Titanic Dataset - Exploratory Data Analysis (EDA) Report

**Objective:** Extract meaningful insights using visual and statistical exploration techniques on the Titanic dataset.

**Tools Used:** Python (Google Collab), Pandas (Data manipulation), Matplotlib & Seaborn (Data Visualization)

# Dataset Overview:

The Titanic dataset contains data on passengers aboard the RMS Titanic. The goal is to explore which factors influenced survival rates.

|  |  |
| --- | --- |
| Feature | Description |
| PassengerId | Unique ID of passenger |
| Survived | Survival (0 = No, 1 = Yes) |
| Pclass | Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd) |
| Name | Full name of the passenger |
| Sex | Gender |
| Age | Age in years |
| SibSp | No of siblings/spouses aboard |
| Parch | No of parents/children aboard |
| Ticket | Ticket number |
| Fare | Ticket fare |
| Cabin | Cabin number |
| Embarked | Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton) |

# Data Summary:

* **Total Rows:** 891
* **Total Columns:** 12
* **Missing Values:**  
   • **Age:** 177 missing  
   • **Cabin:** 687 missing  
   • **Embarked:** 2 missing

# Missing Value Handling

* Age → Filled with median (28.0)
* Embarked → Filled with mode ('S')
* Cabin → Dropped (too many missing values)

# Univariate Analysis

* **Survival**: 62% died, 38% survived. Class imbalance present.
* **Pclass**: Majority in 3rd class
* **Sex:** More males than females (65% male)
* **Age:** Right-skewed; most between 20–40 years old
* **Fare:** Highly skewed; most fares under $50
* **Embarked:** Most passengers boarded from Southampton (S)

# Bivariate Analysis

* **Gender vs Survival:**  
  - Females had ~75% survival  
  - Males had ~19% survival
* **Pclass vs Survival:**  
  - 1st class: 63% survived  
  - 3rd class: 24% survived
* **Age vs Survival:**  
  - Children had higher survival  
  - Adults 20–40 had higher fatalities
* **Fare vs Survival:**- Higher fare = higher survival

# Correlation Matrix

* **Pclass ↔ Survived:** strong negative correlation
* **Fare ↔ Survived:** moderate positive correlation
* No multicollinearity observed

# Pairplot Insights

* Higher fare and 1st class → higher survival
* Clusters observed in Fare, Age, Pclass combinations

# Summary of Key Insights

* **High Impact:**  
  - Sex (female > male)  
  - Pclass (1st class)  
  - Fare (higher fare = higher survival)
* **Other:**  
  - Most were 3rd class males  
  - Port 'S' was most common  
  - Missing values handled logically

# EDA Outcome

* Identified important predictors
* Highlighted class imbalance
* Addressed missing values professionally
* Suggested modeling directions (logistic regression, trees)