot of the variables distributions looks to be a normal distribution. We can see that **AcidIndex**, **STARS**, and **TARGET** are a bit skewed. One thing to note is that the **TARGET** variable has a lot of 0 cases sold. These 0 **TARGET** variables will need to be cleaned during the data prep phase as they can skew the results of the model.

Relationship of Predictors to Target

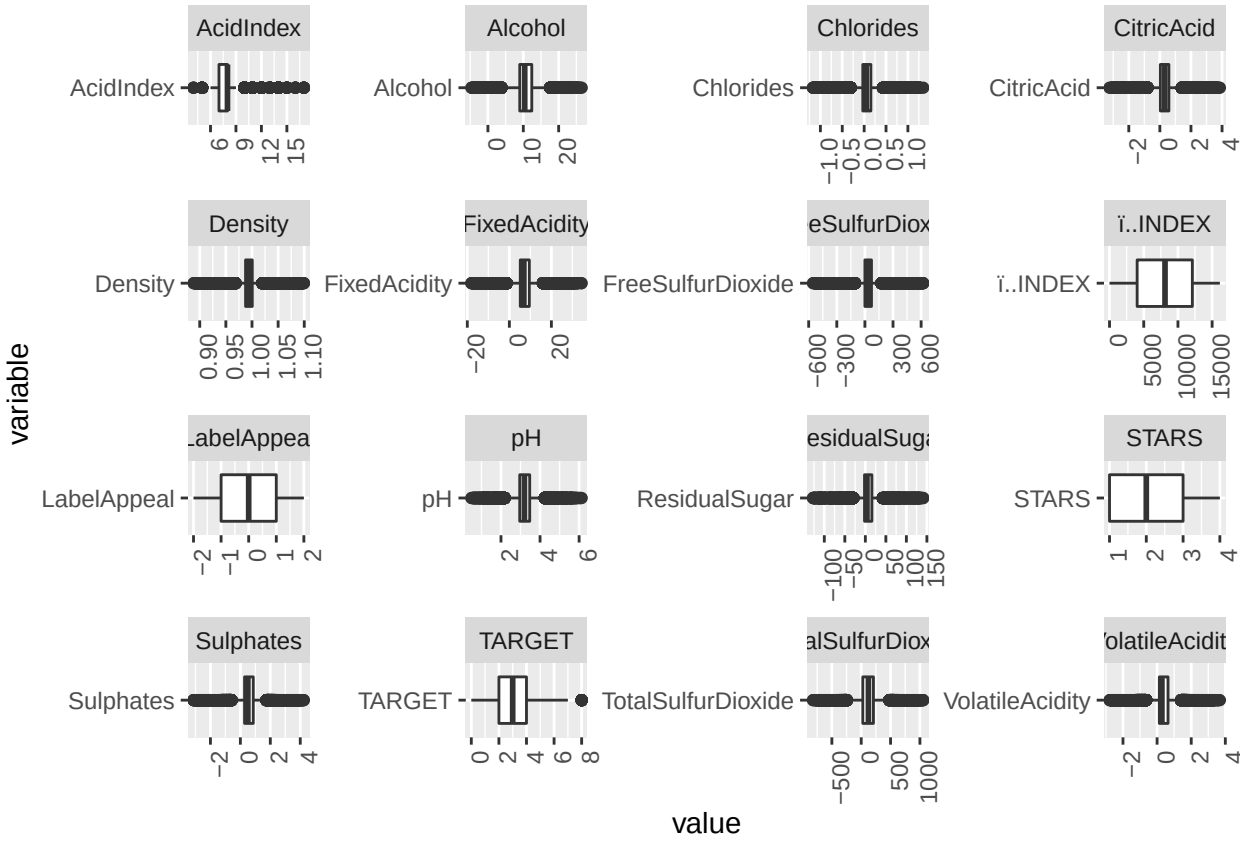
It is useful to assess the plots of each variable against the target variable. Using the GGPairs function from GGally we can plot someof the variables of interest to see if any of the variables correlates with the response variable **TARGET**. We will be making sure to include the variables **STARS** and **LabelAppeal** as it is believed that these two variables affect sales numbers



We can see that **STARS** and **Alcohol** has a bit of correlation with **TARGET**.

Boxplots

After observing our distributions, we can next assess the variables' relationship with our target variable (TARGET).



When looking at the boxplot for **TARGET** we can see a very different picture compared to looking at the histogram. In the histogram it shows all the 0 **TARGETS** which can skew the modeling results while in the histogram one can not easily point that out. This is the reason why we need to look at the data in mulitple different ways.

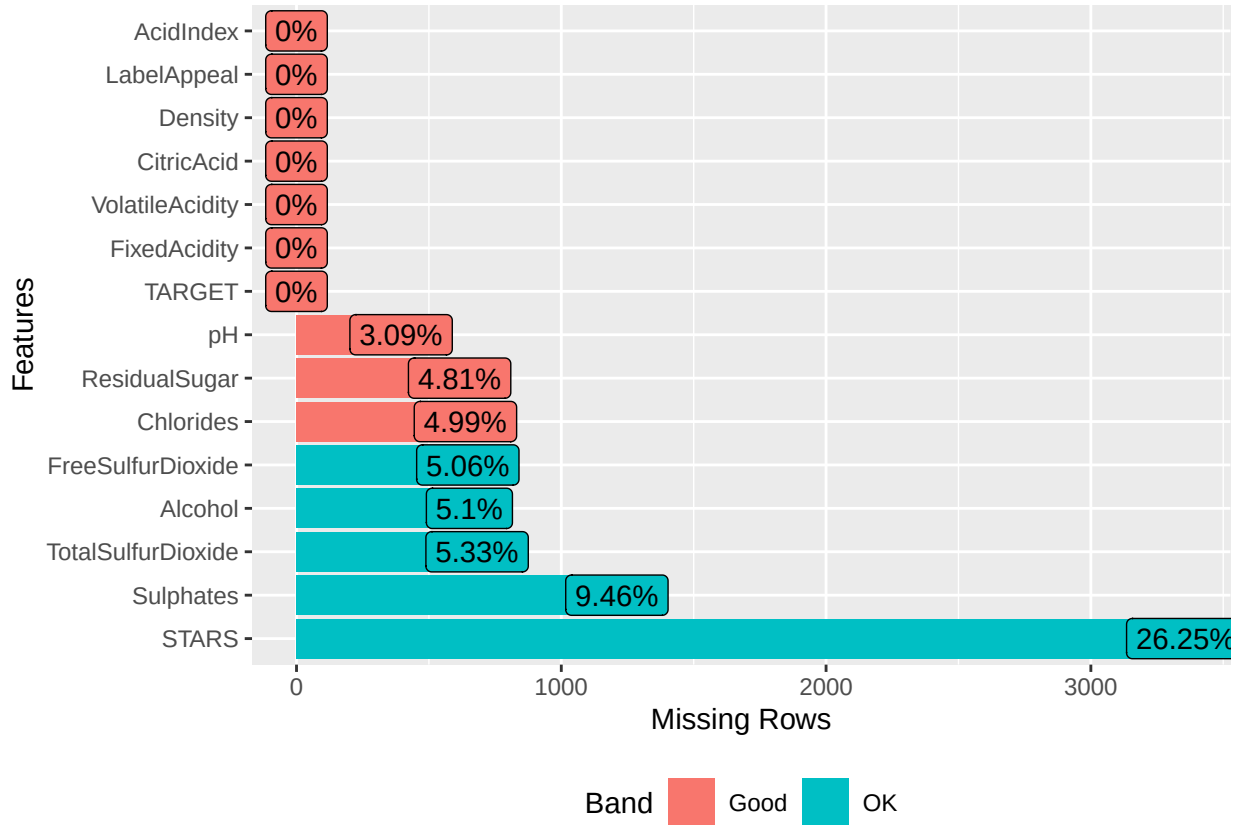
Data Preparation

Identify Missing Values

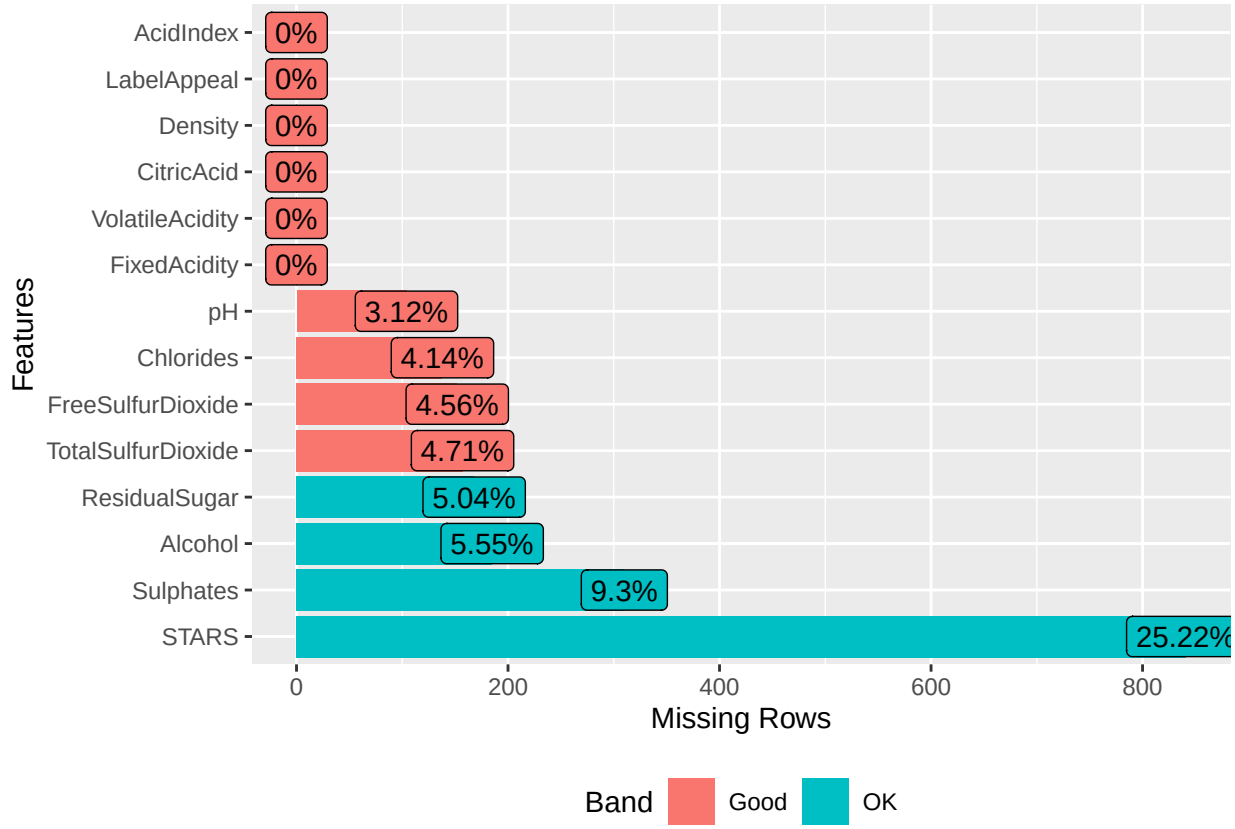
We can see that the same variables for both train data set and evaluation data set contains missing values. The variable that contains the most missing values is the **STAR** followed by **Sulphates**, **Alcohol**, **ResidualSugar**, **TotalSulfurDioxide**, **FreeSulfurDioxide**, **Chlorides**, and **pH**.

```
##    Train Data Set Variales Missing Count
## 1                   TARGET             0
## 2            FixedAcidity             0
## 3         VolatileAcidity             0
## 4              CitricAcid             0
## 5           ResidualSugar           616
## 6               Chlorides           638
## 7       FreeSulfurDioxide           647
## 8      TotalSulfurDioxide           682
## 9                  Density             0
## 10                      pH           395
```

```
## 11               Sulphates          1210
## 12                 Alcohol           653
## 13            LabelAppeal             0
## 14              AcidIndex             0
## 15                  STARS          3359
```




```
##    Train Data Set Variales Missing Count
## 1                  TARGET          3335
## 2           FixedAcidity             0
## 3      VolatileAcidity               0
## 4            CitricAcid              0
## 5         ResidualSugar            168
## 6             Chlorides            138
## 7       FreeSulfurDioxide          152
## 8      TotalSulfurDioxide          157
## 9                 Density             0
## 10                   pH           104
## 11            Sulphates            310
## 12              Alcohol            185
## 13         LabelAppeal               0
## 14           AcidIndex               0
## 15               STARS             841
```

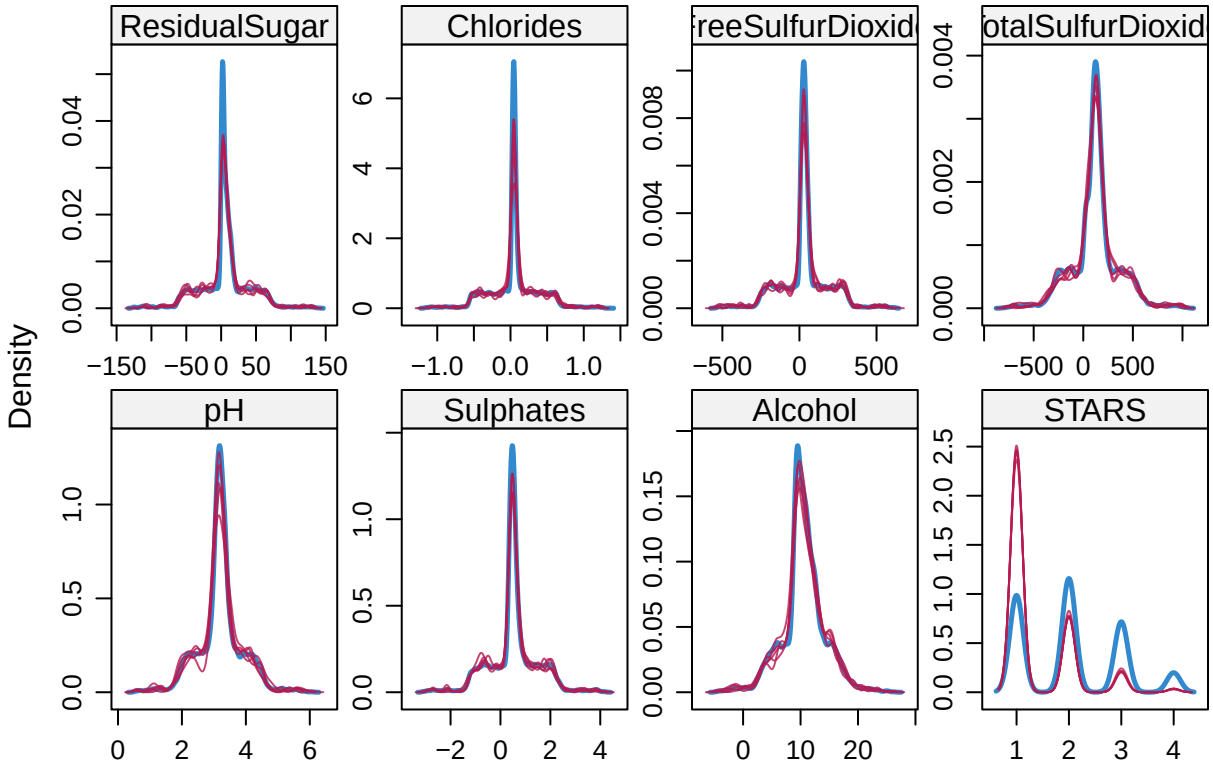


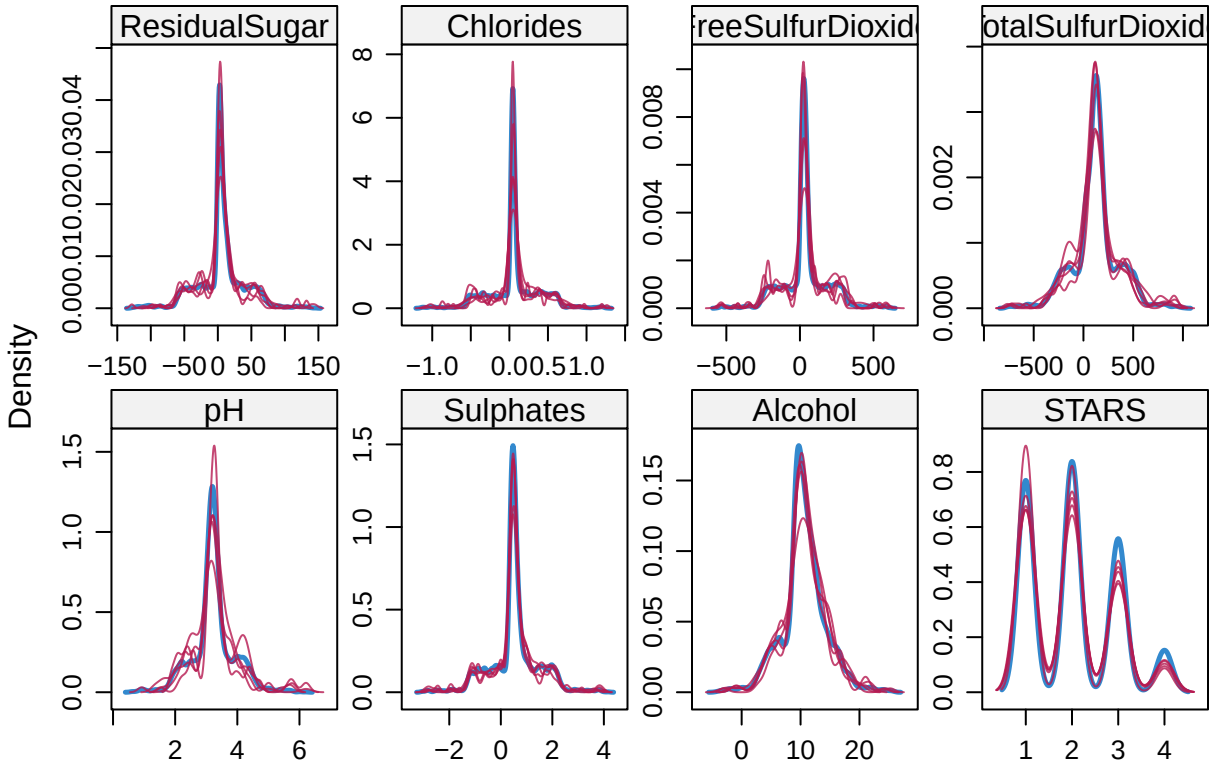
Our chart above does a good job of highlighting the missing values that will no doubt impact our analysis if we don't deal with them. In particular:

- STARS is missing 25% of its records. We could guess that the wines haven't been assessed and rated by experts.
- Sulphates is missing 9% of its records.
- Alcohol is missing 5% of its records. It is unlikely that a 0 here would indicate no alcohol in the wine, given that it's wine, so we can assume these values are missing.
- ResidualSugar is missing 5% of its records. There are some records that have 0 for ResidualSugar so these flagged records most certainly have missing values.
- A few more variables experiences missing values as well such as TotalSulfurDioxide (<5%), FreeSulfurDioxide (<5%), Chlorides (<5%), and pH (<4%).

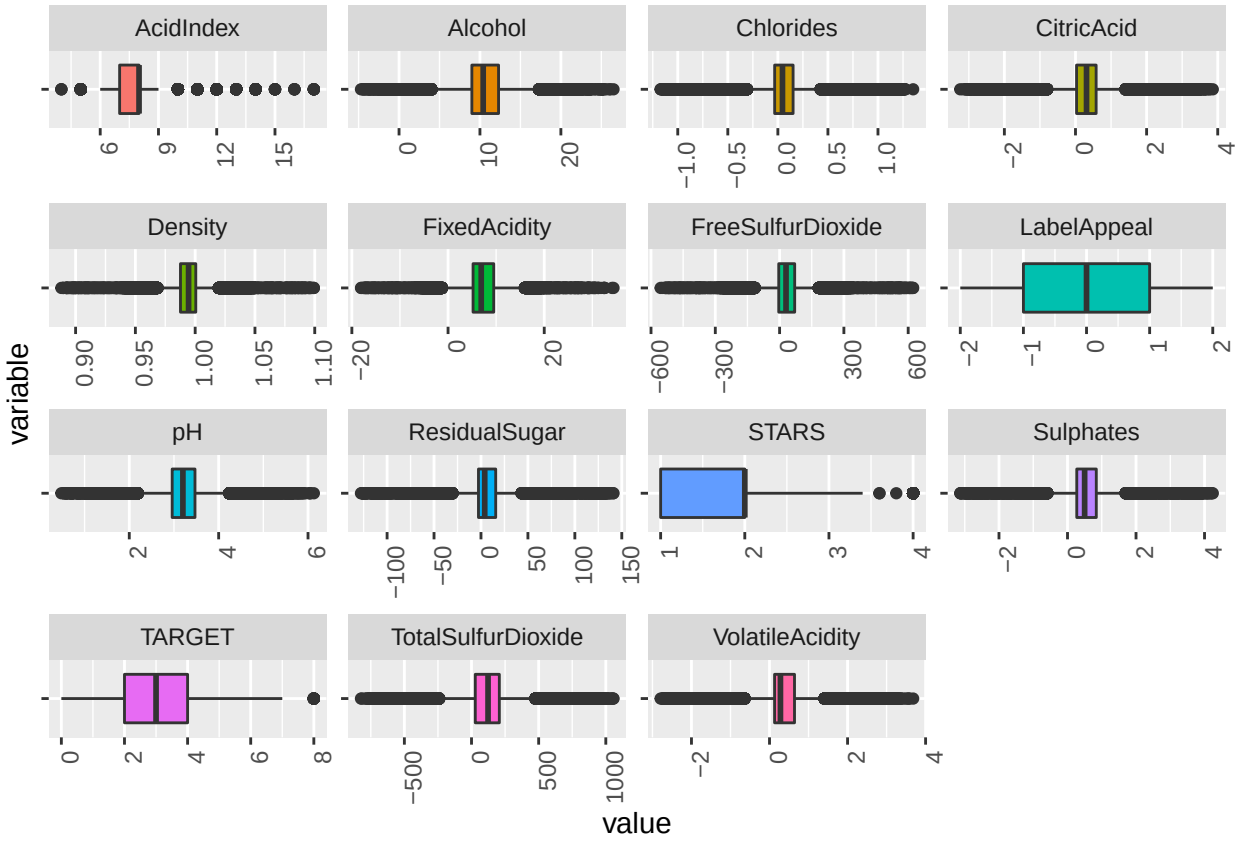
Impute Missing Values

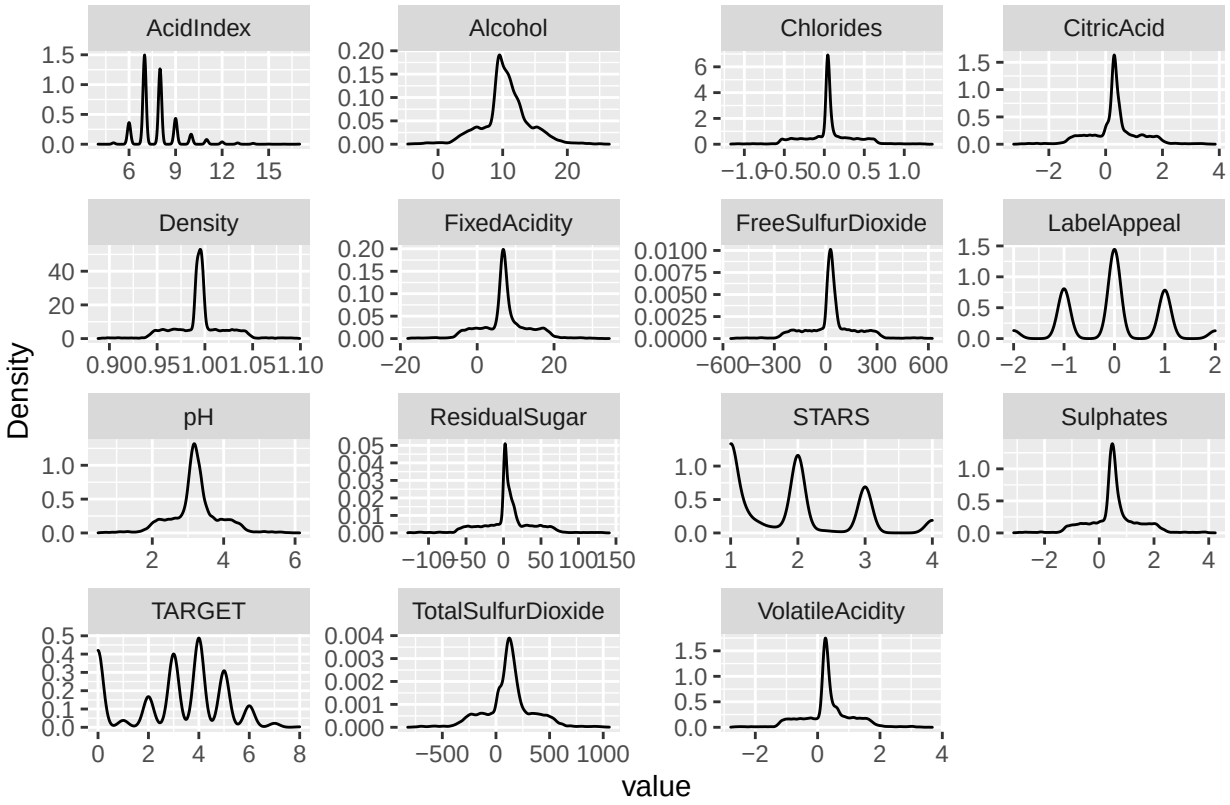
We assume the missing data are Missing at Random and choose to impute. The reason we want to impute the missing data rather than replacing with mean or median because of large number of missing values. If we're replacing with mean or median on the large number of missing values, can result in loss of variation in data. We're imputing the missing data using the MICE package. The method of predictive mean matching (PMM) is selected for continuous variables.


Density
ResidualSugar
Chlorides
reeSulfurDioxid
otalSulfurDioxid
pH
Sulphates
Alcohol
STARS



Next, we take average values of the 5 imputed data set as a final train data set used for building models.

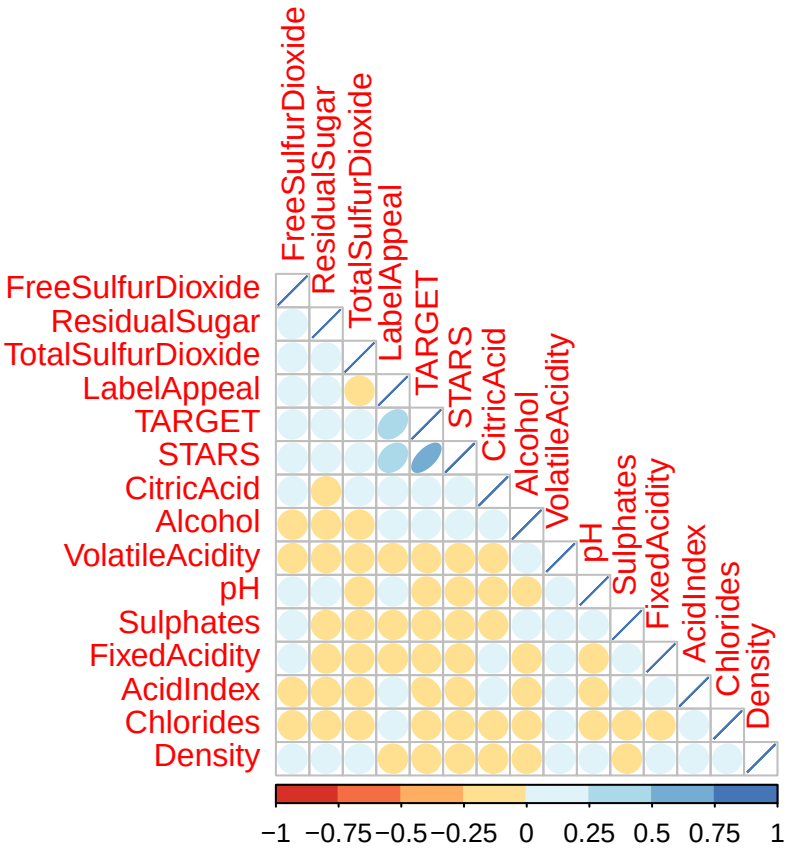

AcidIndex
Alcohol
Chlorides
CitricAcid
Density
FixedAcidity
FreeSulfurDioxide
LabelAppeal
pH
ResidualSugar
STARS
Sulphates
TARGET
TotalSulfurDioxide
VolatileAcidity
variable
value



Because we'll use the Poisson or Negative Binomial regression to build count regression models in GLM approach, transformation for each variable to make them looks normal is not required. Diagnostics of actual outliers or influential points can be identified in the build models section through plots such as residuals vs fitted, standardized residuals vs fitted, etc.

Identifying Multicollinearity

```
##          values                ind
## 1   0.646881500              STARS
## 2   0.356500469        LabelAppeal
## 3   0.063601452            Alcohol
## 4   0.051714048 TotalSulfurDioxide
## 5   0.044495907  FreeSulfurDioxide
## 6   0.015992585      ResidualSugar
## 7   0.008684633         CitricAcid
## 8  -0.009220723                 pH
## 9  -0.035517502            Density
## 10 -0.039429538          Chlorides
## 11 -0.041250213          Sulphates
## 12 -0.049010939       FixedAcidity
## 13 -0.088793212    VolatileAcidity
## 14 -0.246049449          AcidIndex
```



After our EDA and Data Prep, we can render some judgments on the dataset, including that there are some variables with a weak enough relationship with our TARGET that we can plan to drop them. We also need to plan on dealing with the many outliers that could skew our models and pay close attention to multicollinearity, especially as it relates to the relationship between STARS and LabelAppeal. These two features also happen to have the strongest correlation with the TARGET.

Build Models

Poisson Model 1

We start with a quasi-Poisson model using all variables.

```
Model1 <- glm(TARGET ~ ., data=complete_train_data,family=quasipoisson)
summary(Model1)
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = quasipoisson, data = complete_train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8571  -0.6913   0.1207   0.6226   2.6738
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.477e+00  1.833e-01   8.057 8.54e-16 ***
## FixedAcidity        -5.313e-04  7.666e-04  -0.693 0.488320
## VolatileAcidity     -3.910e-02  6.096e-03  -6.414 1.47e-10 ***
## CitricAcid           1.033e-02  5.514e-03   1.873 0.061067 .
## ResidualSugar        5.835e-05  1.439e-04   0.405 0.685132
## Chlorides           -5.080e-02  1.531e-02  -3.319 0.000906 ***
## FreeSulfurDioxide    1.412e-04  3.267e-05   4.323 1.55e-05 ***
## TotalSulfurDioxide   8.434e-05  2.112e-05   3.994 6.53e-05 ***
## Density             -3.413e-01  1.799e-01  -1.897 0.057802 .
## pH                  -1.812e-02  7.118e-03  -2.546 0.010922 *
## Sulphates           -1.532e-02  5.318e-03  -2.881 0.003975 **
## Alcohol              2.676e-03  1.313e-03   2.038 0.041589 *
## LabelAppeal          1.365e-01  5.715e-03  23.880  < 2e-16 ***
## AcidIndex           -9.587e-02  4.231e-03 -22.656  < 2e-16 ***
## STARS                3.582e-01  5.381e-03  66.558  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.874578)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 15774  on 12780  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

We see that the dispersion parameter is close to 1, meaning this regression is close to a regular Poisson regression case.

Using the F test, we check the significance of each of the predictors relative to the full model:

```
drop1(Model1, test="F")
```

```
## Single term deletions
##
## Model:
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +
##     Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
##                     Df Deviance   F value    Pr(>F)
## <none>                   15774
## FixedAcidity         1    15775    0.3403 0.5596804
## VolatileAcidity      1    15810   29.1423 6.845e-08 ***
## CitricAcid           1    15777    2.4863 0.1148631
## ResidualSugar        1    15774    0.1165 0.7328647
## Chlorides            1    15784    7.8048 0.0052185 **
## FreeSulfurDioxide    1    15791   13.2405 0.0002750 ***
## TotalSulfurDioxide   1    15788   11.3018 0.0007766 ***
## Density              1    15777    2.5504 0.1102918
## pH                   1    15780    4.5920 0.0321403 *
## Sulphates            1    15782    5.8794 0.0153328 *
## Alcohol              1    15778    2.9428 0.0862851 .
## LabelAppeal          1    16273  403.8355 < 2.2e-16 ***
```

```
## AcidIndex           1    16246  382.6167 < 2.2e-16 ***
## STARS               1    19543 3053.0393 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Non significant predictors will be dropped and a reduced model is estimated below.

Poisson Model 2

This is the model with only the significant variables at the 5% level.

```
Model2 <- glm(TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + pH + Sulp
                AcidIndex + STARS,data=complete_train_data, family=quasipoisson)
summary(Model2)
```

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + pH + Sulphates + LabelAppeal + AcidIndex +
##     STARS, family = quasipoisson, data = complete_train_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8690  -0.6965   0.1206   0.6212   2.6766
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.169e+00  4.360e-02  26.821  < 2e-16 ***
## VolatileAcidity     -3.939e-02  6.098e-03  -6.459 1.09e-10 ***
## Chlorides           -5.243e-02  1.530e-02  -3.426 0.000614 ***
## FreeSulfurDioxide    1.397e-04  3.267e-05   4.275 1.93e-05 ***
## TotalSulfurDioxide   8.330e-05  2.110e-05   3.947 7.95e-05 ***
## pH                  -1.825e-02  7.119e-03  -2.563 0.010382 *
## Sulphates           -1.539e-02  5.318e-03  -2.894 0.003812 **
## LabelAppeal          1.364e-01  5.717e-03  23.859  < 2e-16 ***
## AcidIndex           -9.654e-02  4.174e-03 -23.130  < 2e-16 ***
## STARS                3.593e-01  5.367e-03  66.942  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.8753499)
##
##     Null deviance: 22861  on 12794  degrees of freedom
## Residual deviance: 15785  on 12785  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

We see there is little practical difference between the two models.

Negative binomial regression

We also estimate a regression using a negative binomial regression.

```
Model3 <- glm.nb(TARGET ~ ., complete_train_data)
summary(Model3)
```

```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = complete_train_data, init.theta = 49193.35395,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8571  -0.6913   0.1207   0.6226   2.6738
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.477e+00  1.960e-01   7.534 4.91e-14 ***
## FixedAcidity        -5.313e-04  8.197e-04  -0.648 0.516935
## VolatileAcidity     -3.910e-02  6.518e-03  -5.998 1.99e-09 ***
## CitricAcid           1.033e-02  5.896e-03   1.752 0.079821 .
## ResidualSugar        5.835e-05  1.539e-04   0.379 0.704534
## Chlorides           -5.080e-02  1.637e-02  -3.104 0.001911 **
## FreeSulfurDioxide    1.412e-04  3.493e-05   4.043 5.28e-05 ***
## TotalSulfurDioxide   8.434e-05  2.258e-05   3.735 0.000188 ***
## Density             -3.413e-01  1.923e-01  -1.774 0.076002 .
## pH                  -1.812e-02  7.611e-03  -2.381 0.017286 *
## Sulphates           -1.532e-02  5.687e-03  -2.694 0.007061 **
## Alcohol              2.676e-03  1.404e-03   1.906 0.056694 .
## LabelAppeal          1.365e-01  6.111e-03  22.332  < 2e-16 ***
## AcidIndex           -9.587e-02  4.525e-03 -21.187  < 2e-16 ***
## STARS                3.582e-01  5.755e-03  62.242  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(49193.35) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 15774  on 12780  degrees of freedom
## AIC: 47749
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  49193
##           Std. Err.:  55792
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -47716.51
```

Zero Inflated Count Models

Considering that the response data has a lot of zeros, we will also estimate a regression using zero inflated count models.

```
Model4<- zeroinfl(TARGET ~ ., data = complete_train_data)
summary(Model4)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ ., data = complete_train_data)
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -1.983021 -0.453229  0.001259  0.403750  6.459031
##
## Count model coefficients (poisson with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.370e+00  2.025e-01   6.767 1.31e-11 ***
## FixedAcidity         2.006e-04  8.416e-04   0.238 0.811564
## VolatileAcidity     -1.158e-02  6.736e-03  -1.719 0.085681 .
## CitricAcid           7.397e-04  6.038e-03   0.122 0.902506
## ResidualSugar       -8.595e-05  1.582e-04  -0.543 0.587026
## Chlorides           -2.231e-02  1.687e-02  -1.322 0.186025
## FreeSulfurDioxide    3.138e-05  3.528e-05   0.890 0.373635
## TotalSulfurDioxide -2.307e-05  2.254e-05  -1.024 0.305910
## Density            -3.042e-01  1.986e-01  -1.531 0.125692
## pH                   5.824e-03  7.850e-03   0.742 0.458106
## Sulphates           -2.918e-04  5.872e-03  -0.050 0.960364
## Alcohol              6.982e-03  1.434e-03   4.869 1.12e-06 ***
## LabelAppeal          2.305e-01  6.370e-03  36.189  < 2e-16 ***
## AcidIndex           -1.671e-02  4.887e-03  -3.420 0.000626 ***
## STARS                1.282e-01  6.413e-03  19.995  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.9047850  1.2126749  -1.571 0.116245
## FixedAcidity         0.0039855  0.0050301   0.792 0.428176
## VolatileAcidity      0.2149929  0.0395880   5.431 5.61e-08 ***
## CitricAcid          -0.0570646  0.0366172  -1.558 0.119136
## ResidualSugar       -0.0006423  0.0009427  -0.681 0.495631
## Chlorides            0.2528988  0.1009827   2.504 0.012267 *
## FreeSulfurDioxide   -0.0008383  0.0002185  -3.837 0.000125 ***
## TotalSulfurDioxide -0.0008014  0.0001397  -5.736 9.70e-09 ***
## Density              0.8825322  1.1936681   0.739 0.459698
## pH                   0.1910630  0.0466618   4.095 4.23e-05 ***
## Sulphates            0.1525424  0.0355848   4.287 1.81e-05 ***
## Alcohol              0.0203647  0.0085429   2.384 0.017135 *
## LabelAppeal          0.6812879  0.0390907  17.428  < 2e-16 ***
## AcidIndex            0.4359032  0.0235896  18.479  < 2e-16 ***
## STARS               -3.4137368  0.0969201 -35.222  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Number of iterations in BFGS optimization: 36
## Log-likelihood: -2.081e+04 on 30 Df
```

As per our text book, we will also estimate a simplified version considering two components: non-count and count variables.

```
Model5 <- zeroinfl(TARGET ~ .|STARS, data = complete_train_data)
summary(Model5)
```

```
##
## Call:
## zeroinfl(formula = TARGET ~ . | STARS, data = complete_train_data)
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -2.14441 -0.50160  0.03202  0.43128  2.16179
##
## Count model coefficients (poisson with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.472e+00  2.020e-01   7.284 3.23e-13 ***
## FixedAcidity        1.122e-04  8.391e-04   0.134   0.8937
## VolatileAcidity    -1.457e-02  6.719e-03  -2.169   0.0301 *
## CitricAcid          1.741e-03  6.019e-03   0.289   0.7723
## ResidualSugar      -6.660e-05  1.577e-04  -0.422   0.6728
## Chlorides          -2.651e-02  1.682e-02  -1.576   0.1150
## FreeSulfurDioxide   4.276e-05  3.519e-05   1.215   0.2243
## TotalSulfurDioxide -9.955e-06  2.248e-05  -0.443   0.6579
## Density            -2.940e-01  1.980e-01  -1.485   0.1377
## pH                  3.058e-03  7.823e-03   0.391   0.6959
## Sulphates          -2.560e-03  5.857e-03  -0.437   0.6621
## Alcohol             6.614e-03  1.430e-03   4.626 3.74e-06 ***
## LabelAppeal         2.203e-01  6.373e-03  34.573  < 2e-16 ***
## AcidIndex          -2.900e-02  5.056e-03  -5.735 9.77e-09 ***
## STARS               1.279e-01  6.438e-03  19.870  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.95313    0.10273   28.75   <2e-16 ***
## STARS       -3.10022    0.08614  -35.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 22
## Log-likelihood: -2.132e+04 on 17 Df
```

Model Selection

In order to select a model, we will compare various metrics for all five models using the training data set. We check models' confusion matrix and its measures such as accuracy, classification error rate, precision, sensitivity, specificity, F1 score, AUC, and also MAE, RMSE, and R-squared.

	Model 1	Model 2	Model 3	Model 4	Model 5
Accuracy	0.1626417	0.1642829	0.1626417	0.2790152	0.2309496
Class. Error Rate	0.8373583	0.8357171	0.8373583	0.7209848	0.7690504
Sensitivity	0.0051207	0.0047549	0.0051207	0.1901975	0.0007315
Specificity	0.8278689	0.8278689	0.8278689	0.7254098	0.8401639
Precision	0.1583255	0.1592698	0.1583255	0.2609380	0.2757318
F1	0.1098039	0.1163399	0.1098039	0.1764706	0.1738562
AUC	0.8722909	0.8727158	0.8722905	0.8819315	0.8884974
MAE	1.2194137	1.2200277	1.2194169	1.0423916	1.1194941
RMSE	1.4640443	1.4648278	1.4640476	1.3306027	1.3914400
R2	0.4260053	0.4253606	0.4260041	0.5231204	0.4790180

In the table above, we see that a our Zero Inflated Count Models have the highest accuracy, R-squared and lowest MAE and RMSE. We'll make our predictions using the Zero Inflated Count - Model4.

Make Predictions

We show below a table of the fitted values, using the predictions based on Model4:

```
##
##    0    1    2    3    4    5    6    7    8
##  194 1490 3598 2704 2964 1226  449  136   34
```

And here it is the training set target value distribution:

```
table(complete_train_data$TARGET)
```

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

Appendix

- Link to full code : https://github.com/ahussan/DATA_621_Group1/blob/main/HW5/HW5.Rmd
- Link to the predicted values over test set : https://github.com/ahussan/DATA_621_Group1/blob/main/HW5/HW5preds.csv