Titanic Dataset - Exploratory Data Analysis (Task 5)

Introduction

This project involves performing Exploratory Data Analysis (EDA) on the Titanic dataset as part of a Data Analyst Internship Task. The dataset contains detailed information about passengers aboard the Titanic, including whether or not they survived.

The aim is to understand survival patterns using data cleaning, univariate/bivariate analysis, visualizations, and derive actionable insights.

Dataset Overview

The Titanic dataset from Kaggle includes features such as PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. The target variable is 'Survived'.

Data Cleaning

- Checked for missing values using df.isnull().sum()
- 'Age' had missing values: filled using the median
- 'Embarked' had missing values: filled using the mode
- 'Cabin' had over 77% missing values and was dropped
- Checked datatypes and summary statistics using df.info() and df.describe()

Univariate Analysis

We analyzed single variables using:

- Countplot for Survived, Sex, Pclass, and Embarked
- Histogram for Age and Fare
- Boxplot for Fare and Age
- Observed distributions and frequency of values

Bivariate Analysis

We compared 'Survived' with other features to find relationships:

- Survived vs Sex: More females survived than males
- Survived vs Pclass: Passengers in 1st class had higher survival
- Survived vs Age: Children had better survival chances
- Survived vs Embarked: Those who boarded from Cherbourg had higher survival
- Survived vs Fare: Higher fare passengers had higher survival

Visualizations Used

- Countplot (Seaborn)
- Histogram
- Boxplot
- Heatmap for correlation
- Pairplot (optional, skipped for performance)
- All plots included clear titles and labels

Summary of Insights

- Females had significantly higher survival rate
- Passengers in 1st class were more likely to survive
- Children and passengers with higher fare had better chances
- 'Cabin' was too sparse to provide meaningful information

Interview Questions Included

At the end of the notebook, commonly asked EDA interview questions were answered, such as:

- What is EDA and why is it important?
- How to detect multicollinearity?
- Difference between heatmap and pairplot
- Univariate vs Bivariate analysis

Conclusion

The Titanic dataset helped me apply EDA techniques on real-world data. From cleaning and visualizing to generating insights, this project showcases beginner-level EDA skills using Python. It strengthened my understanding of data analysis fundamentals and visual storytelling.