Assignment5\_Data621\_MQ

Mubashira Qari

2025-04-13

## Data Exploration

The training dataset has 12,795 observations across 16 columns, including our INDEX and TARGET columns.

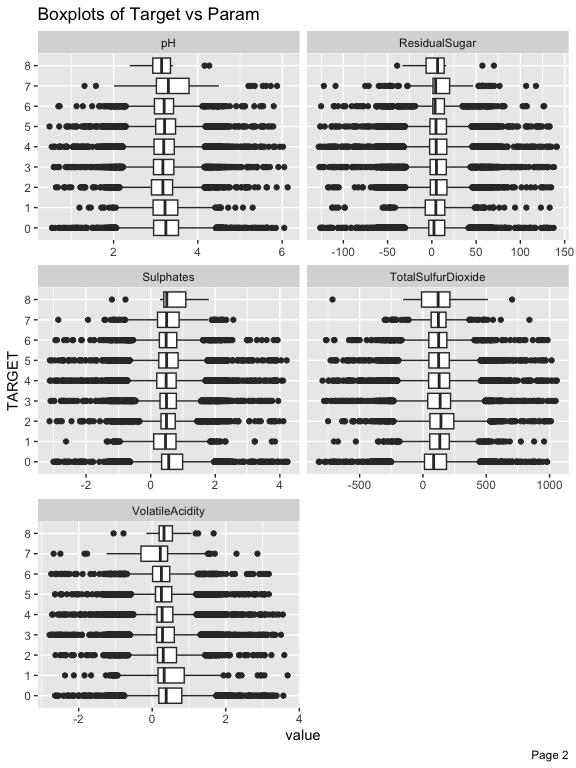
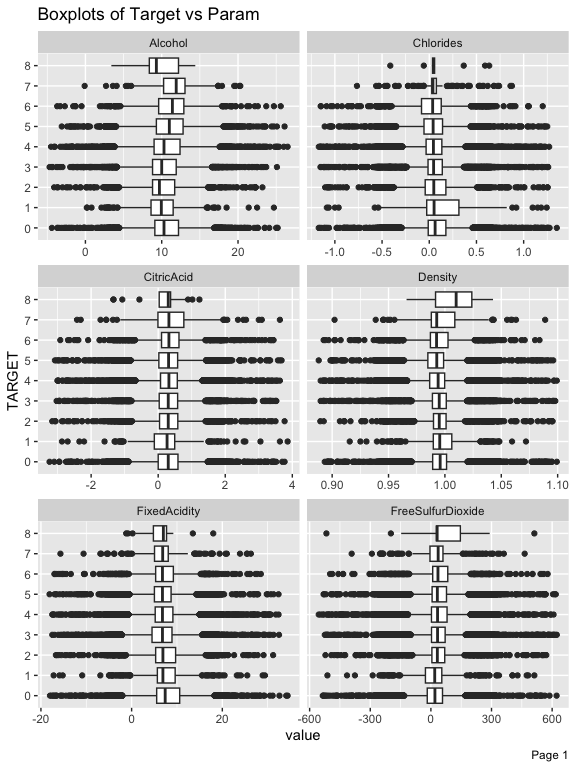
## Rows: 12,795  
## Columns: 15  
## $ TARGET <dbl> 3, 3, 5, 3, 4, 0, 0, 4, 3, 6, 0, 4, 3, 7, 4, 0, 0, …  
## $ AcidIndex <fct> 8, 7, 8, 6, 9, 11, 8, 7, 6, 8, 5, 10, 7, 8, 9, 8, 9…  
## $ Alcohol <dbl> 9.9, NA, 22.0, 6.2, 13.7, 15.4, 10.3, 11.6, 15.0, 1…  
## $ Chlorides <dbl> -0.567, -0.425, 0.037, -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…

## # A tibble: 1 × 0

### Missing Values

Eight parameters had missing values.

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



## Warning: Removed 653 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

## Warning: Removed 638 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

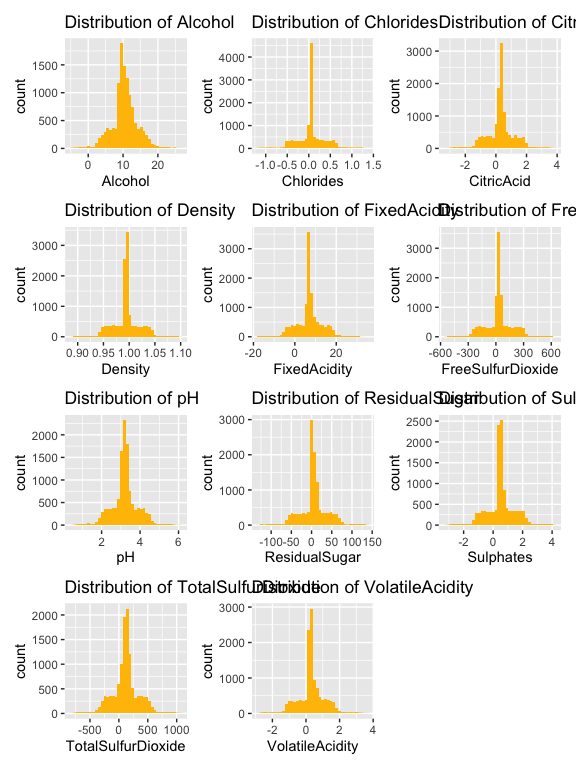
## Warning: Removed 647 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

## Warning: Removed 395 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

## Warning: Removed 616 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

## Warning: Removed 1210 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

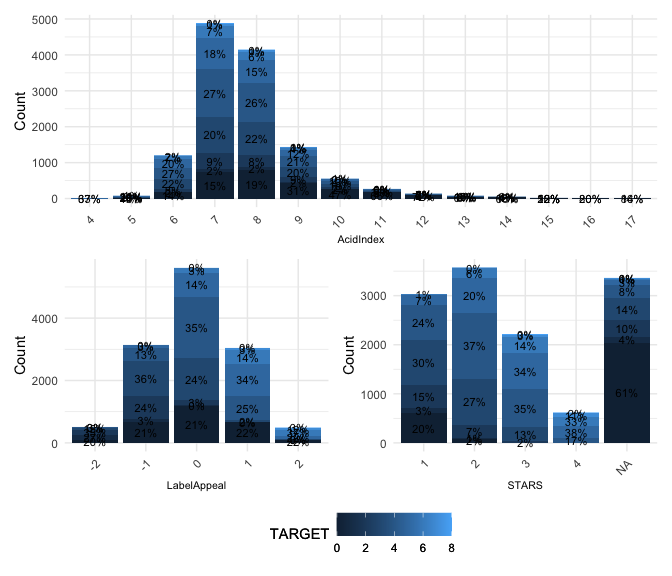
## Warning: Removed 682 rows containing non-finite outside the scale range  
## (`stat\_bin()`).



### Visualizing Distributions for Categorical Variables

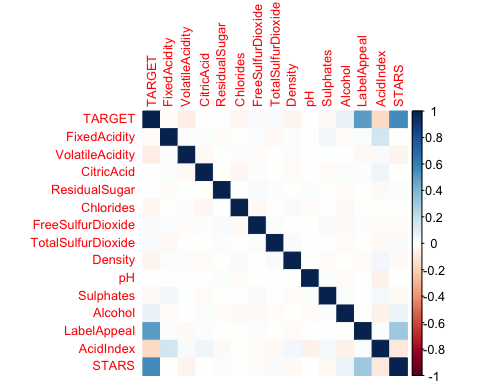
Visualizing the distributions of out categorical variables helps ensure variables are treated as discrete categories, not continuous numbers.

##   
## Counts for AcidIndex :  
##   
## 4 5 6 7 8 9 10 11 12 13 14 15 16 17   
## 3 75 1197 4878 4142 1427 551 258 128 69 47 8 5 7   
## # A tibble: 88 × 5  
## # Groups: group\_var [14]  
## TARGET group\_var count percent label  
## <dbl> <fct> <int> <dbl> <chr>  
## 1 0 4 1 33.3 33%   
## 2 0 5 11 14.7 15%   
## 3 0 6 170 14.2 14%   
## 4 0 7 727 14.9 15%   
## 5 0 8 782 18.9 19%   
## 6 0 9 437 30.6 31%   
## 7 0 10 257 46.6 47%   
## 8 0 11 169 65.5 66%   
## 9 0 12 92 71.9 72%   
## 10 0 13 42 60.9 61%   
## # ℹ 78 more rows  
##   
## Counts for LabelAppeal :  
##   
## -2 -1 0 1 2   
## 504 3136 5617 3048 490   
## # A tibble: 36 × 5  
## # Groups: group\_var [5]  
## TARGET group\_var count percent label  
## <dbl> <fct> <int> <dbl> <chr>  
## 1 0 -2 102 20.2 20%   
## 2 0 -1 671 21.4 21%   
## 3 0 0 1193 21.2 21%   
## 4 0 1 660 21.7 22%   
## 5 0 2 108 22.0 22%   
## 6 1 -2 136 27.0 27%   
## 7 1 -1 89 2.84 3%   
## 8 1 0 19 0.338 0%   
## 9 2 -2 177 35.1 35%   
## 10 2 -1 755 24.1 24%   
## # ℹ 26 more rows  
##   
## Counts for STARS :  
##   
## 1 2 3 4   
## 3042 3570 2212 612   
## # A tibble: 36 × 5  
## # Groups: group\_var [5]  
## TARGET group\_var count percent label  
## <dbl> <fct> <int> <dbl> <chr>  
## 1 0 1 607 20.0 20%   
## 2 0 2 89 2.49 2%   
## 3 0 <NA> 2038 60.7 61%   
## 4 1 1 98 3.22 3%   
## 5 1 2 20 0.560 1%   
## 6 1 <NA> 126 3.75 4%   
## 7 2 1 469 15.4 15%   
## 8 2 2 253 7.09 7%   
## 9 2 3 34 1.54 2%   
## 10 2 <NA> 335 9.97 10%   
## # ℹ 26 more rows

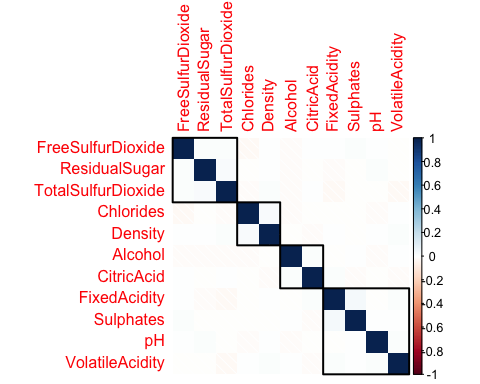


### Visualizing Relationships among Variables:

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 clusters the parameters into groups that appear to have a relationship with each other.



wine\_features <- wine\_training\_df  
  
# Fit linear model using all remaining features to predict TARGET  
vif\_model <- lm(TARGET ~ ., data = wine\_features)  
  
# Compute VIF values  
vif\_values <- car::vif(vif\_model)  
  
# View all VIF values  
print(vif\_values)

## 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

# Optional: Filter features with VIF > 5 (multicollinearity concern)  
high\_vif <- vif\_values[vif\_values > 5]  
print(high\_vif)

## [1] 13

# Column positions of high VIF variables in wine\_features  
which(names(vif\_values) %in% names(high\_vif))

## integer(0)

All very close to 1, ranging from ~1.00 to ~1.10

No variable has a VIF > 5

Final filtered output named numeric(0) and integer(0) means:

No variables exceeded the multicollinearity threshold you set (probably VIF > 5)

### Detect Skewed Variables

# 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))  
  
# 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

AcidIndex 1.65 Highly right-skewed - consider log() or sqrt() STARS 0.56 Moderately right-skewed - Possibly bin or treat as ordinal/factor

# Load libraries  
library(ggplot2)  
library(patchwork)  
  
# Example: If your top skewed vars aren't defined, use dummy fallback  
# Replace with actual top skewed variables as needed  
top\_skewed\_vars <- c("ResidualSugar", "TotalSulfurDioxide", "Chlorides")  
  
# Create 3 ggplot histograms  
p1 <- ggplot(wine\_training\_df, aes\_string(x = top\_skewed\_vars[1])) +  
 geom\_histogram(bins = 30, fill = "#2c7fb8", color = "white") +  
 labs(title = paste("Histogram of", top\_skewed\_vars[1])) +  
 theme\_minimal()

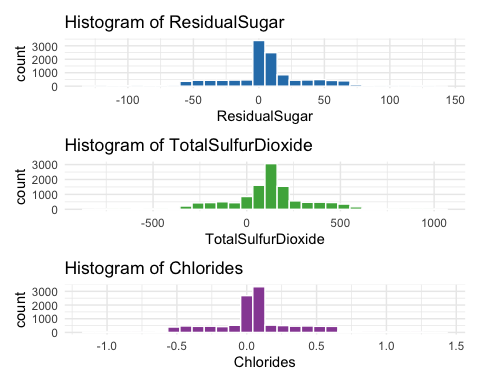
## Warning: `aes\_string()` was deprecated in ggplot2 3.0.0.  
## ℹ Please use tidy evaluation idioms with `aes()`.  
## ℹ See also `vignette("ggplot2-in-packages")` for more information.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.

p2 <- ggplot(wine\_training\_df, aes\_string(x = top\_skewed\_vars[2])) +  
 geom\_histogram(bins = 30, fill = "#4daf4a", color = "white") +  
 labs(title = paste("Histogram of", top\_skewed\_vars[2])) +  
 theme\_minimal()  
  
p3 <- ggplot(wine\_training\_df, aes\_string(x = top\_skewed\_vars[3])) +  
 geom\_histogram(bins = 30, fill = "#984ea3", color = "white") +  
 labs(title = paste("Histogram of", top\_skewed\_vars[3])) +  
 theme\_minimal()  
  
# Display all three plots in one view  
(p1 / p2 / p3) # stacked vertically

## Warning: Removed 616 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

## Warning: Removed 682 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

## Warning: Removed 638 rows containing non-finite outside the scale range  
## (`stat\_bin()`).



# Or side-by-side with: (p1 | p2 | p3)

## Data Preparation

### Handling Missing Values:

# Identify numeric columns with missing values  
numeric\_cols <- sapply(wine\_training\_df, is.numeric)  
  
  
   
  
# Impute missing values in those columns using median  
for (col in names(wine\_training\_df[missing > 0])) {  
 print(col)  
 #median\_val <- median(wine\_training\_df[[col]], na.rm = TRUE)  
 #wine\_training\_df[[col]][is.na(wine\_training\_df[[col]])] <- median\_val  
}

## [1] "Alcohol"  
## [1] "Chlorides"  
## [1] "FreeSulfurDioxide"  
## [1] "pH"  
## [1] "ResidualSugar"  
## [1] "STARS"  
## [1] "Sulphates"  
## [1] "TotalSulfurDioxide"

colSums(is.na(wine\_training\_df))

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

### Transformations

### Log Transform on AcidIndex

log1p(x) is equivalent to log(x + 1) and works better with small or zero values.

This helps reduce right skew and stabilize the variance of AcidIndex.

### To visualize the improvement: