Exploratory Data Analysis and Data Cleaning on Titanic Dataset

Submission Report

Name: Tarun Chincholia
Internship Domain: Data Science
Task Number: Task 2
Organization: Prodigy InfoTech
Date: 23-07-25

Objective:

To perform data cleaning and exploratory data analysis (EDA) on the Titanic dataset in order to identify patterns, trends, and relationships between variables that influenced passenger survival.

Dataset Used:

Source: Kaggle

Link: Titanic Dataset on Kaggle

File Used:train.csv

Technologies Used:

-Python

-Pandas – for data cleaning and transformation

-Matplotlib & Seaborn – for visualizations

-NumPy – for statistical operations

Task-2

In [3]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [14]:

```
data = pd.read_csv("train.csv")
```

In [16]:

```
data.head()
```

Out[16]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/ O2. 3101282
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450

In [17]:

```
data.describe()
```

Out[17]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000

In [18]:

```
data.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
```

In [19]:

```
data.isnull().sum()
```

Out[19]:

```
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 [22]:

```
data.dropna(subset=["Embarked"], inplace=True)
data["Cabin"] = data["Cabin"].fillna("Unknown")
data["Age"] = data["Age"].fillna(data["Age"].mean())
```

In [23]:

```
data.isnull().sum()
```

Out[23]:

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
```

In [24]:

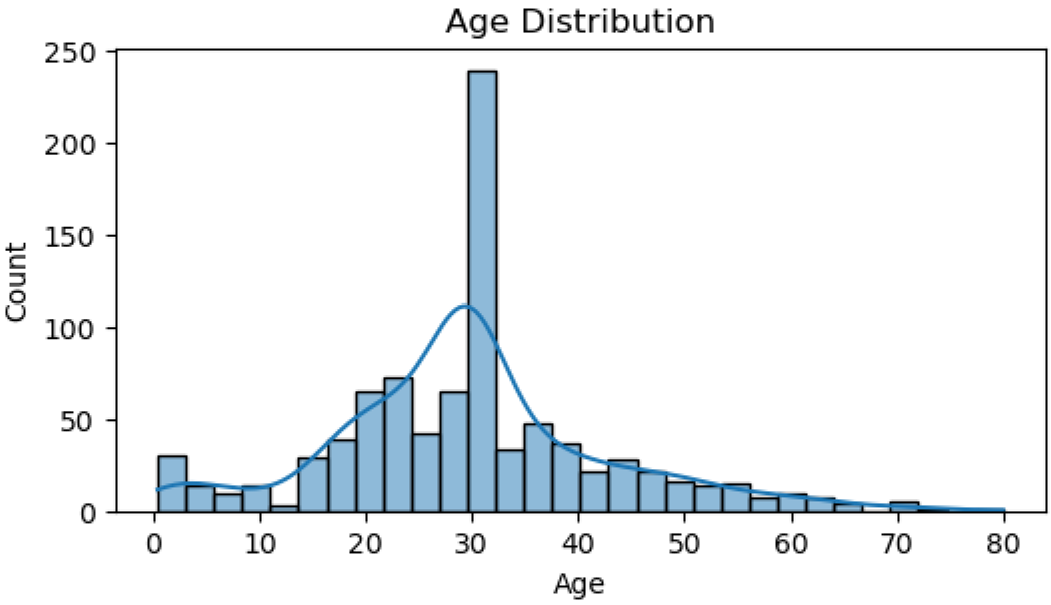
```
data.duplicated().sum()
```

Out[24]:

```
np.int64(0)
```

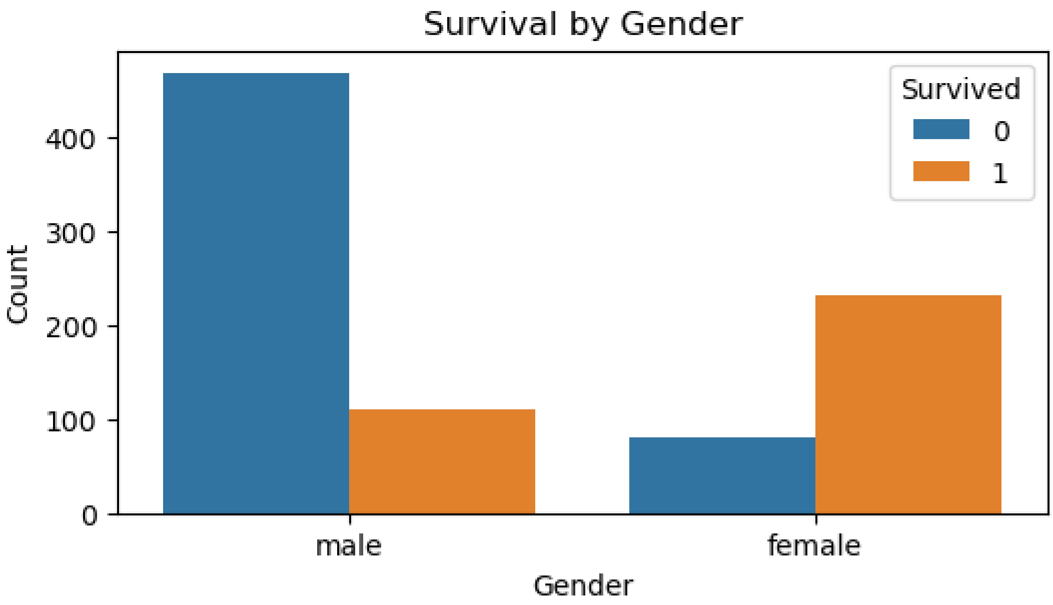
In [25]:

```
plt.figure(figsize=(6, 3))
sns.histplot(data["Age"], kde=True)
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Count")
plt.show()
```



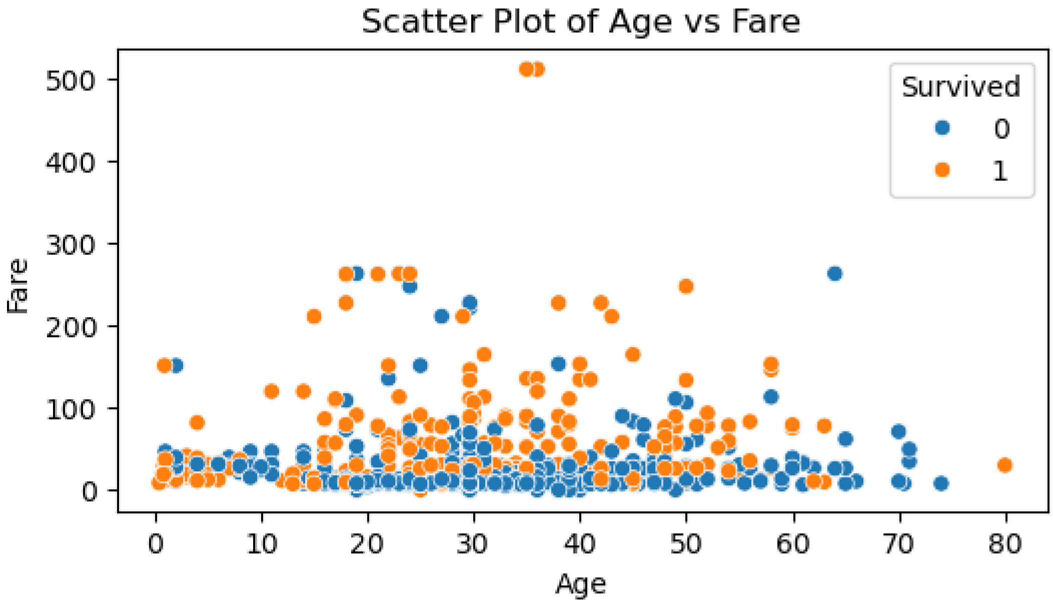
In [26]:

```
plt.figure(figsize=(6, 3))
sns.countplot(data=data, x="Sex", hue="Survived")
plt.title("Survival by Gender")
plt.xlabel("Gender")
plt.ylabel("Count")
plt.legend(title="Survived", loc="upper right")
plt.show()
```



In [27]:

```
plt.figure(figsize=(6, 3))
sns.scatterplot(data=data, x="Age", y="Fare", hue="Survived")
plt.title("Scatter Plot of Age vs Fare")
plt.xlabel("Age")
plt.ylabel("Fare")
plt.legend(title="Survived")
plt.show()
```



In [ ]: