PRACTICAL 7 (Correlation)

Correlation is a statistical measure that describes the strength and direction of the relationship between two variables. It quantifies how changes in one variable are associated with changes in another. The correlation coefficient, denoted as **r**, ranges between -1 and 1.

- **r = 1**: Perfect positive correlation, meaning as one variable increases, the other also increases proportionally.
- **r = -1**: Perfect negative correlation, meaning as one variable increases, the other decreases proportionally.
- **r = 0**: No correlation, meaning there is no linear relationship between the variables.

Types of Correlation Coefficients

1. **Pearson Correlation Coefficient (r):**
   - Measures the linear relationship between two continuous variables.
   - Assumes that the data is normally distributed and there is a linear relationship between the variables.
   - Values close to 1 or -1 indicate a strong correlation, while values close to 0 indicate a weak correlation.
2. **Spearman Rank Correlation:**
   - Non-parametric measure of correlation that assesses how well the relationship between two variables can be described using a monotonic function.
   - Useful when the data does not meet the assumptions of the Pearson correlation (e.g., when the data is ordinal or not normally distributed).
   - Calculated based on the ranks of the data rather than their raw values.

Data pre-processing

```
# Load Libraries
library(dplyr)
library(ggplot2)

# Example dataset
data("mtcars")
df <- mtcars
```

In case if the data doesn't follow normal distribution then we can apply log transformation to achieve normality which is the assumption of Pearson's correlation coefficient

```
df$log_hp <- log(df$hp)
```

```
# Checking normality of the transformed variable
ggplot(df, aes(x = log_hp)) + geom_histogram(bins = 30) + theme_minimal()
```

PRACTICAL 7 (Correlation)

Visualization Techniques for correlation

```
# Load necessary libraries
library(GGally)
library(corrplot)

# Correlation Matrix
cor_matrix <- cor(df_clean)

# Heatmap Visualization
corrplot(cor_matrix, method = "color", type = "lower", tl.col = "black", tl.srt = 45)

# Pairwise scatter plots with correlation coefficients
ggpairs(df_clean, lower = list(continuous = wrap("cor", size = 3)))
```

Practical Application

```
# Stock Market Data is available in this package
library(quantmod)

# Get historical data for Apple (AAPL) and Microsoft (MSFT)
getSymbols(c("AAPL", "MSFT"), src = "yahoo", from = "2022-01-01", to = "2022-12-31")
```

*The `getSymbols()` function from the `quantmod` package fetches **historical** data from a specified period—in this case, from January 1, 2022, to December 31, 2022. This data is based on past stock prices*

```
# Calculate daily returns
aapl_returns <- dailyReturn(Cl(AAPL))
msft_returns <- dailyReturn(Cl(MSFT))
```

dailyReturn(): This function calculates the daily returns of a stock, which is the percentage change in the closing price from one day to the next.
Cl(AAPL): Extracts the closing prices of Apple (AAPL) from the data retrieved.
Cl(MSFT): Extracts the closing prices of Microsoft (MSFT)

```
# Combine the returns into a single dataframe
stock_data <- merge(aapl_returns, msft_returns, all = FALSE)
colnames(stock_data) <- c("AAPL_Returns", "MSFT_Returns")
```

merge(): Combines the two sets of daily returns into a single dataframe. colnames(): Renames the columns in the merged dataframe to "AAPL_Returns" and "MSFT_Returns" for clarity.

```
# Perform correlation analysis
cor(stock_data)

# Plot the relationship between the returns
ggplot(stock_data, aes(x = AAPL_Returns, y = MSFT_Returns)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_minimal()
```

PRACTICAL 7 (Correlation)

Exercise

1. Two judges gave the following rank to a series of eight one act plays in drama competition. Examine the relationship between their judgments.
Judge A 8 7 6 3 2 1 5 4
Judge B 7 5 4 1 3 2 6 8
Write a R program for above problem.

2. Find Karl Pearson's coefficient of correlation for the following
X 62 64 65 69 70 71 72 74
Y 126 125 139 145 165 152 180 208
Write a R program for the above problem.

3. Understand and interpret the correlation between different variables in the following datasets.
   A. airquality
   B. faithful
   C. trees
   D. longley

4. Use the `quantmod` package to retrieve historical stock data for Tesla (TSLA) and Ford (F). Calculate the daily returns for both stocks for the year 2023. Analyze and interpret the correlation between these two stocks. What does the correlation tell you about how these stocks move in relation to each other?