Data 621 - HW5

Devin Teran, Atina Karim, Tom Hill, Amit Kapoor

5/23/2021

Contents

Overview

In this assignment, 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 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 is the description of the variables of interest in the data set.

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
TARGET	Number of Cases Purchased	None
AcidIndex	Proprietary method of testing total acidity of wine by using a weighted average	
Alcohol	Alcohol Content	
Chlorides	Chloride content of wine	
CitricAcid	Citric Acid Content	
Density	Density of Wine	
FixedAcidity	Fixed Acidity of Wine	
FreeSulfurDioxide	Sulfur Dioxide content of wine	
LabelAppeal	Marketing Score indicating the appeal of label design for consumers. High numbers suggest customers like the label design. Negative numbers suggest 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	

Statistics

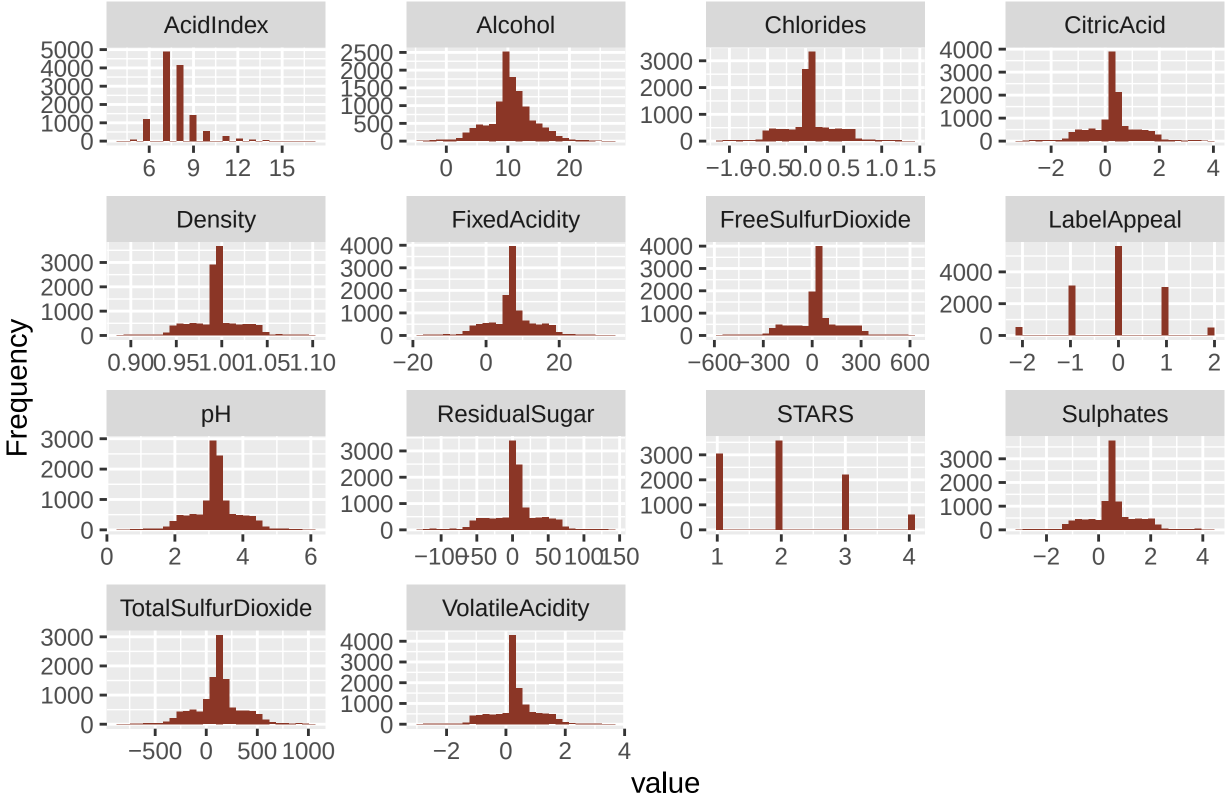
All of the data are numeric and here is the statistics summary of all the predictors.

```
##                     vars     n   mean     sd median trimmed    mad     min
## FixedAcidity           1 12795   7.08   6.32   6.90    7.07   3.26  -18.10
## VolatileAcidity        2 12795   0.32   0.78   0.28    0.32   0.43   -2.79
## CitricAcid             3 12795   0.31   0.86   0.31    0.31   0.42   -3.24
## ResidualSugar          4 12179   5.42  33.75   3.90    5.58  15.72 -127.80
## Chlorides              5 12157   0.05   0.32   0.05    0.05   0.13   -1.17
## FreeSulfurDioxide      6 12148  30.85 148.71  30.00   30.93  56.34 -555.00
## TotalSulfurDioxide     7 12113 120.71 231.91 123.00  120.89 134.92 -823.00
## Density                8 12795   0.99   0.03   0.99    0.99   0.01    0.89
## pH                     9 12400   3.21   0.68   3.20    3.21   0.39    0.48
## Sulphates             10 11585   0.53   0.93   0.50    0.53   0.44   -3.13
## Alcohol               11 12142  10.49   3.73  10.40   10.50   2.37   -4.70
## LabelAppeal           12 12795  -0.01   0.89   0.00   -0.01   1.48   -2.00
## AcidIndex             13 12795   7.77   1.32   8.00    7.64   1.48    4.00
## STARS                 14  9436   2.04   0.90   2.00    1.97   1.48    1.00
##                         max   range  skew kurtosis   se
## FixedAcidity          34.40   52.50 -0.02     1.67 0.06
## VolatileAcidity        3.68    6.47  0.02     1.83 0.01
## CitricAcid             3.86    7.10 -0.05     1.84 0.01
## ResidualSugar        141.15  268.95 -0.05     1.88 0.31
## Chlorides              1.35    2.52  0.03     1.79 0.00
## FreeSulfurDioxide    623.00 1178.00  0.01     1.84 1.35
## TotalSulfurDioxide  1057.00 1880.00 -0.01     1.67 2.11
## Density               1.10    0.21 -0.02     1.90 0.00
## pH                    6.13    5.65  0.04     1.65 0.01
## Sulphates             4.24    7.37  0.01     1.75 0.01
## Alcohol              26.50   31.20 -0.03     1.54 0.03
## LabelAppeal           2.00    4.00  0.01    -0.26 0.01
## AcidIndex            17.00   13.00  1.65     5.19 0.01
## STARS                 4.00    3.00  0.45    -0.69 0.01
```

Numeric Variables

Seeing the distribution plots below of all the predictor variables, it is evident that variables Alcohol, Chlorides, CitricAcid, Density, FixedAcidity, FreeSulphurDioxide, pH, ResidualSugar, Sulphates, TotalSulphurDioxide and VolatileAcidity appear to be symmetrical but non Gaussian since there is a strong spike near the median and not smooth near the tails on either side.

LabelAppeal distribution appears mostly normal while for AcidIndex and STARS seems to follow Poisson distribution.

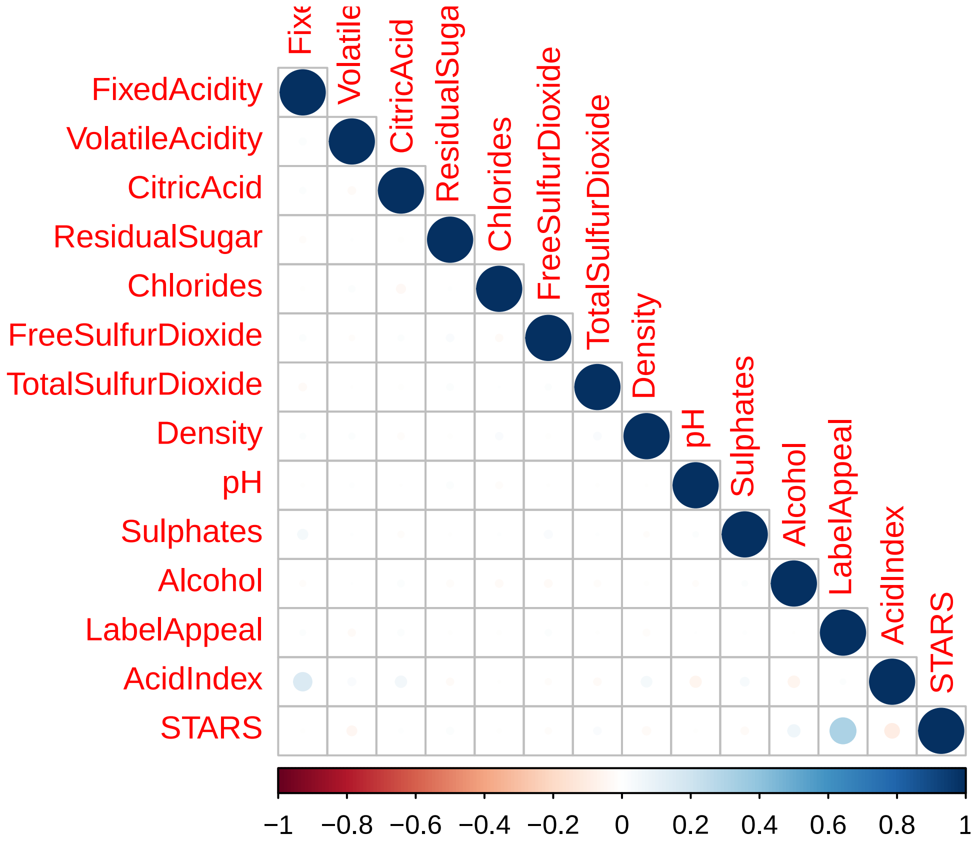


```
## # A tibble: 1 x 15
##   TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
##    <int>        <int>           <int>      <int>         <int>     <int>
## 1      9          470             815        602          2078      1664
## # ... with 9 more variables: FreeSulfurDioxide <int>, TotalSulfurDioxide <int>,
## #   Density <int>, pH <int>, Sulphates <int>, Alcohol <int>, LabelAppeal <int>,
## #   AcidIndex <int>, STARS <int>
```

All variables in this dataset are initially interpretable as numeric data. There are several variables, including AcidIndex, LabelAppeal, and STARS that have few distinct values and may be treated as factors in the future. The target variable, number of cases, also only spans values 0 - 8.

Correlations

The `corrplot` below shows the correlation between predictor variables by ignoring the missing entries.

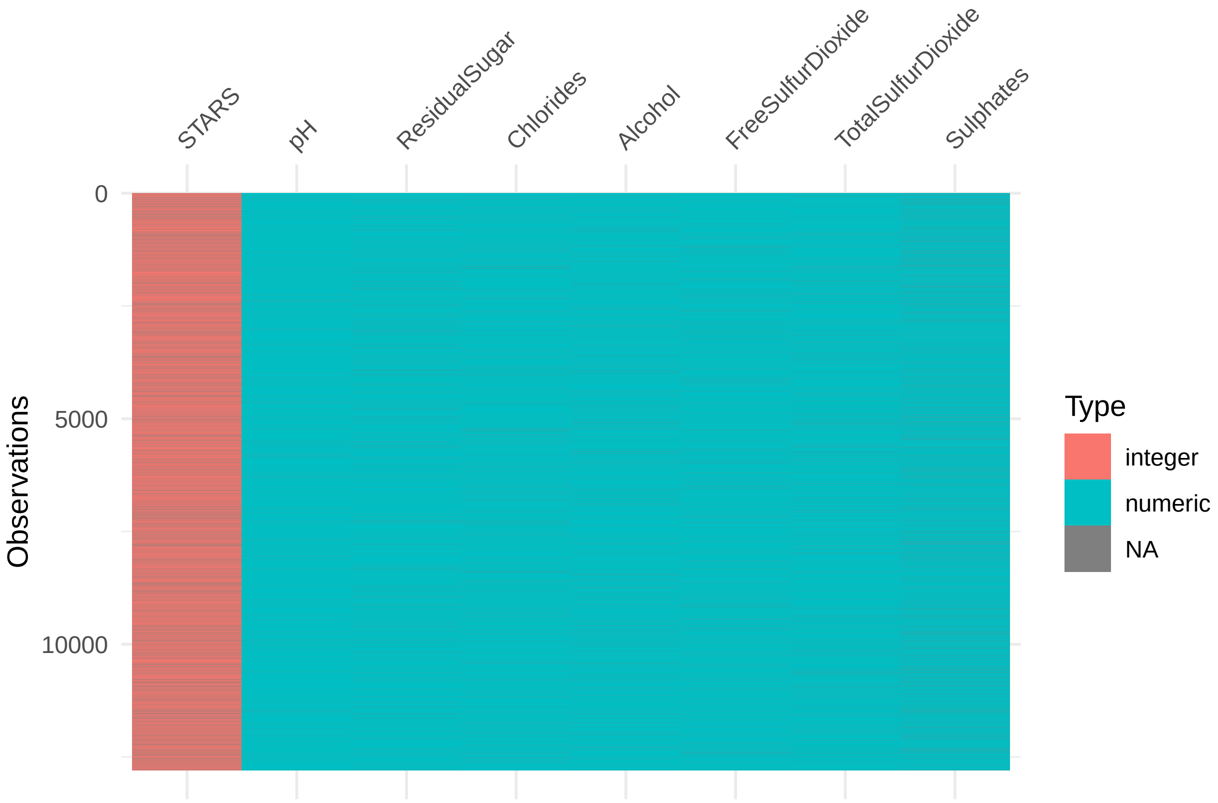


From the above `corrplot`, it is apparent that

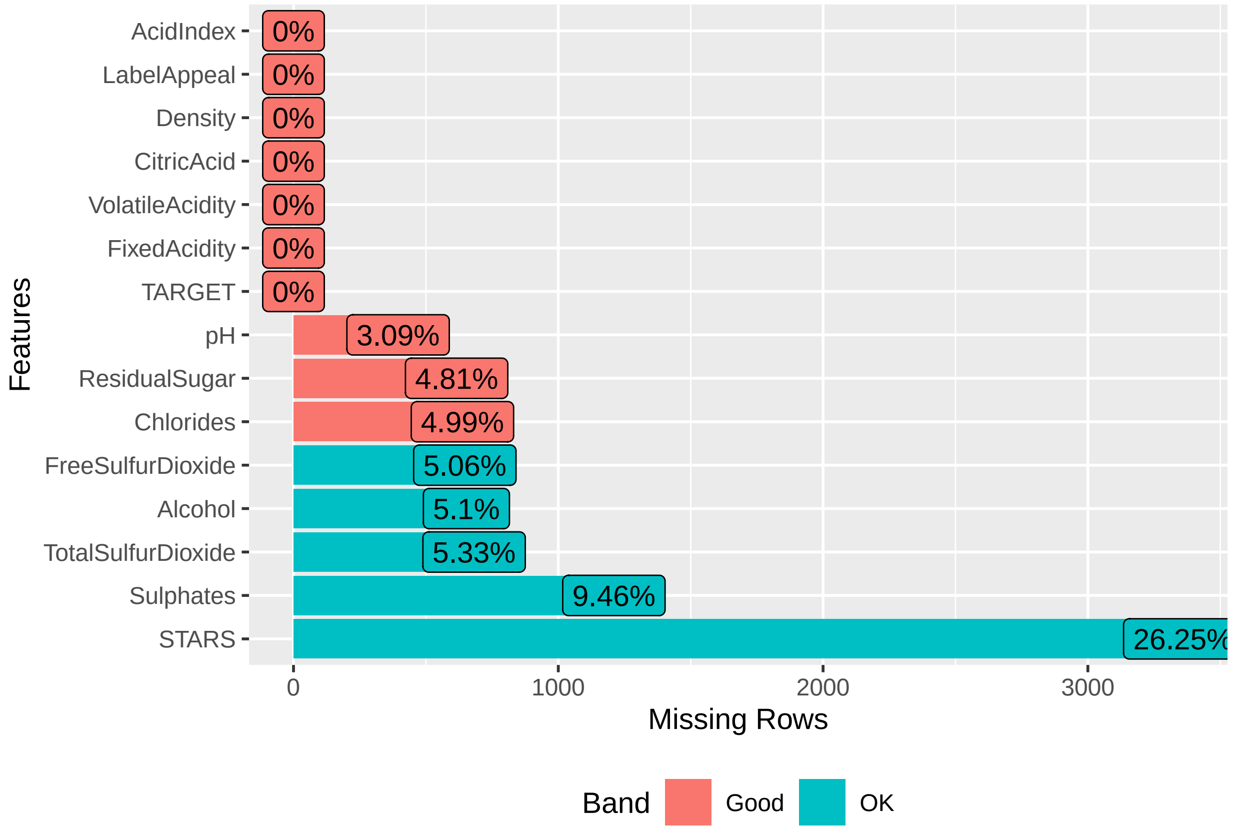
- `AcidIndex` and `FixedAcidity` are positively correlated.
- `STARS` and `LabelAppeal` are positively correlated.
- `STARS` and `AcidIndex` are negatively correlated.

Missing Values

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



The feature with the most misisng variables is STARS, which is a rating between 1-4. It's plausible that the missing values in this case are wine brands that are unrated by STARS. These missing values can potentially be recoded as 'zero' to avoid dropping a substantial proportion of data. There also does not appear to be any apparent pattern in misisng data.



```
## [1] 50.3
```

The remaining missing values comprise less than ten percent of observations separately. Taken together, just over 50% of observations have complete data available.

Data Preparation

Missing Values

We will recode missing values in the predictor STARS as 0.

However, since our ratings range from 1 to 4, we will also impute this variable with the median.

However, imputing with measures of central tendency is that they tend to reduce the variance in the dataset and shrinks standard errors Therefore, our third method of dealing with missing values would be multiple imputation. We will use the MICE package in R to impute via the random forest method.

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

Finally, we will create a data set to impute missing values with the median after we recode missing values in the predictor STARS as 0.

Build Models

Multiple Linear Regression : Model 1

We will first run linear regression with all predictor variables in our dataset.

```
##
## Call:
## lm(formula = TARGET ~ ., data = wine_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0614 -0.5143  0.1240  0.7170  3.2419
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.563e+00  5.530e-01   8.251  < 2e-16 ***
## FixedAcidity        1.685e-03  2.319e-03   0.727   0.4675
## VolatileAcidity    -9.466e-02  1.846e-02  -5.129 3.00e-07 ***
## CitricAcid         -4.836e-03  1.675e-02  -0.289   0.7728
## ResidualSugar      -2.513e-04  4.276e-04  -0.588   0.5567
## Chlorides          -1.134e-01  4.546e-02  -2.494   0.0126 *
## FreeSulfurDioxide   2.264e-04  9.711e-05   2.332   0.0198 *
## TotalSulfurDioxide  7.810e-05  6.288e-05   1.242   0.2142
## Density            -1.281e+00  5.435e-01  -2.357   0.0185 *
## pH                 -9.441e-03  2.121e-02  -0.445   0.6563
## Sulphates          -1.727e-02  1.558e-02  -1.109   0.2676
## Alcohol             1.653e-02  3.887e-03   4.252 2.15e-05 ***
## LabelAppeal         6.442e-01  1.743e-02  36.947  < 2e-16 ***
## AcidIndex          -1.649e-01  1.235e-02 -13.346  < 2e-16 ***
## STARS               7.278e-01  1.710e-02  42.571  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.153 on 6421 degrees of freedom
##   (6359 observations deleted due to missingness)
## Multiple R-squared:  0.445,  Adjusted R-squared:  0.4438
## F-statistic: 367.8 on 14 and 6421 DF,  p-value: < 2.2e-16
```

The adjusted r^2 is 0.4438 and is significant.

Multiple Linear Regression : Model 2

We will now run linear regression with all predictor variables on our dataset with missing values in STARS recoded as 0.

```
##
## Call:
## lm(formula = TARGET ~ ., data = wine_train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5582 -0.9421  0.0522  0.9042  6.0007
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.212e+00  5.446e-01   7.734 1.16e-14 ***
## FixedAcidity        8.444e-04  2.312e-03   0.365 0.714956    
## VolatileAcidity    -9.751e-02  1.822e-02  -5.351 8.96e-08 ***
## CitricAcid          9.147e-03  1.665e-02   0.549 0.582723    
## ResidualSugar      -1.337e-04  4.215e-04  -0.317 0.751066    
## Chlorides          -1.417e-01  4.464e-02  -3.175 0.001502 ** 
## FreeSulfurDioxide   3.240e-04  9.622e-05   3.368 0.000762 ***
## TotalSulfurDioxide  2.682e-04  6.215e-05   4.315 1.61e-05 ***
## Density            -1.019e+00  5.380e-01  -1.895 0.058182 .  
## pH                 -3.611e-02  2.101e-02  -1.719 0.085660 .  
## Sulphates          -2.907e-02  1.535e-02  -1.894 0.058317 .  
## Alcohol             9.400e-03  3.871e-03   2.428 0.015190 *  
## LabelAppeal         4.309e-01  1.669e-02  25.821  < 2e-16 ***
## AcidIndex          -2.066e-01  1.111e-02 -18.602  < 2e-16 ***
## STARS               9.691e-01  1.279e-02  75.776  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.329 on 8660 degrees of freedom
##   (4120 observations deleted due to missingness)
## Multiple R-squared:  0.5194, Adjusted R-squared:  0.5186 
## F-statistic: 668.6 on 14 and 8660 DF,  p-value: < 2.2e-16
```

The adjusted r^2 has gone up to 0.5186 and this is significant.

Multiple Linear Regression : Model 3

We will run the linear regression model on our dataset with the imputed median.

```
## 
## Call:
## lm(formula = TARGET ~ ., data = wine_impute)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.2211 -0.7540  0.3598  1.1254  4.3550 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)         5.355e+00  5.517e-01   9.707  < 2e-16 ***
## FixedAcidity       -1.168e-03  2.315e-03  -0.505 0.613911    
## VolatileAcidity    -1.549e-01  1.838e-02  -8.429  < 2e-16 ***
## CitricAcid          3.976e-02  1.673e-02   2.377 0.017476 *  
## ResidualSugar       4.716e-04  4.371e-04   1.079 0.280670    
## Chlorides          -1.931e-01  4.638e-02  -4.164 3.15e-05 ***
## FreeSulfurDioxide   4.286e-04  9.941e-05   4.312 1.63e-05 ***
## TotalSulfurDioxide  3.098e-04  6.387e-05   4.851 1.25e-06 ***
## Density            -1.274e+00  5.427e-01  -2.347 0.018959 *  
## pH                 -6.387e-02  2.154e-02  -2.965 0.003028 ** 
## Sulphates          -5.485e-02  1.623e-02  -3.380 0.000728 ***
## Alcohol             1.883e-02  3.972e-03   4.739 2.17e-06 ***
## LabelAppeal         5.945e-01  1.686e-02  35.250  < 2e-16 ***
## AcidIndex          -3.259e-01  1.117e-02 -29.169  < 2e-16 ***
## STARS               7.478e-01  1.946e-02  38.431  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.626 on 12780 degrees of freedom
## Multiple R-squared:  0.2879, Adjusted R-squared:  0.2871
## F-statistic: 369.1 on 14 and 12780 DF,  p-value: < 2.2e-16
```

Seems like this decreased out adjusted r^2 to 0.2871.

Multiple Linear Regression : Model 4

Now we will run the linear regression model on our dataset with the data from MICE.

```
##
## Call:
## lm(formula = TARGET ~ ., data = data_imp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2840 -0.9819  0.1864  1.0271  4.7735
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.742e+00  4.870e-01   7.683 1.66e-14 ***
## FixedAcidity       -2.598e-04  2.043e-03  -0.127 0.898823
## VolatileAcidity    -1.274e-01  1.623e-02  -7.851 4.45e-15 ***
## CitricAcid          3.143e-02  1.477e-02   2.128 0.033341 *
## ResidualSugar       7.277e-05  3.764e-04   0.193 0.846698
## Chlorides          -1.474e-01  3.994e-02  -3.690 0.000225 ***
## FreeSulfurDioxide   4.321e-04  8.567e-05   5.044 4.62e-07 ***
## TotalSulfurDioxide  2.588e-04  5.486e-05   4.717 2.42e-06 ***
## Density            -8.520e-01  4.791e-01  -1.778 0.075392 .
## pH                 -4.470e-02  1.874e-02  -2.385 0.017073 *
## Sulphates          -3.339e-02  1.366e-02  -2.444 0.014527 *
## Alcohol             1.342e-02  3.400e-03   3.947 7.97e-05 ***
## LabelAppeal         4.445e-01  1.496e-02  29.713  < 2e-16 ***
## AcidIndex          -2.501e-01  9.942e-03 -25.154  < 2e-16 ***
## STARS               1.119e+00  1.508e-02  74.199  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.436 on 12780 degrees of freedom
## Multiple R-squared:  0.4452, Adjusted R-squared:  0.4446
## F-statistic: 732.7 on 14 and 12780 DF,  p-value: < 2.2e-16
```

Looks like this has brought down our adjusted r^2 to 0.4443. Therefore, it seems like in this case the best model fit was achieved with the dataset where we recoded the missing values as 0 and used all the variables.

Before moving on let's try removing some of the variables to see if we can get a simpler model.

Multiple Linear Regression : Model 5

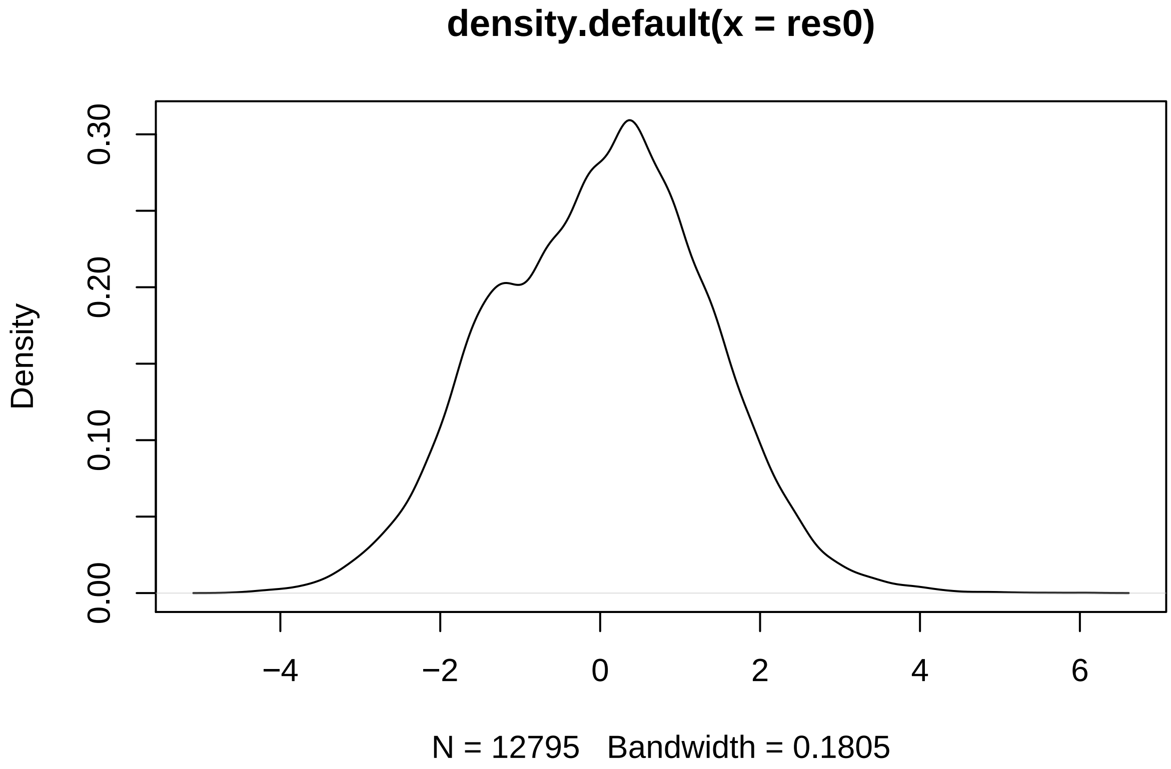
We removed some of the parameters that aren't statistically significant from model2 which so far has the highest adjusted r^2 value. We also chose to remove some variables that were statistically significant (Chlorides,TotalSulfurDioxide,FreeSulfurDioxide and VolatileAcidity) but the coefficients of the model were so small they had little impact. In favor of a simpler model with fewer parameters, these were removed.

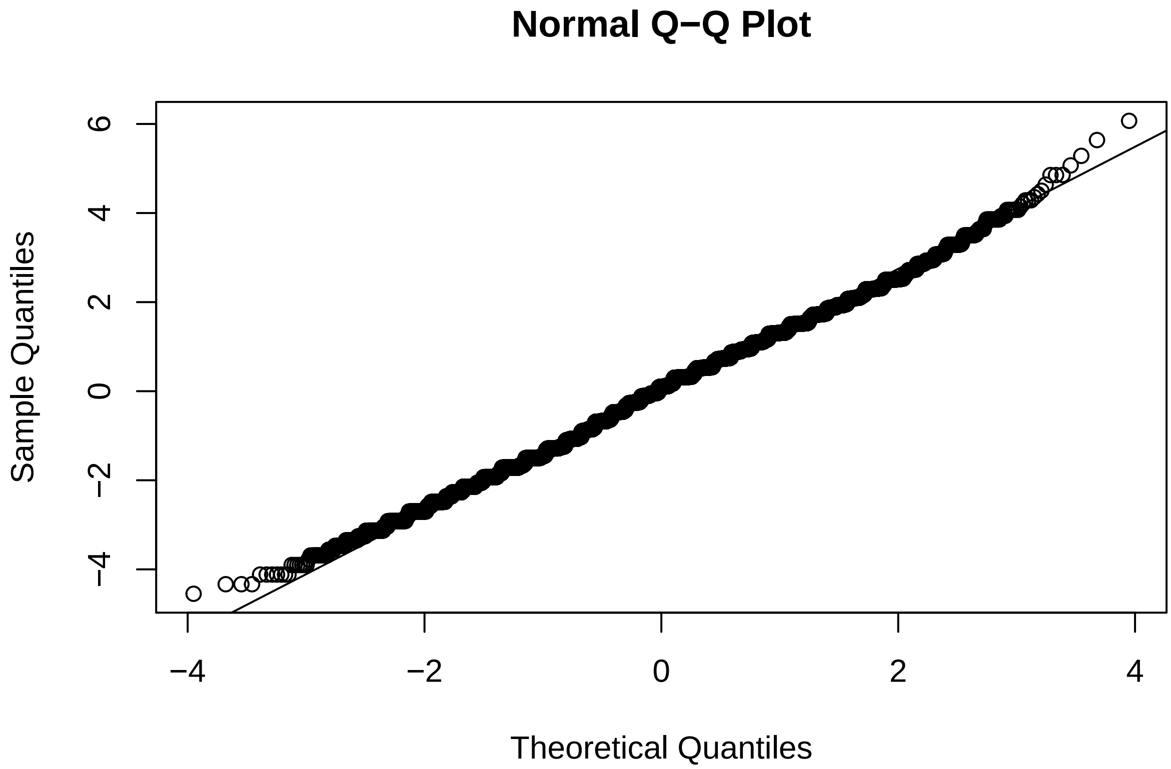
```
##
```

```
## Call:
## lm(formula = TARGET ~ LabelAppeal + AcidIndex + STARS, data = wine_train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5478 -0.9207  0.0973  0.9289  6.0697
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.212216   0.075692   42.44   <2e-16 ***
## LabelAppeal  0.430953   0.013718   31.41   <2e-16 ***
## AcidIndex   -0.214113   0.009037  -23.69   <2e-16 ***
## STARS        0.986226   0.010453   94.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.33 on 12791 degrees of freedom
## Multiple R-squared:  0.5236, Adjusted R-squared:  0.5235
## F-statistic:  4686 on 3 and 12791 DF,  p-value: < 2.2e-16
```

The adjusted r^2 to value is slightly higher at 0.5235

Let’s check the model fit for this model with diagnostic plots:





The density and qq plot for this model indicates that the residuals are normally distributed.

While this model maybe an adequate fit, we are going to develop Poisson Regression models next to see if we can get a better fit.

Poisson Regression STARS = 1 Where Missing: Model 1

First let us use our dataset where we recoded missing values in the predictor STARS as 0.

```
##
## Call:
## glm(formula = TARGET ~ ., family = poisson, data = wine_train1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9803  -0.7083   0.0639   0.5756   3.2351
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.618e+00  2.368e-01   6.830 8.49e-12 ***
## FixedAcidity       -1.785e-04  1.001e-03  -0.178 0.858472
## VolatileAcidity    -3.296e-02  7.888e-03  -4.178 2.94e-05 ***
## CitricAcid          4.358e-03  7.178e-03   0.607 0.543785
## ResidualSugar      -5.403e-05  1.831e-04  -0.295 0.767882
## Chlorides          -4.827e-02  1.939e-02  -2.489 0.012815 *
## FreeSulfurDioxide   1.275e-04  4.173e-05   3.057 0.002239 **
## TotalSulfurDioxide  9.401e-05  2.698e-05   3.484 0.000493 ***
## Density            -3.618e-01  2.332e-01  -1.552 0.120766
## pH                 -1.708e-02  9.073e-03  -1.883 0.059759 .
## Sulphates          -1.092e-02  6.657e-03  -1.640 0.101005
## Alcohol             1.492e-03  1.677e-03   0.890 0.373490
## LabelAppeal         1.324e-01  7.369e-03  17.963  < 2e-16 ***
## AcidIndex          -8.671e-02  5.479e-03 -15.824  < 2e-16 ***
## STARS               3.094e-01  5.532e-03  55.936  < 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: 15334.3  on 8674  degrees of freedom
## Residual deviance:  9962.1  on 8660  degrees of freedom
##   (4120 observations deleted due to missingness)
## AIC: 31705
##
## Number of Fisher Scoring iterations: 5
```

We see that the residual deviance is 9962 on 8660 degrees of freedom. Ideally the ratio of deviance to df should be 1. Otherwise there is overdispersion in the model.

```
##
##  Overdispersion test
##
## data:  poisson1
## z = -9.8815, p-value = 1
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##  0.8686586
```

There is no indication of overdispersion in our data.

Poisson Regression STARS = 1 Where Missing: Model 2

Our next poisson model, we will use the limited variables we found to be relevant in our linear model. Again we are using the dataset where we recoded missing values in the predictor STARS as 0.

```
##
## Call:
## glm(formula = TARGET ~ STARS + LabelAppeal + AcidIndex, family = poisson,
##     data = wine_train1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9872  -0.7168   0.0485   0.5527   3.2791
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.223551   0.036514   33.51   <2e-16 ***
## STARS        0.313946   0.004507   69.65   <2e-16 ***
## LabelAppeal  0.132978   0.006060   21.95   <2e-16 ***
## AcidIndex   -0.088835   0.004462  -19.91   <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: 14804  on 12791  degrees of freedom
## AIC: 46754
##
## Number of Fisher Scoring iterations: 5
```

We see that the residual deviance is 22861 on 12794 degrees of freedom. Ideally the ratio of deviance to df should be 1. Otherwise there is overdispersion in the model.

```
##
##  Overdispersion test
##
## data:  poisson2
## z = -11.701, p-value = 1
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   0.871882
```

There is no indication of overdispersion in our data.

Negative Binomial: Model1

```
##
## Call:
## glm.nb(formula = TARGET ~ ., data = wine_train1, init.theta = 49024.77017,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9802  -0.7083   0.0639   0.5756   3.2350
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.618e+00  2.368e-01   6.830 8.50e-12 ***
## FixedAcidity       -1.785e-04  1.001e-03  -0.178 0.858492
## VolatileAcidity    -3.296e-02  7.888e-03  -4.178 2.94e-05 ***
## CitricAcid          4.358e-03  7.178e-03   0.607 0.543793
## ResidualSugar      -5.402e-05  1.831e-04  -0.295 0.767908
## Chlorides          -4.827e-02  1.939e-02  -2.489 0.012816 *
## FreeSulfurDioxide   1.276e-04  4.173e-05   3.056 0.002240 **
## TotalSulfurDioxide  9.401e-05  2.698e-05   3.484 0.000493 ***
## Density            -3.618e-01  2.332e-01  -1.552 0.120773
## pH                 -1.708e-02  9.074e-03  -1.883 0.059760 .
## Sulphates          -1.092e-02  6.657e-03  -1.640 0.101004
## Alcohol             1.492e-03  1.677e-03   0.890 0.373527
## LabelAppeal         1.324e-01  7.369e-03  17.963  < 2e-16 ***
## AcidIndex          -8.671e-02  5.479e-03 -15.824  < 2e-16 ***
## STARS               3.094e-01  5.532e-03  55.935  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(49024.77) family taken to be 1)
##
##     Null deviance: 15333.7  on 8674  degrees of freedom
## Residual deviance:  9961.8  on 8660  degrees of freedom
##   (4120 observations deleted due to missingness)
## AIC: 31708
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  49025
##           Std. Err.:  61688
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -31675.54
```

Here our output from our first poisson model is exactly the same as our first negative binomial model.

Negative Binomial: Model2

For our second negative binomial we are going to use the function **stepAIC()** to complete forward selection to see how it compares to our other models. We don't want to use the limited set of variables (STARS,LabelAppeal, and AcidIndex) as we've done before because we know that the model output would match our second Poisson model exactly.

The stepAIC cannot handle NA values so we are going to use our wine_train1_imputed_median, which has used the median of the data column for missing data points after setting missing STARS values equal to 0.

```
## Start:  AIC=46700.55
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + ResidualSugar +
##     Chlorides + FreeSulfurDioxide + TotalSulfurDioxide + Density +
##     pH + Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS
##
##                      Df   AIC
## - ResidualSugar       1 46699
## - FixedAcidity        1 46699
## - CitricAcid          1 46700
## <none>                  46701
## - Density             1 46701
## - Alcohol             1 46701
## - pH                  1 46703
## - Sulphates           1 46703
## - Chlorides           1 46705
## - FreeSulfurDioxide   1 46711
## - TotalSulfurDioxide  1 46712
## - VolatileAcidity     1 46725
## - AcidIndex           1 47082
## - LabelAppeal         1 47181
## - STARS               1 51524
##
## Step:  AIC=46698.68
## TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates +
##     Alcohol + LabelAppeal + AcidIndex + STARS
##
##                      Df   AIC
## - FixedAcidity        1 46697
## - CitricAcid          1 46698
## <none>                  46699
## - Density             1 46699
## - Alcohol             1 46699
## - pH                  1 46701
## - Sulphates           1 46702
## - Chlorides           1 46703
## - FreeSulfurDioxide   1 46709
## - TotalSulfurDioxide  1 46710
## - VolatileAcidity     1 46723
## - AcidIndex           1 47080
## - LabelAppeal         1 47179
## - STARS               1 51524
##
## Step:  AIC=46696.82
## TARGET ~ VolatileAcidity + CitricAcid + Chlorides + FreeSulfurDioxide +
##     TotalSulfurDioxide + Density + pH + Sulphates + Alcohol +
##     LabelAppeal + AcidIndex + STARS
##
##                      Df   AIC
## - CitricAcid          1 46697
## <none>                  46697
## - Density             1 46697
## - Alcohol             1 46697
## - pH                  1 46699
## - Sulphates           1 46700
## - Chlorides           1 46701
## - FreeSulfurDioxide   1 46708
## - TotalSulfurDioxide  1 46708
## - VolatileAcidity     1 46721
## - AcidIndex           1 47090
## - LabelAppeal         1 47177
## - STARS               1 51522
##
## Step:  AIC=46696.54
## TARGET ~ VolatileAcidity + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     Density + pH + Sulphates + Alcohol + LabelAppeal + AcidIndex +
##     STARS
##
##                      Df   AIC
## <none>                  46697
## - Density             1 46697
## - Alcohol             1 46697
## - pH                  1 46699
## - Sulphates           1 46699
## - Chlorides           1 46701
## - FreeSulfurDioxide   1 46707
## - TotalSulfurDioxide  1 46708
## - VolatileAcidity     1 46721
## - AcidIndex           1 47088
## - LabelAppeal         1 47177
## - STARS               1 51526
```

	Linear Model #5	Poisson #1	Poisson #2
Description	Fewer variables and missing STARS variables recoded as 0	STARS = 0 where missing	STARS = 0 wher
AIC	43609.4502600085	31705.3531578421	46754.4249914696
MSE	1.76766780724601	0.420870738532527	0.42351532120423

Select Models

Best Linear Model

The best linear model was chosen based on adjusted R^2 value, ~52.4%, and the least number of variables. It accounts for the most variance in our data. The STARS value has the highest impact on the score of a wine. We will compare this model to our poisson and negative binomial models below.

```
## (Intercept) LabelAppeal   AcidIndex       STARS
##   3.2122159   0.4309528  -0.2141127   0.9862259
```

Comparing Models

We're going to use AIC and MSE values to choose the best model

Here we see our Poisson #1 is the best model because it has the lowest AIC and MSE. Our first negative binomial model is the same but we'll choose to use our poisson model moving forward.

Prediction on the Test Data

The coefficients of our best model are:

```
##        (Intercept)       FixedAcidity    VolatileAcidity         CitricAcid
##       1.617560e+00      -1.784975e-04      -3.295863e-02       4.357525e-03
##      ResidualSugar          Chlorides FreeSulfurDioxide TotalSulfurDioxide
##      -5.402792e-05      -4.826940e-02       1.275442e-04       9.401132e-05
##            Density                 pH          Sulphates            Alcohol
##      -3.617887e-01      -1.708142e-02      -1.091699e-02       1.492248e-03
##        LabelAppeal          AcidIndex              STARS
##       1.323698e-01      -8.670583e-02       3.094151e-01
```

Now we will use our poisson model to run our test data.

```
## [1] "The top 6 predictions are:"
```

```
##         1         2         3         4         5         6
## 0.5991587 1.2403192 0.7300546 0.7175504 0.3470400 1.8610080
```

Code Appendix

```
knitr::opts_chunk$set(echo=FALSE, error=FALSE, warning=FALSE, message=FALSE, fig.align = "center")
# Libraries


library(DataExplorer)
library(visdat)
library(dplyr)
library(tidyr)
library(MASS)
library(psych)
library(AER)
library(mlr)
library(mice)
library(imputeTS)

set.seed(621)
# training data
wine_train <- read.csv('https://raw.githubusercontent.com/hillt5/DATA_621/master/HW5/wine-training-data
  dplyr::select(-1)

# test data
wine_test <- read.csv('https://raw.githubusercontent.com/hillt5/DATA_621/master/HW5/wine-evaluation-dat
  dplyr::select(-1)
wine_train %>% dplyr::select(-1) %>% describe()
plot_histogram(wine_train[-1], geom_histogram_args = list("fill" = "tomato4"))
tibble(wine_train %>% summarize_all(n_distinct))
forcorr <- wine_train[complete.cases(wine_train),-1]
corrplot::corrplot(cor(forcorr), type = 'lower')


colSums(is.na(wine_train))


vis_dat(wine_train  %>% dplyr:: select(pH, ResidualSugar, Chlorides, Alcohol, FreeSulfurDioxide , TotalS



plot_missing(wine_train)


100*round((wine_train %>% drop_na() %>% nrow())/nrow(wine_train), 3) ##Number of observations with comp

wine_train1 <- wine_train %>%
  mutate(STARS = replace_na(STARS, 0))  ## Recode missing STARS ratings as '0'


#wine_impute <- mlr:: impute(wine_train, classes = list(numeric = mlr::imputeMedian()))
wine_impute <- imputeTS ::na_mean(wine_train, option = "median")
imp <- mice:: mice(wine_train, method = "rf", m = 1)
# Store data
data_imp <- complete(imp)
wine_train1_imputed_median <- wine_train1
wine_train1_imputed_median <- imputeTS ::na_mean(wine_train1_imputed_median, option = "median")


summary(lm(wine_train, formula = TARGET ~.))

model2<- lm(wine_train1, formula = TARGET ~.)
summary(model2)


summary(lm(wine_impute, formula = TARGET ~ .))

summary(lm(data_imp, formula = TARGET ~.))
```

```
model5 <- lm(wine_train1, formula = TARGET ~ LabelAppeal + AcidIndex + STARS)
summary(model5)

library(ggplot2)
res0 <- resid(model5)
plot(density(res0))
qqnorm(res0)
qqline(res0)
poisson1 <- glm(wine_train1, formula = TARGET ~., family = poisson)
summary(poisson1)

dispersiontest(poisson1)
poisson2 <- glm(wine_train1, formula = TARGET ~ STARS + LabelAppeal + AcidIndex, family = poisson)
summary(poisson2)

dispersiontest(poisson2)
nb1 <- glm.nb(wine_train1, formula = TARGET ~. )
summary(nb1)

nb2 <- glm.nb(wine_train1_imputed_median, formula = TARGET ~ .)

stepmodel <- stepAIC(nb2,selection='forward')

coefficients(model5)
library(dvmisc)
library(kableExtra)

lm5 <- c('Fewer variables and missing STARS variables recoded as 0',AIC(model5),mean(model5$residuals^2
p1 <- c('STARS = 0 where missing',poisson1$aic,mean(poisson1$residuals^2))
p2 <- c('STARS = 0 where missing with fewer variables',poisson2$aic,get_mse(poisson2))
nb1 <- c('STARS = 0 where missing',31708,mean(nb1$residuals^2))
nb2 <- c('STARS = 0 where missing with fewer variables',AIC(stepmodel),mean(stepmodel$residuals^2))

results <- cbind(lm5,p1,p2,nb1,nb2)

colnames(results) <- c('Linear Model #5', 'Poisson #1', 'Poisson #2','Negative Binomial #1', 'Negative
rownames(results) <- c('Description','AIC','MSE')

results %>%
  kable() %>%
  kable_styling()

poisson1$coefficients
wine_test1 <-  wine_test %>%
  mutate(STARS = replace_na(STARS, 0))
wine_test1 <- subset(wine_test1,select = -c(TARGET))
predictions <- predict(poisson1,wine_test1)
print("The top 6 predictions are:")
head(predictions)
```