Task 5 - Exploratory Data Analysis (EDA)

Titanic Dataset

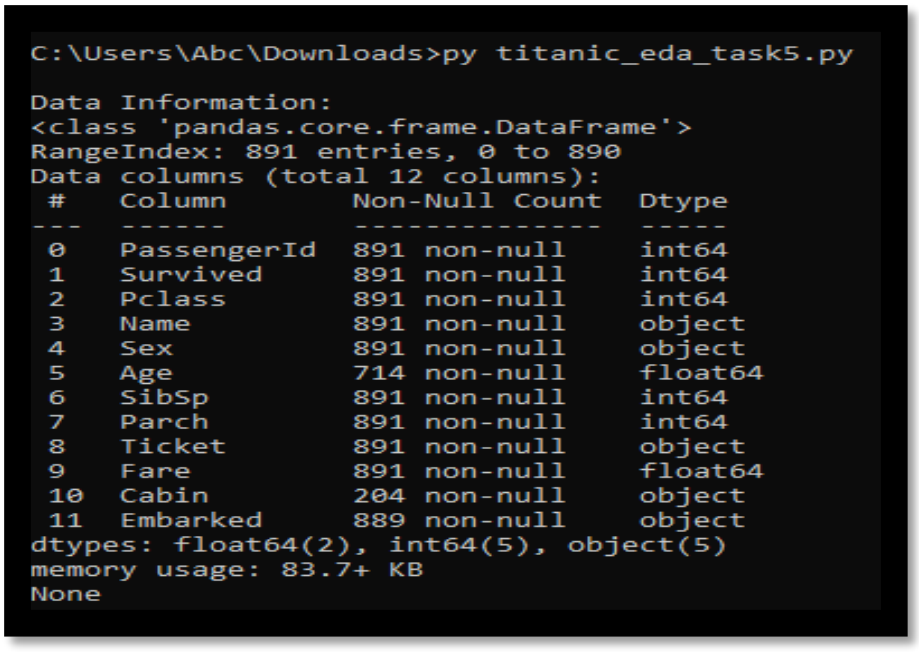
1. First 5 Rows of the Titanic Dataset (df.head())

```
C:\Users\Abc\Downloads>py titanic_eda_task5.py
First 5 rows:
   PassengerId  Survived  Pclass                                               Name     Sex   Age  SibSp  Parch            Ticket     Fare Cabin Embarked
0            1         0       3                            Braund, Mr. Owen Harris    male  22.0      1      0         A/5 21171   7.2500   NaN        S
1            2         1       1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1      0          PC 17599  71.2833   C85        C
2            3         1       3                             Heikkinen, Miss. Laina  female  26.0      0      0  STON/O2. 3101282   7.9250   NaN        S
3            4         1       1       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1      0            113803  53.1000  C123        S
4            5         0       3                           Allen, Mr. William Henry    male  35.0      0      0            373450   8.0500   NaN        S
```

Observation:

- Displays the first 5 records of the dataset.
- Important columns: PassengerId, Survived, Pclass, Name, Sex, Age, etc.
- Some Cabin values are missing (NaN observed).

2. Data Information and Data Types (df.info())



```
C:\Users\Abc\Downloads>py titanic_eda_task5.py

Data Information:
<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
```

Observation:

- Dataset has 891 rows and 12 columns.
- Columns like Age, Cabin, and Embarked have missing values.
- Most columns are integers (int64), few are objects (text).

3. Summary Statistics of Numerical Columns (df.describe())

```
C:\Users\Abc\Downloads>py titanic_eda_task5.py

Summary Statistics:
       PassengerId    Survived      Pclass         Age       SibSp       Parch        Fare
count   891.000000  891.000000  891.000000  714.000000  891.000000  891.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008    0.381594   32.204208
std     257.353842    0.486592    0.836071   14.526497    1.102743    0.806057   49.693429
min       1.000000    0.000000    1.000000    0.420000    0.000000    0.000000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000    0.000000    7.910400
50%     446.000000    0.000000    3.000000   28.000000    0.000000    0.000000   14.454200
75%     668.500000    1.000000    3.000000   38.000000    1.000000    0.000000   31.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000    6.000000  512.329200
```

Observation:

- Average passenger age is about 29.7 years.
- Fare values vary widely, with a maximum of 512.3
- Age values range from 0.42 years to 80 years.
- Most passengers paid around 14–31 fare units.

4. Value Counts for Survived, Sex, and Pclass

```
C:\Users\Abc\Downloads>py titanic_eda_task5.py

Survived Value Counts:
Survived
0    549
1    342
Name: count, dtype: int64

Sex Value Counts:
Sex
male      577
female    314
Name: count, dtype: int64

Pclass Value Counts:
Pclass
3    491
1    216
2    184
Name: count, dtype: int64
```

Observation:

- 549 passengers did not survive, 342 survived.
- 577 passengers were male, 314 were female.
- Most passengers were traveling in third class (Pclass = 3).

5. Missing Values (df.isnull().sum())

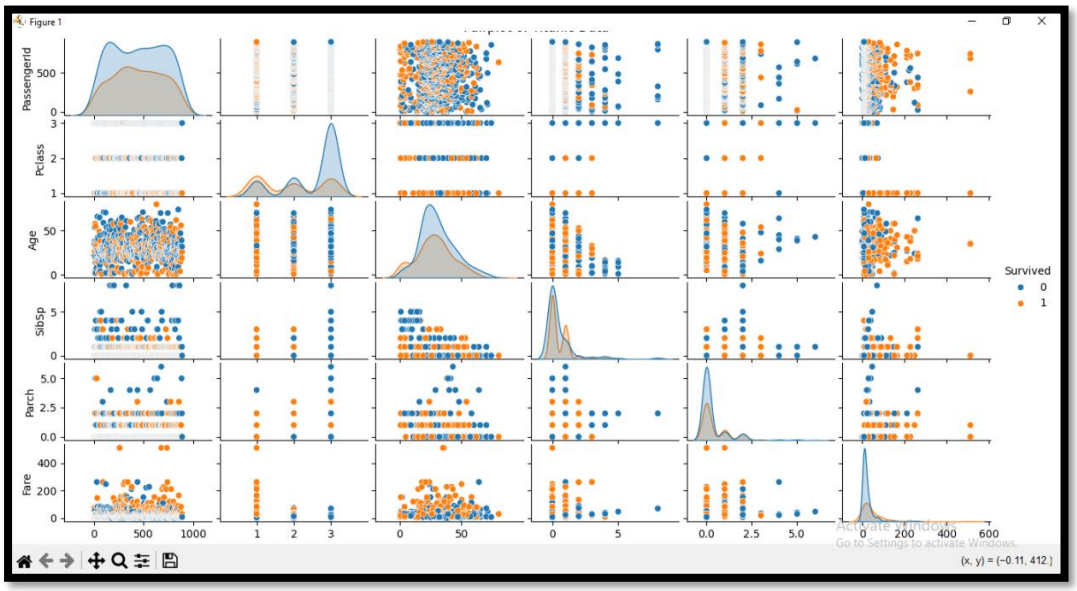
```
C:\Users\Abc\Downloads>py titanic_eda_task5.py

Missing Values:
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Observation:

- Age column has 177 missing values.
- Cabin column has a very large number of missing values (687 missing).
- Embarked has only 2 missing values.

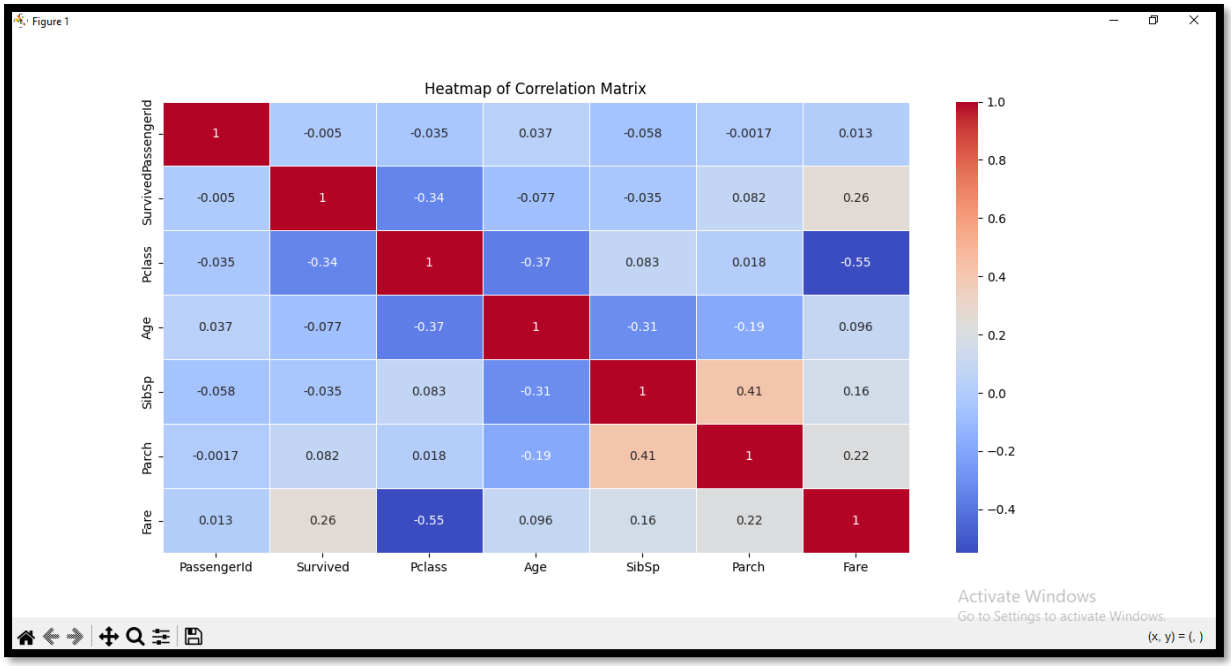
6. Pairplot of Titanic Data



Observation:

- Younger passengers (Age < 20) had higher survival rates.
- First-class passengers had a better chance of survival compared to lower classes.
- Passengers who paid higher fares also had better survival chances.

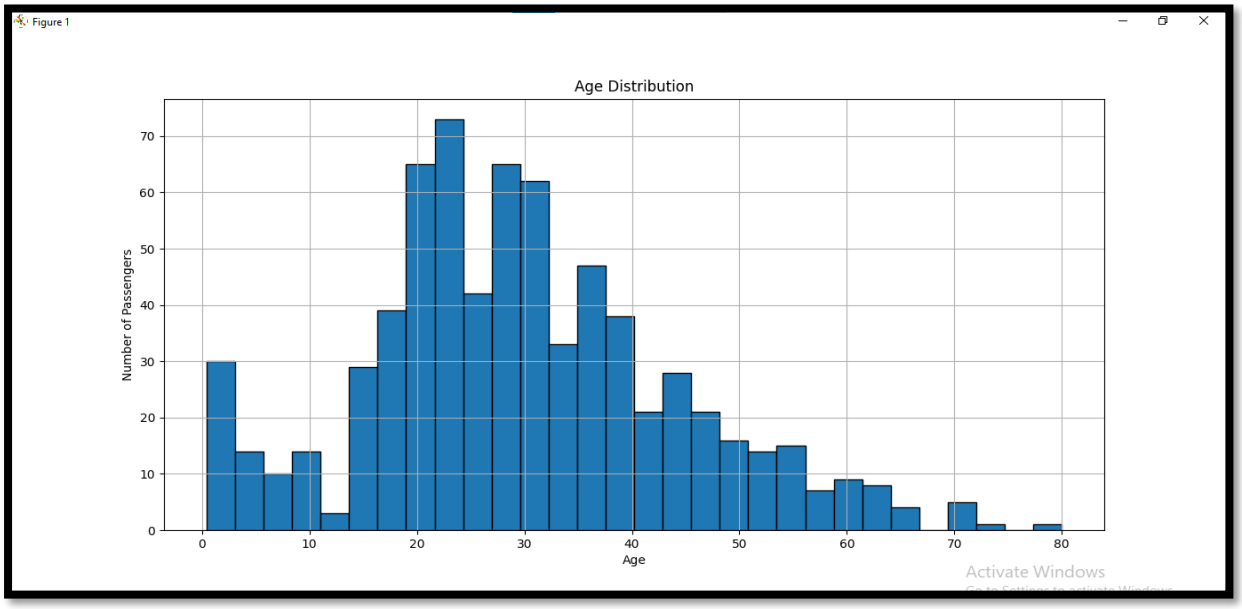
7. Heatmap of Correlation Matrix



Observation:

- Pclass and Fare have a strong negative correlation (-0.55).
- Survived has a moderate positive correlation with Fare (0.26).
- Age has a weak correlation with survival (-0.077).

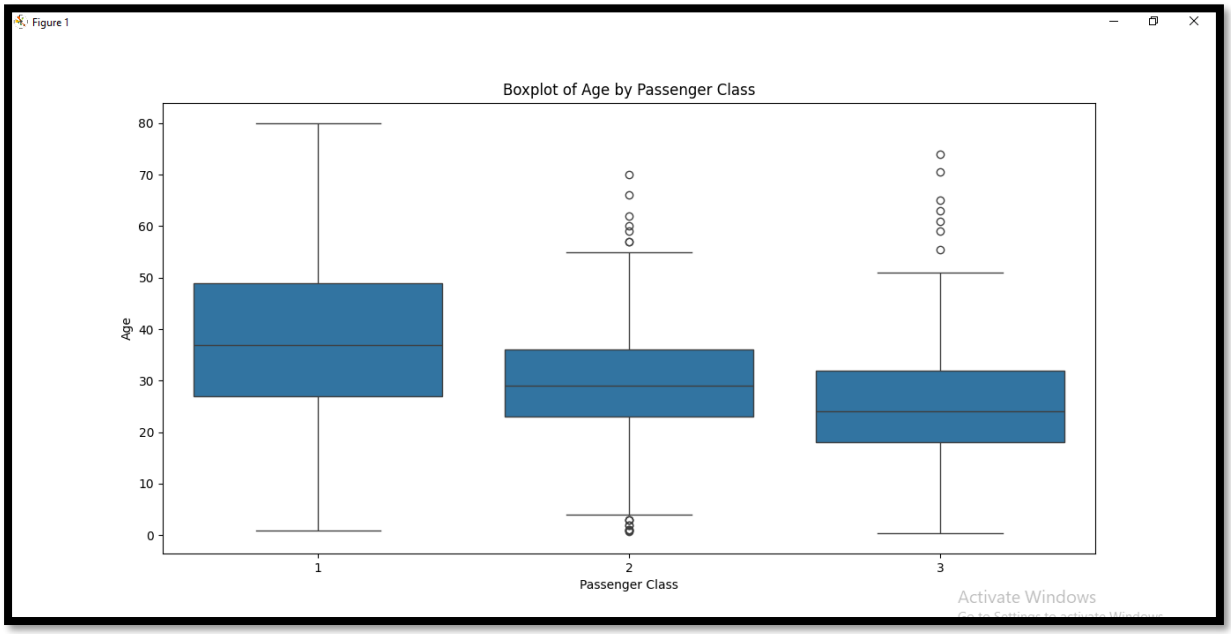
8. Histogram of Age Distribution



Observation:

- Most passengers were between 20 to 40 years old.
- Very few passengers were older than 60.
- There were also a few very young children (under 5 years).

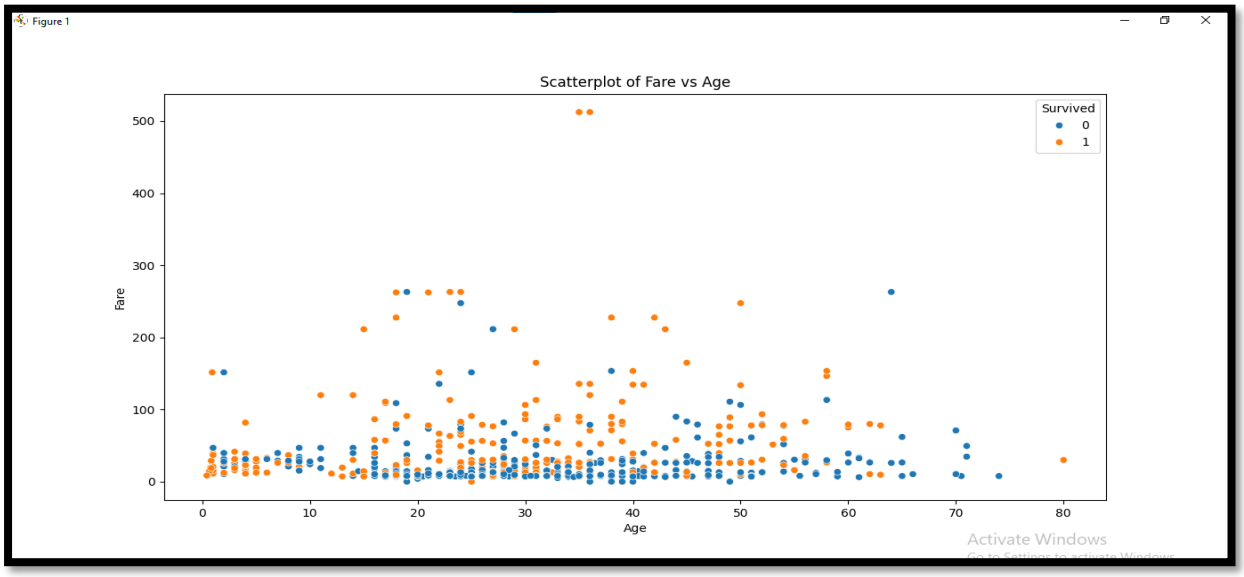
9. Boxplot of Age by Passenger Class



Observation:

- First-class passengers were generally older.
- Third-class passengers were younger on average.
- Outliers are seen in all classes, showing very old or very young passengers.

10. Scatterplot of Fare vs Age



Observation:

- Most passengers paid fares under 100.
- Some passengers (especially older ones) paid extremely high fares (over 200).
- Higher fare is associated with higher survival rates.

Summary of Findings:

- Most passengers were aged between 20 and 40 years.
- Females and first-class passengers had better survival chances.
- Higher fare amounts were positively associated with survival.
- Passenger class strongly affected fare prices.
- Missing values were observed mainly in the Age and Cabin columns.