EXPERIMENT NO: 3B

Hotel Dataset Cleaning and Preprocessing

Aim:

To clean and preprocess the hotel dataset by removing duplicates, correcting invalid and inconsistent values, filling missing data, and standardizing text entries, making it ready for analysis or machine learning.

Algorithm:

1. **Load Data: Read the CSV file into a DataFrame.**
2. **Remove Duplicates: Drop duplicate rows and reset the index.**
3. **Drop Unnecessary Columns: Remove Age_Group.1.**
4. **Handle Invalid Values:**
    - **Replace negative CustomerID, Bill, and EstimatedSalary with NaN.**
    - **Replace NoOfPax values <1 or >20 with NaN.**
5. **Standardize Text Data:**
    - **Correct Hotel names (e.g., 'Ibys' → 'Ibis').**
    - **Normalize FoodPreference values ('Vegetarian'/'veg' → 'Veg', 'non-Veg' → 'Non-Veg').**
6. **Fill Missing Values:**
    - **EstimatedSalary → mean (rounded)**
    - **NoOfPax → median (rounded)**
    - **Rating(1-5) → median (rounded)**
7. **Output: Print the cleaned dataset.**

Program:

```
[3]: import numpy as np
     import pandas as pd
     df = pd.read_csv("C:/Users/vijay/Downloads/Hotel_Dataset.csv")
     df.drop_duplicates(inplace=True)
     df.reset_index(drop=True, inplace=True)
     df.drop(['Age_Group.1'], axis=1, inplace=True)
     df.loc[df.CustomerID < 0, 'CustomerID'] = np.nan
     df.loc[df.Bill < 0, 'Bill'] = np.nan
     df.loc[df.EstimatedSalary < 0, 'EstimatedSalary'] = np.nan
     df.loc[(df['NoOfPax'] < 1) | (df['NoOfPax'] > 20), 'NoOfPax'] = np.nan
     df.Hotel.replace(['Ibys'], 'Ibis', inplace=True)
     df.FoodPreference.replace(['Vegetarian', 'veg'], 'Veg', inplace=True)
     df.FoodPreference.replace(['non-Veg'], 'Non-Veg', inplace=True)
     df.EstimatedSalary.fillna(round(df.EstimatedSalary.mean()), inplace=True)
     df.NoOfPax.fillna(round(df.NoOfPax.median()), inplace=True)
     df['Rating(1-5)'].fillna(round(df['Rating(1-5)'].median()), inplace=True)
     print(df)
```

```
   CustomerID Age_Group  Rating(1-5)      Hotel FoodPreference    Bill  \
0         1.0     20-25            4       Ibis            Veg  1300.0
1         2.0     30-35            5  LemonTree        Non-Veg  2000.0
2         3.0     25-30            6    RedFox            Veg  1322.0
3         4.0     20-25           -1  LemonTree            Veg  1234.0
4         5.0       35+            3       Ibis            Veg   989.0
5         6.0       35+            3       Ibis        Non-Veg  1909.0
6         7.0       35+            4    RedFox            Veg  1000.0
7         8.0     20-25            7  LemonTree            Veg  2999.0
8         9.0     25-30            2       Ibis        Non-Veg  3456.0
9        10.0     30-35            5    RedFox        Non-Veg     NaN

   NoOfPax  EstimatedSalary
0      2.0          40000.0
1      3.0          59000.0
2      2.0          30000.0
3      2.0         120000.0
4      2.0          45000.0
5      2.0         122220.0
6      2.0          21122.0
7      2.0         345673.0
8      3.0          96755.0
9      4.0          87777.0
```

Result:

A cleaned and consistent hotel dataset with no duplicates, corrected text values, invalid numerical entries replaced, and missing values filled, ready for analysis or modeling.