NeuroFlow Data Team Take-Home Project

Goals of this Project

We are asking you to complete this take-home project because we are impressed with you, but would like to see how your skills will translate to our specific context. Please feel free to do any research or preparation you wish, but do not plan to spend more than a few hours on this project. We're not seeking the perfect solution, just trying to get a deeper understanding of your technical skills and how you approach new problems.

Part 1

Our Problem

For this project, we'd like you to consider data from what we call "Subjective Metrics", which are mood, stress, rumination, and sleep tracking. In our iOS and Android apps, we ask users to rate these four elements on a float scale of 0-4 , with 0 being negative and 4 being positive.

Every morning at 8am, we send our patients a notification to rate their mood and their sleep on a circular scale. The five full options are Awful, Bad, Okay, Good, and Great! Awful is 0 and Great is 4. Additionally, patients can enter these subjective measures throughout the day, whenever they feel like it's necessary.

Users can also rate their rumination ("Are you having trouble letting go of something that happened today?") with options "Not at all", "Slightly", "Moderately", "Very", "Extremely" and they can rate their stress ("Are you stressed about something that might happen in the near future?") on the same scale.

The clinical purpose of these activities is to help patients track their sleep, mood, rumination, and stress over time to see if therapy is making a difference. The provider can see this data too, building the basis of a conversation they can have together.

We currently have the problem of not being able to visualize progress well for these four metrics to mental health providers and their patients.

Your Solution

The attached zip contains a file in CSV format with information on each of these four metrics we track.

For each line, the first column represents the time the measurement was made, the second column represents the id of the user submitting the rating, the third column is the type of rating submitted, and the fourth column represents the value of the response.

Given the information you have and any light research you'd like to do on the topic, what insights can you draw? What assumptions have you made about the data? What are 2-3 additional pieces of information that would be important to collect? We are not looking for production-ready code, but we will assess both your approach to visualization and your technical abilities.

When complete, please send us a Github link or other access to a repo to review what you've done, with any necessary instructions on how to run your code locally. Let us know if you have any questions as you work through this assignment.

Part 2

Our Problem

We'd like to see how you design and write SQL for the given questions. Often our business counterparts will ask us for a quick query to answer a question. In this case, the questions are: How many users completed an exercise in their first month grouped by the month of user creation? Which organizations have the most severe patient population?

As context, our platform is a tool that clinical providers use to assign different kinds of exercises to their patients. Typical exercises may be to write a journal entry, complete a meditation session or fill out a clinically validated questionnaire. The purpose of these exercises are to help their patients stay engaged and ultimately feel better faster, so the earlier this feature becomes sticky for the patients, the longer they'll stay engaged.

We want to identify patterns in patient behavior with respect to exercises, compare this over time, and find the driving factors for their exercise completion rates.

Your Solutions

1. How many users completed an exercise in their first month per monthly cohort?

Assume you have two tables in our company's database:

- 'users' table, with columns 'user_id', 'created_at'
- 'exercises' table, with columns 'exercise_id', 'user_id', 'exercise_completion_date'

Write a single SQL query that breaks up the users based on the month that they signed up (their cohort month), and determines the percentage of users that have a completed exercise in their

first month for each monthly cohort (e.g., the 2018 January cohort has x% of users completing an exercise in their first month, 2018 February cohort has x% of users completing an exercise in their first month, etc.).

2. Which organizations have the most severe patient population?

Assume you have two tables in our company's database:

- 'Providers' table that contains 'provider_id', 'organization_id', and 'organization_name'
- 'Phq9' table that contains 'patient_id','provider_id', 'score','datetime_created'

For context, A phq score ranges from 0-27 and anything 20 or above is considered severe. Write a single query that finds the top five organizations that have the highest average phq9 score per patient.

When complete, please send us a Github link, a sql file, or something similar to review what you've done.

Again, let us know if you have any questions as you work through this assignment.