DSC 200 - Data Wrangling

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

Purpose:

This lab aims to develop your ability to extract data from PDF files using Python and enhance your understanding of data acquisition and storage principles. These skills are critical in real-world applications, where data often needs to be gathered from multiple formats and validated for use in research or decision-making.

By the end of this lab, you will:
- Use Python libraries to extract and format data from PDF files.
- Identify relevant data sources for research questions and assess the validity and reliability of those sources.
- Learn how to store and organize datasets effectively for future use.

Objectives:

- Extract data from PDF files using Python.
- Identify the correct sources of data for your research questions.
- Store the data in a structured format for easy access and analysis.

Instructions:

This lab, which is a group assignment, consists of two parts:
Part 1 focuses on data extraction from PDFs, while Part 2 requires identifying potential data sources for a specific research question. Make sure to join one of the groups on Canvas and work with your group members.

Part 1: Data Extraction from PDFs (25 marks)

Task:
You are provided with a PDF file linked in the Canvas assignment. Your task is to:
1. Write a Python script that extracts data from the PDF file and stores it in a CSV file.
2. Name the CSV file using the format: group_[group_number]_Lab5.csv.
3. Ensure the extracted data in the CSV file follows the same structure and formatting guidelines as outlined in Lab 4.
   - The attached PDF contains the same data as the Excel file from Lab 4.

Submission for Part 1:
Submit your Python script with the following naming format:
group_[your_group_number]_Lab5.py

Part 2: Data Acquisition and Storage (25 marks)

Task:
In this part of the lab, you will apply key concepts about data acquisition and validation, covered in class.

1. Formulate a research question:
  - Choose a research question that can be answered using a dataset. The research question can focus on areas such as education, finance, marketing, etc.

2. Identify three possible data sources to answer your research question:
  - For each source, provide the following details:
    - The organization or contact person from which the data is sourced.
    - Information on how the dataset was collected.
    - The frequency of data collection (e.g., yearly, quarterly, monthly).
    - Whom you contacted for the data.

3. Justification for data storage:
  - Explain how you will store the acquired dataset and justify your choice of storage method (e.g., cloud storage, local database). Consider factors like accessibility, security, and ease of use.

Submission for Part 2:
Write a 2-page, single-spaced Word document summarizing the above information. Submit the document using the following filename format:
group_[group_number]_Lab5.docx

Criteria for Success:

For Part 1 (PDF Data Extraction):
- The Python script correctly extracts data from the PDF and outputs it to a CSV file.
- The CSV file follows the format and structure provided in Lab 4.
- The script is named and submitted using the required format.

For Part 2 (Data Acquisition and Storage):
- The research question is clearly stated and well-justified.
- Three valid data sources are identified, with complete and accurate details about the source, collection method, and contact.

- The chosen data storage method is well-explained and justified, with clear consideration of practical factors like security and accessibility.

Relevance to Course/Real-World Application:

This lab prepares you for handling data in professional settings, where extracting, validating, and storing data are critical skills. The ability to extract structured data from unstructured sources (like PDFs) and identifying reliable data sources for research is applicable across various industries, such as finance, education, marketing, and public policy.