# DAI Assignment 1
# Exploratory Data Analysis (EDA)

Nandini
Enrollment No: 23112066

## 1 Introduction

The Titanic dataset is widely used for data analysis and machine learning. This report presents a structured exploratory data analysis (EDA), covering univariate, bivariate, and multivariate analyses. The primary objective is to understand relationships between different variables and factors influencing survival.

## 2 Dataset Overview

The dataset contains 891 passengers with the following key attributes:

- **PassengerId**: Unique identifier for each passenger.
- **Survived**: Survival status (0 = No, 1 = Yes).
- **Pclass**: Ticket class (1st, 2nd, 3rd).
- **Name**: Passenger's full name.
- **Sex**: Gender.
- **Age**: Age of the passenger.
- **SibSp**: Number of siblings/spouses aboard.
- **Parch**: Number of parents/children aboard.
- **Ticket**: Ticket number.
- **Fare**: Fare paid for the ticket.
- **Cabin**: Cabin number.
- **Embarked**: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

## 3 Data Cleaning and Preprocessing

- **Handling Missing Values**: Cabin was dropped due to 77.1% missing values. Age was filled with the median value, while missing values in Embarked were filled with the most frequent value (mode).
- **Feature Engineering**:
  - **Family Size** was created by combining 'SibSp' and 'Parch' to capture family connections.

  - **Group Size** was introduced to track passengers traveling on the same ticket, as some groups were not family.
  - **Fare per Person** was derived by dividing the total fare by the number of people traveling on the same ticket (`Fare_Per_Person = Fare / N_Per_Ticket`). This provides a better representation of an individual's economic status rather than the total fare.

- **Outlier Detection and Treatment**: Outliers in 'Fare' were handled using the IQR method due to a skewed distribution, while 'Age' outliers were treated using the Z-score method since it follows a near-normal distribution.

# 4 Univariate Analysis

- **Age**: Most passengers were between 20-40 years old, with a peak around 30. There were both infants and elderly passengers, showing a diverse age distribution.
- **Fare_Per_Person**: The distribution is heavily skewed to the right, with a few extremely high values.
- **SibSp and Parch**: Most passengers traveled alone, while a smaller proportion had family members aboard.
- **Sex**: About 65% of the passengers were male, and thus comparatively fewer females aboard.
- **Pclass**: Over 55% of passengers belonged to 3rd class, highlighting a socio-economic divide.
- **Embarked**: Southampton was the most common port ( 72% of passengers embarked from there).
- **Survival Rate**: The 38% survival rate on the Titanic tragically highlights the fatal combination of insufficient lifeboats, human error, and the overwhelming power of the disaster.

# 5 Bivariate Analysis

- **Survival by Gender**: Females had a much higher survival rate ( 75%) compared to males ( 18%), reflecting the "women and children first" policy.
- **Survival by Class**: 1st class passengers had a survival rate of 63%, whereas only 24% of 3rd class passengers survived, suggesting socio-economic status played a role.
- **Age and Survival**: Young children, especially those under 10, had higher survival rates, aligning with the priority given to women and children.
- **Fare_Per_Person and Survival**: Higher fare-paying passengers had better survival chances, as they were more likely to be in 1st class with greater access to lifeboats.
- **Embarked and Survival**: Passengers from Cherbourg had a higher survival rate compared to Southampton and Queenstown, possibly due to more first-class passengers boarding from Cherbourg.
- **Family Size and Survival**: Small families (2-4 members) had higher survival rates than individuals traveling alone. Large families had lower survival rates, likely due to evacuation challenges.
- **Fare_Per_Person vs. Pclass**: Passengers in 1st class paid significantly higher individual fares.

# 6 Multivariate Analysis

- **Gender, Class, and Survival**: Females in 1st class had the highest survival rate ( 95%), while males in 3rd class had the lowest ( 13%). This reinforces how class and gender together impacted survival.
- **Age, Fare_Per_Person, and Survival**: Younger passengers who paid higher fares were more likely to survive, suggesting they traveled in first-class sections.
- **Pclass, Fare_Per_Person, and Survival**: The survival rate increased significantly with fare price within each class, reinforcing the advantage of wealth in survival.
- **Group Size and Survival**: Passengers traveling in groups (same ticket) had a slightly higher survival rate than solo travelers, likely due to mutual assistance.

# 7 Conclusion

- Gender, class, and fare were the most significant factors in determining survival chances.
- The "women and children first" policy played a crucial role in survival rates.
- Young children and 1st class passengers had the highest survival rates.
- Traveling alone reduced survival chances compared to traveling in small groups.
- The dataset highlights disparities in survival based on socio-economic status and gender.