Titanic Dataset: Exploratory Data Analysis (EDA) Report

Introduction: This report explores the Titanic dataset to identify patterns related to passenger survival. We analyze numerical and categorical features using statistical summaries and visualizations.

A. Initial Exploration:

1) df.info()

Shows basic structure of the dataset.

What it tells us:

- There are 891 rows and 12 columns.
- Some columns like Age, Cabin, and Embarked have missing values.
- Data types are int64, float64, and object.

---

2) df.describe()

Gives statistics for numeric columns.

Key findings:

- Average age of passengers: ~29.7 years.
- Fare values vary widely – from 0 to over 500.
- Some passengers had 0 siblings/spouse or parents/children on board.

---

3) .value_counts()

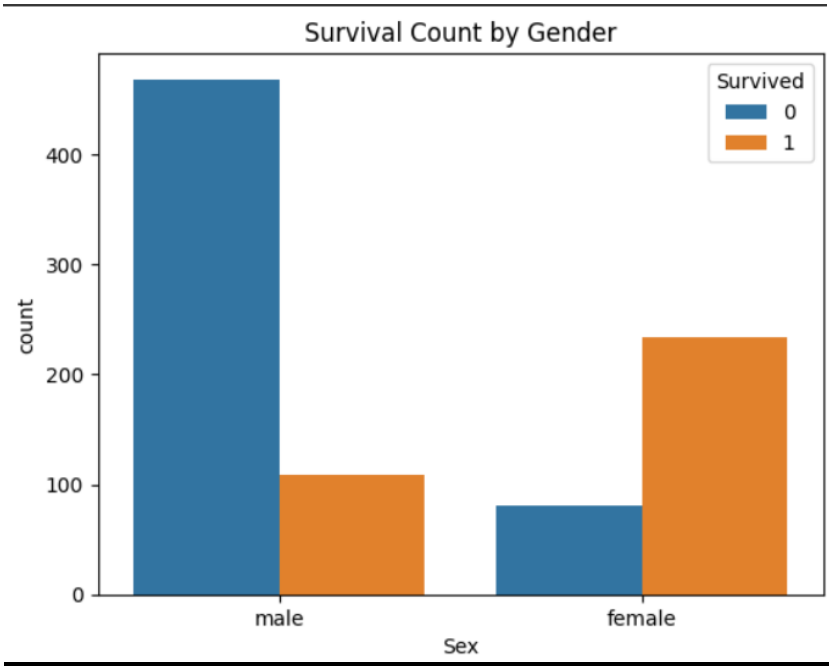
Used on columns like Survived, Pclass, Sex, and Embarked.

What we see:

- Majority didn’t survive (549 out of 891).
- Most passengers were in 3rd class.
- Males were more than females.
- Most people boarded from port S (Southampton).

B. Visualizations & Observations

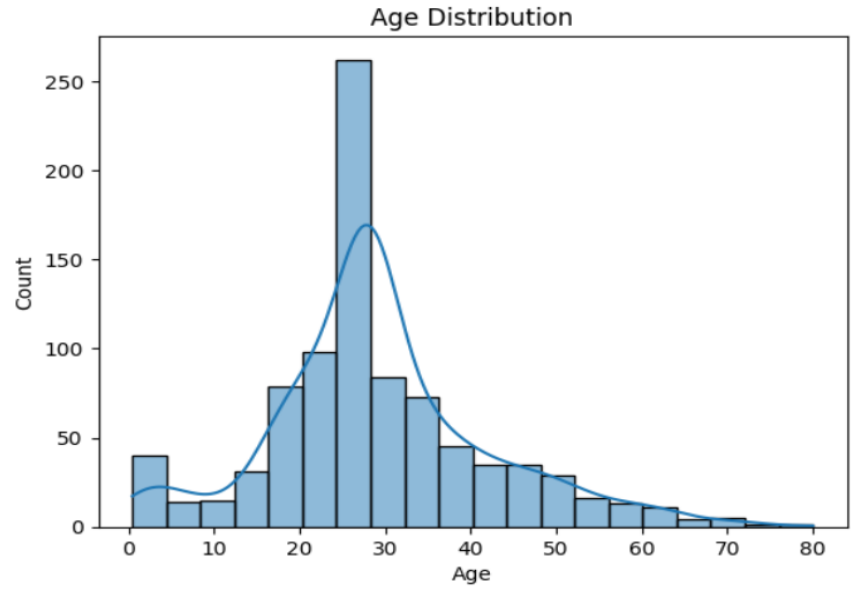
1. Survival Rate by Gender:



Explanation:
This shows how survival varies between males and females.
Observation:

- Most women survived.
- Most men did not survive.

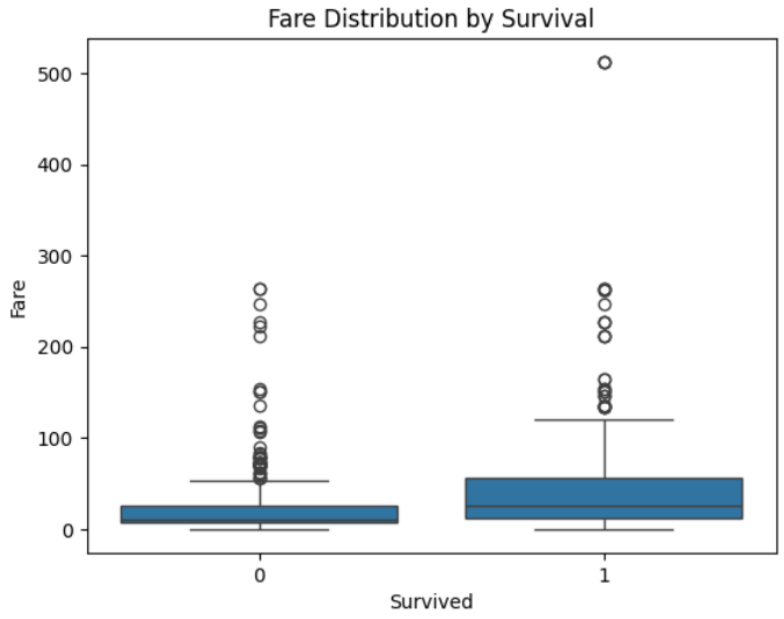
2. Age Distribution:



Explanation:
This shows how many passengers belonged to different age groups.
Observation:

- Most passengers were between **20–40 years**.
- A few children and older adults were on board.

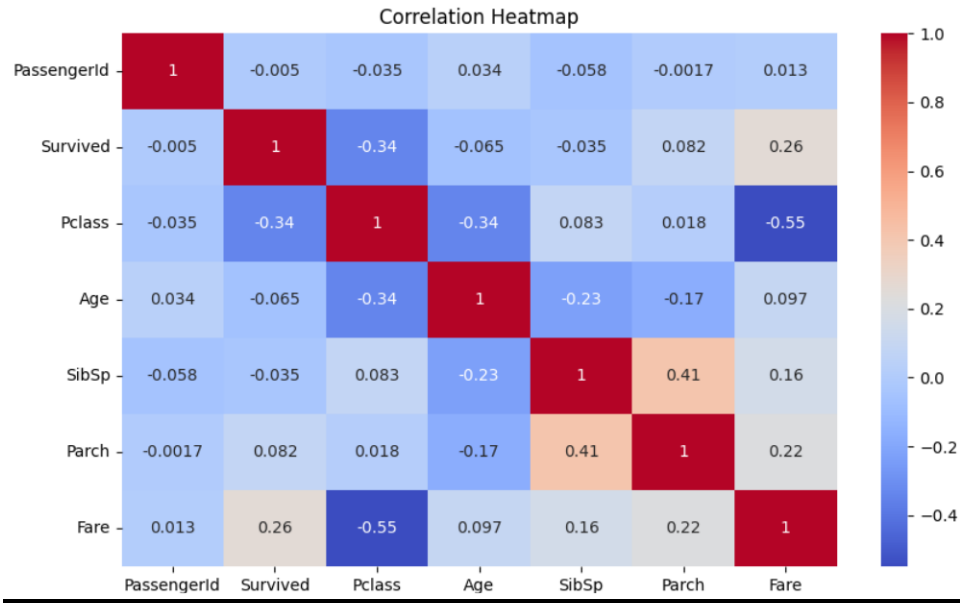
3. Fare vs Survival (Boxplot):



Explanation:
This compares the fares paid by those who survived vs. those who didn't.
Observation:

- Survivors paid **higher fares**.
- Suggests **wealthier people had better chances** of survival.

4. Correlation Heatmap:

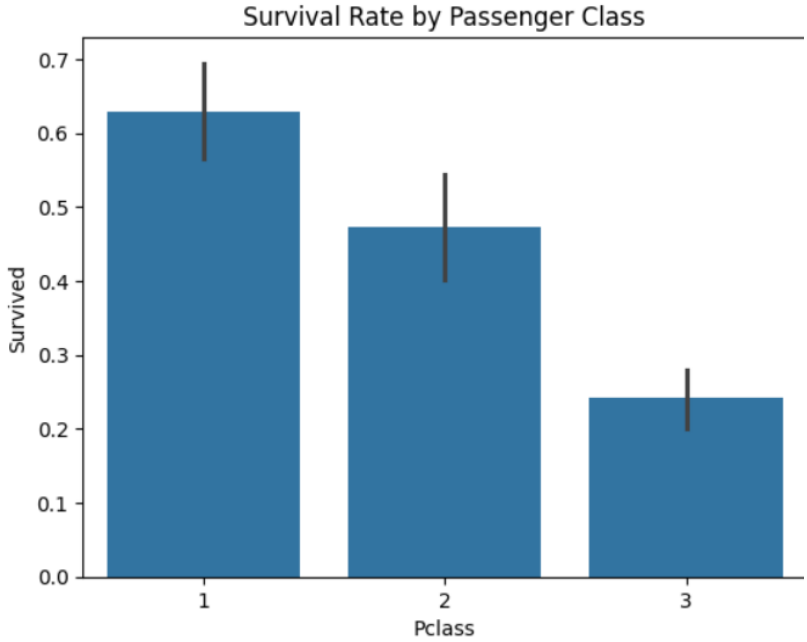


Explanation:
This shows how different numbers relate to each other.

Observation:

- Fare and Survived have a positive correlation.
- Pclass and Survived have a negative correlation.

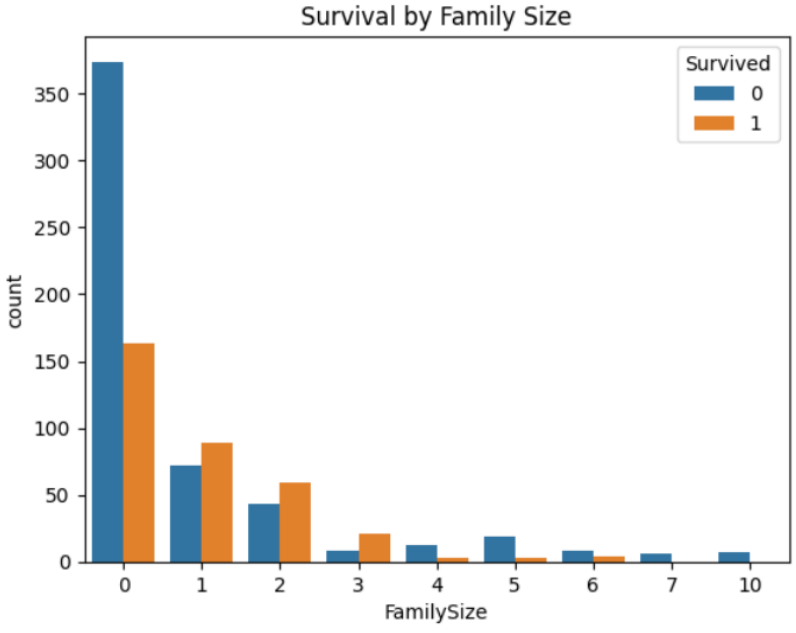
5. Passenger Class vs Survival:



Explanation:
Shows the average survival rate for each class.
Observation:

- 1st class: highest survival
- 3rd class: lowest survival

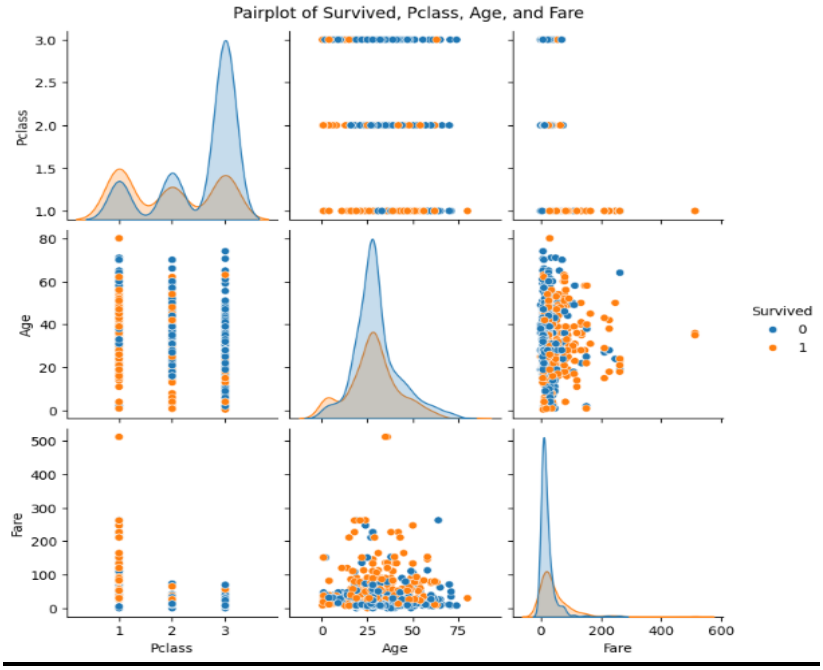
6. Family Size vs Survival



Explanation:
This looks at how survival changes with family size.
Observation:

- Small families (1–3 people) had higher survival rates.
- Very large families had lower survival.

7. Pairplot for Multivariate Analysis:



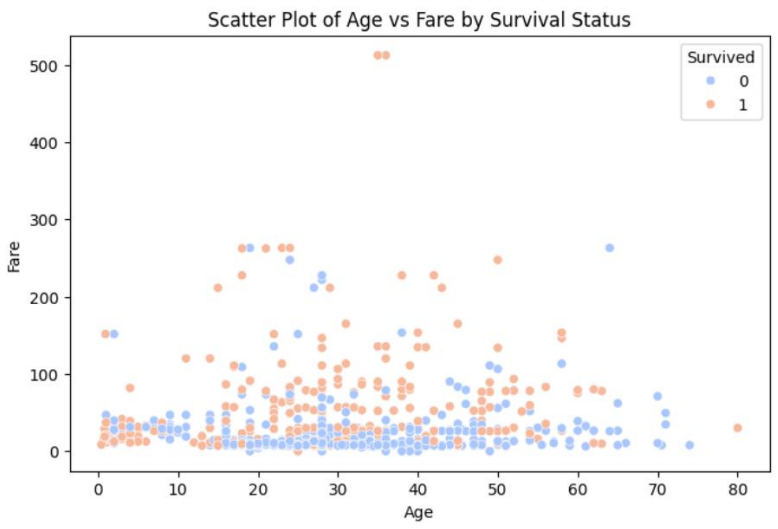
Explanation:

Shows relationships between multiple features at once.

Observation:

- Survivors mostly came from 1st class and paid higher fares.
- Most non-survivors were in 3rd class with low fare.

8. Scatterplot: Fare vs Age:



Explanation:

Looks at age and fare together, and how they affect survival.

Observation:

- Older people who paid more survived more.
- Some young passengers survived across all fare ranges.

Conclusion:

The exploration of the Titanic dataset data resulted in the identification of survival-influencing patterns among passengers. The probability of survival for Titanic passengers depended heavily on gender as well as their passenger class and fare paid and family size. The connection between variables became evident through visual representations that demonstrated their relationships in an understandable way. The obtained insights deepen our comprehension of the disaster while demonstrating that data-based decision-making holds value in practical situations.