# Data Science Project Guidelines

## Objective

The goal of this project is to apply the complete Data Science pipeline — from data collection to preprocessing — on a real-world dataset. You will explore, visualize, and prepare data for potential machine learning applications.

## 1. Dataset Requirements

You must select a dataset that meets the following conditions:

- Source: Kaggle or any open dataset repository (e.g., UCI, data.gov, GitHub, etc.).
- Minimum 500 rows (records).
- At least 2 categorical features (e.g., Gender, City, Education level).
- At least 3 numerical features (e.g., Age, Salary, Score, Price).
- The dataset should be relevant to a real-world problem (e.g., health, education, finance, marketing, sports, environment).

Submit the dataset link to me for approval before you start the analysis before the next lab class.

## 2. Project Tasks

### A. Data Understanding

1. Load the dataset into R.
2. Display the first few rows of the dataset.
3. Show shape (rows × columns).
4. Display data types of each column.
5. Generate basic descriptive statistics (mean, median, mode, std, min, max, count, etc.).
6. Identify categorical and numerical features.

### B. Data Exploration & Visualization

Perform exploratory data analysis (EDA) to understand patterns, relationships, and trends.

1. Univariate Analysis
    - Distribution plots (Histogram, Boxplot, Bar chart)
    - Frequency of categorical variables
2. Bivariate Analysis
    - Correlation matrix (heatmap)
    - Scatter plots for numeric pairs
    - Boxplots between categorical and numeric features
3. Identify patterns, skewness, and possible outliers.

### C. Data Preprocessing

Perform the following preprocessing steps:

1. Handling Missing Values
   - Detect missing values (using code).
   - Replace or remove them (mean, median, mode, or imputation methods).
2. Handling Outliers
   - Identify outliers (using boxplots or IQR method).
   - Decide whether to remove, cap, or transform them.
3. Data Conversion
   - Convert categorical variables using label encoding or one-hot encoding.
   - Ensure all features are in a numeric form for further processing.
4. Data Transformation
   - Normalize or standardize numeric features (e.g., Min-Max Scaling, Z-score normalization).
   - Apply log/sqrt transformation if necessary to fix skewness.
5. Feature Selection
   - Identify and retain the most important features.
   - Use correlation analysis, variance thresholding, or mutual information.

If your dataset does not have missing values or outliers, create them manually (either by modifying a few values or inserting NaN/outliers using code).

## 3. Expected Deliverables

Your final submission should include:

### A. Report (in PDF format)

- Title page (project title, dataset name, student name, ID)
- Dataset source (with link)
- Description of dataset and features
- Data exploration results (with visuals)
- Data preprocessing steps (with explanation)
- Summary of findings and observations

A report template has been uploaded to the shared Google Drive folder. Follow the template strictly to ensure proper formatting of text, images, and tables.

### B. Code File

- R: .R script
- Code must be clean, well-commented, and reproducible
- Data should be read from your Google Drive link or GitHub

**Submission Instructions**

- Upload the PDF report and code file to MS Teams.
- File naming format:
  C-G01-mid-report.pdf
  C-G01-mid-project.R

**Evaluation Criteria**

| Component | Marks |
|---|---|
| Dataset selection & justification | 5 |
| Data understanding & description | 10 |
| Data exploration (visuals & insights) | 10 |
| Preprocessing (missing, outliers, transformation, feature selection) | 15 |
| Report quality & code readability | 10 |
| **Total** | **50 marks** |

**Submission Deadline**

25th November 2025, 11:59 PM

Missing the deadline will result in a **10% deduction** of marks for each day. If you have midterm exams, complete the project before the midterm week. Individual viva sessions will be conducted after the midterm week.

**Note:**

Project progress will be evaluated during the lab sessions of the 4th and 5th weeks for all groups. Therefore, each group should be prepared with their updated code during those lab classes to report their implementation progress.