# DSC 200 – Data Wrangling

Lab 6: Data Cleaning – Investigation, Matching and Formatting

## Learning Objectives:

Upon completing this lab, you should be able to:
1. Identify common data quality issues using Python.
2. Apply data cleaning techniques using the `pandas` library to improve data quality.
3. Demonstrate teamwork skills by collaborating with your group members.

## Assignment Type: Group Assignment

## Instructions:

This lab is a group assignment. It consists of two main parts and requires a single submission file. Follow each part's instructions carefully and review the rubric to ensure your work meets all the criteria.

## Part 1: Data Merging

Goal: Combine multiple datasets into a single file for further analysis.

1. Retrieve Your Datasets: Use the same three datasets you selected in Lab 5, Part 2.
2. Write a Python Function: Create a function that:
   - Inputs: Takes paths to your three datasets.
   - Processes: Merges these datasets into a single data file.
   - Outputs: Saves the merged file for later use in analysis.

3. File Linking or Inclusion: Either:
   - Provide a link within your code to the original dataset locations, or
   - Include the dataset files in your submission (e.g., in a zipped file).

Code Requirements:
- File Name: Ensure your code file is named lab6_[group_number].py.
- Documentation: Include comments explaining each step of your merging process.

*Points: 20*

## Part 2: Data Cleaning

Goal: Identify and fix issues in a provided dataset.

1. Dataset: Download the dataset linked in the assignment on Canvas (Rotten Tomatoes Movies and Critic Reviews). Make sure to unzip the file which contains two files.
   - Link: https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset/data

2. Write a Data Cleaning Function: Create a Python function that:
  - Inputs: Accepts the path to the unzipped data file.
  - Cleans: Corrects issues such as:
    - Duplicate rows
    - Inconsistent or missing data in categorical fields
    - Merge the two files meaningfully, justifying your reasons for the merge
  - Outputs: Saves a cleaned version of the dataset in the same directory as your script.

3. Console Output: Ensure your function:
  - Prints the number of features and observations before and after cleaning.

Code Requirements:

- File Naming Convention: Name the cleaned file in the format lab6_[group_number]_cleaned.csv.
- Documentation: Document each cleaning step in your code, describing what each step does and why it’s necessary.

*Points: 30*

## Submission Requirements

*Points: 10*

1. Submit a Single Python Script: The script should include both functions for Parts 1 and 2.
  - Menu Interface: Create a simple menu in the script that allows users to:
    - Select either Part 1 (Data Merging) or Part 2 (Data Cleaning).
    - Run the selected function.

2. Data Files: If including the original datasets for Part 1, submit them in a zipped folder alongside your Python script.

3. Submission Platform: Upload the zipped folder (if including files) or the Python script directly to Canvas in the linked assignment.