# Introduction to Data Science

## Midterm Project

### Project Description

Apply data preparation steps (which can be applied) for the given data set. Do descriptive statistics (Find the measure of central tendency and measure of spread) for continuous attributes. In this project, we are going to use a modified version of Titanic dataset which can be downloaded from the Teams. The original dataset can be found in the following link where the dataset description is available as well (you may need to log-in to download the dataset).

https://www.kaggle.com/datasets/ibrahimelsayed182/titanic-dataset?resource=download ..

### Project Deliverables

- Submit the implemented R program (R file or Text file) in the Teams. During VIVA session, you will bring this implemented program and we may ask you to execute the program.
- Submit the report in the Teams. See the instruction section below for the report details. **Please bring a printed copy of the submitted report during the VIVA session.**

### Instructions

- The submission deadline for all deliverables is **July 14, 2024** (you must submit the assignment before **11:59 PM**).
- At the beginning of the report, write a short note about the dataset. You will get the dataset details from the above link provided for the dataset.
- For each implemented code segment in the R program, provide the code and its output along with their description in the report. In the description part, only write the content (do not write unnecessary content) that is sufficient to understand the code and its output.
- **Comments are not allowed in the R program**.
- The following topics can be focused to think about the project. **Note that the project is not limited to these topics which are mentioned to get an idea about how to proceed with the project.**
    - If there are any missing values in the dataset, then we will apply all applicable methods from the available options to handle the missing values.
    - We can see missing values on a graph.
    - We can convert the imbalanced data set into the balanced data set.
    - We can find and remove duplicate values.
    - We can apply some filtering methods to filter the data.
    - We can show the mean, median, and mode on a graph.
    - We can convert continuous or numeric attributes into categorical attributes.
    - We can apply the normalization method only for one attribute.
    - If there exist any invalid data/outliers in the data set, then use the appropriate approach to handle those values.