CUNY DATA 621 HW5: Wine

Group 2: Elina Azrilyan, Charls Joseph, Mary Anna Kivenson, Sunny Mehta, Vinayak Patel

May 09, 2020

Data Exploration

```
##   TARGET FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
## 1      3          3.2           1.160      -0.98          54.2    -0.567
## 2      3          4.5           0.160      -0.81          26.1    -0.425
## 3      5          7.1           2.640      -0.88          14.8     0.037
## 4      3          5.7           0.385       0.04          18.8    -0.425
## 5      4          8.0           0.330      -1.26           9.4        NA
## 6      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
```

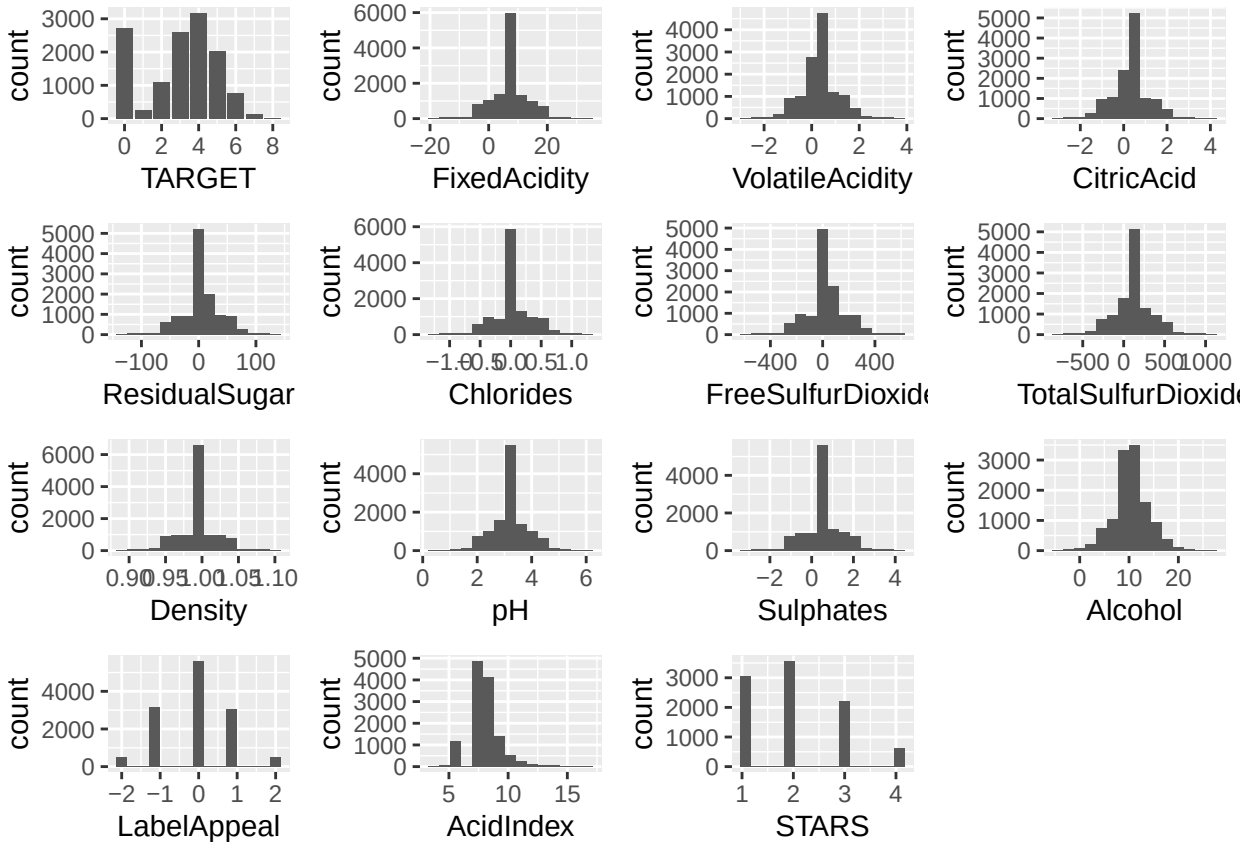
Taking a look at a summary of the data, there seem to be many missing values in the `ResidualSugar`,`Chlorides`,`FreeSulfurDioxide`,`TotalSulfurDioxide`,`pH`,`Sulphate` and `STARS` fields. The `STARS` and `LabelAppeal` columns are both ordinal variables and may need to be transformed into dummy variables.

```
##      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.   :-1.1710   Min.   :-555.00   Min.   :-823.0
##  1st Qu.:  -2.000   1st Qu.:-0.0310   1st Qu.:   0.00   1st Qu.:  27.0
##  Median :   3.900   Median : 0.0460   Median :  30.00   Median : 123.0
##  Mean   :   5.419   Mean   : 0.0548   Mean   :  30.85   Mean   : 120.7
##  3rd Qu.:  15.900   3rd Qu.: 0.1530   3rd Qu.:  70.00   3rd Qu.: 208.0
##  Max.   : 141.150   Max.   : 1.3510   Max.   : 623.00   Max.   :1057.0
##  NA's   :616        NA's   :638       NA's   :647       NA's   :682
##     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   : 0.5271   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
##                   NA's   :395     NA's   :1210      NA's   :653
##   LabelAppeal          AcidIndex          STARS
##  Min.   :-2.000000   Min.   : 4.000   Min.   :1.000
##  1st Qu.:-1.000000   1st Qu.: 7.000   1st Qu.:1.000
##  Median : 0.000000   Median : 8.000   Median :2.000
##  Mean   :-0.009066   Mean   : 7.773   Mean   :2.042
##  3rd Qu.: 1.000000   3rd Qu.: 8.000   3rd Qu.:3.000
##  Max.   : 2.000000   Max.   :17.000   Max.   :4.000
##                                       NA's   :3359
```

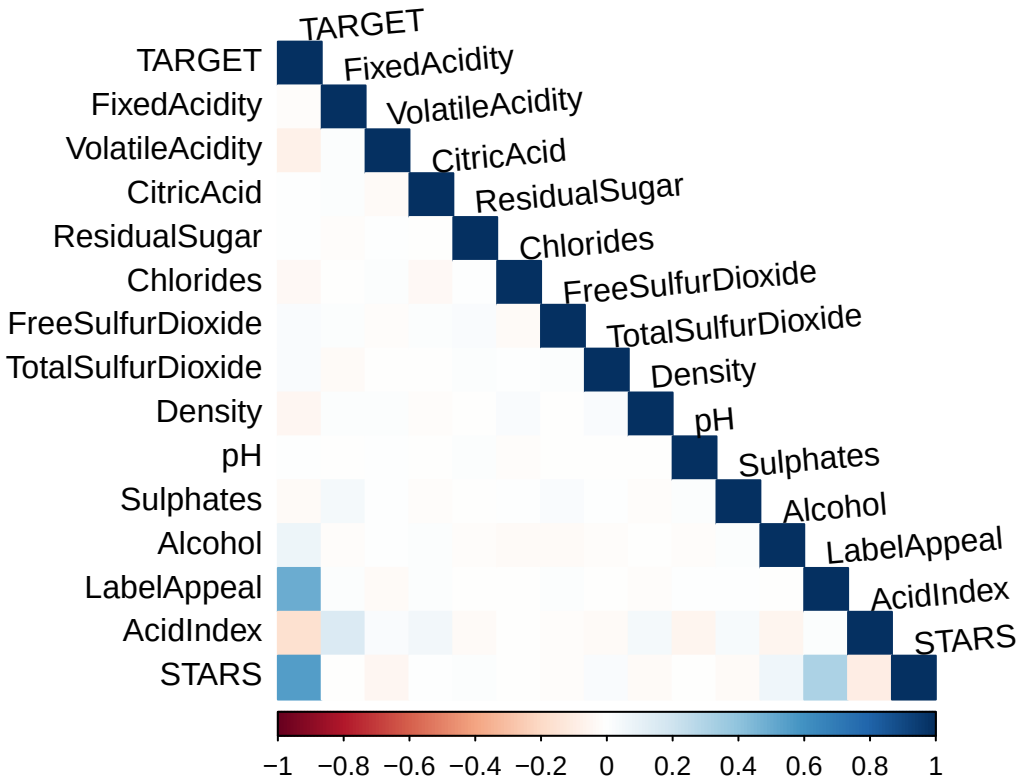
Distributions

The following histograms help visualize the distributions of numerical variables in this dataset. Many of the predictor variables have a narrow spread and have high occurances at the center of the distribution. Normalizing the data may help make the distributions of variables more normal.



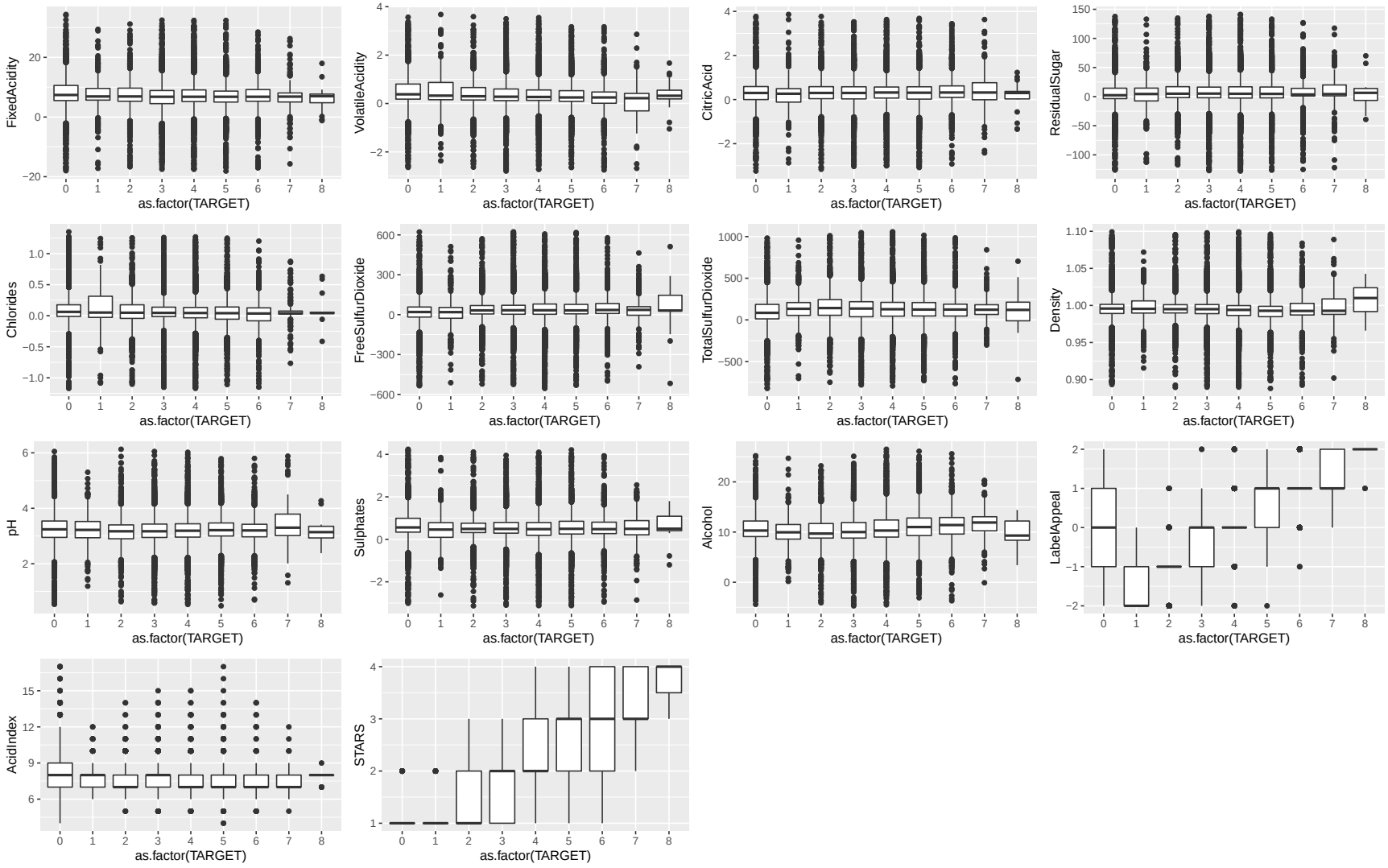
Correlation Plot

This correlation plot shows that there is no multicollinearity in the dataset. The correlations between STARS, AcidIndex, LabelAppeal and TARGET are strong. The remaining predictors have little to no correlation with TARGET.



Box Plots

The weak correlations between most of the predictors and TARGET were suprising. The following box plots provide a more in-depth view at the relationship between predictors and the target variable. The plots confirm that the relationship between target and most of the features appears limited.



Preprocessing

Train Test Split

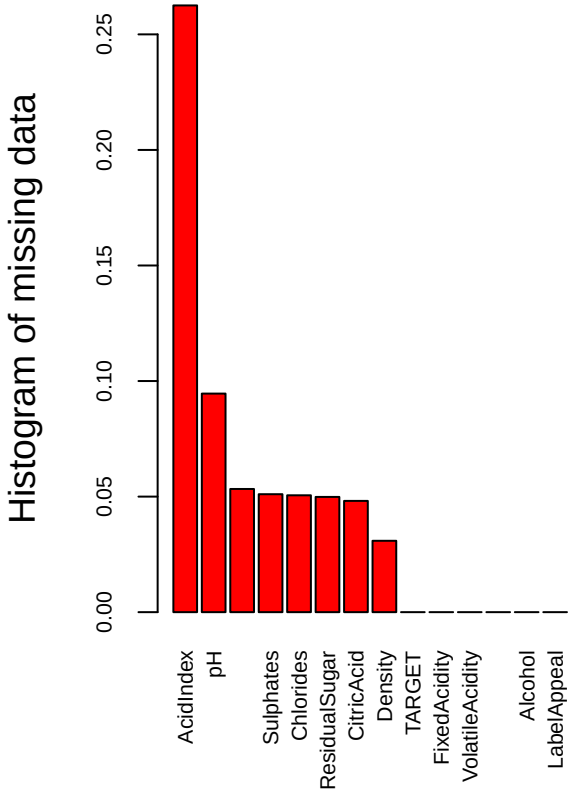
Encoding

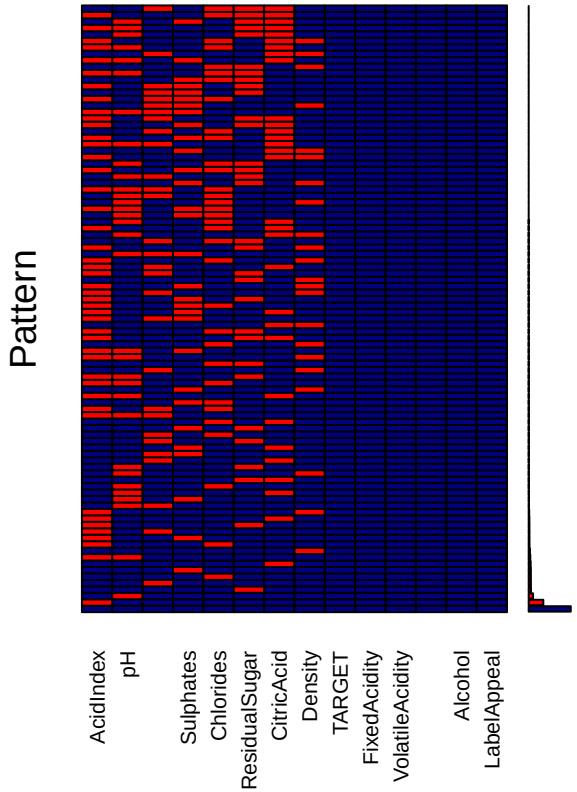
The `STARS` and `LabelAppeal` columns contain ordinal data. Using ordinal variables as-is in a model requires the assumption that categories are equally spaced. Since stars and label appeal are both subjective labels, this assumption may not hold true. To resolve this, these ordinal columns will be encoded into dummy variables.

```
##   FixedAcidity VolatileAcidity CitricAcid ResidualSugar Chlorides
## 1          3.2           1.160      -0.98          54.2    -0.567
## 2          4.5           0.160      -0.81          26.1    -0.425
## 3          7.1           2.640      -0.88          14.8     0.037
## 4          5.7           0.385       0.04          18.8    -0.425
## 5          8.0           0.330      -1.26           9.4        NA
## 6         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
##   AcidIndex STARS.1 STARS.2 STARS.3 STARS.4 LabelAppeal.N2 LabelAppeal.N1
## 1         8       0       1       0       0              0              0
## 2         7       0       0       1       0              0              1
## 3         8       0       0       1       0              0              1
## 4         6       1       0       0       0              0              1
## 5         9       0       1       0       0              0              0
## 6        11       0       0       0       0              0              0
##   LabelAppeal.P1 LabelAppeal.P2
## 1              0              0
## 2              0              0
## 3              0              0
## 4              0              0
## 5              0              0
## 6              0              0
```

Missing Data

The following plots provide a visualization of missing data. There appears to be a patten in the mising values, so it will be useful to include a flag for missing data. KNN imputation is unsupervised, meaning it does not require a target variable. A train test split was performed earlier so that only predictor data is used for imputation.






```
##
##  Variables sorted by number of missings:
##             Variable      Count
##            AcidIndex 0.26252442
##                   pH 0.09456819
##   FreeSulfurDioxide 0.05330207
##            Sulphates 0.05103556
##            Chlorides 0.05056663
##       ResidualSugar 0.04986323
##          CitricAcid 0.04814381
##             Density 0.03087143
```

```
##              TARGET 0.00000000
##        FixedAcidity 0.00000000
##     VolatileAcidity 0.00000000
##  TotalSulfurDioxide 0.00000000
##             Alcohol 0.00000000
##         LabelAppeal 0.00000000
```

To fill missing values, knn imputation will be used. As part of knn imputation, the data will also be centered and scaled. An additional column will be added to identify the percent of missing values in each row.

Model Building

We take the dataframe with missing values now imputed, and first back-convert the binary variables back to 1s and 0s. (We do not want to center and scale binary variables.)

Next we split the data in half, fit a poisson regression model using one half to predict the second half TARGET (out-of-sample) and calculate the RMSE. We do this 100 times and keep track of the RMSE each time.

Below is the mean RMSE from our 100 trials.

```
## [1] 2.589679
```

For comparison sake, we take the original dataset and impute the missing values using simple medians of each columns. But since a missing value for the `STARS` variable may likely be a negative indicator, we impute 0 there instead of the median.

Next we repeat the procedure of splitting our data in half, fitting a poisson regression model on one half to predict the TARGET variable in the second half, and keeping track of the RMSE each time.

Below is the mean RMSE from our 100 trials.

```
## [1] 2.589518
```

Note it is not substantially different from before.

Either way, we have successfully imputed missing data in order to build an effective regression model.