Internship Task Report – Titanic Dataset (EDA)

1. Introduction

For this task, I worked on the famous **Titanic dataset**, which is available on Kaggle. The goal of this project was to perform **Data Cleaning** and carry out **Exploratory Data Analysis (EDA)** to uncover patterns and insights hidden within the data.

2. Understanding the Dataset

The dataset is part of the Kaggle **Titanic Challenge**. It comes with two main files:

- **train.csv** → used for analysis (contains the survival information).
- **test.csv** → used for prediction tasks (not used in this EDA).

Key columns in the dataset:

- **Survived**: Target column (0 = Not Survived, 1 = Survived)
- **Pclass**: Passenger class (1st, 2nd, 3rd)
- **Sex**: Gender of the passenger
- **Age**: Age of the passenger
- **SibSp & Parch**: Family members onboard
- **Fare**: Ticket price
- **Embarked**: Port of embarkation
- **Cabin, Ticket, Name**: Contain too much noise or are less useful for EDA

3. Data Cleaning Process

Before performing analysis, I cleaned the dataset to handle missing or irrelevant values:

- **Age**: Filled missing values with the **median age**.
- **Embarked**: Filled missing values with the **most common port (mode)**.
- **Cabin**: Dropped this column because most values were missing.
- **HasCabin**: Added a new column (1 = passenger had a cabin, 0 = no cabin).
- Converted some categorical features like **Sex, Pclass, Embarked** into categorical data types for better analysis.

4. Exploratory Data Analysis (EDA)

(A) Univariate Analysis

- **Age**: Most passengers were between 20–40 years old.
- **Fare**: The distribution was right-skewed, showing a few passengers paid very high fares.
- **Pclass**: Most passengers belonged to 3rd class.

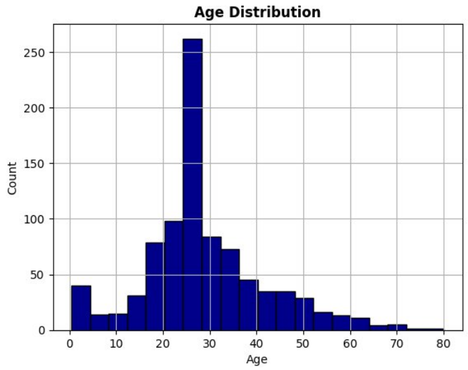
(B) Bivariate Analysis (with Survival)

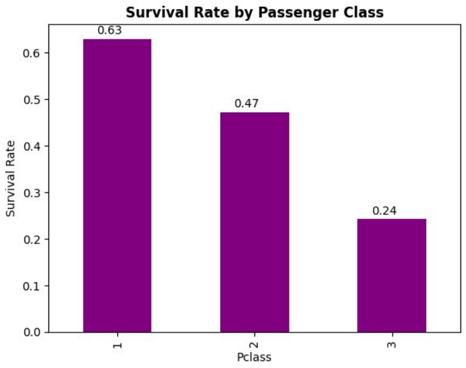
- **Sex**: Survival rate was much higher for women (~74%) compared to men (~19%).
- **Pclass**: 1st class passengers had the highest survival rate (~63%), followed by 2nd class (~47%), and 3rd class (~24%).
- **Embarked**: Passengers who boarded at Cherbourg (Port C) had slightly better chances of survival.
- **Age**: Children under 12 had a higher survival rate compared to adults.

5. Visual Insights

I created multiple charts in Python (using Matplotlib) to visualize the findings:

- Histogram of Age distribution
- Bar charts for survival count
- Survival comparison by **Sex, Pclass, and Embarked**





6. Key Findings

- **Women and children** had significantly higher survival chances.
- Passengers from **wealthier classes (1st class)** survived more compared to those in 3rd class.
- Passengers boarding from **Port C** had better outcomes than others.
- Traveling **with family** improved chances, while being alone reduced survival chances.

7. Conclusion

The Titanic dataset required initial cleaning to handle missing values and irrelevant features. After data cleaning and EDA, I was able to identify strong patterns: survival was closely related to **gender, class, and age**.

These insights not only show the social dynamics during the Titanic tragedy but also lay the foundation for building predictive models in the future.