Exploratory Data Analysis (EDA) – Titanic Dataset

Introduction:

The purpose of this Exploratory Data Analysis (EDA) is to understand the Titanic dataset, identify patterns, detect missing values, analyze relationships between variables, and extract insights that help in model building and decision-making. The Titanic dataset contains passenger details such as age, gender, ticket class, fare, and survival status.

Dataset Overview:

- **Total rows: 891**
- **Total columns: 12**
- **Target variable: Survived**
- **Key features:**
- **Numerical: Age, Fare, SibSp, Parch**
- **Categorical: Sex, Pclass, Embarked, Cabin, Ticket**

Data Cleaning:

Handling Missing Values

- Age: Filled with median age
- Embarked: Filled with mode value (most frequent)
- Fare: Filled with median fare
- Cabin: Filled with “Unknown”
- No rows dropped to avoid data loss

Feature Engineering

Created new features:

- Title extracted from Name
- FamilySize = SibSp + Parch + 1
- IsAlone = 1 if FamilySize == 1 else 0

Univariate Analysis:

Numerical Features

- Age: Right-skewed distribution. Many passengers between age 20–40.
- Fare: Highly right-skewed. Few passengers paid very high fares.
- SibSp & Parch: Most passengers travelled with few or no siblings/parents.

Categorical Features

- Sex: More males than females.
- Pclass: Most passengers were in 3rd class.
- Embarked: Majority boarded at Southampton (S).

Bivariate Analysis:

Survival by Sex

- Females had a significantly higher survival rate than males.

Survival by Passenger Class

- 1st Class passengers survived the most.
- 3rd Class passengers had very low survival.

Survival by Age

- Children (especially age < 12) show higher survival probability.
- Adults show a wider distribution with lower survival.

Survival by Fare

- Higher fare passengers survived more, indicating higher survival of wealthy families.

Multivariate Analysis:

Correlation Heatmap Insights

- Sex has strong positive correlation with Survival (Female = 1).
- Pclass has strong negative correlation with Survival.
- Fare shows positive correlation with Survival.
- FamilySize and IsAlone show mild influence.

Feature Engineering Insights:

Title

- Titles like “Mrs”, “Miss”, “Master” show higher survival.
- “Mr” and rare titles have lower survival.

FamilySize

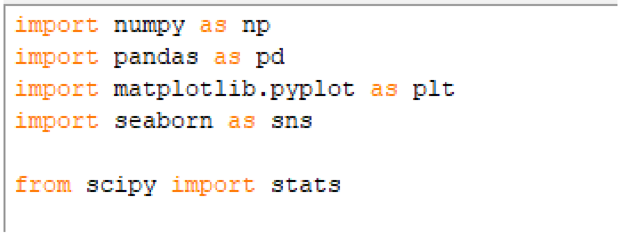
- Medium-sized families have better survival chances.
- Passengers alone have a lower survival rate.

Key Insights & Findings:

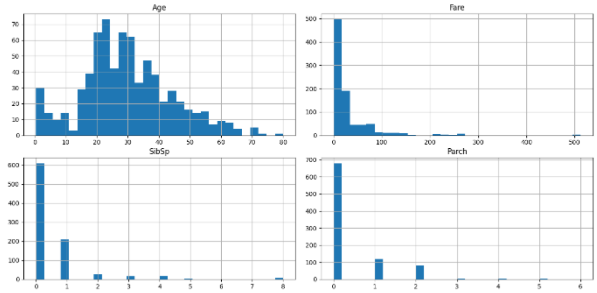
- Sex is the strongest predictor of survival. Women survived more.

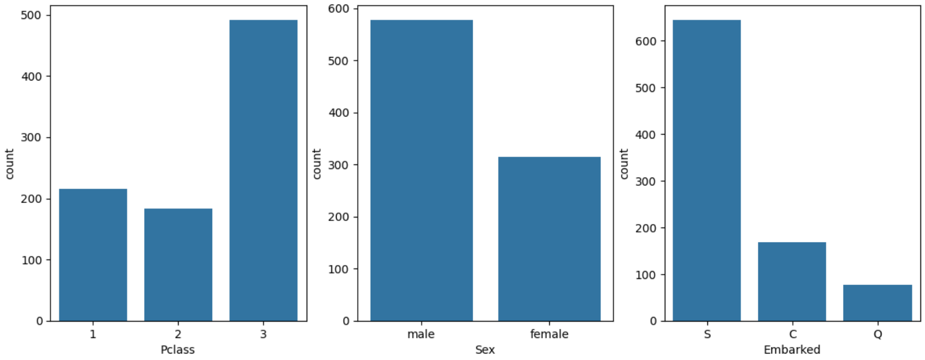
- Pclass strongly influences survival – rich passengers were prioritized.
- Age influences survival – children were prioritized during rescue.
- Fare correlates with survival, reflecting economic status.
- Title, FamilySize, and IsAlone add extra predictive power.
- Most missing values belonged to Cabin, making it unreliable for modeling.

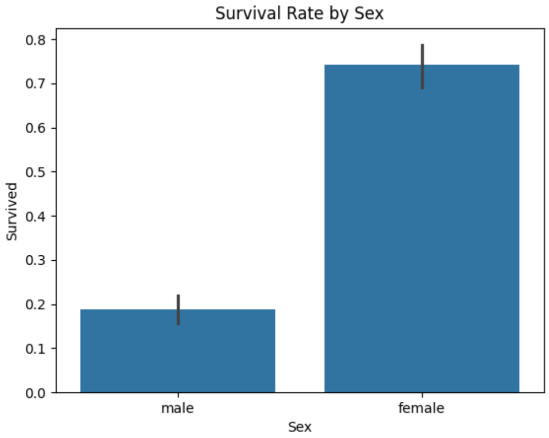
Screenshots(Google colab &IDLE python):

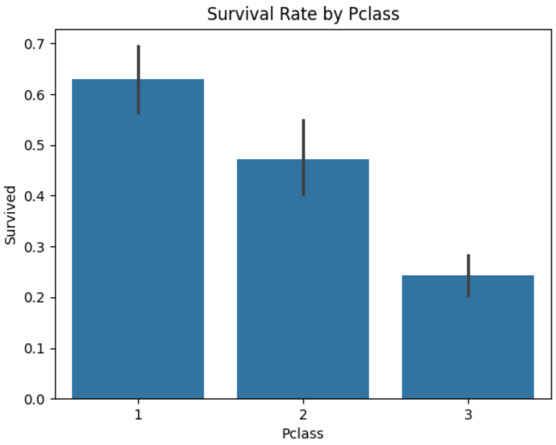


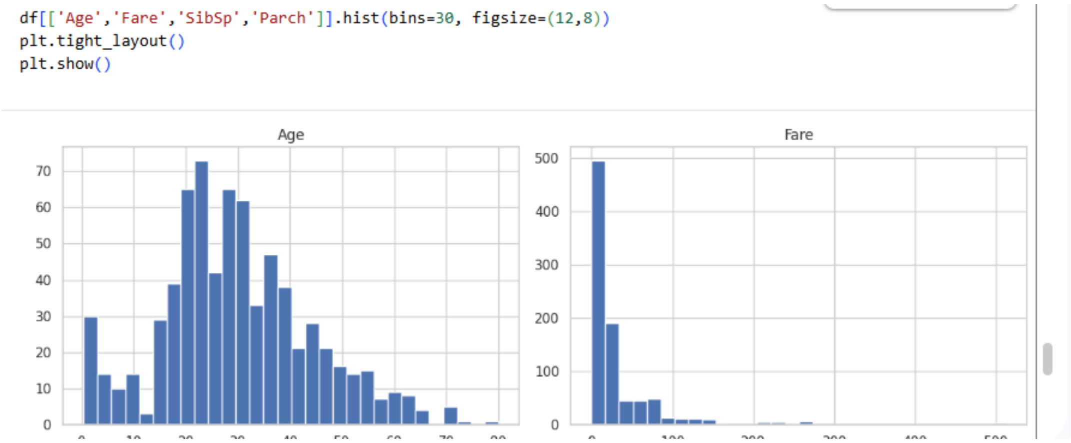
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from scipy import stats
```



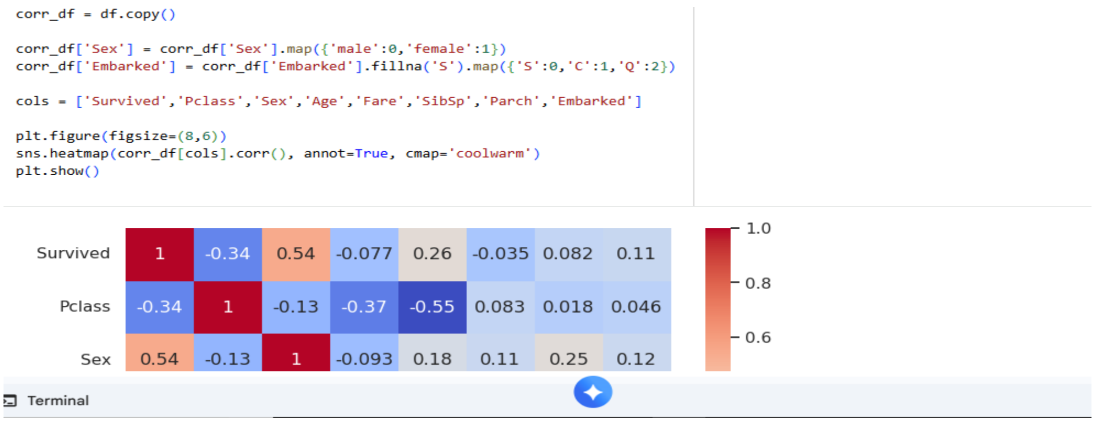






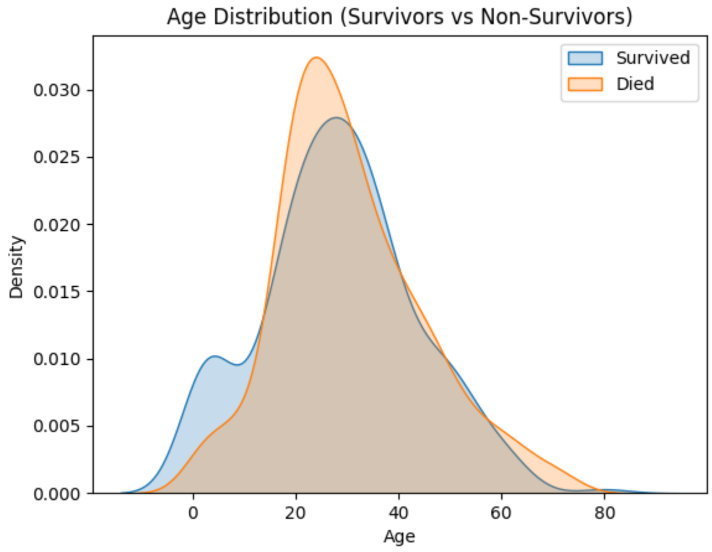


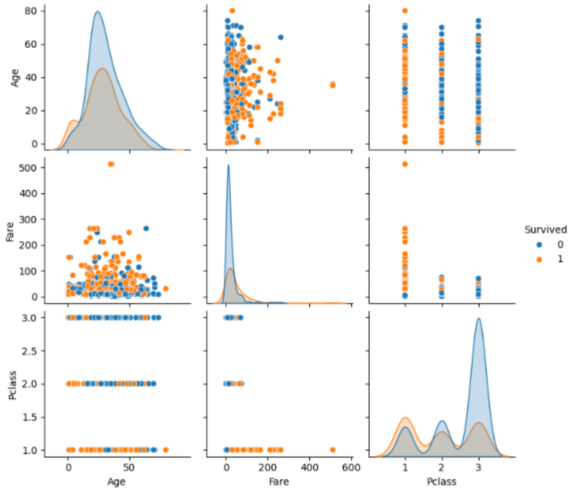
```
df[['Age','Fare','SibSp','Parch']].hist(bins=30, figsize=(12,8))
plt.tight_layout()
plt.show()
```

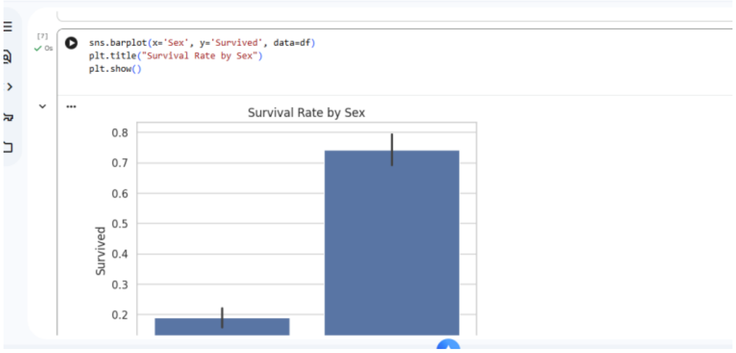


```
corr_df = df.copy()

corr_df['Sex'] = corr_df['Sex'].map({'male':0,'female':1})
corr_df['Embarked'] = corr_df['Embarked'].fillna('S').map({'S':0,'C':1,'Q':2})

cols = ['Survived','Pclass','Sex','Age','Fare','SibSp','Parch','Embarked']

plt.figure(figsize=(8,6))
sns.heatmap(corr_df[cols].corr(), annot=True, cmap='coolwarm')
plt.show()
```

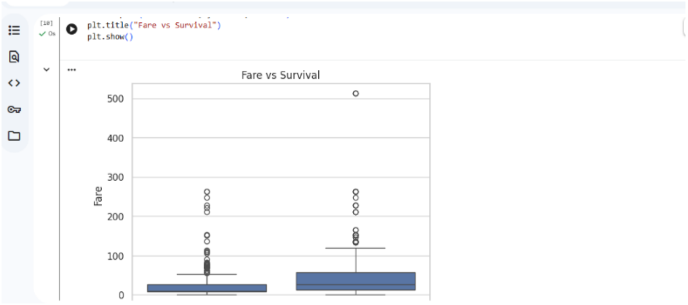
Survived	1	-0.34	0.54	-0.077	0.26	-0.035	0.082	0.11
Pclass	-0.34	1	-0.13	-0.37	-0.55	0.083	0.018	0.046
Sex	0.54	-0.13	1	-0.093	0.18	0.11	0.25	0.12



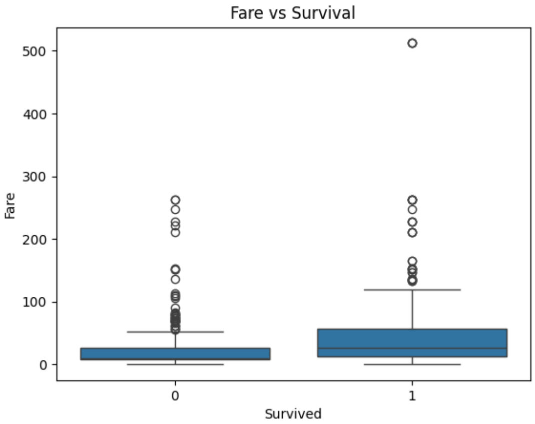


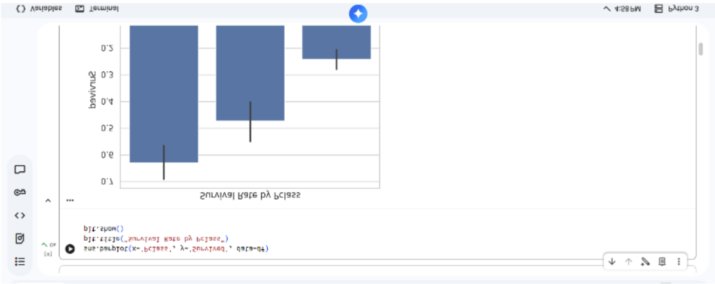


```
sns.barplot(x='Sex', y='Survived', data=df)
plt.title("Survival Rate by Sex")
plt.show()
```

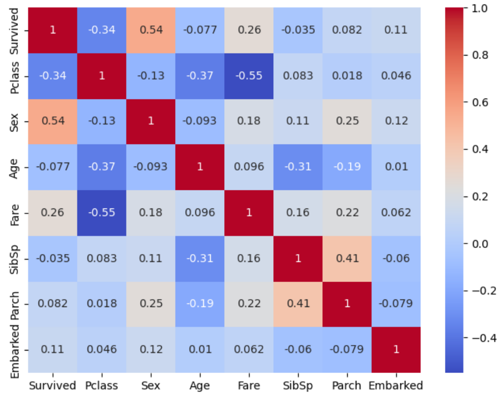


```
plt.title("Fare vs Survival")
plt.show()
```

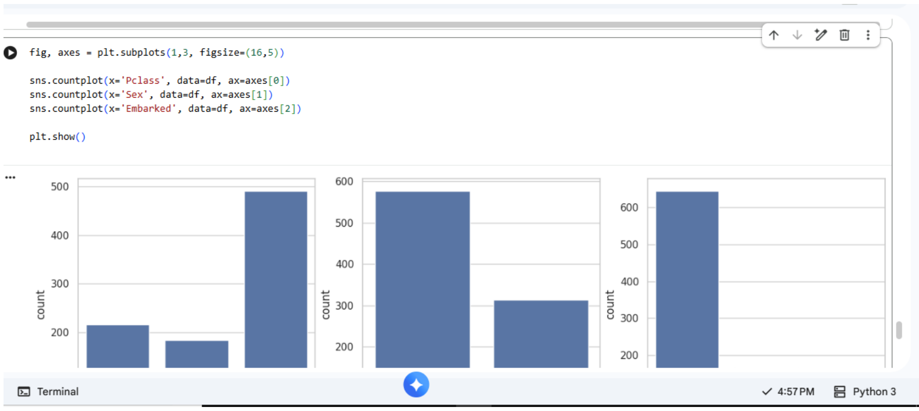




```
sns.barplot(x='Pclass', y='Survived', data=df)
plt.title("Survival Rate by Pclass")
plt.show()
```



	Survived	Pclass	Sex	Age	Fare	SibSp	Parch	Embarked
Survived	1	-0.34	0.54	-0.077	0.26	-0.035	0.082	0.11
Pclass	-0.34	1	-0.13	-0.37	-0.55	0.083	0.018	0.046
Sex	0.54	-0.13	1	-0.093	0.18	0.11	0.25	0.12
Age	-0.077	-0.37	-0.093	1	0.096	-0.31	-0.19	0.01
Fare	0.26	-0.55	0.18	0.096	1	0.16	0.22	0.062
SibSp	-0.035	0.083	0.11	-0.31	0.16	1	0.41	-0.06
Parch	0.082	0.018	0.25	-0.19	0.22	0.41	1	-0.079
Embarked	0.11	0.046	0.12	0.01	0.062	-0.06	-0.079	1



```
fig, axes = plt.subplots(1,3, figsize=(16,5))

sns.countplot(x='Pclass', data=df, ax=axes[0])
sns.countplot(x='Sex', data=df, ax=axes[1])
sns.countplot(x='Embarked', data=df, ax=axes[2])

plt.show()
```

Summary:

The EDA reveals strong influence of gender, class, and age on survival. Feature engineering improved data richness and helped understand passenger categories. The dataset is now clean and insights are ready for model building.