**Exploratory Data Analysis & Visualization of Titanic Dataset**

**Name:** KHANSA AHMAD QURESHI

**Date:** 16th April 2025

**Dataset Used:** Titanic dataset from Seaborn library

# Objective

The goal of this task was to perform exploratory data analysis (EDA) and data visualization on the Titanic dataset to gain insights into passenger demographics, survival trends, missing data, and data distribution.

# Steps Followed

## Importing Libraries & Loading Data

* Used pandas, numpy, seaborn, and matplotlib for data analysis and visualization.
* Loaded the Titanic dataset using sns.load\_dataset('titanic').

## Initial Data Exploration

* Checked basic info using .info(), .describe(), .shape, and .head() methods.
* Identified missing values and duplicate entries.
* Found missing values in:
* age (177 missing)
* embarked (2 missing)
* embark\_town (2 missing)
* deck (688 missing)

# Problems Faced & Solutions

## Missing Values

**Problem:**  
Missing values were found in critical columns such as age, embarked, and embark\_town.

**Solution:**

* Filled age with median value.
* Filled embarked and embark\_town with mode.
* Dropped the deck column due to excessive missing data.
* Reassigned columns directly to avoid FutureWarning from chained assignment.
  1. **Note:** Initially used inplace=True with chained assignment, which caused **FutureWarning**. Solved it by reassigning the column directly without inplace.

## Duplicate Records

**Problem:**  
Found duplicate rows using df.duplicated().sum().

**Solution:**Removed duplicates using df.drop\_duplicates().

## Outliers in Age

**Problem:**  
Detected outliers in the age column using a boxplot.

**Solution:**  
Applied IQR (Interquartile Range) method to filter out outliers from age.

# Visualizations Created

* **Bar Charts:** for Gender Distribution, Passenger Class Distribution, Embarkation Port
* **Histograms:** for Age Distribution, Fare Distribution
* **Boxplot:** to detect and visualize outliers in age
* **Correlation Heatmap:** to understand relationships between numerical features

# Insights

* Most passengers were male and traveled in 3rd class.
* Most passengers embarked from Southampton.
* Younger passengers and females had a higher survival rate.
* Strong correlation found between pclass, fare, and survival likelihood.
* Cleaning missing values and removing outliers improved data quality.

# Conclusion

This project involved a full EDA cycle — from loading the data to cleaning, visualization, and extracting key insights. I successfully handled missing values, removed duplicates and outliers, and visualized important trends that could help in further predictive modeling or business insights.