Titanic Dataset - Exploratory Data Analysis (EDA) Report

Objective

The aim of this EDA is to uncover trends, patterns, and relationships within the Titanic dataset, and understand the factors that influenced passenger survival.

Dataset Description

- Rows: 891 | Columns: 12

- Features: PassengerId, Name, Sex, Age, Fare, Pclass, SibSp, Parch, Embarked, Cabin, Ticket, Survived

- Target: 'Survived' (0 = No, 1 = Yes)

- Missing values found in 'Age', 'Cabin', and 'Embarked'

- 'Cabin' dropped; 'Age' filled with median; 'Embarked' filled with mode

Univariate Analysis

- Most passengers were in 3rd class

- Majority were male (~65%)

- Age distribution peaked between 20-40 years

- Fare was right-skewed, with a few very high-paying passengers

Bivariate Analysis

- Survival rate much higher for females (~75%)

- 1st class passengers had higher survival rate

- Passengers from port 'C' (Cherbourg) had higher survival

- Younger passengers had slightly better survival chances

Multivariate Analysis

- Correlation heatmap showed:

  - 'Fare' and 'Pclass' had moderate negative correlation

  - 'Parch' and 'SibSp' correlated with 'FamilySize'

- New feature 'FamilySize' (SibSp + Parch + 1) showed passengers with family size 2-4 had better survival

Key Insights

- Gender and class were the most influential factors in survival

- Female passengers had significantly higher survival odds

- Being in 1st class increased likelihood of survival

- Fare paid had some positive correlation with survival

- Passengers with moderate family sizes had better survival than those alone or in large families

Data Cleaning Performed

- Filled missing 'Age' values with median

- Filled missing 'Embarked' values with mode

- Dropped 'Cabin' column due to excessive nulls

- Created new feature: 'FamilySize'

Summary

The EDA reveals that gender, passenger class, and family size strongly influenced survival chances. This analysis prepares the dataset for future predictive modeling by addressing missing data, encoding relevant features, and understanding variable relationships.

Next Steps

- Encode categorical variables for modeling

- Train predictive models using cleaned dataset

- Evaluate performance using metrics like accuracy, precision, and recall