Elevate Labs

Task 5: Exploratory Data Analysis

Presented By:

Abijin Suvedha A

Project Overview

This project performs Exploratory Data Analysis (EDA) to understand the key factors that influenced survival. The analysis uses Python, Pandas, Matplotlib, and Seaborn to extract insights through visualizations and statistical summaries.

Objectives

- Explore the structure and contents of the dataset.
- Perform descriptive statistical analysis.
- Visualize passenger distributions and survival patterns.
- Identify relationships and trends across features.
- Summarize findings with key survival factors.

Dataset Description

The dataset consists of three files:

- train.csv → Passenger data with survival labels.
- test.csv → Passenger data without survival labels (for prediction).
- gender_submission.csv → Sample submission file.

Attributes:

- PassengerId: Unique identifier
- Survived: Target variable (0 = No, 1 = Yes)
- Pclass: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- Name: Passenger name
- Sex: Gender
- Age: Age in years
- SibSp: # of siblings/spouses aboard
- Parch: # of parents/children aboard
- Ticket: Ticket number

- Fare: Fare paid by passenger
- Cabin: Cabin number
- Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

4. Data Exploration

4.1 Initial Checks

- .info() → Identified missing values in Age, Cabin, and some in Embarked.
- .describe() → Showed numerical distributions:
    - Average passenger age ≈ 29 years.
    - Fare values varied widely, with outliers at very high values.

4.2 Value Counts

- Survived: More passengers did not survive than survived.
- Pclass: Most passengers traveled in 3rd class.
- Sex: More males than females were aboard.
- Embarked: Majority of passengers embarked from Southampton.

5. Visual Analysis

5.1 Univariate Analysis

- Survival Distribution: Less than 40% of passengers survived.
- Age Histogram: Majority of passengers were between 20–40 years old.
- Fare Distribution: Most fares were below 50, with some extreme outliers.

5.2 Bivariate Analysis

- Survival by Gender: Females had much higher survival rates than males.
- Survival by Class: 1st class passengers had the highest survival rate, followed by 2nd and then 3rd.
- Age vs Survival (Boxplot): Younger passengers (children) showed better chances of survival.
- Age vs Fare (Scatterplot): Higher fare passengers, often in higher classes, had better survival outcomes.

5.3 Multivariate Analysis

- Pairplot: Revealed interactions between Age, Fare, Pclass, SibSp, and Parch with Survival.

- Correlation Heatmap:
    - Positive correlation: Fare ↔ Survival.
    - Negative correlation: Pclass ↔ Survival (higher class = better survival).
    - Weak correlations: SibSp and Parch had minor influence.

Result

1. Gender: Females survived at much higher rates than males.
2. Class: 1st class passengers had better chances of survival compared to 2nd and 3rd class.
3. Age: Children and younger passengers were more likely to survive.
4. Fare: Higher fares (wealthier passengers) correlated with better survival.
5. Family: Small family groups (SibSp + Parch) slightly increased survival chances, while large families reduced chances.
6. Embarkation Port: Passengers from Cherbourg (C) showed slightly higher survival compared to Southampton (S) and Queenstown (Q).

Tools & Libraries Used

- Python 3.9
- Pandas → Data handling and descriptive statistics
- Matplotlib → Visualization
- Seaborn → Advanced plotting and heatmaps

Conclusion

The analysis confirms that gender, class, age and fare were the most significant factors influencing survival. Females, 1st class passengers, younger individuals and wealthier travellers had better odds of survival.