HCES
DATA STORY

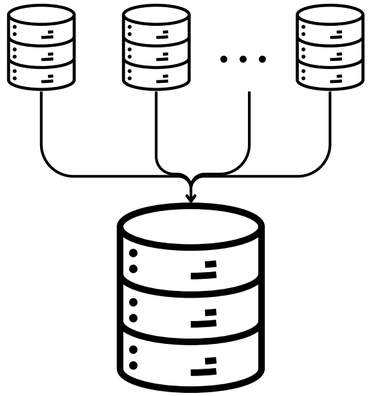
AGENDA

01

How data was collected?

02

How to merge?



03

Problems & Questions

How data was collected?

1

Household Consumption Expenditure Survey: 2023-24

GOVERNMENT

Go & conduct a survey on
the expenditure of Indian households.

3

- India is so big.
- We use the sample.
- In statistics, the sample means randomness

Household

2

- FSU (First Sample Unit)
- Household
- Item Code

1. Overall, this survey report is at the household level, not the person level.
2. FSU, Household, Item Code, we can divide the whole dataset.
3. All the identifiers used in each level are auto-generated.

Survey Hierarchy

4 Understanding the Hierarchy

Survey Identification

↓

Geographic Hierarchy

↓

Sampling Design Variables

↓

First Stage Sampling Unit (FSU)

↓

Second Stage Sampling (Household Level)

↓

Person-Level Identification

Survey Name, Year, Level

State, NSS Region, District, FOD Sub Region

Sector, Stratum, Sub Stratum, Panel, Sub Sample

FSU Serial No, Sample, SU no, Sample Sub Division No

Second Stage Stratum No, Sample Household No, Question No

5 Most of the datasets are HH Level

Person Serial No, Relation to Head

6 Level 02 Data is Person data

Survey Identification

Column	Category
Survey_Name	Survey Metadata
Year	Time
Level	Survey Level Identifier

Geographic Hierarchy

Column	Category
State	Geography
NSS_Region	Geography
District	Geography
FOD_Sub_Region	Geography

Sampling Design Variables

Column	Category
Sector	Sampling Design
Stratum	Sampling Design
Sub_stratum	Sampling Design
Panel	Sampling Design
Sub_sample	Sampling Design

First Stage Sampling Unit (FSU)

Column	Category
FSU_Serial_No	FSU Identifier
Sample_SU_No	FSU Identifier
Sample_Sub_Division_No	FSU Identifier

Second Stage Sampling (Household Level)

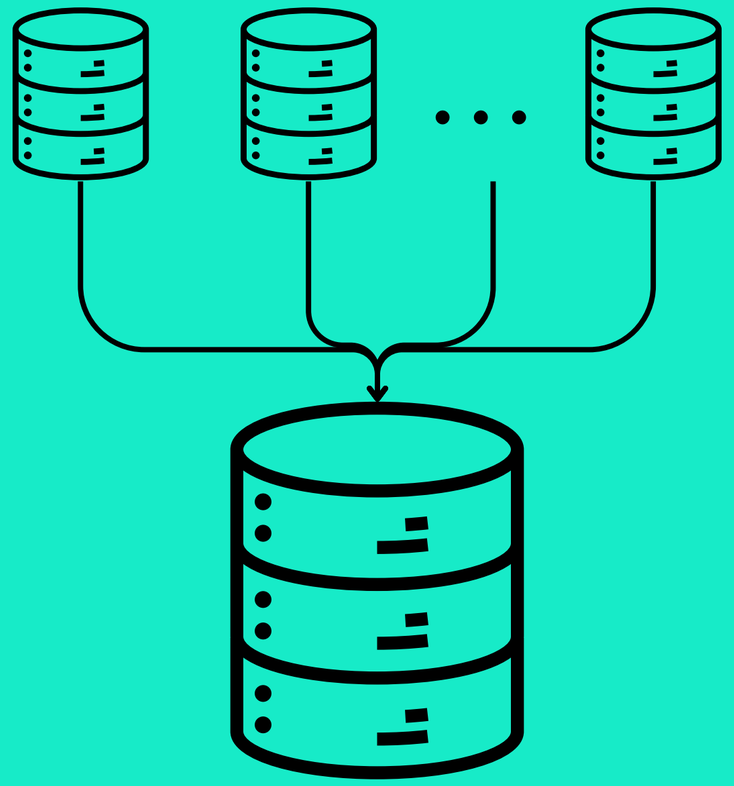
Column	Category
Second_Stage_Stratum_No	Household Sampling
Sample_Household_No	Household Identifier
Questionnaire_No	Household Identifier

Person-Level Identification

Column	Category
Person_Serial_No	Person Identifier
Relation_to_Head	Demographic Attribute

SECTION [1]: Identification of sample FSU

How to merge?



Bihar Dataset	Total Rows	Logic	FSU 27000
BL01	17,255	959 FSU * 18 HH = 17,262	18
BL02	85,751	FSU * HH * Person Serial No	86
BL03	17,255	959 FSU * 18 HH = 17,262	18
BL04	17,255	959 FSU * 18 HH = 17,262	18
BL05	841,515	FSU * HH * Item Code	1042
BL06	113,362		145
BL07	17,255	959 FSU * 18 HH = 17,262	18
BL08	97,939		93
BL09	542,787	FSU * HH * Item Code(*)	630
BL10	63,517		38
BL11	17,255	959 FSU * 18 HH = 17,262	18
BL12	324,076		428
BL13	358,548		418
BL14			
BL15	69,020	FSU * HH * Questionnaire No (4) = 69048	72

I experimented with the BL01, 02, 03, 05, 09, 15

https://github.com/Rudra-G-23/rural-financial-inclusion-govt-scheme-recommendation/tree/main/notebooks/merging-dataset/bihar-data

7 These survey datasets are divided into four parts

Category	Datasets
Person Level	L2
Household Level	L01, 03, 04, 07, 11
Item code Level	L05, 06, 08, 09, 10, 12, 13, 14
Summary	L01, 15

For the Bihar

- We got FSU(First sample unit) auto generated = 959
- Household = 18

In FSU == 27000

- Same Household 18

8

- FSU is telling us the total unique Sample
- HH telling us the total unique sample present in that area

https://github.com/Rudra-G-23/rural-financial-inclusion-govt-scheme-recommendation/blob/main/notebooks/merging-dataset/bihar-data/03-bihar-27000-data.ipynb

Housedhold Level Merging

L01, 03, 04, 07, 11

combine_cols = ['FSU_Serial_No', 'Sector', 'State', 'NSS_Region', 'District', 'Stratum', 'Sub_stratum', 'Panel', 'Sub_sample', 'FOD_Sub_Region', 'Sample_SU_No', 'Sample_Sub_Division_No', 'Second_Stage_Stratum_No', 'Sample_Household_No',]

HH Unique ID

43423110101192327621011303 2
43423110101192327621011303 3
43423110101192327621011303 4
43423110101192327621011303 5
43423110101192327621011303 6

L03

hh_unique_key	Max_Income_Activity	Total_Area_Land_Owned_Acres
27000_01	10	1.2
27000_02	20	1.5
27000_03	30	2
27000_04	40	2.5

L04

hh_unique_key	Ration_Rice	Ration_Wheat
27000_01	10	1.2
27000_02	20	1.5
27000_03	30	2
27000_04	40	2.5

9 Due to this is Household level data
This merges easily without any issues

	L03		L04	
hh_unique_key	Max_Income_Activity	Total_Area_Land_Owned_Acres	Ration Rice	Ration Wheat
27000_01	10	1.2	10	1.2
27000_02	20	1.5	20	1.5
27000_03	30	2	30	2
27000_04	40	2.5	40	2.5

https://github.com/Rudra-G-23/rural-financial-inclusion-govt-scheme-recommendation/blob/main/notebooks/merging-dataset/bihar-data/04-bl-13-hh-merge.ipynb

Item Code Levels Merging

L05, 06, 08, 09, 10, 12, 13, 14

HH	Item_code_key	Item	Out of Home qty	Out of Home
1	27000_01_01	Rice	10	100
1	27000_01_02	Wheat	20	200
2	27000_01_03	Gas	30	300
2	27000_01_04	Oil	40	400

HH	Item_code_key	Item	Total qty	Total Value
1	27000_01_01	Rice	1	10
1	27000_01_06	Sweeper	2	20
2	27000_01_07	Apple	3	30
2	27000_01_08	Tv	4	40

1st Way of Merging

hh	Item_code_key	Item	Out of Home qty	Out of Home Value	Total qty	Total Value
1	27000_01_01	Rice	10	100	1	10
1	27000_01_02	Wheat	20	200	Null	Null

10

- Remove the Duplicate value.
- We lost 99%+ data
- Because very, very few items are asking again during the survey.

https://github.com/Rudra-G-23/rural-financial-inclusion-govt-scheme-recommendation/blob/main/notebooks/merging-dataset/bihar-data/05-bl-59-item-merge.ipynb

2nd Way of Merging

hh	Item_code_key	Item	Out of Home qty	Out of Home Value	Total qty	Total Value
1	27000_01_01	Rice	10	100	Null	Null
1	27000_01_02	Wheat	20	200	Null	Null
2	27000_01_03	Gas	30	300	Null	Null
2	27000_01_04	Oil	40	400	Null	Null
1	27000_01_01	Rice	Null	Null	1	10
1	27000_01_06	Sweeper	Null	Null	2	20
2	27000_01_07	Apple	Null	Null	3	30
2	27000_01_08	Tv	Null	Null	4	40

- Improve the Duplicate.
- Stack the datasets one on another.

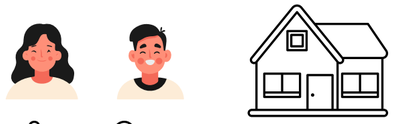
- What about the NULL Value?
- How does our model perform well with null values or duplicates?

https://github.com/Rudra-G-23/rural-financial-inclusion-govt-scheme-recommendation/tree/main/parquet-data/all-level/item-code

Problem & Questions

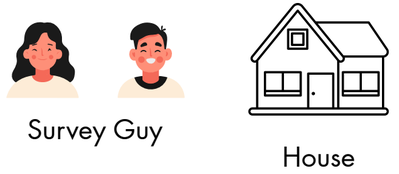
How the data look.

Household Level



Personal Data

Item Data



12

- The dataset is different.
- Categorise into 3 types.

- Merging all the data is not an easy task.
- Datasets already contain null value.

HH	Person_Serial_No
1	1
	2
	3
2	1
	2
	3
	4

HH	Item Code	Qantity	Price
1	2	-	.
	22	-	.
	222	-	.
2	5	-	.
	55	-	.
	555	-	.

What is our END Goal?

Thank You

Rudra Prasad Bhuyan