Job Application Assignment
Sophia Genetics
Rue du Centre 172
1025 Saint-Sulpice

Data Engineer
R&D Department
Due format: Git repository
Due Date: 7 days from reception

Instructions

Please finish this task independently and submit a written report together with accompanying code 7 days after reception (you will receive a confirmation of the deadline upon reception of this document).

Dataset

You will receive a dataset file *data.tsv.gz* containing: measurements of 5 different principal components ("dim1-dim5") for 2750 samples, together with sample names ("sample") two different label assignments for each sample: binary class labels ("sample_type_binary") and multi-class labels of samples ("sample_type").

Expectations

The submitted report is expected to be concise and include results, figures as you deem appropriate, as well as short descriptions of your methods and choices. Please consider the use of interactive report formats.

You are explicitly encouraged to use Python, as well as any freely available tools & libraries of your choice (unless specified otherwise) to address the problems. The submitted format of choice should enable deployment and running of the submitted code to exactly reproduce results and figures.

Creative approaches and demonstration of your insight into the understanding of the problem are highly encouraged. Please be aware that some of the questions are open and you are encouraged to hypothesize on possible strategies as well as discuss any relevant published literature.

Submission

To submit, please create a new *private github repository* and push your *report with accompanying code.* In time for the submission deadline, please

- share access to the repository by adding both Alex Seeholzer (github user *@flinz*) and Christian Pozzorini (github user *@pozzorin*) as collaborators
- send a short email notification to `aseeholzer@sophiagenetics.com`.

1 Database design

Provide an Entity-Relationship model and Data Definition Language (DDL) file (manual or generated; you can use any SQL dialect you would like, or any specific technology) to represent the dataset (*data.tsv.gz*) including both sets of label assignments. Assume in your design that the dataset will later scale up to a volume of 10 million rows and additional dimensions of label assignments for each sample will be added later. Select appropriate indexes. Justify and discuss on your design choices.

2 Deployment & ETL

Provide deployment scripts and Python functions, together with basic testing, that:

- Set up the database structure of question 1 above on a Docker instance
- Extract-Transform-Load (ETL) the rows of the dataset, to populate the database

In the report, comment on and discuss the design choices you have made.

3 Data analysis

Consider in this assigment only the quantitative data in *data.tsv.gz* (dim 1 - dim 5) without considering the labels.

Provide an estimate of how many different clusters of different type the data might contain. Discuss and *justify quantitatively* how you arrived at this conclusion. Discuss the methods you have chosen and possible alternatives.

4 Data visualization

Provide an interactive webpage (can be external to the report) that visualizes the dataset appropriately.