# -:Task 02 Report:-

**Data Cleaning and Exploratory Data Analysis (EDA)**  
**Data Science Project Report**

**Submitted By:**  
Naima Shoukat  
**Submitted To:**  
Prodigy InfoTech  
**Course:**  
Data Science  
**Date:**  
08 August, 2025

**📑 Table of Contents:**

1. Introduction
2. Problem Statement
3. Tools and Technologies
4. Methodology
   * 4.1 Data Collection
   * 4.2 Data Exploration
   * 4.3 Data Cleaning
   * 4.4 Data Visualization
5. Results and Interpretation
6. Conclusion

**1️⃣ Introduction:**

Exploratory Data Analysis (EDA) is a critical step in any data science process. It involves summarizing the main characteristics of a dataset, often using visual methods, and preparing it for further modeling or prediction. In this task, we perform comprehensive data cleaning and EDA on the Titanic dataset to understand trends, correlations, and distributions in passenger data.

**2️⃣ Problem Statement:**

The objective of this task is to clean and explore the Titanic dataset using Python. The focus is on handling missing data, understanding feature distributions, and revealing survival trends using visualizations. These insights will be useful for future modeling and prediction tasks related to passenger survival.

**3️⃣ Tools and Technologies:**

* **Programming Language:** Python
* **Libraries:** pandas, numpy, seaborn, matplotlib
* **Platform:** Jupyter Notebook
* **Visualization Techniques:**
  + Countplot
  + Violin Plot
  + Histogram
  + Correlation Heatmap
* **Data Source:** Titanic dataset (public dataset via GitHub or Kaggle)

**4️⃣ Methodology:**

**🔹 4.1 Data Collection:**

* **Dataset Used:** Titanic Dataset
* **Source:** Local file (titanic.csv), downloaded from public GitHub repository
* **Key Features:**
  + Survived: Target variable (0 = No, 1 = Yes)
  + Sex, Age, Pclass, Embarked, Fare, SibSp, Parch

**🔹 4.2 Data Exploration:**

* Displayed first few rows using .head()
* Used .info() to examine data types and missing values
* Generated summary statistics with .describe()
* Checked missing values using .isnull().sum()

**🔹 4.3 Data Cleaning:**

* Filled missing values in Age using **median**
* Filled missing values in Embarked using **mode**
* Dropped the Cabin column due to excessive missing values
* Verified that the cleaned dataset had **no missing values** left

**🔹 4.4 Data Visualization:**

| **Visualization** | **Variable(s)** | **Purpose** |
| --- | --- | --- |
| Countplot | Survived | To view overall survival counts |
| Countplot | Sex vs Survived | To analyze survival by gender |
| Countplot | Pclass vs Survived | To analyze survival by class |
| Histogram | Age, Fare | To observe passenger age/fare distribution |
| Violin Plot | Age vs Sex & Survival | To see age impact by gender |
| Countplot | Embarked vs Survived | To compare port and survival |
| Heatmap | Numeric features | To understand correlations |

**5️⃣ Results and Interpretation:**

| **Visualization Type** | **Insight** |
| --- | --- |
| Countplot (Survived) | Fewer passengers survived than those who didn't. |
| Gender vs Survival | Females had significantly higher survival rates. |
| Pclass vs Survival | 1st Class passengers had the highest survival rate. |
| Age Histogram | Most passengers were between 20 and 40 years old. |
| Fare Histogram | Many paid lower fares, but high fares correlated with survival. |
| Violin Plot (Age vs Sex) | Younger females had the highest survival rates. |
| Embarked vs Survival | Most passengers boarded from 'S' (Southampton). |
| Heatmap | Strong correlation between Fare and Survival; Pclass also linked. |

**6️⃣ Conclusion:**

**✅ Key Observations:**

* Majority of survivors were **female** and in **1st class**.
* **Age** and **Fare** play a role in determining survival.
* The dataset required careful cleaning due to missing values.
* Visualizations revealed patterns that support future ML modeling.

**⚠️ Limitations:**

* Dataset is limited to Titanic passengers; not generalizable.
* Cabin data had too many nulls and was removed.
* Further insights possible through feature engineering (e.g., title from Name).

**🔄 Process Flowchart:**

┌────────────────────────────┐

│ Load Titanic Dataset │

└────────────┬───────────────┘

↓

┌────────────────────────────┐

│ Explore Dataset (head/info)│

└────────────┬───────────────┘

↓

┌────────────────────────────┐

│ Clean Data (fill/drop NAs) │

└────────────┬───────────────┘

↓

┌────────────────────────────┐

│ Perform EDA (Plots) │

└────────────┬───────────────┘

↓

┌────────────────────────────┐

│ Analyze & Interpret Trends │

└────────────────────────────┘