Project Report

Exploratory Data Analysis on Titanic Dataset

---

1. Introduction

The Titanic dataset is one of the most famous datasets in machine learning and data science. It contains information about passengers who were aboard the Titanic ship, along with whether they survived or not. This project performs **Exploratory Data Analysis (EDA)** to uncover insights about survival patterns, demographics, and ticket classes.

---

2. Dataset Description

The dataset, available on **Kaggle (Titanic: Machine Learning from Disaster)**, includes details of 891 passengers.

- **Columns include:**
  - `PassengerId`: Unique ID for each passenger
  - `Survived`: Survival status (0 = No, 1 = Yes)
  - `Pclass`: Ticket class (1 = First, 2 = Second, 3 = Third)
  - `Name`: Passenger’s name
  - `Sex`: Gender of passenger
  - `Age`: Age of passenger
  - `SibSp`: Number of siblings/spouses aboard
  - `Parch`: Number of parents/children aboard
  - `Ticket`: Ticket number
  - `Fare`: Ticket fare paid
  - `Cabin`: Cabin number (many missing values)
  - `Embarked`: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

---

3. Methodology

3.1 Data Cleaning

- Checked for missing values and filled/imputed where appropriate:
  - Missing `Age` values imputed with median age.

    - Missing `Embarked` values filled with mode.
    - Dropped `Cabin` column due to excessive missing values.
- Converted categorical variables (`Sex`, `Embarked`) into numerical/encoded forms for analysis.

3.2 Exploratory Data Analysis

- Summary statistics were generated for numerical features.
- Group-based analysis conducted:
    - Survival rate by **gender**
    - Survival rate by **passenger class**
    - Survival rate by **age group**
    - Survival rate by **embarkation port**

3.3 Visualization

- **Bar plots** used to show survival by gender and class.
- **Histograms** used to display age distribution of survivors vs non-survivors.
- **Heatmaps** to check correlations between numerical features.

---

4. Key Insights

1. **Gender Factor**: Females had a much higher survival rate compared to males.
2. **Class Factor**: Passengers in **1st class** had significantly better chances of survival than those in 3rd class.
3. **Age Factor**: Children had higher survival rates compared to adults.
4. **Fare Factor**: Higher ticket fares correlated with better chances of survival.
5. **Embarkation**: Passengers embarking from **Cherbourg (C)** had relatively higher survival rates.

---

5. Visualizations

- **Bar Chart**: Survival by gender clearly shows females were prioritized.
- **Stacked Bar Chart**: Combined effect of class and survival.
- **Histogram**: Age distribution highlighting more children survived.
- **Heatmap**: Correlation matrix showing relationships among features (e.g., Fare and Pclass).

---

6. Conclusion

The Titanic dataset highlights how survival was influenced by **gender, social class, and age**. Women, children, and wealthy passengers had higher survival chances due to evacuation policies at the time.

This EDA project demonstrates the importance of **data cleaning, feature understanding, and visualization** in uncovering meaningful insights from real-world datasets.