Task - 1 : Report

EXPLORATORY DATA ANALYSIS ON A PUBLIC DATASET

---

1) Dataset Description

The dataset used in this project is the **Titanic Dataset**, which contains information about passengers aboard the RMS Titanic and whether they survived the disaster.

- **Total Rows:** 891
- **Total Columns:** 12

Features in the Dataset:

Column	Description
PassengerId	Unique passenger identifier
Survived	Survival status (0 = No, 1 = Yes)
Pclass	Passenger class (1st, 2nd, 3rd)
Name	Passenger name
Sex	Gender of passenger
Age	Age in years
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Ticket fare
Cabin	Cabin number
Embarked	Port of embarkation (C, Q, S)

---

2) Objective

The objective of this Exploratory Data Analysis is to:

- Understand the structure and quality of the data
- Handle missing and inconsistent values
- Identify important patterns affecting passenger survival
- Visualize relationships between key variables

---

3) Steps Performed

Data Loading

- Loaded the dataset using **Pandas**
- Checked dataset shape, column names, and data types

Data Cleaning

- Identified missing values in **Age, Cabin, and Embarked**
- Filled missing **Age** values using **median**
- Filled missing **Embarked** values using **mode**
- Dropped **Cabin** column due to excessive missing values
- Checked for duplicate records (none found)

Exploratory Data Analysis

- Univariate analysis on Survived, Sex, Age, and Pclass
- Bivariate analysis between:
    - Survived vs Sex
    - Survived vs Passenger Class
- Distribution analysis of Age and Fare
- Correlation analysis using heatmap

---

4) Visualizations Used

- Count plot for survival distribution
- Count plot for survival based on gender
- Count plot for survival across passenger classes
- Histogram for age distribution
- Correlation heatmap for numerical features

(Visuals created using **Matplotlib and Seaborn**)

---

5) Key Insights

- Female passengers had a **significantly higher survival rate** than males
- Passengers traveling in **1st class** survived more than those in 2nd and 3rd class
- Higher **fare values** were associated with better survival chances
- Younger passengers showed slightly higher survival probability
- Passenger class had a stronger influence on survival than age

---

6) Challenges

- Handling missing values without introducing bias
- Choosing appropriate visualizations for categorical vs numerical data
- Avoiding incorrect conclusions from correlation
- Managing skewed distributions in Fare values

---

7) Conclusion

This EDA helped in understanding passenger demographics, data quality issues, and key factors influencing survival. The insights gained from this analysis can be further extended into **predictive modeling** for survival classification.

Tech Stack Used

- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Jupyter Notebook
