**Assessment Documentation and Answers**

**PARSING**

Before proceeding to resolving the questions, it was important to decipher the given files through Python. Parsing them to a readable format was the very first step of the process.

* NGPR7AFL.DTA

This was a stata file not natively readable by a laptop. In order to avoid installation of any 3rd party applications, this file was converted to a .csv using the ‘read\_stata’ method from the pandas package. The final output was much more readable and had 188010 rows and 358 columns.It is now a .csv file titled ‘**data.csv’**.

* NGPR7AFL.MAP

To understand the abbreviations in the column names in the stata file, the .map file was converted into a .txt file using Python’s write method, to serve as key to understand the meanings of different columns and carry out the rest of the analysis. It is now a .txt file named ‘**output\_data\_dictionary.txt’**

**Q1. How many children under the age of five were tested for malaria with a blood smear**

**test?**

A total of **8351** people were tested for malaria using the blood smear test. The rationale behind this was that column ‘hml34’ corresponding to ‘Bar code for blood smear sample’ & column ‘hc1’ which was age in months . The number of bar codes generated were counted to find out the total number of people tested through blood smear.

**Q2. How many children under the age of five were tested for malaria with a rapid test?**

The rapid test did not have any associated barcode to verify how many people were tested. The data provided has one relevant column, ‘hml35’, which corresponds to the ‘results of rapid test’. The entries in this column can be positive, negative, or blank.

Given this setup:

* If we count all entries, including blanks, a total of **12,867** people were tested.
* If we count only those who received results (positive or negative), **then 11,197** people were tested.

Without additional data, it is not possible to determine the exact number of children under the age of five tested with the rapid test.

**Q3. How many children tested positive for malaria with a microscopy?**

A total of **49,658** children were tested for malaria with microscopy. According to the data, **1,936** children had positive results. The positive results were determined based on the 'hml32' column, which records the results of the blood smear test. Although there is no specific column indicating microscopy results, it is reasonable to infer that the blood smear test, a common method for detecting malaria, is used as a proxy for microscopy results.

**Q4. How many children tested positive for malaria with a rapid test?**

**4258** have shown positive results through rapid test.

**Q5. Make a plot showing the number of U5 children that tested positive for malaria by**

**microscopy and the number of U5 children that tested negative for malaria by**

**microscopy in urban and rural areas.**

In rural areas, **1430** have been recorded positive and **3287** were negative.

In urban areas, **506** were recorded positive and **2921** were negative.

Rural areas display nearly 3x more positive cases.

A green rectangular object with white background

Description automatically generated

**Q6. Make another plot showing the proportion of U5 children that tested positive for**

**malaria by microscopy and the proportion of U5 children that tested negative for**

**malaria by microscopy in urban and rural areas.**

In rural areas, **30.3%** have been recorded positive and **69.7%** were negative.

In urban areas, **14.8%** have been recorded positive and **85.2%** were negative.

A red and blue circle

Description automatically generated

**Q7. Make a map of the number of children that tested positive for malaria by state.**

**A map of nigeria with red and green colors

Description automatically generated**

Any state with cases >=50 has been marked red, otherwise green. We can see Kano has maximum of 119 positive malaria cases and Lagos has the minimum of only 6. Each state’s numbers can be seen when running the cursor on it.

**Q8. Extract the data from the raster file and make a map showing average housing quality**

**values for each Nigerian state.**

A map of nigeria with different colored areas

Description automatically generated

We can see that North Nigeria has poorer housing quality, it improves as we move towards the South. Imo has the best average housing quality out of all states.

**Q9. Identify any website of your choice and scrub their data. Pull it into a csv and make at least one data visualization.**

For this exercise, I have scraped IMDB (Internet Movie Database), where I record my ratings and made 3 graphs.

The first one shows the distribution of the ratings I have recorded on the app and we can see there are majorly 7&8 out of 10.

A blue and white graph

Description automatically generated

The second graph the ratings distribution over time, and on which date they were rated.

A graph showing a line of blue lines

Description automatically generated with medium confidence

Additionally, this bar graph shows how my rating compares to the rating IMDB has recorded for the movie, this offers a contrast.

A graph of a movie

Description automatically generated with medium confidence