# 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 negative values.

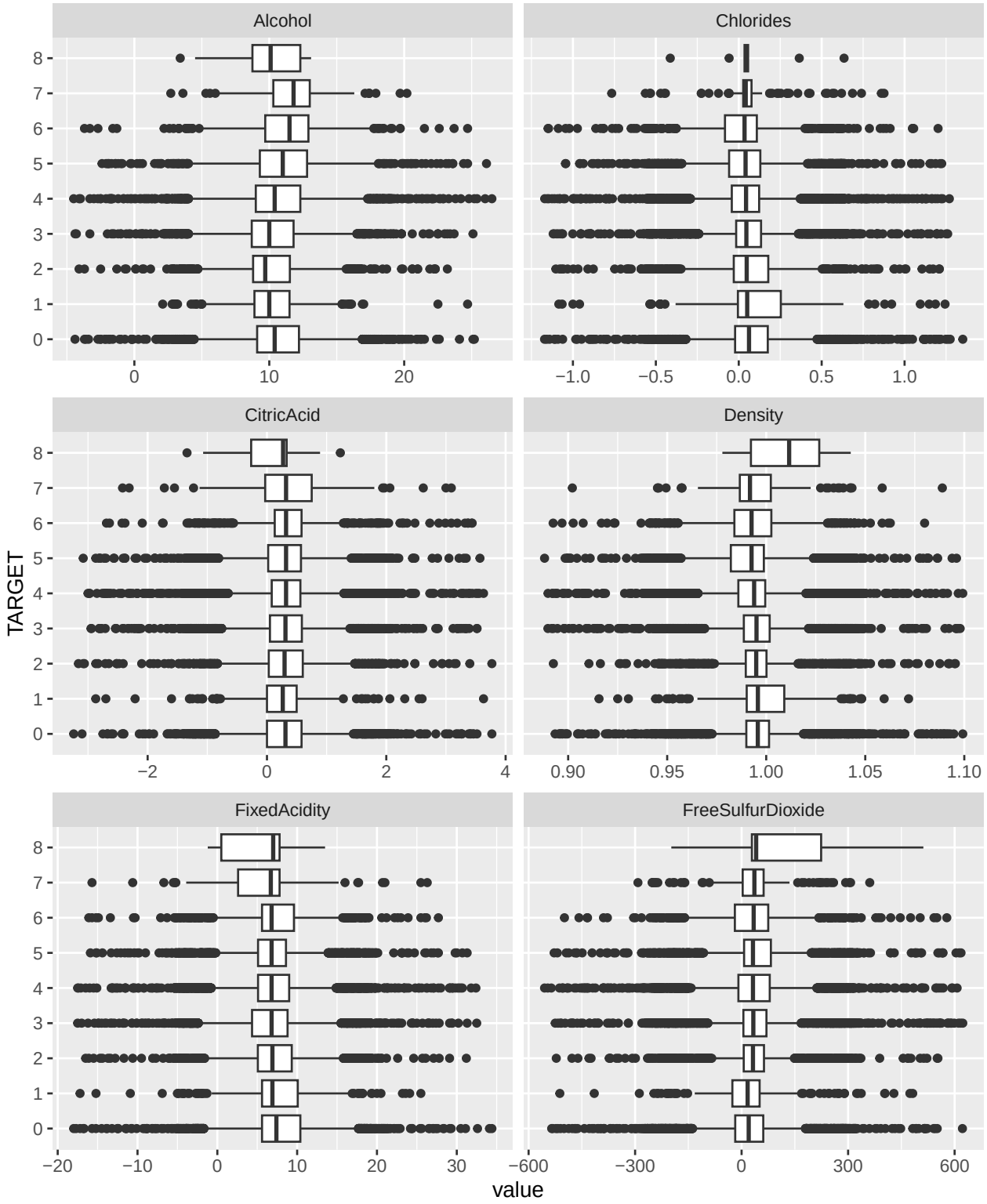
```
##      TARGET         Alcohol        Chlorides         CitricAcid
##  Min.   :0.000   Min.   :-4.70   Min.   :-1.1710   Min.   :-3.2400
##  1st Qu.:2.000   1st Qu.: 9.00   1st Qu.:-0.0310   1st Qu.: 0.0300
##  Median :3.000   Median :10.40   Median : 0.0460   Median : 0.3100
##  Mean   :3.029   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
```

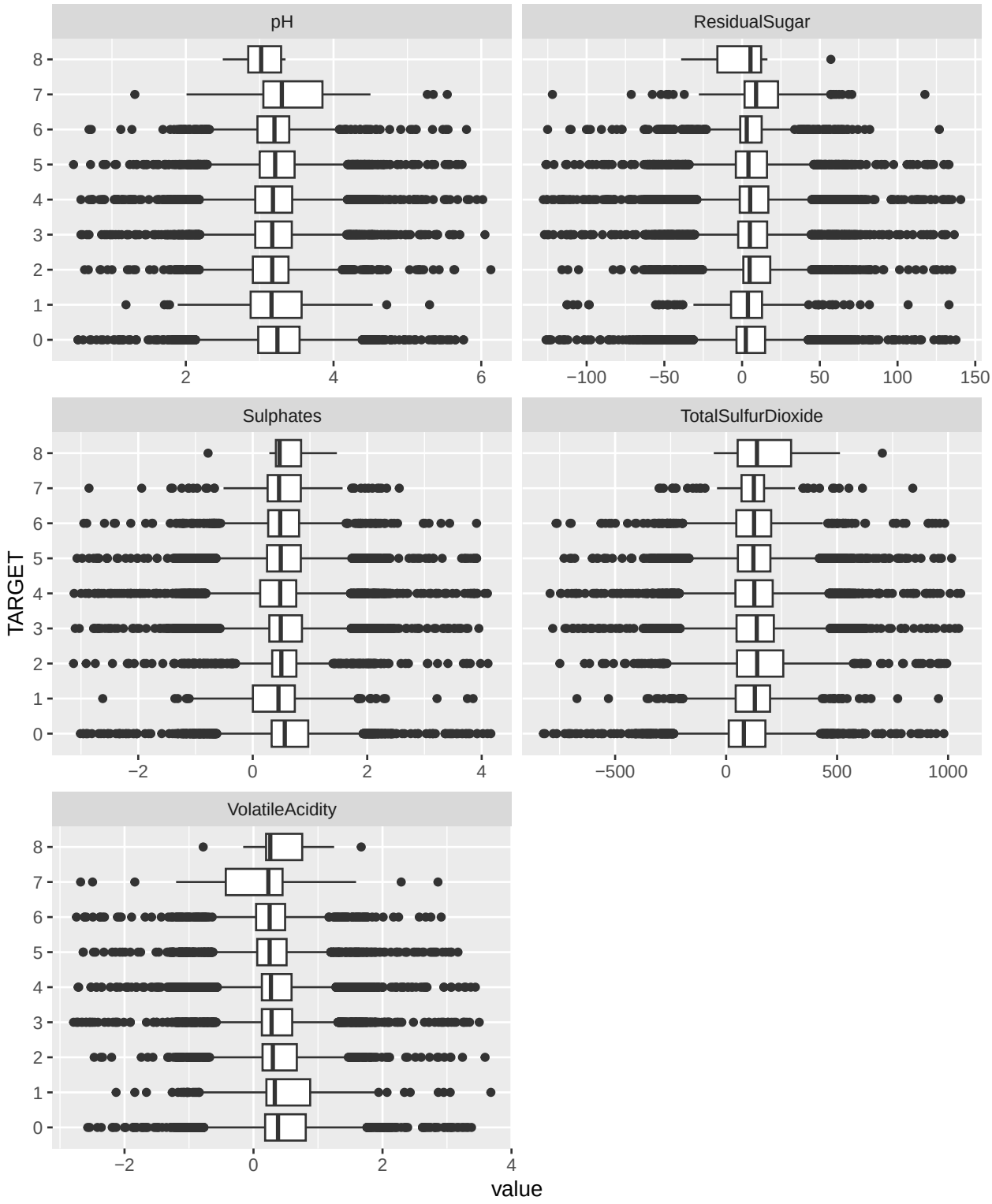
When put in the context of our specific properties of wine they represent, these negative values appear to be erroneous. For example, we expected `Alcohol Content` to have a minimum value of zero instead of a negative value. The same can be said of the other parameters with negative values (Chlorides, 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 | Q25 | Median | Q75 | CommonRange | Rows Below Zero |
|---|---|---|---|---|---|
| TARGET | 2.000 | 3.000 | 4.000 | 0 or higher | 0 |
| Alcohol | 9.000 | 10.400 | 12.400 | 8% – 15% ABV | 118 |
| Chlorides | -0.031 | 0.046 | 0.153 | 0.01 – 0.10 g/L | 3197 |
| CitricAcid | 0.030 | 0.310 | 0.580 | 0 – 1.0 g/L | 2966 |
| Density | 0.988 | 0.994 | 1.001 | 0.990 – 1.005 g/cm$^3$ | 0 |
| FixedAcidity | 5.200 | 6.900 | 9.500 | 4 – 9 g/L | 1621 |
| FreeSulfurDioxide | 0.000 | 30.000 | 70.000 | 10 – 70 mg/L | 3036 |
| pH | 2.960 | 3.200 | 3.470 | 2.9 – 4.0 | 0 |
| ResidualSugar | -2.000 | 3.900 | 15.900 | 0 – 45 g/L | 3136 |
| Sulphates | 0.280 | 0.500 | 0.860 | 0.3 – 1.0 g/L | 2361 |
| TotalSulfurDioxide | 27.000 | 123.000 | 208.000 | 30 – 150 mg/L | 2504 |
| VolatileAcidity | 0.130 | 0.280 | 0.640 | 0.2 – 0.8 g/L | 2827 |

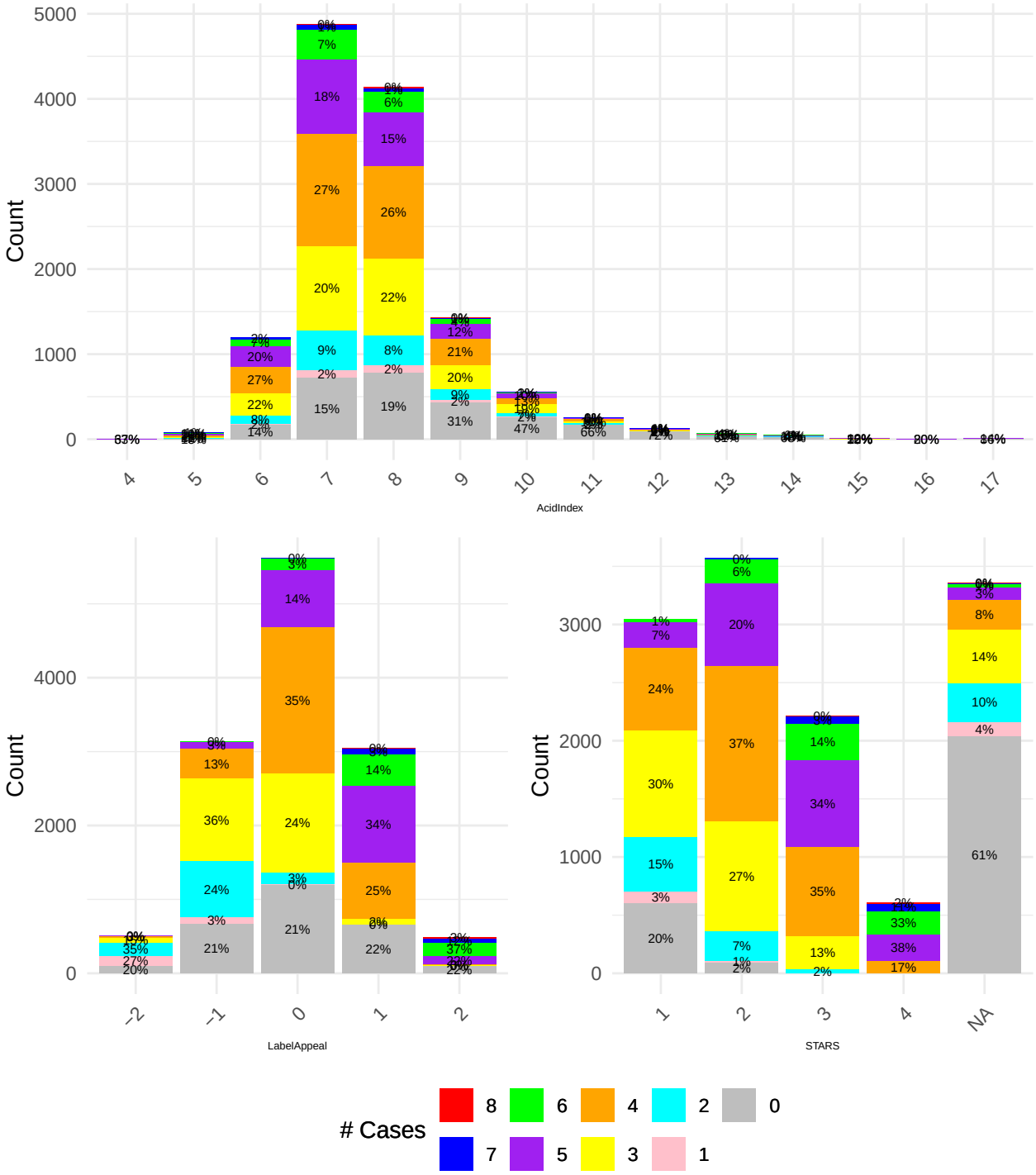
A look at our boxplots shows the IQR's for each parameter are centered around a similar x-axis for each of our case counts. The boxplots confirm the presence of extreme values at lower as well as on the upper ranges. It should be noted the IQRs for the affected variables are in line with their corresponding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


Boxplots of Target vs Param
Alcohol
Chlorides
CitricAcid
Density
FixedAcidity
FreeSulfurDioxide
TARGET
value
0
1
2
3
4
5
6
7
8
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


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

### Examining Categorical Variables

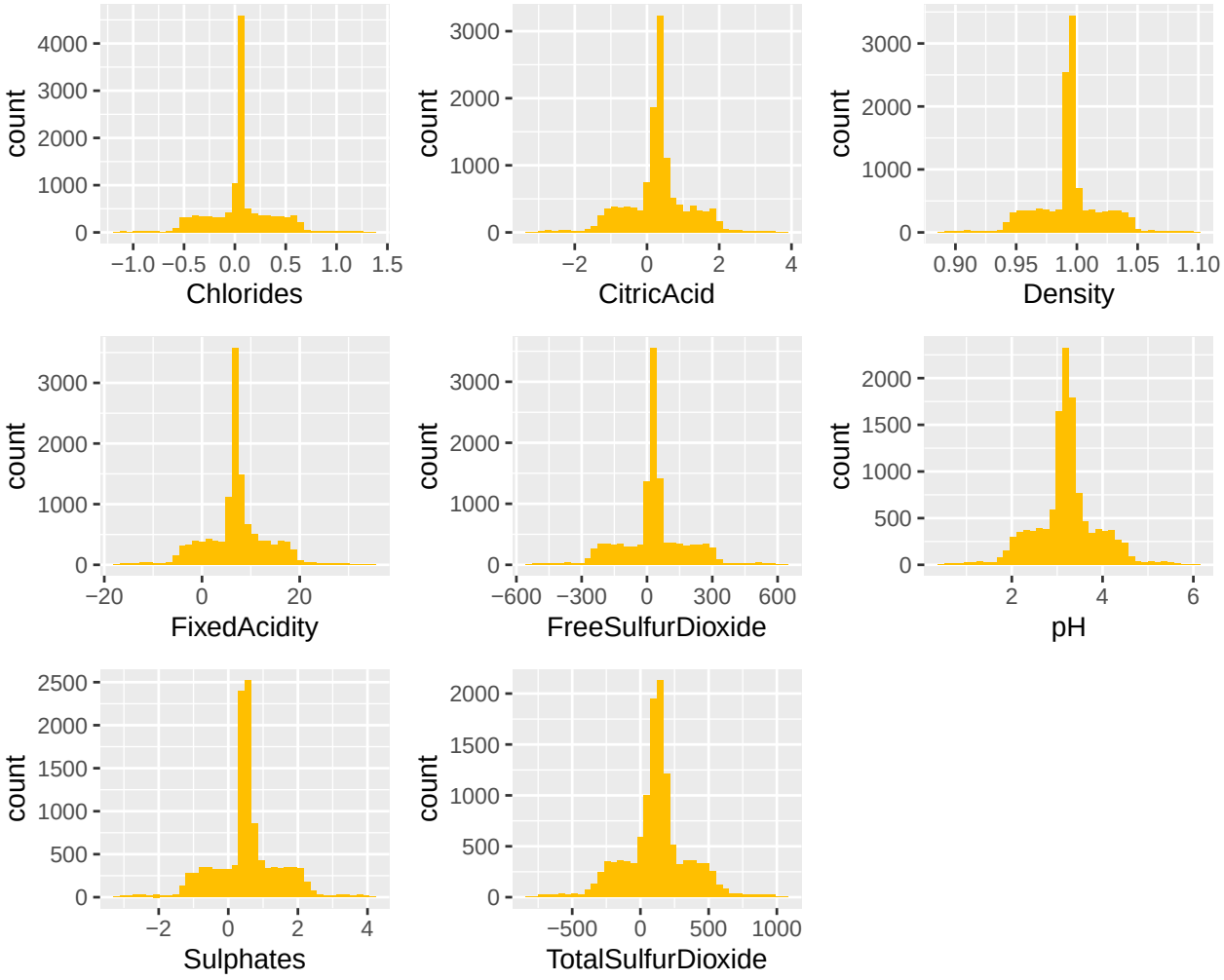
Visualizing the distributions of out categorical variables helps ensure variables are treated as discrete categories, not continuous numbers. Our AcidIndex variable shows that majority of values fall between 6 and 11 with 342 observations collectively making up the remaining values. We may want to bin values for this parameter. Label Appeal has a normal distribution ranging from -2 to 2 and centered around 0. The 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. However, as noted earlier, negative values for many of these parameters would be impossible.



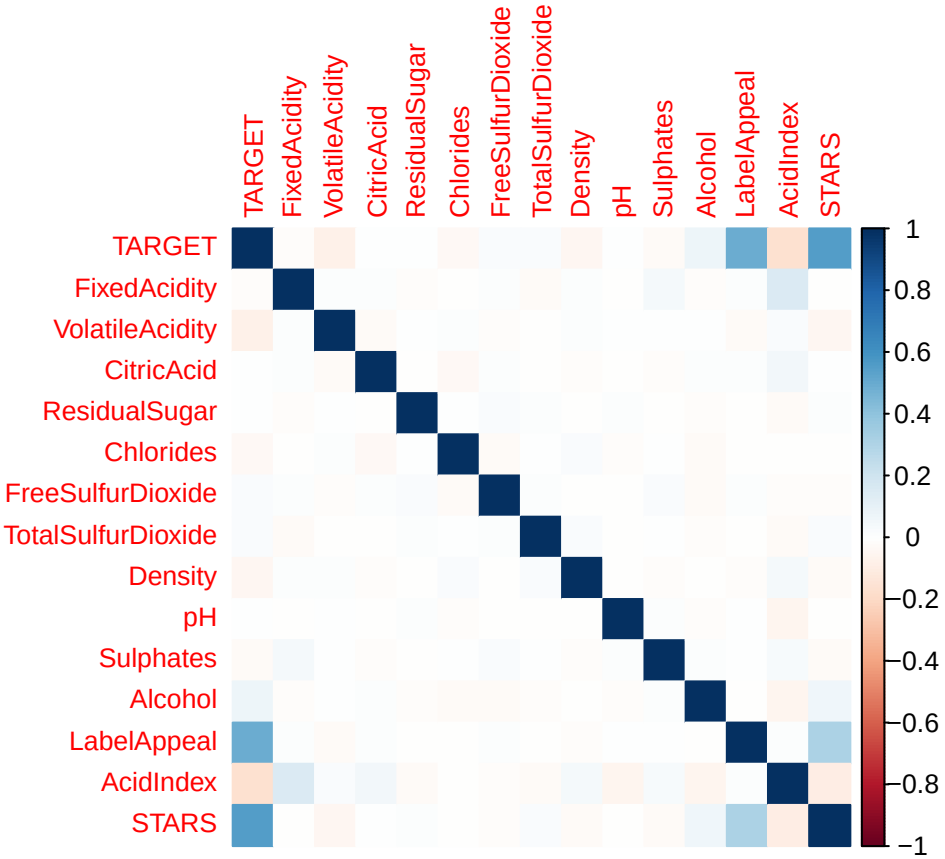
Our skewness test confirms that our numerical variables are nearly symmetrical or almost symmetrical.

Table 2: Skewness

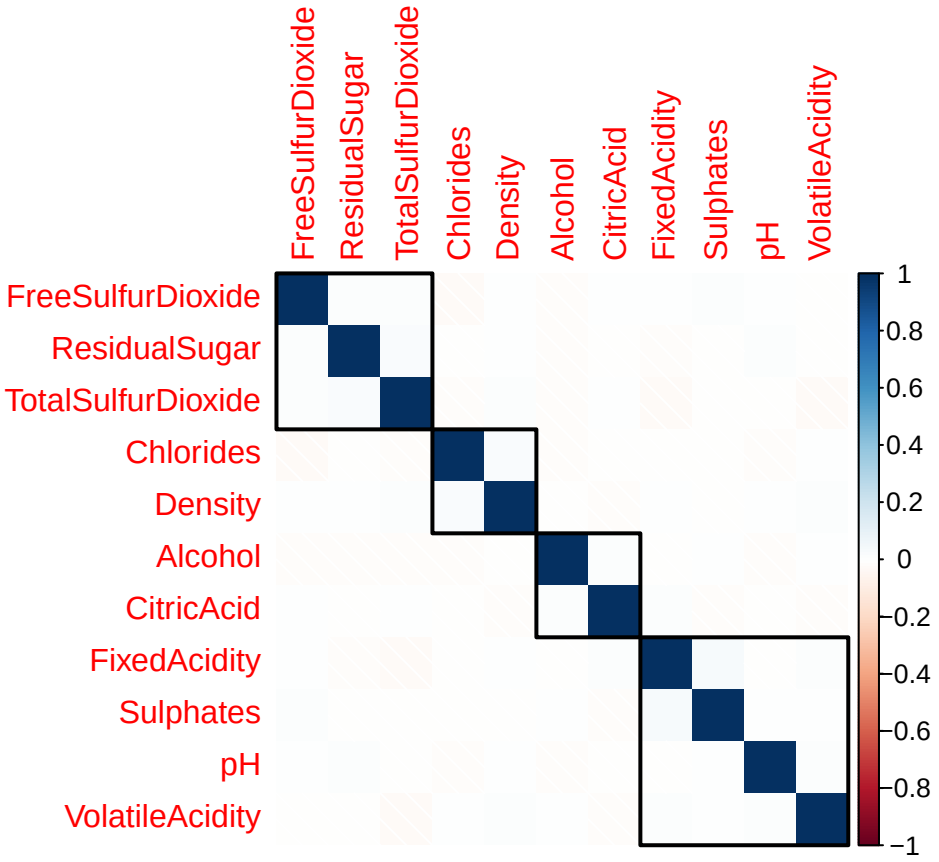
| | Variable | Skewness |
|---|---|---|
| ResidualSugar | ResidualSugar | -0.0531229 |
| CitricAcid | CitricAcid | -0.0503070 |
| pH | pH | 0.0442880 |
| Alcohol | Alcohol | -0.0307158 |
| Chlorides | Chlorides | 0.0304272 |
| FixedAcidity | FixedAcidity | -0.0225860 |
| VolatileAcidity | VolatileAcidity | 0.0203800 |
| Density | Density | -0.0186938 |
| TotalSulfurDioxide | TotalSulfurDioxide | -0.0071794 |
| FreeSulfurDioxide | FreeSulfurDioxide | 0.0063930 |
| Sulphates | Sulphates | 0.0059119 |

### 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 only the 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. 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 small fraction of the total rows in our dataset. 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; dropping these 118 rows may somewhat help reduce this bias.

### Handling Missing Values

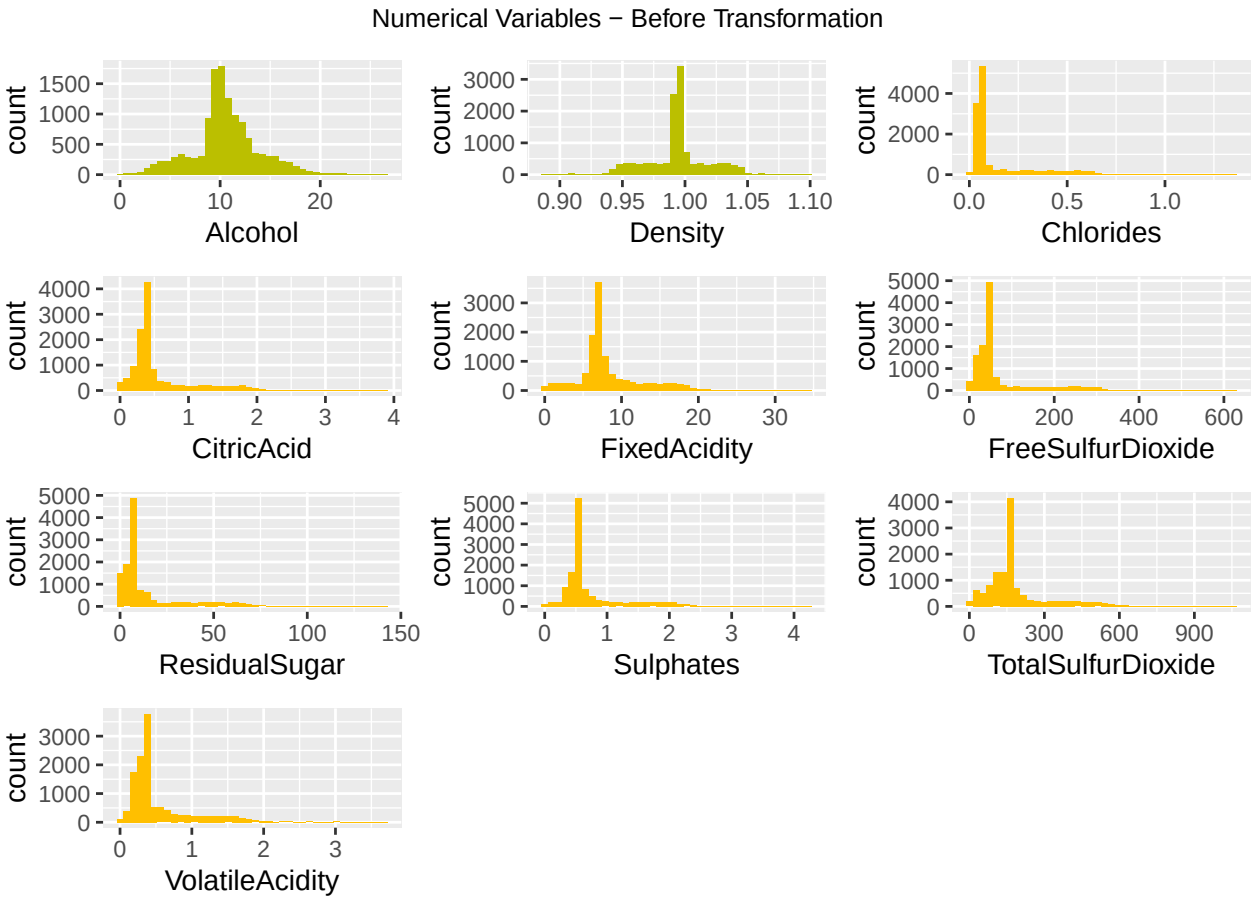
Given that we don't know if our data was previously transformed, we will apply median imputation to our variables with missing data. This method was preferred over multiple imputation or other techniques as it is easier to understand how we have altered the data and simpler to . Additionally, Multiple-imputation may magnify additional bias that may have been introduced if the data was previously transformed. We won't impute values for STARS since this appears to be an indication that a rating was not given to these observations. When exploring our data, we noticed that 61% of the wines without a STAR rating had zero (0) cases purchased. As such, a NA value may represent useful data for our model. We will convert this variable to -1 for interpretability by our model.

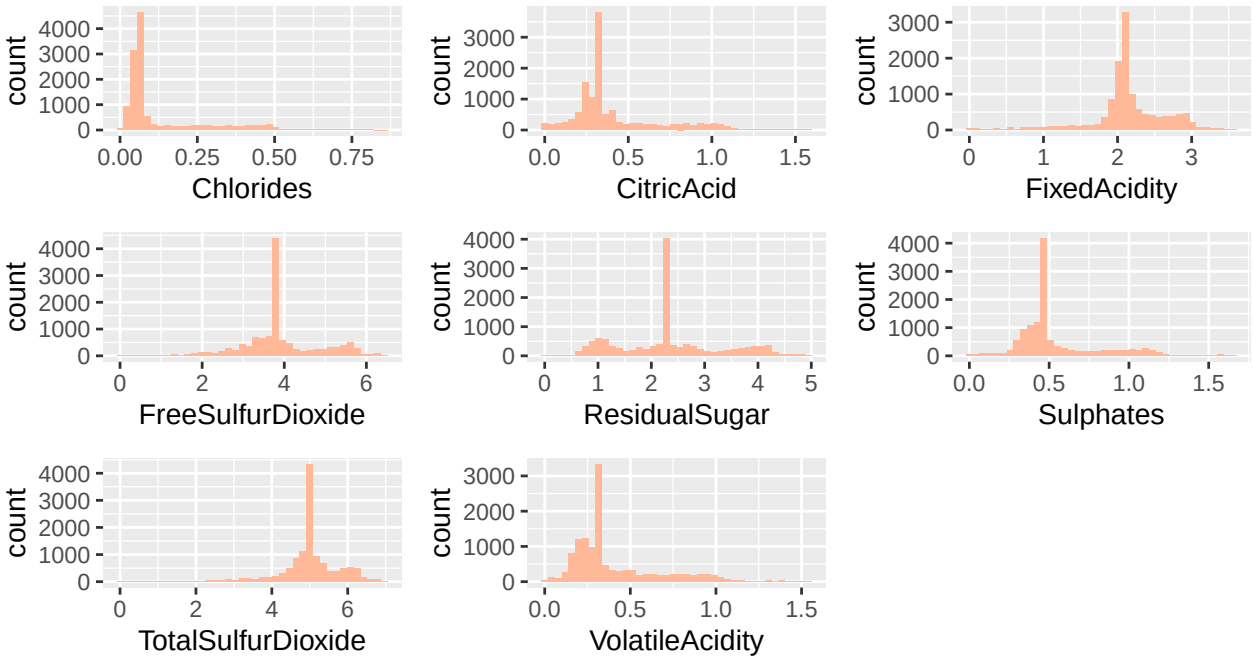
Table 3: Number of missing values

| | Original.Missing.Count | New.Missing.Count | After.Imputation |
|---|---|---|---|
| TARGET | 0 | 0 | 0 |
| AcidIndex | 0 | 0 | 0 |
| Alcohol | 653 | 653 | 0 |
| Chlorides | 638 | 3802 | 0 |
| CitricAcid | 0 | 2933 | 0 |
| Density | 0 | 0 | 0 |
| FixedAcidity | 0 | 1608 | 0 |
| FreeSulfurDioxide | 647 | 3657 | 0 |
| LabelAppeal | 0 | 0 | 0 |
| pH | 395 | 392 | 0 |
| ResidualSugar | 616 | 3714 | 0 |
| STARS | 3359 | 3329 | 0 |
| Sulphates | 1210 | 3534 | 0 |
| TotalSulfurDioxide | 682 | 3151 | 0 |
| VolatileAcidity | 0 | 2795 | 0 |

### Transformations

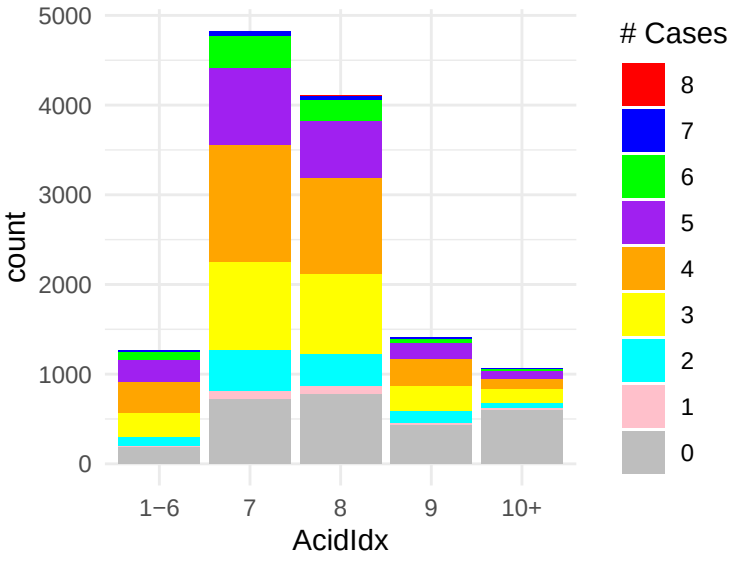
Dropping our negative values and imputing our missing values introduced skewness into most of the numerical variables. All variables but Alcohol, Density and pH have are right—skewed –though the right tail for FixedAcidity and TotalSulfurDioxide are relatively moderate compared the heavy right tails of the other variables. We can apply log transformation to these variables to help address skewness.


Numerical Variables − Before Transformation
count
Alcohol
Density
Chlorides
CitricAcid
FixedAcidity
FreeSulfurDioxide
ResidualSugar
Sulphates
TotalSulfurDioxide
VolatileAcidity


Numerical Variables − Post Transformation
count
Chlorides
CitricAcid
FixedAcidity
FreeSulfurDioxide
ResidualSugar
Sulphates
TotalSulfurDioxide
VolatileAcidity

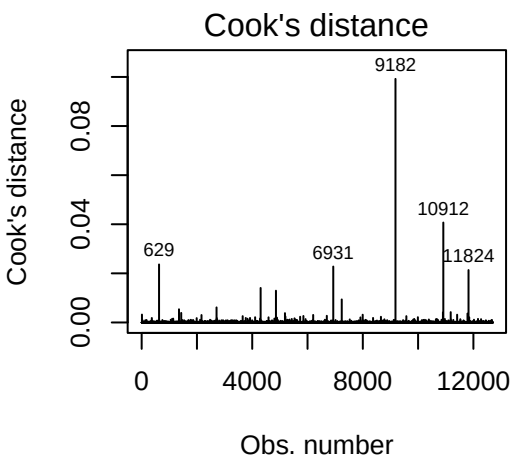
### Binned Transformation

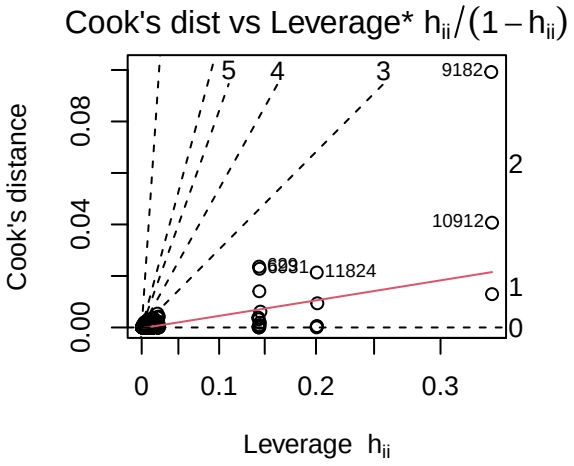
To simplify the effects of AcidIndex, this variable was transformed into categorical bins. This can help reduce the influence of extreme values and better capture non-linear effects in logistic regression.




### Outliers

Diagnostic plots show several points with relatively high Cook's distance values when compared to rest of our data.




## Model Building

## Model Selection