DSC 200 – Data Wrangling

Lab 5: PDF Data Extraction, and Data Acquisition and Storage

Objectives:

- Use python libraries to extract data from PDF files
- Identify the correct source of data for your research questions

Instructions:

This lab is in two parts: data extraction from PDFs and identification of data sources.

Part 1: Data Extraction from PDFs (25 marks)

Given the PDF file linked in the Canvas assignment, you are required to write a program which extracts the data in the PDF file and creates a CSV file containing the extracted data. The file should be named using the format **group_[group_number]_Lab5.csv**. The extracted data (output file) must be properly formatted using the same format as specified in the instructions for Lab 4. The attached file contains the exact same data as in the Excel file you worked with in Lab4.

For par 1, you are required to submit a python script named using the format: **group_[your_group_number]_Lab5.py**.

Part 2: Data Acquisition and Storage (25 marks)

In this part, you will be applying the key concepts discussed in class regarding how to acquire and validate a data set. Your task is to pose a research question that can be answered using a dataset. The research question could be about education, finance, marketing, etc. Identify three possible data sources that will be helpful in answering the question. For each data source, determine the organization from which the data is sourced (or a contact person), information about how the dataset was collected, the frequency of collection and whom you contacted for the data. Also include information about how you will store this dataset and a justification for your choice.

For this part, you are required to submit a 2-page single-spaced Word document that summarizes the information requested. Use the following filename format for this document: **group_[group_number]_Lab5.docx**.