Titanic Dataset - EDA Summary Report

# 1. Introduction

This report presents the Exploratory Data Analysis (EDA) conducted on the Titanic Dataset. The goal is to understand the dataset's structure, identify patterns, handle missing data, detect outliers, and visualize important features to gain insights.

# 2. Dataset Overview

The Titanic dataset consists of 891 entries with 12 features including PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, and Embarked. It contains both categorical and numerical data.

# 3. Data Cleaning

Steps taken during data cleaning:

* - Missing 'Age' values were filled with the median age.
* - Missing 'Embarked' values were filled with the mode.
* - 'Cabin' column was dropped due to high number of missing values.
* - Duplicate rows were removed.
* - Outliers in the 'Fare' column were identified using the IQR method. 116 outliers detected.

# 4. Visualizations

Key plots generated during the EDA:

* - A bar chart showing gender distribution among passengers.
* - A histogram displaying the age distribution.
* - A correlation heatmap revealing relationships among numeric features.

# 5. Insights and Observations

Based on the EDA, the following insights were observed:

* - A large proportion of passengers were in the 20–40 age range.
* - Most passengers were in the third class (Pclass 3).
* - Females had a higher survival rate than males.
* - There was a strong correlation between Fare and Pclass.
* - Siblings/Spouses (SibSp) and Parents/Children (Parch) had some correlation with survival.