Titanic Dataset: Exploratory Data Analysis (EDA)

🌟 Introduction: Setting the Scene.

The sinking of the Titanic is one of the most tragic events in maritime history, and the dataset containing information about the passengers provides a unique opportunity to analyze the factors that influenced survival. The goal of this report is to perform an **Exploratory Data Analysis (EDA)** on the Titanic dataset to uncover insights into the factors that contributed to survival chances, with a focus on **gender, class, age**, and **family size**.

🌟 Objective: Extract insights using visual and statistical exploration.

🧹 Data Cleaning and Preprocessing: Getting Things Ready

Before diving into the analysis, we cleaned and preprocessed the data to ensure accuracy and quality:

- **Handling Missing Values**:
    - The **Age** column had missing values, which were filled with the **mean** of the column to avoid biased results.
    - The **Cabin** column had many missing values and was **dropped** due to its lack of useful data.
    - Missing values in the **Embarked** column were filled with the **mode** (most frequent value).
- **Data Type Fixes**: Ensured that all columns had the correct data types for analysis (e.g., categorical variables like 'Sex' were converted to appropriate formats).

📊 Exploratory Data Analysis (EDA): Uncovering Insights

🧭 Univariate Analysis: Understanding Individual Features

- **Distribution of Age:**
    - The **Age** distribution was **right-skewed**, with most passengers being in their 20s and 30s. Children and young adults were more prominent.

- **Distribution of Fare:**
    - A **right-skewed distribution** for **Fare** indicated that most passengers paid lower fares, while a small group of passengers paid much higher fares for luxury accommodations.
- **Passenger Class (Pclass):**
    - The majority of passengers were in **3rd class**, which could reflect the socioeconomic conditions of the time.
- **Survival Distribution:**
    - **Survival Rate**: Only about 38% of passengers survived. A stark reminder of the disaster's scale.

🔄 Bivariate Analysis: Exploring Relationships

- **Age vs. Pclass:**
    - A **boxplot** revealed that 1st class passengers were generally **older** than 2nd and 3rd class passengers.
- **Pairplot:**
    - **Age vs. Fare**: A slight **positive correlation** was seen. Older passengers tended to pay more for tickets.
    - **Pclass vs. Survived**: Passengers in **1st class** had a much higher survival rate than those in **3rd class**.

💡 Correlation Insights: What Matters Most?

- The **heatmap** visualized the correlation between various features:
    - **Fare and Pclass**: A strong **negative correlation** was found. Higher-class passengers paid more for their tickets.
    - **Survived and Pclass**: A moderate **negative correlation** suggested that passengers in **lower classes** had a lower chance of survival.

🚺 Survival Analysis: Key Factors at Play

👩 Survival Rate by Gender

- **Women vs. Men**:
    - **Women** had a **74% survival rate**, while **men** had only a **19% survival rate**.
    - **Key Insight**: The famous "**Women and children first**" protocol was clearly evident, with a much higher chance of survival for women.

🚃 Survival Rate by Passenger Class

- **Class Disparities**:
    - **1st class passengers** had the highest survival rate, followed by **2nd class**. **3rd class** had the lowest survival rate.
    - **Key Insight**: Class played a major role in survival, with higher-class passengers receiving better treatment and prioritization during the evacuation.

📝 Summary of Key Findings

- **Gender**: Women had a **significantly higher** chance of survival than men.
- **Class**: Higher-class passengers (1st and 2nd) were much more likely to survive than those in 3rd class, likely due to better access to lifeboats and evacuation protocols.
- **Age**: Children had a better survival rate compared to adults.
- **Family Size**: Passengers with smaller families (either alone or with just one family member) had slightly better chances of survival than those traveling in larger family groups.

🟦 Step

Step 1: Data Preprocessing

- **Action: Handling missing values, fixing data types, and feature engineering.**
- **Observations:**
    - **Age had missing values, filled with the mean.**

- **Embarked had missing values, filled with the mode.**
- **Cabin was dropped due to excessive missing values.**

- **Visualization: You can insert a data summary or missing value heatmap here if you generated one during the cleaning process.**

Step 2: Univariate Analysis

- **Action: Exploring individual features.**
- **Observations:**
  - **Age: Most passengers were in their 20s and 30s.**
  - **Fare: Most passengers paid lower fares, but some paid a lot more for luxury.**
  - **Passenger Class (Pclass): Majority were in 3rd class.**
  - **Survival Distribution: 38% survival rate.**
- **Visualizations:**
  - **Histograms for Age and Fare distributions.**
  - **Barplot for Survival distribution.**

Step 3: Bivariate Analysis

- **Action: Exploring relationships between pairs of variables.**
- **Observations:**
  - **Age vs. Pclass: 1st class passengers were older.**
  - **Age vs. Fare: Positive correlation, older passengers paid higher fares.**
  - **Pclass vs. Survived: 1st class had the highest survival rate.**
- **Visualizations:**
  - **Boxplot for Age vs. Pclass.**
  - **Pairplot for correlations between numerical features.**

Step 4: Correlation Analysis

- **Action: Analyzing correlations between numerical features.**

- **Observations:**
    - **Fare and Pclass: Strong negative correlation (-0.55).**
    - **Survived and Pclass: Moderate negative correlation (-0.34).**
    - **Survived and Fare: Positive correlation (0.26).**
- **Visualization: Heatmap showing correlations between numerical features.**

Step 5: Gender and Survival

- **Action: Analyzing survival rate by gender.**
- **Observations:**
    - **Females: ~74% survival rate.**
    - **Males: ~19% survival rate.**
    - **Significant difference in survival rates, highlighting societal protocols like "women and children first."**
- **Visualizations:**
    - **Barplot showing survival rate by gender (with annotated survival rates).**

Step 6: Survival by Passenger Class

- **Action: Analyzing survival rate by passenger class.**
- **Observations:**
    - **1st Class: Highest survival rate.**
    - **3rd Class: Lowest survival rate.**
    - **Reflects the inequality of the time.**
- **Visualizations:**
    - **Barplot showing survival rate by passenger class.**

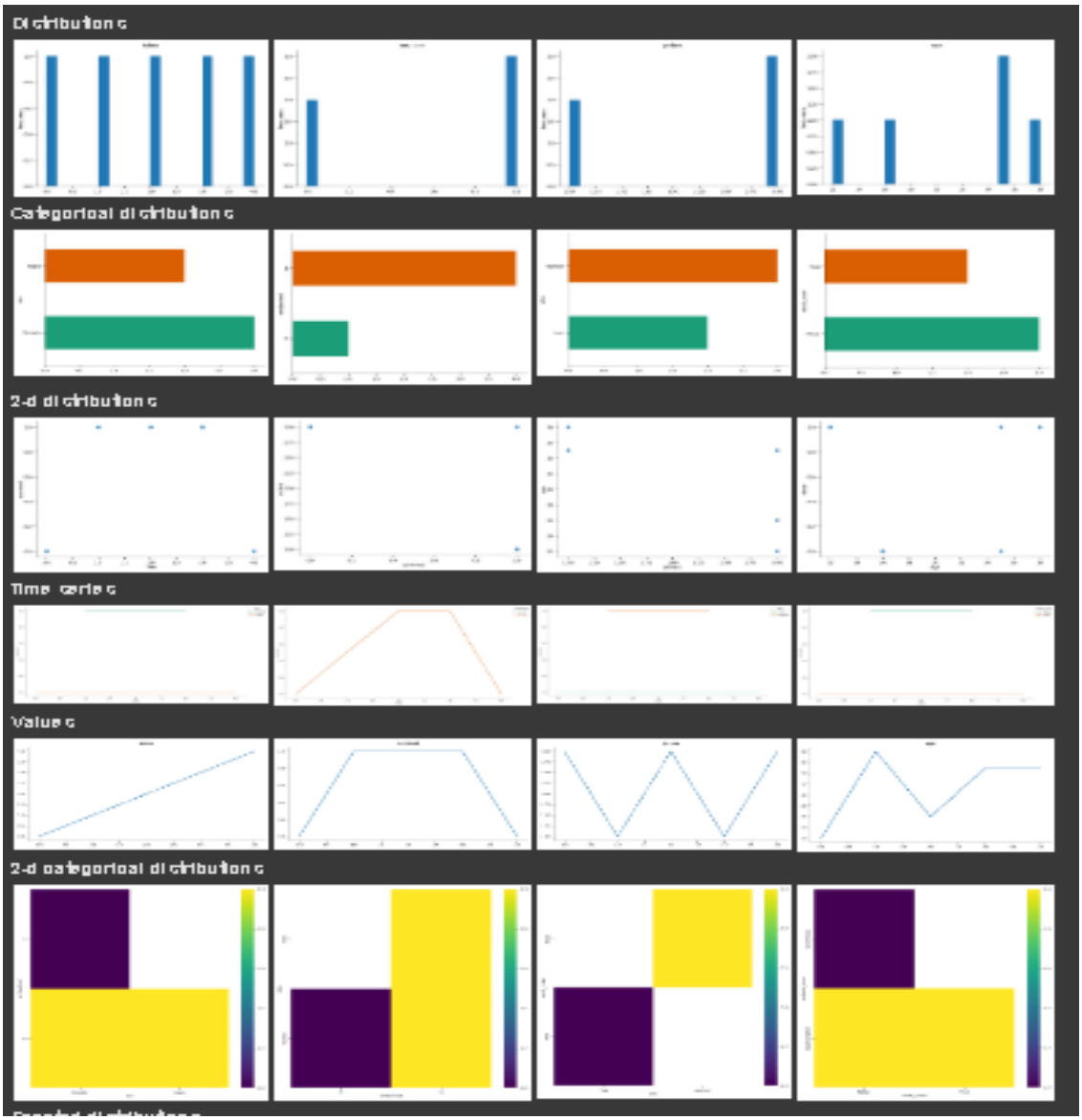
📸 **Visuals and Screenshots**

For each step, you should include relevant screenshots or visualizations such as:

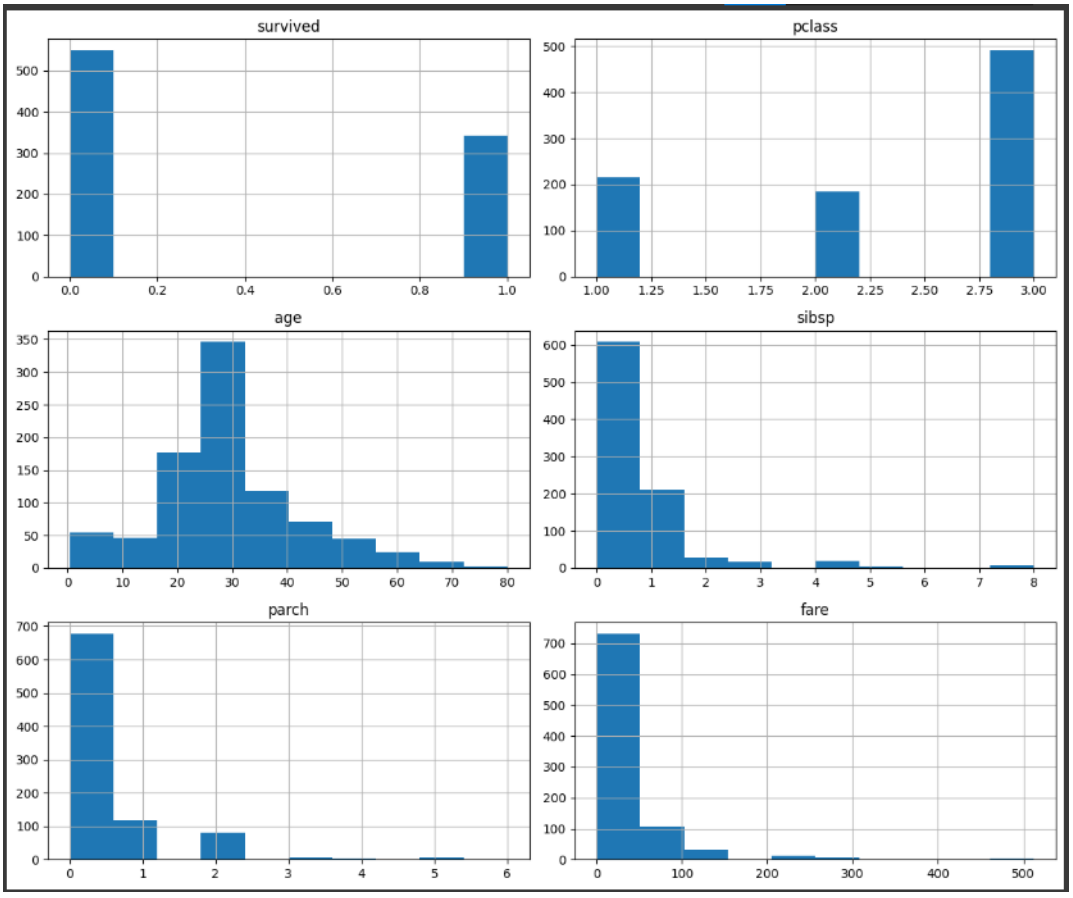
1. Data Summary: Initial look at the dataset before and after preprocessing.

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True

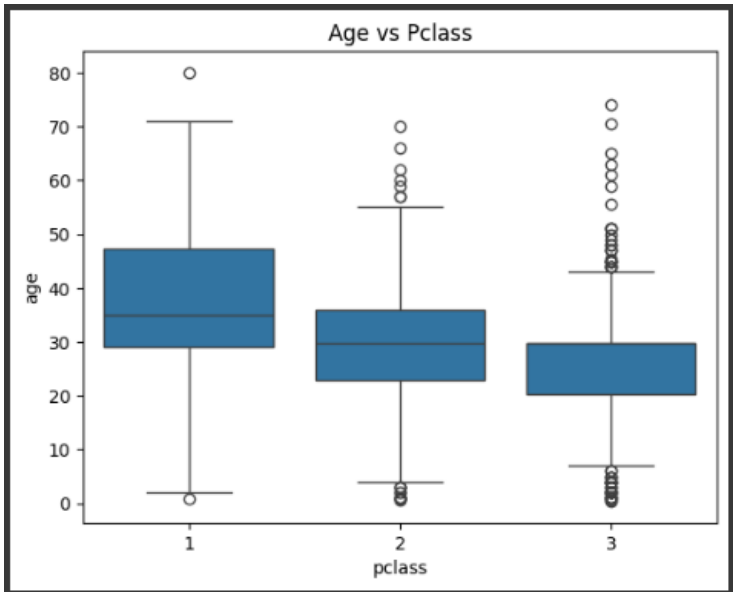
index	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
0	0	3	male	22.0	1	0	7.25	S	Third	man	true	NaN	Southampton	no	false
1	1	1	female	38.0	1	0	71.2833	C	First	woman	false	C	Cherbourg	yes	false
2	1	3	female	26.0	0	0	7.925	S	Third	woman	false	NaN	Southampton	yes	true
3	1	1	female	35.0	1	0	53.1	S	First	woman	false	C	Southampton	yes	false
4	0	3	male	35.0	0	0	8.05	S	Third	man	true	NaN	Southampton	no	true



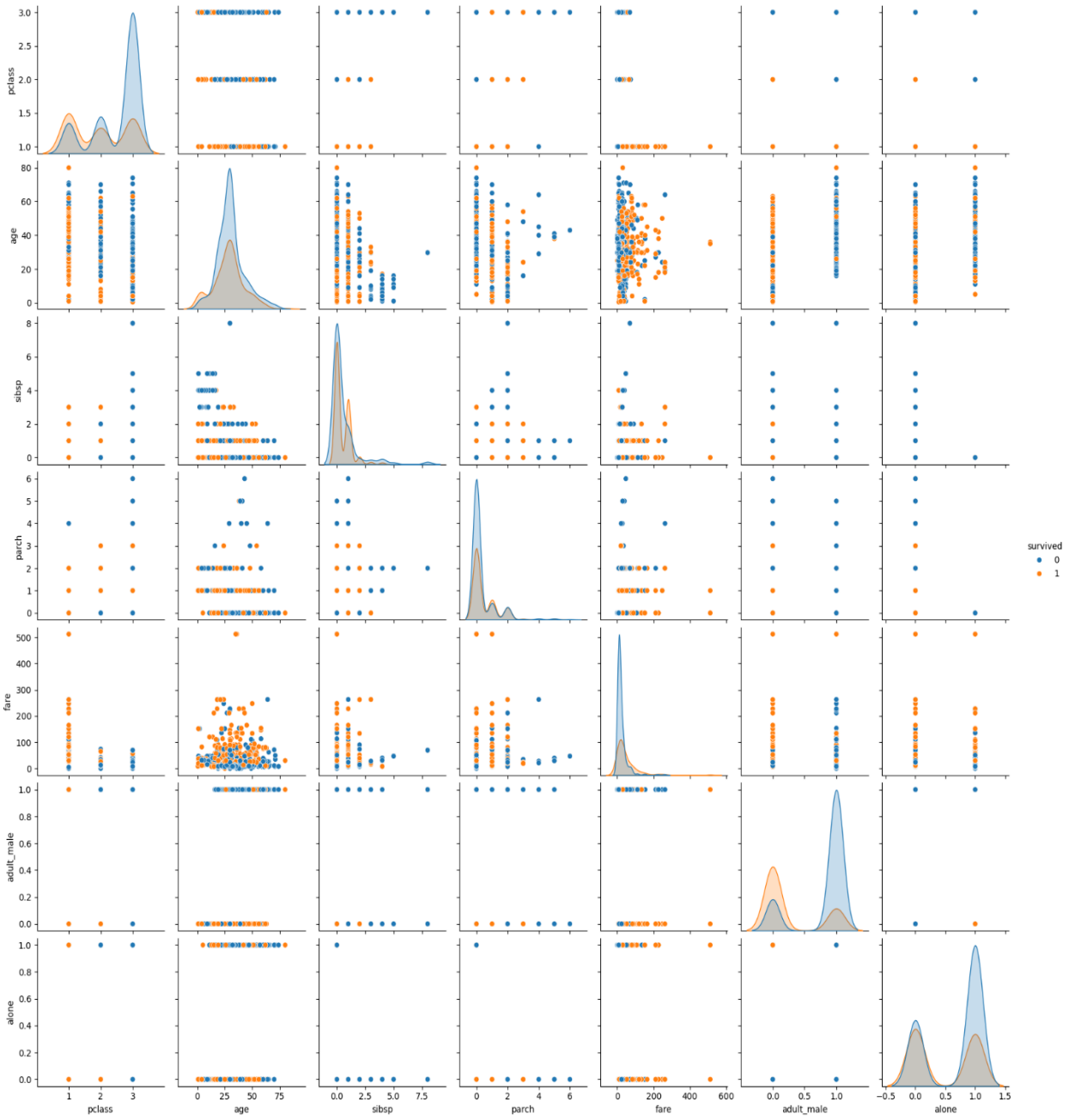
2. **Histograms: For age, fare, and survival distributions.**



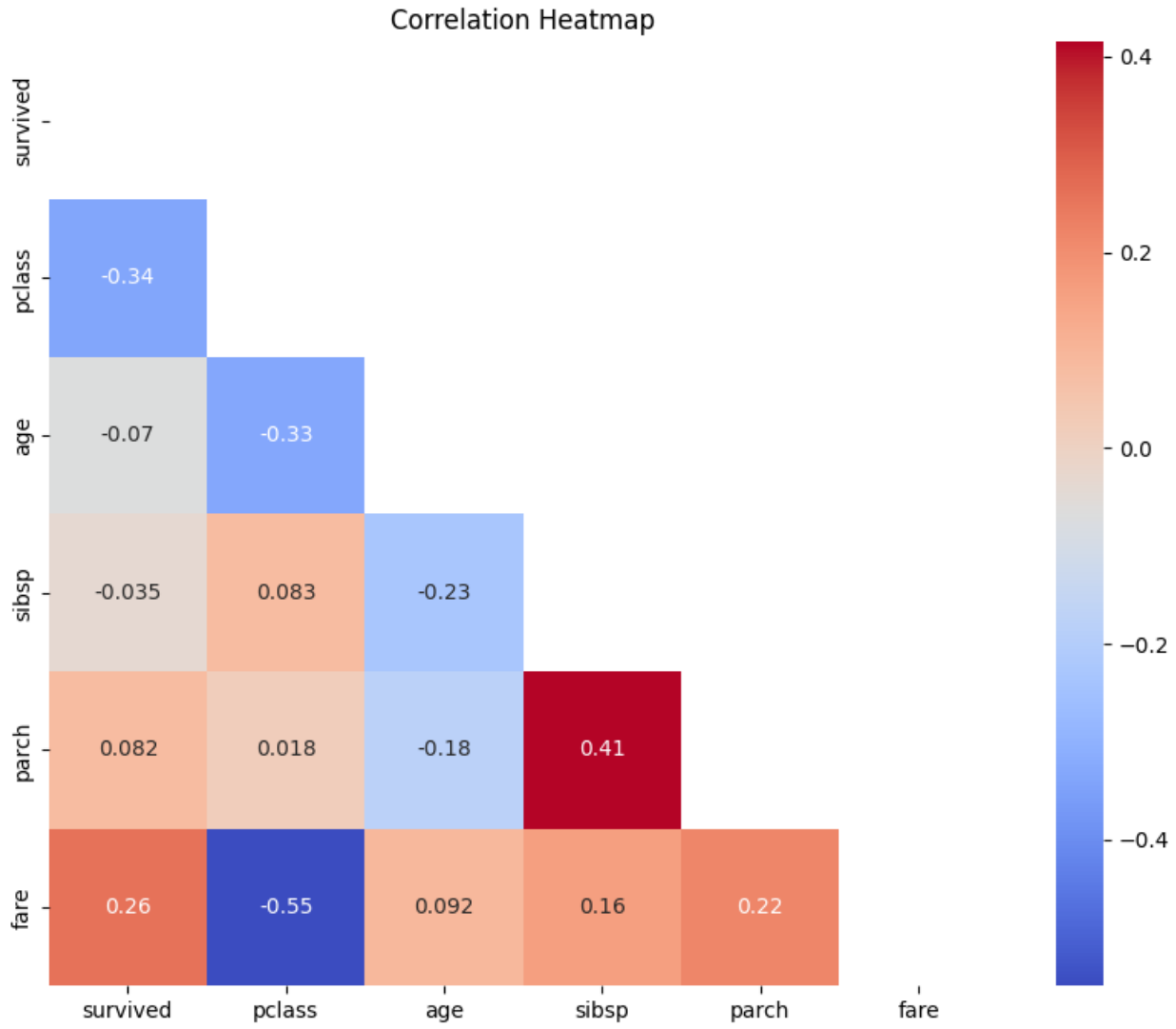
3. **Boxplots: Age vs. Pclass, for example.**



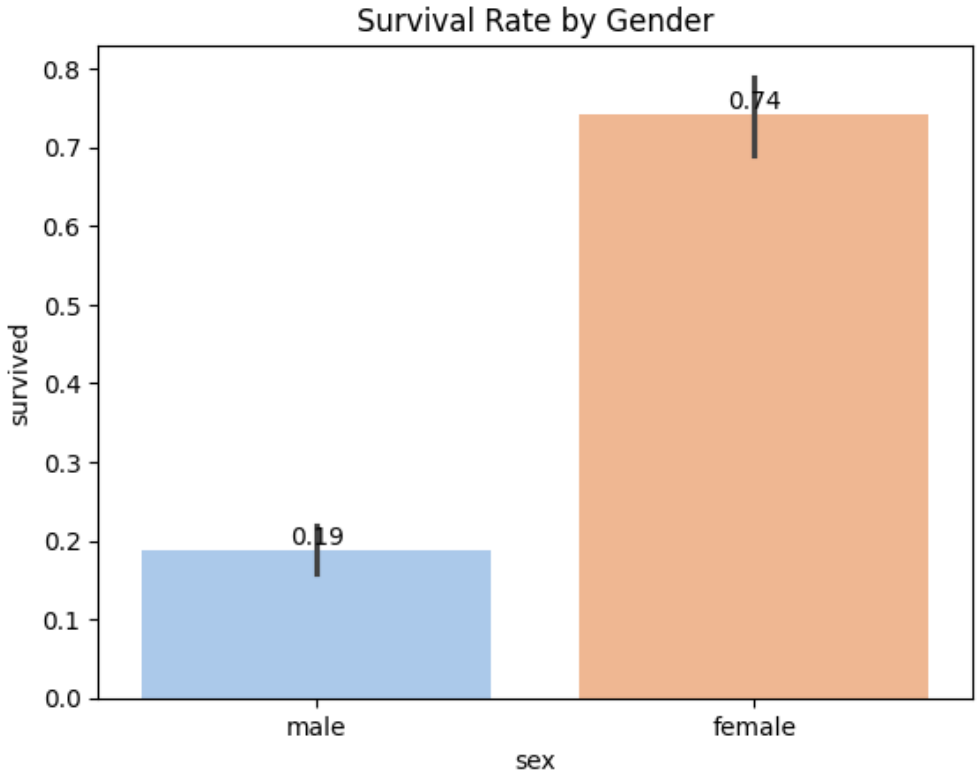
4. Pairplot: To show relationships between variables.

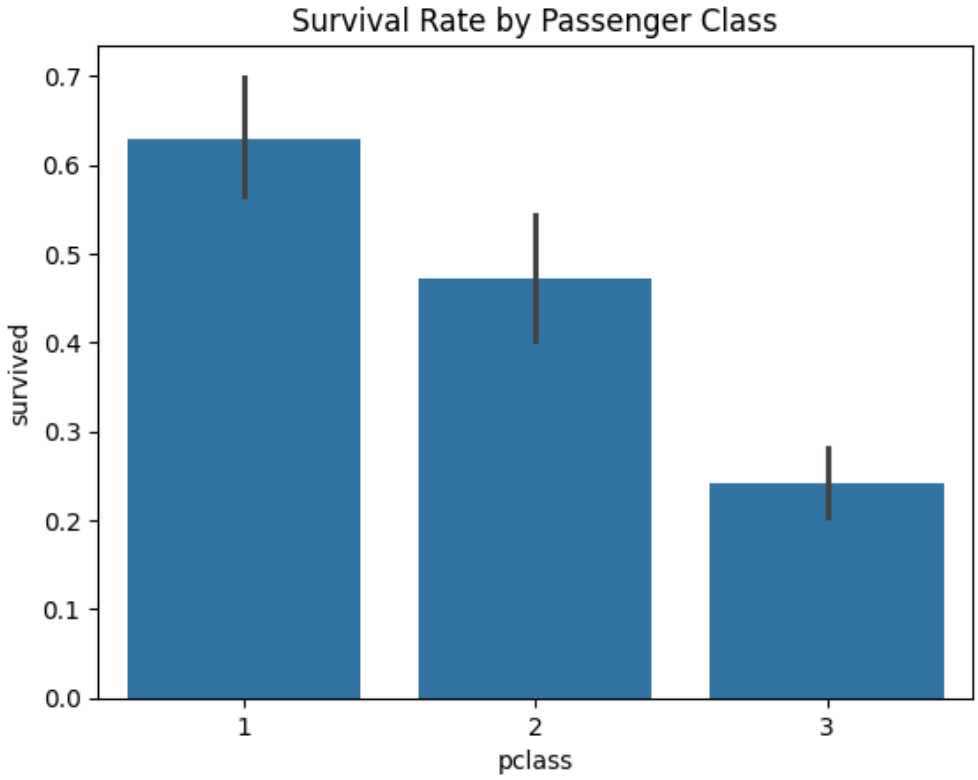


5. **Heatmap: Showing correlations.**



6. **Barplots: Survival rates by gender and class, with annotations showing exact percentages.**





🚀 Conclusion: Lessons from the Past

The **Titanic disaster** wasn't just a tragedy in terms of loss of life, but also a reflection of the **social inequalities** of the time. This analysis highlights how factors like **gender**, **class**, and **age** played a significant role in determining who survived and who didn't.

The insights gained from this analysis shed light on how certain **societal norms** and **social class** disparities influenced the chances of survival, making this a powerful study of **human behavior during crises**.

✍️ Final Thoughts

With this analysis, you now have a deeper understanding of the **Titanic dataset**. These findings pave the way for further exploration, such as creating predictive models based on these insights. This report serves as a foundation for analyzing how **socioeconomic factors** and **social behavior** played critical roles in the survival of passengers on the Titanic.