Titanic Dataset - Exploratory Data Analysis (EDA) Report

Internship Task 2 Report

1. Introduction

This report presents a detailed exploratory data analysis (EDA) performed on the Titanic dataset to understand survival patterns, relationships among variables, and data characteristics. The primary objective is to extract meaningful insights using statistical analysis and visualization techniques.

2. Dataset Overview

The dataset contains passenger information including demographic attributes, travel class, family relations, fare, and survival status. Key features include:
PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Fare, Embarked.

3. Data Cleaning & Preparation

Missing values in Age and Embarked were handled using median and mode imputation. Irrelevant features such as PassengerId and Name were removed. Categorical variables were encoded, and numerical features were standardized to prepare data for analysis.

4. Statistical Summary

Summary statistics such as mean, median, standard deviation, and distribution ranges were generated. Age distribution showed a right skew, while fare values indicated significant variability.

5. Univariate Analysis

Histograms and boxplots revealed:
- Majority passengers were aged between 20 and 40.
- Higher survival rate among females.
- Passengers in 1st class showed higher survival probabilities.

6. Bivariate & Multivariate Analysis

Correlation matrix and pair plots revealed:
- Strong correlation between Pclass and Fare.
- Survival strongly depended on Sex and Pclass.
- Family size moderately influenced survival.

7. Key Insights

- Females had significantly higher survival rates.
- First-class passengers had the highest survival probability.
- Children were more likely to survive.
- Higher fare passengers had better survival chances.

8. Conclusion

This EDA provided valuable insights into the survival patterns of Titanic passengers. The findings highlight the importance of socio-economic and demographic factors in survival prediction and form a strong foundation for future machine learning modeling.