Assignment 1

Vrithik Sibbadi

Load the data set

I have used a section of iris data set

Source: https://gist.github.com/curran/a08a1080b88344b0c8a7#file-iris-csv

Its a classic in data science and contains both types of variables (numerical & categorical)

```
iris <- read.csv("iris.csv")
```

View Data set structure

```
head(iris)
```

```
##   sepal.length sepal.width petal.length petal.width variety
## 1          5.1         3.5          1.4         0.2  Setosa
## 2          4.9         3.0          1.4         0.2  Setosa
## 3          4.7         3.2          1.3         0.2  Setosa
## 4          4.6         3.1          1.5         0.2  Setosa
## 5          5.0         3.6          1.4         0.2  Setosa
## 6          5.4         3.9          1.7         0.4  Setosa
```

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ sepal.length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ sepal.width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ petal.length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ petal.width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ variety     : chr  "Setosa" "Setosa" "Setosa" "Setosa" ...
```

The dataset consists of 150 observations, each with 5 variables. These variables include sepal length, sepal width, petal length, petal width (all four are numerical and measure different parts of the iris flower in centimeters), and the variety of the iris flower (a categorical variable).

Descriptive Statistics

Quantitative variables

```r
quantitative_summary <- summary(iris[,1:4])
print(quantitative_summary)
```

```
##   sepal.length    sepal.width     petal.length    petal.width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Categorical variable

```r
qualitative_summary <- table(iris$variety)
print(qualitative_summary)
```

```
##
##     Setosa Versicolor  Virginica
##         50         50         50
```

- **Sepal Length**: Ranges from 4.3 cm to 7.9 cm, with a mean of approximately 5.84 cm.
- **Sepal Width**: Shows a minimum of 2.0 cm and a maximum of 4.4 cm, with an average width of around 3.06 cm.
- **Petal Length**: Varies between 1.0 cm and 6.9 cm, with an average length of 3.76 cm.
- **Petal Width**: Spans from 0.1 cm to 2.5 cm, with a mean of about 1.2 cm.

Additionally, the dataset evenly represents three varieties of iris flowers - Setosa, Versicolor, and Virginica - each with 50 samples. This balance across species makes it an ideal dataset for classification and comparative studies in flower morphology.

Data Transformation

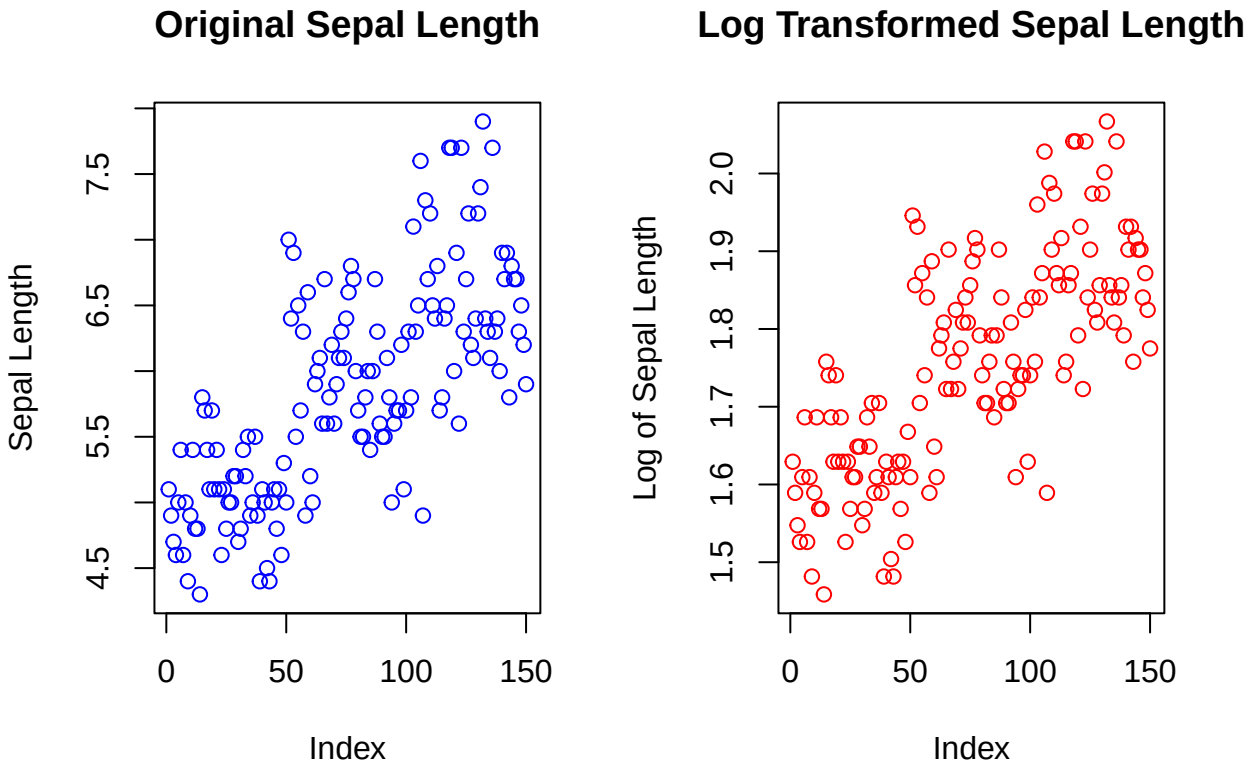
Log transformation is applied to the Sepal Length to normalize data that is not evenly distributed, stabilize variance, and linearize relationships. Make the data more suitable for analysis by reducing the impact of outliers.

```r
# Apply the log transformation to the sepal length
iris$Log_sepal.length <- log(iris$sepal.length)

# Set up the plotting area for two side-by-side plots
par(mfrow = c(1, 2))

# Plot for original sepal length
plot(iris$sepal.length, main = "Original Sepal Length",
     xlab = "Index", ylab = "Sepal Length",
     col = "blue")
```

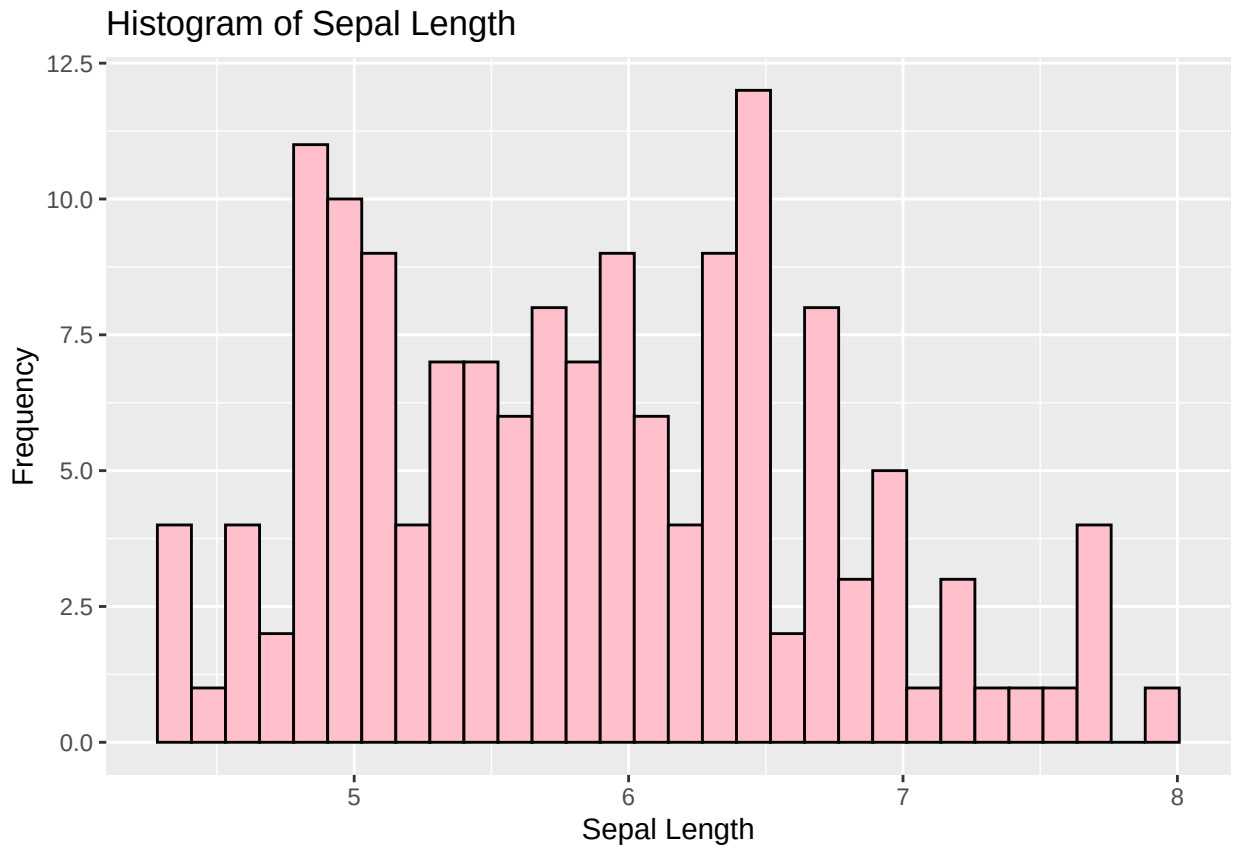
```r
# Plot for log-transformed sepal length
plot(iris$Log_sepal.length, main = "Log Transformed Sepal Length",
     xlab = "Index", ylab = "Log of Sepal Length",
     col = "red")
```



Plotting

Histogram for a Numerical Variable (Sepal Length):

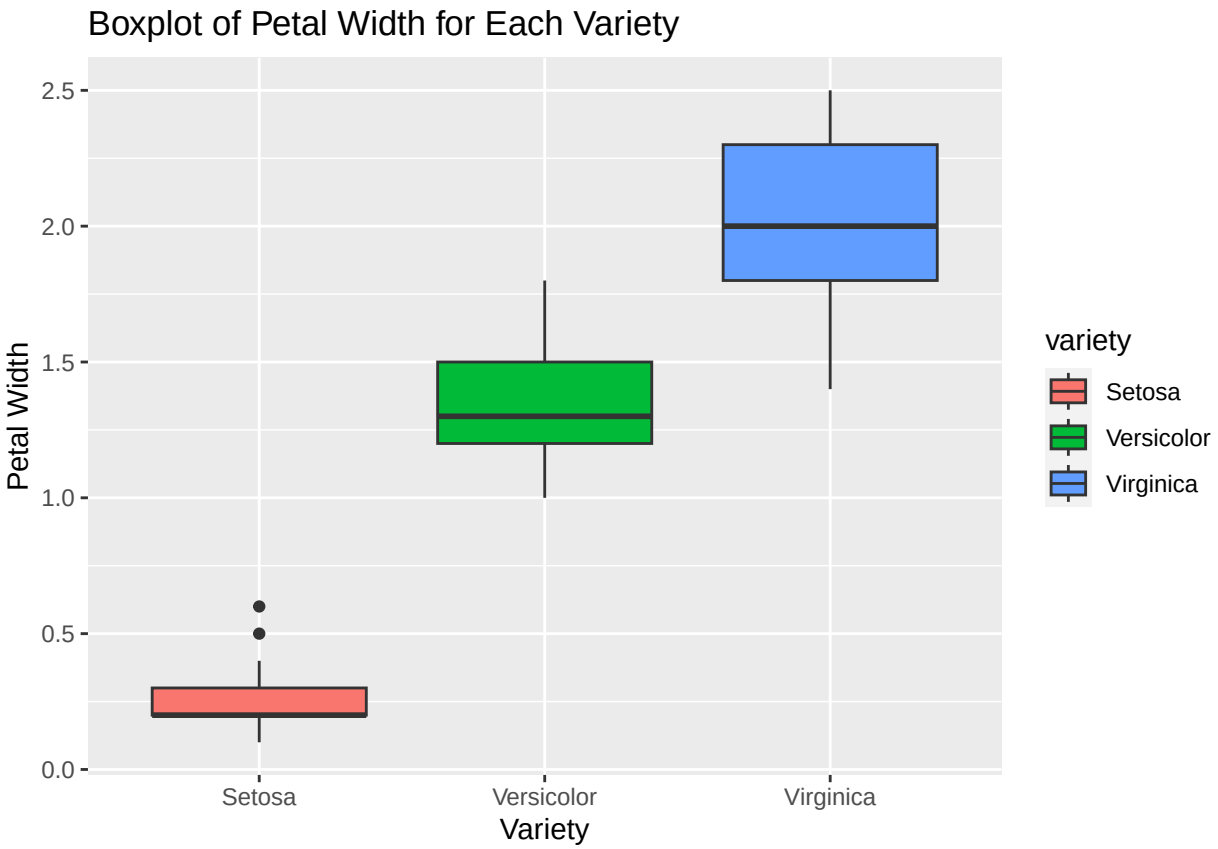
```r
library(ggplot2)
ggplot(iris, aes(x = sepal.length)) +
    geom_histogram(bins = 30, fill = "pink", color = "black") +
    ggtitle("Histogram of Sepal Length") +
    xlab("Sepal Length") +
    ylab("Frequency")
```



This plot shows the frequency distribution of the sepal length across the Iris dataset. The histogram reveals a roughly bimodal distribution, indicating two groups within the data where sepal lengths cluster around certain values.

Boxplot for a Numerical Variable by a Categorical Variable (Petal Width by Variety):

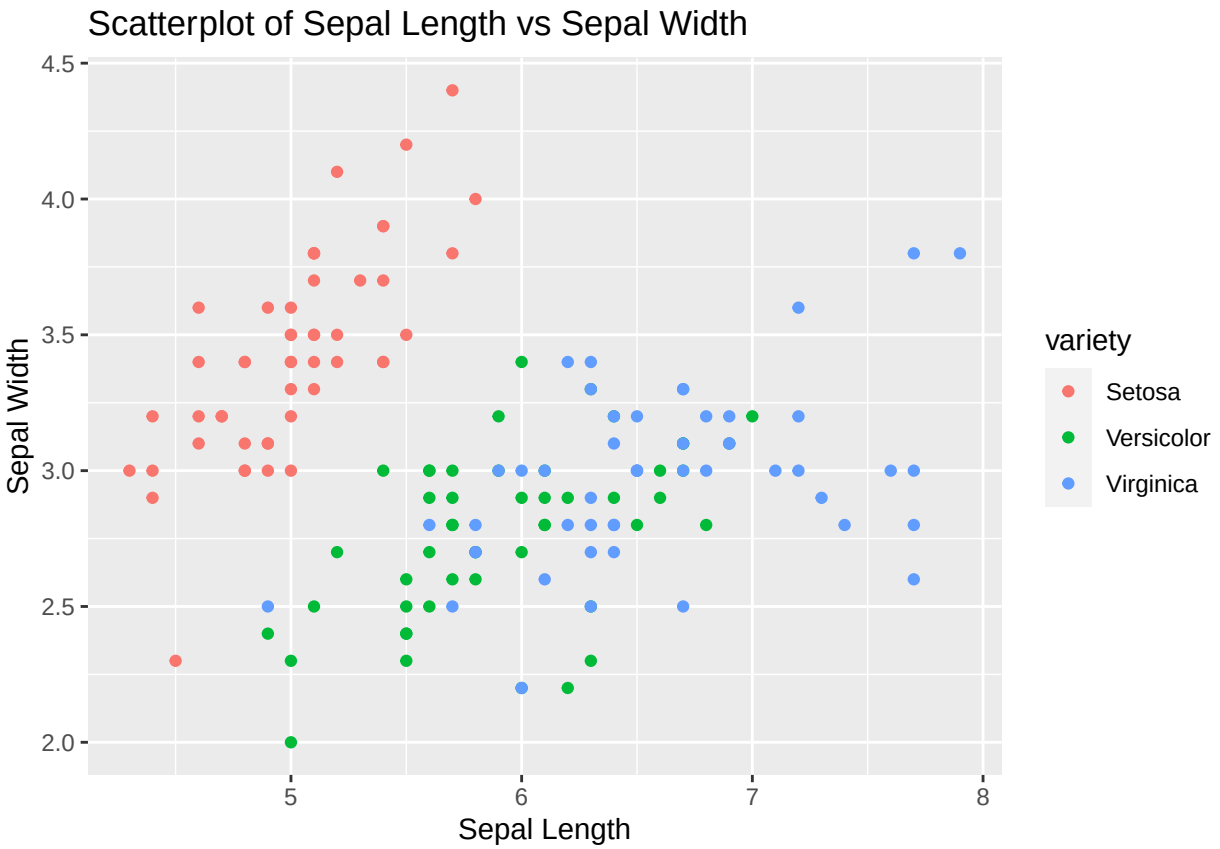
```
ggplot(iris, aes(x = variety, y = petal.width, fill = variety)) +
    geom_boxplot() +
    ggtitle("Boxplot of Petal Width for Each Variety") +
    xlab("Variety") +
    ylab("Petal Width")
```



The boxplot presents the spread and central tendency of petal width for each Iris variety. The median is marked by a line within the box, while the "whiskers" indicate variability outside the upper and lower quartiles. Outliers are represented as individual points.

Scatter plot for Two Numerical Variables (Sepal Length vs Sepal Width):

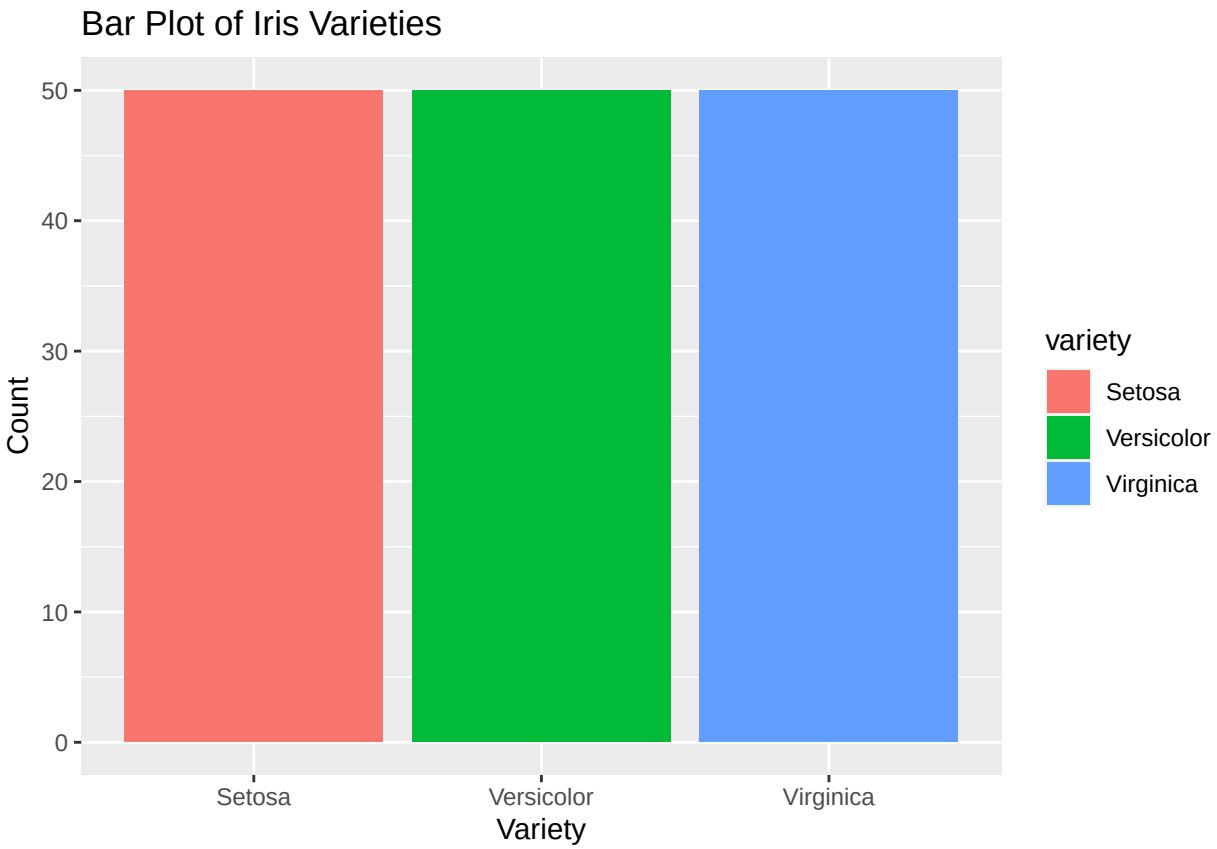
```
ggplot(iris, aes(x = sepal.length, y = sepal.width)) +
    geom_point(aes(color = variety)) +
    ggtitle("Scatterplot of Sepal Length vs Sepal Width") +
    xlab("Sepal Length") +
    ylab("Sepal Width")
```



This scatterplot compares sepal length and width, color-coded by Iris variety. Patterns in the data may suggest relationships between these dimensions, and the color-coding helps to discern if the relationship varies by variety.

Bar Plot for a Categorical Variable (Variety):

```
ggplot(iris, aes(x = variety, fill = variety)) +
    geom_bar() +
    ggtitle("Bar Plot of Iris Varieties") +
    xlab("Variety") +
    ylab("Count")
```



The bar plot displays the count of each Iris variety in the dataset. This is a visual representation of the categorical variable 'variety,' showing the frequency of each category within the data. Each variety's bar is color-coded for easier differentiation. From this plot we can see that there are equal number of counts for each variety