**Theoretical**

1. At first glance, the Master Health Facility Dataset looks cleaner and more arranged than Health Facility Assessment. However, there are many issues to get it to be in the right format before usage with the Health Facility Assessment vice versa. The first step is to carry out Data Quality Checks.

* Addition of variables to Master Health Facility Dataset:  
  Not only does it add variable names to the Master Health Facility Dataset, but it is also imperative to ensure that the variable names to be added to this dataset should be in harmony with the Health Facility Assessment dataset.
* Consistent Value Across the Variables:  
  This will ensure that observations have consistent values. They should be capitalized or small letters used. Any approach adopted should adhere to value uniformity.
* Check for Missing Values in the two datasets
* Harmonize columns containing codes in the two datasets:  
  Again, this will ensure that the two datasets are in sync.
* Checking for a character in numeric and numeric and character columns.Once Data Quality Checks is complete, data pre-processing which includes data cleansing, reduction, and transformation. However, this process is done with the contribution of stakeholders or the final user to ensure that the significant columns are prioritized during this process.

Finally, once the pre-processing stage is complete, Data Validation is carried out

on the two datasets before matching them.

2. Tools to Use.

There are good tools to use to achieve this, but every tool is unique in there right.

However, R/Rstudio would be a great tool to carry out this task because of the rich

libraries explicitly tailored to handle tasks like this. This contains libraries to carry

out data checks, data pre-processing, cleansing, reduction, and transformation.

It also contains tools to check outliers, scatter plots, standard plots, and even

machine learning models on the pre-processed datasets.

Finally, Rstudio can run Python scripts if it needs to be combined to achieve a

purpose.

1. Matching of Datasets

To match the datasets, the Unique feature to use for this purpose is the **Facility name** since the facility name is Unique and present in the two datasets. While the different dataset that we can leverage with the two give is the **Household Surveys.**