Data Analyst Internship: Task 5 EDA Report

1. Introduction and Objective

Objective: The primary goal of this Exploratory Data Analysis (EDA) is to extract meaningful insights and patterns from the Titanic dataset to understand the factors that influenced passenger survival.

Tools Used: Python (Pandas, Matplotlib, Seaborn)

---

2. Data Preparation and Cleaning

The initial inspection revealed missing values and non-predictive columns. The following steps were taken to clean and prepare the data:

Missing Data Imputation:

Column	Missing Count	Handling Strategy
Cabin	687	**Dropped** due to excessive missing data (over 77%).
Age	177	Imputed with the **median** age (≈28.0 years) to maintain the distribution shape while being robust to outliers.
Embarked	2	Imputed with the **mode** ('S' - Southampton), as only two values were missing.

Feature Removal:

The columns `PassengerId`, `Name`, and `Ticket` were dropped as they are identifiers or complex string features that do not directly contribute to survival analysis in a basic EDA.

---

3. Univariate Analysis (Individual Feature Distributions)

3.1 Target Variable (`Survived`)

- **Finding:** The dataset is moderately imbalanced. Approximately **61.6%** of the passengers died (549 passengers), and **38.4%** survived (342 passengers).
    - *Observation:* This imbalance indicates that a simple majority-class prediction model would be 61.6% accurate, setting a baseline for any predictive model.

3.2 Numerical Features (`Age`, `Fare`)

Age: The distribution will be somewhat **normal (bell-shaped)**, but potentially slightly **right-skewed**, centered around the 28-30 year old median.

- **Observation:** The age distribution is roughly normal, with a peak in young adulthood.

Fare: The distribution will be heavily **right-skewed**, concentrated toward lower fares, with many **outliers** extending to very high fare values (above $200).

- **Observation:** The Fare distribution is extremely right-skewed, indicating most passengers paid low fares, but a few paid significantly higher fares, which suggests a clear wealth divide.

3.3 Categorical Features (`Pclass`, `Sex`)

- `Pclass`**:** The majority of passengers (nearly 55%) were in **3rd Class**, highlighting a socio-economic skew toward the lower class among the passengers.
- `Sex`**:** There were significantly **more males** (≈65%) than females (≈35%).

---

4. Bivariate and Multivariate Analysis (Relationships)

4.1 Categorical Predictors vs. Survival

Gender (`Sex`)

- **Finding:** Gender is the single strongest predictor of survival. The survival rate for females (≈74%) was almost four times higher than the rate for males (≈19%).
    - *Observation:* This confirms the historical narrative of "women and children first."

Passenger Class (`Pclass`)

- **Finding:** Survival rate decreased significantly with lower class:
    - **1st Class:** ≈63% survival
    - **2nd Class:** ≈47% survival
    - **3rd Class:** ≈24% survival
    - *Observation:* Class (a proxy for wealth/status and proximity to lifeboats) was a critical factor.

4.2 Numerical Predictors vs. Survival

- **Fare vs. Survived:** Survivors paid significantly **higher fares** on average, and the distribution of their fares had a much higher median and spread compared to non-survivors.

- **Age vs. Survived:** The age distribution for survivors showed a higher concentration of very **young children**, indicating they were prioritized.

4.3 Correlation Analysis (Heatmap)

The heatmap of numerical features revealed the following key linear relationships:

- **Strongest Correlation with Survival:** `Sex_male` ($r \approx -0.54$) and `Pclass` ($r \approx -0.34$).
- **Relationship between features:** There is a strong negative correlation between `Pclass` and `Fare` ($r \approx -0.55$), confirming that lower class (Pclass 3) is associated with lower fares.
- *Observation:* No severe multicollinearity was detected, meaning the features are largely independent.

---

5. Summary of Findings (Conclusion)

Based on the statistical and visual exploration, the primary factors determining survival on the Titanic were social and demographic:

1. **Gender Priority: Sex** was the most influential factor; being female dramatically increased the probability of survival.
2. **Socio-economic Status: Passenger Class** and **Fare** were highly significant. Passengers in **First Class** were over 2.5 times more likely to survive than those in Third Class.
3. **Age Group: Children** had a better chance of survival compared to the general adult population.

These findings suggest that survival was primarily dictated by the rescue protocol ("women and children first") and the socio-economic status of the passenger (Class and Fare).