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. This is a real challenge we are facing, though the data is anonymized or randomized, and simplified.

Part 1

Our Problem

For this project, we'd like you to consider data from GAD-7 validated assessments, which is used to aid in the diagnosis of generalized anxiety disorder as well as screen for panic, social anxiety, and post traumatic stress disorder .

The GAD-7 is a 7 question assessment that asks how often one has been bothered by the problems represented in each question over the last two weeks. The person then scores each question from 0 to 3 using the following scale:

Question Score	Question Label
0	Not at all sure
1	Several days
2	Over half the days
3	Nearly every Day

To score the GAD-7, the values in each question are summed to get the final score.

The final scoring scale is as follows:

GAD7 Score	Severity Label
0-5	Low to Minimal
6-10	Mild
11-15	Moderate
16-21	Severe

When screening for anxiety disorders, a recommended threshold for further clinical evaluation is a score of 10 or greater.

During the course of treatment, patients are asked by their care provider to complete these validated assessments. The cadence at which these are given are typically monthly or twice a month but it can vary.

The clinical purpose of these assessments is to help support clinicians in making a diagnosis, to quantify anxiety symptoms, and to monitor changes 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 this assessment to mental health providers and their patients.

Your Solution

The attached zip contains a file in CSV format with information on the GAD-7 assessment.

For each line, the first column represents the time the measurement was made, the second column represents the id of the patient submitting the assessment, the third column is the type of assessment submitted, the fourth column is the date that the patient was created, and the fifth column represents the final score of the assessment.

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.

Here are some of the key things we look for in your response:

- Clear analysis of business impact
- Investigation of data
- Quality of communication
- Clear and actionable solution

Remember, please limit time to a few hours at most, and feel free to reach out to the hiring manager if you have any blocking questions as you go. If you have minor questions, please simply make a reasonable assumption and call that out in your work so we can follow your logic.

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 assessment. 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. How many users completed a given amount of exercises?

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 returns a frequency distribution of the number of activities each user completed. (Ex: 1000 users completed 1 activity, 500 completed 10 activities, 100 completed 100 activities, etc…)

3. 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 phq9 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, please limit time to a few hours at most, and feel free to reach out to the hiring manager if you have any blocking questions as you go. If you have minor questions, please simply make a reasonable assumption and call that out in your work so we can follow your logic.