TITANIC DATASET: EXPLORATORY DATA ANALYSIS (EDA) REPORT

Introduction:

This report explores the Titanic dataset to uncover patterns affecting passenger survival. Using Python libraries such as Pandas, Seaborn, and Matplotlib, we analyze both numerical and categorical features through statistical summaries and visualizations.

A. Initial Exploration:

1) df.info()
Provides the structure of the dataset.
What it tells us:

- Dataset has 891 rows and 12 columns.
- Columns like Age, Cabin, and Embarked contain missing values.
- Data types include integers, floats, and strings (objects).

2) df.describe()
Gives statistical summaries for numeric columns.
Key findings:

- Average age: ~29.7 years.
- Fares range from free to over 500.
- Most passengers had few or no relatives onboard.

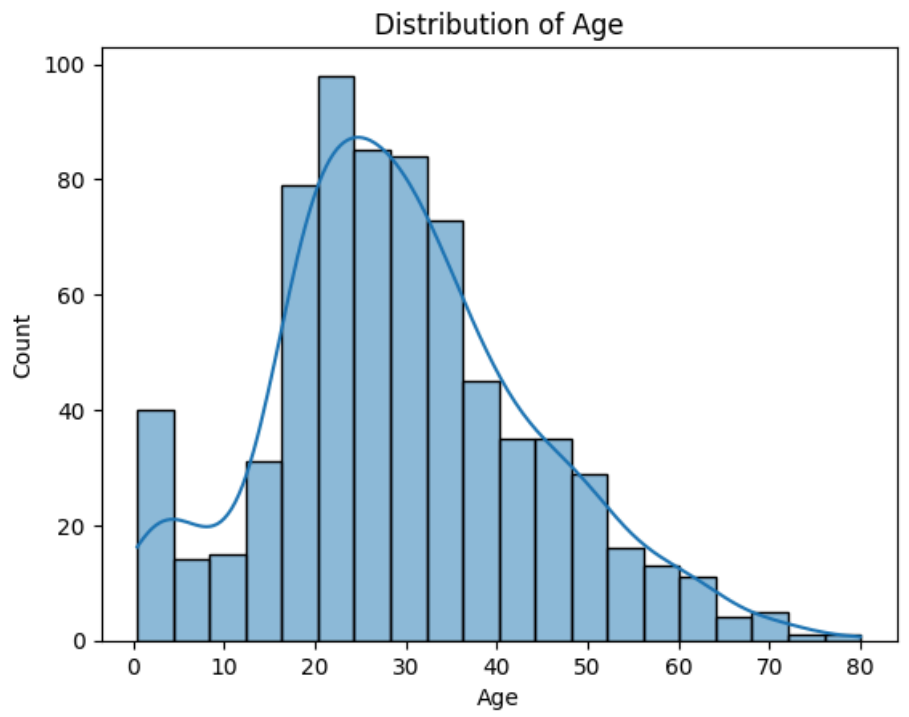
3) df.isnull().sum() and df.nunique()
Identifies missing values and unique counts.
Observations:

- Cabin has many missing values.
- Survived, Pclass, and Sex have a small number of unique values, making them suitable for grouping and comparison.

B. Visualizations & Observations:

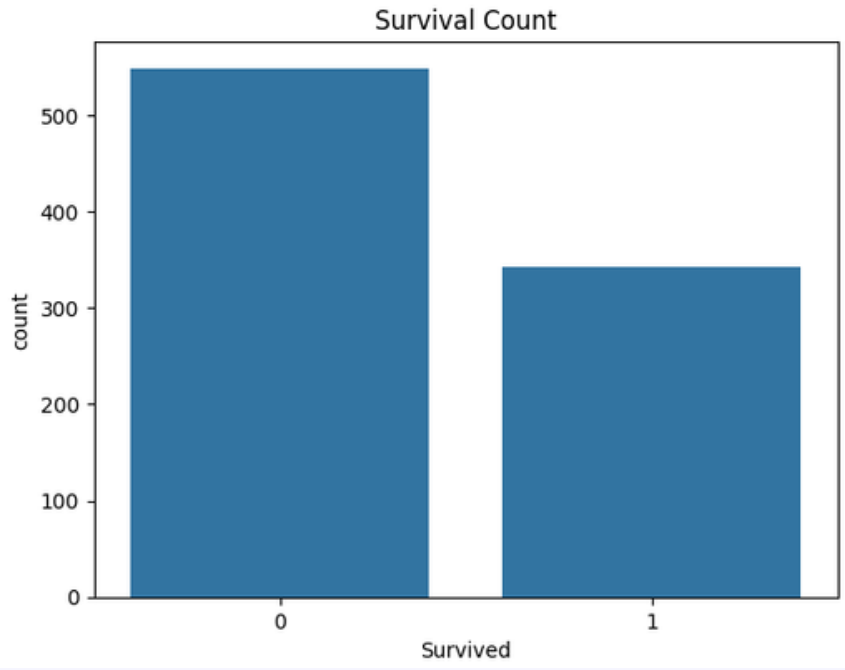
1. Age Distribution
Explanation: Histogram of Age with KDE overlay.
Observation:

- Most passengers were aged 20–40.
- Slight right skew; few passengers above 60.



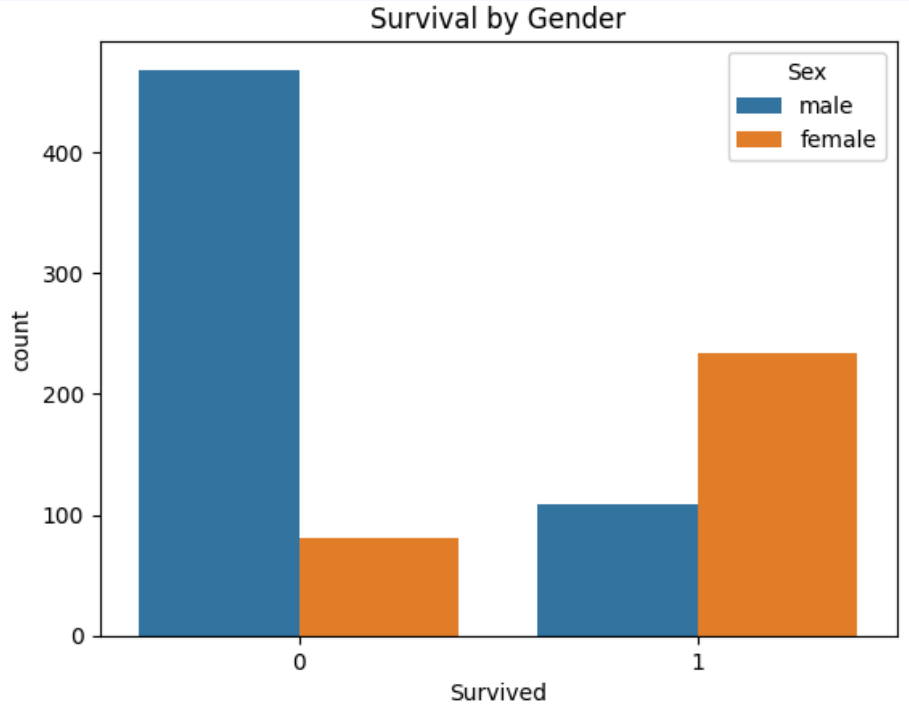
2. Survival Count
Explanation: Bar chart of survivors vs. non-survivors.
Observation:

- Majority did not survive the disaster.



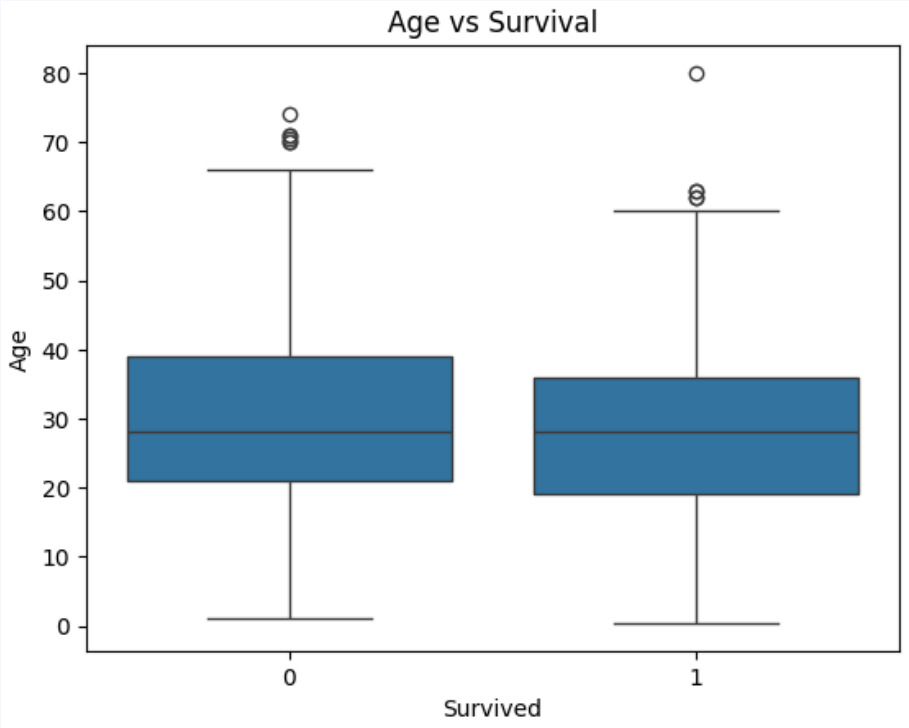
3. Survival by Gender
Explanation: Countplot comparing Survived across Sex.
Observation:

- Most women survived, while most men didn't.



4. Age vs. Survival (Boxplot)
Explanation: Boxplot of Age grouped by survival status.
Observation:

- Median age of survivors slightly lower.
- Survivors include a broader age range.

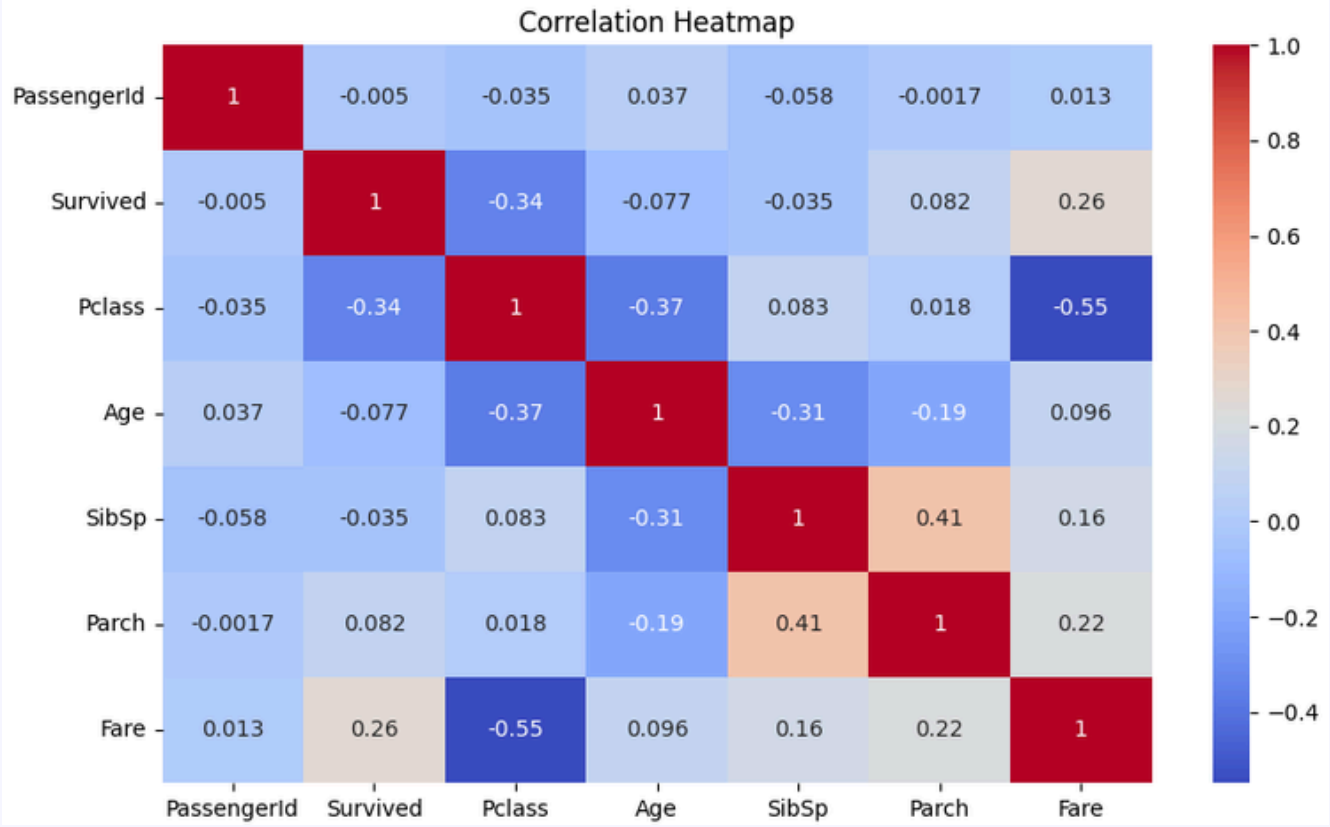


5. Correlation Heatmap

Explanation: Heatmap of numeric feature correlations.

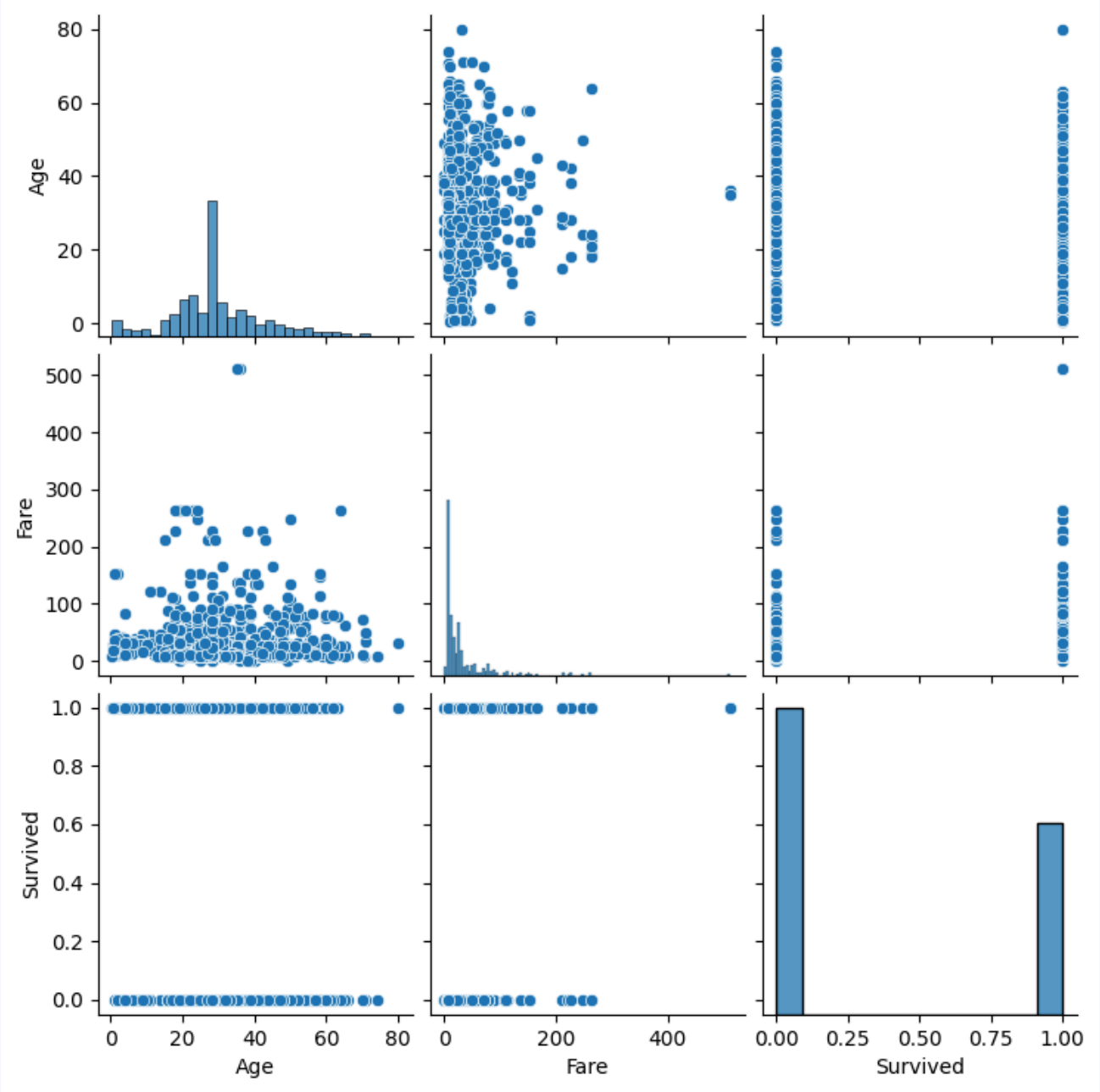
Observation:

- Fare positively correlates with Survived.
- Pclass negatively correlates with Survived.



6. Pairplot (Age, Fare, Survival)
Explanation: Multivariate pairplot for Age, Fare, and Survived.
Observation:

- Survivors tend to have higher fares and come from multiple age groups.
- Clusters suggest class and wealth played a role.



Conclusion

This analysis confirms that survival on the Titanic was influenced by several factors:

- Gender: Women had significantly higher survival rates.
- Fare: Higher fares correlated with survival.
- Class: First-class passengers had better chances.
- Age: Younger individuals showed slightly better survival, but age was less decisive than gender or class.

The combination of statistical summaries and visualizations provided meaningful insight into the tragedy and highlighted patterns valuable for predictive modeling or sociological analysis.