# Data Preparation Guide for HCES Microdata:

## 1. Data Source

- Portal: Download the raw microdata from the NSO’s Microdata Portal:  
 <https://microdata.gov.in/NADA/index.php/home>  
- Survey: Household Consumption Expenditure Survey (HCES) 2022–23.  
- Coverage: 261,746 households (155,014 rural; 106,732 urban) across 8,723 villages and 6,115 urban blocks.  
- Format: Multiple fixed-width text files (“Levels”), plus a metadata layout file that defines column names, start positions, and field widths.

- For more details to download microdata – refer to this link below:  
<https://mospi.gov.in/sites/default/files/main_menu/Guide_to_Download_Data/Guide_to_download_microdata_v2.pdf>

## 2. Metadata Layout

1. Obtain the layout file (often a CSV or TXT) from the portal—this describes, for each Level, the names of variables and their character positions.  
2. Review the schema to understand which Levels correspond to household identifiers, member demographics, asset details, item-level expenditures, etc.

## 3. Parsing Fixed‑Width Files

For each survey Level (e.g., Level 1: household identifiers; Level 3: household characteristics; Level 14: item expenditures):  
1. Identify the subset of layout entries belonging to that Level.  
2. Use the layout to split each fixed‑width text file into columns (according to start position and width).  
3. Convert each parsed file into a tabular format (e.g., CSV or Parquet) for faster downstream use.

## 4. Creating a Unique Household ID

- Select all identification fields (e.g., State code, District code, FSU serial, Sub‑FSU serial, etc.).  
- Concatenate their values (with a delimiter) to form a single HH\_ID.  
- This HH\_ID will serve as the primary key for merging data across Levels.

## 5. Merging Survey Schedules

1. Start with the base household table (Level 1).  
2. Left‑join additional Levels in logical order:  
 - Level 3 (household characteristics)  
 - Level 5 (member demographics)  
 - Level 11 (asset ownership)  
3. Ensure all merges use the same HH\_ID, so every household’s information is consolidated into one master table.

## 6. Calculating Monthly Consumption Expenditure

HCES uses a Modified Mixed Reference Period (MMRP):  
1. Group item‑level expenditures (Level 14) by HH\_ID and recall period (7‑day perishables; 30‑day other items; 365‑day durables).  
2. Sum expenditures within each recall category.  
3. Normalize each sum to a 30‑day equivalent (e.g., multiply 7‑day totals by 30/7; 365‑day totals by 30/365).  
4. Aggregate the normalized sums to compute total monthly consumption for each household.  
5. Merge this MPCE value back into the master household table by HH\_ID.

## 7. Splitting into Training and Test Sets

- Shuffle the list of unique HH\_IDs.  
- Assign 80 % of households to the training set and 20 % to the hold‑out test set.  
- Tag each record in the master table as “train” or “test” based on its HH\_ID to prevent data leakage.

## 8. Ready for Analysis

At this point, you have:  
- A single, consolidated dataset with household identifiers, demographics, assets, and computed MPCE  
- A clear train/test partition at the household level  
- Data stored in a high‑performance format (e.g., Parquet) for rapid loading  
  
This dataset is now ready for feature engineering, model building, validation, and deployment (e.g., into the interactive UI with GIS maps).