# Data Science (R1UC527B)

## Experiment No. 5

Calculate Measures of Central Tendency and Dispersion for Various Columns and Use groupby() to Derive Aggregated Statistics

### Aim:

Calculate measures of central tendency (mean, median, mode) and dispersion (variance, standard deviation, range) for different columns in a dataset. Use groupby() to generate aggregated statistics by categories and uncover initial insights.

### Theory:

Measures of central tendency (mean, median, mode) summarize the center of the data, while measures of dispersion (variance, standard deviation, range) describe the spread. Group-wise statistics are often essential for exploratory data analysis (EDA) to understand how different segments of data behave. Pandas provides built-in methods for these operations, and groupby() helps in grouping the data based on categorical features.

### Key pandas functions:

* mean(), median(), mode() for central tendency
* var(), std() for dispersion
* groupby() to perform aggregation by groups
* agg() to apply multiple aggregation functions at once
* describe() for a quick summary of statistics

### Code:

import pandas as pd  
  
# Load dataset  
df = pd.read\_csv("sample\_data.csv")  
  
# Calculate measures of central tendency  
print("Mean of numeric columns:")  
print(df.mean(numeric\_only=True))  
  
print("\nMedian of numeric columns:")  
print(df.median(numeric\_only=True))  
  
print("\nMode of numeric columns:")  
print(df.mode(numeric\_only=True).iloc[0])  
  
# Calculate measures of dispersion  
print("\nVariance of numeric columns:")  
print(df.var(numeric\_only=True))  
  
print("\nStandard Deviation of numeric columns:")  
print(df.std(numeric\_only=True))  
  
# Group by a categorical column (e.g., 'Department')  
print("\nGrouped Statistics by Department:")  
grouped = df.groupby('Department').agg({  
 'Salary': ['mean', 'median', 'std'],  
 'Age': ['mean', 'std']  
})  
print(grouped)  
  
# Quick descriptive stats  
print("\nDescriptive statistics:")  
print(df.describe(include='all'))

### Input/Output (Sample):

Sample DataFrame:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Name | Department | Age | Salary |
| 1 | Alice | HR | 25 | 50000 |
| 2 | Bob | IT | 30 | 60000 |
| 3 | Charlie | IT | 35 | 70000 |
| 4 | David | HR | 28 | 55000 |
| 5 | Eva | Finance | 40 | 75000 |

Grouped Statistics (Sample):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Department | Salary (mean) | Salary (median) | Salary (std) | Age (mean) | Age (std) |
| Finance | 75000 | 75000 | NaN | 40 | NaN |
| HR | 52500 | 52500 | 3535.53 | 26.5 | 2.12 |
| IT | 65000 | 65000 | 7071.07 | 32.5 | 3.53 |