Titanic Dataset EDA (Pandas Only)

Objective: Perform Exploratory Data Analysis (EDA) on the Titanic dataset using Pandas only, without visualization libraries.

In [1]:
```
import pandas as pd

# Load dataset
df = pd.read_csv("titanic.csv")
```

In [2]:
```
# Dataset shape
print("Shape of dataset:", df.shape)

# Dataset info
df.info()
```

```
Shape of dataset: (891, 12)
<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
```

In [3]:
```
# Summary statistics (numerical)
df.describe()
```

Out[3]:

	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

In [6]:
```
# Summary statistics (categorical)
df.describe(include=['object'])
```

Out[6]:

	Name	Sex	Ticket	Cabin	Embarked
count	891	891	891	204	889
unique	891	2	681	147	3
top	Braund, Mr. Owen Harris	male	347082	G6	S
freq	1	577	7	4	644

In [7]:
```
# Missing values count
df.isnull().sum()
```

Out[7]:
```
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
```

In [8]:
```
# Survival count
df['Survived'].value_counts()
```

Out[8]:
```
Survived
0    549
1    342
Name: count, dtype: int64
```

In [9]:
```
# Passenger class distribution
df['Pclass'].value_counts()
```

Out[9]:
```
Pclass
3    491
1    216
2    184
Name: count, dtype: int64
```

In [10]:
```
# Gender distribution
df['Sex'].value_counts()
```

Out[10]:
```
Sex
male      577
female    314
Name: count, dtype: int64
```

In [11]:
```
# Average age by class
df.groupby('Pclass')['Age'].mean()
```

Out[11]:
```
Pclass
1    38.233441
2    29.877630
3    25.140620
Name: Age, dtype: float64
```

In [12]:
```
# Survival rate by gender
df.groupby('Sex')['Survived'].mean()
```

Out[12]:
```
Sex
female    0.742038
male      0.188908
Name: Survived, dtype: float64
```

In [5]:
```
# Survival rate by class
df.groupby('Pclass')['Survived'].mean()
```

Out[5]:
```
Pclass
1    0.629630
2    0.472826
3    0.242363
Name: Survived, dtype: float64
```

In [4]:
```
# Correlation matrix
df.corr(numeric_only=True)
```

Out[4]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005007	-0.035144	0.036847	-0.057527	-0.001652	0.012658
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307
Pclass	-0.035144	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500
Age	0.036847	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067
SibSp	-0.057527	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651
Parch	-0.001652	0.081629	0.018443	-0.189119	0.414838	1.000000	0.216225
Fare	0.012658	0.257307	-0.549500	0.096067	0.159651	0.216225	1.000000

Observations

1. Most passengers were in 3rd class.
2. Survival rate for females is higher than for males.
3. Higher fare and first class show higher survival chances.
4. There are missing values in Age and Cabin columns.

Summary of Findings

- Majority of passengers were in 3rd class.
- Female passengers had higher survival chances.
- Survival rates were higher in 1st class.
- Fare seems positively correlated with survival.
- Missing values exist in `Age` and `Cabin`.

In [ ]: