Exploratory Data Analysis Report - Titanic Dataset By Arpita Pani

1. Introduction

This report provides an in-depth exploratory data analysis (EDA) of the Titanic dataset. The primary objective is to understand the structure of the data, handle missing values, identify outliers, analyze distributions, and uncover patterns and relationships that influenced passenger survival.

2. Dataset Overview

The dataset consists of 891 passenger records with 12 features, including demographic and travel information.

Column	Description
Survived	Survival (0 = No, 1 = Yes)
Pclass	Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
Name	Name of passenger
Sex	Gender
Age	Age in years
SibSp	No. of siblings/spouses aboard
Parch	No. of parents/children aboard
Ticket	Ticket number
Fare	Passenger fare
Cabin	Cabin number
Embarked	Port of Embarkation (C/Q/S)

3. Missing Values

Missing data visualizations revealed that 'Cabin' has over 77% missing values and 'Age' has about 20%. Most other features are relatively complete. Missing values in 'Age' can be imputed; 'Cabin' may be dropped or categorized.

4. Univariate Analysis

Numerical features such as 'Fare' and 'Age' were analyzed using histograms. 'Fare' is right-skewed with several outliers. 'Age' has a near-normal distribution but with missing entries. Categorical features such as 'Sex' and 'Embarked' show that most passengers were male and embarked from Southampton.

5. Outlier Detection

Box plots revealed high fare outliers, particularly among 1st class passengers. Age outliers exist on both the younger and older ends.

6. Correlation Matrix

The correlation heatmap showed that 'Fare' and 'Pclass' are correlated with 'Survived'. 'Fare' positively correlates with survival; 'Pclass' is negatively correlated.

7. Bivariate Analysis

- Survived vs. Sex: Females had a significantly higher survival rate than males.
- Survived vs. Pclass: 1st class passengers had the highest survival rate.
- Age vs. Survived: Younger passengers were more likely to survive.
- Fare vs. Pclass: Fare increases with passenger class.

8. Key Findings

- Gender and class were strong indicators of survival.
- High-class passengers paid more and had better survival odds.
- Age distribution showed that children had better survival chances.
- Cabin feature is mostly missing and unreliable without preprocessing.

9. Conclusion

The EDA revealed important patterns in the Titanic dataset that could be useful for predictive modeling. Key features such as Sex, Pclass, Fare, and Age should be prioritized. Handling missing values and encoding categorical data are essential steps before model building.