Case study - Data Analyst, Schools

Guidance

As part of your interview process, we would like you to better understand your data skills. This task requires you to clean and analyse some data. You will also use the dataset you create to answer questions. This case study resembles the work you carried 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, R, or SAS 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? Three 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 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 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 to use 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 anonymised 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. This means 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 there is a unique row for each pupil). This file includes the unique school ID, the unique pupil ID, the pupil's grade, the attendance records, and the present records.
    - The attendance records mean the total number of times that a pupil's teacher took attendance.
    - The present records mean the total number of times a pupil was present out of the attendance records.
- *"Pupil scores":* file provided at the pupil*subject level (that means that there is 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.
- *"School information"*: file provided at the school level. It includes the region and province of each school, 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 pupil level with information about their test scores, school information, attendance, and their teacher's lesson completion rate. **Note that this is the main data set 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 pupil). Therefore, we suggest that you start by reshaping the "Pupil scores" file so that each pupil 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 pupils 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)?

  The average Percent Pupils Present is 76.98 %

- 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)?

  The average percent Pupils present is 76.83%

- 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.

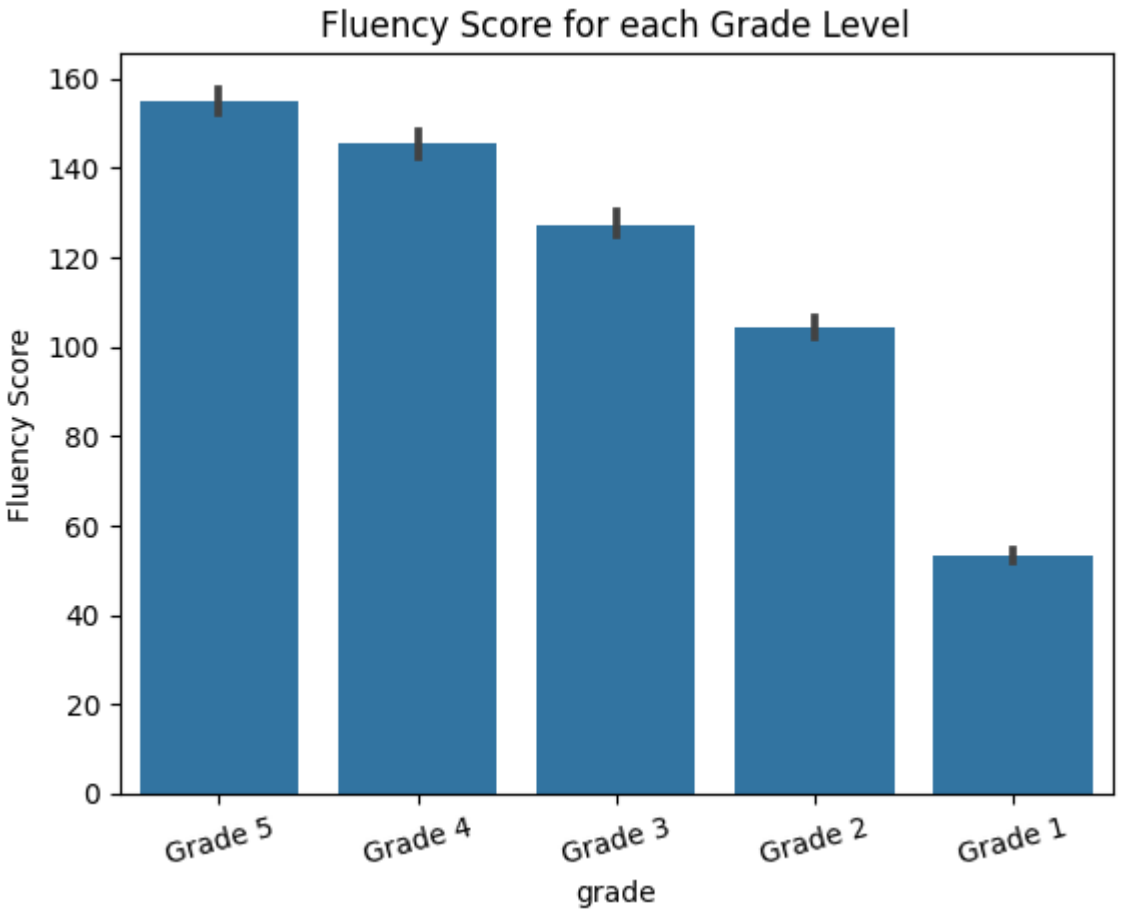
  Having observed that the difference between the average KPI for all students and the average KPI for all students in each school is minimal (0.15). In this case study, the result of both scenarios are similar and won't have much significance on the analysis. However, it will matter if there's a significant difference (1% or more) between the average KPI for all pupils and the average KPI for pupils in each school.

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.



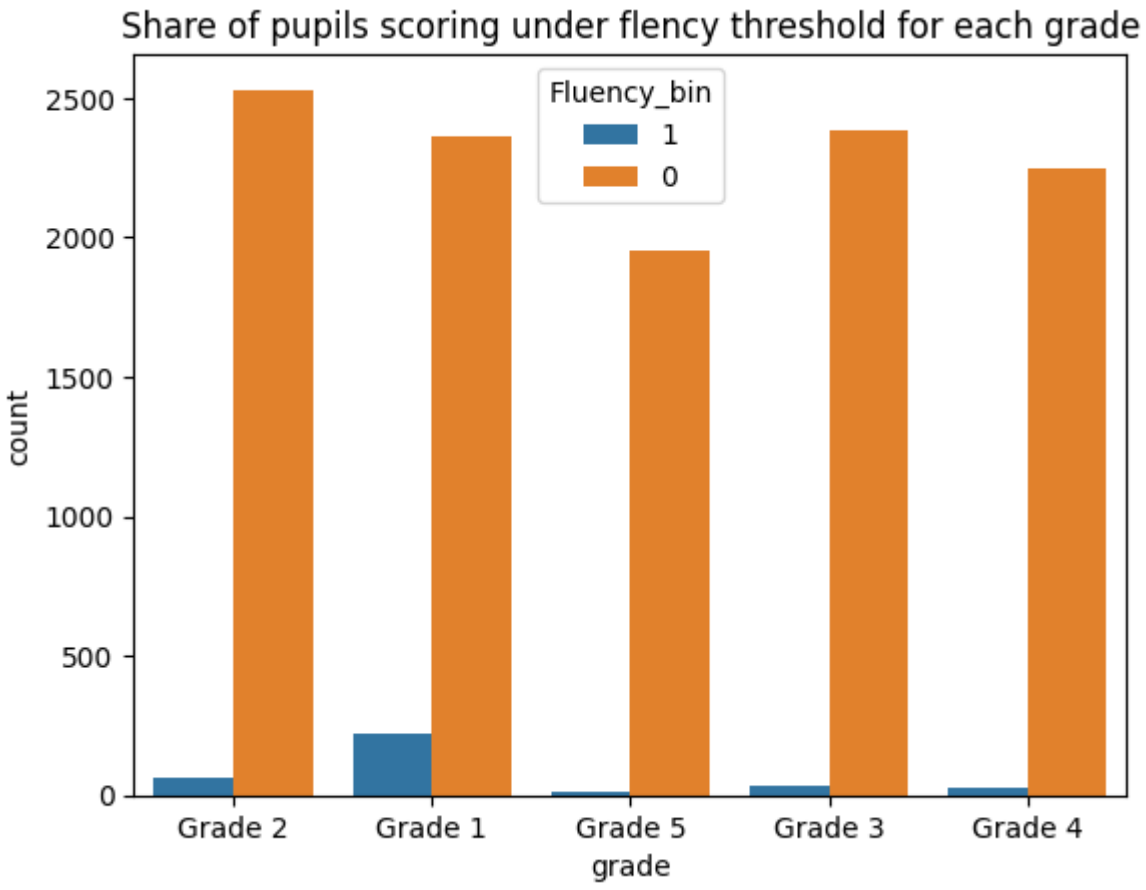
The fluency scores of students increase as they advance from a lower to a higher grade. Students in Grade 1 have an average fluency score of about 60 while those in Grade 5 have a fluency score of about 158. The massive change in fluency scores in pupils in lower grades and pupils in high grades implies that students in higher grades have learned better which is a good indication of an improving syllabus system.

- Which regions (using the “region” variable) have the lowest and highest average fluency score across all grades?

Highest region for fluency score= Machakos

Lowest region for fluency score = Kirinyaga

- 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.



There is a higher level of students with fluency scores greater than 10(category 0) than those with lesser scores across all grades.

- What school has the highest share of pupils scoring under this threshold in grade 3?

school_id 581940 has the highest share of pupils scoring in category 0(score greater than 10) in grade 3

school id 46528 has the highest share of pupils scoring in category 1(score less than 10) in grade 3

Step 4: Writing a Memo (~20 min) - [“Bonus points”]

Our Chief Schools Officer is presenting a brief memo on “Pupil Scores” to the NewGlobe leadership, and you were asked to write a short memo. Using the answers in Step 3, create a memo on page 4 summarising your findings on “Pupil Scores”.

1. Students in higher grades have the highest average fluency score. The higher the grade (from grade 1 to 5) the higher the fluency score. This means students learn to speak or read better as their level or grade advances.

2. Students in Grade 1 have an average fluency score of about 60 while those in Grade 5 have fluency of about 158. The massive change implies that students in higher grades have learned better which is a good indication of the school's improving system.

3. There is a higher level of students with fluency scores greater than 10 than those with lesser across all grades, which is a good indication on student performance across each grade.

4. school_id 581940 has the highest share of pupils scoring in category 0(fluency score greater than 10) and school id 46528 has the highest share of pupils scoring in category 1(fluency score less than 10) in grade 3