# EDA on Titanic Dataset

**Introduction:**
This EDA focuses on the Titanic dataset (891 rows, 12 columns) to analyze survival rates based on passenger information (e.g., age, sex, class) and embarkation details. Key steps include data screening, cleaning, and deriving insights for future predictions.

**Data Overview:**

- **Integer Columns:** PassengerId, SibSp (siblings/spouses), Parch (parents/children), Age, Fare.
- **Character Columns:** Name, Sex, Ticket, Cabin, Embarked, Survived (0 = died, 1 = survived), Pclass (1, 2, 3).

**Missing Values:**

- **Age:** 177 missing values (20%). Impute with mean/median or remove rows.
- **Cabin:** 687 missing values (77%). Remove column due to high irrelevance.

**Outliers:**

- **Age:** 66 outliers (below 0 and above 60). Median: 25-40. Right-skewed distribution.
- **SibSp:** 46 outliers. Most values: 0-1.
- **Parch:** 213 outliers. Most values: 0-1.
- **Fare:** 116 outliers. Median: 15-25. Bimodal distribution (3rd class dominant, 1st class outliers).

**Histograms:**

- **Age:** Right-skewed. Most passengers: 10-40 years (peak at 25-30).
- **SibSp & Parch:** Exponential distribution. Most passengers: 0-1 family members.
- **Fare:** Bimodal. Majority: 3rd class. Outliers: 1st class.

**Barcharts:**

- **Pclass:** Most passengers in 3rd class.
- **Sex:** 60% male, 40% female.
- **Embarked:** Most boarded at Southampton.
- **Survival:** Non-survivors > survivors.

**Correlation Matrix:**

- **Age:** Weak positive with PassengerId (0.034) and Fare (0.097); weak negative with SibSp (-0.23) and Parch (-0.17).
- **SibSp:** Moderate positive with Parch (0.41); weak positive with Fare (0.16).
- **Parch:** Moderate positive with SibSp (0.41); weak positive with Fare (0.22).
- **Fare:** Weak positive with Age (0.097), SibSp (0.16), and Parch (0.22).

**Hypothesis Testing:**

- **Survival by Age:** Non-survivors peak around 30. More non-survivors across age groups.
- **Survival by Gender:** More male non-survivors; more female survivors.
- **Survival by Class:**
    - **1st Class:** Survivors > non-survivors.
    - **2nd Class:** Balanced (slightly more survivors).
    - **3rd Class:** Non-survivors > survivors.

**Conclusion:**
The EDA reveals key trends: higher survival rates among females and 1st-class passengers, with age and family size influencing outcomes. Data cleaning (e.g., handling missing Age, removing Cabin) is essential for accurate predictions.