**GiveDirectly Data Associate Take Home Exercise**

Please use the accompanying CSV files to answer the following questions as directly as possible. Please attach file(s) with your results to your reply message, including all work produced -- code, Excel spreadsheets, etc. **Do not include your name anywhere in the submitted contents.** You have three days to return this assessment, but please try not to spend more than two and a half cumulative hours working on it. Please let us know if you do not anticipate having time within the next three days.

The files contain fake data from a hypothetical GiveDirectly project. It is designed to evaluate your thought process for a real project, but is not representative of a real GiveDirectly project. As with real projects, the data may contain errors and you will have to make judgments about how it can and cannot inform operations.

The **recipients.csv** file contains data on potential recipients for the hypothetical project.

* **recipient\_id** = unique recipient identifier
* **county** = county in which the person lives
* **time\_county** = number of years living in current county
* **age** = person’s age
* **account\_number** = phone number associated with the person’s mobile money account
* **account\_status** = status of their mobile money account
  + Active = ready for payment
  + Not Active = not ready for payment and our field team must follow up to resolve the issue

The **survey\_attempts.csv** file contains data from the recent attempted phone surveys associated with these potential recipients, including calls in which the person did not answer.

* **survey\_id** = unique survey attempt identifier
* **recipient\_id** = unique recipient identifier associated with the survey attempt
* **date** = date the survey attempt was conducted
* **success** = TRUE if survey attempt was successful -- i.e., the correct person answered the phone and was surveyed -- otherwise FALSE

In order to determine next steps, we must assign each recipient to a **stage** based on the data using the following logic in sequential order of priority. Each stage can be thought of as a step in the payment process.

1. **stage** = Start

**success** = FALSE for all surveys associated with the recipient

1. **stage** = Ineligible

**success** = TRUE for one survey associated with the recipient *AND*

**county** is NOT among: County A, County B, County C

1. **stage** = Review:

**success** = TRUE for one survey associated with the recipient *AND*

**county** is among: County A, County B, County C *AND*

**account\_status** = Not Active

1. **stage** = Pay:

**success** = TRUE for one survey associated with the recipient *AND*

**county** is among: County A, County B, County C *AND*

**account\_status** = Active

**Questions**

1. Please evaluate the data in recipients.csv and survey\_attempts.csv to answer the following questions:
   1. How many recipients are in each of the four stages? Please provide the calculation(s) in the spreadsheet or code that you submit.
   2. How many surveys were successfully completed in December, 2020? Please provide the calculation(s) in the spreadsheet or code that you submit.
   3. Did you find any abnormalities in the source data? If so, how did you account for them in your analysis?
2. The program manager has asked for data to help determine whether the field team should focus more effort on calling those in stage Start or following up to resolve issues with those in stage Review. Please write a response to the program manager, including data that may help inform the decision, and some additional factors that you would take into consideration to make the decision. Assume that the program manager’s expertise does not include interpreting data and complex analytics. Please limit your written response to 300 words or less.
3. The country director is considering investing resources into proactively conducting in-person surveys with recipients *in the highest age group* across projects to increase overall survey success rate. They believe that this additional cost might outweigh the current costs of repeated failed phone survey attempts, if we can accurately target those recipients *least likely* to respond to a phone survey.
   1. What analysis would you provide from the provided project data to help make this decision? Please provide the calculation(s) in the spreadsheet or code that you submit.
   2. Are there other factors that might explain the observed survey success rate from this project? Please use your judgment to determine these factors and limit your written response to 400 words or less.