RMIT University

School of Computing Technologies

Practical Data Science with Python

Assignment 1: Data Cleaning and Summarising

Due: 23:59 on the 9th of April, 2023

This assignment is worth 25% of your overall mark.

Introduction

In this assignment, you will examine a (set of) data file(s) 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 (in the Anaconda version specified in the canvas announcement), to load a data file into memory, clean, process, and analyse it. This assignment is intended to give you practical experience with the typical first steps of the data science process.

The "Practical Data Science" Canvas contains further announcements and a discussion board for this assignment. Please be sure to check these on a regular basis – it is your responsibility to stay informed with regards to any announcements or changes. Login through `https://rmit.instructure.com/`.

Where to Develop Your Code

You are encouraged to develop and test your code in the following environments: **Jupyter Notebook on Lab PCs**. (Or please use the Anaconda version as specified in the course announcement)

Jupyter Notebook on Lab PCs

On Lab Computer, you can find Jupyter Notebook via:

Start → All Programs → Jupyter Notebook - Anaconda

Then,

- Select New → Python 3
- The new created '*.ipynd' is created at the following location:
  - C:\Users\sXXXXXXX
  - where sXXXXXXX should be replaced with a string consisting of the letter "s" followed by your student number.

Plagiarism

RMIT University takes plagiarism very seriously. All assignments will be checked with plagiarism-detection software; any student found to have plagiarised will be subject to disciplinary action as described in the course guide. Plagiarism includes submitting code that is not your own or submitting text that is not your own. Allowing others to copy your work is also plagiarism. All plagiarism will be penalised; there are no exceptions and no excuses. For further information, please see the *Academic Integrity* information at `http://www1.rmit.edu.au/academicintegrity`.

Turnitin will be used for Plagiarism Review for this assignment in Canvas.

General Requirements

This section contains information about the general requirements that your assignment must meet. *Please read all requirements carefully before you start.*

- You *must* do the assignment in Jupiter Notebook that are available in Anaconda.
- Parts of this assignment will include a written report, this *must* be in *PDF* format.
- Please ensure that your submission follows the file naming rules specified in the tasks below. File names are case sensitive, i.e. if it is specified that the file name is `gryphon`, then that is exactly the file name you should submit; `Gryphon`, `GRYPHON`, `griffin`, and anything else but `gryphon` will be rejected.

Task 1: Data Preparation (10 marks)

Have a look at the file `Global database on school-age digital connectivity.zip`, which is available in Canvas under the `Assignments/Assignment 1` section of the course Canvas. This data set is provided by UNICEF[1], which contains data about the School-age digital connectivity of children in a school attendance age (approximately 3-17 years old depending on the country) that have internet connection at home.

Being a careful data scientist, you know that it is vital to carefully check any available data before starting to analyse it. 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 checking whether the loaded data is equivalent to the data in the source CSV file. Then, you need to clean the data by using the knowledge we taught in the lectures. You need to deal with all the potential issues/errors in the data appropriately and then write the cleaned data into a comma-separated values (csv) file(s) accordingly, e.g. (file can be named as '**cleaned_Primary.csv**').

[1] https://data.unicef.org/

Task 2: Data Exploration (8 marks)

Explore the provided data based on the following steps:

1. Explore the data about children in primary school: Choose **1** column with *nominal values*, **1** column with *ordinal Values*, and **1** column with *numerical values.* (Please try to explore the columns/attributes of potential importance to the analysis, not just a random choice). Then, create a visualization for each of them.

2. Explore the data about all children in a school attendance age (approximately 3-17 years old depending on the country), and analyze the top 10 countries and areas with the highest total percentage of school-age children (who have internet connection at home) in terms of their Income Group and Residence (Rural or Urban).

3. Please thoroughly compare the percentage of Primary children and Secondary Children that are from a Lower middle income (LM) group with Internet connection at home.

Note, each visualization (graph) should be complete and informative in itself, and should be clear for readers to read and obtain information.

Task 3: Report (7 marks)

Write your report and save it in a file called `report.pdf`, and it must be in PDF format, and must be **at most 6 (in single column format) pages (including figures and references) with a font size between 10 and 12 points**. Penalties will apply if the report does not satisfy the requirement. Moreover, the quality of the report will be considered, e.g. clarity, grammar mistakes, the flow of the presentation.

Remember to clearly cite any sources (including books, research papers, course notes, etc.) that you referred to while designing aspects of your programs.

- Create a heading called "Data Preparation" in your report.

  – Provide a brief explanation of how you addressed the task. For the steps of dealing with the potential issues/errors, please create a sub-section for each type of errors you dealt with (e.g. typos, extra whitespaces, sanity checks for impossible values, and missing values etc), and also explain and justify how you dealt with each kind of errors.

- Create a heading called "Data Exploration" in your report.

  – For each numbered step in Task 2 above, create a sub-section with corresponding numbering.

What to Submit, When, and How

The assignment is due at

23:59 on the 9th of April, 2023.

Assignments submitted after this time will be subject to standard late submission penalties.

You need to submit the following files:

- Notebook file containing your python commands for Task 1 and Task, 'assignment1.ipynb'. **Please use the provided solution template to organise your solutions**: *assignment1_TEMPLATE.ipynb*

# For the notebook files, please make sure to clean them and remove any unnecessary lines of code (cells). Follow these steps before submission:

  1. Main menu → Kernel → Restart & Run All
  2. Wait till you see the output displayed properly. You should see all the data printed and graphs displayed.

- Your `report.pdf` file: **at most 6 (in single column format) pages (including figures and references) with a font size between 10 and 12 points**. Penalties will apply if the report does not satisfy the requirement.

They must be submitted as ONE single zip file, named as your student number (for example, 1234567.zip if your student ID is s1234567). The zip file must be submitted in Canvas:

Assignments/Assignment 1.

Please do NOT submit other unnecessary files.