LAB-1

1. DATA PREPROCESSING AND CLEANING

Load the titanic dataset and convert it into a DataFrame. Explore and Understand the Data set Display the first few rows. Get information about column data types and missing values. Do forward, backward fill on Age. Fill missing values in Cabin with “unknown” and limit to 5. Remove Duplicate Records if any. Encode Categorical Columns Sex using LabelEncoder S cale Numerical Feature Fare using StandardScaler Pair Plot of Selected Features 'Pclass', 'Sex', 'Age', 'SibSp' Display the Correlation Heatmap for 'Pclass', 'Age', 'SibSp', 'Parch', 'Fa re'

CODE:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder, StandardScaler
df = pd.read_csv("train.csv")
print("First 5 rows:")
print(df.head())
print("\nDataset Info:")
print(df.info())
df['Age'] = df['Age'].ffill()
df['Age'] = df['Age'].bfill()
if 'Cabin' in df.columns:
    df['Cabin'] = df['Cabin'].fillna('unknown', limit=5)
else:
```

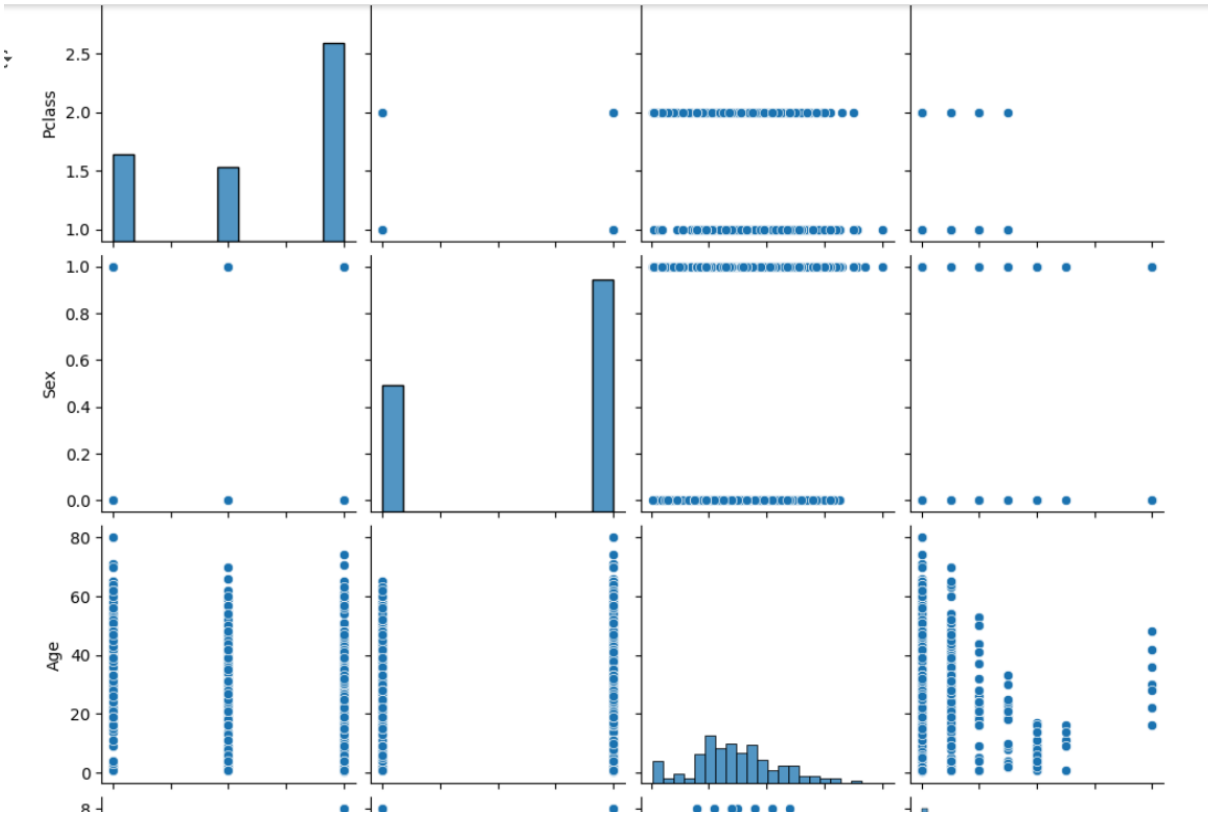
```
    print("\n 'Cabin' column not found in dataset. Skipping this step.")


df.drop_duplicates(inplace=True)
le = LabelEncoder()
df['Sex'] = le.fit_transform(df['Sex'])
scaler = StandardScaler()
df['Fare'] = scaler.fit_transform(df[['Fare']])
selected_features = ['Pclass', 'Sex', 'Age', 'SibSp']
sns.pairplot(df[selected_features])
plt.show()
corr_features = ['Pclass', 'Age', 'SibSp', 'Parch', 'Fare']
plt.figure(figsize=(8,6))
sns.heatmap(df[corr_features].corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

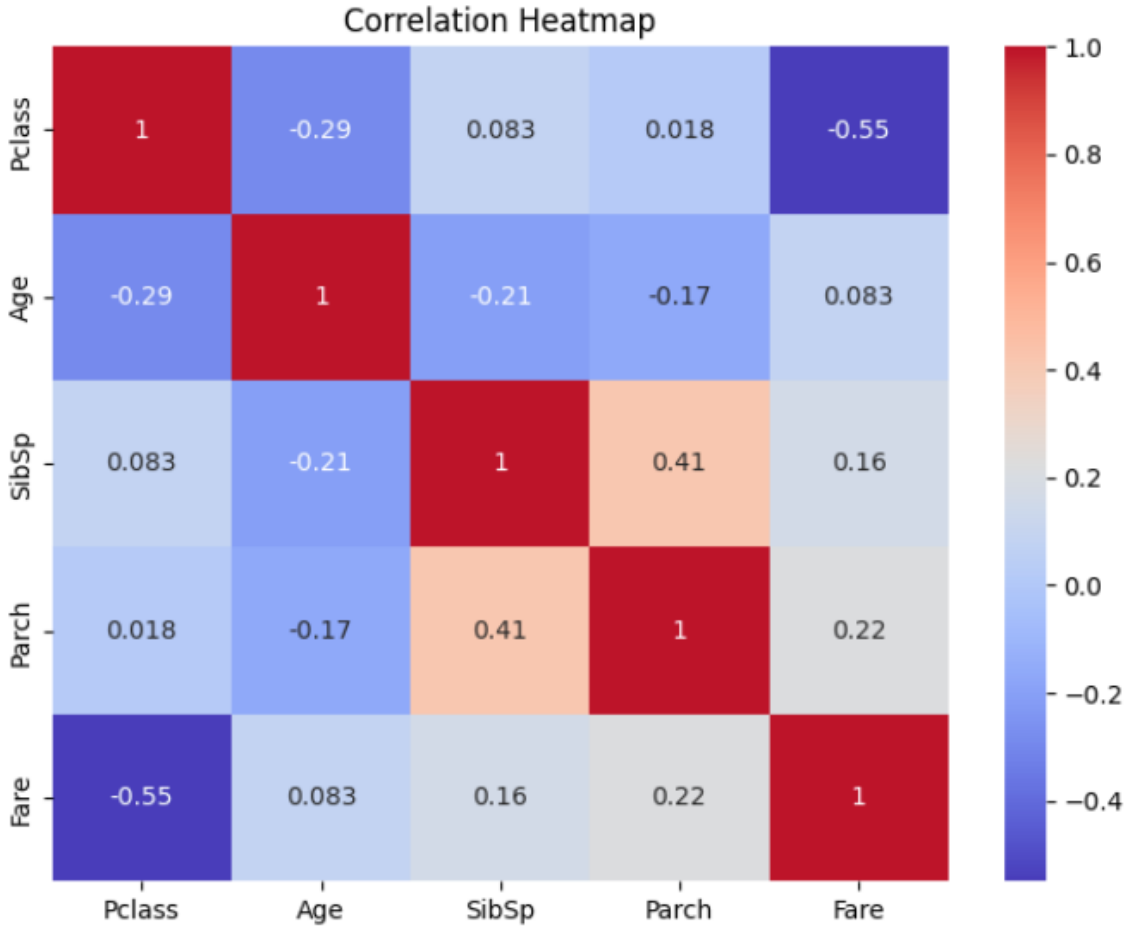
OUTPUT:

```
First 5 rows:
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None
```

231801135


Pclass
Sex
Age


Correlation Heatmap
Pclass
Age
SibSp
Parch
Fare
1
-0.29
0.083
0.018
-0.55
-0.29
1
-0.21
-0.17
0.083
0.083
-0.21
1
0.41
0.16
0.018
-0.17
0.41
1
0.22
-0.55
0.083
0.16
0.22
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

231801135