📄 Exploratory Data Analysis (EDA) Report

Internship Report :

Company: Elevate Labs
Department: Data Analytics
Task 5 : Exploratory Data Analysis

Prepared By:

Name: Purnakam Shrivastava
Position: Data Analytics Intern

Submission Date:

June 2025

Report Objective:

This report presents the Exploratory Data Analysis (EDA) performed on the Titanic dataset. The objective is to explore, visualize, and summarize the dataset to extract meaningful insights and identify patterns that can help in further analysis and modeling.

📚 Table of Contents

1. About Dataset

The Titanic dataset is a widely studied dataset in data science and machine learning. It contains detailed information about passengers aboard the RMS Titanic, including demographic details, ticket information, and survival outcomes. The dataset is frequently used to explore classification techniques and identify factors influencing survival rates.

Dataset Source

This dataset is publicly available on Kaggle: **Titanic Dataset on Kaggle**

Dataset Overview

The dataset comprises 891 records and includes the following columns:

Column Name	Description	Data Type	Non-Null Count
PassengerId	Unique identifier for each passenger	Integer	891
Survived	Survival status (0 = No, 1 = Yes)	Integer	891
Pclass	Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd)	Integer	891
Name	Passenger's name	Object	891
Sex	Gender	Object	891
Age	Age in years	Float	714
SibSp	Number of siblings/spouses aboard	Integer	891
Parch	Number of parents/children aboard	Integer	891

Ticket	Ticket number	Object	891
Fare	Passenger fare	Float	891
Cabin	Cabin number	Object	204
Embarked	Port of embarkation	Object	889

Explanation of Summary Statistics

The summary statistics provide essential insights into the distribution and characteristics of the key numerical variables in the Titanic dataset:

- **Count:** Number of non-missing (non-null) values for each feature. For example, the Age feature contains 714 valid entries, indicating missing data in some rows.

- **Mean:** The average value for each feature. The average age of passengers is approximately 29.7 years.

- **Standard Deviation (Std Dev):** Reflects variability in the data. A larger value indicates more spread. The fare, for instance, has a high standard deviation (49.69), showing a wide range in ticket prices.

- **Minimum (Min):** The smallest value observed. The youngest passenger was just 0.42 years old.

- **25th Percentile (Q1):** The value below which 25% of the observations fall. For fare, 25% of passengers paid less than or equal to 7.91.

- **Median (50th Percentile or Q2):** The middle value splitting the data into two halves. The median age is 28 years.

- **75th Percentile (Q3):** The value below which 75% of the data lies. For example, 75% of passengers paid fares less than or equal to 31.

- **Maximum (Max):** The largest observed value. The highest fare paid was 512.33.

2. Libraries Used

The following Python libraries and tools were utilized for the exploratory data analysis and visualization of the Titanic dataset:

- **NumPy:** For numerical operations and handling arrays efficiently.

- **Pandas:** For data manipulation and analysis, providing powerful data structures.

- **Matplotlib:** A foundational plotting library for creating static visualizations.

- **Seaborn:** Built on Matplotlib, used for enhanced statistical data visualization with improved aesthetics.

- **Scikit-learn (LabelEncoder):** For encoding categorical variables into numeric formats to facilitate machine learning tasks.

- **SciPy:** For conducting statistical tests and advanced scientific computations.

- **Plotly Express:** For creating interactive and dynamic visualizations that enable better exploration of data relationships.

Visualization Settings

To make the charts clear and attractive, these styles were used:

- Seaborn’s **whitegrid** for a clean background with light grid lines.
- Matplotlib’s **ggplot** for a modern, professional look.
- Figure size set to 10 by 6 inches for better visibility.
- Seaborn’s **Set2** color palette for distinct and pleasant colors.

---

3. Missing Data Analysis

During the initial examination of the Titanic dataset, the following columns were found to have missing values:

- **Cabin:** 687 missing values
- **Age:** 177 missing values
- **Embarked:** 2 missing values
- All other columns have no missing values.

Handling Missing Data

- The **Cabin** column was dropped due to the high number of missing entries.

- Missing values in the **Age** column were filled with the median age of the dataset to reduce the impact of outliers.

- Missing values in the **Embarked** column were filled with the mode, which is the most common port of embarkation.

---

4. Feature Engineering

New features were created to improve analysis:

- **Family Size:** Combined siblings/spouses and parents/children plus one.

- **Age Group:** Passengers divided into Child, Teen, Young Adult, Adult, and Senior categories.

- **Encoding:**

    - Gender converted to numeric (male/female).

    - Embarked and Age Group converted into separate binary columns.

---

5. Univariate Analysis

Categorical Features

- **Survival:** 549 did not survive, 342 survived

- **Passenger Class:** 216 in 1st, 184 in 2nd, 491 in 3rd class

- **Gender:** 314 females, 577 males

- **Embarkation Port:** 168 Cherbourg, 77 Queenstown, 646 Southampton

Numerical Features

- **Age Groups:** Young Adult (422), Adult (241), Teen (95), Child (69), Senior (64)
- **Fare:** Majority paid under 20 (496), fewer paid 20-40 (191), and very few above 80 (max 512.33)
- **Family Size:** Mostly solo travelers (537), with smaller counts for larger families up to 11 members

6. Bivariate Analysis

This section explores how survival rates vary across key passenger features, revealing important patterns affecting chances of survival.

Survival Rate by Sex

- Female: Approximately 74–75% survived
- Male: Approximately 18–19% survived

Survival Rate by Passenger Class (Pclass)

- 1st Class: Highest survival rate around 62–64%
- 2nd Class: Moderate survival rate approximately 47–49%
- 3rd Class: Lowest survival rate between 23–25%

Survival Rate by Age Group

- Children (0–12 years): Highest survival rate around 57–58%

- Teens (13–18 years): Survival rate approximately 40–41%
- Young Adults (19–30 years): Survival rate about 32–34%
- Adults (31–50 years): Survival rate around 41–43%
- Seniors (51+ years): Lowest survival rate between 33–35%

These trends highlight the influence of gender, socio-economic class, and age on survival chances aboard the Titanic.

6.2 Chi-Square Test Results

- **Passenger Class vs Survived:**
  Chi-square value = 102.89, p-value = 0.0000
  → Strong, statistically significant association between passenger class and survival.
- **Sex vs Survived:**
  Chi-square value = 260.72, p-value = 0.0000
  → Very strong and significant relationship between gender and survival.
- **Embarked vs Survived:**
  Chi-square value = 25.96, p-value = 0.0000
  → Statistically significant association between port of embarkation and survival.

---

7. Multivariate Analysis

This section explores relationships between multiple variables to better understand factors influencing survival.

7.1 Correlation Heatmap of Numeric Features

- Survived correlates negatively with Pclass (-0.34) and Sex (-0.54), showing lower survival for lower classes and males.
- Fare has a positive correlation with survival (0.26), indicating higher fare passengers had better chances.
- FamilySize strongly correlates with SibSp (0.89) and Parch (0.78), but only weakly with survival (0.02).
- Negative correlations exist between Age and Pclass (-0.34), and Age and FamilySize (-0.25).

7.2 Pairplots Colored by Survival Status

- Survivors tend to be younger and paid higher fares.
- Non-survivors generally had larger family sizes.
- Differences in SibSp and Parch between survivors and non-survivors suggest family structure influenced survival.

7.3 Interaction Analysis: Passenger Class and Gender Survival Rates

- 1st Class Females: 90%–95% survival
- 2nd Class Females: 85%–90% survival
- 3rd Class Females: 50%–53% survival
- 1st Class Males: 37%–39% survival
- 2nd Class Males: 16%–18% survival
- 3rd Class Males: 10%–15% survival

Conclusion: Survival was strongly influenced by both gender and socio-economic class.

---

8. Advanced Visualizations

8.1 Faceted Grid Plots

- **Age Distribution by Sex:**
  - Female peak: Age 25–30, around 70–73 passengers.
  - Male peak: Age 25–30, around 180 passengers.
- **Survival Count by Class and Sex:**
  - **Females:**
    - 1st Class: 75–85 survived.
    - 2nd Class: 60–65 survived.
    - 3rd Class: Survivors and non-survivors nearly equal (~60–65).
  - **Males:**
    - 1st Class: 45–48 survived, 55–60 did not.
    - 2nd Class: 10–15 survived, 80–85 did not.
    - 3rd Class: 45–48 survived, ~300 did not.

8.2 Interactive Scatter Plot

- Shows Age vs Fare, colored by survival status.
- Allows hovering to view Sex and Pclass details.
- Helps explore survival trends, clusters, and outliers quickly.

---

9. Outlier Analysis and Treatment

9.1 Visual Detection of Outliers

Age

- Frequent Range: 21 to 35 years.
- Normal Range: 5 to 55 years.
- Significant Outliers: Near 1 year and around 80 years.

Fare

- Normal Range: 0 to 80 units.
- Frequent Range: 5 to 20 units.
- Extreme Outlier: Highest fare observed at 512 units.

9.2 Outlier Detection and Treatment Decision

Outlier Detection (Using IQR Method)

- Fare Outliers Identified: 116 instances.

- Age Outliers Identified: 66 instances.

Decision:

- Outliers Not Removed: To retain potentially valuable data.
- Fare Transformed: Applied log(1 + Fare) to reduce skewness and manage high-value outliers.
- Age Kept As-Is: Outliers in age considered acceptable and potentially important for survival analysis.