Assessment for Data Steward/Research Software Engineering Role

General explanations:

- There are no “right” and “wrong” answers to the questions. Our main goal is to get a better understanding of how you work, how you approach problems, and how your solutions to problems might look.
- You are free to use any of the following programming languages and methods to explain your solutions:
    - R
    - Python
    - Matlab
    - Text explanations
    - Plots, graphics, tables, etc. — whatever you see fit
- Ensure that you can share your solutions with us, ideally via GitHub. Please document how we can reproduce your code.
- If something about the questions is not clear to you, that’s completely fine. In that case, explain how you would try to resolve this if you encountered this uncertainty on the job.
- Don’t worry about getting your code fully functional if you are lacking time or information. Fill in the blanks by describing how you would continue from where you left off.
- In the second interview, you will **get maximum of 10 minutes** to present your approach and your (partial) solutions to the questions below. After that, we will ask you some further questions.

Task description

A senior researcher approaches you about the FIT study. For this project, participants have been recruited to wear wrist-worn fitness trackers. The fitness trackers record heart rate and steps, among other metrics. The fitness trackers are a commercial product, and the data is stored on the suppliers own cloud systems, for which they offer an API. The data is obtained via a third-party integration that makes use of the API and has been specially developed for Tilburg University. The extraction pipeline tool runs on Tilburg’s own servers and is hosted in Kubernetes and Docker, based on minIO. The data is extracted via a web interface, where users manually enter the time range and study participants they want to download data for. Because this requires quite a bit of manual work, it is typically done by students.

The researcher points out that she has very limited technical understanding of how the data collection, the hosting of the pipeline, and the downloads actually work. So, when she found out that the downloaded data is in a non-human-readable format, she did not know what to do with it and decided to contact you.

Questions:

1. What is your approach in the first meeting with the researcher? What would you prepare in advance and what questions would you ask her? What solution(s) would you pitch to her and how would you explain them?
2. She has sent you the files via SURF file sender. She notes that the participant numbers have this format: FIT04715. The researcher would like to correlate how the maximum heart rate throughout the week correlates with the number of steps taken. For this, she asks you for a single dataset for her 25 participants that she can use for her next paper.
3. Some time later, a follow-up study uses the same method for data collection but involves more participants and a higher sampling frequency of the fitness data. Researchers run into the problem that the cleaning takes much too long. What steps would you explore and suggest to resolve this? What could it depend on, and what would be a viable solution? Explore a couple of alternatives if you can think of any.