Exploratory Data Analysis (EDA) on Titanic Dataset

Introduction

The Titanic dataset contains information about passengers aboard the Titanic, including whether they survived, their class, age, sex, and other details.
The goal of this analysis is to **explore the dataset**, identify **patterns, relationships, and trends**, and summarize insights.

Dataset Overview

- **Number of rows and columns:** (use df.shape)
- **Columns in the dataset:** PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked
- **Data types and missing values:** (use df.info())
- **Summary statistics:** (use df.describe())

Categorical Variable Analysis

Survival:

- 0 = Did not survive, 1 = Survived
- Value counts: (use df['Survived'].value_counts())

Passenger Class (Pclass):

- Classes 1, 2, 3
- Value counts and distribution

Gender (Sex):

- Male vs Female distribution

Port of Embarkation (Embarked):

- C = Cherbourg, Q = Queenstown, S = Southampton

Observations:

- Most passengers did not survive.
- Majority of passengers were in 3rd class.

- More males than females aboard.
- Southampton was the most common embarkation port.

Numeric Variable Analysis

- **Age:** (histogram, boxplot)
- **Fare:** (histogram, boxplot)
- **Observations:**
    - Age ranges from X to Y, with some missing values.
    - Fare is skewed; a few passengers paid very high fares.

Relationships and Trends

Age vs Survival:

- Boxplot shows that younger passengers had slightly higher survival chances.

Pclass vs Survival:

- Passengers in 1st class had higher survival rates than 3rd class.

Sex vs Survival:

- Females had a much higher survival rate than males.

Fare vs Survival:

- Passengers who paid higher fares were more likely to survive.

Correlation Heatmap:

- Shows numeric relationships. Example: Fare positively correlated with Pclass (1st class paying more).

Pairplot:

- Visualizes trends among Age, Fare, Pclass, and Survival.

Histograms, Boxplots, and Scatterplots

- **Histograms:** Show distribution of Age, Fare, etc.
- **Boxplots:** Identify outliers and compare groups (e.g., Age by Survival).

- **Scatterplots:** Show relationship between Age and Fare colored by Survival.

Observations:

- Most passengers are aged 20–40.
- Fare shows high variability, with some extreme outliers.
- Younger females in higher classes had the highest chance of survival.

Summary of Findings

1. Survival was heavily influenced by **Sex** and **Passenger Class**.
2. Females and 1st class passengers had higher survival rates.
3. Younger passengers had slightly better survival odds.
4. Fare is related to class and survival — higher fare = higher survival chance.
5. Embarked port and family size (SibSp + Parch) show minor influence on survival.

Conclusion

EDA reveals clear patterns in the Titanic dataset that can inform predictive modeling:

- **Sex and Class are the strongest predictors** of survival.
- Visualizations help identify outliers, trends, and correlations in the data.
- These insights can be used to improve machine learning models for survival prediction.