**REACH Data analyst Test**



This test includes three parts, 1. General Knowledge, 2. Data Processing and 3. Data Project. Use the spreadsheets Annex 1 to help you answer the questions below.

The test has been designed to take 2:30 hours and examines several competencies regarding data analysis. Read all questions before you begin and note that the three parts can be completed in any order. **When sending back your answers, please share *all scripts, code, files etc.* that you used to solve the exercises. Please also list all websites/external sources you used to answer the questions. Using generative AI to answer the questions is not permitted.** Checks on the use of AI will be performed.

All answers can be noted directly on this answer sheet unless otherwise specified. Please return this document with you answers, together with Annex 1 by email.

**Part 1: General Knowledge**

1. Explain p-values in layman terms. Feel free to use analogies or examples. Keep it simple, but make sure to stay technically accurate.
   1. Why are they important?
      1. P-values given a measure of how precise an estimate is. It tells us how likely we are to see a value at least as large under the assumption that the null hypothesis is true. In the social sciences, a p-value less than or equal to 0.05 is considered to be a statistically significant result.
   2. How can they be interpreted?
      1. They can be interpreted as the probability of seeing a value at least as large for the coefficient under study given that the null hypothesis is true i.e. no effect. This gives some indication that the results are just noise.
   3. What are some common pitfalls/misunderstandings in their use and interpretation?
      1. Common misunderstandings of p-values are that they are the probability that the null-hypothesis is true or that the alternative hypothesis is true. A common misuse of p-values comes up when dealing with researcher degrees of freedom. Performing multiple comparisons of the data can increase the probability of encountering a false positive. There are also issues such as deciding to average certain groups, exclude other observations, and choosing regression predictors that affect the validity of a p-value (See Gelman https://stat.columbia.edu/~gelman/research/published/asa\_pvalues.pdf)
2. When would a Mosaic plot be an appropriate visualization?
   1. A Mosaic plot is good for showing percentages of data in groups. It is a graphical representation of a contingency table. This would be useful for showing the effect of an intervention for treatment and control groups, for example. (See https://www.jmp.com/en\_us/statistics-knowledge-portal/exploratory-data-analysis/mosaic-plot.html#:~:text=Mosaic%20plots%20are%20used%20to,a%20visual%20comparison%20of%20groups.)
3. What is personally identifiable information (PII)? Provide an example. When is it ok to collect PII?
   1. Personally identifiable information is sensitive information that can be used to track down or know the identify of a person. This could be someone’s name, phone number, email, or ID number. This is generally something that should be kept safe. Information should not leave the office nor should it be discussed with unauthorized parties. It could be appropriate when signing up for services where PII is mandated or for following up with a customer.

**Part 2: Data processing**

In the spreadsheet Annex 1, you will find a raw dataset from a recent data collection exercise that was carried out by your team.

1. There are errors in in the dataset. Please identify **at least four** errors by highlighting them in yellow in the excel sheets. In the cleaning log tab, report the cell IDs, variable name and a small explanation on why you think this value can be an error in the comment column.
2. Using the programming language python, create a new variable characterizing the household drinking water source into improved / unimproved source following the classification below. Paste the code / function you used below.

# Create a new variable using the following table mapping

import pandas as pd

import numpy as np

df = pd.read\_csv('REACH\_HR\_TEST\_DATA\_analyst.docx-EmbeddedFile.xlsm - Annex 1 - REACH Assessment Tes.csv')

improved\_list = ['Protected dug well', 'Piped water to yard or plot', 'Piped water into dwelling (house)', 'Bottled Water', 'Tube well or borehold', 'Public tap or standpipe', 'Protected spring']

# Create empty variable

df['improved\_water\_source'] = ''

# Assign value to variable based on criteria in table from Problem 2

df.loc[df['drinking\_water\_source'].isin(improved\_list), 'improved\_water\_source'] = 'Improved water source'

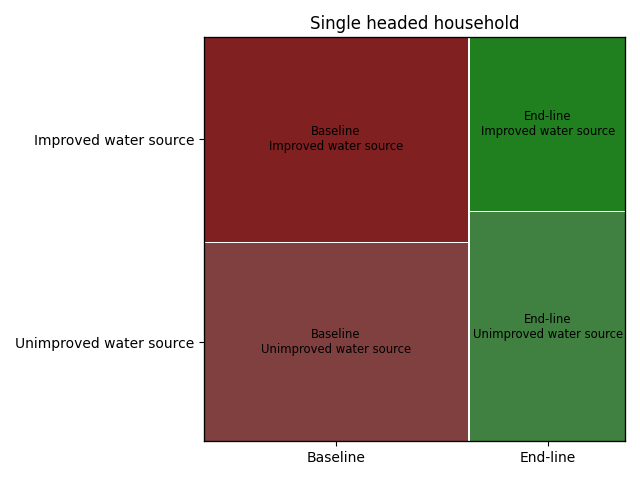
df.loc[~df['drinking\_water\_source'].isin(improved\_list), 'improved\_water\_source'] = 'Unimproved water source'

df.loc[df['drinking\_water\_source'] == 'Other', 'improved\_water\_source'] = np.nan

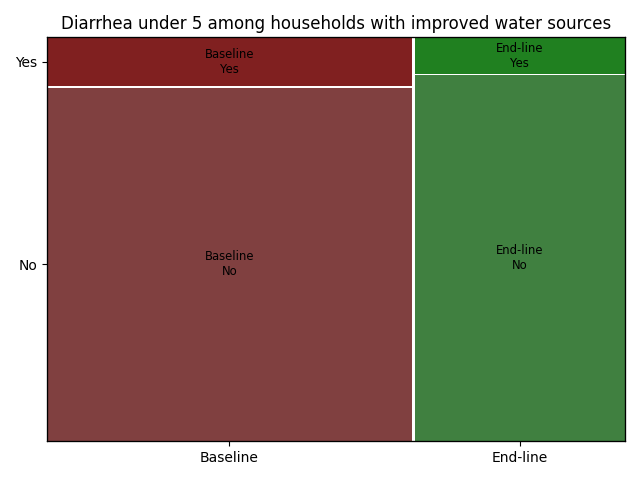
*Table 1: Unimproved / Improved drinking water source categorization*

|  |  |
| --- | --- |
| drinking\_water\_source | Improved water source |
| Protected dug well | Improved water source |
| Piped water to yard or plot | Improved water source |
| Piped water into dwelling (house) | Improved water source |
| Bottled water | Improved water source |
| Tube well or borehole | Improved water source |
| Public tap or standpipe | Improved water source |
| Protected spring | Improved water source |
| Other | NA |
| Cart with small tank or drum | Unimproved water source |
| Unprotected dug well | Unimproved water source |
| Unprotected spring | Unimproved water source |
| Surface water | Unimproved water source |
| Tanker-truck | Unimproved water source |
| Rainwater collection | Unimproved water source |
|  |  |

1. **This exercise requires the results of the previous exercise.** Use any tools, statistics and visualizations that you see fit to analyze the questions below regarding how access to improved water sources changed between the baseline (first data collection round) and the endline (second data collection round, after a water improvement project has been implemented). Records for both rounds are in the same dataset; and come from randomly sampled households in the area of intervention. The column “*data\_collection\_round*” is “Baseline” for records of the first round, and “Endline” for records from the second round. *Please share all code/files used for the analysis.*
   1. Did single headed households receive more/less improvements? (relevant data column: “*single\_headed\_household*”)
      1. Single headed households received fewer improvements in endline than in baseline. There were 137 improved water sources in baseline (50.7% of all water sources) but only 68 improved water sources in endline (43% of all water sources)



* 1. Did the improvements affect cases of diarrhea in children under 5? (relevant data column: “*diarrhea\_under\_5”*)
     1. Yes, diarrhea in children under 5 went down from 72 cases in baseline (4.9%) to 31 cases in endline (3.8%)



**Part 3: Data project**

**Instructions**

Describe in detail the various stages involved in successfully completing a complex data project.

You are supporting a large household survey data collection in KoboToolbox. As part of this project, you want to develop an algorithm that would identify suspicious behavior of enumerator during data collection.

You have it at your disposal.

* the raw data from ODK / KoboToolbox,
* the clean data
* the cleaning log in which cleaning operations are logged (see example in excel Annex 1),
* the audit files from ODK / KoboToolbox that logs enumerator behavior during interview administering the interview. [Form Audit Log - ODK Docs (getodk.org)](https://docs.getodk.org/form-audit-log/)

We want to identify if there are suspicious surveys (fake interviews for example) in the data collection that is coming:

* What are the different stages involved in carrying out such a project?
  1. First download the audit.csv from ODK
  2. We could first look at the distribution of times of interviews. Interviews that are suspiciously short or in the bottom 5% of times taken would be marked for further investigation.
  3. We would also look at the location data to compare it with where the respondent is supposed to be. Assuming that the interview was conducted at the respondent’s home, data that was filled in outside of this would be suspicious.
  4. We would also look at the distribution of the number of changes to response. Enumerators with an unusually high number could be investigated further, perhaps by performing a text analysis of their reasons for changes.
  5. Another way of checking for fake surveys is by conducting back checks. A random 10% of households could be selected to be re-interviewed by a different enumerator. If the survey responses differ markedly from the original survey, this could be evidence of a fraudulent survey.
* For each stage, detail the tasks to be carried out, the methods to be followed, possible avenues and critical points.
* Thoroughly justify your choices.
  1. After downloading audit.csv, a duration variable could be created that is difference between the end and start times of the survey. The elapsed time is accurate even in the start and end timestamps aren’t as mentioned in [Form Audit Log - ODK Docs (getodk.org)](https://docs.getodk.org/form-audit-log/). A histogram of the duration could be be plotted. Surveys with times in the bottom 10%, for example, could be checked to see whether the answers are sensible, how many skipped questions there are, etc. This could indicate whether an enumerator is simply rushing through a survey without regard to skip patterns or filling in responses at random.
  2. For the location data, the GPS coordinates of where the survey was meant to be conducted can be checked against the location where it was actually done. The downside of this is that if the device shuts off or location data is not turned on, this would make it hard to know whether the survey was done at the appropriate location
  3. For the number of changed responses, we could compute the number of changed responses per survey and plot a histogram of the result. Those in the highest 10%, for example, could be set aside for further investigation. We could look at the reasons for the changes and compare those with surveys that are closer to the median.
  4. Finally, for the back check portion, we can randomly sample a subset of interview IDs to be re-interviewed. A trusted enumerator, preferably separate from the rest of the team, will then be sent to re-interview those households using the entire questionnaire or only the questions of concern. The answers will be compared to those from the first survey. One way to compare is to use the *bcstats* tool developed by Innovations for Poverty Action <https://github.com/PovertyAction/ipabcstats>. For more information about back check procecdures, see<https://dimewiki.worldbank.org/Back_Checks>