TASK 5 : Exploratory Data Analysis on the Titanic Dataset

Objective:

- To explore the Titanic dataset using statistical and visual methods to identify patterns, relationships, and anomalies that may help predict passenger survival.

Dataset Description:

- **Source:** Kaggle Titanic – Machine Learning from Disaster dataset
- **Files Used:** train.csv, test.csv, gender_submission.csv
- **Rows & Columns (train):** ~891 rows × 12 columns
- **Main Features:**
    - Survived (target), Pclass, Sex, Age, Fare, Embarked, SibSp, Parch, Cabin, Ticket
- **Date Range:** 1912 Titanic voyage passenger data.

Data Quality Summary:

 Missing values:

- Age: ~20% missing → filled with median age
- Embarked: 2 missing → filled with most frequent port (S)
- Fare (test set): 1 missing → filled with median fare

 Dropped or transformed:

- Cabin mostly missing → excluded from analysis
- Created new features:
    - FamilySize = SibSp + Parch + 1
    - IsAlone = 1 if FamilySize == 1, else 0

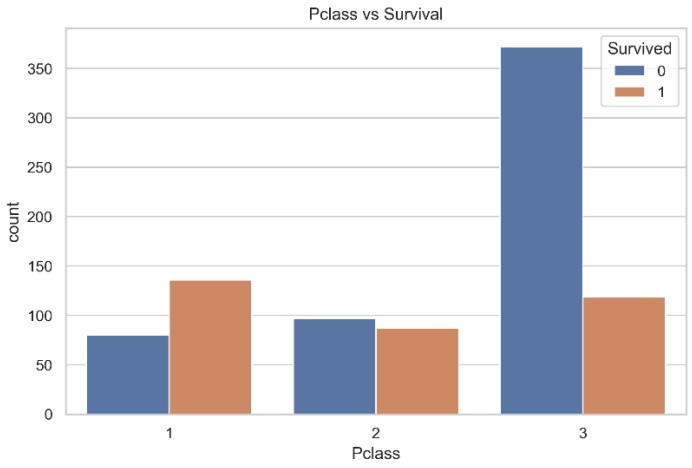
Key Visuals with Observations:

1. Survival Rate by Sex

Females had ~74% survival rate compared to ~19% for males.

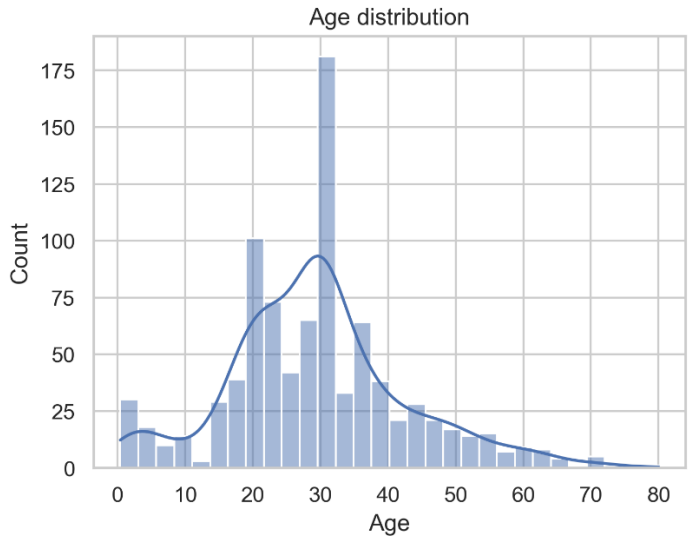
2. Survival Rate by Passenger Class

First-class passengers survived at ~63%, third-class at ~24%.



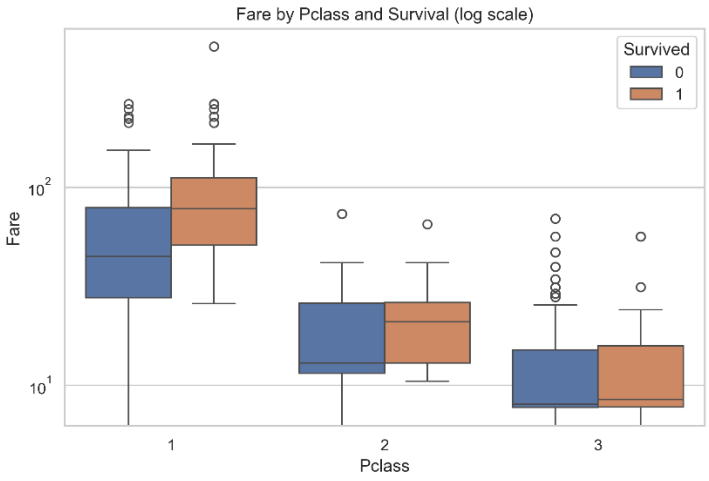
3. Age Distribution by Survival

Children had higher survival chances; survival drops for middle-aged men.



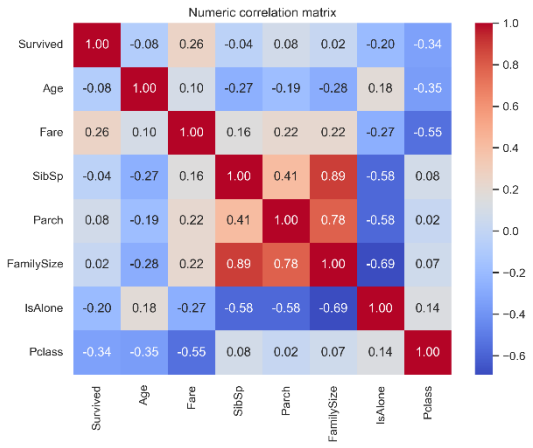
4. Fare vs Survival

Higher fares are associated with higher survival probability.



5. Correlation Heatmap

Pclass is negatively correlated with Fare (-0.55) and survival (-0.34).



6. Family Size vs Survival

Small families had better survival than those alone or with large groups.

Key Insights

- Women were far more likely to survive than men.
- First-class passengers had significantly higher survival rates.

- Age played a role — children survived more often.
- Higher ticket prices were linked to better survival.
- Being alone reduced survival odds; small family groups fared better.

Limitations & Next Steps

- Age imputation by median may oversimplify patterns; could use model-based imputation.
- Cabin data was dropped — extracting deck information may add predictive power.
- Further feature engineering could include ticket group size, fare per person.
- Next step: build a classification model (Logistic Regression, Random Forest, XGBoost) using engineered features.

Conclusions

This EDA revealed clear demographic and socioeconomic patterns in Titanic survival. Sex, class, fare, and family presence strongly influenced survival probability, providing a solid foundation for building predictive models.