Exploratory Data Analysis (EDA) – Titanic Dataset

Task 5 – Data Analysis Internship
Submitted By: *Roshani Chorghe*
Tools Used: Python, Pandas, NumPy, Seaborn, Matplotlib

Introduction

This report presents an Exploratory Data Analysis (EDA) on the Titanic dataset.
The main aim of this analysis is to understand the key factors that affected passenger survival during the Titanic tragedy. Various steps such as data cleaning, visualization, and statistical exploration were performed to derive meaningful insights from the dataset.

Dataset Information

The Titanic training dataset includes:

- **Rows:** 891
- **Columns:** 12
- **Important Attributes:**
    - *Survived* – Indicates whether the passenger survived (1) or not (0)
    - *Pclass* – Travel class of the passenger
    - *Name, Sex, Age*
    - *SibSp, Parch* – Family relationships on board
    - *Fare, Embarked* – Ticket fare and port of boarding

Data Cleaning Performed

The following preprocessing steps were completed before analysis:

- Removed the **Cabin** column due to a large number of missing values.
- Filled missing **Age** values using the **median**, ensuring consistency.
- Filled missing **Embarked** entries using the **mode**.
- Verified that no missing values remained after cleaning.

These steps ensured the dataset was fully prepared for accurate analysis.

Visualizations Included

(All visualizations used in this analysis were created and executed in **Jupyter Notebook**.
They include a variety of plots such as bar charts, count plots, histograms, boxplots, heatmaps,

pairplots, and outlier detection visuals.
Due to the large size of some charts, they are best viewed within the Jupyter Notebook environment where the code was executed.)

Visualizations created:

- Survival Count
- Survival Rate by Gender
- Survival Based on Passenger Class
- Age Distribution
- Boxplot of Age
- Correlation Heatmap
- Pairplot (Survived, Age, Fare, Pclass)
- Outlier Detection for Age and Fare

Key Insights / Findings

1. Gender-based Survival

- Women showed a far higher survival rate compared to men.
- This aligns with the “women and children first” evacuation approach.

2. Effect of Passenger Class

- Passengers in **1st class** had the highest survival likelihood.
- Those in **3rd class** had the lowest.
- This suggests that socio-economic status heavily influenced survival chances.

3. Age-related Patterns

- Children aged **0–15 years** survived at a higher rate.
- The age distribution is slightly **right-skewed**, with more younger passengers.
- A small number of elderly passengers appear as outliers.

4. Fare and Survival

- Higher ticket fare is associated with a higher chance of survival.
- This could indicate better cabin locations or priority access during evacuation.

5. Correlation Overview

- **Fare** shows a positive correlation with survival.
- **Pclass** shows a negative correlation (higher class number = lower survival).

- **Age** has a very weak correlation with survival.

Conclusion

The EDA of the Titanic dataset highlights several strong patterns in survival outcomes. Women, children, and passengers traveling in first class had noticeably higher chances of survival. Ticket fare also played an important role, indicating how social and economic factors influenced survival during the disaster. Overall, the analysis provides a clear understanding of the demographic and socio-economic variables that affected survival rates.