Exploratory Data Analysis (EDA) Report on Titanic Dataset

By Arya Kumari
Date: 25/01/25
OWL-Ai AI Intern

1. Introduction

This report presents an Exploratory Data Analysis (EDA) performed on the **Titanic dataset**, sourced using the Seaborn library. The objective of this analysis is to understand the structure of the dataset, identify patterns, detect missing values, and gain insights into factors influencing passenger survival.

2. Dataset Overview

The Titanic dataset contains information about passengers aboard the RMS Titanic, including demographic details, ticket class, and survival status.

Key Attributes:

- `survived`: Survival status (0 = No, 1 = Yes)
- `pclass`: Ticket class (1st, 2nd, 3rd)
- `sex`: Gender of the passenger
- `age`: Age of the passenger
- `sibsp`: Number of siblings/spouses aboard
- `parch`: Number of parents/children aboard
- `fare`: Ticket fare

- `embarked`: Port of embarkation
- `deck`: Deck level (contains many missing values)

Initial inspection using `head()` and `info()` revealed the dataset structure, data types, and presence of missing values.

3. Summary Statistics

Descriptive statistics were generated for numerical features using `describe()`. This provided insights into:

- Central tendencies (mean, median)
- Spread of data (standard deviation, min, max)
- Distribution of variables such as age and fare

Notably, fare values showed a wide range, indicating potential outliers, while age had missing entries.

4. Missing Value Analysis

Missing values were identified using `isnull().sum()`.

Key Observations:

- `age` had several missing values.
- `deck` had a very high number of missing entries.
- `embarked` had a small number of missing values.

5. Data Visualization and Insights

5.1 Age Distribution

A histogram with a KDE curve was plotted for passenger age.

- Most passengers were between **20 and 40 years old**.
- The distribution was slightly right-skewed.

5.2 Survival Rate by Ticket Class

A bar plot showed survival rates across passenger classes.

- **1st class passengers** had the highest survival rate.
- Survival probability decreased significantly for **3rd class passengers**.

5.3 Survival Rate by Gender

A bar plot comparing survival by gender revealed:

- **Females had a significantly higher survival rate** than males.
- Gender was one of the strongest indicators of survival.

5.4 Correlation Analysis

A heatmap of correlations among numerical variables was generated.

Findings:

- `survived` showed positive correlation with `fare` and negative correlation with `pclass`.
- Strong relationships were observed among family-related features (`sibsp` and `parch`).

6. Data Cleaning Steps

Based on the EDA findings, the following preprocessing steps were applied:

- Missing `age` values were filled using the **median age**.

- The `deck` column was dropped due to excessive missing values.
- Missing `embarked` values were filled using the **mode**.

Post-cleaning checks confirmed that all missing values were appropriately handled.

7. Conclusion

The exploratory analysis revealed that **gender**, **ticket class**, and **fare** were key factors influencing survival on the Titanic. Females and higher-class passengers had a much greater chance of survival. The dataset required moderate cleaning, mainly handling missing age values and removing the deck feature.

This EDA provides a solid foundation for further steps such as feature engineering and predictive modeling.