Handling Large-Scale Unit Level Data Using STATA
Professor. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture No. 24
Combining Datasets in Stata-II

Welcome friends to this NPTEL MOOC module on Handling Large-Scale Data Using Stata and we are here in understanding combining datasets using Stata with the real life datasets. Real life that a particular individual and their narratives have been captured through the proper teams of Government of India with certified teams of Government of India and accordingly the information have been gathered, but information has not been fine tune unless we combine the data. Extraction we did earlier. In the last class we did append to vertically append, add the data, vertically add the observation that I guided you.

Do not get confused by the name merging. Why am I saying merging? because the content for today's lecture is on merging. So, merging is different than that of append. I have guided you categorically. If you have little confusion on it, please go back to my previous lecture and try to understand the information between the lines that will certainly clarify you and we have also given you a schematic graphical presentation of the difference between append and we are guiding on merging. So, without discussing further, I hope you will get the details and you will raise some doubts and for your better, I am happy to clarify you on merging.

The meaning merge, it seems the movement how they are merging as if one is inserted in another one, another is inserted in another one, which way is inserted that does not matter much. We will also clarify that, whether this gets inserted or that gets inserted. Can I say that these get merged? I will say these get added. This is there, this is a vertical, when I say it is, I will say add. When this is there and two different objects are coming there is a possibility of collision that is called not addition, that is called merge.

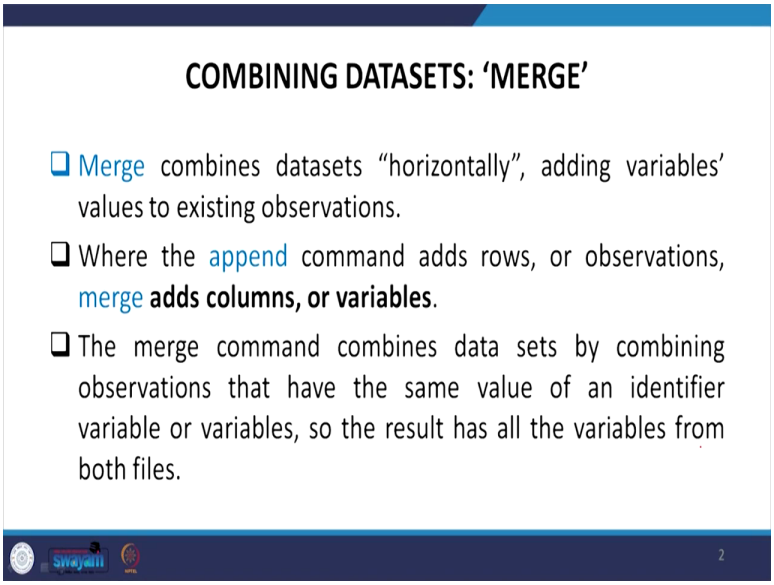
In dataset similarly, we have horizontally available variables and that could be merged. horizontal information as per the given structure of the available datasets they give variables on the columns. These dataset variables are in the column; another dataset variable are in column. On the rows, number of observations are there. Number of observation is not important for us,

because observation will be the same or for the same observation we are going to add different information regarding the observation.

As I told you already that in any like NSS dataset on consumer expenditure we have household information as well as individual information. So, within a household we may expect so many individuals. On the average as in Indian household 3 to 4 members on the average are there in a family. So, in one dataset or one block of the dataset gives individual information only regarding individual characteristics. Individual features are different than that of household features.

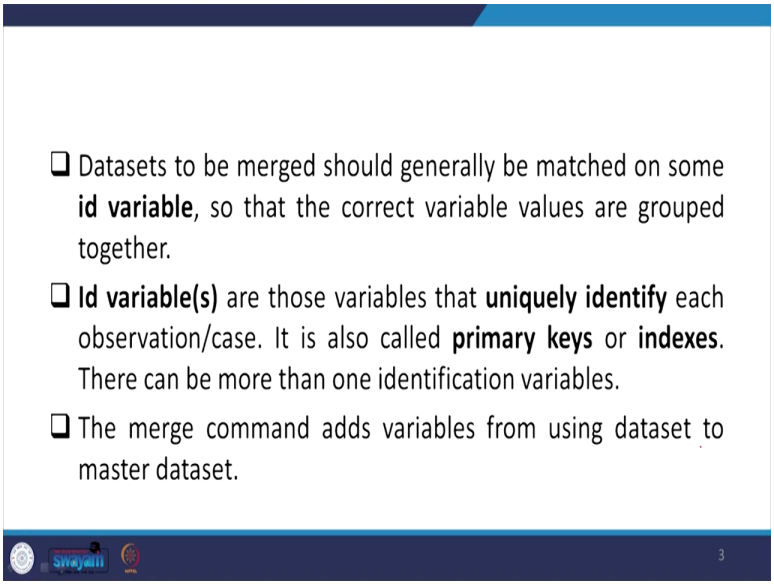
BPL family, for example, BPL status grants, grant is not given to any individual, it is given to a particular household, whether household is under below poverty line or not. Like gas connection, if it is a household having number of individuals within that, so household connection with LPG gas connection refers to a household, not individual. So, household characteristics are different than that of individual characteristics. Individual we mean their age, their sex, so accordingly we are going to guide you differently as compared to append. But the append command adds rows or observation, merge adds columns or variables in simple language.

(Refer Slide Time: 04:47)



The merge command combines datasets by combining observation that have the same value or an identifier variable or variables, so the result has all the variables from both the files.

(Refer Slide Time: 05:04)

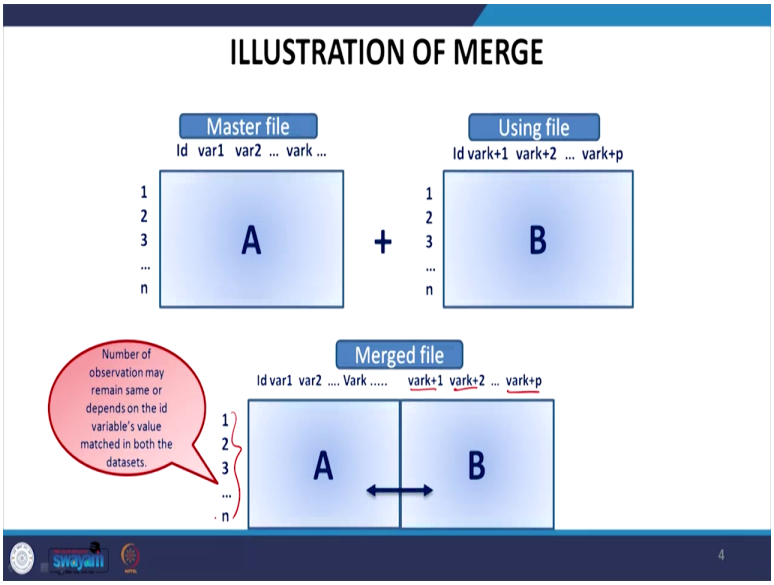


Datasets to be merged should generally be matched or, matched on some ID variable, so that the correct variable values are grouped together. So, ID variable, we have highlighted here, we are going to guide you in detail. So, it is bit tricky to understand what do you mean by ID variable. Generally, in string, we have already guided you from the very beginning all the dataset gives first information related to ID, but ID may not be uniquely identified that is different. We are going to tell you in a short while.

So, ID variable or variables are those variables that uniquely identify each of the observation or case. Single ID variable may not uniquely identify. What do you mean by? again uniquely identifier to be guided you shortly. This ID variable is also called primary keys or the indexes. There can be more than one identification variable as I just said, not just one variable may not represent the entire information. So, ID has to be defined very clearly and there are some tricks to define then only merge is possible.

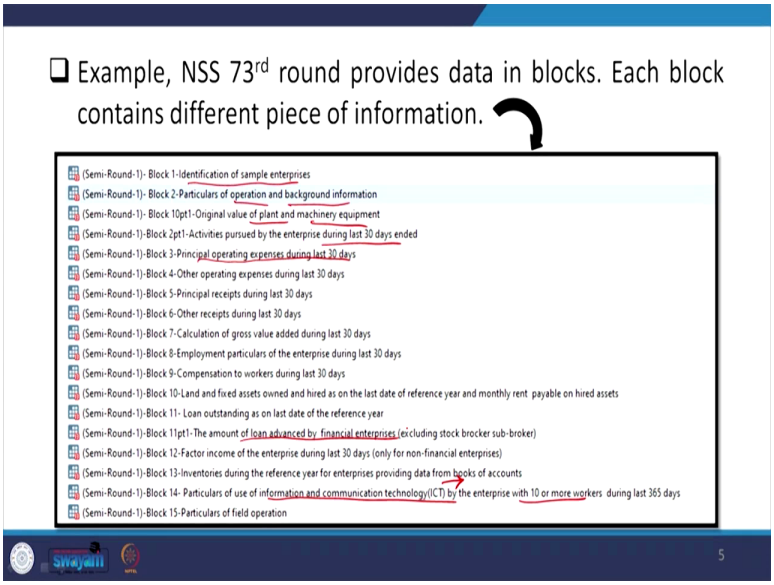
Like an individual belongs to which household, if the individual attach with a household characteristic in the identification variable if you are merging, then that particular individual will carry the information or will be merged into that particular household if the ID is perfectly defined. If the ID is uniquely identified that we are going to guide you. The merge command adds variable from using datasets to master datasets, likewise we explained in the previous lecture on appending.

(Refer Slide Time: 06:58)



So, master here for us is A and the using for us is B as I told you. we are horizontally merging the dataset. Number of observation is 1 to n in both the case. Number of variables, mark the difference very clearly, number of variables I am just highlighting here, number of variables are added. So, number of observation may remain same. Look at this is the same against to the variable, against to the observation now variables are added or more information is available to the corresponding observation. So, ID variables value merge in both the datasets.

(Refer Slide Time: 07:44)



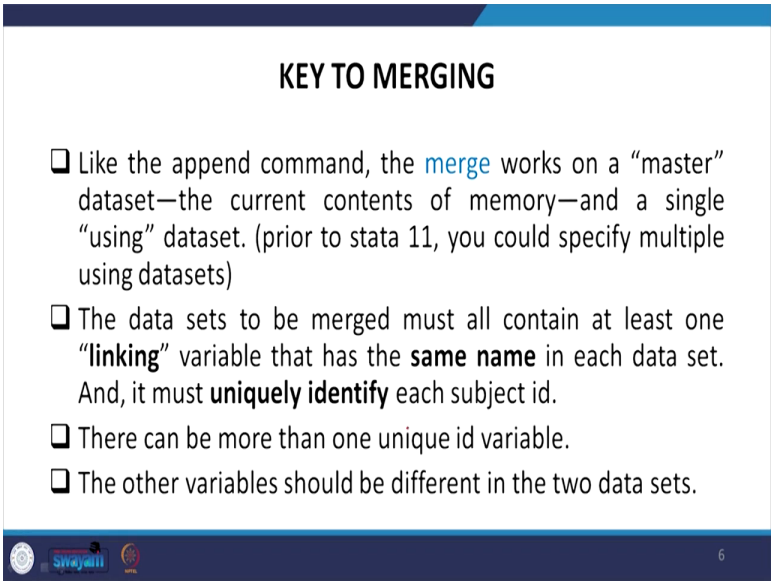
An example is here before us from seventy third round of NSS. So, please recapitulate what is that seventy third round. It is seventy third round of National Sample Survey on unorganized enterprises minus construction so even agriculture is not included that is the round, as I told you earlier, non-agriculture non-construction enterprises are included in this round. So, each block contains different piece of information.

Please mark I have already shown you during the data availability lecture, data explanation lecture, you can go back to that lecture and find out. Once again I am clarifying that. Look at the Block 1, that is why it is called semi-round-1, because entire information is not available in one round so they are naming accordingly. But we have to mark, identification of the sample enterprise is given in Block 1. Block 2 particulars of operation, so operation of the enterprise you will get it from Block 2.

Original value of plant or machinery equipment like enterprise, during last 30 days in Block number 2.1, then Block 3 contains principal operating expenses during last 30 days, similarly, you have many other information. Like some of you might be very interested in understanding information technology. These days the buzzword in the social science research, how the enterprise are equipped with ICT, enterprise with 10 or more workers information is there also.

Like another one is financial inclusiveness, you can also get it, amount of loan advanced by financial enterprises that is also important. There are so many information they have categorized into different blocks. So, we are supposed to merge as per our own requirement. I am going to guide you.

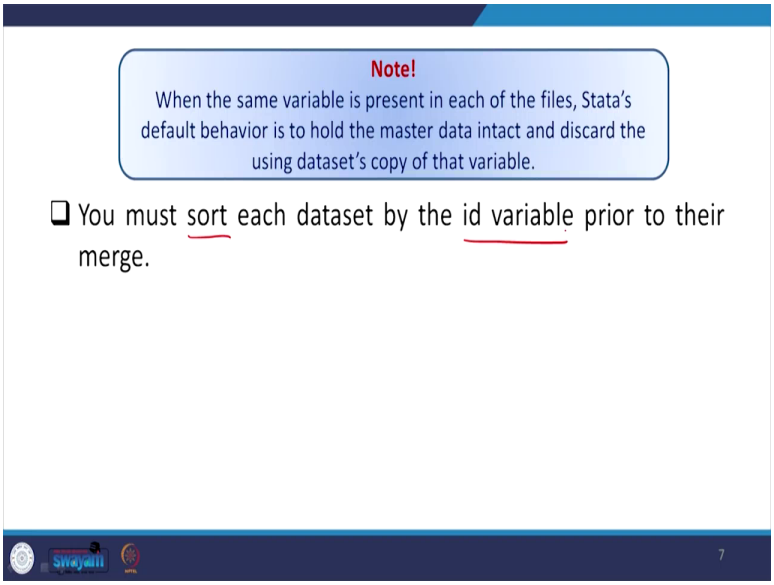
(Refer Slide Time: 09:57)



So, what is the key to merging? Like the append command, the merge works on a "master" dataset, the current contents of the memory and a single "using" dataset. Prior to Stata 11, you could specify multiple using datasets also. The datasets to be merged must all contain at least one "linking" variable that has the same name in each dataset and it must uniquely identify. Like particular variable if it is uniquely identifying means that can be represented in all the merging datasets then only it will recognize that yes this information can be merged from another block. So, variable it is blocked.

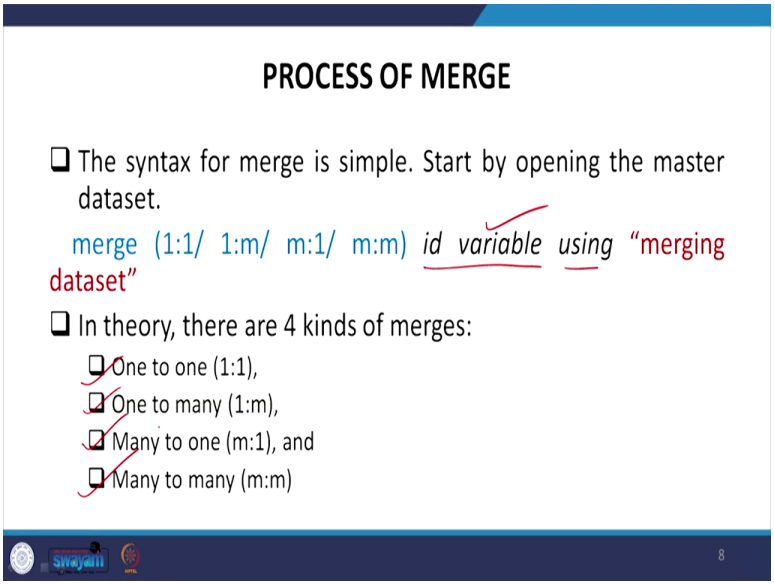
So, there can be more than one unique ID variable as well, just one unique ID that is the variable ID may not be representing the entire blocks of information. So, combination of variables may define that now representing to entire blocks of information given in the data. We will guide you. Please wait for a couple of minutes. You will have those slides and operation. The other variables should be different in two datasets. Please take a note that when the same variable is present in each of the files, Stata's default behavior is to hold the master data intact and discard the using datasets copy of that particular variable.

(Refer Slide Time: 11:44)



Again the way we guided you in the append master dataset is important and some information are continued as per the master dataset. So, we will give you the hands on experience in our Stata operation. You must sort each dataset by the ID variable prior to the merge that is very very important. Sort has to be done of each dataset by the ID variable. So, ID variable we are emphasizing once again. So, what is that ID variable? information you can get it from the report also what is the ID variable. They give the information from their report as well. You can also test whether this is uniquely identifying or not based on some information.

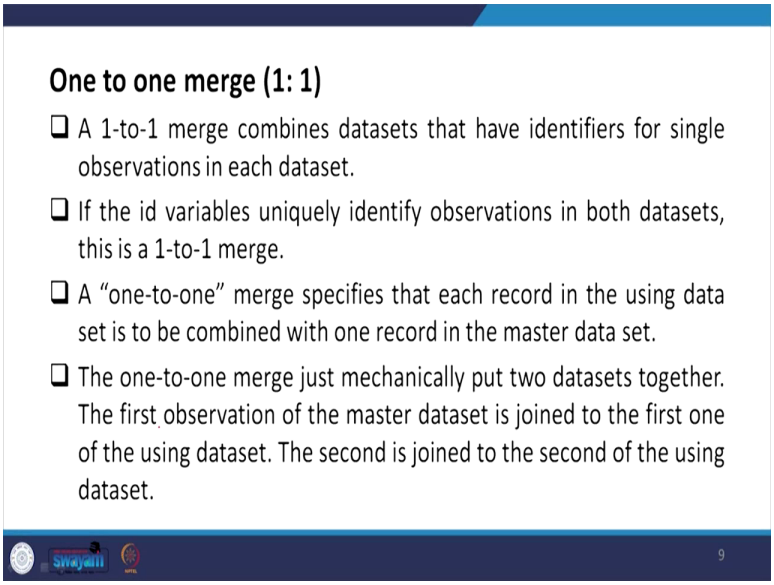
(Refer Slide Time: 12:33)



So, what are the process of merging? The process of merging here is broadly of four types, we are going to discuss. So the syntax for merge is simple. Start by opening the master dataset first then you have to follow the step we are going to guide you. The merge out of these 4 options, one option you have to pick, one to one or one to m, m stands for many, one stands for one, one to many or many to one or many to many ID variable using, you have to give the ID variable here that is very very important to be discussed.

Using file as usual which file you are going to merge. It is not like append. Using file has to be containing other information, not the variable, not the observations. I will guide you how to do that. In theory there are 4 kinds of merges as I said, one to one, one to many, many to one that is m to one and many to many.

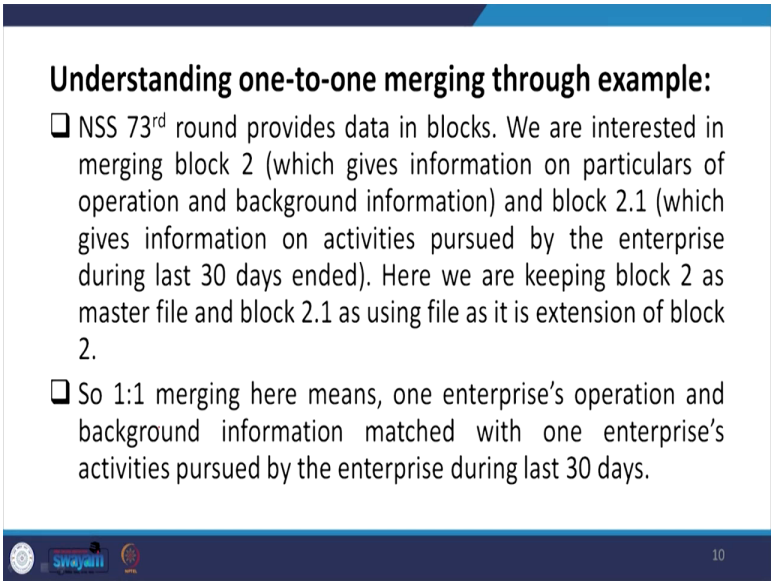
(Refer Slide Time: 13:43)



Let me explain you what do you mean by all those details, one to one merge. So, insert one to one to be entered in the command. A one to one merge combines datasets that have identifier for single observations in each dataset. If the ID variables uniquely identify observation in both datasets, this is a one to one merge. That is basically in both the one if it is uniquely identify observation then that is one to one merge. A one to one merge specifies that each record in the using dataset is to be combined with one record in the master dataset.

So, in one to one merge just mechanically put two datasets together. The first observation of the master dataset is joined to the first one of the using dataset. The second is joined to the second in the using dataset, accordingly it goes.

(Refer Slide Time: 14:43)



let us understand from the real life datasets, the dataset we are referring right now seventy third round of National Sample Survey that provides data in blocks which we have already shown to you. We are now interested in merging block number 2, which gives information on nature of operation and background information regarding the particular unorganized enterprise we are referring to.

And the block 2.1 which we have already shown you, gives information on activities pursued by the enterprise during the last 30 days ended. Here we are keeping block 2 as master file. Block 2 as our master file and the using file is 2.1. Basically, 2.1 information that is all about your activities pursued in last 30 days will be merged with that of the background information and nature of operation of that particular enterprise. So, using file is extension of block 2, that is 2.1 we have already mentioned.

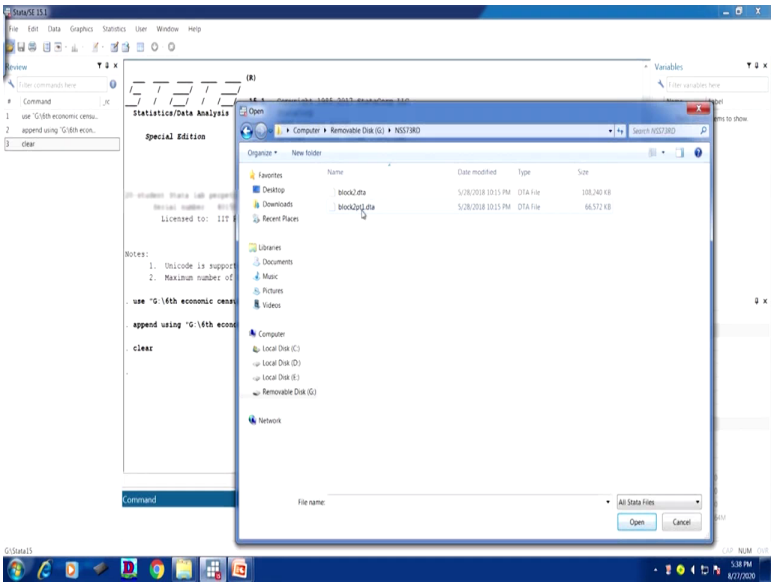
So, one to one merging here means one enterprise operation and background information matched with one enterprise activities pursued by the enterprise during last 30 days. So, one to one merging we are going to discuss.

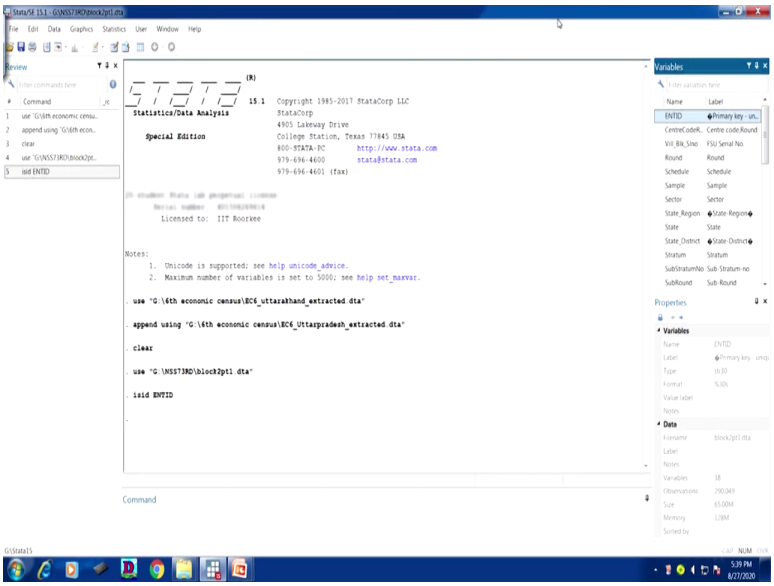
(Refer Slide Time: 16:14)

- Next come the **identifier variable** (Enterprise Id –**ENID** in this case). The identifier variables must exist in both data sets, and have the same names, but in most cases all of the other variables should have different names.
- Now we have to check in both the datasets, ENID is uniquely identifying the observations or not then only one-to-one merging is possible.
- To check whether ENID is unique in both the dataset:

  isid *ENID*

  if ENID is unique in both the datasets, stata will show nothing on the result window else through an error.





Since the identifier variable suggests us to do that, I am going to tell you. Next comes the identifier variable, which we are repeatedly mentioning. In our case, it is ENTID, enterprise ID. So, the identifier variables must exist in both the datasets then only we can do that one to one merging and that have the same names, not just identifier variable but should carry the same name, otherwise how can Stata be able to read that identifier variable with the same in different blocks. See in most cases all of the other variables should have different names then only you are merging information, because different names or different variables are there. It is not similar to append.

We have to check in both the datasets ENID is uniquely identifying the variable, observations or not then only one to one merging is possible. To check whether ENID is unique in both the datasets we have to go for ENID, isid ENID. If ENID is unique in both the datasets, Stata will show nothing on the result window, else otherwise if it is not correct, it will show an error. So, we can test that. So, we are supposed to open the seventy third round of NSS. So let me open that before you. First what I will do, since the dataset has already opened, let me clear that first. So, I am just doing that exercise. The dataset is clear.

So, what we are going to open is seventy third data. For us, we are trying to simplify for you. So, this is sixth economic census opened earlier. What I will open? Seventy third round of NSS. So, there are two files. We will be using block 2, as visible in the screen to you is our master and 2.1

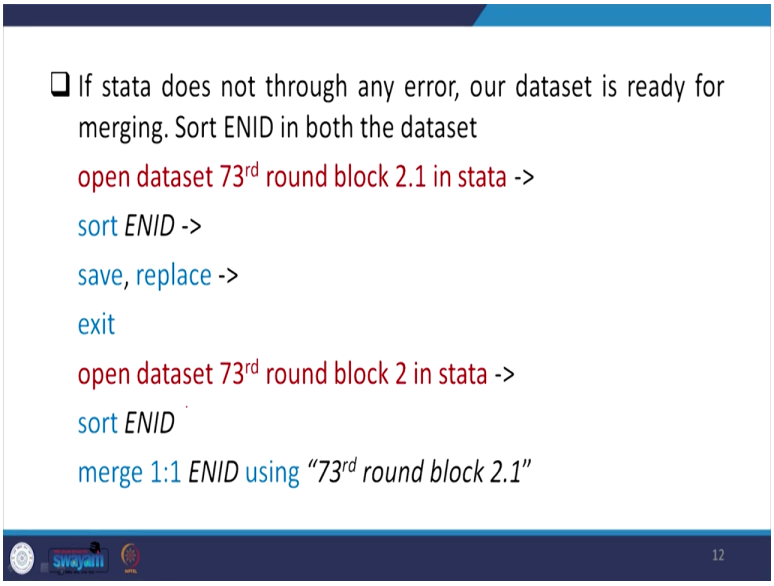
is our using file. But at this moment we are just trying to check whether it is uniquely identifying or not, which is that identifier variable or not.

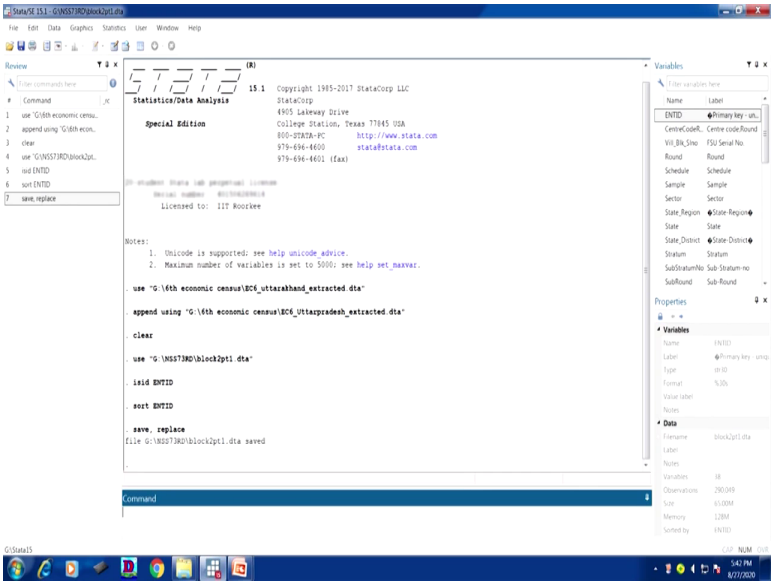
And I have already suggested that we identified from the reports, from the different information of the data that ENID is our identifier variable. We just need to cross check it whether that is there or not. So, let me open this first. I can also open, let me open this first 2.1, because I am going to use this and then I will save this data and I will use it for another master file. The data has already been opened. What is important? look at very carefully that the data is ENID is there, isn't it. This is the one ENTID.

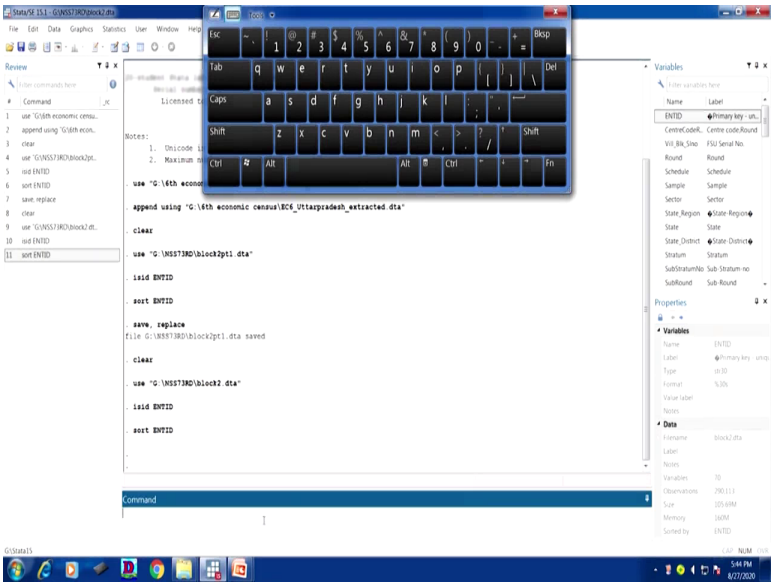
So, I will check with isid command right now. So, isid, here for us is this one, first one. So, let us check whether this is correctly followed or not. As I highlighted from the slide that if it is not throwing any error that means it is uniquely identified. Just a dot information is coming that means it is uniquely identified. So, what we have done so far. We simply checked whether this dataset or this block is uniquely identifying with that particular variable is ENTID in our case.

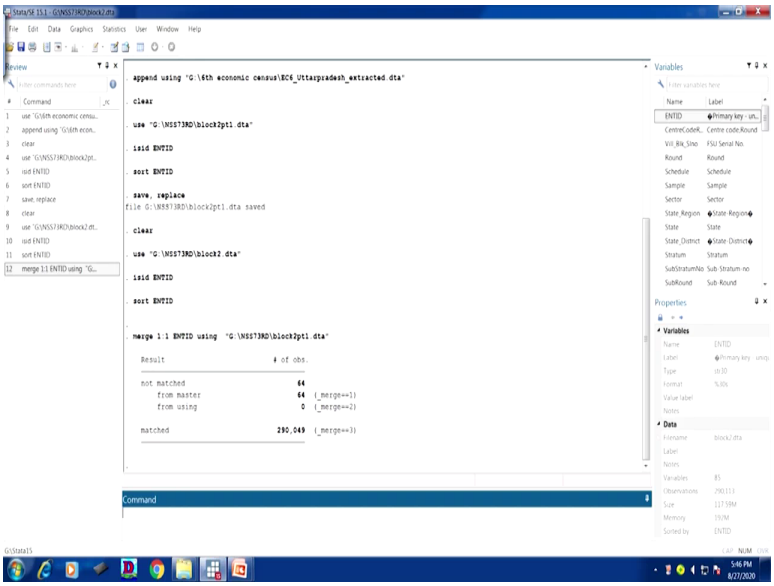
Let us move to our guidance. Similarly, you can check with the master file that is block number 2. I have only shown you the block number 2.1. Why I have opened the using file, there is a reason behind it. I am going to guide you. I have shown you isid. This is very very important to note that you have find it. Here I have entered ENID, please note that in our case in the dataset I have shown you, it is ENTID.

(Refer Slide Time: 20:37)









If the Stata does not through any error, our dataset is ready for merging. Other steps are very important also. So, now sort that as well, it might be the case that first observation is there from the beginning, but in the another file that is at the end. So what I will do, I will sort that ENID or in our case ENTID. We have to sort it. So, what we will do sort that first. So, sort then we will go by this the same then sort then I will enter this. The dataset is sorted by that particular variable.

We will be very sure that dataset is sorted. Now we simply save and replace. You simply save and replace. What do we mean by that? We are going to use this in another master file. So, we are going to save it, save and replace. So, what is left? You need to exit it or simply ok. So, we

need to check the another file also that is the master file, whether it carries with the same name or not. So, open the, now what you will do, you open the dataset of seventy third round block number 2. So, block number 2 let us have a check.

Block number 2, now open, since this is already opened, it might be creating some problem. So, you need not close it, so just the clear is going to add another master file on the same platform, no need to exit or close the data. So, we are going to open the block 2, the master file. Similarly, master file is being opened. The same ID is carrying here ENTID. In the PPT it is ENID, you can correct it, it is ENTID if you wanted to go by the exact command for your clarity.

So, what I will do? the same process. We have to check whether it is uniquely identified or not. So, then you simply add isid then you click this ID then check. There is no error defined. So, that means this is perfectly uniquely identified. Another important aspect, likewise we sort the using file, we have to also sort this as well. Sort this, because the data will be in sort format through this particular filter or variable. This has been sorted. So, both we have processed it for final merging.

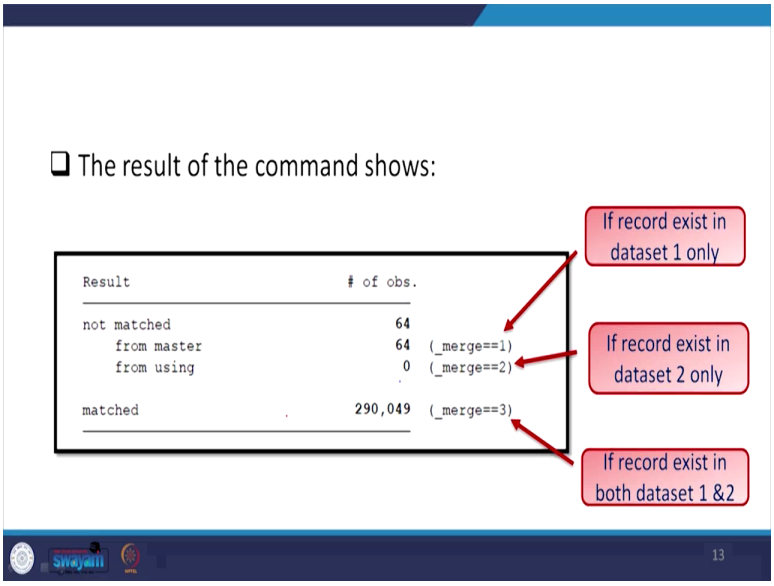
Look at, so I have opened at this moment in front of you of the data that is master block file of the seventy third block number 2. I have also sorted. Now what is left, merge. So, merge one to one, what is that variable we will be using, in this case, ENTID. ENTID, in our case using the block number 2 that is all. So, let me proceed. I will go for merge, isn't it, merge. What is that one is to one? What is to be merged that is your unique ID variable. Rest of the variables will be automatically in order as per the unique ID.

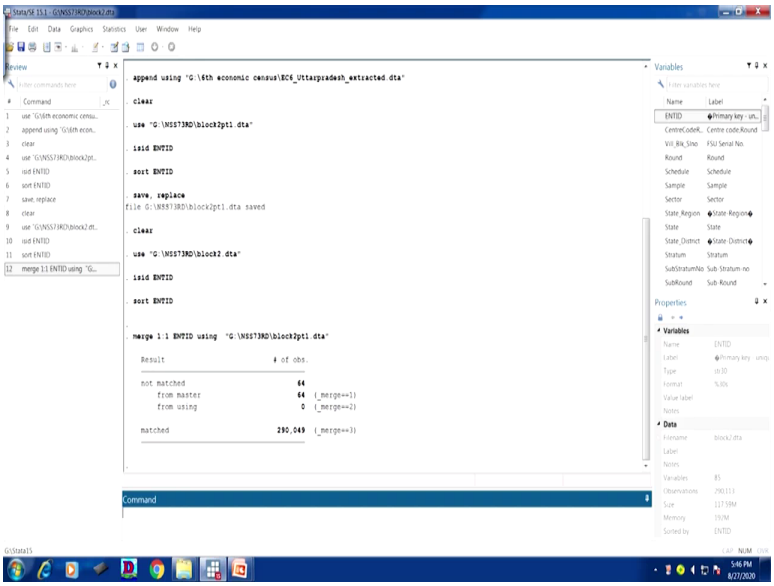
So, you need to define unique ID here. Here unique ID is our ENTID, then using, space, you need to give a space. path has to be defined. So, what is that path? you need to go to that file. It is there. So, here is the path. This is sixth economic census we are going to open the seventy third. So, path is 2.1, so let us copy address, the way we did every time. I am going to paste it. So, the another path is added. You have to add the exact file name path that is you want to copy this. So, we have copied the entire with dot dta has to be copied. Close the document.

You just see, just be careful about it merge one to one then using the path name very correctly. If it is there, finally you have to enter. Look at your number of variables must have been increased, because you are horizontally adding the variables or merging the variables. I am going to show

that through the PPT. Also, it is very important to note, this will give you the result. How many files, it is very important that merge, then two, then merge third. Third is very important the 2,90,049 have been already merged. I will go through the PPT and guide you and accordingly I will show you the result.

(Refer Slide Time: 26:43)

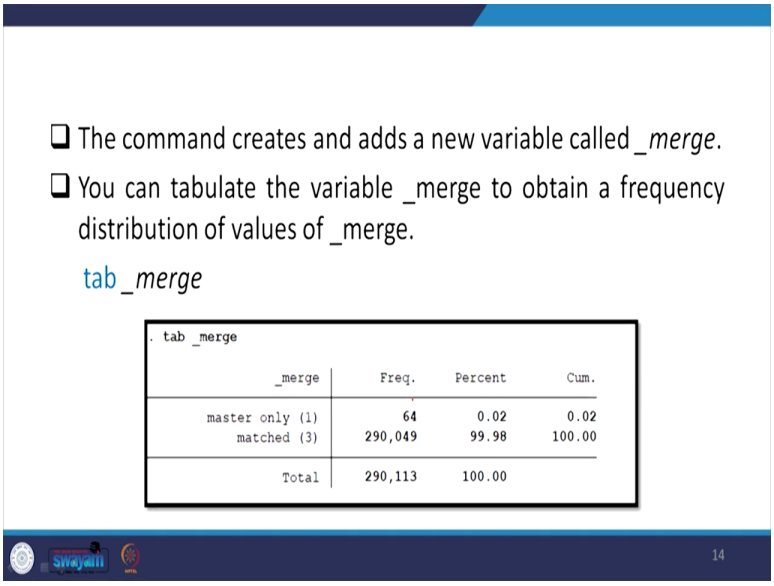


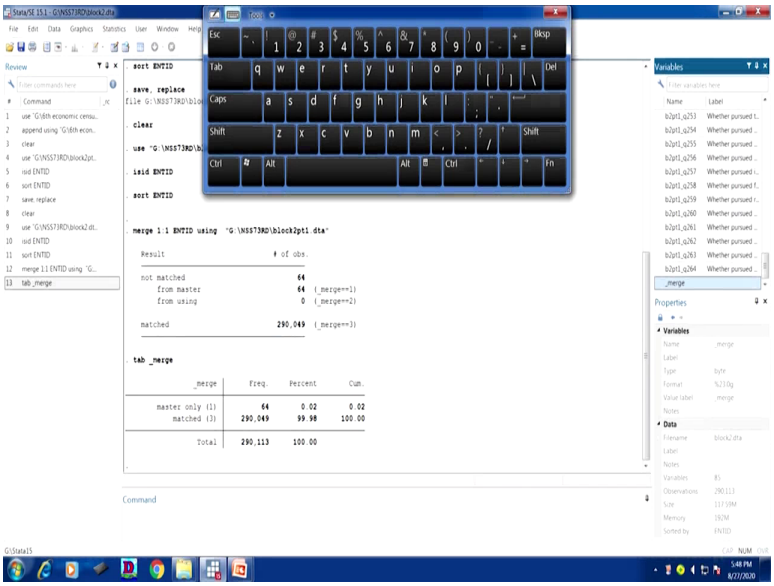


This is what we have also shown you. And you please mark the arrow here. If record exist in dataset 1 that is merge at 1, 64 from the masters are there which are not merged. So from the master it is saying, from the master here, like from the result, here 64 from the master, look at

very carefully, the 64 is not merged. They are different, isn't it. so from the using, 0 is left. That means entire is merged 2,90,049 have been already merged. If record exist in both the datasets it will be 1 and 2 and that is, here it is 3. That has been successfully merged. Since there is no error emerged, so there is no issue.

(Refer Slide Time: 27:36)

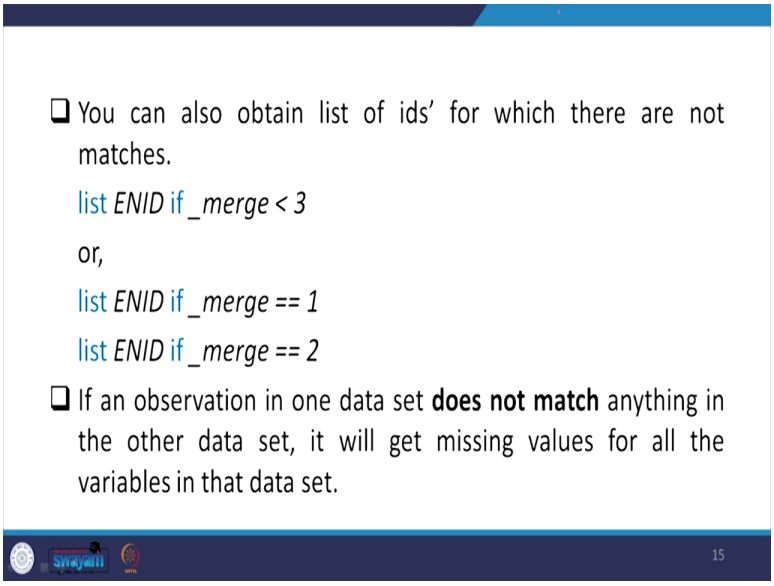




Similarly, you can also check with tab command. A merge variable is defined in the data. You can check that. Here is the merge. You simply tab or ta then you check this. It gives the result here. From the master 64 is still there. Matched one, both the data has been merged from another

using is 2,90,049 has been successfully merged. And so that means we are successfully merging the dataset.

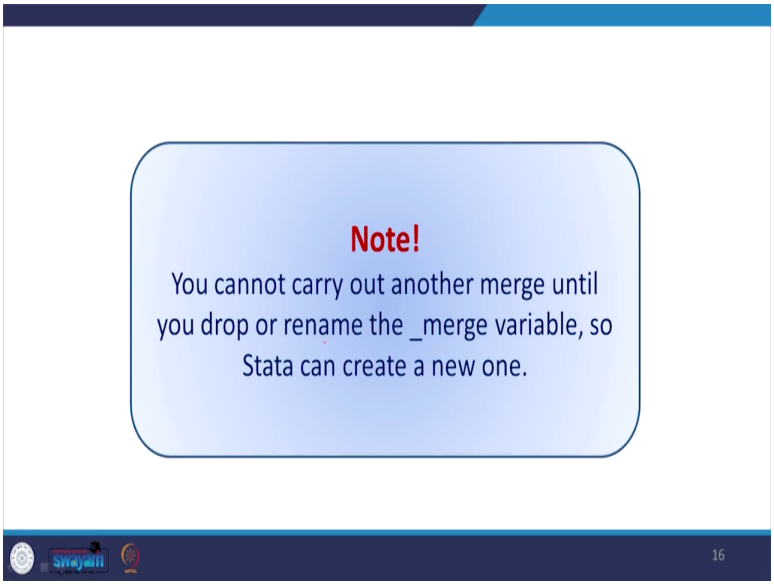
(Refer Slide Time: 28:19)



Let me guide you the other details, so that you need not be confused further. You can also obtain list of IDs for which there are not matches. Like list ENID, here it is list ENTID if merge is less than 2 that is 1 and 2 are saying. So, if it is less than 3 if we are mentioning that means it will automatically take, if it is list out the code of 1 and 2. So, the command, single command you can give less than 3.

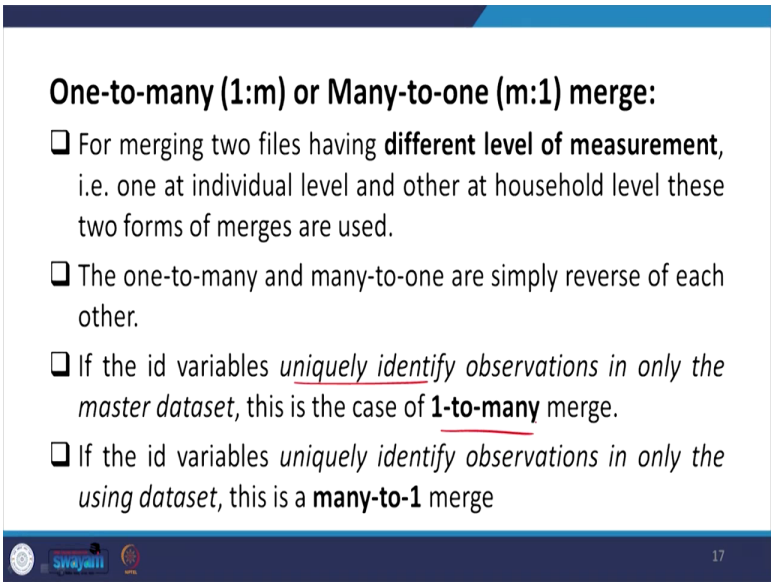
Otherwise, you can individually check that ENTID if merge double equal to one you can check whether it is there or not. Similarly, list ENTID if merge double equal to 2. So, both the category I have already shown you. So, that will show you the result. If an observation in one dataset does not match anything in the other dataset it will give missing values for all the variables that in the dataset. It will certainly show you the missing values.

(Refer Slide Time: 29:23)



Please note that you cannot carry out another merge until you drop or rename the merge variable, so, Stata can create a new one.

(Refer Slide Time: 29:38)



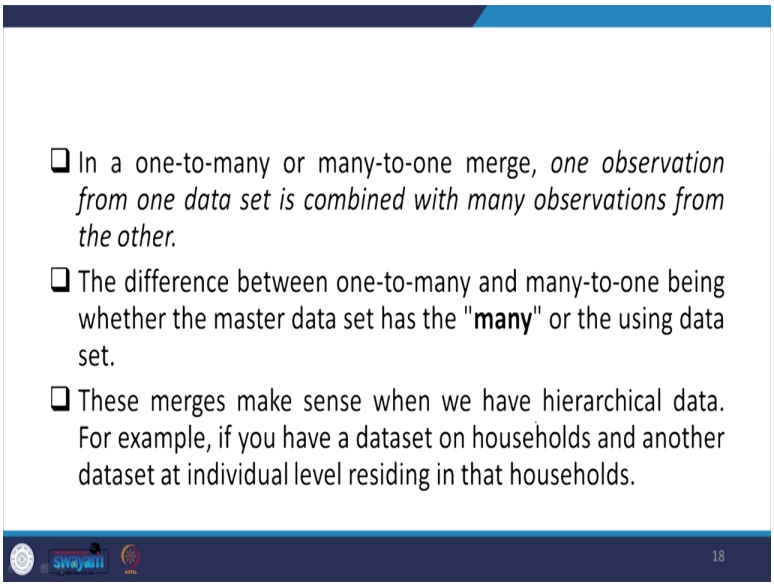
So, I have guided you what is called one to one merging. What is important for us also to know is that one to many. In many case one to many or many to one they are just opposite to each other is important in real life situation while handling the data. For merging two files having different level of measurement that is one at individual level and other at household level as I already told

you, these two forms of merges are used. When one is subset or one is different than that of another one and so somewhere it is many, somewhere it is one. So, in that case many and one or one to many or many to one command is important.

One to many and many to one are simply reverse of each other. If the ID variables uniquely identify observation in only the master dataset this is the case of one to many merge. If the ID variables uniquely identify observations in only the master dataset, the dataset which we wanted to open and merge another one, the first dataset is called the master dataset.

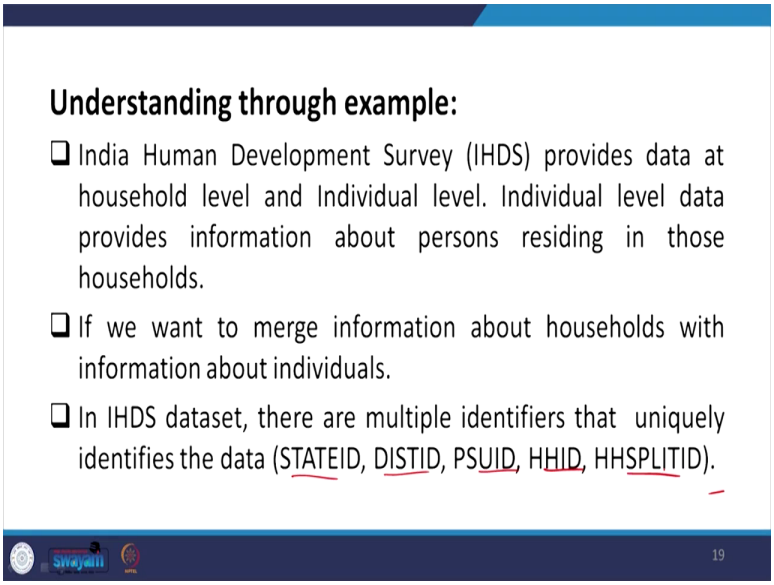
If the ID variables uniquely identify observations in only the using dataset that means you have to provide the reverse approach many to one. So, be very careful about it. Unique ID if it is in master then one to many, if it is in your using file, then it is many to one that is the standard rule to be followed.

(Refer Slide Time: 31:27)



In a one to many or many to one merge, one observation from one dataset is combined with many observations from other. The difference between one to many and many to one, being whether the master dataset has the many or the using dataset. These merges make sense when we have hierarchical data, like individual or household. For example, if you have a dataset on households and another dataset in individual level residing that household in that case it makes really sense for it.

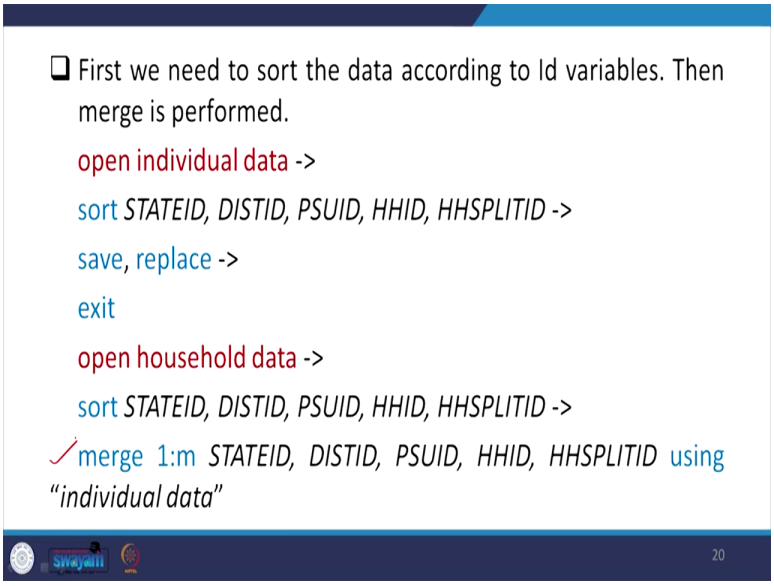
(Refer Slide Time: 32:07)

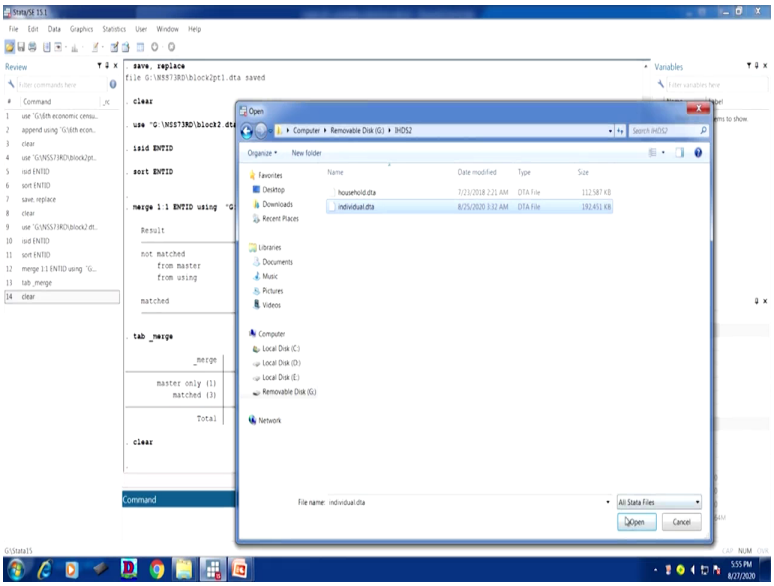


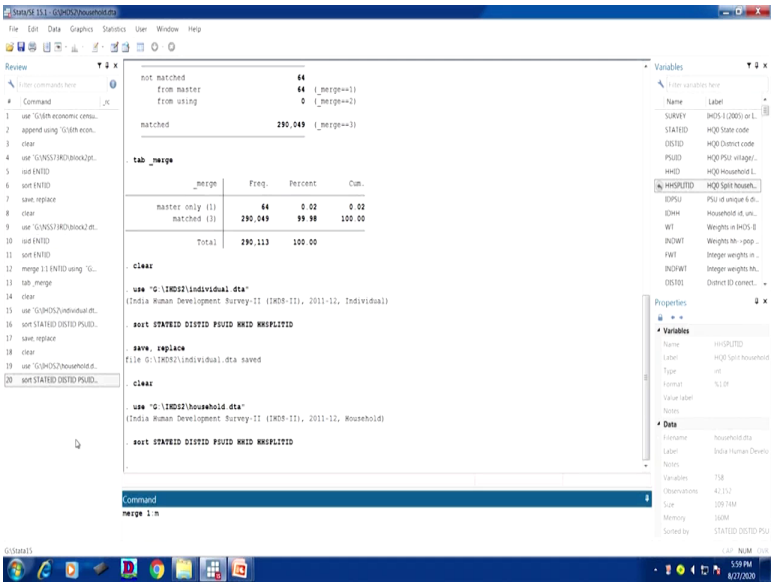
Let us understand through the exact dataset, the real life dataset that is IHDS, India Human Development Survey data in 2011-12 available for public use. It provides data at household level and individual level. Individual level data provides information about persons residing in those households. So, if you we want to merge information about households with information about individuals then this is important.

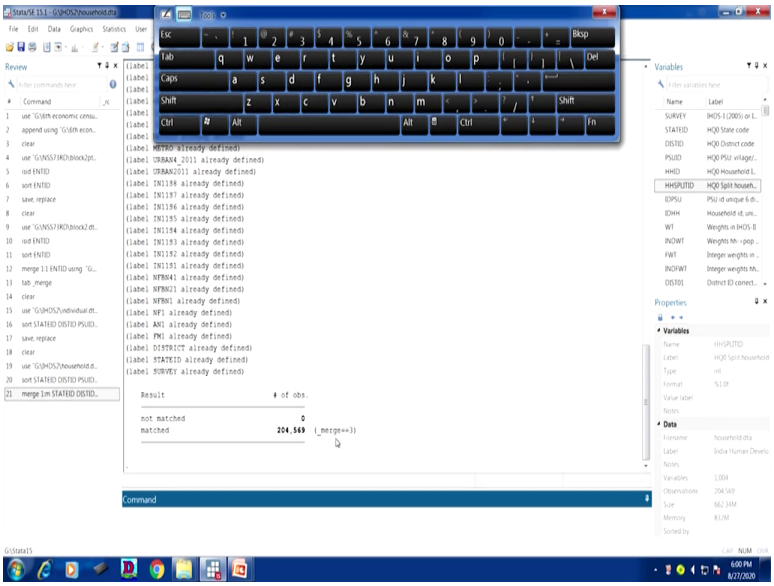
In IHDS dataset there are multiple identifiers that is important to note that uniquely identifies. I repeatedly said earlier that just state ID for example is not uniquely identified in case of the IHDS data. There might be multiple variables or identifier taken together to merge the individual and household information, 1, 2, 3, 4, 5. 5 variables if you are taking together then you can able to merge successfully. So, we are going to guide you through that particular data.

(Refer Slide Time: 33:23)









First, we need to sort the data the way we did. And how to do that, these are the uniquely identify, you have to read the information given in the IHDS website many information, very clarity is already provided. Otherwise, again one by one you have to test. There are number of experiments required to test whether it is individually uniquely identified or it is joint identified.

So, what I do, so open the individual dataset first then, because we need to merge that individual file on the household file. So, let me open that first for you. I am guiding through the dataset of IHDS 2 we are referring. So, what I do, I will open. We have to clear this. If you clear it, it clears the earlier saved data. Your window is free to take or import the fresh dataset. We are opening

the fresh dataset that is IHDS 2. IHDS 2 we are going to guide you. This is third. So, IHDS 2, we are going to open it. So first I will open the individual dataset for you. It is taking little time. It is still searching. Do not get worried of not opening. It takes its own time.

What is important to note here is very very essential for you to understand. State ID, district ID, PSUID, HHID, then HHSPLITID, so these 5 we need to sort because we have already identified from the information that these are the identifiers, uniquely identifiers together, not individually. So, let us sort them. Sort them all together because those identifiers we are going to use it while in merging. So, what I will do, I will copy all those things. Why should I type. So, let me copy this out or I can click there, state ID, district ID, PSUID, HHID and split ID. So, it is there. So, sort state ID, district ID, PSUID, HHID, and household split ID.

Why household split ID, because the IHDS gives information about some split household in the next round. All the households are not covered in the IHDS 2. If you go and check, I have already mentioned this during the lectures of our module on IHDS, please check that lecture once again. I also told to you that time which are called unique identifiers. So, once again we are explaining. We are trying to explain you what is called sort here of the identifier variables. So, we explain how to sort all those identifiers and we have entered the command called sort.

What is important? Then once that is sorted, what you will do, you have to again save and replace, then similarly clear you have to do it, the way we did. So, save comma replace. Then now it has saved, you need to clear. Because we are going to open the master file next and accordingly the same approach we will follow. So, it will take time to clear. And it is first save and clear the dataset. It has not yet cleared the window, still running.

So, till then let me guide you that, now it has cleared. I will guide you that we have filtered as per the requirement we have sorted the variable as per the unique identifier variables. That is very essential. We are going to open what master data file that is household in our case. You can also go by the reverse approach. I will tell you in between. So, where is the master one? Master file is here. So, master is your household. So, we have already opened this. So, again the same approach we will follow.

Then sort the unique IDs, the same unique ID. It has to be the same unique ID, otherwise it will be problematic. Since there is no error coming that means it has identified, but isid we are not

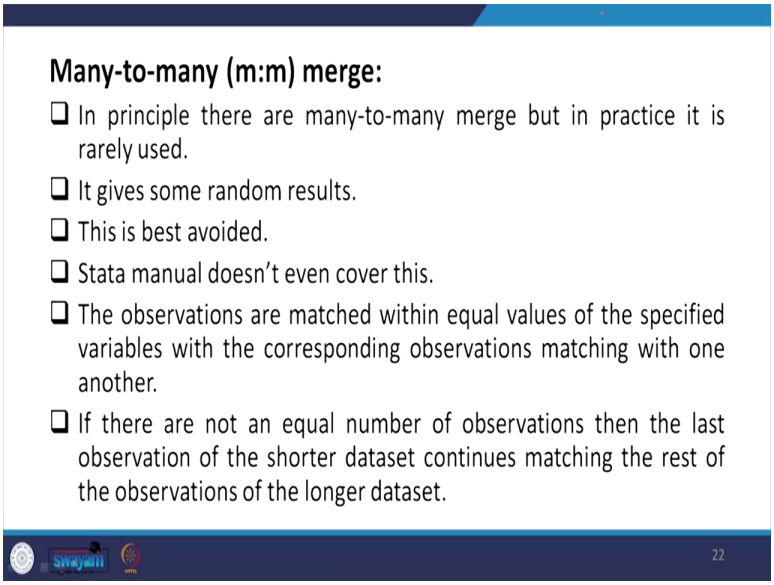
going to check it. We can check it as well the way we did it, but since we know it we are not going to check it. So, this part is done. So what is left for us, we have sort both the files, now we will merge. So, merge one to m, likewise it is mentioned here. This is merge one to m then all the identifier variables using the individual data, because the individual data is our using data. So, let us go by that.

So, merge one to m then all those unique identifiers must have been mentioned because they together jointly explaining the merging or using the identifier, using file we have to go by the path. So, path is there. Path is our individual, because we have opened the household file. So, household file is the one household with many information we are including, many information because individual has many information than that of the household. So, that is why we are using the command called one is to many.

We have added the path name that is available in our file is G, for you it might be D, might be C, you have to mark this carefully and copy that the way we used to do it. Once you enter, this gives the merging information. Look at the number of variables. This has been increased. and since there is no error there is no problem. matched, look at the matched file, the merge command is going to give you the entire information that is the total number of observation that is 2 lakhs 4 is completely merged and 0 is not matched that means there is no problem, nothing is left unmatched. So, this is successfully merged.

What is important for us to understand that you can do like this. The reverse approach you can follow. You can follow with merge m to one, but which is m here, m is your individual file. So, you open your individual file as your master file, then you use your household file as your using file. Then in that case merge m to one, then rest of the unique ID has to be same that we have already given the command sort and there will be no problem. Vice versa is also possible with m to one command, so both will give the same result.

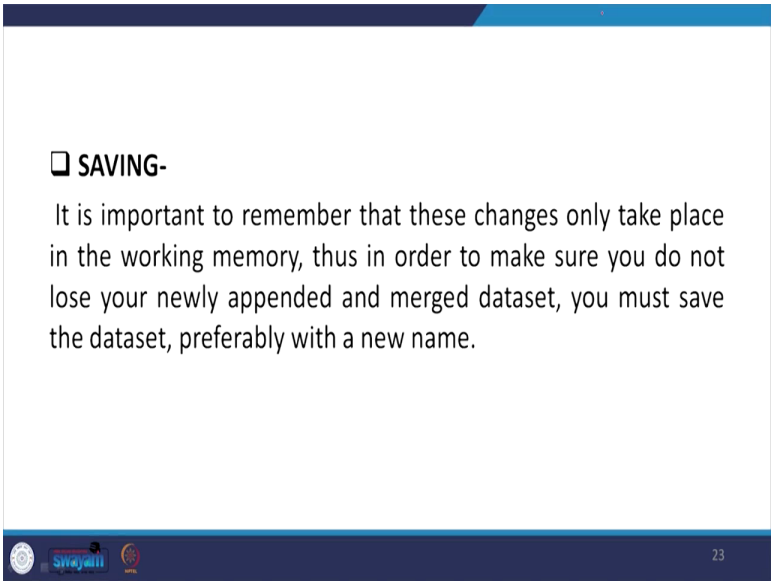
(Refer Slide Time: 41:21)



So, last one here to be discussed is called many to many merging. In principal, there are many to many merge, but in practice it is rarely used, because many information to many information rarely is not there. We are going to merge with one to many or many to one in most of the cases. Since there is no availability, we are not really giving you the hands on experience, but we will guide you for sure. So, it gives some random result. This is best avoided as possible. Stata manual does not cover this, if you go by the original Stata manual, which we guided already.

The observations are matched within equal values of the specified variables with the corresponding observation matching with one another. If there are not an equal number of observations, then the last observation of the shorter dataset continues matching the rest of the observation of the longer dataset that is the crux of the m to m merging.

(Refer Slide Time: 42:19)



So, saving is important here. It is important to remember that these changes only take place in the working memory, but not in the final file or path folder memory, thus in order to make sure that you want to reuse your data later and make sure that you do not lose your newly created, appended or merge dataset, you must save the dataset with a new name, preferably a convenient name must have been given the way we guided earlier.

So, that is all for merging and we have guided you NSS, we have guided you IHDS and also we have given you enough logical binding between the merging, especially one to m and m to one and the steps we have categorically emphasized and m to m even Stata since not guiding we are not giving in detail and better to avoid that also indicator we mentioned that.

So, I hope it is not creating any further hurdle as a learner. I know that Stata is very sensitive. So, please try to practice on your own. We are providing the practice dataset as well, along with the PPT in the command, and I am 100 percent sure you can able to extract, merge, append any database in front of you. Still if there are doubts, we are there to help you for sure and that is our pride. Thank you so much.