Titanic Dataset – Exploratory Data Analysis (EDA) Report

1. Introduction

This project performs Exploratory Data Analysis (EDA) on the Titanic dataset to understand key factors influencing passenger survival.
The dataset includes demographic, socio-economic, and travel-related information such as class, gender, age, fare, and family size.

---

2. Data Overview

Loaded Columns

- PassengerId
- Survived
- Pclass
- Name
- Sex
- Age
- SibSp
- Parch
- Fare
- Embarked

Key Notes

- **Survived**: 1 = Survived, 0 = Did not survive
- **Pclass**: Socio-economic class (1 = Upper, 2 = Middle, 3 = Lower)
- **SibSp/Parch**: Family members aboard
- **Fare**: Ticket price

---

3. Data Cleaning

- Missing values identified in: **Age** and **Embarked**
- Age filled using statistical techniques (mean/median or modeling)
- Embarked filled with mode ("S")
- Data types corrected where needed

---

4. Univariate Analysis

Survival Count

- More passengers died than survived.
- Imbalanced dataset highlighting the disaster's scale.

Distribution of Passenger Class

- Majority passengers were **3rd class**
- Reflects Titanic's demographic (more lower-class travelers)

Age Distribution

- Most passengers were between **20–40 years**
- Age distribution is slightly right-skewed

Fare Distribution

- Right-skewed distribution
- Indicates wide fare variation based on class and cabin type

---

5. Bivariate & Multivariate Analysis

1. Survival vs Gender

- **Females had a much higher survival rate**
- "Women and children first" policy reflected

2. Survival vs Passenger Class

- Survival strongly linked with socio-economic status:
    - **1st Class: Highest survival**
    - **3rd Class: Lowest survival**

3. Survival vs Age

- Children (≤10 years) had higher survival
- Older passengers (>50) had lower survival

4. Fare vs Class & Survival

- Higher-class passengers paid significantly higher fares
- **Survivors generally had higher fares**, showing economic advantage
- Fare is a strong indicator of both class and survival probability

5. Family Size Impact

- Passengers traveling with **1–3 family members** had better survival
- Solo travelers and very large groups had lower chances

---

6. Correlation Analysis

- **Fare** positively correlates with survival
- **Pclass** negatively correlates with survival
- Age has weak correlation
- Heatmap shows survival influenced more by socio-economic attributes than age or family size

---

7. Visual Analysis Summary

- **Bar charts** show clear survival differences across class and gender
- **Histograms** confirm skewed Fare and Age distributions
- **Boxplots** describe class-wise fare variation
- **Count plots** highlight dominance of 3rd class passengers
- **Violin and swarm plots** reveal patterns between Fare, Class, and Survival
- **FacetGrid plots** show age distribution differences among survivors and non-survivors
- **Heatmap** confirms strongest predictors: Class, Fare, Sex

---

8. Key Insights

1. **Gender** is the strongest survival predictor → Females survived more.

2. **Class** plays a major role → 1st class highest survival, 3rd class lowest.
3. **Fare** reflects both class and survival probability.
4. **Age** influences survival — children favored.
5. **Traveling with small families** increases survival chances.
6. Economic and social factors played a larger role than physical factors like age.

---

9. Conclusion

The Titanic tragedy shows a clear pattern of survival inequality based on gender and socio-economic status.
This analysis highlights how wealth, gender norms, and family presence impacted survival during the disaster.

---

10. Future Improvements

- Feature engineering (Title extraction from names, family size categories)
- Predictive modeling (Logistic Regression, Random Forest, XGBoost)
- Hyperparameter tuning
- Deployment-ready classification pipeline