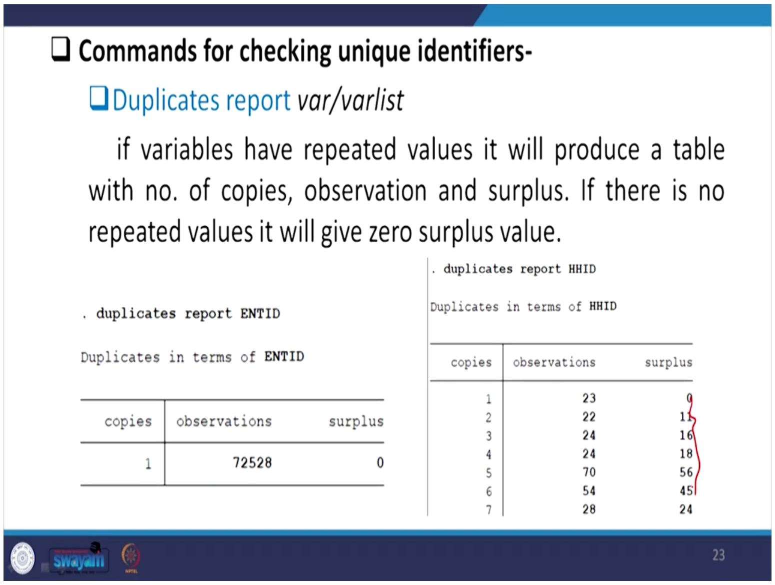
Handling Large-Scale Unit Level Data Using STATA

Professor Pratap C. Mohanty
Department of Humanities and Social Sciences,
Indian Institute of Technology, Roorkee

Lecture 14
Managing Data in Stata - II

Welcome friends once again to the NPTEL module on Handling Large-Scale Data Using Stata. We are trying to explain the use of Stata with the core data, the original data. I have already guided you in the last class about some commands, some basic commands, like summarize, codebook, listing, browsing. All those things we have done in the last class. So as a continuous into our previous lecture, we have mentioned Managing Data in Stata II. So here we are going to guide you further. Last class we have shown you codebook and its compact details.

(Refer Slide Time: 01:22)



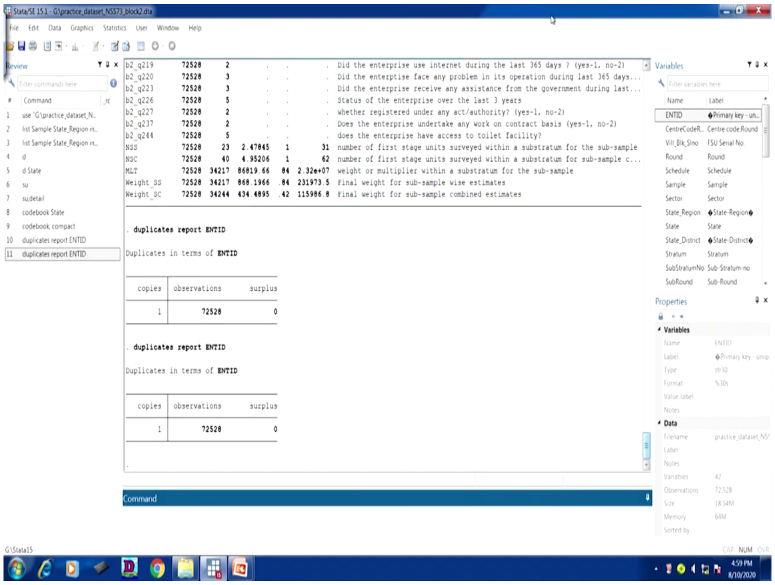
❑ Commands for checking unique identifiers-

❑ Duplicates report *var/varlist*

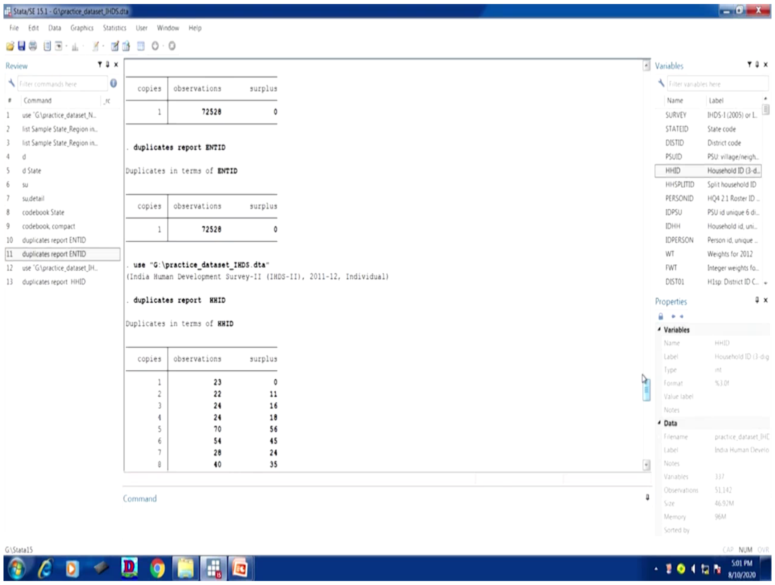
if variables have repeated values it will produce a table with no. of copies, observation and surplus. If there is no repeated values it will give zero surplus value.

```
. duplicates report ENTID

Duplicates in terms of ENTID

--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |        72528             0
--------------------------------------
```

```
. duplicates report HHID

Duplicates in terms of HHID

--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |           23             0
        2 |           22            11
        3 |           24            16
        4 |           24            18
        5 |           70            56
        6 |           54            45
        7 |           28            24
```



```
b2_q219      72528      2          .    .          .  Did the enterprise use internet during the last 365 days ? (yes-1, no-2)
b2_q220      72528      3          .    .          .  Did the enterptise face any problem in its operation during last 365 days...
b2_q223      72528      3          .    .          .  Did the enterprise receive any assistance from the government during last...
b2_q226      72528      5          .    .          .  Status of the enterprise over the last 3 years
b2_q227      72528      2          .    .          .  whether registered under any act/authority? (yes-1, no-2)
b2_q237      72528      2          .    .          .  Does the enterprise undertake any work on contract basis (yes-1, no-2)
b2_q244      72528      5          .    .          .  does the enterprise have access to toilet facility?
NSS          72528     23    2.47845    1         31  number of first stage units surveyed within a substratum for the sub-sample
NSC          72528     40    4.95206    1         62  number of first stage units surveyed within a substratum for sub-sample c...
MLT          72528  34217   86819.66   84   2.32e+07  weight or multiplier within a substratum for the sub-sample
Weight_SS    72528  34217   868.1966  .84   231973.5  Final weight for sub-sample wise estimates
Weight_SC    72528  34244   434.4895  .42   115986.8  Final weight for sub-sample combined estimates


. duplicates report ENTID

Duplicates in terms of ENTID

--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |        72528             0
--------------------------------------

. duplicates report ENTID

Duplicates in terms of ENTID

--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |        72528             0
--------------------------------------

.
```



```
--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |        72528             0
--------------------------------------

. duplicates report ENTID

Duplicates in terms of ENTID

--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |        72528             0
--------------------------------------

. use "G:\practice_dataset_IHDS.dta"
(India Human Development Survey-II (IHDS-II), 2011-12, Individual)

. duplicates report  HHID

Duplicates in terms of HHID

--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |           23             0
        2 |           22            11
        3 |           24            16
        4 |           24            18
        5 |           70            56
        6 |           54            45
        7 |           28            24
        8 |           40            35
```

I am going to tell you the commands for checking unique identifiers. I told you that identifier is very important for analysis. Which are the unique identifiers, how duplicate reports are important for analysis. Like, if you enter duplicate report, let me clarify what you mean by that. If a variable has a repeated value, it will produce a table with number of copies, there is observations, likewise, visible in our table here, these are sample result we developed, copies, observation and surplus. If there are no repeated values, it will simply give you zero surplus.

So zero surplus means all the observations are uniquely identified, likewise this. So duplicate report, I am just going to click that for you to operate duplicate, there is a mistake, let me clarify, just a minute, so duplicates report space, then the variable we wanted to check, now enter. So this gives information about the copies, the variable we wanted to check has zero surplus. So zero surplus means no duplicates are there in our data.

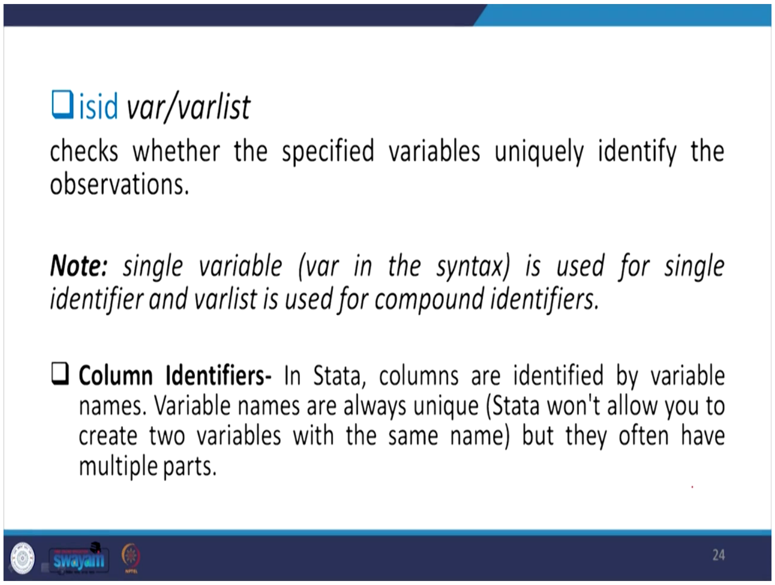
Let me go through once again. We can also take this with our IHDS dataset. How many duplicates are there? So, IHDS dataset we can check like here. Let me open that in our window. So here is the IHDS one, let me click ok. So here is the IHDS sample data we have just opened. from the IHDS data, again if I go by the same command duplicates report, what you do, you can go by page up command. I have the command here duplicates report here. instead of this, what I do, I will enter the ID for us.

If the second one is ID, HHID is there. So if I just enter it gives me the information like this. Here how many duplicates copies are generated. Look at so many duplicate copies are generated. So it suggests that there are duplicates. So just one variable is not uniquely identified, we have to go by a combination of variables for finding the unique ID.

So, it is not coming so many are there. what I suggest you to go through this interpretation, like if it is there like one copy, 23 observation, and zero surplus that means 23 observation are clearly unique ID. But there are two copies of such, 22 of such observations having two copies are generated. So two copies divided by two, 22 divided by two, we have 11 surplus left.

Similarly, for three copies, type of three copies are of 24, if I divided by three, so it will be eight. Eight already. two copies of 22 observations are there. If I just multiply then with that eight into two, 16 will be the surplus. So these kind of surplus information are visible before us. You can find out and rest if you operate, you will find out differences.

(Refer Slide Time: 06:31)



❑ isid *var/varlist*

checks whether the specified variables uniquely identify the observations.

Note: *single variable (var in the syntax) is used for single identifier and varlist is used for compound identifiers.*

- ❑ **Column Identifiers-** In Stata, columns are identified by variable names. Variable names are always unique (Stata won't allow you to create two variables with the same name) but they often have multiple parts.

Let us understand the easiest way of tracking the unique IDs. Let me point out very clearly that it is very important for you to merge the data with another dataset. If you want to merge, unique IDs are very essential. Or even sometimes if you try to make merging for panel information, it is very very essential or some blocks in NSS data there are so many blocks information they give, if you wanted to simply merge those blocks, you need unique ID as well.

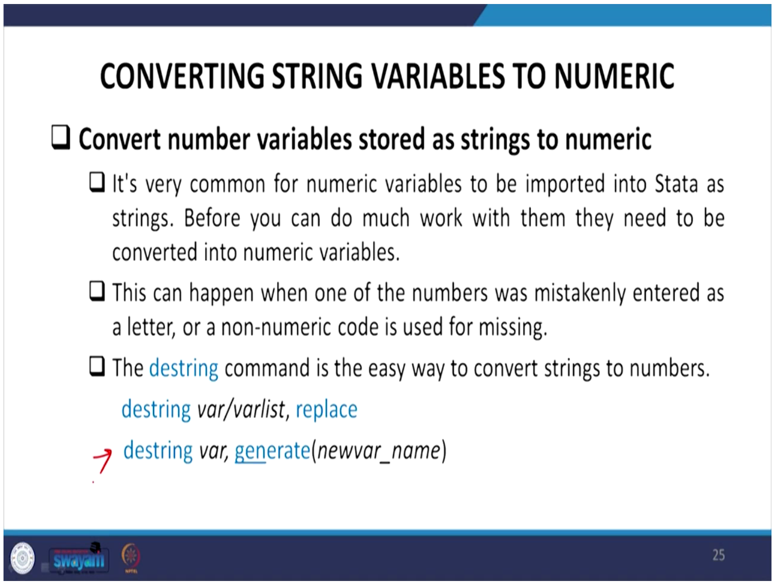
Now, check this ISID checks whether the specified variables uniquely identify the observation or not. So we have already told you if surplus is there that means those are not uniquely identified, that number of copies are generated, so those are not uniquely identified. One, let me take a note here that single variable that is variable in the syntax command var command or one variable in the command is used for single identifier.

And if you include varlist, it include the compound identifiers, list of variables if you give that may give that. Combinedly they are the uniquely identifier variables that generate a single variable which is uniquely identified. But individually there may not be, let us test it. In column identifiers, what do you mean by column identifiers? Rows we have already told you they are the observations and in Stata columns are identified by variable names. Variables are generally entered in columns. Variable names are always unique. You cannot have same variable names together in Stata. It always popped up with an error command or to suggest you to change the name since the same name already exist.

So Stata would not allow you to create two variables of the same name. But they often have multiple parts, like the name is there, if you just add another variable with an underscore or with another addition, you have another part of that particular variable to identify whether that is a different variable or not, but connected with the original variable.

We will operate everything steadily in our successive classes. We will keep on using those commands. But at this moment, I am just guiding you which are the simplest way of clarifying some entries.

(Refer Slide Time: 09:18)

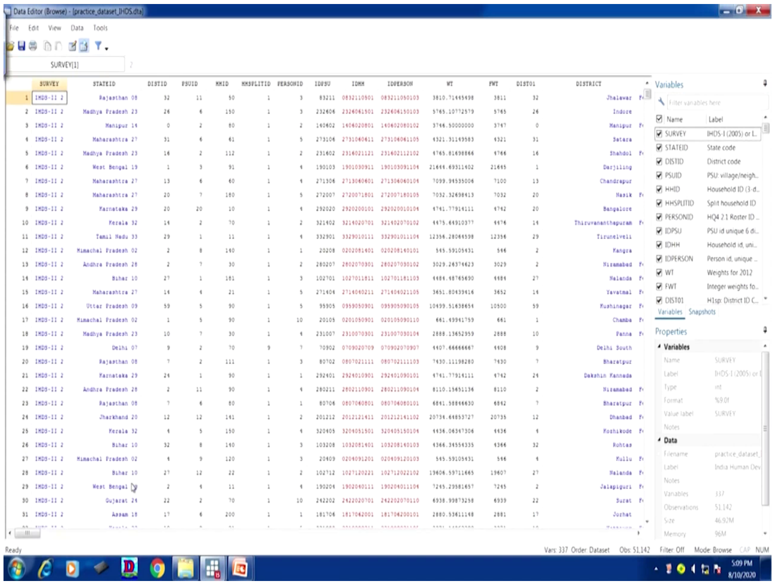


CONVERTING STRING VARIABLES TO NUMERIC

- ❑ Convert number variables stored as strings to numeric
  - ❑ It's very common for numeric variables to be imported into Stata as strings. Before you can do much work with them they need to be converted into numeric variables.
  - ❑ This can happen when one of the numbers was mistakenly entered as a letter, or a non-numeric code is used for missing.
  - ❑ The destring command is the easy way to convert strings to numbers.

    destring *var/varlist*, replace

    destring *var*, generate(*newvar_name*)



	SURVEY	STATEID	DISTID	PSUID	HHID	HHSPLITID	PERSONID	IDPSU	IDHH	IDPERSON	WT	FWT	DIST01	DISTRICT	
1	IHDS-II 2	Rajasthan 08	32	11	50	1	3	83211	0832110501	083211050103	3810.71445498	3811	32	Jhalawar	F
2	IHDS-II 2	Madhya Pradesh 23	26	6	150	1	3	232606	2326061501	232606150103	5765.10772579	5765	26	Indore	
3	IHDS-II 2	Manipur 14	0	2	80	1	2	140602	1406020801	140602080102	3746.50000000	3747	0	Manipur	F
4	IHDS-II 2	Maharashtra 27	31	6	61	1	5	273106	2731060611	273106061105	4321.31149583	4321	31	Satara	
5	IHDS-II 2	Madhya Pradesh 23	16	2	112	1	2	231602	2316021121	231602112102	4765.81698866	4766	16	Shahdol	F
6	IHDS-II 2	West Bengal 19	1	3	91	1	4	190103	1901030911	190103091104	21644.69311402	21645	1	Darjiling	
7	IHDS-II 2	Maharashtra 27	13	6	60	1	4	271306	2713060601	271306060104	7099.94535006	7100	13	Chandrapur	
8	IHDS-II 2	Maharashtra 27	20	7	180	1	5	272007	2720071801	272007180105	7032.32698413	7032	20	Nasik	F
9	IHDS-II 2	Karnataka 29	20	20	10	1	4	292020	2920200101	292020010104	4741.77914111	4742	20	Bangalore	
10	IHDS-II 2	Kerala 32	14	2	70	1	2	321402	3214020701	321402070102	4475.64910377	4476	14	Thiruvananthapuram	F
11	IHDS-II 2	Tamil Nadu 33	29	1	11	1	4	332901	3329010111	332901011104	12356.28064598	12356	29	Tirunelveli	
12	IHDS-II 2	Himachal Pradesh 02	2	8	140	1	1	20208	0202081401	020208140101	545.59105431	546	2	Kangra	
13	IHDS-II 2	Andhra Pradesh 28	2	7	30	1	2	280207	2802070301	280207030102	3029.26374623	3029	2	Nizamabad	F
14	IHDS-II 2	Bihar 10	27	1	181	1	3	102701	1027011811	102701181103	4484.48765690	4484	27	Nalanda	F
15	IHDS-II 2	Maharashtra 27	14	4	21	1	5	271404	2714040211	271404021105	3651.80439416	3652	14	Yavatmal	F
16	IHDS-II 2	Uttar Pradesh 09	59	5	90	1	5	95905	0959050901	095905090105	10499.51638654	10500	59	Kushinagar	F
17	IHDS-II 2	Himachal Pradesh 02	1	5	90	1	10	20105	0201050901	020105090110	661.49941759	661	1	Chamba	F
18	IHDS-II 2	Madhya Pradesh 23	10	7	30	1	4	231007	2310070301	231007030104	2888.13652959	2888	10	Panna	F
19	IHDS-II 2	Delhi 07	9	2	70	9	7	70902	0709020709	070902070907	4407.66666667	4408	9	Delhi South	
20	IHDS-II 2	Rajasthan 08	7	2	111	1	3	80702	0807021111	080702111103	7430.11198280	7430	7	Bharatpur	
21	IHDS-II 2	Karnataka 29	24	1	90	1	1	292401	2924010901	292401090101	4741.77914111	4742	24	Dakshin Kannada	
22	IHDS-II 2	Andhra Pradesh 28	2	11	90	1	4	280211	2802110901	280211090104	8110.15651136	8110	2	Nizamabad	F
23	IHDS-II 2	Rajasthan 08	7	6	80	1	1	80706	0807060801	080706080101	6841.58846630	6842	7	Bharatpur	F
24	IHDS-II 2	Jharkhand 20	12	12	141	1	2	201212	2012121411	201212141102	20734.64853727	20735	12	Dhanbad	F
25	IHDS-II 2	Kerala 32	4	5	150	1	4	320405	3204051501	320405150104	4436.06347306	4436	4	Kozhikode	F
26	IHDS-II 2	Bihar 10	32	8	140	1	3	103208	1032081401	103208140103	4366.34554335	4366	32	Rohtas	
27	IHDS-II 2	Himachal Pradesh 02	4	9	120	1	3	20409	0204091201	020409120103	545.59105431	546	4	Kullu	F
28	IHDS-II 2	Bihar 10	27	12	22	1	2	102712	1027120221	102712022102	19606.59711665	19607	27	Nalanda	F
29	IHDS-II 2	West Bengal 19	2	4	11	1	4	190204	1902040111	190204011104	7245.29581657	7245	2	Jalapiguri	F
30	IHDS-II 2	Gujarat 24	22	2	70	1	10	242202	2422020701	242202070110	6938.99873258	6939	22	Surat	F
31	IHDS-II 2	Assam 18	17	6	200	1	1	181706	1817062001	181706200101	2880.53611148	2881	17	Jorhat	

☑ Name	Label
☑ SURVEY	IHDS-I (2005) or I...
☑ STATEID	State code
☑ DISTID	District code
☑ PSUID	PSU: village/neigh...
☑ HHID	Household ID (3-d...
☑ HHSPLITID	Split household ID
☑ PERSONID	HQ4 2.1 Roster ID ...
☑ IDPSU	PSU id unique 6 di...
☑ IDHH	Household id, uni...
☑ IDPERSON	Person id, unique ...
☑ WT	Weights for 2012
☑ FWT	Integer weights fo...
☑ DIST01	H1sp: District ID C...

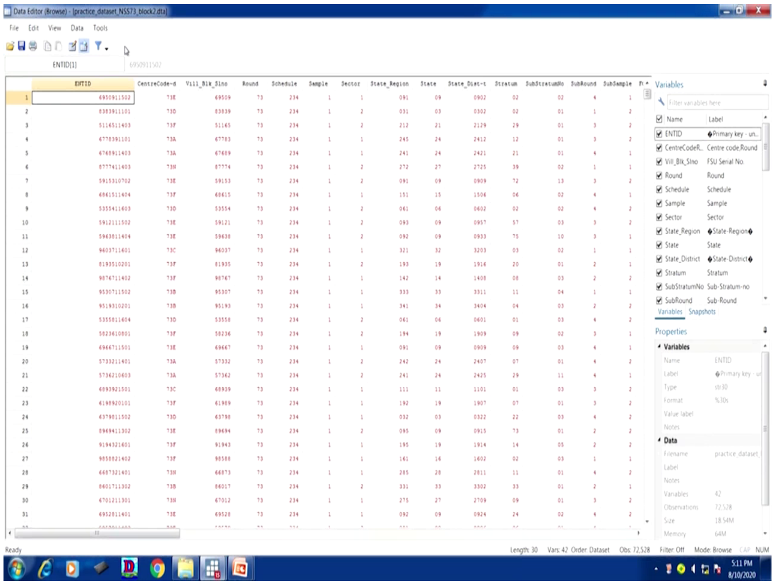
Variables

Name	SURVEY
Label	IHDS-I (2005) or I
Type	int
Format	%9.0f
Value label	SURVEY
Notes	

Data

Filename	practice_dataset_I
Label	India Human Dev
Notes	
Variables	337
Observations	51,142
Size	46.92M
Memory	96M

Ready   Vars: 337 Order: Dataset   Obs: 51,142   Filter: Off   Mode: Browse   CAP   NUM


Data Editor (Browse) - [practice_dataset_NSS73_block2.dta]
File Edit View Data Tools

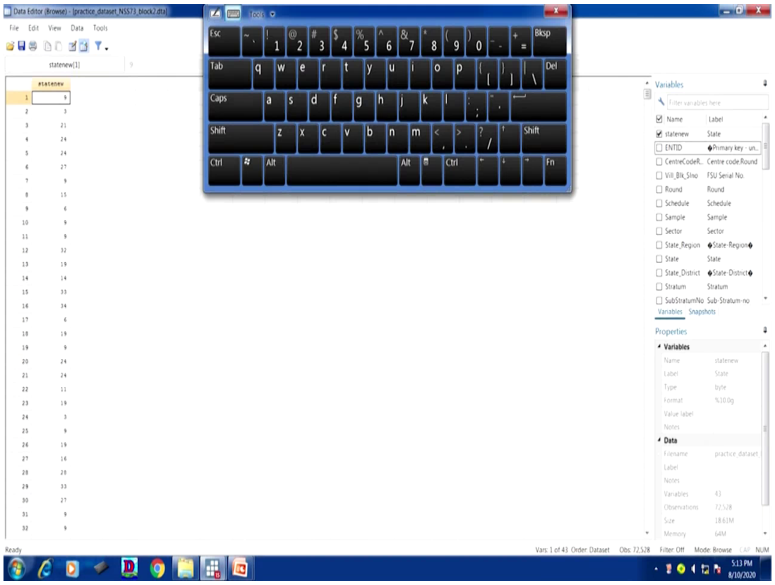
ENTID[1] 6950911502

	ENTID	CentreCode~d	Vill_Blk_Slno	Round	Schedule	Sample	Sector	State_Region	State	State_Dist~t	Stratum	SubStratumNo	SubRound	SubSample
1	6950911502	73E	69509	73	234	1	1	091	09	0902	02	02	4	1
2	8383911101	73D	83839	73	234	1	2	031	03	0302	02	01	1	2
3	5116511403	73F	51165	73	234	1	2	212	21	2129	29	01	3	2
4	6778391101	73A	67783	73	234	1	1	245	24	2412	12	01	3	2
5	6768911403	73A	67689	73	234	1	1	241	24	2421	21	01	4	1
6	8777411403	73N	87774	73	234	1	2	272	27	2725	39	02	1	1
7	5915310702	73E	59153	73	234	1	2	091	09	0909	72	13	3	2
8	6861511404	73F	68615	73	234	1	1	151	15	1506	06	02	4	1
9	5355411603	73D	53554	73	234	1	2	061	06	0602	02	02	4	2
10	5912111502	73E	59121	73	234	1	2	093	09	0957	57	03	3	2
11	5963811404	73E	59638	73	234	1	2	092	09	0933	75	10	3	1
12	9603711601	73C	96037	73	234	1	1	321	32	3203	03	02	1	1
13	8193510201	73F	81935	73	234	1	2	193	19	1916	20	01	2	1
14	9876711402	73F	98767	73	234	1	1	142	14	1408	08	03	2	2
15	9530711502	73B	95307	73	234	1	1	333	33	3311	11	04	1	1
16	9519310201	73B	95193	73	234	1	1	341	34	3404	04	03	2	2
17	5355811604	73D	53558	73	234	1	2	061	06	0601	01	03	4	2
18	5823610801	73F	58236	73	234	1	2	194	19	1909	09	02	3	1
19	6966711501	73E	69667	73	234	1	1	091	09	0909	09	03	4	1
20	5733211401	73A	57332	73	234	1	2	242	24	2407	07	01	4	2
21	5736210603	73A	57362	73	234	1	2	241	24	2425	29	11	4	1
22	6893921501	73C	68939	73	234	1	1	111	11	1101	01	03	3	2
23	6198920101	73F	61989	73	234	1	1	192	19	1907	07	01	3	2
24	6379811502	73D	63798	73	234	1	1	032	03	0322	22	03	4	2
25	8969411302	73E	89694	73	234	1	2	095	09	0915	73	01	2	2
26	9194321601	73F	91943	73	234	1	1	195	19	1914	14	05	2	2
27	9858821402	73F	98588	73	234	1	1	161	16	1602	02	03	1	1
28	6687321401	73N	66873	73	234	1	1	285	28	2811	11	01	4	2
29	8601711302	73B	86017	73	234	1	2	331	33	3302	33	01	2	1
30	6701211301	73N	67012	73	234	1	1	275	27	2709	09	01	3	2
31	6952811401	73E	69528	73	234	1	1	092	09	0924	24	02	4	2

☑ Name	Label
☑ ENTID	�Primary key - un...
☑ CentreCodeR...	Centre code,Round
☑ Vill_Blk_Slno	FSU Serial No.
☑ Round	Round
☑ Schedule	Schedule
☑ Sample	Sample
☑ Sector	Sector
☑ State_Region	�State-Region�
☑ State	State
☑ State_District	�State-District�
☑ Stratum	Stratum
☑ SubStratumNo	Sub-Stratum-no
☑ SubRound	Sub-Round

Variables	
Name	ENTID
Label	�Primary key - un
Type	str30
Format	%30s
Value label	
Notes	
Data	
Filename	practice_dataset_
Label	
Notes	
Variables	42
Observations	72,528
Size	18.54M
Memory	64M

Ready Length: 30 Vars: 42 Order: Dataset Obs: 72,528 Filter: Off Mode: Browse CAP NUM


Data Editor (Browse) - [practice_dataset_NSS73_block2.dta]
File Edit View Data Tools
Tools
Esc ~ ` ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 _ - + = Bksp
Tab q w e r t y u i o p { [}] | \ Del
Caps a s d f g h j k l : ; " ' ↵
Shift z x c v b n m < , > . ? / Shift
Ctrl Alt Alt Ctrl Fn

statenew[1] 9

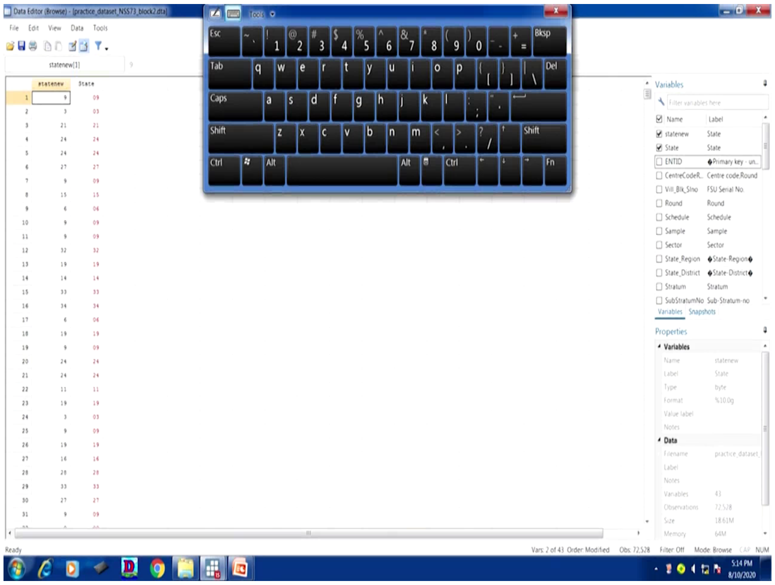
	statenew
1	9
2	3
3	21
4	24
5	24
6	27
7	9
8	15
9	6
10	9
11	9
12	32
13	19
14	14
15	33
16	34
17	6
18	19
19	9
20	24
21	24
22	11
23	19
24	3
25	9
26	19
27	16
28	28
29	33
30	27
31	9
32	9

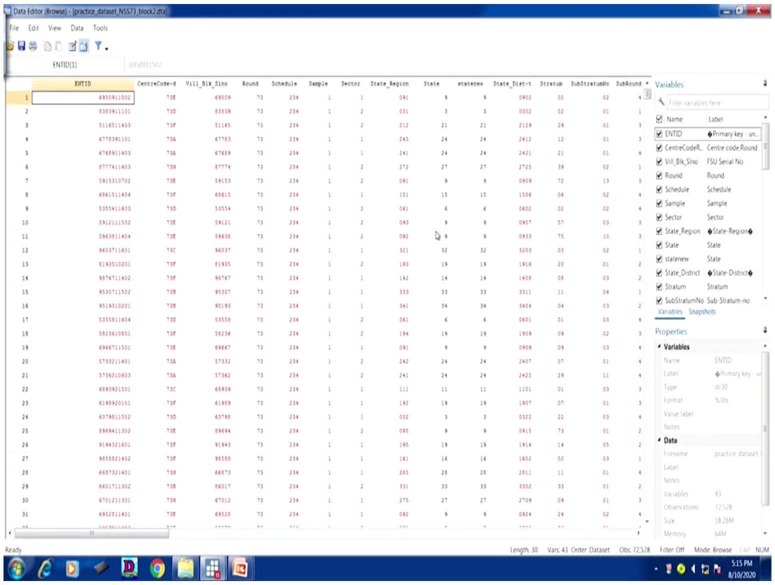
Variables
Filter variables here

☑ Name	Label
☑ statenew	State
□ ENTID	�Primary key - un...
□ CentreCodeR...	Centre code,Round
□ Vill_Blk_Slno	FSU Serial No.
□ Round	Round
□ Schedule	Schedule
□ Sample	Sample
□ Sector	Sector
□ State_Region	�State-Region�
□ State	State
□ State_District	�State-District�
□ Stratum	Stratum
□ SubStratumNo	Sub-Stratum-no

Variables	
Name	statenew
Label	State
Type	byte
Format	%10.0g
Value label	
Notes	
Data	
Filename	practice_dataset_
Label	
Notes	
Variables	43
Observations	72,528
Size	18.61M
Memory	64M

Ready Vars: 1 of 43 Order: Dataset Obs: 72,528 Filter: Off Mode: Browse CAP NUM


Data Editor (Browse) - [practice_dataset_NSS73_block2.dta]
File Edit View Data Tools
statenew[1]
statenew
State
Variables
Properties
Ready


Data Editor (Browse) - [practice_dataset_NSS73_block2.dta]
File Edit View Data Tools
ENTID[1]
ENTID
Variables
Properties
Ready

So let us convert string variable to numeric variable. As I already told you repeatedly that string variable, mathematical operations are just not possible directly. We need to convert first then can operate. Let me remind you once again that, there are some categorical variables also entered as string. So that can be further converted to label numeric values. So convert number variables stored as string to numeric. Basically those are numeric variables or number variables, but stored as string. We need to convert them to numeric to do our operation.

So it is very common for numeric variables to be imported into Stata as string. Before you can do much work with them, they need to be converted into numeric variable as I just mentioned. This can happen when one of the numbers was mistakenly entered as a later or a non-numeric code is used for missing. There are various ways of getting a string variable in our data. Besides the manual entry as string, we have also other possibilities like mentioned here. So mistakenly we can enter letter also in our data. So that converts the data to string. Similarly, a non-numeric code also uses for missing also read as string.

So let us understand how to convert it. The most important command here is destring. Destring command is highlighted in blue color, is the easiest way to convert string to the numbers. Variable is string and by various ways you can understand whether it is string, through codebook you can understand whether it is string or not, through summary statistics also you can understand whether it is string or not, simply browsing the data in the browser window, you can get whether it is string or not by color. if it is red, you will get string.

Once you have observed that variable is string, you need to destring it. You need to convert it to a numeric value. So the command is destring, then the variable name or variable list if you have so many variables and you simply replace, that will replace the original value and it will give you the new that is the destring value. But when you want both the variables together, you need to generate a new variable, let me operate here.

The simplest way to check whether it is string or not, I am just going to the data. This is the IHDS data. It is a bigger one. Let me start with, string variables are highlighted in red color. the blue color is in labeled numeric and the black color are in numeric values or continuous variables.

I can just change, suppose I know that state ID, I know that IDHH and ID person, these are in string data. But let me operate from the NSS at this moment. We will clarify later on from the IHDS data. At this moment, let me operate from the National Sample Survey, because we have already sampled and filtered the data for quick understanding.

So what I do, I will simply destring. I told you already destring, then variable name, then replace. So then you just go through destring, like this destring, then variable name. The variable are string or not, once again you can check. This is the variable. Let me come back once again I will open that, open once again the variable I will open the NSS then I will operate.

Let me check once that whether that is a string or not. And I know that there are so many variables in string. State, for example, state is in string. Let me convert it to a destring one. So the command is here destring, then state you just click on state. if I just add comma to it, comma then replace, it will replace.

It is always suggested that you keep both the variables. If you have already filtered and shortlisted your data, your dataset is very minimal, so try to keep both. Otherwise, if you are doubly sure your string data is not going to be useful at all and you go by a new variable then you may continue with that, continue with this command.

Let me first start with a destring then variable then generate. Then I will show you, like for example here, let me put a comma here, then generate, I am trying to generate a new variable. Then within bracket a variable name state, so statenew. Bracket is not there. Let me give the bracket, shift bracket, then bracket close. If I just enter here, destring state generate a new variable enter. look at another statenew variable is defined here. I am just putting my cursor. It might be visible to you, state then next statenew.

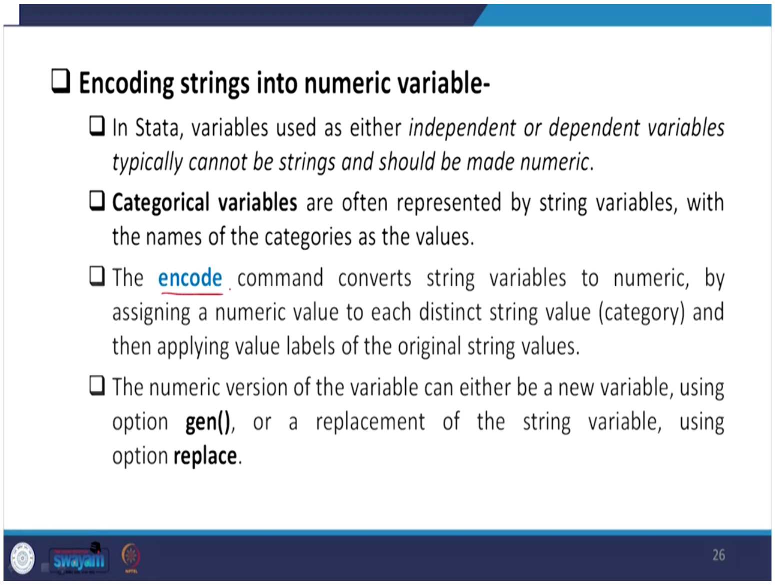
I will just show you from the browse window or for simplicity, for your simple understanding, br then statenew enter. It has shown you in numeric. this is no longer in red color. Whereas you just compare these two together instead of one just compare two together, like here you just enter state once again and enter, we will get both together. Initially it was state with red color, now statenew is in black color that means it has confirmed us that your data is converted to a numeric number. Now you can apply your mathematical operations for results.

Let me move on further. for your clarity, if I just do the same command destring, it is already defined. I am just going to do it, like if I just say replace instead of generating a new one, it will convert. Let us experiment it. So, destring state replace, what it gives, it has replaced the original string to the new state now is in numeric format. Look at this state is now here state you compare state and newstate. State since we have already replaced with a numeric variable, this is highlighted in black color.

Let me proceed further we have already clarified. So we will suggest that you please go by this then later on you can drop that particular variable if you want or you can replace later. But at this moment a new variable is suggested to generate and operate likewise we did.

It is also interesting to note that the string the way we did is not enough. In case of some categorical values are entered in the string format or string variable, then in that case encoding is important. Encoding string into numeric variable is very very important just string is not enough. String command is not going to work.

(Refer Slide Time: 19:25)



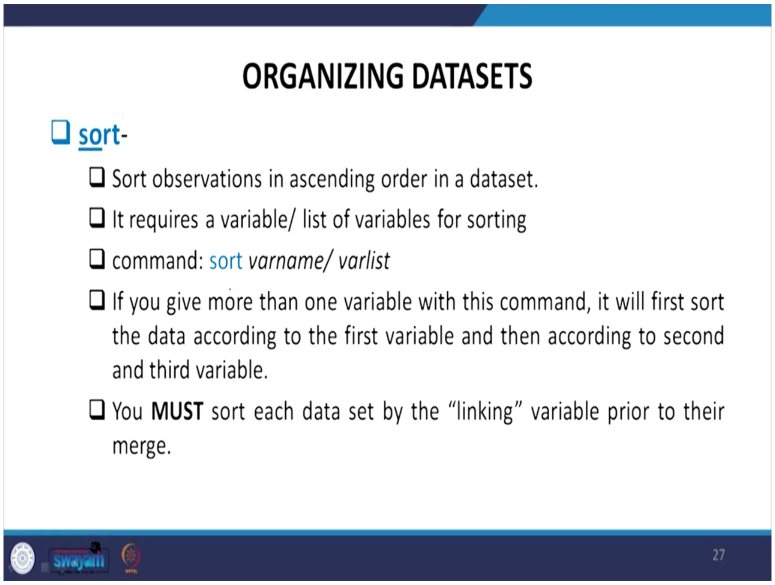
- ❑ **Encoding strings into numeric variable-**
  - ❑ In Stata, variables used as either *independent or dependent variables typically cannot be strings and should be made numeric.*
  - ❑ **Categorical variables** are often represented by string variables, with the names of the categories as the values.
  - ❑ The **encode** command converts string variables to numeric, by assigning a numeric value to each distinct string value (category) and then applying value labels of the original string values.
  - ❑ The numeric version of the variable can either be a new variable, using option **gen()**, or a replacement of the string variable, using option **replace**.

In Stata, variables used as either independent or dependent variable typically cannot be strings and should be made numeric. Typically, the string operation is not going to make string. But can be made numeric. These are generally categorical variables and are often represented by string variables, which we have already clarified in the last lecture with the names of categorical, categories as the values. The categories as the values are entered, but not numeric one.

So the encode command is in place of destring we did in the last slide. Encode, rest of the commands are same, encode is going to help you. Encode command converts string variables to numeric, by assigning a numeric value to each string that is categories and then applying value labels of the original string values. Once it has converted to a numeric value, we can then apply the label, on our own convenience based on the given variable description in the original data. So once the label is defined, we can also clarify in our converted data also.

The numeric version of the variable can either be a new variable using option, like the way we did generate the new variable name within brackets or a replacement or replace we did in the last slide likewise this you can do it for the string variable using an option replace.

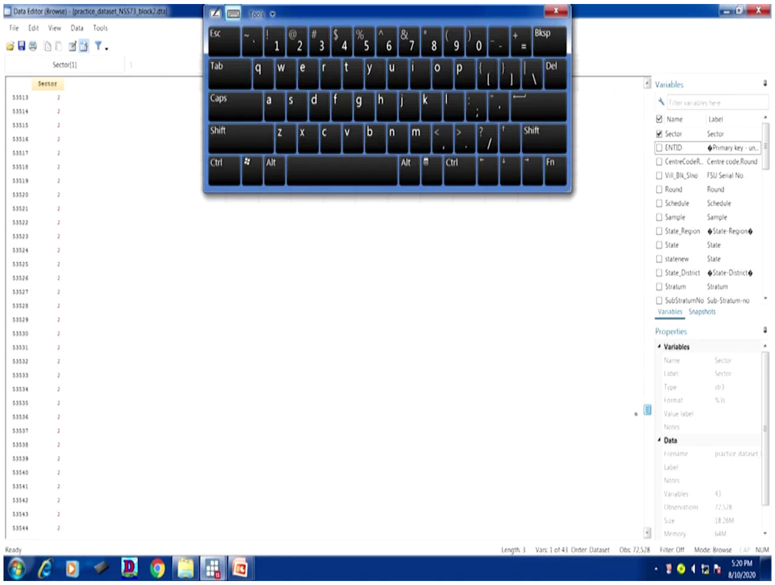
(Refer Slide Time: 21:12)



ORGANIZING DATASETS

❑ sort-

- ❑ Sort observations in ascending order in a dataset.
- ❑ It requires a variable/ list of variables for sorting
- ❑ command: sort *varname/ varlist*
- ❑ If you give more than one variable with this command, it will first sort the data according to the first variable and then according to second and third variable.
- ❑ You **MUST** sort each data set by the "linking" variable prior to their merge.



Data Editor (Browse) - [practice_dataset_NSS73_block2.dta]

File
Edit
View
Data
Tools
Esc ~ ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 - + = Bksp
Tab q w e r t y u i o p { [}] | \ Del
Caps a s d f g h j k l : ; " '
Shift z x c v b n m < , > . ? / Shift
Ctrl Alt Alt Ctrl Fn
Tools

Sector[1]

	Sector
53513	2
53514	2
53515	2
53516	2
53517	2
53518	2
53519	2
53520	2
53521	2
53522	2
53523	2
53524	2
53525	2
53526	2
53527	2
53528	2
53529	2
53530	2
53531	2
53532	2
53533	2
53534	2
53535	2
53536	2
53537	2
53538	2
53539	2
53540	2
53541	2
53542	2
53543	2
53544	2

Variables

Filter variables here

Name	Label
☑ Sector	Sector
☐ ENTID	�Primary key - un...
☐ CentreCodeR...	Centre code,Round
☐ Vill_Blk_Slno	FSU Serial No.
☐ Round	Round
☐ Schedule	Schedule
☐ Sample	Sample
☐ State_Region	�State-Region�
☐ State	State
☐ statenew	State
☐ State_District	�State-District�
☐ Stratum	Stratum
☐ SubStratumNo	Sub-Stratum-no

Properties

Variables

Name	Sector
Label	Sector
Type	str3
Format	%3s
Value label	
Notes	

Data

Filename	practice_dataset
Label	
Notes	
Variables	43
Observations	72,528
Size	18.26M
Memory	64M

Ready   Length: 3   Vars: 1 of 43   Order: Dataset   Obs: 72,528   Filter: Off   Mode: Browse   CAP   NUM

We did so far one interesting part that is called string. Many researchers usually commit that mistake and get frustrated with the data, that the data is becoming difficult to convert. Some errors are coming. They repeatedly search here and there I think the last couple of slides are going to help you a lot in understanding the data and converting to a numeric. The standard rule is always to convert to a numeric number for quicker operation. There are some advanced commands that can also continue with the string value as well, but we are not dealing with that part.

Organizing dataset is another important aspect of understanding Stata and the large dataset. The sort variable, sort command of the variable is going to help you a lot. Sorting of, likewise in Stata, likewise in Microsoft Excel, we used to sort the variables. If there are so many variables just need to analyze together, we need to sort one and accordingly match whether others are following or not.

So sort observations in ascending order in a dataset is generally essential. It requires a variable or list of variables for sorting. If there are so many variables you are sorting together, it by default consider first as your sorting variable. And on the basis of the first variable, it sorts other variables also. You might have operated in excel also.

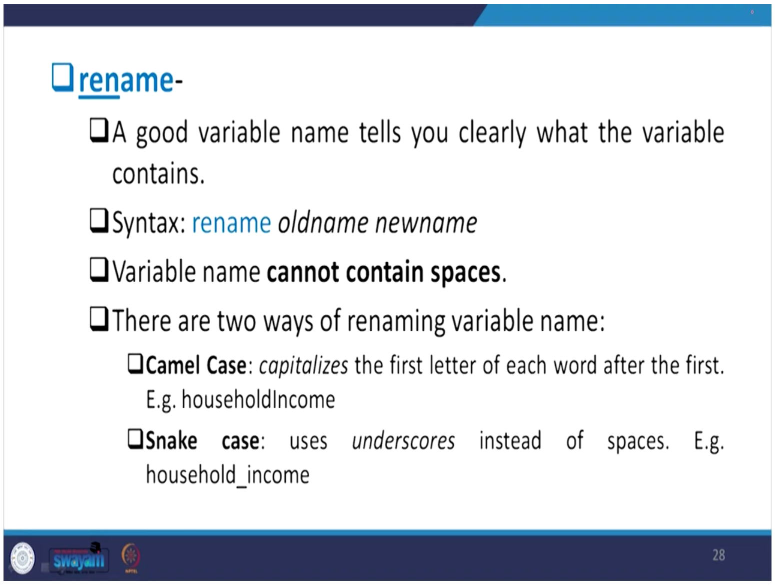
So the command for sort is sort then variable name. Let the variable is sector. Then how it is like this. Now look at what I do browse then sector, it will show you it has already sorted look at this I am going down, one then two is there. So since it is in ascending order it has already sorted.

So let me move on. If you give more than one variable with this command, it will by default sort the first variable and then accordingly saved the data with second and third variable as per the first variable. So you must sort each dataset by the linking variable prior to that of merging.

I think we will also clarify while merging and appending and we have some dedicated lectures also for it. We will clarify that it is an essential step before merging that you need to sort the data. Without sort, the variable, the uniquely identified variables cannot be sorted, it cannot match the corresponding unique ID of another dataset.

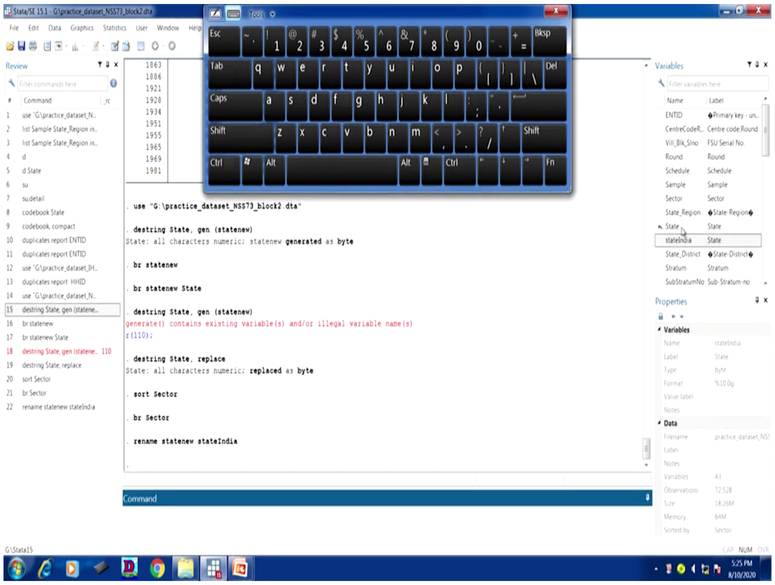
So once you have sorted it that means by each row all are same in both the datasets. So then only linking variables as well as the original variables both should be sorted before merging and then only merging is going to be successful.

(Refer Slide Time: 24:48)



❑ rename-

- ❑ A good variable name tells you clearly what the variable contains.
- ❑ Syntax: rename *oldname newname*
- ❑ Variable name **cannot contain spaces**.
- ❑ There are two ways of renaming variable name:
  - ❑ **Camel Case**: *capitalizes* the first letter of each word after the first. E.g. householdIncome
  - ❑ **Snake case**: uses *underscores* instead of spaces. E.g. household_income



Stata/SE 15.1 - G:\practice_dataset_NSS73_block2.dta

#	Command	_rc
1	use "G:\practice_dataset_N...	
2	list Sample State_Region in...	
3	list Sample State_Region in...	
4	d	
5	d State	
6	su	
7	su,detail	
8	codebook State	
9	codebook, compact	
10	duplicates report ENTID	
11	duplicates report ENTID	
12	use "G:\practice_dataset_IH...	
13	duplicates report HHID	
14	use "G:\practice_dataset_N...	
15	destring State, gen (statene...	
16	br statenew	
17	br statenew State	
18	destring State, gen (statene...	110
19	destring State, replace	
20	sort Sector	
21	br Sector	
22	rename statenew stateIndia	

```
     1863
     1886
     1921
     1928
     1934
     1951
     1955
     1965
     1969
     1981

. use "G:\practice_dataset_NSS73_block2.dta"

. destring State, gen (statenew)
State: all characters numeric; statenew generated as byte

. br statenew

. br statenew State

. destring State, gen (statenew)
generate() contains existing variable(s) and/or illegal variable name(s)
r(110);

. destring State, replace
State: all characters numeric; replaced as byte

. sort Sector

. br Sector

. rename statenew stateIndia

.
```

Name	Label
ENTID	�Primary key - un...
CentreCodeR...	Centre code,Round
Vill_Blk_Slno	FSU Serial No.
Round	Round
Schedule	Schedule
Sample	Sample
Sector	Sector
State_Region	�State-Region�
State	State
stateIndia	State
State_District	�State-District�
Stratum	Stratum
SubStratumNo	Sub-Stratum-no

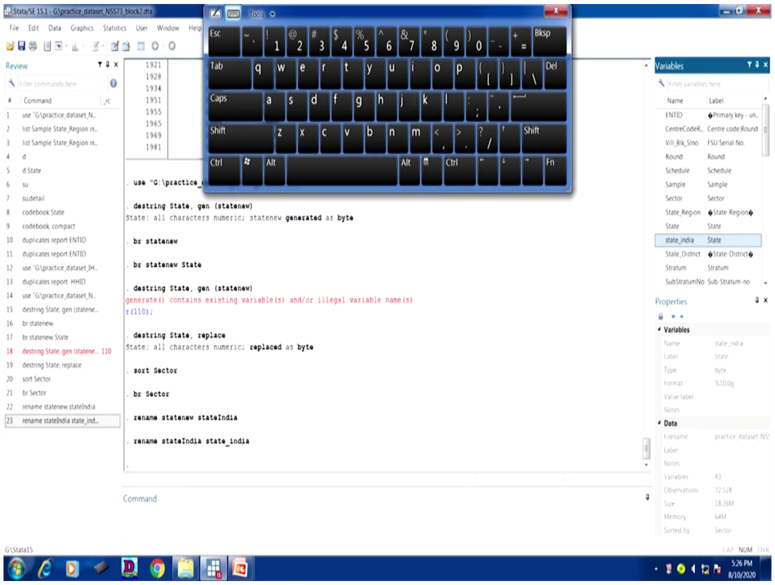
Variables

Name	stateIndia
Label	State
Type	byte
Format	%10.0g
Value label	
Notes	

Data

Filename	practice_dataset_NS
Label	
Notes	
Variables	43
Observations	72,528
Size	18.26M
Memory	64M
Sorted by	Sector

G:\Stata15



Stata/SE 15.1 - G:\practice_dataset_NSS73_block2.dta

#	Command	_rc
1	use "G:\practice_dataset_N...	
2	list Sample State_Region in...	
3	list Sample State_Region in...	
4	d	
5	d State	
6	su	
7	su,detail	
8	codebook State	
9	codebook, compact	
10	duplicates report ENTID	
11	duplicates report ENTID	
12	use "G:\practice_dataset_IH...	
13	duplicates report HHID	
14	use "G:\practice_dataset_N...	
15	destring State, gen (statene...	
16	br statenew	
17	br statenew State	
18	destring State, gen (statene...	110
19	destring State, replace	
20	sort Sector	
21	br Sector	
22	rename statenew stateIndia	
23	rename stateIndia state_ind...	

```
       1921
       1928
       1934
       1951
       1955
       1965
       1969
       1981

. use "G:\practice_

. destring State, gen (statenew)
State: all characters numeric; statenew generated as byte

. br statenew

. br statenew State

. destring State, gen (statenew)
generate() contains existing variable(s) and/or illegal variable name(s)
r(110);

. destring State, replace
State: all characters numeric; replaced as byte

. sort Sector

. br Sector

. rename statenew stateIndia

. rename stateIndia state_india

.
```

Name	Label
ENTID	�Primary key - un...
CentreCodeR...	Centre code,Round
Vill_Blk_Slno	FSU Serial No.
Round	Round
Schedule	Schedule
Sample	Sample
Sector	Sector
State_Region	�State-Region�
State	State
state_india	State
State_District	�State-District�
Stratum	Stratum
SubStratumNo	Sub-Stratum-no

Variables	
Name	state_india
Label	State
Type	byte
Format	%10.0g

Let us move on understanding some other operations like renaming the variable. A good variable name tells you very clearly that what the variable contains. If you have named very appropriately, for example, you are working with GDP, but you have named a state, but, probably you will be confused later on when you open and you have already operated with many other information.

State maybe, state name is there, but if you have not mentioned GDP, you will be confused. But, like agriculture, output is there, but you have given gender. Generally, this mistake we do not do but some close name we generally give, but still make sure that this is very unique in your understanding that most appropriate value must be given for better and quicker results.

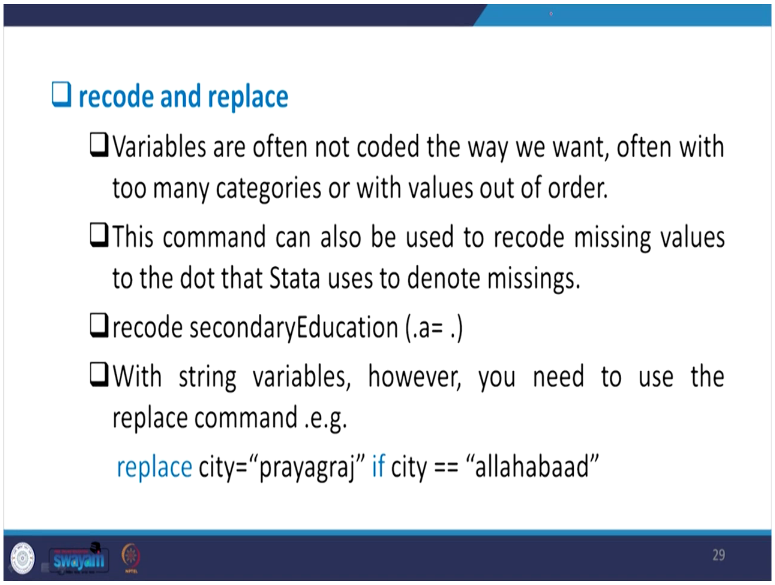
For our syntax like rename is the command you are supposed to give. Rename old, basically if an old name is there, you need to rename that old name. The dataset has given certain by default names as per their convenience, but you wanted to continue with a new name that name maybe very suitable as per your explanation, your model so you add a new name. So, old name will be renamed with a new one, I will tell you right now.

The variable name cannot contain spaces as I told you, there are two ways of renaming variables. One is renaming is camel case, another is called snake case. Likewise camel has a continuity, so each name you need not give spaces, you just capitalize the second starting letter. It is not a continual flow, there is a capital letter. And in case of snake, there is a clear space. This is different than that of another one. So you need to give a underscore. Underscore identifies for separating two different words in the name, we will clarify right now.

So let us operate. Suppose I just wanted to rename here, statenew is there. I already defined statenew. Let me rename it to another one like stateone or maybe stateIndia if you are working with other countries together, your data containing other countries, state of India, like state, two approaches as I told you, camel view, so India will be capital like this. So if you enter it will be renamed like this. Now it has already been renamed, you can just have a look from your variable view. I have pointed out, I have put my cursor here. It is getting visible to you.

Otherwise, you can also do underscore operation like the snake view. You just give underscore India. The name has already been changed. So statenew is not there, it is now stateIndia. So I am just going to change that stateIndia to state_india. Now this has changed to state_india. So what is important now the variable has been changed with a new name as per your own convenience. And this is very essential and usually very useful.

(Refer Slide Time: 29:26)



❑ recode and replace

- ❑Variables are often not coded the way we want, often with too many categories or with values out of order.
- ❑This command can also be used to recode missing values to the dot that Stata uses to denote missings.
- ❑recode secondaryEducation (.a= .)
- ❑With string variables, however, you need to use the replace command .e.g.

  replace city="prayagraj" if city == "allahabaad"

Then recode and replace. Recoding and replace is equally important for work and for operation in our module. Like, for example, if you have a data like gender, for example, gender is there in your variable, 1 is coded for male and 2 is for female, but you want it to, when your result, your interpretation will be higher, positively linked towards, since higher weight is, as per the numeric number higher weightage with female. But it might be becoming very inconvenient for you.

Like, if I go by how many males are impacting the society in any particular policy framework, your data is related to any particular policy analysis, you wanted to understand or interpret through male, not female. But where female is important, you wanted to answer through female, your weight is higher in female. there is no harm in dealing with that, but for some convenience, if you recode it to, like female you make 1, recode to 1 and male at 2.

Suppose you wanted to do that, now your regression coefficient is coming out to be positive. your likelihood analysis will be when the society is more of female, the likelihood of getting advantage or disadvantage is positively linked because of higher male members. Since you have already given a positive or higher value to male, so your relationship is expected to be positive there. But depending upon, it is not a standard rule at all.

There are many sort of example I will do it in between let me go by that. How we are recoding it, I will explain it later. Recoding variables are often not coded the way we want, as I just said. Like in some dataset education attainments they give it in reverse order. Educational attainment like class 10th they have given 1, 12 they have given 2, for higher education 3. If that is there, higher weightage is there already. If the reverse order is there already in the data like 3 weight is for 10th, then 2 is for secondary and 1 is for higher secondary, then that might be inconvenient for the researcher to analyze so you need to recode it.

So, this command, the recode command can also be used to recode missing values to the dot that Stata uses to denote missings. Like as I already pointed out earlier that Stata usually reads dot as the missing values. But if there are some other entries, by mistake also there might be some entries, some alphabet are entered in the numeric data and you have checked that some alphabet has been entered and converted to a string value and you have detected that.

So it is very difficult that which number has replaced with that alphabet. So in that case the easiest way to convert that number to when you do not know the exact replacement, you simply convert it to a missing value as dot. So recode that particular entry as dot. So recoding command is going to be very useful.

I will tell you some of the sample entries for our clarification. So recode secondary education, if secondary education is our variable name, your command will be recode then the variable name, whatever the variable name you have. If you have a .a entry, as I just pointed out a couple of minutes back that you just wanted to read this as a missing value, you simply read as .a equal to simply dot. Recode, in the command recode .a or if 98 is entered in the data or usually 8888 are considered to be the missing value in different dataset, 88, 89, even these are the standard missing values in our different datasets. I think I told you already while guiding the datasets. We have already guided you, reading the missing values in the data. So you can recode it to dot.

So its string variables, however, you need to use the replace command, not the recode. If it is numeric, so we can convert it. But if it is a string variable, the standard recode command is not going to work. In that case replace that. If city is a string variable and with some entry Allahabad, you want to replace it by Prayagraj you have to enter as replace, not recode. Recode is going to be erroneous. It is not going to give you the right result.

So if you want to change it to another one, just change it by that name. But again it is a text. It is in string variable still now. Even if you have recoded, it still converted to a string variable. What we will do for your analysis, we have to convert the other way for analysis to destring, in that case you need to go by the encode command, or sometimes wherever it is only in string but not in label, not in categorical labeling or categorical values in that case if it is only in string format with characters, your conversion will be through destring I have already pointed out.

So here for recode and replace, recode is useful for numeric variables, replace is useful for the variables which are string in nature. We will continue rest of this from keep and drop command and there are other entries in our analysis in our next class. So thank you very much for patient listening.