# Introduction to Data Science

**Midterm Project**

## Project Overview

In this project, you will develop your own dataset and apply various data exploration and preprocessing techniques. This will help you gain practical experience in handling real-world data.

## Dataset Requirements

- Create a dataset with at least **5 columns**, including:
    - **4 feature columns** (at least one categorical, at least two numerical)
    - **1 target column** (categorical)
- The dataset should contain **at least 50 rows** for each label
- Save the dataset in **CSV format**

Your dataset should follow a structure similar to:

| Col1 | Col2 | Col3 | Col4 | Target Column |
|---|---|---|---|---|
| Data1 | Data2 | Data3 | Data4 | Category1 |
| Data5 | Data6 | Data7 | Data8 | Category2 |
| … | | | | |

## Tasks to Perform

Once your dataset is ready, complete the following steps:

### 1. Data Exploration

- Perform **univariate analysis** (summary statistics, histograms, boxplots, etc.)
- Perform **multivariate analysis** (scatter plots, correlation, pair plots, etc.)

### 2. Handling Missing Values

- Remove one or more values from some columns
- Demonstrate methods to handle missing values (e.g., imputation, deletion)

### 3. Data Type Conversion

- Identify necessary data type conversions (e.g., converting categorical variables into numerical form)
- Apply appropriate transformations

**4. Data Transformation**

- Apply scaling, normalization, or encoding as required

**5. Outlier Detection**

- Identify outliers in individual columns
- Check for outlier records (rows with multiple unusual values)
- Use appropriate techniques to handle them

**Submission Guidelines**

- Submit the following files:
    1. **Your dataset (CSV format)**
    2. **A well-documented R Notebook** containing:
        - Code for data exploration and preprocessing
        - Explanation of each step
        - Visualizations and observations
    3. **A project report (PDF format)** explaining the entire process, including:
        - Dataset creation
        - Data preprocessing steps
        - Key findings from analysis
        - Justification for transformations and outlier handling
- Submit your work as a **ZIP file** on **MS Teams**
- Filename format: **StudentID_MidtermProject.zip** (e.g., 22-47178-1_MidProject.zip)
- Deadline: 15th April. No extension of deadline will be allowed.

**Project Evaluation: Total 50 marks**

Implementation (Code): 20
Viva: 20
Report: 10

**Important Note:**
If your dataset is collected from an online source and it is found to be identical to another student's dataset, your project will be disqualified. Ensure that your dataset is unique and independently created.