Homework-5

Waheeb Algabri, Joe Garcia, Lwin Shwe, Mikhail Broomes

2024-04-29

Introduction

Data is the key to making informed decisions and achieving success in modern business. We'll begin our analysis by examining the dataset for outliers, missing data, potential encoding errors, multicollinearity etc., then, we'll implement any required data cleaning procedures. Once we've prepared a reliable dataset, we'll construct and assess three distinct linear models to forecast sales. Our dataset comprises both training and evaluation data; we'll train the models using the primary training dataset and then assess their performance against the separate evaluation dataset. Finally, we'll choose a best model that strikes the optimal balance between accuracy and simplicity.

Data Exploration

```
train_df <- read.csv('https://raw.githubusercontent.com/waheeb123/Data-621/main/Homeworks/Homework%205/
```

```
## Warning: `as.tibble()` was deprecated in tibble 2.0.0.
## i Please use `as_tibble()` instead.
## i The signature and semantics have changed, see `?as_tibble`.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
evaluate_df <- read.csv('https://raw.githubusercontent.com/waheeb123/Data-621/main/Homeworks/Homework%20
train_df$INDEX <- NULL # Remove index column
evaluate_df$IN <- NULL
#evaluate_df$TARGET <- NULL
str(train_df)
```

```
## tibble [12,795 x 15] (S3: tbl_df/tbl/data.frame)
##  $ TARGET            : int [1:12795] 3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity      : num [1:12795] 3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity   : num [1:12795] 1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
##  $ CitricAcid        : num [1:12795] -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...
##  $ ResidualSugar     : num [1:12795] 54.2 26.1 14.8 18.8 9.4 ...
##  $ Chlorides         : num [1:12795] -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...
##  $ FreeSulfurDioxide : num [1:12795] NA 15 214 22 -167 -37 287 523 -213 62 ...
##  $ TotalSulfurDioxide: num [1:12795] 268 -327 142 115 108 15 156 551 NA 180 ...
##  $ Density           : num [1:12795] 0.993 1.028 0.995 0.996 0.995 ...
##  $ pH                : num [1:12795] 3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...
```

```
##  $ Sulphates         : num [1:12795] -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...
##  $ Alcohol           : num [1:12795] 9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...
##  $ LabelAppeal       : int [1:12795] 0 -1 -1 -1 0 0 0 1 0 0 ...
##  $ AcidIndex         : int [1:12795] 8 7 8 6 9 11 8 7 6 8 ...
##  $ STARS             : int [1:12795] 2 3 3 1 2 NA NA 3 NA 4 ...
```

```
str(evaluate_df)
```

```
## tibble [3,335 x 15] (S3: tbl_df/tbl/data.frame)
##  $ TARGET            : logi [1:3335] NA NA NA NA NA NA ...
##  $ FixedAcidity      : num [1:3335] 5.4 12.4 7.2 6.2 11.4 17.6 15.5 15.9 11.6 3.8 ...
##  $ VolatileAcidity   : num [1:3335] -0.86 0.385 1.75 0.1 0.21 0.04 0.53 1.19 0.32 0.22 ...
##  $ CitricAcid        : num [1:3335] 0.27 -0.76 0.17 1.8 0.28 -1.15 -0.53 1.14 0.55 0.31 ...
##  $ ResidualSugar     : num [1:3335] -10.7 -19.7 -33 1 1.2 1.4 4.6 31.9 -50.9 -7.7 ...
##  $ Chlorides         : num [1:3335] 0.092 1.169 0.065 -0.179 0.038 ...
##  $ FreeSulfurDioxide : num [1:3335] 23 -37 9 104 70 -250 10 115 35 40 ...
##  $ TotalSulfurDioxide: num [1:3335] 398 68 76 89 53 140 17 381 83 129 ...
##  $ Density           : num [1:3335] 0.985 0.99 1.046 0.989 1.029 ...
##  $ pH                : num [1:3335] 5.02 3.37 4.61 3.2 2.54 3.06 3.07 2.99 3.32 4.72 ...
##  $ Sulphates         : num [1:3335] 0.64 1.09 0.68 2.11 -0.07 -0.02 0.75 0.31 2.18 -0.64 ...
##  $ Alcohol           : num [1:3335] 12.3 16 8.55 12.3 4.8 11.4 8.5 11.4 -0.5 10.9 ...
##  $ LabelAppeal       : int [1:3335] -1 0 0 -1 0 1 0 1 0 0 ...
##  $ AcidIndex         : int [1:3335] 6 6 8 8 10 8 12 7 12 7 ...
##  $ STARS             : int [1:3335] NA 2 1 1 NA 4 3 NA NA NA ...
```

The training data set has 12,795 rows and 15 columns: 14 features and 1 response variable, TARGET. The variable `INDEX` is used for observed identification. Twelve of the features describe the chemical properties of wine. The remaining two predictors are rating variables: LabelAppeal refers to the perceived attractiveness of a wine's product label, while `STARS` is an assessment of wine quality. The output variable, `TARGET`, is a count measure indicating the number of wine case purchases by distributors.

```
library(vtable)
st(train_df)
```

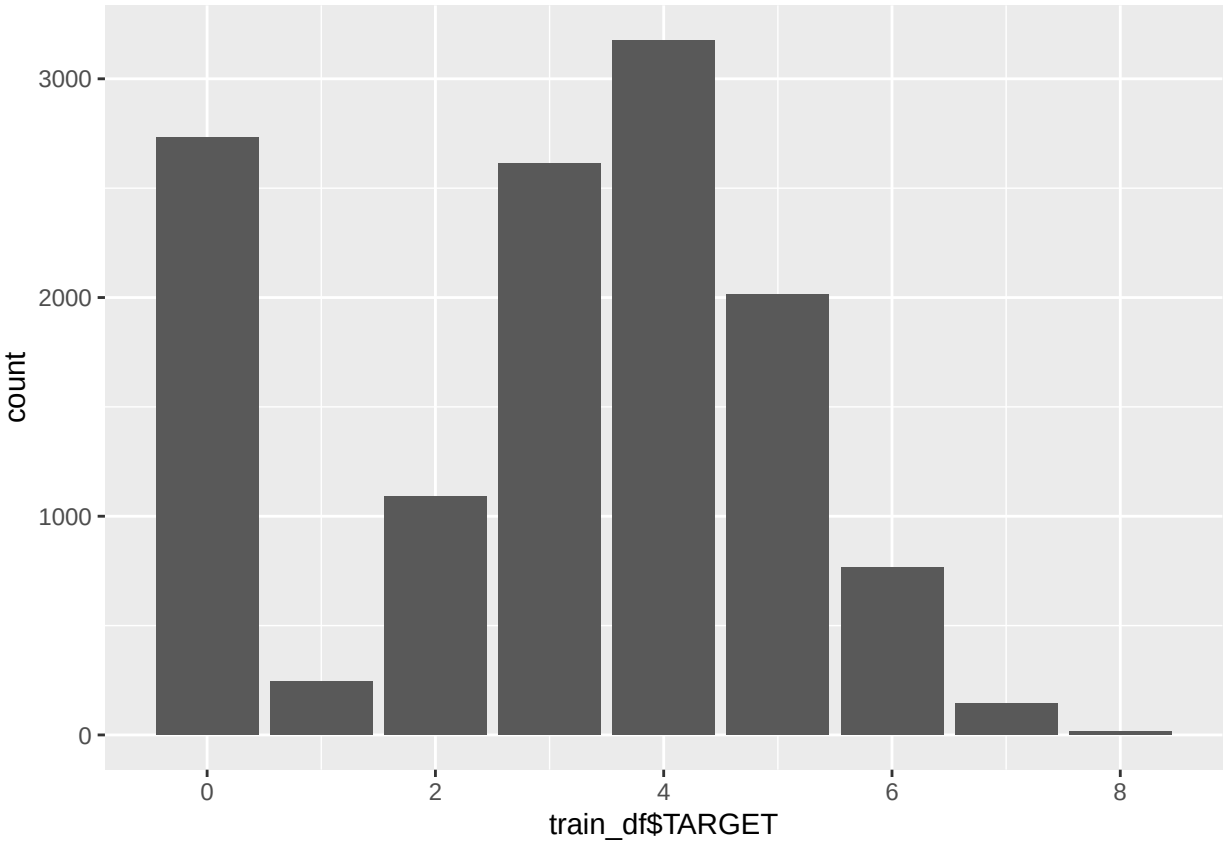
Predictor Variables

Of the 14 feature columns, 8 of them such as `ResidualSugar`, `Chlorides`, `FreeSulfurDioxide`, `TotalSulfurDioxide`, `pH`, `Sulphates`, `Alcohol`, and `STARS` variables contain several missing values. `LabelAppeal`, `AcidIndex`, and `STARS` are discrete variables (i.e categorical) and the rest are continuous. We also noticed that several numerical features representing chemical quantities in the wine exhibit negative minimum values. We hypothesize that the original chemical measurements might have been normalized (potentially via a log transform), allowing for negative values. However, from a physical standpoint, negative concentrations shouldn't be possible. Despite this, we've opted to retain these values as they are.

Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
TARGET	12795	3	1.9	0	2	4	8
FixedAcidity	12795	7.1	6.3	-18	5.2	9.5	34
VolatileAcidity	12795	0.32	0.78	-2.8	0.13	0.64	3.7
CitricAcid	12795	0.31	0.86	-3.2	0.03	0.58	3.9
ResidualSugar	12179	5.4	34	-128	-2	16	141
Chlorides	12157	0.055	0.32	-1.2	-0.031	0.15	1.4
FreeSulfurDioxide	12148	31	149	-555	0	70	623
TotalSulfurDioxide	12113	121	232	-823	27	208	1057
Density	12795	0.99	0.027	0.89	0.99	1	1.1
pH	12400	3.2	0.68	0.48	3	3.5	6.1
Sulphates	11585	0.53	0.93	-3.1	0.28	0.86	4.2
Alcohol	12142	10	3.7	-4.7	9	12	26
LabelAppeal	12795	-0.0091	0.89	-2	-1	1	2
AcidIndex	12795	7.8	1.3	4	7	8	17
STARS	9436	2	0.9	1	1	3	4

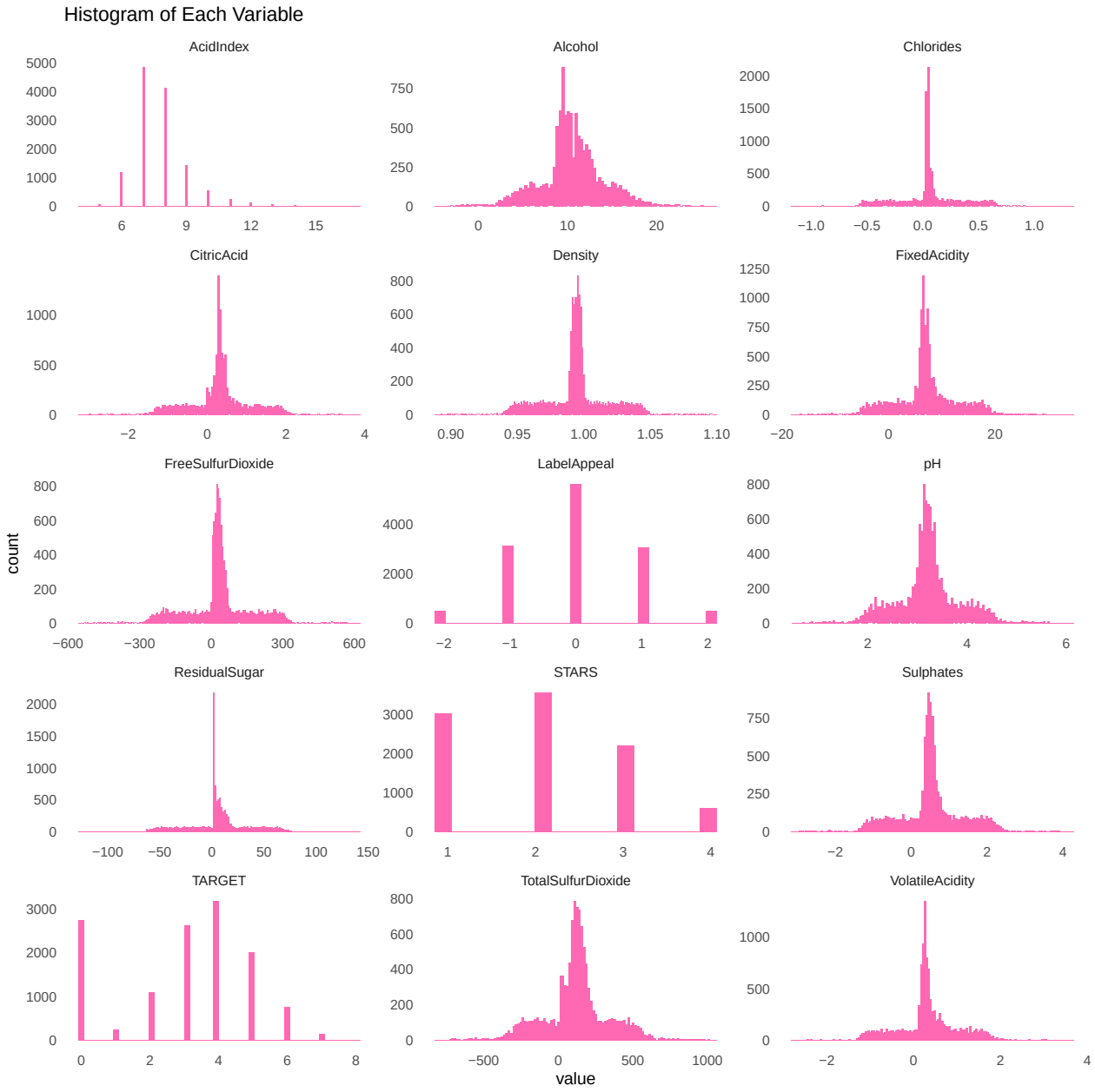
Response Variables



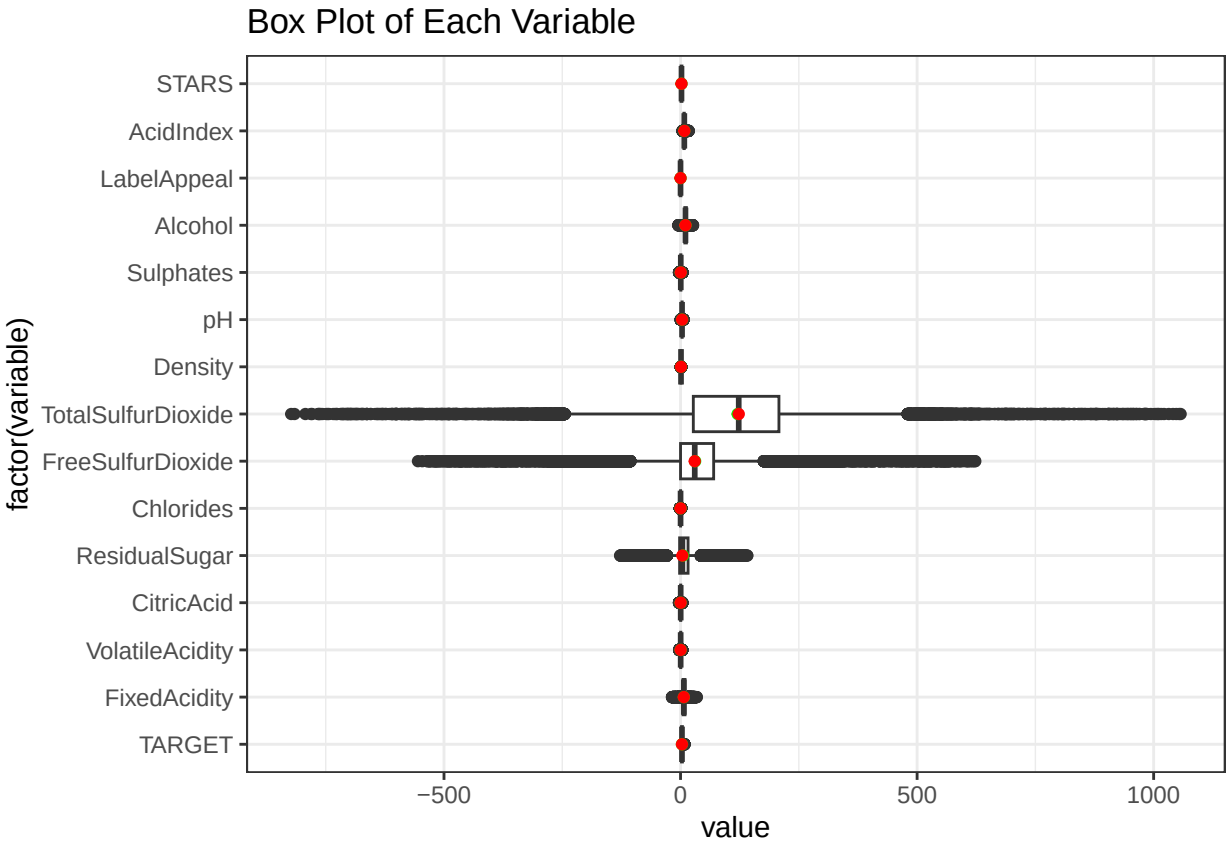
We see that the response variable, `TARGET` value is always between 0 and 8, which makes sense as this is the "Number of Cases of Wine Sold". In addition, the distribution of wine cases sold, given at least one sale, exhibits symmetry and approximates normality. The target variable, number of cases, is shown below. The

data shows a large number of zero values.

Variables Distributions

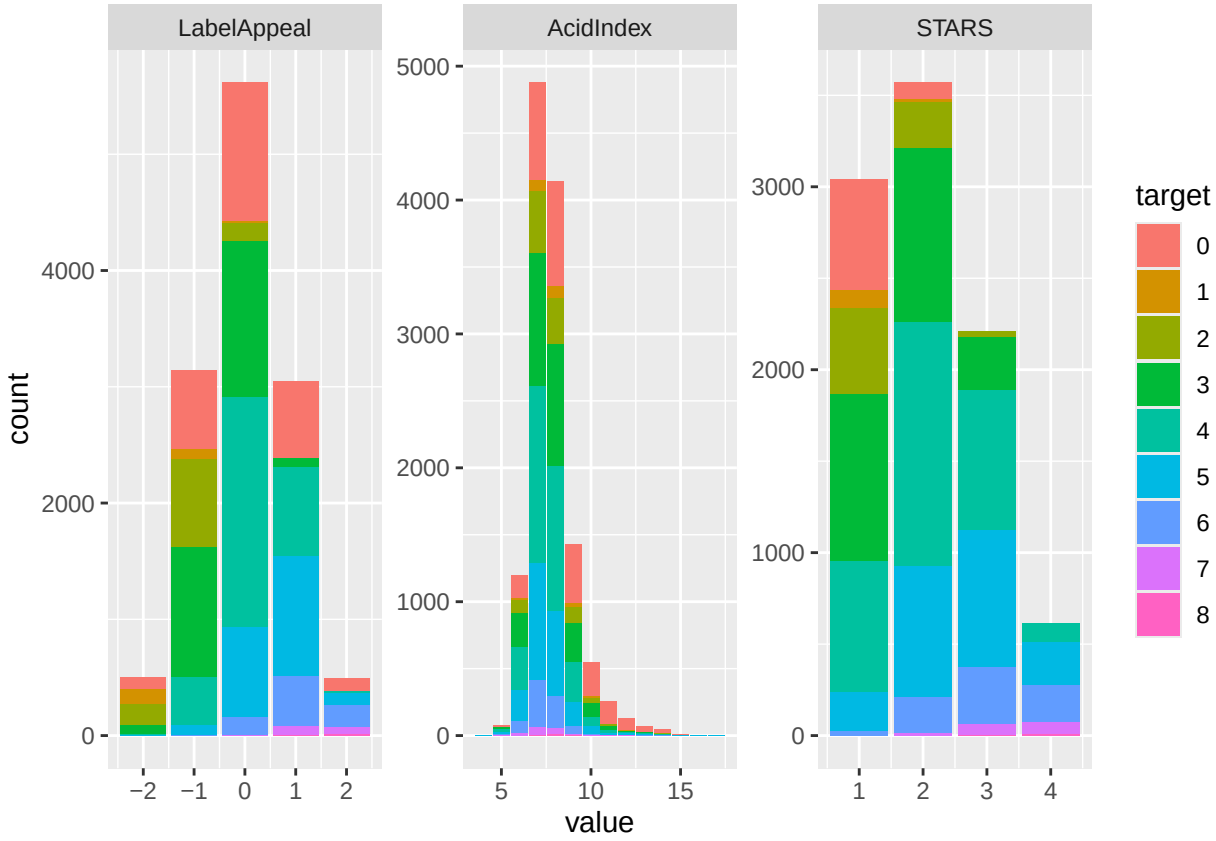


We observe that continuous variables exhibit a somewhat normal steep distribution. However, variables such as `AcidIndex` and `STARS` display right skewness.



In the box plot, there are not many outliers in the variables. However, `TotalSulfurDioxide`, `FreeSulfurDioxide`, and `ResidualSugar` variables have large ranges compared to other variables.We can tell a high number of variables have numerous outliers.

Relationship Between Categorical and Response Variables



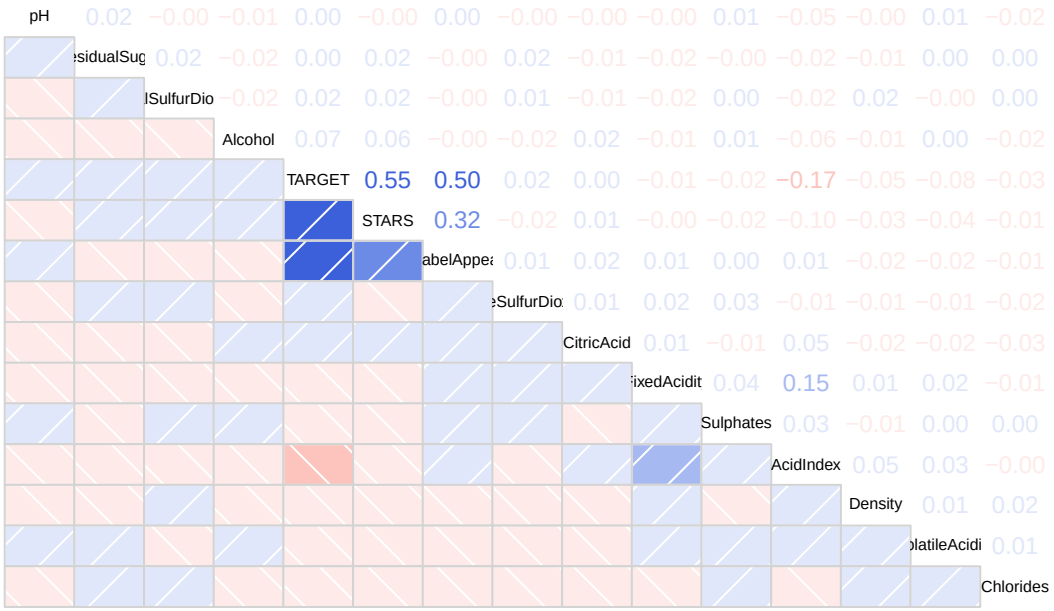
The bar charts compare the three discrete categorical variables to the `TARGET` variable. `AcidIndex` shows large quantity of wine were sold with the index number 7 and 8. `LabelAppeal` shows us generic label does yield higher number of wine samples per order. Lastly, `STARS` shows high star wine bottles have high price tags. For each of these predictors, there appears to be a significant relationship between the ordered levels and the number of wine cases sold.

Multicolinearity

	x
TARGET	1.0000000
FixedAcidity	-0.0125381
VolatileAcidity	-0.0759979
CitricAcid	0.0023450
ResidualSugar	0.0035196
Chlorides	-0.0304301
FreeSulfurDioxide	0.0226398
TotalSulfurDioxide	0.0216021
Density	-0.0475989
pH	0.0002199
Sulphates	-0.0212204
Alcohol	0.0737771
LabelAppeal	0.4979465

AcidIndex	-0.1676431
STARS	0.5546857

Correlation Plot



In the correlation table, we can see that `STARS` and `LabelAppeal` are most positively correlated variables with the response variable. Also, we some mild negative correlation between the response variable and `AcidIndex` variable.

Data Preparation

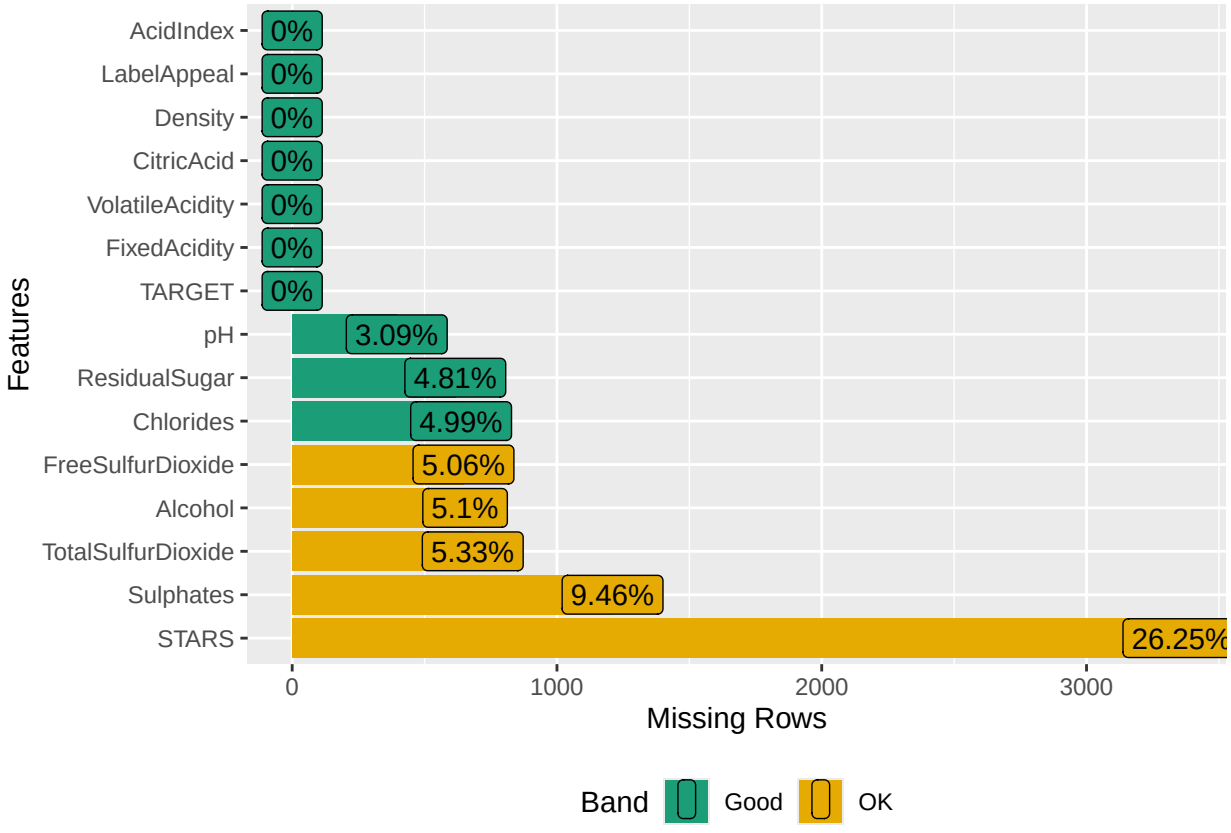
Negative Values

The data has some wine quality measures that are negative that should not be. We will simply take the absolute value of these for now. The alternative would be to center by adding the min of each variable. Since we are given little information about the source of this dataset, and why these quality measures are so off, it is difficult to ascertain the best overall approach. For `LabelAppeal`, we will add the min.

```
train_df$FixedAcidity <- abs(train_df$FixedAcidity)
evaluate_df$FixedAcidity <- abs(evaluate_df$FixedAcidity)

train_df$VolatileAcidity <- abs(train_df$VolatileAcidity)
evaluate_df$VolatileAcidity <- abs(evaluate_df$VolatileAcidity)

train_df$CitricAcid <- abs(train_df$CitricAcid)
evaluate_df$CitricAcid <- abs(evaluate_df$CitricAcid)

train_df$ResidualSugar <- abs(train_df$ResidualSugar)
evaluate_df$ResidualSugar <- abs(evaluate_df$ResidualSugar)

train_df$Chlorides <- abs(train_df$Chlorides)
```

```r
evaluate_df$Chlorides <- abs(evaluate_df$Chlorides)

train_df$FreeSulfurDioxide <- abs(train_df$FreeSulfurDioxide)
evaluate_df$FreeSulfurDioxide <- abs(evaluate_df$FreeSulfurDioxide)

train_df$TotalSulfurDioxide <- abs(train_df$TotalSulfurDioxide)
evaluate_df$TotalSulfurDioxide <- abs(evaluate_df$TotalSulfurDioxide)

train_df$Sulphates <- abs(train_df$Sulphates)
evaluate_df$Sulphates <- abs(evaluate_df$Sulphates)

train_df$LabelAppeal <- train_df$LabelAppeal + abs(min(train_df$LabelAppeal))
evaluate_df$LabelAppeal <- evaluate_df$LabelAppeal + abs(min(evaluate_df$LabelAppeal))
```

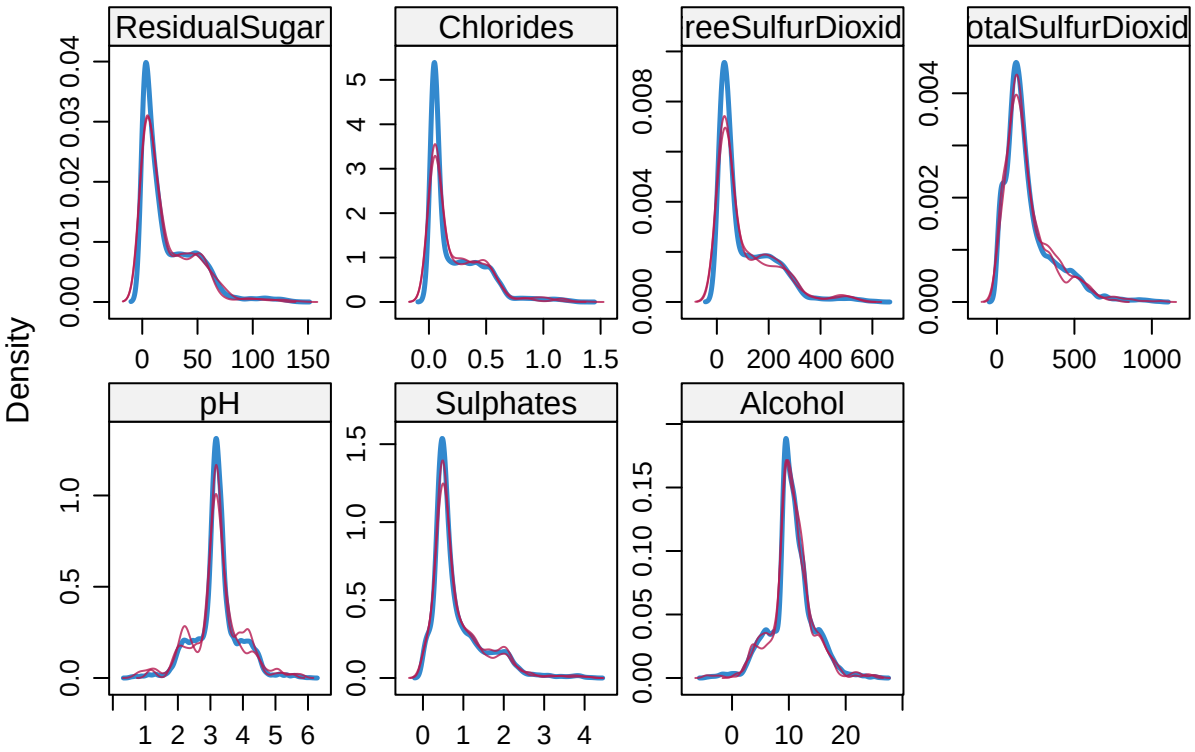
Missing Data



According to the graph, the data set has multiple variables with missing variables. The `STARS` variable has the most NA values. The `Sulphates` variable records missing values in roughly 10% of observations, while the remaining six predictors have missing values ranging from 3% to 5%. These missing values will be imputed later on during the data preparation using the MICE package and random forest prediction method.

```r
train_df$STARS[is.na(train_df$STARS)] <- 0
evaluate_df$STARS[is.na(evaluate_df$STARS)] <- 0
# Perform multiple imputation
mice_imputes <- mice(train_df, m = 2, maxit = 2, print = FALSE)
```

```
# Visualize the imputed values with density plot
densityplot(mice_imputes)
```

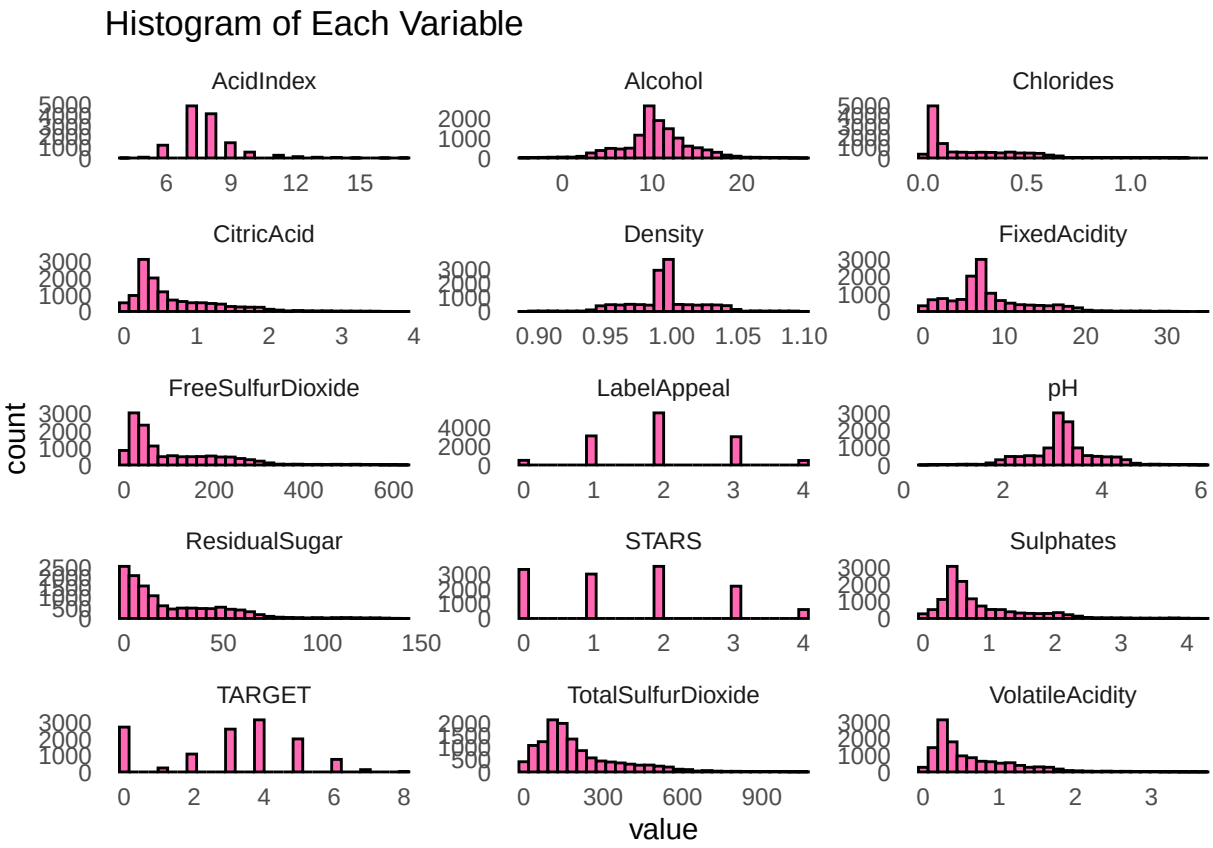


We can see that each of the remaining variables with missing values seem to be MAR, as the mice imputation distributions roughly match the existing. We'll also run the mice imputation again on both the train and test set. Instead of using it for our models, however, we'll simplify our run and fill in our data. Finally, after our analysis, we can use it in in our model, we'll update STARS to become a factor variable.

```
mice_train <-  mice(train_df, m = 1, maxit = 1, print = FALSE)
cleaned_train <- complete(mice_train)

mice_evaluate <- mice(evaluate_df, m = 1, maxit = 1, print = FALSE)
cleaned_evaluate <- complete(mice_evaluate)

cleaned_train$STARS <- as.factor(cleaned_train$STARS)
cleaned_evaluate$STARS <- as.factor(cleaned_evaluate$STARS)
```

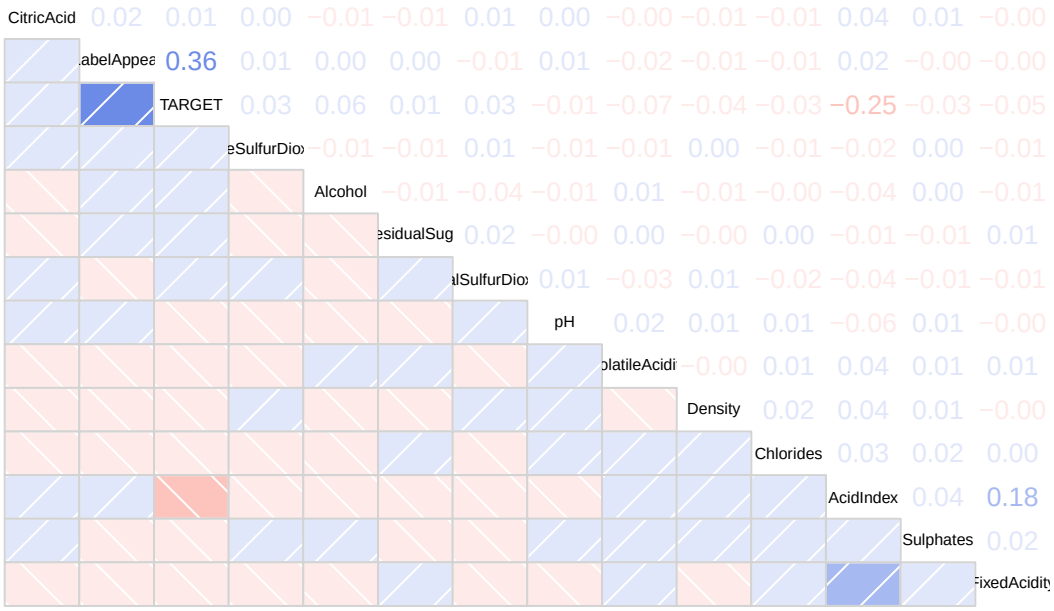
Descriptive Summaries and Correlation Review

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



	Mean	Std.Dev	Min	Q1	Median	Q3	Max	M
AcidIndex	7.7727237	1.3239264	4.00000	7.00000	8.00000	8.00000	17.00000	1.482
Alcohol	10.4894078	3.7362394	-4.70000	9.00000	10.40000	12.40000	26.50000	2.520
Chlorides	0.2227773	0.2344037	0.00000	0.04600	0.09900	0.36800	1.35100	0.100
CitricAcid	0.6863150	0.6060052	0.00000	0.28000	0.44000	0.97000	3.86000	0.326
Density	0.9942027	0.0265376	0.88809	0.98772	0.99449	1.00052	1.09924	0.009
FixedAcidity	8.0632513	4.9961186	0.00000	5.60000	7.00000	9.80000	34.40000	2.965
FreeSulfurDioxide	106.5577569	108.0547504	0.00000	28.00000	56.00000	172.00000	623.00000	60.786
LabelAppeal	1.9909340	0.8910892	0.00000	1.00000	2.00000	3.00000	4.00000	1.482
pH	3.2052739	0.6790213	0.48000	2.95000	3.20000	3.47000	6.13000	0.385
ResidualSugar	23.4076397	24.9744650	0.00000	3.60000	12.90000	38.70000	141.15000	16.308
Sulphates	0.8443970	0.6548955	0.00000	0.43000	0.59000	1.09000	4.24000	0.326
TARGET	3.0290739	1.9263682	0.00000	2.00000	3.00000	4.00000	8.00000	1.482
TotalSulfurDioxide	204.4533411	163.3559961	0.00000	99.00000	154.00000	263.00000	1057.00000	102.299
VolatileAcidity	0.6410856	0.5556141	0.00000	0.25000	0.41000	0.91000	3.68000	0.326

Revised Correlation



Split the Sample data Set

With our transformations complete, we can now add these into our `cleaned_train` dataframe and continue on to build our models. To better measure each model performance, we split our data into a training and testing data set. We will train using the first, then measure model performance again the testing hold out set.

```
## [1] "Number of Training Samples:  10238"
```

```
## [1] "Number of Testing Samples:  2557"
```

Build Models

Poisson Regression Model 1

In this first model, we include all available features: `FixedAcidity`, `VolatileAcidity`, `CitricAcid`, `ResidualSugar`, `Chlorides`, `FreeSulfurDioxide`, `TotalSulfurDioxide`, `Density`, `pH`, `Sulphates`, `Alcohol`, `LabelAppeal`, `AcidIndex`, `STARS`

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = trainingData)
##
## Coefficients:
##                       Estimate  Std. Error z value             Pr(>|z|)
## (Intercept)         0.89651161  0.21882996   4.097            0.0000419 ***
## FixedAcidity        0.00044060  0.00116861   0.377              0.70615
## VolatileAcidity    -0.03886226  0.01041277  -3.732              0.00019 ***
```

```
## CitricAcid          0.01073449  0.00926350   1.159              0.24654
## ResidualSugar       0.00012727  0.00022874   0.556              0.57795
## Chlorides          -0.03134181  0.02446772  -1.281              0.20021
## FreeSulfurDioxide   0.00001196  0.00005234   0.228              0.81927
## TotalSulfurDioxide  0.00006708  0.00003500   1.916              0.05534 .
## Density            -0.39456503  0.21372859  -1.846              0.06488 .
## pH                 -0.01019817  0.00841827  -1.211              0.22573
## Sulphates          -0.00733133  0.00890821  -0.823              0.41052
## Alcohol             0.00438273  0.00154021   2.846              0.00443 **
## LabelAppeal         0.15626849  0.00687182  22.740 < 0.0000000000000002 ***
## AcidIndex          -0.07910604  0.00508904 -15.544 < 0.0000000000000002 ***
## STARS1              0.77512942  0.02193040  35.345 < 0.0000000000000002 ***
## STARS2              1.09769400  0.02048368  53.589 < 0.0000000000000002 ***
## STARS3              1.21759722  0.02150641  56.616 < 0.0000000000000002 ***
## STARS4              1.34370972  0.02720736  49.388 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 18209  on 10237  degrees of freedom
## Residual deviance: 10872  on 10220  degrees of freedom
## AIC: 36515
##
## Number of Fisher Scoring iterations: 6
```

```
##            RMSE        Rsquared             MAE             aic             bic
##     7.396644537     0.002272751     5.611462200 36514.546196098 36644.755704304
```

Poisson Regression Model 2

In this Model 2, we only include the most predictive features : `VolatileAcidity`, `TotalSulfurDioxide`, `Alcohol`, `LabelAppeal`, `AcidIndex`, `STARS`

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + TotalSulfurDioxide +
##     Alcohol + LabelAppeal + AcidIndex + STARS, family = poisson,
##     data = trainingData)
##
## Coefficients:
##                       Estimate  Std. Error z value             Pr(>|z|)
## (Intercept)         0.47032351  0.04970854   9.462 < 0.0000000000000002 ***
## VolatileAcidity    -0.03903866  0.01041153  -3.750             0.000177 ***
## TotalSulfurDioxide  0.00006701  0.00003498   1.916             0.055411 .
## Alcohol             0.00438791  0.00153971   2.850             0.004374 **
## LabelAppeal         0.15628244  0.00686826  22.754 < 0.0000000000000002 ***
## AcidIndex          -0.07879652  0.00501598 -15.709 < 0.0000000000000002 ***
## STARS1              0.77645950  0.02192325  35.417 < 0.0000000000000002 ***
## STARS2              1.09897982  0.02047100  53.685 < 0.0000000000000002 ***
## STARS3              1.21951647  0.02149474  56.736 < 0.0000000000000002 ***
## STARS4              1.34483074  0.02719986  49.443 < 0.0000000000000002 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 18209  on 10237  degrees of freedom
## Residual deviance: 10881  on 10228  degrees of freedom
## AIC: 36508
##
## Number of Fisher Scoring iterations: 6
```

```
##            RMSE        Rsquared             MAE             aic             bic
##     7.396644537     0.002272751     5.611462200 36507.877709108 36580.216324778
```

Negative Binomial Regression Model 3

Similar to Poisson Model 1, the predictors for the following model are: `FixedAcidity`, `VolatileAcidity`, `CitricAcid`, `ResidualSugar`, `Chlorides`, `FreeSulfurDioxide`, `TotalSulfurDioxide`, `Density`, `pH`, `Sulphates`, `Alcohol`, `LabelAppeal`, `AcidIndex`, `STARS`

```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = trainingData, init.theta = 40817.06846,
##     link = log)
##
## Coefficients:
##                       Estimate  Std. Error z value             Pr(>|z|)
## (Intercept)         0.89654613  0.21883979   4.097            0.0000419 ***
## FixedAcidity        0.00044063  0.00116866   0.377              0.70615
## VolatileAcidity    -0.03886366  0.01041322  -3.732              0.00019 ***
## CitricAcid          0.01073468  0.00926393   1.159              0.24655
## ResidualSugar       0.00012727  0.00022875   0.556              0.57797
## Chlorides          -0.03134249  0.02446883  -1.281              0.20022
## FreeSulfurDioxide   0.00001196  0.00005234   0.228              0.81926
## TotalSulfurDioxide  0.00006708  0.00003501   1.916              0.05533 .
## Density            -0.39457493  0.21373827  -1.846              0.06488 .
## pH                 -0.01019903  0.00841865  -1.211              0.22571
## Sulphates          -0.00733176  0.00890860  -0.823              0.41051
## Alcohol             0.00438267  0.00154028   2.845              0.00444 **
## LabelAppeal         0.15626750  0.00687212  22.739 < 0.0000000000000002 ***
## AcidIndex          -0.07910845  0.00508924 -15.544 < 0.0000000000000002 ***
## STARS1              0.77512839  0.02193085  35.344 < 0.0000000000000002 ***
## STARS2              1.09769300  0.02048414  53.587 < 0.0000000000000002 ***
## STARS3              1.21759697  0.02150699  56.614 < 0.0000000000000002 ***
## STARS4              1.34371048  0.02720854  49.386 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(40817.07) family taken to be 1)
##
##     Null deviance: 18208  on 10237  degrees of freedom
## Residual deviance: 10872  on 10220  degrees of freedom
## AIC: 36517
##
```

```
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  40817
##           Std. Err.:  38243
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -36478.88
```

```
##            RMSE        Rsquared             MAE             aic             bic
##     7.396644537     0.002272751     5.611462200 36516.883744203 36654.327113977
```

Negative Binomial Regression Model 4

Similar to Poisson Model 2, the predictors for the following model are: `VolatileAcidity`, `FreeSulfurDioxide`, `TotalSulfurDioxide`, `Alcohol`, `LabelAppeal`, `AcidIndex`, `STARS`

```
##
## Call:
## glm.nb(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide +
##     TotalSulfurDioxide + Alcohol + LabelAppeal + AcidIndex +
##     STARS, data = trainingData, init.theta = 40793.87401, link = log)
##
## Coefficients:
##                       Estimate  Std. Error z value             Pr(>|z|)
## (Intercept)         0.46905886  0.05000564   9.380 < 0.0000000000000002 ***
## VolatileAcidity    -0.03901671  0.01041250  -3.747             0.000179 ***
## FreeSulfurDioxide   0.00001242  0.00005231   0.237             0.812285
## TotalSulfurDioxide  0.00006687  0.00003499   1.911             0.055949 .
## Alcohol             0.00438761  0.00153980   2.849             0.004379 **
## LabelAppeal         0.15624854  0.00687001  22.744 < 0.0000000000000002 ***
## AcidIndex          -0.07878939  0.00501649 -15.706 < 0.0000000000000002 ***
## STARS1              0.77643351  0.02192395  35.415 < 0.0000000000000002 ***
## STARS2              1.09887845  0.02047584  53.667 < 0.0000000000000002 ***
## STARS3              1.21951530  0.02149541  56.734 < 0.0000000000000002 ***
## STARS4              1.34477616  0.02720218  49.436 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(40793.87) family taken to be 1)
##
##     Null deviance: 18208  on 10237  degrees of freedom
## Residual deviance: 10881  on 10227  degrees of freedom
## AIC: 36512
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  40794
##           Std. Err.:  38220
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -36488.16
```

```
##            RMSE        Rsquared             MAE             aic             bic
##     7.396644537     0.002272751     5.611462200 36512.158926885 36598.965265690
```

Multiple Linear Regression Model 5

The predictors for the following model are: `FixedAcidity`, `VolatileAcidity`, `CitricAcid`, `ResidualSugar`, `Chlorides`, `FreeSulfurDioxide`, `TotalSulfurDioxide`, `Density`, `pH`, `Sulphates`, `Alcohol`, `LabelAppeal`, `AcidIndex`, `STARS`

```
##
## Call:
## lm(formula = TARGET ~ ., data = trainingData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6213 -0.8628  0.0255  0.8484  6.2298
##
## Coefficients:
##                       Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)         3.10535322  0.49433268   6.282       0.000000000348 ***
## FixedAcidity        0.00230525  0.00263284   0.876              0.38128
## VolatileAcidity    -0.11216401  0.02310797  -4.854       0.000001228421 ***
## CitricAcid          0.03355295  0.02118897   1.584              0.11334
## ResidualSugar       0.00018510  0.00052089   0.355              0.72233
## Chlorides          -0.09375842  0.05542373  -1.692              0.09074 .
## FreeSulfurDioxide   0.00005049  0.00011953   0.422              0.67273
## TotalSulfurDioxide  0.00020595  0.00007955   2.589              0.00964 **
## Density            -1.17500504  0.48553899  -2.420              0.01554 *
## pH                 -0.02742550  0.01908601  -1.437              0.15076
## Sulphates          -0.01604593  0.01998160  -0.803              0.42197
## Alcohol             0.01452909  0.00347430   4.182       0.000029152843 ***
## LabelAppeal         0.45568948  0.01524494  29.891 < 0.0000000000000002 ***
## AcidIndex          -0.20066084  0.01012834 -19.812 < 0.0000000000000002 ***
## STARS1              1.37556161  0.03683294  37.346 < 0.0000000000000002 ***
## STARS2              2.40897671  0.03583666  67.221 < 0.0000000000000002 ***
## STARS3              2.97990262  0.04125005  72.240 < 0.0000000000000002 ***
## STARS4              3.71958526  0.06624345  56.150 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.306 on 10220 degrees of freedom
## Multiple R-squared:  0.5398, Adjusted R-squared:  0.5391
## F-statistic: 705.3 on 17 and 10220 DF,  p-value: < 0.00000000000000022
```

```
##            RMSE        Rsquared             MAE             aic             bic
##     7.396644537     0.002272751     5.611462200 34547.360364751 34684.803734524
```

Multiple Linear Regression Model 6

For the final Linear Model, we leverage `stepAIC` on our Linear Model 5 to choose the most important features.

```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     TotalSulfurDioxide + Density + pH + Alcohol + LabelAppeal +
##     AcidIndex + STARS, data = trainingData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6038 -0.8624  0.0274  0.8492  6.2080
##
## Coefficients:
##                       Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)         3.11692999  0.49360418   6.315       0.000000000282 ***
## VolatileAcidity    -0.11223904  0.02310355  -4.858       0.000001202855 ***
## CitricAcid          0.03294389  0.02117970   1.555              0.11987
## Chlorides          -0.09534539  0.05538496  -1.722              0.08519 .
## TotalSulfurDioxide  0.00020726  0.00007951   2.607              0.00916 **
## Density            -1.17929697  0.48544237  -2.429              0.01514 *
## pH                 -0.02750368  0.01908215  -1.441              0.14952
## Alcohol             0.01446469  0.00347347   4.164       0.000031481886 ***
## LabelAppeal         0.45569546  0.01524095  29.899 < 0.0000000000000002 ***
## AcidIndex          -0.19953964  0.00997542 -20.003 < 0.0000000000000002 ***
## STARS1              1.37606290  0.03681807  37.375 < 0.0000000000000002 ***
## STARS2              2.40906765  0.03581169  67.270 < 0.0000000000000002 ***
## STARS3              2.98033924  0.04123736  72.273 < 0.0000000000000002 ***
## STARS4              3.72069114  0.06621835  56.188 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.306 on 10224 degrees of freedom
## Multiple R-squared:  0.5398, Adjusted R-squared:  0.5392
## F-statistic: 922.3 on 13 and 10224 DF,  p-value: < 0.00000000000000022

##            RMSE        Rsquared             MAE             aic             bic
##     7.396644537     0.002272751     5.611462200 34541.069711693 34649.577635198
```

Select Models

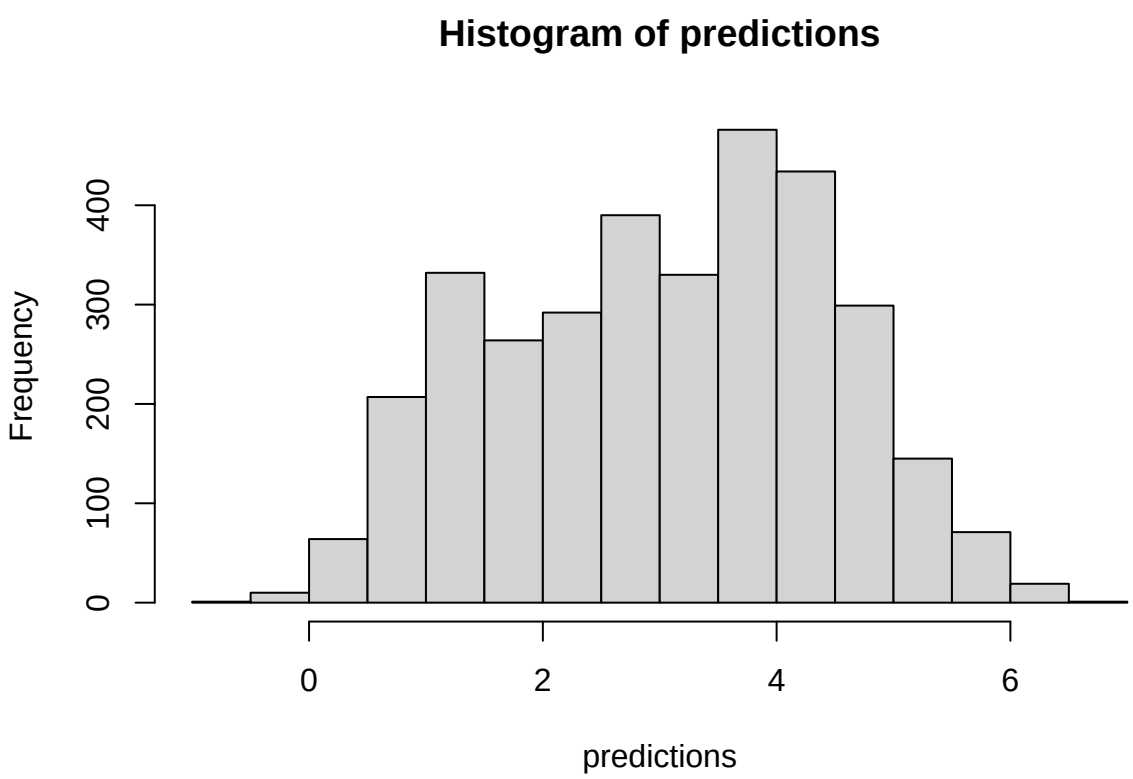
	RMSE	Rsquared	MAE	aic	bic
Poission_Evaluate1	7.396644	0.0022728	5.611462	36514.55	36644.76
Poission_Evaluate2	7.396644	0.0022728	5.611462	36507.88	36580.22
Negative_Binomial_eval3	7.396644	0.0022728	5.611462	36516.88	36654.33
Negative_Binomial_eval4	7.396644	0.0022728	5.611462	36512.16	36598.97
Linear_Regression_eval5	7.396644	0.0022728	5.611462	34547.36	34684.80
Linear_Regression_eval6	7.396644	0.0022728	5.611462	34541.07	34649.58

This table summarizes the **RMSE**, **RSQUARED**, **MAE**, **AIC** and **BIC** for all SIX models. The Linear regressions (**Linear Model 5** and **Linear Model 6**) had the overall best performance based on **RMSE** and **RSQUARED**; as well as **Linear Model 6** had the best performance based on **AIC** and **BIC**.

Finally, we chose **Multiple Linear Regression Model 6** as our final model since it had a far lower **AIC** and **BIC**.

Prediction

```
eval_data <- cleaned_evaluate %>% dplyr::select(-TARGET)
predictions <- predict(lm6, eval_data)

eval_data$TARGET <- predictions

hist(predictions)
```




```
write.csv(eval_data, 'DATA621_HW5_Predictions.csv', row.names=FALSE)

head(eval_data)
```

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

```
##   LabelAppeal AcidIndex STARS    TARGET
## 1           1         6     0 1.2393948
## 2           2         6     2 4.0952884
## 3           2         8     1 2.3896560
## 4           1         8     1 2.3256759
## 5           2        10     0 0.8120265
## 6           3         8     4 5.5798664
```

The histogram shows that our predictions have a similar shape to our training Target variable, the means and medians are almost identical, and the kurtosis values are close.

The predicted file is uploaded to Github: https://github.com/waheeb123/Data-621/blob/main/Homeworks/Homework%205/DATA621_HW5_Predictions.csv

Conclusions

The Linear Regression model that we chose as the best model has an adjusted R2 value of 0.55. All predictors and levels within each categorical predictor are significant at the 6% level. The model coefficients make intuitive sense. Sales of wine decrease with each increase in the `AcidIndex` value. Having a `STARS` rating results in more sales compared to no ratings, and higher STARS ratings are associated with higher sales. Finally, increases in alcohol content are associated with higher sales.The interpretation of the coefficients aligns with the insights gained during exploratory data analysis.