28th September

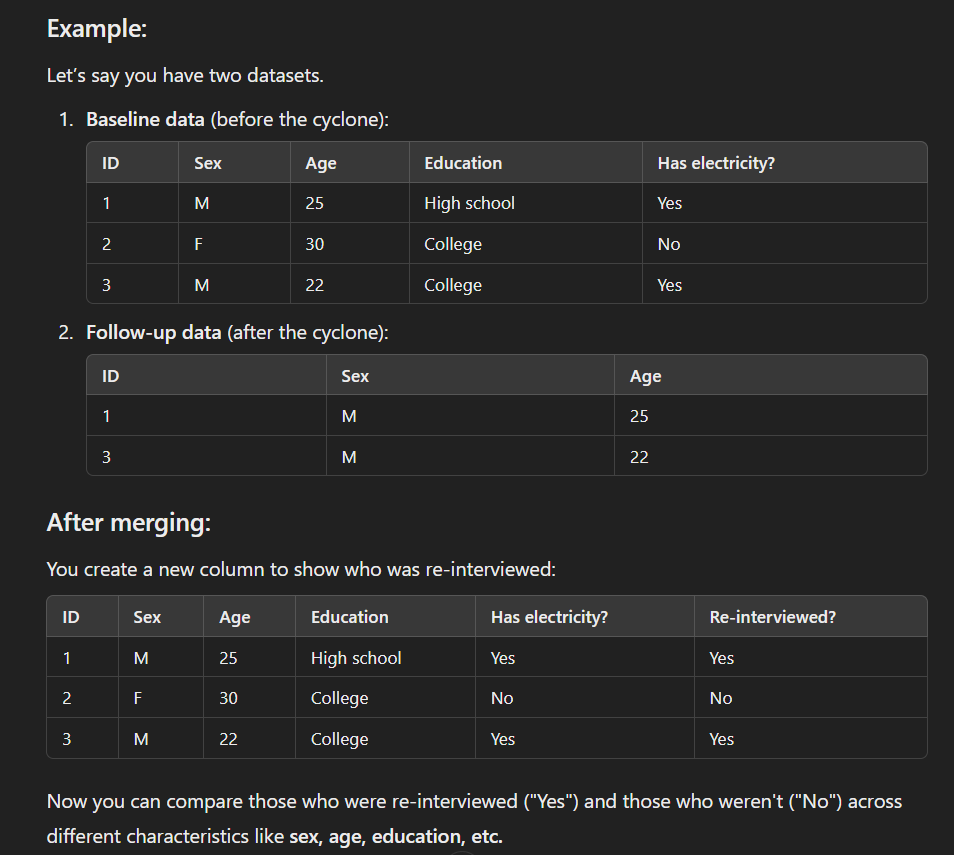
* Datasets are in Stata format
* “pre\_cyclone\_baseline.dta” contains information on individuals surveyed prior to cyclone Freddy, which took place in Feb/Mar 2023
  + Each row is an individual phone user, whereas each column gives information about this user (e.g., their sex, age etc...)
  + All columns have a name with the "pre\_" prefix
  + In total, the number of individuals in this data set is 8,909
  + Data is labelled; variables have labels, and most values have a label
* “post\_cyclone\_followup.dta” contains information on the subset of individuals in the “pre\_cyclone\_baseline” data who were re-interviewed after the cyclone
  + Collected using the follow-up questionnaire that is in the shared folder
  + The variable names match the question numbers in the questionnaire
  + In total, this data set has 5,218 individuals
  + It can be merged with the pre\_cyclone\_baseline data set using the "caseid" variable. This variable uniquely identifies an individual in both datasets
* Haven package can be used to open Stata data
* Labelled package allows use of the metadata that Stata attaches to variables (e.g. value labels)
* In simple terms, your professor wants you to compare two groups of people:
  + Those who were interviewed again after the cyclone.
  + Those who were not interviewed again after the cyclone.
* Steps:
  + Merge the data: You have two sets of data. The first one has information about everyone interviewed before the cyclone, and the second one has information about a smaller group of people who were interviewed again after the cyclone. You need to combine these two datasets to figure out who was interviewed both times and who was only interviewed once
  + Create a new variable: Once you merge the data, you’ll make a new column (or variable) that indicates whether a person was re-interviewed or not. For example, if someone was interviewed both times, you might mark them with a "Yes," and if they were only interviewed once, you might mark them with a "No."
  + Compare key characteristics: After that, you’ll look at how these two groups (re-interviewed vs. not re-interviewed) differ based on important factors like age, sex, education level, and household conditions (like if they have electricity or water).

str(pre\_cyclone):

* 8,909 individuals (rows), 23 columns (variables)

**Variables:**

1. **caseid**:
   * **Type**: Character (chr)
   * **Description**: Unique ID number for each individual in the dataset.
   * **Example values**: "A10", "A10001", etc.
2. **pre\_outcome**:
   * **Type**: Character (chr)
   * **Description**: The outcome of the interview, likely indicating whether it was completed.
   * **Example values**: "Completed interview".
3. **pre\_attempts**:
   * **Type**: Character (chr)
   * **Description**: Number of call attempts made to the individual.
   * **Example values**: "1", "2", etc.
4. **pre\_enumerator\_language**:
   * **Type**: Character (chr)
   * **Description**: Languages spoken by the interviewer.
   * **Example values**: "English, Chichewa", "English, Chichewa, Chitumbuka".
5. **pre\_language**:
   * **Type**: Character (chr)
   * **Description**: Language of the interview.
   * **Example values**: "Chichewa".
6. **pre\_district**:
   * **Type**: Character (chr)
   * **Description**: District of residence for the individual.
   * **Example values**: "Mangochi", "Blantyre City", etc.
7. **pre\_Region**:
   * **Type**: Numeric with labels (dbl+lbl)
   * **Description**: Region of residence, with numeric codes and labels attached to identify the specific region.
   * **Values and Labels**:
     + 1 = "Northern"
     + 2 = "Central"
     + 3 = "Southern"
8. **pre\_age**:
   * **Type**: Numeric (num)
   * **Description**: Age of the individual in completed years.
   * **Example values**: 21, 20, 26, etc.
9. **pre\_Education**:
   * **Type**: Numeric with labels (dbl+lbl)
   * **Description**: Educational level of the respondent.
   * **Values and Labels**:
     + 1 = "None/primary school"
     + 2 = "Secondary school"
     + 3 = "Higher education"
10. **pre\_sex**:
    * **Type**: Numeric with labels (dbl+lbl)
    * **Description**: Sex of the respondent.
    * **Values and Labels**:
      + 1 = "Male"
      + 2 = "Female"
11. **pre\_kinship**:
    * **Type**: Character (chr)
    * **Description**: Relationship to the head of household.
    * **Example values**: "SON/DAUGHTER-IN-LAW", "OTHER RELATIVE", "HEAD", etc.
12. **pre\_UR**:
    * **Type**: Numeric with labels (dbl+lbl)
    * **Description**: Whether the place of residence is urban or rural.
    * **Values and Labels**:
      + 1 = "Urban"
      + 2 = "Rural"
13. **pre\_Electricity**:
    * **Type**: Numeric with labels (dbl+lbl)
    * **Description**: Whether the household has electricity.
    * **Values and Labels**:
      + 1 = "Yes"
      + 2 = "No"
14. **pre\_Roofing**:
    * **Type**: Numeric with labels (dbl+lbl)
    * **Description**: Type of roofing material used in the household.
    * **Values and Labels**:
      + 1 = "Thatch"
      + 2 = "Improved roof material"
15. **pre\_Water**:
    * **Type**: Numeric with labels (dbl+lbl)
    * **Description**: Source of water for the household.
    * **Values and Labels**:
      + 1 = "Unprotected source"
      + 2 = "Protected source"
16. **pre\_SVY\_duration**:
    * **Type**: Numeric (num)
    * **Description**: Duration of the interview, in seconds.
    * **Example values**: 1758, 1290, etc.
17. **pre\_call\_duration**:
    * **Type**: Numeric (num)
    * **Description**: Duration of the phone call, in seconds.
    * **Example values**: 1451, 1213, etc.
18. **pre\_household**:
    * **Type**: Numeric with labels (dbl+lbl)
    * **Description**: Number of household members.
    * **Values and Labels**:
      + 1 = "0-4 members"
      + 2 = "5-8 members"
      + 3 = "9+ members"
19. **pre\_GO**:
    * **Type**: Numeric with labels (dbl+lbl)
    * **Description**: Age group of the respondent.
    * **Values and Labels**:
      + 1 = "18-29"
      + 2 = "30-39"
      + 3 = "40-49"
      + 4 = "50-64"
20. **pre\_interview**:
    * **Type**: Date
    * **Description**: The date when the interview was conducted.
    * **Example values**: "2022-01-31", "2022-01-25", etc.
21. **pre\_group**:
    * **Type**: Numeric (num)
    * **Description**: Identifies the RaMMPS round (i.e., which survey round this interview belongs to).
    * **Example values**: 1, 1, 1, etc.
22. **pre\_WR**:
    * **Type**: Numeric (num)
    * **Description**: Sampling weight for the individual. This helps adjust the analysis to ensure the sample is representative of the population.
    * **Example values**: 0.127, NA, etc.
23. **pre\_WT**:
    * **Type**: Numeric (num)
    * **Description**: Another sampling weight variable to adjust the analysis.
    * **Example values**: 0.103, 0.284, etc.

* Several variables are categorical but have been encoded as numeric with labels (like pre\_sex, pre\_Electricity, etc.). These labels need to be interpreted properly when analyzing the data
* Some variables, such as sampling weights (pre\_WR, pre\_WT), are used to ensure the data is representative of the population.
  + For example, if one person represents 100 people in the population, their responses will count more in your analysis
* 

Str(post\_cyclone)

* 5218 rows (survey responses) and 584 columns (variables/questions)
* Each column is followed by information about its **type** (e.g., character strings chr, numbers num, or date-time POSIXct) and examples of the data values stored in it.
* Some columns have **attributes** such as labels (label), formats (format.stata), or specific **named levels** (e.g., 1 = YES, 2 = NO).

 The data includes:

* **Timestamps** for when the survey started (starttime), ended (endtime), and was submitted (submissiondate).
* **Interviewer** details (Interviewer), along with variables like the **duration** of the interview (duration), **call number** (call\_num), and **call status** (last\_call\_status).
* Questions and responses related to **personal information** (e.g., l5 asks if the respondent has previously been interviewed) and demographic details like **age** (e3), **gender** (e2), and **marital status** (b1).
* Questions about **housing status**, **damage from the cyclone** (a3a, a3b), and **location** both pre- and post-cyclone (region, a2\_region).
* It also includes responses related to **post-cyclone impacts** and reasons for moving (a6), **sources of drinking water** (b8), and **electricity access** (b7).