Simprints Technical Assessment

General advice

- This assessment contains 1 task with 4 parts and should take between 45-60 minutes - please specify how long it took you to finish. Please don't spend more than 60 minutes - the task is open-ended (one of our engineers took nearly 4 hours and still had unanswered questions) and so we don't expect you to go down the rabbit hole to completion!
- All files you need for this assessment can be found here(`https://drive.google.com/drive/folders/1YfrxmMNbDxvgbfubCmkpS8HbBEidxRaI`)
- If you have any thoughts or feedback regarding this assessment, please give it to us! We'll give you personalised feedback for your assessment if you do this.
- Submit your solution by replying to the email which linked you here. You have freedom as to how you do that (PDF report of a Jupyter notebook, knitted RMarkdown document, link to a private repository etc), but bear in mind that the people reviewing your solutions may not have the same set-up as you, so try and avoid Apple/Windows specific file formats.

Investigative Programming Task (20 points)

A healthcare program, *DevHealthOrg*, uses biometrics to ensure the unique identity of its beneficiaries. They have 3 databases `users.txt`, `areas.csv`, and `visits.csv` that correspond to a project in the region of Syldavia.

- `users.txt` contains information about the users in *DevHealthOrg* (community health workers), and the beneficiaries who are assigned to them. Format for each line: `<UserID>: <BeneficiaryID>,<BeneficiaryID>,<BeneficiaryID>,...`.
- `areas.csv` contains information regarding where the beneficiaries were registered. Format for each row: `<BeneficiaryID>,<DistrictID>`.
- `visits.csv` contains a row every time a user met a beneficiary and delivered healthcare services. Format for each row: `Date,Type,<BeneficiaryID>`. `Type` denotes whether or not the meeting happened in the beneficiary's house (`Home`) or at a healthcare clinic (`Clinic`), and `<BeneficiaryID>` denotes which beneficiary was receiving healthcare. `Date` denotes when the beneficiary was visited.

DevHealthOrg wants to diagnose any systematic problems in healthcare delivery services, and needs your help as a data scientist. Write a script (in a language of your choice) to analyse the data and provide the insights they are after.

Note that each user provides healthcare services to the beneficiaries that they are assigned, and all users are expected to work the same number of hours. For the purposes of this task, a good proxy for "amount of healthcare being received" for a given beneficiary is the number of meetings between a beneficiary and a user (irrespective of whether they happened at a clinic or at home). Besides this, feel free to make assumptions about the data and the program as long as you state them explicitly.

Please provide

- **(A) which districts have beneficiaries that are not receiving sufficient amounts of healthcare (bonus points for a visual representation)**
- **(B) your source code to answer (A)**
- **(C) any assumptions you make**
- **(D) possible explanations for why you see what you see, and potential recommendations to the program**

You will be assessed on your ability to (i) capture trends in the data by asking questions, (ii) communicating your insights visually, and (iii) providing feedback to the program. An example investigation to get you started would be "Are all users doing the same amount of work? (and why)?".

Other

Please also submit

- **(E) how long it took you to do the task in this assessment**
- **(F, optional) any feedback you might have for us about this assessment**