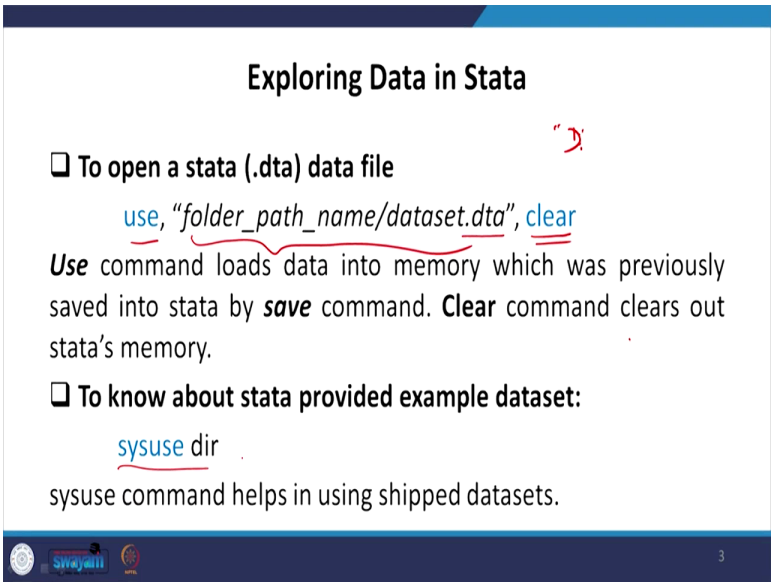
Handling Large-Scale Unit Level Data Using STATA
Professor. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture No. 25
Review of Commands

So, friends once again welcoming you to this forum of understanding the large scale unit level data with Stata. After understanding so many command over last two weeks, we wanted to enable you with the review of such commands. I am deliberately keeping a small session on review of commands in order to make you used of all the command so far explained. Otherwise, what really happens in my experience that some of the commands we usually forget. So, in order to keep you in touch with all the important commands, we have kept a session on reviewing it. Let us go together.

Why we are doing it, because we are heading for analyzing unit level data and we need to recall all the commands so far we have used. And the commands are very useful in understanding the data for analysis and interpretation. And this gives a better background and structure of our data and also helps in understanding our variables correctly and converting it as per our requirement. let us understand how to explore the data in Stata which are the commands.

(Refer Slide Time: 01:44)

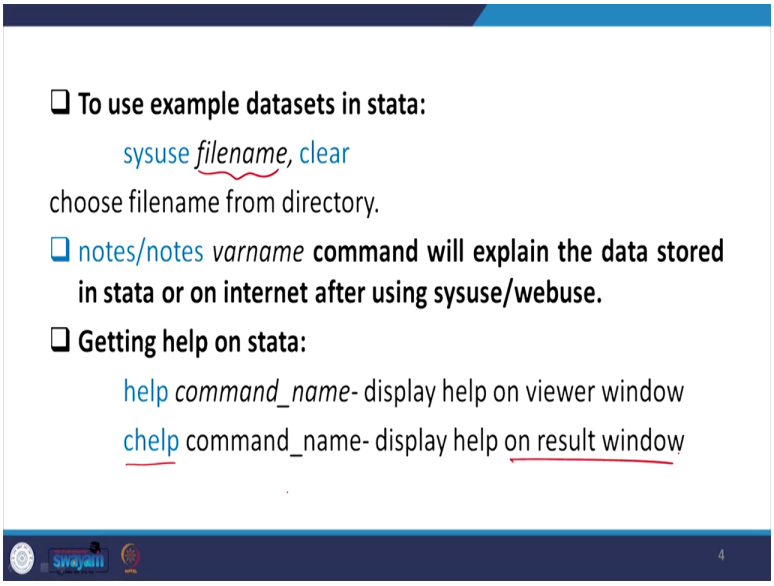


So, we have discussed that the use, if I type in the use command then the folder path, I am not going to operate in Stata because we have already operated. I am simply saying it so that you can recall. So, use is important and the path name is important. So, path name usually comes if your data is in D folder, so it comes with D. Path name starts with double inverted comma D colon, it comes with D. D then other path name, then folder name is there.

It automatically, I told you, you just go to that folder and click on the top bar and copy it, then you copy that path name of that particular folder and ended with dot dta, generally comes with dot dta if you have a Stata data then at the end you enter with, end with the inverted comma, closing inverted comma, then a clear. Clear basically clears the earlier existing memory, so that you can start with your fresh option.

So what we do then I told you that where to experiment your data, how to know it. Initially we told you that if you do not have any database, you do not know where to start, I suggested you to go by the directory, system data, system use that is sysuse directory. So, we opted for life expectancy, you can also go for system use auto data. So, there are different datasets we showed to you already and those will be very very useful for quick operations.

(Refer Slide Time: 03:43)



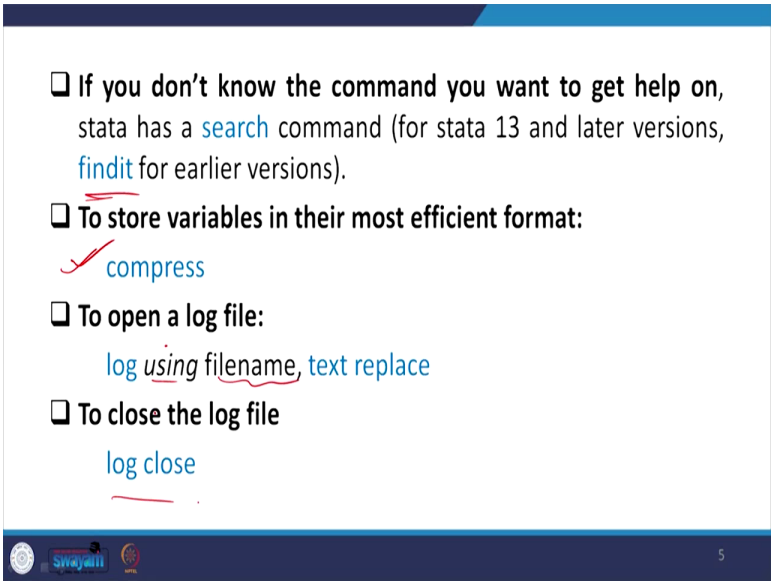
Then the directory out of the so many directory, if life expectancy which we operated many times if you simply type that and comma clear that will open the example data for you. Now

comes to the first command called notes. Notes and the variable name, variable name so many variables if there, it gives you the information about that particular variable. So, notes is also important to start with understanding the data.

There is some help command and chelp command that is also important to note. Help command most often used when our computer gets stopped and some commands are not running correctly, so help commands opens in a new window. And help with the particular command if you are confused then you simply type that. Sometimes it is not installed in the version of the Stata not in the ado file in the Stata, it gets attaches with the ado version, ado file of your Stata.

But chelp, if you go by chelp, not help, chelp displays the result on the same window. So, that displays on the result window which we are operating. So, that is also important.

(Refer Slide Time: 05:17)



If you do not know the command you want to get help on, so in that case you have to search, search command. So, Stata has a search command. Basically search command is for the 13 and later versions and findit for the earlier versions. findit was the command we used to give for the previous version than 13.

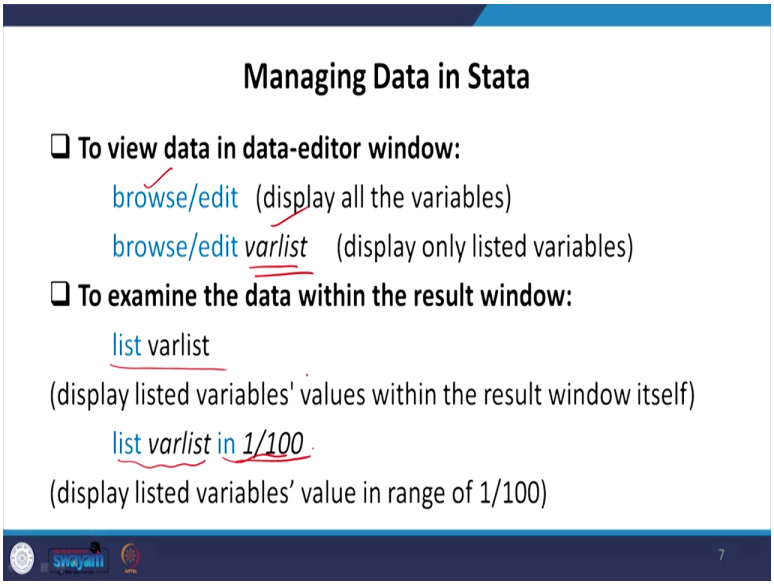
To store variables in their most efficient format, it is always suggested every time, you simply go by a compress command. So, usually the dataset we are dealing with, they already come up with compressed data. But when we do our own experiment, we do survey and enter the data so it is

better that we should write down compress and it will automatically compressed to its base format.

To open a log file, how to open a log file, I told you that you go to either in manual format or you simply type log using a file name, give a file name and then text replace, in that case it opens the log file. To close the log file, simply you type that log close. Log close will close the log file which you have opened. So, all the commands we are operating will automatically be saved.

Then save the modify data. Suppose, generally while we click on the close, it asks for save. So, it is better to save and give a file name. Type save and file name and then you can exit the Stata version for better operation. So far we have discussed how to explore the Stata.

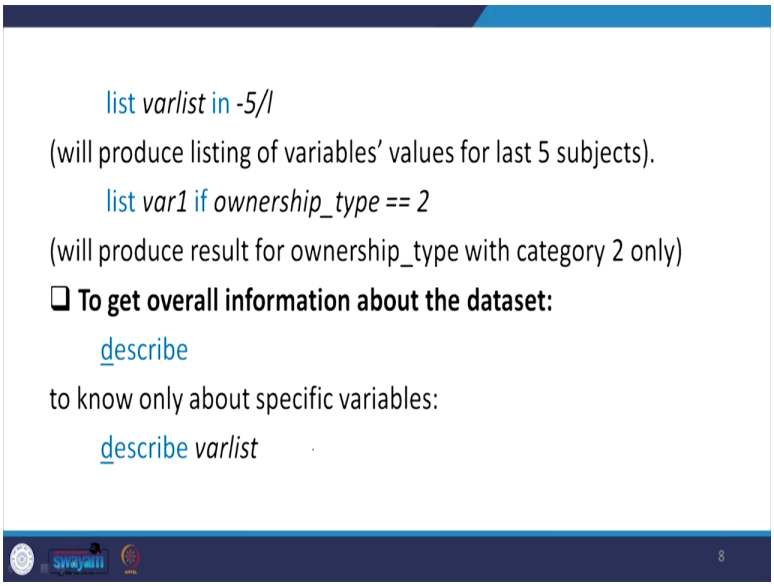
(Refer Slide Time: 07:11)



We are now going to talk about how to manage the data in Stata. So, to manage the data in data editor window that is the window we are referring every time, so either you go by browse or edit. I told you br in short or generally we suggest to go for br. Br will not de-stored the data. But when you are doubly sure you are going to edit the data, change somewhere then you can go for edit the data. And then once you do that, br you simply click browse or edit, displays entire variable, whereas if you write down the specific variable for display it will only specify and show you those variables in the browse window.

To examine the data within the result window, you need to list it. So, list then variable name, it will show you within the result window. Then this displays listed variable values within the result window itself. It will not deviate or it will not open in another window. But br, browse window gives the result in another new window. Come to an important part. If you wanted to get certain variable within 100, 1 starting till 100 they need to specify. Specify that the list, list variable in that is 1 out of 100. So, this basically gives the range from 1 to 100.

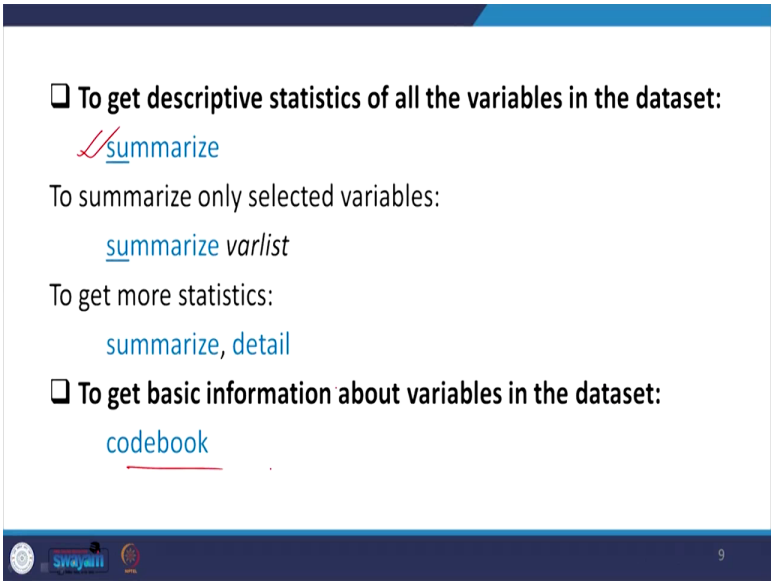
(Refer Slide Time: 08:58)



Similarly, the minus indicates the last five subjects. So, if you want the last indicators that will give you the last. Will produce you listing of variables' values for the last five subjects. So, list variable 1 if we are doubly sure that we wanted to get the list of that variable with a particular category out of the ownership type. We have already shown you that if you wanted to get the particular category that is 2, so it will only produce the ownership with the category the code with 2.

Then today also, in our other lectures also we discussed many other command like describe is most often used. So, to get overall information about the dataset, overall information about the data regarding the entry and byte space or entry in numeric form or in long form or in string form those information can be derived through describe window. Describe if you give it or des or only d that gives you the information. To only about specific variables then you enter the variable list. Every time I am saying that variable list wherever required you just enter accordingly.

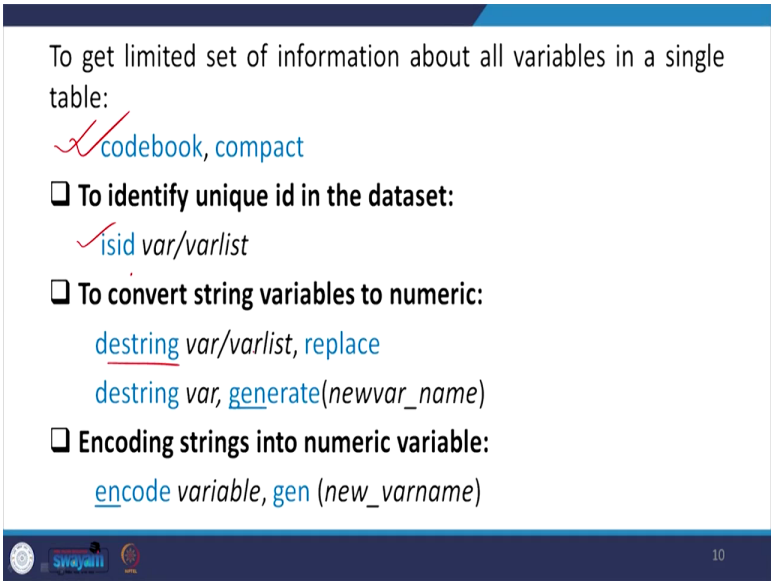
(Refer Slide Time: 10:29)



After getting this summarize is another most important aspect. Then summarize with variable list, sum. It is very essential for paper to be communicated and the important variable you are using for your model; you should summarize it. Sum and the variable list it will give you summary like the average, it gives the minimum, maximum, standard deviation of it. So, those details are displayed through summarize command. So, if you wanted to get even further details regarding the summary, you have to attach the comma with a detailed information.

Let us come to some other information like codebook. What are the codes taken for that particular variable, which type of codes are given? So, codebook if you type it, it gives the entire variable and their codes.

(Refer Slide Time: 11:29)



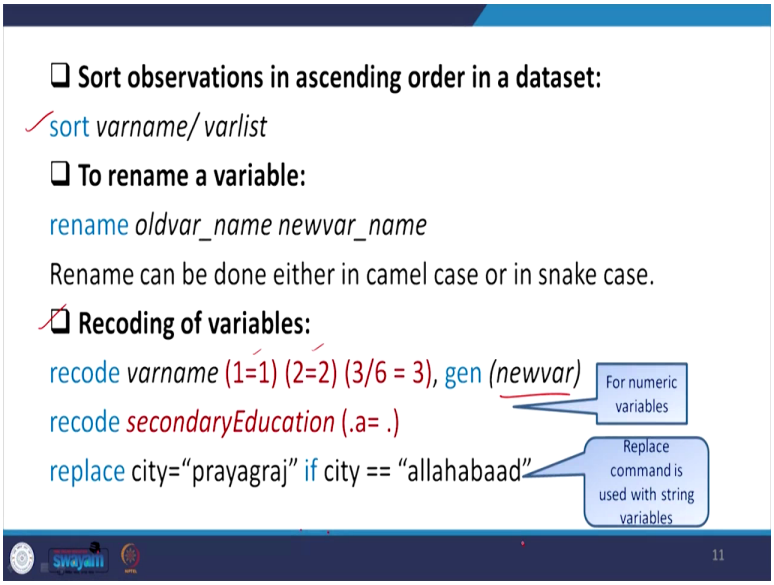
So, codebook, compact displays you within a compacted table with their relevant codes, variable information and so many things are displayed and for better comparison. So, codebook, compact I will suggest you to go for instead of just codebook. Codebook displays a different variable separately. It might be very difficult to compare them together. And to identify unique ID of the dataset that is most important especially while we go for combining the data or making a panel data.

So, in that case isid command is important to understand whether that particular variable is uniquely identified across the different groups of information given by the dataset which particular variables or the variables' information if you can take that will combine the dataset. So, isid command identifies whether the variable is key or not. The principal variable or not that can be identified we have already explained to you.

Several times I kept on mentioning that if the variable is in non-numeric form, mathematical operations are difficult, so you need to destring the variable. To destring the variable we have to enter the command like destring then variable list then replace, it replace the earlier memory. Then destring then variable, suppose you wanted to write a new variable name after destringing because we also require the string variable as well for further operations. So, it is suggested that you destring the variable and then generate a new variable with your new convenient name.

Encoding the string is also required, string into numeric variable, encoding of string into numeric variable required in that case we need to encode the variable and accordingly we can generate a new name to it.

(Refer Slide Time: 13:33)



After understanding all those minimum operations specially in case of merging we need to sort the variable name. If you want to keep it in ascending order or descending order, generally in ascending order we go by sort. Sort the variable name it will sort it. To rename the variable, most often we use it, rename a variable. Simply you just type rename then the old variable name then you type the new variable name you want to rename. We want to give the new name to the old variable name, that will rename.

So, rename can be done either in camel case or in snake case that I mentioned several times. So, either continuous, there should be continuous space. You should not give any space in between. Or if you wanted to separate two names, you need to start with the new name with a capital letter. So, that you can do it in the rename. But the name should not have any space in between.

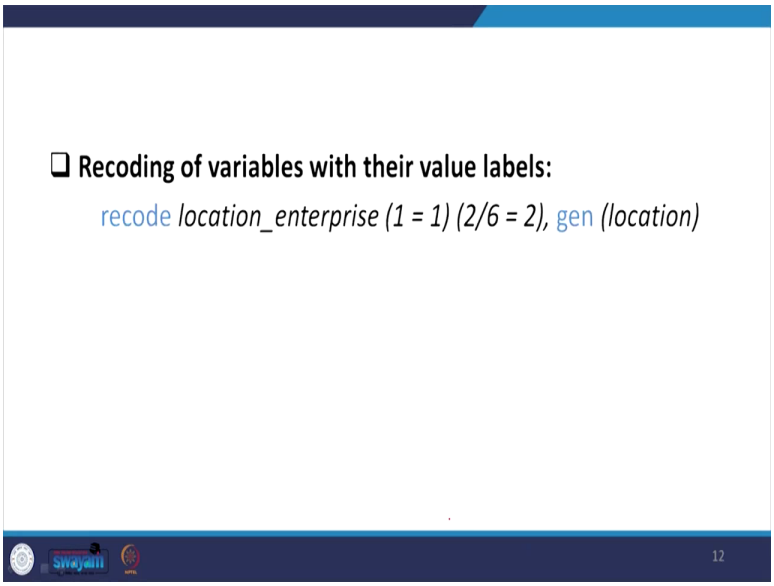
So, comes to recoding. Recoding is important, recoding of the variable with their codes. Like in this case which we have already explained recode the variable name 1 is equal to 1, 2 equal 2 or specially if 3 to 6 we wanted to club into 3 only. There are 3, 4, 5, 6 any entries are there, we

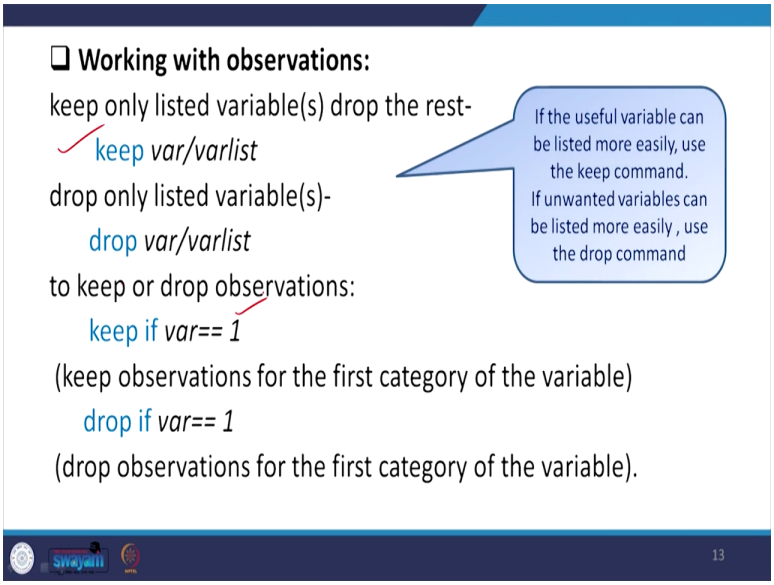
simply wanted to make it into three categories, 1 will remain as 1 keep as 1, 2 as 2 and 3 to 6 we wanted to make it 3 then that will recode.

Since we have recoded a new variable, it is in a new format, better to give a new name of that variable. Recode of the variable name with dot a, so basically what we wanted to say that for numeric variables we can able to recode like this, but like if dot a is there like some missing format is there simply what you do because that is in non-numeric format so you simply make it to only dot, that reads in the numeric in the mathematical operation. If you do that, that will convert it or recode it.

Replace is the command, can replace with a particular entry with a new entry. So, replace command is used with string variables generally. So, the string variables here is Prayagraj, if the city name is Allahabad, you can name it to another one.

(Refer Slide Time: 16:28)

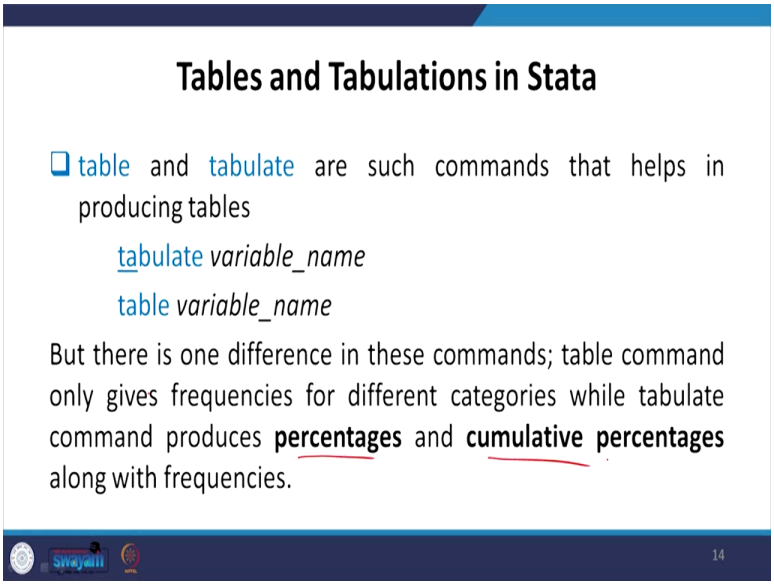




Recoding of variables with their value labels. Value labels that we have already mentioned, so I need not mention separately. You can go to that particular respective lecture and you will identify. Coming to the important variable and non-important variable. When these variable are very important to me and those numbers are very less then keep command is very useful. Simply keep those variables, other variables should automatically be dropped. Otherwise, when very specific variables are to be dropped, you then go by drop command. Other important variables will be kept in the file to make your space very comfortable for your work.

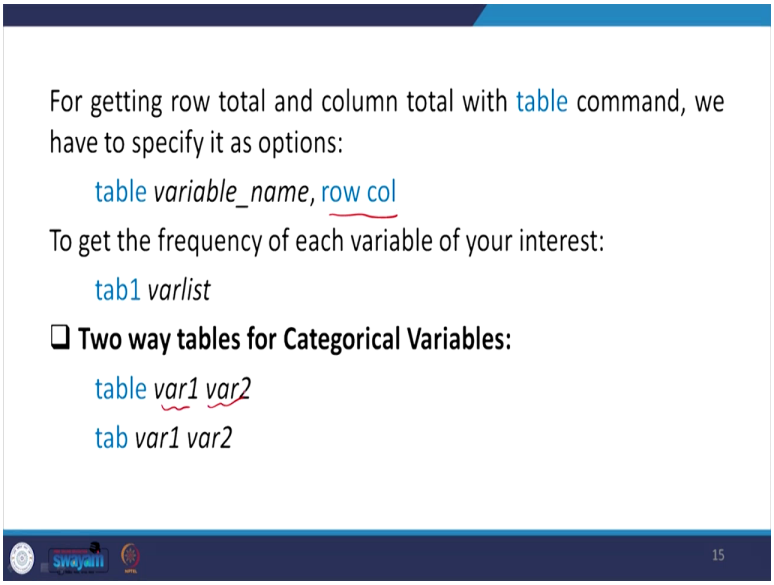
And similarly, keep with, keep variable of a particular category can also be mentioned. When you know that particular category to be continue and others to be dropped then you simply keep with that particular category. Similarly, drop as well.

(Refer Slide Time: 17:28)



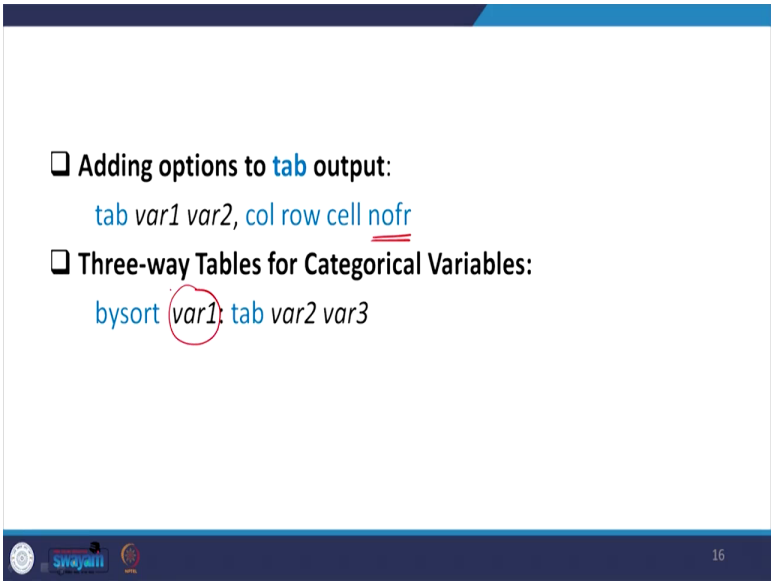
Some couple of aspects like tables and tabulations in Stata, how to do tables and tabulations. So, the tabulate that is tab or ta of the variable name will give you the frequency distribution. But the table gives information about, let me mention that, there is one difference in these two commands, that table command only gives frequencies for different categories, while tabulate command produces percentages as well, percentage and cumulative percentages. So, that is the difference between tabulate and table. We generally go for the tabulate command.

(Refer Slide Time: 18:20)



For getting row total and column total with table command, we have to specify it as option like this, either you have to specify row or column at the end so that it will displays accordingly that we have already shown you. If you have any difficulties, go to the respective class and watch the video, you will find out. So, to get the frequency of each variable of your interest, then you simply write down tab1 then variable list, it will give you correctly. Two-way table if you wanted to, if I am interested for two-way table, then you have to specify two variables together then it will give you the correct result.

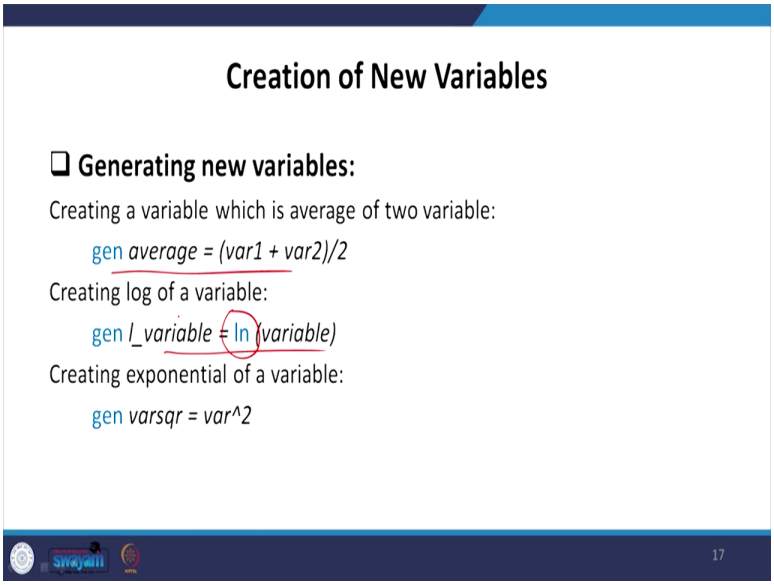
(Refer Slide Time: 19:01)



So, adding options, if you have any other options like, as I just mentioned for row wise frequency percentage and column wise percentages, then col, row or cell, if you are only interested for no frequencies then frequency will not be displayed, only percentage will be displayed, then nofr option should be mentioned correctly.

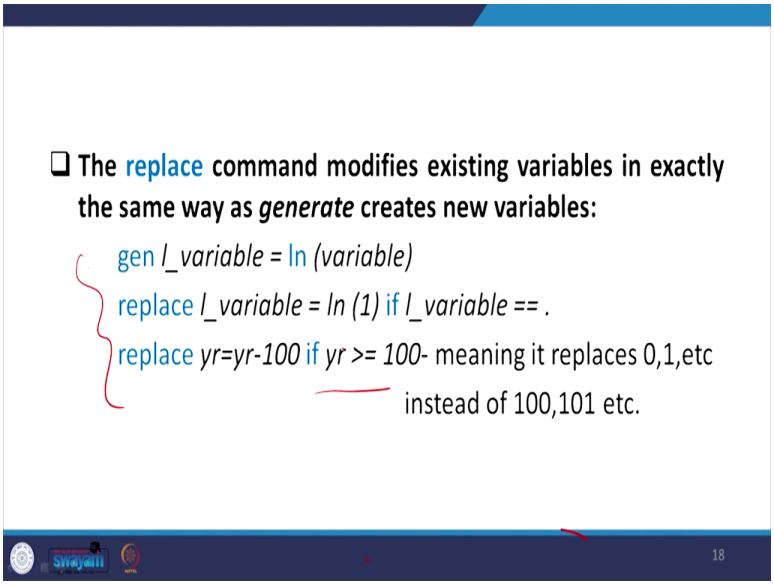
three-way table, basically two way and two cross tabulation and within that another variable, if you wanted to include these two information with another information, we wanted to get the information within another third variable, in that case the third variable we have to bysort. We start with the variable of interest that is variable one, let it be, then you sort it through the variable 1, then you can tab variable 2 and variable 3 which we have guided.

(Refer Slide Time: 20:02)



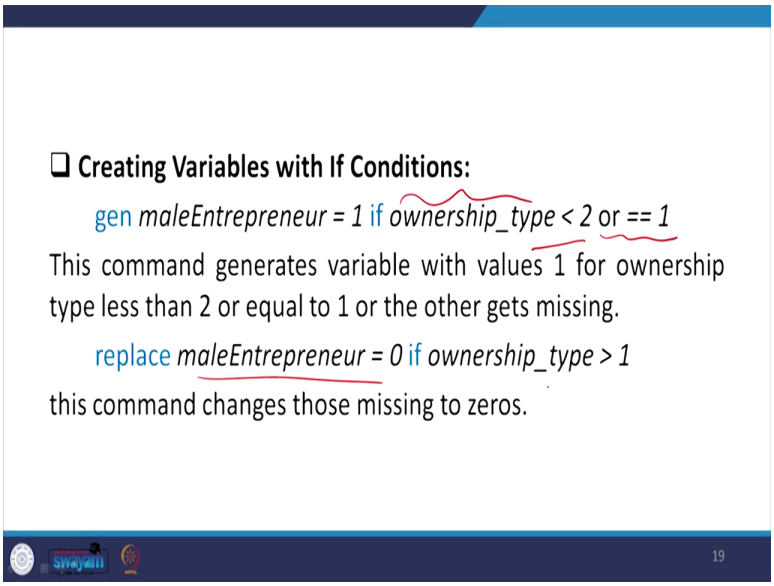
Coming to creation of a new variable, if you wanted to create a new variable, then the most important command we generally go for, why to connect a variable because sometimes we want averages of two variables, sometimes we want logarithmic transformation of the old variable and we wanted to get a new variable simply you generate and take that name if you are doing average then you go by this command. In logarithmic transformation you go by this command natural log and the variable name. Similarly, square of a variable and other commands you can do it.

(Refer Slide Time: 20:43)



So generate, replace and other aspect within 100 or like we discussed in our case the years of operation we have already mentioned you go and check you will find out other details.

(Refer Slide Time: 21:01)

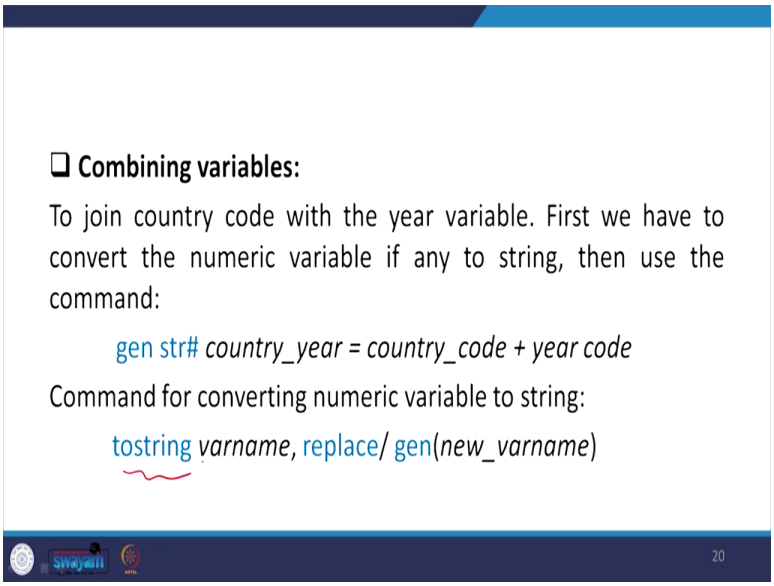


I just wanted to mention another aspect if you generate a variable that is if it is male entrepreneur with 1, if the ownership basically that is male entrepreneurs we know that it is with the code 1 in your original variable if we have, we know that it is less than 2. So, either you enter the command as less than 2 or you enter the command that is double equal to 1 that is ownership

type double equal to 1 it will give you the information with a new variable that is male entrepreneur with 1 as the entry.

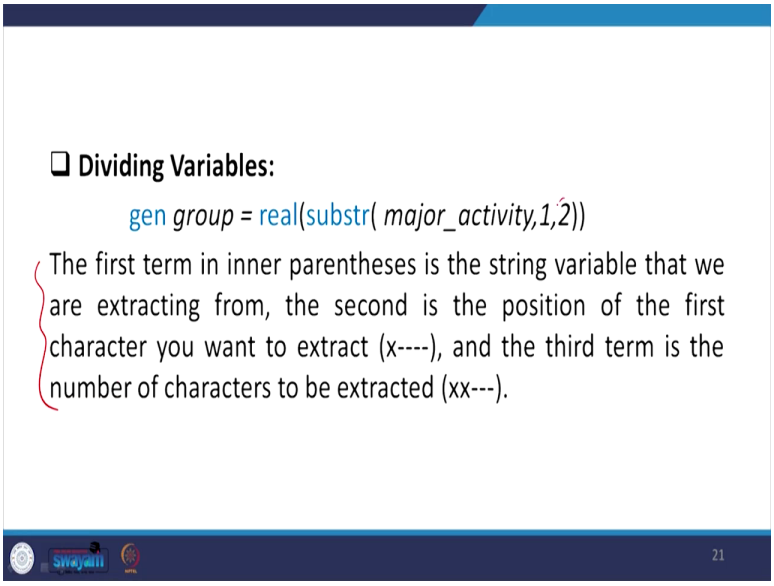
So, similarly, this command can generate variable with values 1 for ownership type less than 2. And similarly, for replace command, we can enter as 0 if ownership type is greater than 1. So, if it is greater than one, we simply want to replace all other 1 as to 0 because we are interested for male entrepreneurs. So, it will replace all other values to be 0. So, that means whatever is left is nothing but the male entrepreneurs. So, this command changes those missing to 0.

(Refer Slide Time: 22:16)



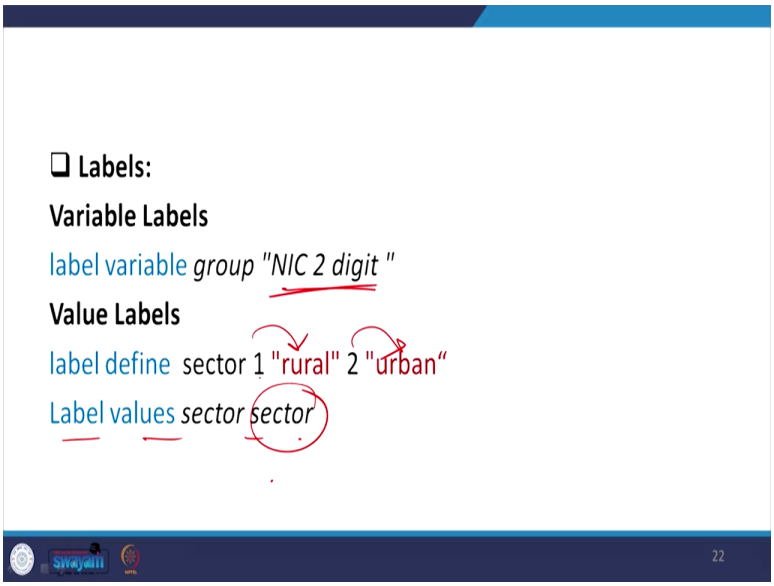
So, now combining variables, we can combine variables through either if like to join country code with we have shown, given you the country example country code and their years, country code with the years information, first we have to convert the numeric variable if any to string then you use the command. Generate then string with the year number is equal to country code and year code. Then command for converting numeric variable to string and accordingly we can mention tostring command that you can see it through the data.

(Refer Slide Time: 23:02)



Similarly, to divide the variable between major activities with its command, I think we have already guided you, you go and check its space, which space to be counted and that we have mentioned correctly. I am not going into the further details.

(Refer Slide Time: 23:16)

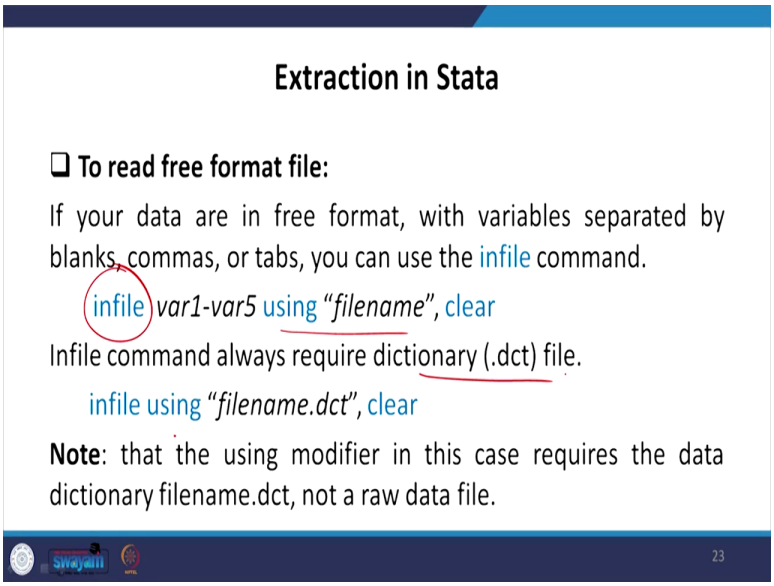


And I am coming back the explanation of labeling a variable. Labeling a variable, labeling variable group number NIC 2 digit, so if I just label that it gives the variable with this

information correctly, NIC 2-digit. Against to that variable it will show it. Similarly, value labels and variable labels, there are two information we clarified very clearly.

In case of value label, the entries that is one or two values have given that value, if only one is there, we label it with rural and two with urban then it will show with the label that is rural or urban. So, label values then you have to enter another command label values sector and sector, if your sector is defined as a label variable then we can enter, that we have guided during that lecture.

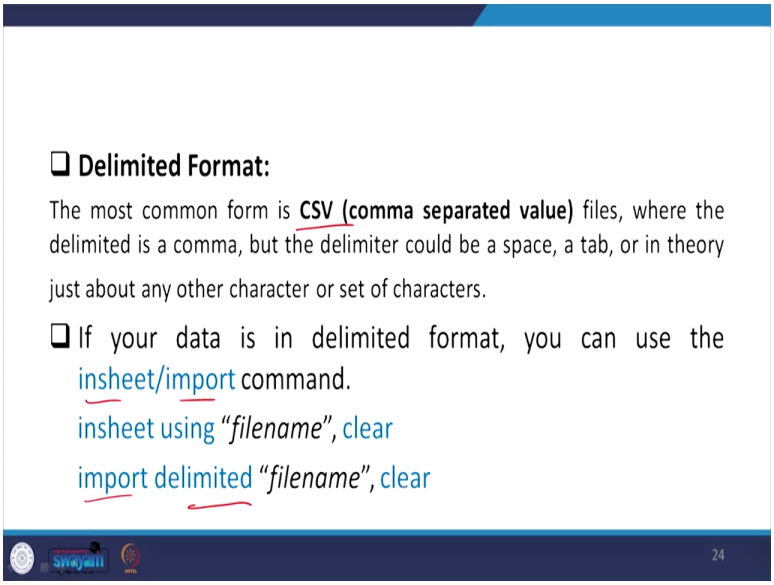
(Refer Slide Time: 24:16)



So, couple of things to be mentioned for ending this particular review lecture, just we wanted to mention that we already explained the extraction and merging lectures. So, for extraction, in case of free format file, free format data if it is there, in that case, we use the command if your interest variables are one to five in that command we use infile command if it is a free format type. Then using the file name, which file to be used we have already discussed.

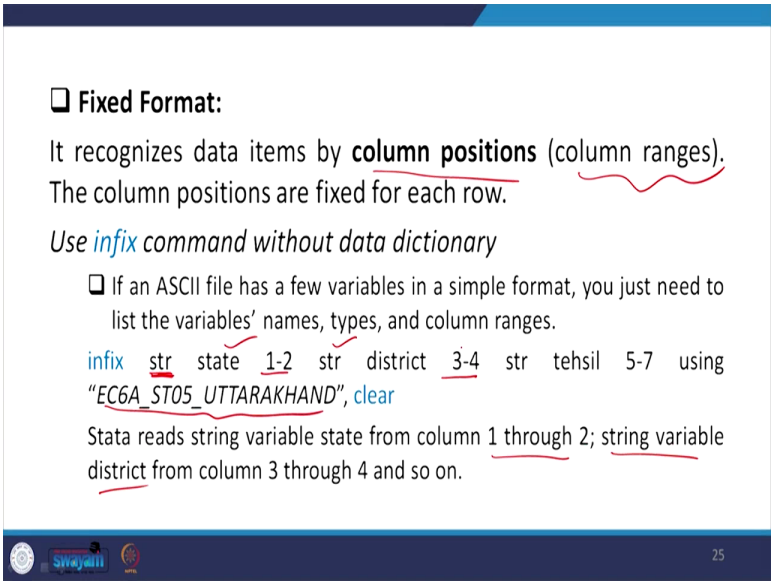
So, basically in our data in a free format type with a variable separated by blanks or commas or tabs you can use the infile command. So, infile command always require a dictionary file. Wherever a dictionary file dot dct file is available infile command is going to be very useful.

(Refer Slide Time: 25:19)



So, let me move. There is another format called delimited data format. So, delimited data is most common form with CSV comma separated value files. They come with CSV files, where the delimited is a comma format, all the entries are separated by comma, the delimited could be a space also, could be a tab or in theory just about any other character or set of characters. If your data is in delimited format, you can use the insheet command, insheet or even import command. So, insheet using that file name then clear, it will give you the extraction. Another approach is insheet delimited then the file name, it will extract it.

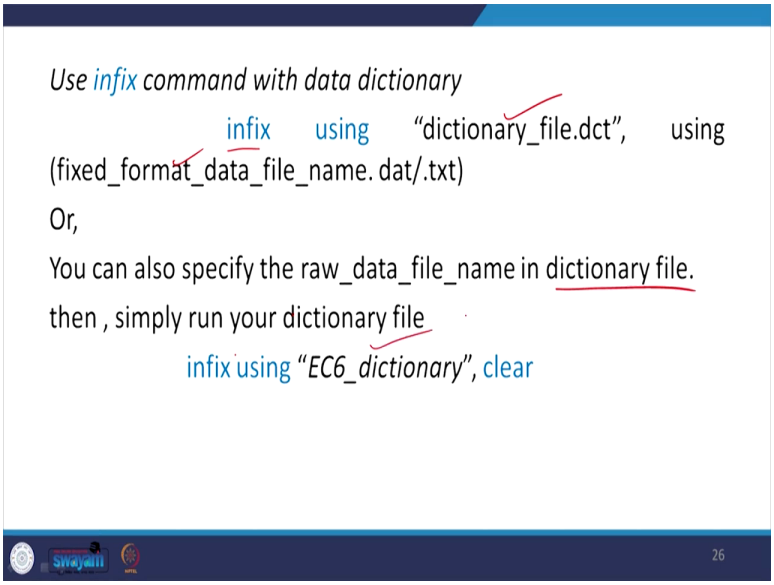
(Refer Slide Time: 26:13)



So, last one to be guided so far as extraction is concerned, fixed format. Usually this is very common in our dataset. It recognizes data item by column position only. And the column positions are given in our instruction of the data and the column ranges are also very clearly mentioned. So, column positions are fixed for each row. It is very clearly mentioned in which extent the position is defined. So, in that case we need to infix command without data dictionary.

In that case, if an ASCII format which has the clear column positions given, few variables in a simple format, you just need to list the variables names like this, infix then in that case if you want to get it in string then the state and its bytes position is 1 to 2 then district in string its position is 3 to 4 and accordingly you can mention then using where is the file name, the raw data file might have given with this name which we have already shown to you, once you do it, it gets converted, gets extracted. Stata reads string variable state from column 1 to 2, 1 through 2 and string variables district from 3 to 4 that we have already mentioned.

(Refer Slide Time: 27:46)



So, use infix command with data dictionary. Previously we said if there is no data dictionary that was the approach we said. If you have a clear data dictionary and its format is understood to you that we have already explained during the lecture, then simply you infix using the data dictionary and second using the raw data, that will give you. Data dictionary already enters with all the systematic entries with their byte space, the variable and their byte space.

So, you can also specify the raw data file name in dictionary file as well. You can clearly specify the file name as well, then simply run your dictionary file itself. Simply if you run the dictionary file and that is infix of dictionary file contains the raw data file already then infix using that dictionary that will also extract. So, couple of other things you need to also review related to merging. So, that merging most importantly requires the key variables to be merged, in that case, most importantly, the isid command is very very useful and then you need to sort the isid variables the command variable then you can merge, and you have to go by a master file and a using file.

Sometimes it is one to one, one to m, m to one or m to m. Usually, one to m or m to one are inverse to each other and you need to be careful which master file you are converging. So, we generally avoid extracting, converging the dataset or combining dataset m to m because there are some results expected to be erroneous, one to m or m to one is very very useful. So, whatever technicalities I have used right now, if you just follow the merging lecture you will certainly able to understand very clearly.

I think we have reviewed till twenty fifth lecture very clearly and the window is open to you and you start using the important other applications. Enjoy it.