Introduction

This report presents a statistical analysis of banana data. The goal is to explore the relationships between various characteristics of bananas and their quality.

Data Loading

```
# Load the banana data
banana_data <- read_csv("banana.csv")
```

```
## Rows: 8000 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Quality
## dbl (7): Size, Weight, Sweetness, Softness, HarvestTime, Ripeness, Acidity
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(banana_data)
```

```
## # A tibble: 6 x 8
##     Size Weight Sweetness Softness HarvestTime Ripeness Acidity Quality
##    <dbl>  <dbl>     <dbl>    <dbl>       <dbl>    <dbl>   <dbl> <chr>
## 1 -1.92   0.468    3.08      -1.47       0.295     2.44   0.271 Good
## 2 -2.41   0.487    0.347     -2.50      -0.892     2.07   0.307 Good
## 3 -0.358  1.48     1.57      -2.65      -0.647     3.09   1.43  Good
## 4 -0.869  1.57     1.89      -1.27      -1.01      1.87   0.478 Good
## 5  0.652  1.32    -0.0225    -1.21      -1.43      1.08   2.81  Good
## 6 -2.81   1.14     3.45      -1.71      -2.22      2.08   2.28  Good
```

Exploratory Data Analysis (EDA)

Summary Statistics

Let's begin with some summary statistics for each variable.

```
summary(banana_data)
```

```
##       Size              Weight           Sweetness          Softness
##  Min.   :-7.9981   Min.   :-8.2830   Min.   :-6.4340   Min.   :-6.95932
##  1st Qu.:-2.2777   1st Qu.:-2.2236   1st Qu.:-2.1073   1st Qu.:-1.59046
##  Median :-0.8975   Median :-0.8687   Median :-1.0207   Median : 0.20264
##  Mean   :-0.7478   Mean   :-0.7610   Mean   :-0.7702   Mean   :-0.01444
##  3rd Qu.: 0.6542   3rd Qu.: 0.7755   3rd Qu.: 0.3110   3rd Qu.: 1.54712
##  Max.   : 7.9708   Max.   : 5.6797   Max.   : 7.5394   Max.   : 8.24155
##   HarvestTime         Ripeness          Acidity            Quality
##  Min.   :-7.5700   Min.   :-7.4232   Min.   :-8.226977   Length:8000
##  1st Qu.:-2.1207   1st Qu.:-0.5742   1st Qu.:-1.629450   Class :character
```

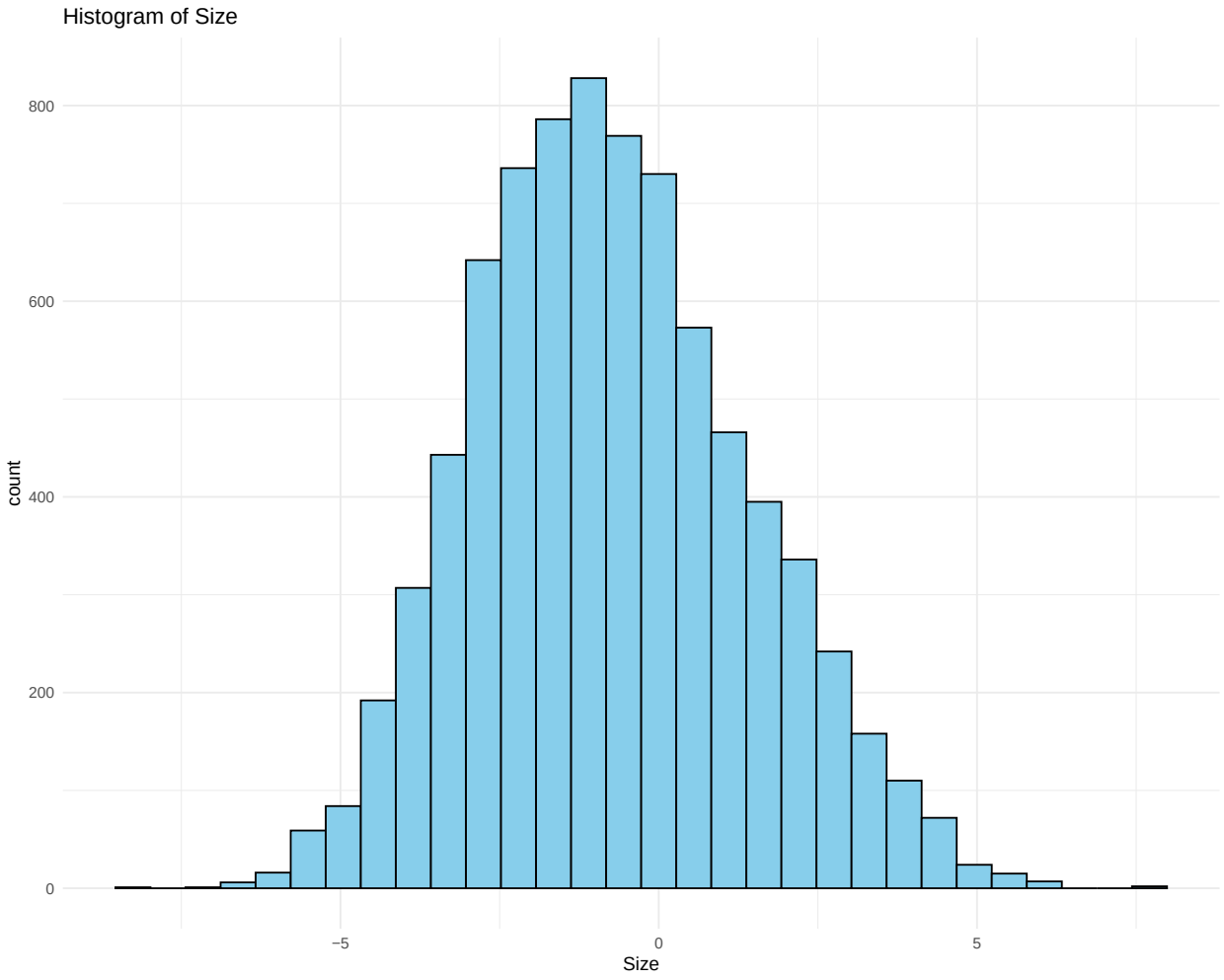
```
##  Median :-0.9342   Median : 0.9650   Median : 0.098735   Mode  :character
##  Mean   :-0.7513   Mean   : 0.7811   Mean   : 0.008725
##  3rd Qu.: 0.5073   3rd Qu.: 2.2616   3rd Qu.: 1.682063
##  Max.   : 6.2933   Max.   : 7.2490   Max.   : 7.411633
```

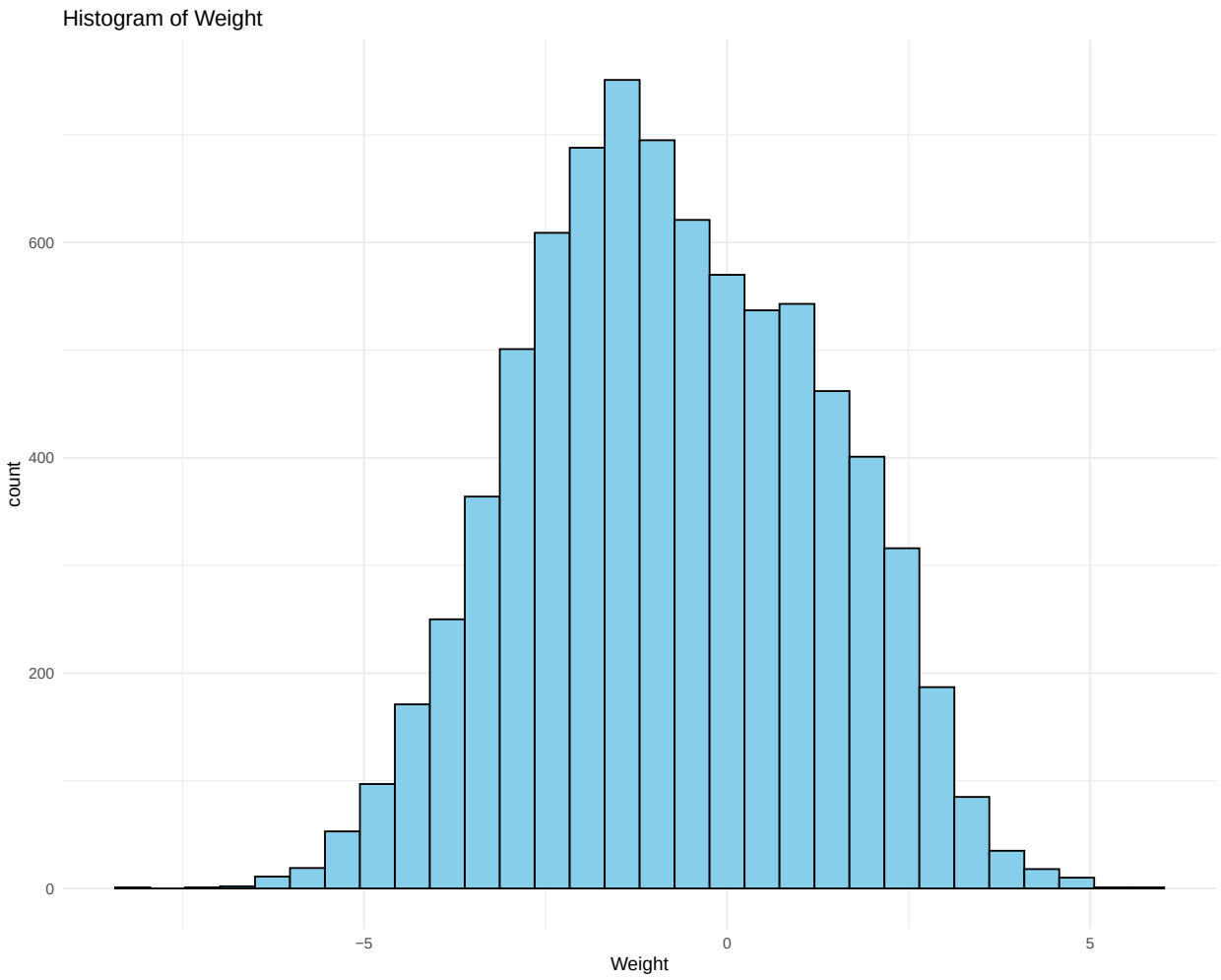
Visualizing Distributions with Histograms

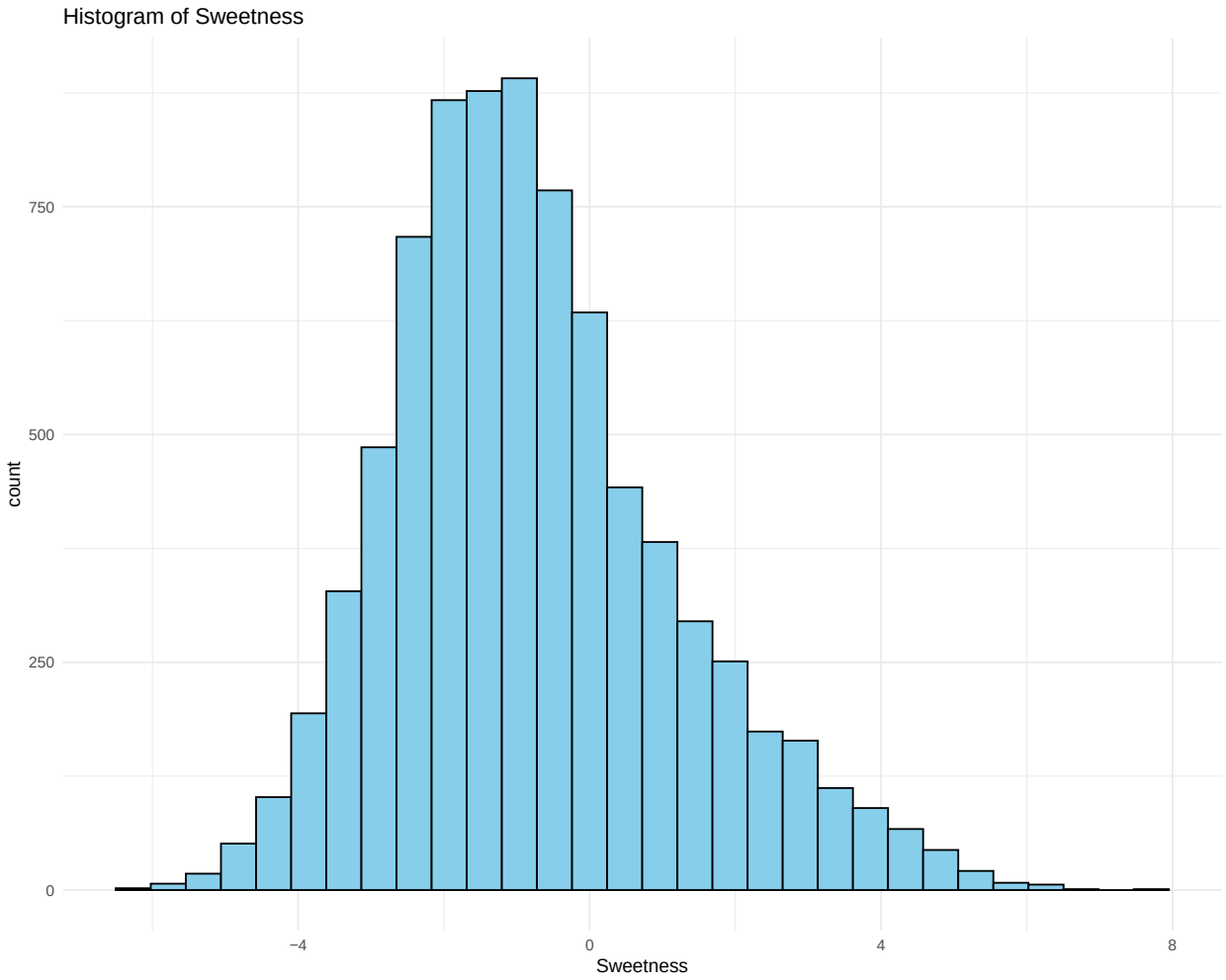
Histograms provide a quick way to visualize the distribution of each numerical variable.

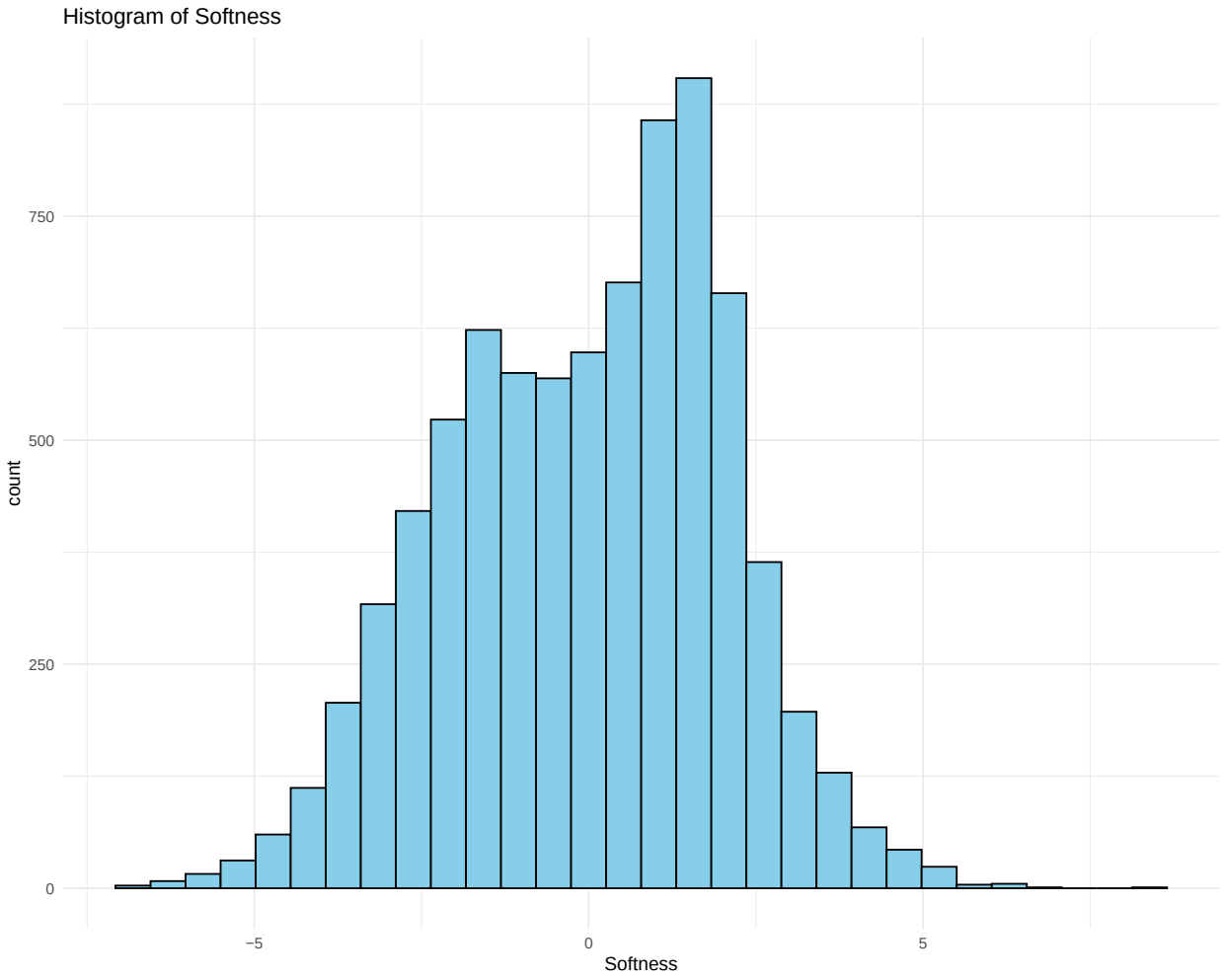
```r
# Plotting histograms for each numerical variable
num_vars <- c("Size", "Weight", "Sweetness", "Softness", "Ripeness", "Acidity", "HarvestTime")
for (var in num_vars) {
  print(ggplot(banana_data, aes_string(x = var)) +
          geom_histogram(bins = 30, fill = "skyblue", color = "black") +
          theme_minimal() +
          ggtitle(paste("Histogram of", var)))
}
```

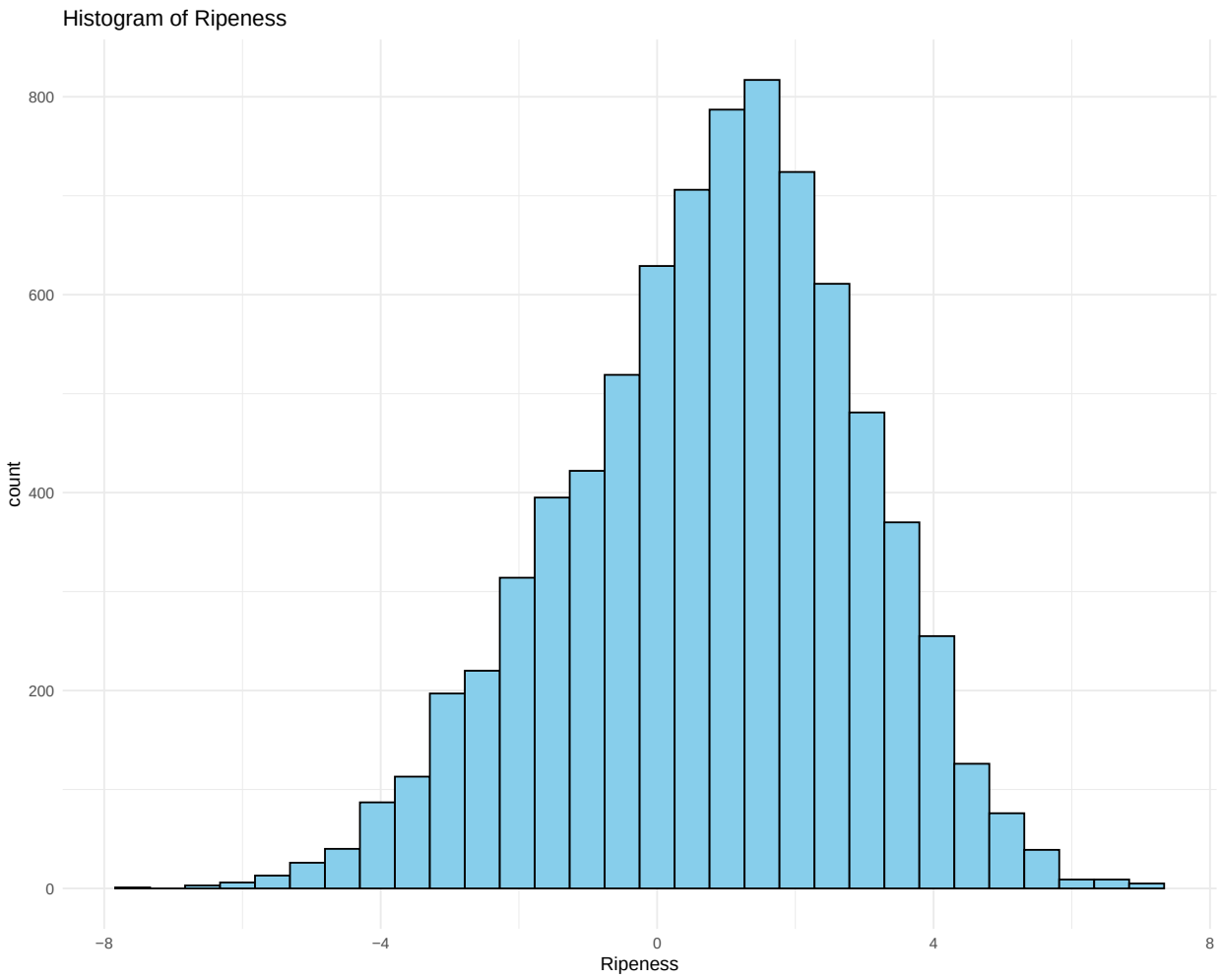
```
## Warning: ‘aes_string()‘ was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with ‘aes()‘.
## i See also ‘vignette("ggplot2-in-packages")‘ for more information.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this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

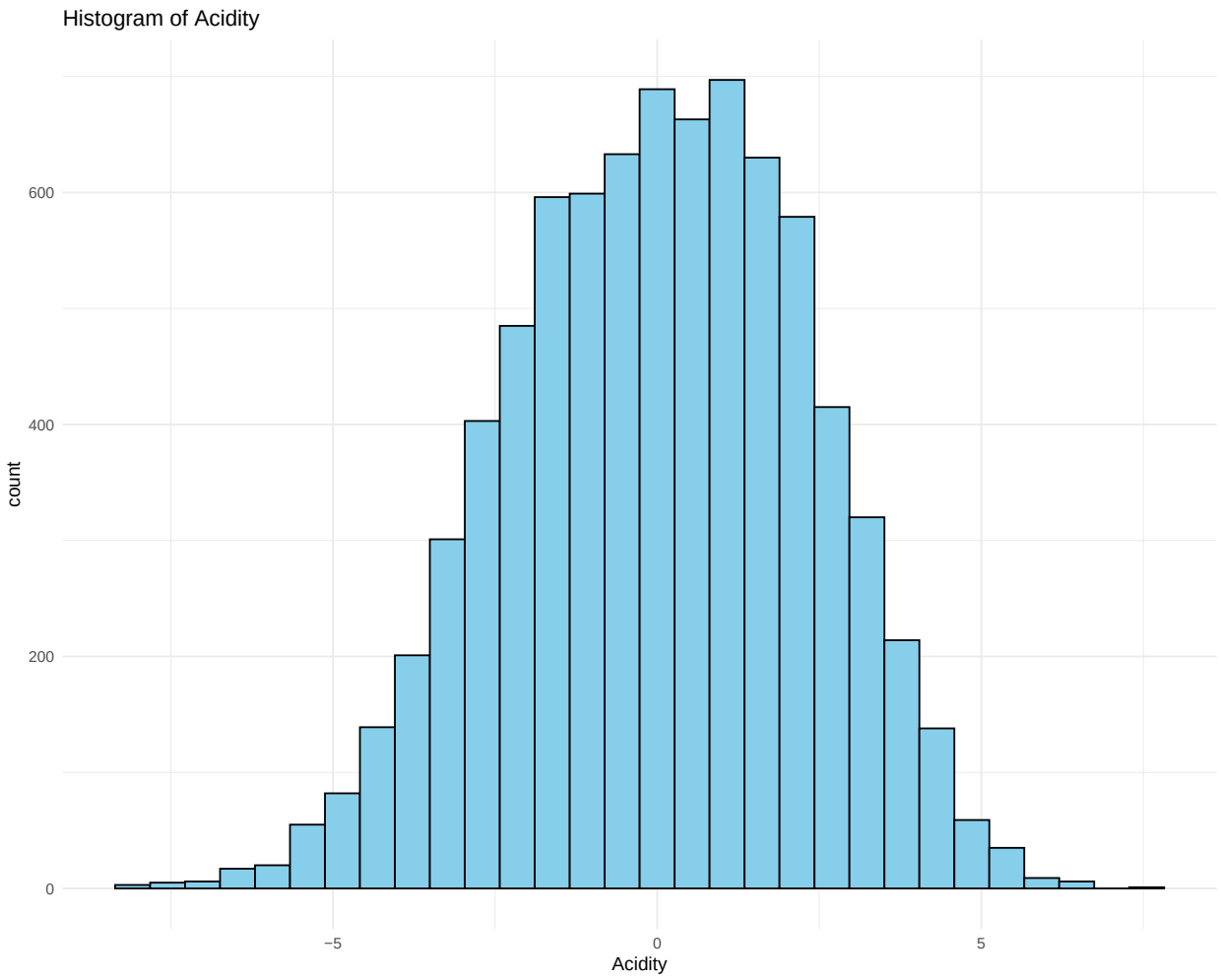

Histogram of Size
800
600
400
200
0
count
−5
0
5
Size

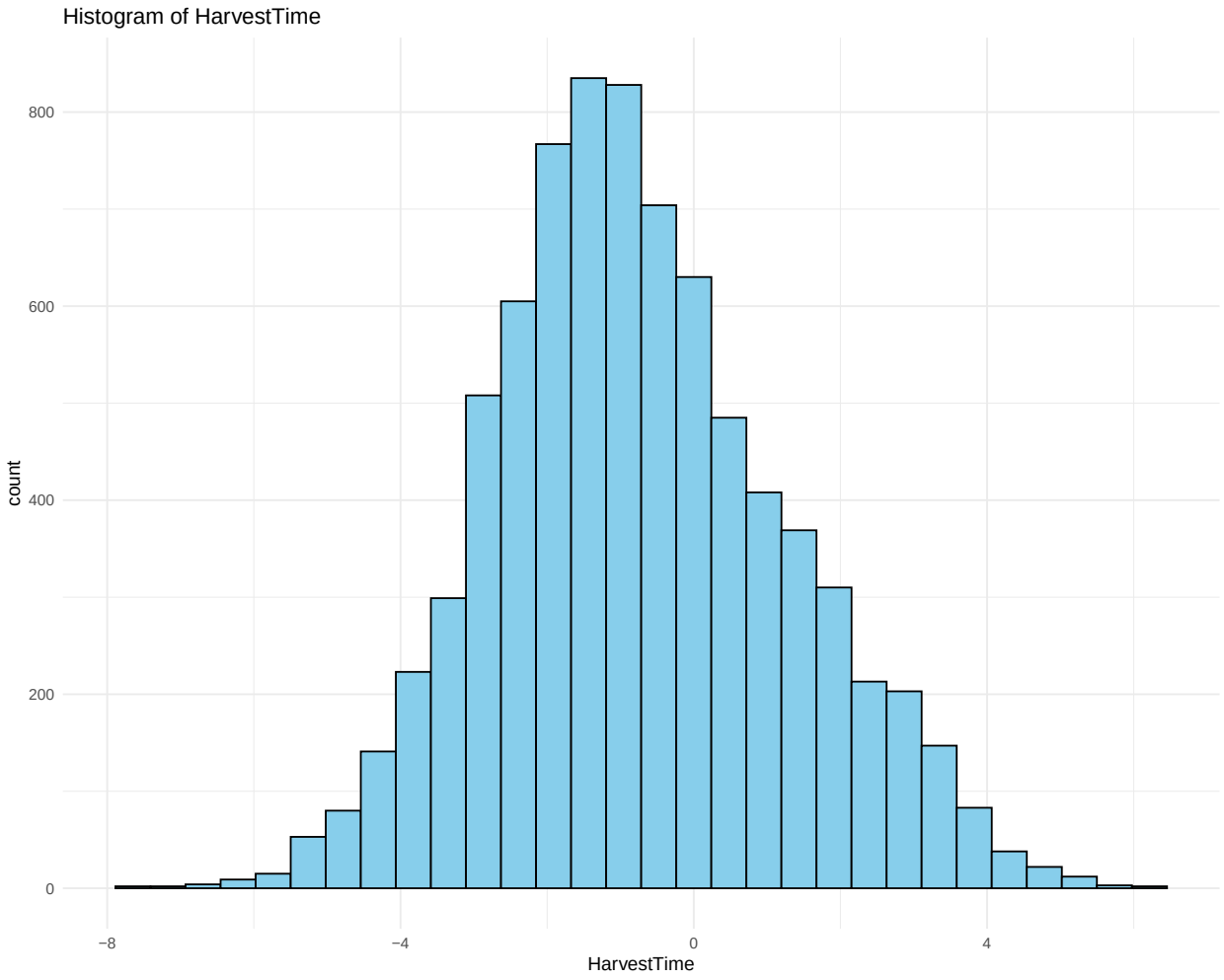

Histogram of Weight
count
600
400
200
0
−5
0
5
Weight


Histogram of Sweetness
count
750
500
250
0
−4
0
4
8
Sweetness


Histogram of Softness
count
750
500
250
0
−5
0
5
Softness


Histogram of Ripeness
count
800
600
400
200
0
−8
−4
0
4
8
Ripeness

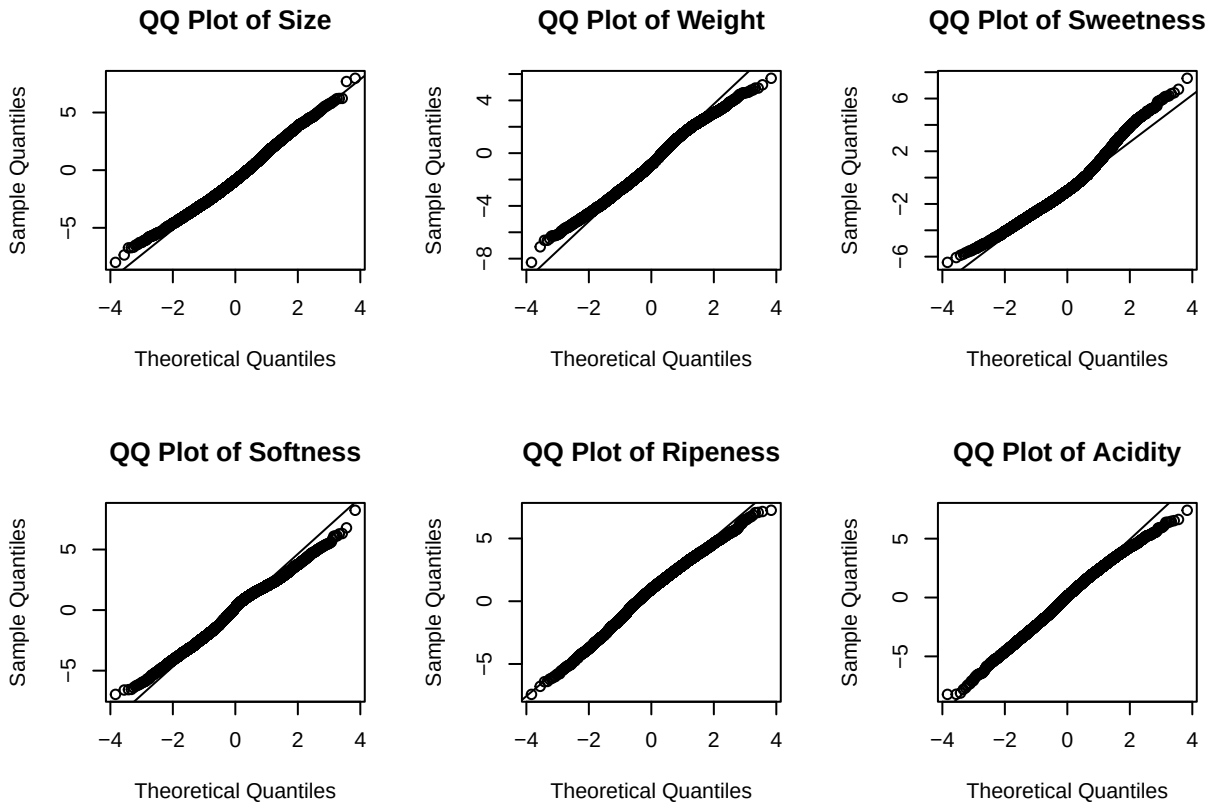

Histogram of Acidity
count
600
400
200
0
−5
0
5
Acidity



QQ Plots for Assessing Normality

QQ plots help in assessing if the variables follow a normal distribution.

```
# Generating QQ plots
par(mfrow = c(2, 3))
for (var in num_vars[-7]) { # Excluding 'HarvestTime'
  qqnorm(banana_data[[var]], main = paste("QQ Plot of", var))
  qqline(banana_data[[var]])
}
```

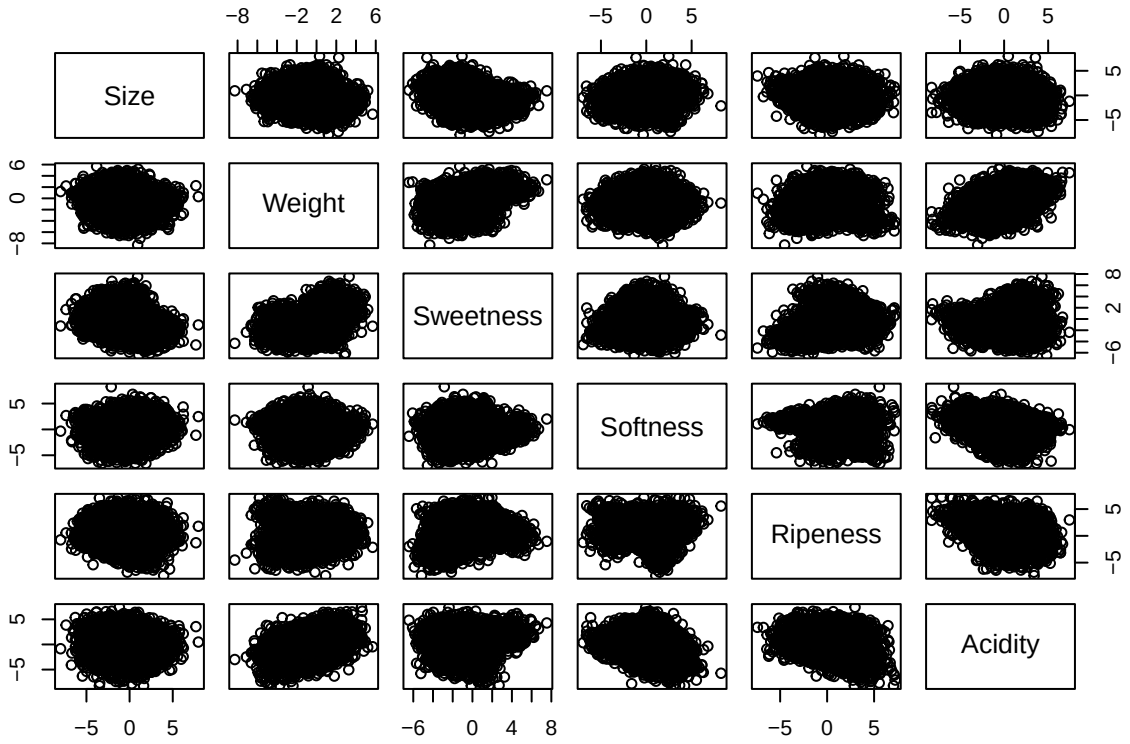



```
par(mfrow = c(1, 1))
```

Pairwise Scatterplots

Investigate the relationships between variables with pairwise scatterplots.

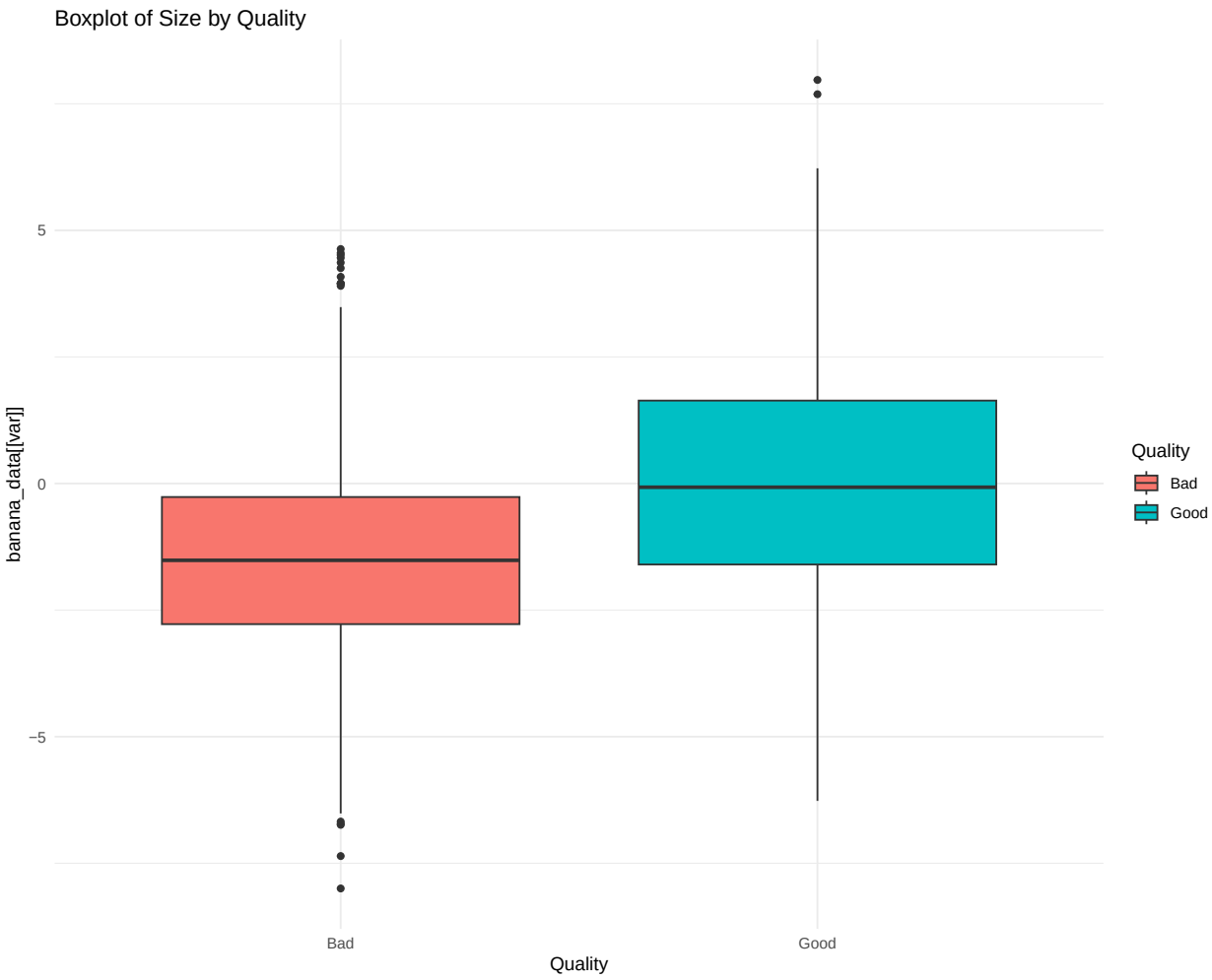
```
pairs(~ Size + Weight + Sweetness + Softness + Ripeness + Acidity, data = banana_data)
```

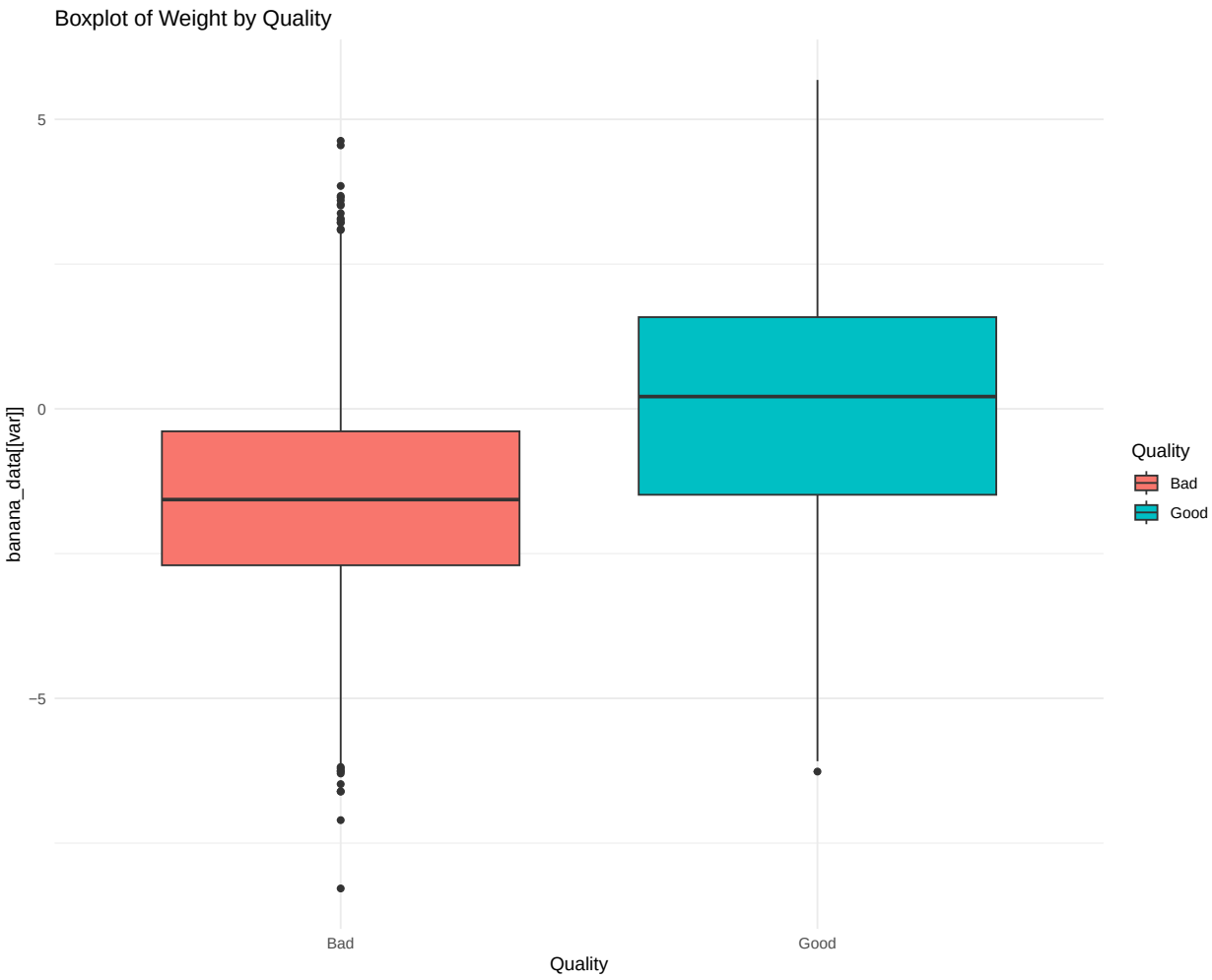


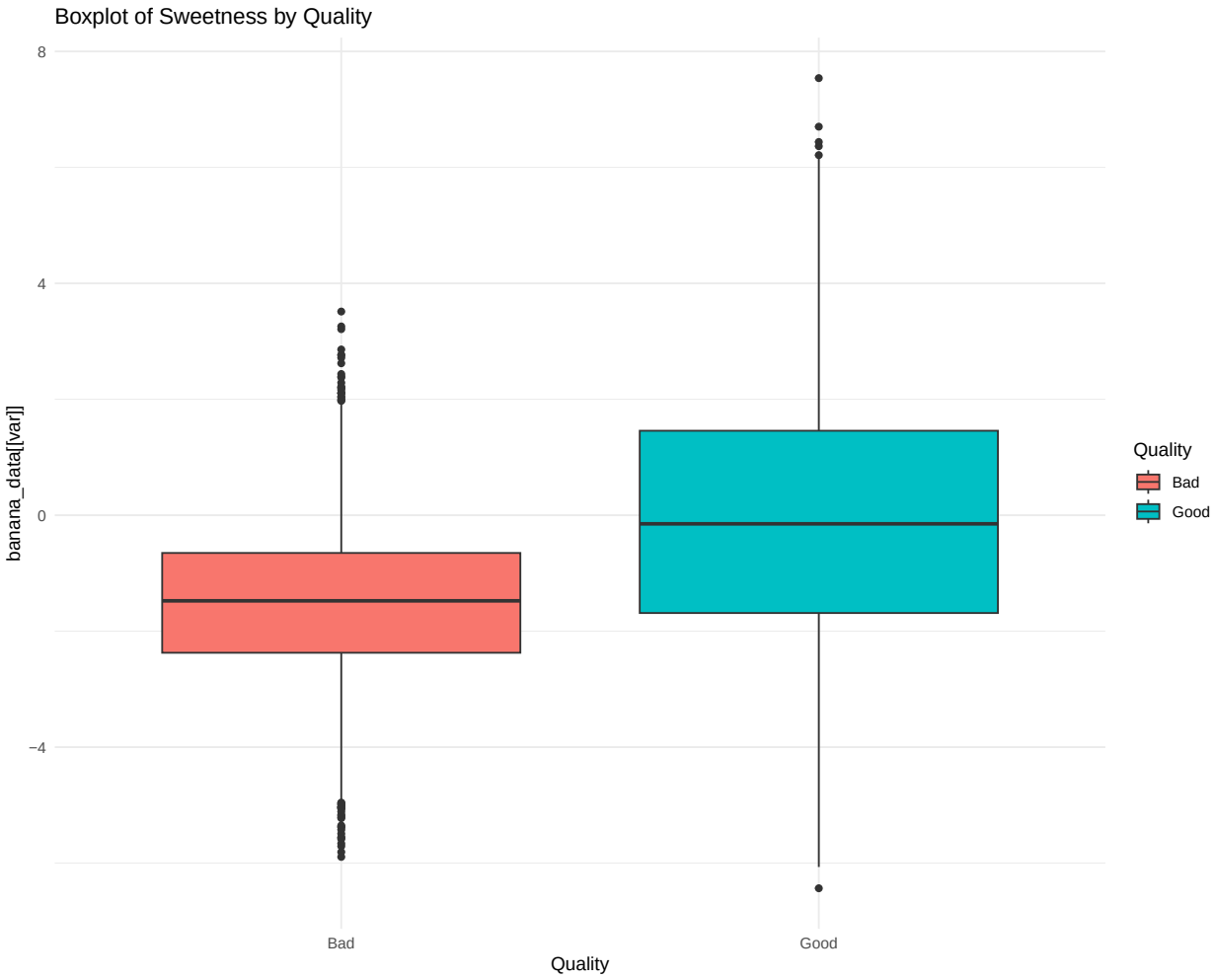
Boxplots by Quality

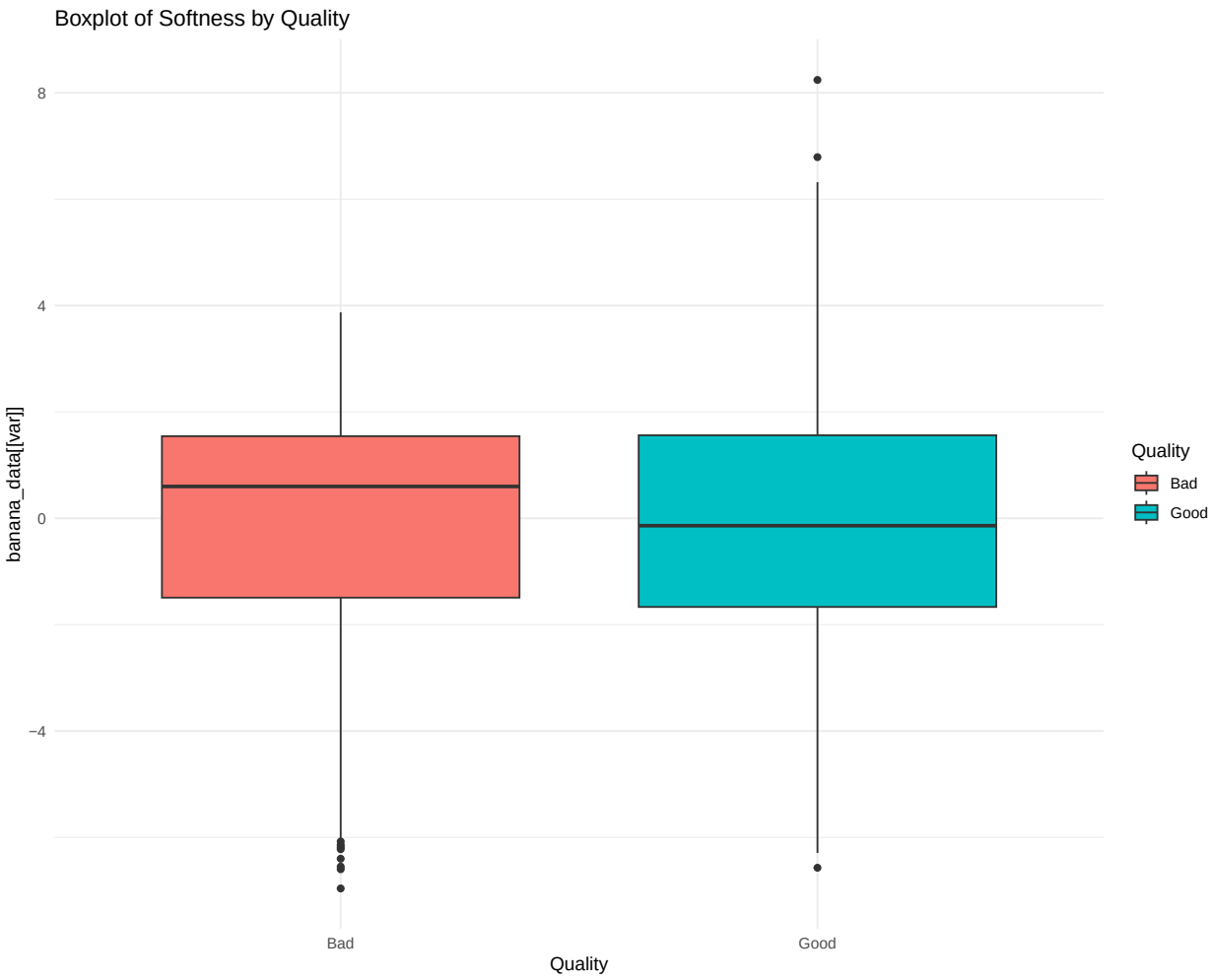
Boxplots are excellent for comparing distributions between two or more groups.

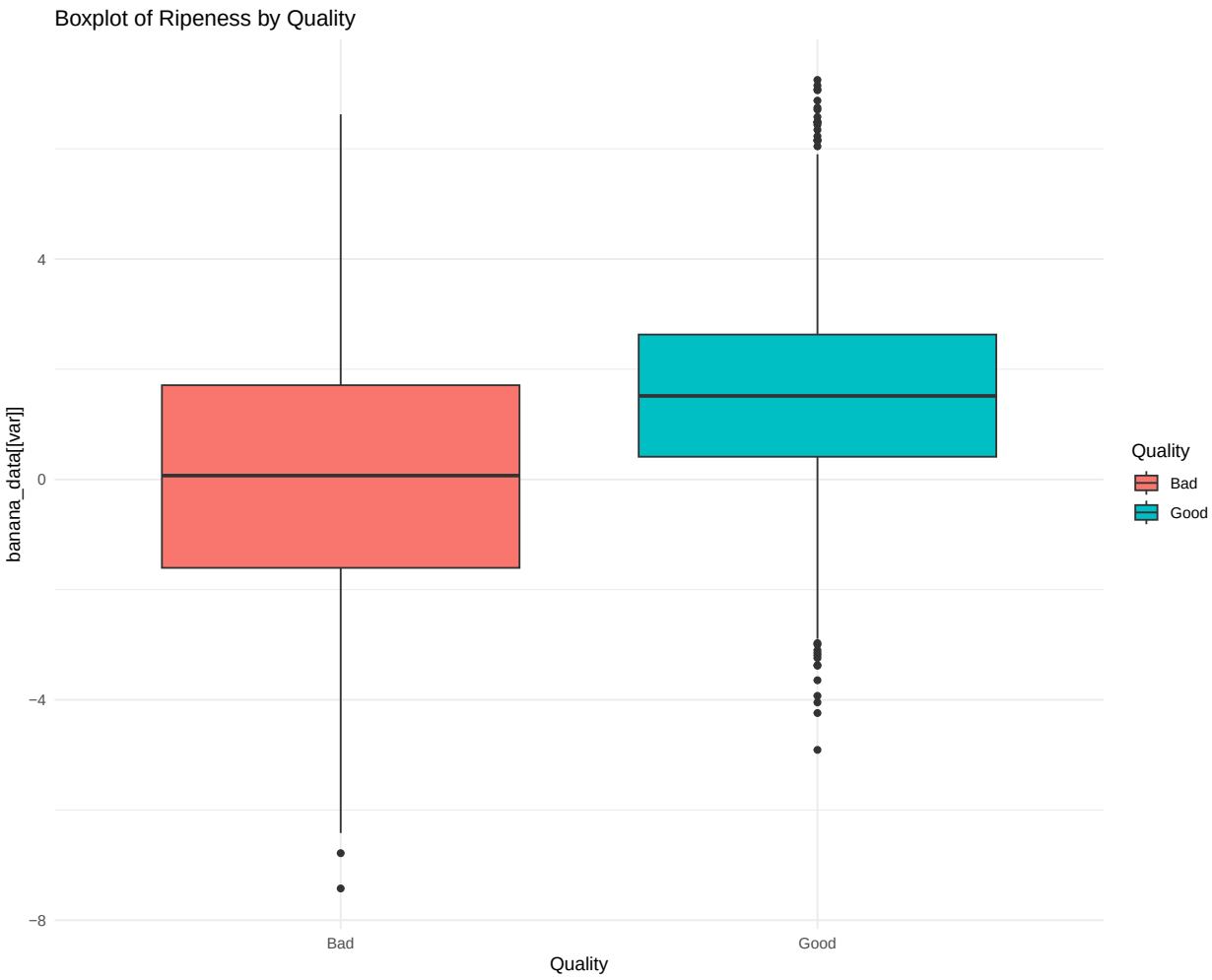
```
for (var in num_vars[-7]) { # Excluding 'HarvestTime'
  print(ggplot(banana_data, aes(x = Quality, y = banana_data[[var]], fill = Quality)) +
          geom_boxplot() +
          theme_minimal() +
          ggtitle(paste("Boxplot of", var, "by Quality")))
}
```

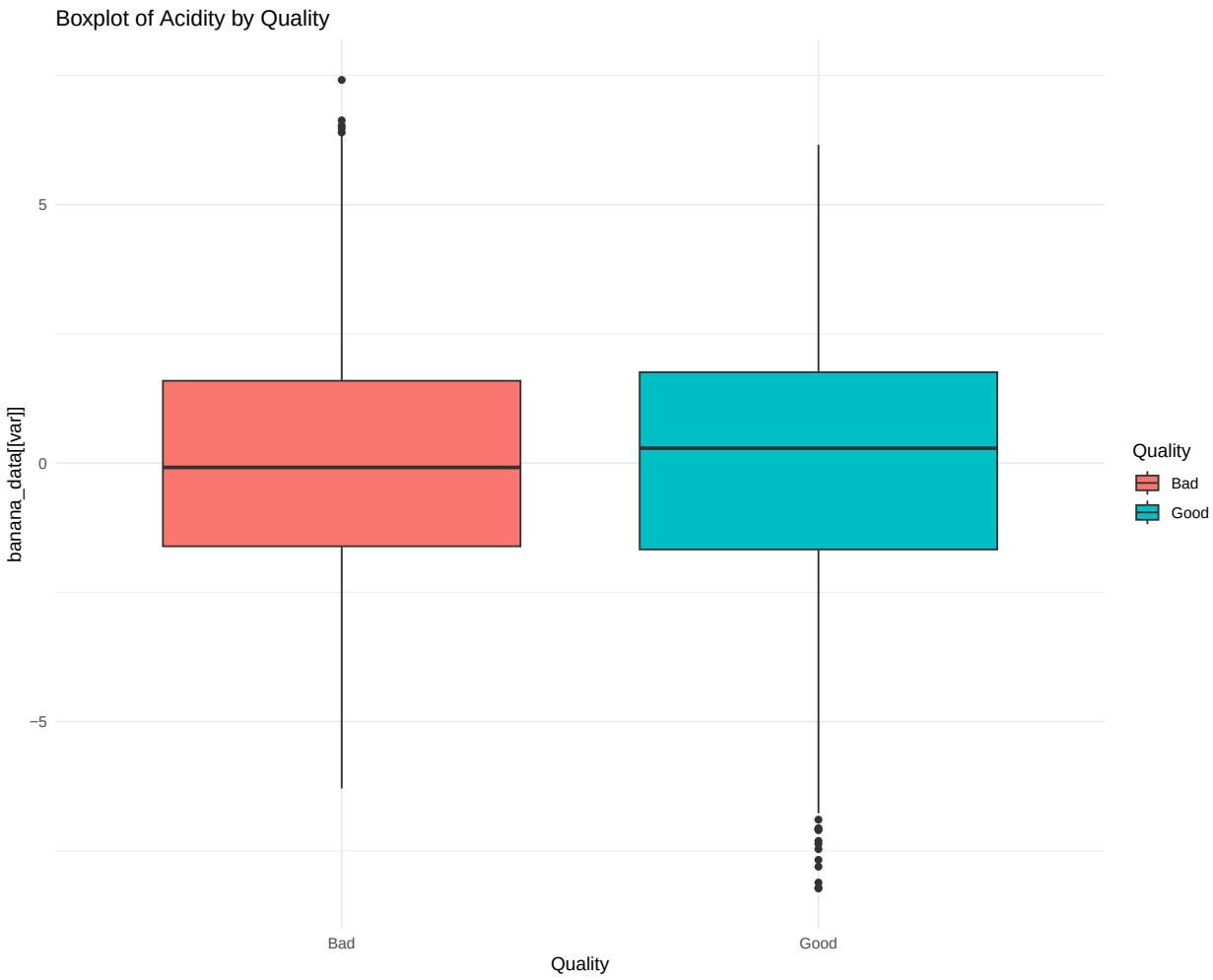

Boxplot of Size by Quality
banana_data[[var]]
5
0
−5
Bad
Good
Quality
Quality
Bad
Good


Boxplot of Weight by Quality
banana_data[[var]]
5
0
−5
Bad
Good
Quality
Quality
Bad
Good


Boxplot of Sweetness by Quality
8
4
0
−4
banana_data[[var]]
Bad
Good
Quality
Quality
Bad
Good


Boxplot of Softness by Quality
8
4
0
−4
banana_data[[var]]
Bad
Good
Quality
Quality
Bad
Good


Boxplot of Ripeness by Quality
banana_data[[var]]
4
0
−4
−8
Bad
Good
Quality
Quality
Bad
Good

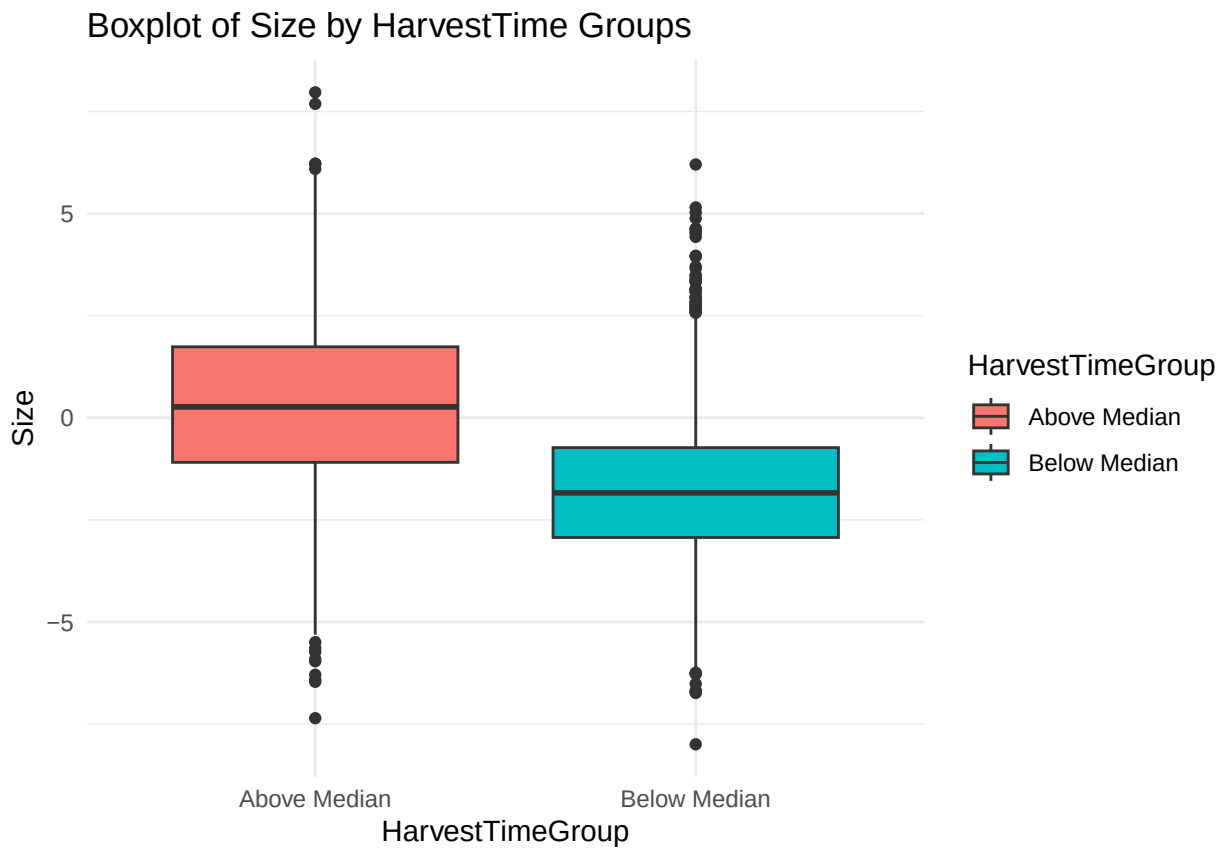


Further Analysis

Based on the initial EDA, you might want to explore specific hypotheses or relationships further. For instance, comparing the characteristics of bananas with 'HarvestTime' above and below the median.

```
# Example: Comparing 'Size' based on 'HarvestTime' median split
median_harvest_time <- median(banana_data$HarvestTime)
banana_data$HarvestTimeGroup <- ifelse(banana_data$HarvestTime > median_harvest_time, "Above Median", "

ggplot(banana_data, aes(x = HarvestTimeGroup, y = Size, fill = HarvestTimeGroup)) +
  geom_boxplot() +
  theme_minimal() +
  ggtitle("Boxplot of Size by HarvestTime Groups")
```



Conclusion

In this report, we conducted a preliminary exploration of the banana dataset, focusing on the relationships between various characteristics and banana quality. Further analyses are recommended to explore identified trends and hypotheses.

“‘

Notes for Running the Rmd Document:

- Make sure to install the necessary packages (`readr`, `ggplot2`, `dplyr`) using `install.packages()` if you haven't already.
- Adjust the file path to `"banana.csv"` as needed based on where your data file is located.
- This Rmd document is structured to render as an HTML document, which is versatile and easy to share. You can change the output to PDF or Word by modifying the `output:`