

## Case study - Analyst [M&E/Data Analytics Teams]

### Guidance

As part of your interview process, we would like you to better understand your data skills. This task will require you cleaning and analyzing some data, and using this dataset that you create to answer some questions that resemble some of the work you would carry out at NewGlobe.

### Instructions

- This case should take you between 1-2 hours to complete. Please do not take longer than 2.5 hours to complete this task.
- You may use Stata or R for this exercise.
- Please complete all coding exercises in a well-annotated and well-organized script.
- What should you send us when you have completed this exercise? Two things:
  - Please make a copy of this document and change its name to "Case study - Data Analytics Team [YOUR NAME]". Then write all your answers, and add any tables or figures you'd like to show (if at all) in this document. Share that link with us.
  - In a folder, please share with us all the code you may have used, and any clean or intermediate files you may have created.

### Some tips

- Feel free to Google commands, functions, or coding help, but *we ask that you do not consult your answers or analytical approach with anyone else.*
- We value high-quality code, but your creativity in tackling questions, and providing nuanced, short, and insightful answers is even more valuable.
- Two additional points that go together: (1) We are evaluating applications holistically, and (2) some of these questions may be hard/cryptic on purpose. If you can't answer something, that is okay! Please do not spend a lot of time figuring it out. We want to see candidates as a whole, and the fact that you can't answer one or more of these questions does not mean we will immediately discard your application. After all, we are all learning!

### Case study

We will be using data from our Bridge Kenya programme. You can find that raw data in this [folder](#) (in Excel and Stata formats - feel free whichever format you prefer). We will ask you to complete tasks that involve four crucial skills for our analyst: (1) data cleaning, (2) calculation of key performance indicators (KPIs), (3) descriptive analysis, and (4) impact evaluation.



## Some context

According to the datasets provided, our Bridge Kenya programme operates in 111 schools, in 7 provinces across 31 regions in Kenya. You will get anonymized data for ~13,000 pupils from grades 1-5 from the end of an undisclosed school term in the past five years. (Note that each school year has three terms, and they consist of ~3-3.5 months each).

## The data you received

You have received four files, all in .dta and .xlsx formats, so you can use whichever format you prefer. These files are the following:

- *“Lesson completion”*: file provided at the teacher level, meaning that there is a unique row for each teacher. The file contains the grade that each teacher teaches, and the average lesson completion rate over the term of interest.
- *“Pupil attendance”*: file provided at the pupil level (that means that there is a unique row for each pupil). This file includes the unique school ID, unique pupil ID, the pupil’s grade, the attendance records, and the present records.
  - The attendance records means the total number of times that a pupil’s teacher took attendance.
  - The present records means the total number of times that a pupil was present, out of the attendance records.
- *“Pupil scores”*: file provided at the pupil\*subject level (that means that there are more than one row per pupil). This file includes the unique school ID, unique pupil ID, the pupil’s grade, the subject for this assessment, and the score obtained in this assessment.
- *“School information”*: file provided at the school-level. It includes the region and province where each school is located, the unique school ID, and the “treatment status” (yes/no) for a given tutoring program.

## Step 1: Data cleaning (~45 min)

Please create a file at the student-level which has information about their test scores, school information, their attendance, and their teacher’s lesson completion rate. **Note that this is the main data set that we expect you to share with us.**

**Hint:** note that the four data sets you will use are all presented at different “levels” of the data (e.g., “School information” is at the level of the school, but “Pupil scores” is at the level of the student). Therefore, we suggest that you start by reshaping the “Pupil scores” file so that each student only has one row in the data, with different columns for their scores in math, fluency, and Kiswahili. Use this as your “base file”, and start merging all the other files to this. Be careful with how you merge things: since there are many students to a school or even a teacher, some of these merges will need to be “many-to-one” (but not all).



## Step 2: Calculating KPIs (~25 min)

One of our main KPIs within the Schools Vertical is “Percent Pupils Present”. The “layman’s definition” of this KPI is “The percentage of pupils who were present, out of all pupils - across all days in the term to date”. In other words, the percentage of pupils who were present (for each pupil in the “Pupil attendance” file, this is displayed in the “present\_records” variable), out of pupils who had attendance records (the “attendance\_records” variable in the same file).

- The first task is to translate this KPI into the data. We will calculate this KPI in two different ways. First, calculate this KPI for all pupils at once. What is the network-level average Percent Pupils Present (use two decimal points)?
- Now, please calculate this percentage for each school, and create an average at the school-level. What is the average Percent Pupils Present now (use two decimal points)?
- How does the interpretation of the KPI change between the two approaches? Does it matter in this case? When would it matter, (i.e., when would one be more appropriate than the other?) 2-4 sentences max.

## Step 3: Descriptives (~30 min)

Let’s dig into the reading fluency scores in your current data set. These came from the “*Pupil scores*” data, but you will need the data set you created in Step 1 above to answer these questions. Please answer the following questions as succinctly as possible.

- Please create a figure or a table, whichever you prefer, which shows average fluency scores for each of the five grades.
- Which regions (using the “region” variable) have the lowest and highest average fluency score across all grades?
- Please create a binary variable that is 1 if a given child reads at 10 or lower, and 0 otherwise. Please create a bar chart with grades on the x-axis, and the share of pupils scoring under this threshold for each grade.
- What school has the highest share of pupils scoring under this threshold in grade 3?

## Step 4: Impact evaluation (~20 min) - [“Bonus points”]

During this term, we rolled out an intensive after-school tutoring program in 55 schools. The selection to be a part of the 55 schools was randomly assigned - in other words, these schools were part of a randomized controlled trial (RCT). The “School Information” data set has a binary variable for whether each school was part of the program or not.

- Our Chief Academic Officer would like to know whether this program had any effects on test scores in math, Kiswahili, fluency, and/or student attendance. Please conduct any calculations you see fit to answer his questions.
- After conducting the impact evaluation, we have heard anecdotally that teachers in schools that received tutoring felt more motivated and were completing their lessons at a faster pace. Hence, we could worry that the effects that we see are not (solely) due to the tutoring program, but also due to the higher lesson completion rate. Does this hypothesis hold up in the data?