Name: **Gargi Devendra Shintre**
Div: **D15C** Roll No. **69**

Practical No. 2

Aim: To perform Data Preprocessing using Python.

Theory:

In real-world datasets, raw data often contains missing values, duplicates, incorrect data types, categorical variables, and outliers. These issues need to be addressed before applying any machine learning or statistical models.

The following preprocessing techniques are applied in this experiment:

1. **Handling Missing Values:** Missing data can bias analysis. A common approach is replacing missing numerical values with the median of the column.
2. **Removing Duplicates:** Duplicate rows inflate data size and distort analysis, hence they should be removed.
3. **Encoding Categorical Variables:** Since machine learning models require numerical input, categorical values must be converted to numeric codes. Label Encoding is used here.
4. **Fixing Data Types:** Incorrect data types can cause computational errors. For example, converting "Salary" into float ensures accurate calculations.
5. **Handling Outliers:** Outliers such as unrealistic ages (e.g., >100) are replaced with the median age to maintain data consistency.

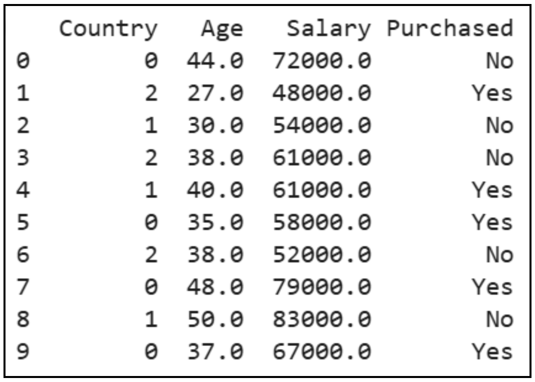
Procedure:

1. Import necessary libraries (pandas, numpy, sklearn.preprocessing).
2. Load the dataset (Data.csv) into a pandas DataFrame.
3. Replace missing values in the Age column with the median age.
4. Remove duplicate rows from the dataset.
5. Encode categorical values (e.g., Country) into numeric format using LabelEncoder.
6. Convert the Salary column to float datatype.
7. Detect and handle outliers in the Age column by replacing values greater than 100 with the median age.

Code:

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder

df = pd.read_csv(“Data.csv”)
print(df)

df.fillna({‘Age’ : df[‘Age’].median()}, inplace = True)

df.drop_duplicates(inplace=True)

le = LabelEncoder()
df[‘Country’] = le.fit_transform(df[‘Country’])

df[‘Salary’]=df[‘Salary’].astype(float)

median_age = df[‘Age’].median()
df.loc[df[‘Age’] > 100, ‘Age’] = median_age
```

Output:



```
   Country   Age   Salary Purchased
0        0  44.0  72000.0        No
1        2  27.0  48000.0       Yes
2        1  30.0  54000.0        No
3        2  38.0  61000.0        No
4        1  40.0  61000.0       Yes
5        0  35.0  58000.0       Yes
6        2  38.0  52000.0        No
7        0  48.0  79000.0       Yes
8        1  50.0  83000.0        No
9        0  37.0  67000.0       Yes
```

Conclusion:

The dataset was successfully cleaned and preprocessed. The applied steps ensured the data is consistent, free from duplicates, properly encoded, and ready for further analysis or machine learning tasks. Handling missing values, duplicates, datatype corrections, categorical encoding, and outlier treatment are essential preprocessing steps in any real-world data analysis pipeline.