Foundations of R Software
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Data Handling
Lecture - 46
Data Handling - Importing and reading EXCEL and other data files

Hello friend welcome to the course Foundations of R Software and you can recall that in the last lecture we had studied about the topic that how you can import different types of data files in the R Software and how you can do different types of operation. So, in the last lecture we had learned that how you can import the CSV and tabular data files in the R software and for that you had learned a couple of options also that how the R can read them correctly and can bring them inside the R software

So, now in this lecture today we are going to learn one more type of file which we want to import in the R software and this file is the spreadsheet which is created in the Microsoft Excel software. So, you know that Microsoft excel is a software in which these spreadsheets can be created and the files can be stored in TXT or CSV and various other formats.

So, now in this lecture our very modest objective is that how to read those excel files in the inside the R software. Just for your information means earlier in this Microsoft excel there were two options to save the file that is dot XLS and another was dot XLSX. And when you are trying to read the excel file in the R software earlier they are used to be different packages and with the help of those packages we used to read the file in the R software.

But now with the latest development now there is only one package which is used to read both types of files which are created in the excel file and yeah these things may change actually. For example, this TXT and CSV file formats they are the they have a built in facility inside the R software.

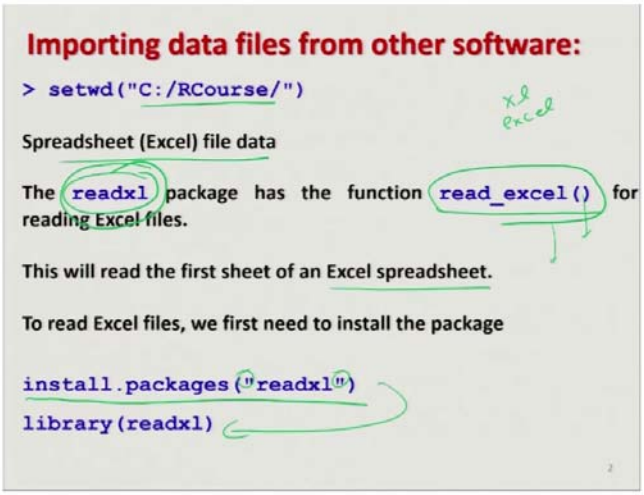
So, if the R software gets updated these commands will be there, but in case if something else is happening you have to be very careful when you are trying to read such files from some external sources and you have to always cross check whether the package which you use earlier is it continuing or not, right. So, now with this objective we begin our

lecture and we try to understand how we can import a data file which is created in the MS excel software.

And then you have to just keep in mind that we are not going to do something new today. We already have learned that what are the different options which can be used in the in reading such files. So, the similar type of options we are going to use here and I will try to keep the discussion at an elementary level so that you can understand.

And after that there are many other options which are available which can be used to read the excel files. I would request you that you please try to look into the help menu and try to understand those option. So, let us begin our lecture, right ok.

(Refer Slide Time: 03:36)



So, as we discussed in the last lecture we had set up our working directory on C colon as a folder RCourse and the same thing we will continue in this lecture also. And you also need to first set your working directory where you are going to store your file, because in order to explain you as I did in the last lecture today also I have created a very small excel file from which I will try to read the data.

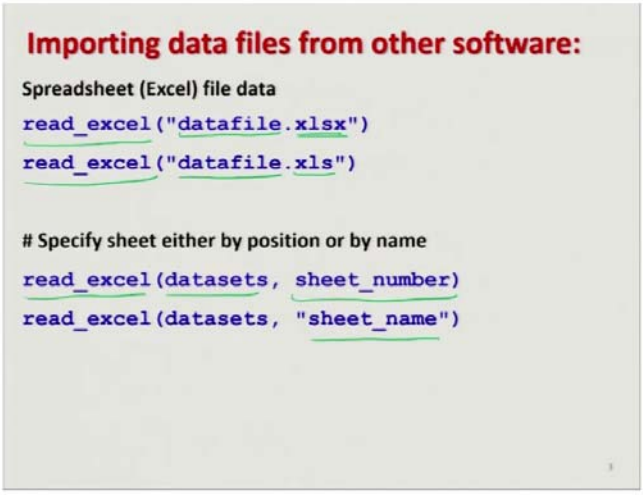
Before we move forward let me inform you that in order to read the spreadsheet from the excel software we need to install a special package which is readxl, all in lower case alphabets. So, this is the package and after that we will try to use the command read underscore excel.

Now, you have to be very careful in my pronunciation because when I am trying to say excel whether this is xl or this is e x c e l that is difficult to discriminate. So, you have to just focus on the slide set that where I am trying to mark my pen. So, I will repeat once again. We need the readxl package to the spreadsheet which are created inside the Microsoft Excel software and for that we need a command here read underscore e x c e l and inside the parenthesis we will try to give the file name and some options, right.

So, when you are trying to read it and you know that in the excel spreadsheet sheets are within the same file you can have more than one sheets, right, those sheets can be even a number they can be identified as an index or they can also be given some name. So, when you are trying to use this command read underscore excel then the first sheet of the spreadsheet is going to be read, right and after that you have to give some option give some command. So, that you can read a particular excel file, right so ok.

So, first I try to do that I would rather request you that you please try to install this package by the command install i n s t a l l dot packages p a c k a g e s and within double quotes you try to write down here read xl readxl and after installing it you just load it using the command library l i b r a r y and inside the parenthesis readxl, yeah. I already have done it on my computer; so, right ok.

(Refer Slide Time: 06:02)



.

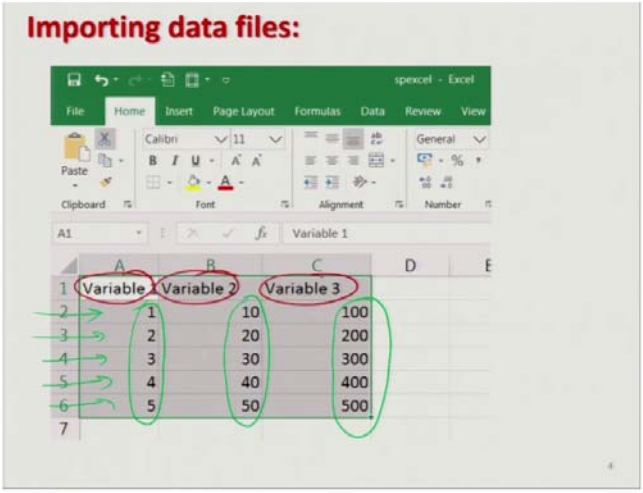
Suppose there is a file whose name is datafile and excel software there are two options that you can have a file with the name xlsx as an extension and another file name as

datafile with an extension dot xls. In order to read both types of file we have the same command here read underscore e x c e l, right that you have to keep in mind.

And in case if you want to read any particular sheet number or sheet name that also can be given here by the command that you try to use the same command read underscore excel and give here the file name. And then after that you have to use if you are trying to use the sheet number you have to give the command here a sheet underscore number and in case if you are trying to read the name of the sheet, then you have to write within double quotes sheet underscore name that whatever is the number or whatever is the name, right.

So, I will try to show you this is an example so that you can understand it more clearly ok.

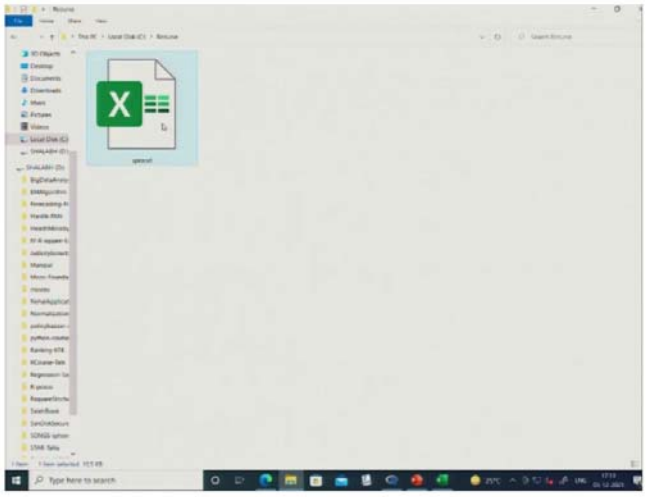
(Refer Slide Time: 07:12)



So, now if you try to see I have created a file here whose name is s p e x c e l and this is the same file similar type of file which I used in the case of psv and text while that it has got say this 5 values here and there are 3 columns and the values are here like 1 2 3 4 5 in the first column; 10 20 30 40 50 in the second column; 100 200 300 400 and 500 in the third column and the first mean there are rows and in which the data is like here 1 2 3 4 5 and then 10 20 30 40 50 and so on.

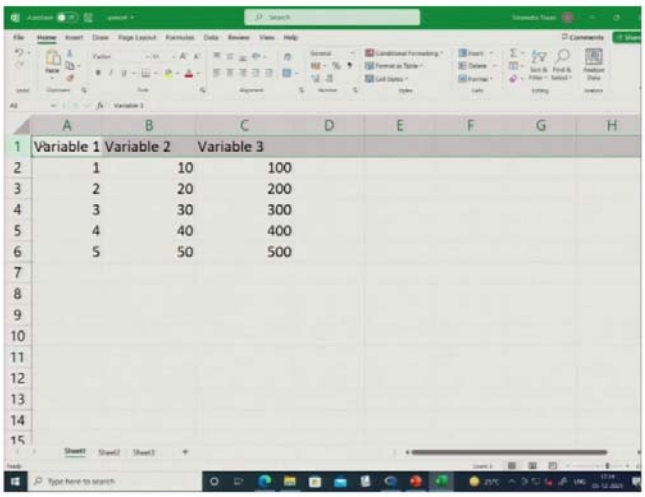
And now I have done here one more thing that I have given here are name to these columns which are here Variable 1, Variable 2 and Variable 3, right.

(Refer Slide Time: 08:05)



And just to show you I mean I can show you this file is located on my this Rcourse folder you can see it this is here.

(Refer Slide Time: 08:10)



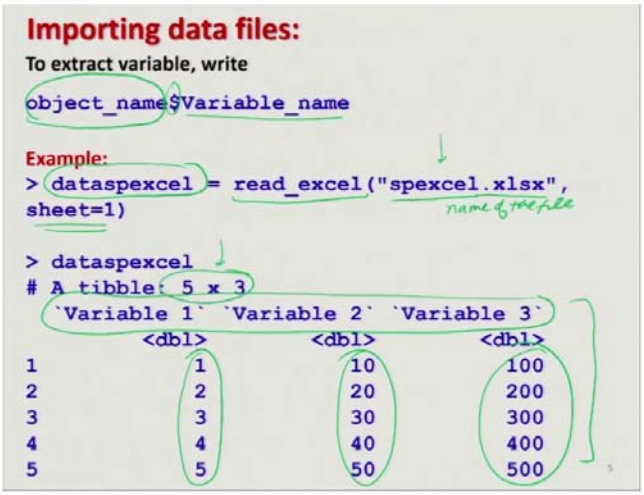
.

And if you try to open it this will look like here this, right. So, I have just taken a screenshot to explain you. So, you can see here this is my first row where I have written

Variable 1, Variable 2, Variable 3 and I want to indicate in the R software that this is going to be the header, right.

Do you remember that in the last lecture we had considered this option that whether there is a header in the file or not and then this is my first column, this is second column, this is my third column. So, anyway we come to our slide and try to understand it from here.

(Refer Slide Time: 08:40)

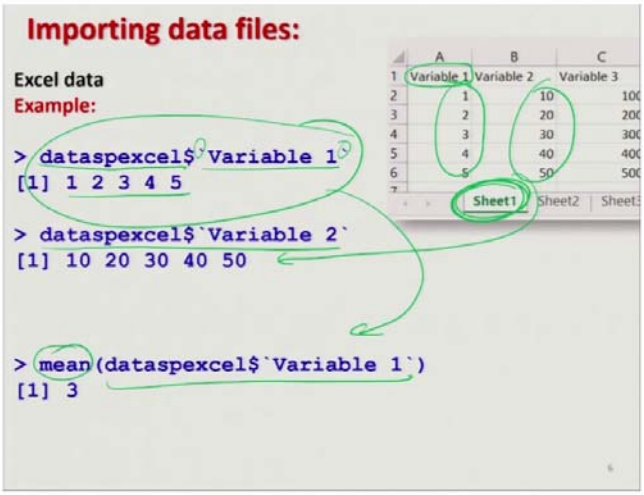


So, and yeah after that I had explained you in the last lecture also, but that means I will just tell you here once again that once you read the file after that you have to use the standard commands to do the usual operation. For example, if you want to access a particular variable from this file, you have to just use the same approach that object name try to write in which you are trying to store the file and then write down the dollar operator and after that you try to write down the variable name whatever you have given in the file, right.

So, ok suppose now in this file what I just shown you I have here only one sheet, right and the name of the file here is spexcel, right. So, after that the extension is dot xlsx. So, I try to use here the command read underscore excel, then the name of the file exactly in the same way as you did earlier and then comma sheet is equal to 1, right. And then I try to store the outcome in an object whose name is dataspexcel, right. So, that is indicating the name that whatever is the data in the file name spexcel that is stored here in the dataspexcel, right.

So, now if you try to see here this type of outcome will come here. So now it will show you here that is the first row here this is here like this; Variable 1, Variable 2, Variable 3 and after that you have here these values in the first column, second column and third column, right. So, that is, right exactly the same way as you did earlier and it is trying to show you here that it is a table of order 5 by 3 means 5 rows and 3 columns, right ok.

(Refer Slide Time: 10:25)



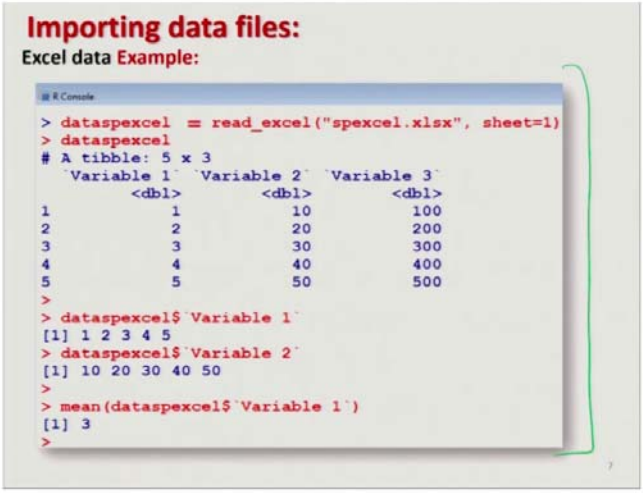
Now, suppose if you want to import the data under the Variable 1. So now, you have to be a little bit watchful that how are you going to do it. If you try to see here in your excel file the variable name was simply v a r i a b l e and then 1, right, but here you have to give here this quotes, right. And so I will write down here the object name here as a dataspexcel dollar and then this variable name and you will see here this is giving you here this 1 2 3 4 5.

Similarly, if you try to find out the values in the Variable 2, similarly if you write down the data object name dataspexcel dollar and within quotes Variable 2. So, you will get here the data which is stored here under 10 20 30 40 50 and so on, right and you can see here this is here sheet number here 1.

So, if the data is in sheet number 2 or 3 you have to accordingly use, right the appropriate number. And now after that if you want to find out the mean of the values which are stored here in this dataspexcel Variable 1 you simply have to write down here

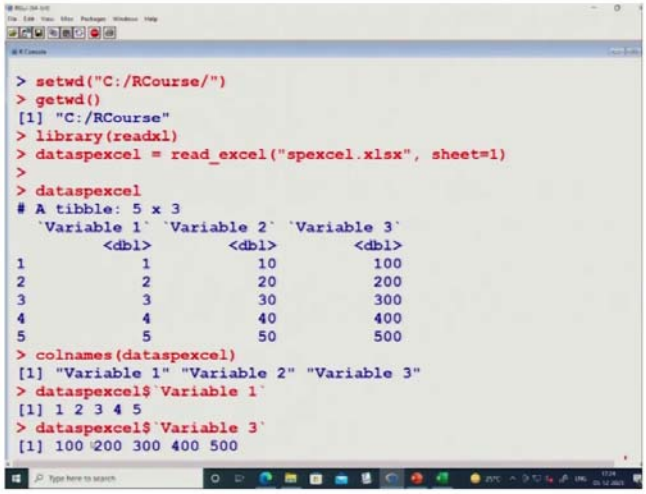
the variable name and then you have to write down here mean and you can do the similar operation.

(Refer Slide Time: 11:44)



And you can see here this is here the screenshot of all these observations, but let me try to show you these things on the R console so that you become more confident here.

(Refer Slide Time: 12:03)

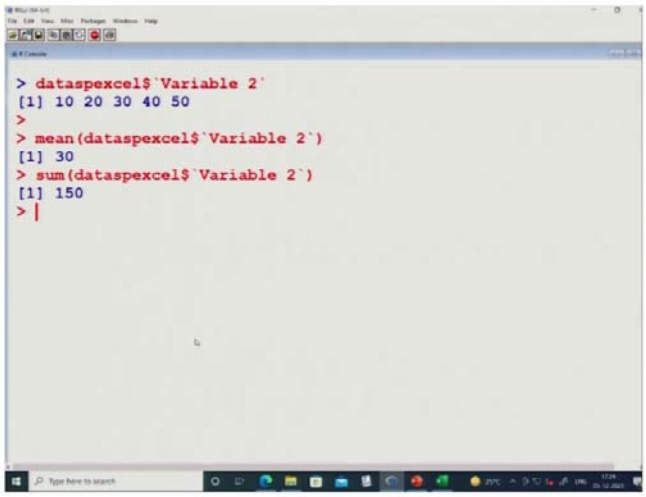


So, ok, so the first thing is so first I need to set my working directory. So, you can see here this is now here like this C colon RCourse and then after that I have to upload the package. So, I already have installed it on my computer, but you need to install it, right.

So, this is uploaded and then I have to give here this file name dataspexcel here and you can see here as soon as I enter this data is stored in the dataspexcel.

So, if you try to see here dataspexcel if you try to write here you can see here this is here the variable like this. And similarly if you try to if you want to really access any particular variable my advice will be once again the same that you try to copy this name from here or you can also use the command like here colnames (Refer Time: 13:01) here dataspexcel and then you can see here these are the values here like this, right.

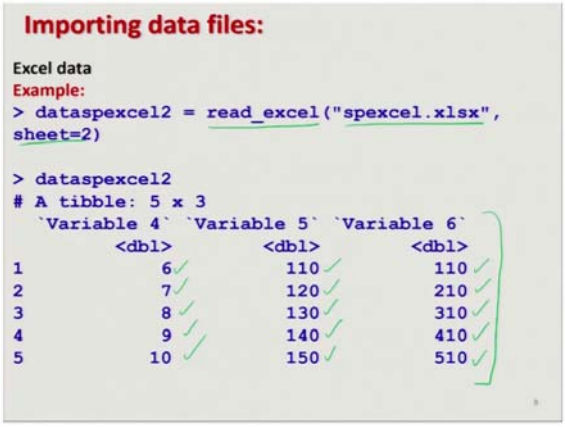
(Refer Slide Time: 13:07)



And in case if you for example, if you are want to see the values in the column number 2 it is here like this and if you really suppose if you want to find out the mean of the values in this data you can simply use it here like this. Similarly, if you wish if you want to find out here the sum of these values which are stored in the variable number 2 you can find it out, this is 150.

So, you can see that it is not a very difficult thing to get all these operation done just by extracting a particular variable from here, right.

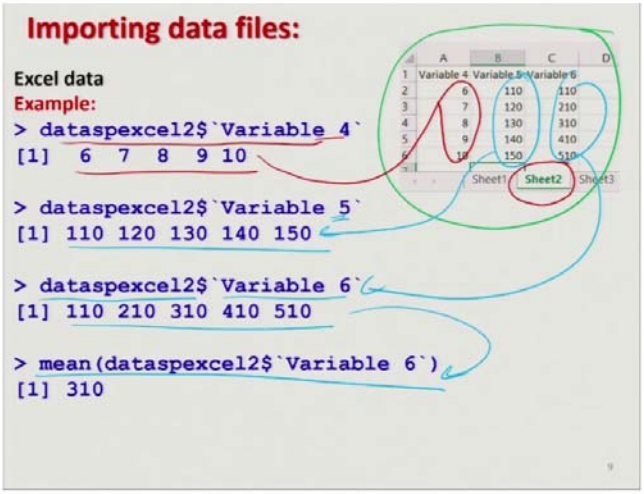
(Refer Slide Time: 13:42)



So, now I try to give you here one thing here more that suppose if I try to create here sheet number here 2 in which there are some values here and if you want to read these values from the sheet number 2 what we have to give here that is only that read underscore excel give the file name and then try to write down here sheet equal to 2.

So, in this case you will get here the values like here. This one you can see here this is here 6 7 8 9 10; 10 values are here. The second column here are 110, 120, 130, 140, 150 and yeah in the and in the column number 6 they are here like this one 110, 210, 310, 410, 510, right.

(Refer Slide Time: 14:29)

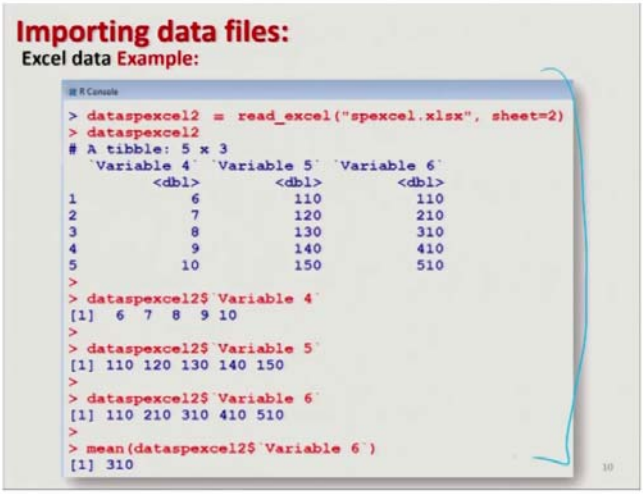


So, well now you know that it is not a very difficult job here and similarly if you try to see here this is the screenshot of this sheet number here 2. So, you can see here where very clearly here that it is here sheet number 2 which have these values. So, if you try to use here the same operation here I mean the data object name and then the variable name like as here 4 if you give it here it is giving you the same observations here.

And similarly if you try to take here the second column here of 110, 120 etc. that you can access by this by the same command just by replacing the variable number 4 to variable number 5 and similarly you can access the third variable here by this command, the name of the data object and then variable 6 inside the quotes.

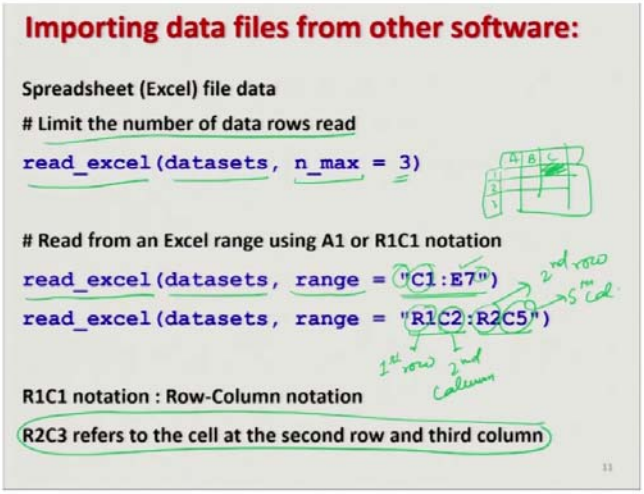
So, you have here this thing and then if you want to find out the mean of this value just try to use here mean of these values. So, these are the very simple operations, right. So, that is what I wanted to show you.

(Refer Slide Time: 15:23)



And you can see here this is the screenshot of the same operation, right.

(Refer Slide Time: 15:29)



So, now after this I tried to show you some more commands and can I will try to come back to the R software, right. These are pretty simple commands. Now, in case if you want to limit the number of data rows it to be read, for example, you want that ok, you just want to read only the means at the most three rows, then you have a command here try to read down you the same command read underscore excel name of the file and then use hidden option n underscore max is equal to for example, 3, right.

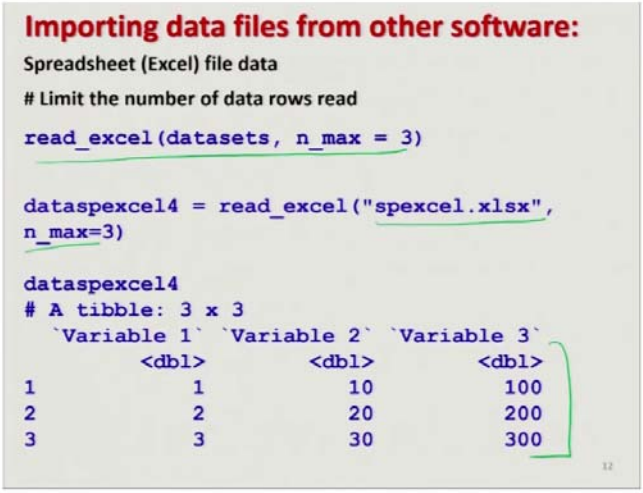
And then similarly if you want to read only a particular range of this data file, then what you can do here? Then you can use the same command, but then you have to use a particular option. So, the command here is the same read underscore excel the name of the data file and now you have to give it here range r a n g e all in lower case alphabets and now you have to give here this type of address within the double quotes.

So, this is the address of the cell. Means you know in the MS Excel you will have here row number 1 2 3 and then here you have here column A B C etc. So, it is giving you here the address of this column or the cell C1 which is the say this column C and row number here 1 like this.

So, you can see here from C1 to E7 you want to read the range or the second option is like is here row column notation like as here from R1C2 to R2C5; that means, R1 is like as here 1st row and C2 is here the 2nd column and similarly this here R2 is the 2nd row and C5 is the 5th column, right. So, you can give this here range and this is how actually

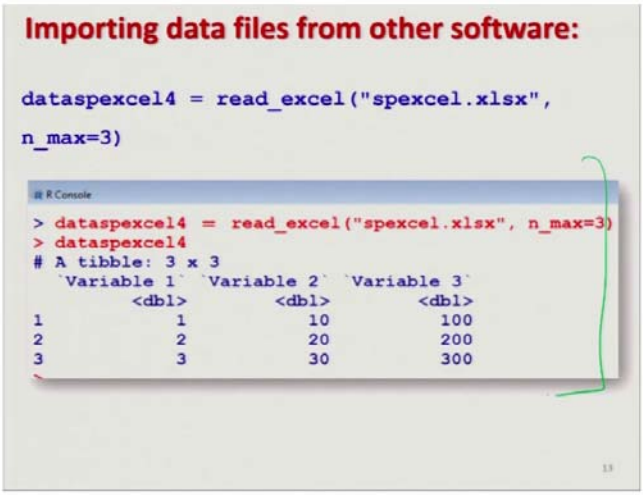
it means when I try to write an R2C3; that means, the cell at the 2nd row and the 3rd column. So, this is the way it tries to read it.

(Refer Slide Time: 17:41)



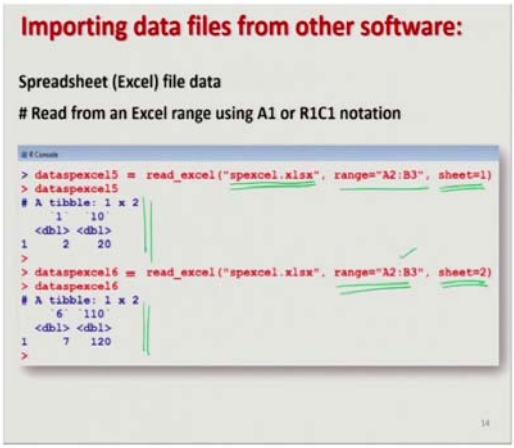
And then yeah you can do it very easily. For example, if I try to say here read underscore excel dataset. Whatever is this if I want to read the number of row is equal to here 3 and I try to use here the file name spexcel dot xlsx and I try to give here the option n underscore m a all in lower case alphabet equal to 3 then it will try to read only the first three rows of this data file, right.

(Refer Slide Time: 18:09)



And you can see here this is the screenshot of the same operation, right.

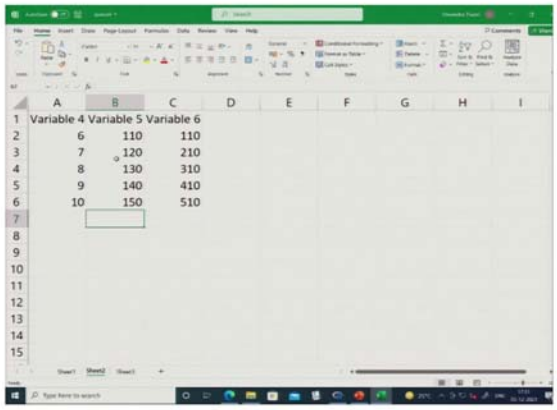
(Refer Slide Time: 18:12)



And similarly if you want to read here say here any particular file any particular range of the data set in the file, so, you can see here I am trying to give you here this file name and then range is equal to A2 colon B3 sheet number 1. Or if you want to give the range in the sheet number here 2, so, you can write down the range here within double quotes A2 colon B3 and sheet number here 2 and so on and you can see here these values are obtained here, right.

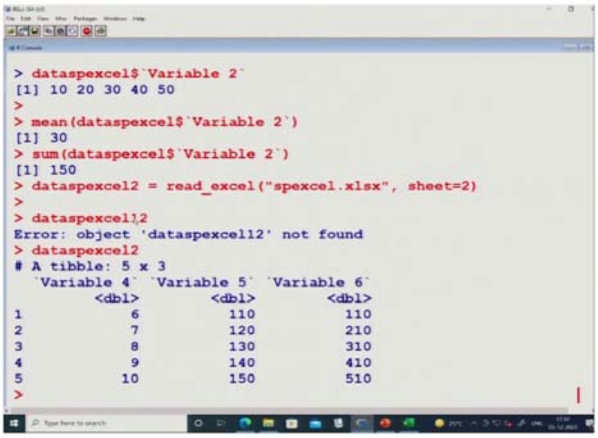
So, let me try to show you these operations first on the R console and then I will try to move ahead ok. So, now, let me try to read the data from the sheet number 2.

(Refer Slide Time: 19:02)



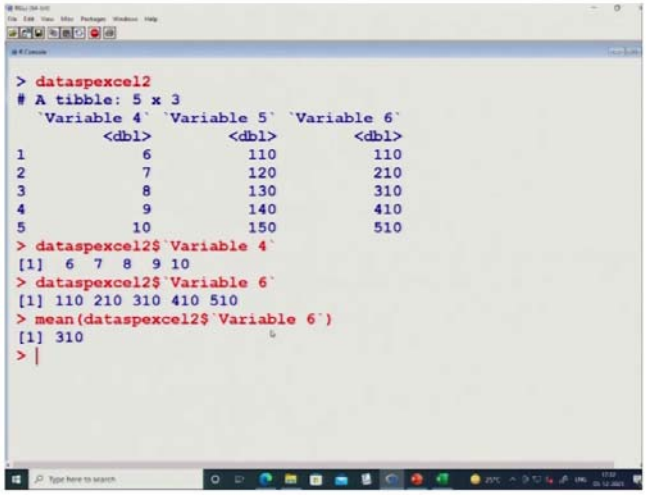
And I will try to show you here that in the sheet number 2 here this is here the data I can read, you can see here. This is in the this is a sheet number 2 you can see here in the bottom where I am trying to move my cursor and sheet number 1 was like this, right.

(Refer Slide Time: 19:23)



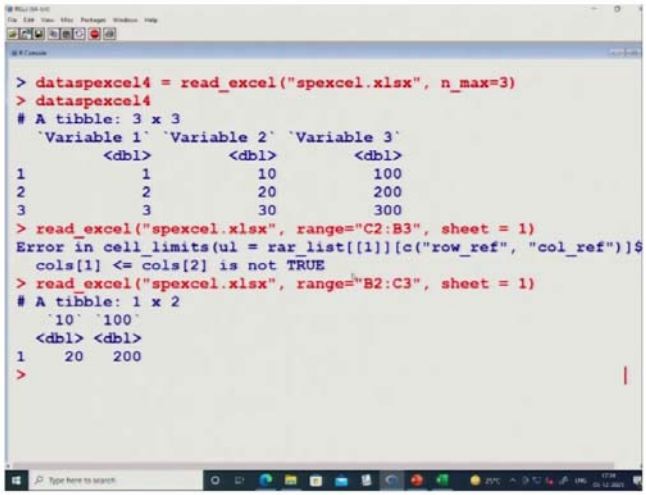
So, now you can see here this data is there. So, now, I simply have to use this command here like that and then it will try to see here this is dataspexcel12 and you can see here; there is 12 here like this, you can see here like this, right. So, this data is here actually, right. Here I had given wrong name. So, you have to be very careful that when you are trying to give a wrong name it will not read it ok.

(Refer Slide Time: 19:52)



Now, if you try to read a particular value from this variable you can simply write down here you can simply copy this variable name here and then you try to write down dollar and then here variable name. You can see here this is here and similarly if you want to write down the value and the variable 6 these are here and if you try to find out the mean of these values you can very easily find out exactly in the same way, right, ok.

(Refer Slide Time: 20:18)

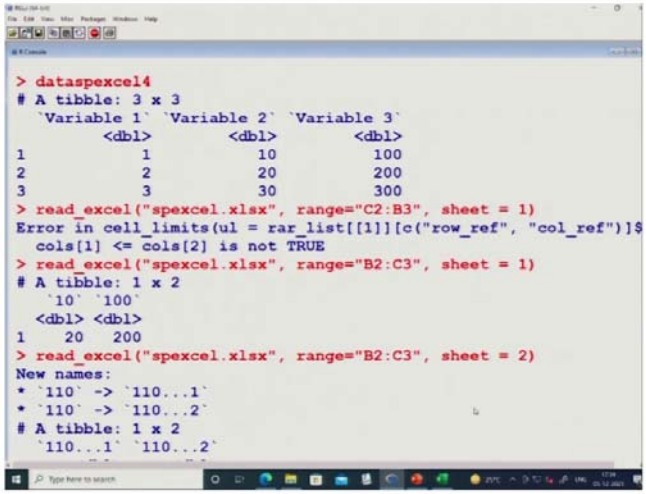


So, now after this I come to your here next execution. I am trying to read here this number of rows. So, you can see here I am trying to read down here like this say n max equal to 3 and you can see here this is here like this and if you try to see here dataspexcel4 this is like this. So, you can see here it is reading only the first three rows in the data vector, right.

And in case if you try to obtain here quickest part here that you want to write down here range is equal to A2 colon B3. So, you can use here for example, the same thing here and if I try to show you here I simply try to modify my earlier command and I try to write down here range is equal to so, within double quotes C2 to B3 and if you try to see here and then I try to write down here sheet number, number 1.

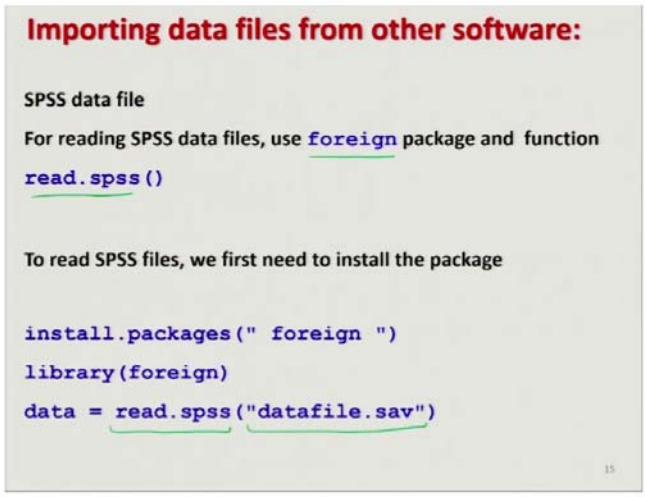
You can see here this is error. Why if you try to see? Because you are trying to give the range here C2 to B3 in the opposite direction, but if you try to make it here B2 to C3 then it will make sense, yeah. If you try to see here, now it will read.

(Refer Slide Time: 21:41)



And if you try to read it from the sheet number here 2, you can see here it gives you here these values here, right like this, ok anyway.

(Refer Slide Time: 21:59)



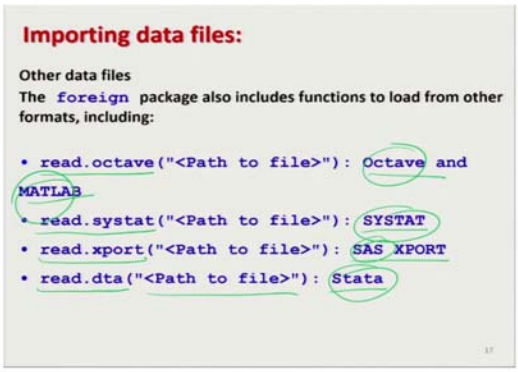
So, now let us come back to our slides and try to understand that what are the other things. So, now, you see once you have learned how to read excel file, how to read txt file, how to read means csv file now there is no end. So, now, I will very quickly give you that how you can read the data files which are generated another software.

For example, so, there is one package here foreign f o r e i g n and you will need to install this package if you want to read the files which are generated in the SPSS software, right then after that you have to use the command here read dot spss and then whatever is the file name datafile dot sav you have to give and then it will read it and it will work exactly in the same way as you have done with the xls file.

(Refer Slide Time: 22:46)

And similarly if you want to read the HTML tables those data files which are in the form of HTML data files. So, you have to use here a package XML and after that you have to use the command read this HTMLTable, but you have to see here this r e a d they are in lower case HTML and T they are in upper case and a b l e that is in lower case, right. So, you try to install it and after that the command is similar, try to write down this command and then within double quotes the file name, right.

(Refer Slide Time: 23:21)



And similarly if you want to read here some here Octave and MATLAB file then you then your command is read dot octave. If you want to read the files which are generated in the SYSTAT software then the command is read dot systat. If you want to read the file from which are generated in the SAS XPORT software then the command here read r e a d dot xport.

And then if you want to read the data files which are generated in the Stata software then the command here is read dot dta, after that you have to give the file name. So, actually this Octave, MATLAB, SYSTAT, SAS Stata etc. these are different software which are used in statistics. So, people are trying to get the data from there and from there they will try to use these files over here.

Now, we come to an end to this lecture and then I will also stop with this data handling and now I am pretty confident that you can understand very easily how you can means read different type of data files which are generated from some external software. But now you know now more responsibility is coming towards your head as we are moving further because now you have learned how to read different types of file.

And then after that now you can read any data any value and now you know lots of data handling functions and commands. So, now you are more mature more experienced more and you have done lot of practice. So, now, if you try to experiment those thing what you were doing earlier before learning the R software with the data files and try to do them in the R software possibly that will make you more confident and a better programmer.

So, you try to take some example, try to practice it and I will see you in the next lecture with some more commands on the data handling, till then goodbye.