LAB – 4 AND LAB-5 ASSIGNMENT

AI &ML

NAME – RAVI TIWARI

SAP ID – 500124009

EMAIL – Ravi.124009@stu.upes.ac.in

BATCH – 50

LAB:4) Write a python program to import and export data using Pandas and show the details of the dataset like number of rows, columns, first five rows, size, number of missing values, sum, average, min and max values from the numerical columns.

Objective: The goal of this lab is to write a Python program that imports and exports data using the Pandas library. You will also analyse the dataset by displaying key details such as the number of rows, columns, size, missing values, and summary statistics for numerical columns.

Instructions:

1. Import the dataset using Pandas from a file (e.g., a CSV file).

2. Display the following details about the dataset:

   o Number of rows and columns.

   o The first five rows of the dataset.

   o Total size of the dataset (number of elements).

   o Number of missing values in each column.

   o Summary statistics of numerical columns (sum, average, minimum, and maximum values).

```
clang > AIML > 🐍 lab4.py > ...
 1    # Import necessary libraries
 2    import pandas as pd
 3
 4    # Read the dataset into a DataFrame (Modify the file path as per your dataset)
 5    file_path = 'your_dataset.csv'
 6    df = pd.read_csv(file_path)
 7
 8    # Display number of rows and columns
 9    print(f"Number of rows: {df.shape[0]}")
10    print(f"Number of columns: {df.shape[1]}")
11
12    # Display first five rows of the dataset
13    print("\nFirst five rows of the dataset:")
14    print(df.head())
15
16    # Display size (total number of elements in the dataset)
17    print(f"\nSize of the dataset (total number of elements): {df.size}")
18
19    # Display the number of missing values for each column
20    print("\nNumber of missing values in each column:")
21    print(df.isnull().sum())
```

```python
22
23    # Display sum, average, min, and max values of numerical columns
24    print("\nSummary statistics for numerical columns:")
25    print(df.describe())
26
27    # Export the data to a new file (e.g., a CSV file)
28    output_file = 'exported_dataset.csv'
29    df.to_csv(output_file, index=False)
30    print(f"\nDataset exported to {output_file}")
31
```

Lab 5:) Exploratory Data Analysis (EDA) of the Imported Dataset.

Objective:
In this lab, you will perform exploratory data analysis (EDA) on the dataset imported in Lab 4. The goal of EDA is to understand the dataset's structure, identify patterns, and detect any anomalies or outliers.

Task:

1. Write a Python program to perform the following EDA tasks:

   o Check and display the data types of each column.

   o Check for and display missing values in the dataset.

   o Display the statistical summary of numerical columns.

- Plot the distribution of numerical columns (using histograms).

- Create a boxplot to detect outliers in numerical columns.

- Generate a correlation matrix heatmap for numerical columns.

- Display value counts of categorical columns (if applicable).

- Display the number of unique values in each column.

```python
1    # Import necessary libraries
2    import pandas as pd
3    import matplotlib.pyplot as plt
4    import seaborn as sns
5
6    # Read the dataset into a DataFrame (Modify the file path as per your dataset)
7    file_path = 'your_dataset.csv'
8    df = pd.read_csv(file_path)
9
10   # 1. Check the data types of each column
11   print("\nData types of each column:")
12   print(df.dtypes)
13
14   # 2. Check for missing values in the dataset
15   print("\nNumber of missing values in each column:")
16   print(df.isnull().sum())
17
18   # 3. Statistical Summary of Numerical Data
19   print("\nStatistical summary of numerical columns:")
20   print(df.describe())
21
```

```python
22    # 4. Visualize the distribution of numerical columns (Histograms)
23    df.hist(figsize=(10, 8))
24    plt.suptitle("Distribution of Numerical Columns", fontsize=16)
25    plt.show()
26
27    # 5. Boxplot for outliers detection
28    plt.figure(figsize=(10, 6))
29    sns.boxplot(data=df.select_dtypes(include=['float64', 'int64']))
30    plt.title("Boxplot for Numerical Columns (Outliers Detection)")
31    plt.show()
32
33    # 6. Correlation Matrix (for numerical columns)
34    plt.figure(figsize=(10, 8))
35    corr = df.corr()
36    sns.heatmap(corr, annot=True, cmap='coolwarm', fmt=".2f")
37    plt.title("Correlation Heatmap")
38    plt.show()
39
40    # 7. Count of categorical values (if any categorical columns exist)
41    categorical_columns = df.select_dtypes(include=['object']).columns
42    if not categorical_columns.empty:
43        for col in categorical_columns:
```

```python
44            print(f"\nValue counts for {col}:")
45            print(df[col].value_counts())
46
47    # 8. Check for unique values in each column
48    print("\nUnique values in each column:")
49    print(df.nunique())
50
```