Day 08

Topic Covered: Built-in Datasets in R-Studio

Summary:

Today's session focused on built-in datasets in R-Studio, which are preloaded collections of structured data available for practice and analysis. These datasets are widely used in data analytics, visualization, and machine learning to test methods and understand real-world data structures.

In R, these datasets behave like regular data frames, meaning we can apply all data frame functions to them. Exploring built-in datasets helps in understanding how data is organized, cleaned, and prepared for analysis. Popular datasets such as iris, mtcars, airquality, and AirPassengers are often used for classification, regression, and time-series forecasting tasks.

New Concepts Learned:

1. **Built-in Datasets**
   R provides more than 180 built-in datasets. Some datasets discussed were:

- iris: Contains measurements of different iris flower species, commonly used for classification.
- mtcars: Includes automobile data with variables like horsepower and mileage, useful for regression and clustering.
- airquality: Contains New York air quality measurements, used in environmental and time-series analysis.
- AirPassengers: A classic time-series dataset recording monthly airline passenger numbers from 1949 to 1960.

2. **Key Functions**
   Important functions used to explore built-in datasets:

- data(): Access built-in datasets
- data(package = .packages(all.available = TRUE)): List all available datasets

- dim(): View number of rows and columns
- names(): Display column names
- summary(): Show dataset summary statistics
- ?: View dataset documentation
- min(), max(), mean(), median(): Calculate descriptive statistics on dataset columns
- Mode calculation using table() and sorting manually

Activity:

- Explored built-in datasets such as iris and mtcars using data()
- Used summary(), dim(), and names() to understand dataset structure
- Calculated mean, median, and mode for selected columns in the iris dataset
- Listed all available datasets in R using data(package=...)

Challenges Faced:

Finding the mode manually without a built-in function was slightly confusing at first. Understanding how to read dataset documentation using the help command also required some practice.

Key Takeaway:

Built-in datasets are useful for learning data manipulation and statistical analysis in R. They allow beginners to practice data exploration and prepare for more advanced tasks such as model building, forecasting, and visualization.