Laterite Technical Assessment

Data Quality Analyst

September 2023

laterite

DATA | RESEARCH | ANALYTICS

INSTRUCTIONS

The purpose of this exercise is to test your capacity to process a dataset, the proficiency, efficiency and replicability of your code and your ability to clearly explain all the steps you take when processing data. This is a task that you will routinely encounter in your role as a Data Quality Analyst.

To complete the assessment, use STATA. You may consult any resources you like, **except other people**. Please remember to include a list of any resources that you consulted in your submission.

You are expected to complete this task **within 24 hours** of receiving it. Submissions received after this time will not be considered.

Please submit your completed assessment documents using the link in the email that was sent to you informing you of the Technical Assessment.

You should send the following documents when submitting the completed assessment:

- All the STATA code that you have used in the data cleaning i.e., STATA do file
- Any dataset you might have used and/or produced.

Please make sure to send us back all the documents in a compressed zip file.

TASK

Prior to the launch of any data collection project, Laterite conducts listing activities to recruit potential survey participants.

Laterite recently finished listing activities for a project on household incomes. The study was conducted in the 3 districts that make up Kigali City. In each of the districts, a list of households in ubudehe category 1 was collected.

For each household, additional information was collected on the household head.

You have been provided with 2 datasets for each district:

- a main dataset, and
- a roster dataset.

The main datasets contain information on:

i. where the households on which ubudehe data was collected are located,
ii. the local authority in charge of ubudehe data, and
iii. the number of households in ubudehe category 1.

The roster dataset contains information on:

i. the household head in each of the households listed.

A detailed description of all the variables is given in Table 1 below.

The Data Quality Manager wants you to undertake the following sub-tasks and output relevant files:

1. Using the variables names and values labels provided in Table 1 below, please label:
   a. all variables in the main and roster datasets
   (10 marks)
   b. the values of all categorical variables in both the main and roster datasets
   (5marks)

2. Using the values and value labels in **Table 1 (Variable description**) below, create a single variable for each of the location variables (province, district, sector, cell, and village) that displays the value label but also has the value embedded.
   (15 marks)

3. Please create a dataset combining the main and roster dataset so that we have one observation per village with all household heads as well as their personal information contained in the roster.

   Save this dataset as "*merged_yourname_yyyymmdd*".

   - ***yourname*** means both your names starting with first name then last name.
   - ***yyymmdd*** means year, month, date e.g., 20230920

   (20marks)

4. Organize the data set.

a. Order the variables as per **Table 1 (Variable description**) below.

(5 marks)

b. Make sure all variables are labeled properly as per **Table 1 (Variable description**) below.

Save this dataset as "*clean_yourname_yyyymmdd*".

- ***yourname*** means both your names starting with first name then last name.
- ***yyymmdd*** means year, month, date e.g., 20230920

(10 marks)

5. To aid the field team in drafting a data collection field plan, please extract lists of households per district in .xlsx format, with each sector in each of the 3 districts on a separate worksheet.

   Save your do-file as *"yourname_yyymmdd".*

   - ***yourname*** means both your names starting with first name then last name.
   - ***yyymmdd*** means year, month, date e.g., 20230920

   (25 marks)

Additional marks will be given for the quality and clarity of your code.

(10 marks)

TABLE 1: VARIABLE DESCRIPTION

Variable name	Variable Type	Values	Value Labels	Variable Label
province_id	Numeric	N/A	N/A	Province code
province_name	String	N/A	N/A	Province of residence
district_id	Numeric	N/A	N/A	District code
district_name	String	N/A	N/A	District of residence
sector_id	Numeric	N/A	N/A	Sector code
sector_name	String	N/A	N/A	Sector of residence
cell_id	Numeric	N/A	N/A	Cell code
cell_name	String	N/A	N/A	Cell of residence
village_id	Numeric	N/A	N/A	Village code
Village_name	String	N/A	N/A	Village of residence
list_avail	Numeric	1	Yes	If the ubudehe list is available
		0	No	
list_source	Numeric	1	Sector Executive Officer	From whom the ubudehe list was obtained
		2	Cell Executive Officer	
		3	Village Leader	
ubudehe_contact	String	N/A	N/A	Ubudehe list contact person
contact_role	Numeric	1	Village leader	Role of the ubudehe list contact person
		2	Village in-charge of security	
		3	Village Community Health Worker	
contact_phone	Numeric	N/A	N/A	Phone number of ubudehe list contact person
num_households	Numeric	N/A	N/A	Number of households in ubudehe category 1
hh_head_position	Numeric	N/A	N/A	Position of household head on ubudehe list
name_hh_head	String	N/A	N/A	Name of the household head
nid_hh_head	String	N/A	N/A	National ID number of the household head
phone_hh_head	Numeric	N/A	N/A	Phone number of household head
gps	Numeric	N/A	N/A	GPS coordinates of where list was obtained
parent_key	String	NA	NA	Unique ID

laterite
DATA | RESEARCH | ANALYTICS

From data to policy

Rwanda | Ethiopia | Kenya | Sierra Leone | Tanzania | Uganda | The Netherlands

www.laterite.com