**21CSS303T DATA SCIENCE**

**Assignment - II**

1. Discuss the steps involved in data wrangling: cleaning, transforming, and merging data.
2. Describe how pivoting helps in reshaping datasets. Give an example.
3. You are analysing patient data from a health study. The data is spread across multiple files: patient demographics, test results, and treatment outcomes. Each dataset uses different keys and has inconsistencies in formats.

 Describe how you would **merge and combine** these datasets properly. Include cases of merging on indexes and concatenation.

 If two datasets have **overlapping information** (e.g., addresses updated differently), how would you **combine with overlap** while preserving the most accurate information?

1. Given the following Data Frame, handle missing values by filling them with the median of their respective columns.

*import pandas as pd*

*import numpy as np*

*data = {*

*'A': [1, 2, np.nan, 4],*

*'B': [5, np.nan, 7, 8],*

*'C': [9, 10, 11, np.nan]*

*}*

*df = pd.DataFrame(data)*

1. *import pandas as pd*

*import numpy as np*

*data = {*

*'Customer\_ID': ['C100', 'C101', 'C102', 'C103', 'C104'],*

*'Name': ['john doe', 'Alice\_Green', 'BOB SMITH', 'diya patel', 'Eva\_Lee'],*

*'Age': [25, 33, 28, 45, 22],*

*'Purchase\_Amount': [120, 250, 300, 50, 5000],*

*'City': ['New York', 'los angeles', 'Chicago', 'chicago', 'NEW YORK'],*

*'Rating': ['4.5', '3.2', '4.0', '2.5', '4.8']*

*}*

*df = pd.DataFrame(data)*

Write python code for the following

1. Convert all names to Title Case
2. Find mean and median *'Purchase\_Amount'*
3. Number of unique cities
4. Convert the 'Age' column into 3 bins: "Young (18-25)", "Adult (26-40)", "Senior (41+)"
5. *import pandas as pd*

*df1 = pd.DataFrame({*

*'ID': [1, 2, 3],*

*'Name': ['Alice', 'Bob', 'Charlie'],*

*'Dept': ['HR', 'Tech', 'Finance']*

*})*

*df2 = pd.DataFrame({*

*'ID': [2, 3, 4],*

*'Salary': [50000, 60000, 45000],*

*'Join\_Date': ['2020-01-15', '2019-05-20', '2021-11-10']*

*})*

* Concatenate *df1* and *df2* (*axis = 0 and axis = 1*)
* Merge *df1* and *df2* (*inner, left, right and outer*)
* Join *df1* and *df2*

**Date of Submission: 05.05.2025**