EXP NO: 1	Data Preprocessing and Cleaning
DATE: 22/7/25	

Aim:

To preprocess the Titanic dataset by handling missing values, encoding categorical data, scaling numeric features, and visualizing relationships between attributes.

Program:

Step 1: Import Required Libraries

```
import pandas as pd import seaborn as sns import
matplotlib.pyplot as plt from sklearn.preprocessing import
LabelEncoder, StandardScaler Step 2: Load the Dataset

df = sns.load_dataset('titanic')

print("--- Initial DataFrame Head ---") print(df.head())

print("\n--- DataFrame Info ---") df.info()
```

Step 3: Handle Missing Values

```
# Fill missing 'age' values using forward and backward fill
df['age'] = df['age'].ffill().bfill()

#Add a new category for missing 'deck' values and fill them as 'Unknown'
df['deck'] = df['deck'].cat.add_categories('Unknown') df['deck'] =
df['deck'].fillna('Unknown') Step 4: Remove Duplicate Records
df.drop_duplicates(inplace=True)
```

Step 5: Encode Categorical Variables

```
# Encode 'sex' column to numeric values
le = LabelEncoder() df['sex'] =
le.fit_transform(df['sex'])
```

Step 6: Scale Numerical Columns

```
# Standardize the 'fare' column scaler
= StandardScaler()
df['fare'] = scaler.fit_transform(df[['fare']])
```

Step 7: Display Processed Data

```
print("\n--- DataFrame Head After Preprocessing ---")
print(df.head())
```

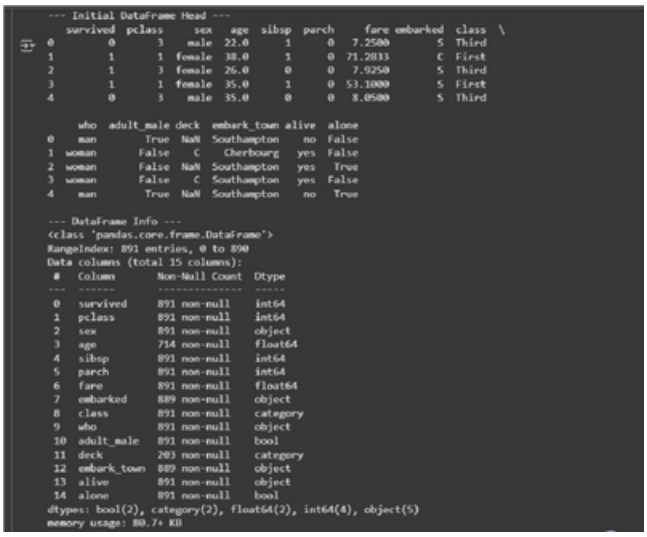
Step 8: Generate Pair Plot

```
pair_plot_features = ['pclass', 'sex', 'age', 'sibsp']
sns.pairplot(df[pair_plot_features])
plt.suptitle("Pair Plot of Selected Titanic Features", y=1.02)
plt.show()
```

Step 9: Generate Correlation Heatmap

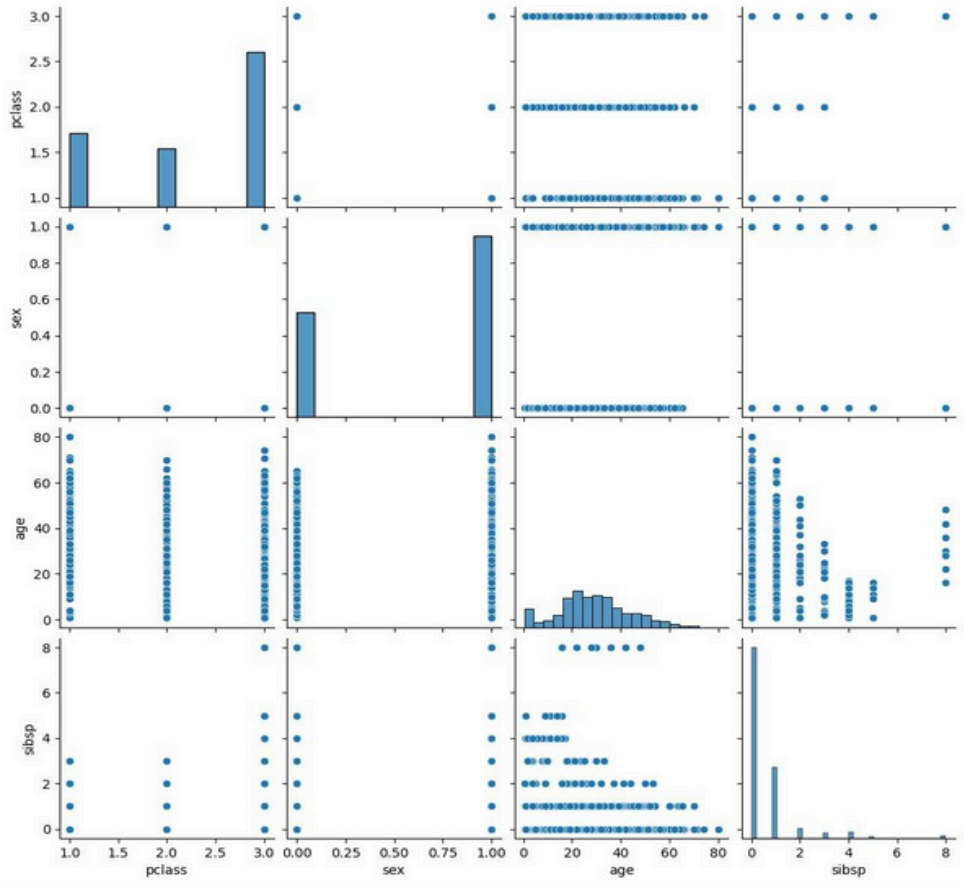
```
corr_features = ['pclass', 'age', 'sibsp', 'parch', 'fare']
corr_matrix = df[corr_features].corr()


plt.figure(figsize=(8, 6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap of Titanic Dataset') plt.show()
```

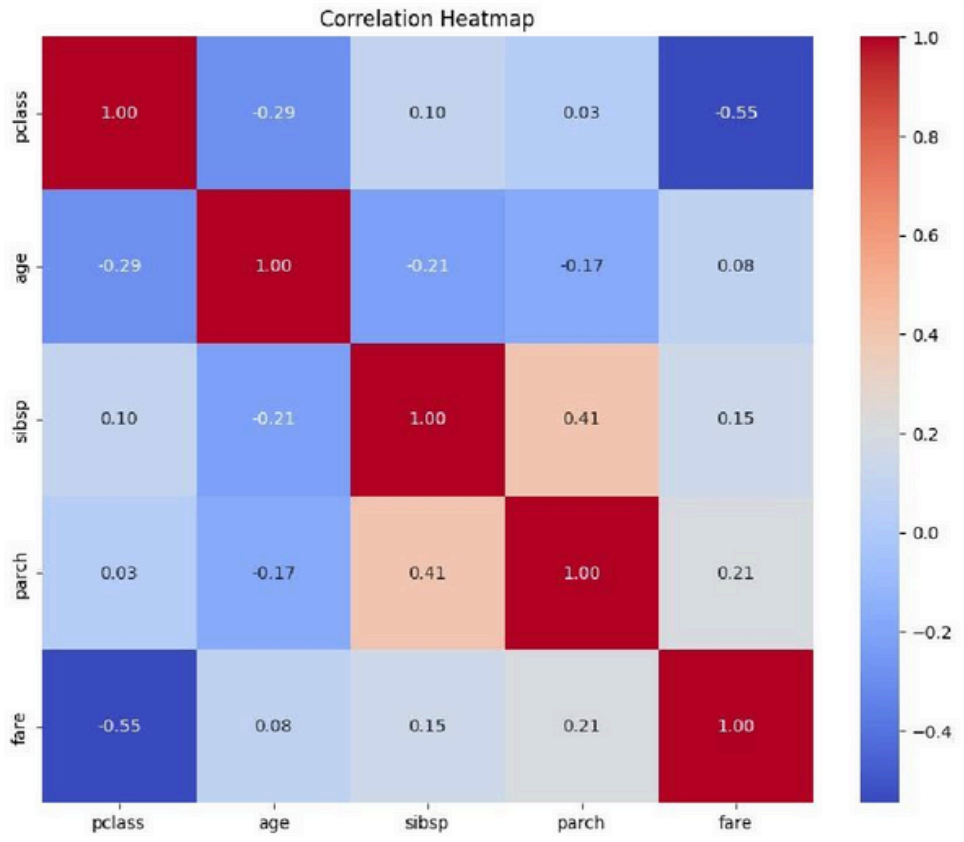
Output:



```
--- Initial DataFrame Head ---
   survived  pclass     sex   age  sibsp  parch     fare embarked  class  \
0         0       3    male  22.0      1      0   7.2500        S  Third
1         1       1  female  38.0      1      0  71.2833        C  First
2         1       3  female  26.0      0      0   7.9250        S  Third
3         1       1  female  35.0      1      0  53.1000        S  First
4         0       3    male  35.0      0      0   8.0500        S  Third

     who  adult_male deck  embark_town alive  alone
0    man        True  NaN  Southampton    no  False
1  woman       False    C    Cherbourg   yes  False
2  woman       False  NaN  Southampton   yes   True
3  woman       False    C  Southampton   yes  False
4    man        True  NaN  Southampton    no   True

--- DataFrame Info ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   survived     891 non-null    int64
 1   pclass       891 non-null    int64
 2   sex          891 non-null    object
 3   age          714 non-null    float64
 4   sibsp        891 non-null    int64
 5   parch        891 non-null    int64
 6   fare         891 non-null    float64
 7   embarked     889 non-null    object
 8   class        891 non-null    category
 9   who          891 non-null    object
 10  adult_male   891 non-null    bool
 11  deck         203 non-null    category
 12  embark_town  889 non-null    object
 13  alive        891 non-null    object
 14  alone        891 non-null    bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.7+ KB
```

```
--- DataFrame Head After Preprocessing ---
   survived  pclass  sex   age  sibsp  parch      fare embarked  class    who  \
0         0       3    1  22.0      1      0 -0.516916        S  Third    man
1         1       1    0  38.0      1      0  0.740208        C  First  woman
2         1       3    0  26.0      0      0 -0.503664        S  Third  woman
3         1       1    0  35.0      1      0  0.383227        S  First  woman
4         0       3    1  35.0      0      0 -0.501210        S  Third    man

   adult_male     deck  embark_town alive  alone
0        True  Unknown  Southampton    no  False
1       False        C    Cherbourg   yes  False
2       False  Unknown  Southampton   yes   True
3       False        C  Southampton   yes  False
4        True  Unknown  Southampton    no   True
```


pclass
sex
age
sibsp
pclass
sex
age
sibsp


Correlation Heatmap
pclass
age
sibsp
parch
fare
1.00 -0.29 0.10 0.03 -0.55
-0.29 1.00 -0.21 -0.17 0.08
0.10 -0.21 1.00 0.41 0.15
0.03 -0.17 0.41 1.00 0.21
-0.55 0.08 0.15 0.21 1.00
pclass
age
sibsp
parch
fare
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Result:

The dataset was successfully cleaned, encoded, normalized, and visualized using pair plots and a correlation heatmap.