Titanic Dataset - EDA Summary Report

Objective

The aim of this analysis is to explore the Titanic dataset to uncover patterns and insights related to passenger survival, using statistical summaries and visualizations.

Dataset Overview

- **Total Records**: 891 passengers
- **Key Columns**: Survived, Pclass, Sex, Age, SibSp, Parch, Fare, Embarked
- **Target Variable**: Survived (0 = No, 1 = Yes)

Data Cleaning

- **Missing Values**:
    - Age: 177 missing → filled with median age
    - Embarked: 2 missing → filled with mode ('S')
    - Cabin: 687 missing → column dropped
- **New Feature**:
    - FamilySize = SibSp + Parch + 1

Univariate Analysis

- **Survival Count**:
    - ~38% survived, ~62% did not
    - Females had higher survival rate than males
- **Passenger Class**:
    - Most passengers belonged to 3rd class
    - 1st class had better survival rate
- **Fare Distribution**:
    - Right-skewed; majority of fares under 100
    - A few high-fare outliers existed
- **Age Distribution**:
    - Most passengers between 20-40 years
    - Children also present; median age ~28

Bivariate Analysis

- **Survival vs. Sex**:
    - ~75% of females survived vs. ~19% of males
- **Survival vs. Pclass**:
    - 1st class passengers had the highest survival rate
    - 3rd class passengers had the lowest
- **Fare vs. Pclass**:
    - 1st class had highest median fare
    - Clear stratification of fare by class
- **Survival vs. Family Size**:
    - Medium family sizes (2–4) had better survival chances
    - Very large or solo travelers had lower survival rates
- **Age vs. Survival**:
    - Younger children had better survival
    - Elderly passengers had lower survival rates

Correlation Insights

- **Strongest Positive Correlation**: Fare ↔ Pclass (negatively correlated due to class order)
- **Survival Correlates**: Weak correlations observed with Age, Fare, Pclass, and Sex (encoded)

Key Takeaways

- **Gender** and **Class** were the strongest indicators of survival.
- **Higher fare** generally corresponded to better survival outcomes.
- **Family size** influenced survival, with moderate-sized families faring better.
- Handling missing values and feature engineering (like FamilySize) improved analysis depth.