EXPLORATORY DATA ANALYSIS REPORT

Student Name: Kalvacherla Reethika

Date: 19/02/2026

Task: Exploratory Data Analysis (EDA)

Tools Used: Python, Pandas, Matplotlib, Seaborn

1️⃣ Introduction

The objective of this project is to perform Exploratory Data Analysis (EDA) on the Titanic dataset to identify patterns, trends, relationships, and anomalies that influenced passenger survival.

EDA helps in understanding data structure, detecting missing values, visualizing distributions, and identifying important features before building predictive models.

2️⃣ Dataset Overview

The dataset contains information about passengers aboard the Titanic, including:

- PassengerId
- Survived (Target Variable)
- Pclass (Passenger Class)
- Name
- Sex
- Age
- SibSp (Siblings/Spouses aboard)
- Parch (Parents/Children aboard)
- Ticket
- Fare
- Cabin
- Embarked

Dataset Shape:

- Rows: 891
- Columns: 12

3️⃣ Data Cleaning & Preparation

- Checked for missing values using .info() and .isnull().sum()
- Age column contains missing values.
- Cabin column contains many missing values.
- Embarked has few missing values.
- Verified data types using .info()
- Used .describe() for statistical summary.

4️⃣ Univariate Analysis

4.1 Survival Distribution

- Approximately 38% of passengers survived.
- 62% did not survive.
- The dataset is slightly imbalanced.

4.2 Gender Distribution

- Majority passengers were male.
- Female passengers were fewer in number.

4.3 Passenger Class Distribution

- Most passengers were from 3rd class.
- 1st class had the least passengers.

4.4 Age Distribution

- Most passengers were between 20–40 years.
- Some missing age values exist.
- Distribution slightly right-skewed.

4.5 Fare Distribution

- Fare is highly right-skewed.
- Few passengers paid very high fares (outliers).

---

5 Bivariate Analysis

5.1 Survival vs Gender

- Females had significantly higher survival rate.
- Majority of males did not survive.
- Gender strongly influenced survival.

5.2 Survival vs Passenger Class

- 1st class passengers had highest survival rate.
- 3rd class had lowest survival rate.
- Socioeconomic status played an important role.

5.3 Age vs Survival

- Children had better survival chances.
- Age alone is not a strong predictor.

5.4 Fare vs Survival

- Higher fare passengers had better survival probability.
- Strong link between economic status and survival.

---

6 Multivariate Analysis

Correlation Heatmap Findings

- Fare positively correlated with Survival.
- Pclass negatively correlated with Survival.
- Weak correlation between Age and Survival.
- No strong multicollinearity observed.

---

7 Key Insights

1. Survival rate was around 38%.

2. Gender is the strongest predictor of survival.
3. Passenger class significantly impacts survival.
4. Higher fare increases survival chances.
5. Children had better survival probability.
6. Missing data present in Age and Cabin columns.

8️⃣ Conclusion

The analysis clearly shows that survival was strongly influenced by:

- Gender
- Passenger Class
- Fare

Passengers from higher socioeconomic classes and female passengers had significantly better survival rates. These features can be effectively used for predictive modeling.

9️⃣ Outcome

Through this project, the following skills were developed:

- Performing data exploration using Pandas
- Creating visualizations using Matplotlib and Seaborn
- Identifying trends and relationships
- Writing analytical observations
- Summarizing findings professionally