Introductory Stata Training
Lesson 2 – Transforming Data

Harry Gimiseve, Rishabh Mukerjee, Emily Schmidt

Port Moresby, Papua New Guinea

March 2025

International Food Policy Research Institute (IFPRI)

Lesson 2 Agenda

1. Review of Lesson 1

Lesson 2 Agenda

1. Review of Lesson 1

2. Transforming existing data

Lesson 2 Agenda

1. Review of Lesson 1
2. Transforming existing data
3. Creating new variables

Lesson 2 Agenda

1. Review of Lesson 1
2. Transforming existing data
3. Creating new variables
4. Help window

Lesson 2 Agenda

1. Review of Lesson 1
2. Transforming existing data
3. Creating new variables
4. Help window
5. Collapse and merge

1. Review of Lesson 1

- Today we are going to work mostly with the household roster dataset.

1. Review of Lesson 1

- Today we are going to work mostly with the household roster dataset.
- First, let me show you the household roster module in the questionnaire (Module 1.1)

Module 1 – Household characteristics

Section 1.1: Household demographics, current household members

We would like to record all of the individuals that are members of this household. **A household member is "a part of a group of people who live together, may pool some of their money, and eat most meals from the same pot".** A member can also be someone who lives most of the time (at least 8 months of the year) and shares meals with the family, but is currently away (e.g., for work, ceremony, travel, etc.) It is possible that there are individuals who are household members that are not related by blood or family ties, but they are still a part of the household because they share in the household resources (such as a servant that lives in the household and eats from the same food stock).

Person ID Code	1 Please [NAME] everyone in your household starting with yourself.	2 What is the relation-ship of [NAME] to the head? **[code: rel_vl]**	3 What is the sex of [NAME] ? **[code: gender _vl]**	4a What is the age of [NAME]? [Mark 0 if [NAME] is less than 1 year.] *If age is <= 5 → skip to Q5* *If age is > 5 and < 12 → skip to Q10* *If age ≥ 12 → skip to Q7* If age is 'Don't know' please write 888 → answer 4b	4b. I understand that you do not know the age of [NAME], but do you think you could give me an estimate of the age within these categories? **[code: agecategory_vl]**	4c – 4e When was [NAME] born?			5a. For children <5: Is the **biological** mother of the child a current household member? If no →skip to 5c	5b. For children <5: Who is the mother? [PID]	5c. For children <5: Is the father of the child a current household member? If no →skip to 6	5d. For children <5: Who is the father? [PID]	6. Do you have a clinic card / book for this child? **[code: yesno_dk_vl]** [ENUMERATOR: Please ask to see the immunization card, you will need it for future modules.]
						4c. Day If age is 'Don't know' please write 888	4d. Month **[code: months _vl]**	4e. Year					
1													
2													
3													
4													
5													
6													
7													

1. Review of Lesson 1

- Today we are going to work mostly with the household roster dataset.
- First, let's turn to the household roster module in the questionnaire (Module 1.1 – page 5)
- This module asks questions about each person in the household.
- Earlier, the data was at the *household level*

1. Review of Lesson 1

- Earlier, the data was at the *household level*, and now it is at the *person level*

1. Review of Lesson 1

- Earlier, the data was at the *household level*, and now it is at the *person level*

- Do you think we will have more or less observations in our dataset today, compared to previous lesson?

1. Review of Lesson 1

- Yesterday, the data was at the *household level*, and today it is at the *person level*
- Do you think we will have more or less observations in our dataset today, compared to previous lesson?
- How can we find out how many observations are in the new dataset?

1. Review of Lesson 1

- Yesterday, the data was at the *household level*, and today it is at the *person level*
- Do you think we will have more or less observations in our dataset today, compared to yesterday?
- How can we find out how many observations are in the new dataset? *Count*

1. Review of Lesson 1

- Let's double-click to open *s11_long.dta*

1. Review of Lesson 1

- Let’s double-click to open *s11_long.dta*
- How many observations are there?

1. Review of Lesson 1

- Let's double-click to open *s11_long.dta*
- How many observations are there? *Count*

1. Review of Lesson 1

- Let's double-click to open *s11_long.dta*
- How many observations are there? *Count*

```
. count
  16,192
```

- *16,192 observations (or 16,192 people)*

1. Review of Lesson 1

- The hh_3_4.dta dataset, contained 2,699 household observations

1. Review of Lesson 1

- The hh_3_4.dta dataset, contained 2,699 household observations

- Now, the household roster dataset, there are 16,192 person observations (*person-level*)

1. Review of Lesson 1

- The hh_3_4.dta dataset, contained 2,699 household observations
- Now, the household roster dataset, there are 16,192 person observations (person-level)
- So there are many more observations in this dataset than in the previous data

1. Review of Lesson 1

- Let’s take a deeper look at some of the variables in this new dataset

1. Review of Lesson 1

- Let’s take a deeper look at some of the variables in this new dataset
- Question: How many men are in this dataset?

1. Review of Lesson 1

- Let’s take a deeper look at some of the variables in this new dataset
- Question: How many men are in this dataset?
- Looking at Window 3 (Variables Window), which variable can tell us this information?

1. Review of Lesson 1

- Let's take a deeper look at some of the variables in this new dataset
- Question: How many men are in this dataset?
- Looking at Window 3 (Variables Window), which variable can tell us this information? *a3 – "3. What is the sex of ${a1}?"*

1. Review of Lesson 1

- Let's take a deeper look at some of the variables in this new dataset
- Question: How many men are in this dataset?
- Looking at Window 3 (Variables Window), which variable can tell us this information? *a3 – "3. What is the sex of a1?"*
- a1 was the name of the HH member, so a3 is asking their sex

1. Review of Lesson 1

- Question: How many men are in this dataset?
- What code can we use on a3 to see how many men are in this dataset?

1. Review of Lesson 1

- Question: How many men are in this dataset?

- What code can we use on a3 to see how many men are in this dataset? *tab a3*

1. Review of Lesson 1

- <u>Question:</u> How many men are in this dataset?

- What code can we use on a3 to see how many men are in this dataset? *tab a3*
  - *Codebook a3* could also work, but tabulate is the best option because it will also tell us the percentage in addition to the frequency

1. Review of Lesson 1

- Question: How many men are in this dataset?

1. Review of Lesson 1

- Question: How many men are in this dataset?

```
. tab a3

 3. What is |
 the sex of |
          ? |      Freq.     Percent        Cum.
------------+-----------------------------------
       Male |      8,459       52.24       52.24
     Female |      7,733       47.76      100.00
------------+-----------------------------------
      Total |     16,192      100.00
```

1. Review of Lesson 1

- Question: How many men are in this dataset?

- Answer: *8,459*

```
. tab a3

 3. What is |
 the sex of |
          ? |      Freq.     Percent        Cum.
------------+-----------------------------------
       Male |      8,459       52.24       52.24
     Female |      7,733       47.76      100.00
------------+-----------------------------------
      Total |     16,192      100.00
```

1. Review of Lesson 1

- Question: How many men are in this dataset?
- Answer: *8,459*
- Question: What percent of this dataset is male?

```
. tab a3

 3. What is |
 the sex of |
          ? |      Freq.     Percent        Cum.
------------+-----------------------------------
       Male |      8,459       52.24       52.24
     Female |      7,733       47.76      100.00
------------+-----------------------------------
      Total |     16,192      100.00
```

1. Review of Lesson 1

- Question: How many men are in this dataset?
- Answer: *8,459*
- Question: What percent of this dataset is male? 52.24%

```
. tab a3
```

3. What is the sex of ?	Freq.	Percent	Cum.
Male	8,459	52.24	52.24
Female	7,733	47.76	100.00
Total	16,192	100.00	

1. Review of Lesson 1

- Let’s look at another variable in this new dataset.
- <u>Question:</u> What is the average age in this dataset?

1. Review of Lesson 1

- Let’s look at another variable in this new dataset.
- Question: What is the average age in this dataset?
- Looking at Window 3 (Variables Window), which variable can tell us this information?

1. Review of Lesson 1

- Let’s look at another variable in this new dataset.
- Question: What is the average age in this dataset?
- Looking at Window 3 (Variables Window), which variable can tell us this information? *a4a – “4. What is the age of ${a1}?”*

1. Review of Lesson 1

- Question: What is the average age in this dataset?
- What code can we use on a4a to see what the average age is?

1. Review of Lesson 1

- Question: What is the average age in this dataset?

- What code can we use on a4a to see what the average age is? *sum a4a*
  - Can also use *sum a4a, det* or *codebook a4a*

1. Review of Lesson 1

- Question: What is the average age in this dataset?

- What code can we use on a4a to see what the average age is? *sum a4a*
  - Can also use *sum a4a, det* or *codebook a4a*

```
. sum a4a

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         a4a |     16,192     125.095    279.0221          0        888
```

1. Review of Lesson 1

- <u>Question:</u> What is the average age in this dataset?
- What code can we use on a4a to see what the average age is? *sum a4a*
  - Can also use *sum a4a, det* or *codebook a4a*

```
. sum a4a

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         a4a |     16,192     125.095    279.0221          0        888
```

- What's wrong with this output?

1. Review of Lesson 1

- Question: What is the average age in this dataset?
- What code can we use on a4a to see what the average age is? *sum a4a*
  - Can also use *sum a4a, det* or *codebook a4a*

```
. sum a4a

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         a4a |     16,192     125.095    279.0221          0        888
```

- What's wrong with this output? *Average age appears to be too high!*

1. Review of Lesson 1

- What's wrong with this output? *Average age appears to be too high!*

- Why is the average age too high?

```
. sum a4a

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         a4a |     16,192     125.095    279.0221          0        888
```

1. Review of Lesson 1

- What's wrong with this output? *Average age appears to be too high!*
- Why is the average age too high? *Maximum age is an impossibly large number, 888.*

```
. sum a4a

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         a4a |     16,192     125.095    279.0221          0        888
```

1. Review of Lesson 1

- What's wrong with this output? Average age appears to be too high!
- Why is the average age too high? Maximum age is an impossibly large number, 888.

```
. sum a4a

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         a4a |     16,192     125.095    279.0221          0        888
```

- If we look at the questionnaire, we see that data collectors were told to write 888 if age is 'don't know'

2. Transforming Existing Data

- Now, we will learn how we can transform/change the data, to find the real average age

2. Transforming Existing Data

- Now, we will learn how we can transform/change the data, to find the real average age

- But first, let's talk about ***missing values***

2. Transforming Existing Data
Missing Values

- Which command tells us how many observations have a missing value for a variable?

2. Transforming Existing Data
Missing Values

- Which command tells us how many observations have a missing value for a variable? *Codebook*

2. Transforming Existing Data
Missing Values

- So let’s see how many missing observations there are in our age variable. What do we type?

2. Transforming Existing Data
Missing Values

- So let's see how many missing observations there are in our age variable. What do we type? *Codebook a4a*

2. Transforming Existing Data
Missing Values

- So let's see how many missing observations there are in our age variable. What do we type? *Codebook a4a*

- How many missing observations are there?

```
. codebook a4a

-------------------------------------------------------------------------------
a4a                                                        4a. What is the age of
-------------------------------------------------------------------------------

                  Type: Numeric (int)

                 Range: [0,888]                     Units: 1
         Unique values: 97                      Missing .: 0/16,192

                  Mean: 125.095
             Std. dev.: 279.022

           Percentiles:     10%       25%       50%       75%       90%
                              5        10        22        45       888
```

2. Transforming Existing Data
Missing Values

- So let's see how many missing observations there are in our age variable. What do we type? *Codebook a4a*
- How many missing observations are there? *0*

```
. codebook a4a

a4a                                                        4a. What is the age of

                  Type: Numeric (int)

                 Range: [0,888]                    Units: 1
         Unique values: 97                     Missing .: 0/16,192

                  Mean: 125.095
             Std. dev.: 279.022

           Percentiles:     10%       25%       50%       75%       90%
                              5        10        22        45       888
```

2. Transforming Existing Data
Missing Values

```
. codebook a4a

a4a                                                  4a. What is the age of

                  Type: Numeric (int)

                 Range: [0,888]                    Units: 1
         Unique values: 97                     Missing .: 0/16,192

                  Mean: 125.095
              Std. dev.: 279.022

           Percentiles:     10%       25%       50%       75%       90%
                              5        10        22        45       888
```

- But we just said that there are missing observations, and they were input as 888

2. Transforming Existing Data
Missing Values

```
. codebook a4a

a4a                                                    4a. What is the age of

                  Type: Numeric (int)

                 Range: [0,888]                    Units: 1
         Unique values: 97                     Missing .: 0/16,192

                  Mean: 125.095
             Std. dev.: 279.022

           Percentiles:     10%       25%       50%       75%       90%
                              5        10        22        45       888
```

- But we just said that there are missing observations, and they were input as 888
- So, we need to change 888 so that they are recognized as missing.

2. Transforming Existing Data
Missing Values

- Missing values appear in 2 different ways in Stata, depending on the variable's data type:

2. Transforming Existing Data
Missing Values

- Missing values appear in 2 different ways in Stata, depending on the variable's data type:

  - Numeric/categorical variables
    - Missing values are coded as a period (.)
    - Code: replace [varname]=. if …

2. Transforming Existing Data
Missing Values

- Missing values appear in 2 different ways in Stata, depending on the variable's data type:
  - Numeric/categorical variables
    - Missing values are coded as a period (.)
    - Code: replace [varname]=. if …
  - String variables
    - Missing values are coded as a blank ("")
    - Code: replace [varname]="" if …

2. Transforming Existing Data
Missing Values

- We need to know what value to change the 888 to in the a4a variable, so Stata recognizes it as missing

2. Transforming Existing Data
Missing Values

- We need to know what value to change the -888 to in the a4a variable, so Stata recognizes it as missing

- What type of data is a4a? What code can we type to find out?

2. Transforming Existing Data
Missing Values

- We need to know what value to change the -888 to in the a4 variable, so Stata recognizes it as missing

- What type of data is a4a? What code can we type to find out? *Codebook a4a – a4a is numeric*

2. Transforming Existing Data
Missing Values

- We need to know what value to change the 888 to in the a4a variable, so Stata recognizes it as missing
- What type of data is a4a? What code can we type to find out? *Codebook a4a – a4a is numeric*
- So what type of missing value do we want to change the 888 to?

2. Transforming Existing Data
Missing Values

- We need to know what value to change the 888 to in the a4a variable, so Stata recognizes it as missing.

- What type of data is a4a? What code can we type to find out? *Codebook a4a – a4a is numeric*

- So what type of missing value do we want to change the 888 to? *A period (.)*

2. Transforming Existing Data
Recoding Values

- We need to change all 888 to . in variable a4a. There are 2 ways to do this:

2. Transforming Existing Data
Recoding Values

- We need to change all 888 to . in variable a4a. There are 2 ways to do this:

- Code: *recode a4a 888*=.

2. Transforming Existing Data
Recoding Values

- We need to change all 888 to . in variable a4a. There are 2 ways to do this:

- Code: *recode a4a 888=.*

- Code: *replace a4a=. if a4a==888*

2. Transforming Existing Data
Recoding Values

- We need to change all 888 to . in variable a4a. There are 2 ways to do this:

- Code: *recode a4a 888=.*

- Code: *replace a4a=. if a4==888*

- Choose one of these codes, and run it

2. Transforming Existing Data
Recoding Values

- Now, let's run the codebook command again for a4a and see if we succeeded in changing them to missing:

2. Transforming Existing Data
Recoding Values

- Now, let's run the codebook command again for a4a and see if we succeeded in changing them to missing:

```
. codebook a4a

a4a                                                        4a. What is the age of

                  Type: Numeric (int)

                 Range: [0,106]                      Units: 1
         Unique values: 96                       Missing .: 1,904/16,192

                  Mean: 23.4313
             Std. dev.: 18.0899

           Percentiles:     10%       25%       50%       75%       90%
                              4         9        18        34        51
```

2. Transforming Existing Data Recoding Values

- Now, let's run the codebook command again for a4a and see if we succeeded in changing them to missing:

```
. codebook a4a

a4a                                                               4a. What is the age of

                  Type: Numeric (int)

                 Range: [0,106]                       Units: 1
         Unique values: 96                        Missing .: 1,904/16,192

                  Mean: 23.4313
             Std. dev.: 18.0899

           Percentiles:     10%       25%       50%       75%       90%
                              4         9        18        34        51
```

- How many ages are unknown/missing?

2. Transforming Existing Data
Recoding Values

- Now, let's run the codebook command again for a4a and see if we succeeded in changing them to missing:

```
. codebook a4a

------------------------------------------------------------------------------------------
a4a                                                                    4a. What is the age of
------------------------------------------------------------------------------------------

                  Type: Numeric (int)

                 Range: [0,106]                    Units: 1
         Unique values: 96                     Missing .: 1,904/16,192

                  Mean: 23.4313
             Std. dev.: 18.0899

           Percentiles:     10%       25%       50%       75%       90%
                              4         9        18        34        51
```

- How many ages are unknown/missing? *1,904 (REMEMBER THIS, it will be used later)*

2. Transforming Existing Data
Recoding Values

- Another way to see how many missing values there are is to use the “missing” option on the tabulate command

2. Transforming Existing Data
Recoding Values

- Another way to see how many missing values there are is to use the “missing” option on the tabulate command

- *tab a4a, missing*

2. Transforming Existing Data
Recoding Values

- Another way to see how many missing values there are is to use the "missing" option on the tabulate command

- *tab a4a, missing*

- The beginning and end of the output is shown here (it's a very long output because there are so many different ages!)

```
. tab a4a, m

4a. What is |
 the age of |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        317        1.96        1.96
          1 |        255        1.57        3.53
          2 |        349        2.16        5.69
          3 |        311        1.92        7.61
          4 |        327        2.02        9.63
          5 |        260        1.61       11.23
```

```
         91 |          1        0.01       88.20
         92 |          1        0.01       88.21
         93 |          2        0.01       88.22
         96 |          1        0.01       88.23
        105 |          1        0.01       88.23
        106 |          1        0.01       88.24
          . |      1,904       11.76      100.00
------------+-----------------------------------
      Total |     16,192      100.00
```

2. Transforming Existing Data
Recoding Values

- Another way to see how many missing values there are is to use the "missing" option on the tabulate command

- *tab a4a, missing*

- The beginning and end of the output is shown here (it's a very long output because there are so many different ages!)

- What percent of ages are missing?

```
. tab a4a, m

4a. What is
the age of |      Freq.     Percent        Cum.
-----------+-----------------------------------
         0 |        317        1.96        1.96
         1 |        255        1.57        3.53
         2 |        349        2.16        5.69
         3 |        311        1.92        7.61
         4 |        327        2.02        9.63
         5 |        260        1.61       11.23
```

```
        91 |          1        0.01       88.20
        92 |          1        0.01       88.21
        93 |          2        0.01       88.22
        96 |          1        0.01       88.23
       105 |          1        0.01       88.23
       106 |          1        0.01       88.24
         . |      1,904       11.76      100.00
-----------+-----------------------------------
     Total |     16,192      100.00
```

2. Transforming Existing Data
Recoding Values

- Another way to see how many missing values there are is to use the "missing" option on the tabulate command

- *tab a4a, missing*

- The beginning and end of the output is shown here (it's a very long output because there are so many different ages!)

- What percent of ages are missing? 11.76%

```
. tab a4a, m
```

4a. What is the age of	Freq.	Percent	Cum.
0	317	1.96	1.96
1	255	1.57	3.53
2	349	2.16	5.69
3	311	1.92	7.61
4	327	2.02	9.63
5	260	1.61	11.23
91	1	0.01	88.20
92	1	0.01	88.21
93	2	0.01	88.22
96	1	0.01	88.23
105	1	0.01	88.23
106	1	0.01	88.24
.	1,904	11.76	100.00
Total	16,192	100.00	

2. Transforming Existing Data
Recoding Values

- Now we can find the average age in the dataset.

- What code do we use?

2. Transforming Existing Data
Recoding Values

- Now we can find the average age in the dataset.

- What code do we use? *sum a4a*

```
. sum a4a

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         a4a |     14,288    23.43127    18.08986          0        106
```

2. Transforming Existing Data
Recoding Values

- Now we can find the average age in the dataset.

- What code do we use? *sum a4a*

```
. sum a4a

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         a4a |     14,288    23.43127    18.08986          0        106
```

- What is the average age?

2. Transforming Existing Data Recoding Values

- Now we can find the average age in the dataset.
- What code do we use? *sum a4a*

```
. sum a4a

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         a4a |     14,288    23.43127    18.08986          0        106
```

- What is the average age? *23.43 years*

2. Transforming Existing Data
Recoding Values

- Now we can find the average age in the dataset.

- What code do we use? *sum a4a*

```
. sum a4a

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         a4a |     14,288    23.43127    18.08986          0        106
```

- How many observations are being included in this summarize command?

2. Transforming Existing Data
Recoding Values

- Now we can find the average age in the dataset.

- What code do we use? *sum a4a*

```
. sum a4a

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         a4a |     14,288    23.43127    18.08986          0        106
```

- How many observations are being included in this summarize command? *14,288*

2. Transforming Existing Data
Recoding Values

- Now we can find the average age in the dataset.
- What code do we use? *sum a4a*

```
. sum a4a
```

Variable	Obs	Mean	Std. dev.	Min	Max
a4a	14,288	23.43127	18.08986	0	106

- How many observations are being included in this summarize command? *14,288*
  - *This is because 1,904 are now missing (they are not included)*

2. Transforming Existing Data
Renaming variable and changing labels

- What if we want our age variable to have a more intuitive name and label?

2. Transforming Existing Data
Renaming variable and changing labels

- What if we want our age variable to have a more intuitive name and label?

- Renaming a variable
  - Code: *rename a4 age*

2. Transforming Existing Data
Renaming variable and changing labels

- What if we want our age variable to have a more intuitive name and label?

- Renaming a variable
  - Code: *rename a4 age*

- Relabeling a variable
  - Code: *label variable age "Age of household member"*

2. Transforming Existing Data
Renaming variable and changing labels

- What if we want our age variable to have a more intuitive name and label?

- Renaming a variable
  - Code: *rename a4 age*

- Relabeling a variable
  - Code: *label variable* ***age*** *“Age of household member”*
  - Remember, we just changed the name of our variable, so we have to refer to it here as “age” instead of “a4”

2. Transforming Existing Data
Renaming variable and changing labels

- Now let’s visualize our newly cleaned age variable

2. Transforming Existing Data
Renaming variable and changing labels

- Now let's visualize our newly cleaned age variable

- Do you remember the code we learned yesterday to show the density of a variable as bar graphs?

2. Transforming Existing Data
Renaming variable and changing labels

- Now let's visualize our newly cleaned age variable

- Do you remember the code we learned yesterday to show the density of a variable as bar graphs? *Histogram*

2. Transforming Existing Data
Renaming variable and changing labels

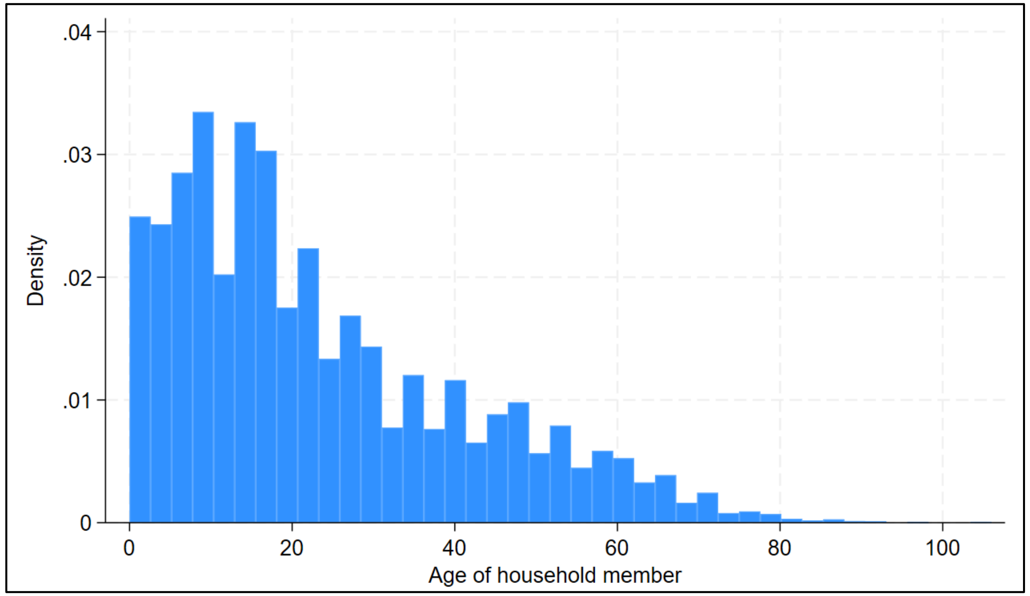
- Now let's visualize our newly cleaned age variable

- Do you remember the code we learned yesterday to show the density of a variable as bar graphs? *Histogram*

- Code: *hist age*

2. Transforming Existing Data
Renaming variable and changing labels

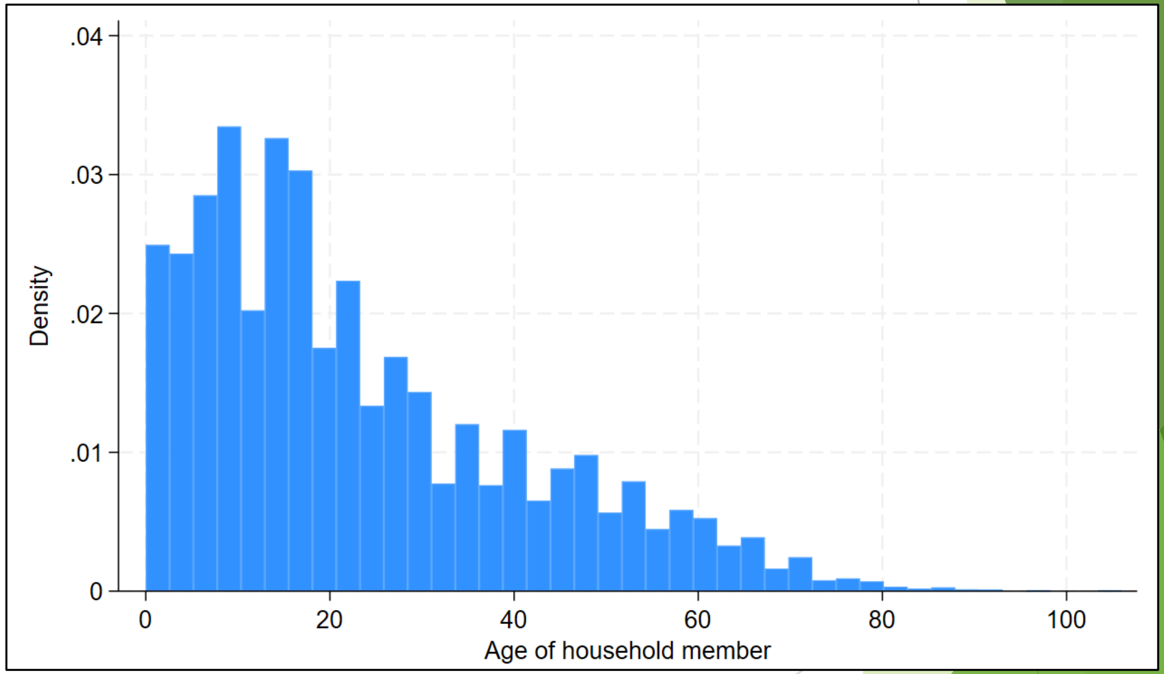
- *hist age*



2. Transforming Existing Data
Renaming variable and changing labels

- *hist age*

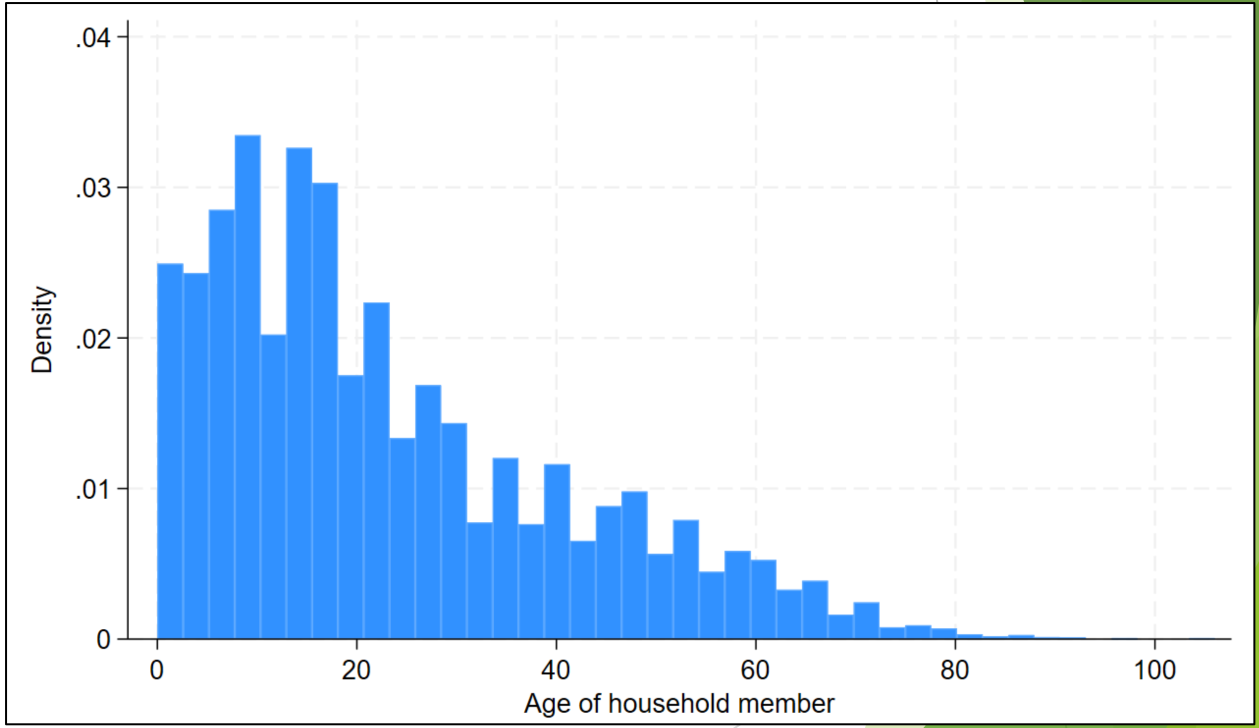
- What can we learn from this figure?



2. Transforming Existing Data
Renaming variable and changing labels

- *hist age*

- What can we learn from this figure?
  - *The majority of people in the dataset are 25 years old or younger*
  - *There are very few people in the dataset older than 80*
  - *The sample comprise of more young population*



3. Creating New Variables
Dummy Variables

- Now we are going to learn how to create a new variable

3. Creating New Variables Dummy Variables

- Now we are going to learn how to create a new variable
- The type of variable we are going to create is called a ***dummy variable***

3. Creating New Variables
Dummy Variables

- A dummy variable (also known as an indicator variable or a binary variable) takes the value **0 or 1** to indicate the absence or presence of some categorical effect

3. Creating New Variables Dummy Variables

- A dummy variable (also known as an indicator variable or a binary variable) takes the value 0 or 1 to indicate the absence or presence of some categorical effect

- For example, it may be useful to have a variable that easily indicates whether or not someone on the household roster is a child (15 years or younger)

3. Creating New Variables
Dummy Variables

- A dummy variable (also known as an indicator variable or a binary variable) takes the value 0 or 1 to indicate the absence or presence of some categorical effect

- For example, it may be useful to have a variable that easily indicates whether or not someone on the household roster is a child (15 years or younger)

- Examples of what this would let us easily see:
  - How many children are in the sample
  - What percent of children are in school

3. Creating New Variables
Generate

- Let’s make a dummy variable for whether or not someone is a child (15 years or younger)

3. Creating New Variables
Generate

- Let's make a dummy variable for whether or not someone is a child (15 years or younger)

- First, we will create the variable and set all values equal to missing (.)

3. Creating New Variables
Generate

- Let's make a dummy variable for whether or not someone is a child (15 years or younger)
- First, we will create the variable and set all values equal to missing (.)
- Code: *generate child*=.
  - Shorthand: *gen child*=.

3. Creating New Variables

- Let’s explore our new variable:

3. Creating New Variables

- Let’s explore our new variable: *codebook child*

3. Creating New Variables

- Let's explore our new variable: *codebook child*

```
. codebook child

child                                                              (unlabeled)

                  type:  numeric (float)

                 range:  [.,.]                        units:  .
         unique values:  0                        missing .:  16,192/16,192

            tabulation:  Freq.  Value
                        16,192  .
```

3. Creating New Variables

- Next, let's change all observations to 1 if the person is 15 years or younger.

3. Creating New Variables

- Next, let's change all observations to 1 if the person is 15 years or younger.
- What code did we learn to recode or change values?

3. Creating New Variables

- Next, let’s change all observations to 1 if the person is 15 years or younger.

- What code did we learn to recode or change values?
  replace…if

3. Creating New Variables

- Next, let's change all observations to 1 if the person is 15 years or younger.
- What code did we learn to recode or change values? *replace…if*
- Code: *replace child=1 if age<=15*
  - Also works: *replace child=1 if age<16*

3. Creating New Variables

- Let’s explore our new variable again: *codebook child*

3. Creating New Variables

- Let's explore our new variable again: *codebook child*

```
. codebook child

child                                                                  (unlabeled)

                  type:  numeric (float)

                 range:  [1,1]                        units:  1
         unique values:  1                        missing .:  10,132/16,192

            tabulation:  Freq.  Value
                         6,060  1
                        10,132  .
```

3. Creating New Variables

- Next, let’s change all observations to 0 if the person is an adult (16 years or older)

3. Creating New Variables

- Next, let’s change all observations to 0 if the person is an adult (16 years or older)
- Code:

3. Creating New Variables

- Next, let’s change all observations to 0 if the person is an adult (16 years or older)

- Code: *replace child=0 if age>=16*
  - Also works: *replace child=0 if age>15*

3. Creating New Variables

- Let’s explore our new variable again: *codebook child*

3. Creating New Variables

- Let's explore our new variable again: *codebook child*

```
. codebook child

--------------------------------------------------------------------------------
child                                                               (unlabeled)
--------------------------------------------------------------------------------

                  type:  numeric (float)

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/16,192

            tabulation:  Freq.  Value
                        10,132  0
                         6,060  1
```

3. Creating New Variables

- Let's explore our new variable again: *codebook child*

```
. codebook child

-------------------------------------------------------------------------------
child                                                               (unlabeled)
-------------------------------------------------------------------------------

                  type:  numeric (float)

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/16,192

            tabulation:  Freq.   Value
                        10,132   0
                         6,060   1
```

- How many missing values do we have?

3. Creating New Variables

- Let's explore our new variable again: *codebook child*

```
. codebook child

child                                                          (unlabeled)

                  type:  numeric (float)

                 range:  [0,1]                    units:  1
         unique values:  2                    missing .:  0/16,192

            tabulation:  Freq.  Value
                        10,132  0
                         6,060  1
```

- How many missing values do we have? *0*

3. Creating New Variables

- Let's explore our new variable again: *codebook child*

```
. codebook child

child                                                                  (unlabeled)

                  type:  numeric (float)

                 range:  [0,1]                       units:  1
         unique values:  2                       missing .:  0/16,192

            tabulation:  Freq.   Value
                        10,132   0
                         6,060   1
```

- How many missing values do we have? *0 – this is a problem! Does anyone know why?*

3. Creating New Variables

- How many missing values do we have? *0 – this is a problem! Does anyone know why?*

- How many missing values were changed when we recoded the 888 to missing in the “age” variable?

3. Creating New Variables

- How many missing values do we have? *0 – this is a problem! Does anyone know why?*

- How many missing values were changed when we recoded the 888 to missing in the “age” variable? *1,904*

3. Creating New Variables

- How many missing values do we have? *0 – this is a problem! Does anyone know why?*
- How many missing values were changed when we recoded the -888 to missing in the “age” variable? *1,908*
- So, there are 1904 people who we don’t know the ages of, but now they are coded as either “yes” a child or “no”, not a child in our new “child” variable…

3. Creating New Variables

- How many missing values do we have? *0 – this is a problem! Does anyone know why?*

- Stata treats missing values in numeric variables as “infinite”, and so when we use > or >=, all missing values are included.

3. Creating New Variables

- How many missing values do we have? *0 – this is a problem! Does anyone know why?*
- Stata treats missing values in numeric variables as “infinite”, and so when we use > or >=, all missing values are included.
- How can we change these values in the “child” variable back to missing, if they were missing in the “age” variable?

3. Creating New Variables

- How many missing values do we have? *0 – this is a problem! Does anyone know why?*
- Stata treats missing values in numeric variables as “infinite”, and so when we use > or >=, all missing values are included.
- How can we change these values in the “child” variable back to missing, if they were missing in the “age” variable? *replace child=. if age==.*

3. Creating New Variables

- How many missing values do we have? *0 – this is a problem! Does anyone know why?*

- Stata treats missing values in numeric variables as "infinite", and so when we use > or >=, all missing values are included.

- How can we change these values in the "child" variable back to missing, if they were missing in the "age" variable? *replace child=. if age==.*

Note the use of the single = compared to the double ==

3. Creating New Variables

- Let’s explore our new variable again: *codebook child*

3. Creating New Variables

- Let's explore our new variable again: *codebook child*

```
. codebook child

child                                                          (unlabeled)

                  type:  numeric (float)

                 range:  [0,1]                    units:  1
         unique values:  2                    missing .:  1,904/16,192

            tabulation:  Freq.  Value
                         8,228  0
                         6,060  1
                         1,904  .
```

3. Creating New Variables

- Let's explore our new variable again: *codebook child*

```
. codebook child

child                                                                    (unlabeled)

                  type:  numeric (float)

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  1,904/16,192

           tabulation:  Freq.   Value
                        8,228   0
                        6,060   1
                        1,904   .
```

- Now what do we still need to add?

3. Creating New Variables

- Let's explore our new variable again: *codebook child*

```
. codebook child

child                                                                  (unlabeled)

                  type:  numeric (float)

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  1,904/16,192

            tabulation:  Freq.   Value
                         8,228   0
                         6,060   1
                         1,904   .
```

- Now what do we still need to add? *Labels!*

3. Creating New Variables

- Let’s first add a variable label. What is the code for this?

3. Creating New Variables

- Let’s first add a variable label. What is the code for this?

- *Label variable child “Is the household member 15 years or younger?”*
  - Shorthand: *lab var child “Is the household member…”*

3. Creating New Variables

- Now, let's label the 0 values as "No" and the 1 values as "Yes"

3. Creating New Variables

- Now, let’s label the 0 values as “No” and the 1 values as “Yes”
- First, we have to create a value label definition

3. Creating New Variables

- Now, let's label the 0 values as "No" and the 1 values as "Yes"
- First, we have to create a value label definition
- Code: *label define no_yes 0 "No" 1 "Yes"*

3. Creating New Variables

- Now, let’s label the 0 values as “No” and the 1 values as “Yes”
- First, we have to create a value label definition
- Code: *label define no_yes 0 “No” 1 “Yes”*
- Right now this label definition is only saved in Stata’s memory - it has not been applied to any variables yet

3. Creating New Variables

- Now we will write the code to apply this value label to our new “child” variable

3. Creating New Variables

- Now we will write the code to apply this value label to our new "child" variable

- Code: *label values child no_yes*

3. Creating New Variables

- Now we will write the code to apply this value label to our new "child" variable
- Code: *label values child no_yes*
- This code applies our new label definition "no_yes" to the variable "child"

3. Creating New Variables

- Now we will write the code to apply this value label to our new “child” variable
- Code: *label values child no_yes*
- This code applies our new label definition “no_yes” to the variable “child”
- Let’s look one more time at our new variable: *codebook child*

child
No
No
No
No
No
Yes
Yes
No
No
Yes

3. Creating New Variables

- Let's explore our new variable: *codebook child*

```
. codebook child

-------------------------------------------------------------------------------
child                                 Is the household member 15 years or younger?
-------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  noyes

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  1,904/16,192

            tabulation:  Freq.   Numeric  Label
                         8,228         0  No
                         6,060         1  Yes
                         1,904         .
```

3. Creating New Variables

Variable label

- Let’s explore our new variable: *codebook child*

```
. codebook child

--------------------------------------------------------------------------------
child                                     Is the household member 15 years or younger?
--------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  noyes

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  1,904/16,192

            tabulation:  Freq.   Numeric  Label
                         8,228         0  No
                         6,060         1  Yes
                         1,904         .
```

3. Creating New Variables

- Let’s explore our new variable: *codebook child*

```
. codebook child

child                                    Is the household member 15 years or younger?

                  type:  numeric (float)
                 label:  noyes

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  1,904/16,192

           tabulation:  Freq.   Numeric  Label
                        8,228         0  No
                        6,060         1  Yes
                        1,904         .
```

Variable label

Value labels

3. Creating New Variables

- Let's explore our new variable: *codebook child*

Variable label

```
. codebook child

-------------------------------------------------------------------------------
child                                 Is the household member 15 years or younger?
-------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  noyes

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  1,904/16,192

            tabulation:  Freq.   Numeric  Label
                         8,228         0  No
                         6,060         1  Yes
                         1,904         .
```

Value labels

- How many children are in the dataset?

3. Creating New Variables

- Let's explore our new variable: *codebook child*

```
. codebook child

--------------------------------------------------------------------------------
child                                    Is the household member 15 years or younger?
--------------------------------------------------------------------------------

                  type:  numeric (float)
                 label:  noyes

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  1,904/16,192

           tabulation:  Freq.   Numeric  Label
                        8,228         0  No
                        6,060         1  Yes
                        1,904         .
```

Variable label

Value labels

- How many children are in the dataset? *6,060*

3. Creating New Variables

- Question: What percent of children have ‘student’ listed as their primary activity?

3. Creating New Variables

- Question: What percent of children have ‘student’ listed as their primary activity?

- What code might help us find this answer?

3. Creating New Variables

- Question: What percent of children have ‘student’ listed as their primary activity?
- What code might help us find this answer? *Tabulate*

3. Creating New Variables

- Question: What percent of children have 'student' listed as their primary activity?
- What code might help us find this answer? *Tabulate*
- Code: *tab a8 if child==1*

3. Creating New Variables

- Question: What percent of children have ‘student’ listed as their primary activity?

```
. tab a8 if child == 1
```

b8 : What is the current PRIMARY activity / occupation for ${b_name2}? --------	Freq.	Percent	Cum.
Any agricultural (farmer, gardener, sub	125	10.37	10.37
Forestry or fishery labor	1	0.08	10.46
Transport and storage laborers	1	0.08	10.54
Services and sales worker	1	0.08	10.62
Security guard	1	0.08	10.71
Student	1,041	86.39	97.10
Domestic work/housewife	7	0.58	97.68
Unemployed and NOT looking for work	6	0.50	98.17
Unemployed and ACTIVELY looking for wor	1	0.08	98.26
Community leader (counselor, peace medi	1	0.08	98.34
Not applicable	18	1.49	99.83
Other	2	0.17	100.00
Total	1,205	100.00	

3. Creating New Variables

- Question: What percent of children have ‘student’ listed as their primary activity?

```
. tab a8 if child == 1

      b8 : What is the current PRIMARY |
  activity / occupation for ${b_name2}? |
                               -------- |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
Any agricultural (farmer, gardener, sub |        125       10.37       10.37
              Forestry or fishery labor |          1        0.08       10.46
         Transport and storage laborers |          1        0.08       10.54
              Services and sales worker |          1        0.08       10.62
                         Security guard |          1        0.08       10.71
                                Student |      1,041       86.39       97.10
                 Domestic work/housewife |          7        0.58       97.68
       Unemployed and NOT looking for work |          6        0.50       98.17
Unemployed and ACTIVELY looking for wor |          1        0.08       98.26
Community leader (counselor, peace medi |          1        0.08       98.34
                         Not applicable |         18        1.49       99.83
                                  Other |          2        0.17      100.00
----------------------------------------+-----------------------------------
                                  Total |      1,205      100.00
```

- Answer:

3. Creating New Variables

- Question: What percent of children have 'student' listed as their primary activity?

```
. tab a8 if child == 1

      b8 : What is the current PRIMARY
   activity / occupation for ${b_name2}?
                               --------  |      Freq.     Percent        Cum.
-----------------------------------------+-----------------------------------
Any agricultural (farmer, gardener, sub  |        125       10.37       10.37
                   Forestry or fishery labor |          1        0.08       10.46
            Transport and storage laborers |          1        0.08       10.54
                  Services and sales worker |          1        0.08       10.62
                             Security guard |          1        0.08       10.71
                                    Student |      1,041       86.39       97.10
                    Domestic work/housewife |          7        0.58       97.68
         Unemployed and NOT looking for work |          6        0.50       98.17
Unemployed and ACTIVELY looking for wor  |          1        0.08       98.26
Community leader (counselor, peace medi  |          1        0.08       98.34
                             Not applicable |         18        1.49       99.83
                                      Other |          2        0.17      100.00
-----------------------------------------+-----------------------------------
                                      Total |      1,205      100.00
```

- Answer: *86.39% of children aged 15 and younger are primarily students*

3. Creating New Variables

- Challenge question: Which province has the highest percentage of children?

3. Creating New Variables

- Challenge question: Which province has the highest percentage of children?

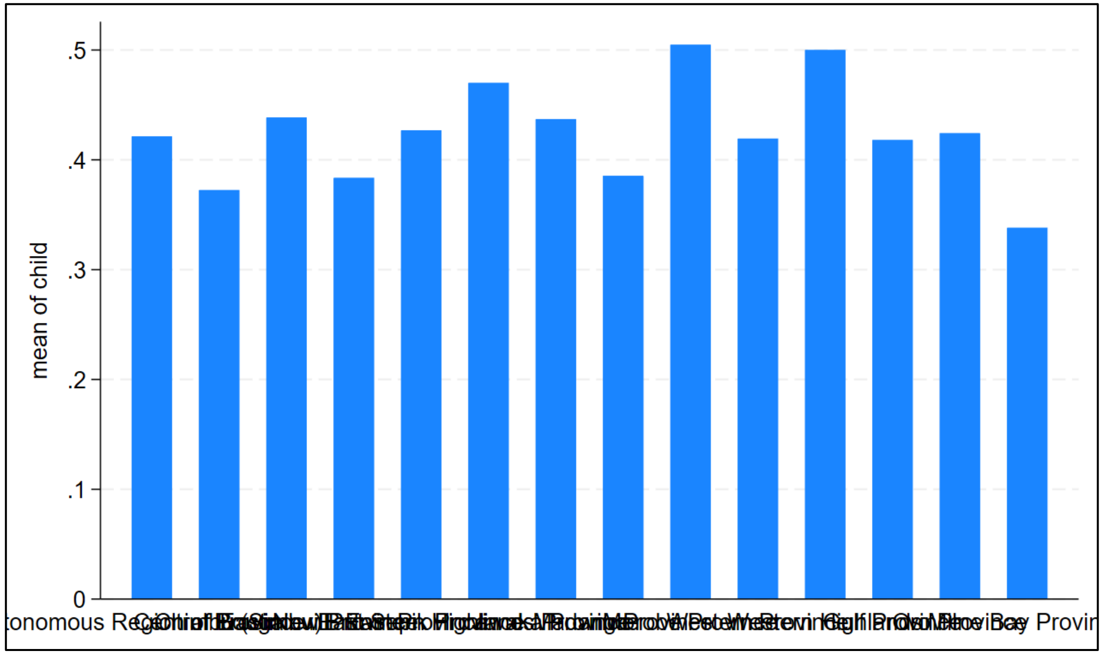
- Let's use a figure to find the answer!

3. Creating New Variables

- Challenge question: Which province has the highest percentage of children?
- Let's use a figure to find the answer!
- Code: *graph bar child, over(prov)*

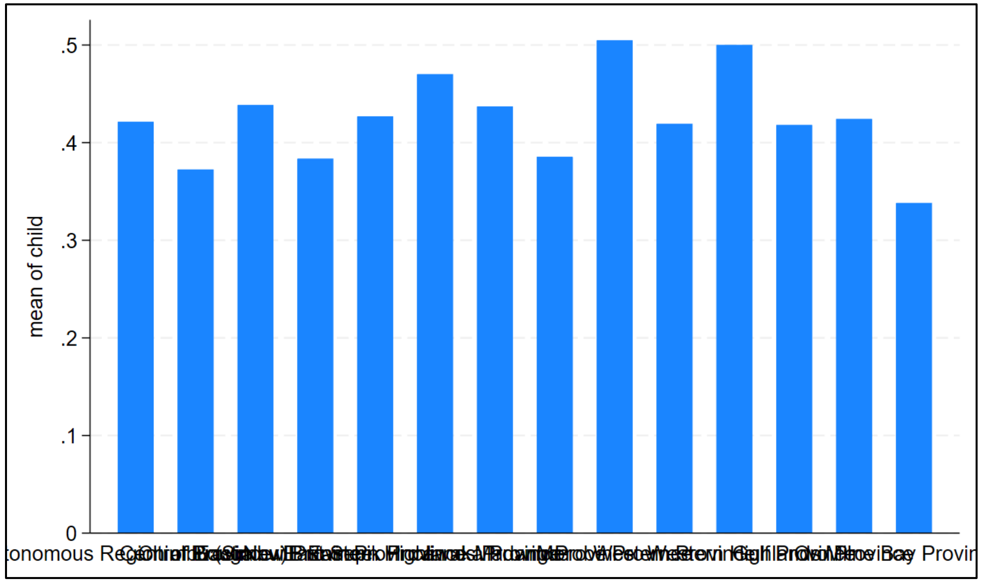
3. Creating New Variables

- *graph bar child, over(prov)*



3. Creating New Variables

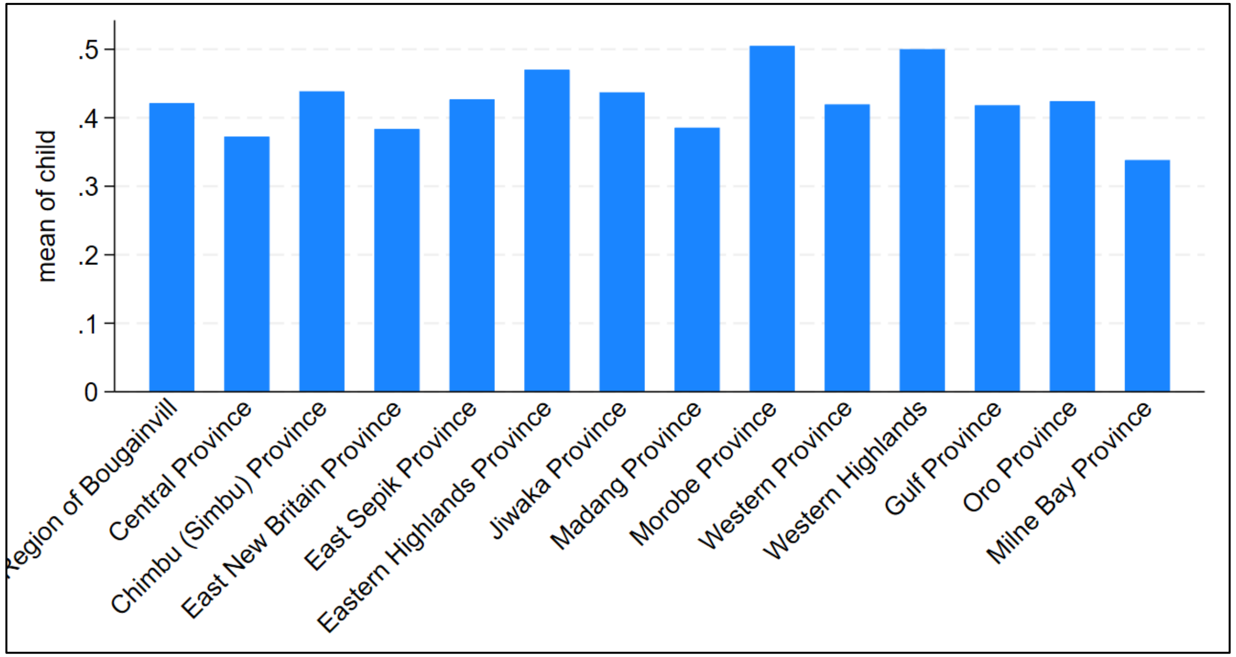
- *graph bar child, over(prov)*



- *Notice that the bar labels on x-axis are on top of each other. To produce a neat graph, we can use the following code to avoid the overlap:*
- *graph bar child, over(prov, label(angle(45)))*

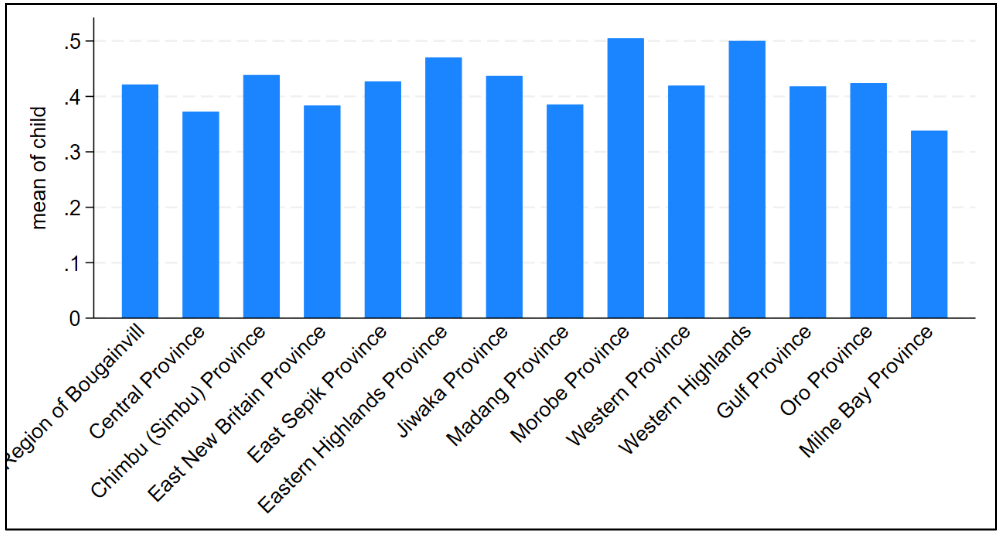
3. Creating New Variables

- *graph bar child, over(prov, label(angle(45)))*



3. Creating New Variables

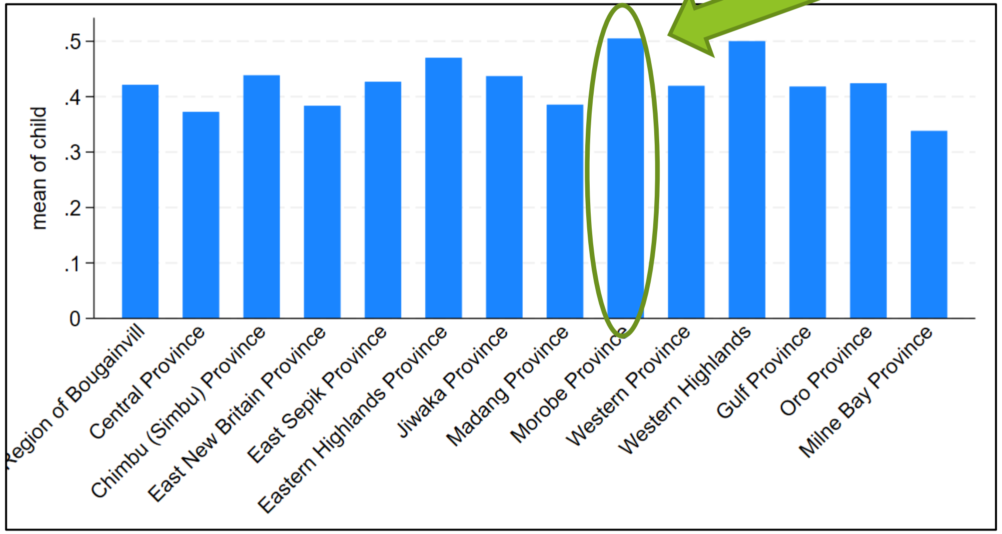
- *graph bar child, over(prov, label(angle(45)))*



- Because values are 0 and 1, a bar graph (ranging from 0 to 1) shows the prevalence of the dummy variable (1=100%)

3. Creating New Variables

- *graph bar child, over(prov)*



Morobe has the highest prevalence of children

- Because values are 0 and 1, a bar graph (ranging from 0 to 1) shows the prevalence of the dummy variable (1=100%)

3. Creating New Variables

- Now, let’s try to graph the prevalence of women by province.

3. Creating New Variables

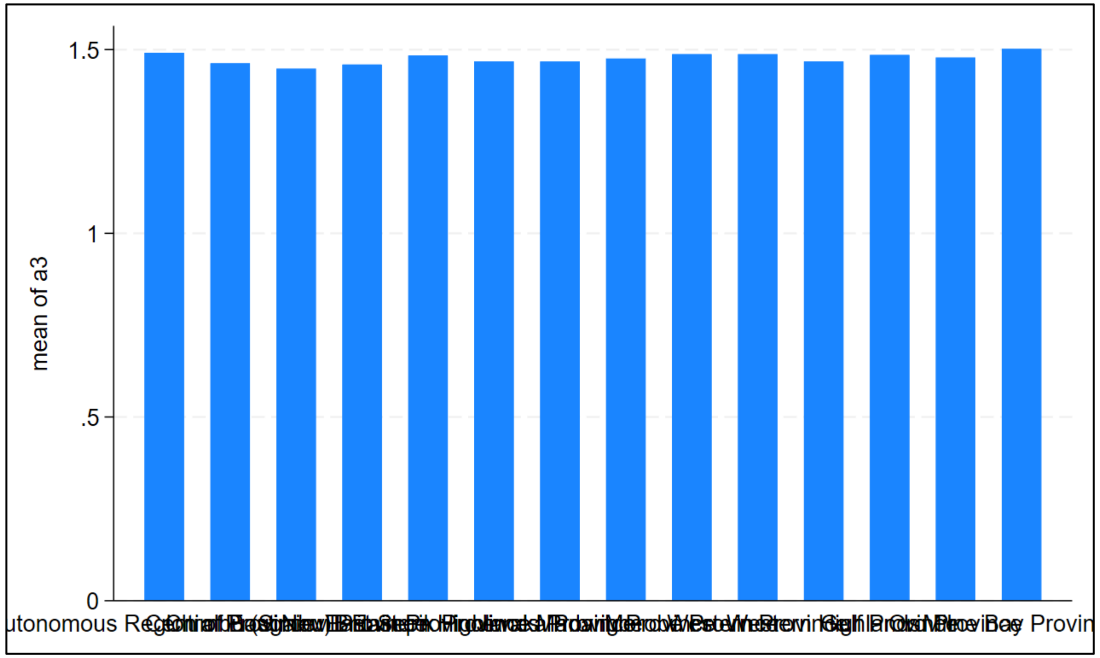
- Now, let’s try to graph the prevalence of women by province.
- How might we write the code?

3. Creating New Variables

- Now, let's try to graph the prevalence of women by province.
- How might we write the code? *graph bar a3, over(prov)*

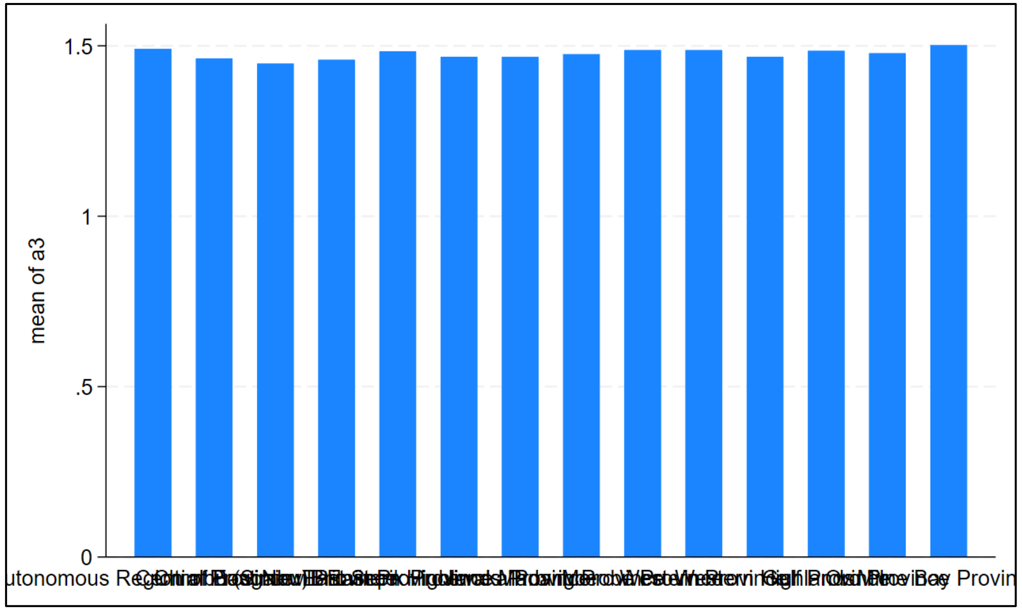
3. Creating New Variables

- *Graph bar a3, over(prov)*



3. Creating New Variables

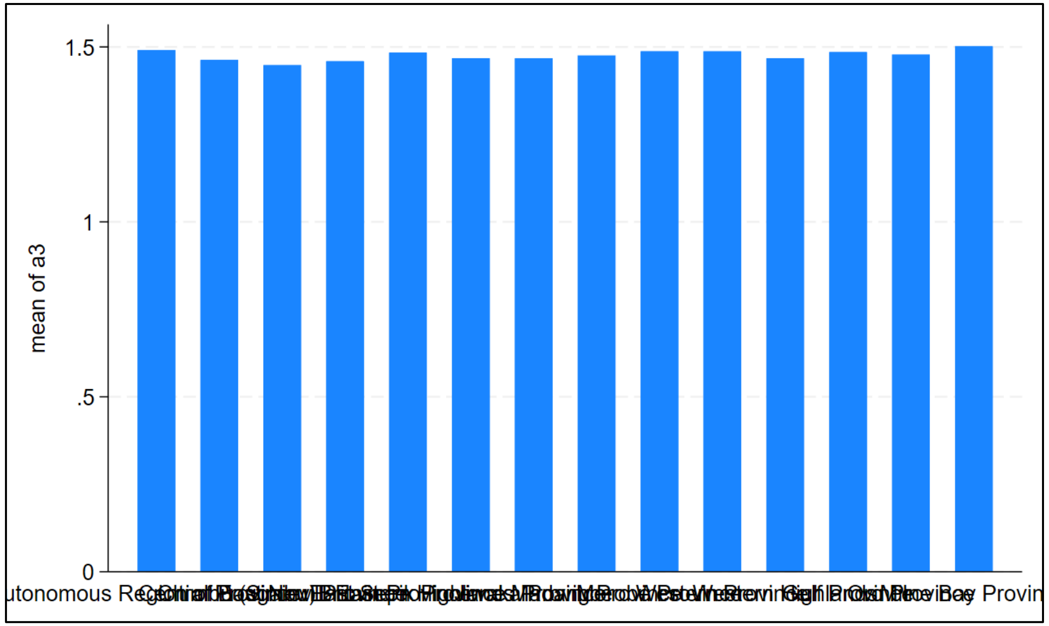
- *Graph bar a3, over(prov)*



- Does this show the prevalence of women in each province?

3. Creating New Variables

- *Graph bar a3, over(prov)*



- Does this figure show the prevalence of women in each province? *No – all of the bars are over 1!*

3. Creating New Variables

- Does this figure show the prevalence of women in each province? *No – all of the bars are over 1!*
- Why is this figure different from the other? Let's look again at our gender variable.

3. Creating New Variables

- Does this figure show the prevalence of women in each province? *No – all of the bars are over 1!*

- Why is this figure different from the other? Let's look again at our gender variable. *Codebook a3*

3. Creating New Variables

- Does this figure show the prevalence of women in each province? *No – all of the bars are over 1!*

- Why is this figure different from the other? Let’s look again at our gender variable. *Codebook a3*

```
. codebook a3

a3                                                      3. What is the sex of ?

                  Type: Numeric (byte)
                 Label: b3

                 Range: [1,2]                         Units: 1
         Unique values: 2                         Missing .: 0/16,192

            Tabulation: Freq.   Numeric  Label
                        8,459         1  Male
                        7,733         2  Female
```

3. Creating New Variables

- Does this figure show the prevalence of women in each province? *No – all of the bars are over 1!*
- Why is this figure different from the other? Let's look again at our gender variable. *Codebook a3*

```
. codebook a3

a3                                                    3. What is the sex of ?

                  Type: Numeric (byte)
                 Label: b3

                 Range: [1,2]                          Units: 1
         Unique values: 2                          Missing .: 0/16,192

            Tabulation: Freq.   Numeric  Label
                        8,459         1  Male
                        7,733         2  Female
```

The value labels are 1/2 instead of 0/1

3. Creating New Variables

- Let’s make a new dummy variable for whether someone is a woman.

3. Creating New Variables

- Let’s make a new dummy variable for whether someone is a woman.
- What’s first?

3. Creating New Variables

- Let's make a new dummy variable for whether someone is a woman.
- What's first? *generate woman=.*

3. Creating New Variables

- Let’s make a new dummy variable for whether someone is a woman.
- What’s first? *generate woman=.*
- How do we set the values to 1 for woman and 0 for man?

3. Creating New Variables

- Let’s make a new dummy variable for whether someone is a woman.
- What’s first? *generate woman=.*
- How do we set the values to 1 for woman and 0 for man?
  - *Replace woman=1 if a3==2*
  - *Replace woman=0 if a3==1*

3. Creating New Variables

- Let’s create variable label:

3. Creating New Variables

- Let’s create variable label: *lab var woman “Is the household member a woman?”*

3. Creating New Variables

- Let’s create variable label: *lab var woman “Is the household member a woman?”*

- Remember, we already have the value label definition “no_yes” saved in Stata. Will this value definition work for this variable too?

3. Creating New Variables

- Let’s create variable label: *lab var woman “Is the household member a woman?”*

- Remember, we already have the value label definition “no_yes” saved in Stata. Will this value definition work for this variable too? *Yes, it would work for all dummy variables*

3. Creating New Variables

- Let's create variable label: *lab var woman "Is the household member a woman?"*
- Remember, we already have the value label definition "no_yes" saved in Stata. Will this value definition work for this variable too? *Yes, it would work for all dummy variables*
- How can we apply this value label definition to our new "woman" variable?

3. Creating New Variables

- Let’s create variable label: *lab var woman “Is the household member a woman?”*

- Remember, we already have the value label definition “no_yes” saved in Stata. Will this value definition work for this variable too? *Yes, it would work for all dummy variables*

- How can we apply this value label definition to our new “woman” variable? *label values woman no_yes*

3. Creating New Variables

- Let’s look at our new variable:

3. Creating New Variables

- Let’s look at our new variable: *codebook woman*

3. Creating New Variables

- Let's look at our new variable: *codebook woman*

```
woman                                         Is the household member a woman?

                  Type: Numeric (float)
                 Label: no_yes

                 Range: [0,1]                       Units: 1
         Unique values: 2                       Missing .: 0/16,192

            Tabulation: Freq.   Numeric  Label
                        8,459         0  No
                        7,733         1  Yes
```

3. Creating New Variables

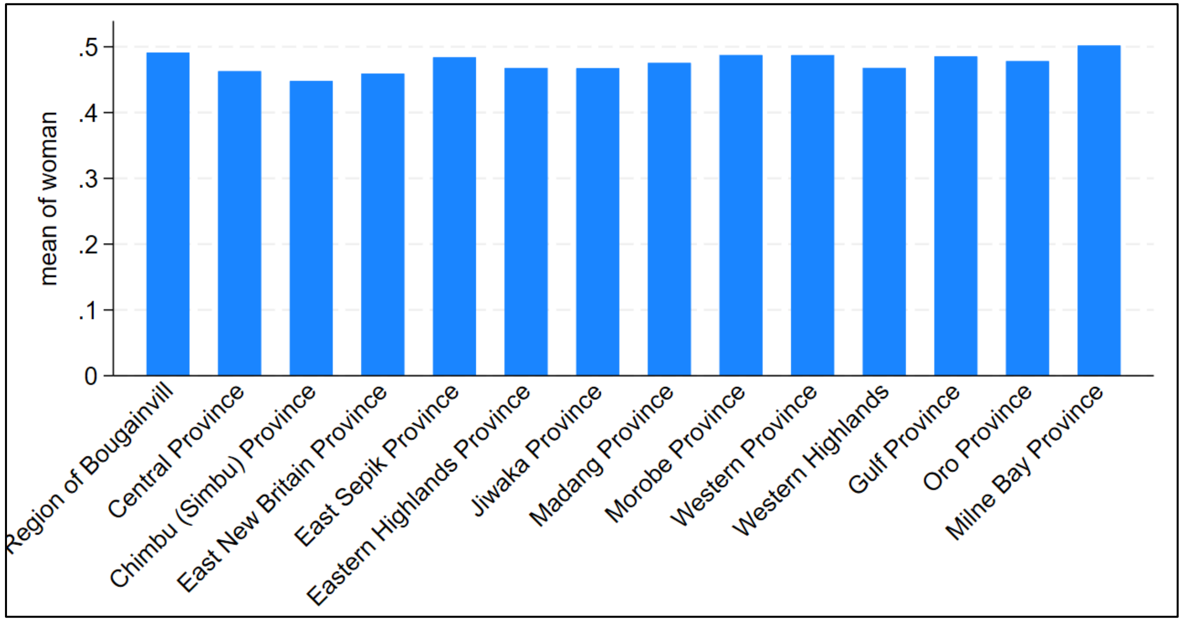
- Let’s try the bar graph again with our new variable:

3. Creating New Variables

- Let’s try the bar graph again with our new variable: *graph bar woman, over(prov, label(angle(45)))*

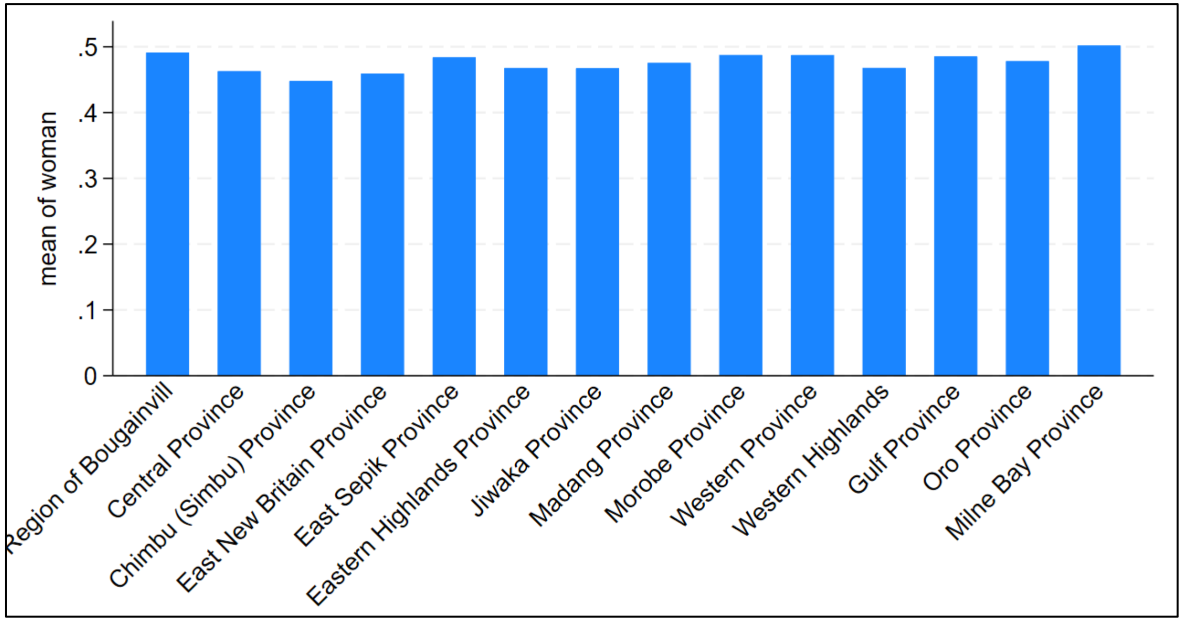
3. Creating New Variables

- Let's try the bar graph again with our new variable: *graph bar woman, over(prov, label(angle(45)))*



3. Creating New Variables

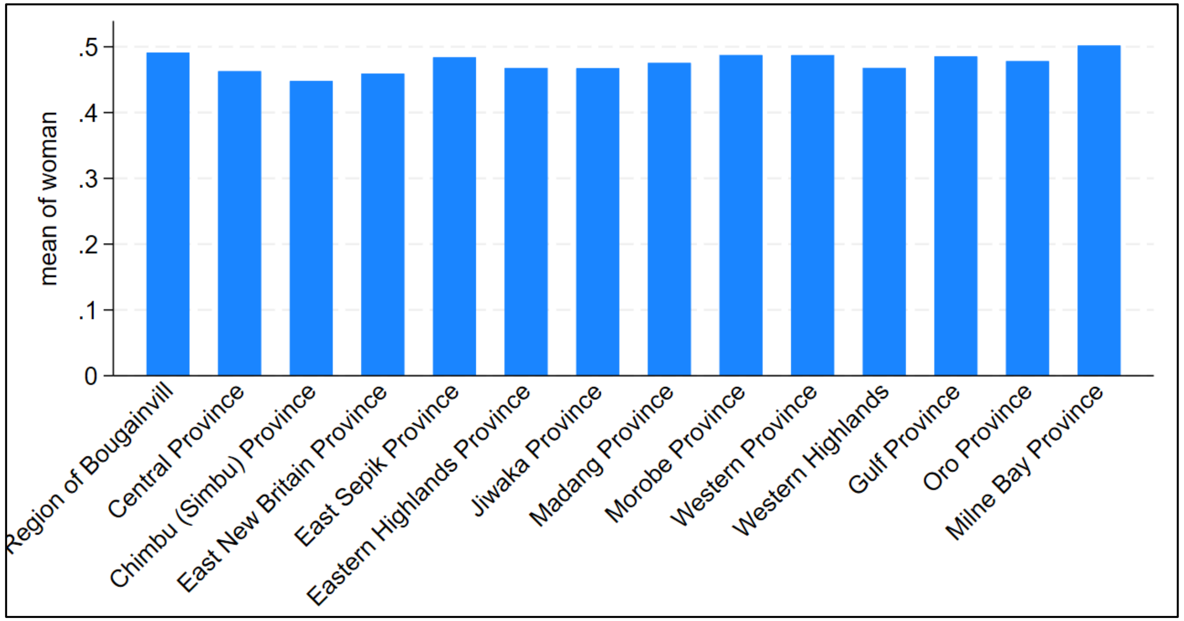
- Let's try the bar graph again with our new variable: *graph bar woman, over(prov, label(angle(45)))*



- Does it show prevalence now?

3. Creating New Variables

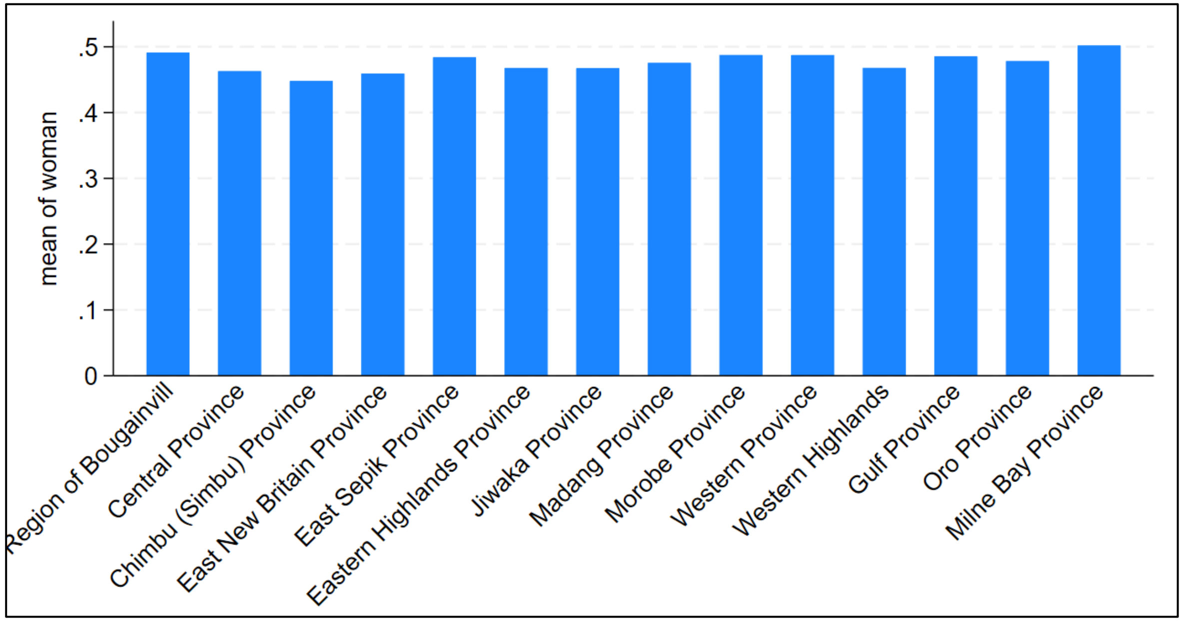
- Let’s try the bar graph again with our new variable: *graph bar woman, over(prov, label(angle(45)))*



- Does it show prevalence now? *Yes!*

3. Creating New Variables

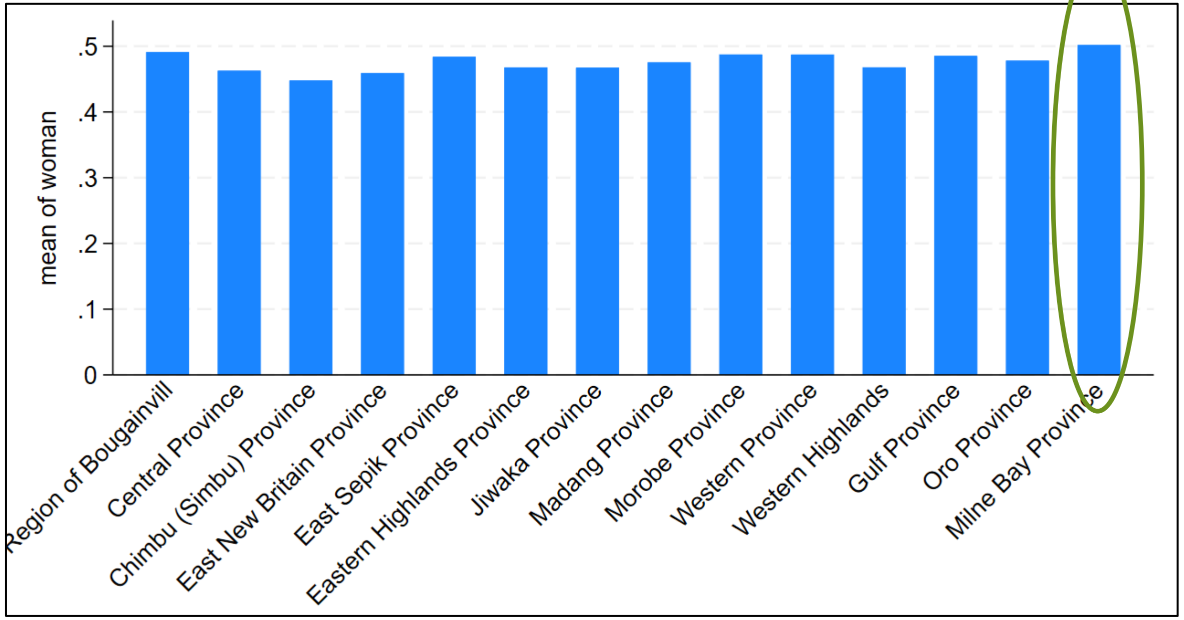
- Let's try the bar graph again with our new variable: *graph bar woman, over(prov, label(angle(45)))*



- Does it show prevalence now? *Yes!*
- Which province has the highest prevalence of women in the sample?

3. Creating New Variables

- Let's try the bar graph again with our new variable: *graph bar woman, over(prov, label(angle(45)))*



- Does it show prevalence now? *Yes!*

- Which province has the highest prevalence of women in the sample? *Milne Bay province*

4. Help Window

- Help windows explain how to use each code, provide different options, examples, etc.

4. Help Window

- Help windows explain how to use each code, provide different options, examples, etc.
- Code: *help [command]*

4. Help Window

- Help windows explain how to use each code, provide different options, examples, etc.
- Code: *help [command]*
- Example: *help tabulate twoway*

4. Help Window

- Example: *help tabulate twoway*
  - ***Syntax***

```
[R] tabulate twoway — Two-way table of frequencies
                      (View complete PDF manual entry)


Syntax

    Two-way table

        tabulate varname1 varname2 [if] [in] [weight] [, options]
```

4. Help Window

- Example: *help tabulate twoway*
  - *Syntax*
  - ***Options***

options	Description
Main	
chi2	report Pearson's chi-squared
exact[(#)]	report Fisher's exact test
gamma	report Goodman and Kruskal's gamma
lrchi2	report likelihood-ratio chi-squared
taub	report Kendall's tau-b
V	report Cramér's V
cchi2	report Pearson's chi-squared in each cell
column	report relative frequency within its column of each cell
row	report relative frequency within its row of each cell
clrchi2	report likelihood-ratio chi-squared in each cell
cell	report the relative frequency of each cell

4. Help Window

- Example: *help tabulate twoway*
  - *Syntax*
  - *Options*
  - ***Examples***

```
     Two-way table of frequencies
          . tabulate region agecat

     Include row percentages
          . tabulate region agecat, row

     Include column percentages
          . tabulate region agecat, column

     Include cell percentages
          . tabulate region agecat, cell
```

4. Help Window

- Now let's look at the help window for *summarize*

- Code:

4. Help Window

- Now let's look at the help window for *summarize*

- Code: *help summarize*

4. Help Window

- Now let's look at the help window for *summarize*

- Code: *help summarize*
  - ***Syntax***
  - ***Options***

```
[R] summarize — Summary statistics
                (View complete PDF manual entry)


Syntax

        summarize [varlist] [if] [in] [weight] [, options]

    options              Description
    ---------------------------------------------------------------------------
    Main
      detail             display additional statistics
      meanonly           suppress the display; calculate only the mean; programmer's option
      format             use variable's display format
      separator(#)       draw separator line after every # variables; default is separator(5)
      display_options    control spacing, line width, and base and empty cells
```

5. Collapse and Merge
Collapse

- What if we want to know some things at the household level that need to be constructed from this person-level data?

5. Collapse and Merge
Collapse

- What if we want to know some things at the household level that need to be constructed from this person-level data?
  - Household size?

5. Collapse and Merge
Collapse

- What if we want to know some things at the household level that need to be constructed from this person-level data?
  - Household size?
  - Age of youngest household member?

5. Collapse and Merge
Collapse

- What if we want to know some things at the household level that need to be constructed from this person-level data?
  - Household size?
  - Age of youngest household member?
  - Percent of children and percent of women in the household?

5. Collapse and Merge
Collapse

- What if we want to know some things at the household level that need to be constructed from this person-level data?
  - Household size?
  - Age of youngest household member?
  - Percent of children and percent of women in the household?

- We can use the *collapse* command to find these things!

5. Collapse and Merge
Collapse

- *Collapse* creates a dataset of the summary statistics that you specify, at the level that you specify

5. Collapse and Merge
Collapse

- *Collapse* creates a dataset of the summary statistics that you specify, at the level that you specify

- Examples of summary statistics that can be used are: *mean*, *median*, *sum*, *count*, *max*, *and min*

5. Collapse and Merge
Collapse

- *Collapse* creates a dataset of the summary statistics that you specify, at the level that you specify

- Examples of summary statistics that can be used are: *mean, median, sum, count, max, and min*

- We will collapse this dataset to the household level.

5. Collapse and Merge
Collapse

- *Collapse* creates a dataset of the summary statistics that you specify, at the level that you specify

- Examples of summary statistics that can be used are: *mean, median, sum, count, max, and min*

- We will collapse this dataset to the household level.

- Which summary statistic and variable would we use to find the number of household members?

5. Collapse and Merge
Collapse

- *Collapse* creates a dataset of the summary statistics that you specify, at the level that you specify

- Examples of summary statistics that can be used are: *mean, median, sum, count, max, and min*

- We will collapse this dataset to the household level.

- Which summary statistic and variable would we use to find the number of household members? *(count) a_pid*

5. Collapse and Merge
Collapse

- *Collapse* creates a dataset of the summary statistics that you specify, at the level that you specify
- Examples of summary statistics that can be used are: *mean, median, sum, count, max, and min*
- We will collapse this dataset to the household level.
- Which summary statistic and variable would we use to find the age of the youngest HH member?

5. Collapse and Merge
Collapse

- *Collapse* creates a dataset of the summary statistics that you specify, at the level that you specify

- Examples of summary statistics that can be used are: *mean, median, sum, count, max, and min*

- We will collapse this dataset to the household level.

- Which summary statistic and variable would we use to find the age of the youngest HH member? *(min) age*

5. Collapse and Merge
Collapse

- *Collapse* creates a dataset of the summary statistics that you specify, at the level that you specify

- Examples of summary statistics that can be used are: *mean, median, sum, count, max, and min*

- We will collapse this dataset to the household level.

- Which summary statistic and variables would we use to find the percent of women & children in each HH?

5. Collapse and Merge
Collapse

- *Collapse* creates a dataset of the summary statistics that you specify, at the level that you specify
- Examples of summary statistics that can be used are: *mean, median, sum, count, max, and min*
- We will collapse this dataset to the household level.
- Which summary statistic and variables would we use to find the percent of women & children in each HH? *(mean) child woman*

5. Collapse and Merge
Collapse

- Code: *collapse (count) a_pid (min) age (mean) child woman, by(hhid)*

5. Collapse and Merge
Collapse

- Code: *collapse (count) a_pid (min) age (mean) child woman, by(hhid)*

- How many observations do we have now in the dataset?

5. Collapse and Merge
Collapse

- Code: *collapse (count) a_pid (min) age (mean) child woman, by(hhid)*

- How many observations do we have now in the dataset?
  2,699

5. Collapse and Merge
Collapse

- Now, let’s look at what happened in our Window 3 (Variable Window) after the collapse

5. Collapse and Merge
Collapse

- Now, let’s look at what happened in our Window 3 (Variable Window) after the collapse

Variables

Filter variables here

Name	Label
hhid	Household identifier
pid	(count) pid
age	(min) age
child	(mean) child
woman	(mean) woman

5. Collapse and Merge
Collapse

- Now, let's look at what happened in our Window 3 (Variable Window) after the collapse

Variables

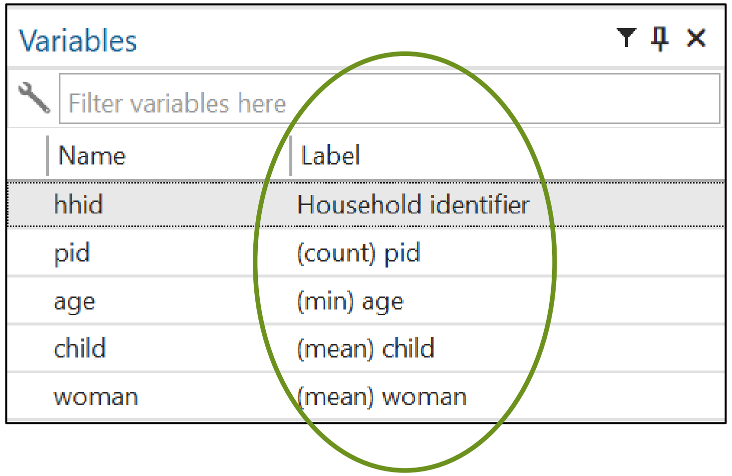
Filter variables here

Name	Label
hhid	Household identifier
pid	(count) pid
age	(min) age
child	(mean) child
woman	(mean) woman

- Variable names remained the same

5. Collapse and Merge
Collapse

- Now, let’s look at what happened in our Window 3 (Variable Window) after the collapse



- Variable names remained the same
- Variable labels tell us the summary statistic used in the collapse (instead of keeping the old labels)

5. Collapse and Merge
Collapse

- Let’s create more meaningful variable names and labels:

5. Collapse and Merge
Collapse

- Let’s create more meaningful variable names and labels:

- What does the variable *a_pid* tell us now?

5. Collapse and Merge
Collapse

- Let’s create more meaningful variable names and labels:

- What does the variable *a_pid* tell us now? *The count of people in the household = household size!*

5. Collapse and Merge
Collapse

- Let’s create more meaningful variable names and labels:

- What does the variable *a_pid* tell us now? *The count of people in the household = household size!*

- *rename a_pid hhsize*

5. Collapse and Merge
Collapse

- Let’s create more meaningful variable names and labels:
- What does the variable *a_pid* tell us now? *The count of people in the household = household size!*
- *rename a_pid hhsize*
- *lab var hhsize “Household size”*

5. Collapse and Merge
Collapse

- Let's create more meaningful variable names and labels:

- What does the variable *age* tell us now?

5. Collapse and Merge
Collapse

- Let's create more meaningful variable names and labels:

- What does the variable *age* tell us now? *The minimum age out of the ages of each person in the HH = the age of the youngest person in the HH*

5. Collapse and Merge
Collapse

- Let's create more meaningful variable names and labels:

- What does the variable *age* tell us now? *The minimum age out of the ages of each person in the HH = the age of the youngest person in the HH*

- *rename age age_youngest*

5. Collapse and Merge
Collapse

- Let’s create more meaningful variable names and labels:
- What does the variable *age* tell us now? *The minimum age out of the ages of each person in the HH = the age of the youngest person in the HH*
- *rename age age_youngest*
- *lab var age_youngest “Age of the youngest household member”*

5. Collapse and Merge
Collapse

- Let's create more meaningful variable names and labels:

- What does the variable *child* tell us now?

5. Collapse and Merge
Collapse

- Let's create more meaningful variable names and labels:

- What does the variable *child* tell us now? *The average/mean of the dummy variable "child" = the percent of children in each household*

5. Collapse and Merge
Collapse

- Let's create more meaningful variable names and labels:
- What does the variable *child* tell us now? *The average/mean of the dummy variable "child" = the percent of children in each household*
- *rename child perc_children*

5. Collapse and Merge
Collapse

- Let’s create more meaningful variable names and labels:
- What does the variable *child* tell us now? *The average/mean of the dummy variable “child” = the percent of children in each household*
- *rename child perc_children*
- *lab var perc_children “Percent of children 15 years and younger in the household”*

5. Collapse and Merge
Collapse

- Let’s create more meaningful variable names and labels:

- What does the variable *woman* tell us now?

5. Collapse and Merge
Collapse

- Let's create more meaningful variable names and labels:

- What does the variable *woman* tell us now? *The average/mean of the dummy variable "woman" = the percent of women in each household*

5. Collapse and Merge
Collapse

- Let’s create more meaningful variable names and labels:

- What does the variable *woman* tell us now? *The average/mean of the dummy variable “woman” = the percent of women in each household*

- *rename woman perc_women*

5. Collapse and Merge
Collapse

- Let’s create more meaningful variable names and labels:
- What does the variable *woman* tell us now? *The average/mean of the dummy variable “woman” = the percent of women in each household*
- *rename woman perc_women*
- *lab var perc_women “Percent of women in the household”*

5. Collapse and Merge
Collapse

- Question: What is the average household size?

5. Collapse and Merge
Collapse

- Question: What is the average household size?

- *sum hhsize*

5. Collapse and Merge
Collapse

- <u>Question:</u> What is the average household size?
- *sum hhsize*

```
. sum hhsize

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      hhsize |      2,699    5.999259    2.603812          1         22
```

- <u>Answer:</u>

5. Collapse and Merge
Collapse

- Question: What is the average household size?

- *sum hhsize*

```
. sum hhsize

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      hhsize |      2,699    5.999259    2.603812          1         22
```

- Answer: *5.9 people per household*

5. Collapse and Merge
Collapse

- Question: What is the average age of the youngest household member?

5. Collapse and Merge
Collapse

- Question: What is the average age of the youngest household member?
  - *Sum age_youngest*

5. Collapse and Merge
Collapse

- <u>Question:</u> What is the average age of the youngest household member?
  - *Sum age_youngest – 8.7 years old*

5. Collapse and Merge
Collapse

- Question: What is the average age of the youngest household member?
  - *Sum age_youngest – 8.7 years old*

- Question: What is the average prevalence of children in a household?

5. Collapse and Merge
Collapse

- Question: What is the average age of the youngest household member?
  - *Sum age_youngest – 8.7 years old*

- Question: What is the average prevalence of children in a household?
  - *Sum perc_children*

5. Collapse and Merge
Collapse

- Question: What is the average age of the youngest household member?
  - *Sum age_youngest – 8.7 years old*

- Question: What is the average prevalence of children in a household?
  - *Sum perc_children – 40.8% children*

5. Collapse and Merge
Collapse

- Question: What is the average age of the youngest household member?
  - *Sum age_youngest – 8.7 years old*

- Question: What is the average prevalence of children in a household?
  - *Sum perc_children – 40.8% children*

- Question: What is the average prevalence of women in a household?

5. Collapse and Merge
Collapse

- Question: What is the average age of the youngest household member?
  - *Sum age_youngest – 8.7 years old*

- Question: What is the average prevalence of children in a household?
  - *Sum perc_children – 40.8% children*

- Question: What is the average prevalence of women in a household?
  - *Sum perc_women*

5. Collapse and Merge
Collapse

- Question: What is the average age of the youngest household member?
  - *Sum age_youngest – 8.7 years old*

- Question: What is the average prevalence of children in a household?
  - *Sum perc_children – 40.8% children*

- Question: What is the average prevalence of women in a household?
  - *Sum perc_women – 47.2% women*

5. Collapse and Merge
Merge (same number of observations)

- What if we want to know if household size is associated with the type of roof that a household has?

5. Collapse and Merge
Merge (same number of observations)

- What if we want to know if household size is associated with the type of roof that a household has?

- We have household size in this dataset, but roof type is in the dataset that we worked with in Lesson 1

5. Collapse and Merge
Merge (same number of observations)

- What if we want to know if household size is associated with the type of roof that a household has?

- We have household size in this dataset, but roof type is in the dataset that we worked with in Lesson 1

- We can merge the two datasets together – and now they have the same number and type of observations *(2,699 households)*

5. Collapse and Merge
Merge (same number of observations)

s11_long (collapsed; n=2,699)	
hhid	age_youngest
0100RSPN	11

5. Collapse and Merge
Merge (same number of observations)

s11_long (collapsed; n=2,699)	
hhid	age_youngest
0100RSPN	11

hh_3_4 (n=2,699)	
hhid	i1
0100RSPN	metal roof

5. Collapse and Merge
Merge (same number of observations)

s11_long (collapsed; n=2,699)	
hhid	age_youngest
0100RSPN	11

hh_3_4 (n=2,699)	
hhid	i1
0100RSPN	metal roof

merged data (n=2,699)		
hhid	age_youngest	i1
0100RSPN	11	metal roof

5. Collapse and Merge
Merge (same number of observations)

s11_long (collapsed; n=2,699)	
hhid	age_youngest
0100RSPN	11

hh_3_4 (n=2,699)	
hhid	i1
0100RSPN	metal roof

merged data (n=2,699)		
hhid	age_youngest	i1
0100RSPN	11	metal roof

5. Collapse and Merge
Merge (same number of observations)

- *Help merge*

5. Collapse and Merge
Merge (same number of observations)

- *Help merge*

```
[D] merge —— Merge datasets
             (View complete PDF manual entry)


Syntax

    One-to-one merge on specified key variables

        merge 1:1 varlist using filename [, options]
```

5. Collapse and Merge
Merge (same number of observations)

- *Help merge*

```
[D] merge — Merge datasets
            (View complete PDF manual entry)


Syntax

    One-to-one merge on specified key variables

        merge 1:1 varlist using filename [, options]
```

- We are merging *one household to one household* (we have exactly one observation for each household in both datasets) – *so we will use "1:1" syntax*

5. Collapse and Merge
Merge (same number of observations)

- Which variable will we be using to match observations from one dataset to the other?

5. Collapse and Merge
Merge (same number of observations)

- Which variable will we be using to match observations from one dataset to the other? *hhid – “Household Identifier”*

5. Collapse and Merge
Merge (same number of observations)

- Which variable will we be using to match observations from one dataset to the other? *hhid – "Household Identifier"*

- Code: *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta"*

5. Collapse and Merge
Merge (same number of observations)

- *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta"*

```
.                  merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh  Mukerjee\Stata_t
> raining_PNG2025\Stata_training_png_March2025\Datasets\hh_3_4.dta"       //Merge the dataset with
>  the s33 dataset from Lesson 1, matching based on hhid (household ID)
(label community already defined)

    Result                      Number of obs
    -----------------------------------------
    Not matched                             0
    Matched                             2,699  (_merge==3)
    -----------------------------------------
```

5. Collapse and Merge
Merge (same number of observations)

- *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta"*

```
.                  merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh  Mukerjee\Stata_t
> raining_PNG2025\Stata_training_png_March2025\Datasets\hh_3_4.dta"       //Merge the dataset with
>  the s33 dataset from Lesson 1, matching based on hhid (household ID)
(label community already defined)

    Result                      Number of obs
    -----------------------------------------
    Not matched                             0
    Matched                             2,699  (_merge==3)
    -----------------------------------------
```

- Stata output tells us how many observations were and were not matched. How many were matched?

5. Collapse and Merge
Merge (same number of observations)

- *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta"*

```
.                 merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh  Mukerjee\Stata_t
> raining_PNG2025\Stata_training_png_March2025\Datasets\hh_3_4.dta"       //Merge the dataset with
>  the s33 dataset from Lesson 1, matching based on hhid (household ID)
(label community already defined)

    Result                      Number of obs
    -----------------------------------------
    Not matched                             0
    Matched                             2,699  (_merge==3)
    -----------------------------------------
```

- Stata output tells us how many observations were and were not matched. How many were matched? *2,699 (all of them!)*

5. Collapse and Merge
Merge (same number of observations)

- *merge 1:1 hhid using “C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta”*

```
.                 merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh  Mukerjee\Stata_t
> raining_PNG2025\Stata_training_png_March2025\Datasets\hh_3_4.dta"       //Merge the dataset wit
>  the s33 dataset from Lesson 1, matching based on hhid (household ID)
(label community already defined)

    Result                      Number of obs
    -----------------------------------------
    Not matched                             0
    Matched                             2,699  (_merge==3)
    -----------------------------------------
```

- With every merge, Stata creates a new variable called *_merge*

5. Collapse and Merge
Merge (same number of observations)

- *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta"*

```
.                  merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh  Mukerjee\Stata_t
> raining_PNG2025\Stata_training_png_March2025\Datasets\hh_3_4.dta"       //Merge the dataset with
>  the s33 dataset from Lesson 1, matching based on hhid (household ID)
(label community already defined)

    Result                      Number of obs
    -----------------------------------------
    Not matched                             0
    Matched                             2,699  (_merge==3)
    -----------------------------------------
```

- With every merge, Stata creates a new variable called *_merge*
  - *_merge* equals 1 or 2 for un-matched observations
  - *_merge* equals 3 for matched observations

5. Collapse and Merge
Merge (same number of observations)

- *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta"*

```
.                 merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh  Mukerjee\Stata_t
> raining_PNG2025\Stata_training_png_March2025\Datasets\hh_3_4.dta"       //Merge the dataset with
>  the s33 dataset from Lesson 1, matching based on hhid (household ID)
(label community already defined)

    Result                      Number of obs
    -----------------------------------------
    Not matched                             0
    Matched                             2,699  (_merge==3)
    -----------------------------------------
```

- With every merge, Stata creates a new variable called *_merge*
  - *_merge* equals 1 or 2 for un-matched observations
  - *_merge* equals 3 for matched observations
  - If all matched, it's a good practice to *drop _merge*

5. Collapse and Merge
Merge (same number of observations)

- Question: On average, do households with a thatched roof or a corrugated roof have more children?

5. Collapse and Merge
Merge (same number of observations)

- Question: On average, do households with a thatched roof or a corrugated roof have more children?

- Option 1: two "sum…if" codes

5. Collapse and Merge
Merge (same number of observations)

- Question: On average, do households with a thatched roof or a corrugated roof have more children?

- Option 1: two "sum…if" codes
  - *sum perc_children if i1==1*
  - *sum perc_children if i1==2*

5. Collapse and Merge
Merge (same number of observations)

- Question: On average, do households with a thatched roof or a corrugated roof have more children?

- Option 1: two "sum…if" codes
  - *sum perc_children if i1==1 – on average in households with thatched roofs, 43% of the household members are aged 15 years or younger*
  - *sum perc_children if i1==2 - on average in households with corrugated metal roofs, 38% of the household members are aged 15 years or younger*

5. Collapse and Merge
Merge (same number of observations)

- Question: On average, do households with a thatched roof or a corrugated roof have more children?

- Option 2: bysort
  - Bysort repeats a Stata command on a subset of data

5. Collapse and Merge
Merge (same number of observations)

- Question: On average, do households with a thatched roof or a corrugated roof have more children?

- Option 2: bysort
  - Bysort repeats a Stata command on a subset of data
  - We can repeat the "summarize perc_children" command on all of the different values of i1

5. Collapse and Merge
Merge (same number of observations)

- Question: On average, do households with a thatched roof or a corrugated roof have more children?

- Option 2: bysort
  - Bysort repeats a Stata command on a subset of data
  - We can repeat the "summarize perc_children" command on all of the different values of i1
  - Code: *bysort i1: sum perc_children*

5. Collapse and Merge
Merge (same number of observations)

- Question: On average, do households with a thatched roof or a corrugated roof have more children?

- *bysort i1: sum perc_children*

```
. bysort i1: sum perc_children

-> i1 = Thatched roof (grass, palm fronds, sago leaves, etc.)

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
perc_child~n |      1,387      .43039    .2671725          0          1

-> i1 = Corrugated metal roof

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
perc_child~n |      1,209    .3842051    .2353648          0          1
```

5. Collapse and Merge
Merge (same number of observations)

- Question: On average, do households with a thatched roof or a corrugated roof have more children?

- Answer:

5. Collapse and Merge
Merge (same number of observations)

- Question: On average, do households with a thatched roof or a corrugated roof have more children?

- Answer: *Descriptive statistics suggest that households with thatched roofs have a higher percentage of children than households with corrugated metal roofs.*

5. Collapse and Merge
Merge (different number of observations)

- What if we want to know how many people in our dataset live in a house with a thatched roof?

5. Collapse and Merge
Merge (different number of observations)

- What if we want to know how many people in our dataset live in a house with a thatched roof?

- We would want to look at the roof type data at the person-level

5. Collapse and Merge
Merge (different number of observations)

- What if we want to know how many people in our dataset live in a house with a thatched roof?
- We would want to look at the roof type data at the person-level
- We can merge the housing characteristics dataset to the household roster dataset, maintaining the dataset at the *person-level (16,192 people)*

5. Collapse and Merge
Merge (different number of observations)

Hh_3_4 (n=2,699)	
hhid	i1
0100RSPN	metal roof

5. Collapse and Merge
Merge (different number of observations)

Hh_3_4 (n=2,699)	
hhid	i1
0100RSPN	metal roof

s11_long (n=16,192)		
hhid	a_pid	a4
0100RSPN	1	M
0100RSPN	2	F
0100RSPN	3	F
0100RSPN	4	F
0100RSPN	5	M
0100RSPN	6	M
0100RSPN	7	M

5. Collapse and Merge
Merge (different number of observations)

Hh_3_4 (n=2,699)	
hhid	i1
0100RSPN	metal roof

s11_long (n=16,192)		
hhid	a_pid	a4
0100RSPN	1	M
0100RSPN	2	F
0100RSPN	3	F
0100RSPN	4	F
0100RSPN	5	M
0100RSPN	6	M
0100RSPN	7	M

merged data (n=6,057)			
hhid	a_pid	a4	i1
0100RSPN	1	M	metal roof
0100RSPN	2	F	metal roof
0100RSPN	3	F	metal roof
0100RSPN	4	F	metal roof
0100RSPN	5	M	metal roof
0100RSPN	6	M	metal roof
0100RSPN	7	M	metal roof

5. Collapse and Merge
Merge (different number of observations)

Hh_3_4 (n=2,699)	
hhid	i1
0100RSPN	metal roof

s11_long (n=16,192)		
hhid	a_pid	a4
0100RSPN	1	M
0100RSPN	2	F
0100RSPN	3	F
0100RSPN	4	F
0100RSPN	5	M
0100RSPN	6	M
0100RSPN	7	M

merged data (n=6,057)			
hhid	a_pid	a4	i1
0100RSPN	1	M	metal roof
0100RSPN	2	F	metal roof
0100RSPN	3	F	metal roof
0100RSPN	4	F	metal roof
0100RSPN	5	M	metal roof
0100RSPN	6	M	metal roof
0100RSPN	7	M	metal roof

5. Collapse and Merge
Merge (different number of observations)

- Let’s start with the household characteristics dataset this time

5. Collapse and Merge
Merge (different number of observations)

- Let's start with the household characteristics dataset this time
  - Close out of Stata

5. Collapse and Merge
Merge (different number of observations)

- Let's start with the household characteristics dataset this time
  - Close out of Stata
  - Double click on *hh_3_4.dta* to open this file

5. Collapse and Merge
Merge (different number of observations)

- Let's start with the household characteristics dataset this time
  - Close out of Stata
  - Double click on *hh_3_4.dta* to open this file
  - What level are the observations in this dataset?

5. Collapse and Merge
Merge (different number of observations)

- Let's start with the household characteristics dataset this time
  - Close out of Stata
  - Double click on *hh_3_4.dta* to open this file
  - What level are the observations in this dataset? *Household-level*

5. Collapse and Merge
Merge (different number of observations)

- Let's start with the household characteristics dataset this time
  - Close out of Stata
  - Double click on *hh_3_4.dta* to open this file
  - What level are the observations in this dataset? *Household-level*

- We are merging it with s11_long.dta (the household roster file we have been working with today). What level are the observations in this dataset?

5. Collapse and Merge
Merge (different number of observations)

- Let's start with the household characteristics dataset this time
  - Close out of Stata
  - Double click on *hh_3_4.dta* to open this file
  - What level are the observations in this dataset? *Household-level*

- We are merging it with s11_long.dta (the household roster file we have been working with today). What level are the observations in this dataset? *Person-level*

5. Collapse and Merge
Merge (different number of observations)

- Before, our merge command was: *merge 1:1 hhid using [datafile]*

5. Collapse and Merge
Merge (different number of observations)

- Before, our merge command was: *merge 1:1 hhid using [datafile]*

- What part of the command do you think we will change for this merge?

5. Collapse and Merge
Merge (different number of observations)

- Before, our merge command was: *merge 1:1 hhid using [datafile]*

- What part of the command do you think we will change for this merge? *Change 1:1 to 1:m*

5. Collapse and Merge
Merge (different number of observations)

- Before, our merge command was: *merge 1:1 hhid using [datafile]*
- What part of the command do you think we will change for this merge? *Change 1:1 to 1:m*
- It's now "1:m" because we merging ONE household observation (in the HH characteristics file) to MANY household observations (one per person in the household roster file)

5. Collapse and Merge
Merge (different number of observations)

Hh_3_4 (n=2,699)	
hhid	i1
0100RSPN	[illegible]

ONE household obs.

s11_long (n=16,192)		
hhid	a_pid	a4
0100RSPN	1	M
0100RSPN	2	F
0100RSPN	3	F
0100RSPN	4	F
0100RSPN	[illegible]	[illegible]
0100RSPN	[illegible]	[illegible]
0100RSPN	[illegible]	[illegible]

MANY household obs.

merged data (n=6,057)			
hhid	a_pid	a4	i1
0100RSPN	1	M	metal roof
0100RSPN	2	F	metal roof
0100RSPN	3	F	metal roof
0100RSPN	4	F	metal roof
0100RSPN	5	M	metal roof
0100RSPN	6	M	metal roof
0100RSPN	7	M	metal roof

5. Collapse and Merge
Merge (different number of observations)

- Code: *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta"*

5. Collapse and Merge
Merge (different number of observations)

- Code: *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta"*

- We are still using *hhid* as our matching identifier across the two datasets

5. Collapse and Merge
Merge (different number of observations)

- *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta"*

```
Result                      Number of obs
-----------------------------------------
Not matched                             0
Matched                            16,192  (_merge==3)
-----------------------------------------
```

5. Collapse and Merge
Merge (different number of observations)

- *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta"*

```
Result                      Number of obs
-----------------------------------------
Not matched                             0
Matched                            16,192  (_merge==3)
-----------------------------------------
```

- <u>Question:</u> Did all of the observations match?

5. Collapse and Merge
Merge (different number of observations)

- *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta"*

```
Result                          Number of obs
---------------------------------------------
Not matched                                 0
Matched                                16,192  (_merge==3)
---------------------------------------------
```

- <u>Question:</u> Did all of the observations match? *Yes (all 16,192)*

5. Collapse and Merge
Merge (different number of observations)

- *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\ STATA_training_PNG2025\ Stata_training_png_March2025\datasets\hh_3_4.dta"*

```
Result                          Number of obs
---------------------------------------------
Not matched                                 0
Matched                                16,192  (_merge==3)
---------------------------------------------
```

- Question: Did all of the observations match? *Yes (all 16,192)*
- Question: What are the values of the new *_merge* variable?

5. Collapse and Merge
Merge (different number of observations)

- *merge 1:1 hhid using "C:\Users\RMukerjee\IFPRI Dropbox\Rishabh Mukerjee\STATA_training_PNG2025\Stata_training_png_March2025\datasets\hh_3_4.dta"*

```
Result                           Number of obs
-----------------------------------------------
Not matched                                  0
Matched                                 16,192  (_merge==3)
-----------------------------------------------
```

- Question: Did all of the observations match? *Yes (all 16,192)*
- Question: What are the values of the new *_merge* variable?

All 3s

5. Collapse and Merge
Merge (different number of observations)

- Question: How many people in the dataset live in houses with a thatched roof?

5. Collapse and Merge
Merge (different number of observations)

- Question: How many people in the dataset live in houses with a thatched roof?

- Code:

5. Collapse and Merge
Merge (different number of observations)

- Question: How many people in the dataset live in houses with a thatched roof?

- Code: *tab i1*

5. Collapse and Merge
Merge (different number of observations)

- Question: How many people in the dataset live in houses with a thatched roof?
- Code: *tab i1*
- Answer:

5. Collapse and Merge
Merge (different number of observations)

- Question: How many people in the dataset live in houses with a thatched roof?
- Code: *tab i1*
- Answer: *8,328 people (51.43% of people in the sample live in houses with a thatched roof)*

5. Collapse and Merge
Merge (different number of observations)

- Question: How many *women* in the dataset live in houses with a thatched roof?

5. Collapse and Merge
Merge (different number of observations)

- Question: How many *women* in the dataset live in houses with a thatched roof?

- Code:

5. Collapse and Merge
Merge (different number of observations)

- Question: How many *women* in the dataset live in houses with a thatched roof?

- Code:
  - *tab i1 if a3==2*
    - Can use codebook a3 to remember how “woman” is coded
  - *tab i1 a3, column*
  - *tab a3 i1, row*

5. Collapse and Merge
Merge (different number of observations)

- Question: How many *women* in the dataset live in houses with a thatched roof?
- Code:
  - *tab i1 if a3==2*
    - Can use codebook a3 to remember how “woman” is coded
  - *tab i1 a3, column*
  - *tab a3 i1, row*
- Answer:

5. Collapse and Merge
Merge (different number of observations)

- Question: How many *women* in the dataset live in houses with a thatched roof?
- Code:
  - *tab i1 if a3==2*
    - Can use codebook a3 to remember how "woman" is coded
  - *tab i1 a3, column*
  - *tab a3 i1, row*
- Answer: *3,961 women live in a house with a thatched roof (51.22% of women in the sample live in a house with a thatched roof)*