Lab 10

R Programming Team Activity

Dataset Visualization and Collaboration using GitHub

---

1. Team Details

Name	Roll Number
Rudresh Baban Achari	2330
Sangana Ibrampurkar	2310
Shripad Chodankar	2317
Unnat Umarye	2303
Sanika Hoble	2309
Adhit Amonkar	2304
Rakshita Kubal	2319

Team Name: `Obsidian`
GitHub Repository Link:https://github.com/sangana03/Lab_10-obsidian

---

2. Objective

To use a standard dataset in R, generate multiple types of plots for data visualization, and collaboratively manage the project using GitHub for version control and teamwork.

---

3. Dataset Introduction

Dataset Used: *iris* (built-in R dataset)

Description:
The **iris** dataset contains measurements of 150 iris flowers from three different species — *setosa*, *versicolor*, and *virginica*.
It includes four numeric variables:

- Sepal.Length
- Sepal.Width
- Petal.Length
- Petal.Width

Purpose:
To explore relationships and patterns among the flower measurements using visual analysis.

---

4. Plots and Explanations

Plot 1: Histogram – Sepal Length Distribution (2330)

Code Snippet:

```
# Histogram for Sepal.Length
hist(iris$Sepal.Length,
     main = "Distribution of Sepal Length in Iris Dataset",
     xlab = "Sepal Length (cm)",
     ylab = "Frequency",
     col = "lightblue",
     border = "black")

# Enhanced version using ggplot2
library(ggplot2)
ggplot(iris, aes(x = Sepal.Length, fill = Species)) +
  geom_histogram(binwidth = 0.3, color = "black", alpha = 0.7) +
```

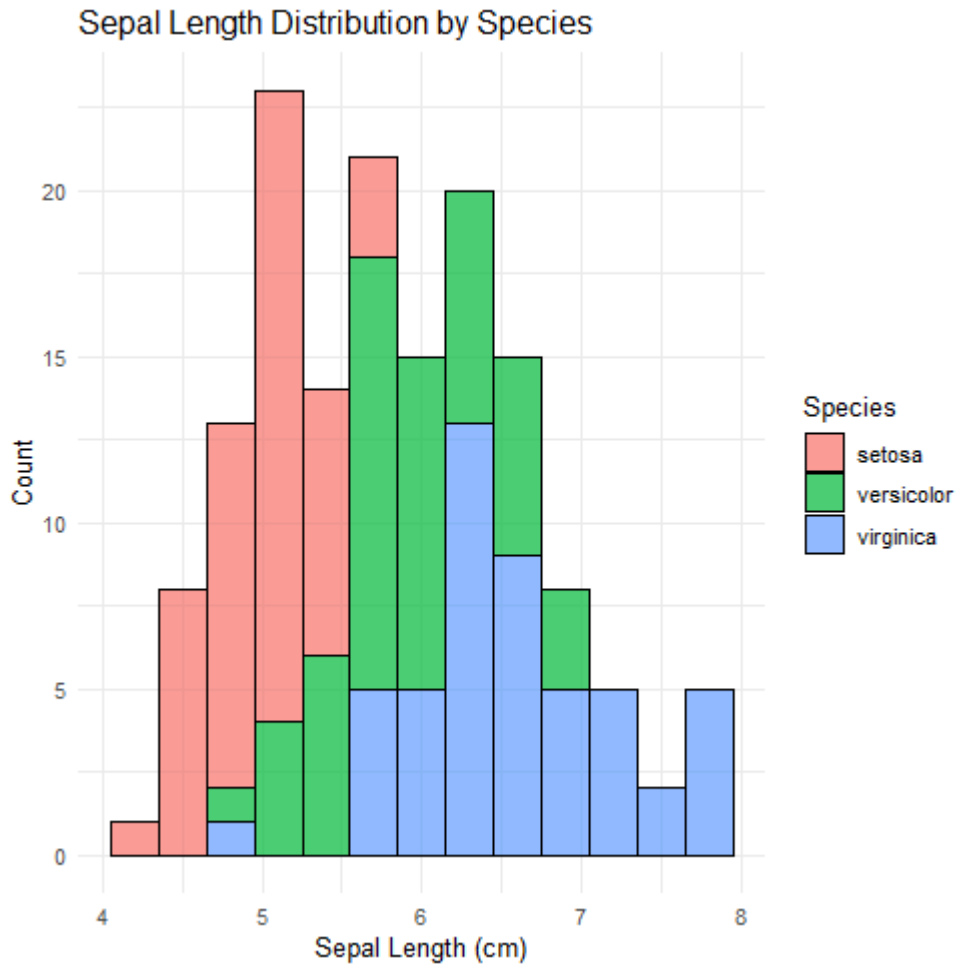
```
  labs(title = "Sepal Length Distribution by Species",
       x = "Sepal Length (cm)",
       y = "Count") +
  theme_minimal()
```

Explanation:
The histogram represents the **distribution of Sepal Length values** across the three species of the *Iris* flower — *setosa*, *versicolor*, and *virginica*. Each color corresponds to one species, making it easy to visually differentiate their ranges and frequencies. The x-axis shows the sepal length measured in centimeters, while the y-axis indicates the number of flower samples that fall within each length interval.

From the plot, it is evident that **Setosa** flowers (shown in red) generally have smaller sepals, with most values concentrated between 4.5 cm and 5.5 cm. **Versicolor** species (green) display intermediate sepal lengths, typically ranging from 5.5 cm to 6.5 cm. **Virginica** flowers (blue) possess longer sepals overall, extending up to 7.5 cm. This clear separation between the species demonstrates how sepal length can serve as a distinguishing feature in flower classification.

The histogram also highlights how the data is **not uniformly distributed** — each species occupies a distinct segment of the overall range, reflecting natural variations in their morphology. The use of color enhances interpretability, and adding the ggplot2 version makes the visualization more polished and publication-ready. Overall, this plot effectively illustrates inter-species differences and reinforces the usefulness of exploratory data visualization in understanding dataset patterns.

Screenshot:



Plot 2: Scatter Plot – Sepal vs Petal Length

Code Snippet:

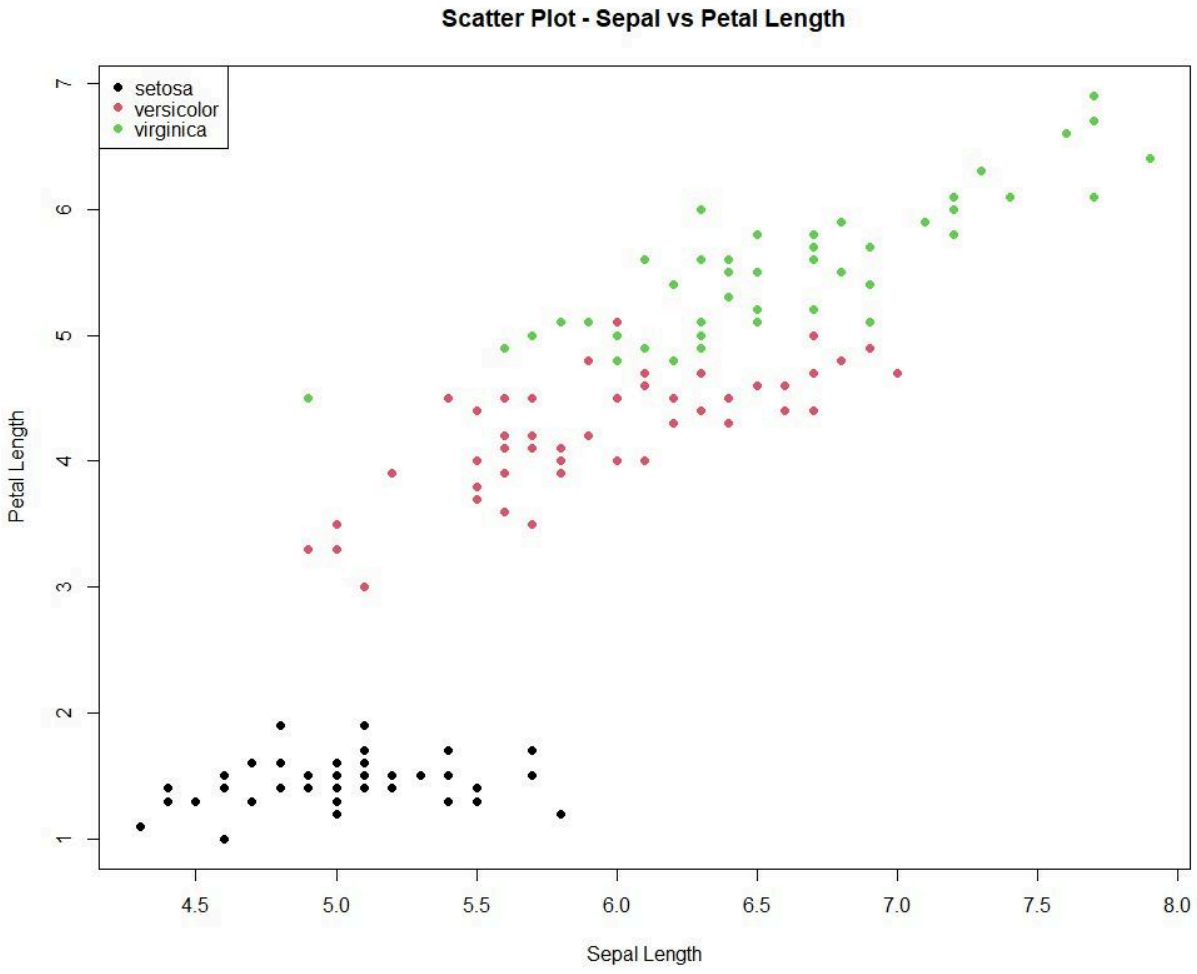
```
# Scatter Plot - Sepal vs Petal Length
plot(iris$Sepal.Length, iris$Petal.Length,
     main = "Scatter Plot - Sepal vs Petal Length",
     xlab = "Sepal Length",
     ylab = "Petal Length",
     col = iris$Species,
     pch = 19)
```

```
legend("topleft", legend = unique(iris$Species), col = 1:3, pch
= 19)
```

Explanation:
The provided R code generates a scatter plot using the famous `iris` dataset. In this plot, each point represents an individual flower sample, where the x-axis indicates the sepal length and the y-axis indicates the petal length. The points are colored according to the species of the flower—black for setosa, red for versicolor, and green for virginica—making it easy to visually distinguish between the three species. The `plot` function creates the scatter plot and labels the axes and the chart, while the `legend` function adds a key in the top left corner to explain the color coding for each species.

This graph illustrates the relationship between sepal length and petal length across different iris species. It shows that setosa samples (represented by black dots) tend to have both shorter sepals and petals and form a tight cluster at the bottom left. In contrast, versicolor and virginica samples (red and green) spread towards higher values on both axes, and virginica generally exhibits the largest measurements. There is a clear trend where, for all three species, as the sepal length increases, petal length also tends to increase. The clustering and separation among the colored groups highlight how these two measurements can help in distinguishing between the species within the `iris` dataset.

Screenshot:



Plot 3: Box Plot – Sepal Width by Species

Code Snippet:

```
# Box Plot to compare Petal Width across Species
boxplot(Petal.Width ~ Species,
        data = iris,
        main = "Box Plot of Petal Width by Species",
        xlab = "Species",
        ylab = "Petal Width (cm)",
        col = c("lightcoral", "lightgreen", "lightblue"),
        border = "darkblue",
```

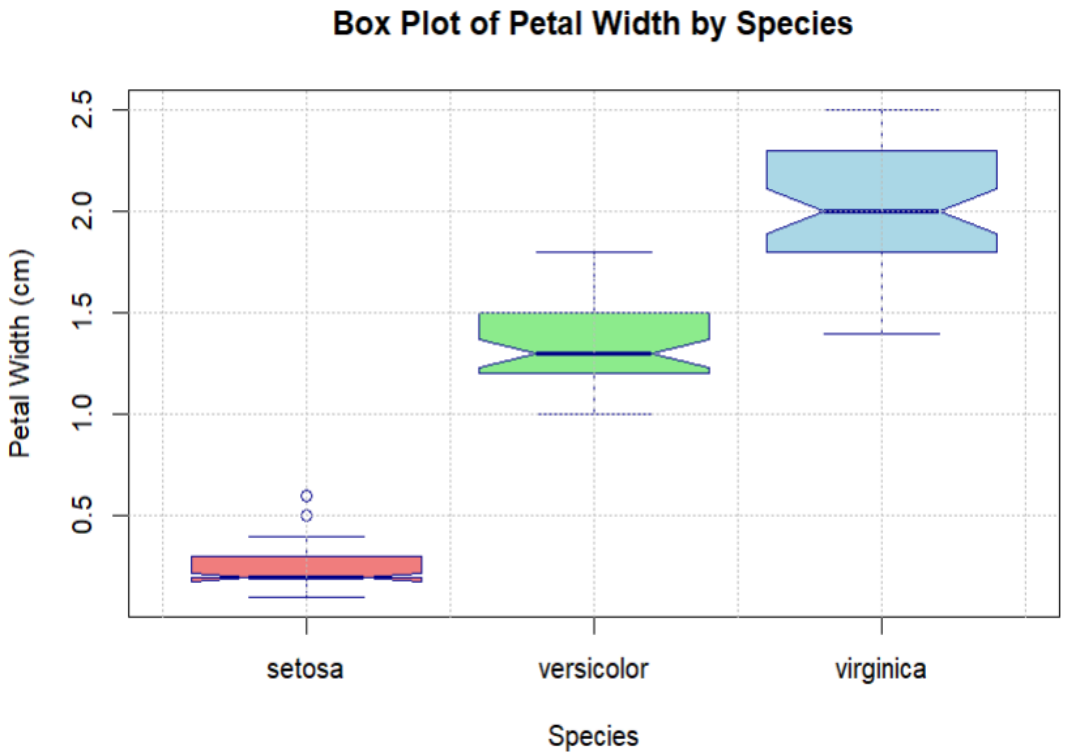
```
        notch = TRUE)  #Adds notches to show median confidence
interval

# Add grid lines for better readability
grid(nx = NULL, ny = NULL, col = "grey", lty = "dotted")
```

Explanation:-

This box plot visualizes the distribution of Petal Width for each species in the iris dataset. It helps identify the spread, median, and any potential outliers.The box plot illustrates the distribution of **Petal Width** for the three iris species — *setosa*, *versicolor*, and *virginica*. Each box represents the **interquartile range (IQR)**, which contains the middle 50% of the data for that species. The **horizontal line inside each box** shows the **median** petal width, representing the central or typical value. The **whiskers** extending from the box indicate the range of data values outside the IQR, showing how the measurements vary within each species. Any **points beyond the whiskers**, if present, are considered **outliers**, which highlight unusually high or low values compared to the rest of the dataset.

From the plot, it is clear that **Iris-setosa** has the **smallest petal widths** with **very little variation**, meaning most flowers of this species have consistently narrow petals. **Iris-versicolor** shows a **moderate spread** in petal width values, indicating more variation within this species. On the other hand, **Iris-virginica** displays the **largest petal widths** and the **widest range**, showing that this species tends to have significantly thicker petals. Overall, the box plot effectively demonstrates how petal width increases gradually from *setosa* to *versicolor* and finally to *virginica*, reflecting a clear pattern of growth among the three species.

Screenshot:



Plot 4: Line Plot – Comparing Average petal and sepal length of all 3 species

Code Snippet:

```
ggplot(iris, aes(x = Species, group = 1)) +


  stat_summary(aes(y = Petal.Length, color = "Petal Length"),
               fun = mean, geom = "line", lwd = 1) +
  stat_summary(aes(y = Petal.Length, color = "Petal Length"),
               fun = mean, geom = "point", size = 3) +


  stat_summary(aes(y = Sepal.Length, color = "Sepal Length"),
               fun = mean, geom = "line", lwd = 1) +
  stat_summary(aes(y = Sepal.Length, color = "Sepal Length"),
```

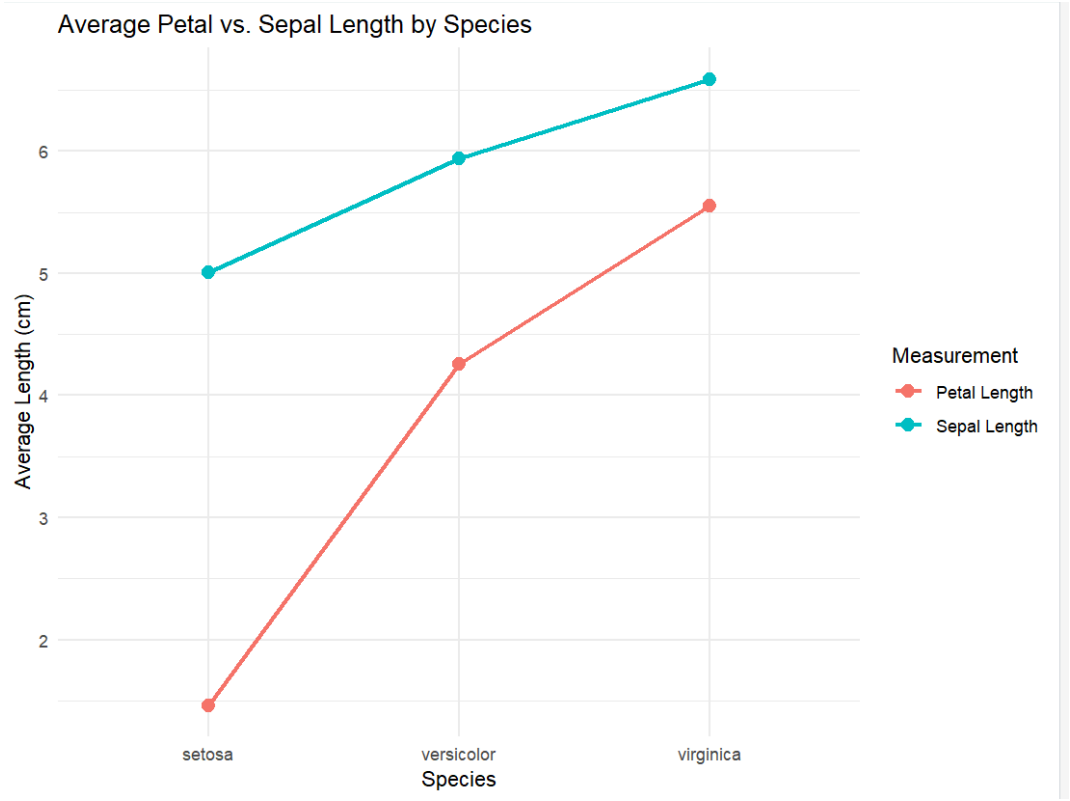
```
                fun = mean, geom = "point", size = 3) +


  labs(title = "Average Petal vs. Sepal Length by Species",
       y = "Average Length (cm)",
       color = "Measurement") +
  theme_minimal()
```

Explanation:

This R code uses `ggplot2` and the built-in `iris` dataset to compare the average Petal Length and Sepal Length for each species. It uses `stat_summary(fun = mean)` to calculate averages and plots them as both points and lines to clearly show differences across species. Labels and minimal styling are added for clarity.

The graph shows that both petal and sepal lengths increase from setosa → versicolor → virginica. Sepal length is consistently higher than petal length across all species, and virginica has the largest average measurements. This visualization helps quickly compare growth patterns among the three iris species.

Screenshot:



Plot 5: Bar Plot – Count of Species

Code Snippet:

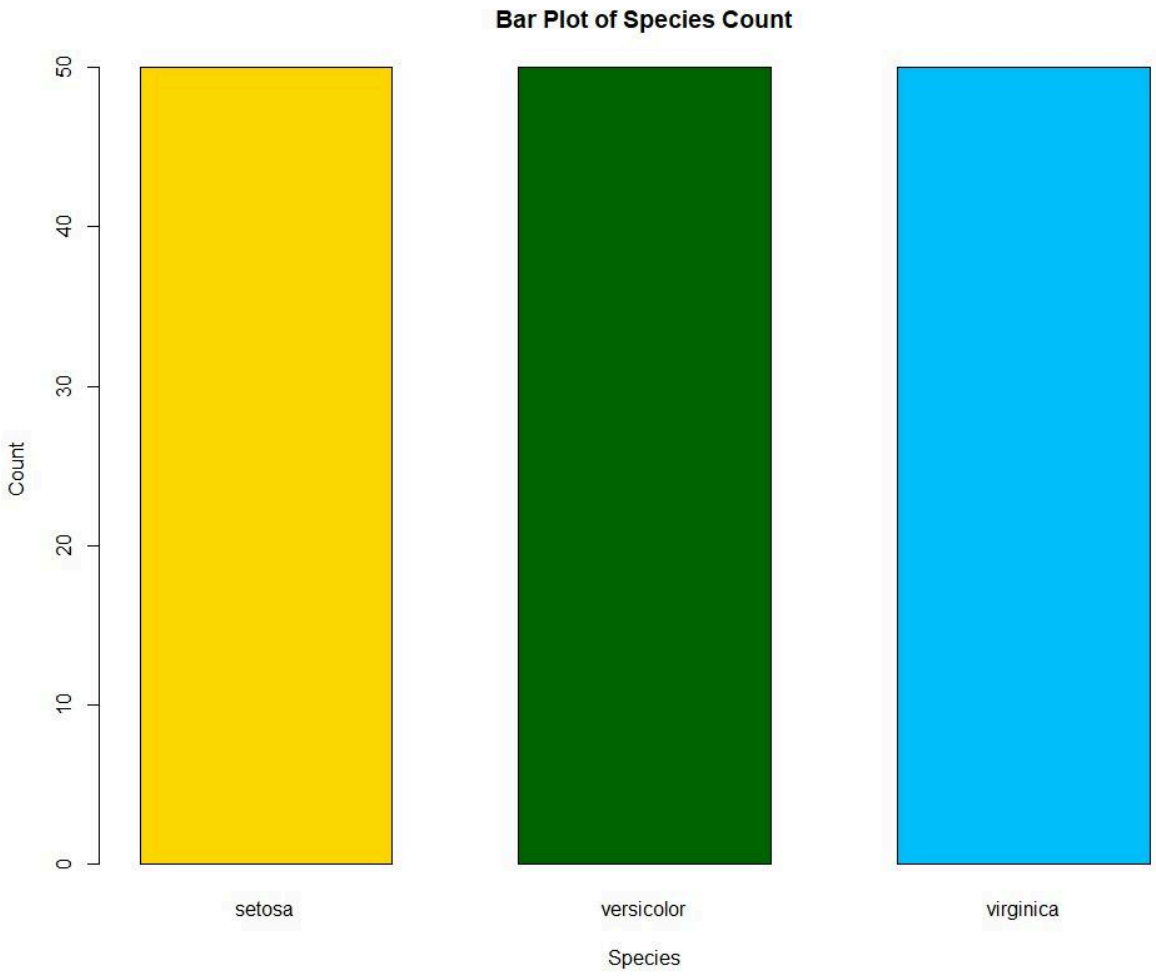
```
species_count <- table(iris$Species)

barplot(species_count,
        main = "Bar Plot of Species Count",
        xlab = "Species",
        ylab = "Count",
        col = c("gold", "darkgreen", "deepskyblue"),
        border = "black",
        space = 0.5)
```

Explanation:

This bar plot represents the number of flower samples for each species in the iris

dataset. The x-axis displays the three types of iris flowers — Setosa, Versicolor, and Virginica, while the y-axis shows the number of samples belonging to each species. Each bar's height corresponds to the count of flowers for that particular species. The bars are colored in gold, dark green, and deep sky blue, with a black border and slight spacing between them for a neat appearance. From the graph, it is clear that all three bars are of equal height, indicating that each species has 50 samples in the dataset. This shows that the iris dataset is balanced, meaning each flower species is equally represented.

Screenshot:



Plot 6: Pie Chart –

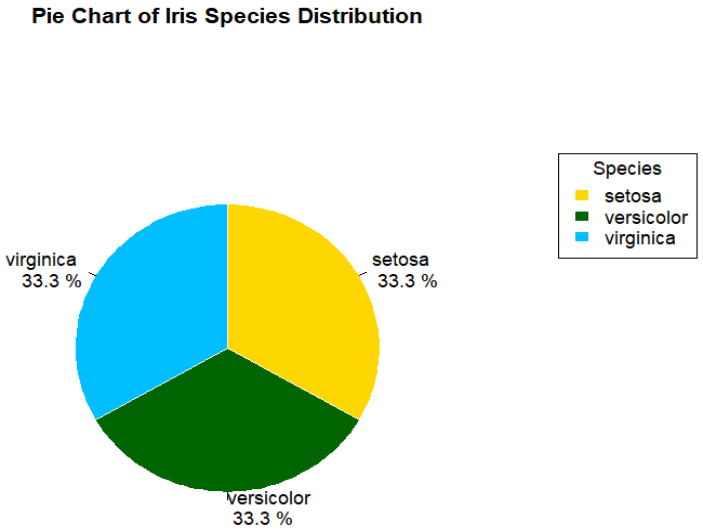
Code Snippet:

```
# 6. Pie Chart - Proportion of Each Species
species_count <- table(iris$Species)

# Create pie chart
pie(species_count,
    labels = paste(names(species_count), "\n", round(100 *
species_count / sum(species_count), 1), "%"),
    main = "Pie Chart of Iris Species Distribution",
    col = c("gold", "darkgreen", "deepskyblue"),
    border = "white",
    clockwise = TRUE)

# Add legend
legend("topright",
       legend = names(species_count),
       fill = c("gold", "darkgreen", "deepskyblue"),
       border = "white",
       title = "Species")
```

Explanation:

The pie chart illustrates the distribution of the three iris species—setosa, versicolor, and virginica—in the iris dataset. Each species occupies an equal portion of the chart, representing 33.3% of the total observations. This indicates that the dataset is perfectly balanced, containing 50 samples per species out of a total of 150. Such an even distribution ensures that comparisons among species for various features like sepal length, petal length, and petal width can be made without any sampling bias, providing a fair basis for statistical and visual analysis.

Screenshot:



Plot 7: Heat Map –

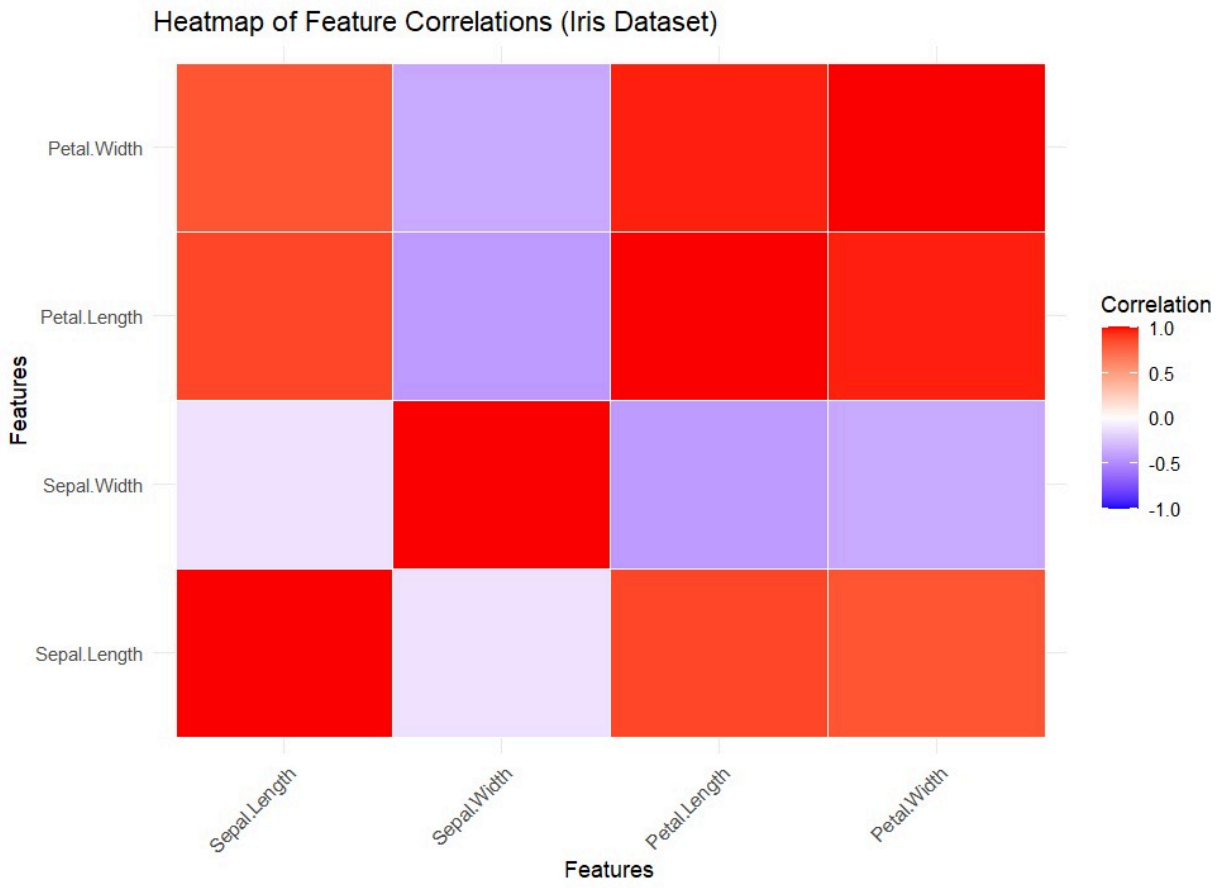
Code Snippet:

```
library(reshape2)

# Compute correlation matrix
cor_matrix <- cor(iris[, 1:4])

# Melt correlation matrix for ggplot2
melted_cor <- melt(cor_matrix)

# Create heatmap
ggplot(melted_cor, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid =
"white",
```

```
                      midpoint = 0, limit = c(-1, 1),
                      name = "Correlation") +
  theme_minimal() +
  labs(title = "Heatmap of Feature Correlations (Iris Dataset)",
       x = "Features", y = "Features") +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust
= 1))
```

Explanation:
The heatmap displays the correlation matrix of the four numeric features in the Iris dataset. Colors indicate the strength and direction of relationships: blue for strong negative, red for strong positive, and white for near zero correlation. This visualization highlights feature interdependencies, such as the strong positive correlation between petal length and petal width, assisting in understanding data structure, guiding feature selection.

Screenshot:



---

5. Collaboration on GitHub

- Repository initialized and managed by: **Sangana Ibrampurkar**
- Each member contributed through:
    - Uploading code updates
    - Adding explanations and screenshots
    - Merging commits to the main branch

- Instructor's GitHub account (**https://github.com/SwapnilFadte**) added for access.

Sample Commit Log Evidence:
(Insert screenshot or snippet of GitHub contributions)

---

6. Conclusion

Through this lab activity, we successfully:

- Used the **iris dataset** for exploratory visualization.
- Generated multiple plots using **R base plotting functions**.
- Practiced **team collaboration via GitHub**, including commits, merges, and version tracking.

This exercise enhanced our understanding of R's visualization capabilities and the importance of version control in collaborative data projects.

---

7. References

- R Documentation: `?iris`, `?plot`, `?hist`
- https://www.rdocumentation.org/
- GitHub Guides – Collaborating with GitHub