**Cleaning and Appending CMIE Consumer Pyramids Data**

Author: *Ewan Nikhil Thomas*

Script to be used: *subsetting\_appending\_data.py*

This document explains in detail the processes undertaken in cleaning and appending monthly consumption expenditure data from September 2020 to August 2021. The process was undertaken as part of the INFUSION project under ‘BIPP\_Diet’ grant.

There were 12 raw data files which were subjected to the following changes via a loop in the above script:

1. Each data file was subset to the first 87 variables which deals with monthly consumption expenditure for food and edible substances.
2. All variable names were lowercased.
3. The column ‘month’ was a string with first 3 alphabets for month and year put together, eg: “Aug 2021”. This column was split into month and year and both the resultant columns were retained as strings.
4. The data was further subset to remove all household observations which had “Non Response” under the column ‘response\_status’. This removes all observation with value “-99” for all varibales.
5. Further, the following variables, namely ‘response\_status’, ‘reason\_for\_non\_response’, ‘hr’, ‘psu\_id’, ‘stratum’ and ‘month\_slot’ were removed.
6. Certain numerical columns had value zero which were replaced with NaN.
7. State names were corrected to remove special characters from strings.
8. Checks for uniqueness of ‘hh\_id’ was conducted prior to append for all data sets.

In the final appended long data, each observation can be uniquely identified using a combination of ‘hh\_id’, ‘month’ and ‘year’.