CLINICAL DATA MANAGEMENT
COURSEWORK 2 - GROUP WORK
2022

OVERVIEW AND INSTRUCTIONS

Description

This coursework constitutes your second piece of assessment for the Clinical Data Management module. Its purpose is to improve your Python skills and test your knowledge on the data sharing and anonymisation component of the module.

What you will need

- Jupyter Notebook and Python3 installed on your computer.
- Access to the internet.
- Your lecture notes and relevant practicals.
- Data provided in the Data folder of the coursework. Please see the Data section below for further details.

Important Note - Terms and Conditions

This coursework constitutes a **group** piece of assessment. For some of the questions you will need to do research online to discover how to perform certain tasks and this is expected. Imperial College London has strict policy on academic misconduct. Please see the course handbook, which can be found on the course blackboard page, to understand what constitutes the different forms of academic misconduct.

The problem

The CEO of an insurance company, iInsureU123, wants to understand if she can increase the policy fee for customers with a particular gene variant - the gene DRD4, which is known as the Wanderlust gene. Her hypothesis is that customers with this gene variant travel more - and consequently are at greater risk - than those that don't have this variant. So, she has asked her team of junior researchers to investigate if they can find any evidence in the data that she has collected and that could help her justify this increase. However, her junior researchers need some help so she has asked some of her former colleagues at Imperial College to help her with this research project.

On another note, she is also collaborating with the government on a larger insurance project and wants to share her data with the government. The government wants to understand if people with this Wanderlust gene have anything in common from an educational or geographical perspective. She has been told that the data will be made available online for anyone in the public domain to access it. She is not helping the government with their analysis. She is just sharing the data as part of her collaboration contract.

What is the problem? The CEO realises she can not just share her data in its raw format as that would be a breach of her customer's trust. However, her junior researchers have confessed to her and they do not know anything about data anonymisation or data sharing. She has come across the concept of `k-anonmity` and thinks that calculating this for her anonymised dataset would be relevant but she does not have the time to currently understand how it is calculated in practice.

What you need to do - as a group

Help the CEO of iInsureU123 with this part of the process. Specifically, you will need to anonymise the data and calculate the k-anonymity of the dataset, so that she can share it with the researchers at Imperial and also with her collaborators in government, in a secure and appropriate manner.

The data

The data for this coursework is provided in the Data folder of the Coursework Resources folder. It relates to the study above, done by iInsureU123. The customers included in the study are all Caucasian, live in the United Kingdom and have all been born overseas. Some of these customers have a particular variant of the gene DRD4, which is known as the Wanderlust gene.

When signing up to iInsureU123, customers agreed to undergo a set of tests and were asked to fill in a questionnaire with some personal details. The data for the study has been compiled and is contained in the file `customer_information.csv`.
The file contains the following variables on each of the customers included in the study:

- `given_name` - first name
- `surname` - last name
- `gender` - gender (F or M)
- `birthdate` - date of birth
- `country_of_birth` - country of birth
- `current_country` - location where the individual is currently living
- `phone_number` - personal phone number
- `postcode` - residential postcode
- `national_insurance_number` - national insurance number
- `bank_account_number` - bank account number
- `cc_status` - binary variable indicating if the customer has the identified gene variant (0 - No, 1 - Yes).
- `weight` - weight (kgs) of the customer at the time of registration

- `height` - height (m) of the customer at the time of registration
- `blood group` - blood type of the customer
- `avg_n_drinks_per_week` - average number of drinks in alcohol units per week
- `avg_n_cigret_per_week` - average number of cigrets per week
- `education level` - level of education (primary, secondary, bachelor,masters, pHd, other)
- `n_countries_visted` - number of countries visited up until the registration date.

What you will need to submit

To complete the assessment you must submit by **Thursday 15th of December at 1pm** one zipped folder with the following two sub-folders:

1. `Anonymised_data` - This folder contains the instructions on how to access the anonymised dataset in a secure fashion along with relevant data documentation for the future users. You will need to consider what is the best approach for this and use the resources you currently have at your disposal (this does not include the HPC server).

2. `Supporting_material` - This folder contains all supporting material that was used to create your anonymised data and is not shared with the either set of collaborators. This folder should include at least the following items:

    - A pdf copy of your Jupyter-notebook where you display all of the code used to answer the problem along with the resulting output from that code. Please use the text mode (`Esc+M`) or the code mode (`Esc + Y`) to separate your Jupyter-notebook into meaningful sections.
    - A python script (`.py`) created from the Jupyter-notebook.
    - Slides for your presentation.

Other items the `Supporting_material` folder can include are things such as the key or salt file, hash function script and so on.

Please zip your files into a folder titled *CDM_CW2_GyourGroupNumber* and upload your zipped folder rather than individual files. So if your group number is 1 then it would be *CDM_CW2_G1*. Please make sure that the individual files do not contain individual names.

How you will be assessed

The assessment for this coursework consists of a group oral presentation of 10 minutes, 5-10 minutes for questions and the delivery of all the items specified.

You will be assessed against the presentation rubric which is presented in the handbook of this module. In addition we will be checking the files that you submitted and marks will also be awarded for correct anonymisation and procedures.

Things to consider

Below are some questions that you should be thinking about when approaching this problem:

- Who is going to be receiving the anonymised data?
- How are they going to be using the anonymised dataset?
- What are their intentions for the data?
- How can you make the sharing of the data more secure?
- Can you encrypt the data file in some fashion before sharing it?
- What about the balance between data utility and data privacy? Does is it apply to this problem?

Some things to include in your presentation

In addition to what is mentioned in the rubric please consider including the following in your presentation:

- Background of the problem.
- Description of the anonymisation process (method or methods) your group implemented.
- Description of the calculation of the k-anonymity.
- Description of how the anonymised data is shared with the CEO's collaborators and justification of this specific approach.

GOOD LUCK!