**2311cs020228-Day(21)**

**Part 1: Data Exploration**

1. Load the dataset and display the first few rows. What are the independent and dependent variables?
2. Check for missing values. How would you handle them if any exist?
3. Generate summary statistics (mean, median, standard deviation) for numerical variables.
4. Create a correlation heatmap and interpret the correlation between Construction Cost and other factors.

Code:

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

# Load the dataset (replace 'file\_path' with the actual path of your dataset)

data = pd.read\_csv('Civil\_Engineering\_Regression\_Dataset.csv')

# Display the first few rows of the dataset

print("First few rows of the dataset:")

print(data.head())

# Identify the dependent and independent variables (assuming 'Construction Cost' is the dependent variable)

dependent\_variable = 'Construction Cost'

independent\_variables = [col for col in data.columns if col != dependent\_variable]

print("\nIndependent variables:", independent\_variables)

print("Dependent variable:", dependent\_variable)

missing\_values = data.isnull().sum()

print("\nMissing values in each column:")

print(missing\_values)

# Handle missing values: Fill missing numerical values with the median

data.fillna(data.median(), inplace=True)

summary\_stats = data.describe()

print("\nSummary statistics:")

print(summary\_stats)

correlation\_matrix = data.corr()

plt.figure(figsize=(10, 8))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)

plt.title("Correlation Heatmap")

plt.show()

construction\_cost\_correlation = correlation\_matrix['Construction Cost']

print("\nCorrelation with Construction Cost:")

print(construction\_cost\_correlation)|