# Project 1: Data Acquisition, Cleaning, Preprocessing and Feature Engineering for Exploratory Analysis

## Project Overview

In this project, you will acquire data from various sources and perform essential data preparation tasks, including cleaning, exploratory data analysis (EDA), preprocessing, and feature engineering. The final deliverable will be a well-documented report summarizing the findings and methodology. The objective is to gain hands-on experience in handling real-world data challenges and deriving meaningful insights.

The project will help you understand the complete data pipeline, from raw data collection to generating actionable insights. It will provide opportunities to work with different data sources, tackle real-world inconsistencies, and implement industry-standard techniques for data preprocessing and feature engineering. By engaging in this project, you will improve your ability to communicate data-driven results effectively and gain practical knowledge applicable to various domains in data science and analytics.

You are encouraged to be creative in your data selection and analysis approaches. The final report should not only highlight technical work but also showcase the significance of the findings and how they can be utilized in decision-making or further predictive modeling tasks.

## Project Tasks

1. **Data Acquisition**
   You must obtain a dataset from one or more of the following sources (using multiple sources/files is highly encouraged):
   - Public data repositories (e.g., Kaggle; see Lecture 01 for the full list)
   - Web scraping (e.g., using rvest, BeuatifulSoap, Selenium, etc.)
   - APIs (e.g., Twitter API, OpenWeather API, Census API, etc.)
   - Other open-access sources

2. **Data Cleaning and Handling Inconsistencies**
   - Detect and rectify data inconsistencies (e.g., incorrect data types, inconsistent labels, etc.)
   - Ensure uniform formatting across variables
   - Address duplicate data entries
   - Identify and handle outliers

- Identify and handle missing values using appropriate techniques (imputation, removal, etc.)

3. **Exploratory Data Analysis (EDA)**
    - Generate summary statistics and descriptive insights
    - Visualize distributions, relationships, and trends
    - Identify patterns, correlations, and anomalies
    - Provide initial interpretations of data insights

4. **Data Preprocessing and Feature Engineering**
    - Normalize/standardize numerical variables as needed
    - Encode categorical variables appropriately
    - Create meaningful new features that could improve the dataset's usability
    - Justify preprocessing steps taken

5. **Report Preparation**
    - Document all steps taken in data acquisition, cleaning, EDA, preprocessing, and feature engineering
    - Present insights using well-structured visualizations and explanations

## Project Deliverables [Due on February 19th at 11:59PM]

1. **Dataset**: The raw and final clean, processed datasets (these files can be submitted in a GitHub repository).

2. **Code Files**: A well-commented Python and/or R script(s) containing the full workflow (these files should be submitted in a GitHub repository with proper documentation; include a README file with instructions on how to run the code).

3. **Report**: A structured report in Quarto/RMarkdown (converted into a PDF file) that includes
    - Introduction and dataset description
    - Data acquisition methodology
    - Cleaning and preprocessing steps
    - Exploratory Data Analysis (EDA)
    - Feature engineering process and justification
    - Summary of key findings
    - Challenges faced and future recommendations

- Link to your GitHub repository. Make sure it is public and that I can access all the files stored in it. If preferred, you can add me to the list of the repository members (ap4347)
- Each member's contribution to the project

You must submit the report on Courseworks (one member of a team can submit the report on behalf of the team).

## Evaluation Rubrics

1. **Data Acquisition, EDA, Pre-processing, Feature Engineering [0 – 32pt]:**
   - **Data Selection [0 – 8pt]:**
     - **Basic [2pt]:** Chooses a dataset without considering its complexity or data quality. Dataset has minimal/no data quality issues, making it relatively straightforward to work with.
     - **Intermediate [5pt]:** Selects a dataset with moderate complexity and some data quality issues.
     - **Advanced [8pt]:** Selects a dataset with high complexity and significant data quality challenges.

     NOTE: Data complexity encompasses factors such as data quality, data structures (e.g., structured or unstructured data), and data sources (e.g., public repositories, web scraping), among others.

   - **Data Exploration [0 – 8pt]:**
     - **Basic [2pt]:** Performs basic data exploration with limited visualizations and insights.
     - **Intermediate [5pt]:** Conducts thorough data exploration with appropriate visualizations, identifying key patterns and trends.
     - **Advanced [8pt]:** Conducts extensive data exploration, utilizing advanced visualizations and statistical techniques to uncover nuanced insights and relationships.
   - **Data Pre-processing [0 – 8pt]:**
     - **Basic [2pt]:** Implements basic data cleaning techniques but overlooks some issues.
     - **Intermediate [5pt]:** Applies standard data preprocessing techniques effectively, handling most issues and optimizing data for modeling.
     - **Advanced [8pt]:** Implements advanced data preprocessing techniques with meticulous attention to detail, effectively addressing all data quality issues and optimizing data for optimal model performance.

- **Feature Engineering [0 – 8pt]:**
    - **Basic [2pt]:** Implements basic/no feature engineering techniques without considering the full potential of feature manipulation.
    - **Intermediate [5pt]:** Applies standard feature engineering techniques effectively, including normalization and standardization, and handling missing values appropriately.
    - **Advanced [8pt]:** Implements advanced feature engineering techniques with a high level of creativity and sophistication. Demonstrates a deep understanding of feature engineering principles and applies them effectively to enhance the predictive power of the dataset.

2. **Written Report [0 – 15pt]:**
    - **Basic [5pt]:** Provides a simple written report that lacks detail and coherence; lacks clarity in conveying key points; fails to provide sufficient explanation or analysis of the project tasks, methodologies used, and findings.
    - **Intermediate [10pt]:** Prepares a clear and organized written report that effectively summarizes the project's approach, methodologies, and findings. Presents information in a logical manner, provides a moderate level of detail and analysis, offering some insight into project tasks and outcomes.
    - **Advanced [15pt]:** Produces a comprehensive and well-articulated written report that demonstrates a deep understanding of the project tasks and methodologies. Presents detailed explanations and insightful analysis of each project task, including challenges faced and solutions implemented. Communicates findings effectively, providing clear conclusions.