Task Report 2: Data Cleaning and Exploratory Data Analysis (EDA)

# Introduction

The objective of this task **(Task 2)** is to perform **data cleaning and exploratory data analysis (EDA**) on the Titanic dataset from **Kaggle**. This process involves handling missing values, converting data types, and exploring relationships between variables to uncover patterns and trends. The goal is to gain a comprehensive understanding of the dataset to inform further analysis or modeling.

# Tools Used

I chose the **Kaggle** platform to generate the dataset and selected the **'Titanic dataset**.' I used the following tools: **Jupyter** and the **Python** programming language to complete the task.

# Steps to Accomplish the Task

* Import the necessary libraries for data analysis and visualization.
* Load the Titanic dataset into a **DataFrame**.
* Generate the header of the **DataFrame**.
* Generate descriptive statistics for the **DataFrame**.
* Display summary information about the DataFrame, including data types and non-null values.
* Display the data types of each column in the **DataFrame**.
* Examine each column to understand its values, distribution, and potential issues (e.g., missing values, outliers).
* Count the number of missing (null) values in each column of the **DataFrame**.
* Display all rows in the **DataFrame** where specific columns have missing (null) values.
* Create copies of columns as needed for manipulation and cleaning.
* Handle missing (null) values by appropriate imputation or removal.
* Remove or correct any incorrect or outlier values identified during the exploration.
* Convert categorical columns (e.g., 'Sex', 'Embarked') to numerical values for analysis.
* Plot visualizations (e.g., histograms, bar charts) for each column to understand their distributions.
* Explore relationships between variables using visualizations such as scatter plots and correlation matrices.