Exploring Survey Data on Health Care

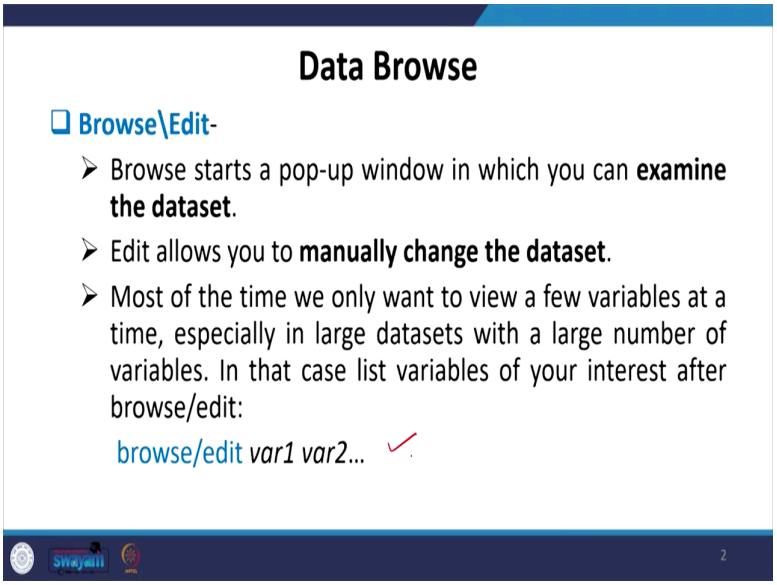
Prof. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture - 19
Data Browse and Basic Statistics - I

Welcome friends to the NPTEL MOOC module on Handling Healthcare Data. We are on the 4th week, explaining data software; specially identifying various operations with the help of Stata. In this particular lecture, we are explaining you or we are trying to give you the very basic statistics through the Data Browse. There are lots of hand holding required. So, this is what the lecture is meant for.

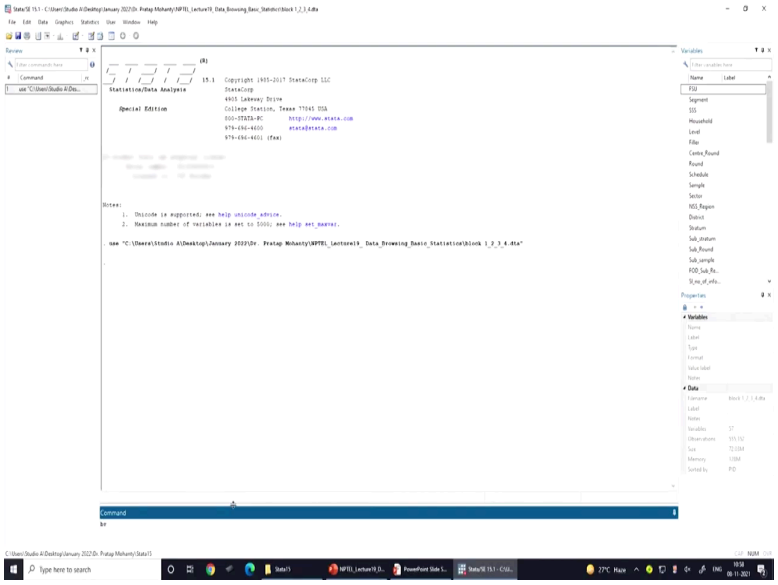
Now, without missing our time minutes let us move on and clarify.

(Refer Slide Time: 01:19)



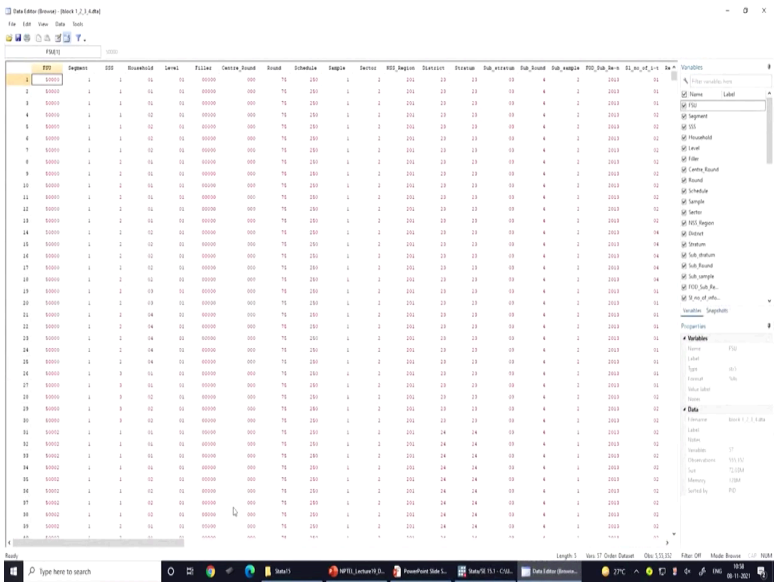
So, starting from the very basics that is called data browse, what is the meaning of it like once you have got the data on your screen; how you can be able to get your data browsed.

(Refer Slide Time: 01:28)



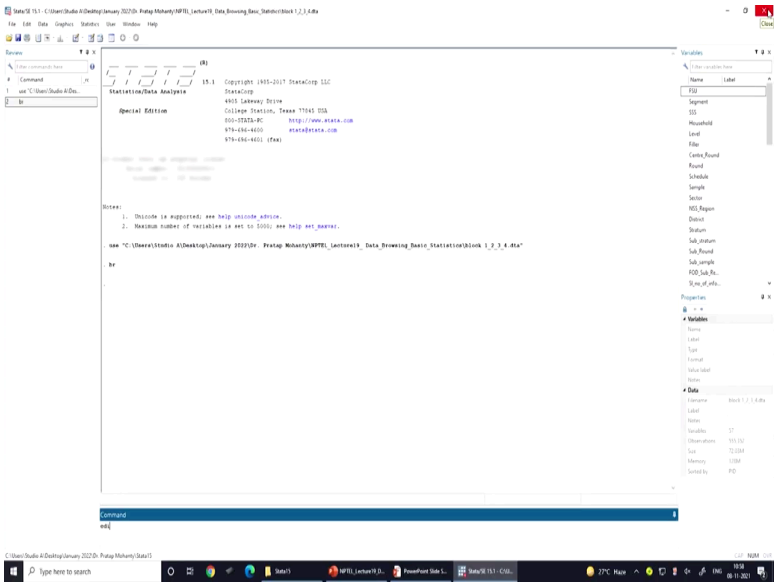
So, straight away I will show you something from the beginning which will be very useful to you. Like from the starting if I simply say browse here as ***br*** in my command and then Enter.

(Refer Slide Time: 01:39)



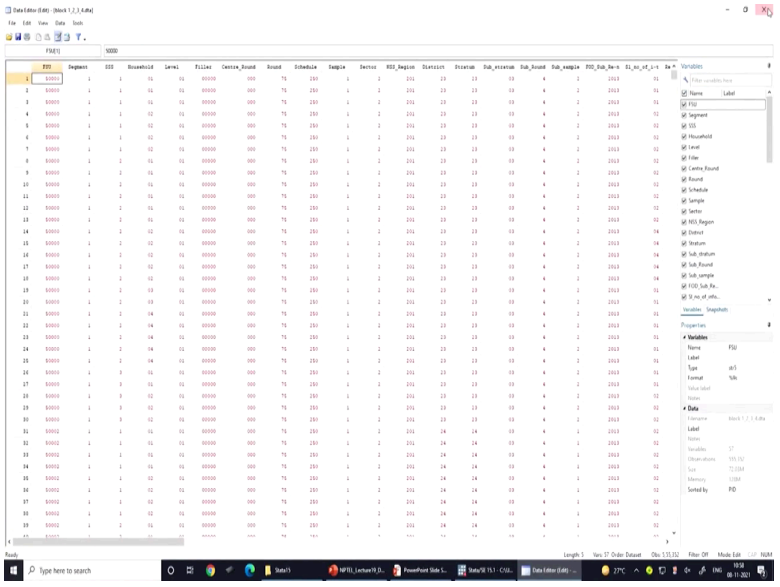
So, this is what is going to give you as the data browse.

(Refer Slide Time: 01:48)



Now, if instead of that I simply type edit here, it will give you data browse edit only.

(Refer Slide Time: 01:54)



So, that is now the difference between this and the previous command that browse command is only give you the display of the data whereas, the edit command in fact, helps you to edit the data or to change the data.

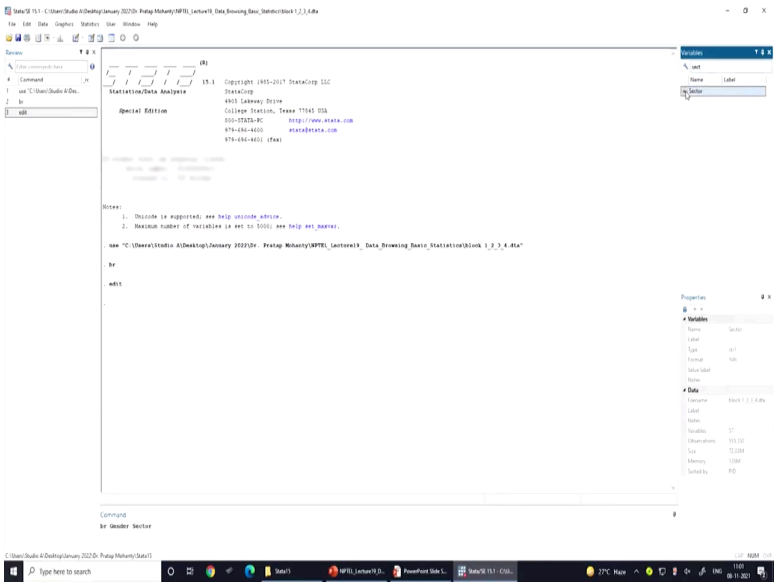
Now, each of the data points like that there are some missing entries on your data. Those missing entries could be modified (could be changed), but if you do not open your data in the edit mode, you cannot be able to change the data.

Now, let us come back to our presentation and I will clarify rest of the details. So, this is from the starting I have used the Stata window. Now, I am just explaining how we can go for all those things for those are very new to Stata they might get confused. So, I am just clarifying point by point.

So, Browse starts a pop-up window in which you can examine the data set. So, another pop-up window opened and from there we examine our data set. Whereas Edit allows you to manually change the data set if there are any sort of change is required.

Most of the time we only want to view a few variables at a time, especially in large data sets with a large number of variables. In that case least variables of your interest after browse or edit are important. Initially, you can specify with browse/edit with your required variable names, later on we can go and clarify about listing of the variables.

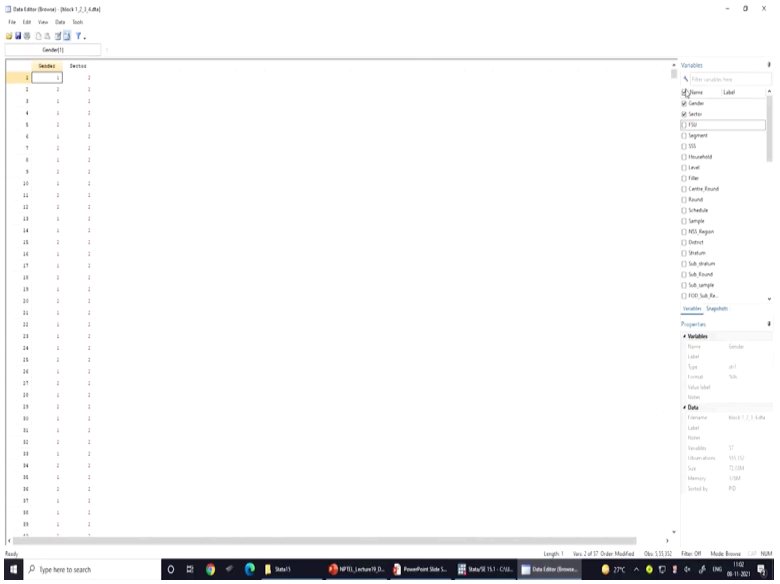
(Refer Slide Time: 04:01)



So, first of all, I have already explained you about this browse and edit, like if once again go through the screen this is here. So, I will browse two specific variable for which I require br and then two variables.

So, any two variables I am just writing; let it be on gender and I also want to know another one is on our sector (rural or urban).

(Refer Slide Time: 04:23)



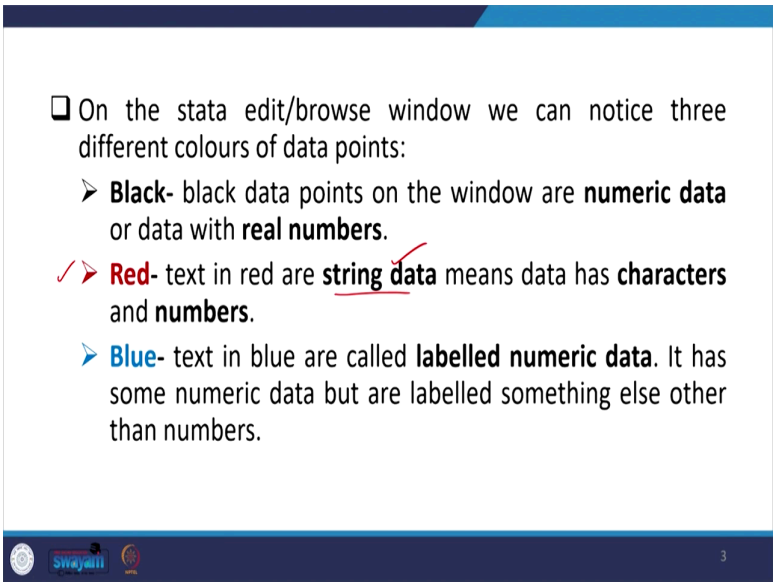
If I simply enter only two variables and their details are displayed. So, now, further aspects on these you may observe eventually, but at this moment I just wanted to show you that the variables and their values are entered; specially the values are entered in red color red font. These shows that the variables entries are of special type. The values are of special type. These are here I have explained earlier as well.

Once again, I am saying, since this is in red font this suggests that the entries are in string; these are called string variable. So, what do you mean by string variable? I think at the time of the explanation of data, different type of data sets I explained you.

String basically is the composition of characters symbols, some features are given together that does not explain any numeric interpretations. So, like your Pan card or Aadhar number these are all called string variable, each are with certain meaning or codes.

So, I am just moving back to the slide once again and this is what I have already explained to you.

(Refer Slide Time: 05:54)

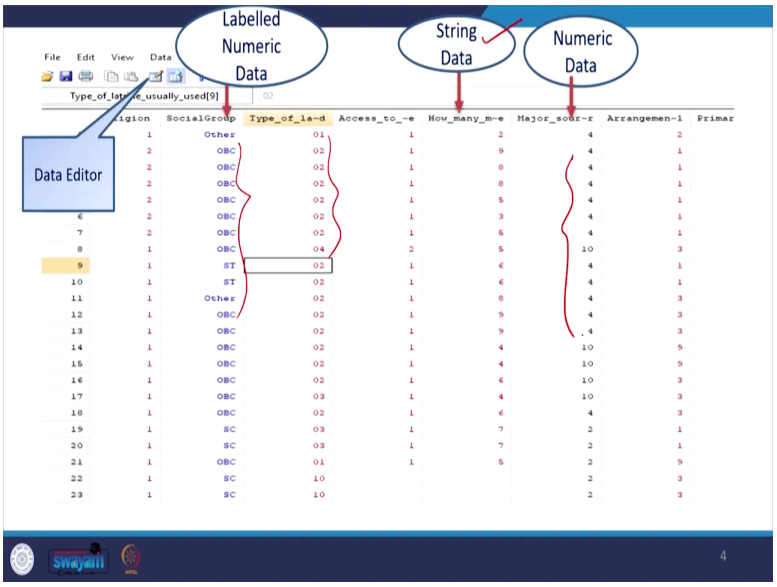


Let us discuss on another aspect like on the browsing window we can get a red font which I have already mentioned to you. Red represents string data, these are characters and numbers, but represents certain meaning not exactly by the value of the entry or of the real number. Black is in fact, if the entries are in black, they are simply called the real numbers you can have numeric estimation out of it.

There might be blue entries as well blue once you have to destring the data, string we will use it sometimes that may convert it to black or blue. If blue font is visible; that means, your data are of labeled new numeric; not just not only numeric they are called labeled numeric.

Labeled means those are value, but have certain level, certain code labellings are defined. Still some numeric forms of estimation possible like median value can be generated, then those have certain meaning. So, blue font data if any are visible on the screen there are some numeric operations possible, but in case of red they are not possible.

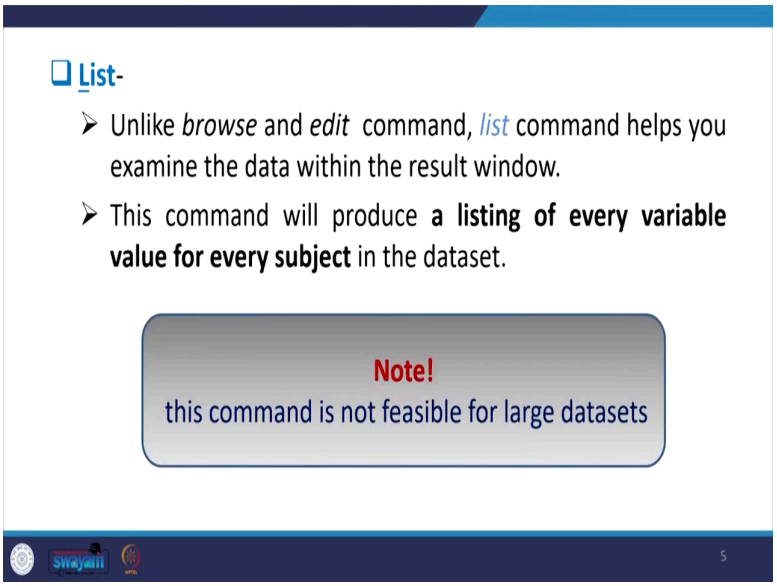
(Refer Slide Time: 07:27)



So, this is what we have captured from our own screen. You can see that here a blue in it is given alphabet. A blue could be also in numeric number as well, but if it is in numeric number that represents certain label, not with the exact value.

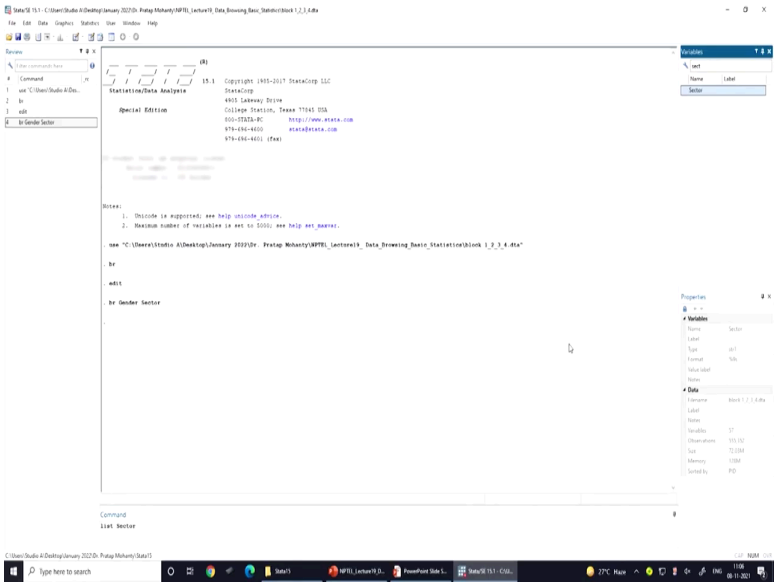
So, like you can find out median value, you can find out mode of those value, but not of course, the average. In case of string, you cannot do it.

(Refer Slide Time: 08:10)



Coming to another operation called list; a list of list command unlike *browse* command and edit command list command helps you examine the data within the result window. So, that is we have already shown on the screen.

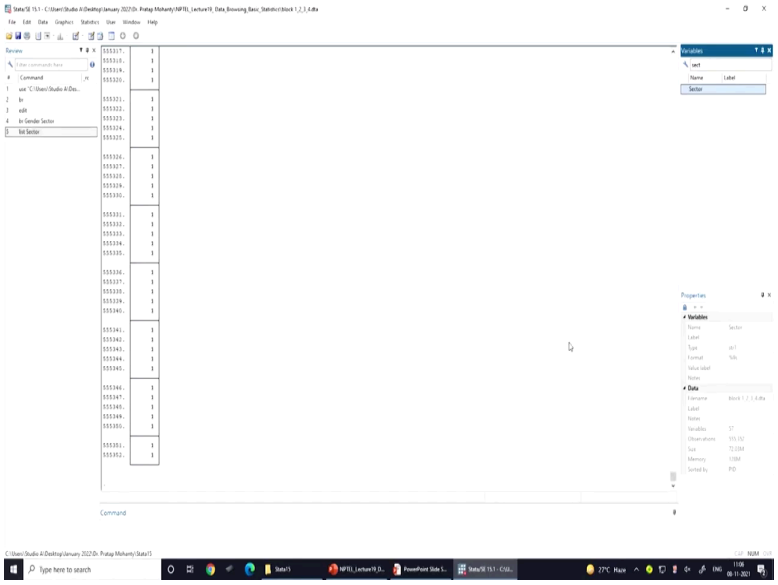
(Refer Slide Time: 08:30)



I am just closing here; suppose I say this is your data and variable are available here. What you can do, like initially we browsed and/or edited that displays result in another window. It is a new popup window comes out, but in case of list.

So, I am like writing here as list, list and variable name list. In our case list and the variable name may be sector.

(Refer Slide Time: 09:10)

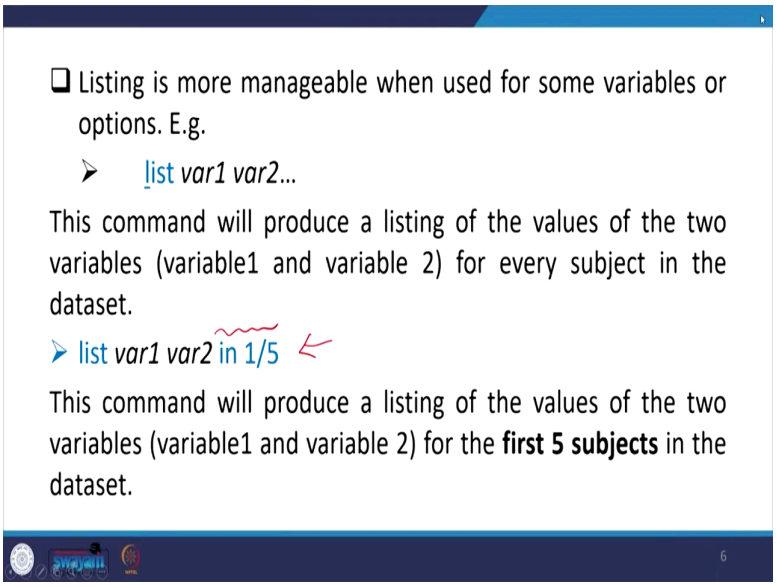


So, this is going to display the result or the table on the screen only. This gives the window or the outcome window on the screen. This is in fact, the difference between the browse and list. Like here this command will produce a listing of every variable value for every subject in the data set.

One note, we have to give it to you that is this command is not feasible when your data set is very large. When your display outcome is going to be very large; it is it is better not to use this command.

Next one is listing is more manageable when used for some variable or options. So, if it is specific one, that is it is perfectly fine.

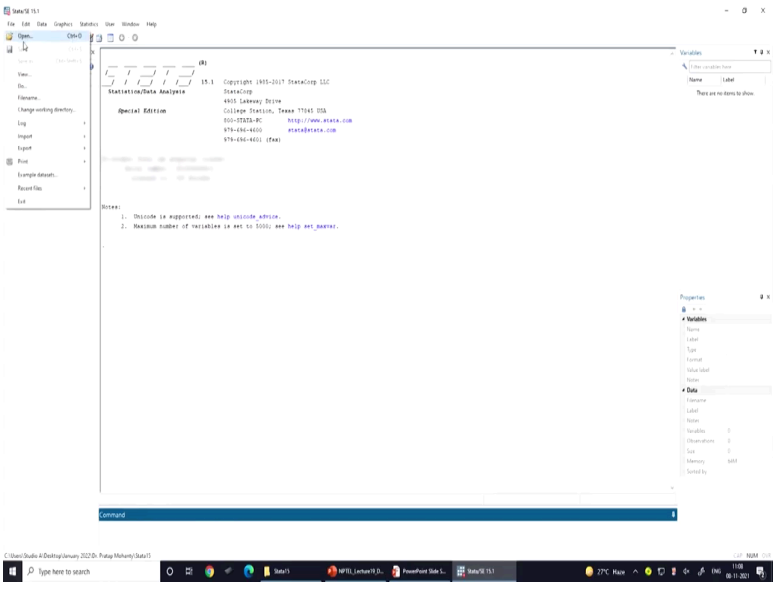
(Refer Slide Time: 10:01)



If options are given, then list is fine. This command will produce a listing of the values of two variables for every subject in the data set; like why we said listing is important for specific command, specific variables, and specific options. Options we have taken like variable 1 and 2 we have mentioned, listing we have given, but we are specifying ways 1 till 5, which will give you observations first till fifth units.

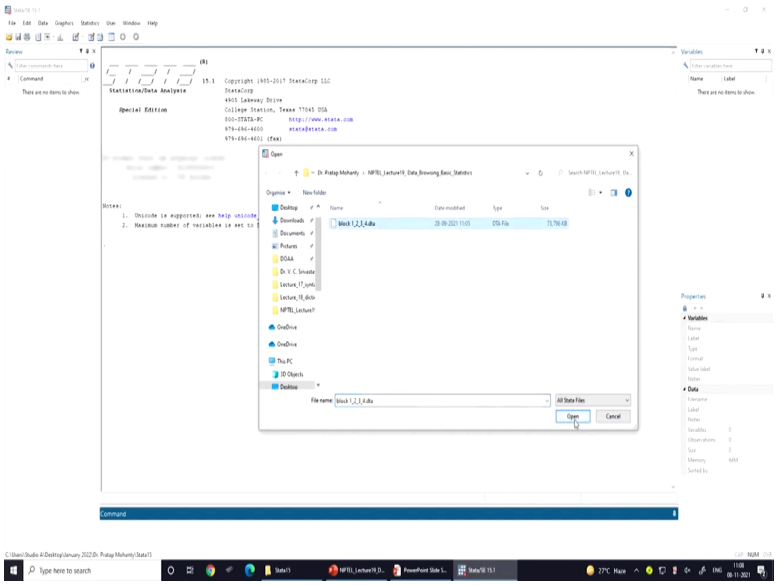
So, least the variables 1 and 2 with options like from first to five are listed not all. So, only first to five will be listed on your screen ; that we can easily do it with this command.

(Refer Slide Time: 10:52)

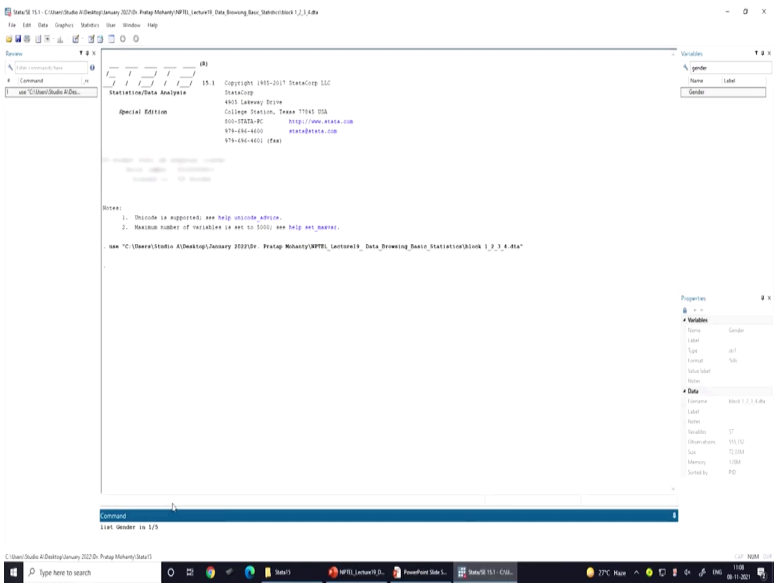


Once again, we have to open that. So, now, we will load the data data here.

(Refer Slide Time: 10:59)

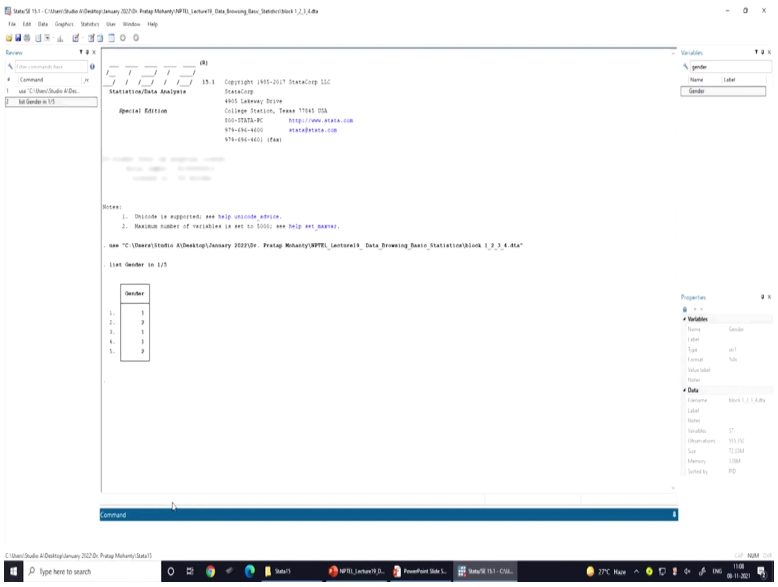


(Refer Slide Time: 11:03)



It is a sample data not exactly the data you will be working. Now, the variable we can list it; list the variable name in 1/5. So, now, this is going to list you the data from the starting till fifth units.

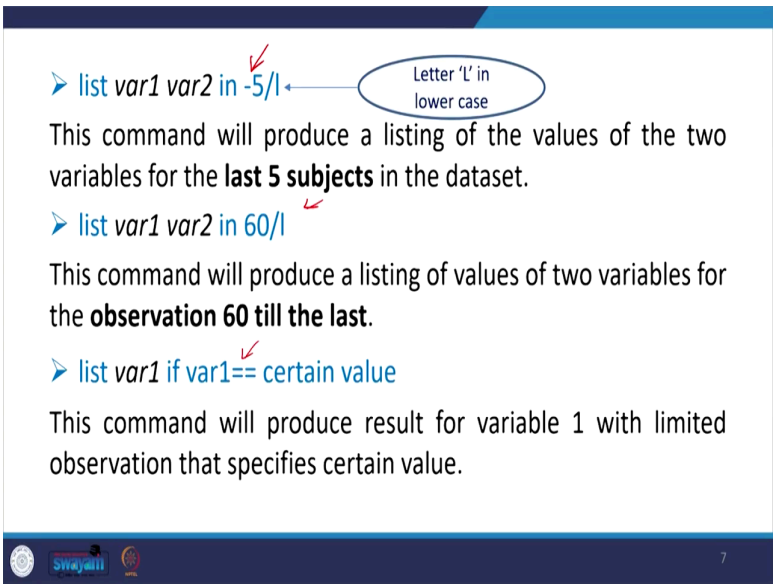
(Refer Slide Time: 11:33)



Now, enter. So, now, you can see the difference when I say that it is going to display the first five entries on the data screen. This has given you how these are actually entered.

Similarly, other entries you can do it last 5 or last 10 or first 10; anything you can do it for your further clarification. So, the first 5 subjects in the data sets are displayed.

(Refer Slide Time: 12:03)



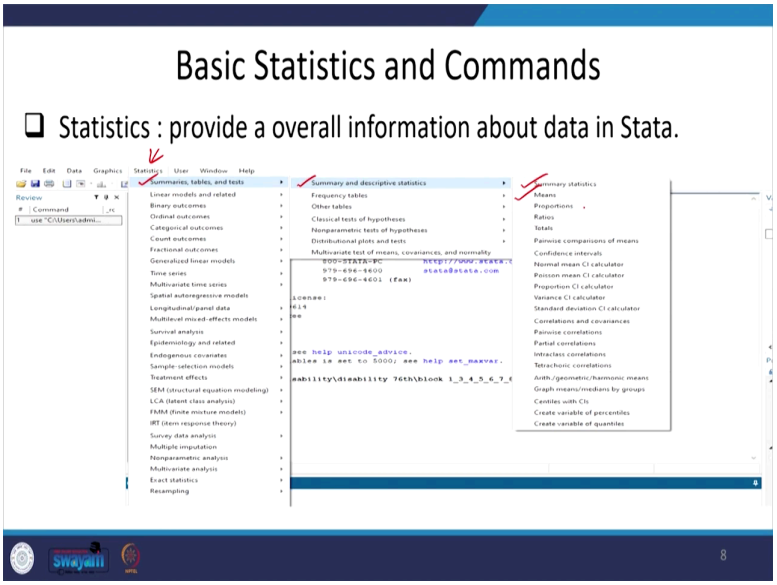
Similarly, if you wanted to look at or observe what exactly the entries in the last 5 subjects; so, it should be carrying with a minus sign 5 divided by l, in lower case. Once you give

command for last 5 it will list it accordingly. Similarly, if you want observation 60 till the last 60 till the last. So, first 60 is not going to be listed, but 60 till the last is going to be listed.

This command will produce like command list 60 and this command will produce a listing of values of two variables for the observation 60 till the last. Then if that you have to get the listing on your screen for certain value, you wanted to check how certain exact value we wanted to check how many numbers are available.

So, you can specify with double equal to sign. With a value this command will produce result for variable 1 with limited observation that specify certain value.

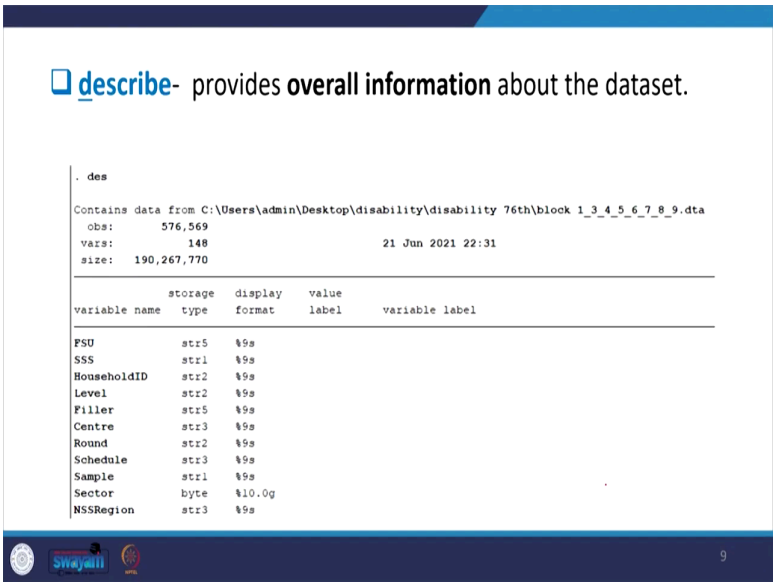
(Refer Slide Time: 13:19)



So, these are all so far about browse, edit and enlisting. Some basic statistics you can get it from the drop-down menu as well. Drop down menu on the Stata browser; we have Statistics on the top menu.

We have Statistics on the very first statistics there are first point is on Summaries, tables and test. Within summary you will get another link Summary and descriptive statistics, then we can get many things out of it maybe Summary statistics, may be Mean, maybe Ratio anything etc.

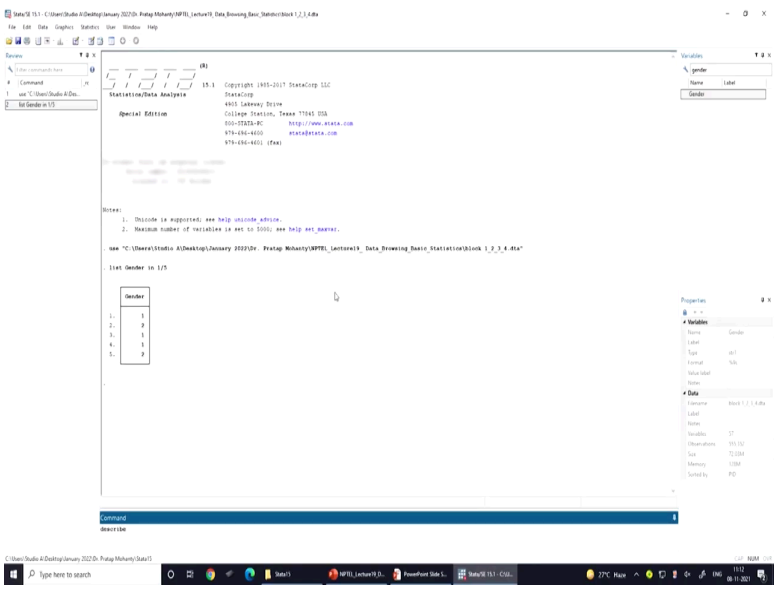
(Refer Slide Time: 14:08)



So, we are now going to operate it through the Stata command like once we have the data; the data which you have shown is from the 75th round of NSS on social consumption on health care. In the data, we will be explaining you about the describe; how we will describe the data.

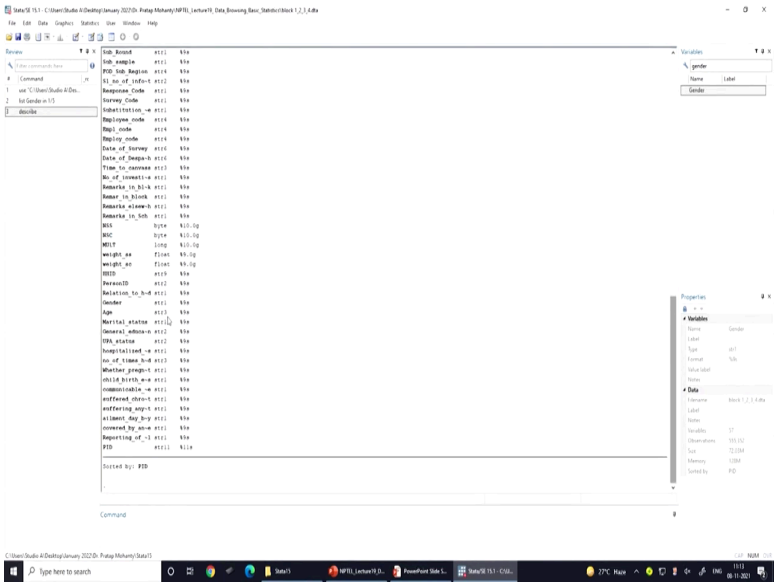
So, describe the data does mean that it gives information about the variable name and the storage type which kind of storage whether it is a string or in a numeric or not in byte space or not those clarification. Similarly, the display format is also mentioned, then value label etc. you will get to know about it. That we can do it just to describe for on the screen.

(Refer Slide Time: 15:11)

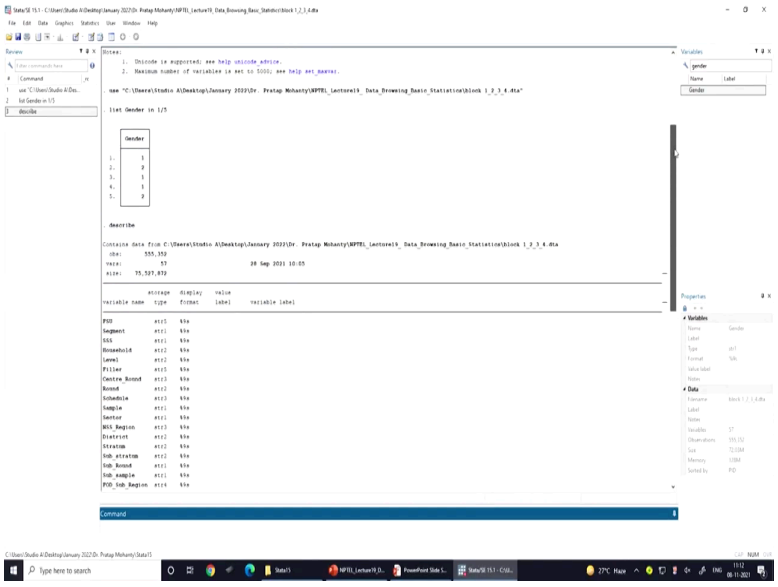


So, now I will write down here as des or complete describe whatever you want; one any variable you can do it or simply describe. If I can type describe and enter it will describe all the variables and its details, Enter.

(Refer Slide Time: 15:30)



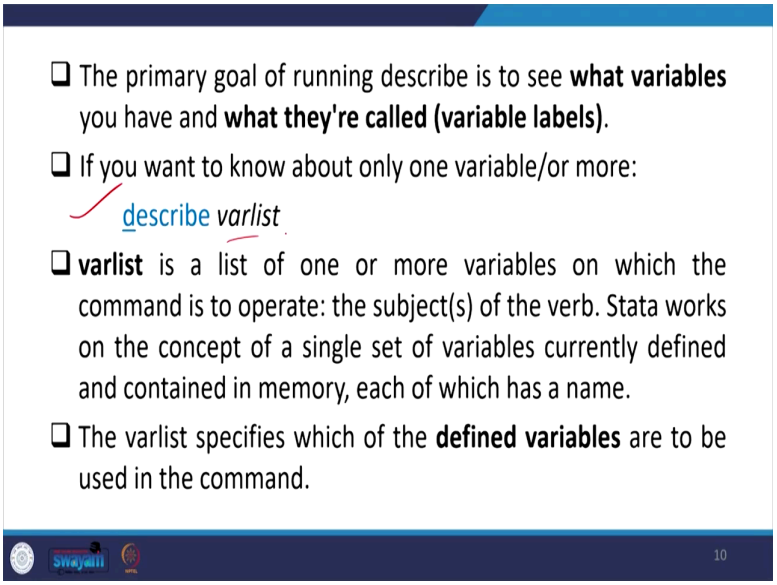
(Refer Slide Time: 15:35)



Yes, now this has given all the details of all the variables on our list. Here, number of observations it has given, it has given how many variables, what is the size of this database and most importantly about the variable name; like we are operating with state sector, then gender.

So, gender, age, marital status, hospitalized or not, number of times hospitalized etc. Since we extracted the data into string variable, so, these are all in string. Then it is other details are mentioned we will explain when we are going to use it later. So, from the describe we have got all those information.

(Refer Slide Time: 16:23)

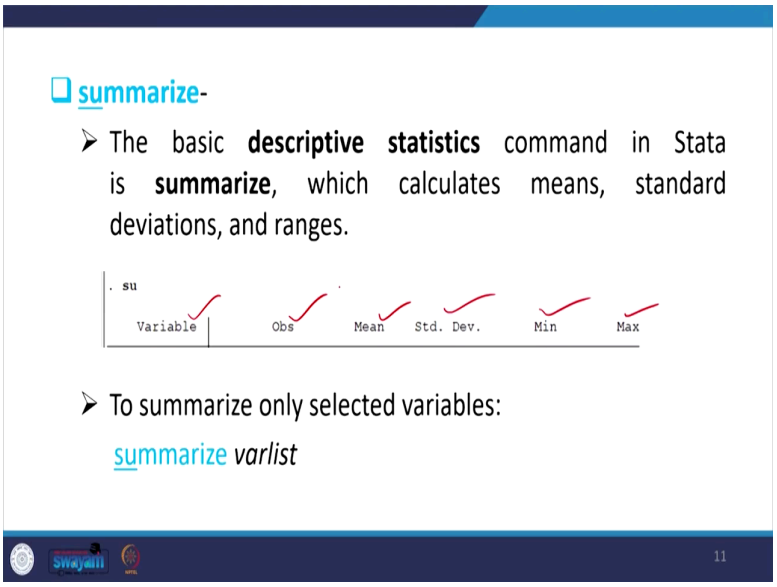


Now, the primary goal of running describe is to see what variables you have and what they are called and their variable labels. If you want to know about only one variable or more, then you can describe des or describe that particular variable name.

varlist or variable list is a list of one or more variables on which the command is to operate the subject of the var. Stata works on the concept of a single set of variables currently defined and contained in memory, each of which has a name.

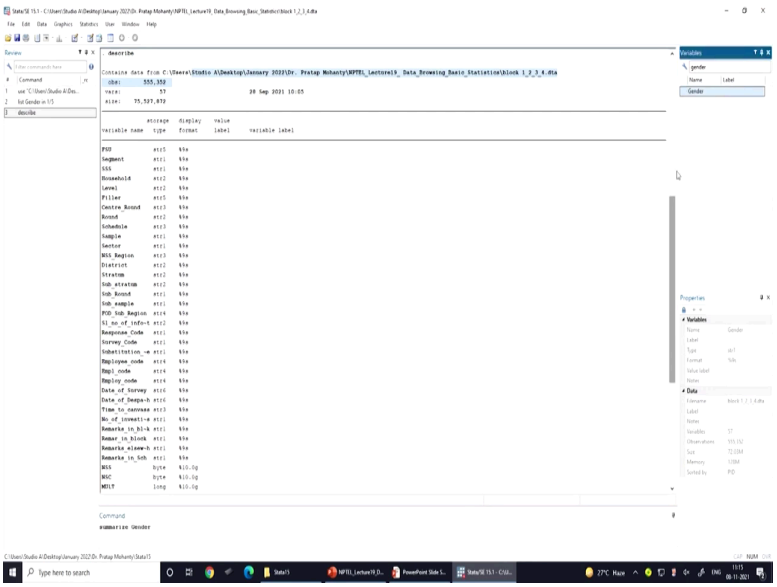
The variable list specifies which of the defined variable are to be used in the command; this is what we have explained. So, the font that is in blue is your command name and rest is your variable name, alright.

(Refer Slide Time: 17:25)



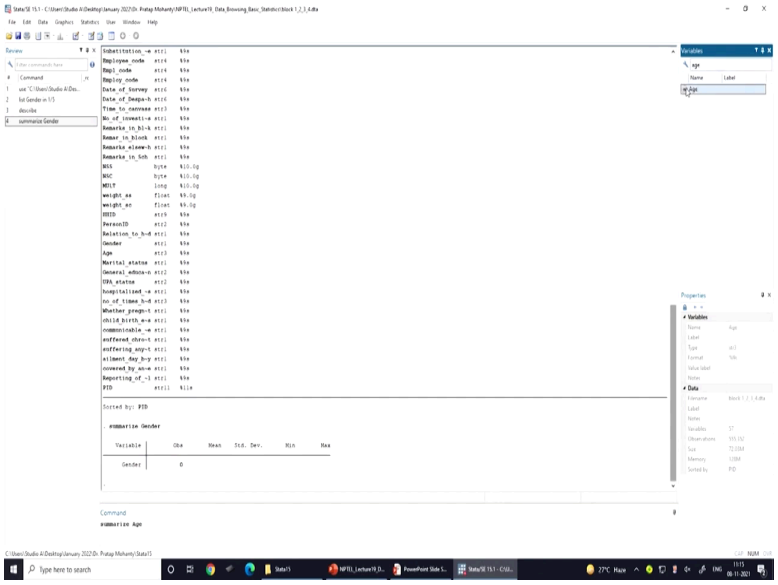
So, now next one is to understand summary of the data; first we describe the data, then we summarize the data. Summarize the data gives this information about the variable, then number of observations mean, standard deviation, minimum and maximum. Now, summarize some of the variables we can do it on the screen as well, then we will discuss about it.

(Refer Slide Time: 17:48)



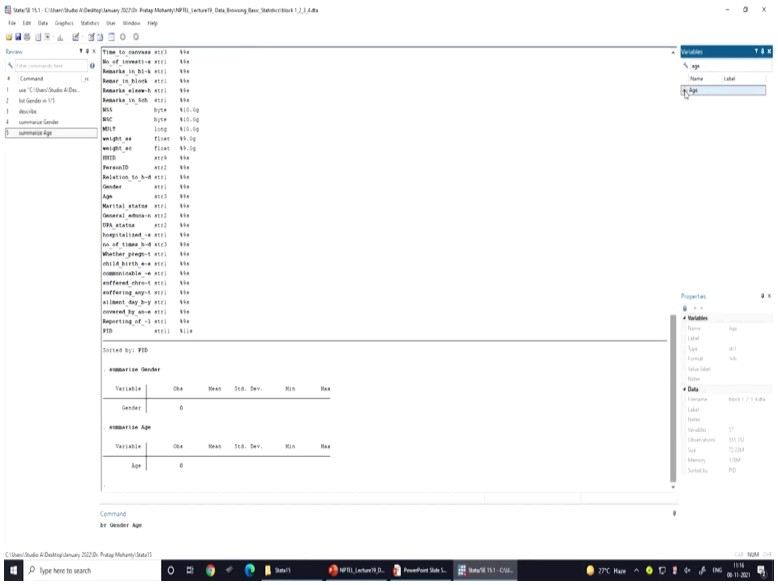
So, summarize here, then may be one variable only.

(Refer Slide Time: 18:00)



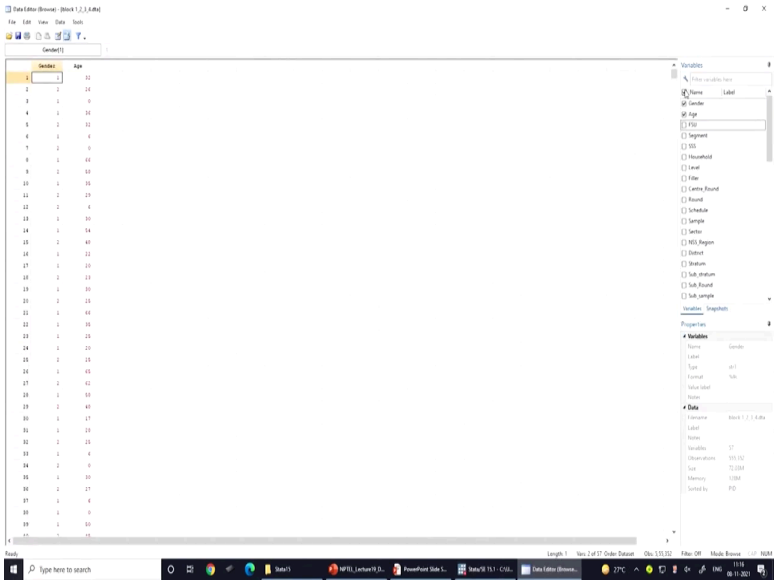
Gender as we know is a categorical one. This is a categorical data. But in this case there is no observation, we can go to check another variable. So, summarize in any numeric maybe age of the person. So, we need to open it once again and then operate it.

(Refer Slide Time: 18:31)



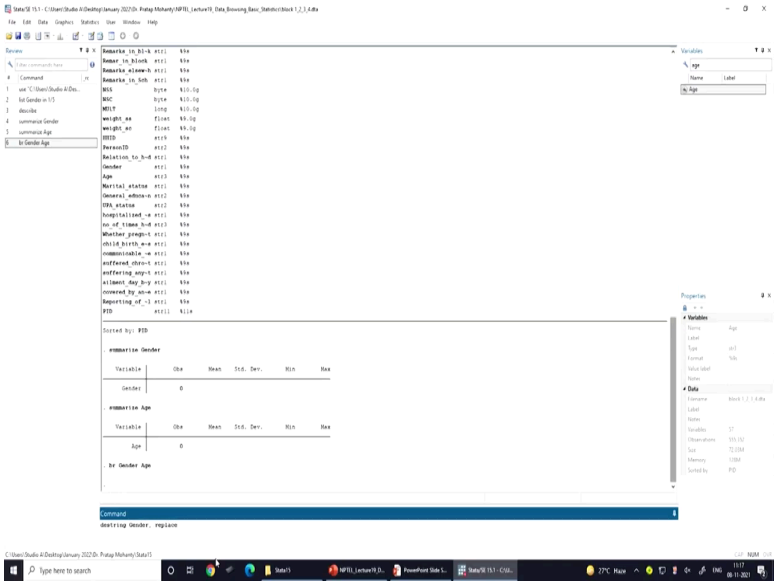
So, one problem is there that our data is in string format. So, that is why the result is not displaying rightly. So, what we will do? We will first check the browse, br of variable gender and age i.e., br gender age.

(Refer Slide Time: 19:14)



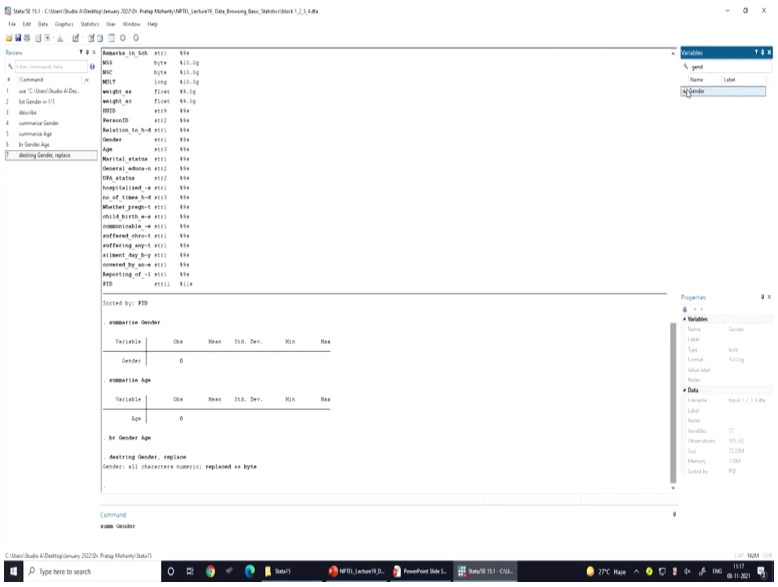
So, now enter we can see that these are in string. Data is in string. So, this in red color. So, what is required here we need to distinct first. Since we are going for summarize. So, in case of summarize it extract information about mean and minimum maximum etc. So, that requires destringing of the variables. First, we will destring the variable.

(Refer Slide Time: 19:42)



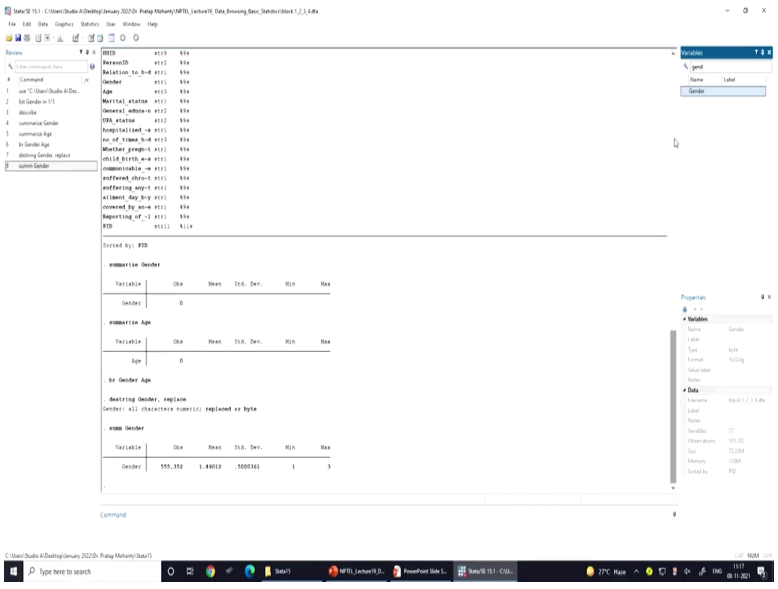
So, destring then these two variables at this moment you can destring entire variable as well.

(Refer Slide Time: 19:58)



Now, we can operate the summarize command here summary; sum only, then this variable gender.

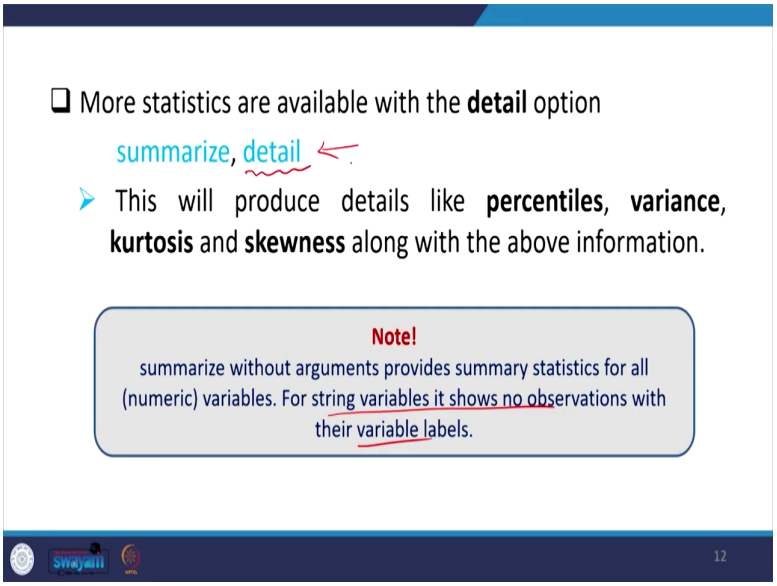
(Refer Slide Time: 20:15)



So, now, this is going to give us the result. So, we know that this is a categorical variable initially I thought it is categorical. So, it is not giving the result, but still it could have given you the result; even if it is categorical, but since it was a string variable and string variable where we cannot have numerical operations.

So, from the command summarize or summary statistics, we get a number of observations about the variable, then mean of that of the variable standard deviation minimum and maximum values. So, let us come back to our, explanation once again.

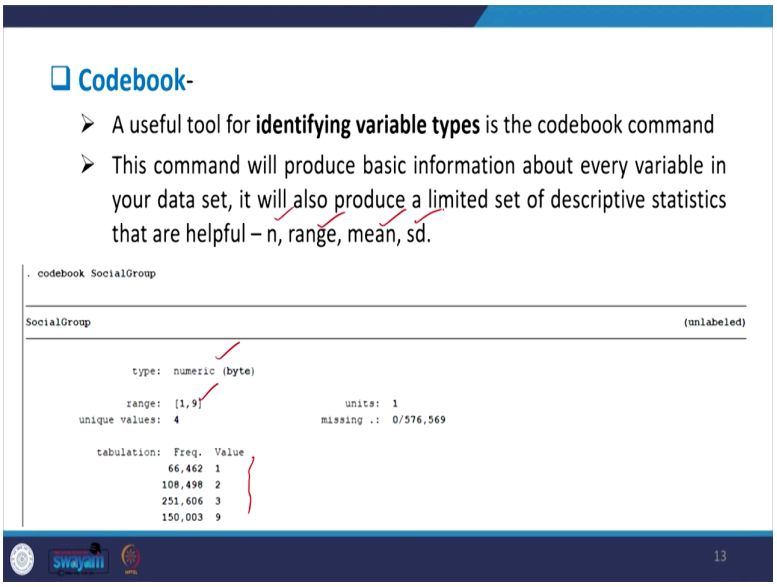
(Refer Slide Time: 21:05)



So, summarize in variable list we have done it then this is very useful. We can also get furthermore options if we apply summarize comma detail then this gives more options like percentile, variance, kurtosis, skewness along with the above information that we explained to you.

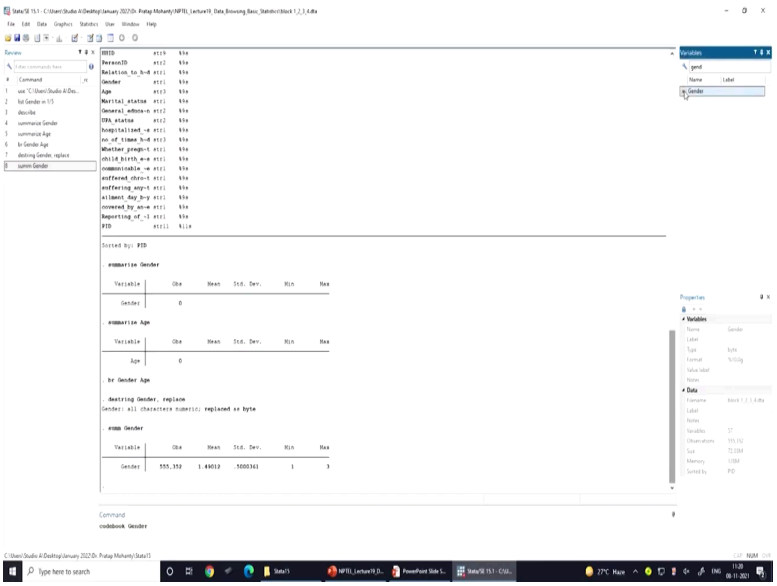
Summarize without argument provide summary statistics for all numeric variables. For string variables, it shows no observations with their variable labels this is what we have already mentioned in our PPT. So, we are not going to do it you can operate from the sample data. We are going to keep it a sample data for your operation. So, this you can just try once how other information you can get it from it.

(Refer Slide Time: 21:54)



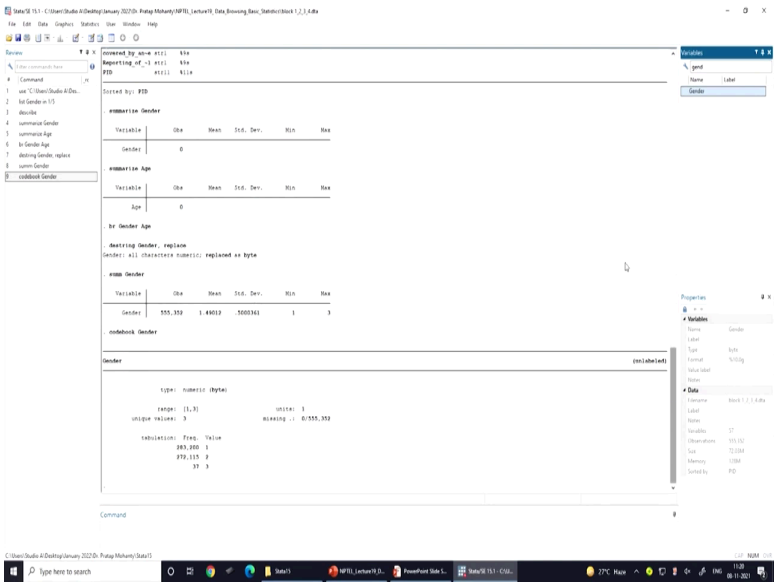
The next aspect is hugely important because this is often accessed or applied by researchers; that is codebook. Codebook identifies variable types, which type of coding are taken for; different variables. Like codebook if I give the variable name it can give us the information about whether it is a numeric or not whether what is the range of that, what are the codes entered in that particular variable.

(Refer Slide Time: 22:38)



Let us check that once codebook and gender.

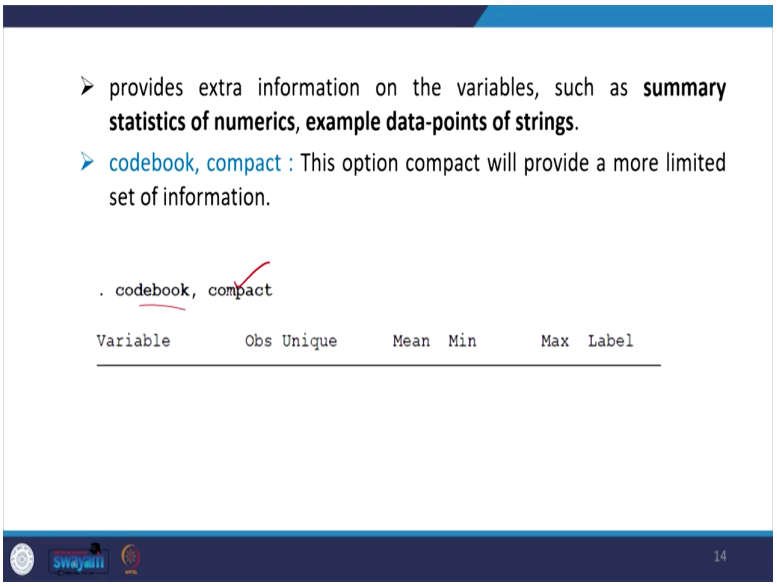
(Refer Slide Time: 22:46)



We can get to all those detail codebook and gender. So, now, you can get enough information about it; since we have already destring. So, now, it has converted to a numeric value or numeric variable. The range is from 1 to 3. 1 2 and 3 codes for gender because third gender is also taken, then the value the codebook are 1 2 and 3 and in each category what are the frequency is also displayed like on the first 283200, second and third you can also note it accordingly.

So, code book we have explained. Then this command will produce basic information about every variable in your data set and it will also produce a limited set of descriptive statistics that are helpful, usually a descriptive statistics like n range of the data, mean and standard deviation you can get it.

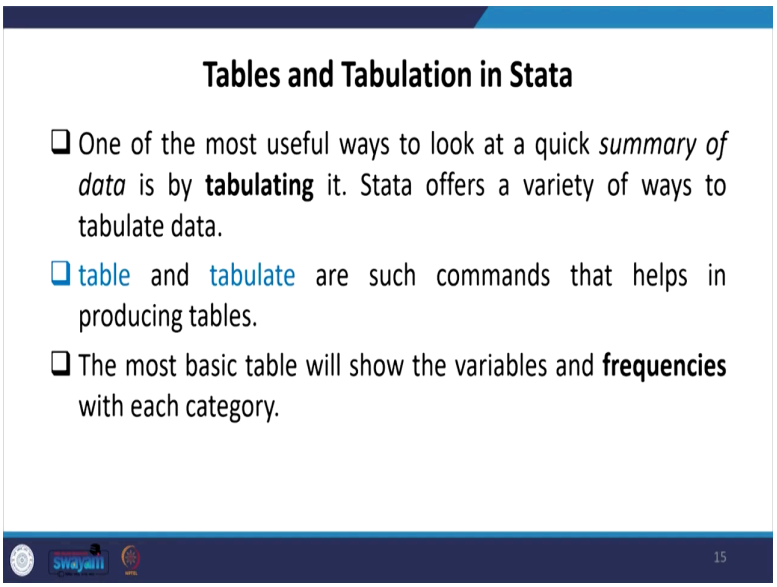
(Refer Slide Time: 23:48)



Now, this provides extra information on the variables such as summary statistics of numeric, example data points of string i.e., what string entries are there, what are their label. Codebook compact command will provide a more limited set of information as well. You can also try on your own codebook and compact.

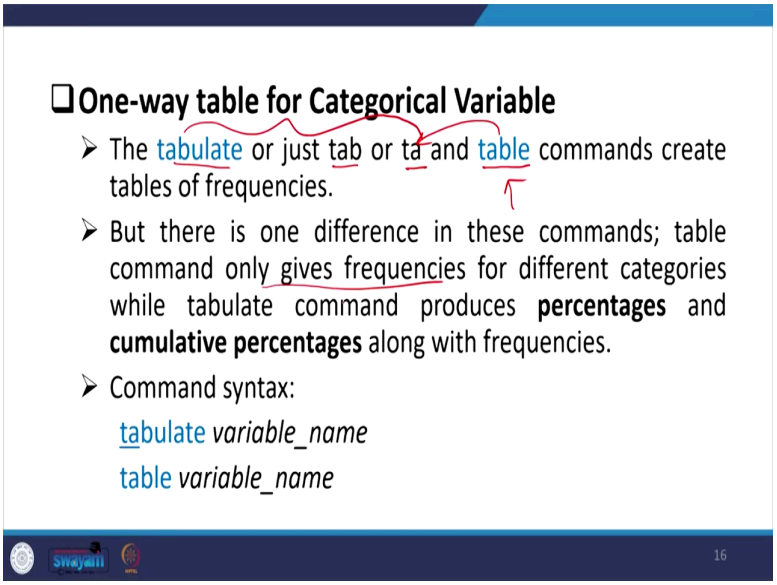
Compact is the word, codebook the variable name has to be code booked variable name or if you simply enter compact, it gives entire variable name.

(Refer Slide Time: 24:32)



Then tables and tabulations in Stata is one of the most useful ways to look at a summary data. Stata offers a variety of ways to tabulate data. Table and tabulate are such commands that helps in producing tables. The most basic table will show the variables and frequencies with each category.

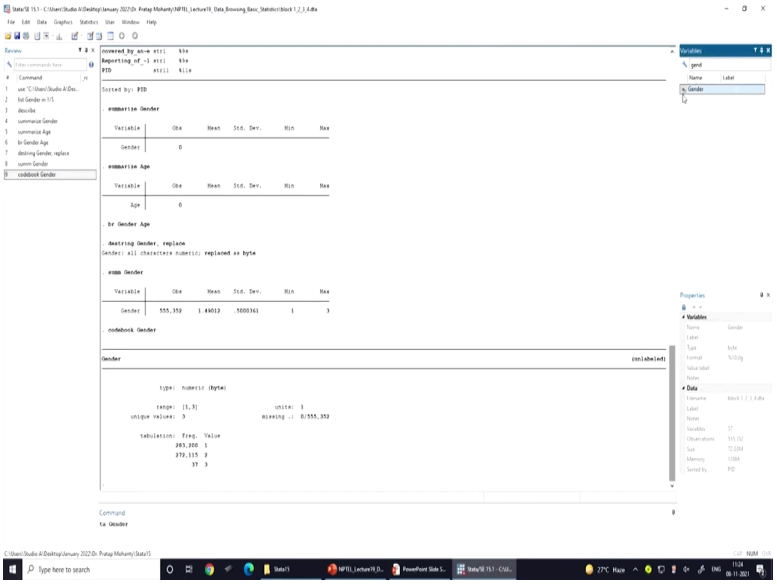
(Refer Slide Time: 25:02)



Like one way table if you want to just get one way table it is in their frequencies of one variable not two variables simultaneously. So, in that case the tabulate or just the tab or ta, these are all commands you can enter. These commands create tables of frequencies. But there is difference between table and ta or tabulate (these are all same command). This is going to give little different information. Tabulate gives some more some different information that only gives frequencies not with the percentages. There is one difference in these two these commands like table command only gives frequencies. This is what I have already just mentioned for different categories; whereas, tabulate gives percentage and their cumulative percentages as well along with the frequencies.

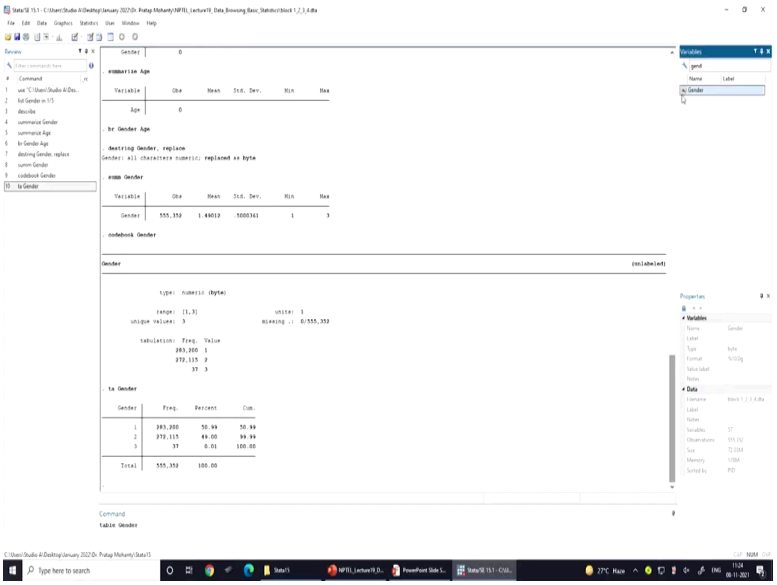
The command is for you is like tabulate or table we will experiment we will show you both on the screen.

(Refer Slide Time: 26:28)



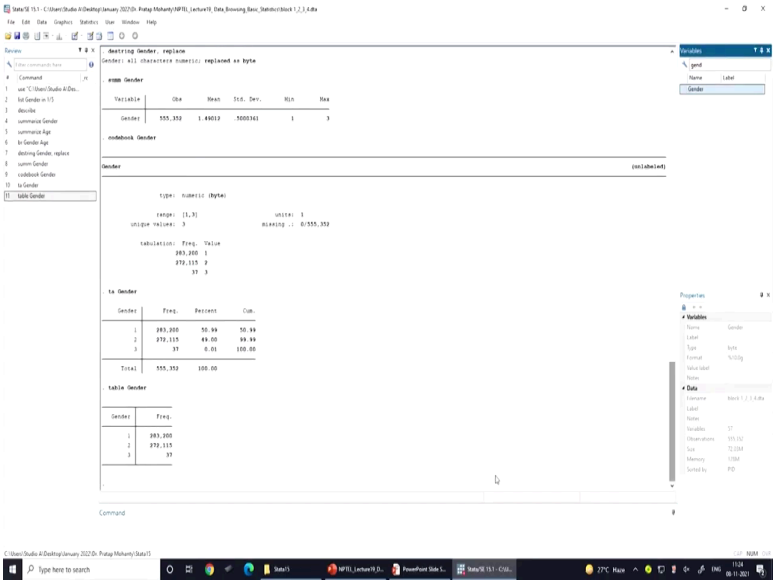
So, now, I will type first tab or ta; then any variable let it be gender.

(Refer Slide Time: 26:43)



So, now, this gives information about its frequency's percent and cumulative percentage. Now, if I type table, this only gives frequencies (table gender).

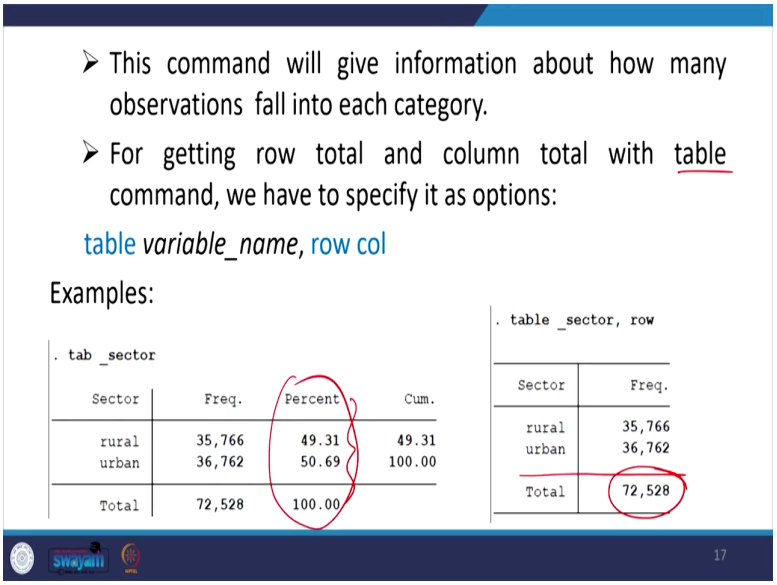
(Refer Slide Time: 27:01)



So, now you have got the information about the frequencies above of the same variable. But so, the tabulate is more important. Now, if you want to list so many variables together on the screen and then table command is more important.

If you are applying tabulate with so many variables more than two variables, then there it gets complicated, and it may not give you the right result. That is basically called cross tabulation we are going to show it that is basically two-way table, we will explain it here. So, this is done.

(Refer Slide Time: 27:44)

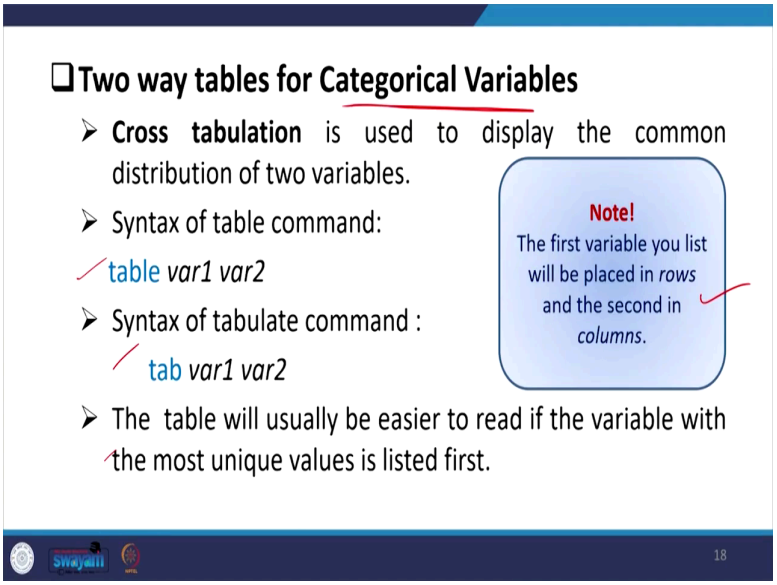


This command will give information about how many observations fall into each category. I think that is fine another one like if you are including row and column tab table variable name row and column for getting row total and column total with table command, not tabulate command.

So, the table command, row total and the column total is also going to give you like this this is your row total, this is your column total. So, if you have more variables then it is going to give you the row total as well. That is all about your table command, but we are more concerned about tabulate because it not only gives frequencies it also gives the percentage respective percentage and then cumulative frequencies.

Percentages are in fact, more important for interpretation, for writing research papers.

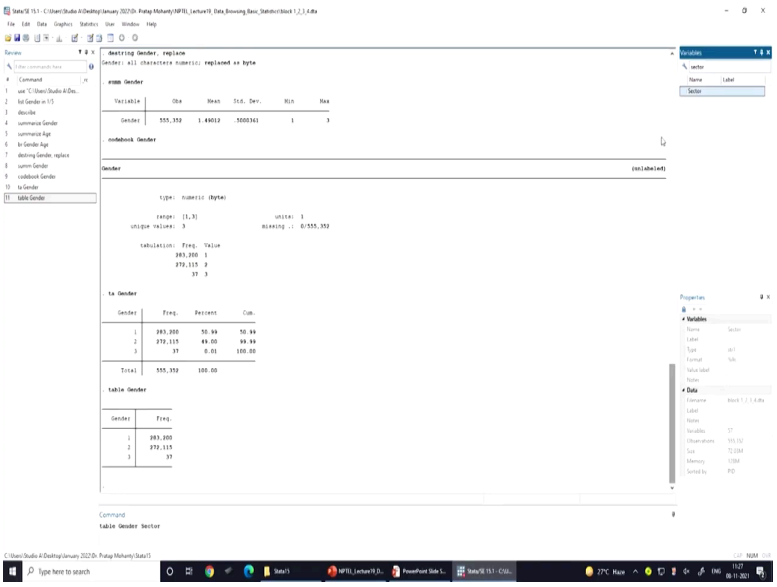
(Refer Slide Time: 28:56)



So, like I will experiment here with two-way table, for categorical variables. We have explained already what do you mean by categorical variables?

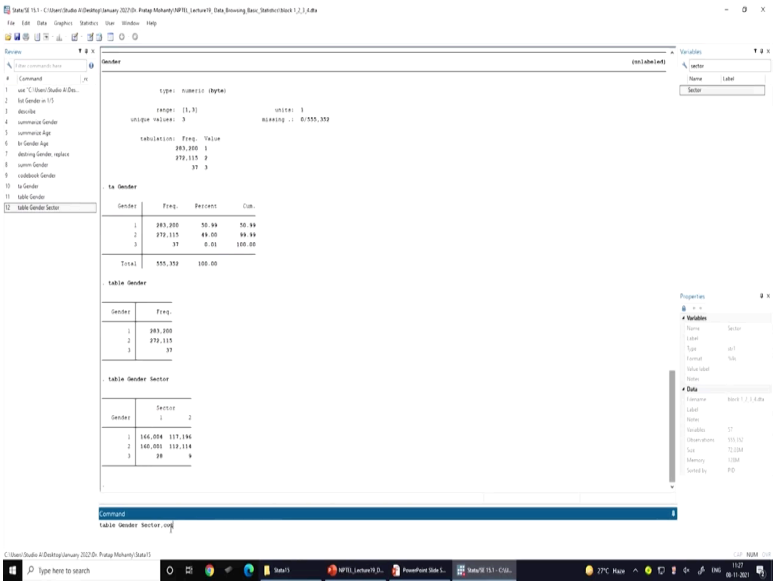
Now, the tab command and tabulate command; we will operate both here and we will see how it works.

(Refer Slide Time: 29:15)



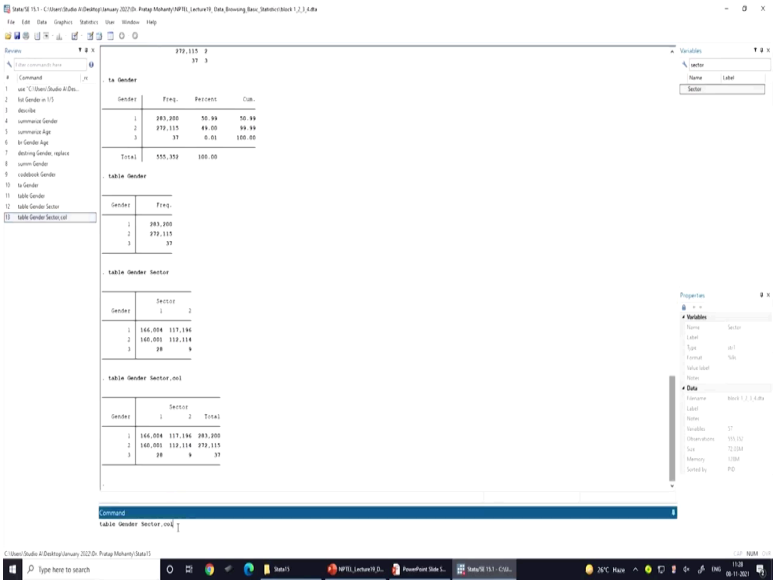
So, here two variables we will take first table: table two variables gender and sector.

(Refer Slide Time: 29:33)



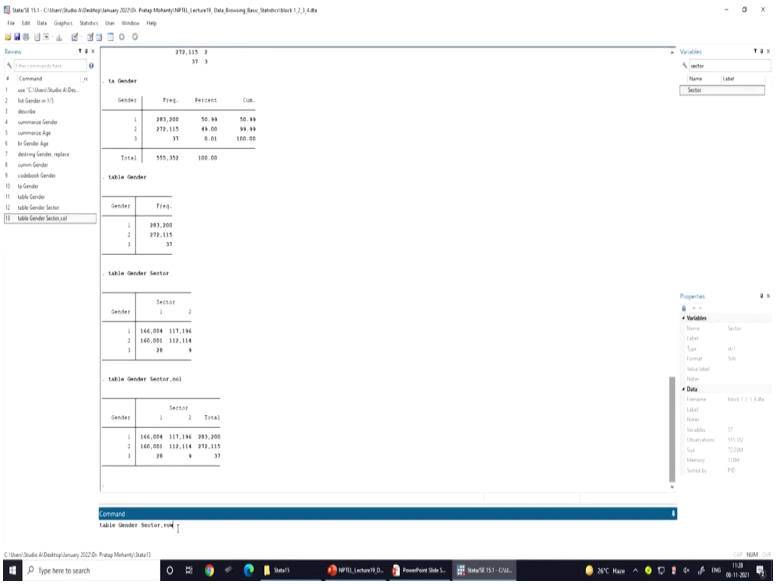
Now, this is going to give you their frequencies by absolute number in each category like one gender 1 and belong to one that is sector may be rural area that is 166004. Like on the same command; if I just go back again with the same command and comma columns.

(Refer Slide Time: 30:09)



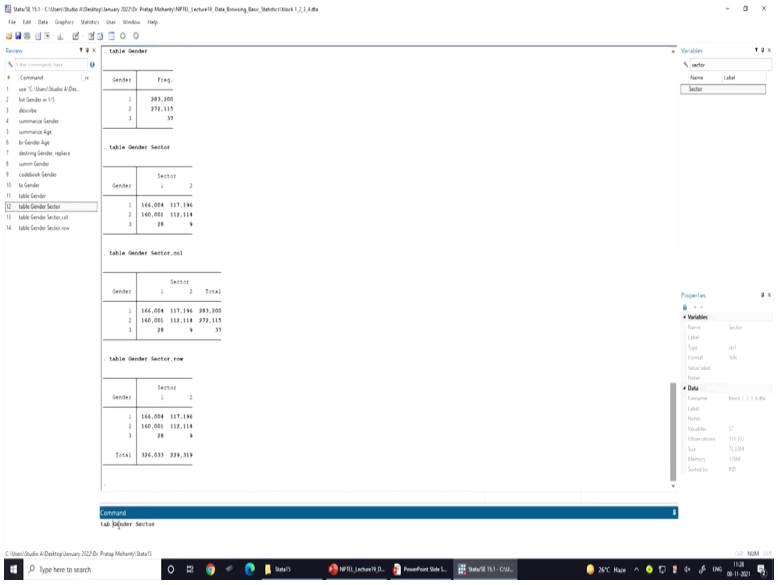
If I just add comma and column i.e., col then it gives you the column total. You can just have a look; this is going to give you the total by this way. So, this way it gets added. If I just add a row, then all the row will be added. Basically, here column wise added on the column total.

(Refer Slide Time: 30:41)



But, here if you are giving the command with row; so, all the row will be added.

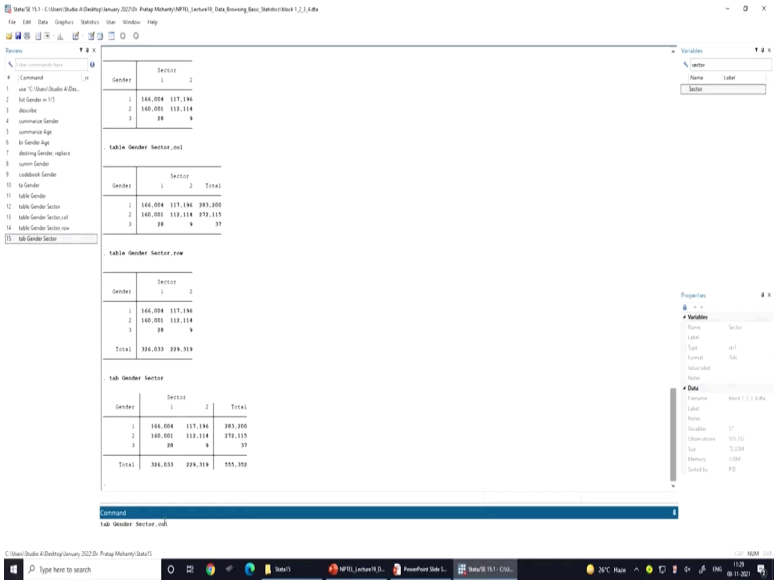
(Refer Slide Time: 30:45)



Yes, this row is added and that is why it is called row total. This is called row total because first row and second row one that is 2; 166000, 160000 something and 28 is added to 2; 326000.

So, that is all about your table command. Now, we are just changing the command, instead of table we are just changing it to tab; tab that is that is tabulate.

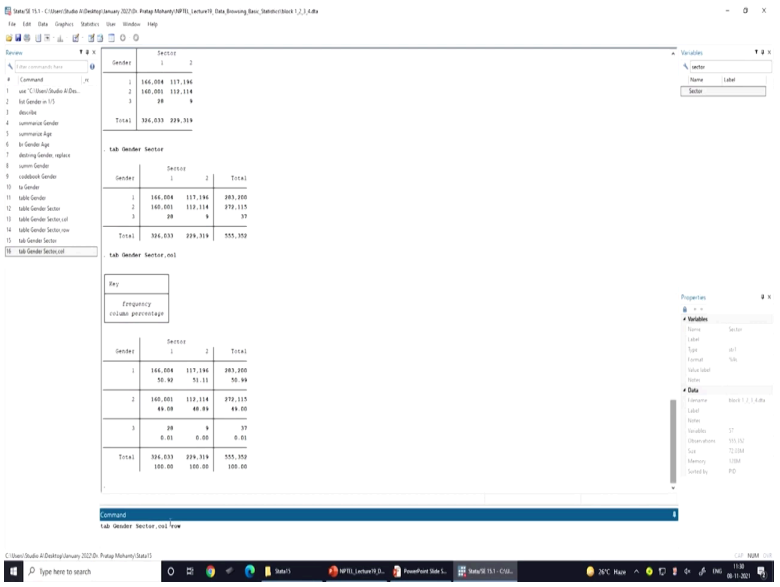
(Refer Slide Time: 31:24)



So, now, with this command we get here almost similar to the table command by its frequencies since we have not added their percentages requirement. Here, one additional aspect displayed that is it is row total and the column total.

So, row total is by default with this command derived. Now, if you are interested in getting percentages by column, then we can add extra option, on the same command, we will be adding comma column (col).

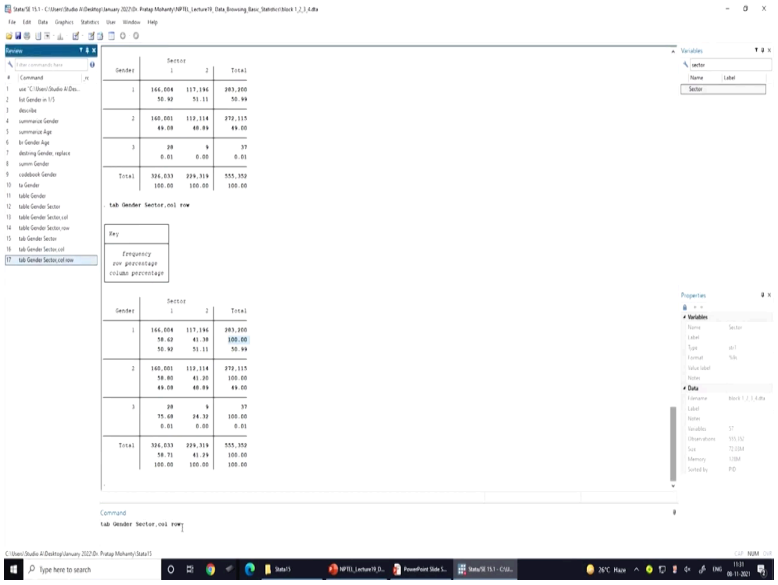
(Refer Slide Time: 32:12)



So, now, this has given us percentages by their column. So, out of these 100 we have got 50 percent for 1 and 1 is gender and 49.08 for the 2nd gender. Similarly, for the 3rd gender as well.

Now, if we add both column and row it will also give us both result column, as well as row and the absolute frequencies as well.

(Refer Slide Time: 32:48)

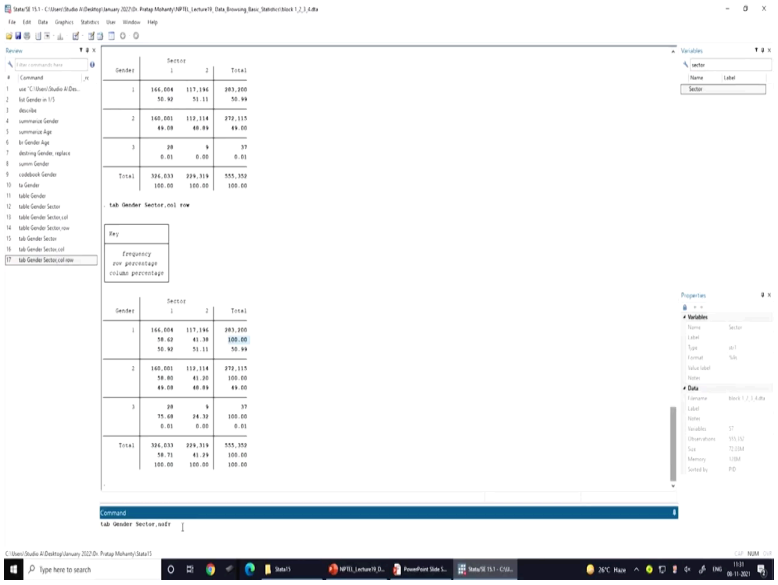


So, the row and column wise total both are displayed. This 100, this is by row percentage, this is the interpretation like; if I explain by the row total, what is the answer here that 58.62 percents or you are about 59 percent population belong to male; one is the code that may be male or the rural area in the rural area if sector 1 is rural.

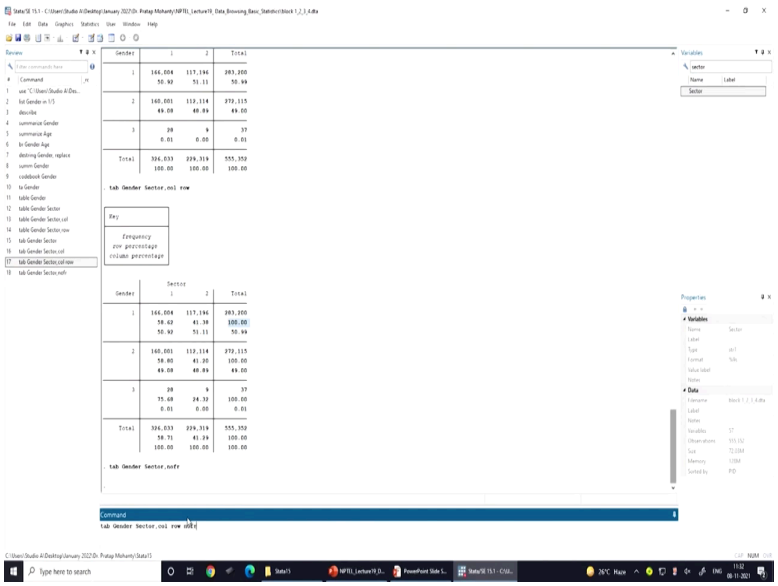
So, now on the reverse, if I interpret the column total what I say out of the total persons in rural areas out of the total persons respondents in rural areas 50.92 percent belong to male or gender. So, accordingly you can interpret.

Now, what I will do; it seems as if more complicated we can avoid these things and only consider the frequencies. So, nofr command i.e., no frequencies basically we will avoid frequency and only we want the percentages.

(Refer Slide Time: 34:08)

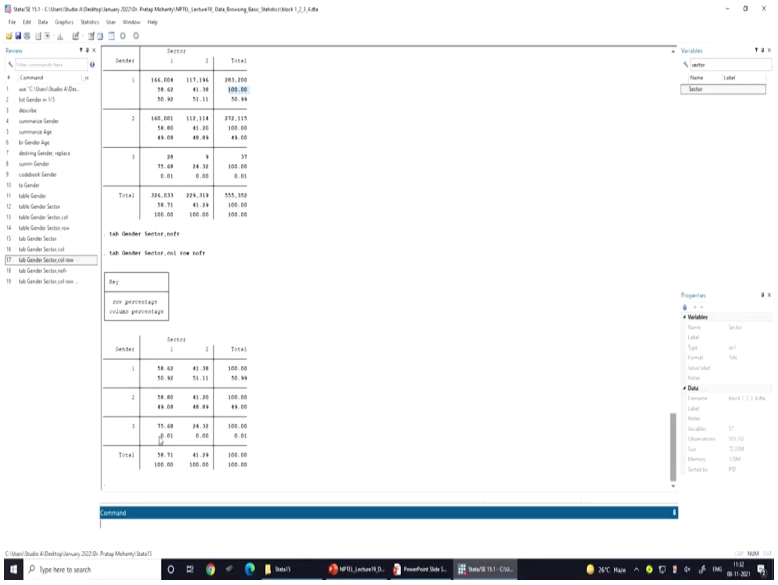


(Refer Slide Time: 34:11)



So, we will only type the nofr. So, nofr is going to give us only percentages not the frequencies. There must be a space that has taken the command already.

(Refer Slide Time: 34:36)



So, first we need to give both one right column and row; it does not read whether you want by column or row. It has to be like if you are not attaching column or row, Stata does not read that whether you want in frequency or not in frequencies.

So, first of all we have to keep the command as column and row that that indicates your you want in percentages, again we defined as no frequencies; that means, all the frequencies are avoided only the percentages displayed. Another, one we can go for cell percentages as well. Here I will just guide you without wasting time.

One note is there the first variable you list will be placed in row; second one will be on the column for that is most important. First variable which you have entered in your command is going to be placed on the row and the second one will be on the column, and this is important to note.

So, I think you must have understood the concept very well. We have tried our best to give you very basic directions as if a newcomer is going to understand everything. And, if you have any difficulties, please do not hesitate and come back to us.

With this thank you very much.