Practical Data Science (with Python)

COSC2670/COSC2738

Assignment 1

Semester 2, 2024

Type	Individual assignment
Due Date	Sunday, 1 September 2024 by 23:59
Weighting	25%
Marking Criteria	See the course Canvas shell >> Assignments
Mark/Feedback	Will be available in Canvas

1. Introduction

In this assignment, you will examine a data file in CSV and carry out the first steps of the data science process, including the cleaning and exploring of data. You will need to develop and implement appropriate steps, in Jupyter Notebook (available in Anaconda), to load a data file into memory, clean, process, and analyse it. In this assignment you may need to use Python packages/libraries such as Pandas, NumPy, Matplotlib, and/or Seaborn. This assignment is intended to give you practical experience with the typical first steps of the data science process.

2. General Requirements

Please read the following requirements and information carefully before you start.

- You must do the assignment in Jupyter Notebook.
- You must choose (by yourself) and use the *appropriate* Python functions (or methods) to complete the tasks.
- You must provide *sufficient (or detailed), precise* comments and/or annotations in your code, explaining what each part does.
- You must use the provided template A1code-Template.ipynb to organise your code.
- Ensure that your submission follows the file naming rules specified in this document. Replace YourStudentNumber with your Student Number (e.g. s1234567) wherever applicable.
- Responsible use of AI tools is allowed. You must appropriately acknowledge and reference the use of any AI tools and their outputs. Failure to reference the use of these tools can result in academic misconduct. For more instructions, see: https://rmit.libguides.com/referencing_AI_tools
- Plagiarism is *never* okay. Relevant tools (like Turnitin) might be used for plagiarism check for this assignment.
- There are rules in place to support you uphold the academic integrity of RMIT. See: https://www.rmit.edu.au/students/my-course/assessment-results/academic-integrity

3. Tasks

Task 1: Data Preparation

Download and have a look at the file A1data.csv, which is available in Canvas under the *Assignments >> Assignment 1* section of the course Canvas shell. This dataset contains data about the digital connectivity information of children in a school attendance age that have internet connection at home. Check the file Readme-A1data.txt for details about the dataset.

Your task is to prepare the provided data for analysis. You will start by loading the CSV data from the file (using appropriate Pandas functions) and then clean the data (using Python, *not* manually). You need to identify and address all the potential issues/errors in the data properly and write the cleaned data into a CSV file, named as YourStudentNumber-cleaned-A1data.csv. Note that there are at least **four types** of errors/issues of the data. You can presume the first 4 columns (ISO3, Countries and areas, Region, Sub-region) don't contain any error.

Task 2: Data Exploration

Use the cleaned data YourStudentNumber-cleaned-A1data.csv (which you obtained in the above Task) and complete the following subtasks.

2.1. Consider the overall/total percentage of children in a school attendance age that have internet connection at home (i.e., column **Total**). Create (and display) the side-by-side boxplot (as one graph/chart) having the data separated/grouped by **Region**. Compute the Median (of the total percentage) for each Region.

2.2. Compute the Mean (of the percentage of school-age children who have internet connection at home) for Wealth quintile (Poorest) and Wealth quintile (Richest), respectively. Display/list the top 10 countries with the highest percentages for Wealth quintile (Poorest) and Wealth quintile (Richest), respectively.

2.3. Consider the data about children that are from the **Lower middle income (LM)** group. Compare the percentages of different categories of Residence (Rural versus Urban), using at least three statistics measures (of your choice).

Task 3: Written Report

Write your report using the template (A1report-Template.docx) provided in Canvas. Your report must be at most 5 pages (including everything). For references, if any, use the APA 7th edition referencing style (see: https://www.lib.rmit.edu.au/easy-cite/).

The report should comprise the following sections:

- **Data Preparation**: Provide a concise explanation of how you addressed Task 1. Create a sub-section for each type of errors. Explain and justify the approach taken to address each kind of errors.
- **Data Exploration**: For each subtask in Task 2, create a sub-section with corresponding numbering; explain and justify how you explored the data as required, and summarise the results (i.e., findings).
- **Use of AI Tools**: Discuss how (generative) AI tools (e.g. Val and ChatGPT) can help with completing Tasks 1 and 2. Explain how you used AI tool(s) in this assignment; if you didn't use any AI tool, explain why.
- **References** (if needed): Include a list of cited sources, if any.

4. Submission

The following files must be submitted (uploaded one by one, in one submission) in Canvas:

- **Jupyter Notebook file** named as YourStudentNumber-A1code.ipynb: containing your Python code (and comments) for Task 1 and Task 2.

    - For the Jupyter Notebook code, please make sure to clean them and remove any unnecessary lines of code (cells). Comments are required. Follow these steps before submission:
        - Main menu → Kernel → Restart & Run All
        - Wait till you see the output displayed correctly. You should see all the data outputs printed and graphs displayed.
- **Your cleaned data file** named YourStudentNumber-cleaned-A1data.csv.
- **Your written report file in PDF**, named as YourStudentNumber-A1report.pdf.

Resubmission without penalty is allowed until the deadline. Please do NOT submit any unnecessary files.