04-638 A: Programming for Data Analytics

Analytics Assignment I: Exploratory Data Analysis and Reporting [Group Assignment]

Release Date: 8th November 2023 **Due Date: 16th November 2023, 11.59pm CAT**

Points Total: 100 pts

To submit: Jupyter Notebook file, and PDF report (Max: 2 pages)

Submission Mechanism: Submit on Canvas.

TASK CONTEXT:

You will work in groups (of 3-4 students each). The groups have been formed for you. You will obtain the CO2 emissions estimates dataset from the UN website (https://data.un.org/) under the environment category. You are required to perform exploratory data analysis (EDA) on the data and produce a report summarizing the insights.

All tasks, apart from the accompanying PDF report, should be accomplished within a Jupyter Notebook file. The task only entails exploratory data analysis (a.k.a. data exploration) and reporting. No machine learning processing should be attempted.

Exploratory data analysis allows the analyst to detect anomalies, identify patterns, understand the data, and identify preliminary insights from the data. Methods can be quantitative, graphical, or a combination of both.

TASK OBJECTIVES:

a) Understand the dataset.
b) Understand the attributes/variables, their types, relevance and significance, and their relationships.
c) Prepare and clean the dataset in readiness for analytics.
d) Communicate insights.

Toolkit: Basic set: numpy, pandas, and matplotlib; Advanced set: any other tools chosen by team, e.g., Seaborn, JavaScript, etc.

TASK DESCRIPTION:

1) From https://data.un.org/, select the CO2 emissions estimates dataset and download and save the relevant CSV file.

	Data Set Categories:	**Specific Datasets**

	Environment	CO2 emissions estimates

2) Perform a cursory examination of the CSV file to get an idea of the kind of data you are dealing with – the fields and the actual data. Some of the information obtained during this stage will be useful in subsequent tasks.

3) Based on the cursory examination, identify a set of ten (10) to twelve (12) preliminary questions that you would like to answer with the dataset. Add these questions to your Jupyter Notebook. You may also add more questions as your understanding of the data improves. Later, your EDA should seek to answer at between five (5) to eight (8) of these questions.

4) Load the data into appropriate pandas data frames. You are free to decide on which and how many data frames to use, in line with the cursory examination in 2) above.

5) ***Perform exploratory data analysis (EDA)*** on the data (already in data frames) and produce insights centered around the following issues. Each issue should be described in the Jupyter Notebook using a Markdown/Heading cell. You should provide *relevant graphics (where appropriate) and effective narrative descriptions for each significant observation* as you perform the tasks in this section. See *Useful References* below for *effective data storytelling*.

   a) **[8 Pts]** Understanding the data- attributes/fields, data types, data size, significant fields, etc.

   b) **[10 Pts]** Understanding the basic statistics (if necessary) and the distribution of numerical data (if any).

   c) **[15 Pts]** Using relevant visualizations to help understand and communicate insights from the data.

   d) **[7 Pts]** Check for missing values and outliers (and, where necessary, handle missing values).

   e) **[15 Pts]** Determine trends, patterns, and relationships among the data e.g., increase in gross value added in Rwanda over the last 3 years for *services and industry* and a corresponding decline in gross value added attributed to *agriculture*; reduced agricultural production in West Africa accompanied by a marginal increase in food production; poor access to safe water and sanitation in East Africa compared to Central Africa; increased balance of trade in North Africa compared to Southern Africa; etc.

   f) **[20 Pts]** Answer the questions in 3) above.

6) **[25 Pts]** Produce a summary writeup that contains the following information:

   - ***Metadata***: Course Name and Code; Instructor; Assignment Title (see top of this document); Report Title (Choose a title for your group's work); Group Members; Submission Date.
   - ***Summary of insights that includes the following sections [Max(Strictly): 1300 words, 11pt Times New Roman]***: Abstract, Background and Problem Description, Methods, Results and Discussion, Conclusion, References (at least six(6)). The references do not count in the 1300 word limit.

REPORT GRADING:

Abstract (3 pts); Background and Problem Description (4 pts); Methods(4 pts); Results and Discussion(8 pts); Conclusion(3 pts); References(1.5 pts); formatting (1.5 pts).

USEFUL REFERENCES

1. Nussbaumer Knaflic, Cole. Storytelling with Data: A Data Visualization Guide for Business Professionals. 1st ed. Hoboken: John Wiley & Sons, Incorporated, 2015. Print. Link: https://cmu.primo.exlibrisgroup.com/permalink/01CMU_INST/6lpsnm/alma991019811317904436
2. Knaflic, Cole Nussbaumer, and Cathy Madden. *Storytelling with Data: Let's Practice!* 1st edition. Hoboken, New Jersey: Wiley, 2020. Print. Link: https://cmu.primo.exlibrisgroup.com/permalink/01CMU_INST/6lpsnm/alma991019695933004436
3. Vora, Sejal. *The Power of Data Storytelling*. First Edition. New Delhi: SAGE Publications India Pvt, Ltd, 2019. Web. Link: https://cmu.primo.exlibrisgroup.com/permalink/01CMU_INST/8lb6it/cdi_proquest_miscellaneous_2188046294