Assignment5_Data621

Mubashira Qari, Marco Castro

2025-04-13

Data Exploration

The training dataset has 12,795 observations across 15 columns and an additional INDEX column. For the purposes of our analysis, we will drop the INDEX column. The parameters STARS, AcidIndex, LabelAppeal appear to be categorical, while the remaining 12 variables, including our TARGET variable, appear numerical. The number of cases purchased (TARGET) ranges from 0-8. Roughly 21.4% (2734) of our observations had a TARGET value of zero.

```
## Rows: 12,795
## Columns: 15
## $ TARGET             <dbl> 3, 3, 5, 3, 4, 0, 0, 4, 3, 6, 0, 4, 3, 7, 4, 0, 0, ~
## $ AcidIndex          <fct> 8, 7, 8, 6, 9, 11, 8, 7, 6, 8, 5, 10, 7, 8, 9, 8, 9~
## $ Alcohol            <dbl> 9.9, NA, 22.0, 6.2, 13.7, 15.4, 10.3, 11.6, 15.0, 1~
## $ Chlorides          <dbl> -0.567, -0.425, 0.037, -0.425, NA, 0.556, 0.060, 0.~
## $ CitricAcid         <dbl> -0.98, -0.81, -0.88, 0.04, -1.26, 0.59, -0.40, 0.34~
## $ Density            <dbl> 0.99280, 1.02792, 0.99518, 0.99640, 0.99457, 0.9994~
## $ FixedAcidity       <dbl> 3.2, 4.5, 7.1, 5.7, 8.0, 11.3, 7.7, 6.5, 14.8, 5.5,~
## $ FreeSulfurDioxide  <dbl> NA, 15, 214, 22, -167, -37, 287, 523, -213, 62, 551~
## $ LabelAppeal        <fct> 0, -1, -1, -1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 2, 0, 0, ~
## $ pH                 <dbl> 3.33, 3.38, 3.12, 2.24, 3.12, 3.20, 3.49, 3.20, 4.9~
## $ ResidualSugar      <dbl> 54.20, 26.10, 14.80, 18.80, 9.40, 2.20, 21.50, 1.40~
## $ STARS              <fct> 2, 3, 3, 1, 2, NA, NA, 3, NA, 4, 1, 2, 2, 3, NA, NA~
## $ Sulphates          <dbl> -0.59, 0.70, 0.48, 1.83, 1.77, 1.29, 1.21, NA, 0.26~
## $ TotalSulfurDioxide <dbl> 268, -327, 142, 115, 108, 15, 156, 551, NA, 180, 65~
## $ VolatileAcidity    <dbl> 1.160, 0.160, 2.640, 0.385, 0.330, 0.320, 0.290, -1~
```

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

Missing Values

Additionally, eight parameters had missing values ranging from 395 missing values (pH) to 3359 missing values (STARS). Below is the full list of variables with missing values:

```
##            Alcohol          Chlorides  FreeSulfurDioxide                 pH 
##                653                638                647                395 
##      ResidualSugar              STARS          Sulphates TotalSulfurDioxide 
##                616               3359               1210                682
```

Examining Numerical Variables

A review of the summary statistics reveals issues with our data. In particular, nine of the 11 numeric variables show minimum values below zero. Table 1 shows number of missing values.

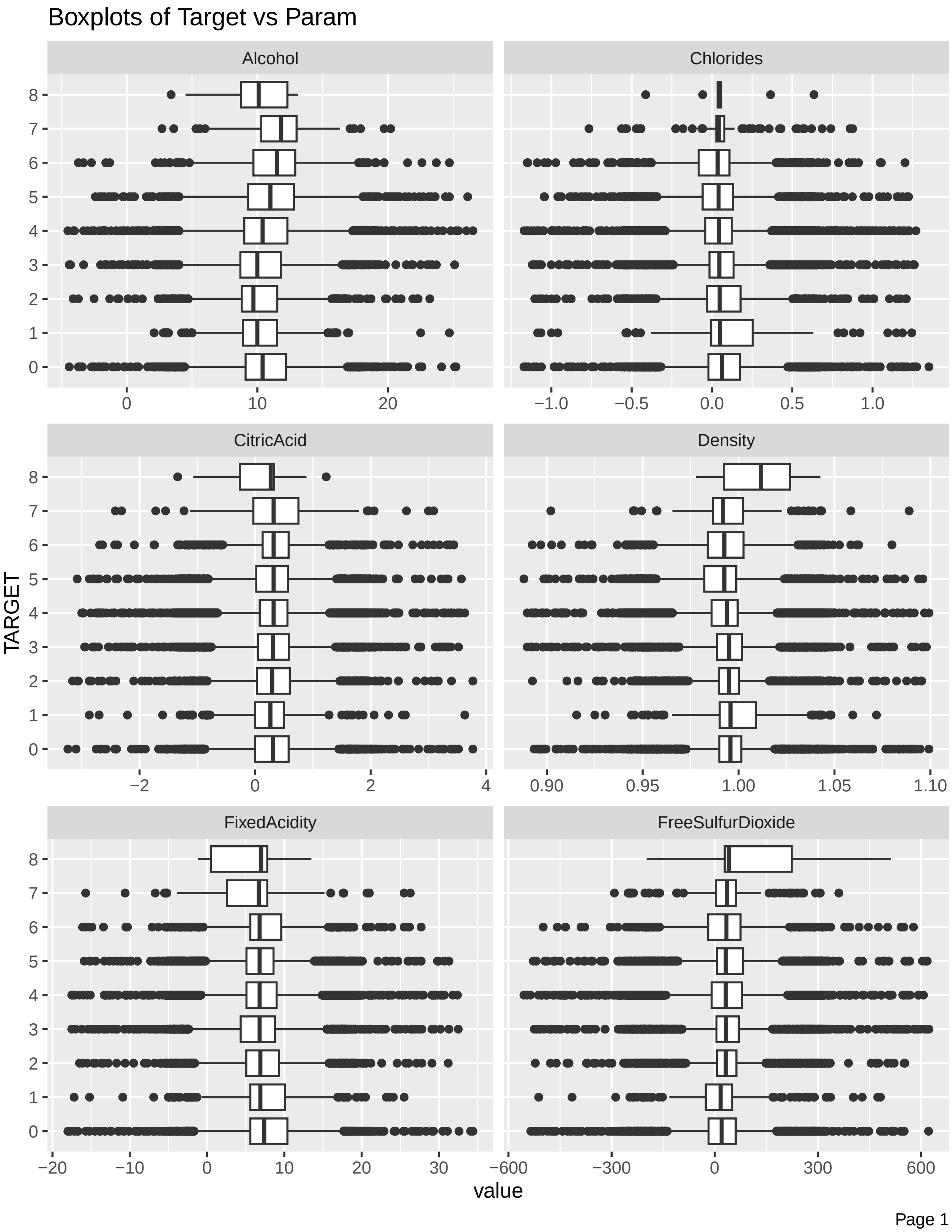
```
##      TARGET         Alcohol        Chlorides         CitricAcid
##  Min.   :0.000   Min.   :-4.70   Min.   :-1.1710   Min.   :-3.2400
##  1st Qu.:2.000   1st Qu.: 9.00   1st Qu.:-0.0310   1st Qu.: 0.0300
##  Median :3.000   Median :10.40   Median : 0.0460   Median : 0.3100
##  Mean   :3.029  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

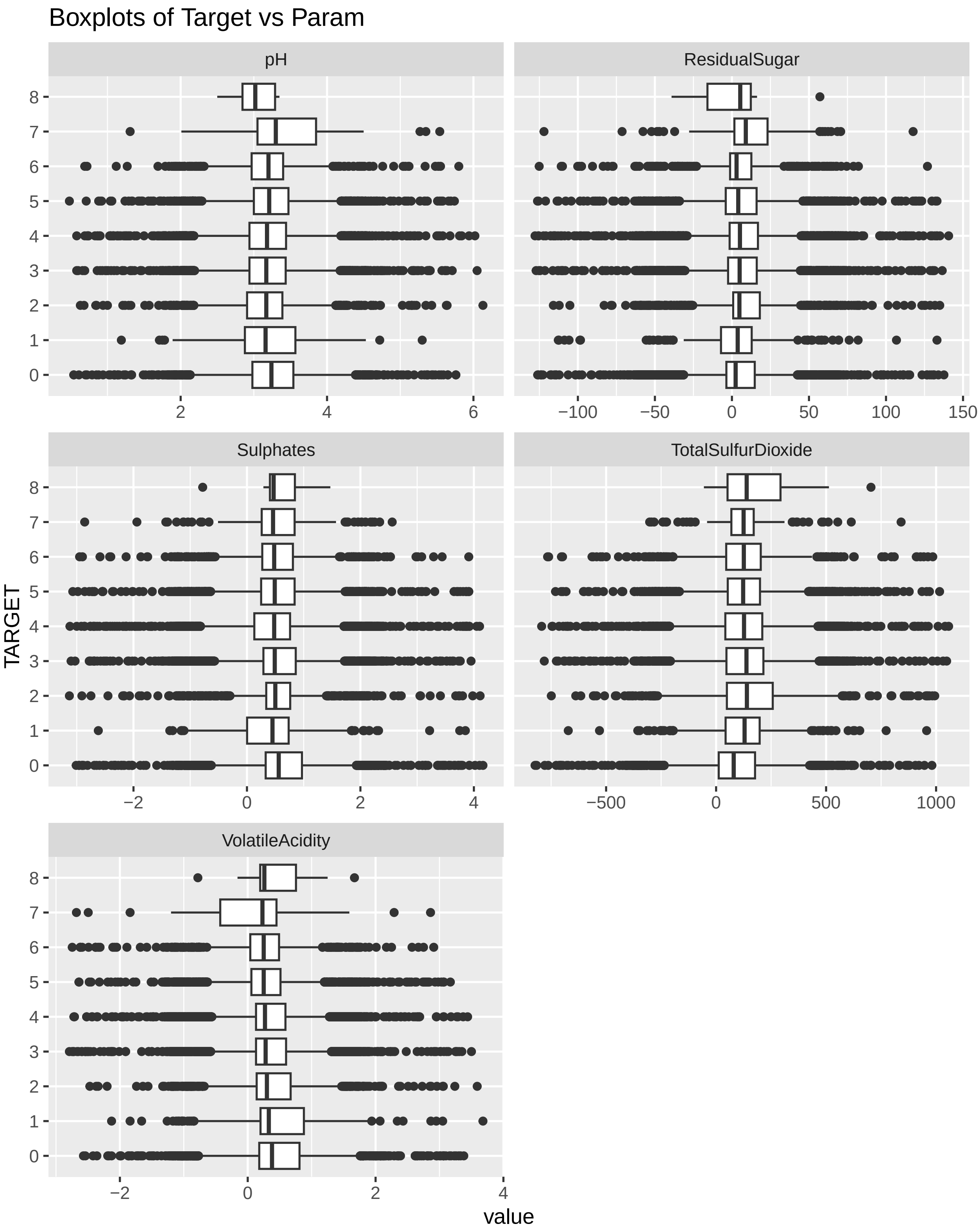
When put in the context of our specific properties of wine they represent, these negative values appear to be erroneous. For example, we expected `Alcohol Content` to have a minimum value of zero instead of a negative value. The same can be said of the other parameters with negative values (Chlorides, Citric Acid, Fixed Acidity, Free Sulfur Dioxide, Residual Sugar, Sulphates, Total Sulfur Dioxie, and Volatile Acidity). This suggests possible data entry errors or normalization that shifted our actual values to the left.

Table 1: Number of negative values

Variable	Rows Below Zero	NormalRange
TARGET	0	0 or higher
Alcohol	118	8% – 15% ABV
Chlorides	3197	0.01 – 0.10 g/L
CitricAcid	2966	0 – 1.0 g/L
Density	0	0.990 – 1.005 g/cm^3
FixedAcidity	1621	4 – 9 g/L
FreeSulfurDioxide	3036	10 – 70 mg/L
pH	0	2.9 – 4.0
ResidualSugar	3136	0 – 45 g/L
Sulphates	2361	0.3 – 1.0 g/L
TotalSulfurDioxide	2504	30 – 150 mg/L
VolatileAcidity	2827	0.2 – 0.8 g/L

A look at our boxplots shows the IQR's for each parameter are centered around a similar x-axis for each of our case counts. The boxplots confirm the presence of extreme values at lower as well as on the upper ranges. It should be noted the IQRs for the affected variables are in line with their corresponding typical ranges according to VineEnology.com as shown in Table 1.


Boxplots of Target vs Param
Alcohol
Chlorides
CitricAcid
Density
FixedAcidity
FreeSulfurDioxide
TARGET
8
7
6
5
4
3
2
1
0
0
10
20
−1.0
−0.5
0.0
0.5
1.0
−2
0
2
4
0.90
0.95
1.00
1.05
1.10
−20
−10
0
10
20
30
−600
−300
0
300
600
value
Page 1

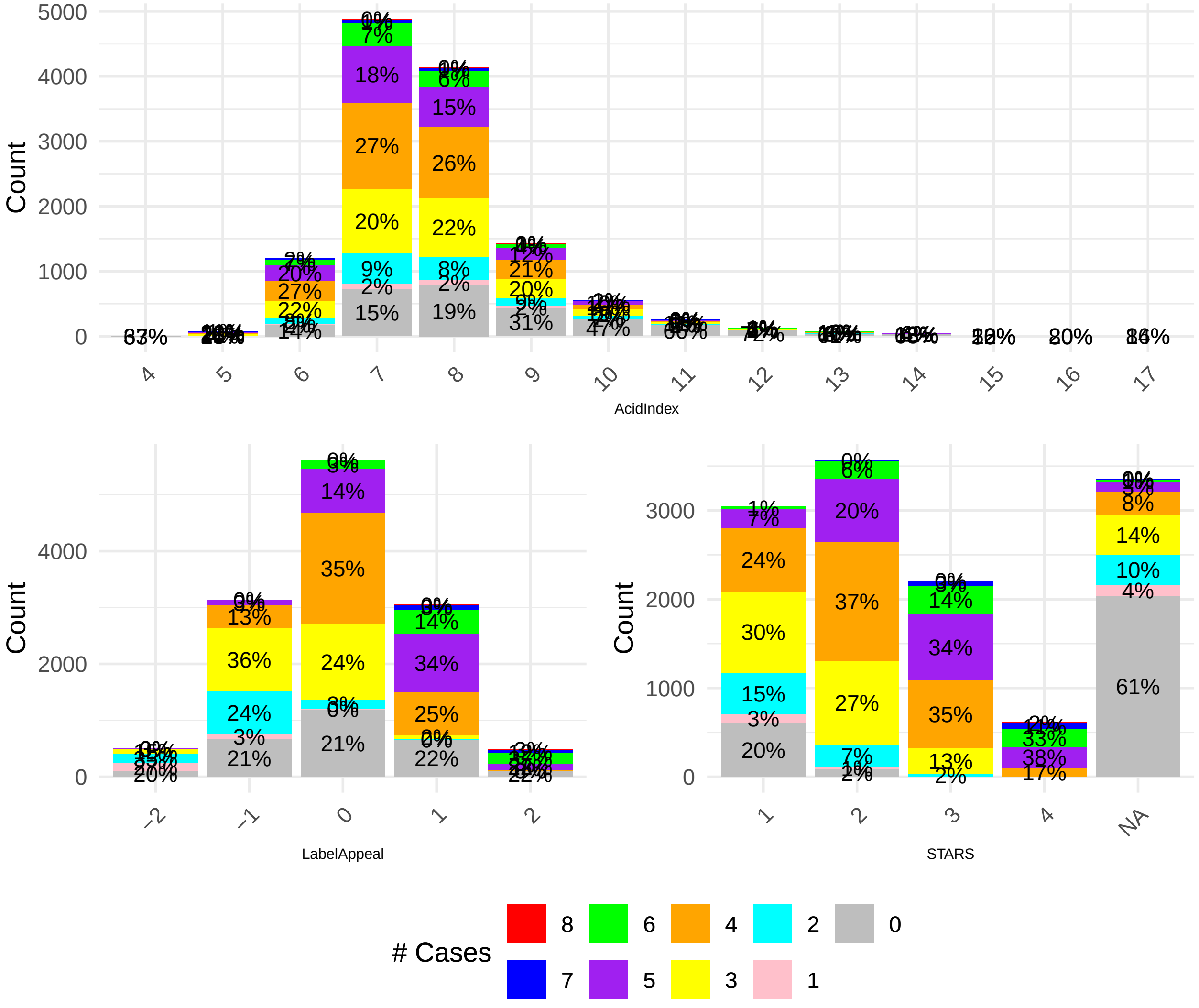

Boxplots of Target vs Param
pH
ResidualSugar
Sulphates
TotalSulfurDioxide
VolatileAcidity
TARGET
value
0
1
2
3
4
5
6
7
8
2
4
6
−100
−50
0
50
100
150
−2
0
2
4
−500
0
500
1000
−2
0
2
4

Page 2

Examining Categorical Variables

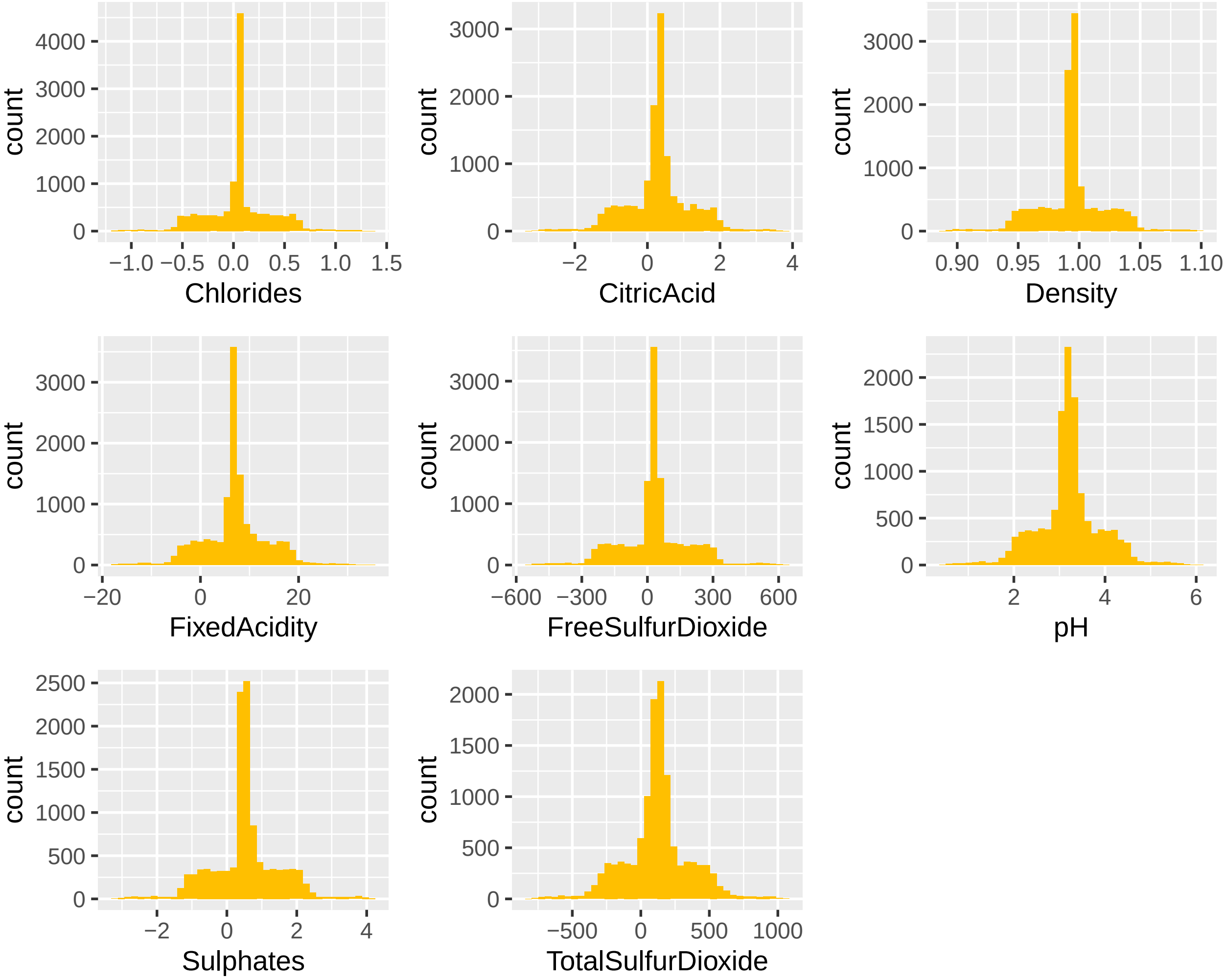
Visualizing the distributions of out categorical variables helps ensure variables are treated as discrete categories, not continuous numbers. Our AcidIndex variable shows that majority of values fall between 6 and 11 with 342 observations collectively making up the remaining values. We may want to bin values for this parameter. Label Appeal has a normal distribution ranging from -2 to 2 and centered around 0. The our Wine rative variable STARS has the most missing values (3359) of any variable; of missing value, 61% of rows had 0 cases purchased. As would be expected, the majority of observations have a low STARS value (1/2), while few observations have a perfect value of 4.

```
##    AcidIndex     LabelAppeal   STARS
##  7       :4878   -2: 504     1   :3042
##  8       :4142   -1:3136     2   :3570
##  9       :1427   0 :5617     3   :2212
##  6       :1197   1 :3048     4   : 612
##  10      : 551   2 : 490     NA's:3359
##  11      : 258
##  (Other): 342
```



Visualizing Distributions

Next we will visualize the distributions for our numeric variables. Using the histograms, we can quickly spot skewness, check distribution and value ranges, and identify variables with spikes or unusual spread. The histograms show symmetric unimodal distributions strongly peaked with thin tails across our numeric variables.



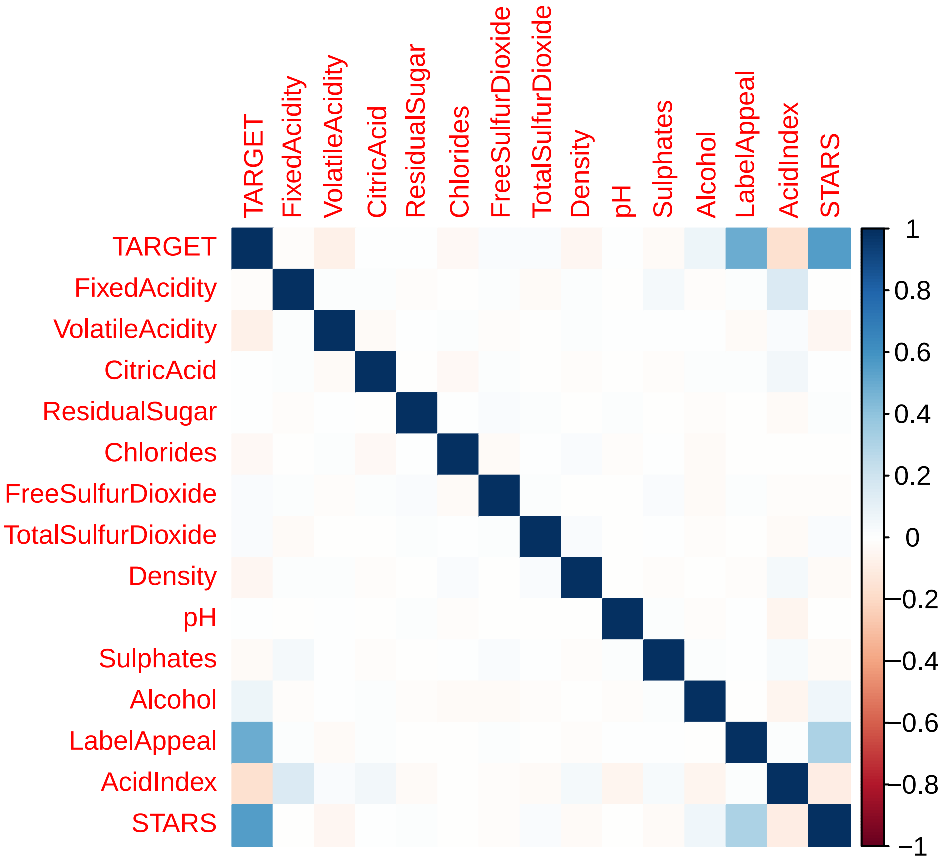
Our skewness test confirms that our numerical variables are nearly symmetrical or almost symmetrical.

```r
# Calculate skewness for each numeric variable (using original values)
skew_vals <- sapply(numeric_df |> subset(select=-c(TARGET)), function(x) skewness(x, na.rm = TRUE, type

# Create a dataframe with skewness
skew_df <- data.frame(
  Variable = names(skew_vals),
  Skewness = skew_vals
)

# Sort by highest absolute skewness
skew_df <- skew_df[order(-abs(skew_df$Skewness)), ]

# Show top 10 most skewed variables (untransformed)
head(skew_df, 10)
```

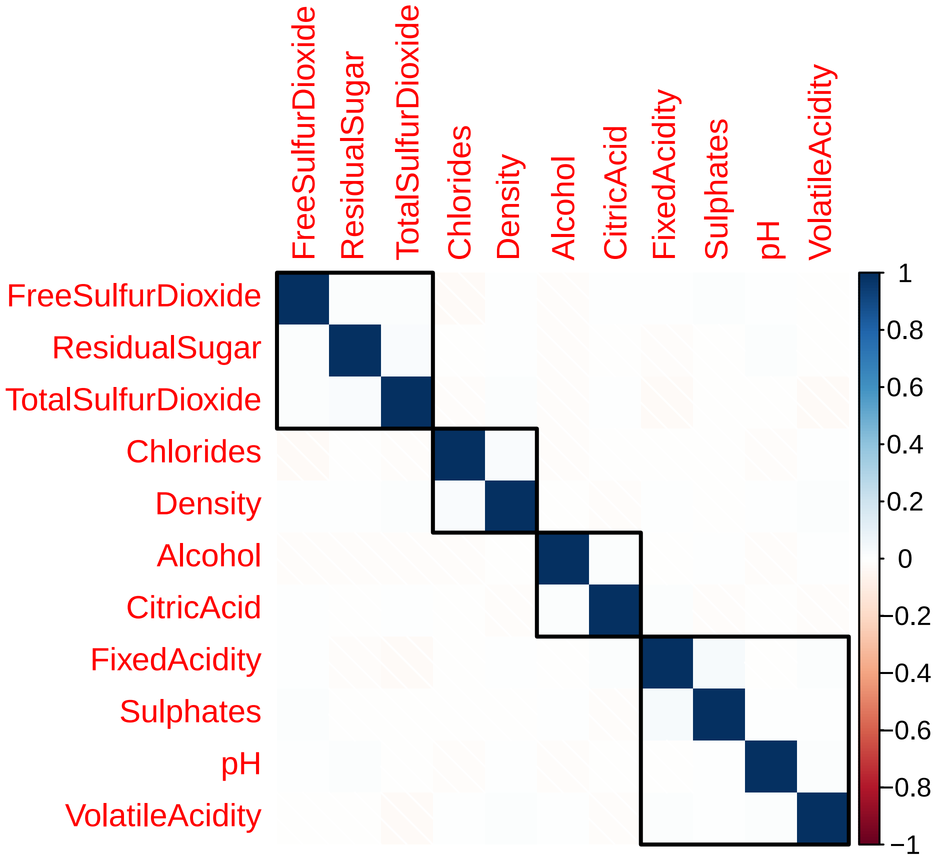
```
##                              Variable     Skewness
## ResidualSugar           ResidualSugar -0.053122905
## CitricAcid                 CitricAcid -0.050307040
## pH                                 pH  0.044288014
## Alcohol                       Alcohol -0.030715836
## Chlorides                   Chlorides  0.030427175
## FixedAcidity             FixedAcidity -0.022585961
## VolatileAcidity       VolatileAcidity  0.020379965
## Density                       Density -0.018693764
## TotalSulfurDioxide TotalSulfurDioxide -0.007179351
## FreeSulfurDioxide   FreeSulfurDioxide  0.006393010
```

Visualizing Relationships Among Variables

Correlation plots help us understand variable relationships and potential multicollinearity. A correlation plot for all variables shows moderate correlation between our dependent variable TARGET and the variables STARS and LabelAppeal and weak correlation between TARGET and AcidIndex.



STARS, LabelAppeal, and AcidIndex are also three parameters that we identified to be ordinal. The following correlation plot shows the correlation between the remaining numerical parameters. This view also groups the parameters into four clusters that appear to have a relationship with each other.



A Variance Inflation Factor (VIF) test confirms that no major multicollinearity present in between our variables.

```
##                        GVIF Df GVIF^(1/(2*Df))
## AcidIndex          1.100734 13        1.003698
## Alcohol            1.013464  1        1.006709
## Chlorides          1.007238  1        1.003612
## CitricAcid         1.008134  1        1.004059
## Density            1.009057  1        1.004518
## FixedAcidity       1.031541  1        1.015648
## FreeSulfurDioxide  1.006411  1        1.003200
## LabelAppeal        1.146432  4        1.017229
## pH                 1.007787  1        1.003886
## ResidualSugar      1.005833  1        1.002912
## STARS              1.162161  3        1.025363
## Sulphates          1.007464  1        1.003725
## TotalSulfurDioxide 1.007607  1        1.003796
## VolatileAcidity    1.005933  1        1.002962
```

Data Preparation

Handling Negative Values

In the Data Exploration phase, we discovered that nine out of 11 numerical variables had negative values. While Poisson and Negative Binomial Regression will allow negative predictor values, we know that these values are impossible in the read world. Thus, ignoring these values could lead to biased coefficient estimates that could introduce highly misleading relationships in our models. We will assume that these erronous values may be the result of data entry or normalization errors and attempt to address them.

As we do not know what transformations may have been applied if normalization occurred, we will need to address the negative values through another method. From our earlier observations, we noted that nearly all of the affected parameters had thousands of affected records, with Chlorides having the most affects rows (3197). This means that nearly 1/4 of our 12,795 may be affected one way or another and we would loose too much data if we were to drop the affected records. We will instead only drop the 118 records with for Alcohol Content as it is a low percentage of the dataset since some of these records may also have negative values; we will then set the remaining negative values to N/A, allowing the values to be imputed if desired. Imputing may introduce some bias into our results, but will retain much of our data.

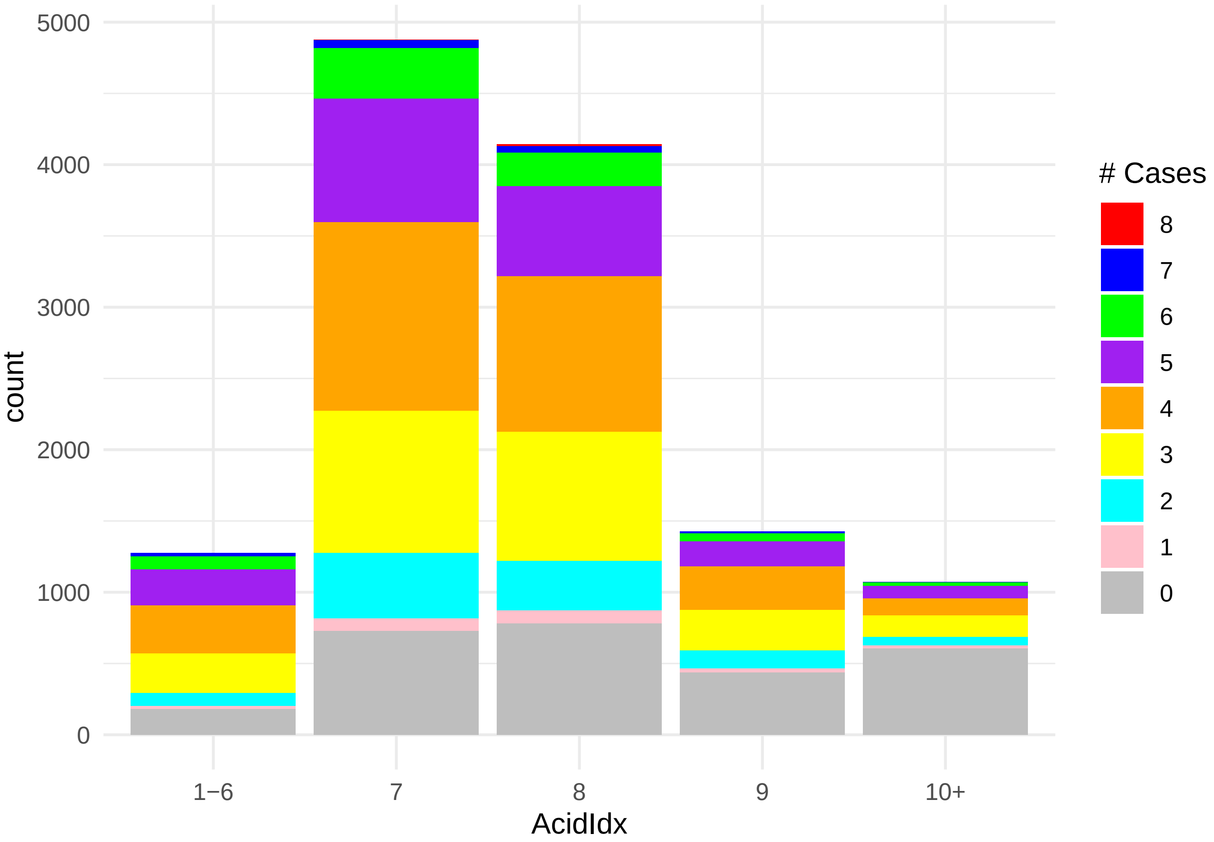
Handling Missing Values

Table 2: Number of missing values

	Original.Missing.Count	New.Missing.Count	After.Imputation
TARGET	0	0	0
AcidIndex	0	0	0
Alcohol	653	0	0
Chlorides	638	3603	0
CitricAcid	0	2772	2772
Density	0	0	0
FixedAcidity	0	1532	1532
FreeSulfurDioxide	647	3473	0
LabelAppeal	0	0	0
pH	395	368	0
ResidualSugar	616	3515	0
STARS	3359	3149	0
Sulphates	1210	3358	0
TotalSulfurDioxide	682	2973	0
VolatileAcidity	0	2651	2651

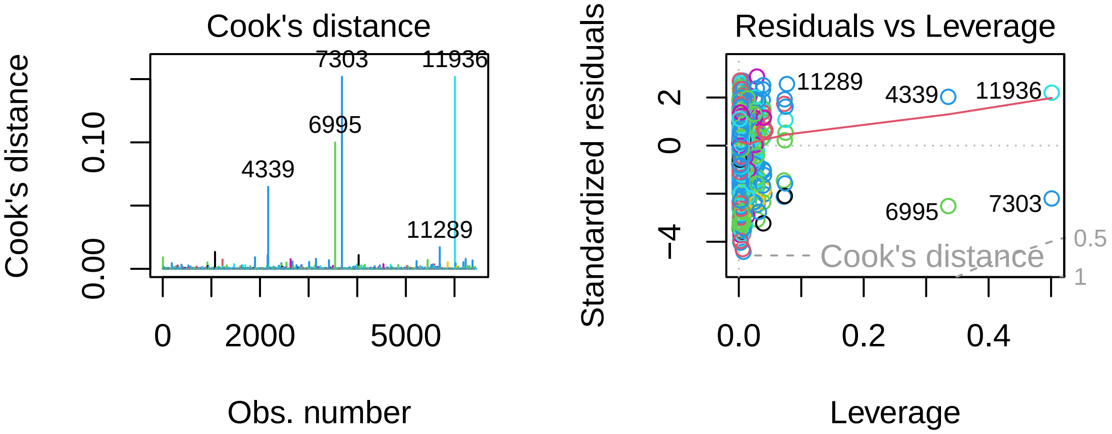
Binned Transformation

To simplify the effects of AcidIndex, this variable was transformed into categorical bins. This can help reduce the influence of extreme values and better capture non-linear effects in logistic regression.



Transformations

```
## Warning: not plotting observations with leverage one:
##   1351, 2469
```



Model Building

Model Selection