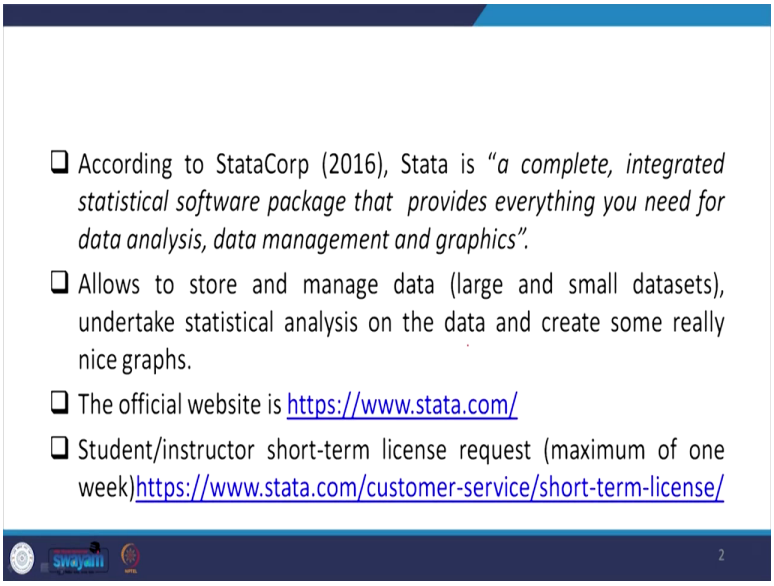
Handling Large-Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences,
Indian Institute of Technology, Roorkee
Lecture 11
Introduction to Stata

Welcome friends once again to the NPTEL MOOC program on Handling Large-Scale Data Using Stata. As part of the course, we have already started. This is our 11th lecture in this module. And we have so far discussed largely on understanding large-scale data. Over last two weeks, if you remember, we discussed three important databases, where we identified the core aspects of the data to be used in Stata. So, this is the 11th lecture, where we will be discussing how to initiate the Stata application, and accordingly, we have specified the title of the lecture as Introduction to Stata.

So now onwards you will be given the module on handling Stata. So, this week particularly this session is purely dedicated to different aspects of Stata, different features of Stata and there are so many commands in Stata we will be discussing in detail. So let me proceed for our detailed discussion.

(Refer Slide Time: 02:20)



Let me give you the backdrop of Stata as a starting point of our analysis of our application. As you all know that it is the company Stata with the name Statcorp. In 2016 mentioned quote and

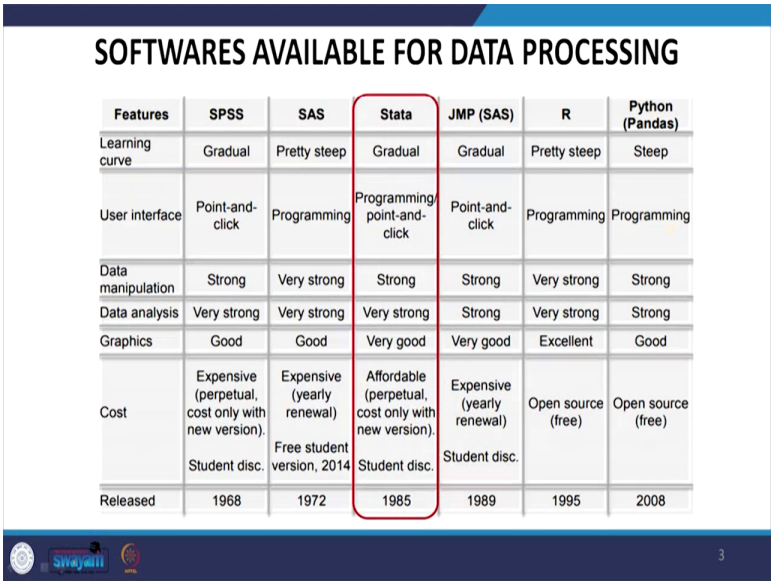
unquote, I have to quote this phrase, because this encompasses the detailed information of Stata. It is a complete integrated statistical software package that provides everything you need for the data analysis, management and their graphics. So in two or three lines we are expecting all those answers using the Stata also with the large dataset. Now this allows us to store, manage large as well as small datasets, large we have already discussed, large and small datasets earlier we discussed. But in simple meaning it is basically covering a huge number of variables as well as observations. So row by column in excel sheet if you are working, large and small differences can be easily understood.

since this package allow us to store, manage and also help in undertaking many statistical analysis and also we can generate very good graphs out of it. The website, as a starting point for you to look at is www.stata.com. If I just click it, it will be redirected to the screen, but I think you can handle on your own. So I am not spending time for it.

Since I am going to use Stata software, this is a licensed version and you cannot operate Stata on your own without a licensed version. This is the first requirement. It is better to go for a licensed version and these are paid. But what Stata has done, StataCorp or the company has done, for the learner like students and instructor, they give a short-term license request depending upon your requirement and they generally allow one week permission.

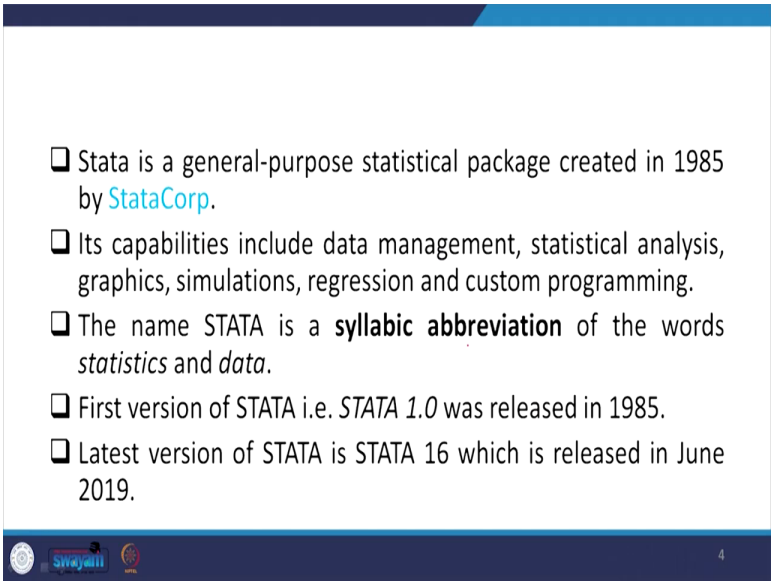
And I have already attach the information for you to fill up the form and get the license for one week and it is the prerogative of Stata to provide for this particular software. And I will suggest that before reading our other commands in these slides and in our successive slides, you should procure one copy of Stata and that will be very useful for better understanding.

(Refer Slide Time: 05:30)



Let us move to the understanding of various datasets, various statistical packages available for social scientist or even for the corporates, for the managers in their different companies to handle different information or different datasets. For us, Stata we have highlighted here you can look at why data is important and why it is so necessary based on how easy you can handle and how better data manipulation you can do related to data analysis, also affordability of that particular software. So you just have a comparison for your idea, for your knowledge that will be useful.

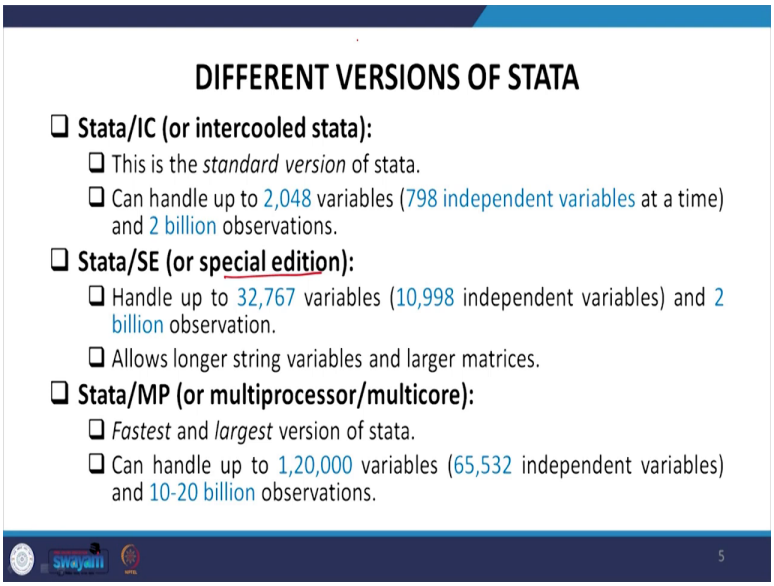
(Refer Slide Time: 06:29)

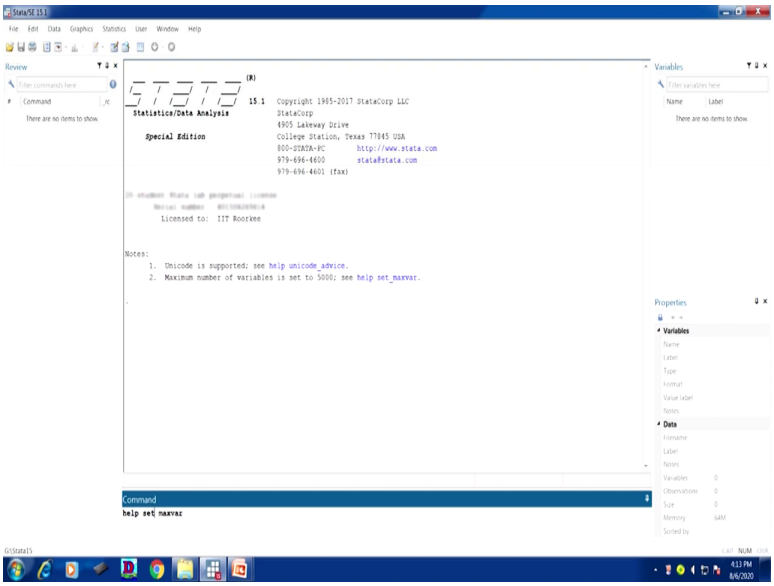


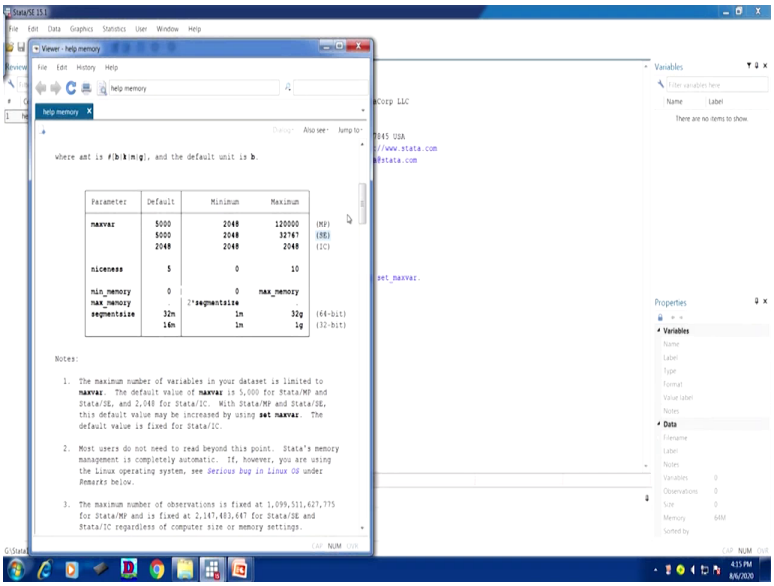
let me move on to the next set of information for you as a starting point for the learner, for the first time users of Stata. As I already mentioned, this package was initially created in 1985 by StataCorp and its capabilities include, management of data, analysis with statistical tools, graphics, then simulations, we usually apply regression techniques and custom, many customs programming.

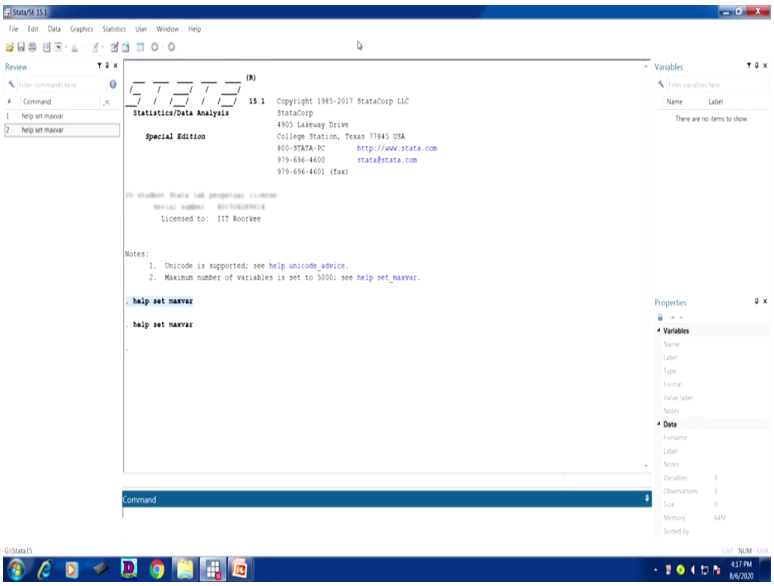
So the name Stata must be intriguing to you. Why it is Stata? It is a syllabic abbreviation of the word statistics and data. So, both the information are combined together for the acronym Stata. The first version of Stata, as I already pointed out, that is Stata 1.0 was released in 1985. The latest version which you should procure for your better understanding and analysis is Stata 16, and this is released already by StataCorp in June 2019.

(Refer Slide Time: 07:22)




Stata/SE 15.1
File Edit Data Graphics Statistics User Window Help
Review
Command
There are no items to show.
Variables
Name Label
There are no items to show.
Properties
Variables
Data
Command
help set maxvar
G:\Stata15
CAP NUM OVR
4:13 PM
8/6/2020


Stata/SE 15.1
Viewer - help memory
File Edit History Help
help memory
where amt is #[b|k|m|g], and the default unit is b.
Parameter Default Minimum Maximum
maxvar 5000 2048 120000 (MP)
5000 2048 32767 (SE)
2048 2048 2048 (IC)
niceness 5 0 10
min_memory 0 0 max_memory
max_memory 2*segmentsize
segmentsize 32m 1m 32g (64-bit)
16m 1m 1g (32-bit)
Notes:
1. The maximum number of variables in your dataset is limited to maxvar. The default value of maxvar is 5,000 for Stata/MP and Stata/SE, and 2,048 for Stata/IC. With Stata/MP and Stata/SE, this default value may be increased by using set maxvar. The default value is fixed for Stata/IC.
2. Most users do not need to read beyond this point. Stata's memory management is completely automatic. If, however, you are using the Linux operating system, see Serious bug in Linux OS under Remarks below.
3. The maximum number of observations is fixed at 1,099,511,627,775 for Stata/MP and is fixed at 2,147,483,647 for Stata/SE and Stata/IC regardless of computer size or memory settings.
Variables
Properties
CAP NUM OVR
4:15 PM
8/6/2020



Let me also introduce to you with a different version of Stata, are generally three version of Stata are available, one is Stata/IC, IC stands for intercooled version, and second one is special addition, and the third one is multiprocessor or multicore version.

So, let me also tell you that in Stata, once you open that link, you can be able to understand the Stata, if you open the Stata software, this will look like this. The Stata version will look like this and we have already installed in our system and this is the version. I want to introduce all its points, all its entries, different windows I will discuss very clearly to you.

But for your clarity, if you want to understand which different versions are available, what maximum space are available in different versions, just for your clarity, I am mentioning here that you please do one thing. I will help, why am I doing all those operations. So I am just going by one operation, one command for some clarity that will help us, it will come as a set and without that also I think it will come.

So I am just writing help. Help is the command we are going to use many times in order to understand Stata in detail. So I have retained for something to show you that help set maxvar. So maximum variability is given here. It is visible in the result window. It will just come in a short while. I am just entering the command in the screen. What is being popped up, you just look at. what is being popped up here for you.

So, I am just looking at which type of versions are there for us. Let me make a bigger screen for you, it is here. So let me close this for a minute. So this is the MP version, look at here MP version, then special edition version and intercooled version.

IC stands for intercooled, then special edition, then multiprocessor version. Each version has different space. Space in the sense, it occupies different byte space as well as different variables, observations. I will clarify in that PPT. So, for you, just for your knowledge, you can have an idea of this particular table. You just type this command. I will clarify later, so you will understand.

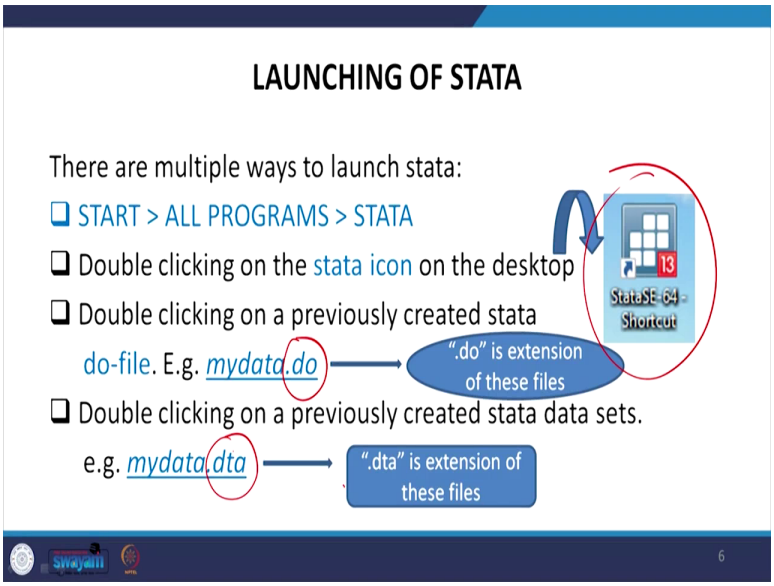
Let me move to our slide once again. So the very standard version of Stata is none other than intercooled version. The intercooled version is the version which can handle up to 2,048 variables and which is at maximum covering 798 independent variables at a time and 2 billion observations.

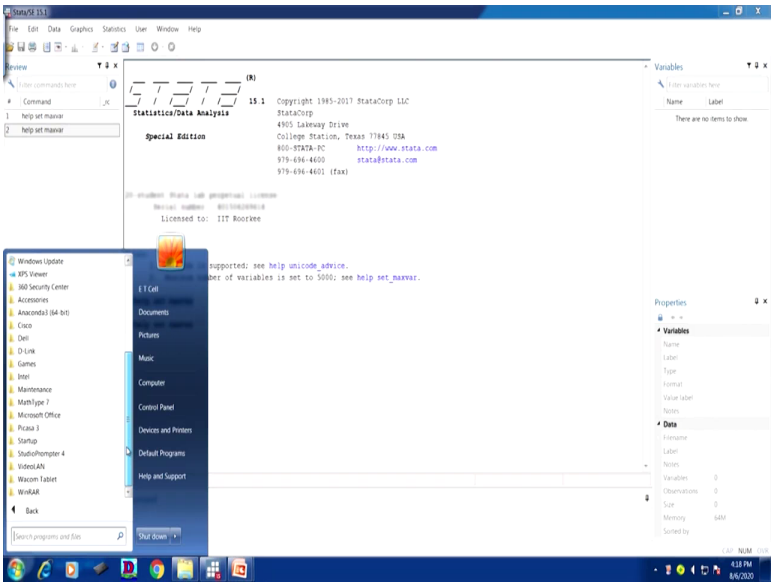
Whereas, the special edition, is mentioned here, that can handle 32,767. This was already given in that window. I have already mentioned this was there in the window. I am again, for your clarity, I am just putting a double click here. It will give you the same 32,767, then 2,048, then 1,20,000, these are there I have already explained to you.

Let me move on to the slide. So these are the information we derived from the Stata itself. We are not writing on our own. These are derived from the version. The third version is called the MP version that covers 1,20,000 variables, that is of 65,532 independent variables. We will clarify all those operations and how much independent variables are covered and how much you can set, which command to set, like for our version, let me show you one thing here for your clarity.

Look at the window of the Stata which is loaded in our system, you can get the demo version and do it. On the left, top left corner, it is clearly mentioned the Stata/SE 15.1. So this is the special edition of Stata which covers 32,767 variables. So, this is the version we are operating right now. I think most of our research can be done with the help of the special edition. If you cannot able to procure the MP version, you can able to do it with this version.

(Refer Slide Time: 13:51)





Let me move on to the next slide clarifying the very basics command of Stata. basically how to launch the Stata. There are multiple ways to do that. One is if you just go to the start button of your computer and all programs, let me show you in a short while. Here is the start button. If you go to the start then all programs, then you will get the Stata. Since Stata is in our system, in our computer, so it will be popped up. Any kind of software is there; it will be automatically come.

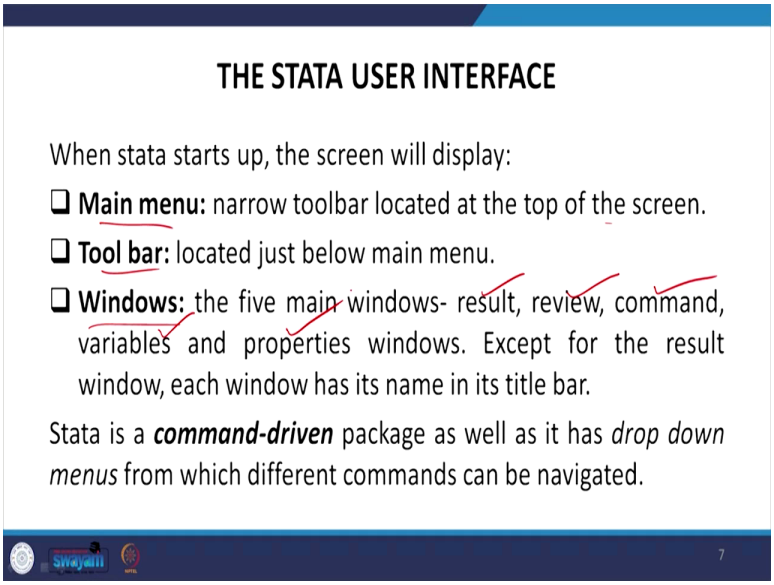
Once you load it, it will come and we are operating from our system and let me now move to how to do it. Once you load it, you will certainly get that. Another option for you is to start it

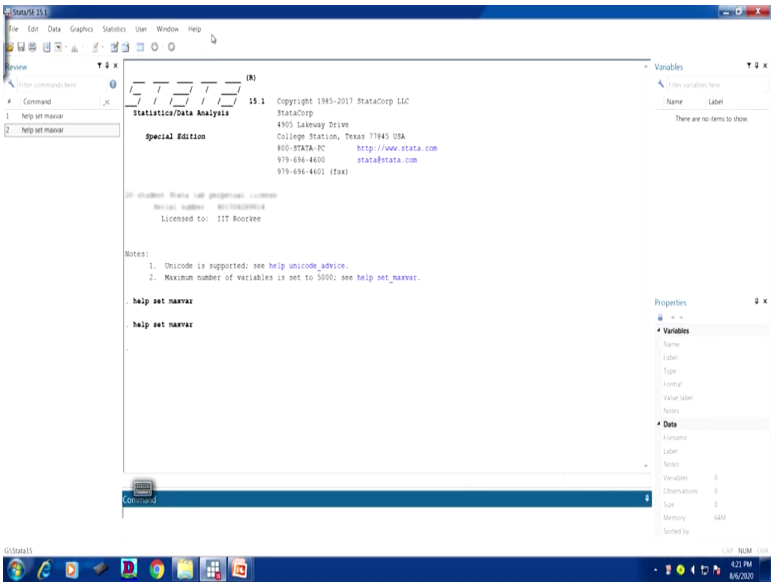
through double clicking on the icon. Once you have installed it or loaded in your system, may be on desktop or in your computer or laptop, This type of icon will be there for sure.

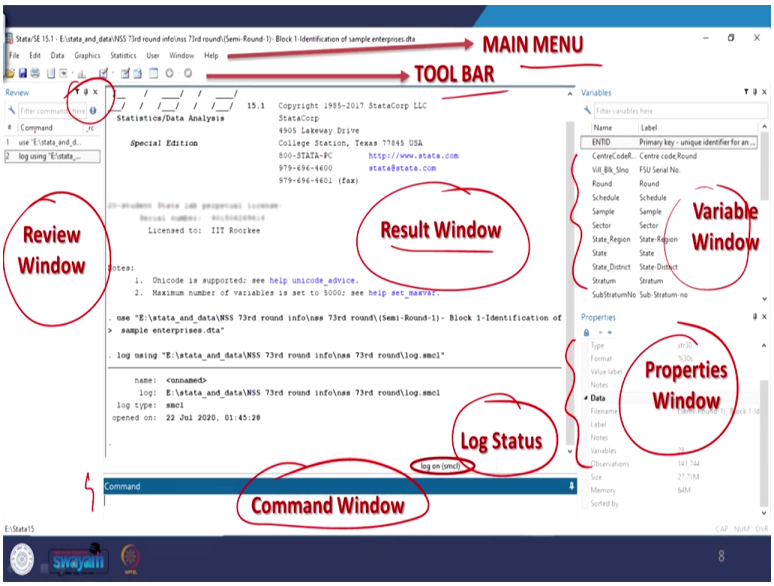
And once you double click on it, that particular screen will be coming for sure. second one is by double clicking on a previously created Stata do-file, you can get that window also. Like I will tell you what is that do-file and how to go for that do-file, I will clarify in my successive slides.

Similarly, double clicking on a previously created datasets, like you have a dataset, we have already shown you three datasets, IHDS, NFHS and NSS datasets, each dataset set is available with .dta file is the extension of Stata file. Do-file, we will clarify what you mean by do-file. Do-file usually we save our commands or syntax and those can also be operated just by double clicking on that particular file. We will read it for you and clarify it to you.

(Refer Slide Time: 16:24)







And now let me also clarify some of the other important aspects of the interface as part of our introduction lecture. In our next lecture we have the dataset, screen of the dataset, different variables are mentioned, some of the operations will be made. Let me quickly go to the interface, the window of that particular version of the Stata once you have installed it.

There are different parts of the interface. One is called menu bar. let me first tell you, then we will come back to this particular one. Here is the screen. The menu bar, the toolbar, then there are different windows of our operation related activities in the Stata, like review window, result

window, variable window. Variable window is here, result window is here, then review window is here, then properties window is here, then command window, then log status. We will clarify.

And better to first start with the main menu. Main menu is mentioned here and toolbar is mentioned here. Just try to see it from the Stata software. So main menu I have already clarified to you. Tool menu is just available below to that of the main menu. And the review is written here. You can find out Review is here. Then variable window, then properties window, then command window, then result window and the log window is not visible. Once we go for log file, I will clarify what you mean by log file.

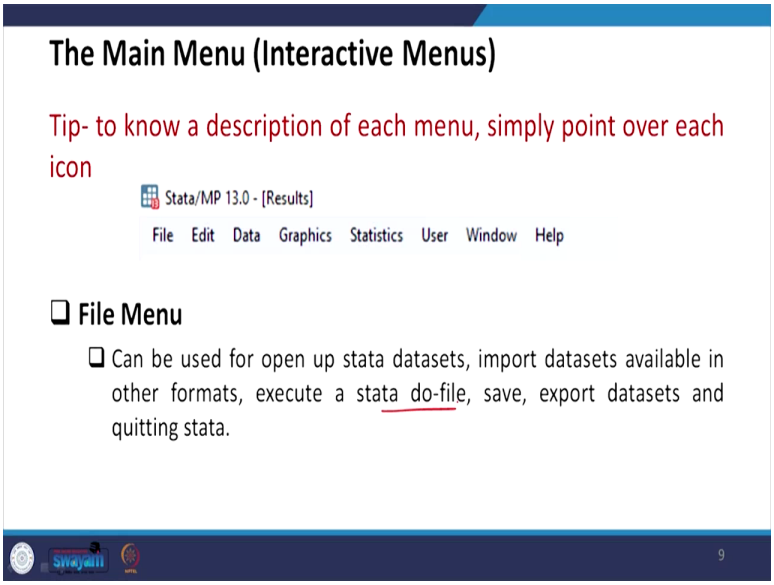
The log file has certain options to save. So once you save that, it will automatically pop-up, whether it is log file is on or it is off. I will clarify it. So let me go to the previous slide once again. As I already mentioned, it has main menu, then toolbar, windows, then let me just clarify something inside it, like the main menu that is the narrow toolbar located at the top screen, then toolbar that is located just below the main menu, I have already mentioned.

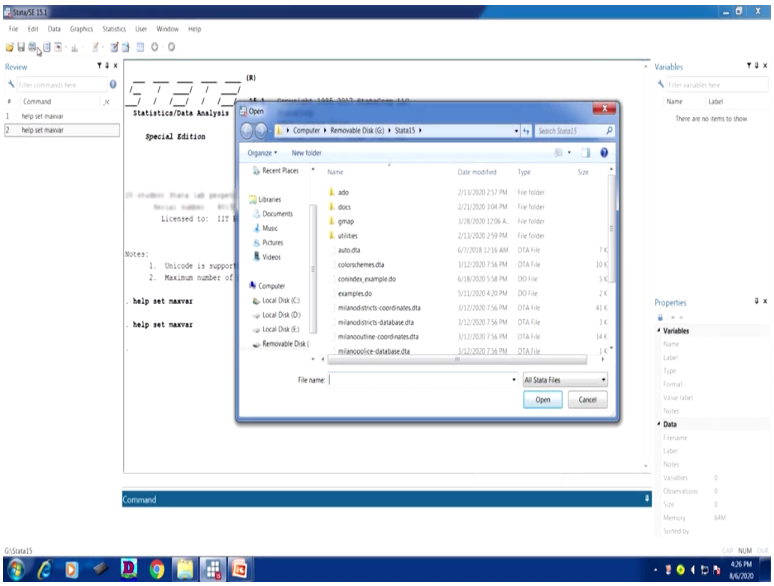
I have already shown to you, there are five main windows, that is result, review, command, variables, properties window. So, broadly five windows are going to be very useful while operating with Stata. And except for the result window, each window has its name in its title bar. So result window it does not have, look at result window is visible here, it does not have a title bar, otherwise all other have their own title bar.

Stata is a command driven package as well as it has drop down menus. So it has a command based menus. We can type command and we can also operate through the drop down menus like these. Drop down menus we referring like, if I just click file, it will give you other options, like click edit it will give other options, if you click data it will give so many options. Similarly, graphs, statistics, user, window, there are so many operations will come and we will clarify one by one.

Let me move into the next aspect. I think I have already clarified this. We will also decode each of the explanation given in variable, even in properties, even in command, even in the small notations inside it, small symbols inside it. There are so many aspects will be discussed in our slides as well as in our lecture.

(Refer Slide Time: 20:45)



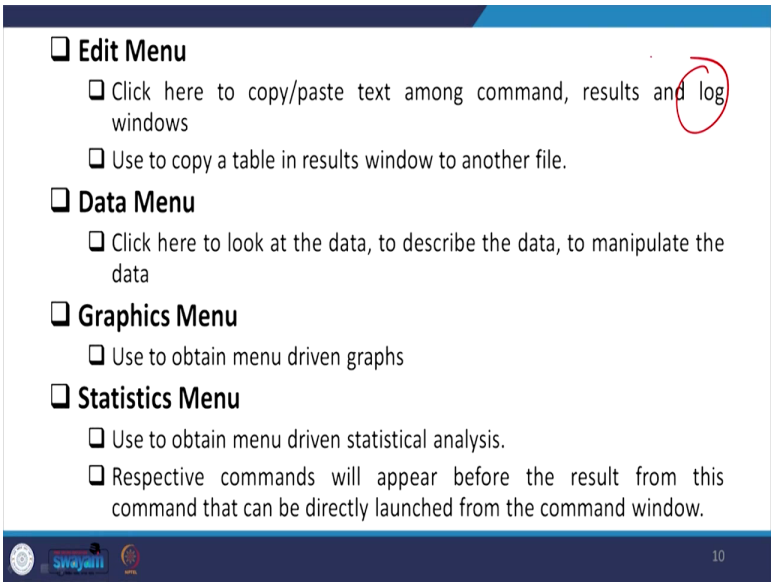


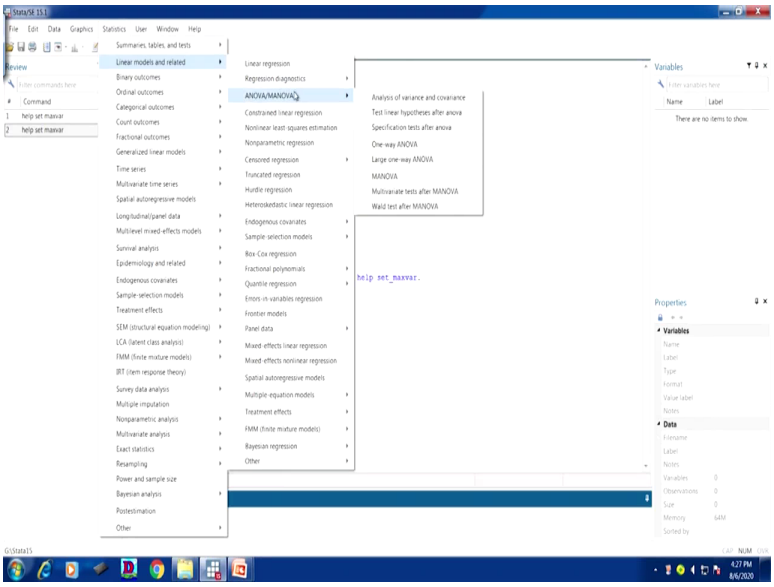
The main menu, as I already mentioned, that is also called the interactive menus. And so simply if you can point over that particular main menu it has popped up, it has resulted in several other indicators, options. These are the options, main menu options are given file, edit, data I have already shown to you. what do you mean by file menu?

The file menu can be used to open datasets. Also it helps importing datasets available in other formats, executing a do-file. Do-file we will tell you. This is bit technical. And we can save that. Those operations you usually do in other software, other apps. So, but do-file is technical. We

will tell you. Similarly, data export, exporting dataset and quitting the Stata at what how you can able to quit let me just show you here. Like in the file menu, look at this. You can open the dataset wherever you have saved it will open. If you have a .dta extension based files or files are there, it will automatically, if you click on it will come to your window. We are not loading right now. We have other options. I will clarify which other options are also easy for you to operate.

(Refer Slide Time: 22:23)



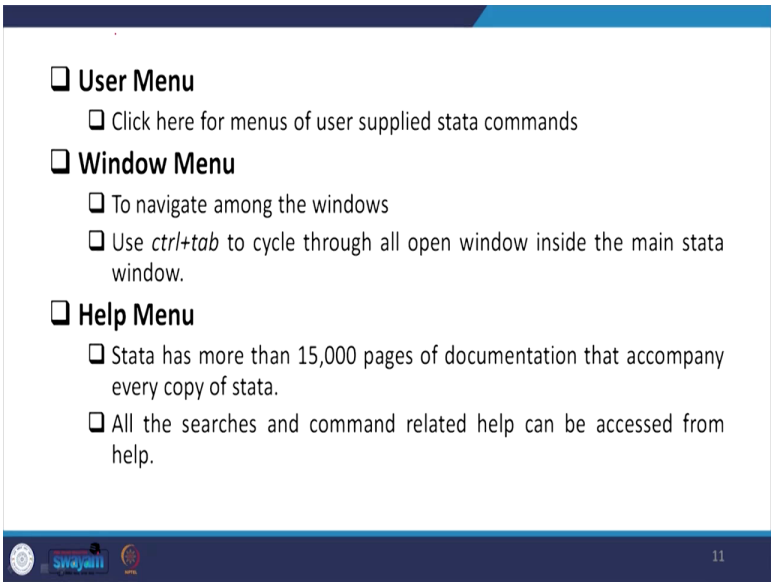


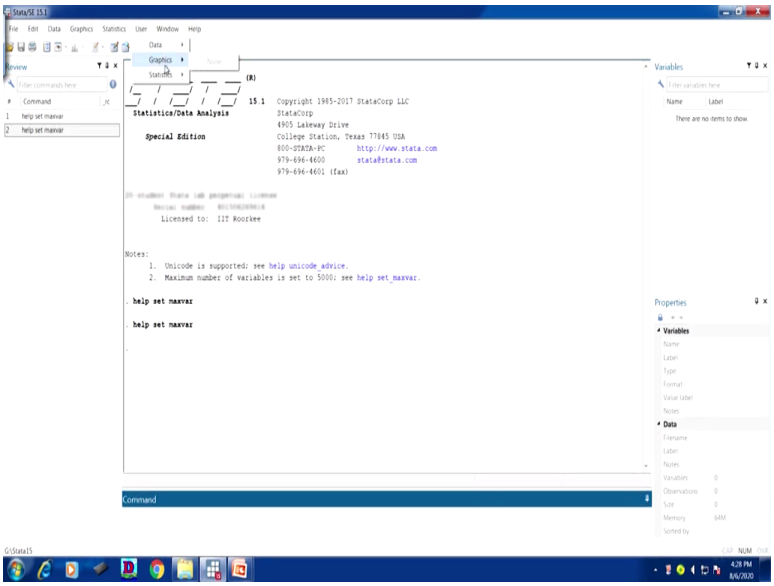
Let me move on to other edit option, like copying, pasting of the command or result, log windows. Log, this is technical. We will clarify in our successive slides. Similarly, it is useful to

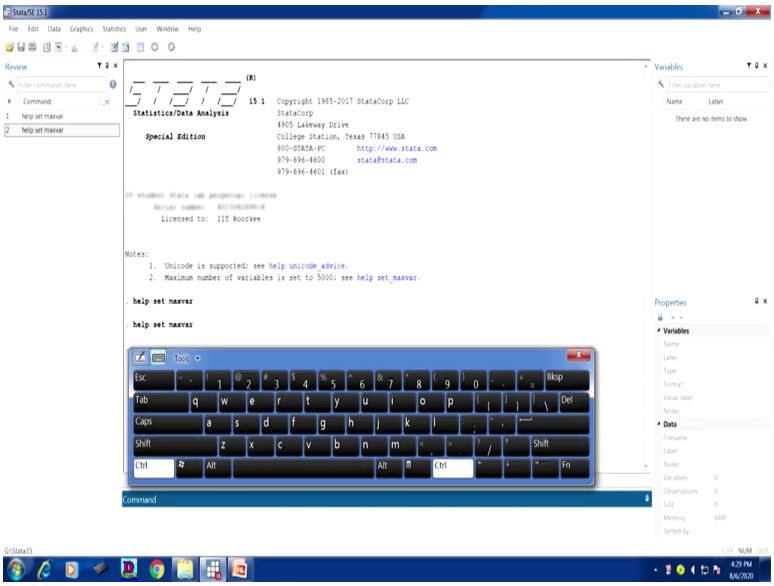
copy a table in result window to another file. You wanted to take the result to your word file or to excel file. So this is going to be very useful. In which version you should copy, we will clarify.

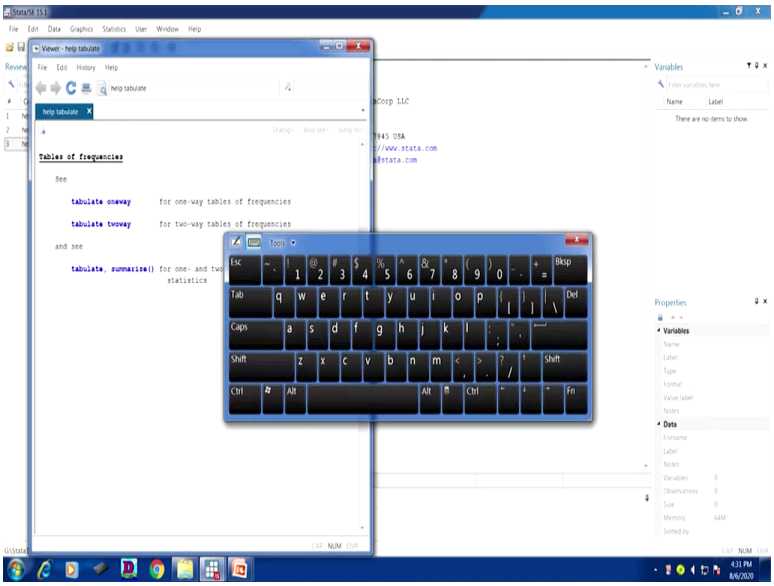
Similarly, data menu, it helps in describing the data, manipulating the data, then graphics, similarly graphics menu, statistics menu, it contains different statistics. For your knowledge, for your interest, let me just point over on particular statistics. Look at so many things, like summaries, tables, test, linear models. We usually go for regression model, linear regression diagnostics tests are there, ANOVA, MANOVA. There are so many operations, panel data and binary outcomes, like logistic regressions and so many. Those who go for experimental design, they have also some options. So, you will explore later on.

(Refer Slide Time: 23:51)




Stata/SE 15.1
File Edit Data Graphics Statistics User Window Help
Data
Graphics
Statistics
Review
Command
1 help set maxvar
2 help set maxvar
(R)
Statistics/Data Analysis
15.1
Special Edition
Copyright 1985-2017 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
Licensed to: IIT Roorkee
Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5000; see help set_maxvar.
. help set maxvar
. help set maxvar
Variables
Filter variables here
Name Label
There are no items to show.
Properties
Variables
Name
Label
Type
Format
Value label
Notes
Data
Filename
Label
Notes
Variables 0
Observations 0
Size 0
Memory 64M
Sorted by
Command
G:\Stata15
CAP NUM OVR
4:28 PM
8/6/2020


Stata/SE 15.1
File Edit Data Graphics Statistics User Window Help
Review
Command
1 help set maxvar
2 help set maxvar
(R)
Statistics/Data Analysis
15.1
Special Edition
Copyright 1985-2017 StataCorp LLC
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
Licensed to: IIT Roorkee
Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5000; see help set_maxvar.
. help set maxvar
. help set maxvar
Tools
Esc Bksp Tab Del Caps Shift Shift Ctrl Alt Alt Ctrl Fn
Variables
There are no items to show.
Properties
Command
G:\Stata15
CAP NUM OVR
4:29 PM
8/6/2020



Let me proceed further. Others like user menu, user menu contains users supplied stata commands like this and it contains some you will get the, if you have the data, it will show it and if you have graphics, it will show it. Statistics, if you have already done it, will show it . But right now since we have not yet started, I cannot show you. But user menu is also going to be useful.

Similarly, window menu, that helps in navigating among the windows. If you just click control and tab, it will automatically navigate from table, then review, then variables, then properties, then command automatically. Let me try for you, control and tab. I am just going to do it here for you. You can just have a look, control, then tab. This is tab.

So, tab review is being popped up, control then tab, variable is being pooped up. Then tab is showing properties and other if you control and tab, it will show you the, further details of the properties, then control tab, a command window is being highlighted.

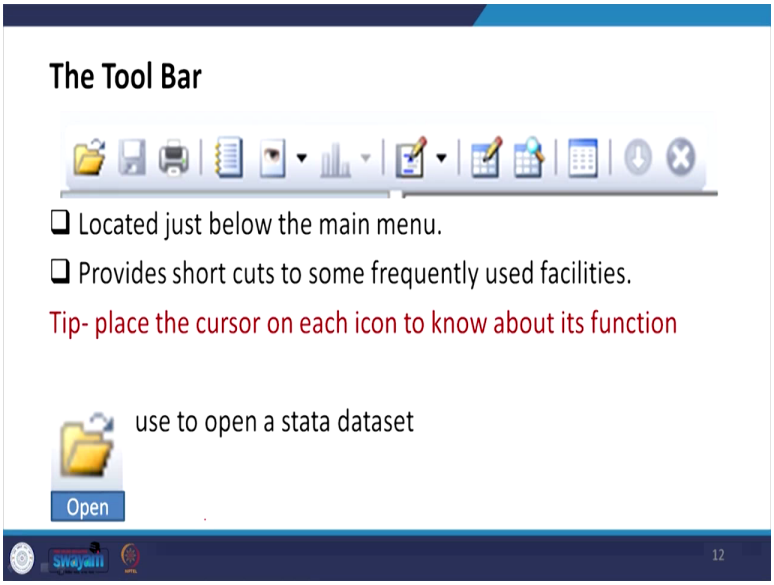
If you again do control and tab, it will again come back to the review. And these are the command you must remember. It help you to fasten your operation and it makes you very easy to operate data for quicker results. Generally, it is suggested for the users to remember and keep it with a handy mode.

Let me go to the help menu. Help menu, as I already told you, if you go for help, it shows so many commands, like if you simply type help here, help of anything, like you want to tabulate,

help tabulate, it will give you so many options, like one-way tabulation, two-way tabulation or other way tabulation, as well as summarize it will give you.

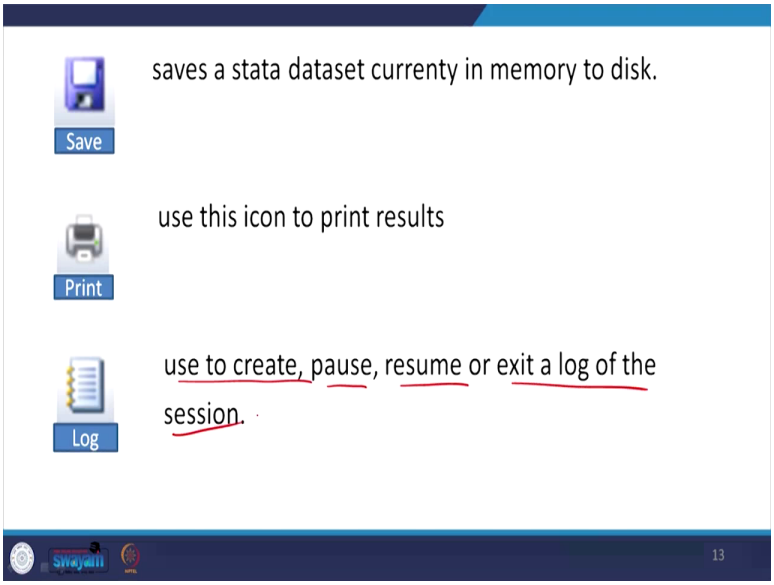
Similarly, if you want to get help, you want some data, like ready data available. So one command we are going to tell you system use, but right now, let me skip it. We have all those details in our slide. I will explain it later.

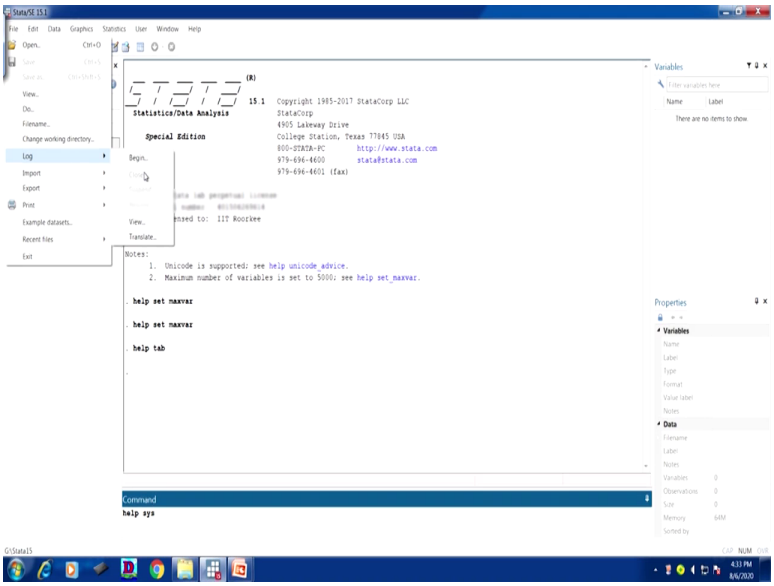
(Refer Slide Time: 27:16)



So let me move into next slide. It has all those coverages. What the toolbar contains? The toolbar contains all those things. I have already shown you, these are the toolbar below to that upper main menu. So toolbar contains this. So it provides shortcuts to some frequently used facilities. place the cursor on each of the icon to know about its functions. the way I have done it. Like there is a folder, if you just click on it, will redirect you to the available dataset in your computer. We can open that.

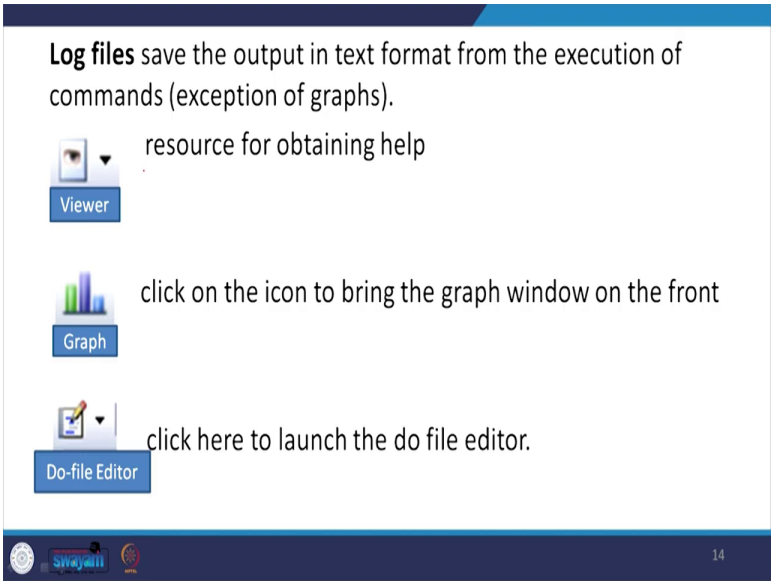
(Refer Slide Time: 27:51)

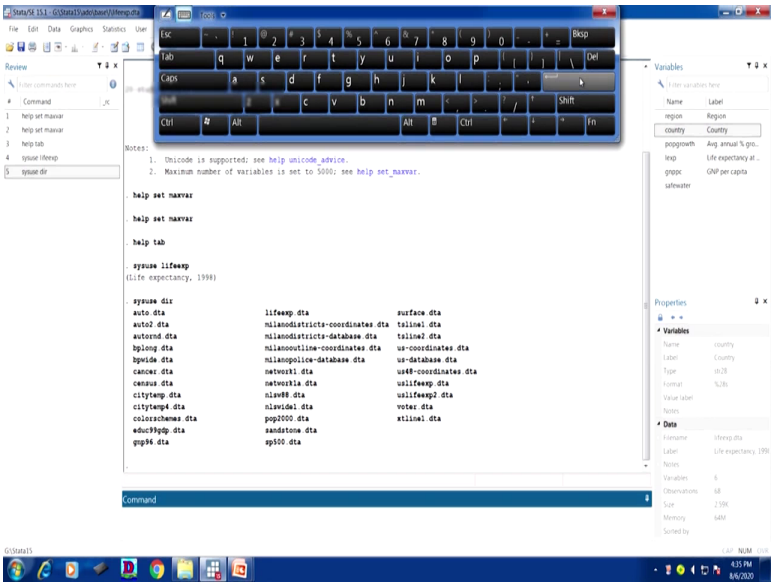


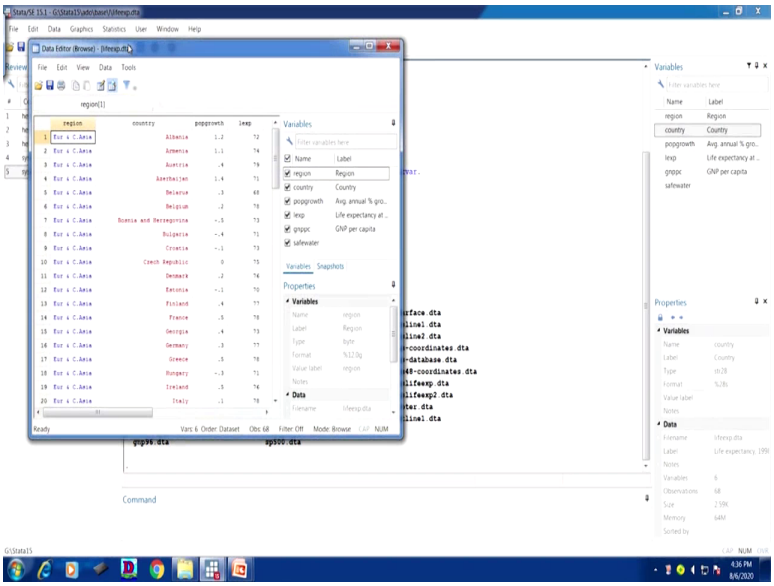


Let me move into the next, similarly for save, print, log file. We are going to tell you in detail, use of create log file, pause, resume, exit a log of a session. Like let me just tell you that is the way. Here is the log file icon. I have simply clicked on it. Once again, I have simply pointed over that particular log. It shown as log, log begin, close, suspend, resume is written. Otherwise, from the file also, from the file, we have also a log option. Here is the log. Begin, then close if you have already opened the log, it will be there for sure.

(Refer Slide Time: 28:44)







So, let me move into further details. Otherwise I will be completely lost in that particular command. There are so many other options. Similarly, so resource for obtaining help in detail are available. So log file, as I already mentioned, it saves the output in text format from the execution of commands exception of graphs. Graphs cannot be saved in text format as you all know. So there, you need to save it in different window. Graphs popped up in different window.

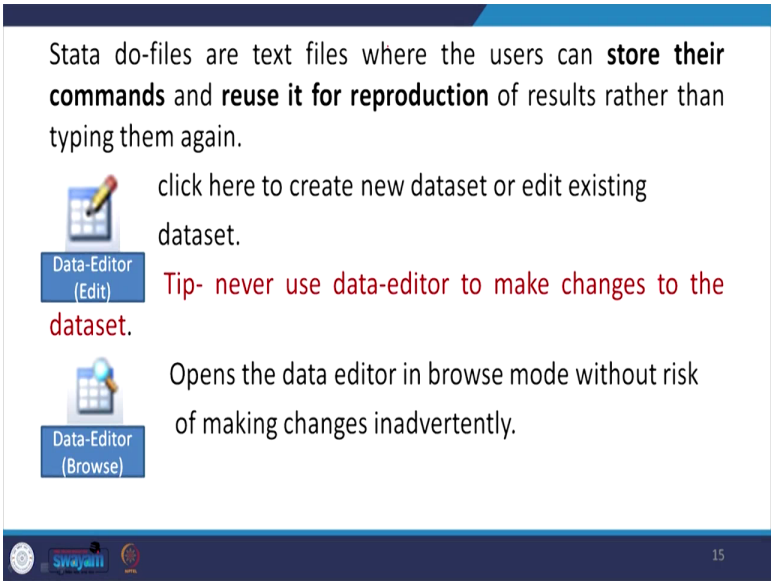
Similarly, another one is the do-file editor. Click here to launch the do-file editor. It is always suggested to all of you that you please start with the do-file editor. here is the do-file. Since no data loaded in this software, we are going to load it. If I use system use, just for your clarity I am trying to do it within one minute that let me be explicit. I will just load a data for you. sysuse is the command. We will tell you. I am deliberately trying to get one dataset, lifeexp. I know the database where it is located. So I will tell you also.

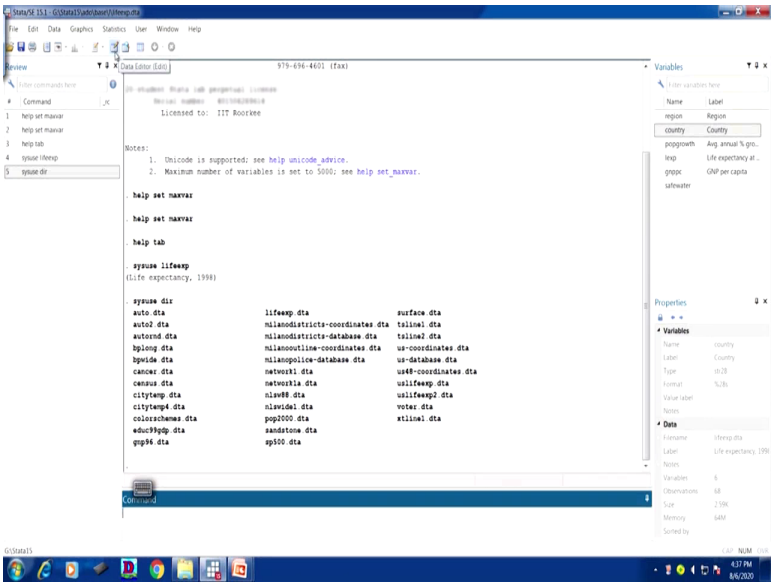
The data file is loaded with its variables; with its different nature of data I am going to clarify in a short while. But otherwise, for your clarity, if you want to open the database directory for you, so you just simply go to the sysuse and you simply type directory, dir. It will give you all option datasets which are already available in Stata directory. I have used, for the timing I have used the life expectancy data for the analysis. So here is the data, the second, the first one, lifeexp.dta. So you can do it when we are going to use it, you can use it simultaneously.

Let me proceed to our explanation. So we are going to do it do-file. Once data file is loaded, since we have already come up with the data, you can create the do-file like this, then it should

pop-up. Do-file should come, right. Anyway it is there but I am going to get it, in our slide I will get it for sure. Otherwise, it should be there. But let me also show you the data which we are trying to refer. This is the data we are trying to refer. And let me try it. So we have those details. Do-file editor details are there. I will open it for sure.

(Refer Slide Time: 32:03)



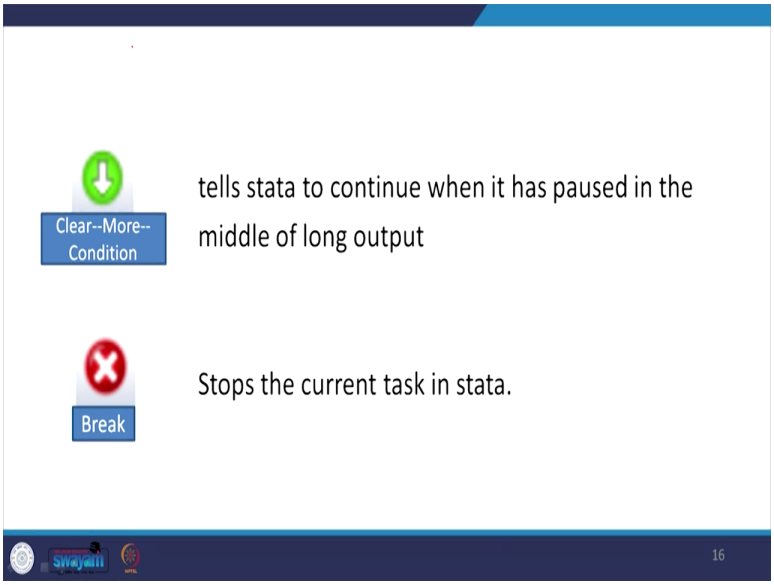


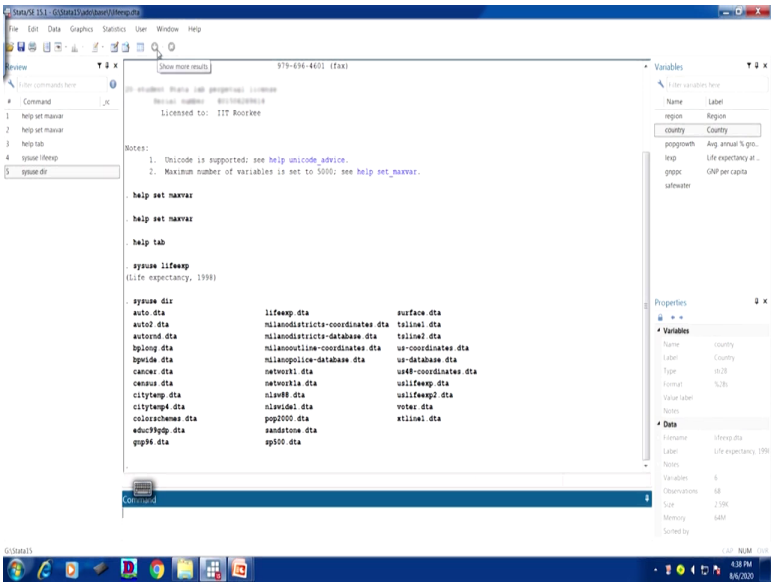
Stata do-files are, let me clarify what do we mean by do-files? Stata do-files are text files where the users can store their commands and use it for reproduction of results rather than typing them again. Similarly, it has two options. One is regarding data, we have two options, one is data edit,

another is called data browse. It has this column. These two are there. I am just putting my cursor here. This is data edit and this is data browse. It is always suggested to go for data browse instead of data editor.

If unless you do not require some kind of recoding, editing of the data and where to edit, if you do not have those capture, do not have those necessities, better not to open the data edit. Always better to check the data browser. It is there. All those details are mentioned in this particular slide. You can explore.

(Refer Slide Time: 33:14)



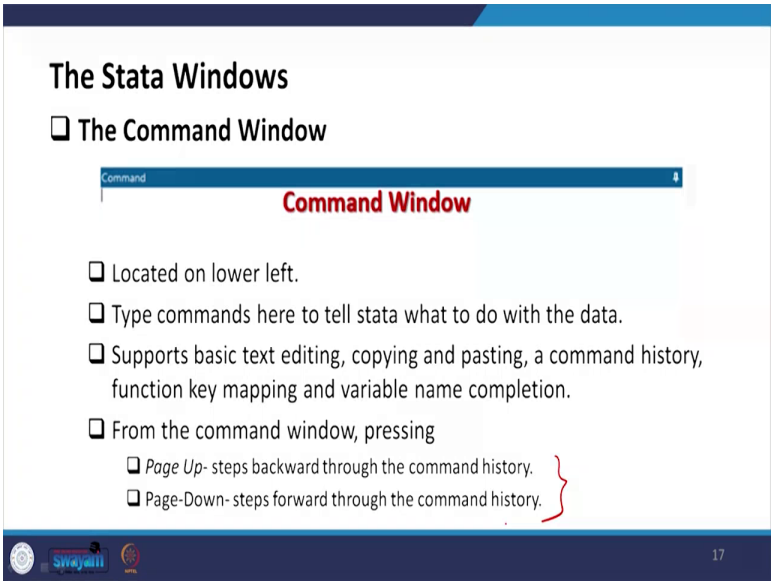


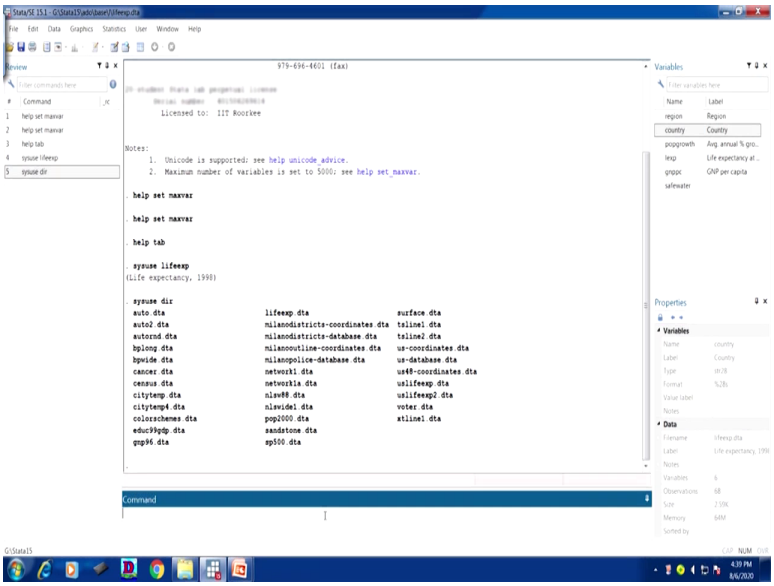
Similarly, once you have opened the data, you have operated with certain command, but somewhere it is being paused. So, what you do, you click on this one. Here is the arrow, the down arrow. Next to the close, the break, one this is a down arrow. It will pop-up with a color. And once you click it, it will work.

If there is any kind of pause due to the system problem and it will operate correctly. Similarly, break, if it is completely problematic and you have already run a command which is wrongly redirected you and that too it is taking huge time, you wanted to get rid of these problems and

you wanted to come at the end of the result and it may be the case that it is not so essential you simply break that. Break and start your next operation.

(Refer Slide Time: 34:10)



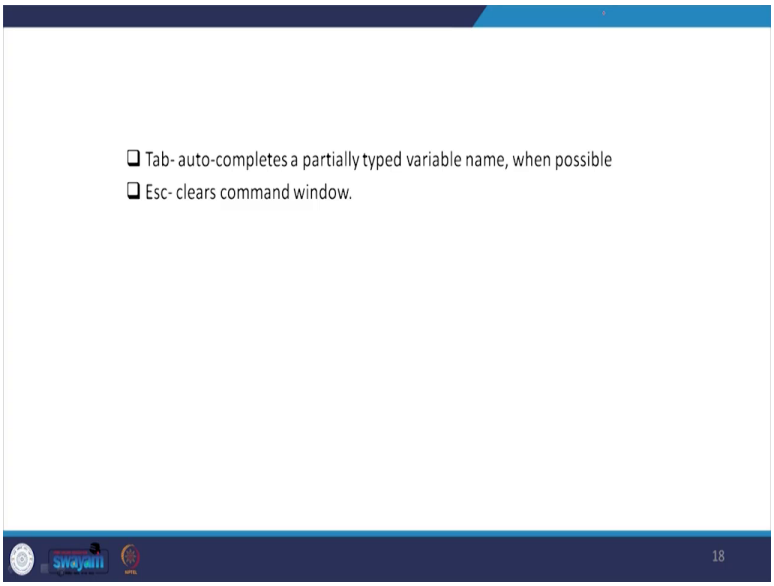


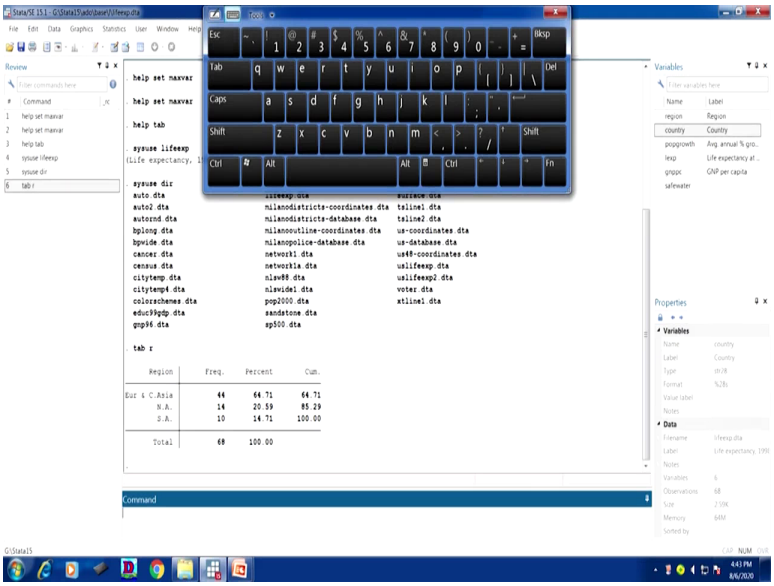
Let me tell you the command window. Command we have already shown you. Here is the command at the bottom. I am not spending more time on it. This command has so many options. I will operate. So I am not spending more time on it, because without command it is not that, not possible to operate data.

Otherwise, even it is possible there are some dropdown menus. If you click on it, also it is still possible, but better to go by the command based operation. It is faster. So it supposed copying, pasting command, going for the command history, functions. Similarly, page up from the command window, pressing page up steps backward through the command history, page own like I will operate, when I operate, I will tell you. You just remember this.

I am going to tell you in detail that if you just page up, there in the keyboard, these options are there. Since this is a soft keyboard if you open, let me tell you in this particular one, it is not there since I am using the digital version of it, and in your keyboard page up is there. If you click the page up, it will pop-up with the previous command and page down it will give you this next command. You just open your keyboard; you will find out. Next to all those arrows, you will get those commands.

(Refer Slide Time: 35:49)

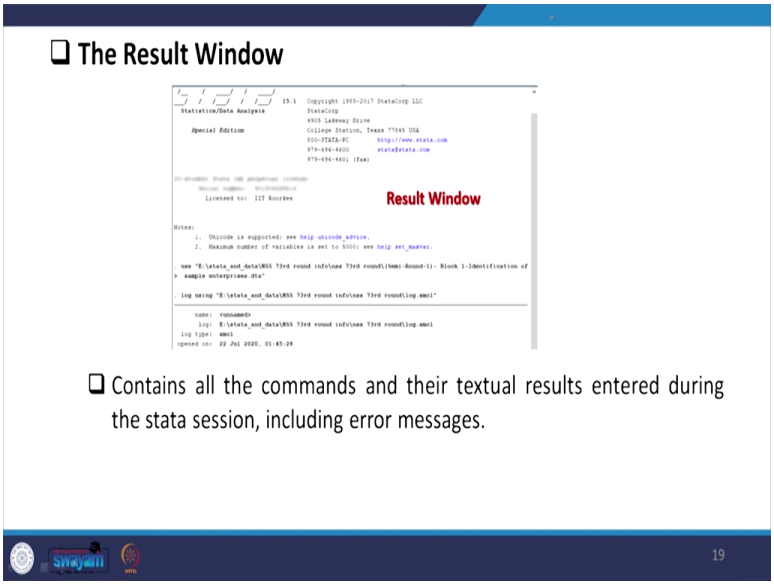




Similarly, tab-auto completes a partially typed variable name when possible. Similarly, if you have already entered something, some wrong command like wrong command you have already entered there and you want to escape that, you would want to get rid of instead of typing or erasing all, simply press escape in your keyboard it would automatically escape.

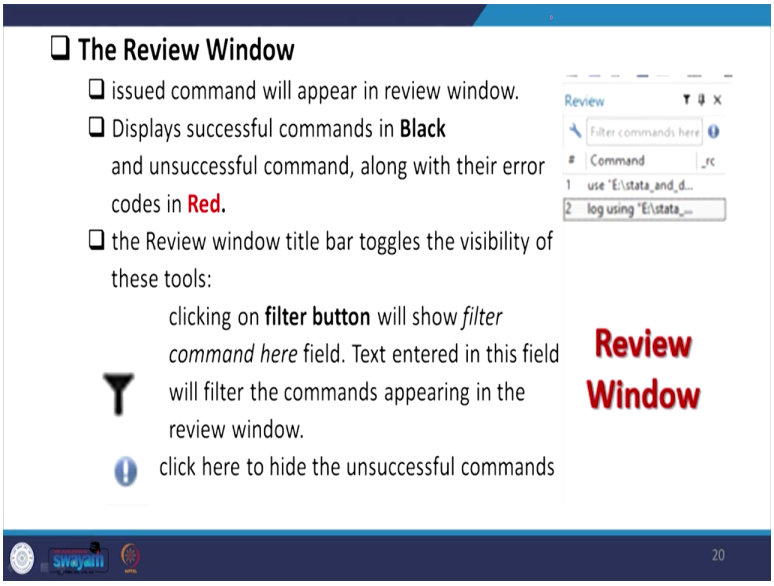
Like this, if you have written something wrong Esc, go for it. Like escape is there, but in the keyboard, there are some problems. If I simply click escape, so like you have written something you want to get rid of that particular command and you simply press escape it will erase all the command. And so that is how it is important.

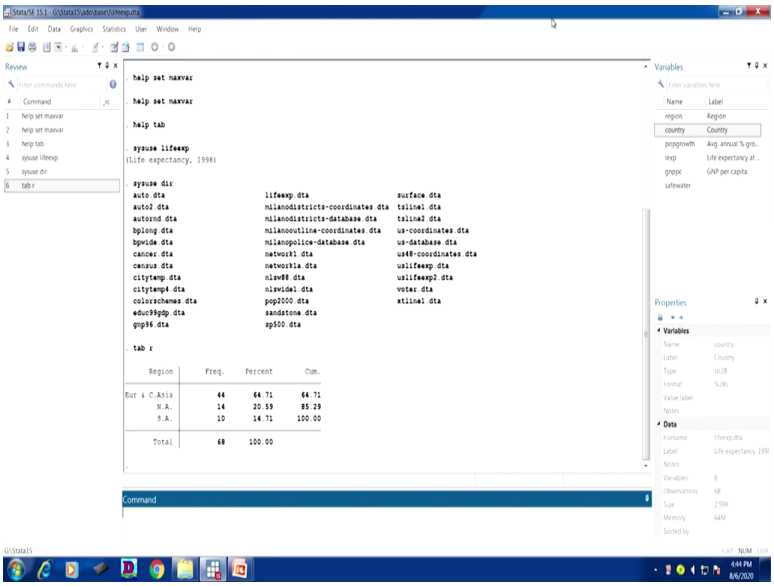
(Refer Slide Time: 37:01)



In the result window, this is the result window. We have not yet explained. It contains all the command information that has already been entered. And there are textual results entered during the Stata session, including error messages. For example, any error, wrong command you have entered, the error message will also be displayed in the command. We will show you all those details later on.

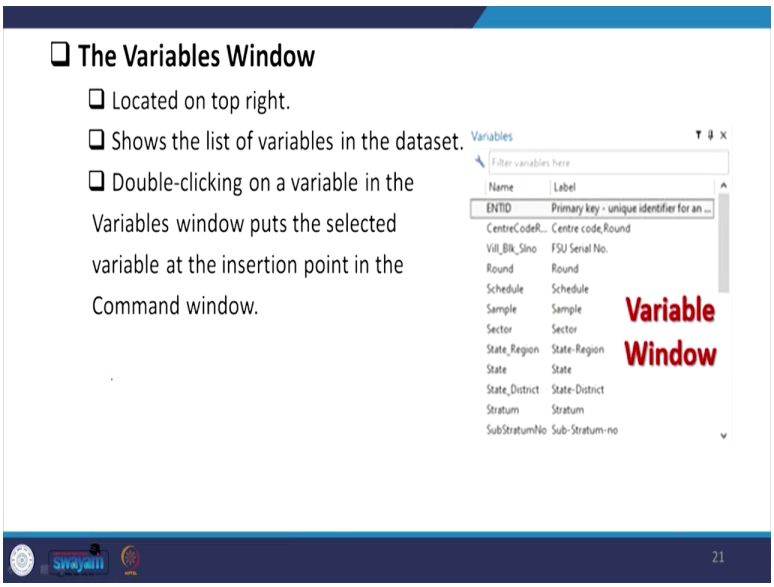
(Refer Slide Time: 37:30)





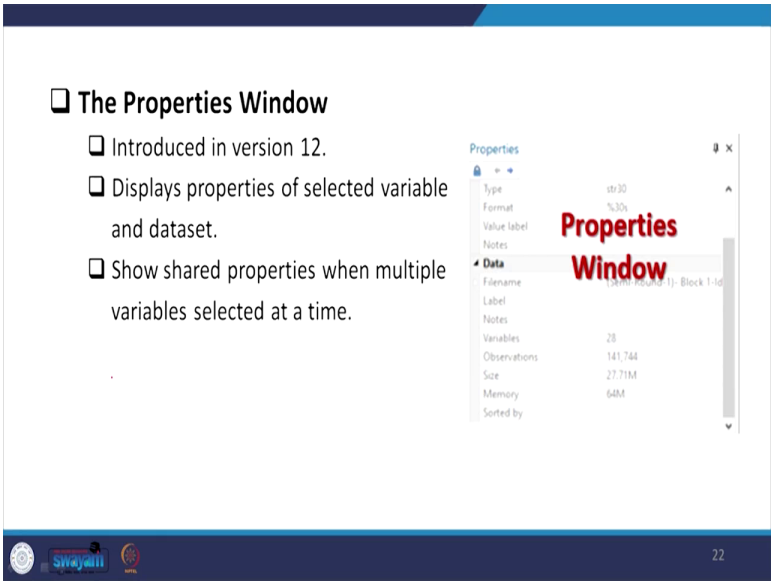
Similarly, the review command, review command I have already shown you at the left hand side of the screen here on the left, on the very left side of the command. For example, you have given any wrong command here, it will highlight with a red color. You will operate on your own what I said, otherwise all those correct reviews are highlighted in black color. If you want to filter further, these filter buttons are there. I have already shown you with its symbol and those filter can be also saved for further use. We will explain all those things later.

(Refer Slide Time: 38:14)



Variable window at the right side is being displayed. So many variables are there, but for our inbuilt data from the StataCorp we have taken that contains four or five variables. We are now showing a different variable source. So, you can get it further details while we are in operation.

(Refer Slide Time: 38:42)



And similarly, the properties window, it gives switch kind of variabilities, whether it is a string variable, whether it is a numeric variable or not, all those details will be displayed. This is particularly introduced in version 12 onwards and so. If you are clicking on that particular variable, it will show you the variable details. But if you have clicked so many variables, it will give a shared information.

Since we have discussed the background of Stata, we have shown you different buttons of Stata, different points of Stata, some basic operation of Stata we did, like I do remember that we are going to explain you further on log files, we are going to explain you further on do-files. Do-files was not opening because our data was not clearly loaded in the system. So in the next one we are going to show you with those operations.

With this, let me close here. We look forward to the next class. Thank you.