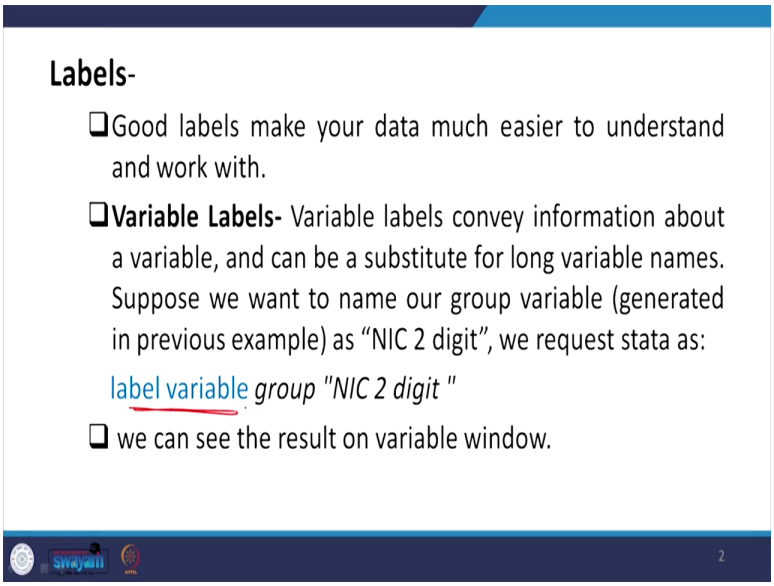
Handling Large-Scale Unit Level Data Using STATA
Professor. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture No. 20
Variable and Value Labels

Welcome friends once again to the NPTEL MOOC module on Handling Large-Scale Data Using Stata. We are almost entered into the handling of Stata software using the core dataset which usually social science and management guys use and we are trying our best to explain you every time with the help of that core data with Stata. This time we are continuing from the last lecture. I think in my understanding there are couple of doubts what I can foresee related to variable and value labels. So, accordingly we have titled it for your better clarity. Let me explain you once again.

(Refer Slide Time: 01:14)

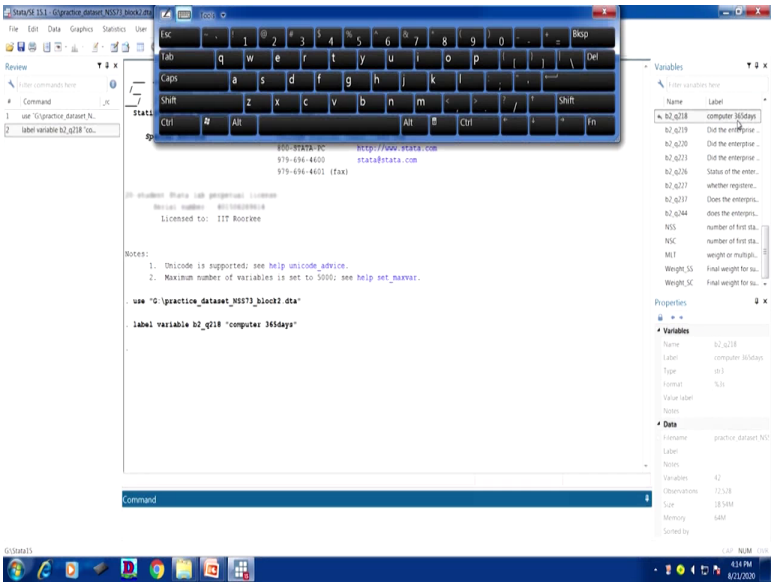


So, good label, as I already mentioned that it makes the data much easier to understand and also it helps in better to work with. Then relating to variable label, like labeling is always good, but there are two kinds of labeling we are going to guide you here. One is variable labeling and another one is value labeling. The variable labeling conveys the information about a particular variable and can be a substitute for long variable names and it can be also helpful for a very short variable to better comfortable variable on our wish.

Suppose, for example, in this case, if you want to name our group variable generated in the previous example that we have given to you that we are referring to NSS seventy third round, where our data has the coverage of one of the variable called NIC 5-digit level, NIC classification 5-digit level information is there. We wanted to convert it and filter it by our approach we have made it to NIC two-digit classification, but if you generate a new variable with two digit one, let us name that correctly. Let it be, name it as NIC 2 digit.

We request Stata as per our Stata command, it is label variable with the variable name as group NIC 2 digit. I will show you right now. We can see the result in a variable window. Let us test it or let us have a clear understanding through our Stata. So, let me handle this particular variable labeling with the help of the exact data of Stata.

(Refer Slide Time: 3:34)



So, this is our Stata software and the window have accordingly opened. Let me open a data and guide you how to label the variable. So, I am straight away trying my best to open the practice dataset and let me open this. And this will be available to you because we have made, tried our best to reduce the volume of the data and the can be easily opened.

And so I am just going to rename any variable, but I will try with that particular NIC dataset it is. But let me first rename any variable here. So, for example, not rename, I am going to give you the labeling. Suppose this is there, bigger name there. I wanted to reduce it to a smaller name.

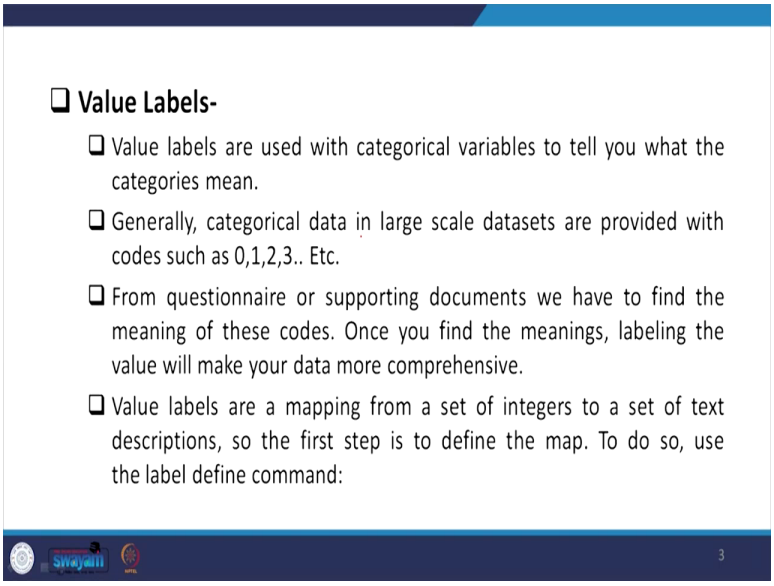
Did enterprise use computers during the last 365 days, simply computer last 365 days, if I just want to label it accordingly, if I just try to do that, just look at this. What I will do, I will label variable, type label variable.

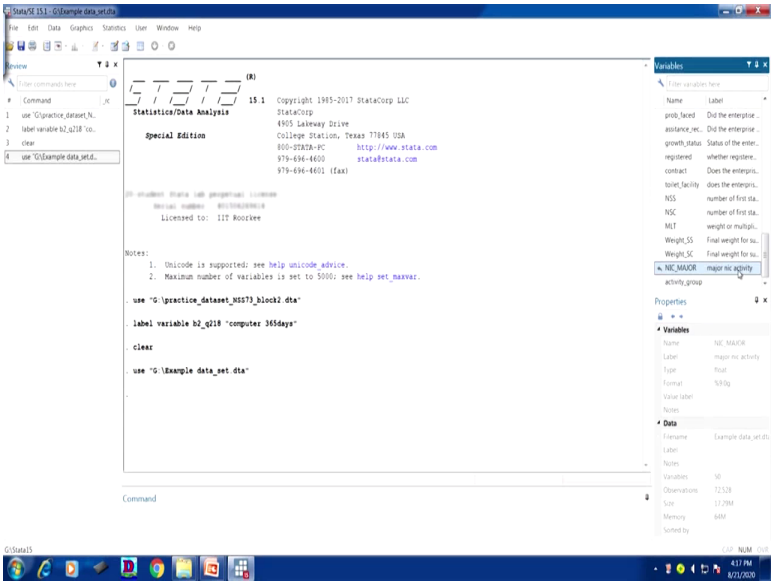
So, here is our label variable then I will click on this as I just said 365 days' computer where was that, which of our variable you may go through like computer. So, I will label it. Label is already there, did the enterprise, this is the label. Under the label it is clearly written. I just wanted to change this. What I will do, I will just put a space. Within double inverted comma, I will write down on my own comfortable wording, so that I can easily understand.

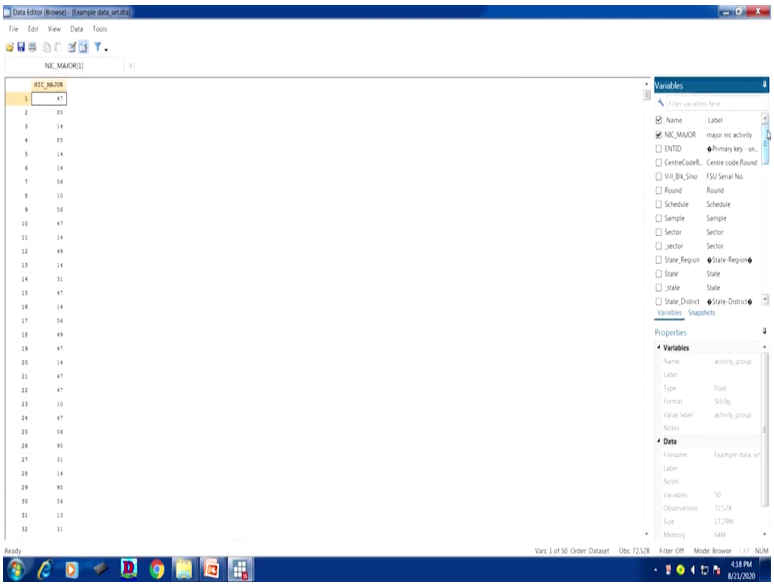
So, as I told you, if this is written with such a big sentence or big phrase, I will write down computer 365 days. Suppose I write it down, computer. you can use space, there is no problem, because this is labeling not the name of that particular variable, computer 365 days. you close the inverted comma, then you enter. look at this. It is clearly visible computer 365 days instead of the long labeling data, you can short it to your comfortable naming. Any time you just open this data, you have to save it, make sure that you have saved this it will show with this comfortable name as per your own understanding and that is generally very good for researcher to operate very quickly.

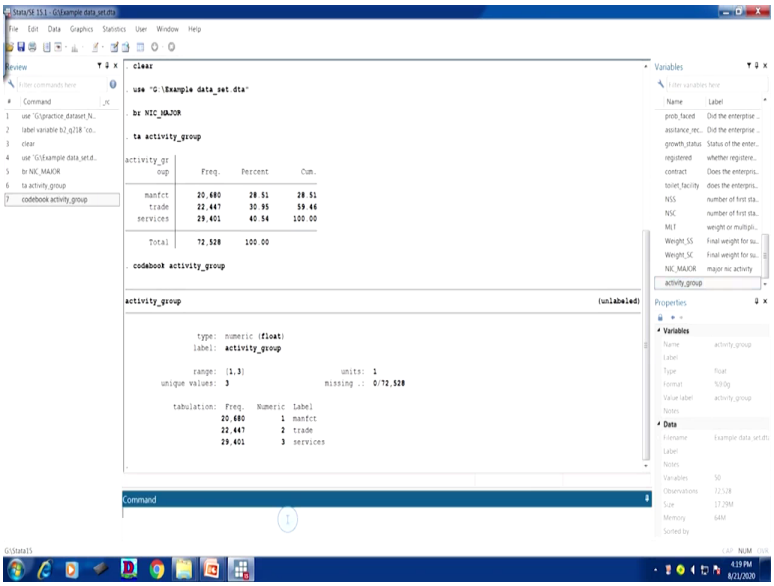
Similarly, in our data, you can write down accordingly. The sample I have already mentioned to you NIC 2 digit, if it is not written, then you can write it down accordingly. If you write down two instead of 2 any type of name you can keep it based on your own understanding and own comfortability. So, let us make a move. So, this is all about label variable that is more important. So, label variable is very important for our understanding. So, mark carefully we have explained so far label variable.

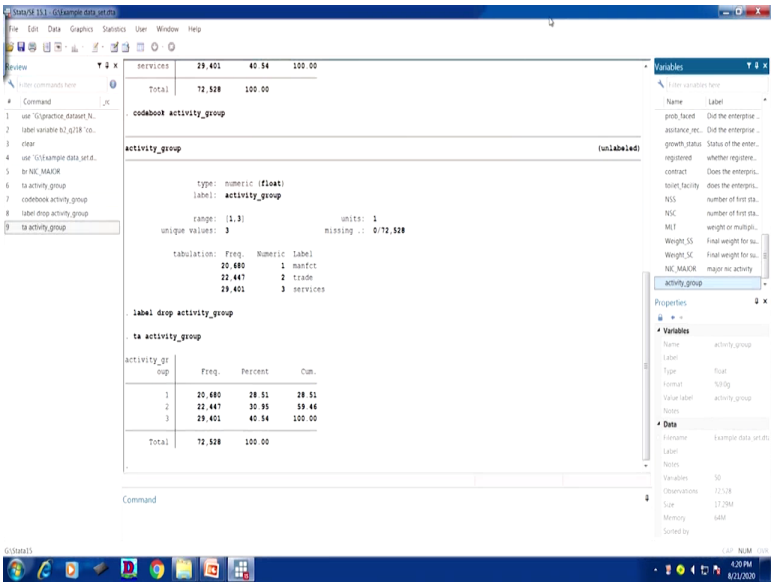
(Refer Slide Time: 07:42)






Data Editor (Browse) - [Example data_set.dta]
File Edit View Data Tools
NIC_MAJOR[1]
47
NIC_MAJOR
Variables
Name
Label
NIC_MAJOR
major nic activity
Properties
Variables
Data
Ready
Vars: 1 of 50 Order: Dataset Obs: 72,528 Filter: Off Mode: Browse
4:18 PM
8/21/2020


Stata/SE 15.1 - G:\Example data_set.dta
File Edit Data Graphics Statistics User Window Help
Review
Command
1 use "G:\practice_dataset_N...
2 label variable b2_q718 "co...
3 clear
4 use "G:\Example data_set.d...
5 br NIC_MAJOR
6 ta activity_group
7 codebook activity_group
. clear
. use "G:\Example data_set.dta"
. br NIC_MAJOR
. ta activity_group
activity_group | Freq. Percent Cum.
manfct | 20,680 28.51 28.51
trade | 22,447 30.95 59.46
services | 29,401 40.54 100.00
Total | 72,528 100.00
. codebook activity_group
activity_group (unlabeled)
type: numeric (float)
label: activity_group
range: [1,3] units: 1
unique values: 3 missing .: 0/72,528
tabulation: Freq. Numeric Label
20,680 1 manfct
22,447 2 trade
29,401 3 services
Command
Variables
Properties
Variables
Data
G:\Stata15
CAP NUM OVR
4:19 PM
8/21/2020



Another aspect of labeling is called value labeling. So, value labeling, there are different ways of doing it. But why you are going to label the values. Let us be the case that in our data, let me open a comfortable data and I will show you that why we require it and what is the necessity of value labeling. I am just going to open our example dataset. This is already opened. Even the NIC classification is also visible. You can have a look. So, it is there. So, by major activities we have name the variable name as NIC major. So, it has that particular name.

Major NIC activities we have labeled it accordingly as per our own understanding. But if you wanted to do it you can do it, but you need to check which digit is it available you can just have a check on your own of this particular variable NIC major. look at, it is in two-digit label, isn't it? So, it is in two-digit label, but the name is, it is in two-digit label, NIC major once again, I am coming to that activity group right now.

But let me just guide you, here suppose you wanted to name as per your own understanding NIC, it is there at the above. NIC major, if you wanted to, instead of major NIC activity if you wanted to keep your own name as I just guided, you can do it accordingly which have already been guided. We are discussing value labeling. So, how to label value is important. So, we are sticking to this. Activity group, there are some categories, some, I will tell you what is the necessity. Let me just have a check of this particular group. What is that?

There are 3 categories inside this particular variable activity group. The another clarity you can have on your own which kind of codes are already entered. If you wanted to go for it, just fine.

So, codes are 1, 2, 3 are there already given. These labeling for 1, what do you mean by 1, 1 is for manufacturing, 2 is for trade and 3 is for services is already given.

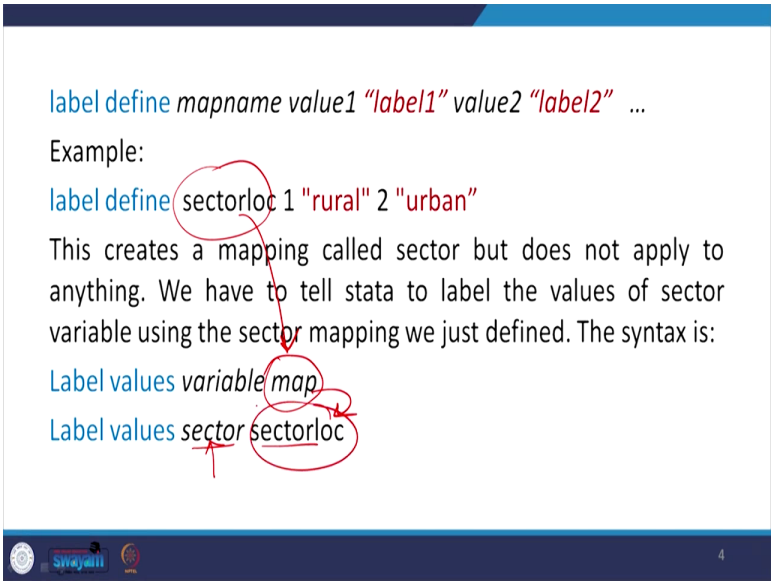
Now we have labeled it for our own understanding. But if you want to label on your own wish separately instead of manfct, manufacturing, you wanted to write it correctly, manufacturing, complete name of it. So, since I have already labeled what I will do, I also tell you, so simple aspect here to do it label drop. First I will do label drop and then I will show you. This is related to label, first label comes, then drop. Since labeling has already been made for our own understanding, let me first drop it and then I will clarify.

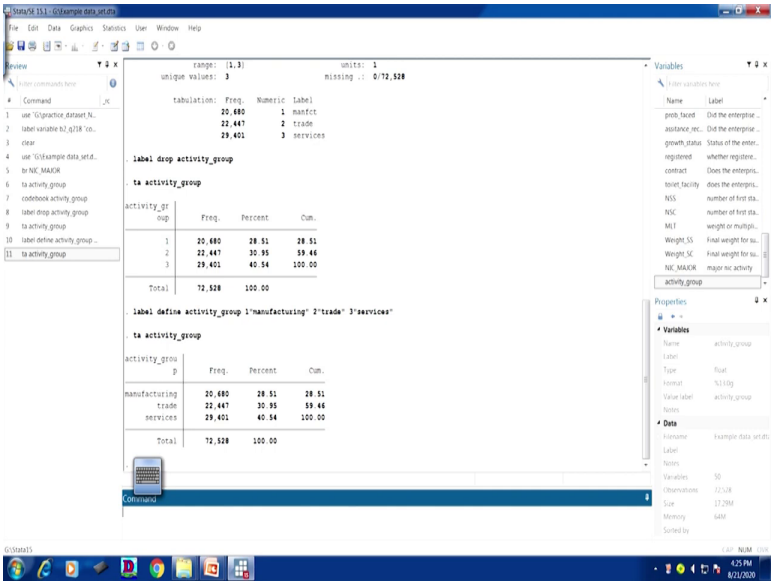
Here if I just check after dropping the label, look at now no labeling is given, labeling of the values are not given, 1, 2, 3 is now coming. In previous case before dropping the label, it was manufacturing then trade and services. Now we are going to label it. There are, let me just guide it from the PPT. There are important ways of handling. You just mark carefully how we are going to handle it. So, label define is the approach I am going to guide you. So, generally, categorical data in large scale datasets are provided with codes. So, large scale dataset with categories wherever are there those categories are represented with codes.

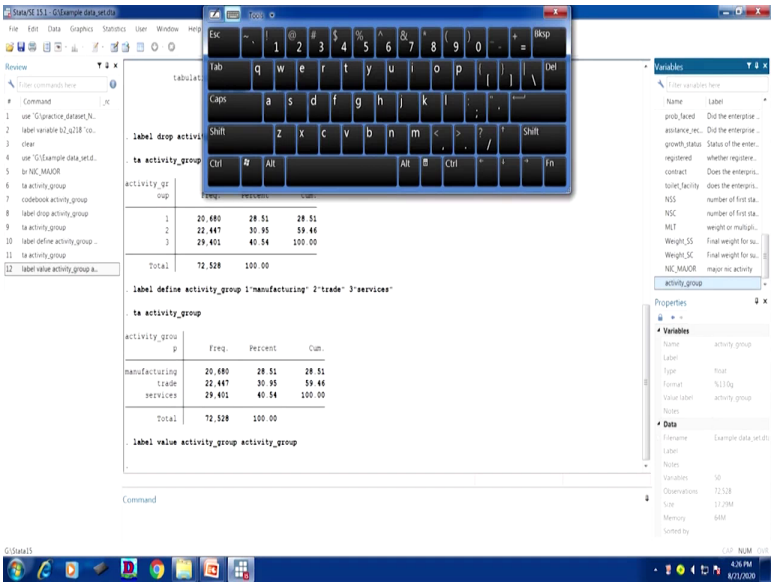
Like in our example for a major activities group we have 1, 2, 3 as the codes for different activities. And from questionnaire or supporting documents we have to find the meaning of these codes. What do we mean by these codes? these are clearly given in our questionnaire or there are some supporting documents. We have already guided earlier also at the time of extracting today or in another lecture we are going to guide you clearly.

Once you find the meaning and labeling the value will make your data more comprehensive. Value labels are mapping from a set of integers to set of text descriptions, so the first step is to define the map, what do you mean by that labeling map. To do so, use a label define command.

(Refer Slide Time: 13:54)







So, label define command must have been there. Like here the command is being shown in front of you. What I will do, I will write down like this. I will go to activity group, so what I will do? I will start with label define, then what label define is required? So, what label define of that particular variable. Within that variable there are some values given and it is in only 1, 2, 3. In integers we do not know what is the meaning of it, but once we read our questionnaire and some supporting document we know that 1 stand for manufacturing, 2 stands for trade and 3 stands for services. So, let us do that.

So, this is the one variable we wanted the value of it, we wanted to label, isn't it? 1 as very clearly so you have to do manufacturing, then 2 trade, then 3 services, any enterprise we are involving this kind of activities are labeled. this is one way. What I do, we have got the labeling variable, labeling of the values. And just have a check of it. It has already been labeled.

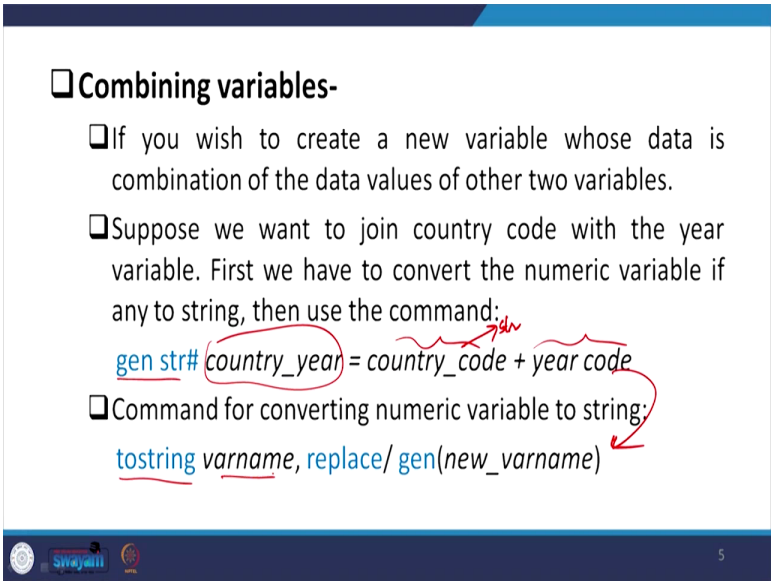
There are some important aspects within it, I am going to guide you. So, this after label define of that particular values, we need to check some important aspects. So, this creates a mapping called sector, but does not apply to anything. We have to tell Stata to label the values of sector, here in our example it is sector loc and sector data like rural, urban we discussed already, sector and location of that particular sector and we have to tell Stata to label the values of sector variable using the sector mapping we just defined. The syntax is in front of you. We will also provide you the particular aspect.

The way we did right now for label define of a particular variable called, that we just define, activity group, isn't it? So, what I do here, for me label values is important. Since I have defined, what I do, second one is, first I defined the label values, but I have to enter the values within that labeling. So, label values are also equally important, label value. Then for the, look at this, for me the sector and sector the mapping variable which I have already created is, in this case, in our example it is activity group. So, what I will do, I will enter the activity group here.

I will tell you, if you have a separate coding and variable we have already made and value label we have made, then the second one should be your mapping variable, the way it has been guided. If I just enter, your value has already been labeled at that particular variable called activity group. you just mark carefully what we wanted to say. The mapping variable, wherever we have done label define that variable should come here.

Like there are number of variables we require with binary coding, yes-no, yes-no type or dummy variable type. And every time you need not type the same 1 for yes, 2 for no. You simply map, you just add that particular variable which you have created just now, you just add that variable against to the variable you wanted to label. You wanted to label this right now, label the value as per the mapping variable you have created called sectorloc, location. Once you have done that, you will get the exact labeling of the values.

(Refer Slide Time: 19:04)



What is also important after understanding the labeling, combining variables, there are some intricacies involved, some variable, how to combine it. Not necessarily all variables can be combined on its own wish. if you wish to create a new variable whose data is combination of data values of other two variables, suppose we just want to join country code, in some data country codes are there. Even in our data state codes are there with the year variable. Year maybe in numeric number is clearly given in this year that year numeric has given, but country code cannot be numeric, isn't it?

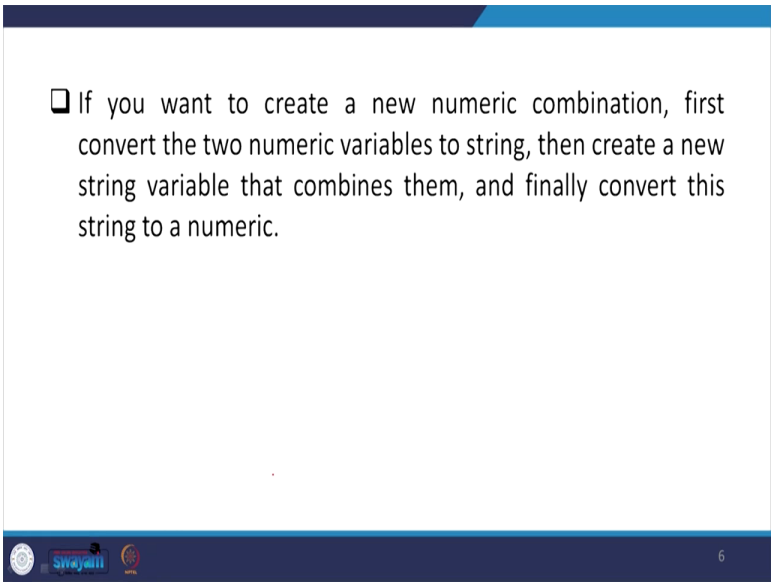
First what we need, to convert that numeric to string then only it will be combined. So, the numeric variable to be convertible to string first. So, what we do, convert the numeric to string, then use this particular command, generate string, but if your year one, you wanted to country and year if the characters, number of characters you enter here as the str and its number of characters then is equal to two variable we wanted to combine this and this, isn't it?

How to make it string, the right command is tostring command, tostring, then the variable name, country code is in, if it is in, if your year code is in numeric and country code is already in string, this is in string, in our easy understanding country code should be in string. Year code, year is generally a numeric, if they have given even it in string code, then you need not change it. You need to check whether they are in string or not by clicking browse as I told you. Browse data, if

the color comes out to be red then those are the string. If it is not in string, then you convert this variable with this command.

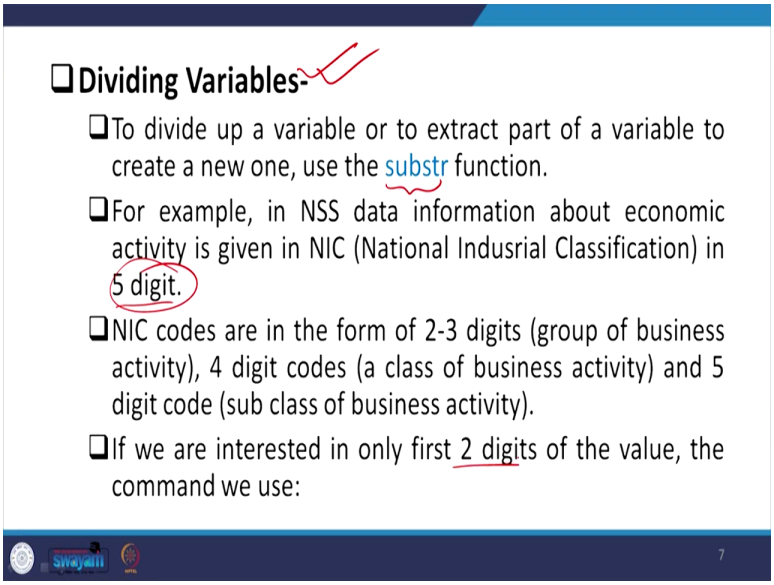
You just add this variable name which is in numeric, convert to string then you change that replace with a new name, because since you are changing as I already suggested that whenever anything you are going to change, please change it to a new name. Give your new comfortable name, maybe year, you can say year string. You can convert accordingly, now you can add or combine these two variables.

(Refer Slide Time: 21:51)



If you want to create a new numeric combination, firs convert the two numeric values to string then create a string variable that combines them and finally convert them to a numeric. Now your combination will be of string one. We have already guided you destring command is going to be useful. You apply destring then that variable, it converts to a numeric variable and then you can able to work out with all kind of mathematical applications or command you wanted to do you can do it.

(Refer Slide Time: 22:27)

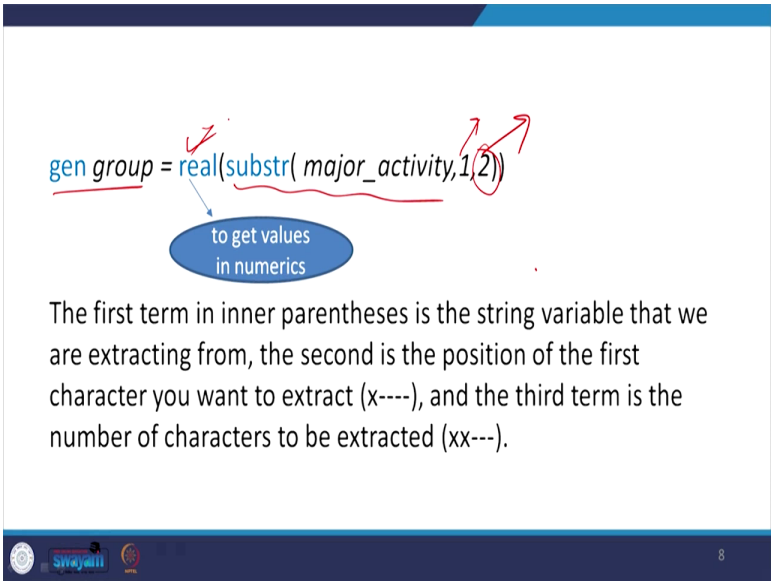


This is also important, dividing variable is also quite important in Stata and that too in our databases whichever are available. To divide a variable or to extract part of a variable to create a new one use the sub-string function. Like let me just tell you very clearly, I have already guided you that our data comes with NIC classification, let me guide you first, NIC classification of the enterprises we are considering for the NSS seventy third round, it gives in 5-digit level.

But 5-digit level, what do you mean by 4-digit label is there must be some meaning. The 5-digit level that composed of like 2 to 3 digit combinations that basically of group of business activities, or 4 digit codes those are, they are out of that 5 digits, generally 2 to 3, till the 3 digits they keep in group of business activities, then 4 digit a class of business activities, then fifth one is the sub-code, the sub-classification of the business activities. At a very disaggregated level they give the information if it is of 5-digit classification.

But if it is 5-digit classification, the number of enterprises are so many. It is very difficult to track. The activities are very difficult to track and that too very difficult to interpret. So, if you can make it to 2 to 3 digit or 2-digit classification, the broad classification of the enterprise, which type of enterprise they are operating if you can make it, so 2-digit classification is quite important for you. So, generally, first two digit if it is important for you, then you sub-string that and as per our command this, you will get the result.

(Refer Slide Time: 24:38)



Generate the variable name you define two digit. Sub-string is important, I am going to break it here. Sub-string and in our variable it is major activity which you have just seen, I have just shown to you. ours is the, most importantly in the parenthesis what is important you need to mention. Parenthesis you have to mention which the string variable you wanted to sub-string. Here it is, for us, it is major activity. The second after comma this one, the second one is the position of the first character you wanted to extract. So, out of that 5 and that you want 2 digit level from the beginning or from the end, it depends.

Generally, for the 2 digit level it is given from the beginning of our data, broad classification given from the beginning. So, you just specify one digit, one there and after the comma that is the third term is how much digit it has occupied. The 2 digit level classification wanted to extract from the data, it has occupied till second space, second character. So, the second character, you have to mention two stands for here second character. There are 5 characters in that string. So, the second character you have to mention accordingly.

Once you have done it, your sub-string is ready. Your data is extracted to a string data with two-digit classification. But why to again run another command called destring to numeric for analysis, for mathematical operations, so a single command can help you here. Just simply add real then within bracket sub-string then the inner parenthesis should be with the information. If you do that, your data is now ready with numeric data.

And so basically we have guided you here today that how to label the variable, the values and understanding string variable with numeric variable to combine those variables which is very, very important for analysis. With this, let me close this. And we will start a new chapter in the next week on understanding combining data, bigger database like extracting the data, merge, appending, merging. So, we are going to do it in the next class. Thank you.