Report for Titanic\_EDA\_project

# Introduction

The Titanic dataset is a well-known dataset used for predictive modeling, containing information about passengers who were aboard the ill-fated Titanic voyage in 1912. The dataset includes both numerical and categorical data regarding the passengers, and it is commonly used to predict the survival rate of passengers based on various features.

The dataset consists of the following columns:

* PassengerId: A unique identifier for each passenger.
* Pclass: The passenger’s class (1st, 2nd, or 3rd).
* Name: The name of the passenger.
* Sex: The gender of the passenger (Male/Female).
* Age: The age of the passenger.
* SibSp: The number of siblings or spouses aboard the Titanic.
* Parch: The number of parents or children aboard the Titanic.
* Ticket: The ticket number.
* Fare: The fare paid by the passenger.
* Cabin: The cabin where the passenger stayed.
* Embarked: The port at which the passenger boarded (C = Cherbourg; Q = Queenstown; S = Southampton).
* Survived: Whether the passenger survived or not (0 = No, 1 = Yes).

# Objective:

The objective of this exploratory data analysis (EDA) is to analyze the Titanic dataset and understand the factors influencing the survival rates of passengers. Specifically, we aim to explore how different features such as age, gender, passenger class, fare, and embarked location contribute to a passenger's chances of survival. By visualizing and correlating various features, we aim to identify patterns and relationships that could explain survival rates.

# Data Cleaning

In the data cleaning phase, several tasks were performed to prepare the dataset for further analysis:

1. **Removal of Irrelevant Columns**:
   * **Parch** (Number of Parents/Children aboard) and **Cabin** were removed from the dataset. These columns were either redundant or contained too many missing values to be useful for the analysis.
   * **Embarked** was also removed because we decided not to analyze embarkation locations for this particular analysis, though it can be useful in other case.
2. **Handling Missing Data**:
   * **Age**: The **Age** column had missing values, so these were handled by filling the missing entries with the **mean** age. This was done because the mean is a reasonable estimate for this numerical feature.
   * **Fare**: The **Fare** column had one missing value, which was replaced with the **mean fare** of the dataset.
   * After handling the missing values, we verified that no null values remained in the dataset.

# Performing EDA

1. **Descriptive Statistics**:
   * First, we generated the basic statistical summary of the dataset using the .describe() function. This provided information such as:
     + **Mean, Median, and Mode** of numerical columns like **Age** and **Fare**.
     + **Count, Standard Deviation**, and other relevant metrics.
     + This helped us get an overall sense of the distribution and range of values in the dataset.
2. **Info and Summary of Data**:
   * The **.info()** method was used to check the data types, non-null counts, and memory usage of each column. This helped ensure the integrity of the dataset and allowed us to verify if all missing values were properly handled.
3. **Data Aggregation**:
   * We calculated the **mean**, **mode**, and **average** for numerical columns like **Age** and **Fare**, which provided additional insights into central tendencies in the dataset.
4. **Charts and Visualizations**: We created several visualizations to better understand the relationships between different variables and the survival rate.
   * **Age vs Survival**: We plotted survival rates across different **age groups** to understand if certain age groups had higher survival chances.
   * **Gender vs Survival**: A plot showing survival rates across genders (Male vs Female).
   * **Pclass vs Survival**: We visualized survival rates across different passenger classes.
5. **Correlation Matrix**:
   * We used the **.corr()** function to generate a correlation matrix for the numerical features. This matrix helped us identify relationships between various features in the dataset.
     + For example, **Pclass** had a negative correlation with **Fare**, meaning higher class passengers paid higher fares.
     + We ensured that only numerical columns were included in the correlation matrix, as columns like **Sex** or **Embarked** (which are categorical) cannot be directly correlated with numerical columns.