Exploratory Data Analysis (EDA) - Titanic Dataset

1. Introduction

The Titanic dataset contains information about passengers, including demographics, ticket details, and survival status.
The goal of this EDA is to understand the data structure, identify patterns, and prepare it for modelling.

2. Dataset Overview

- Rows: 891
- Columns: 12
- Target Variable: `Survived`

Features:

- Pclass: Passenger class (1 = First, 2 = Second, 3 = Third)
- Sex: Gender of passenger
- Age: Age in years
- SibSp: Number of siblings/spouses aboard
- Parch: Number of parents/children aboard
- Ticket: Ticket number
- Fare: Passenger fare
- Cabin: Cabin number
- Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

3. Data Cleaning

- Checked for missing values in `Age`, `Cabin`, and `Embarked`
- Imputed missing `Age` values with median
- Filled missing `Embarked` with mode
- Created new feature `HasCabin` from `Cabin`

4. Univariate Analysis

- Categorical Variables: Used `countplot` to visualize `Sex`, `Pclass`, `Embarked`, `Survived`
- Numerical Variables: Used `histplot` and `describe()` for `Age` and `Fare`

5. Bivariate Analysis

- Categorical vs Target: Survival rates by `Sex`, `Pclass`, and `Embarked`
- Numerical vs Target: Boxplots of `Age` and `Fare` against `Survived`

6. Correlation Analysis

- Generated correlation matrix and plotted with `sns.heatmap()`
- Created `sns.pairplot()` to visualize pairwise relationships between numerical variables

7. Key Insights

- Females had a much higher survival rate than males
- First-class passengers had the highest survival rate
- Younger passengers had a slightly better chance of survival
- Higher fare was positively correlated with survival

---

8. Conclusion

The dataset shows clear patterns in survival based on passenger class, gender, and fare.