College Roll…………………………………………Student Name:……………………………………………………………

DSC3171-Lab Assignments – Day 2

You need to complete assignments in 1hr 45 mins in the Lab

Day 2: Part-I: Data Cleaning and Summary Statistics

Objective: Perform basic EDA with a dataset using pandas.
Task:

- Load two datasets in sequence (e.g., 1) Salary_Data.csv and 2) Used_Car_Data.csv).
- Check for null values and remove/fill them appropriately.
- Display summary statistics: mean, median, mode, standard deviation.
- Identify categorical and numerical columns.
- Save the cleaned data to new CSV files.

Day 2: Part-II: Outlier Detection and Handling

Objective: Detect and handle outliers in a dataset (use the above dataset).
Task:

- Use the IQR method or Z-score method to detect outliers.
- Visualize data using boxplots to show outliers.
- Handle outliers by removing or transforming them.
- Describe how the data distribution changes.

---- Hint on IQR method or Z-score method to detect outliers

Both the IQR (Interquartile Range) and Z-score methods are used for outlier detection, but they have different strengths and weaknesses. The IQR method is generally preferred for skewed or non-normal data because it's robust to outliers, while the Z-score method is more suitable for data that follows a normal distribution.

- IQR Method: **Implementation in Python:**

```
import numpy as np
    import pandas as pd
    # Sample data
    data = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
    df = pd.DataFrame(data, columns=['Value'])

    Q1 = df['Value'].quantile(0.25)
    Q3 = df['Value'].quantile(0.75)
    IQR = Q3 - Q1

    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    outliers_iqr = df[(df['Value'] < lower_bound) | (df['Value'] > upper_bound)]
    print("Outliers using IQR method:\n", outliers_iqr)
```

- Z-score Method: **Implementation in Python:**

```
from scipy.stats import zscore
    import pandas as pd

    # Sample data
    data = [10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
    df = pd.DataFrame(data, columns=['Value'])

    df['Z_score'] = zscore(df['Value'])
    outliers_zscore = df[abs(df['Z_score']) > 3] # Common threshold of 3
    print("Outliers using Z-score method:\n", outliers_zscore)
```