Titanic Dataset - Exploratory Data Analysis (EDA)

Author: *Swati Raina*
Date: *28 April, 2025*

---

1. Introduction

This report presents an exploratory data analysis (EDA) of the Titanic dataset.
The objective is to extract meaningful insights into the survival patterns of passengers based on features such as age, gender, fare, and class.
The tools used are Python libraries: Pandas, Matplotlib, and Seaborn.

---

2. Dataset Overview

The Titanic dataset contains information about 891 passengers, including:

- **Features**: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked
- **Target Variable**: Survived (0 = No, 1 = Yes)

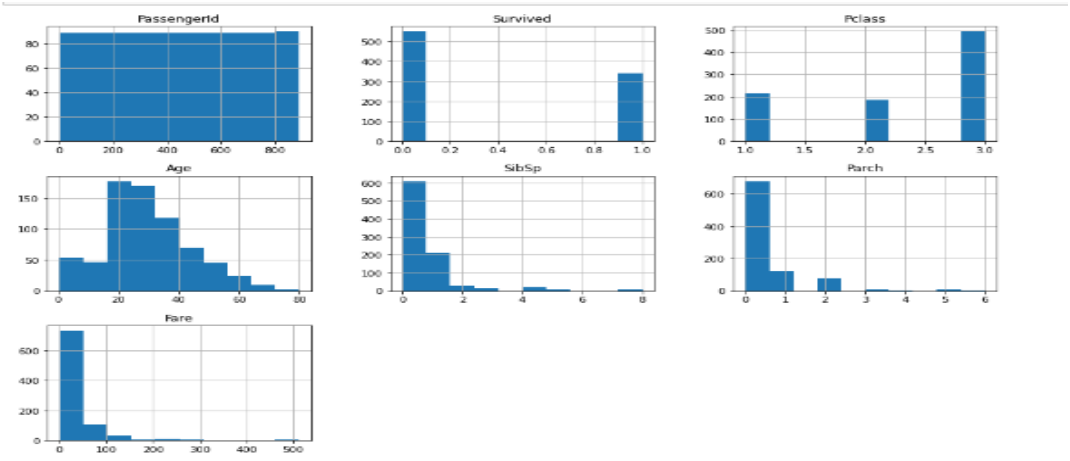
Data Summary:

- **Numerical Features**: Age, SibSp, Parch, Fare
- **Categorical Features**: Sex, Pclass, Embarked, Cabin
- **Missing Values**:
    - Age - some missing
    - Cabin - many missing
    - Embarked - few missing

---

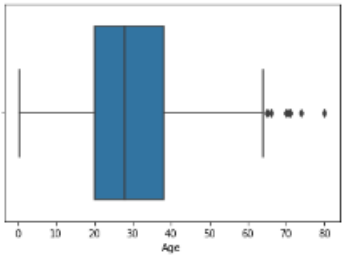
3. Univariate Analysis

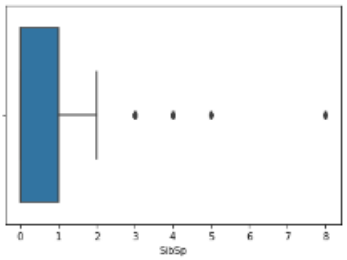
3.1 Histogram (All Numeric Features)

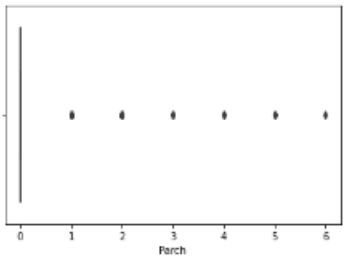


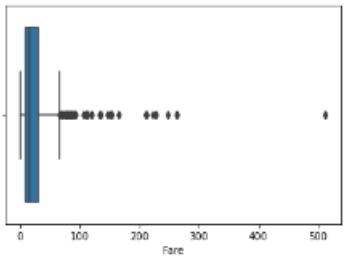
- Age is approximately normally distributed but slightly right-skewed.
- Fare is highly right-skewed with many low values and some extremely high values (outliers).
- SibSp and Parch mostly have small values (0–2), indicating most people traveled alone or with few family members.

3.2 Boxplots (All Numeric Features)







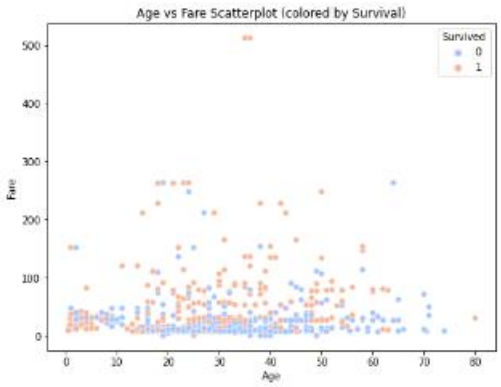


- **Age** has mild outliers (young children and elderly).
- **Fare** shows **strong outliers**, with some fares much higher than typical passengers.
- **SibSp** and **Parch** show extreme values but for a very small number of passengers.
- Boxplots help visualize spread and skewness in the data.

---

4. Bivariate Analysis

4.1 Age vs Fare Scatterplot (Colored by Survival)

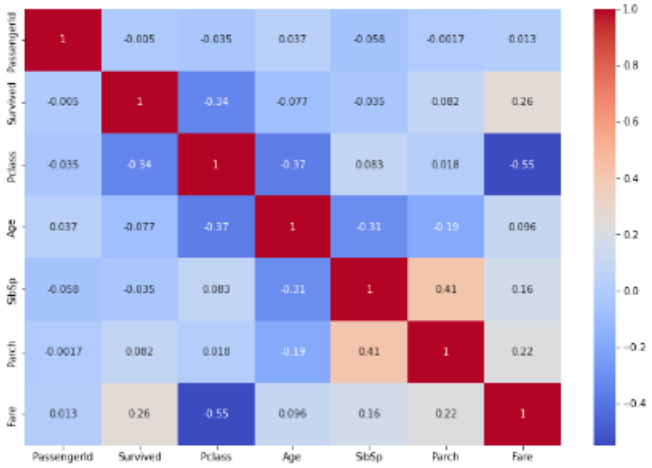


- Passengers who paid higher fares had a higher survival rate (more red points at higher fares).
- Survivors (Survived = 1) are more clustered at **higher Fare values**.
- There is no strong visible relationship between Age and Fare directly.
- Younger and middle-aged passengers exist across all fare ranges.

---

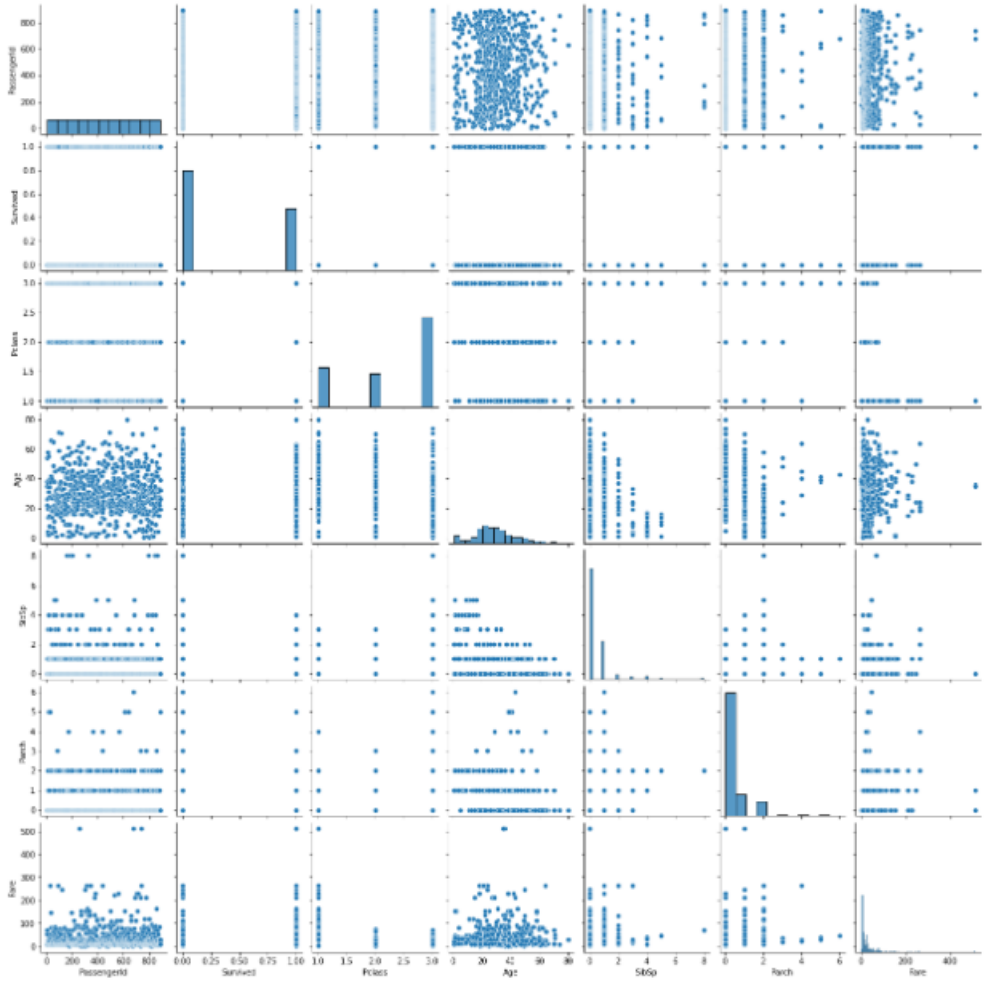
5. Multivariate Analysis

5.1 Heatmap of Feature Correlations



- **Fare** is **positively correlated** with **Survival**.
- **Pclass** is **negatively correlated** with **Survival**.
- SibSp and Parch are **positively correlated** with each other (families traveling together).

5.2 Pairplot



- There is no strong visible linear relationship between most variables.
- Some clustering is seen based on survival in features like Fare and Pclass.
- Outliers are visible, especially for Fare (some very high fare passengers)

---

6. Summary of Key Findings

Aspect	Key Insight
Survival by Fare	Higher fare passengers had better survival chances.
Survival by Class	1st Class passengers had much higher survival rates than 2nd and 3rd Class.
Age Distribution	Majority of passengers were between 20–40 years old.
Family Size	Traveling with family members affected survival slightly.
Missing Data	Many missing values in Cabin; some missing in Age and Embarked.

---

7. Conclusion

The EDA reveals that **socio-economic status** (represented by ticket class and fare) played a major role in determining survival chances aboard the Titanic.
Younger passengers and those from 1st Class had better chances of survival.
Future modeling could focus on predicting survival using engineered features like FamilySize, Title from Name, and imputation of missing values in Age.