REPORT

Credora Internship – Data Science

WEEK 2 -Task 02

[Data Cleaning & Exploratory Data Analysis – Titanic Dataset]

Submitted by: AKANKSHA BHOSLE

DATE: 26- 5-2025

Introduction:

In this task, I explored the Titanic dataset to understand survival patterns. I cleaned the data by handling missing values and converting categories, then used Python libraries like Pandas, Matplotlib, and Seaborn to create visualizations. This helped me learn how to prepare real-world data and find insights through charts**.**

Data Discription:

analysis was performed using three datasets:

1. **train.csv**

-Main dataset used for cleaning, analysis, and visualizations. Includes survival labels.

2. **test.csv**

-Dataset with the same structure as train.csv, but without survival outcomes.

3. **test.csv**

-Dataset with the same structure as train.csv, but without survival outcomes

---

Key Columns in the Dataset:

- PassengerId – Unique ID for each passenger
- Pclass – Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- Name, Sex, Age – Personal details
- SibSp, Parch – Number of siblings/spouses or parents/children aboard
- Ticket, Fare, Cabin – Ticket information
- Embarked – Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Tools & Libraries Uses:

Tools:

- Google Colab – For writing and running Python code in a cloud-based notebook
- GitHub – To store and share the project files and code
- Kaggle – Source of the Titanic dataset

Libraries:

- Pandas – For loading, cleaning, and exploring the data
- NumPy – For numerical operations
- Matplotlib – For creating basic plots and charts
- Seaborn – For more advanced and stylish data visualizations
- Survived – 0 = No, 1 = Yes (whether the passenger survived)

Data Cleaning & Preprocessing:

To prepare the Titanic dataset for analysis, I applied the following data cleaning and preprocessing steps:

Cleaning Steps:

- **Handled Missing Values:**
  - Age: Filled with median age
  - Embarked: Filled with most frequent value (mode)
  - Cabin: Dropped due to too many missing values
- **Dropped Irrelevant Columns**:
  - Removed Ticket, Name, and Cabin columns as they didn't contribute directly to analysis

Preprocessing Steps:

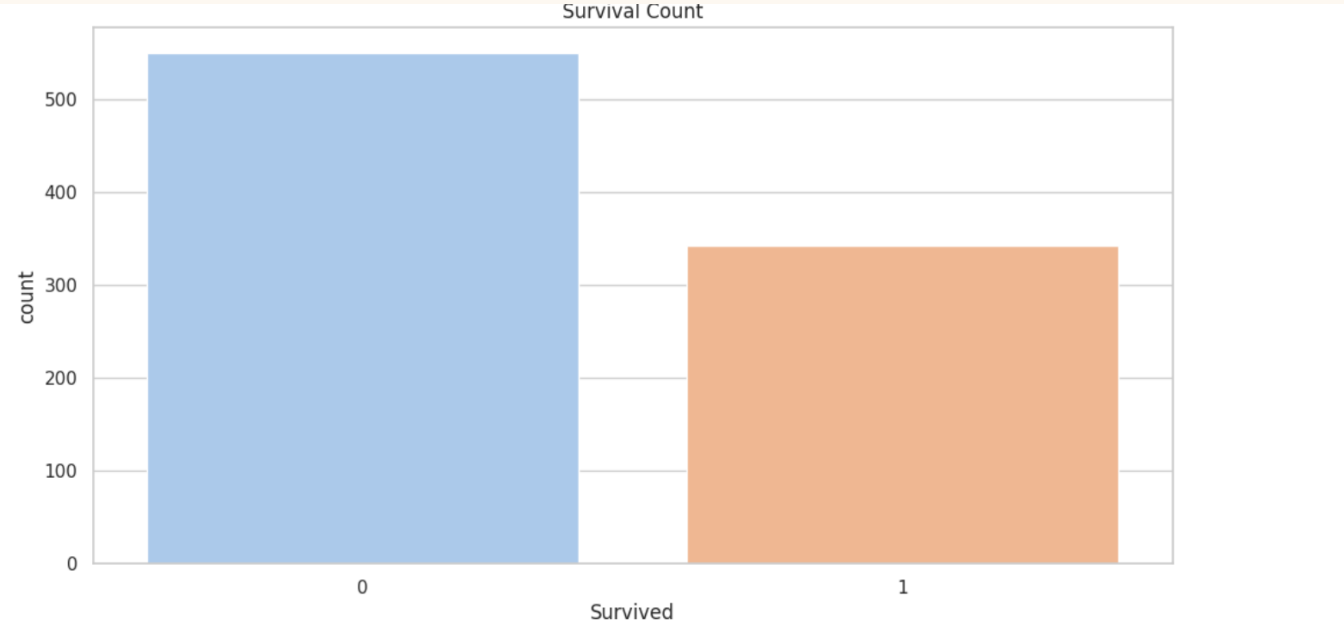
- **Converted Categorical to Numerical:**
  - Sex: Mapped to 0 (male) and 1 (female)
- **One-Hot Encoded:**
  - Embarked: Used get_dummies() to convert into numeric columns (Embarked_Q, Embarked_S)

- **Checked for Duplicates and Consistency:**
  - Verified there were no duplicate rows or inconsistent data types
  - Sex: Mapped to 0 (male) and 1 (female)
- **One-Hot Encoded:**
  - Embarked: Used get_dummies() to convert into numeric columns (Embarked_Q, Embarked_S)
- **Checked for Duplicates and Consistency:**
  - Verified there were no duplicate rows or inconsistent data types

---

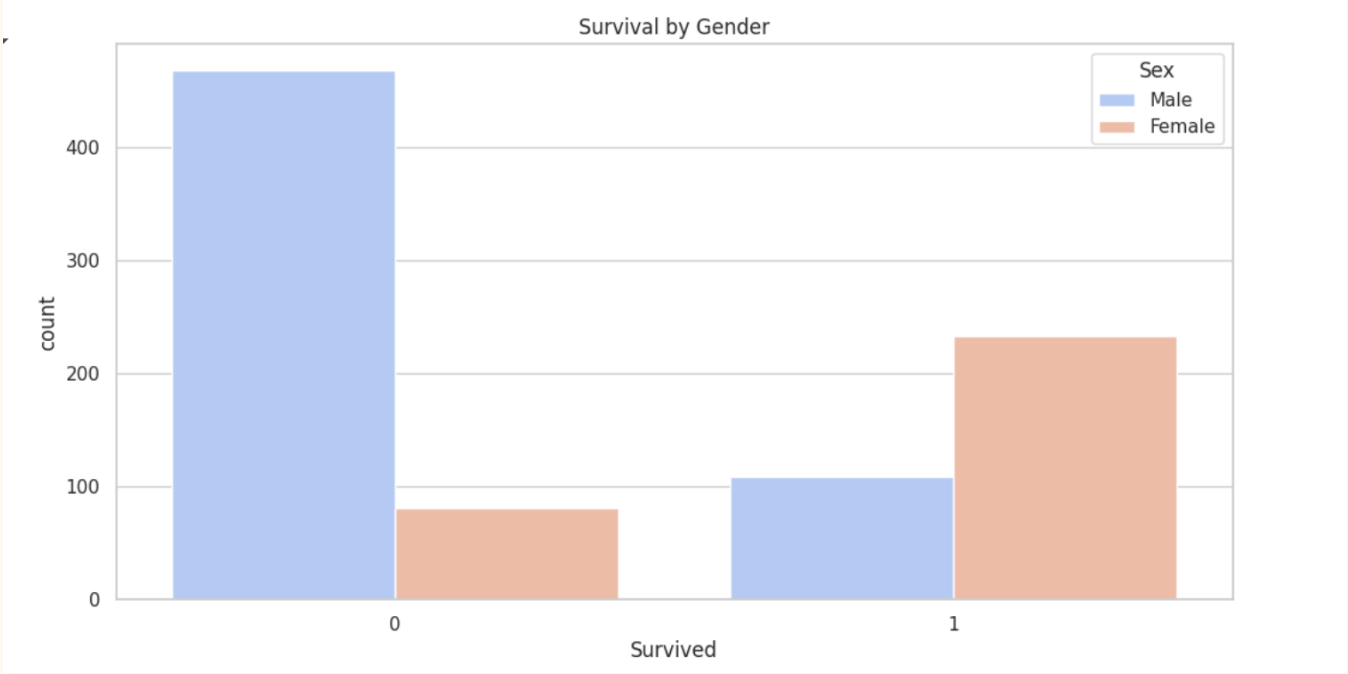
Visualizations & Insight:

1.Survival Count:



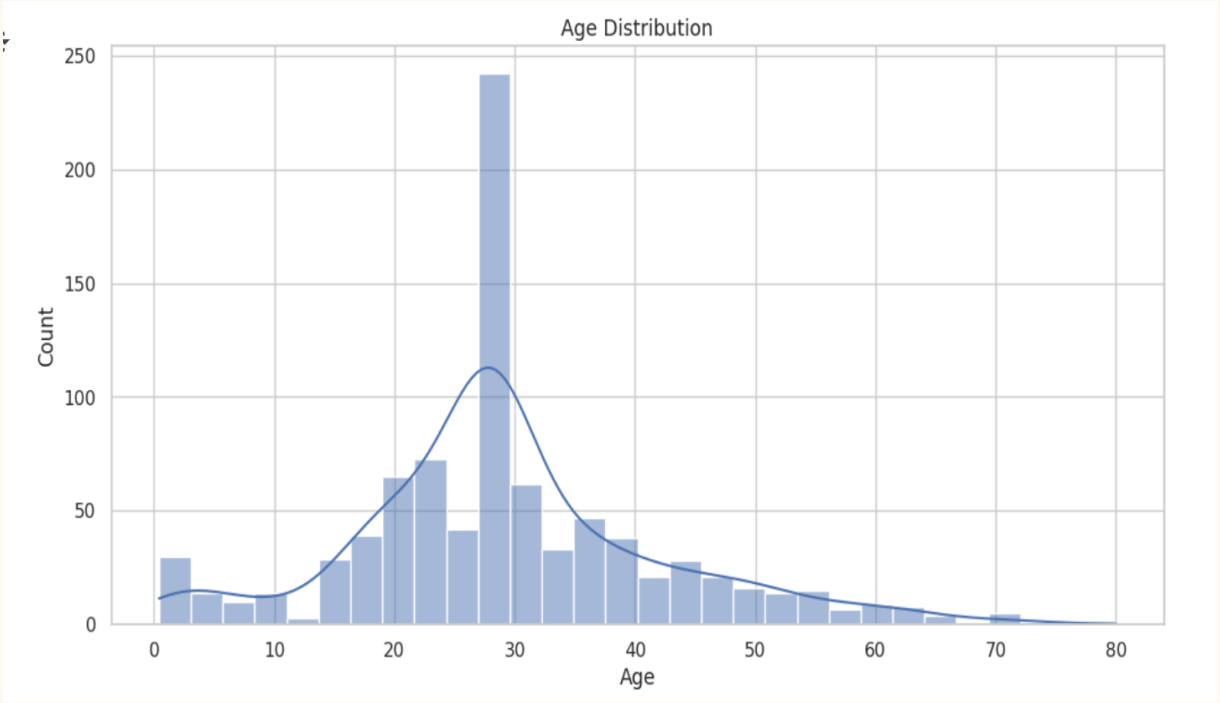
This chart shows the number of passengers who survived (1) and did not survive (0). It is clear that more passengers did not survive the Titanic disaster.

2.Survival by Gender:



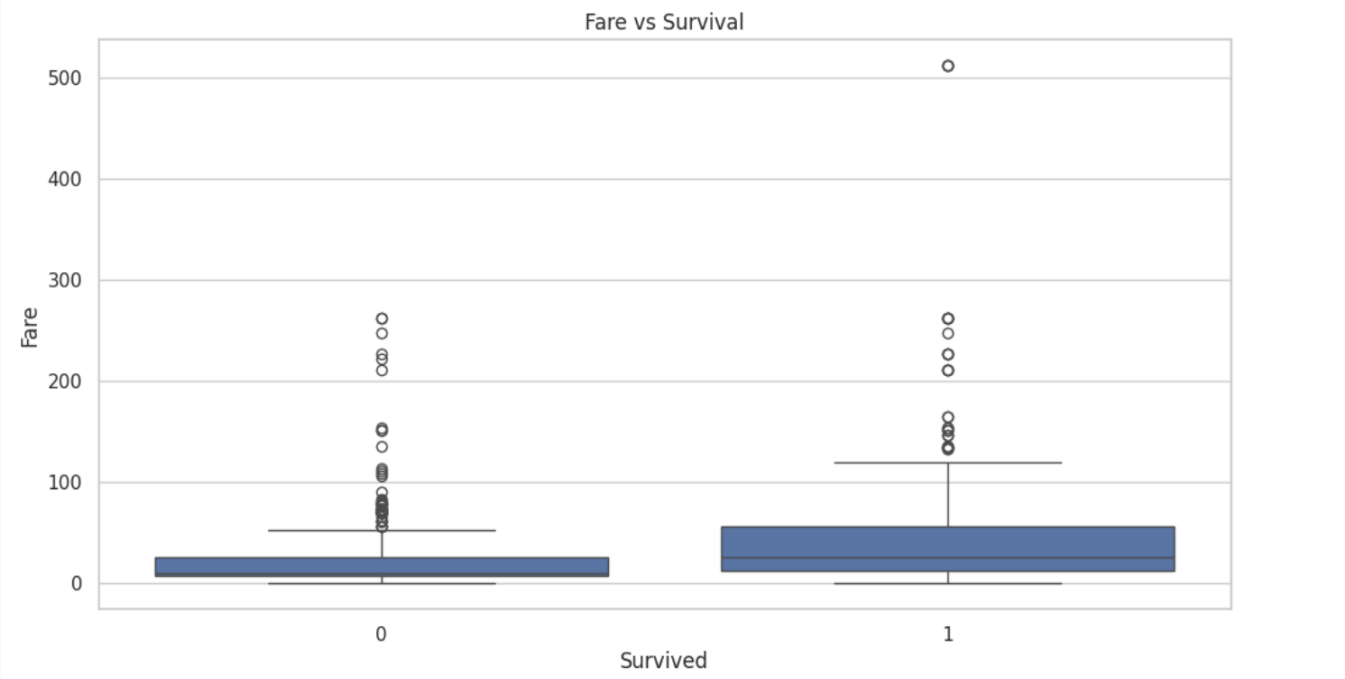
This bar chart shows the survival count by gender. A higher percentage of females survived compared to males.

3.Age Distribution:



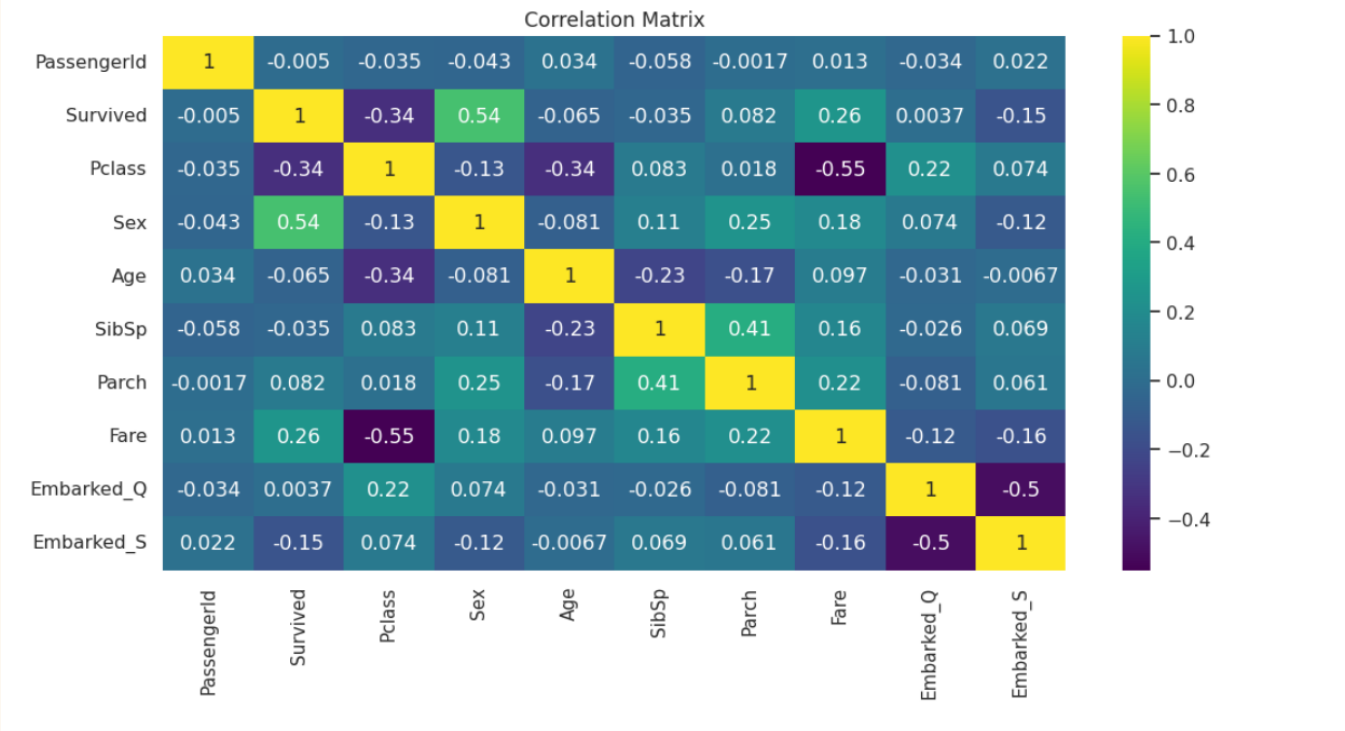
The age histogram displays the distribution of passenger ages. Most passengers were between 20 and 40 years old.

4.Fare vs Survival:



The boxplot shows the distribution of fares paid by passengers, split by survival status. Those who paid higher fares had better chances of survival.

5.Correlation Matrix:



The heatmap displays correlation between features. Survival has strong correlation with Sex, Pclass, and Fare.

Insights:

Survival Count

- The majority of passengers did not survive.
- Survival rate was less than 40%.

Survival by Gender

- Females had a much higher survival rate than males.
- Most survivors were women.

Age Distribution

- Most passengers were aged between 20 and 40 years.
- Children under 10 also had a noticeable survival count.

Fare vs Survival (Boxplot)

- Survivors generally paid higher fares.
- Many non-survivors were in the lower fare range.

Correlation Heatmap

- Strong positive correlation between Fare and Survived.
- Sex and Pclass were also highly correlated with survival chance.

Challenges & Solutions:

Challenge	Solution
Missing values in Age and Embarked	Filled Age with the median value and Embarked with the mode (most frequent value).
Too many null values in Cabin	Dropped the Cabin column completely as it had too many missing entries.
Categorical data like Sex and Embarked	Converted Sex into binary (0 and 1) and used one-hot encoding for Embarked.
Unnecessary columns (Name, Ticket)	Removed these columns as they didn't add value to the analysis.
Preparing data for visualization	Ensured all relevant columns were numeric and cleaned before plotting.

Project Links

- 📁 GitHub Repository: [CREDORA-TASK]
- 📓 Google Colab Notebook: [CREDORA-COLAB-TASK2]

Contact Details

- 👩‍💻 Name: Akanksha Bhosle
- 💼 LinkedIn: [linkdin-akanksha]
- 📧 Email: *akanshabhosle31@gmail.com*