TASK 2

Title: Exploratory Data Analysis (EDA) on the Titanic Dataset

Objective

The objective of this task is to perform **data cleaning** and **exploratory data analysis (EDA)** on the Titanic dataset from Kaggle to discover patterns, trends, and relationships between different variables, which can be used for predictive modeling.

Dataset Information

- **Dataset Name:** Titanic - Machine Learning from Disaster
- **Source:** Kaggle Titanic Dataset

- **Files Used:**

- train.csv – includes passenger details and survival status

- test.csv – includes details without survival status (for prediction)

- gender_submission.csv – sample submission format

1: Data Cleaning

The dataset contains missing values and categorical data, which must be handled before analysis.

Steps performed:

- Filled missing Age values with median age.

- Dropped the Cabin column due to excessive missing data.

- Filled missing values in Embarked with the most frequent value.

- Converted Sex and Embarked columns to numerical format.

- Dropped irrelevant columns like Ticket, PassengerId (for some analyses).

2: Exploratory Data Analysis (EDA)

Univariate Analysis

- **Age & Fare:** Histograms showed a majority of passengers were young adults, and most fares were under $100.

- **Pclass & Sex:** Count plots highlighted that 3rd class had the highest number of passengers, and there were more males than females.

- **Survived:** Count plot revealed that more people died than survived.

Bivariate Analysis

- **Survival Rate by Sex:** Females had a significantly higher survival rate.

- **Survival Rate by Pclass:** 1st class passengers had better survival odds than 2nd and 3rd class.

- **Boxplot of Fare vs Survived:** Higher fares slightly correlated with higher survival.

Multivariate Analysis

- Created a correlation heatmap between numerical variables.

- Grouped survival data by Sex, Pclass, and Embarked to spot trends.

Key Insights

- **Gender:** Females had much higher survival rates than males.

- **Class:** Higher class = higher survival.

- **Age:** Children had better chances of survival than older adults.

- **Embarkation Port:** Passengers from Cherbourg (C) had a better survival rate than those from Southampton (S) or Queenstown (Q).

Tools & Technologies Used

- **Python**: Pandas, NumPy, Matplotlib, Seaborn

- **IDE**: Google Colab / Jupyter Notebook

- **Data Source**: Kaggle

Conclusion

EDA revealed clear relationships between features like **Sex, Pclass, and Age** with survival outcomes. Data cleaning and visualization were essential in understanding how these variables influenced survival and prepared the data for machine learning models.

Link to Dataset

🔗 https://www.kaggle.com/c/titanic/data