Data 621 - Homework 5

Group 2: William Aiken, Donald Butler, Michael Ippolito, Bharani Nittala, and Leticia Salazar

December 11, 2022

Contents

Overview:

We 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 larger 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

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 face that a variable is missing is actually predictive of the target.

Description

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

VARIABLE NAME:	DEFINITION:	THEORETICAL EFFECT:
INDEX	Identification Variable (do not use)	None
TARGET	Number of Cases Purchased	None
AcidIndex	Proprietary method of testing totalacidity of wine by using a weighted average	
Alcohol	Alcohol Content	
Chorides	Cholride content of wine	
CitricAcid	Citric Acid Content	
Density	Density of Wine	
FixedAcidity	Fixed Acidity of Wiine	
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 customers don't like the design.	Many consumers purchase based on the visual appeal of the wine label design. Higher numbers suggest better sales.
ResidualSugar	Residual Sugar of wine	
STARS	Wine rating by a team of experts: 4 Stars = Excellent, 1 Star = Poor	A high number of stars suggests high sales
Sulphates	Sulfate content of Wine	
TotalSulfurDioxide	Total Sulfur Dioxide of Wine	
VolatileAcidity	Volatile Acid content of wine	
pH	pH of wine	

Load Libraries:

These are the libraries used to explore, prepare, analyze and build our models

```
library(tidyverse)
library(dplyr)
library(corrplot)
library(skimr)
library(DataExplorer)
library(ggplot2)
library(hrbrthemes)
library(mice)
library(MASS)
library(dvmisc)
library(gridExtra)
library(lattice)
```

Load Data set:

We have included the original data sets in our GitHub account and read from this location. Below we are showing the training data set:

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

Data Exploration:

Using the `summary()` function lets start exploring the training and evaluation data.

Training:

```
##      INDEX           TARGET       FixedAcidity     VolatileAcidity
##  Min.   :    1   Min.   :0.000   Min.   :-18.100   Min.   :-2.7900
##  1st Qu.: 4038   1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300
##  Median : 8110   Median :3.000   Median :  6.900   Median : 0.2800
##  Mean   : 8070   Mean   :3.029   Mean   :  7.076   Mean   : 0.3241
##  3rd Qu.:12106   3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400
##  Max.   :16129   Max.   :8.000   Max.   : 34.400   Max.   : 3.6800
##
##    CitricAcid      ResidualSugar       Chlorides       FreeSulfurDioxide
##  Min.   :-3.2400   Min.   :-127.800   Min.   :-1.1710   Min.   :-555.00
##  1st Qu.: 0.0300   1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00
##  Median : 0.3100   Median :   3.900   Median : 0.0460   Median :  30.00
##  Mean   : 0.3084   Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85
##  3rd Qu.: 0.5800   3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00
##  Max.   : 3.8600   Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00
##                    NA’s   :616        NA’s   :638       NA’s   :647
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-823.0     Min.   :0.8881   Min.   :0.480   Min.   :-3.1300
##  1st Qu.:  27.0     1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800
##  Median : 123.0     Median :0.9945   Median :3.200   Median : 0.5000
##  Mean   : 120.7     Mean   :0.9942   Mean   :3.208   Mean   : 0.5271
##  3rd Qu.: 208.0     3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600
##  Max.   :1057.0     Max.   :1.0992   Max.   :6.130   Max.   : 4.2400
##  NA's   :682                         NA's   :395     NA's   :1210
##     Alcohol       LabelAppeal          AcidIndex          STARS
##  Min.   :-4.70   Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.: 9.00   1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median :10.40   Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :10.49   Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.:12.40   3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   :26.50   Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##  NA's   :653                                          NA's   :3359
```

Evaluation:

```
##        IN         TARGET         FixedAcidity     VolatileAcidity
##  Min.   :    3   Mode:logical   Min.   :-18.200   Min.   :-2.8300
##  1st Qu.: 4018   NA's:3335      1st Qu.:  5.200   1st Qu.: 0.0800
##  Median : 7906                  Median :  6.900   Median : 0.2800
##  Mean   : 8048                  Mean   :  6.864   Mean   : 0.3103
##  3rd Qu.:12061                  3rd Qu.:  9.000   3rd Qu.: 0.6300
##  Max.   :16130                  Max.   : 33.500   Max.   : 3.6100
##
##    CitricAcid      ResidualSugar       Chlorides        FreeSulfurDioxide
##  Min.   :-3.1200   Min.   :-128.300   Min.   :-1.15000   Min.   :-563.00
##  1st Qu.: 0.0000   1st Qu.:  -2.600   1st Qu.: 0.01600   1st Qu.:   3.00
##  Median : 0.3100   Median :   3.600   Median : 0.04700   Median :  30.00
##  Mean   : 0.3124   Mean   :   5.319   Mean   : 0.06143   Mean   :  34.95
##  3rd Qu.: 0.6050   3rd Qu.:  17.200   3rd Qu.: 0.17100   3rd Qu.:  79.25
##  Max.   : 3.7600   Max.   : 145.400   Max.   : 1.26300   Max.   : 617.00
##                    NA's   :168        NA's   :138        NA's   :152
##  TotalSulfurDioxide    Density             pH          Sulphates
##  Min.   :-769.00    Min.   :0.8898   Min.   :0.600   Min.   :-3.0700
##  1st Qu.:  27.25    1st Qu.:0.9883   1st Qu.:2.980   1st Qu.: 0.3300
##  Median : 124.00    Median :0.9946   Median :3.210   Median : 0.5000
##  Mean   : 123.41    Mean   :0.9947   Mean   :3.237   Mean   : 0.5346
##  3rd Qu.: 210.00    3rd Qu.:1.0005   3rd Qu.:3.490   3rd Qu.: 0.8200
##  Max.   :1004.00    Max.   :1.0998   Max.   :6.210   Max.   : 4.1800
##  NA's   :157                         NA's   :104     NA's   :310
##     Alcohol       LabelAppeal         AcidIndex          STARS
##  Min.   :-4.20   Min.   :-2.00000   Min.   : 5.000   Min.   :1.00
##  1st Qu.: 9.00   1st Qu.:-1.00000   1st Qu.: 7.000   1st Qu.:1.00
##  Median :10.40   Median : 0.00000   Median : 8.000   Median :2.00
##  Mean   :10.58   Mean   : 0.01349   Mean   : 7.748   Mean   :2.04
##  3rd Qu.:12.50   3rd Qu.: 1.00000   3rd Qu.: 8.000   3rd Qu.:3.00
##  Max.   :25.60   Max.   : 2.00000   Max.   :17.000   Max.   :4.00
##  NA's   :185                                         NA's   :841
```

Using the `DataExplorer` package we use the `create_report` function which pulls a full data profile from our training data set and create an html file with basic statistics, structure, missing data, distribution visualizations, correlation matrix and principal component analysis for our data. You can find these output in our github.

```
# Do not render since it will produce a separate html file
# Remove TARGET from eval report since it will contain all
# NAs and will make the correlation plot fail to render
DataExplorer::create_report(dftrain, output_file = "training_report.html")
DataExplorer::create_report(dfeval %>%
    select(-TARGET), output_file = "eval_report.html")
```

Based on this our training data includes 12795 records and 16 variables whereas the evaluation data includes 3335 records and 16 variables.

Training:

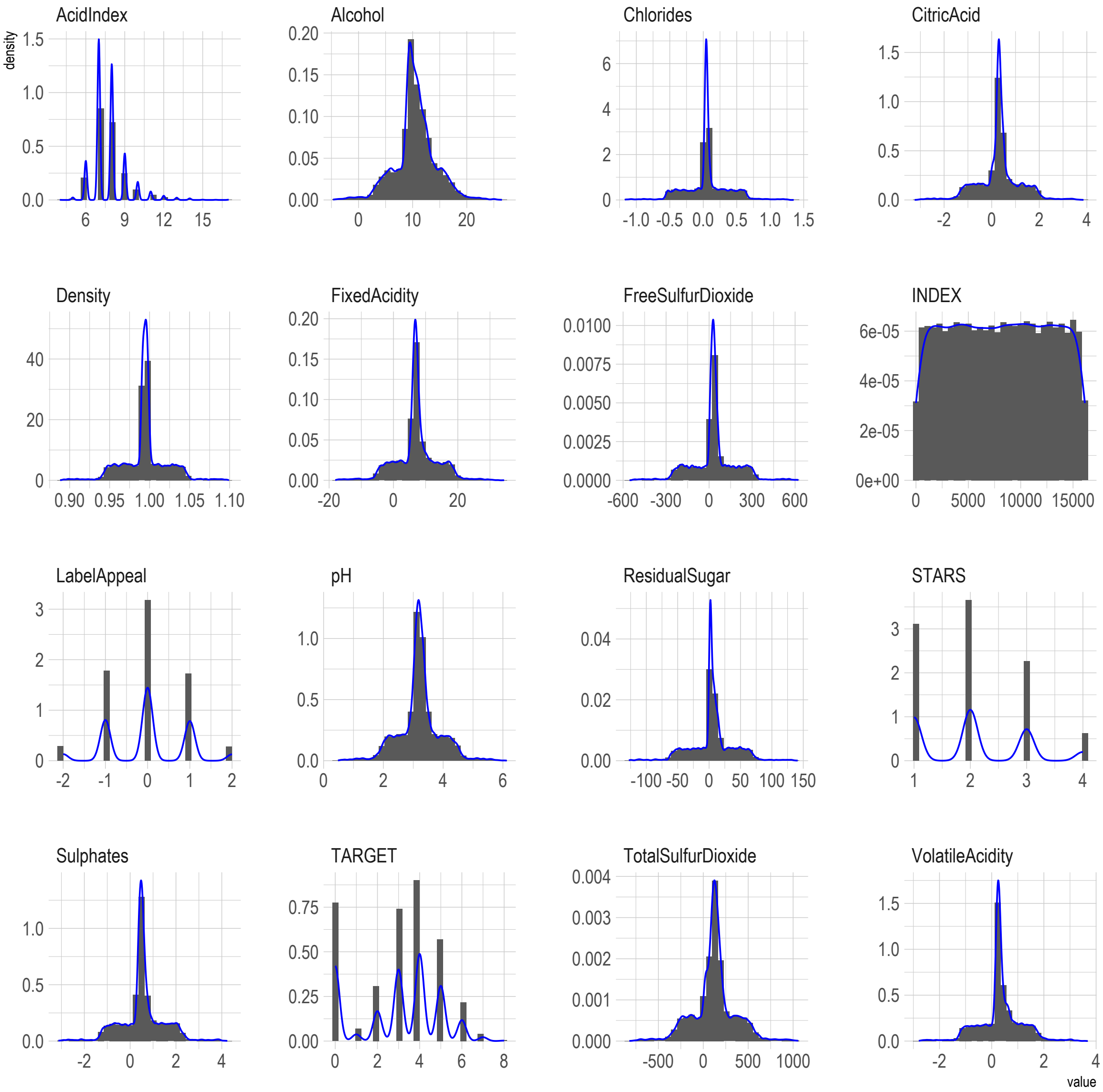
```
## 'data.frame':    12795 obs. of  16 variables:
##  $ INDEX             : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET            : int  3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity      : num  3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity   : num  1.16 0.16 2.64 0.385 0.33 0.32 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 ...
```

Evaluation:

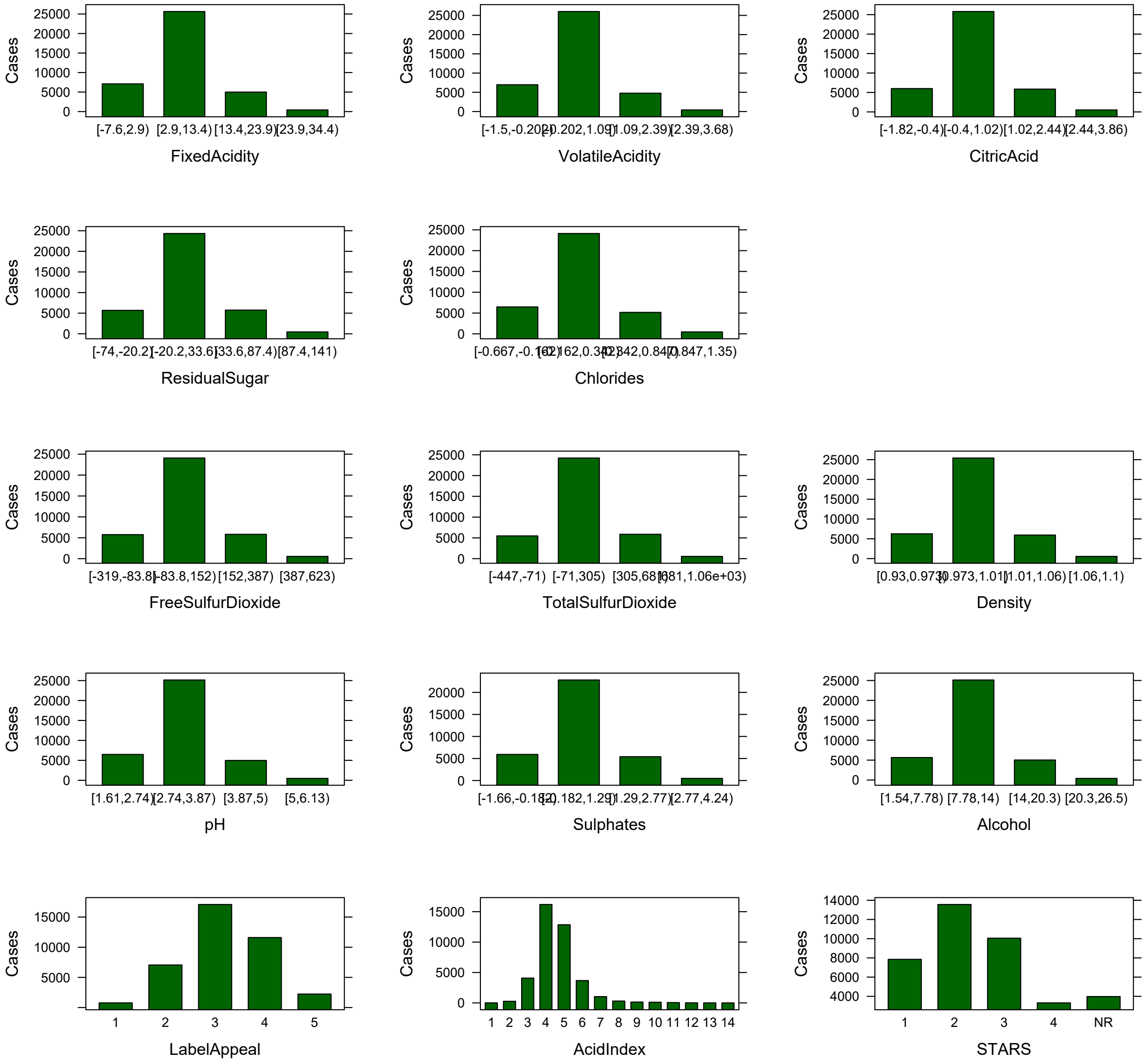
```
## 'data.frame':    3335 obs. of  16 variables:
##  $ IN                : int  3 9 10 18 21 30 31 37 39 47 ...
##  $ TARGET            : logi  NA NA NA NA NA NA ...
##  $ FixedAcidity      : num  5.4 12.4 7.2 6.2 11.4 17.6 15.5 15.9 11.6 3.8 ...
##  $ VolatileAcidity   : num  -0.86 0.385 1.75 0.1 0.21 0.04 0.53 1.19 0.32 0.22 ...
##  $ CitricAcid        : num  0.27 -0.76 0.17 1.8 0.28 -1.15 -0.53 1.14 0.55 0.31 ...
##  $ ResidualSugar     : num  -10.7 -19.7 -33 1 1.2 1.4 4.6 31.9 -50.9 -7.7 ...
##  $ Chlorides         : num  0.092 1.169 0.065 -0.179 0.038 ...
##  $ FreeSulfurDioxide : num  23 -37 9 104 70 -250 10 115 35 40 ...
##  $ TotalSulfurDioxide: num  398 68 76 89 53 140 17 381 83 129 ...
##  $ Density           : num  0.985 0.99 1.046 0.989 1.029 ...
##  $ pH                : num  5.02 3.37 4.61 3.2 2.54 3.06 3.07 2.99 3.32 4.72 ...
##  $ Sulphates         : num  0.64 1.09 0.68 2.11 -0.07 -0.02 0.75 0.31 2.18 -0.64 ...
##  $ Alcohol           : num  12.3 16 8.55 12.3 4.8 11.4 8.5 11.4 -0.5 10.9 ...
##  $ LabelAppeal       : int  -1 0 0 -1 0 1 0 1 0 0 ...
##  $ AcidIndex         : int  6 6 8 8 10 8 12 7 12 7 ...
##  $ STARS             : int  NA 2 1 1 NA 4 3 NA NA NA ...
```

Lets take a look at the distribution of each variables in the training data set.

Based on the plots below, we can tell that most of the variables seem to be normally distributed with the exception of `AcidIndex` and `STARS` being right skewed. `INDEX` shows a uniform distribution but has no effect on our data so during the data preparation stage we will be removing it.



The fact that some wines are not rated could be a potential predictor. We'll treat NAs as its own star rating. We'll also look at the number of cases of wine sold against the predictors.



As shown, more cases of wine are sold for mid-range values of all categories of acidity, sugar, chlorides, the dioxides, density, pH, sulphates, and alcohol. Surprisingly, more cases were sold for labels that had mid-range label appeal. A lower acid index seemed to indicate more cases sold. And more cases were sold for wines rated only two stars, indicating that consumers may consider higher-starred wines as too pricey.

Data Preparation:

Data preparation was performed on both the training and evaluation data sets but will only be displayed for the training data. We'll also need to removing the `INDEX` variable.

Now we'll impute missing values using R's Multiple Imputation by Chained Equations (MICE) package. We'll avoid imputing the STARS variable as the absence of a star rating may be a significant predictor.

```
##
##  iter imp variable
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   1   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   1   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   1   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   1   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   2   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   2   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   2   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   2   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   3   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   3   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   3   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   3   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   4   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   4   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   4   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   4   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   5   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   5   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   5   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   5   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
```

```
##
##  iter imp variable
##   1   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   1   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   1   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   1   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   1   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   2   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   2   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   2   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   2   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   2   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   3   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   3   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   3   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   3   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   3   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   4   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   4   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   4   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   4   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   4   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   5   1  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   5   2  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   5   3  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   5   4  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
##   5   5  ResidualSugar  Chlorides  FreeSulfurDioxide  TotalSulfurDioxide  pH  Sulphates  Alcohol
```

Lets look at another summary to make sure there aren’t any NAs where we’re not expecting them.

Training data:

```
##      TARGET       FixedAcidity     VolatileAcidity     CitricAcid
##  Min.   :0.000   Min.   :-18.100   Min.   :-2.7900   Min.   :-3.2400
##  1st Qu.:2.000   1st Qu.:  5.200   1st Qu.: 0.1300   1st Qu.: 0.0300
##  Median :3.000   Median :  6.900   Median : 0.2800   Median : 0.3100
##  Mean   :3.029   Mean   :  7.076   Mean   : 0.3241   Mean   : 0.3084
##  3rd Qu.:4.000   3rd Qu.:  9.500   3rd Qu.: 0.6400   3rd Qu.: 0.5800
##  Max.   :8.000   Max.   : 34.400   Max.   : 3.6800   Max.   : 3.8600
##  ResidualSugar        Chlorides        FreeSulfurDioxide TotalSulfurDioxide
##  Min.   :-127.800   Min.   :-1.17100   Min.   :-555.00   Min.   :-823.0
##  1st Qu.:  -2.100   1st Qu.:-0.02900   1st Qu.:  -1.00   1st Qu.:  27.0
##  Median :   3.900   Median : 0.04600   Median :  30.00   Median : 124.0
##  Mean   :   5.428   Mean   : 0.05525   Mean   :  30.84   Mean   : 120.7
##  3rd Qu.:  16.000   3rd Qu.: 0.15250   3rd Qu.:  70.00   3rd Qu.: 208.0
##  Max.   : 141.150   Max.   : 1.35100   Max.   : 623.00   Max.   :1057.0
##     Density             pH          Sulphates          Alcohol
##  Min.   :0.8881   Min.   :0.480   Min.   :-3.1300   Min.   :-4.70
##  1st Qu.:0.9877   1st Qu.:2.960   1st Qu.: 0.2800   1st Qu.: 9.00
##  Median :0.9945   Median :3.200   Median : 0.5000   Median :10.40
##  Mean   :0.9942   Mean   :3.208   Mean   : 0.5269   Mean   :10.49
##  3rd Qu.:1.0005   3rd Qu.:3.470   3rd Qu.: 0.8600   3rd Qu.:12.40
##  Max.   :1.0992   Max.   :6.130   Max.   : 4.2400   Max.   :26.50
##   LabelAppeal          AcidIndex         STARS
##  Min.   :-2.000000   Min.   : 4.000   Length:12795
##  1st Qu.:-1.000000   1st Qu.: 7.000   Class :character
##  Median : 0.000000   Median : 8.000   Mode  :character
##  Mean   :-0.009066   Mean   : 7.773
##  3rd Qu.: 1.000000   3rd Qu.: 8.000
##  Max.   : 2.000000   Max.   :17.000
```

Evaluation data:

```
##   FixedAcidity     VolatileAcidity     CitricAcid      ResidualSugar
##  Min.   :-18.200   Min.   :-2.8300   Min.   :-3.1200   Min.   :-128.300
##  1st Qu.:  5.200   1st Qu.: 0.0800   1st Qu.: 0.0000   1st Qu.:  -2.600
##  Median :  6.900   Median : 0.2800   Median : 0.3100   Median :   3.600
##  Mean   :  6.864   Mean   : 0.3103   Mean   : 0.3124   Mean   :   5.225
##  3rd Qu.:  9.000   3rd Qu.: 0.6300   3rd Qu.: 0.6050   3rd Qu.:  17.150
##  Max.   : 33.500   Max.   : 3.6100   Max.   : 3.7600   Max.   : 145.400
##    Chlorides       FreeSulfurDioxide TotalSulfurDioxide    Density
##  Min.   :-1.1500   Min.   :-563.00   Min.   :-769.0     Min.   :0.8898
##  1st Qu.: 0.0155   1st Qu.:   3.00   1st Qu.:  28.0     1st Qu.:0.9883
##  Median : 0.0470   Median :  29.00   Median : 124.0     Median :0.9946
##  Mean   : 0.0624   Mean   :  34.34   Mean   : 123.9     Mean   :0.9947
##  3rd Qu.: 0.1740   3rd Qu.:  79.00   3rd Qu.: 210.0     3rd Qu.:1.0005
##  Max.   : 1.2630   Max.   : 617.00   Max.   :1004.0     Max.   :1.0998
##        pH          Sulphates          Alcohol       LabelAppeal
##  Min.   :0.600   Min.   :-3.0700   Min.   :-4.20   Min.   :-2.00000
##  1st Qu.:2.980   1st Qu.: 0.3300   1st Qu.: 9.00   1st Qu.:-1.00000
##  Median :3.210   Median : 0.5000   Median :10.40   Median : 0.00000
##  Mean   :3.235   Mean   : 0.5326   Mean   :10.59   Mean   : 0.01349
##  3rd Qu.:3.480   3rd Qu.: 0.8150   3rd Qu.:12.50   3rd Qu.: 1.00000
##  Max.   :6.210   Max.   : 4.1800   Max.   :25.60   Max.   : 2.00000
##    AcidIndex         STARS             TARGET
##  Min.   : 5.000   Length:3335        Mode:logical
##  1st Qu.: 7.000   Class :character   NA's:3335
##  Median : 8.000   Mode  :character
##  Mean   : 7.748
##  3rd Qu.: 8.000
##  Max.   :17.000
```

Build Models:

Based on the data, we'll try two model types: a poisson general linear model and a Gaussian multiple linear model.

Poisson Models:

- Possion Model 1

```
##
## Call:
## glm(formula = TARGET ~ ., family = "poisson", data = cleandf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2780  -0.6619  -0.0015   0.4504   3.7616
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.880e+00  1.951e-01   9.636  < 2e-16 ***
## FixedAcidity        5.726e-05  8.196e-04   0.070 0.944295
## VolatileAcidity    -3.058e-02  6.528e-03  -4.684 2.81e-06 ***
## CitricAcid          4.970e-03  5.896e-03   0.843 0.399265
## ResidualSugar       7.227e-05  1.507e-04   0.479 0.631607
## Chlorides          -4.361e-02  1.609e-02  -2.710 0.006735 **
## FreeSulfurDioxide   9.543e-05  3.402e-05   2.805 0.005034 **
## TotalSulfurDioxide  8.066e-05  2.215e-05   3.641 0.000271 ***
## Density            -2.730e-01  1.918e-01  -1.423 0.154601
## pH                 -1.289e-02  7.550e-03  -1.707 0.087742 .
## Sulphates          -1.284e-02  5.474e-03  -2.346 0.018956 *
## Alcohol             3.470e-03  1.375e-03   2.523 0.011626 *
## LabelAppeal         1.595e-01  6.127e-03  26.031  < 2e-16 ***
## AcidIndex          -7.973e-02  4.573e-03 -17.434  < 2e-16 ***
## STARS2              3.220e-01  1.434e-02  22.454  < 2e-16 ***
## STARS3              4.405e-01  1.562e-02  28.203  < 2e-16 ***
## STARS4              5.556e-01  2.167e-02  25.640  < 2e-16 ***
## STARSNR            -7.666e-01  1.954e-02 -39.234  < 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: 13647  on 12777  degrees of freedom
## AIC: 45625
##
## Number of Fisher Scoring iterations: 6
```

- Possion Model with stepwise AIC approach

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS, family = "poisson", data = cleandf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2803  -0.6604  -0.0027   0.4510   3.7603
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.882e+00  1.951e-01   9.646  < 2e-16 ***
## VolatileAcidity    -3.071e-02  6.527e-03  -4.706 2.53e-06 ***
## Chlorides          -4.380e-02  1.609e-02  -2.721 0.006503 **
## FreeSulfurDioxide   9.585e-05  3.402e-05   2.817 0.004842 **
## TotalSulfurDioxide  8.085e-05  2.214e-05   3.651 0.000261 ***
## Density            -2.750e-01  1.918e-01  -1.434 0.151591
## pH                 -1.280e-02  7.548e-03  -1.696 0.089814 .
## Sulphates          -1.289e-02  5.472e-03  -2.355 0.018525 *
## Alcohol             3.487e-03  1.375e-03   2.536 0.011209 *
## LabelAppeal         1.595e-01  6.127e-03  26.040  < 2e-16 ***
## AcidIndex          -7.945e-02  4.518e-03 -17.585  < 2e-16 ***
## STARS2              3.222e-01  1.434e-02  22.475  < 2e-16 ***
## STARS3              4.405e-01  1.562e-02  28.210  < 2e-16 ***
## STARS4              5.558e-01  2.167e-02  25.650  < 2e-16 ***
## STARSNR            -7.668e-01  1.954e-02 -39.244  < 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: 13648  on 12780  degrees of freedom
## AIC: 45620
##
## Number of Fisher Scoring iterations: 6
```

Multiple Linear Regression Models:

- MLR Model 1

```
##
## Call:
## lm(formula = TARGET ~ ., data = cleandf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8479 -0.8590  0.0251  0.8458  6.1615
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.056e+00  4.419e-01  11.440  < 2e-16 ***
## FixedAcidity        5.629e-04  1.858e-03   0.303 0.761966
## VolatileAcidity    -9.472e-02  1.477e-02  -6.412 1.49e-10 ***
## CitricAcid          1.700e-02  1.343e-02   1.266 0.205593
## ResidualSugar       2.473e-04  3.420e-04   0.723 0.469637
## Chlorides          -1.337e-01  3.640e-02  -3.673 0.000241 ***
## FreeSulfurDioxide   2.829e-04  7.758e-05   3.647 0.000266 ***
## TotalSulfurDioxide  2.317e-04  4.984e-05   4.648 3.39e-06 ***
## Density            -7.980e-01  4.357e-01  -1.831 0.067053 .
## pH                 -3.304e-02  1.706e-02  -1.937 0.052746 .
## Sulphates          -3.394e-02  1.239e-02  -2.740 0.006147 **
## Alcohol             1.156e-02  3.114e-03   3.713 0.000205 ***
## LabelAppeal         4.674e-01  1.363e-02  34.299  < 2e-16 ***
## AcidIndex          -1.997e-01  9.097e-03 -21.949  < 2e-16 ***
## STARS2              1.031e+00  3.256e-02  31.671  < 2e-16 ***
## STARS3              1.600e+00  3.765e-02  42.510  < 2e-16 ***
## STARS4              2.292e+00  5.965e-02  38.422  < 2e-16 ***
## STARSNR            -1.361e+00  3.291e-02 -41.369  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.306 on 12777 degrees of freedom
## Multiple R-squared:  0.5412, Adjusted R-squared:  0.5406
## F-statistic: 886.7 on 17 and 12777 DF,  p-value: < 2.2e-16
```

- MLR Model 2

```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS, data = cleandf)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8483 -0.8620  0.0239  0.8436  6.1561
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.064e+00  4.419e-01  11.460  < 2e-16 ***
## VolatileAcidity    -9.509e-02  1.477e-02  -6.438 1.25e-10 ***
## Chlorides          -1.344e-01  3.640e-02  -3.693 0.000223 ***
## FreeSulfurDioxide   2.848e-04  7.756e-05   3.672 0.000241 ***
## TotalSulfurDioxide  2.328e-04  4.983e-05   4.672 3.02e-06 ***
## Density            -8.060e-01  4.357e-01  -1.850 0.064311 .
## pH                 -3.300e-02  1.706e-02  -1.935 0.053009 .
## Sulphates          -3.414e-02  1.238e-02  -2.757 0.005835 **
## Alcohol             1.158e-02  3.113e-03   3.722 0.000199 ***
## LabelAppeal         4.674e-01  1.363e-02  34.302  < 2e-16 ***
## AcidIndex          -1.984e-01  8.939e-03 -22.198  < 2e-16 ***
## STARS2              1.032e+00  3.255e-02  31.702  < 2e-16 ***
## STARS3              1.601e+00  3.764e-02  42.525  < 2e-16 ***
## STARS4              2.293e+00  5.965e-02  38.441  < 2e-16 ***
## STARSNR            -1.362e+00  3.290e-02 -41.384  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.306 on 12780 degrees of freedom
## Multiple R-squared:  0.5411, Adjusted R-squared:  0.5406
## F-statistic:  1077 on 14 and 12780 DF,  p-value: < 2.2e-16
```

Select Models:

In this section, an optimal model will be selected based on its performance when trained on the data. To select the models, we'll use AIC and MSE to measure accuracy of the predicted values.

Below, the Poisson and Multiple Linear Regression models have been compared to select the model with the lowest AIC.

Comparison of Poisson Models:

We'll need to compare the AIC's of each Poisson Model.

Poisson Model 1:

```
## [1] 45625.22
```

Poisson Model 2:

```
## [1] 45620.16
```

Poisson Model 2 proves to have the lower AIC of the two, with a 33947.74 AIC. Below is the formula for Poisson Model 2.

```
## [[1]]
## TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     Density + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex +
##     STARS
```

Comparsion of Multiple Linar Models:

We'll need to compare the Adjusted R Squares of each Linear Model.

Linear Model 1:

```
## [1] 0.5406183
```

Linear Model 2:

```
## [1] 0.5406471
```

Linear Model 2 proves to have the higher Adjusted R Squares, with a value of 0.4544041. Below is the formula for Linear Model 2.

```
## [[1]]
## TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     Density + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex +
##     STARS
```

Mean Square Error:

The Mean Square Error measures the averaged square different between the estimated values and the actual value. The lower the value of the MSE, the more accurately the model is able to predict the values.

$$\text{MSE} = \frac{1}{n}\sum(y-\hat{y})^2$$

Comparison of Possion and Gaussian Linear Models:

By evaluating the AIC's and MSE's of each model, we can choose the best one be looking at the lowest AIC and lowest MSE.

	Possion Model 1	Possion Model 2	Linear Model 1	Linear Model 2
MSE	6.7060661648579	6.70614723918115	1.70471690181985	1.70461008756085
AIC	45625.2226362434	45620.1579756524	43155.4653254402	43151.6674641144

Based on the above, the linear model has better model statistics than the poisson model.

Prediction from optimal multiple linear regression model:

```
## # A tibble: 10 x 15
##    Fixed~1 Volat~2 Citri~3 Resid~4 Chlor~5 FreeS~6 Total~7 Density    pH Sulph~8
##      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <dbl>
##  1     5.4  -0.86     0.27   -10.7   0.092      23     398   0.985  5.02    0.64
##  2    12.4   0.385   -0.76   -19.7   1.17       -37      68   0.990  3.37    1.09
##  3     7.2   1.75     0.17   -33     0.065       9      76   1.05   4.61    0.68
##  4     6.2   0.1      1.8       1    -0.179     104      89   0.989  3.2     2.11
##  5    11.4   0.21     0.28     1.2   0.038      70      53   1.03   2.54   -0.07
##  6    17.6   0.04    -1.15     1.4   0.535    -250     140   0.950  3.06   -0.02
##  7    15.5   0.53    -0.53     4.6   1.26        10      17   1.00   3.07    0.75
##  8    15.9   1.19     1.14    31.9  -0.299     115     381   1.03   2.99    0.31
##  9    11.6   0.32     0.55   -50.9   0.076      35      83   1.00   3.32    2.18
## 10     3.8   0.22     0.31    -7.7   0.039      40     129   0.906  4.72   -0.64
## # ... with 5 more variables: Alcohol <dbl>, LabelAppeal <int>, AcidIndex <int>,
## #   STARS <chr>, TARGET <dbl>, and abbreviated variable names 1: FixedAcidity,
## #   2: VolatileAcidity, 3: CitricAcid, 4: ResidualSugar, 5: Chlorides,
## #   6: FreeSulfurDioxide, 7: TotalSulfurDioxide, 8: Sulphates
```

Appendix:

```
# load libaries
library(tidyverse)
library(dplyr)
library(corrplot)
library(skimr)
library(DataExplorer)
library(ggplot2)
library(hrbrthemes)
library(mice)

# load data
dftrain <- read.csv("https://raw.githubusercontent.com/letisalba/Data_621/master/Homework_5/csv/wine-tra
dfeval <- read.csv("https://raw.githubusercontent.com/letisalba/Data_621/master/Homework_5/csv/wine-eval
head(dftrain)

# summary of training and evaluation data sets
summary(dftrain)
summary(dfeval)

# Do not render since it will produce a separate html file
# Remove TARGET from eval report since it will contain all
# NAs and will make the correlation plot fail to render
DataExplorer::create_report(dftrain, output_file = "training_report.html")
DataExplorer::create_report(dfeval %>%
    select(-TARGET), output_file = "eval_report.html")

# structure of training and evaluation data
str(dftrain)
str(dfeval)
```

```r
# plotting distribution of training data
plot_train <- dftrain %>%
    gather(key = "variable", value = "value")

ggplot(plot_train) + geom_histogram(aes(x = value, y = ..density..),
    bins = 30) + geom_density(aes(x = value), color = "blue") +
    theme_ipsum() + facet_wrap(. ~ variable, scales = "free",
    ncol = 4)

# Create logical variable to indicate whether there is a
# star rating for this wine
dftrain <- dftrain %>%
    mutate(STARS = ifelse(is.na(STARS), "NR", STARS))
dfeval <- dfeval %>%
    mutate(STARS = ifelse(is.na(STARS), "NR", STARS))

# Look at the number of cases of wine sold against the
# predictors.
plt <- vector("list", ncol(dftrain) - 1)
for (i in seq(3, 16)) {
    # skip INDEX and TARGET variables
    if (class(dftrain[, i]) == "numeric") {
        tmpmin <- min(dftrain[, i], na.rm = T)
        tmpinterval <- (max(dftrain[, i], na.rm = T) - tmpmin)/5
        tmpcuts <- c()
        for (j in seq(1, 5)) {
            tmpcuts <- c(tmpcuts, tmpmin + (j * tmpinterval))
        }
        # dftrain$x <- dftrain[, i] %>% cut(breaks=5,
        # ordered_result=T, right=F)
        dftrain$x <- dftrain[, i] %>%
            cut(breaks = tmpcuts, ordered_result = T, right = F)
    } else {
        dftrain$x <- dftrain[, i]
    }
    dftmp <- dftrain %>%
        group_by(x) %>%
        summarize(ct = sum(TARGET))
    plt[[i]] <- barchart(dftmp$ct ~ dftmp$x, horiz = F, col = "darkgreen",
        xlab = colnames(dftrain)[i], ylab = "Cases")
}
dftrain <- subset(dftrain, select = -x)  # remove temporary variable
grid.arrange(grobs = plt[3:7], ncol = 3, nrow = 2)
grid.arrange(grobs = plt[8:13], ncol = 3, nrow = 2)
grid.arrange(grobs = plt[14:16], ncol = 3, nrow = 2)

# Removing INDEX from training and eval data For some
# reason R renamed the INDEX column to 'ï..INDEX'
dftrain <- dftrain %>%
    dplyr::select(-ï..INDEX)
dfeval <- dfeval %>%
    dplyr::select(-IN)
```

```
# Impute missing values in training data
dftrain_imputed <- mice(dftrain, m = 5, maxit = 5, method = "pmm")
cleandf <- complete(dftrain_imputed) %>%
    mutate(STARS = dftrain$STARS)

# Impute missing values in eval data (except for TARGET)
dfeval_imputed <- mice(dfeval %>%
    select(-TARGET), m = 5, maxit = 5, method = "pmm")
cleandf_eval <- complete(dfeval_imputed) %>%
    mutate(STARS = dfeval$STARS, TARGET = dfeval$TARGET)

# Look at another summary to make sure there aren't any NAs
# where we're not expecting them
summary(cleandf)
summary(cleandf_eval)

# Poisson model
p_mod1 <- glm(TARGET ~ ., family = "poisson", data = cleandf)
summary(p_mod1)

# Possion Model with stepwise AIC approach
p_mod2 <- stepAIC(p_mod1, trace = F)
summary(p_mod2)

# Multiple Linear Regression Models:

# MLR Model 1
lm_mod1 <- lm(TARGET ~ ., data = cleandf)
aic_lm_mod1 = AIC(lm_mod1)
summary(lm_mod1)

# MLR Model 2
lm_mod2 <- stepAIC(lm_mod1, trace = F)
aic_lm_mod2 = AIC(lm_mod2)
summary(lm_mod2)

# Select Models:

# Comparison of Poisson Models:

# Poisson Model 1:
aic_p_mod1 <- p_mod1$aic
aic_p_mod1

# Poisson Model 2:
aic_p_mod2 <- p_mod2$aic
aic_p_mod2

# Poisson - Minimum AIC
c(p_mod1$formula, p_mod2$formula)[which.min(c(p_mod1$aic, p_mod2$aic))]

# Comparsion of Multiple Linar Models:
```

```r
# Linear Model 1:
r2_lm_mod1 <- summary(lm_mod1)$adj.r.squared
r2_lm_mod1

# Linear Model 2:
r2_lm_mod2 <- summary(lm_mod2)$adj.r.squared
r2_lm_mod2

# Multiple Linear Regression Model - Highest Adjusted R
# Squared
c(formula(lm_mod1), formula(lm_mod2))[which.max(c(summary(lm_mod1)$adj.r.squared,
    summary(lm_mod2)$adj.r.squared))]

# Mean Square Error:
mse <- function(df, model) {
    mean((df$TARGET - predict(model))^2)
}
mse_p_mod1 <- mse(cleandf, p_mod1)
mse_p_mod2 <- mse(cleandf, p_mod2)
mse_lm_mod1 <- get_mse(lm_mod1)
mse_lm_mod2 <- get_mse(lm_mod2)

# Comparison of Possion and Negative Binomial Model's:
models <- c("Possion Model 1", "Possion Model 2", "Linear Model 1",
    "Linear Model 2")
# rows <- c('Models', 'MSE', 'AIC')
MSE <- list(mse_p_mod1, mse_p_mod2, mse_lm_mod1, mse_lm_mod2)
AIC <- list(aic_p_mod1, aic_p_mod2, aic_lm_mod1, aic_lm_mod2)
knitr::kable(rbind(MSE, AIC), col.names = models)

# Prediction from optimal multiple linear regression model
prob2 <- predict(lm_mod2, cleandf_eval, interval = "prediction")
cleandf_eval$TARGET <- prob2[, 1]
cleandf_eval %>%
    head(10) %>%
    as_tibble()
write.csv(cleandf_eval, "wine_predictions2.csv", row.names = FALSE)
```

References:

https://englianhu.files.wordpress.com/2016/01/faraway-extending-the-linear-model-with-r-e28093-2006.pdf