

# Exploratory Data Analysis (EDA) of Titanic Dataset

## Steps Performed:

1. Imported libraries NumPy, Pandas, Matplotlib, and Seaborn
2. Extracted data into Jupyter Notebook using `.read_csv`
3. Split "Name" column into "Title", "FirstName", and "LastName"
4. Used `.head()` and `.tail()` to check changes in dataframe
5. Found the position of the original "Name" column, dropped the column, and shifted the columns "Title", "FirstName", and "LastName" into the same position using `.get_loc()`, `.drop()`, and `.insert()`, and `.pop()` respectively.
6. Edited column names "Pclass", "Parch", "Ticket", and "Fare" to "PClass", "ParCh", "TicketNo", and "TicketFare" for consistent format and enhanced readability
7. Used functions `.info()` for column details and `.desc()` for preliminary statistical data
8. Used `.isna().sum()` to find out null values in any column
9. Filled null values in "Age", "Cabin", and "Embarked" columns with "0" and empty string, respectively
10. Removed rows without an embarkation point

## Observations:

1. *Derived "survival rate by passenger class", "survival rate by embarkation point", "survival rate by gender", and "survival rate by age" values. Plotted subplots with the 4 data.*

### ➤ Survival rate by passenger class (Bar Graph)

#### • Key Elements in the Plot

- **X-axis (PassengerClass)** → Represents the passenger class.
- **Y-axis (SurvivalRate)** → Represents passenger survival rate.
- **Color (Shades of blue ["#00008B", "#0000FF", "#4169E1"])**

#### • Insights derived:

Passengers in the First Class had the highest survival rate, followed by Class Two and Class Three passengers.

➤ **Survival rate by embarkation point (Bar Graph)**

- **Key Elements in the Plot**

- **X-axis (Embarked)** → Represents the embarkation point.
- **Y-axis (SurvivalRate)** → Represents passenger survival rate.
- **Color** (Shades of green ["#006400", "#008000", "#228B22"])

- **Insights derived:**

Passengers who embarked from "C" i.e., Cherbourg (France) had the highest survival rate. It was followed by passengers from "S" i.e., Southampton (England) and "Q" i.e., Queenstown (Ireland) respectively.

➤ **Survival rate by passenger gender (Bar Graph)**

- **Key Elements in the Plot**

- **X-axis (Sex)** → Represents the passenger genders (Male or Female).
- **Y-axis (SurvivalRate)** → Represents passenger survival rate.
- **Color** (Shades of red ["#8B0000", "#B22222", "#FF0000"])

- **Insights derived:**

Female passengers were among the highest survivors.

➤ **Survival rate by passenger age group (Histogram)**

- **Key Elements in the Plot**

- **X-axis (AgeGroup)** → Represents the passenger age groups (0-10, 11-20, 21-30, and so on).
- **Y-axis (SurvivalRate)** → Represents passenger survival rate.

- **Insights derived:**

Passengers with highest survival rate were aged between 31 to 60 years at the time of the incident followed by the age group 11 to 30.

2. ***Created a pair plot to find out the correlation between the different values based on survival state***

### 3. *Found out the age distribution of the passengers on board the Titanic*

- **Key Elements in the Plot**

- **X-axis (Age)** → Represents the passenger genders (Male or Female).
- **Y-axis (SurvivalRate)** → Represents passenger survival rate.
- **Color (Shade of purple (#800080))**

- **Insights derived:**

- **Peak around ages 0-10:** Indicates a significant number of young passengers on board.
- **Smaller peaks at ages 20, 30, and 40:** Suggests that many passengers were in these age brackets.
- **Gradual decline in older ages:** Fewer passengers were in the 60+ age group.

### 4. *A box plot was created to plot the fare distribution among the passengers.*

- **Insights derived:**

- There is **high variability** in ticket prices, with fares ranging from low-cost to extremely expensive.
- Some passengers paid **significantly more**, as seen by the presence of **outliers**.

### 5. *A pie chart was created to find the passenger distribution based on embarkation points.*

- **Key Elements in the Plot**

- **Legends:**

- S = Southampton (England)
- C = Cherbourg (France)
- Q = Queenstown (Ireland)

- **Insights derived:**

- The **highest percentage of passengers boarded from Southampton (England) (72.44%)**, the main departure point of the Titanic, followed by **Cherbourg (France) (18.9%)**, a major transatlantic port where many wealthy passengers boarded.

6. ***A scatter plot depicts the Fare vs. Passenger Class analysis based on Survival.***

- **Key Elements in the Plot**

- **X-axis (TicketFare)** → Represents the ticket prices.
- **Y-axis (PClass)** → Represents passenger classes (e.g., 1st, 2nd, and 3rd class).
- **Color (Survived)** → Shows survival status (color bar helps interpret survival rates).

- **Insights derived:**

- Passengers in **1st class (PClass = 1)** generally paid **higher fares**.
- Lower-class passengers (PClass = 3) paid **significantly lower fares**.
- Passengers with higher ticket fares and higher class had the most chances of surviving.

### **Summary of Findings:**

- **First Class passengers** had the highest survival rate, followed by Second and Third Class.
- **Female passengers** and those aged **31–61 years** had the best survival chances.
- Passengers who **embarked from Cherbourg** had a higher likelihood of survival compared to those from Southampton or Queenstown.
- The majority of passengers boarded at **Southampton (72.44)**.
- **Ticket fares varied greatly**, with higher fares and First Class travel correlating strongly with survival.
- The **age distribution** showed peaks in young (0–10) and adult (20s–40s) age groups, with fewer elderly passengers.
- Overall, **higher class, higher fare, female gender, and middle-age range** were key factors linked to survival.