Assessment

Questions are encouraged. If you have any questions about the task or data, please reach out to us (at: *alvin[at]antarahealth[dot]com*) and we will respond within 24 hrs. The purpose of this exercise is to test your programming and critical thinking ability and give you an idea of what some of the work you will encounter on the job might look like.

Task

Context

Imagine you are a data something working for a company that provides a digital platform for patient engagement and communication. The company collaborates with various healthcare organizations to deliver targeted health-related campaigns to specific patient populations via messaging apps like WhatsApp.

These campaigns aim to educate, motivate, and support patients with different health conditions, such as hypertension. The messages are tailored to the target audience and may include informational content, motivational messages, or reminders to encourage adherence to treatment plans and promote better health outcomes.

Over the past four weeks, the team responsible for running these campaigns has been diligently collecting data on a campaign to motivate and educate people living with diabetes. They now need your help building some visuals to help them understand the performance of the campaign.

Data

Within `antara_assessment.zip`, the folder `data/` contains two folders:

- *campaign_trackers*
    - Within this folder are 4 files where each file tracks metadata around health-related campaign, with details about the campaign type and associated campaign report.
    - Each file indicates when the campaign run. For example “campaign_trackers - week_1.csv” indicates that this file contains metadata for the campaign’s first campaign.

    - Each of the file contains the following columns:
        - **campaign_id:** A unique identifier for each campaign.
        - **campaign_cohort:** Indicates the target audience and type of message for the campaign. For example, "chronic__motivational_message" indicates the campaign is aimed at people with chronic conditions and contains a motivational message.
        - **campaign_report_file:** The filename of the report associated with each campaign.
        - **campaign_sent_at:** The timestamp indicating when the campaign was sent, including the date and time in UTC format (Coordinated Universal Time).
    - The format of these files changes week over week but will always be either of the following formats: `.csv`, `.xlsx`, `.tsv` .
- *campaign_reports*
    - Within this folder are 4 files where each file contains detailed information about the messages sent to the cohort; to whom the messages were sent and their delivery status.
    - Each of the report contains the following columns:
        - **message_id:** A unique identifier for each message sent in the campaign.
        - **user_attributes:** A set of key-value pairs containing information about the user to whom a message was sent, such as their WhatsApp display name, phone number, User Id, and organization.
        - **message_status:** Indicates the current status of each message sent in the campaign. The possible status values are:
            - *responded:* The user has responded to the message.
            - *read:* The user has read the message but not responded.
            - *delivered:* The message has been successfully delivered to the user but not yet read.
            - *invalid_user:* The message could not be delivered because the user is unable to receive it.
        - **message_delivery_failure_reason:** If the message_status is "invalid_user", this column provides a brief explanation of why the message could not be delivered.
    - The format of these files does not change and will always be `.csv`.

Deliverables

- Using the data above please:
    - Prepare a funnel (visual) of Sent vs Delivered vs Read vs Responded messages. The visual should help answer what was the conversion stage-over-stage.
    - Prepare a histogram (visual) of Count of Responses. The visual should help answer the question how many respondents replied *n* number of times.
- The COO comes by and asks you to identify the highest performing campaign week. You are left to your own devices to define what "highest performing" means. Please provide how you would define which campaign week was highest performing and why. **Do not write code for this**, a short paragraph with your thoughts is more than enough ツ

Constraints

While we do not have any firm expectations on the format of your final deliverable, we request that you do the following:

- Use a combination of either of the following languages:
    - *Python and SQL*
    - *R and SQL*
- For the SQL bit:
    - Use duckdb *(Python/R)*
    - Use it in at least one 'core' part of your work, an example of this is using SQL to aggregate the data that helps you build a visual.
- Submit your assignment as a compressed folder within which at least one file should either be a Jupyter Notebook or R Notebook containing your working (depending on the language you choose).

Apart from the above constraints please feel free to use whatever libraries/packages you are most comfortable *(pandas, polars, tidyverse packages, plotly etc)*.

Rubric

We will be evaluating submissions on the following criteria:

- The submission is self-contained and can run on any environment where Python or R is installed.
- The code is accurate (produces what it intends to produce, without errors).
- The code is easy to read and interpret, including commenting where warranted.

- The code shows an eye for reusability and maintainability.
- The approach is valid (any math, aggregation, or other analysis).
- Comments are clearly written, and suited to their audience (a fellow analyst).
- The response to the COO demonstrates professionalism and business thinking.