Task:-

1. code exclusively in either python or R for each question. Make sure to write clean, clear code and comment properly.

**Done in python**

1. Include any explanations for critical decisions made in the code in this document.
2. **Rural dataset is bigger than urban one by 4 times.**
3. **Big data so little DSA concept is enforced.**

Libraries used:-

**pyreadstat :- .dta file to dataframe conversion**

**pandas :- for dataframe work**

Questions:

1. Read ec05st02.txt file based on the data dictionary in Documentation/Record layout.doc. Make sure to read correctly based on rural/urban status. Let’s call this the EC datafile.
   1. Were there any rows that you could not read correctly?

**Yes there were many such line having spaces “ ” i.e. no record. Included all of them in:-**

**“urban file containing error.txt”**

**“rural file containing error.txt”**

* 1. Code non-numeric variables correctly. For example, factor variables are not numeric variables.

**Okay\*\***

1. We will now add a geographic identifier for a village/town (or groups of villages/towns) that can be used to include Economic Census data from other years on the same geographical unit. This identifier is called “shrid” and is available from the Development Data Lab 1 . You do not need to download these datafiles as they have been placed in the folder already.
   1. In the EC datafile you read in part 1, which group of variables identify a village? Which variables identify a town?

**Village:-**

left=["STATE-UT code","Village code"]

right=["ec05\_state\_id","ec05\_village\_id"]

**Town:-**

left=["STATE-UT code","DISTRICT code","Town code"]

right=["ec05\_state\_id","ec05\_district\_id","ec05\_town\_id"]

* 1. Use these variables to merge the EC data file with shrug\_ec05\_rural and shrug\_ec05\_urban data files.

See file named **“rural final merged shrid.csv”** and **“urban final merged shrid.csv”**

Link:- <https://drive.google.com/drive/folders/1CvnZouT81ET0F_Do2XvsfKcbN2t4M1yM?usp=sharing>

\*\* **Note:-** Open the .csv file with notepad instead of excel to analyze the correct data type.