Ex: 01
22/09/25

DATA PREPROCESSING AND CLEANING

Aim:

To load the titanic dataset and convert it into a DataFrame and explore it

Program code

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder,
                                  StandardScaler

titanic = sns.load_dataset('titanic')
print(titanic.head())
print("Data info:")
print(titanic.info())
print("\n Missing values per column:")
print(titanic.isnull().sum())
titanic['age'] = titanic['age'].fillna(method
               = .fillna(method='bfill')

titanic['deck'] = titanic['deck'].cat.add_
                  categories(['unknown'])

titanic.loc[missing_deck_indices, 'deck']
                = 'unknown'
```

Output

	Survived	pclass	age	sibsp	parch	fare	class
0	0	3	22.0	1	0	7.2500	Third
1	1	1	38.0	1	0	71.2833	First
2	1	3	26.0	0	0	7.9250	Third
3	1	1	35.0	1	0	53.1000	First
4	0	3	35.0	0	0	8.0500	Third

Missing values per column

Survived : 0
pclass : 0
sex : 0
age : 177
sibsp : 0
parch : 0

Duplicates removed : 86

```
le = LabelEncoder()
titanic['sex_encoded'] = le.fit_transform
                         (titanic['sex'])

scaler = StandardScaler()
titanic['fare_scaled'] = scaler.fit_transform
                         (titanic[['fare']])

sns.pairplot(titanic, vars=['pclass', 'sex_encoded',
                            'age', 'sibsp'])

plt.suptitle('pairplot of selected features',
             y=1.02)

corr_features = ['pclass', 'age', 'sibsp',
                 'parch', 'fare']

plt.figure(figsize=(8,6))
sns.heatmap(corr, annot=True, cmap='coolwarm',
            fmt=".2f")
plt.title('correlation Heatmap')
plt.show()
```

Thus the data preprocessing and cleaning using titanic dataset has been executed successfully