Exploratory Data Analysis

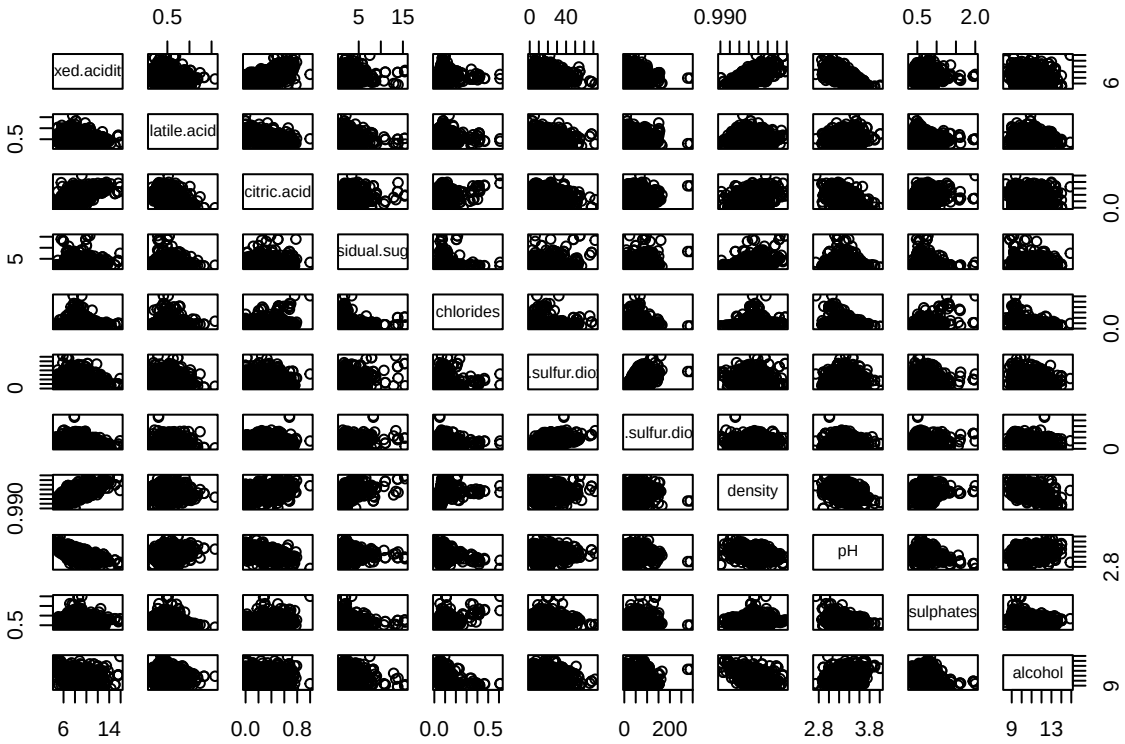
```
# Exploratory Data Analysis

# Read your CSV data
data <- read.csv("winequality-red.csv", sep=';')

# Summary statistics
summary_stats <- summary(data)

# Features
features <- c("fixed.acidity", "volatile.acidity", "citric.acid",
              "residual.sugar", "chlorides", "free.sulfur.dioxide",
              "total.sulfur.dioxide", "density", "pH", "sulphates",
              "alcohol")

# Pairwise scatterplots
pairwise_scatterplots <- pairs(data[, features])
```




```
# Histograms for continuous variables
histograms <- lapply(features, function(var) {
  ggplot(data, aes(x = get(var))) +
    geom_histogram(binwidth = 1, fill = "blue", color = "black", alpha = 0.7) +
    labs(title = paste("Histogram of", var), x = var, y = "Frequency") +
    theme_minimal()
})
```

```r
# Boxplots - Separated out sulfur since scales much larger
# Extracting the variables for separate box plots
sulfur_variables <- c("total.sulfur.dioxide", "free.sulfur.dioxide")
other_variables <- setdiff(features, sulfur_variables)

# Boxplot for sulfur variables
boxplot_sulfur <- ggplot(data %>% pivot_longer(cols = sulfur_variables),
                         aes(x = name, y = value)) +
  geom_boxplot(fill = "blue", color = "black", alpha = 0.7, width = 0.5) +
  labs(title = "Boxplots of Sulfur Variables", x = "Variable", y = "Value") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(expand = expansion(mult = c(0.05, 0.2)))

# Boxplot for other variables
boxplot_other <- ggplot(data %>% pivot_longer(cols = other_variables),
                        aes(x = name, y = value)) +
  geom_boxplot(fill = "blue", color = "black", alpha = 0.7, width = 0.5) +
  labs(title = "Boxplots of Other Variables", x = "Variable", y = "Value") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_y_continuous(expand = expansion(mult = c(0.05, 0.2)))

# Quantiles, means, medians, and standard deviations
quantiles_means_medians_sds <- data %>%
  summarise(across(features, list(quantiles = ~quantile(.),
                                  mean = ~mean(.),
                                  median = ~median(.),
                                  sd = ~sd(.)))))

# Print summary statistics
print(summary_stats)
```

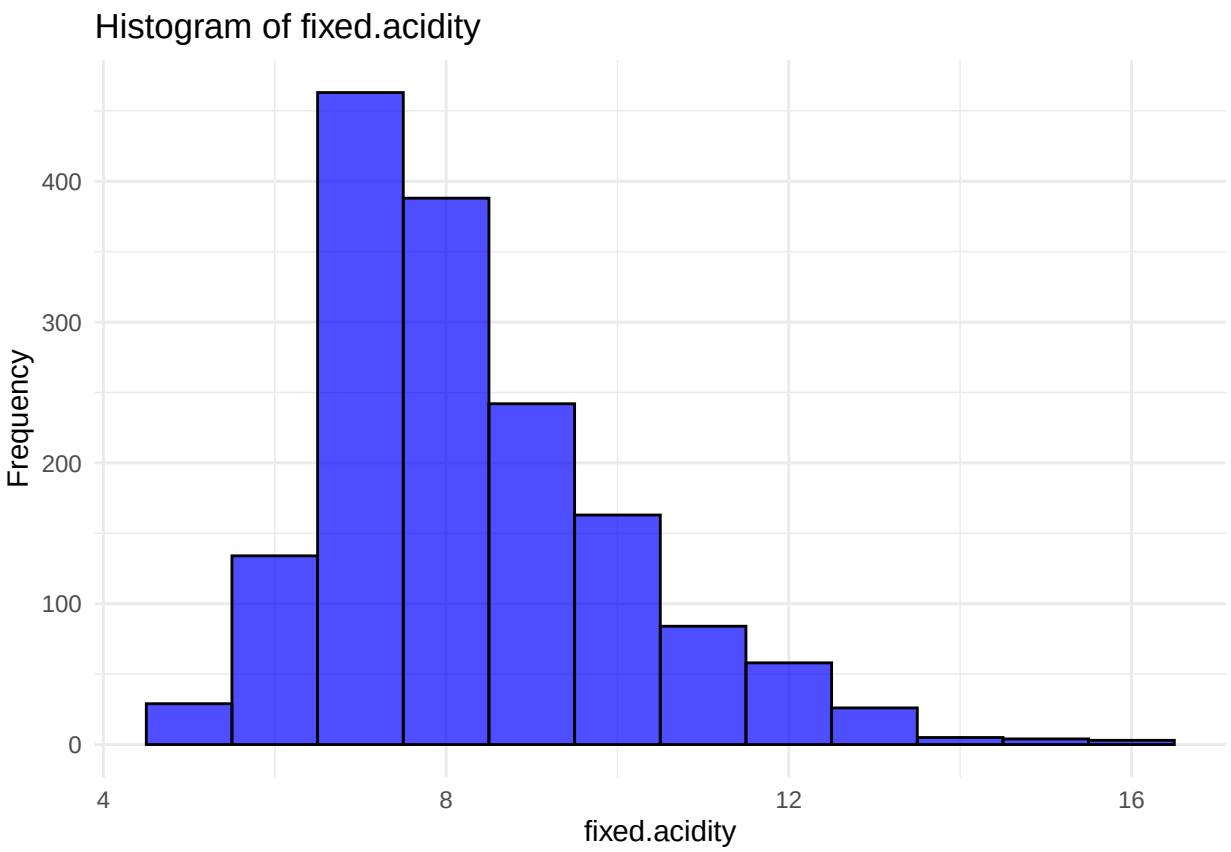
```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
```

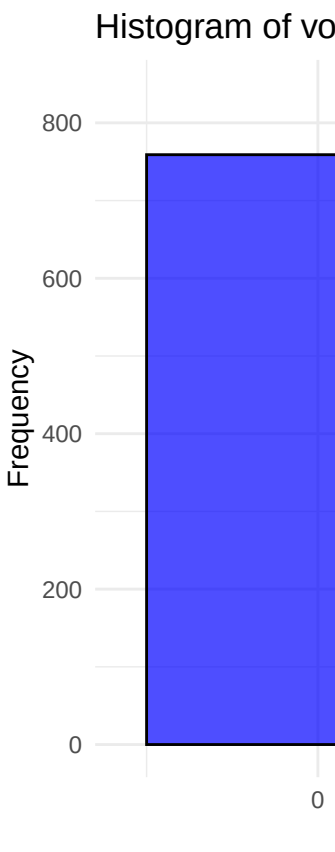
```
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

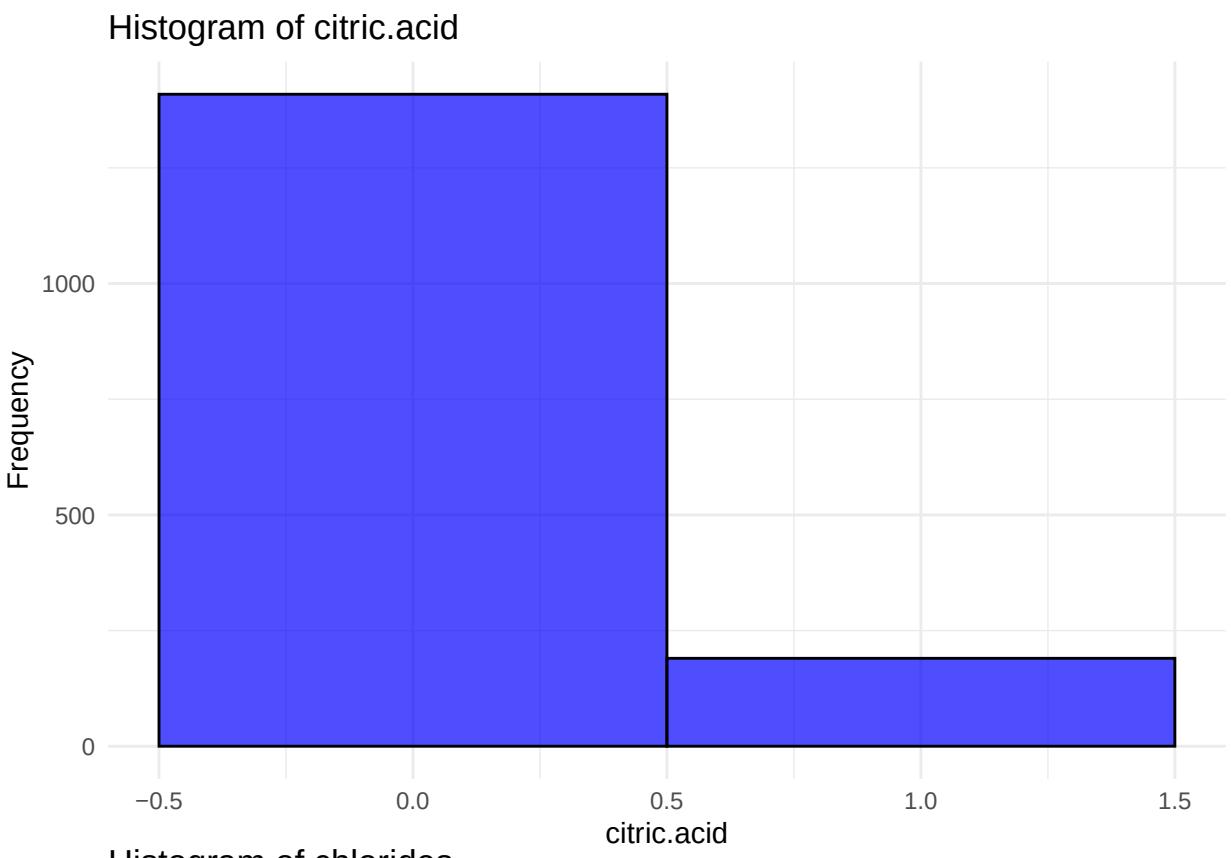
```
# Display pairwise scatterplots
print(pairwise_scatterplots)
```

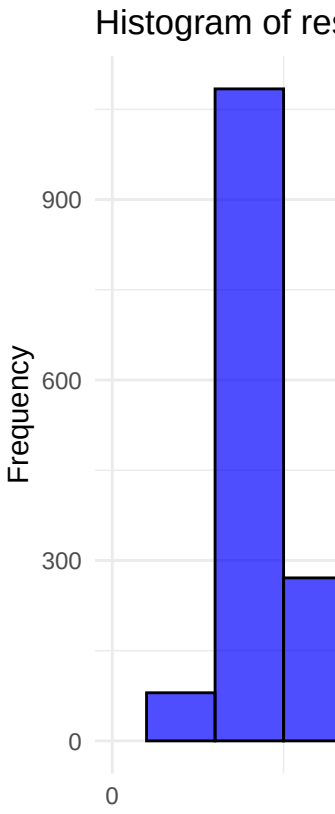
```
## NULL
```

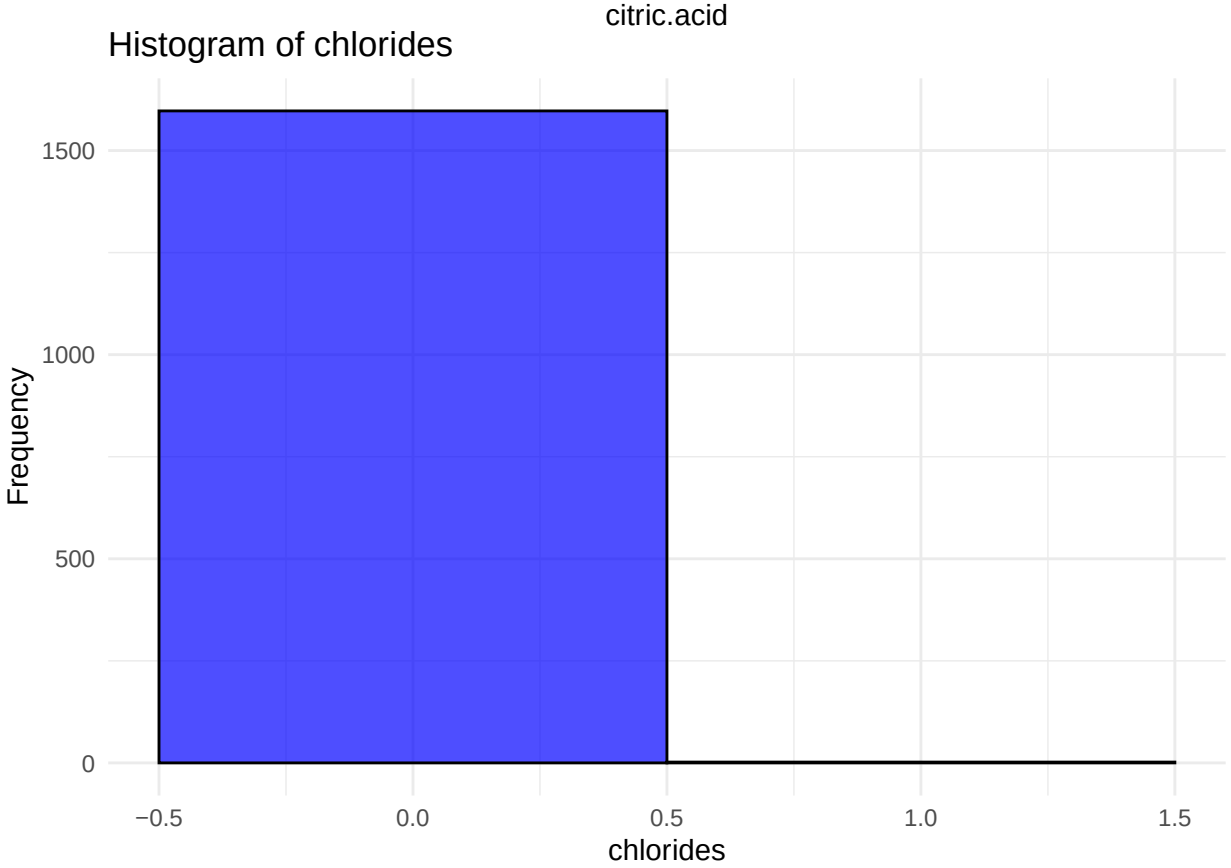
```
# Display histograms
for (hist_plot in histograms) {
  print(hist_plot)
}
```

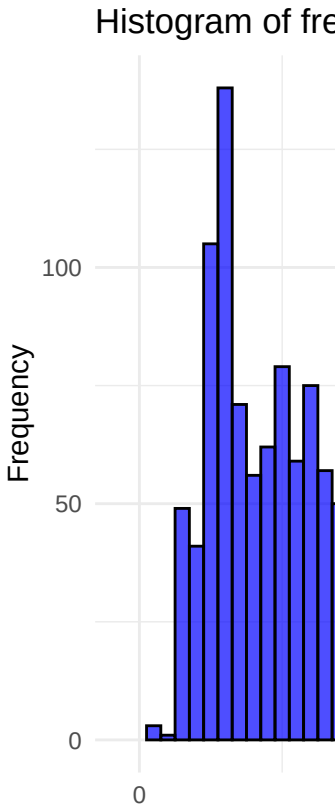


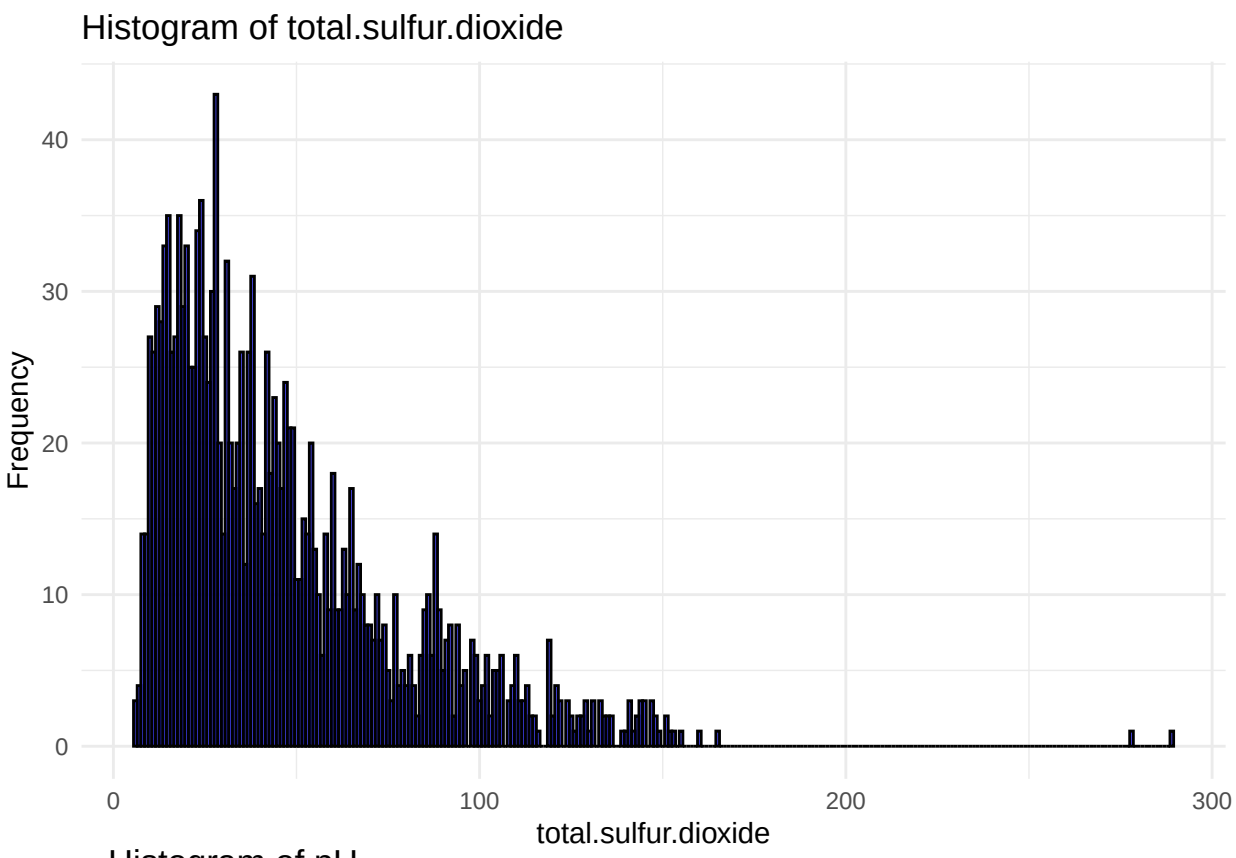


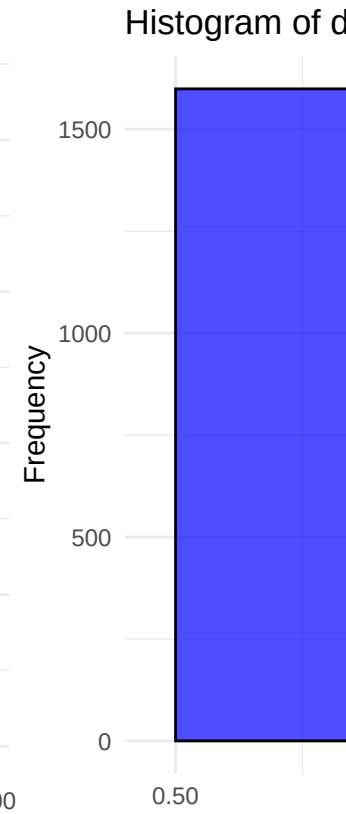

Histogram of citric.acid
Frequency
1000
500
0
−0.5
0.0
0.5
1.0
1.5
citric.acid

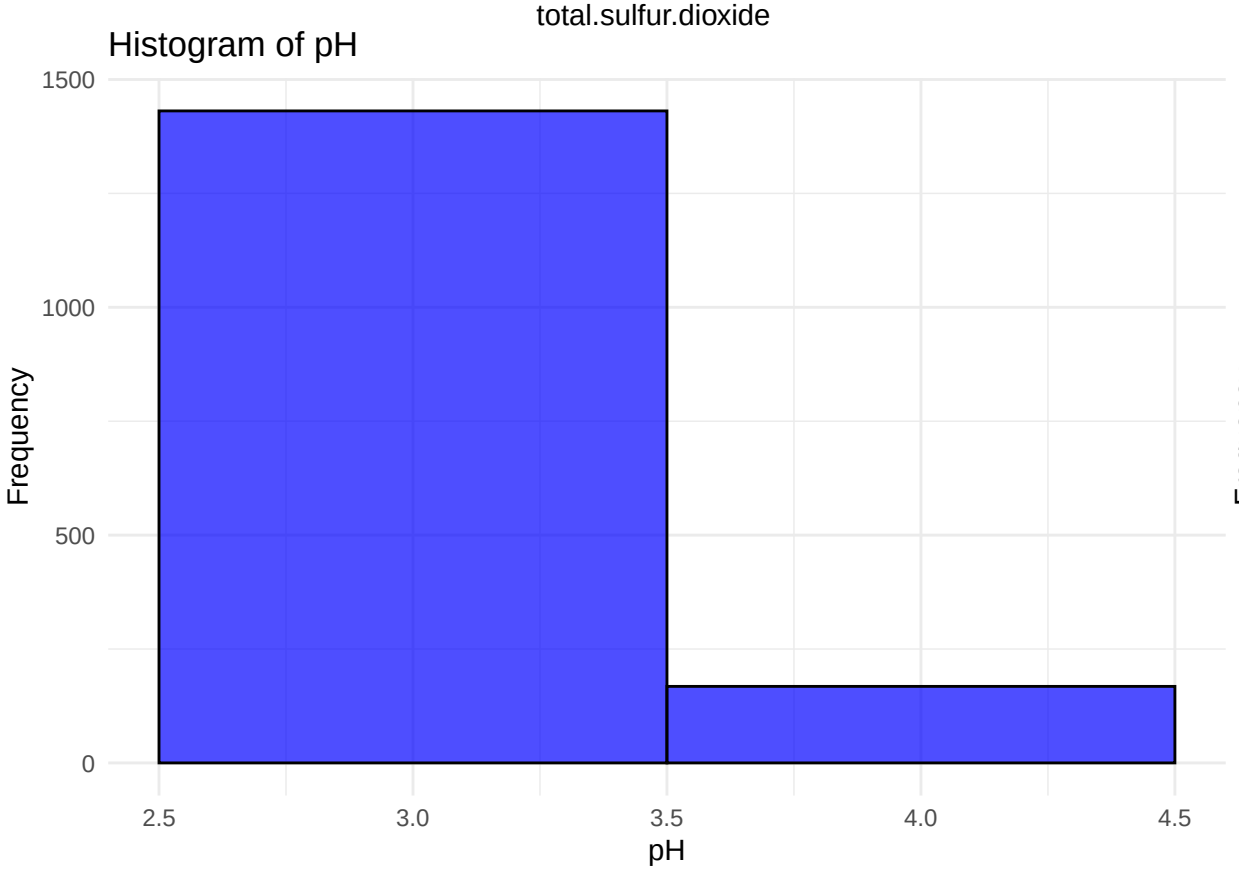

Histogram of re
Frequency
900
600
300
0
0

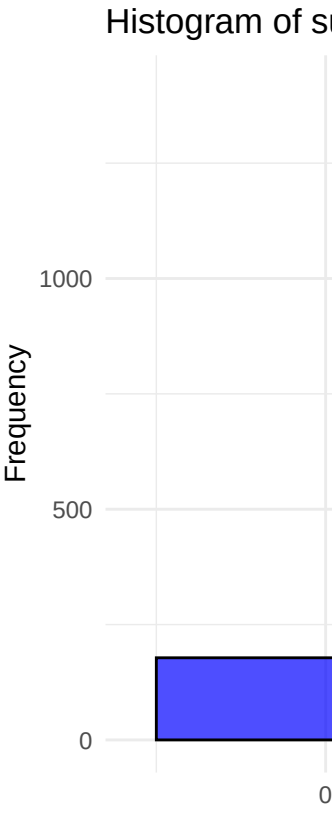

Histogram of chlorides
Frequency
1500
1000
500
0
−0.5
0.0
0.5
1.0
1.5
chlorides

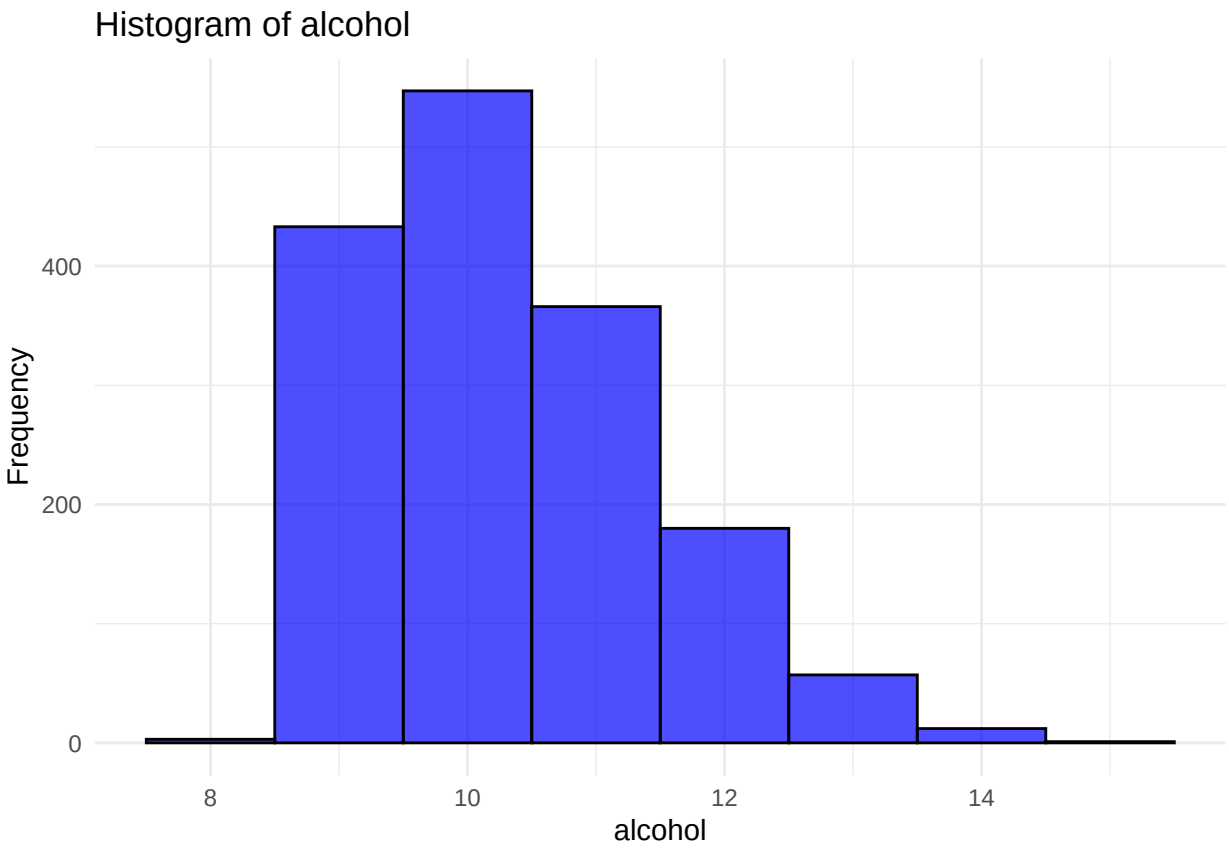

Histogram of fre
Frequency
100
50
0
0


Histogram of total.sulfur.dioxide
Frequency
total.sulfur.dioxide
0
10
20
30
40
0
100
200
300

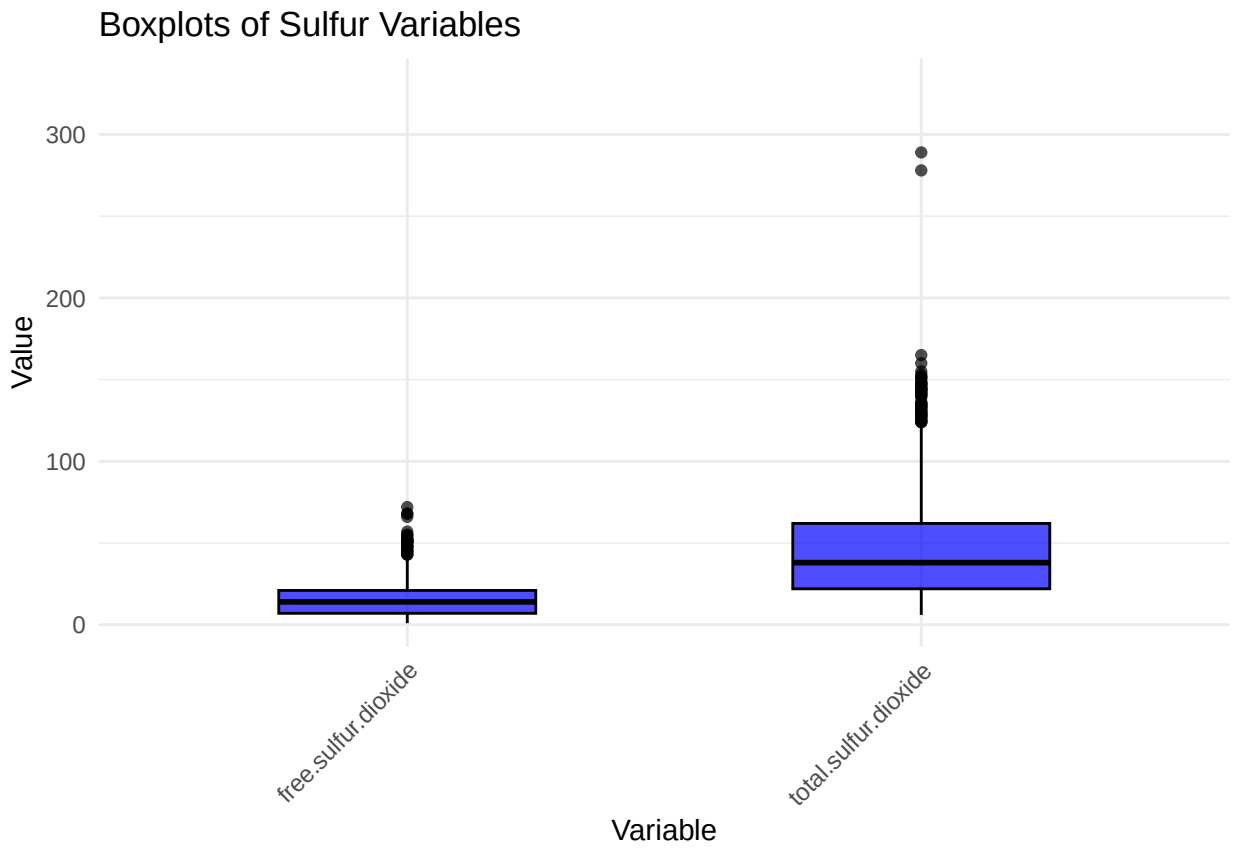

Histogram of d
Frequency
0
500
1000
1500
0.50


Histogram of pH
Frequency
pH
0
500
1000
1500
2.5
3.0
3.5
4.0
4.5

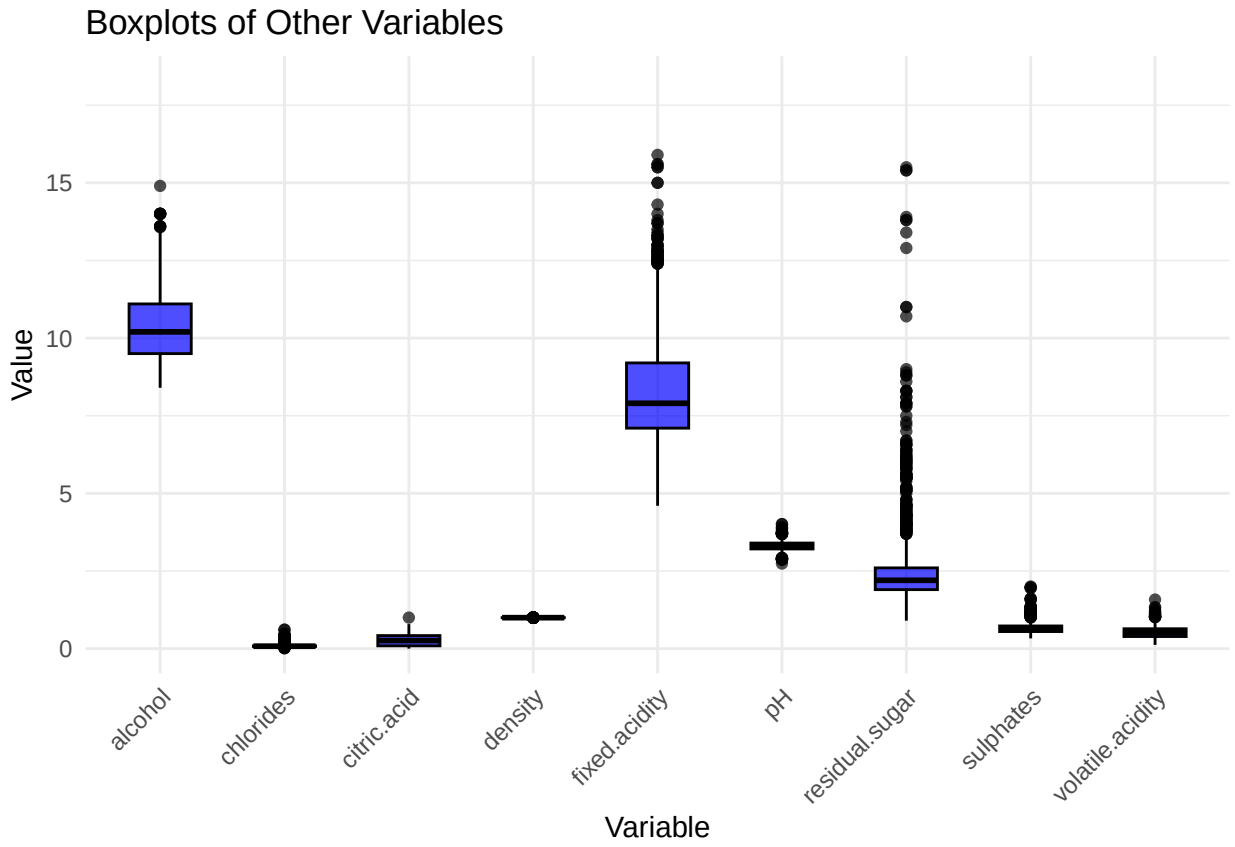

Histogram of su
Frequency
0
500
1000




```
# Display boxplot
print(boxplot_sulfur)
```




```
print(boxplot_other)
```



```
# Display quantiles, means, medians, and standard deviations
print(quantiles_means_medians_sds)
```

```
##   fixed.acidity_quantiles fixed.acidity_mean fixed.acidity_median
## 1                     4.6          8.319637                 7.9
## 2                     7.1          8.319637                 7.9
## 3                     7.9          8.319637                 7.9
## 4                     9.2          8.319637                 7.9
## 5                    15.9          8.319637                 7.9
##   fixed.acidity_sd volatile.acidity_quantiles volatile.acidity_mean
## 1         1.741096                      0.12            0.5278205
## 2         1.741096                      0.39            0.5278205
## 3         1.741096                      0.52            0.5278205
## 4         1.741096                      0.64            0.5278205
## 5         1.741096                      1.58            0.5278205
##   volatile.acidity_median volatile.acidity_sd citric.acid_quantiles
## 1                    0.52          0.1790597                  0.00
## 2                    0.52          0.1790597                  0.09
## 3                    0.52          0.1790597                  0.26
## 4                    0.52          0.1790597                  0.42
## 5                    0.52          0.1790597                  1.00
##   citric.acid_mean citric.acid_median citric.acid_sd residual.sugar_quantiles
## 1        0.2709756               0.26      0.1948011                      0.9
## 2        0.2709756               0.26      0.1948011                      1.9
## 3        0.2709756               0.26      0.1948011                      2.2
## 4        0.2709756               0.26      0.1948011                      2.6
## 5        0.2709756               0.26      0.1948011                     15.5
##   residual.sugar_mean residual.sugar_median residual.sugar_sd
## 1            2.538806                  2.2          1.409928
## 2            2.538806                  2.2          1.409928
## 3            2.538806                  2.2          1.409928
## 4            2.538806                  2.2          1.409928
## 5            2.538806                  2.2          1.409928
##   chlorides_quantiles chlorides_mean chlorides_median chlorides_sd
## 1               0.012     0.08746654            0.079    0.0470653
## 2               0.070     0.08746654            0.079    0.0470653
## 3               0.079     0.08746654            0.079    0.0470653
## 4               0.090     0.08746654            0.079    0.0470653
## 5               0.611     0.08746654            0.079    0.0470653
##   free.sulfur.dioxide_quantiles free.sulfur.dioxide_mean
## 1                             1                15.87492
## 2                             7                15.87492
## 3                            14                15.87492
## 4                            21                15.87492
## 5                            72                15.87492
```

```
##   free.sulfur.dioxide_median free.sulfur.dioxide_sd
## 1                         14               10.46016
## 2                         14               10.46016
## 3                         14               10.46016
## 4                         14               10.46016
## 5                         14               10.46016
##   total.sulfur.dioxide_quantiles total.sulfur.dioxide_mean
## 1                              6                  46.46779
## 2                             22                  46.46779
## 3                             38                  46.46779
## 4                             62                  46.46779
## 5                            289                  46.46779
##   total.sulfur.dioxide_median total.sulfur.dioxide_sd density_quantiles
## 1                          38                32.89532          0.990070
## 2                          38                32.89532          0.995600
## 3                          38                32.89532          0.996750
## 4                          38                32.89532          0.997835
## 5                          38                32.89532          1.003690
##   density_mean density_median  density_sd pH_quantiles  pH_mean pH_median
## 1    0.9967467        0.99675 0.001887334         2.74 3.311113      3.31
## 2    0.9967467        0.99675 0.001887334         3.21 3.311113      3.31
## 3    0.9967467        0.99675 0.001887334         3.31 3.311113      3.31
## 4    0.9967467        0.99675 0.001887334         3.40 3.311113      3.31
## 5    0.9967467        0.99675 0.001887334         4.01 3.311113      3.31
##       pH_sd sulphates_quantiles sulphates_mean sulphates_median sulphates_sd
## 1 0.1543865                0.33      0.6581488             0.62     0.169507
## 2 0.1543865                0.55      0.6581488             0.62     0.169507
## 3 0.1543865                0.62      0.6581488             0.62     0.169507
## 4 0.1543865                0.73      0.6581488             0.62     0.169507
## 5 0.1543865                2.00      0.6581488             0.62     0.169507
##   alcohol_quantiles alcohol_mean alcohol_median alcohol_sd
## 1               8.4     10.42298           10.2   1.065668
## 2               9.5     10.42298           10.2   1.065668
## 3              10.2     10.42298           10.2   1.065668
## 4              11.1     10.42298           10.2   1.065668
## 5              14.9     10.42298           10.2   1.065668
```

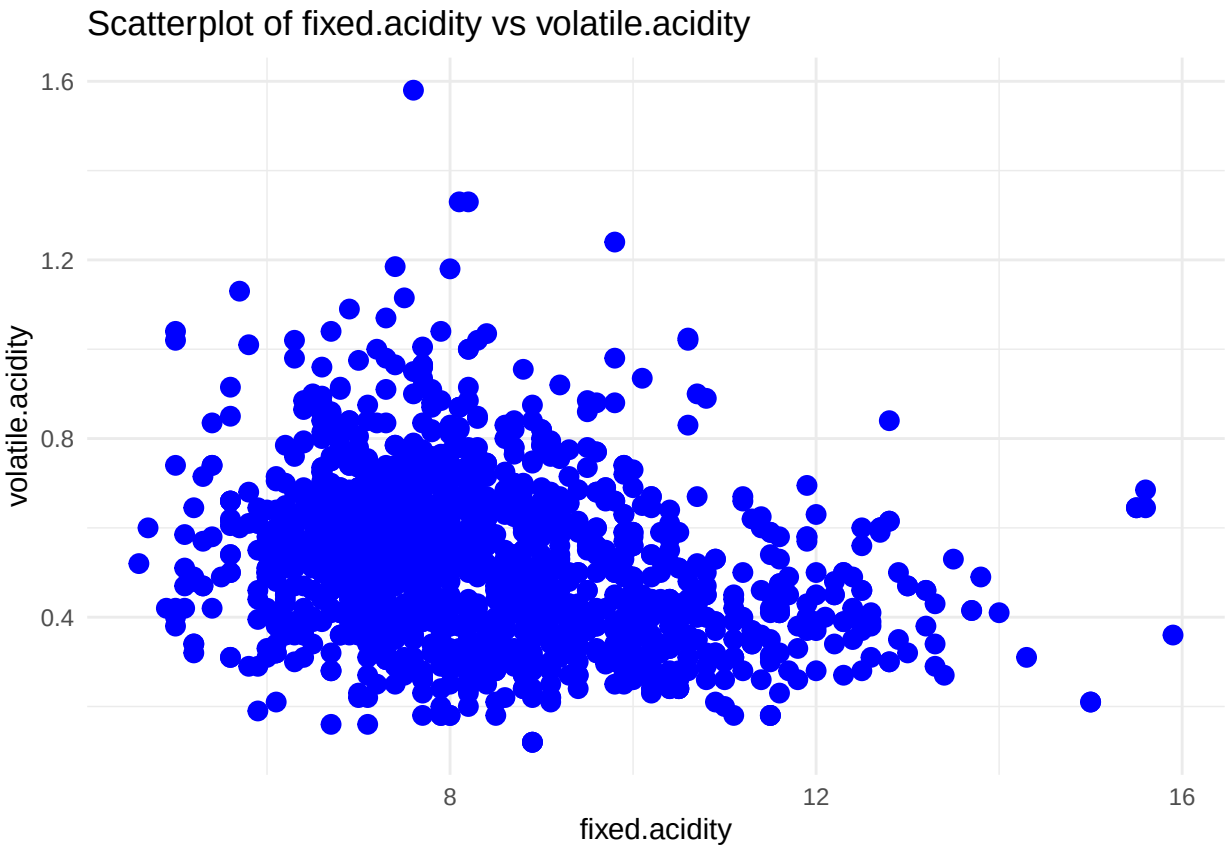
```r
# Function to create a scatterplot of two features based on scatterplot matrix
feature1 <- data$fixed.acidity
feature <- data$pH

create_scatterplot <- function(data, feature1, feature2) {
  ggplot(data, aes_string(x = feature1, y = feature2)) +
    geom_point(color = "blue", size = 3) +
    labs(title = paste("Scatterplot of", feature1, "vs", feature2),
```

```
         x = feature1, y = feature2) +
    theme_minimal()
}

# Example: Create a scatterplot for "fixed.acidity" vs "volatile.acidity"
scatterplot_example <- create_scatterplot(data, "fixed.acidity", "volatile.acidity")

# Print the scatterplot
print(scatterplot_example)
```




```
# Frequentist Ordinal Regression

# Assuming you have a data frame named "data" containing your features and labels
set.seed(123)  # Set seed for reproducibility
split_index <- createDataPartition(data$quality, p = 0.8, list = FALSE)

# Train/Test Split
train_data <- data[split_index, ]
test_data <- data[-split_index, ]

# Backward elimination for the most parsimonious model
all_features <- setdiff(names(train_data), "quality")
selected_features <- all_features
```

```
best_accuracy <- 0
best_feature_set <- NULL
best_num_variables <- Inf  # Initialize with a large value

while (length(selected_features) >= 1) {
  current_accuracy <- 0
  current_num_variables <- length(selected_features)
  worst_feature <- NULL
  exit_loop <- FALSE  # Flag to control loop exit

  for (feature in selected_features) {
    current_features <- setdiff(selected_features, feature)

    if (length(current_features) == 0) {
      exit_loop <- TRUE
      break  # Exit the inner loop when only one feature is left
    }

    # Train the model with the current set of features
    model <- polr(factor(quality) ~ ., data = train_data[, c("quality", current_features)], Hess = TRUE)

    # Predict using the trained model on the test set
    # (Ensure test_data is also a data frame)
    predicted_labels <- predict(model, newdata = as.data.frame(test_data))

    # Convert predicted and true labels to integers (if not already)
    predicted_labels <- as.integer(predicted_labels)
    true_labels <- as.integer(test_data$quality)

    # Compute accuracy
    accuracy <- sum(predicted_labels == true_labels) / length(true_labels)

    # Update the current accuracy, worst feature, and number of variables if needed
    if (accuracy > current_accuracy) {
      current_accuracy <- accuracy
      worst_feature <- feature
      current_num_variables <- length(current_features)
    }
  }

  if (exit_loop) {
    break  # Exit the outer loop when only one feature is left
  }
```

```
  # Remove the worst feature from the selected features
  selected_features <- setdiff(selected_features, worst_feature)

  # Update the best feature set, accuracy, and number of variables if needed
  if (current_accuracy > best_accuracy ||
      (current_accuracy == best_accuracy && current_num_variables < best_num_variables)) {
    best_accuracy <- current_accuracy
    best_feature_set <- selected_features
    best_num_variables <- current_num_variables
  }

  cat("Selected features:", selected_features, "\n")
  cat("Current accuracy:", current_accuracy, "\n")
  cat("Current number of variables:", current_num_variables, "\n\n")

}
```

```
## Selected features: fixed.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxid
## Current accuracy: 0.01572327
## Current number of variables: 10
##
## Selected features: fixed.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxid
## Current accuracy: 0.02515723
## Current number of variables: 9
##
## Selected features: citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density pH a
## Current accuracy: 0.02515723
## Current number of variables: 8
##
## Selected features: citric.acid chlorides free.sulfur.dioxide total.sulfur.dioxide density pH alcohol
## Current accuracy: 0.02515723
## Current number of variables: 7
##
## Selected features: citric.acid chlorides free.sulfur.dioxide total.sulfur.dioxide pH alcohol
## Current accuracy: 0.02515723
## Current number of variables: 6
##
## Selected features: citric.acid chlorides total.sulfur.dioxide pH alcohol
## Current accuracy: 0.02515723
## Current number of variables: 5
##
## Selected features: citric.acid chlorides total.sulfur.dioxide alcohol
## Current accuracy: 0.02515723
## Current number of variables: 4
```

```
##
## Selected features: citric.acid total.sulfur.dioxide alcohol
## Current accuracy: 0.02515723
## Current number of variables: 3
##
## Selected features: total.sulfur.dioxide alcohol
## Current accuracy: 0.02201258
## Current number of variables: 2
##
## Selected features: alcohol
## Current accuracy: 0.02515723
## Current number of variables: 1
```

```
cat("Best feature set:", best_feature_set, "\n")
```

```
## Best feature set: alcohol
```

```
cat("Best accuracy:", best_accuracy, "\n")
```

```
## Best accuracy: 0.02515723
```

```
cat("Best number of variables:", best_num_variables, "\n")
```

```
## Best number of variables: 1
```