**Titanic Dataset - Exploratory Data Analysis (EDA) Report**

**🔍 1. Introduction**

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, the Titanic sank after hitting an iceberg, resulting in the deaths of over 1,500 passengers and crew. The goal of this analysis is to explore the passenger data to extract meaningful insights about survival patterns.

This report is part of a data analyst internship task focusing on **Exploratory Data Analysis (EDA)**, where we aim to discover patterns, detect anomalies, and test assumptions through visual and quantitative methods.

We will use the Titanic dataset available from Kaggle, which includes information like age, gender, passenger class, and fare, and whether the passenger survived.

**🛠 2. Libraries Used**

The following Python libraries were used for this analysis:

* **Pandas**: for data manipulation and cleaning.
* **NumPy**: for numerical operations.
* **Matplotlib**: for creating basic plots and visualizations.
* **Seaborn**: for advanced statistical plots with aesthetic themes.

Example import code:

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

**📂 3. Dataset Overview**

* **Source**: [Kaggle Titanic Dataset](https://www.kaggle.com/c/titanic/data)
* **Total Rows**: 891
* **Total Columns**: 12
* **Target Variable**: Survived (0 = No, 1 = Yes)

**Key Features:**

| **Column** | **Description** |
| --- | --- |
| PassengerId | Unique identifier for each passenger |
| Survived | Survival status (0 = No, 1 = Yes) |
| Pclass | Passenger class (1 = 1st, 2 = 2nd, 3 = 3rd) |
| Name | Passenger's name |
| Sex | Gender |
| Age | Age in years |
| SibSp | # of siblings/spouses aboard |
| Parch | # of parents/children aboard |
| Ticket | Ticket number |
| Fare | Amount paid for ticket |
| Cabin | Cabin number |
| Embarked | Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton) |

**📉 4. Data Cleaning & Missing Values**

* **Age**: 177 missing values — imputed with median based on passenger class and gender.
* **Cabin**: 687 missing values — dropped due to high missing rate.
* **Embarked**: 2 missing values — filled with mode ('S').

**📊 5. Univariate Analysis**

**Age Distribution**

* Most passengers are aged between **20 and 40** years.

**Gender Distribution**

* Around **65% of passengers were male**.

**Passenger Class**

* The majority of passengers were from **3rd class**.

**🔄 6. Bivariate & Multivariate Analysis**

**Gender vs Survival**

* **Females had a much higher survival rate** (~74%) compared to males (~19%).

**Class vs Survival**

* Passengers in **1st class had the highest survival rate**, followed by 2nd and then 3rd.

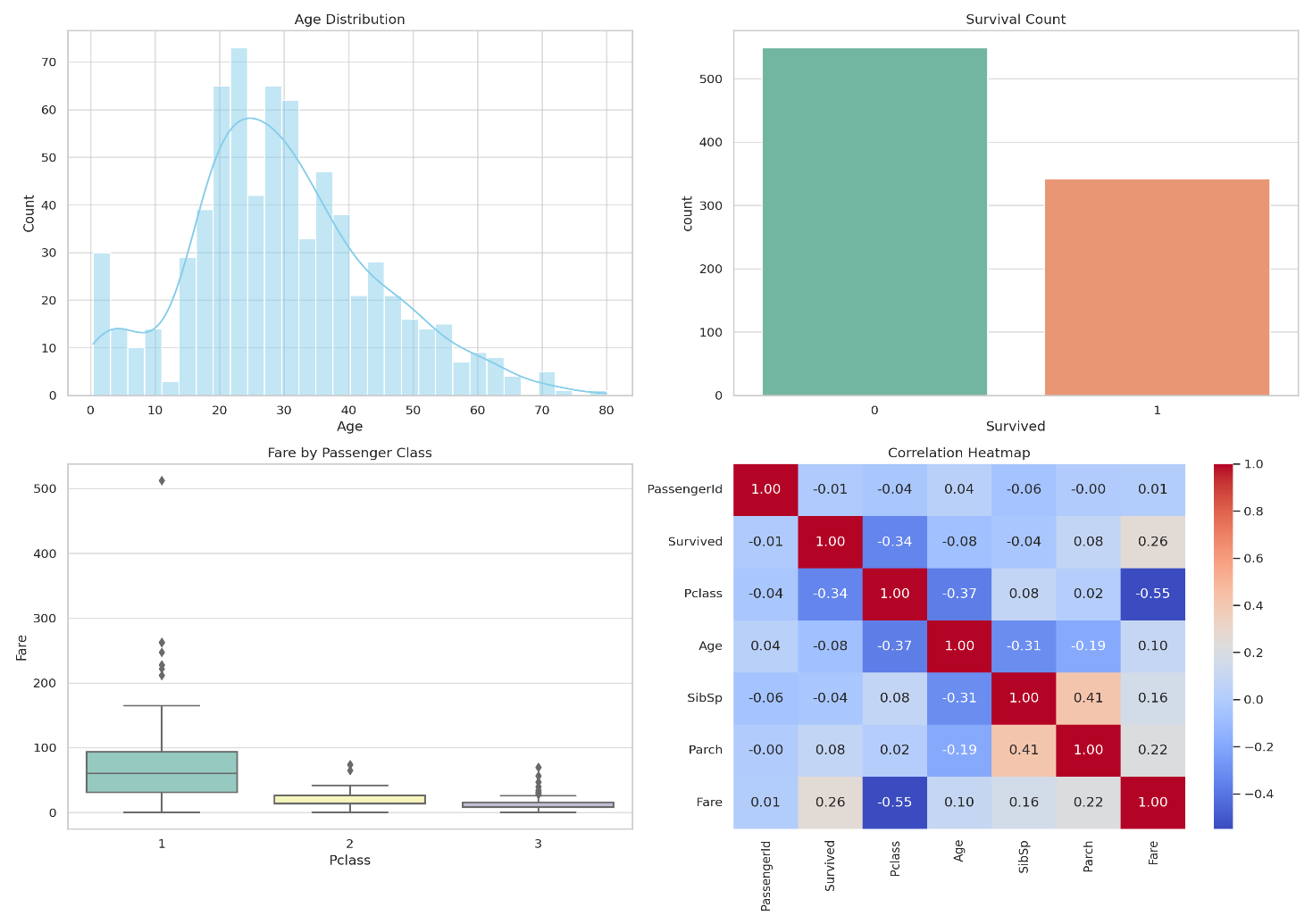
**Correlation Matrix**

* Positive correlation between Fare and Survived.
* Negative correlation between Pclass and Survived.

**Fare vs Class**

* 1st class passengers paid significantly more.

**📸 7. Visual Analysis**



Include this image showing:

* Histogram of Age
* Survival Count
* Fare by Class Boxplot
* Correlation Heatmap

**🧠 8. Insights & Recommendations**

* **Women and children first**: Survival highly favored females.
* **Wealth matters**: 1st class passengers had better chances of survival.
* **Fare influences survival**: Higher fares slightly correlated with survival.
* **Cabin info missing**: High rate of missing data; drop or process carefully.

**✅ 9. Conclusion**

This EDA exercise on the Titanic dataset shows the power of statistical and visual analysis. We identified trends in survival based on gender, class, and fare, and handled missing data appropriately. These findings can support machine learning modeling and further feature engineering tasks.