Lab Program - 3

R V Abhishek

2025-08-26

Basic Statistical Operations on Open-Source Datasets

Objective: This program emphasizes the application of statistical concepts on real-world datasets and visualization of the data.

```
# Load necessary
library(dplyr)  # For data manipulation
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(ggplot2)  # For visualization
library(moments)  # For skewness and kurtosis
library(palmerpenguins) # For Palmer Penguins dataset
```

```
Attaching package: 'palmerpenguins'

The following objects are masked from 'package:datasets':

    penguins, penguins_raw
```

```
data(iris)  # Load Iris dataset

data(penguins)  # Load Palmer Penguins
```

```
# Function to calculate mode
calc_mode <- function(x) {
  return (as.numeric (names (sort (table (x), decreasing = TRUE)) [1] ))
}
```

```r
# Perform Statistical Analysis on Iris Dataset
print("----- Iris Dataset Analysis -----")
```

```
[1] "----- Iris Dataset Analysis -----"
```

```r
# Mean
iris_mean <- sapply (iris[, 1:4], mean, na.rm = TRUE )
print(paste("Mean of Iris dataset : ", iris_mean))
```

```
[1] "Mean of Iris dataset :  5.84333333333333"
[2] "Mean of Iris dataset :  3.05733333333333"
[3] "Mean of Iris dataset :  3.758"
[4] "Mean of Iris dataset :  1.19933333333333"
```

```r
#Median
iris_median <- sapply(iris[, 1:4], median, na.rm = TRUE )
print(paste("Median of Iris dataset : ", iris_median))
```

```
[1] "Median of Iris dataset :  5.8"  "Median of Iris dataset :  3"
[3] "Median of Iris dataset :  4.35" "Median of Iris dataset :  1.3"
```

```r
#Mode
iris_mode <- sapply(iris[, 1:4], calc_mode )
print(paste("Mode of Iris dataset : ", iris_median))
```

```
[1] "Mode of Iris dataset :  5.8"  "Mode of Iris dataset :  3"
[3] "Mode of Iris dataset :  4.35" "Mode of Iris dataset :  1.3"
```

```r
#Variance
iris_variance <- sapply(iris[, 1:4], var, na.rm = TRUE )
print(paste("Variance of Iris dataset : ", iris_variance))
```

```
[1] "Variance of Iris dataset :  0.685693512304251"
[2] "Variance of Iris dataset :  0.189979418344519"
[3] "Variance of Iris dataset :  3.11627785234899"
[4] "Variance of Iris dataset :  0.581006263982103"
```

```r
#Standard Deviation
iris_sd <- sapply(iris[, 1:4], sd, na.rm = TRUE )
print(paste("Standard Deviation of Iris dataset : ", iris_sd))
```

```
[1] "Standard Deviation of Iris dataset :  0.828066127977863"
[2] "Standard Deviation of Iris dataset :  0.435866284936698"
[3] "Standard Deviation of Iris dataset :  1.76529823325947"
[4] "Standard Deviation of Iris dataset :  0.762237668960347"
```

```r
#Skewness
iris_skewness <- sapply(iris[, 1:4], skewness, na.rm = TRUE )
print(paste("Skewness of Iris dataset : ", iris_skewness))
```

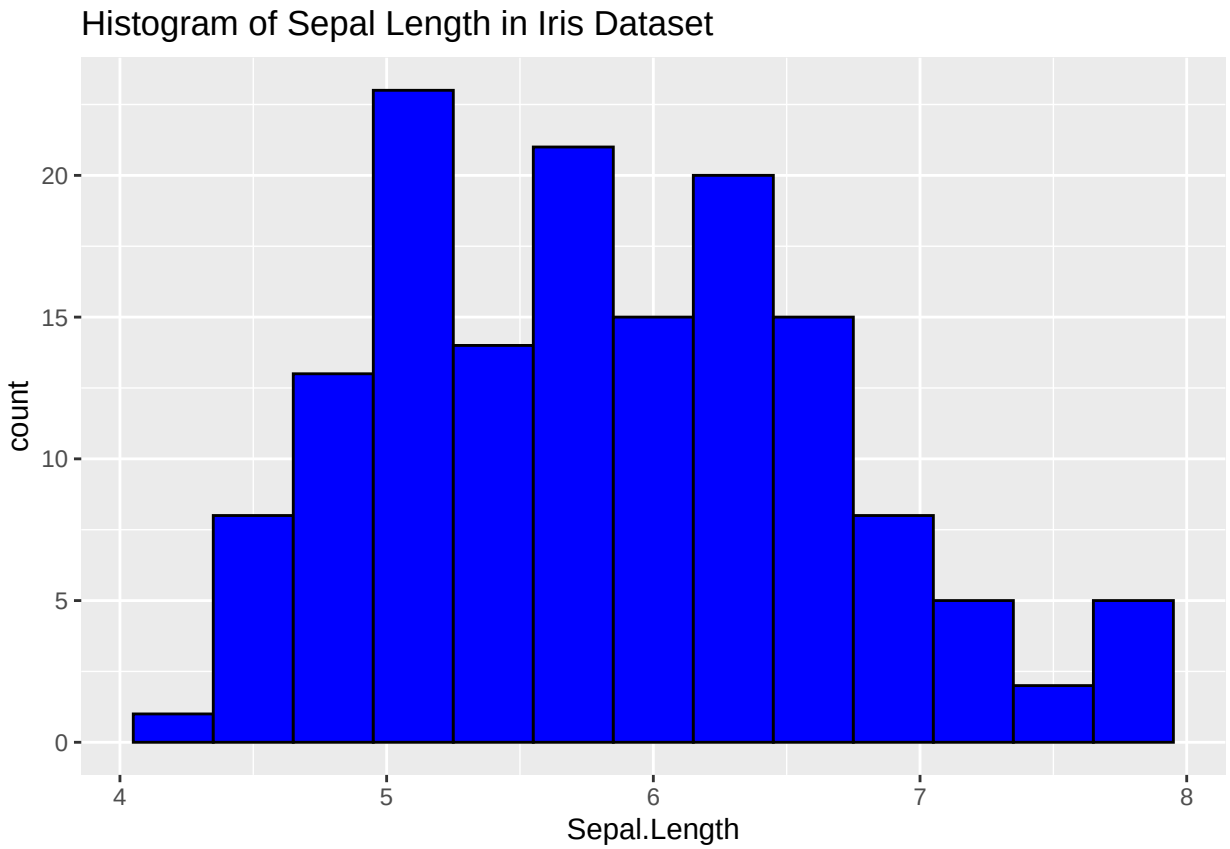
```
[1] "Skewness of Iris dataset :  0.311753058502296"
[2] "Skewness of Iris dataset :  0.315767106338938"
[3] "Skewness of Iris dataset :  -0.272127666456721"
[4] "Skewness of Iris dataset :  -0.101934206565599"
```

```r
# Hypothesis Testing (t-test) between Sepal.Length of Setosa and Versicolor
setosa <- subset(iris, Species == "setosa")$Sepal.Length
versicolor <- subset(iris, Species == "versicolor")$Sepal.Length
t_test <- t.test(setosa, versicolor)
print(t_test)
```

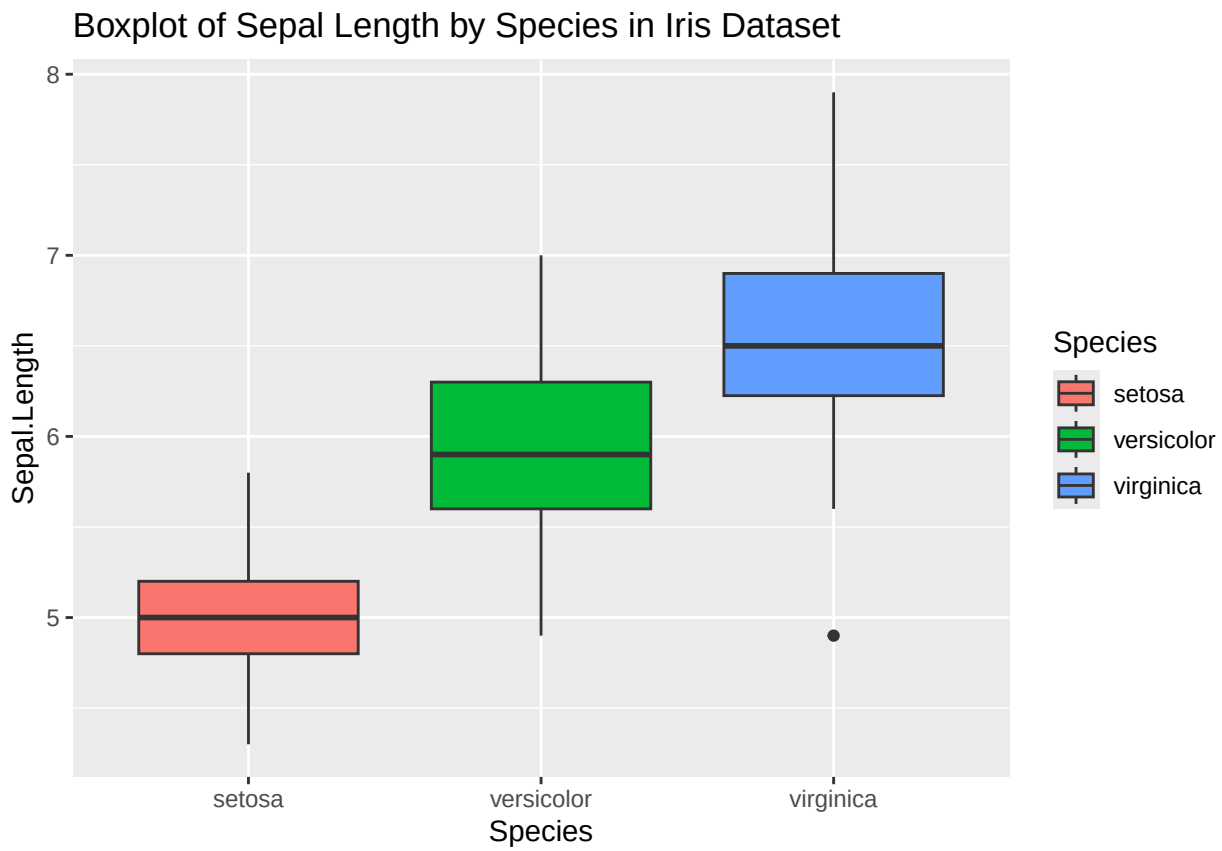
```
    Welch Two Sample t-test

data:  setosa and versicolor
t = -10.521, df = 86.538, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.1057074 -0.7542926
sample estimates:
mean of x mean of y
    5.006     5.936
```

```r
# Visualization of Iris Dataset
# Histogram for Sepal.Length
ggplot(iris, aes(x = Sepal.Length)) +
  geom_histogram(binwidth = 0.3, fill = "blue", color = "black") +
  ggtitle("Histogram of Sepal Length in Iris Dataset")
```




```
# Boxplot for Sepal.Length across Species
ggplot(iris, aes(x = Species, y = Sepal.Length, fill = Species)) +
  geom_boxplot() +
  ggtitle("Boxplot of Sepal Length by Species in Iris Dataset")
```




```
print("----- Palmer Penguins Dataset Analysis -----")
```

```
[1] "----- Palmer Penguins Dataset Analysis -----"
```

```
# Remove rows with missing values
penguins_clean <- na.omit(penguins)

# Mean
penguins_mean <- sapply(penguins_clean[, 3:6], mean, na.rm = TRUE)
print(paste("Mean of Palmer Penguins dataset:", penguins_mean))
```

```
[1] "Mean of Palmer Penguins dataset: 43.9927927927928"
[2] "Mean of Palmer Penguins dataset: 17.1648648648649"
[3] "Mean of Palmer Penguins dataset: 200.966966966967"
[4] "Mean of Palmer Penguins dataset: 4207.05705705706"
```

```
# Median
penguins_median <- sapply(penguins_clean[, 3:6], median, na.rm = TRUE)
print(paste("Median of Palmer Penguins dataset:", penguins_median))
```

```
[1] "Median of Palmer Penguins dataset: 44.5"
[2] "Median of Palmer Penguins dataset: 17.3"
[3] "Median of Palmer Penguins dataset: 197"
[4] "Median of Palmer Penguins dataset: 4050"
```

```r
# Mode
penguins_mode <- sapply(penguins_clean[, 3:6], calc_mode)
print(paste("Mode of Palmer Penguins dataset:", penguins_mode))
```

```
[1] "Mode of Palmer Penguins dataset: 41.1"
[2] "Mode of Palmer Penguins dataset: 17"
[3] "Mode of Palmer Penguins dataset: 190"
[4] "Mode of Palmer Penguins dataset: 3800"
```

```r
# Variance
penguins_variance <- sapply(penguins_clean[, 3:6], var, na.rm = TRUE)
print(paste("Variance of Palmer Penguins dataset:", penguins_variance))
```

```
[1] "Variance of Palmer Penguins dataset: 29.9063334418756"
[2] "Variance of Palmer Penguins dataset: 3.87788830999674"
[3] "Variance of Palmer Penguins dataset: 196.441676616375"
[4] "Variance of Palmer Penguins dataset: 648372.487698542"
```

```r
# Standard Deviation
penguins_sd <- sapply(penguins_clean[, 3:6], sd, na.rm = TRUE)
print(paste("Standard Deviation of Palmer Penguins dataset:", penguins_sd))
```

```
[1] "Standard Deviation of Palmer Penguins dataset: 5.46866834264756"
[2] "Standard Deviation of Palmer Penguins dataset: 1.9692354633199"
[3] "Standard Deviation of Palmer Penguins dataset: 14.0157652882879"
[4] "Standard Deviation of Palmer Penguins dataset: 805.215801942897"
```

```r
# Skewness
penguins_skewness <- sapply(penguins_clean[, 3:6], skewness, na.rm = TRUE)
print(paste("Skewness of Palmer Penguins dataset:", penguins_skewness))
```

```
[1] "Skewness of Palmer Penguins dataset: 0.0451359779776739"
[2] "Skewness of Palmer Penguins dataset: -0.149044996398334"
[3] "Skewness of Palmer Penguins dataset: 0.358523654622741"
[4] "Skewness of Palmer Penguins dataset: 0.470116171418382"
```

```r
# Kurtosis
penguins_kurtosis <- sapply(penguins_clean[, 3:6], kurtosis, na.rm = TRUE)
print(paste("Kurtosis of Palmer Penguins dataset:", penguins_kurtosis))
```

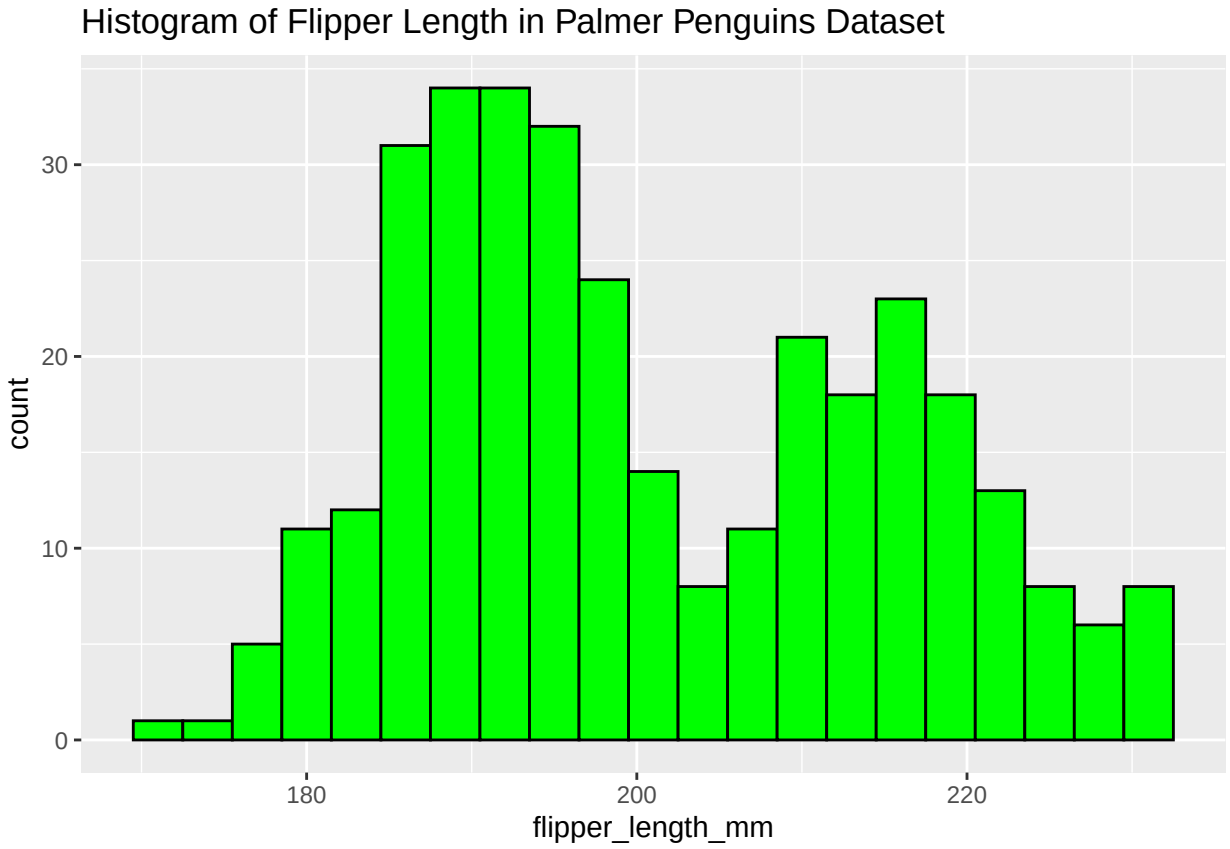
```
[1] "Kurtosis of Palmer Penguins dataset: 2.11182658541194"
[2] "Kurtosis of Palmer Penguins dataset: 2.10341274887238"
[3] "Kurtosis of Palmer Penguins dataset: 2.03516741259049"
[4] "Kurtosis of Palmer Penguins dataset: 2.25951411974012"
```

```r
# Hypothesis Testing (t-test) between flipper_length_mm of Adelie and Gentoo species
adelie <- subset(penguins_clean, species == "Adelie")$flipper_length_mm
gentoo <- subset(penguins_clean, species == "Gentoo")$flipper_length_mm
t_test_penguins <- t.test(adelie, gentoo)
print(t_test_penguins)
```

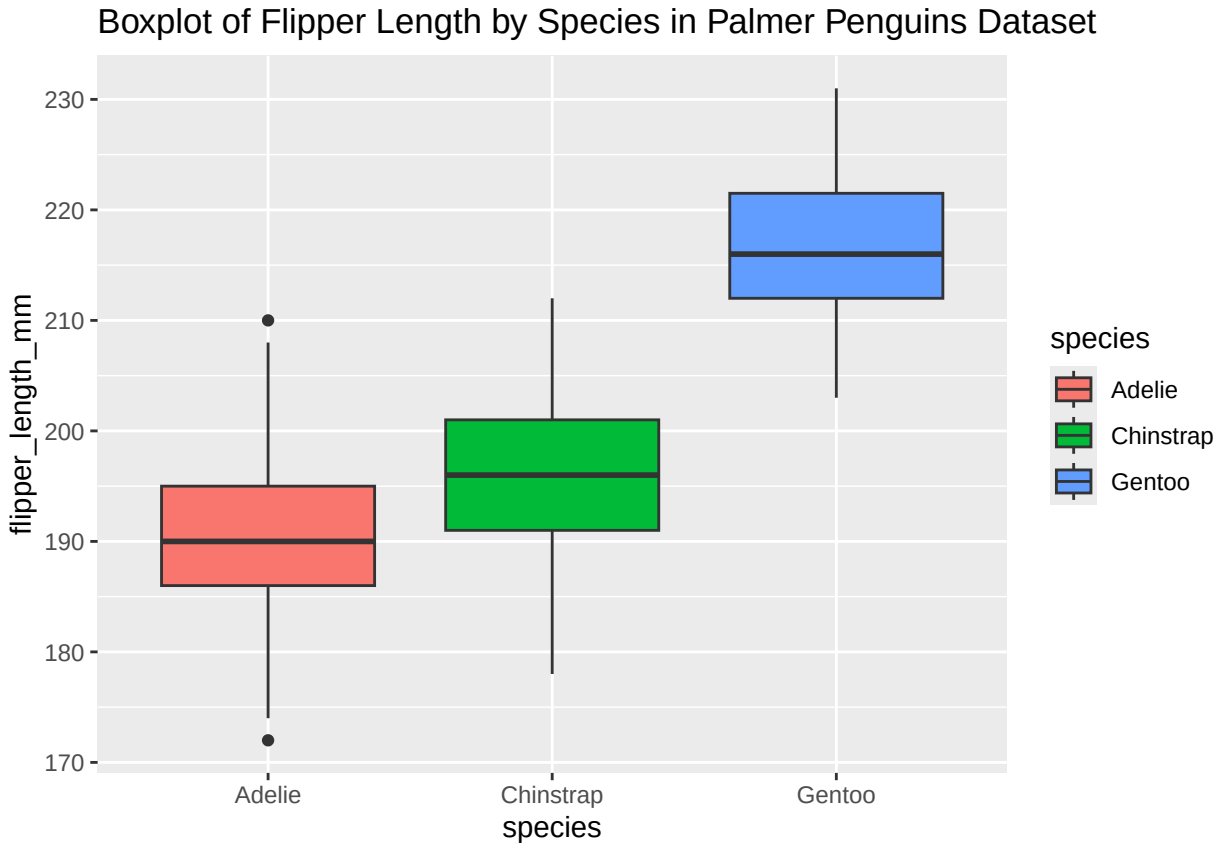
```
    Welch Two Sample t-test

data:  adelie and gentoo
t = -33.506, df = 251.35, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -28.72740 -25.53771
sample estimates:
mean of x mean of y
 190.1027  217.2353
```

```r
# Visualization of Palmer Penguins Dataset
# Histogram for flipper_length_mm
ggplot(penguins_clean, aes(x = flipper_length_mm)) +
  geom_histogram(binwidth = 3, fill = "green", color = "black") +
  ggtitle("Histogram of Flipper Length in Palmer Penguins Dataset")
```




```r
# Boxplot for flipper_length_mm across Species
ggplot(penguins_clean, aes(x = species, y = flipper_length_mm, fill = species)) +
  geom_boxplot() +
  ggtitle("Boxplot of Flipper Length by Species in Palmer Penguins Dataset")
```


Boxplot of Flipper Length by Species in Palmer Penguins Dataset
flipper_length_mm
230
220
210
200
190
180
170
Adelie
Chinstrap
Gentoo
species
species
Adelie
Chinstrap
Gentoo