Day 10

Topic Covered: Data Manipulation in R

Summary:

Today's session focused on data manipulation in R, which is an important step in data analytics for cleaning and preparing datasets. The dplyr library is commonly used for efficient data manipulation, including filtering, selecting, sorting, renaming, mutating, and handling missing values.

The session covered two main areas: data manipulation on data frames and data manipulation on datasets. These techniques help transform raw data into structured, meaningful information required for analysis.

New Concepts Learned:

1. **Data Manipulation on Data Frames**
   Using the example data frame:
   dataframe <- data.frame(st_id = c(1, 2, 3), St_name = c("john", "merry", "april"), age = c(20, 21, 23))

Key functions learned:

- filter(): Selects rows based on given conditions
  Example: filter(dataframe, age > 21)
- select(): Chooses specific columns
  Example: select(dataframe, St_name)
- arrange(): Sorts rows based on a column
  Example: arrange(dataframe, age)
- rename(): Renames columns
  Example: rename(dataframe, St_name = dist)
- mutate(): Creates or modifies new columns
  Example: mutate(dataframe, speed_n = speed^2)

2. **Handling Missing Values**
   Example data frame:
   employee <- data.frame(emp_id = c(1, 2, 4, 2, NA), emp_name = c("john","merry",NA,"henry","walter"), salary = c(NA,2000,5000,NA,2000), age = c(20,NA,NA,24,35))

Key functions learned:

- is.na(): Checks for missing values
  Example: missing_frame <- is.na(employee)
- fill(): Fills missing values in selected columns
  Example: fill(employee, emp_name, .direction = "updown")
  fill(employee, age, .direction = "downup")

Activity:

- Created sample data frames for students and employees
- Used filter, select, arrange, rename, and mutate functions to manipulate data
- Checked for missing values using is.na()
- Filled missing values using fill() with different directions

Challenges Faced:
Remembering the correct syntax for dplyr functions required attention. Choosing the appropriate method to fill missing values was challenging. Ensuring that mutate and rename operations did not overwrite important columns also needed careful handling.

Key Takeaway:
The dplyr package provides efficient and clear functions for data manipulation. Proper handling of missing values is important for accurate analysis. Learning these functions helps in cleaning, transforming, and preparing data for further analytics work.