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

Lecture - 01
Why R and Installation Procedure

Hello friends, welcome to the course Foundations of R Software and from this lecture we are going to formally start the lecture for the learning of R software. But, before going forward let me try to clarify one thing that in 2017, I had floated a similar course Introduction to R Software and now over a period of time many things have changed. So, this course is practically an updation and revision of the earlier course introduction to R software.

So, you may find many things, many topics which are common and as I said in the earlier lecture that this is a fundamental course, this is an introductory course. So, when you are trying to teach the introductory part you are going to teach the basic fundamental things. So, they are going to remain more or less the same, but definitely I have tried my best to add up the new things which have been developed and which are important and necessary for the understanding of the basic fundamentals of the R software.

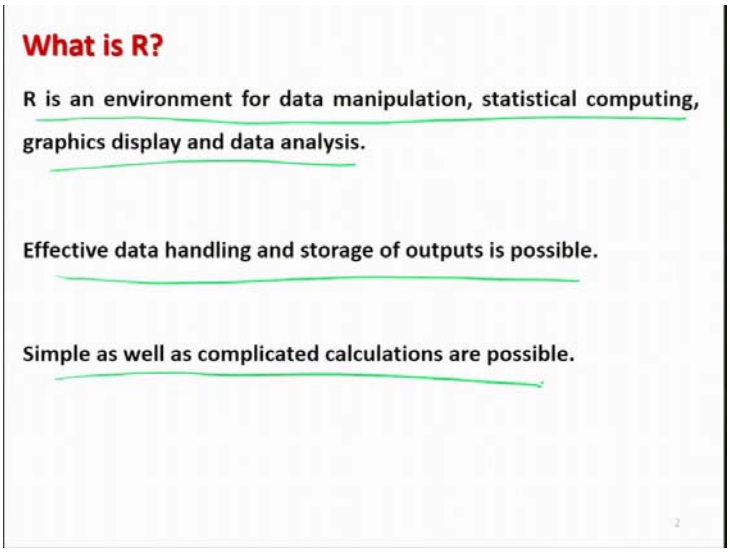
So, with this introduction and with this information I would like to now start this lecture. So, whenever you are going to learn any topic, the first question which comes to our mind is why, why should I learn this topic? And, then because of curiosity you would like to know the background that how it happened, why it happened and many questions crop up.

For example, this is the free software, who developed it, why developed it, why it is free and what are the advantages and disadvantages compared to any other software and is it going to be a good software or we are going to lack something etc. etc.

So, in this lecture I am just going to address all these basic queries which comes to a human mind whenever we are trying to learn something new. And, then the next question comes from where I am going to get it and then how I am going to begin the learning of this software. So, that is my another thing. So, this lecture is going to be very elementary, a story telling type lecture and then I will try to show you how you can

download the software, from where you can download the software, how are you going to install it, so that when we come to the next lecture, we are ready to learn the fundamentals and foundations of R software. So, let us begin our lecture. So, in this lecture we are going to talk about that why should I learn R and how are we going to install the software on my computer for the learning of the fundamentals and foundations.

(Refer Slide Time: 03:08)



So, I am sure that you must have learned about the software. Now, this is the time where the students are trying to learn about the software, computer programming etc. right from the childhood. They are taught about this computers, programming language etc. in their schools starting from elementary classes.

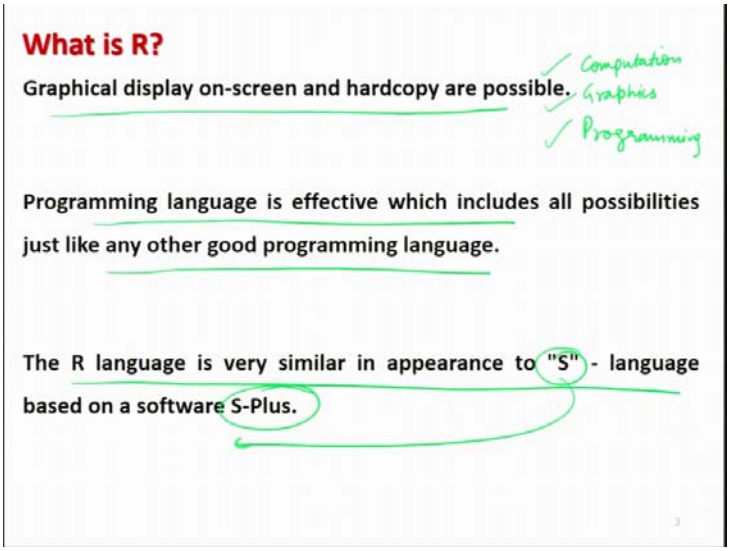
So, this is not a question that what is R software, as you can understand very easily from its terminology that R is simply a software, just like you have used many other software. So, this is also a software, but the main question is what it is doing and for what it is going to be used?

So, in that case I can explain you in very simple language that is R is an environment for data manipulation, statistical computing, graphic display, data analysis and various type of computations. When we are trying to talk about the software, there are some software which are used for typing, there are some software which are used for mathematical

typing, chemical typing, chemical symbols are typed in a different way. There are some software which are trying to do some statistical analysis, there are some software which are trying to do some mathematical analysis etc.

So, similarly R also has got a role and in this R software effective data handling and the storage of output is possible. And, when we are talking about the calculations, the simple as well as the complicated calculations both are possible without any problem.

(Refer Slide Time: 04:40)



And, whenever we are talking of any statistical software or any mathematical software usually there are couple of things in which we are interested. The first thing is the computation that what type of computation, what different varieties of computation the software can perform.

What about the graphics? Graphics means can the software create different types of graphical display and when we are trying to create such graphics whether the software can also give the soft copy or the hard copy, is it also possible to save them. And, after that what about the programming? Programming means there are two types of software: one software which are just based on some click, click, click type of buttons that you go to those buttons, you click there and the execution is done.

And, second type of software are where you have to go to some prompt and you have to execute the command, you have to type it and then the things are done. So, what happens

in R? So, now, I can assure you that in R all the things are possible, you can execute different types of computations, you can execute different types of graphics and you can also do the programming. And, when we are working in the R the graphical display which is on your screen as well as the hardcopy, both are possible.

And, as I said that in R software the programming is also possible and the programming language is quite effective and this includes all sorts of possibilities which are possible in any other good programming language. And, now a very interesting story that why this R came into existence, what really happened, what motivated the people to develop this R software. So, I am the witness of that era and I have seen the birth of this R software and the reasons why possibly R software was developed.

Well, when I was doing my undergraduate and post graduation at that was the time when the computers started coming to India. Yes, that looks, that appears to be a very old story for all the young candidates, but that was the truth. So, when these software started coming to India and computers coming came to India then at that time there was only prompt means command prompt, DOS what you call, there was no Windows.

So, even if you want to copy a file mean the way you do it today that you take your mouse and you say right click and say copy or paste that was not possible. But, we used to use the command like control C to copy the command or control V to copy or paste something or copying from file from one directory to another directory, there used to be command and we were taught that how to write the path etc.

Then came the Windows platform and in Windows, it was possible that instead of writing the typing the commands using the mouse we can simply go to that block that is your window and there I can click and can execute the basic commands, basic functionalities. So, what really happened, for example, if somebody wants to find out the arithmetic mean of some numbers. So, in the in those software which are based on the DOS, someone has to write a program, those who have done the basic programming they might remember that we used to write sum is equal to 0, num is equal to 0, sum is equal to sum plus num etc. But, when the windows started then people wrote those commands and it was possible that I enter the data and then I press on the button mean and or say sum then automatically I will get an output.

But, the disadvantage was that suppose if I want to find out the square of the sum or square of the mean then in Window based software where there was no provision of programming, then it was not possible whereas, in case if you go back to the DOS based software then it was possible. And, then there ware some software which started coming to the market, they were usually the paid software for which you have to pay and they were quite expensive.

And, during that time among various software there was one software which became very popular that was S Plus, capital S and then P l u s plus that was the name of the software. The reason or say I would say one of the reasons why it became very popular was that it had a sort of hybrid mode means somebody has written a program for finding the mean. So, in case if somebody has got a set of numerical values and if the arithmetic mean has to be found, someone has to all simply write mean m e a n and then within the parenthesis write down the values and it will give you the value of the mean.

And, for that you don't have to write down the entire program like sum is equal to 0, num is equal to 0 etc. And, then there was another possibility that one can also write own program that if somebody wants to write a program for the mean using sum is equal to 0 or num is equal to 0, sum is equal to sum plus num, num is equal to num plus 1 etc., it was also possible.

So, now here was the advantage that in case if I want to simply compute the square of the mean value, I can simply find out the mean and whatever is the value I can simply square it and I do not have to write the entire program and this gained popularity. Many people started working in this software S Plus and S Plus has a different type of style, the assignment operator and some other things were different. The way we used to write down the program that was different in the S Plus software.

So, people started learning this S Plus software and it was very popular, but it was very expensive, very expensive means at least for me I can say it was not possible to pay from my pocket from my salary to buy those things.

So, then there were people like me around the world and when they realized that this is a very important software that is a very good software then couple of people gathered from

different academic communities from all over the world and they thought that ok why not to developed a similar software like S Plus.

And, that was the idea then people started working together and they developed this R software and gradually it was available for the common people to use. But, there is one issue, this was a free software and they decided that they will not charge any money for this. So, now you know we all are human being whenever is given to us free, our human minds ask us that we have to doubt on it.

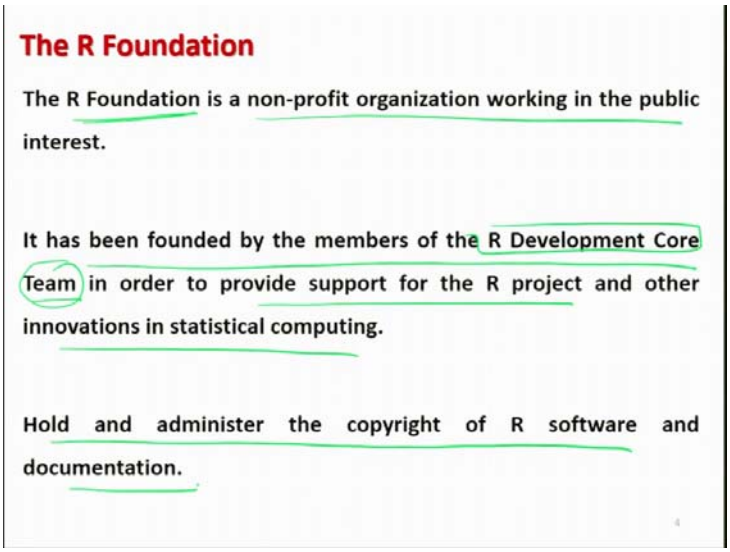
It is just like if your friend comes to you and if that friend ask you ok let me offer you a very good coffee today, expensive coffee today. This is very natural that instead of enjoying that your friend is offering you a wonderful or a good cup of coffee, you start thinking first ok what is happening, why he or she is trying to give you a cup of coffee, what is the reason, what is the intention behind it and the same thing happened with the R software also.

So, initially I saw means I am the witness that when the software was available for using, initially including me, we were hesitant in believing that whether this software is giving us the right calculation right outcome or not. But, gradually over a period of time, R established its authority and people started believing on it and then people also realized the importance of this software and what type of freedom do they have with the software.

And, that is how this software became very popular and then later on very established publishers, they also started publishing or the books on R software and even they started a complete series on the R software like statistics with R, computation with R etc. and this is how this R software became very popular. So, when this S Plus software was introduced then the programming language of that S Plus software was called as S language, capital S language.

So, there is a belief that ok that because of that thing people call the language of the R software which is used for programming as R language. Well, there is another reason that I will try to explain you, but this is the my experience of growing with R. So, let us try to come back to our slides and try to understand the further things. So, this R language is very similar in appearance to S language and this S language was based on the software which was called as S Plus, S P l u s right

(Refer Slide Time: 14:45)



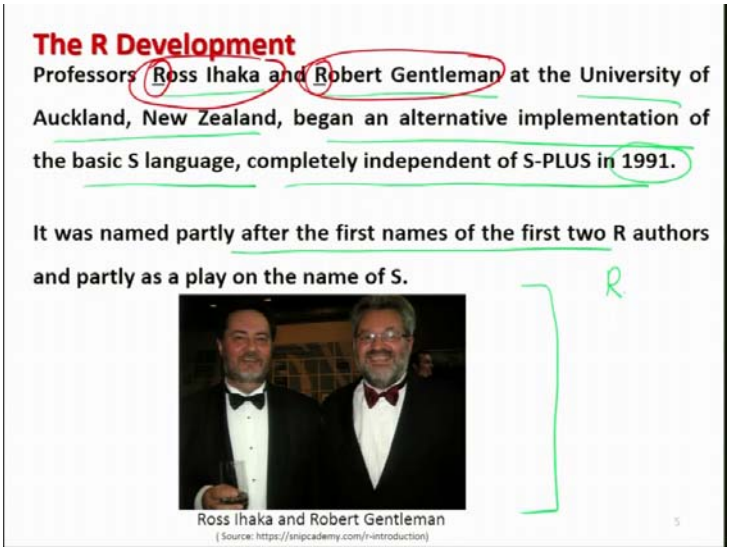
So, now this will give you an idea that how are you going to do it and those people who started working together for the development of the R software, they grouped together and then in order to use this software and to provide a platform the R foundation was created. So, this R foundation is a non-profit organization which is working in the public interest because it is providing you the software for free.

And, this foundation has been founded by the members of the R Development Core Team. What is this R development core team? This is the group of people who are involved in the development of the R software. You see R software is a huge software, that is the very big software and it has different types of possibilities that someone can develop own package and can contribute it and then there are different types of things which have to be done in the software.

So, this group of people from all over the world from this academic community they work together and the group of those members, this is called as R development core team. So, this R development core team provides the support for the R project. R project means well you are trying to develop a big software, huge software. So, that is why this is called as the project and they also try to provide different types of innovation in the statistical computing, mathematical computing.

And, it is not even that now in the last decade this R has developed into different dimensions right and this R foundation holds and administrated the copyright of the R software and its documentation. Well, someone has to take the responsibility that whatever is there is correct or something has been scrutinized right.

(Refer Slide Time: 16:26)



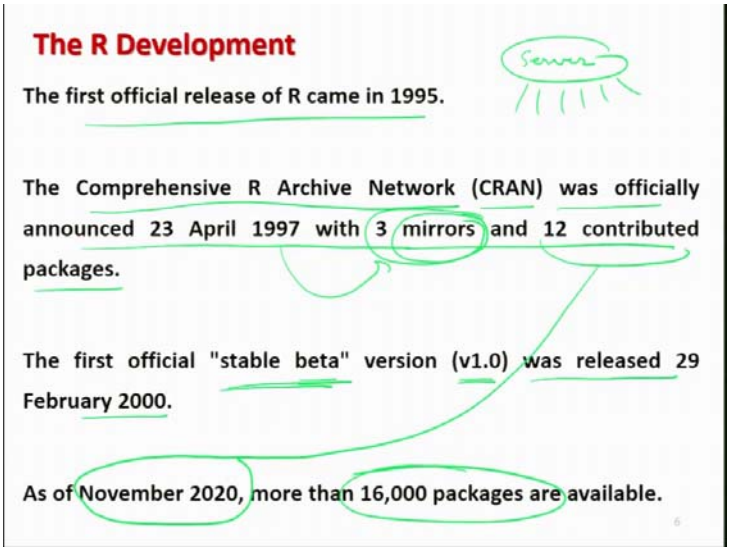
So, just for your information, the one question comes that who started this R software? So, there were two academician at the University of Auckland in New Zealand, their names are Professor Ross Ihaka and Professor Robert Gentleman. Well, I have given here their photograph also, right. So, yeah I got this photograph and I duly acknowledge the help from this website snipcademy dot com, where I got a wonderful photograph of these two academicians and thanks to them for bringing their joint photograph for us.

So, these two professors they came together and they began an alternative implementation of the basic S language which was completely independent of the S plus software and all this begin in 1991. So, as I said 1991 was the time when I was doing my MSc and I was in the 1st year of my MSc program. So, that is what I said that I have witnessed the development of this software.

So, people do say that the name of this software was put as R because, R is the first letter of the name of these two professors. You can see here this is this software was partly named after the first names of the first two R authors and yeah partly it is also because

earlier we had S Plus software and its language was called as S language. So, this was called as R.

(Refer Slide Time: 17:56)



So, these people started working on this project and the first official release of this R software came in 1995. And, then the Comprehensive R Archive Network which is briefly called as CRAN, C R A N that was officially announced on 23rd April 1997 with 3 mirrors and 12 contributed packages. What does this mean?

You see when this R software was developed and then it was to be distributed for the people. So, that they can download and they can use it so; obviously, if this software is uploaded on one server and if many many people are suddenly downloading from the same server then the load of the server will become quite high and there is a possibility of crashing the server.

So, this load was divided and different academic institutions all over the world they agreed that they can also host this software on their website. So, these different websites from different academic institution who in some way provide a copy of this software for downloading they are called as mirrors. So, in 1997, three people possibly agreed to mirror this software.

And, then in R software there is a possibility to contribute the packages. What does this mean? We will try to discuss in the forthcoming lecture, but here I can briefly tell that

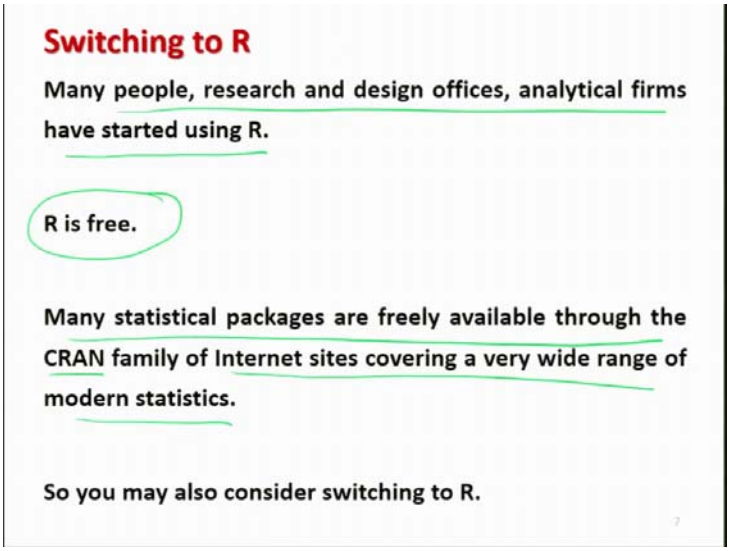
there is a possibility in the R software that if I want to develop some program for doing some specific job then this program can be uploaded on the website of the R software and people can download it and can use it.

So, these are called in layman's language as contributed packages. So, when this R started in 1997 there were 12 contributed packages only. Now, you can guess how many packages are there, I will try to tell you later on and then people started using it. So, we know that whenever the software is introduced this is in a sort of experimental mode. So, this is called as beta version that is the terminology that we use.

So, people started using it, there were some issues, people tried to correct it and then several iterations possibly worked and finally, the first official stable beta version which was version 1.0 was released on 29th of February 2000, right. So, you can see that it took almost a decade to get a stable version of this free software and now if you try to see the based on the data what I have, this R software begin with 12 contributed packages which has now in November 2020 there are more than 16,000 packages which are available.

So, you can see the growth in the last two decades which happened when in the R software; 16,000 packages means you can do 16,000 different things from the same software, right.

(Refer Slide Time: 20:51)



And, then many people, research, offices, design offices, analytical firms, they got motivated, they got confident that ok R is providing us the good result and they started using this software and they started switching to the R software. So, now the other question comes for you that why should you switch to the R software, means I am sure that everybody is using some software, but then why should you come to a new software, why should you learn this new software, what are the advantages?.

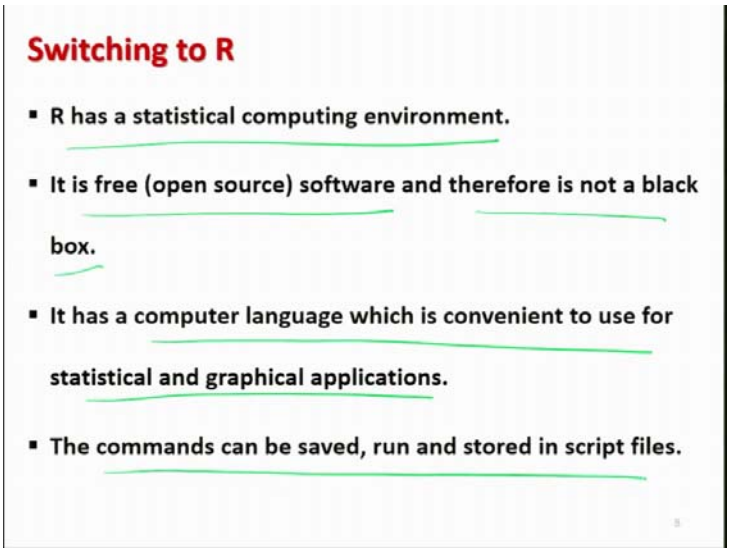
So, let us try to understand this thing. So, the first biggest advantage is that R is a free software. We always listen from different people that ok, that whenever there is a new version of the software they have to buy a new software and then they have to pay once again etc. Many students cannot afford to buy those software, many institutions also cannot afford to buy those software or they cannot afford to pay the recurring cost etc.

Well, I am not talking about those people who are rich, but I am talking about person like me who cannot afford to buy a software from my own salary, right. So, this is the big advantage for a person like us that we are getting the software for free which can do the same thing what others software's are doing and it is giving me the liberty of doing many more things.

And, whenever there is a new version of the software I can get it for free. So, and beside those thing many statistical packages are freely available through this comprehensive R archive network that is the CRAN family. So, they are uploaded on the internet site and they cover a variety and a wide range of the tools which are used in the modern statistics.

Whenever there is a new development in and if you want to use it, you can simply download that package and can use this package for your given set of data. So, you may also consider switching to R, right.

(Refer Slide Time: 22:50)

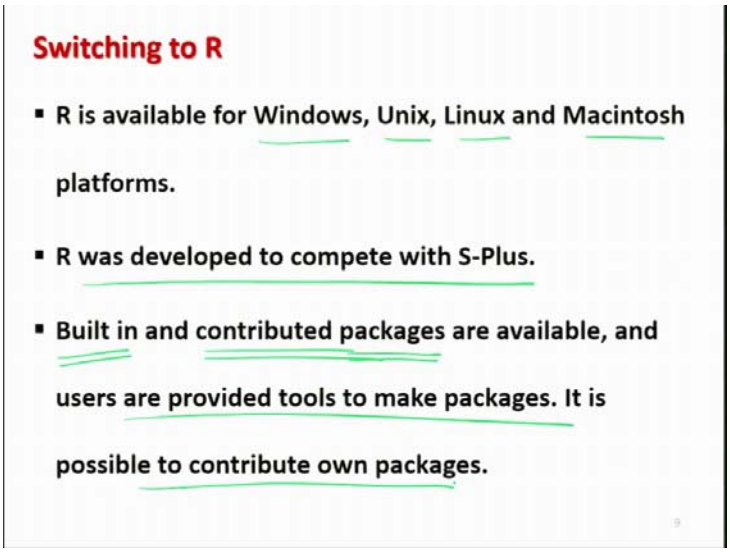


Now, when you take a call that ok you want to switch to R you are not confident, you are hesitant because you do not know what is there inside the R. So, I can assure you here that R has a statistical computing environment. This is a free open source software and therefore, it is not a black box means you can see what is there, how the computation are being done, what type of programming has been done; if you are using any algorithm, you can see that what the algorithm is really doing.

And, it has got a computer programming language which is convenient to use for statistical and graphical applications, without any problem you can do the programming for different types of computation, simulation, calculations and you can also produce wonderful graphics also. And, in case, if you try to see in any standard programming language there is an option that you can write the program, you can save the program and you can store the program and whenever you need it, you can call them and can use them again.

So, all those things are possible in the R software, all the commands can be saved, run and stored in the script files; script file means simple language, it is programming file.

(Refer Slide Time: 23:56)



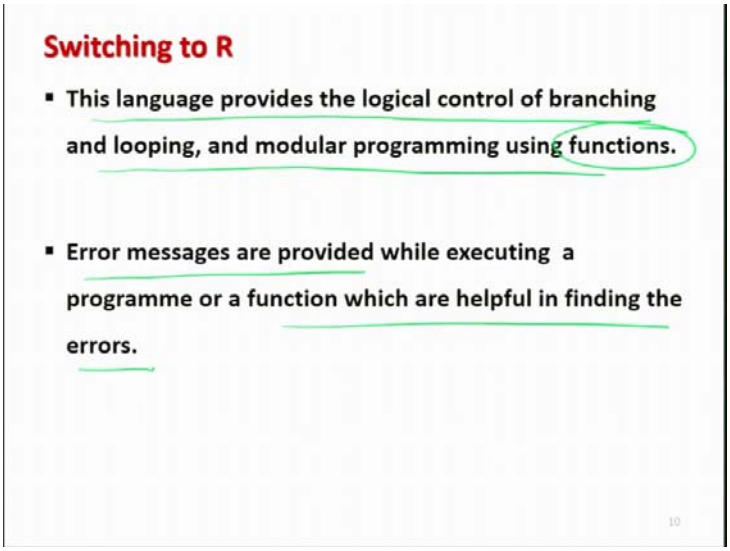
Then there comes another question that people are using different types of operating system like as Windows, Unix, Linux, Macintosh etc. So, this R is available for all such platforms. There may be some small changes in the instruction that how do you define the path in Windows, Unix, Linux etc. But, those people who are working in these environments, they are very much familiar with these things, right.

So, this is not an issue, whatever you want that is available, whatever platform you want to use for this R software, the R software is available for that and R was developed as I said to compete with the S Plus software. So, that I already have explained you and whenever you are trying to do for example, any statistical analysis. So, there are some types of operations which are common which most of the people would like to use and there are some type of operations which everybody may not like to use.

And, then there are some other type of operation because people are doing research; so, they are trying to develop different types of statistical tool. So, if they try to develop any new statistical tool people would like to use it. So, how to use it? So, that researcher can write the program and can contribute to R software. So, that is why we have here two types of packages; packages means in simple words you can assume, you can understand the packages mean in order to do something that is the specific program.

So, there are built in packages which comes with the R software and some other programs which are called as contributed packages both are available. And, in case if you want to develop something, if you want to do something you can also write a program, you can also write a package and the tools for development of such packages are also available. So, that you can also contribute your own package to the R software.

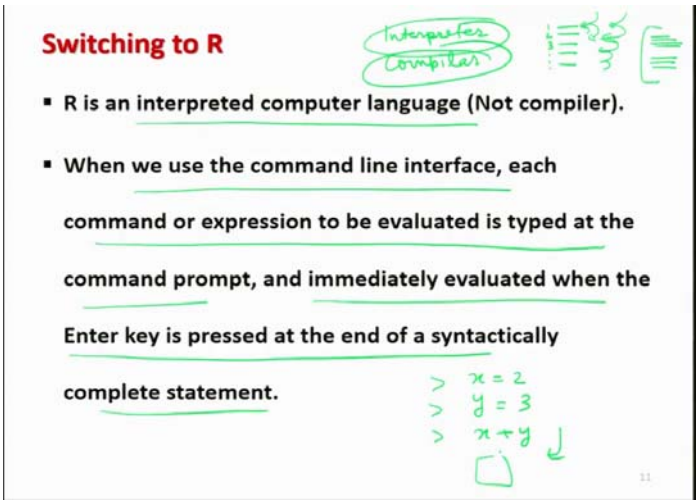
(Refer Slide Time: 25:52)



Now, in case if you try to think of any good programming language that has certain types of instructions like a branching, looping, logical control, modular programming etcetera; [FL] just like any good programming language, this language also provides the logical control of branching, looping, modular programming etcetera using the functions. What is function that we will try to understand and that was one of the beauty of this R language that they have defined the concept of functions that is going to be extremely useful for us.

And, whenever we are trying to write a program we always get some error messages. So, in the R software also whenever we are trying to do the programming we also get some error messages and the language of such messages is quite convenient. So, that helps us in finding out that where we have made the mistake in the program, right.

(Refer Slide Time: 26:48)



Now, when you are trying to work on any programming language, those who are working actually they know that whenever we try to type the program then after that in order to execute it, we have to use the platform and then there are two types of approaches.

One is interpreter and another is compiler. Well, those who have good knowledge of programming they might be understanding, but those who don't have for them I can give you a brief idea. So, whenever we are trying to write down a program something like a line number 1, line number 2, line number 3 and so on. So, when we try to execute the program then the control will come here and if there is some error into the line number 1 then there are two options; either the program will stop here or the program will come to line number 2. So, in case if the program stops here; that means, first the programmer has to look into the program and the mistake in the line number 1 has to be removed. And, then the program is re executed and in case if the mistake in the line number 1 is removed, then the control will come to line number 2 and then in case if there is no mistake then it will come to line number 3.

And, if there is any mistake then at line number 3 that will stop again and the person has to look inside the program and the mistake in the line number 3 has to be removed and then the program will move further. So, this is what happens in the interpreter and another approach is that the control comes to line number 1, then it find if everything is fine, no issue it will come to line number 2 and suppose it finds some mistake in the line

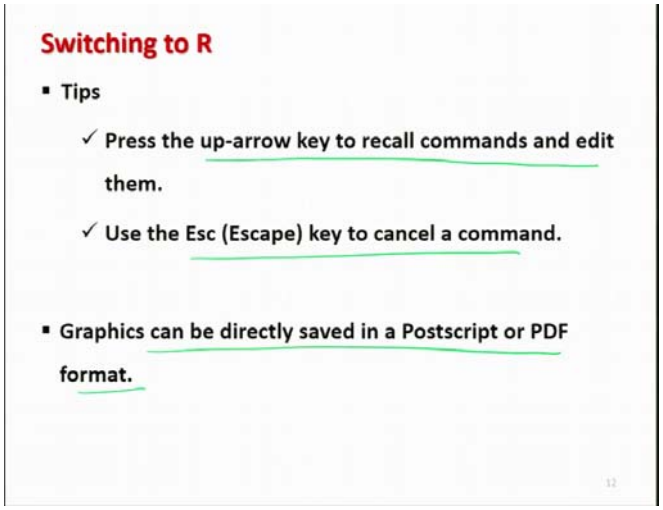
number 2, it will record somewhere. Then it will come to line number 3, suppose there is another mistake in line number 3, it will record somewhere here and it will come to line number 4 etc. etc. and it will come up to n. And, then it will show you the list of mistakes which are inside the programming language. Now, one can go to the program and can rectify these mistakes in the entire program because all the mistakes in the programs are known to us.

So, this is the way the compiler works. So, R is an actually interpreted computer language, it is an interpreter. So, whenever you are trying to execute any program, if there is any mistake in the line number 1 the program will not go to line number 2 unless and until you clear the mistake or you rectify the mistake in the line number 1, right and whenever we use the command line interface.

What is command line interface, that I will try to show you that this is the place where we try to type the command, then each command or expression whatever is could be evaluated is typed at that command prompt. And, then it is immediately evaluated after this using the enter key on the keyboard, right and this completes the entire statement.

So, what does this mean? It is as simple as that suppose if I try to write here x equal to 2 and then y equal to 3 and if I try to write down here x plus y and as soon as I try to press here enter yeah, I am writing it on the R software. So, and then you will get here a value right.

(Refer Slide Time: 29:46)



So, I will try to show you on the R software and yeah just like any other software for example, you can up arrows, down arrows etc. keys to recall the commands and edit them. And, you can use the escape key to cancel a command or in case of the program is running you can immediately cancel it by pressing the escape key.

So, that is just like what is possible in any other good language and whatever are you graphics they can be directly saved into the Postscript file, PDF file, JPEG format etc. So, whatever the ways which are available in any standard software they are available in the R software also.

(Refer Slide Time: 30:22)

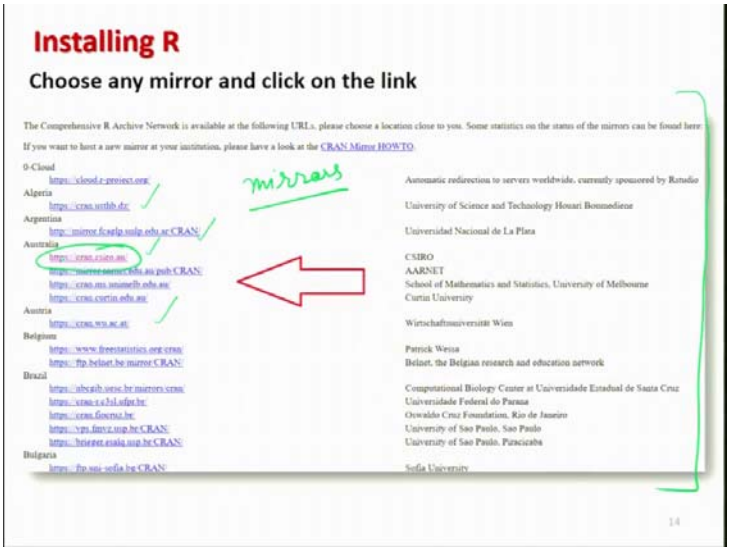
So, now this is the brief background to convince you that you should not feel that if you are trying to use R software which is available for free, you are not compromising on anything and you are going to get the same quality of outcome which will be available from any other software. So now, once you are convinced then the next question comes here- how are you going to install this software, from where you are going to obtain the software right?

So, there is a website www r hyphen project p r o j e c t dot o r g. So, you can download the software for any platform Windows, Macintosh, Unix, Linux etc. from this website. So, I have given you here a screenshot of this website from where you can download it. So, you can see here this is the address of the website and if you go there, there is a link

here something like download R and then you can see that when I am trying to record this lecture at this moment, the latest version which is available is R version 4.1.2.

So, we are going to work in this version only and now you can recall that earlier I told you that there was version 1.0 that was released in 90s, right. So, now, it has come to 4.1.2.

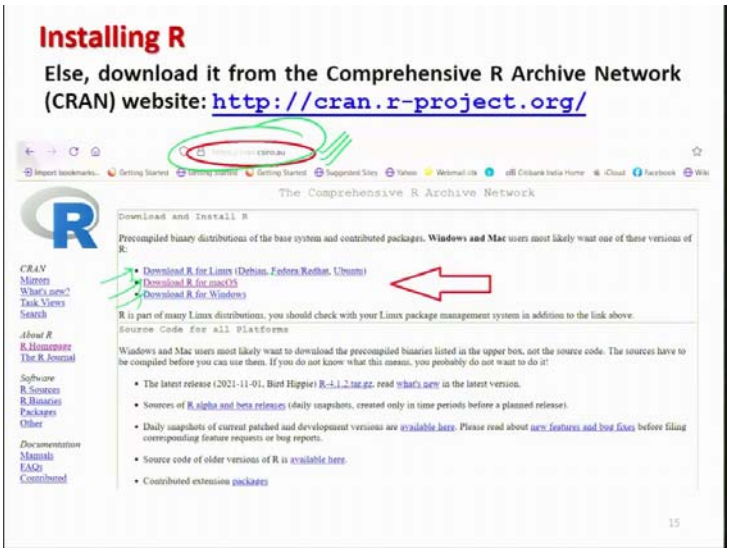
(Refer Slide Time: 31:48)



So, you simply click over here at this button download R and then you will get here this software and then actually what will happen that after you press on this button download R, then it will show you this type of screen. Actually, these are the different mirrors; that means, now you can see now there are many many institutions across all the world in different countries who have hosted this website. And, all those R software and their packages they can be obtained from any of these websites.

So, you can see here there is a site in Algeria, Argentina, Australia, Austria etc.; they are arranged in an alphabetical order of the list of the name of the country right. So, now here I would like to; so, you can actually click here any of this issue and then after that you will come to this site. For example, I try to press here on this in Australia c r a n dot c s i r o dot a u, if I click here I can I get here at this address same address.

(Refer Slide Time: 32:40)



And, this screen is going to be the same even if you try to click at any other address and you will get here this type of screen download R for Linux, for mac, for this Windows etc. And, whatever is your requirement you try to type here and then you try to simply click here and then you will get the software and then after that you have to simply do click, click, click etc. etc.

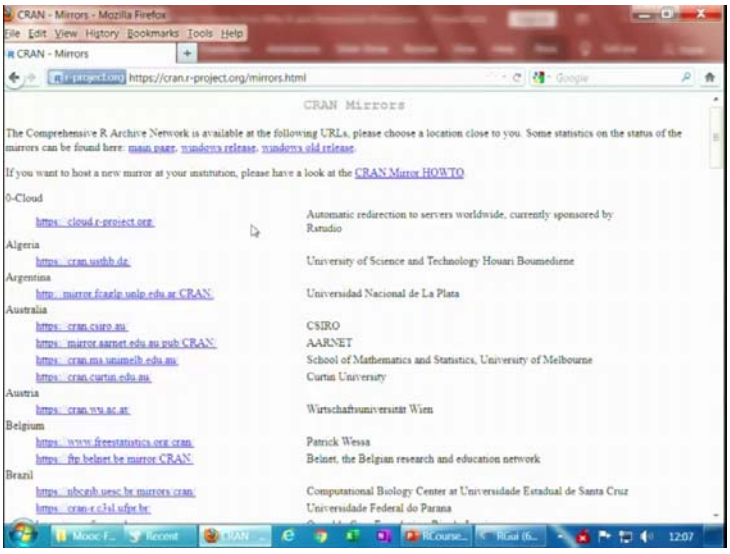
And, then you can install it in the usual way, the way by which you try to install any other software. One question which I would like to address here before I try to show you that how these things are being done that this R software is available on different websites which are hosted by different academic institutions across the world. So, sometime people think that that which of the website is going to give us the good software.

So, I want to clear this myth that all the software there is a long list that you will see in the R software, they have the same software number 1. Number 2, either you download it from the site in Australia or Austria or Brazil, it doesn't make any difference, right. Sometimes people do say that ok, if you try to download the software from some neighboring country possibly that helps. Well, I have no reason to claim this or to get convinced with this claim right. So, you can download it from any site and then you can work on it.

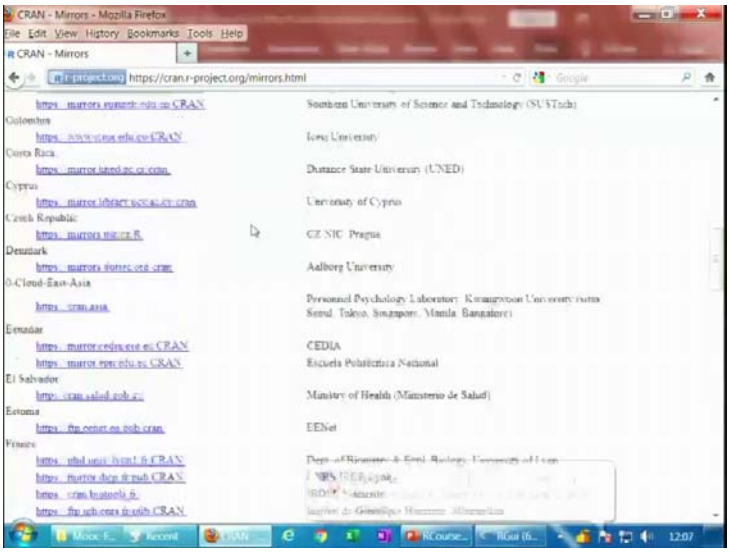
(Refer Slide Time: 34:10)

So, let me try to show you here that when you try to do these things in the internet how are you going to get. So, you can see here that I have means click on this site w w w r hyphen project dot r dot o r g that you can see here, I can increase the font size. So, you can see here if you click here, you will get here and then here you can see where I am trying to move my control cursor, you can see here there is a site here download R.

(Refer Slide Time: 34:39)

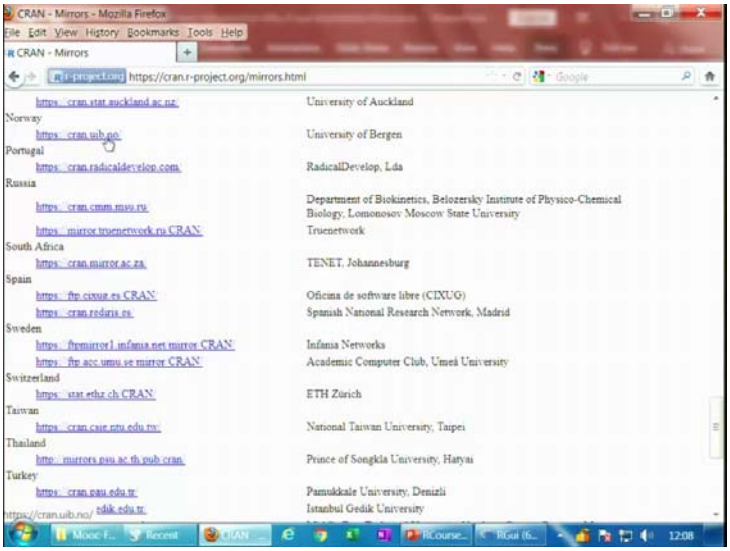


(Refer Slide Time: 34:44)



So, if you try to put here download R and then you will see here that it will try to bring to the different type of this mirrors which are CRAN mirrors and you can see here that this is a long list of different countries which are available, right.

(Refer Slide Time: 34:50)



And, I will say here that you just click on any country say Norway or anywhere you get here, suppose I click on click here in Norway and then it will bring you to the same site which I have shown you here and then you can install this software. So, with this

detailed discussion and after giving you a reason that why should you switch to R and how are you going to install it, I come to an end to this lecture.

And, I hope that I was successful in making you understand that what was the story behind this R software. And, then sometimes people do ask me why people are doing it for free? You see we are working in this academic community and in academic community, money is not everything. We always try to do some research, we always try to publish our research papers, many people think that that once our research papers are published we get some money, this is wrong; that's our job.

And, that is our duty towards the society that we have to try our best so, that we can develop the things which help the mankind, humankind. So, that is the very simple modest objective of those academicians who group together and they are still working, there are many people across the world who are working for the development of different types of packages for the R software. So, one should not doubt on their integrity, one should not doubt on their intentions. They are doing it for the welfare of the people that is all.

So, with this comment I conclude this lecture and from the next lecture I will try to show you something more on the R software. Till then good bye.