Exploratory Data Analysis (EDA) on Titanic Dataset

1. Introduction

Exploratory Data Analysis (EDA) is a crucial step in any data science or machine learning project. It helps in understanding the structure of the dataset, identifying patterns, detecting anomalies, handling missing values, and extracting meaningful insights before applying any predictive models.

In this project, EDA is performed on the Titanic Dataset, which is a well-known real-world dataset containing information about passengers who traveled on the RMS Titanic. The primary objective of this analysis is to study passenger characteristics and identify the key factors that influenced survival during the Titanic disaster.

The analysis uses Python, along with popular data analysis and visualization libraries such as Pandas, Matplotlib, and Seaborn.

2. Dataset Description

The Titanic dataset consists of 891 rows (passengers) and 12 columns (features). Each row represents one passenger, and each column provides specific information about that passenger.

Attributes in the Dataset

- PassengerId: Unique identifier for each passenger
- Survived: Survival status (0 = No, 1 = Yes)
- Pclass: Passenger class (1st, 2nd, 3rd)
- Name: Passenger name
- Sex: Gender of the passenger
- Age: Age of the passenger
- SibSp: Number of siblings/spouses aboard
- Parch: Number of parents/children aboard
- Ticket: Ticket number
- Fare: Fare paid for the ticket
- Cabin: Cabin number
- Embarked: Port of embarkation (C, Q, S)

The dataset contains a mix of numerical and categorical variables, making it ideal for exploratory analysis.

3. Data Loading and Initial Exploration

The dataset was loaded using the Pandas library. The first few rows were displayed to understand the data format and observe sample values.

Initial inspection showed that:

- Some columns contained missing values.
- Data types were correctly assigned (integers, floats, and objects).
- The dataset size was manageable for analysis.

Basic statistical measures such as mean, median, minimum, maximum, and standard deviation were examined using summary statistics to understand the distribution of numerical features like age and fare.

4. Understanding Dataset Structure

The dataset structure was analyzed using info() and describe() methods.

Key Observations

- The Age column had missing values.
- The Cabin column had a large number of missing values.
- The Embarked column had a few missing values.
- No duplicate records were present.
- Numerical columns such as Fare and Age showed a wide range of values, indicating variability among passengers.

This step helped in identifying which columns required preprocessing before visualization and analysis.

5. Data Cleaning and Preprocessing

Data preprocessing is essential to ensure the quality and reliability of analysis results.

5.1 Handling Missing Values

- Age: Missing values were filled using the median age. The median is preferred over the mean because it is less affected by outliers.
- Embarked: Missing values were filled using the mode, as it is a categorical variable.
- Cabin: The Cabin column had 687 missing values out of 891, which is approximately 77%. Due to this excessive missing data, the column was dropped from the dataset.

5.2 Duplicate Data Handling

- Duplicate records were checked and none were found.

After preprocessing, the dataset contained no missing values, making it clean and suitable for visualization and further analysis.

6. Exploratory Data Analysis and Visualization

Visualization plays a key role in understanding relationships and patterns in data. Various plots were created using Matplotlib and Seaborn.

6.1 Survival Distribution

A count plot was used to visualize the number of passengers who survived and those who did not.

Observation:

- The number of passengers who did not survive was higher than those who survived.
- This indicates that survival was not common and depended on certain factors.

6.2 Age Distribution

A histogram with a kernel density estimate (KDE) was plotted for passenger ages.

Observation:

- Most passengers were between 20 and 40 years old.
- There were fewer children and elderly passengers.
- Age distribution was slightly right-skewed.

6.3 Survival by Gender

A count plot was created to compare survival rates between males and females.

Observation:

- Females had a significantly higher survival rate than males.
- A large number of male passengers did not survive.
- This supports the historical rule of "women and children first" during evacuation.

6.4 Survival by Passenger Class

Passenger class was analyzed against survival using a count plot.

Observation:

- Passengers in 1st class had the highest survival rate.
- Survival decreased as passenger class moved from 1st to 3rd.
- Passengers in 3rd class had the lowest survival rate.

This suggests that socio-economic status played an important role in survival.

7. Key Insights from the Analysis

Based on the EDA, the following insights were derived:

1. Gender Influence
   Females were more likely to survive than males.
2. Passenger Class Influence
   Higher-class passengers had better survival chances.
3. Age Factor
   Younger passengers had slightly higher survival rates.
4. Economic Factor
   Higher ticket fares were associated with better survival outcomes.
5. Data Quality Importance
   Proper handling of missing data significantly improved the reliability of analysis.

8. Conclusion

This Exploratory Data Analysis on the Titanic dataset successfully uncovered important patterns and relationships affecting passenger survival. The analysis highlighted that gender, passenger class, age, and fare were key factors influencing survival outcomes.

The project demonstrated the importance of:

- Data cleaning and preprocessing
- Visual exploration of data
- Interpreting patterns before applying machine learning models

Overall, this EDA provides a strong foundation for building predictive models such as survival classification using machine learning algorithms.