Titanic Dataset – Exploratory Data Analysis (EDA)

Internship Task 5 Submission
By: RAVI OJHA
Tools Used: Python, Pandas, Seaborn, Matplotlib

Objective

The objective of this task was to explore the Titanic dataset and extract insights using statistical and visual analysis. This involved identifying patterns, trends, and outliers that help understand the factors affecting passenger survival.

1. **Dataset Loaded**: **Used Pandas to load the Titanic dataset (CSV file).**
2. **Initial Exploration.**

- **Viewed first few rows using `df.head()`**
- **Used `df.info()` and `df.describe()` to understand data structure**

3. Missing Values Checked:

- **Found missing values in Age, Cabin, and Embarked**
- **Filled missing Age values with median (to handle outliers)**
- **Converted Age column from float to integer using .round().astype(int)**

Chart Type	Column(s) Analyzed	Purpose / Why Used
Countplot	Survived	To show how many passengers survived
Countplot	Sex vs Survived	To see survival based on gender
Countplot	Pclass vs Survived	To check survival based on class
Histogram	Age, Fare	To check distribution & skewness
Boxplot	Pclass vs Fare	To analyze fare spread across classes
Boxplot	Survived vs Age	To check age variation in survival
Heatmap	Numeric Correlation	To check correlation between numeric columns
Pairplot	Age, Fare, Pclass, Survived	For multivariate relationship visualization
Stacked Bar	Sex & Survived	To show survival by gender clearly
Countplot	Embarked & Survived	To show survival by embarkation port

Key Analysis & Why It Matters

- **More people traveled in 3rd class: Found by using value_counts() and countplot. This is likely because 3rd class was most affordable.**
- **Average Fare by Class: 1st class passengers paid ~₹84, while 3rd class paid ~₹13. This huge difference explains why more people were in 3rd class.**
- **Gender-based Survival: Females had much higher survival rate than males, shown using hue='Survived' in countplot.**
- **Age Distribution: Majority of passengers were between 20–40 years. Children and teenagers had higher survival rates.**
- **Fare and Age Outliers: Found using boxplots — especially visible in 1st class.**
- **Embarked Port Analysis: Most passengers boarded from Port 'S' (Southampton).**

Final Insights (Summary)

- 💰 **Fare is a major reason why most people chose 3rd class.**
- 👩 **Gender plays a huge role in survival — females survived more.**
- 👶 **Younger passengers (especially children) had better chances of survival.**
- 🚢 **Port 'S' had the highest number of passengers.**
- 📊 **There is a slight correlation between Fare, Pclass, and Survival.**

Conclusion

This EDA task helped me apply real-world data analysis techniques using Python libraries. I used visualization effectively to interpret hidden insights from data, handled missing values properly, and made evidence-based observations. This task prepared me to handle more advanced data projects and present findings in a clean, professional format.

Thank you for the opportunity!