Ex 3 EDA-Data Cleaning

AIM:

To clean the dataset by handling missing values, removing duplicates, performing data type conversion, and normalizing data using standardization and min-max scaling.

1. Handlingmissing values: detection, filling, and dropping

```
import pandas as pd
import numpy as np

data = {
    'Name': ['Alice', 'Bob', 'Charlie', 'Alice', 'Eve', None],
    'Age': [25, np.nan, 30, 25, 45, 35],
    'Score': [85.0, 90.5, np.nan, 85.0, 70.0, 95.0],
    'Gender': ['F', 'M', 'M', 'F', 'F', 'M']
}

df = pd.DataFrame(data)
print("Original Dataset:\n", df)
```

```
Original Dataset:
      Name   Age  Score Gender
0    Alice  25.0   85.0      F
1      Bob   NaN   90.5      M
2  Charlie  30.0    NaN      M
3    Alice  25.0   85.0      F
4      Eve  45.0   70.0      F
5     None  35.0   95.0      M
```

```
print("\nMissing Values:\n", df.isnull().sum())

df['Age'] = df['Age'].fillna(df['Age'].mean())
df['Score'] = df['Score'].fillna(df['Score'].mean())

df['Name'] = df['Name'].fillna(df['Name'].mode()[0])

print("\nAfter Handling Missing Values:\n", df)
```

```
Missing Values:
 Name      1
Age        1
Score      1
Gender     0
dtype: int64

After Handling Missing Values:
      Name   Age  Score Gender
0    Alice  25.0   85.0      F
1      Bob  32.0   90.5      M
2  Charlie  30.0   85.1      M
3    Alice  25.0   85.0      F
4      Eve  45.0   70.0      F
5    Alice  35.0   95.0      M
```

2. Removing duplicates and unnecessary data

```
print("\nBefore Removing Duplicates:", df.shape)

df = df.drop_duplicates()

print("After Removing Duplicates:", df.shape)
```

```
Before Removing Duplicates: (6, 4)
After Removing Duplicates: (5, 4)
```

2. Data type conversion and ensuring consistency

```
[ ]  df['Gender'] = df['Gender'].astype('category')

     print("\nData Types After Conversion:\n", df.dtypes)
```

```
Data Types After Conversion:
 Name         object
Age          float64
Score        float64
Gender      category
dtype: object
```

3. Normalize data (e.g., standardization, min-max scaling).