EDA REPORT

Title: Exploratory Data Analysis (EDA) Report — Titanic Dataset
Task: Data Analyst Internship — Task 5
Dataset: train.csv

---

1. Introduction

This report presents a detailed Exploratory Data Analysis (EDA) of the Titanic dataset, which contains information about 891 passengers aboard the RMS Titanic. The goal of this analysis is to explore the structure of the dataset, identify key trends, and understand the factors associated with passenger survival.

---

2. Dataset Overview

The dataset includes passenger demographics, ticket details, fare information, family relations, cabin data, and survival status.

Key Columns

- **Survived:** 0 = No, 1 = Yes
- **Pclass:** Passenger class (1, 2, 3)
- **Name:** Passenger name
- **Sex:** Gender
- **Age:** Passenger age
- **SibSp:** Number of siblings/spouses aboard
- **Parch:** Number of parents/children aboard
- **Ticket:** Ticket number
- **Fare:** Ticket fare
- **Cabin:** Cabin number (highly missing)
- **Embarked:** Port of embarkation (C, Q, S)

---

3. Data Cleaning

During the preprocessing stage:

- Missing **Age** values were filled using the median age.
- Missing **Embarked** values were filled with the mode.
- **Cabin** contained too many missing values, so it was excluded from deep analysis.

- Dataset was inspected for duplicates and inconsistencies.

---

4. Univariate Analysis

Age

- Most passengers were **20 to 40 years** old.
- Very few elderly passengers.
- Age distribution is right-skewed.

Fare

- Highly right-skewed distribution.
- Majority of passengers paid **below 100**.
- A small number of passengers paid very high fares (mostly first-class).

SibSp & Parch

- Most passengers traveled **alone**.
- High family sizes were uncommon.

---

5. Bivariate Analysis

Sex vs Survived

- Females had a **significantly higher survival rate**.
- Gender appeared to be one of the strongest predictors of survival.

Pclass vs Survived

- First-class passengers survived the most.
- Third-class passengers had the lowest survival rate, indicating a socioeconomic impact.

Age vs Fare

- Higher-fare passengers showed a higher likelihood of survival.
- Survivors clustered around mid-age groups but were present across all ages.

---

6. Correlation Analysis

Key Findings

- **Pclass** and **Fare** showed strong correlation.
- **Fare** was positively correlated with survival.

- **Pclass** was negatively correlated with survival (lower class → lower survival).
- Family-related variables (**SibSp**, **Parch**) had weak correlation with survival but gave insights on group travel.

---

7. Key Insights

- **Gender and class** were the most influential factors for survival.
- **Children** had slightly better survival chances than adults.
- **Wealthier passengers** (indicated by higher fare and 1st class) survived more.
- **Traveling alone** often reduced survival chances.
- Missing data treatment did not distort distributions significantly.

---

8. Conclusion

The Titanic EDA reveals that survival depended heavily on **gender**, **passenger class**, and **ticket fare**. These findings create a strong foundation for predictive modeling and provide historical context on priorities during the Titanic evacuation.