EXP NO: 2	# Data Discovery and Preparation
DATE: 22/7/25	

Aim:

To explore, clean, and prepare the Titanic dataset for analysis by handling missing values, performing data exploration, and splitting the dataset for modelling.

Program:

Step 1: Import Required Libraries

```
import pandas as pd import numpy as np import seaborn
as sns import matplotlib.pyplot as plt from
sklearn.impute importSimpleImputer from
sklearn.model_selectionimport train_test_split Step
df=pd.read_csv('titanic.csv') Step
```

2: Load the Dataset

3: Understand the Data

```
print("\nShape of the    dataset:") print(df.shape)

print("\nInformation     about the dataset:")
df.info()
print("\nDescriptive     statistics    of    the    dataset:")
print(df.describe())
```

Step 4: Handle Missing Values

```
# Replace missing 'Age' values with mean imputer
= SimpleImputer(strategy='mean') df['Age'] =
imputer.fit_transform(df[['Age']])


#    Fill    missing    'Cabin'    values    with    'Unknown'
df['Cabin'].fillna('Unknown', inplace=True)

 # Fill missing 'Embarked' values with most frequent value
mode_embarked          =          df['Embarked'].mode()[0]
df['Embarked'].fillna(mode_embarked, inplace=True) Step 5:
```

Visualize Passenger Class Distribution

```
plt.figure(figsize=(8, 6))
sns.countplot(x='Pclass', data=df)

plt.title('Passenger Count by Class') plt.show()
```

Step 6: Display Female Passengers Who Survived

```
female_survivors = df[(df['Sex'] == 'female') & (df['Survived'] == 1)]
print(female_survivors[['Name', 'Sex', 'Survived']].head()) Step 7:
```

Display 3rd Class Passengers Under18

```
third_class_under_18 = df[(df['Pclass'] == 3) & (df['Age'] < 18)]
print(third_class_under_18[['Name', 'Pclass', 'Age']].head())
```

Step 8: Display 1st Class Passengers Older than 40

```
first_class_over_40 = df[(df['Pclass'] == 1) & (df['Age'] > 40)]
print(first_class_over_40[['Name', 'Pclass','Age']].head())
```

Step 9: Survivors from the Above Category (1st Class, >40)

```
survivors_first_class_over_40 =
first_class_over_40[first_class_over_40['Survived'] == 1]
print(survivors_first_class_over_40[['Name', 'Pclass', 'Age',
'Survived']].head())
```

Step 10: Male Passengers with Fare > 100

```
male_high_fare = df[(df['Sex'] == 'male') & (df['Fare'] > 100)]
print(male_high_fare[['Name', 'Sex', 'Fare']].head())
```

Step 11: Passengers from Cherbourg ('C') in 2nd Class

```
cherbourg_second_class = df[(df['Embarked'] == 'C') & (df['Pclass'] == 2)]
print(cherbourg_second_class[['Name', 'Embarked', 'Pclass']].head())
```

Step 12: Passengers with More than 2 Siblings/Spouses

```
large_families_sibsp = df[df['SibSp'] > 2]
print(large_families_sibsp[['Name',
'SibSp']].head())
```

Step 13: Passengers Who Did Not Survive and Had No Family

```
died_alone = df[(df['Survived'] == 0) & (df['SibSp']
                                        == 0) & (df['Parch'] == 0)]
print(died_alone[['Name', 'Survived', 'SibSp', 'Parch']].head())
```

Step 14: T
Oldest Survivors

```
oldest_survivors = df[df['Survived']          ==           1].sort_values(by='Age',
ascending=False).head(5) print(oldest_survivors[['Name', 'Age',
'Survived']])
```

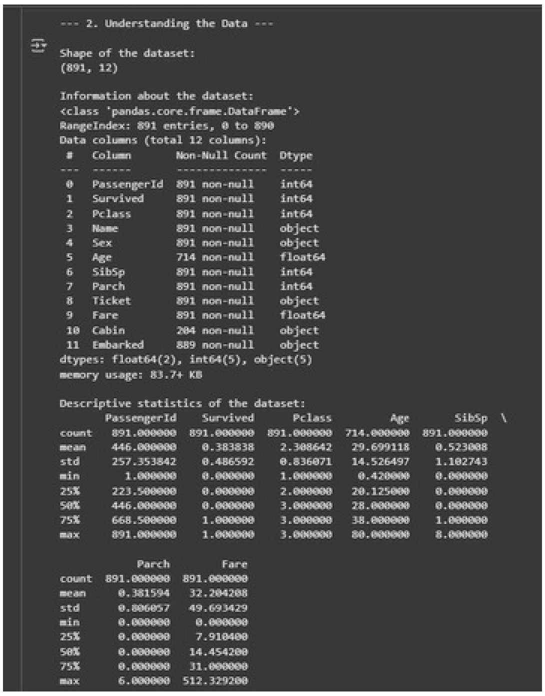
Step 15: Passengers with Zero Fare

```
zero_fare_passengers      =      df[df['Fare']      ==      0]
print(zero_fare_passengers[['Name', 'Fare']])
```

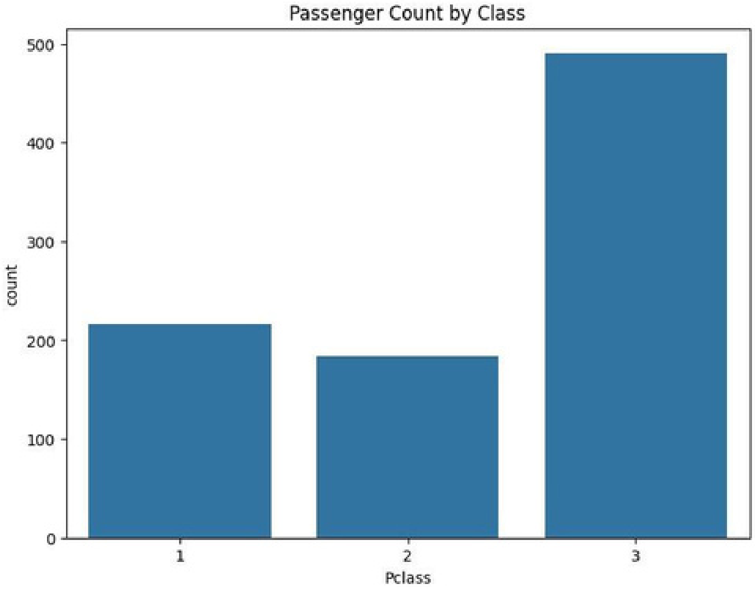
Step 16: Split Dataset into Train and Test Sets

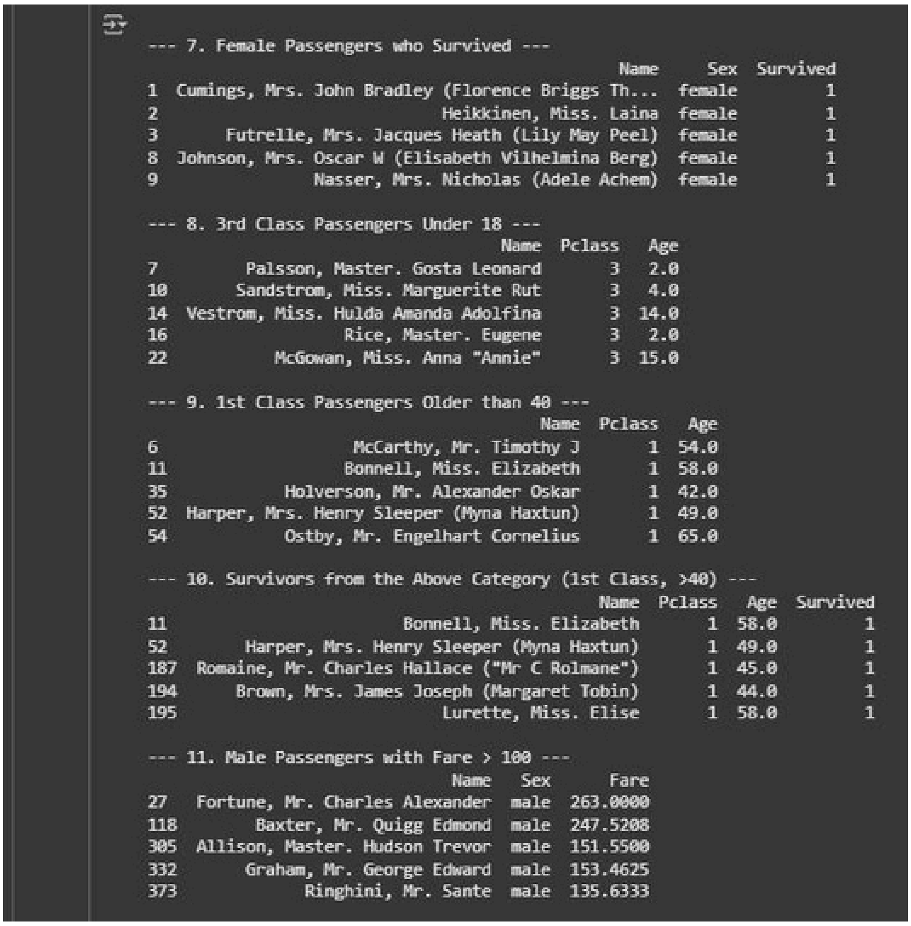
```
df_cleaned = df.drop(['Name', 'Ticket', 'Cabin', 'Embarked', 'Sex'], axis=1)
X = df_cleaned.drop('Survived', axis=1) y
= df_cleaned['Survived']

X_train,  X_test,  y_train,  y_test  =  train_test_split(X,  y,  test_size=0.2,
random_state=42)
 print("Training set shape (X_train):", X_train.shape)
print("Testing set shape (X_test):", X_test.shape)
print("Training labels shape (y_train):", y_train.shape)
print("Testing labels shape (y_test):", y_test.shape)
```

Output:

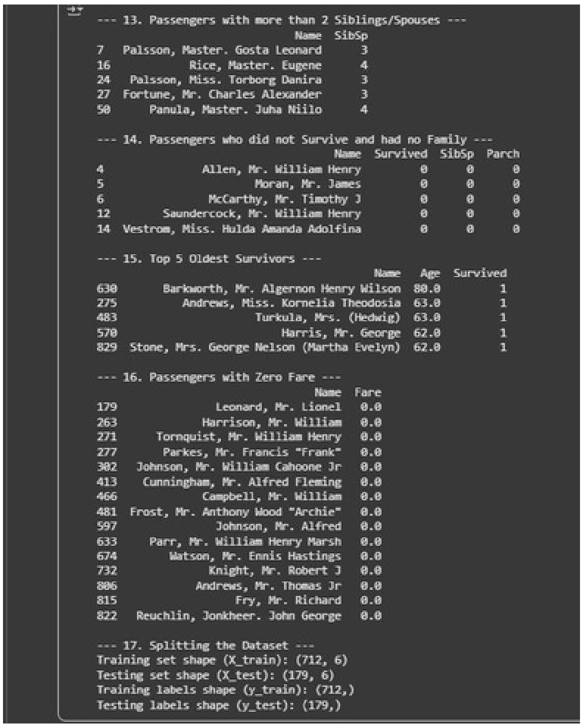


```
--- 2. Understanding the Data ---

Shape of the dataset:
(891, 12)

Information about the dataset:
<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

Descriptive statistics of the dataset:
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000   
mean    446.000000    0.383838    2.308642   29.699118    0.523008   
std     257.353842    0.486592    0.836071   14.526497    1.102743   
min       1.000000    0.000000    1.000000    0.420000    0.000000   
25%     223.500000    0.000000    2.000000   20.125000    0.000000   
50%     446.000000    0.000000    3.000000   28.000000    0.000000   
75%     668.500000    1.000000    3.000000   38.000000    1.000000   
max     891.000000    1.000000    3.000000   80.000000    8.000000   

            Parch        Fare  
count  891.000000  891.000000  
mean     0.381594   32.204208  
std      0.806057   49.693429  
min      0.000000    0.000000  
25%      0.000000    7.910400  
50%      0.000000   14.454200  
75%      0.000000   31.000000  
max      6.000000  512.329200  
```





```
--- 7. Female Passengers who Survived ---
                                               Name     Sex  Survived
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female         1
2                             Heikkinen, Miss. Laina  female         1
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female         1
8  Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)  female         1
9                Nasser, Mrs. Nicholas (Adele Achem)  female         1

--- 8. 3rd Class Passengers Under 18 ---
                                Name  Pclass   Age
7         Palsson, Master. Gosta Leonard      3   2.0
10        Sandstrom, Miss. Marguerite Rut      3   4.0
14   Vestrom, Miss. Hulda Amanda Adolfina      3  14.0
16                    Rice, Master. Eugene      3   2.0
22           McGowan, Miss. Anna "Annie"      3  15.0

--- 9. 1st Class Passengers Older than 40 ---
                                    Name  Pclass   Age
6                    McCarthy, Mr. Timothy J       1  54.0
11                  Bonnell, Miss. Elizabeth       1  58.0
35            Holverson, Mr. Alexander Oskar       1  42.0
52  Harper, Mrs. Henry Sleeper (Myna Haxtun)       1  49.0
54            Ostby, Mr. Engelhart Cornelius       1  65.0

--- 10. Survivors from the Above Category (1st Class, >40) ---
                                          Name  Pclass   Age  Survived
11                    Bonnell, Miss. Elizabeth       1  58.0         1
52        Harper, Mrs. Henry Sleeper (Myna Haxtun)       1  49.0         1
187  Romaine, Mr. Charles Hallace ("Mr C Rolmane")       1  45.0         1
194     Brown, Mrs. James Joseph (Margaret Tobin)       1  44.0         1
195                         Lurette, Miss. Elise       1  58.0         1

--- 11. Male Passengers with Fare > 100 ---
                            Name   Sex      Fare
27   Fortune, Mr. Charles Alexander  male  263.0000
118        Baxter, Mr. Quigg Edmond  male  247.5208
305  Allison, Master. Hudson Trevor  male  151.5500
332        Graham, Mr. George Edward  male  153.4625
373              Ringhini, Mr. Sante  male  135.6333
```



```
--- 13. Passengers with more than 2 Siblings/Spouses ---
                            Name  SibSp
7   Palsson, Master. Gosta Leonard      3
16            Rice, Master. Eugene      4
24   Palsson, Miss. Torborg Danira      3
27  Fortune, Mr. Charles Alexander      3
50      Panula, Master. Juha Niilo      4

--- 14. Passengers who did not Survive and had no Family ---
                                  Name  Survived  SibSp  Parch
4             Allen, Mr. William Henry         0      0      0
5                     Moran, Mr. James         0      0      0
6              McCarthy, Mr. Timothy J         0      0      0
12        Saundercock, Mr. William Henry         0      0      0
14  Vestrom, Miss. Hulda Amanda Adolfina         0      0      0

--- 15. Top 5 Oldest Survivors ---
                                       Name   Age  Survived
630       Barkworth, Mr. Algernon Henry Wilson  80.0         1
275        Andrews, Miss. Kornelia Theodosia  63.0         1
483                    Turkula, Mrs. (Hedwig)  63.0         1
570                        Harris, Mr. George  62.0         1
829  Stone, Mrs. George Nelson (Martha Evelyn)  62.0         1

--- 16. Passengers with Zero Fare ---
                               Name  Fare
179              Leonard, Mr. Lionel   0.0
263             Harrison, Mr. William   0.0
271        Tornquist, Mr. William Henry   0.0
277         Parkes, Mr. Francis "Frank"   0.0
302   Johnson, Mr. William Cahoone Jr   0.0
413    Cunningham, Mr. Alfred Fleming   0.0
466             Campbell, Mr. William   0.0
481  Frost, Mr. Anthony Wood "Archie"   0.0
597              Johnson, Mr. Alfred   0.0
633      Parr, Mr. William Henry Marsh   0.0
674        Watson, Mr. Ennis Hastings   0.0
732              Knight, Mr. Robert J   0.0
806            Andrews, Mr. Thomas Jr   0.0
815                 Fry, Mr. Richard   0.0
822   Reuchlin, Jonkheer. John George   0.0

--- 17. Splitting the Dataset ---
Training set shape (X_train): (712, 6)
Testing set shape (X_test): (179, 6)
Training labels shape (y_train): (712,)
Testing labels shape (y_test): (179,)
```

Result:

The dataset was successfully analyzed, cleaned, and divided into training and testing sets, ready for further machine learning tasks.