Exploratory Data Analysis (EDA) Report

Titanic Dataset

---

1. Introduction

Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their main characteristics using statistical methods and visualizations. The goal of EDA is to understand the structure of data, detect patterns, identify anomalies, and extract meaningful insights before building any predictive model.

In this project, EDA was performed on the Titanic dataset to understand factors that influenced passenger survival.

---

2. Dataset Description

The dataset consists of three CSV files:

- **train.csv** – 891 rows and 12 columns
- **test.csv** – 418 rows and 11 columns
- **gender_submission.csv** – 418 rows and 2 columns

The training dataset contains passenger information such as:

- PassengerId
- Survived
- Pclass
- Name
- Sex
- Age
- SibSp
- Parch
- Ticket
- Fare
- Cabin
- Embarked

---

3. Data Inspection and Cleaning

3.1 Shape of Dataset

- Train: (891, 12)
- Test: (418, 11)

This step was performed to understand the size of the dataset.

3.2 Checking Data Types and Missing Values

Using .info() method, the following missing values were identified:

Train Dataset:

- Age has missing values
- Cabin has many missing values
- Embarked has 2 missing values

Test Dataset:

- Age has missing values
- Fare has 1 missing value
- Cabin has many missing values

This step is important to identify data quality issues before analysis.

3.3 Statistical Summary

Using .describe() method:

- Average Age ≈ 29.7 years
- Maximum Fare ≈ 512.33
- Survival Mean ≈ 0.38 (38% survived)

This helped understand central tendency and spread of numerical variables.

---

4. Univariate Analysis

4.1 Age Distribution

- Most passengers were between 20 and 40 years.
- Age distribution is slightly right-skewed.
- Few elderly passengers above 65.

Purpose: To understand age concentration and detect skewness.

4.2 Fare Distribution

- Highly right-skewed.
- Majority paid lower fares (below 50).
- Few extreme values above 200 and 500.

Purpose: To identify fare spread and detect outliers.

4.3 Boxplot Analysis

Age Boxplot:

- Moderate spread.
- Few high-value outliers.

Fare Boxplot:

- Many extreme outliers.
- Strong right skewness.

Purpose: To detect outliers and understand variability.

---

5. Bivariate Analysis

5.1 Survival vs Gender

- Female survival rate is significantly higher.
- Majority of male passengers did not survive.

Conclusion: Gender played a major role in survival.

5.2 Survival vs Passenger Class

- 1st class had highest survival rate.

- 3rd class had highest death rate.
- Survival decreases as class number increases.

Conclusion: Socioeconomic status influenced survival probability.

---

6. Correlation Analysis

A heatmap was generated to analyze relationships between numerical variables.

Key findings:

- Pclass has negative correlation with Survived (-0.34)
- Fare has positive correlation with Survived (0.26)
- Strong negative correlation between Pclass and Fare (-0.55)

Interpretation:
Passengers with lower class number (1st class) and higher fares had higher survival probability.

---

7. Multivariate Analysis

Pairplot was used to analyze relationships among Survived, Pclass, Age, and Fare.

Observations:

- Clear survival pattern with Pclass.
- Higher fare passengers show more survival.
- Age does not strongly separate survival.

Purpose: To visually examine multiple variable interactions.

---

8. Final Summary of Findings

- Total passengers: 891
- Survival rate: 38%
- Majority passengers were male and 3rd class
- Female passengers had significantly higher survival rate
- 1st class passengers had higher survival probability
- Fare positively impacts survival

- Passenger class negatively impacts survival
- Fare distribution is highly skewed
- Age is not a strong independent survival predictor

---

9. Conclusion

Exploratory Data Analysis revealed that gender and passenger class were the most influential factors affecting survival in the Titanic dataset. Higher fare and first-class passengers had better survival chances. Age had weaker influence compared to other variables.

EDA helps uncover hidden patterns and supports better decision-making in predictive modeling.