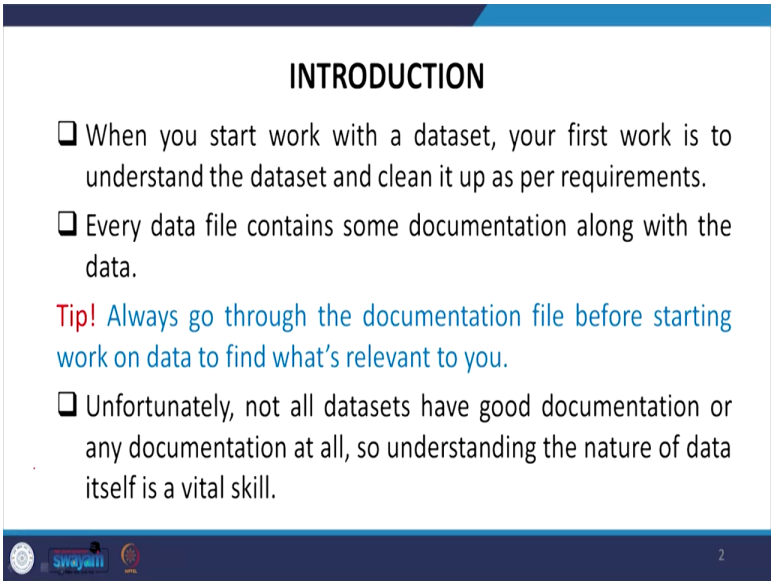
Handling Large-Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences,
Indian Institute of Technology, Roorkee
Lecture 13
Managing Data in Stata - I

Welcome once again dear learners to the NPTEL module on Handling Large-Scale Data Using Stata, and we are explaining how to manage data with the Stata. So far in last two lectures we have been trying our best to introduce you with Stata and there are many pointer options or command options we discussed, but now onwards it will be more practical oriented. Though earlier classes were also practical oriented, but the appetite of understanding Stata was through data. So we introduced data earlier. Now onwards we are using data with the help of Stata software. So, deliberately the title is defined as Managing Data in Stata.

So, once again, for your understanding, I am going to use the Stata 15 software and I am again suggesting you to operate yourself through the demo version or if you have a purchased version or licensed version, please open that and start working and go through our instructions and also interpretations.

(Refer Slide Time: 01:53)

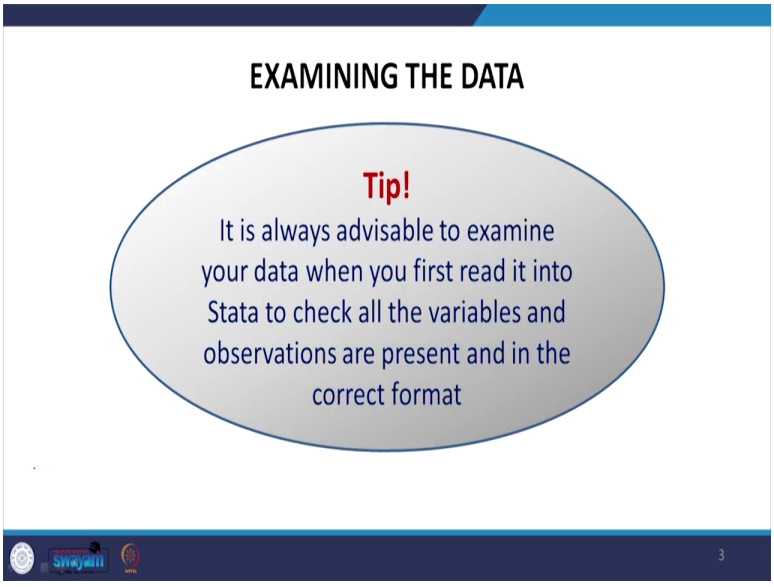


When you start your work with the dataset, first of all, it is always good to understand the dataset and try to clean up as much as possible for smooth functioning and smooth interpretation of the

results. So each data file contains some documentation, because without any structured presentation of the data, which is provided by the surveyor or the organization which make data, if they do not make it in a structure, it is very difficult to understand. So to make it, put it in a structure, they sometimes compress, they go by systematic commands and so that the data is in structured work for us that just cannot be extracted and used. You have to mine the data or simplify the data a bit and start explaining it. So here we are to help you how to simplify the data and make it in a structure.

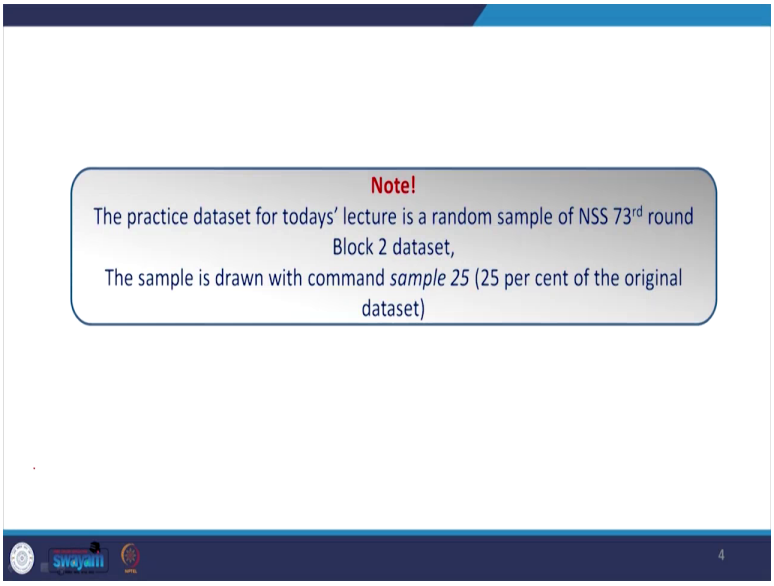
So the tip here is given that always go through the documentation file, every data. We have already shown those documentations with NSS, National Sample Survey, IHDS and National Family Health Survey data. Please check our previous lectures, you will be guided with the documentation very clearly. And before starting to work, it is once again suggested you to please check the documentation. We will also guide you side by side. Unfortunately, not all datasets have good documentation or any documentation at all, so understanding the nature of data itself is a vital skill for all of us.

(Refer Slide Time: 03:48)



As highlighted it is always advisable to examine your data when you first read it into Stata to check all the variables and observations are present and in the correct format that is very essential.

(Refer Slide Time: 04:05)

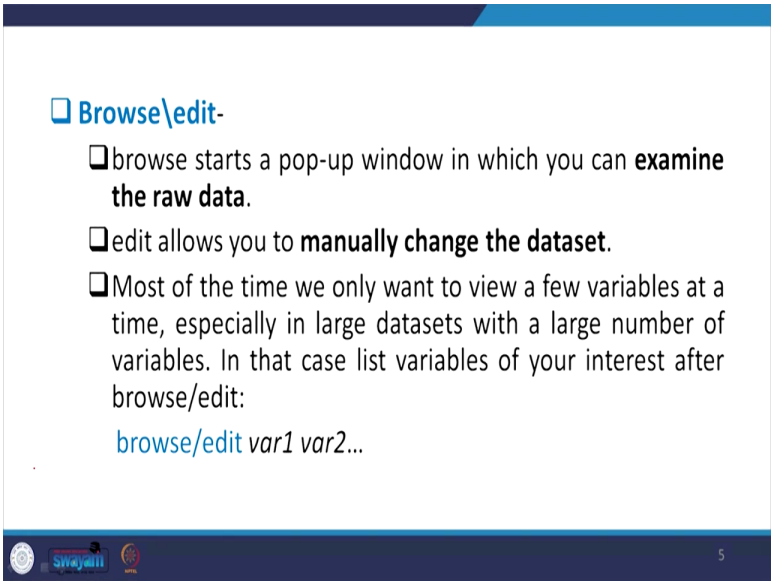


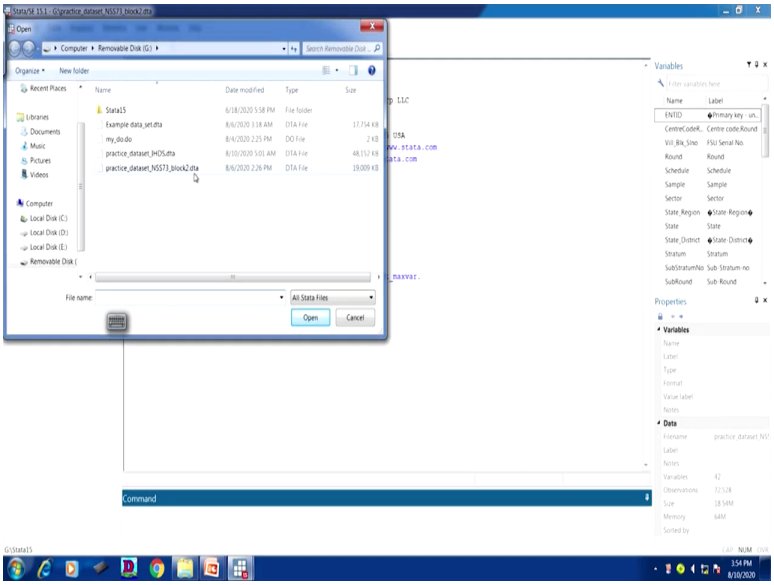
The practice dataset for today's lecture we are going to use is National Sample Survey 73rd round that was published in 2015-16 and we are focusing on block number two. There is a purpose behind using this. Our title of the module is handling large-scale unit level data with Stata. Since the title is large-scale, we are trying to equip you all with the exact and core dataset which are available for social science discipline as well as for management discipline largely. Others can also get the advantage out of it.

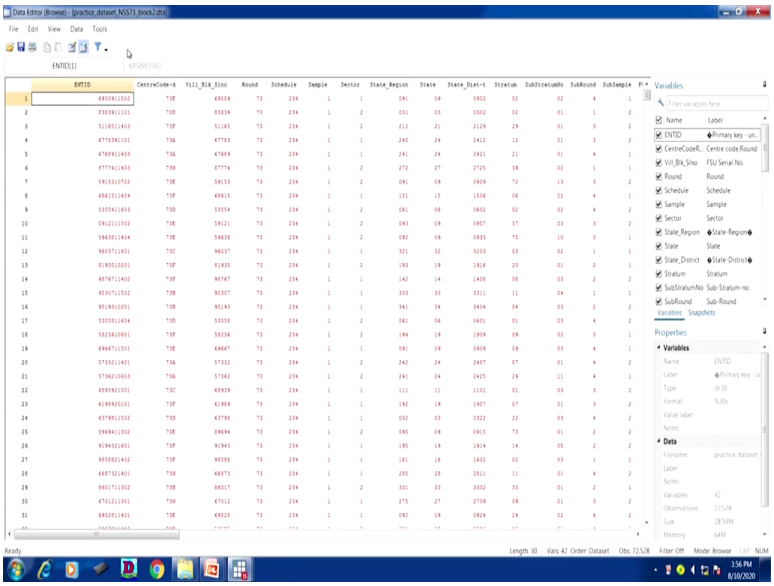
What I am suggesting you, we will also provide whatever we are using, we also provide these database to you. These are going to be your added materials. NSS 73rd round block number two. This data is also available as I already guided you in earlier lectures, this is already available but all the blocks are available. If you go by all the blocks, the data will be very large enough and it consumes huge byte space, the memory is consumed. So when you operate, it takes time. So, we will guide you side by side how to filter and simplify.

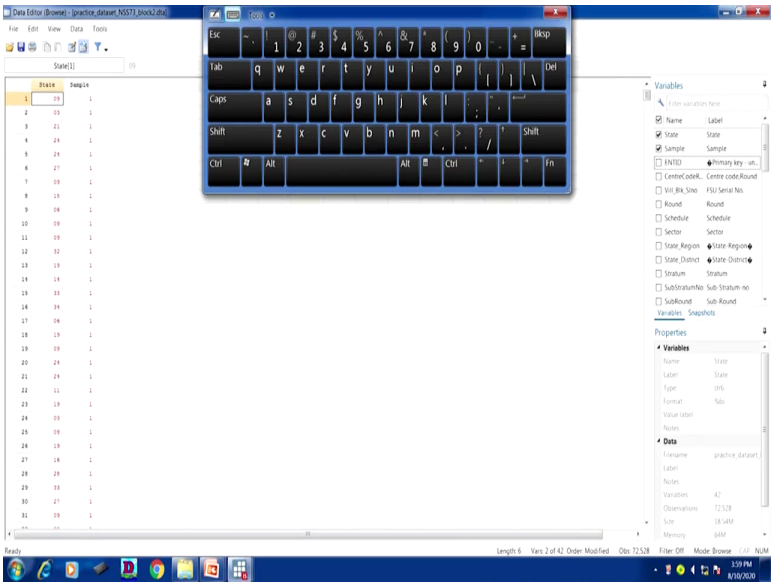
So, for your simplicity, we have deliberately drawn a sample of 25% and the command is given. You can experiment on your own. So from the original dataset we have reduced to 25% of the total sample for easy operation and understanding of the data.

(Refer Slide Time: 06:00)









Let me proceed to start working with Stata. Let me open the dataset first. We will also provide you the data, the 25% of the sample data to you along with this PPT and since you have already downloaded the Stata version. I assume that you have downloaded so far. The demo version you have got so far or you have purchased.

So, please keep three things together. First is one sample data. Otherwise, if you do not have sample data, it is becoming difficult for you. I guided you in the last class about Stata directory and we used life expectancy data. You can also operate from there as well. This is a very simple

small dataset and very quickly you can understand applying the command we are going to suggest.

Let me tell you what you are supposed to collect at this moment. A Stata version must be there. Maybe 13 onwards version, but at this moment even older version is going to be useful except few commands, like I told you already help command, search command, find it command was earlier used. But still most of the operation you can do with the earlier version. If those are freely available by any reason, you can also procure that. You can write it to StataCorp and they will guide you.

So here is the window for you to start working with the data I have opened. Once again let me tell you, I will just guide you here since I have already opened here. So you go to the file and file menu, then you open. Once you open, I have opened here the practice_dataset_NSS 73_block2.dta. I have already guided you .dta file must be there. Otherwise, it might be problematic. Some data can be also imported with different versions, like excel or maybe other whichever are supported, you can also open that. But at this moment I am working with .dta. So I have this.

I am also going to use some other aspect to it. Later on, I will be discussing about IHDS data, practice dataset IHDS. I will tell you later, but at this moment I am using NSS 73rd. So these, after opening that all the variables list are visible at the right hand side under variable heading and their respective properties are also getting visible to all of you.

Here is the data review file as I told you, if you click on tab, it keeps on rotating with the availability of review, then variables, then properties, then command, then result. Result is not visible through tab, but others are visible. So let me proceed. Start with understanding browse or edit, as I already told you. Browse starts a pop-up window in which you can examine the raw data. Edit allows you to manually change the dataset.

So, in the last lecture, I specifically mentioned that if you just want to check the data and their nature, try to go by the browse not by the edit. So if you want to change the data, then you click on edit or you type edit then the variable or you wanted to change the particular value within the variable, then you go by edit that is separate, but always it is suggested to open the browse data unless it is necessary.

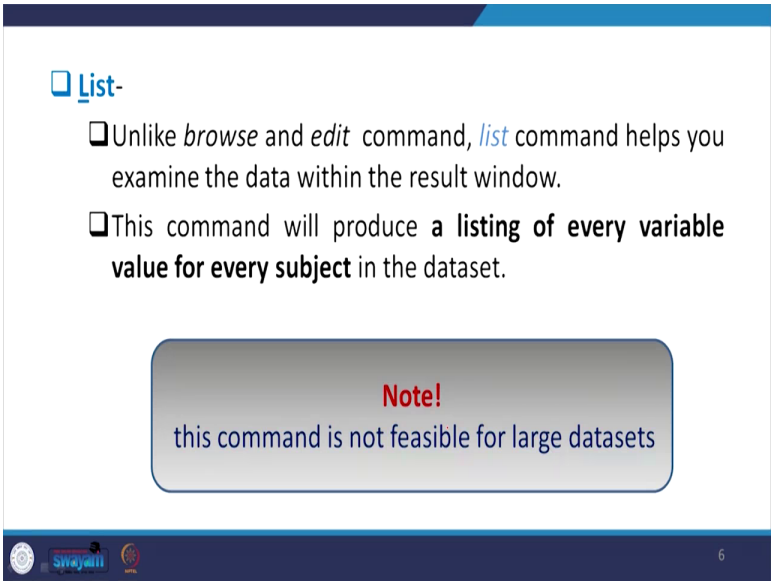
So most of the time, we only want to view a few variables at a time. All the variables like let me start with. I will operate now. There are two options, simply you can click. At the tool bar menu, there are two. First one is I can highlight here data edit, then data browse, that is one option. If I just click on browse it will show you the data. The data fitted in .dta format and it is now visible to us.

Instead of open that easiest way you can simply type br and enter, it will also direct you to the same window. Else if I just click on edit, it will also give you the same data. But there is a difference, here you can change the data, but in br command, browse command, you are not going to change the data. You can only see the data. So, as I told you already that seeing the data is different than that of editing the data. It is safe to see the data. When you are doubly sure that you are going to change the data then open the edit command.

So, let me move on to our next operation. As I told you, usually when you open the data there are so many variables. You look at this. I have already told you that just by br command I have opened the dataset before you. There are so many variables listed. You can just scroll to right you will find so many variables.

But so many variables you may not be interested to look at a given time. You may simply type br, any two variable name or three variable name or even one variable name. Whatever you are interested to carefully check, just click on or type that particular variable. I am just putting a double click here. I am there to switch the enter button. So it has given me only two variables information in the browse. Similarly, in the edit command as well you can go through.

(Refer Slide Time: 12:52)

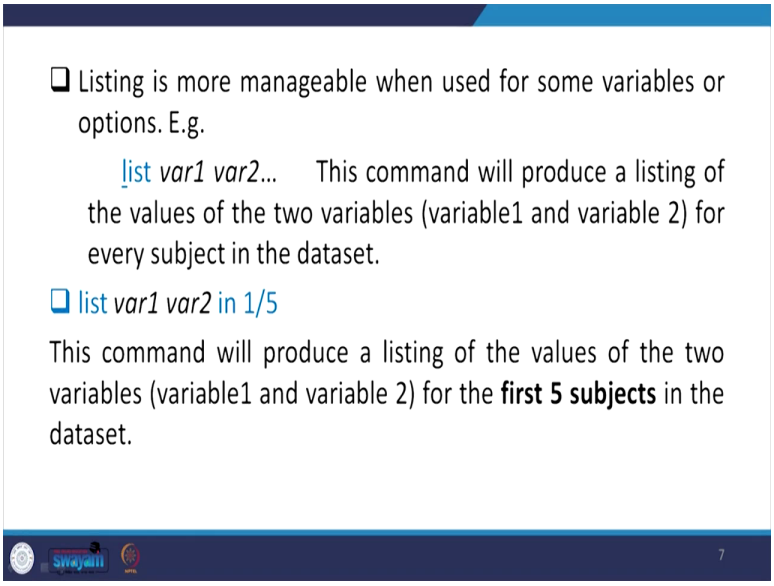


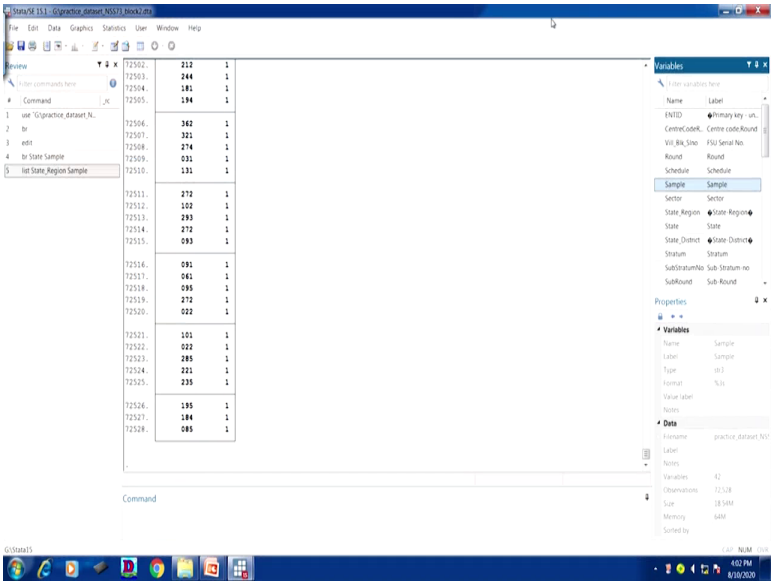
Let me proceed further to other details. So here is the detail. next important aspect for our understanding is list, listing unlike browse and edit command, list command helps you to examine the data within the result window. Like help window and chelp window we told you earlier one is in another new window will open up and in the c command, chelp, it gives you the command within that same window.

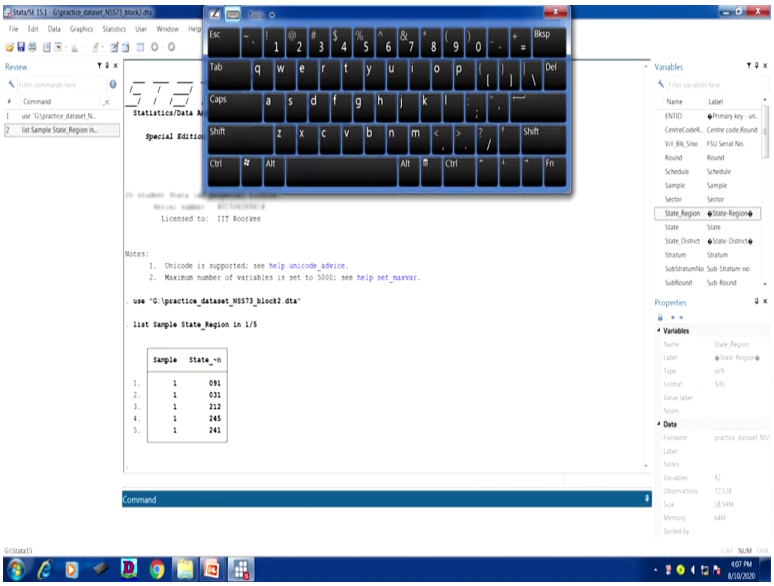
So similarly list command helps you to examine the data within the result window. Like when I browse or edit another window popped up. But if I am interested to look at some variables and a specific aspect of it, I can go by list, because it will be visible on the same window, result window. Result window we mean here only on your screen. So, I will operate and let all this.

So this command will produce a listing of every variable, value for every subject in the dataset. This command is not visible for large datasets. Why? Because so many information will be visible on the screen and all your screen space is blocked and it will be very difficult to scroll down till the end and it will be very uneasy. That does not mean it will not be visible. It will be very uneasy to check. I am not going to do that all together. If I do it all will be opening on the same screen it will be suggested not to open that if you want to take entire, better to open in the new window. So, list is important.

(Refer Slide Time: 14:47)







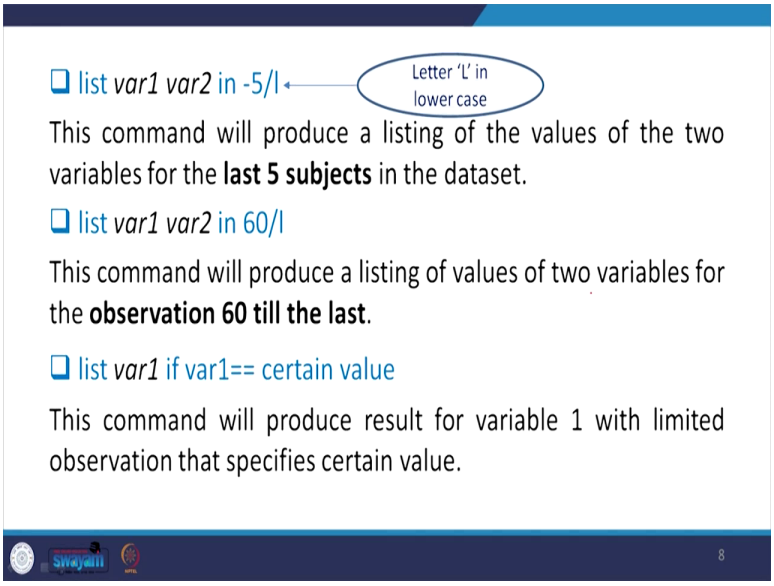
We can do the listing more manageable when used for some variables or options. As I just told you a couple of minutes back, options are important. List variable one and two, you will get the listing of the values of two variables for every subject in the dataset like this. Suppose I do here for you. So I just typed list variable one here, then variable two let it be here, I just enter. Now look at this, these are visible on the same screen. But I am interested in so many variables I wanted to check I will suggest you to go by the browse.

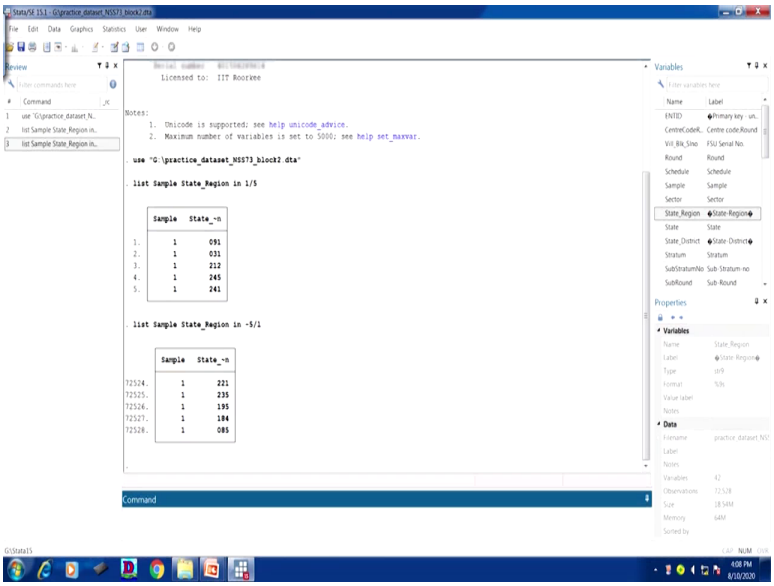
List variable, this is going to be more interesting. , if there are some lakhs of observations, usually the dataset we are handling has more than 2 lakhs or even 6 lakhs observations are there. So, 6 lakhs observations and you are going to show by the variable on the screen is also going to be cumbersome. So what do we do, if some clarity are required particularly from five observation or maybe last five observation or maybe some portion of the observation, we have some separate command.

Here this command will enable you to check a listing of the values of two variables for the first five subjects. Two variables were put here as variable one and variable two, so two variables are there for you. For the first five subjects in the dataset with the command in 1/5, list variable in 1/5. Let me operate it for you. So what we have been dealt so far is understanding some portion of our data, let it be 1/5. So I am deliberately trying to open the Stata once again to show you how 1/5 is being displayed in the command. How it is important, let me clarify before you. Yes, it is here.

So, how to do it? So I will discuss list then two variables, then in 1/5. So let me click here by list, then as I told you two variables, just double click on two variables, this and this let it be two variables, then I am using in 1/5. So, let us enter. So it has given us the details of 1/5 of the observations.

(Refer Slide Time: 18:14)





Let me understand all the details from our explanation. Similarly, if you want to list the values of the two variables for the last five subjects in the datasets, last five subjects instead of 1/5 here last means l, let me help you, l is important. So it is even highlighted letter l in lowercase, not in

capital, because Stata is very sensitive with the exact command. If you are deviating from it, there may be some errors.

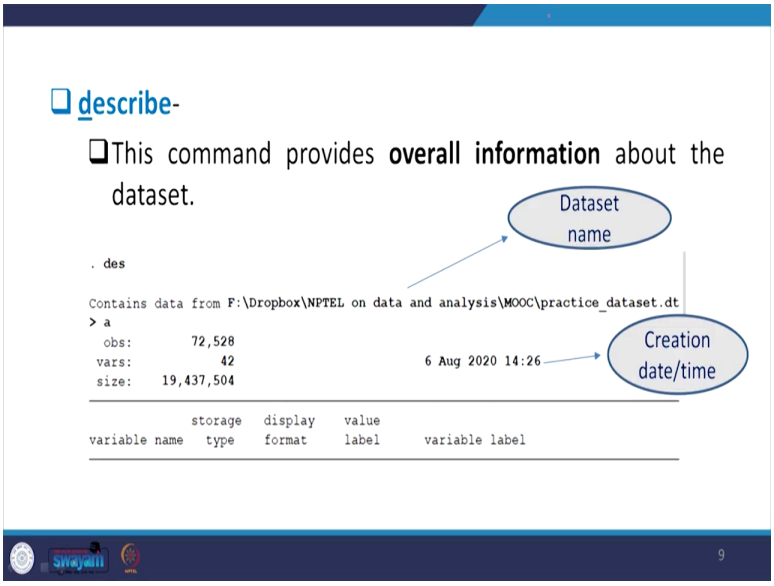
So if I just show you once again with the command like what I try to do, I will go by the previous command. If you use page up, you will get that command but here also I can show you page up. So page up command is now visible to me. For you just click on the page up key in your keyboard you will get that. Now I am going to delete this 5, I will just show you minus 5 slash small l, now enter.

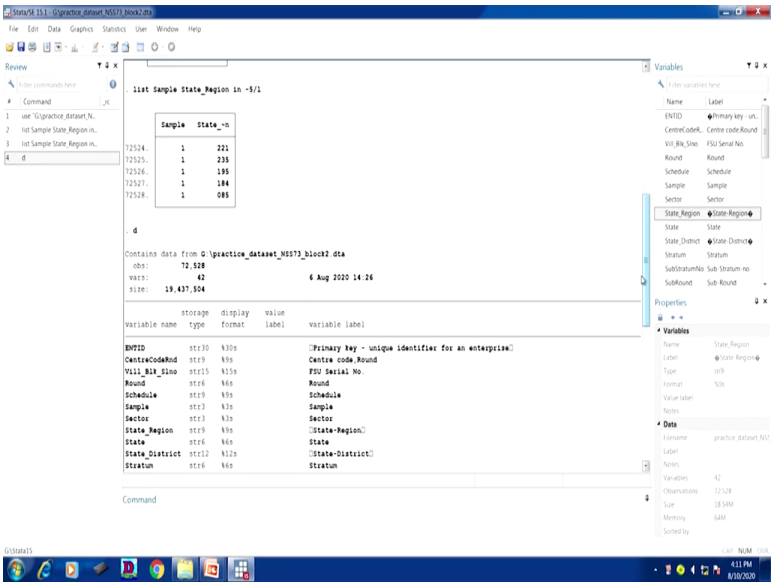
It has given you the last five information. So it is very important, though it is visible as not so important, but you have to be very handy in your operation. You will certainly use it. Those who wish to have a mastery on Stata, these commands are very very important for better explanation.

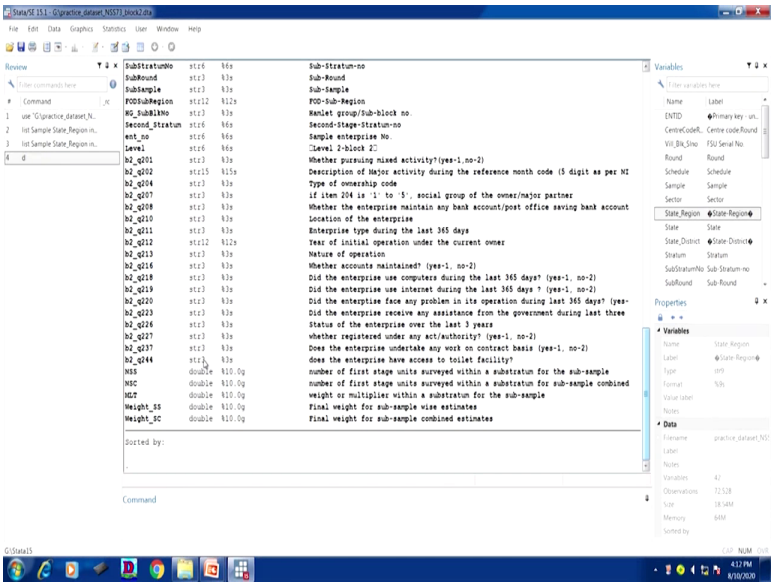
Similarly, list if you want observation 60 till the last, then 60 slash l is going to be important. small l, l in lowercase we are suggesting you to enter. Similarly, list variable one if variable one double equal to certain value. If you want to get certain values and how variable one looks like with certain values is very important. This command will produce result for variable one with limited observation that specifies certain values, any value as a center and you will certainly get that.

But if command is essential, because you are conditioning to a particular value. So I am not going to explain you. If you are stuck, we will clarify in the doubt class.

(Refer Slide Time: 21:00)







Other details are there, like describe. You might have heard in many articles, many journal papers, description is very important. Though variable description is a different one, but some information we can derive from the describe table, describe table is always very important. The command is highlighted in blue color. For you, I am going to show you how describe is going to clarify you.

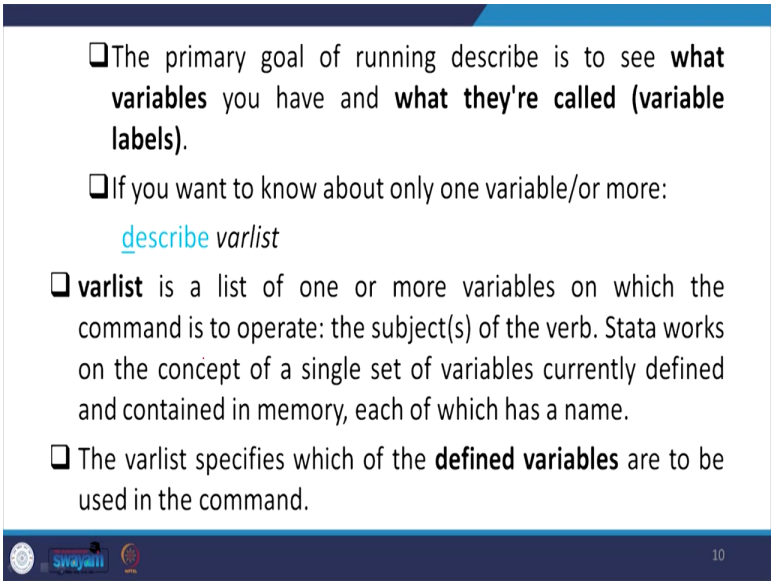
So, let me type it here, simply d. D is also going to give you the result. underscore of d was there. So you just click on, you will get the data or ds or describe whatever you type you will get the information for sure. Look at how clearly you get the information of your data which we are trying to work. Look at this, the data is visible before you generated on August 6, 2020 at 2:26 or 14:26 hours. So, it is also showing that practice dataset NSS 73rd which I have already told you.

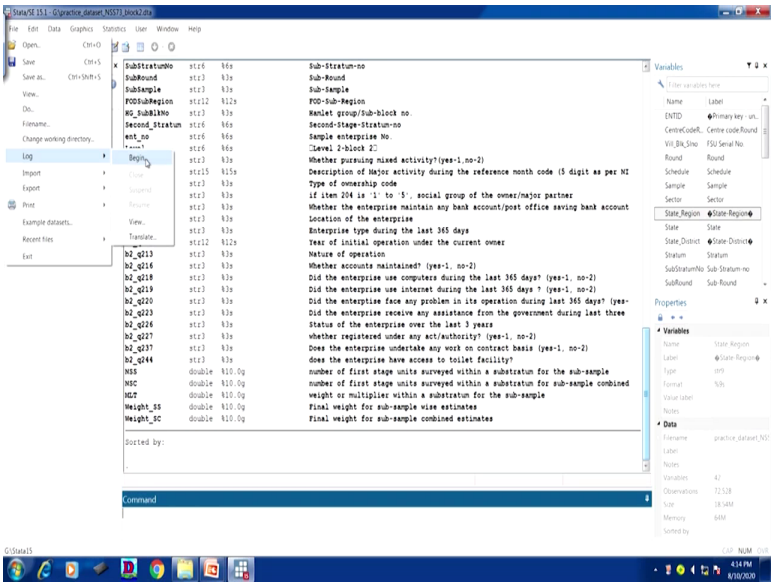
It also gives information like storage types, whether a string or in double entry or in numeric all those details are also going to be there. So, the second column is explaining that. So variable name on the first column, but variable label is very very important for a researcher to understand the data, labeling of the variables, which label, which name is there and you can also rename those labeling. I will guide you in between.

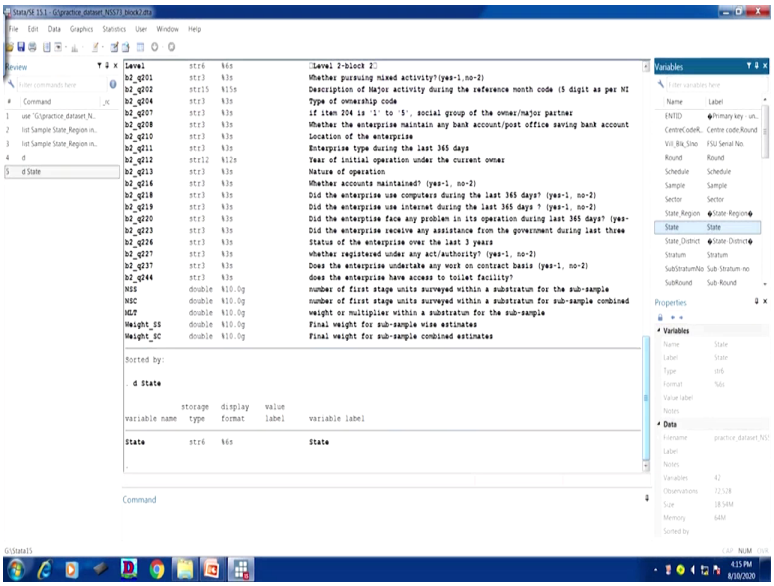
Also look at other entries here, like all the variables information are given, double, as I told you, string 3, I told you one string str1 till 2,045 entries are there for string values, string l. Sometimes the long string, large string is coded as strL. So, other details we will guide you in between, let

me proceed. Even the dataset name as I already told you with their location, its directory or the file name is clearly highlighted here.

(Refer Slide Time: 23:49)







The primary goal of running describe before you is to find what variables you have and what they are called, variable labels as I already guided you. If you want to know about only one variable or more to describe variable list, like in the screen which we have just shown you describes entire variable.

You might forget one thing, last class I started guiding you, when you operate any kind of trial and error, even any kind of operation or management with Stata, I always suggest that you start with a log file. I will guide you in between do we mean by log file. Last class I started, but I did not guide you in detail.

Here is the option I am just pointing out it will show you log begin, close. Otherwise, file then log then you begin. If you do that, and there are two options, I will guide you all the details are going to be dealt in this lecture only, like log files comes with smcl or in text format. Text format is suggested because we generally take those result to word format or excel format for further operations, for further manipulations. So I will guide you all detail, but I will suggest you just click on this and save with your filename. Rest of the details I will guide you in due course of time.

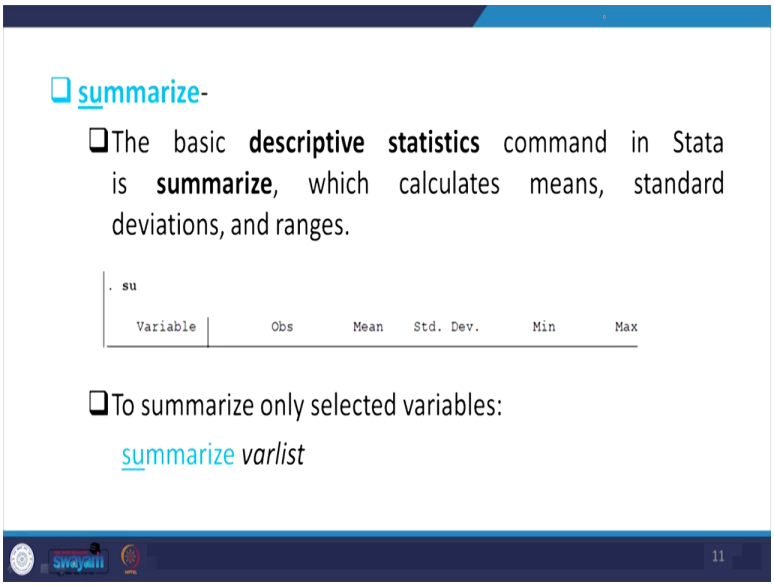
Variable list, as I told you, variable list or the describe variable, as I already told you, start with a d, enter with a d command or des command or even describe command it will give you the result. If you are interested to go for only one variable, you simply click on that variable, like click here, yes I will again type d or you just go by your page up command, you will get that

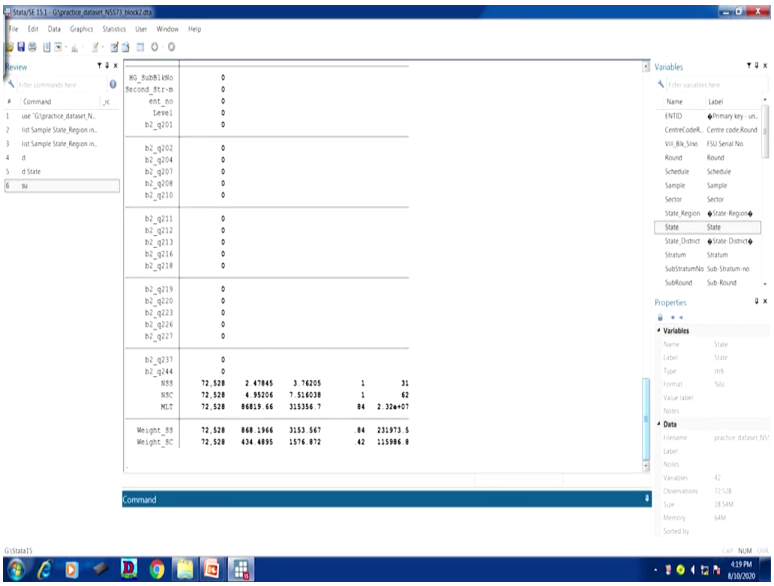
same command, but now I am only using one variable. One variable, let me use another variable called state, now enter. It gives the information of one variable.

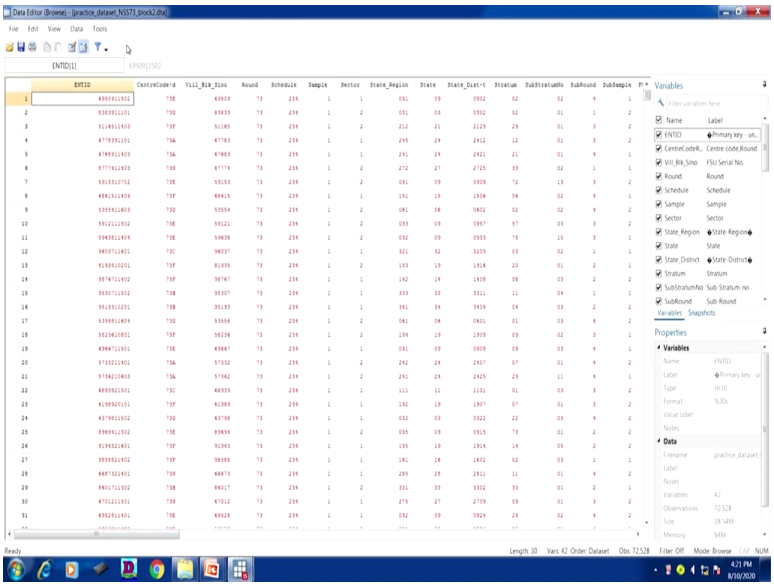
When you are quite sure you wanted to operate with a particular variable, so why to check so many. So many may not be going to be very important, one is important, you just click on that and find out. Otherwise, if less number of variables are there you can check at a go. So, variable list is a list of one or more variables on which the command is to operate, the subject of the verb.

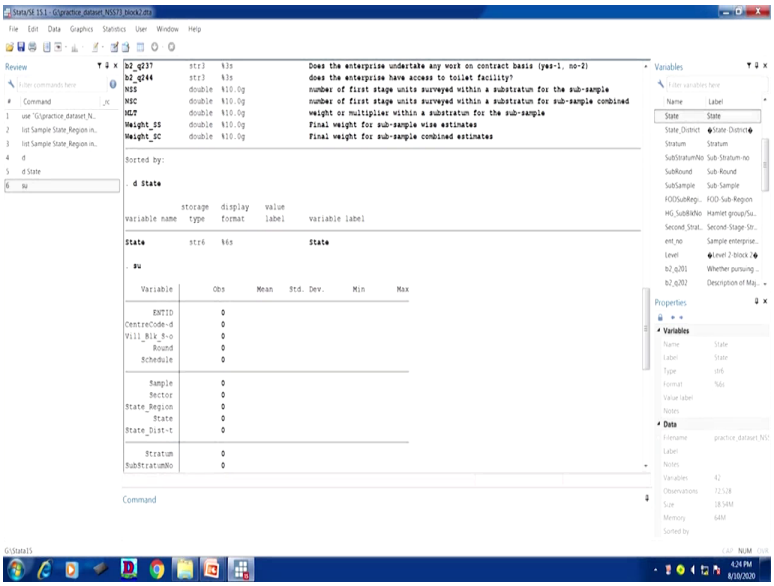
Stata works on the concept of a single set of variable currently defined and contained in memory, each of which has a name. We have already mentioned that basically this suggests that you click on the variable, you will get the list. If you want more variables, you click more variables then you will get the information. The varlist specifies which of the defined variable are to be used in the command.

(Refer Slide Time: 27:33)




Stata/SE 15.1 - G:\practice_dataset_NSS73_block2.dta
File Edit Data Graphics Statistics User Window Help
Review
Command
Variables
Properties
Command
G:\Stata15


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



So, this is the window we have already shown. Similar to this kind of window, we have shown you, but this is a little different, because we are going to operate with the summarize. Earlier one was describe. This kind of window we have shown you. I am going to talk about summarize.

So, summarize gives you descriptive statistics and this is going to be very very useful for all the researchers while writing papers and communicating to journals, even for books or even for assignment. If you do not give a descriptive statistic of all your variables you are using for modeling, I think your modeling is completely redundant. How another author is going to understand your variable descriptions?

This carries with different name, variable descriptions, descriptive statistics. And usually this gives information like how many observations are there, the mean, standard deviation, minimum and maximum. So far as standard command that is summarize you are going to enter. This is called a standard command for summarize your descriptive statistics. There may be a detailing of the summary descriptions as well.

Sometimes some researcher in the paper write a summary. So in a table, in the heading they simply write summary. So summary also give you some kind of description as highlighted here summarize, you can write summarize or only su. Underscore su is mentioned or underlined with su, so with the variable list we will give you the detailed information about it.

Now, let me guide you first before going to the detailing of it. If I just type su or summarize, sum also, usually in my understanding many researchers usually enter sum. So let me exercise with su. If I simply start with su, it gives all the variables and their summary. How to read those, just have a look, how to read these? Look at this. As I told you, it gives information like observation, mean, standard deviation, mean maximum.

Another important thing is to be noted here, why observations boils down to zero because of string information. As I told you, wherever you have numeric values or labeled numeric values, you will get the result, because there are, then only we can able to enter some statistics on those data, some mathematical operations are made on the numeric data or even on the labeled numeric data, but if it is a string variable, it only represents characteristics, or some coding with a picturization of it or the time, roman characters are there, it is very difficult to do any mathematical operations.

So here the summary does not give information about the string variable, like observation is now showing zero. I will tell you when we go by the numeric value, it gives you observe, like numeric value here for NSS, NSC. As I told you already, here number of first stage, you need units surveyed within a stratum for sub-sample then number of first stage unit surveyed within a stratum for sub-sample combined. So these are also MLT weight or the multiplier within a stratum, then weight of those indicators these are a numeric entry.

Even we can, you can check with our data browse. So, as I told you, if it is in red color, these are clearly the string data. If you browse through other variables, I am going by the right scroll, not the bottom or vertical, I am going by the horizontal scroll. Look at the black fonts, as I already guided you in the last class that this indicates numeric values. Had it been blue color, these are called labeled numeric, still operations are possible with this command. So we will clarify further details.

Let me now go to the window once again. What I wanted to say here once again to you that if you are not interested to summarize all the variables, you are interested into the particular variable, then you can simply write down the variable name. But usually when you have already filter and you have used the particular variable for your model and these variables are important,

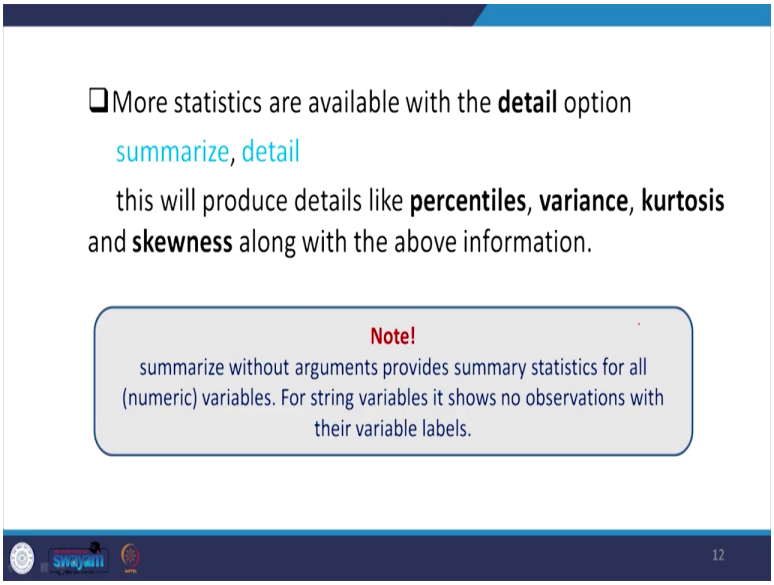
so better to include all the variables, which variables are important for you, like in your dataset so many variables are available.

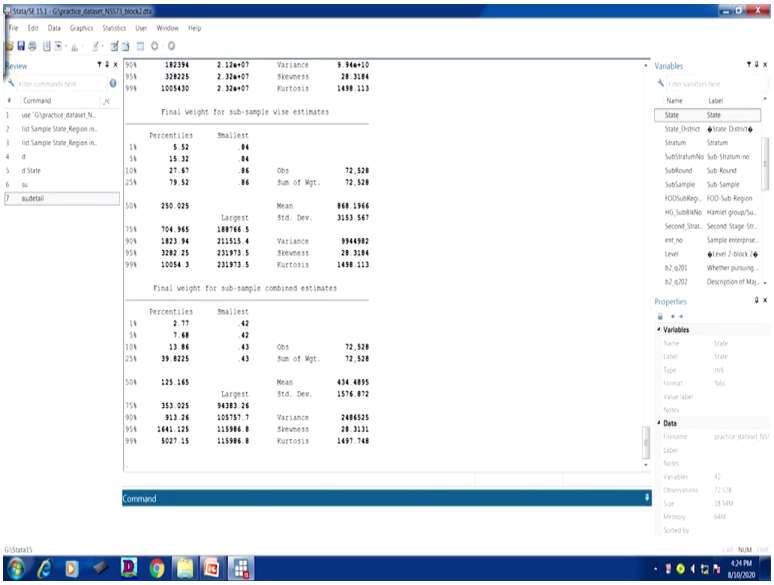
If you simply go by sum or su, it gives the variable descriptive statistics, description, not variable description, descriptive statistics of entire variable available in the data. But in your model, like in a regression model you have used only selected few. Why to include the descriptive statistics of entire data you are interested for only those lists, so enter only those selected variables. So that is suggested at this moment.

If you want further details, like some data are categorical in nature, I told you from here, is numerical, I told you, some data are categorical by string, string categorical. But if string categorical, then you cannot operate it. There are certain other approaches to do it. I will show you in a short while.

But usually the standard variable or the summary statistics is not going to give you the right result, like categorical data, our result only gives mean value. The mean value of categorical data is redundant. It is not right. Mean value of, for example, gender of a person, mean value of a gender of a person is going to give you absurd result. Like gender of a person are generally coded as 1, 2, 3, so average might be boils down to 1.39. So 1.39 indicates nothing. So you require some detailed command, detailed summarize, summary statistics.

(Refer Slide Time: 34:49)





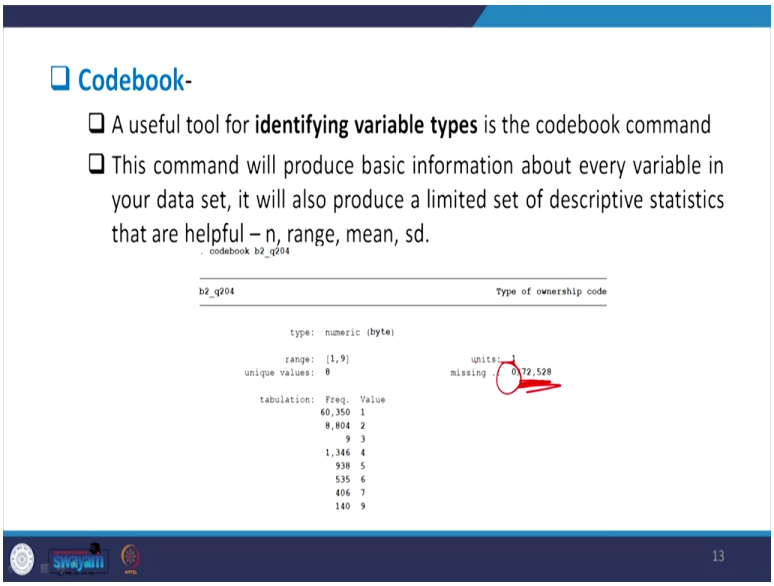
Here we are suggesting some detailing of it, detailing like su then detail. Summarize a detail. I think let me again go through that, like I have done here. I told you that su then you, key here, so here is a key I have entered then detail, so enter. What it gives? it gives other information as well. Like let me show you, since there are so many variables, I have gone through entire one, so many pages of ours unnecessarily occupied.

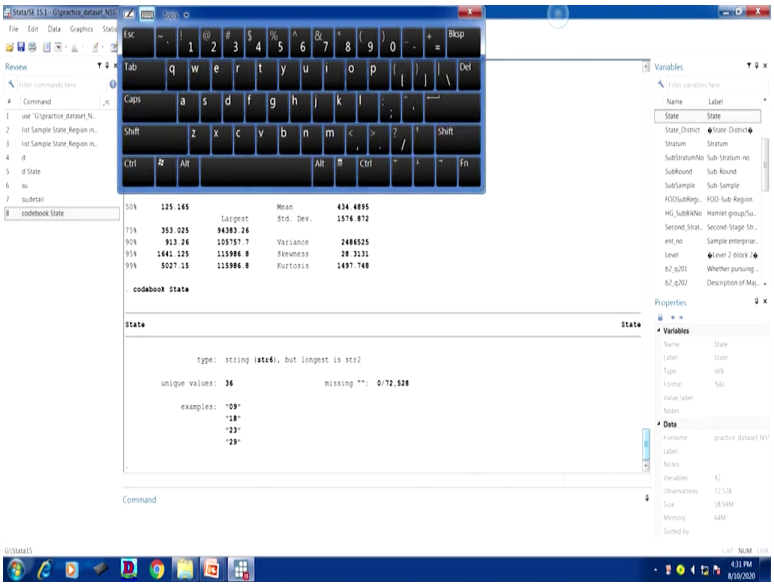
So what I do, I will suggest you stick to the particular variable and operate this, like su the variable you want, like 1, 2, 3, 4, whatever you just simply click on your variable tab, then you

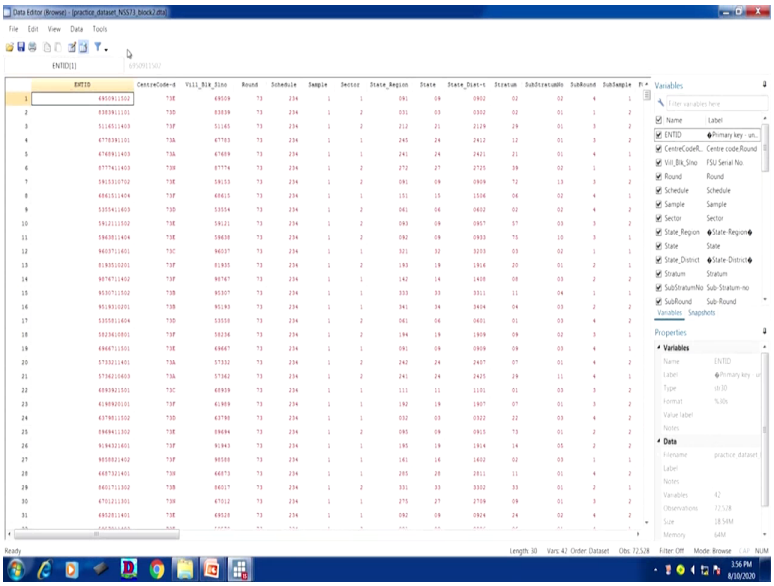
comma detail, it gives you more information like percentiles, then it gives like skewness, kurtosis it is visible here in the window. I am just showing you variance, mean standard deviation we have already shown you in the standard command. But in addition to that It gives percentile, variance, kurtosis, skewness along with the above information as I already guided.

So, summarize without argument, please take a note I am going to read this carefully. Summarize without arguments provide summary statistics for all numeric variables, like for non-numeric value variables zero was the result. For string variable, it shows no observation with their variable labels.

(Refer Slide Time: 37:49)







So, let me proceed further. let us understand codebook. So far we discussed browse, edit, then listing, then we discussed summarize. We are trying to guide you, codebook. How codebook, these commands are most often used by the users. We usually use all those commands simultaneously in our everyday life for working with the data or while working with the data.

So, codebook, let me guide you, is a useful tool for identifying variable types and is the codebook for identifying variables with a codebook command. This command will produce basic information about every variable in your dataset. It also produced a limited set of descriptive statistics that are helpful, like n number of observation, range, mean, standard deviation.

Look at the screen that is visible here. This is different than that of the summary statistics we have. These keeps like starting with which kind of data it is, then its range maximum, minimum value, then the unique values, then the frequency of the particular variable and their individual value with label is clearly given. Sometimes those labeling are not clearly written, only in value they have written. So labeling should have been mentioned. Usually the researcher will label, I will tell you.

This also give you information like missing values out of total observation you just look at carefully here, total observation information is there, how many are missing? Zero is missing. And this is important, if missing is there, instead of zero you have got some information like out of 72,528 samples, you are near about 50% is missing, around 30,000 or 40,000 is missing. So it must be noted as a serious issue.

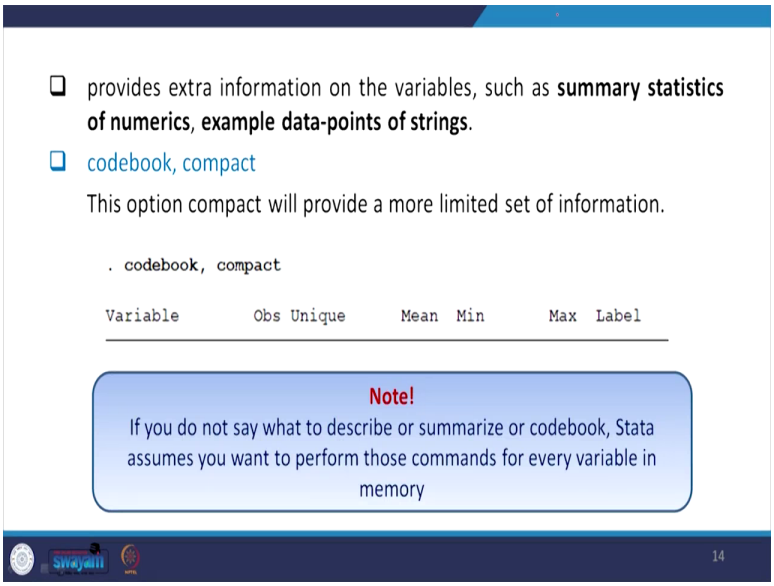
And you must cross check your data or if it is like 2% or 1% is missing and this number is given like out of this, let it be 7 or 10 out of this much is given so better to handle this missing data. Usually Stata reads missing as dot in the data. I will show you if it is available in our data.

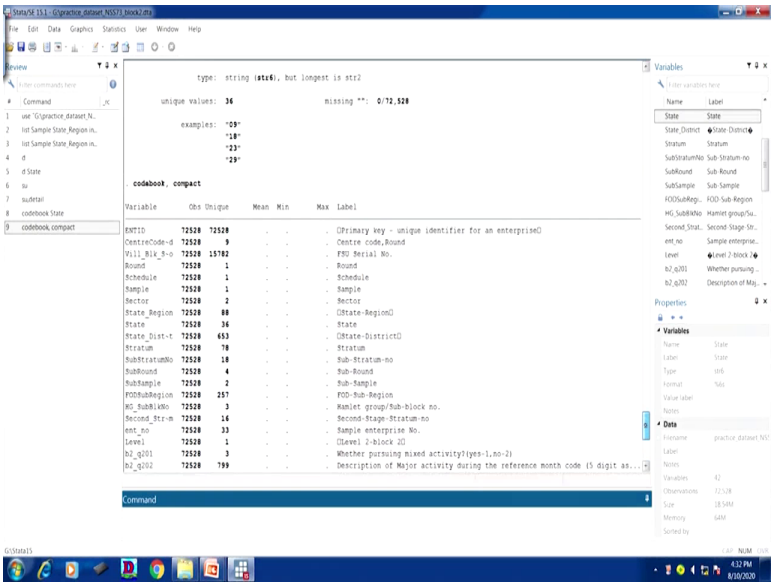
You just have a look, since we have filter for our use, I think we have already corrected the dot. All the entries are filled up. There is no dot here. But if any places you have seen any dot, only dot is entered instead of any other entries, then please note that this is a missing data. also let me operate this before you. Likewise, codebook, variable name if you give, it gives you the information.

Let me go back to the start up window once again and operate. So I will start with codebook here, if I simply enter codebook it gives all the variables and their codebooks. So I am interested for a particular variable, let me just go by state again. So let me enter.

So it gives me the information of, let us say string data and its different examples or entries are also visible and their frequency or their labeling, since labeling is not mentioned here, so not giving, but most importantly to be noted here, the sample and there is no missing value. Similarly, you can check please do experiment with other variables, you will find some interesting part of it and you may raise this question in our discussion time.

(Refer Slide Time: 41:50)





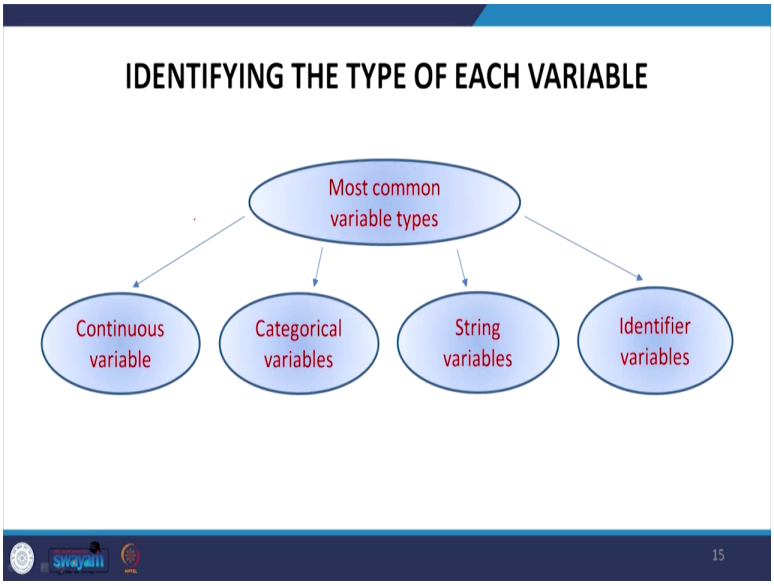
So let me explain you another interesting part that most of the researchers generally do not know. I claim this as important because there are certain important aspects for you to do some advance analysis or advance simplification of your data. For better simplification of our data or operating our data this command is going to be useful. Codebook provides extra information on the variables such as summary statistics of numerics, example data points of strings as well.

Codebook, compact is another aspect where all the variables are opened in another window with all such information, variable, then observation, unique ID and information mean, minimum value, then maximum value and the labeling. I will tell you right now.

Codebook comma compact, this has given me information in the window, in the same window that how result is, the data, the codes can be read in different approach like a number of observations is visible, variable and whether it is unique or not. Look at the first one ID. ID is expected to be unique, but not necessarily if there are other way of checking whether it is unique or not, I will tell you in a short while. But at this moment these two are same. I am going to tell you why this is going to be important.

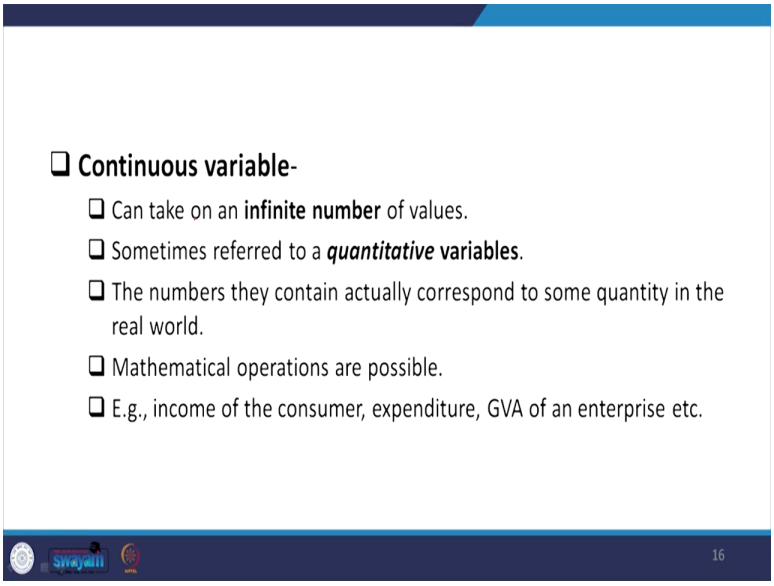
It seems that this variable is uniquely identified, but I am emphasizing now on the compact information. This gives in a vertical window of all that information and their labeling also. At a go, you can check all those details. So let me come back to our explanation. Please take a note here, if you do not say what to describe or summarize or codebook, Stata assumes that you want to perform those commands for every variable in the memory. So, if you specify, then it will give you. Otherwise, all details will be displayed on the window.

(Refer Slide Time: 44:32)



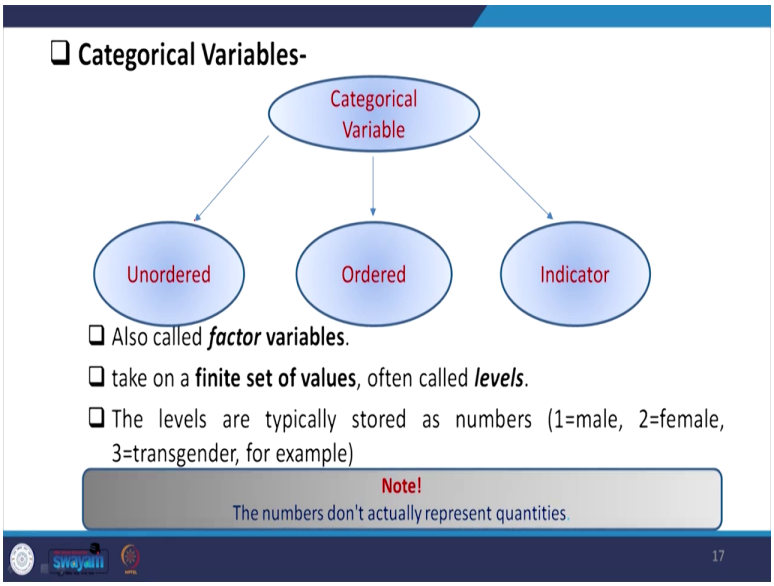
This is going to be very interesting for all of us to note that identifying the type of each variable, like I told you that some colors are visible, colors in red, colors in black, colors in blue. But let us understand, what are those variables? Why string? Where to get the string? How to read a string? So most common variable types are continuous, categorical, string and identifier variable. Usually these four categories are available in the existing datasets we are going to work with.

(Refer Slide Time: 45:25)



So the continuous variables mean that can take infinite numbers or values. Those also referred to as quantitative variables, they contain correspond to some quantity in the real world and mathematical operations are really possible, because these are continuous variable and these are quantitative in nature. Some example we are already discussed like income of the consumer, expenditure, gross value added of an enterprise etc. are called continuous variable.

(Refer Slide Time: 46:02)



Categorical variable, the word categorical again of three types. the meaning category maybe in order or may not be in order, so unordered or ordered or indicator variable. So like these are also called, categorical variable, also called factor variables. And take on finite set of values, often called levels. So since its ordered or unordered or even indicator, some boundary limits are defined.

Based on the boundary limits the ordering or non-ordering is possible and this is separate kind of variable. So I am using these words for clarifying the concept of categorical variable. These are sometimes also called close ended variables. The end is closed it is not open ended unlike the continuous variable.

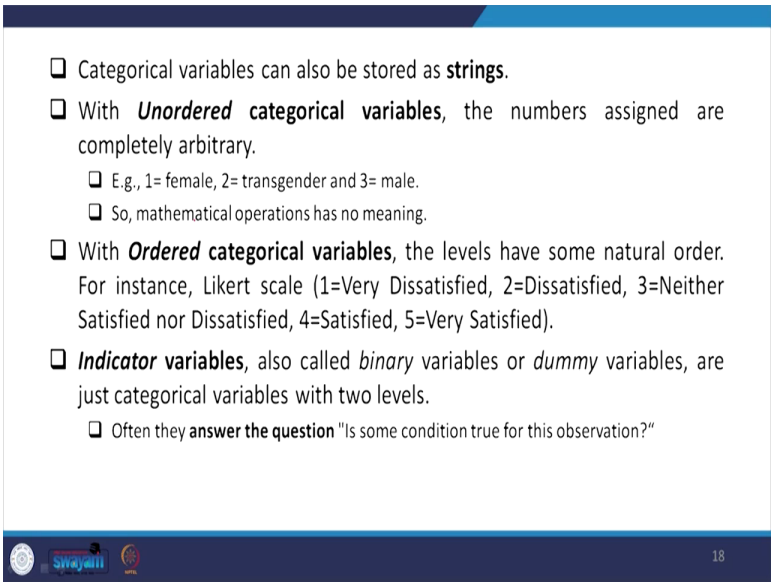
The levels are typically stored as numbers, basically labeled sometimes with one or two if it is taken gender as male and female or third is transgender. In our data, we are including all, the

National Sample Survey has already included transgender, for your clarity, so three is already there. You may operate with that or you may also operate with another one.

Please take a note that the numbers do not represent quantities. These are just though coded or labeled with a number, but they are never represents the magnitude of the quantity or the degree of the quantity or the value of the quantity. Two does not mean higher than that of one. It is not greater than one.

So please do not read in case of order or even unordered, ordered generally have certain higher values, is usually higher values ordered or in lesser value ordered, but in case of unordered one, any number, those number never represents any value. So mathematical operations just by this number is not possible. There are some specific mathematical operations suggested. We will tell you in that particular lecture.

(Refer Slide Time: 48:36)



Categorical variable can also be stored as strings. As I mentioned this word while understanding string variable, categorical variables like education in case of standard of education, first standard, second standard, third standard, those are categorical. But first, instead of one, two, three, they can carry with first roman letter, second roman letter, third roman letter. So if those entries are there, you cannot just operate all conventional statistical or stata commands will give

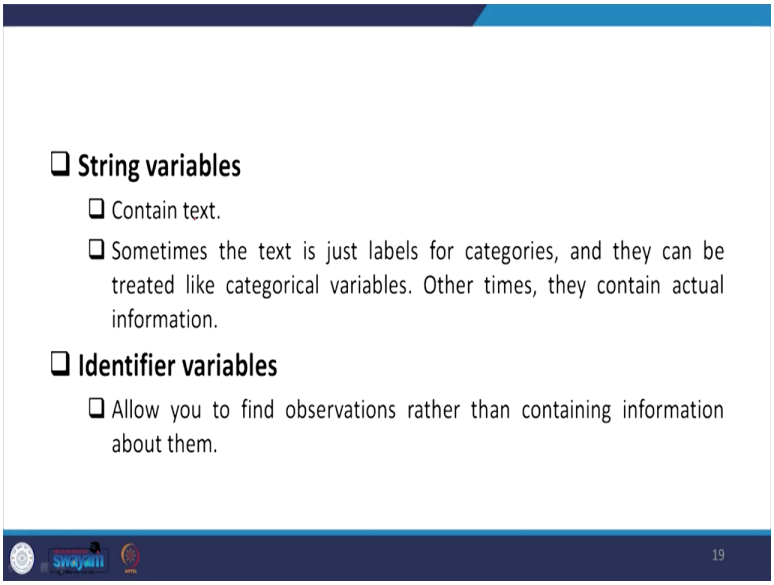
you the result is a wrong notion. You have to convert those string to numeric values for better operation.

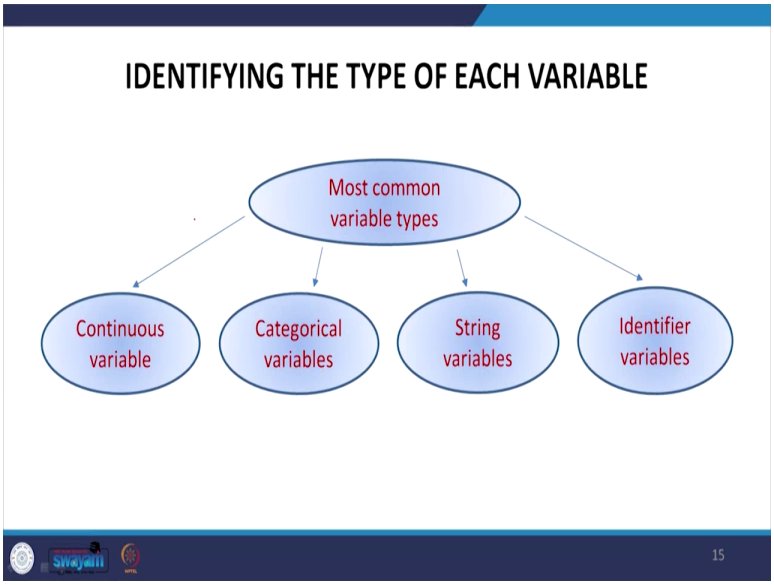
With unordered categorical variables, I have already discussed about categorical, if it is unordered one, the numbers assigned are completely arbitrary. It is completely arbitrary, like one for female, for gender I have already mentioned. So, mathematical operation has no meaning. I have already clarified that.

With ordered categorical variable, like standard as I told you or maybe customer care and their ratings usually in 5 point Likert scale, 7 point Likert scale these are in order, in different order. And so far as indicator variables are concerned, like if we are asking this type of question, are you satisfied with our service. So you have to indicate one.

Though options are generally yes and no, but you have to indicate one. It is nothing to do with the ordering or with its category. It does signal an indicator, so generally called dummy variables or binary variables. So often the answering the question is like some condition true for this observation or options will be yes or no.

(Refer Slide Time: 51:00)





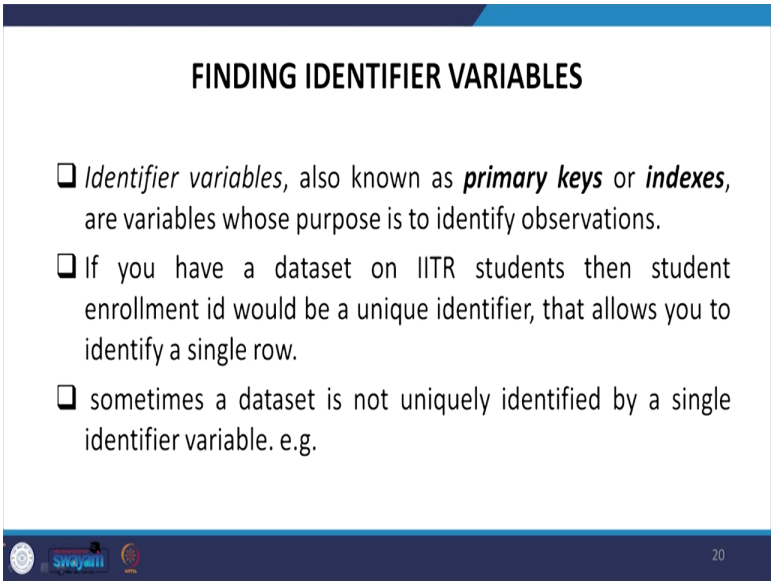
String variable, string variables contain, some further clarity to string variable is essential. I have already mentioned two types here out of four, continuous, categorical we explained, we are explaining string variables. Then we will explain identifier variables also. The string variables contain text, as I clarified several times.

Sometimes the text is just labels for categories like, the categories can be simply labeled and put in label, but the labels are not numeric values. So they can be treated like categorical variables only. Other times, they contain actual information. Even if actual information, some description information is there, descriptions are there, numerical operations just based on string variable is not possible. We have to operate in a different way, we will suggest in a short while.

So last one, understanding the variables which are entered in the data is called identifier variables. This identifier variable allows you to find observation rather than containing information about them. So, regarding observation, whether this is the identifier or not, is very much understood through the identifier variables. It has nothing to do with the number or nothing to do with the string number, nothing to do with the categories or not, only identifier, where this identify the particular row or identify the particular person or the household is very very important for understanding the data. I think I am little faster for it, but since I am providing everything together, friends you can understand from my slides.

So I will tell you through the data. I have already given you the background of different type of variables, different type of data. I have shown you, data in the sense the colors in the data. In the browsing window I have shown you the data. So I am clarifying the identifier variables.

(Refer Slide Time: 53:19)

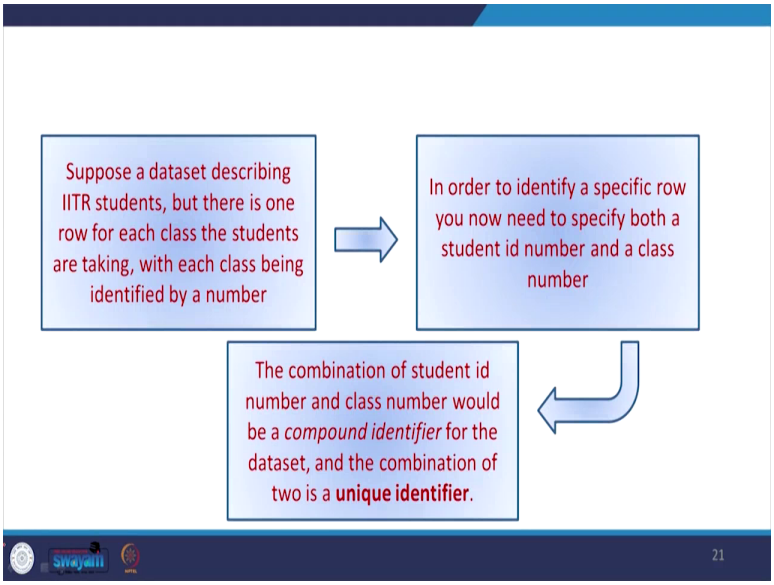


Identifier variables also known as primary key or the indexes and are the variables whose purpose is to identify the particular observation. In the row which particular observation is unique or there are so many observations of the same category. If it is not unique, then there will be some problem. We are going to tell you please hold for some time. In our lecture, we will certainly clarify.

If you have a dataset on IITR student, one example, IIT Roorkee students we are citing here, student enrollment ID would be a unique identifier, like student ID number, unique ID they define is generally the unique identifier, which allows you to identify as a single row. If all the students and their information are mentioned, but the starting one is your enrollment number against to that your age, your class name and your batch name, your subject name, your teacher name or your previous percentage whatever is written, but unique ID is already defined then that is perfectly fine.

But sometimes the dataset may not be having with a single unique ID. It might be coming with a compound variable. You may have to go by composite variable to define a unique ID. In the IITR student, for example, here let me read out once then I will clarify.

(Refer Slide Time: 55:00)



Suppose a dataset describing IITR students, but there is one row for each class the students are taking, with each class being identified by a number. In order to identify a specific row you need to specify both a student ID number and a class number. Student ID is there, but you need to identify which particular class.

Class, you wanted to add further information to the ID, like this represents also the class number, also the student ID number. If we are also attaching some information to the ID from, if I am the administrator, I wanted to track the person about his particular class, if that is attached with the unique ID, some additional information is attached with the unique ID, then from the new unique ID, you can easily track the particular observation or the student.

So basically what I wanted to say that combining this two information that is unique student ID with the class number. Combination of student ID and the class number would be a compound identifier for the dataset and the combination of the two is a unique ID or unique identifier. Similarly, in our dataset like IHDS data, there are so many information required, not just one ID,

the first column usually the ID, but that is not enough for observation to be defined as unique. I will guide you. It is there in our content.

So we have tried to clarify the previous one where how to define the unique ID. I will also test it through ISID command we have in our PPT, in our operation. I will tell you in our next class. So let me stop here and thank you so much we will continue from the next class.