Titanic Dataset EDA Report

1. Introduction

This report provides an exploratory data analysis (EDA) of the Titanic dataset from Kaggle. The goal is to uncover key insights and patterns related to passenger survival.

2. Dataset Overview

The dataset contains 891 rows and 12 columns including demographic and travel details like Name, Age, Sex, Ticket, Fare, Cabin, and Survival status. Some columns contain missing values.

3. Data Cleaning

Missing values in 'Age' were handled using median imputation. Rows missing 'Embarked' were dropped. 'Cabin' was excluded due to too many missing values.

4. Univariate Analysis

Key categorical variables such as 'Sex', 'Pclass', and 'Embarked' showed distinct survival distributions. Female passengers had a much higher survival rate. Most survivors were in first class.

5. Feature Engineering

We created new features like 'Family_Size' (SibSp + Parch) and 'Alone' (indicator for traveling alone). These features helped uncover survival trends linked to family presence.

6. Multivariate Insights

Heatmaps and pairplots were used to visualize relationships. 'Sex', 'Pclass', and 'Fare' were most correlated with survival. Higher fare and younger age groups had better chances.

Titanic Dataset EDA Report

7. Summary of Findings

- Women and children had higher survival rates.

- First-class passengers were more likely to survive.

- Passengers traveling with small families fared better.

- High fare values correlate with increased survival.

8. Next Steps

Future steps include building predictive models such as logistic regression and decision trees, and exploring more granular cabin-level survival data if available.