

Health Economics

Dr Pratap C Mohanty

Department of Humanities and Social Sciences,

Indian Institute of Technology Roorkee

Week - 11

Lecture 56- Health Data Handling Packages: Open Source

Welcome readers to this NPTEL MOOC module on Health Economics. As we know, this week is meant to guide you on how you can go for different data packages. In this lecture, we are clarifying the packages that are open source on health data or the health data handling packages, which are open source. So, largely all the sources of the data packages are shown-



All are written, but here, they are not all open source. We will discuss this (Python), (R), and DEAP. We will emphasize these three in this lecture.

So, in the last lecture, we explained linear regression or different econometric modelling, especially in healthcare research, such as linear regression or non-linear regression. We briefly described the research design, including randomized control trials, quasi- or natural experiments, and also different estimation methods. Here, we will discuss the type of data processing software that can be used in health economic analysis. We will introduce you with Python and R and re-emphasize the discussion we did on DEAP. So, DEAP, we did it already. But we will remind you to follow again; that is also an open-source software. We will give you very step-by-step operations that you can easily understand, and you can at least start working with the very basics of Python and R.

I am sure that you might have heard of these studios and this software. These are largely used by scholars, and especially in corporate research, they take the help of these tools.

Software available for data processing

Feature	SAS	JMP (SAS)	SPSS IBM-2009	Stata	R	Python (Pandas)	DEAP
Learning Curve	Pretty Steep	Gradual	Gradual	Gradual	Pretty Steep	Steep	Pretty Steep
User Interface	Programming	Point-&-Click	Point-&-Click	Programming/ Point-&-Click	Programming	Programming	Programming
Data Manipulation	Very strong	Strong	Strong	Strong	Very strong	Strong	Strong
Data Analysis	Very Strong	Strong	Very strong	Very strong	Very strong	Strong	Very strong
Graphics	Good	Very Good	Good	Very Good	Excellent	Good	Very Good
Cost	Expensive (yearly renewal) Free Student version, 2014	Expensive (yearly renewal) Student Disc.	Expensive (perpetual, cost only with new version) Student Disc.	Affordable (perpetual, cost only with new version) Student Disc.	Open Source (Free)	Open Source (Free)	Open Source (Free)
Released	1972	1989	1968	1985	1995	2008	1996

Already discussed, please refer previous unit, i.e., → 10

Software available for data processing is largely on SAS, the JMP version of SAS, then SPSS of IBM, and another one is called Stata, but these are all paid versions. We were going to discuss the last three (i.e., R, Python, DEAP), but we have already discussed the DEAP (data envelopment analysis package) in the previous unit, i.e., number 10. Suppose you compare all these together on the basis of learning curve, user interface, data manipulation, data analysis, graphics, and cost-effectiveness. In that case, we have a special advantage of these: free software (these three: R, Python and DEAP). Otherwise, even in all categories, you will see that these three are very strong regarding their data manipulation and analysis. In particular, R is very strong, while Python and DEAP are strong, so far as data manipulation is concerned, and graphics are excellent in R.

So, R is largely suggested in almost all directions. I think I am not explaining here on DEAP. I will start here, first with Python, and then we will explain R. Python is a self-interpreted, object-oriented, high-level programming language with dynamic semantics. The basic syntax is 'Print ()', where the common data type is required. A data type is a means of classifying a value and determining what operations can be performed. All objects have a data type. So, what does it look like? The data type looks like-

Data Type

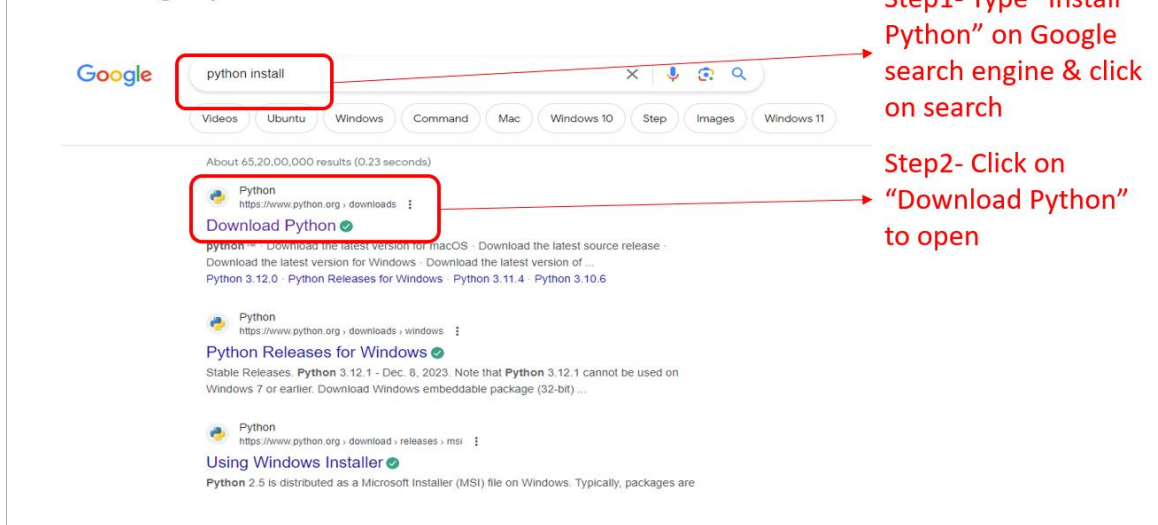
Class	Description
bool	Boolean value
int	integer (arbitrary magnitude)
float	floating-point number
list	mutable sequence of objects
tuple	immutable sequence of objects
str	character string
set	unordered set of distinct objects
frozenset	immutable form of set class
dict	associative mapping (aka dictionary)

Here it considers its Boolean values, then integer values as int, floating-point number, the list, tuple, string characters, set, frozen-set, and dict (basically associated with mapping as dictionary). I will tell you. There are operators as well. Operators are larger in the symbols and are used to carry out specific functions and computations. So, we just highlighted operators here for your reference-

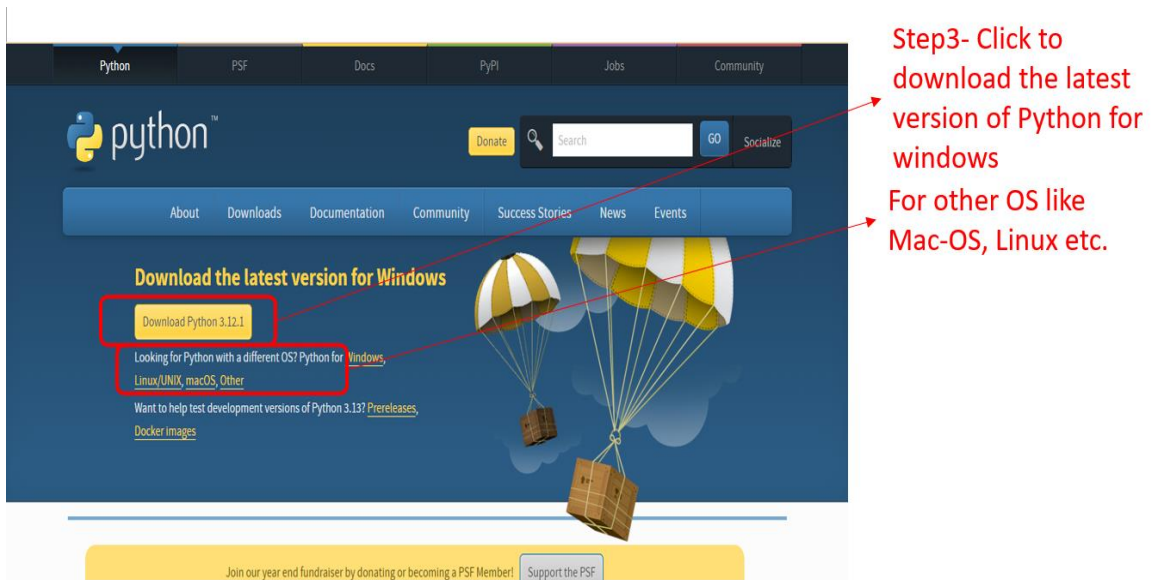
Operator	Description
**	Exponentiation (raise to the power)
~ + -	Ccomplement, unary plus and minus (method names for the last two are +@ and -@)
* / % //	Multiply, divide, modulo and floor division
+ -	Addition and subtraction
>> <<	Right and left bitwise shift
&	Bitwise 'AND'
^	Bitwise exclusive 'OR' and regular 'OR'
<= < > >=	Comparison operators
<> == !=	Equality operators
= %= /= // = -= +=	Assignment operators
*= **=	Assignment operators
is is not	Identity operators
in not in	Membership operators
not or and	Logical operators

How do we install Python, and where do we get it? This lecture will be very useful for the one who is just starting, and hearing about this, and going to start the basic data analysis.

Installing Python



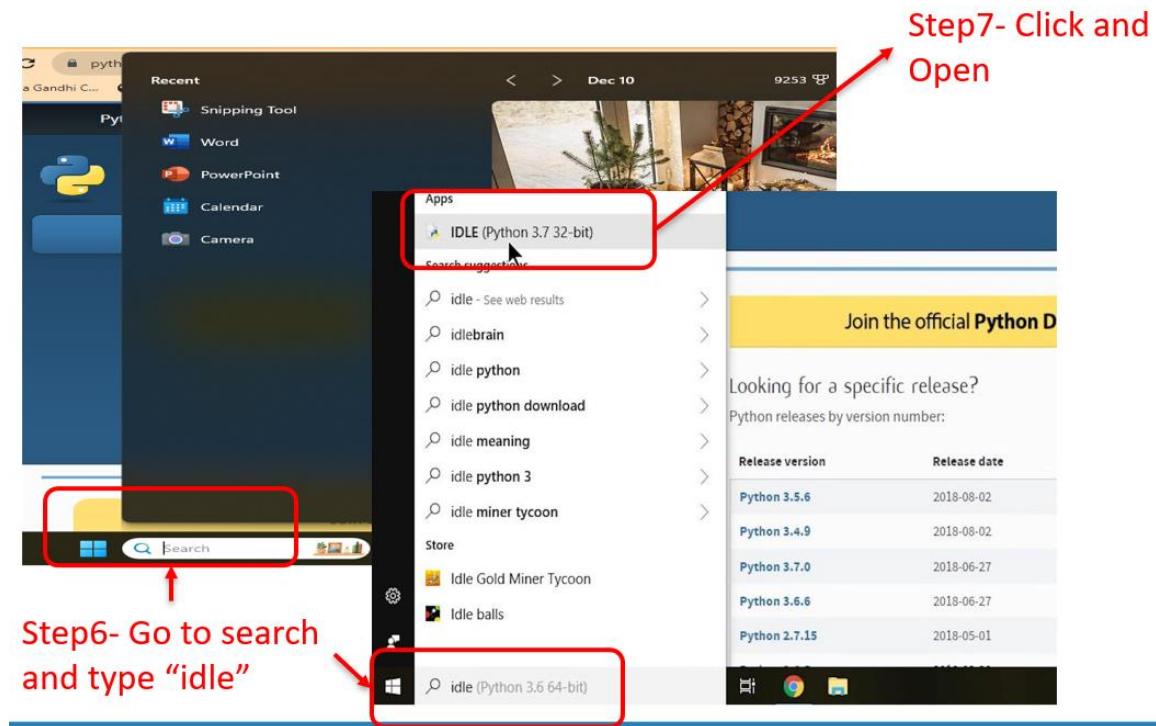
So, here, if you search on Google, 'Install Python,' it will give you the directions. So, the first step is to type on Google and search (we have just highlighted here). And in the second step, what do you need to do? You need to click 'download Python link' to open. And, then, it directs you for the window to download the latest version-



And the latest version is here. You can download either for Windows or for OS. You have to just mark carefully. If you require Python for Windows, then click on download Python. If you require for the operating system like MacBook or Mac or Linux etc., you have to install a separate one (below link). Then this version is with us now: 3.12.1, 64 bits.

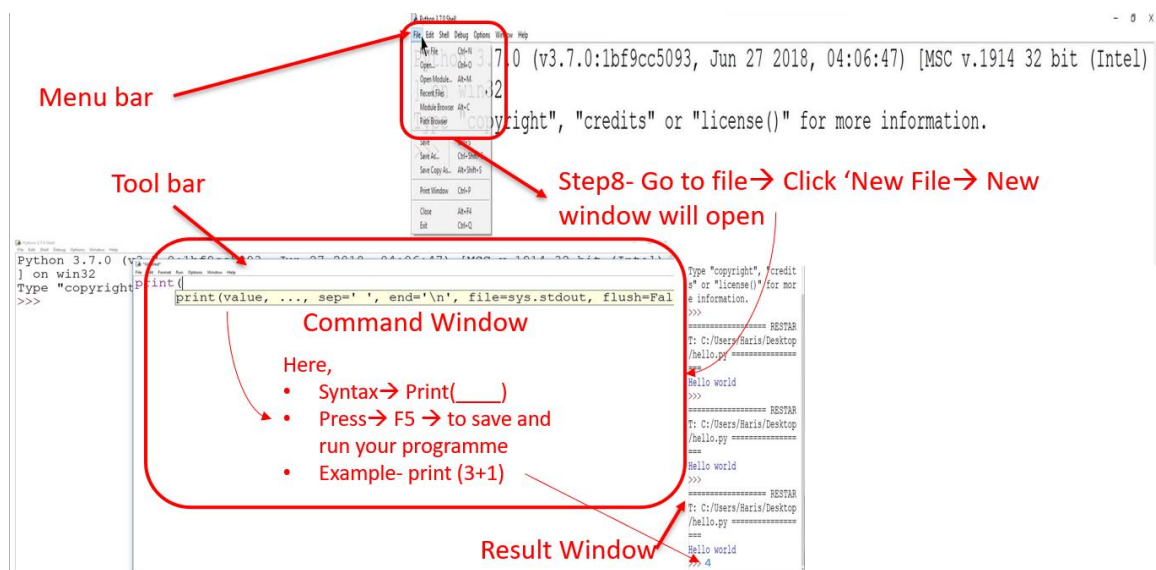
Install instructions are there. Just click on that, but make sure that you have been marked as the admin. Check mark (both of these)- 'admin' and 'add Python.exe to path'. Then you just

install it. And once this is made, you will see this is the one (app), and in your screen you can just search this as 'idle'.



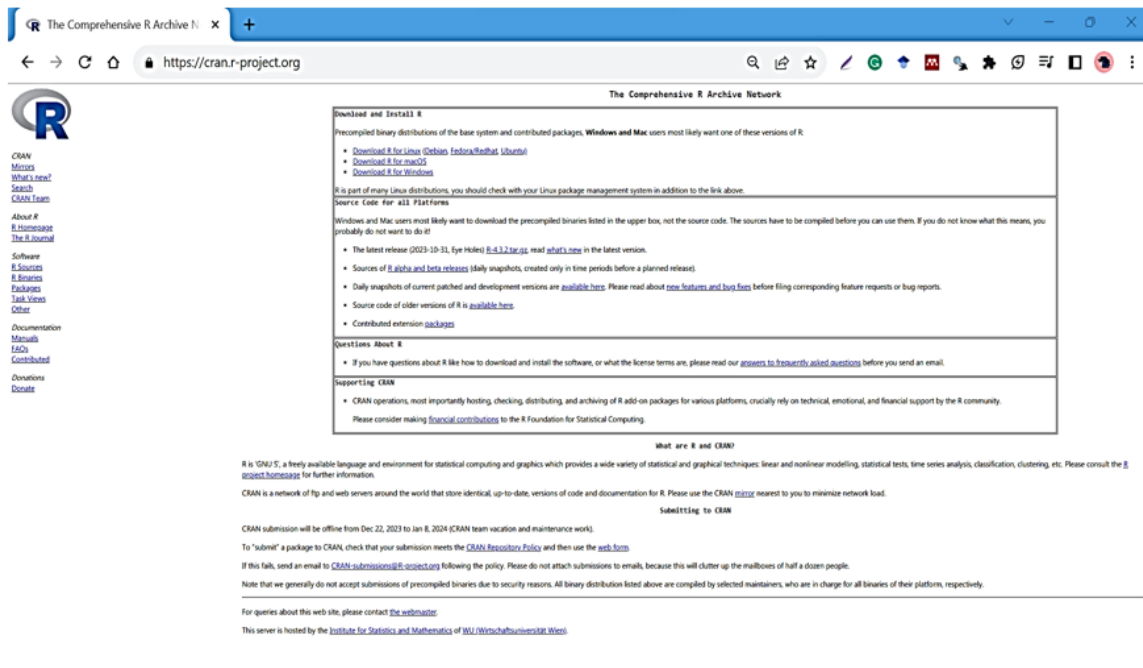
And it will display you with the one which we are working for the data. Then you simply, once it is displayed like this, you click on it, and then open.

And the window will look like this-



Starting with it as a menu bar, we are just highlighting which are the menu bar. Then, if you want a new file, then a new window will open. Then, it will give you another new window. In that window, you will see it has toolbar, and the results (however, the results details are displayed just to the right), and on this new window we will see next to the toolbar, you will see the command window. And in the command window, we are supposed to write the syntax. Like syntax here, for example, syntax print (___), then you press F5 to save and run your program. For example, print (3 + 1) = 4. So, it is just displayed as the result on the right-hand side as 4 (we have just highlighted here). Then you need to also save, and also run the program. So, like in Python, we have just given you the very basics of Python.

Similarly, like Python, R is also an open-source software, and the downloading process is given below-



You can follow the steps. So, now, download R for Linux, then R for Mac, or Windows, all options are given. So, you can click on the respective one.

Download and Install R
Precompiled binary distributions of the base system and contributed packages. Windows and Mac users most likely want one of these versions of R.

- [Download R for Linux \(Debian, Fedora, Redhat, Ubuntu\)](#)
- [Download R for macOS](#)
- [Download R for Windows](#)

It is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for All Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it.

- The latest release (2023-10-11, Eye Holes) [R 4.3.2 for gcc](#); read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).

Supporting CRAN

- CRAN operations, most importantly hosting, checking, distributing, and archiving of R add-on packages for various platforms, crucially rely on technical, emotional, and financial support by the R community. Please consider making [financial contributions](#) to the R Foundation for Statistical Computing.

Step 1: Download R from <https://cran.r-project.org/>

You are supposed to download the R from this website- <https://cran.r-project.org/>. Then, you install the RStudio, which is downloaded from that website.

posit PRODUCTS SOLUTIONS LEARN & SUPPORT EXPLORE MORE PRICING

1: Install R

RStudio requires R 3.3.0+. Choose a version of R that matches your computer's operating system.

[DOWNLOAD AND INSTALL R](#)

2: Install RStudio

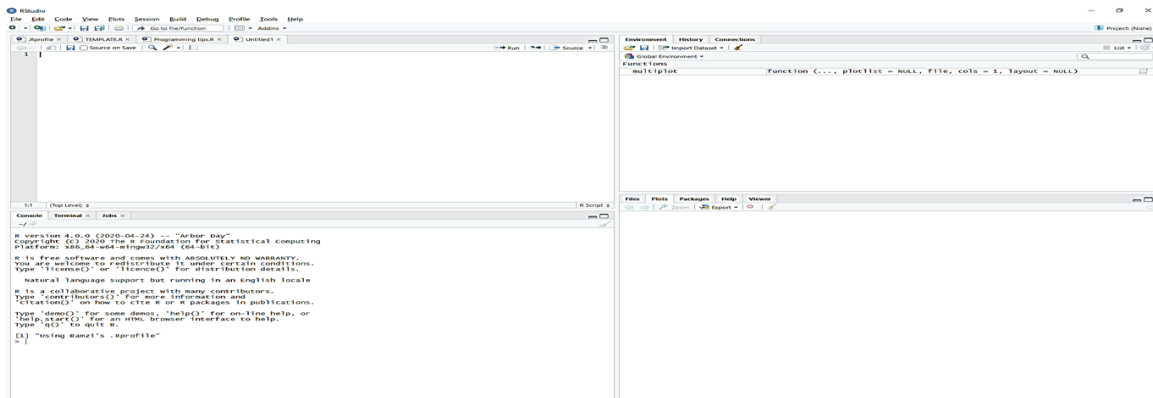
[DOWNLOAD RSTUDIO DESKTOP FOR WINDOWS](#)

Size: 214.34 MB | SHA-256: FE62B784 | Version: 2023.09.1+494 | Released: 2023-10-17

Step 2: Download RStudio from <https://posit.co/download/rstudio-desktop/>

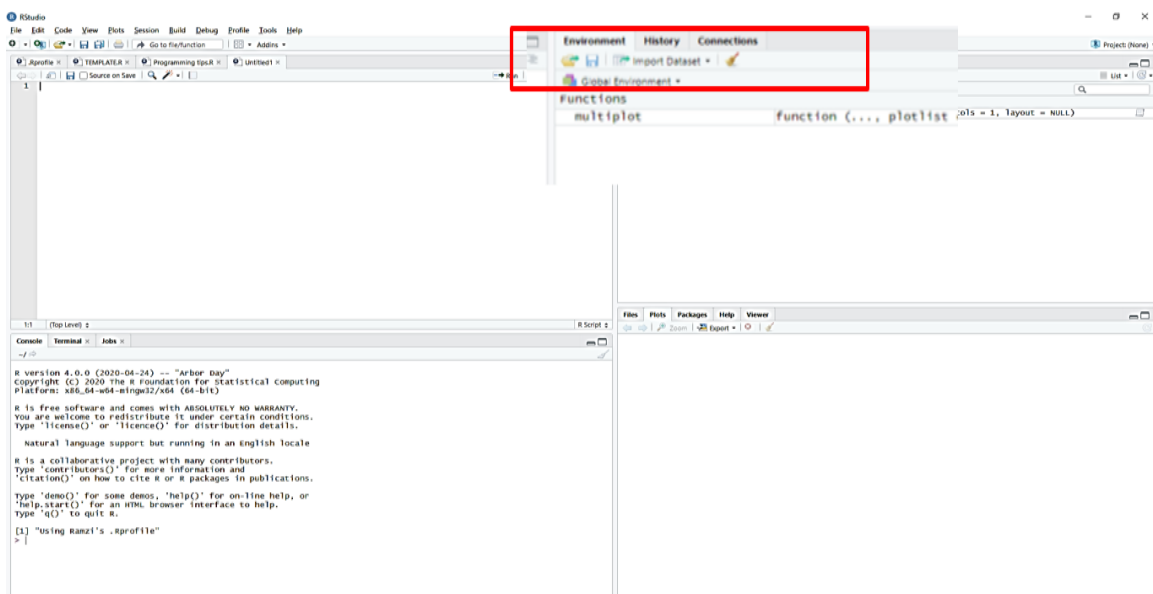
You have to install it, especially if you are working with desktop Windows. Then you have to install this, download it, and install it.

RStudio is an integrated development environment (IDE) for R, a programming language for statistical computing and graphics. You can run the R on RStudio with ease. Other IDE available for R are- Jupyter Notebook, Visual Studio Code, Emacs, and ESS, etc. Those who are the advanced user, they used to go for it. It usually displays four windows simultaneously, so you can work. When you open RStudio, you will see four panes, each with multiple tabs-

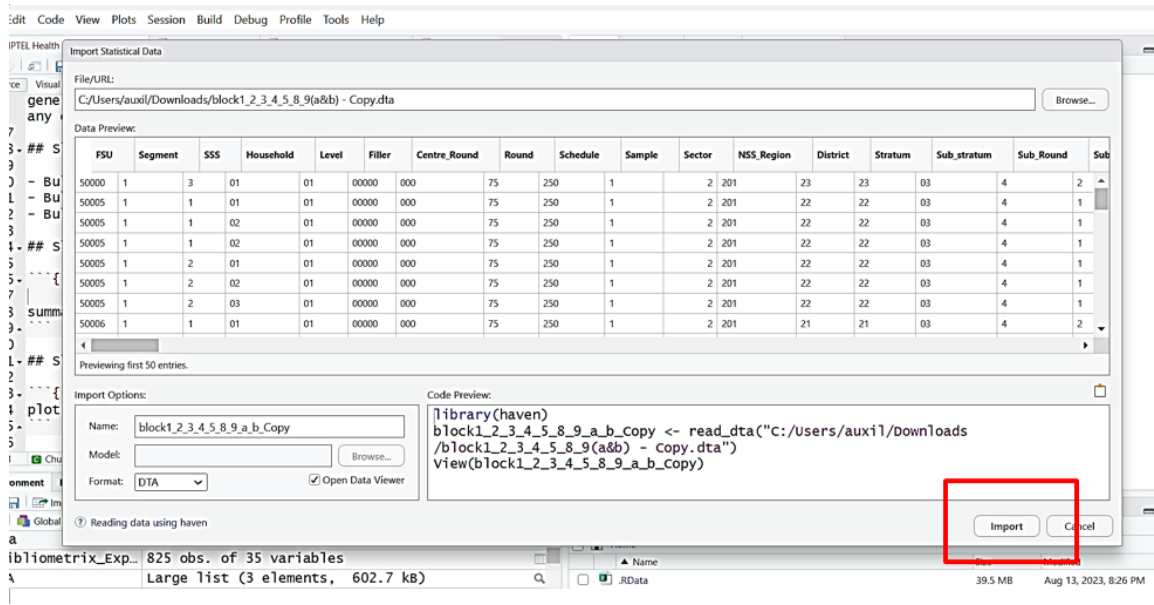


The upper left (the top one) is a script file. The upper right is the environment file, which will list any objects you create. The lower left is the console. This is where you type the R code to run. And the lower right are the plots. This is where the graphics you create will appear.

If you want to work with some databases, of course, databases are required. So, how to import the database? So, click on import data in the environment pane (this is how we are just highlighting the environment pane)-

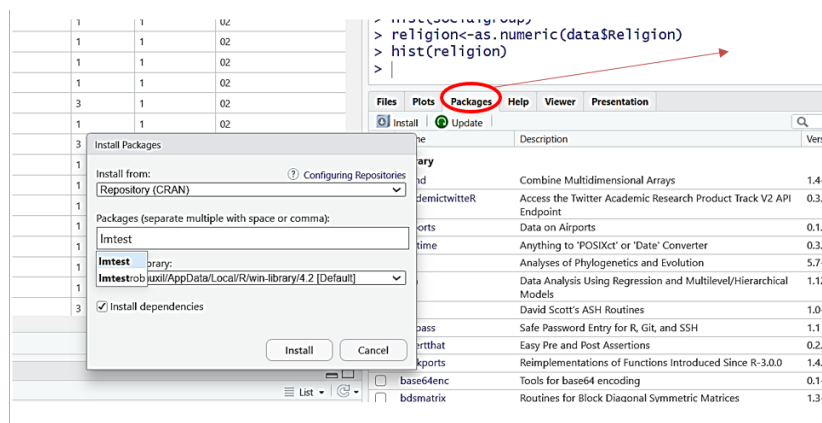


Then you browse your data, browse option is there, browse your folder where you might have stored your data, and you simply click on import.



Once imported, your data is imported to R, and this is how it looks like. You import it, import bottom is there, and then your data will be imported.

You can download packages from your data analysis.



- You can download packages for your data analysis.
- Step 1: Click on Packages in Lower Right.
 - Step 2: Click on Install
 - Step 3: Type name in space provided (as shown in e.g. **lmtest**)
 - Step 4: Click Install

Click on packages in the fourth pane in the lower right (which already discussed), and click on the install. Then you type the name in the space provided, where you have the data, etc.

Then, there are some packages that we have just discussed.

The screenshot shows the CRAN website at <https://cran.r-project.org>. The page is titled "Topics" and features a large "R" logo. On the left, there are several navigation menus: "CRAN" (with links for Mirrors, What's new?, Search, CRAN Team), "About R" (with links for R Homepage, The R Journal), "Software" (with links for R Sources, R Binaries, **Task Views**, Other), and "Documentation" (with links for Manuals, FAQs, Contributed). A red circle highlights the "Task Views" link. A red arrow points from this link to a text box on the right. The main content area lists various topics such as Actuarial Science, Agriculture, Bayesian Inference, Causal Inference, ChemPhys, Clinical Trials, Cluster, Databases, Differential Equations, Distributions, Econometrics, Environmental Data, Epidemiology, Experimental Design, Extreme Value Analysis, Finance, Functional Data, Graphical Models, High-Performance Computing, Hydrology, Machine Learning, Medical Imaging, Meta-Analysis, Missing Data, and Mixed Models.

You can explore Task Views in CRAN website. It has list of packages you will need for a particular task. Example Econometrics; CausalInference etc.

Extra packages for better analysis, if you go through, you will be guided. You can also explore task views in CRAN website, which we have highlighted here. You will need it for a particular task, for example, maybe econometrics, causal inference etc. So, you may track from there. So, all the list of topics are given. So, you can follow it accordingly.

So, what is in the next lecture to be discussed? The next lecture will be on the licensed data handling packages, especially we will be emphasizing on health data, and you can follow our lecture for interesting aspects. So, we will clarify very step by step using the license packages, maybe Stata and SPSS. So, here are the suggested sources for your reading, and you can follow this for the details.

So, thank you. I think I should stop here. Thank you.