2486 Programming Module Assessment 2022

You have been provided with two data files in a comma-separated value (CSV) file format. One file (births.csv) contains individual data on 500 births in a hospital in London. The other file (score.csv) contains data information on a numerical score. The ID identifier is the same in both data files.

Your task is to write a program to:

1. Read in the data from the two files and capture it in a suitable data structure;
2. Examine the births dataset, and using its codebook, transform the categorical variables in the right format with labels; clean the dataset by excluding records with missing observations.
3. Combine the births and score datasets keeping both matching and unmatching records and examine the final dataset. Comment on the result of the operation.
4. Reorder the combined dataset by the variable score.
5. Examine the correlation between (truly) continuous variables, and the two-way distribution of categorical variables (the latter can be represented as a table). Comment on the results.
6. Create a new variable highscore that identifies a score higher than 150.
7. Create an aggregated version of the dataset that reports the average birthweight by highscore and sex.
8. Visualise the distribution of the variables with correlation higher than 0.5 from Step 5 in a suitable plot. Save the graph in pdf format using appropriately named files.
9. Export the final dataset in a format readable from Excel.

This must be repeated in both Python and R. The suggestion is to write a program in one of the two languages first, and then to translate the code in the other language.

The code is expected to be written in a code editor (script) and submitted as a set of .py, .r, Jupyter Notebook, or Rmarkdown files. Comments must be included in the scripts or markdown in notebooks to describe the steps and results. A README file must be included detailing how to run the code submitted as a plain text file. All the **files should be submitted in a single compressed (using zip) folder**. The number of lines of code is not expected to exceed 150 lines including lines of comment. The README file should contain sufficient information for the examiner to run the programs on their local machine, including the location of the data files. The zip folder should include the data files as well as the scripts and README file.

Ensure that you put your student number as a comment in each file and within the file name. Also, please complete an assessment cover sheet (available on Moodle) and include that in the compressed folder as well.

You should aim to demonstrate an excellent programming style in both Python and R: good variable naming, correct language syntax and layout. Code should be appropriately annotated, with a comprehensive set of comments detailing the processes being carried out. Potential errors should be handled efficiently, and the code should run to completion as directed in the accompanying text file. Output files and figures used to display primary results should be appropriately labelled and laid out.

Deadline for submission: 16:00 GMT Thursday 17th November 2022

Submission portal: A submission portal will be open to upload your files to on the main module Moodle page.