📝 Titanic EDA Summary Report

📌 Objective

The goal of this exploratory data analysis (EDA) is to extract key insights from the Titanic dataset using statistical summaries and visualizations.

📁 Dataset Overview

- **Source**: Kaggle Titanic Competition
- **Total Entries**: 891
- **Main Features**: `Age`, `Sex`, `Pclass`, `Fare`, `SibSp`, `Parch`, `Embarked`, `Cabin`, `Survived`
- **`Target Variable`**: `Survived (0 = No, 1 = Yes)`
-

✅ Data Cleaning (Manual)

- **Age**: Missing values replaced with median value (28).
- **Cabin**: Missing values replaced with **"Unknown"**, acknowledging that ~77% data is missing but retaining column for potential future use.

🔍 Exploratory Steps

1. Basic Info

- `info()` showed:
  Several non-null categorical and numerical columns.
  `Cabin` had many missing entries.
- `.describe()` showed:
  Age ranged from 0.42 to 80 years.
  Fare ranged from 0 to over 500, with high skew.

2. Missing Values

Column	% Missing
Age	0% (cleaned)
Cabin	~77% → replaced with `"Unknown"`
Embarked	~0.2%

📊 Visual Analysis

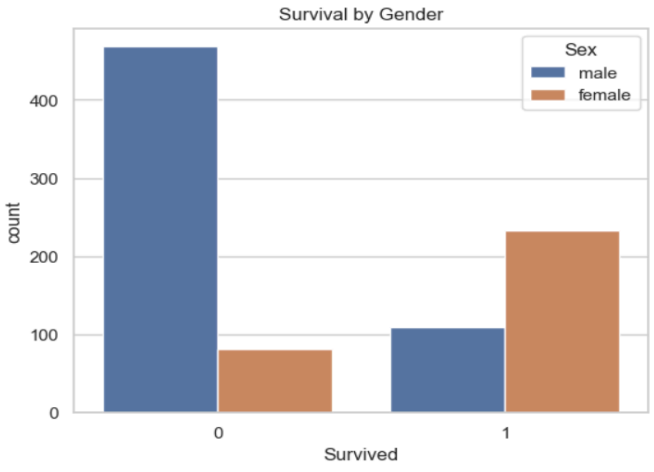
🔹 Survival Count

- **More people died than survived**.
- Only ~38% survived overall.

🔹 Survival by Gender

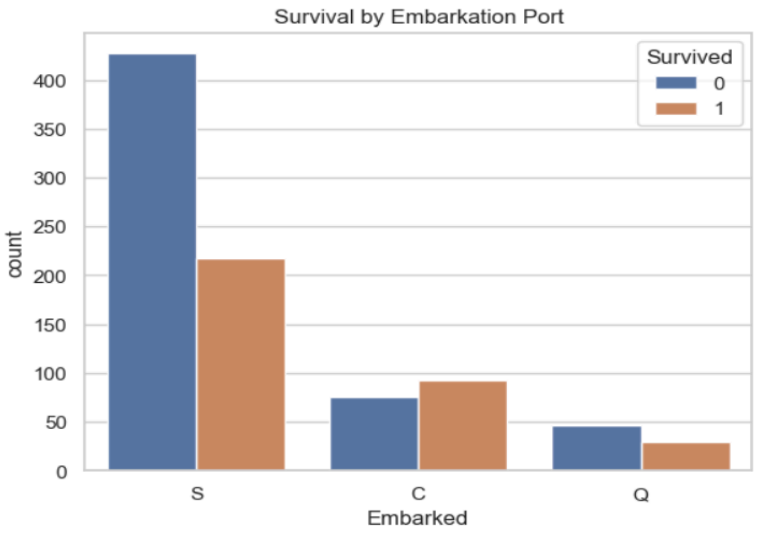
Using `sns.countplot()`:

- **Female survival rate**: ~74% ("women and children first" principle.)
- **Male survival rate**: ~19%



🔹 Survival by Class

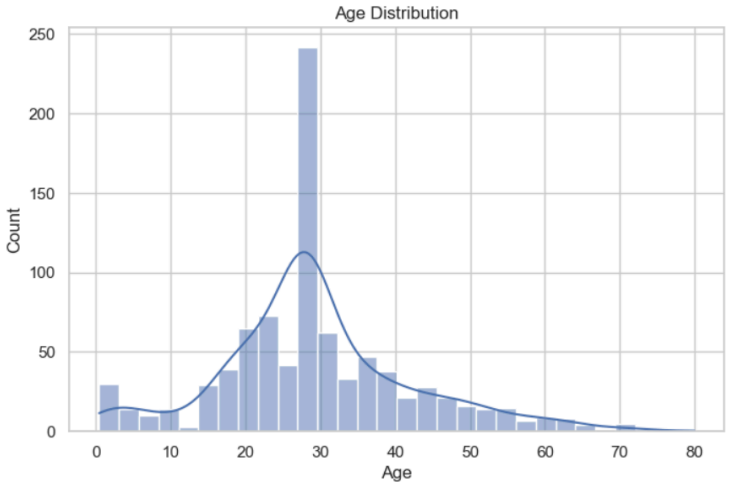
- **Higher class = higher chance of survival**
- 1st Class: ~63% survived
- 3rd Class: ~24% survived



🔹 Age Distribution

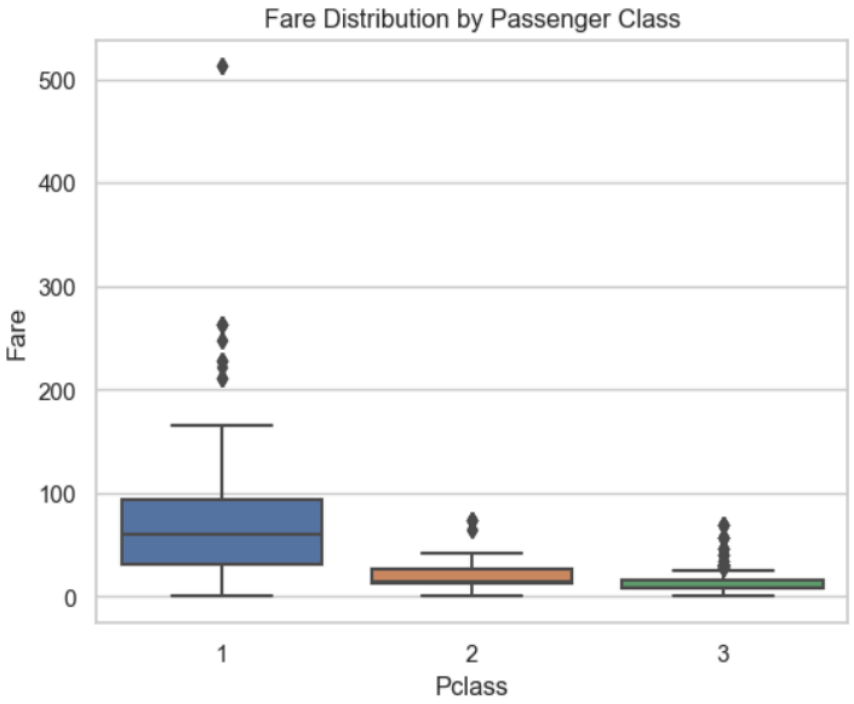
Using histogram (`sns.histplot()`):

- Most passengers are between **20-40 years old**.
- Survivors had slightly lower median age.



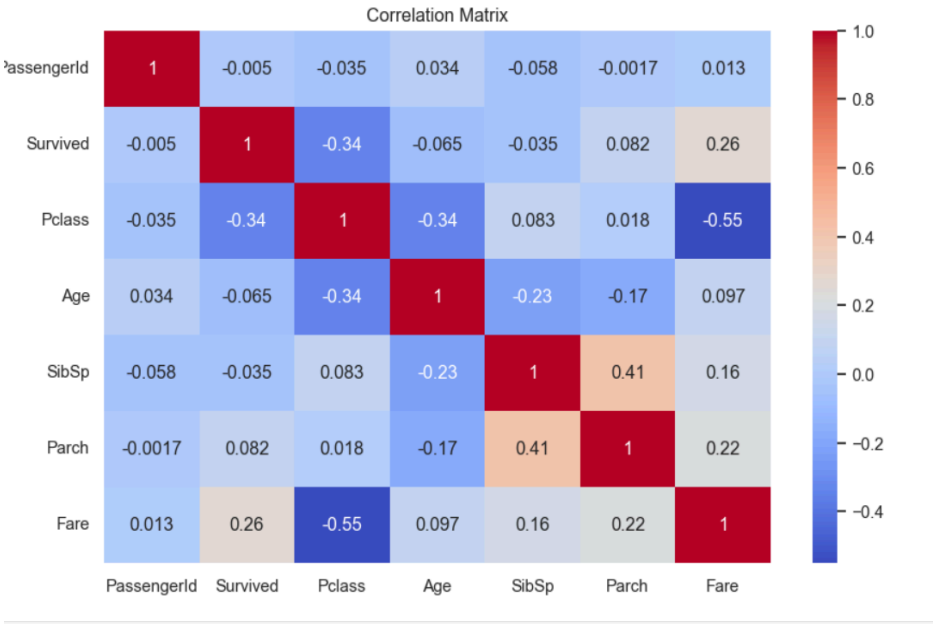
🔹 Fare Distribution

- The **Fare** column is right-skewed.
- A few passengers paid very high fares (up to 500+).
- Median fare was around 14.45.



🔹 Correlation Heatmap

- Strong positive correlation between **Fare and Survival**.
- Strong negative correlation between **Pclass and Survival** (higher class = lower number = higher chance).
- Sex (encoded) also correlated well with survival.



📌 Key Insights

- Sex and Pclass are strong predictors of survival.
- Children and women in higher classes had the best survival rates.
- Many features like `Cabin` may hold potential but are too sparse.

🧾 Conclusion

This EDA confirms well-known trends: survival on the Titanic was influenced by **gender**, **passenger class**, and **fare**. Missing data was handled thoughtfully to preserve statistical integrity, making the analysis reliable for further modeling or storytelling.