UCSC

SCS2211 - LABORATORY II
R Lab Practical Sheet - 11

Instructions

- **Do the Activities and save in a .ipynb file**
- **File name should be <Index number>.rmd (Eg: 2000000.ipynb) and upload to the given link.**
- **Any form of plagiarism or collusion is not allowed**

Histogram Representation in RStudio

A histogram is a graphical representation that organizes a group of data points into specified ranges (bins). It helps to visualize the distribution of a dataset. Below is a guide to creating histograms in RStudio with examples and corresponding code.

Example 1: Basic Histogram

```
# Sample Data
weights <- c(10, 20, 30, 40, 50, 60, 70, 80, 90, 100)
# Creating Histogram
hist(weights, main = "Weight Distribution", xlab = "Weight (Kg)", col = "blue")
```

Example 2: Customized Histogram

```
# Sample Data
weights <- c(5, 12, 19, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80)
# Creating Histogram with Custom Bins
hist(weights, breaks = 7, main = "Customized Histogram", xlab = "Weight (Kg)", col =
"green", border = "black")
```

Example 3: Histogram with Axis and Color Customization

```
# Sample Data
weights <- c(8, 15, 22, 29, 35, 41, 47, 53, 59, 65, 71, 77, 83, 89, 95)
# Creating Histogram with Axis Customization
hist(weights, breaks = seq(0, 100, by = 10),
     main = "Weight Distribution",
     xlab = "Weight (Kg)",
     col = "red",
     xlim = c(0, 100),
     ylim = c(0, 5),
     border = "black")
```

Example 4: Histogram with Specific Bar Width

```
# Sample Data
weights <- c(9, 14, 21, 26, 32, 37, 42, 48, 53, 58, 63, 69, 74, 79, 85)
# Creating Histogram with Defined Bar Width
hist(weights, breaks = seq(0, 90, by = 10),
     main = "Histogram with Defined Bar Width",
     xlab = "Weight (Kg)",
     col = "purple",
     border = "black")
```

Key Parameters in hist() function:

- breaks: Defines the number of bins.
- col: Specifies the color of bars.
- xlab & ylab: Labels the x-axis and y-axis.
- main: Sets the title of the histogram.
- xlim & ylim: Defines the x-axis and y-axis limits.
- border: Specifies the border color of bars.

Stem-and-Leaf Diagram Representation in RStudio

A stem-and-leaf diagram is a method of displaying numerical data that maintains the original values while organizing them in a structured way. It helps to quickly visualize the distribution of a dataset.

Example 1: Basic Stem-and-Leaf Plot

```
# Sample Data
weights <- c(10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 35, 40, 45, 50)
# Creating Stem-and-Leaf Plot
stem(weights)
```

Example 2: Stem-and-Leaf Plot with Modified Scale

```
# Sample Data
weights <- c(5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75)
# Creating Stem-and-Leaf Plot with Scale Adjustment
stem(weights, scale = 2)
```

Key Parameters in stem() function:

- scale: Adjusts the spacing of the plot.

- The first column represents the stem (leading digits), and the second column represents the leaves (trailing digits).

By using stem-and-leaf diagrams, you can quickly understand the distribution and shape of a dataset while preserving individual data values.

Generating a Random Numeric Dataset

Use rnorm(), runif(), or sample() for generating random numeric data.

Example: Generate 100 random numbers following a normal distribution

```
set.seed(123)  # Set seed for reproducibility
random_data <- rnorm(100, mean = 50, sd = 10)  # 100 values with mean 50 and SD 10
print(random_data)
```

Example: Generate 50 random numbers between 1 and 100 (Uniform Distribution)

```
random_uniform <- runif(50, min = 1, max = 100)
print(random_uniform)
```

Example: Generate a dataset with 20 random integers from 1 to 100

```
random_integers <- sample(1:100, 20, replace = TRUE)
print(random_integers)
```

Generating a Random Categorical Dataset

Use sample() to create categorical values.

Example: Generate a dataset with 30 random Gender values

```
gender <- sample(c("Male", "Female"), 30, replace = TRUE)
print(gender)
```

Boxplot Creation and Analysis in RStudio

A boxplot is a graphical representation of the distribution of a dataset that shows the median, quartiles, outliers, and skewness. It helps in identifying data spread and detecting anomalies.

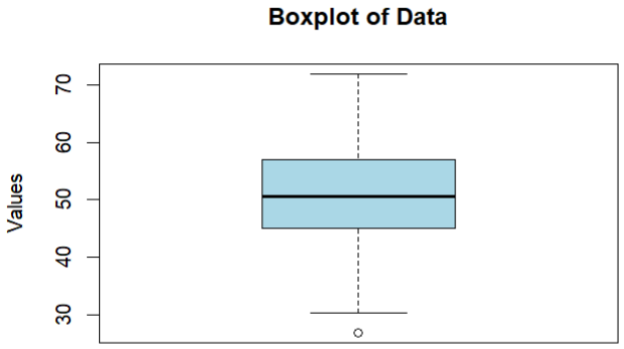
Identifying the Median and Five-Number Summary in a Boxplot in RStudio

A boxplot provides a graphical representation of the five-number summary, which consists of:

- Minimum – Smallest value in the dataset (excluding outliers)
- First Quartile (Q1) – 25th percentile (lower quartile)
- Median (Q2) – Middle value (50th percentile)
- Third Quartile (Q3) – 75th percentile (upper quartile)
- Maximum – Largest value in the dataset (excluding outliers)

Generate a Boxplot and Find the Median

```
# Sample Data
set.seed(123)
data <- rnorm(100, mean = 50, sd = 10)
# Creating Boxplot
boxplot(data, main = "Boxplot of Data", ylab = "Values", col = "lightblue")
# Finding the Median
median_value <- median(data)
print(paste("Median:", median_value))
```



- The thick horizontal line inside the box represents the median.
- The box itself represents the interquartile range (IQR) from Q1 to Q3.
- The whiskers extend to the minimum and maximum values (excluding outliers).

Identify Outliers

Outliers in a boxplot are values outside the whiskers, calculated as:

Lower Bound = Q1 - 1.5 * IQR

Upper Bound = Q3 + 1.5 * IQR

You can extract outliers using:

```
outliers <- boxplot.stats(data)$out
print(outliers)
```

This will list any extreme values outside the normal range.

Identify Skewness

- If the median is centered inside the box ⟶ Symmetric distribution
- If the median is not centered in the box, the data is skewed.
- If the median is closer to Q1 (lower part of the box) ⟶ Right-skewed : If the upper whisker is longer(positive skew)
- If the median is closer to Q3 (upper part of the box) ⟶ Left-skewed : If the lower whisker is longer (negative skew)

To check skewness numerically:

```
install.packages("moments")
library(moments)
skewness(data)
```

Positive value ⟶ Right-skewed

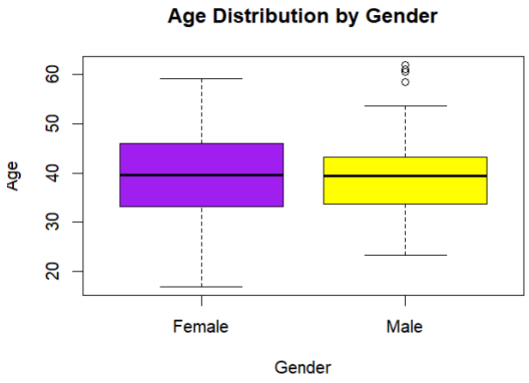
Negative value ⟶ Left-skewed

Near zero ⟶ Symmetric

Example 1: Boxplot of Age vs. Gender

Creating a Boxplot for Age grouped by Gender

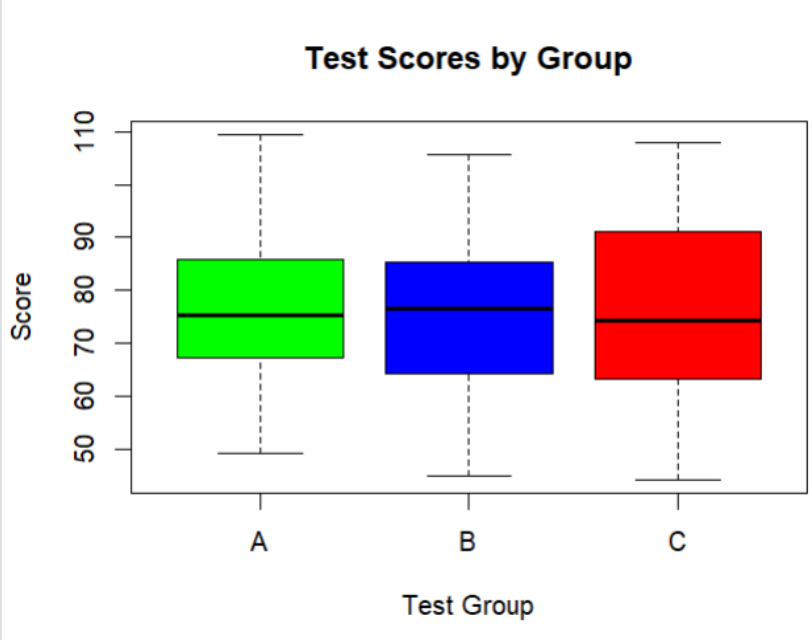
```
# Sample Data
set.seed(123)
data1 <- data.frame(
  Gender = sample(c("Male", "Female"), 100, replace = TRUE),
  Age = round(rnorm(100, mean = 40, sd = 10), 1)
)
# Creating Boxplot
boxplot(Age ~ Gender, data = data1,
        main = "Age Distribution by Gender",
        xlab = "Gender",
        ylab = "Age",
        col = c("purple", "yellow"))
```



Example 2: Boxplot of Scores vs. Group

Creating a Boxplot for Scores grouped by Test Group

```
# Sample Data
data2 <- data.frame(
  Group = sample(c("A", "B", "C"), 150, replace = TRUE),
  Score = round(rnorm(150, mean = 75, sd = 15), 1)
)
# Creating Boxplot
boxplot(Score ~ Group, data = data2,
        main = "Test Scores by Group",
        xlab = "Test Group",
        ylab = "Score",
        col = c("green", "blue", "red"))
```



Outlier and Skewness Analysis

- Check for individual points outside the whiskers (outliers).
- Skewness direction based on median position.

Example 3: Random Data Boxplot and Five-Number Summary

Generating a Random Dataset and Creating a Boxplot

```
# Generating Random Data
set.seed(123)
mydata1 <- rnorm(120, mean = 50, sd = 12)
# Creating Boxplot
boxplot(mydata1, main = "Random Data Distribution",
        ylab = "Values",
        col = "cyan")
```

Finding the Five-Number Summary

```
# Five-Number Summary
summary(mydata1)
# Calculate Range
range_val <- range(mydata1)
range_val
```

Scatterplot Creation and Analysis in RStudio

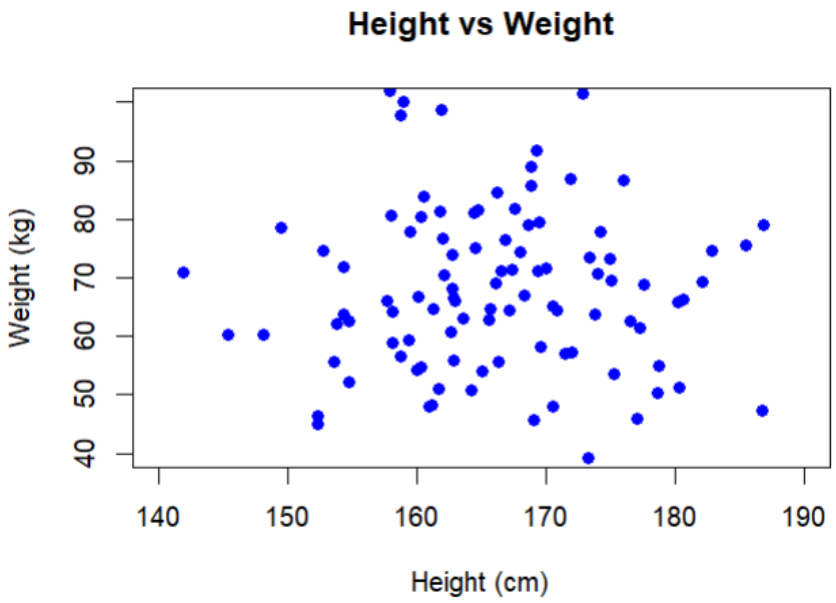
A scatterplot is used to observe relationships between two numerical variables. It helps to detect correlations, clusters, and outliers.

Creating a Scatterplot in RStudio

Example 1: Scatterplot of Height vs Weight

```
# Generate Sample Data
```

```
set.seed(123)
data <- data.frame(
  Height = rnorm(100, mean = 165, sd = 10),
  Weight = rnorm(100, mean = 70, sd = 15)
)
# Creating Scatterplot
plot(data$Height, data$Weight,
     main = "Height vs Weight",
     xlab = "Height (cm)",
     ylab = "Weight (kg)",
     xlim = c(140, 190),
     ylim = c(40, 100),
     col = "blue",
     pch = 16)
```

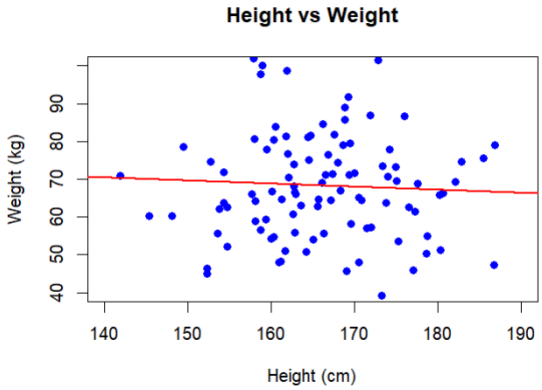


Observations:

- Each point represents an individual's height and weight.
- The spread of points indicates the strength of correlation.
- Color and markers can enhance visualization.

Adding Regression Line to Scatterplot

```
# Adding a Regression Line
model <- lm(Weight ~ Height, data = data)
abline(model, col = "red", lwd = 2)
```



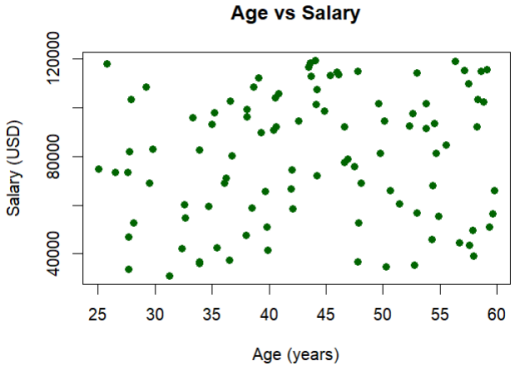
Interpretation:

- The red line represents the linear regression fit.
- The slope indicates how Weight changes with Height.

Customizing Scatterplots

Example 2: Scatterplot of Age vs Salary

```
# Generate Sample Data
set.seed(456)
data2 <- data.frame(
  Age = runif(100, min = 25, max = 60),
  Salary = runif(100, min = 30000, max = 120000)
)
# Creating Scatterplot
plot(data2$Age, data2$Salary,
     main = "Age vs Salary",
     xlab = "Age (years)",
     ylab = "Salary (USD)",
     col = "darkgreen",
     pch = 19)
```



Customization Options:

- Change colors: Use col = "red" or any color.
- Change markers: Use pch for different point shapes.
- Modify axis ranges: Use xlim and ylim to adjust axis limits.

Activity 01

Consider the following set of weights of certain parcels in (Kg)
14,22,33,45,56,23,12,56,45,34,23,11,17,3,5,23,34,38,54,6,7,24,48,46

- Create a histogram using above set of weights
- Name X axis as weight
- Use column color - yellow
- For the x axis range use 0-70
- For the y axis range use 0-10
- Width of the bar 5

Activity 02

Consider the following table

Type	No of Likes
Comedy	40
Action	50
Romance	60
Drama	10
SiFi	30

1. Represent the above table in a pie chart
    a. Using the Rainbow color palette
    b. Title of the pie chart - “Favorite type of Movie”
2. Represent the above piechart, where it shows the percentage except the Types labels.
3. Create a bar chart for the above table.
    a. Name the y axis as “No of Likes”, x axis as “Movie Type”
    b. Title - “Favourite type of Movie”
    c. Bar color - red, outline - Yellow

Activity 03

1. Take a random data set from R as mydata
2. Create a stem and leaf plot

Activity 04

Dataset 1 :https://www.kaggle.com/fedesoriano/heart-failure-prediction
Dataset 2 :https://www.kaggle.com/datasnaek/chess

1. Download the Above given Data sets
2. Consider the Dataset 1
    a. Create Box Plots graph for the relation between the MaxHR and Sex.(Taking sex for X axis and MaxHR for Y axis)
    b. 3. Name the X axis as “Sex”, Y axis as “Maximum Heart Rate” and name the graph as Heart Rates
    c. Use the colors purple and yellow
    d. Consider each boxplot, Are there any outliers in the plot, If present then in which boxplot
    e. Consider each boxplot, Is there any skewness, If then How the plot is skewed
3. Consider the Dataset 2
    a. 7. Create Boxplots graph for the relation between the Winner and turns.(Taking Winner For X axis and turns for Y axis)
    b. Name the X axis as “Winner of the game”, Y axis as “No of turns” and name the graph as “Chess game summary”
    c. Use the colors green, blue and red
    d. Consider each boxplot, Are there any outliers in the plot, If present then in which boxplot
    e. Consider each boxplot, Is there any skewness, If then How the plot is skewed
    f. Take a random data set from R as mydata1
    g. Create a boxplot graph for the above dataset in 14.
    h. Find the five number summary (Minimum, Maximum, First Quartile, Third Quartile, and median), Range, Skewness.
    i. Are there any outliers in the plot?

Activity 05

Use the Dataset “USArrests” in R and Draw a Scatterplot.

a. X axis - Murder
b. Y axis - Assault
c. X axis name - Murders
d. Y axis name - Assaults
e. Draw X axis from 8.0-14.0 and y axis from 150-300
f. Title - USA arrest rates