Ex 2

5|8|27

TITANIC DATASET ANALYSIS AND PREPROCESSING USING PANDAS AND SIMPLE IMPUTER.

Aim:

To load titanic dataset from csv, handle missing values using simple imputer, analyze key passanger features.

procedure / Alg:

Step 1: Load titanic.csv into a data frame

Step 2: Explore dataset shap, info and summary statistics.

Step 3: Use simple Imputer to fill missing age.

Step 4: fill missing Cabin with "unknown" and embarked with mode

Step 5: visualize passanger class distribution with countplot.

Step 6: Identity top oldest.

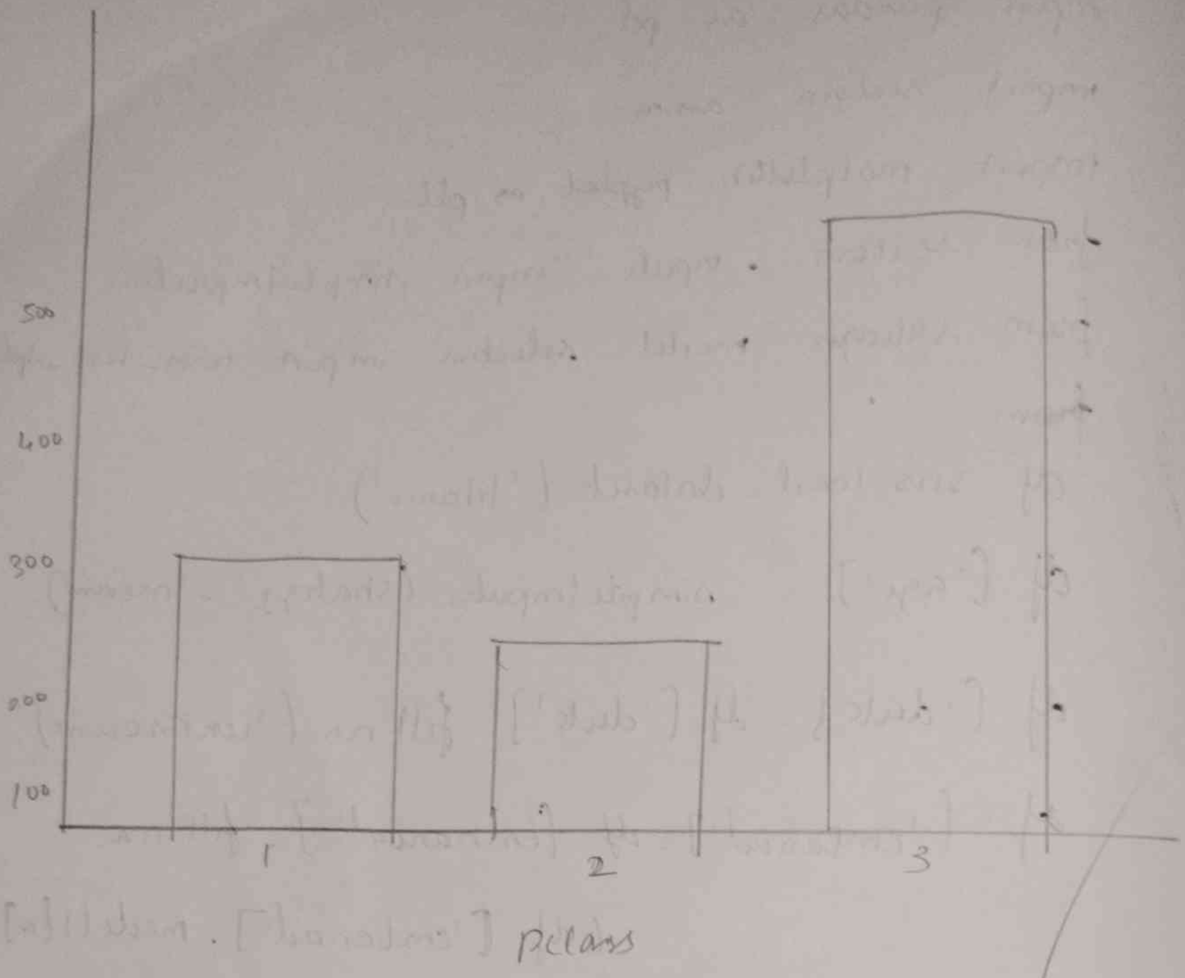
Step 7: Split training and testing splits.

program:

```
import pandas as pd.
import seaborn assns
import matplotlib . pyplot as plt.
from seaborn . inputs import simpleImputer.
from sklearn model - selection import train_test_split
~~from~~    .
 df = sns load_dataset ('titanic')
 df ['age'] = simpleImputer (strategy = 'mean')

 df ['deck] = df ['deck'] . fill na ('unknown')
 df ['embarked'] = df ['embarked'] . fill na
                (df ['embarked'] . mode()[0])
 sns . countplot (x = 'pclass', data = df) plt-
       title ('passanger class distribution')
     plt show().

print ("females who survied :"; df ((df.sex ==
 "female") & (df . survived == 1)] . index . list().
```


500
400
300
200
100
1
2
3
pclass

```
print ("male passangers who paid fare >
            100:", df [df.sex == 'male'] &
        (df.fare >100)].index.tolist())

print ("passengers embarked of 'c' and in
      class 2:", df ((df.embarked == 'c') &
      (df.pclass == 2)].
```

Result.

The program successfully identifies passengers with zero fare and efficiently splits the datasets into 80% training and 20% testing sets.