# Exploratory Data Analysis (EDA) Report

- **Titanic Dataset**

## 1. Introduction

Exploratory Data Analysis (EDA) is an important step in the data analysis process. It helps in understanding the dataset, identifying patterns, trends, relationships, and anomalies using statistical methods and visualizations.

In this project, EDA is performed on the **Titanic dataset**, which contains information about passengers who travelled on the Titanic. The main objective is to analyze the dataset and identify the factors that influenced passenger survival.

## 2. Objective of the Project

The main objectives of this project are:

- To understand the structure and quality of the Titanic dataset
- To perform statistical analysis on numerical and categorical variables
- To visualize data using different plots
- To identify relationships between features and survival
- To summarize key insights obtained from the analysis

## 3. Dataset Description

The Titanic dataset consists of passenger information such as:

- Passenger ID
- Survival status
- Passenger class
- Name
- Gender
- Age
- Number of siblings/spouses aboard
- Number of parents/children aboard
- Fare paid
- Port of embarkation

The dataset includes both **numerical** and **categorical** variables, making it suitable for exploratory analysis.

## 4. Tools and Technologies Used

The following tools and libraries were used in this project:

- **Python** – Programming language
- **Pandas** – Data manipulation and analysis
- **Matplotlib** – Data visualization
- **Seaborn** – Statistical data visualization
- **Jupyter Notebook** – Interactive analysis environment

## 5. Data Understanding and Inspection

Initial data inspection was carried out using functions such as:

- .head() – to view the first few rows
- .info() – to understand data types and missing values
- .shape() – to identify the number of rows and columns
- .describe() – to get statistical summaries

**Observations:**

- The dataset contains missing values in columns such as **Age**, **Cabin**, and **Embarked**
- The **Survived** column is the target variable
- Fare values show a wide range, indicating possible outliers

## 6. Exploratory Data Analysis

### 6.1 Univariate Analysis

Univariate analysis was performed to understand individual variables.

**Survival Distribution:**

- More passengers did **not survive** compared to those who survived

**Gender Distribution:**

- The number of male passengers was higher than female passengers

**Age Distribution:**

- Most passengers were between **20 and 40 years old**

### 6.2 Bivariate Analysis

Bivariate analysis was used to understand relationships between two variables.

**Survival vs Gender:**

- Female passengers had a **higher survival rate** than males

**Survival vs Passenger Class:**

- First-class passengers had better survival chances
- Third-class passengers had the lowest survival rate

**Fare vs Survival:**

- Passengers who paid higher fares were more likely to survive

### 6.3 Multivariate Analysis

Multivariate analysis was carried out using:

- **Pair plots** to analyze interactions between numerical features
- **Correlation heatmap** to identify relationships between variables

**Observations:**

- Fare has a positive correlation with survival
- Passenger class has a negative correlation with survival
- Higher socio-economic status increased survival chances

## 7. Key Insights and Findings

- Gender played a significant role in survival, with females surviving more
- Passenger class strongly influenced survival probability
- Higher fares were associated with higher survival rates
- Children and younger passengers had better survival chances
- Socio-economic factors were crucial in determining survival

## 8. Conclusion

This exploratory data analysis of the Titanic dataset revealed that survival was not random but influenced by multiple factors such as gender, passenger class, age, and fare. Visualization techniques helped uncover hidden patterns and trends in the data.

The analysis demonstrates how EDA is a powerful technique for understanding data before applying machine learning models or further analysis.

## 9. Future Scope

- Handle missing values more effectively
- Perform feature engineering
- Build machine learning models to predict survival
- Compare multiple classification algorithms

## 10. References

- Titanic Dataset – Kaggle
- Python Documentation
- Pandas, Matplotlib, and Seaborn Documentation