Task 3: Exploratory Data Analysis (EDA) on Titanic Dataset

1. Introduction

Exploratory Data Analysis (EDA) is a fundamental step in the data science process. It involves examining datasets to summarize their main characteristics, detect patterns, identify anomalies, and generate insights using statistical techniques and visualizations.

In this task, EDA was performed on the Titanic dataset obtained from Kaggle. The dataset contains demographic and travel-related information about passengers aboard the Titanic, including age, gender, ticket class, fare, and survival status. The objective of this analysis is to identify key factors that influenced passenger survival.

2. Dataset Overview

The dataset consists of **891 entries and 12 columns**, including both numerical and categorical variables .

Key Features:

- PassengerId
- Survived (Target variable)
- Pclass (Passenger Class)
- Name
- Sex
- Age
- SibSp (Siblings/Spouses aboard)
- Parch (Parents/Children aboard)
- Ticket
- Fare
- Cabin
- Embarked

Data Types:

- Numerical: Age, Fare, Pclass, SibSp, Parch
- Categorical: Sex, Embarked, Cabin, Ticket

3. Data Cleaning and Preprocessing

3.1 Missing Values

From the dataset summary :

- Age: 177 missing values
- Cabin: 687 missing values
- Embarked: 2 missing values

Handling Strategy:

- Age was filled using the median value.
- Embarked was filled using the mode.
- Cabin was dropped due to excessive missing values.

This ensured the dataset was clean and ready for analysis.

4. Univariate Analysis

4.1 Survival Distribution

The survival count visualization (Page 3) shows:

- Majority of passengers did not survive.
- Approximately 60% perished.
- Around 40% survived.

This indicates an imbalanced survival distribution.

4.2 Gender Distribution

From the gender distribution plot (Page 4) :

- Male passengers significantly outnumbered female passengers.
- However, further analysis is needed to assess survival by gender.

4.3 Age Distribution

The age histogram (Page 6) reveals:

- Most passengers were between 20 and 40 years old.
- The distribution is slightly right-skewed.
- Fewer elderly passengers were present.

5. Bivariate Analysis

5.1 Survival by Gender

Although more males were onboard, females had a higher survival proportion. This suggests that evacuation protocols may have prioritized women.

5.2 Survival by Passenger Class

The survival by passenger class chart (Page 5) shows:

- First-class passengers had the highest survival rate.
- Third-class passengers had the lowest survival rate.
- Socio-economic status strongly influenced survival chances.

This demonstrates inequality in survival probability based on travel class.

6. Correlation Analysis

The correlation heatmap (Page 7) provides deeper insights:

Key Observations:

- Pclass has a negative correlation with Fare (-0.55).
- Survived has moderate correlation with Fare (0.26).
- Survived has negative correlation with Pclass (-0.34).

Interpretation:

- Higher-class passengers paid higher fares.
- Higher fare correlates with better survival probability.
- Lower class (higher Pclass value) correlates with lower survival.

7. Multivariate Analysis

Age vs Fare by Survival (Page 8)

The scatterplot shows:

- Passengers who paid higher fares had better survival probability.
- Survival is more common among first-class passengers.
- Age alone does not strongly determine survival, but interacts with fare and class.

8. Key Insights from EDA

1. Gender played a crucial role — females had higher survival rates.
2. Passenger class significantly influenced survival chances.
3. Higher ticket fare increased probability of survival.
4. Age showed moderate influence but was not the strongest predictor.
5. Socio-economic factors were major determinants of survival.

9. Conclusion

This Exploratory Data Analysis on the Titanic dataset revealed that survival was not random but influenced by multiple demographic and socio-economic factors. Passenger class, gender, and fare were the strongest predictors of survival. First-class passengers and females had significantly better survival rates, highlighting the impact of evacuation priorities and economic status.

EDA proved essential in uncovering these patterns before building predictive models. The insights gained provide a strong foundation for further machine learning modeling and classification tasks.