Session 1 - Introduction to data work in R

R training - Georgia RS-WB DIME

Marc-Andrea Fiorina, Luis Eduardo San Martin
The World Bank | WB Github

April 2023

Table of contents // სარჩევი

About this training // ამ ტრენინგის შესახებ

About this training // ამ ტრენინგის შესახებ

- This is an **introduction** to data work and statistical programming in R
- The course does not require any background in statistical programming
- The course requires a computer with R and RStudio installed

About this training // ამ ტრენინგის შესახებ

Learning objectives

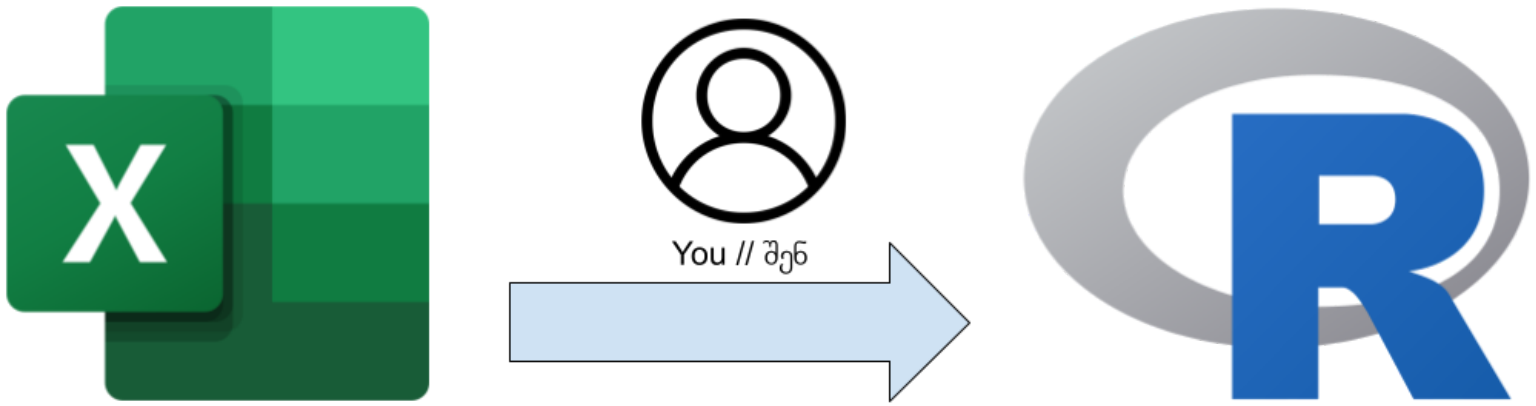
By the end of the training, you will know:

- How to write **basic** R code
- A notion of how to conduct data work in R and how it differentiates from Excel

About this course // ამ კურსის შესახებ

Previous knowledge

- We assume that you have some experience working with data in Excel
- The idea of this short training is to use that knowledge and "translate" some of it to R

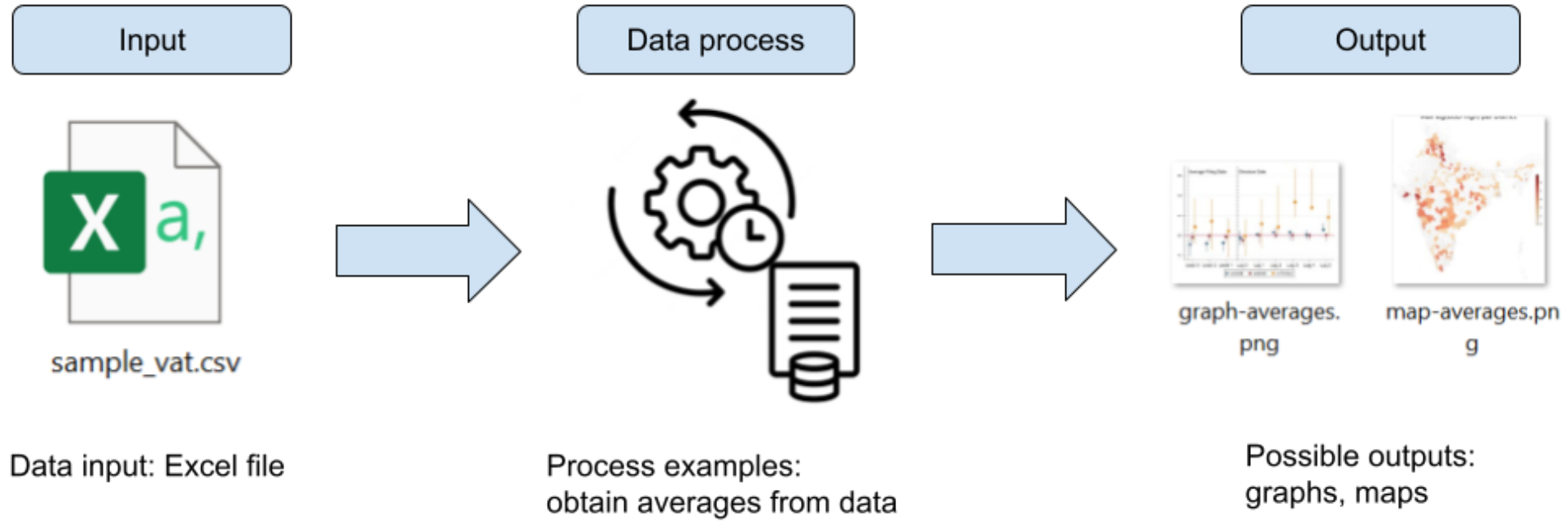


Data work // მონაცემთა მუშაობა

Data work // მონაცემთა მუშაობა

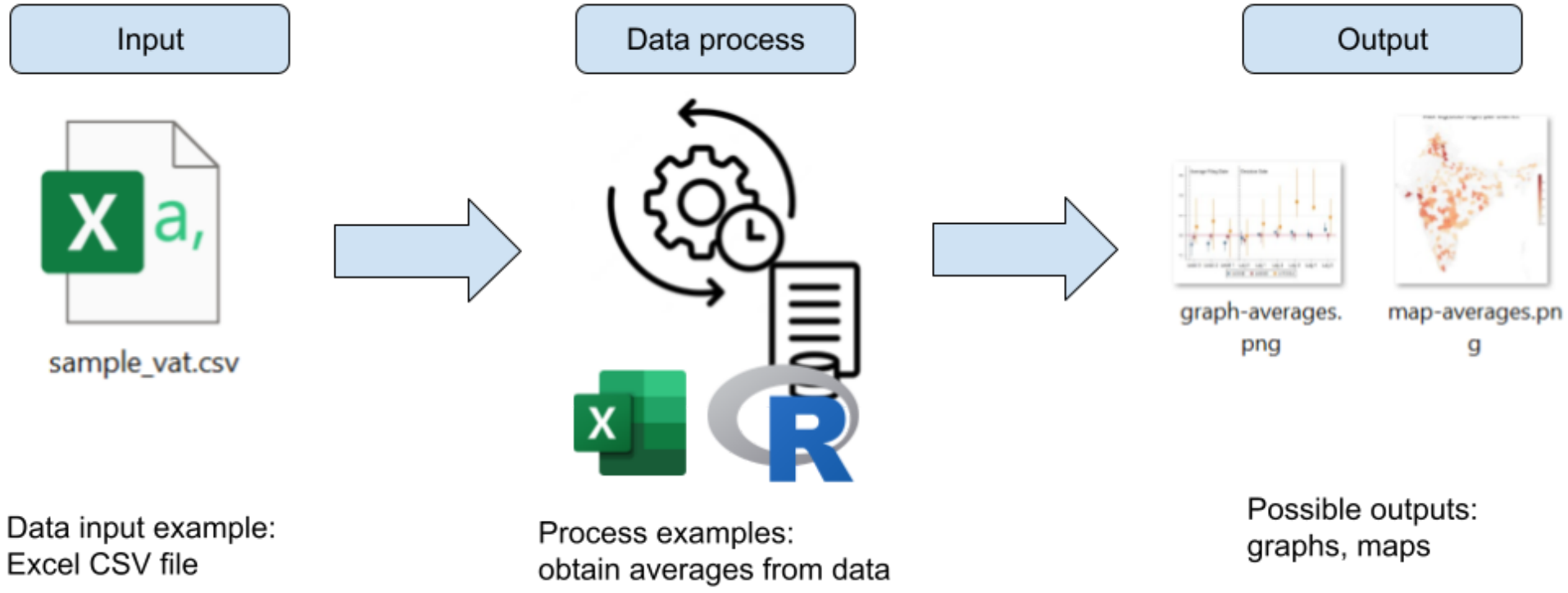
For the context of this training, we'll call data work everything that:

1. Starts with a data input
2. Runs some process with the data
3. Produces an output with the result



Data work // მონაცემთა მუშაობა

- It's also possible to do data work with Excel
- However, we will show in this training why using statistical programming (through R) is a better way of conducting data work

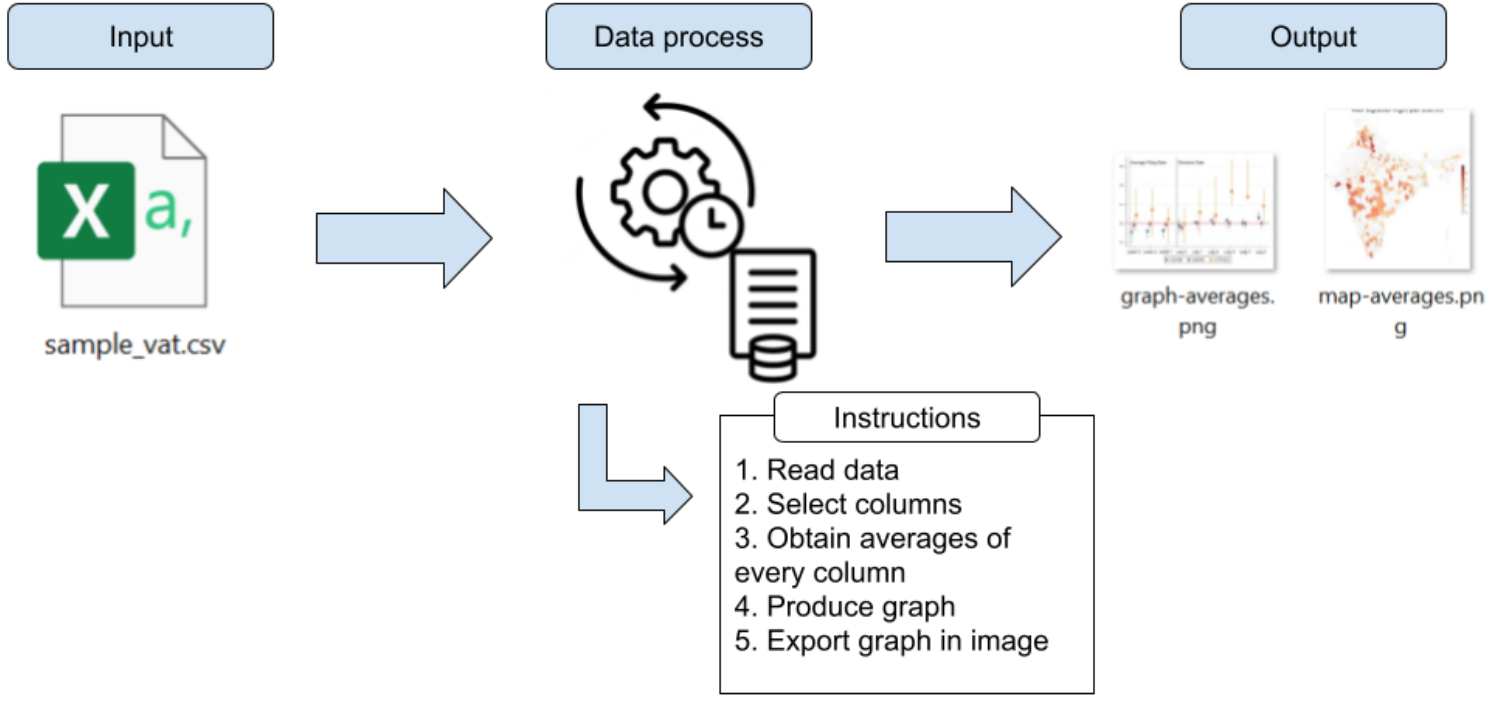


Statistical programming // სტატისტიკური პროგრამირება

Statistical programming

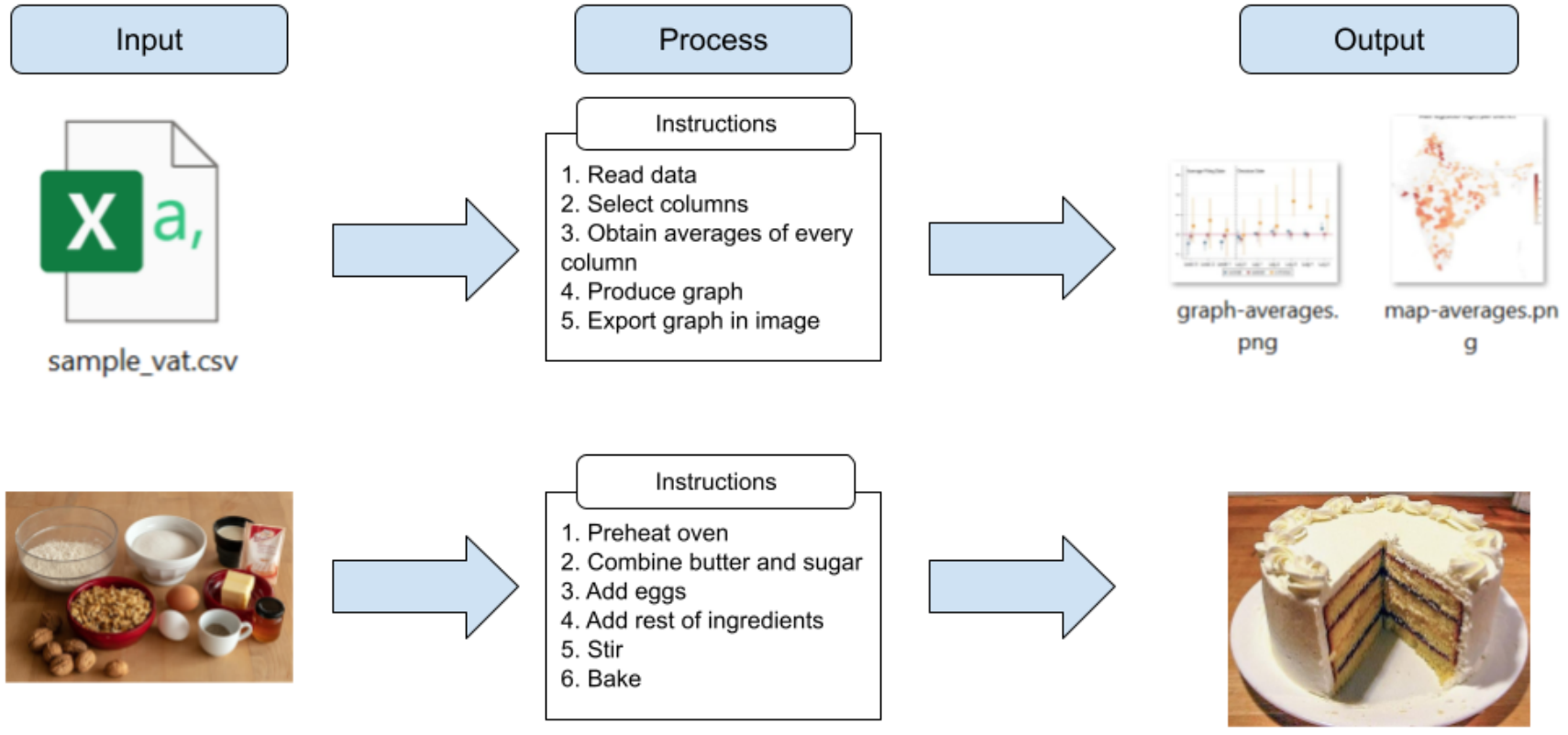
What do we mean by statistical programming?

- Programming consists of producing instructions to a computer to do something
- In the context of data work, that "something" is statistical analysis or mathematical operations
- Hence, statistical programming consists of producing instructions so our computers will conduct statistical analysis on data



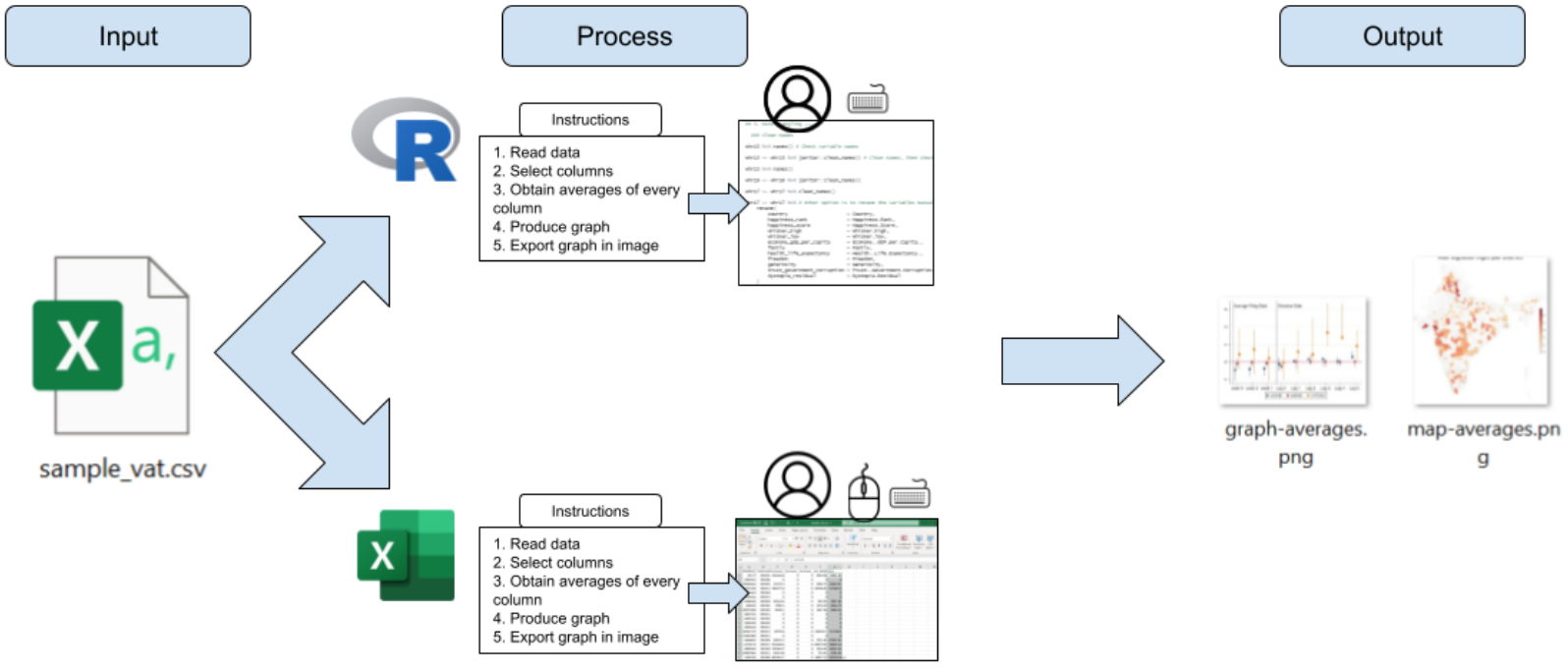
Statistical programming

- You can think of statistical programming as writing a recipe



How is R different than Excel?

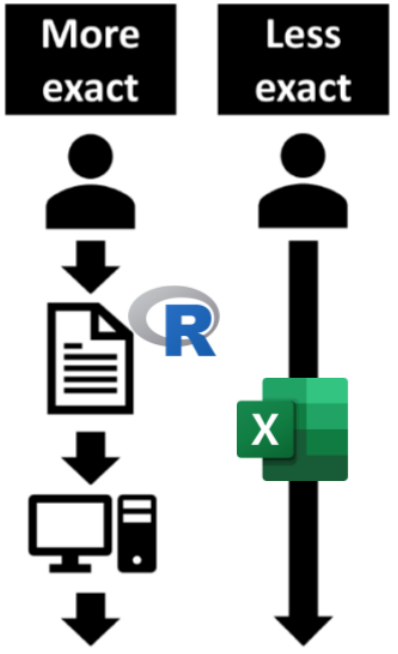
- In Excel, you usually work on the data directly. The input changes as you work on it
- In R, you produce code which contains the instructions your computer follows to conduct the process you want to do

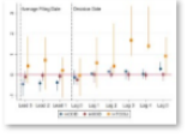


Statistical programming

How can this benefit my work?

- Data work should be very exact to be correct
- Programming consists of giving computers "instructions" about what to do with data. This is more exact than manually working directly with the data, as we do in Excel
- Computers are **very exact**. In programming, they will do exactly what you tell them to do
- This means that you can generate code that is an exact record of how a result was generated





graph-averages.png

Statistical programming

Why use R

- Statistical programming can be implemented through many different software. Other options are Stata and Python
- We recommend using R for these reasons:
  - R is free
  - R was designed specifically for statistical programming
  - There is a large worldwide community of R users. This means you can easily look for help or examples of code in the internet

How to write R code?

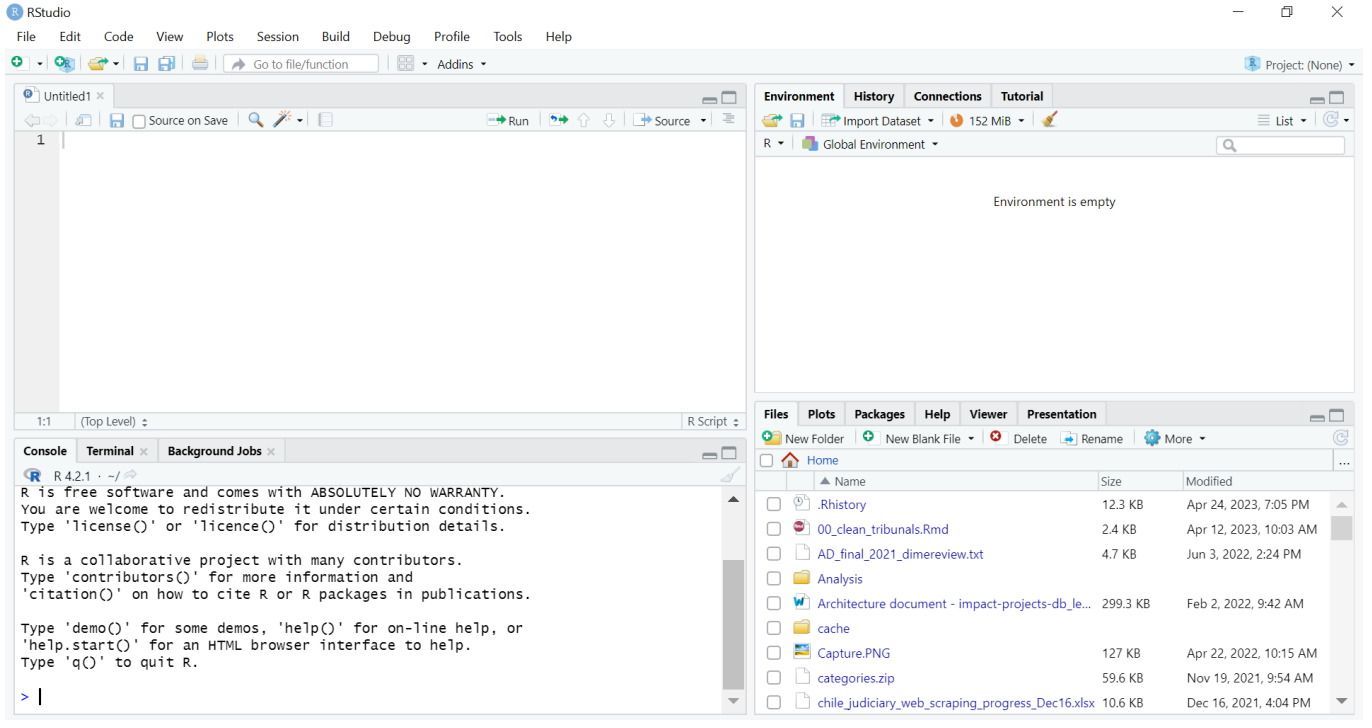
- The rest of today's session focuses on the basics of writing R code
- We'll use RStudio to write R code in this training

How to write R code?

- Now open RStudio in your computer
- Please make sure you're opening RStudio and not R

How to write R code?

- Now open RStudio in your computer
- Please make sure you're opening RStudio and not R

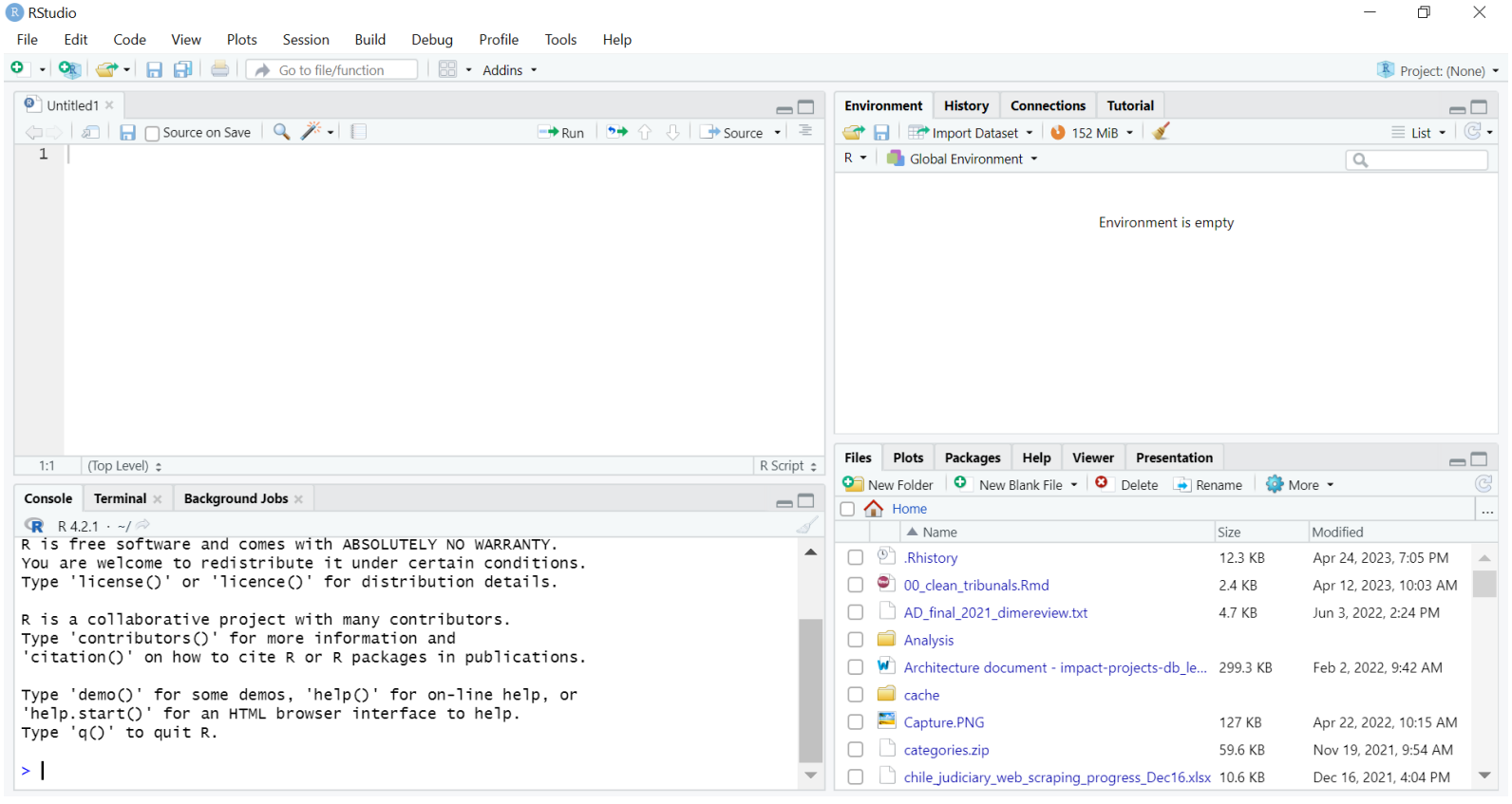


Questions? // კითხვები?

Writing R code // R კოდის დაწერა

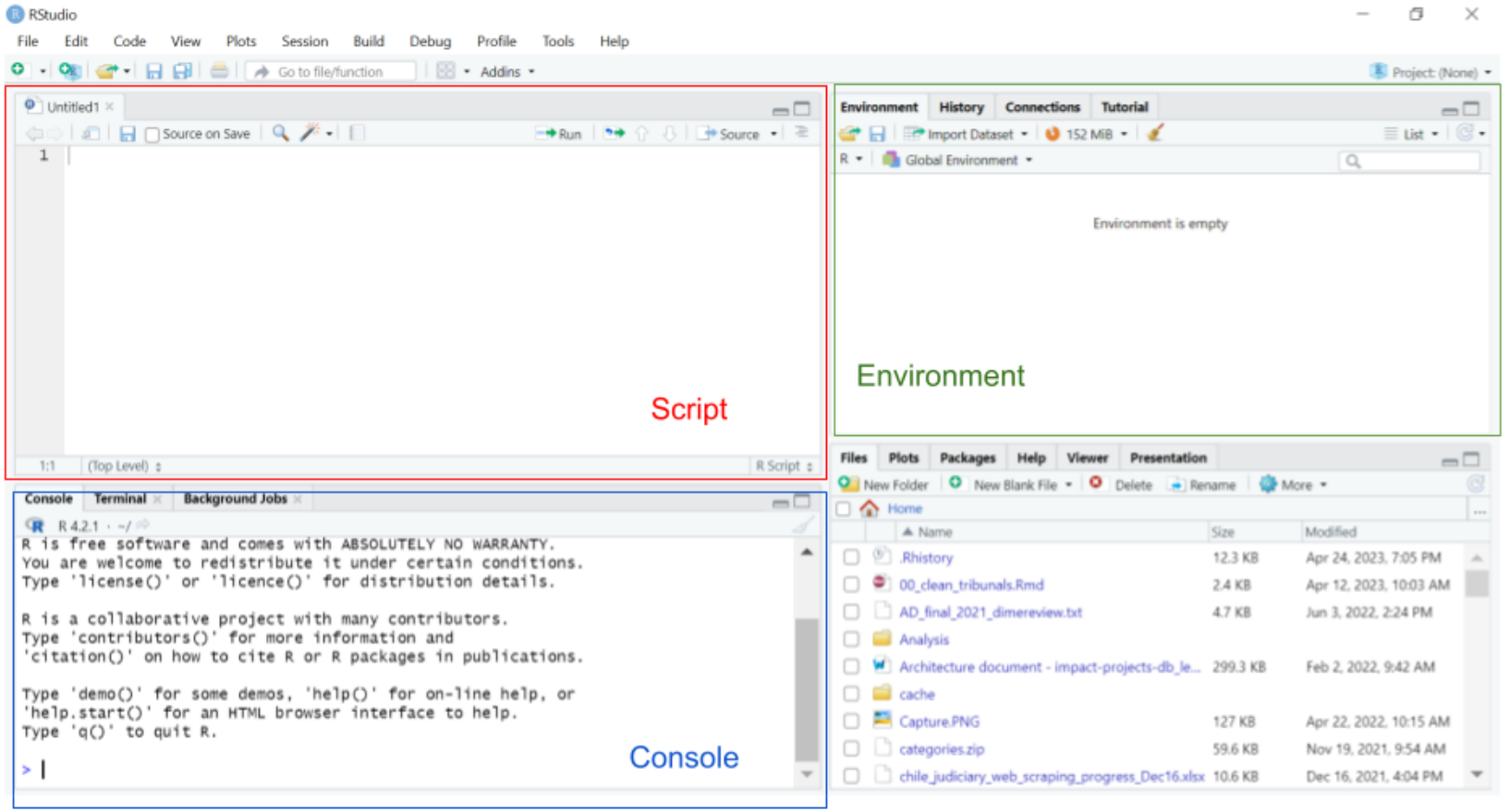
Writing R code // R კოდის დაწერა

RStudio interface



Writing R code // R კოდის დაწერა

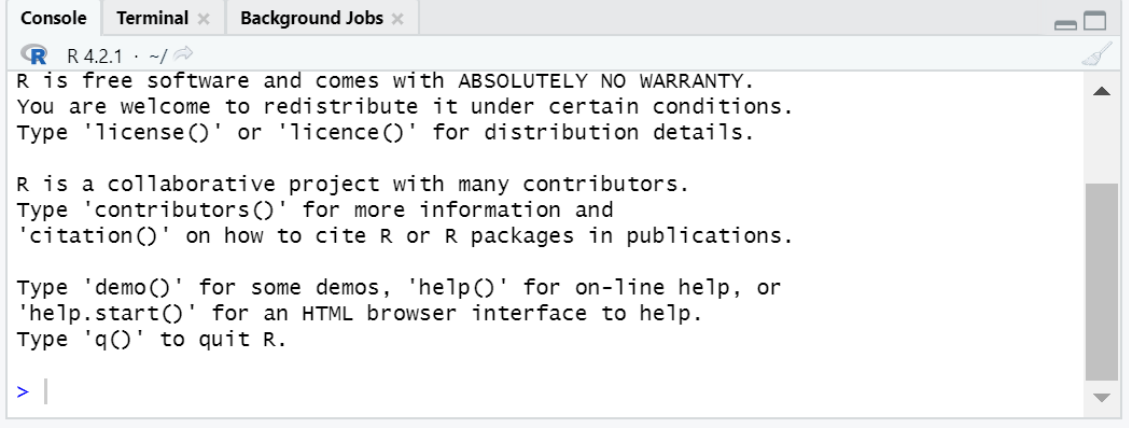
RStudio interface



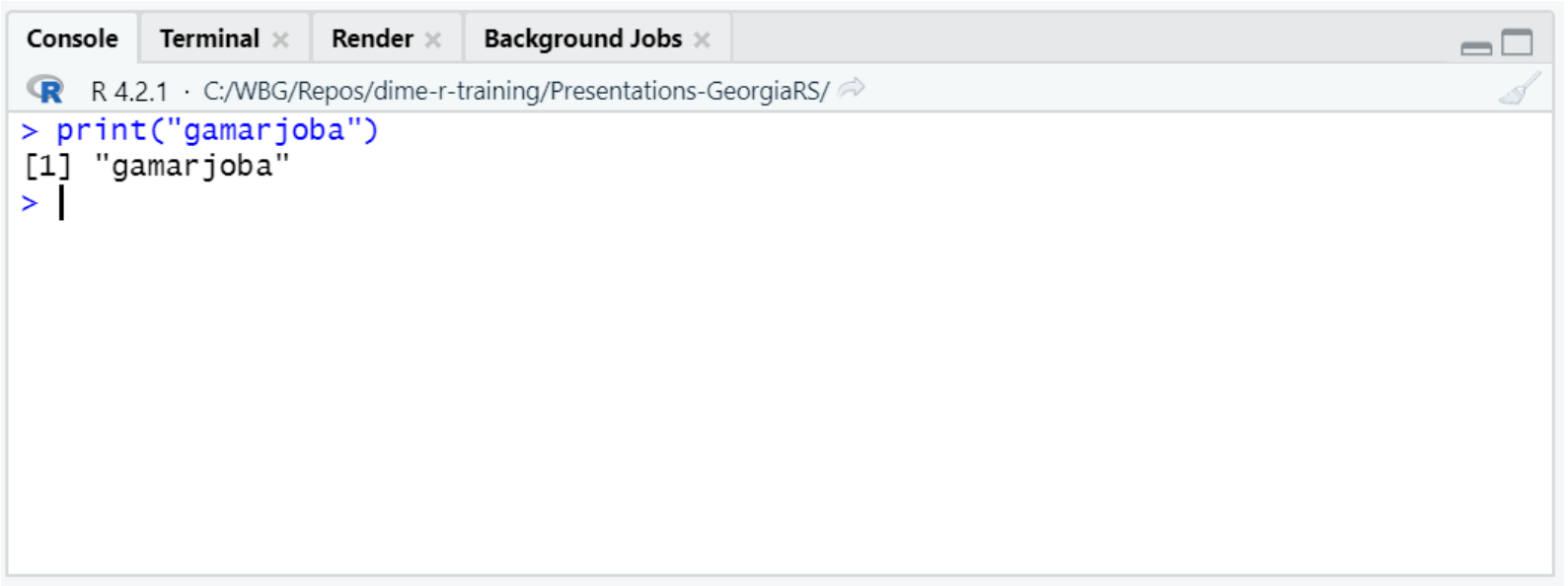
Writing R code // R კოდის დაწერა

Exercise 1: writing code in the console

1. Write the following code in the console of RStudio
    - `print("gamarjoba")`
    - Make sure to include the quotes: `" "`
2. Press Enter to run the code



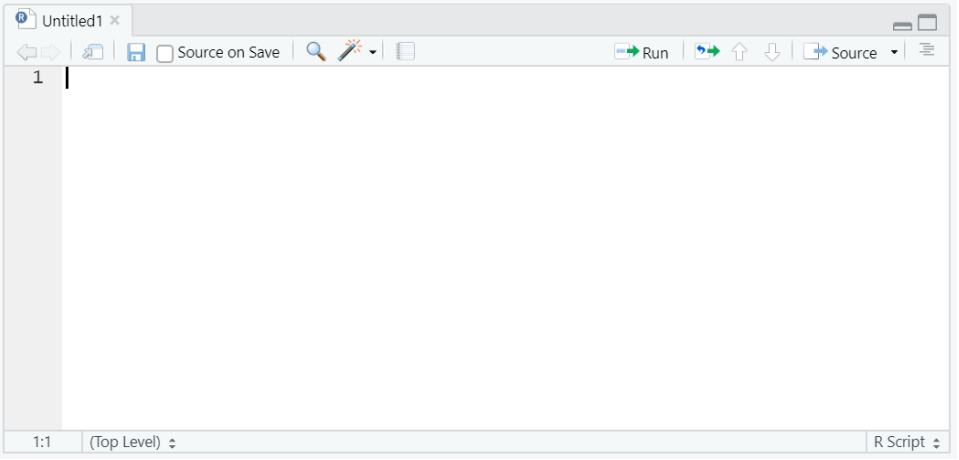
Writing R code // R კოდის დაწერა



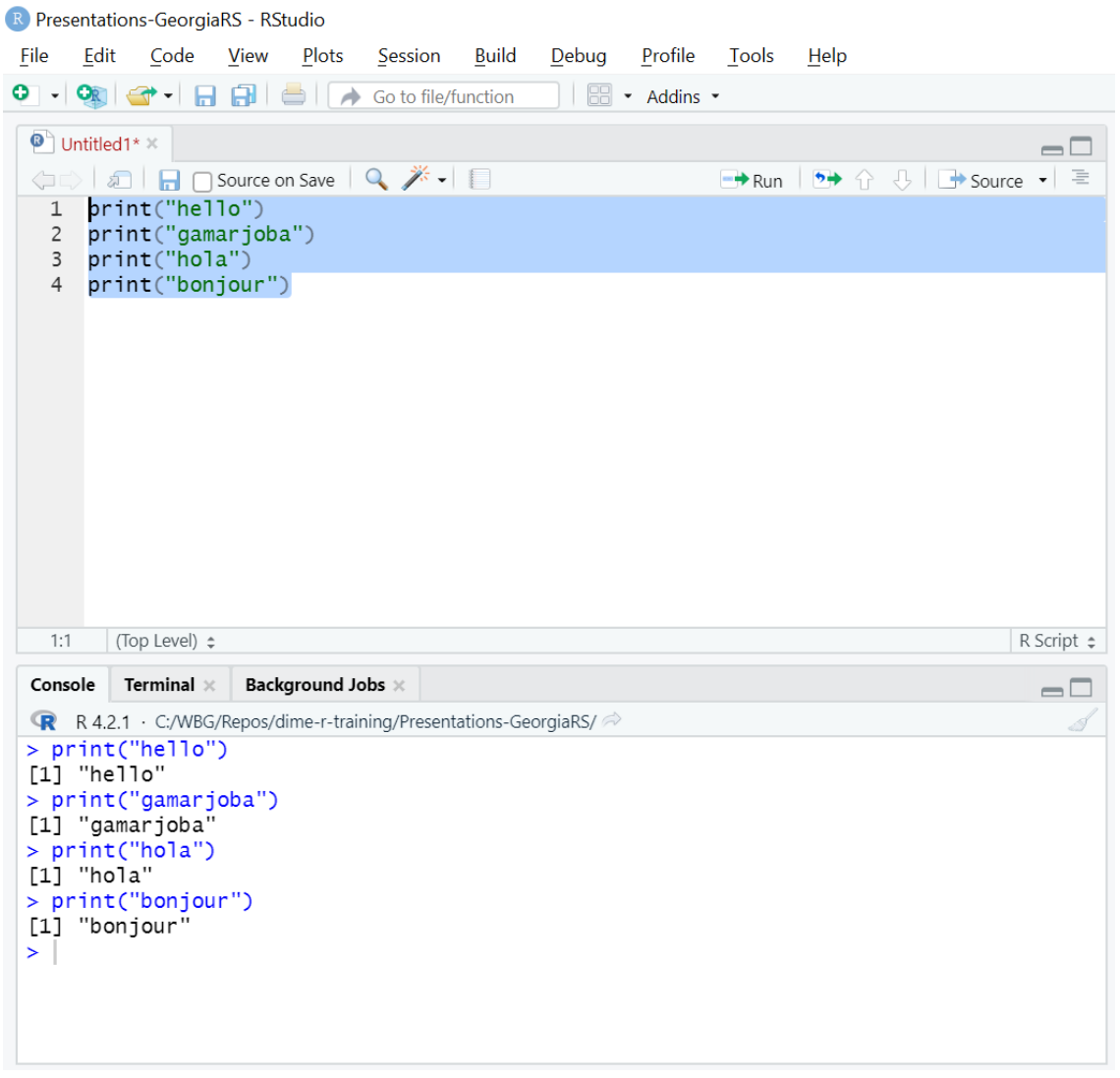
Writing R code // R კოდის დაწერა

Exercise 2: writing a short script

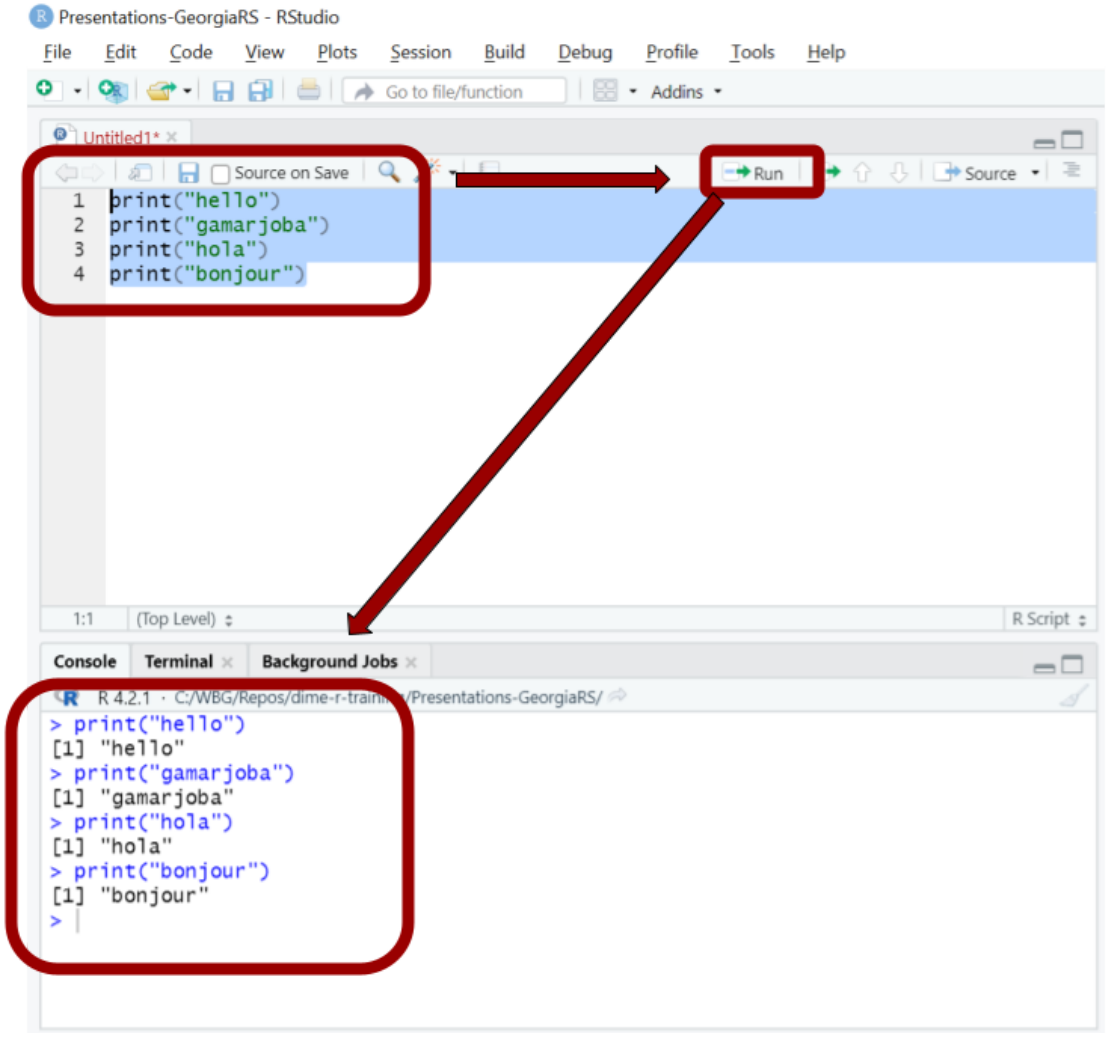
1. Write or copy the following text into the script section of RStudio
    - `print("gamarjoba")`
    - `print("hello")`
    - `print("hola")`
    - `print("bonjour")`
2. Select the text you introduced with your mouse
3. Press "Run"



Writing R code // R კოდის დაწერა



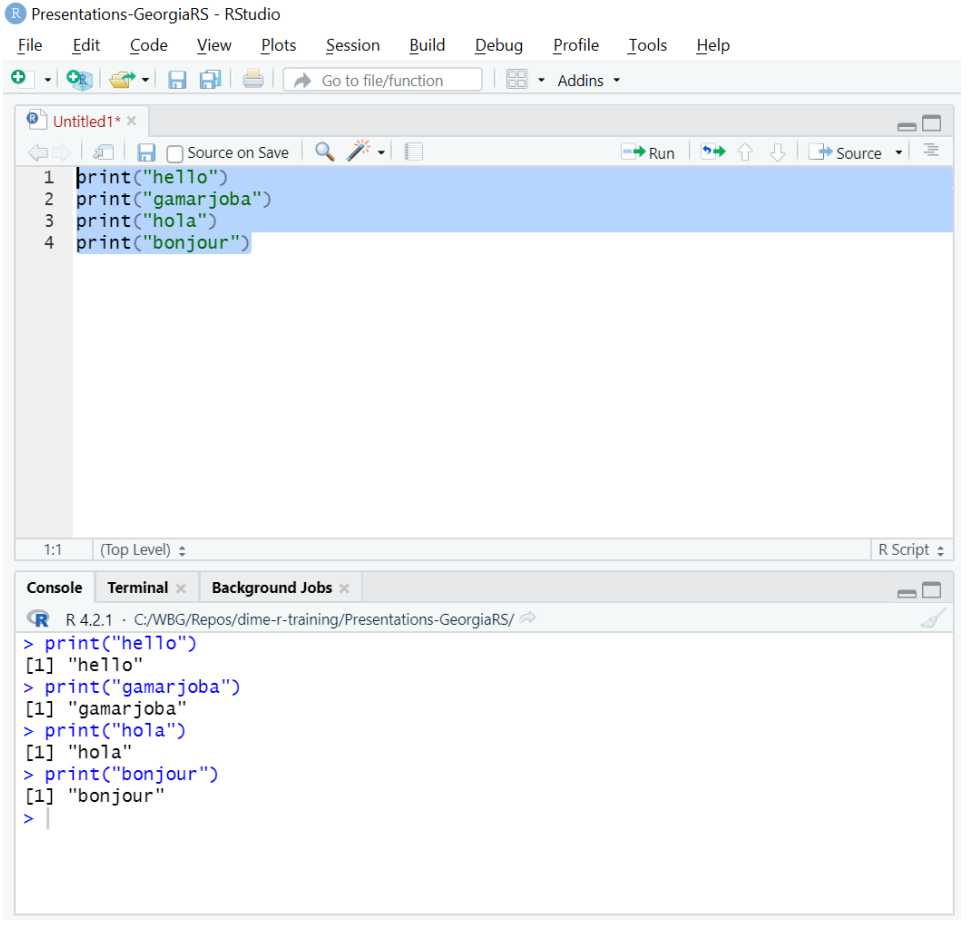
Writing R code // R კოდის დაწერა



Writing R code // R კოდის დაწერა

R scripts

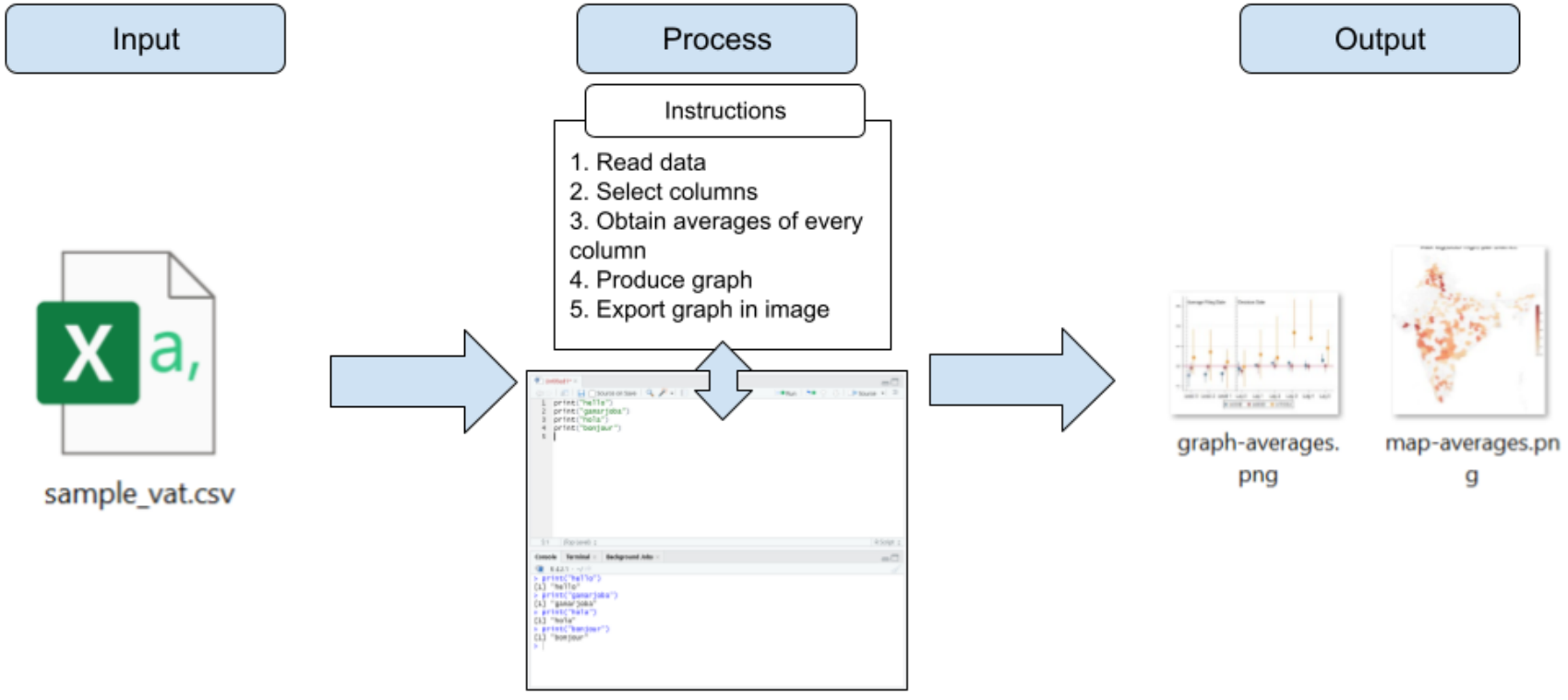
- Writing and running code from the console will execute it immediately
- Writing code in the script panel allow us to write multiple lines of code and execute them later
    - Each line is executed in order
    - The line and the results will show in the console
- **Important:** for the rest of the training, remember to always introduce your code in the script (and not in the console) so you can keep record of what you did



Writing R code // R კოდის დაწერა

R scripts

- In other words: scripts contain the instructions you give to your computer when doing data work



Writing R code // R კოდის დაწერა

Creating objects in R

- Remember we also mentioned the environment panel? that's where R keeps track of objects
- Objects are representations of data that currently exist in R's memory
    - A single number can be an object
    - A word can be an object
    - Even an entire data file can be an object
- We create objects in R with the arrow operator (`<-`)
- Example: creating an object called `x`

```
x <- 10
```

Creating objects in R

- After an object is created, we can refer to it using its name

```
x <- 10
print(x)
```

```
## [1] 10
```

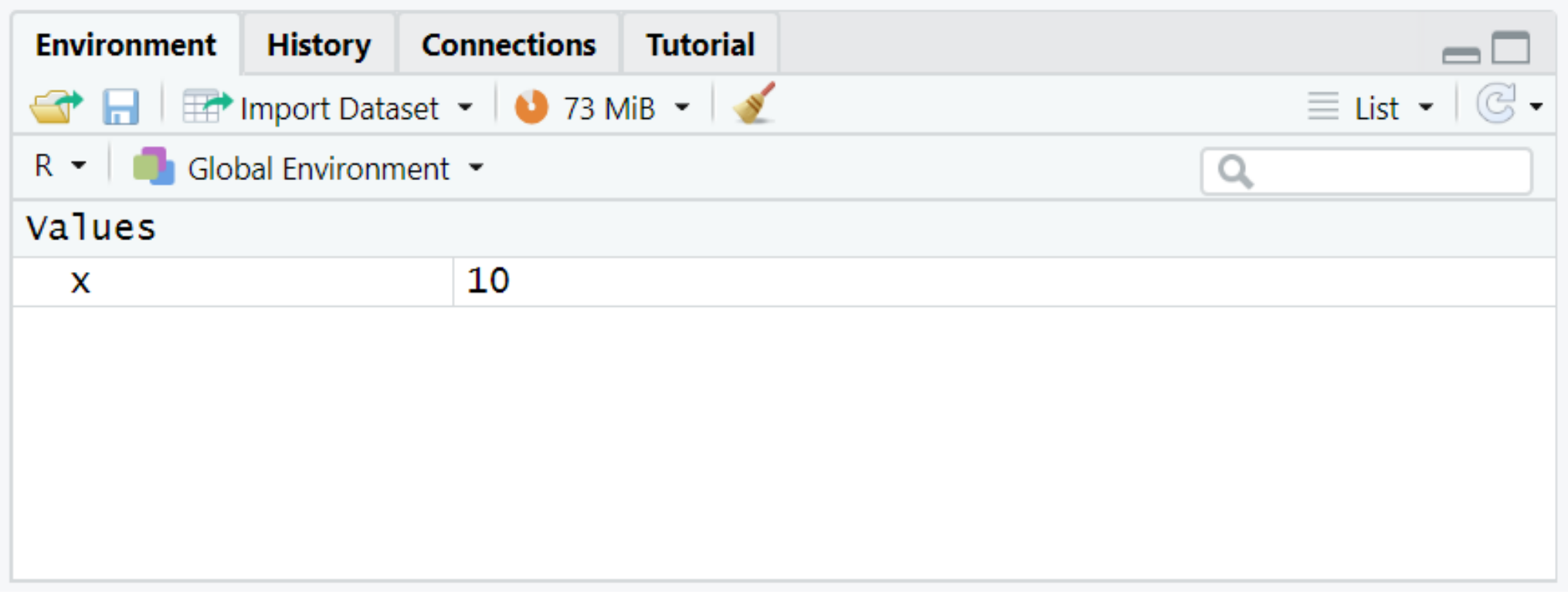
```
print(x + 5)
```

```
## [1] 15
```

Writing R code // R კოდის დაწერა

Creating objects in R

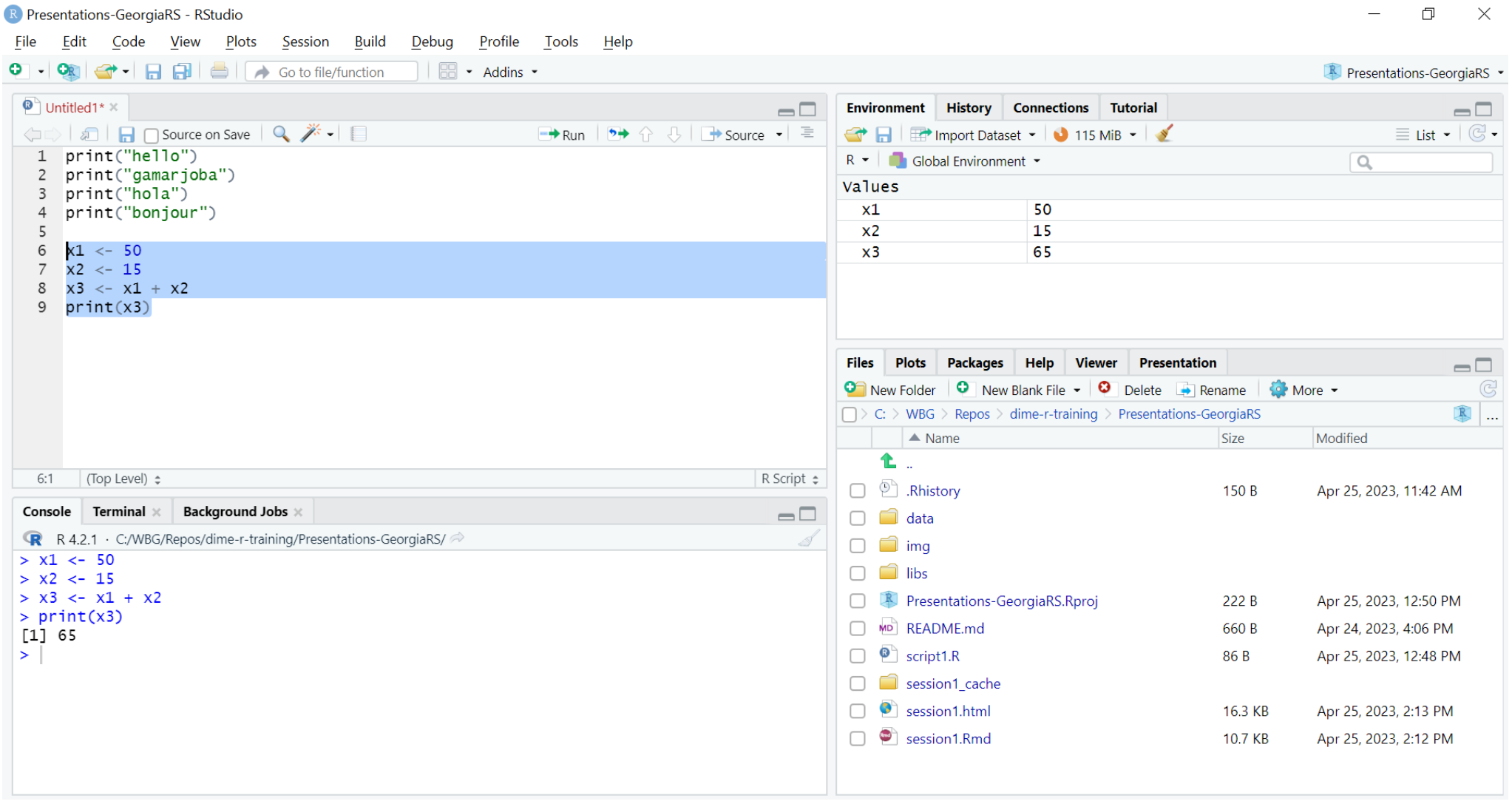
- After an object is created, it will show in the environment panel



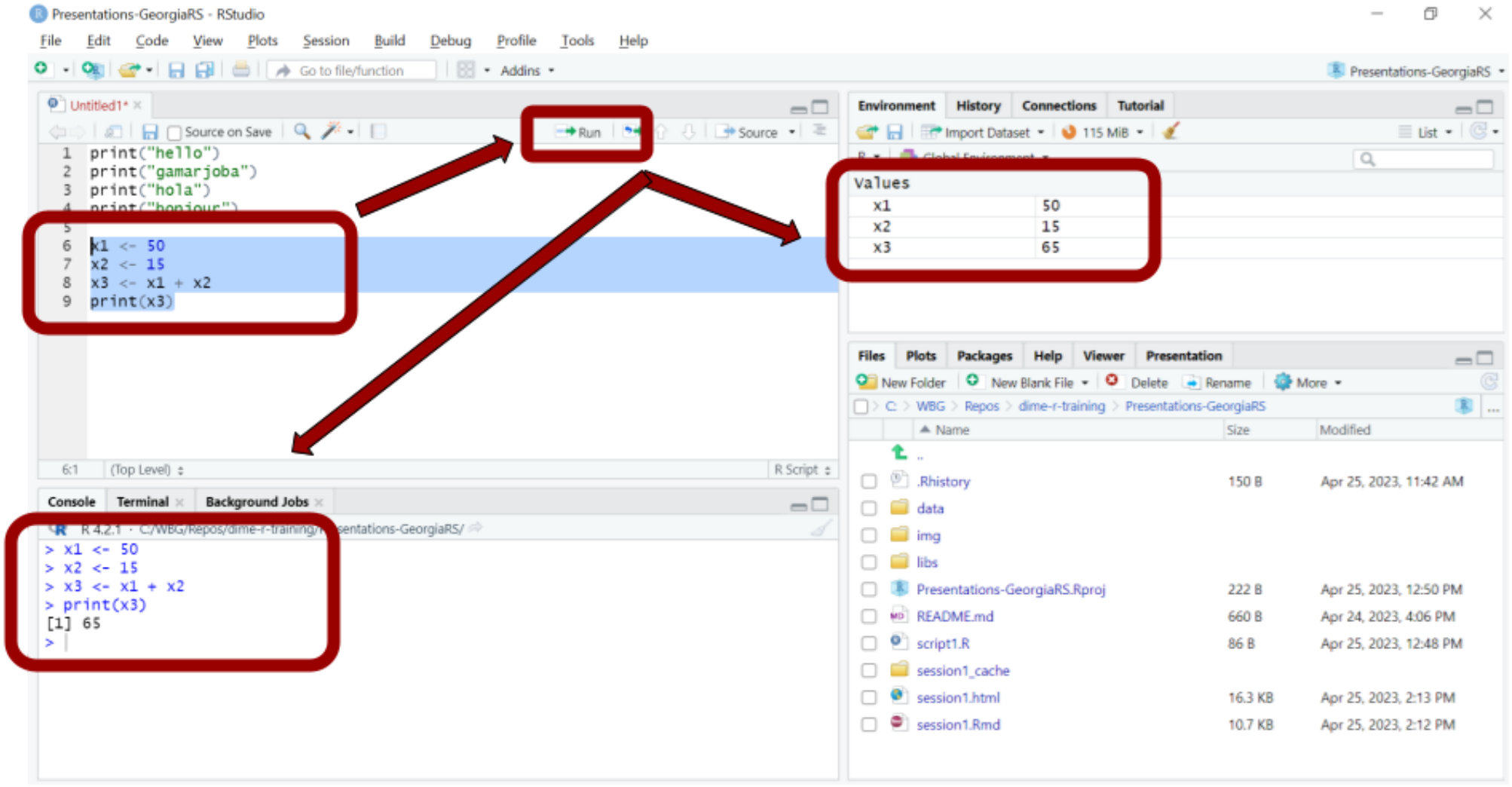
Exercise 3: operations with objects

1. Create an object called `x1` in your script and assign to it the value of 50
2. Create an object called `x2` and make it equal to 15
3. Create an object called `x3` and make it the sum of x1 and x2
4. Print `x3`. It should show the value 65

Writing R code // R კოდის დაწერა



Writing R code // R კოდის დაწერა



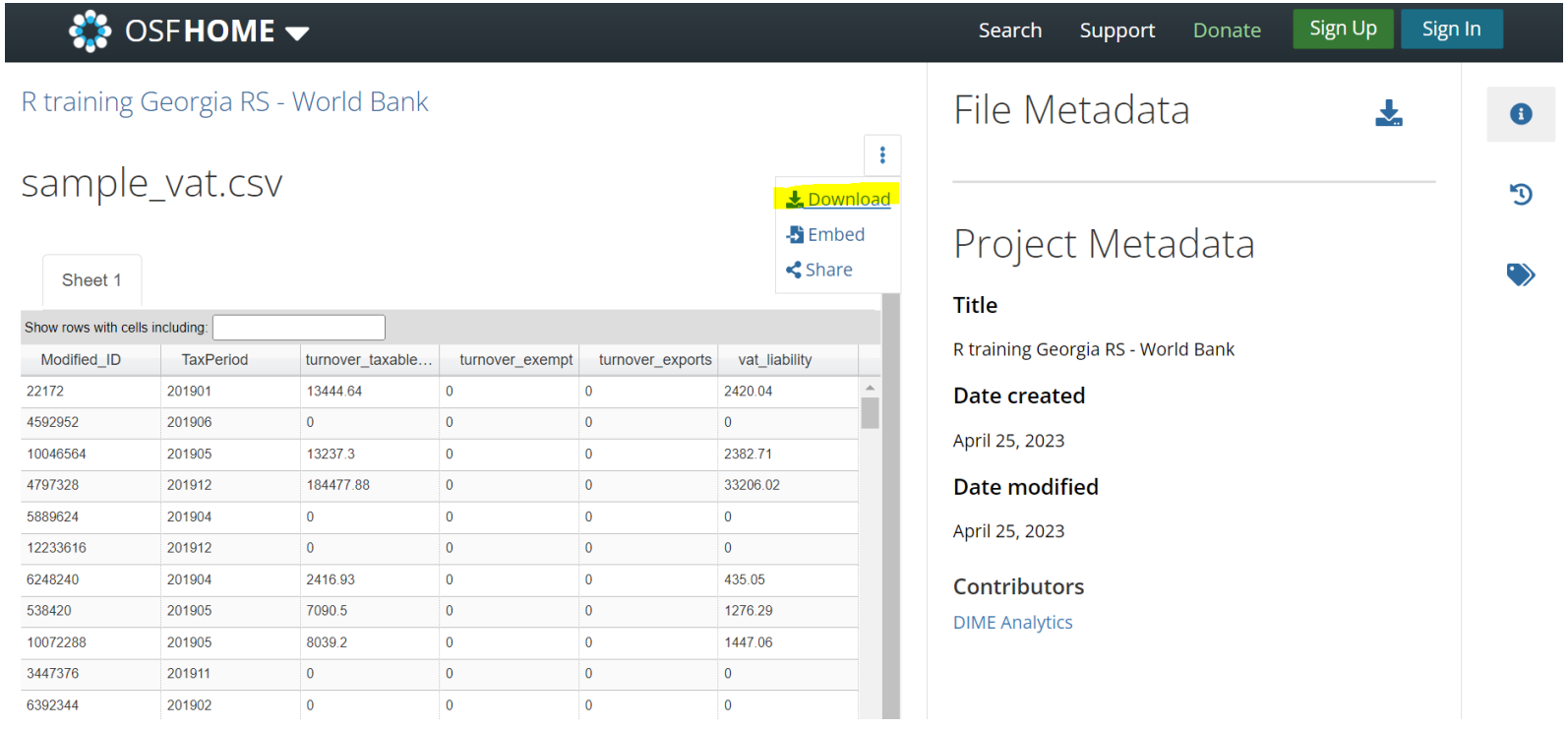
Writing R code // R კოდის დაწერა

- Now we know how to use RStudio to write R code and produce scripts
- However, we haven't still introduced the data to our data work. That comes next

Data in R // მონაცემები R

Exercise 4: Loading data into R

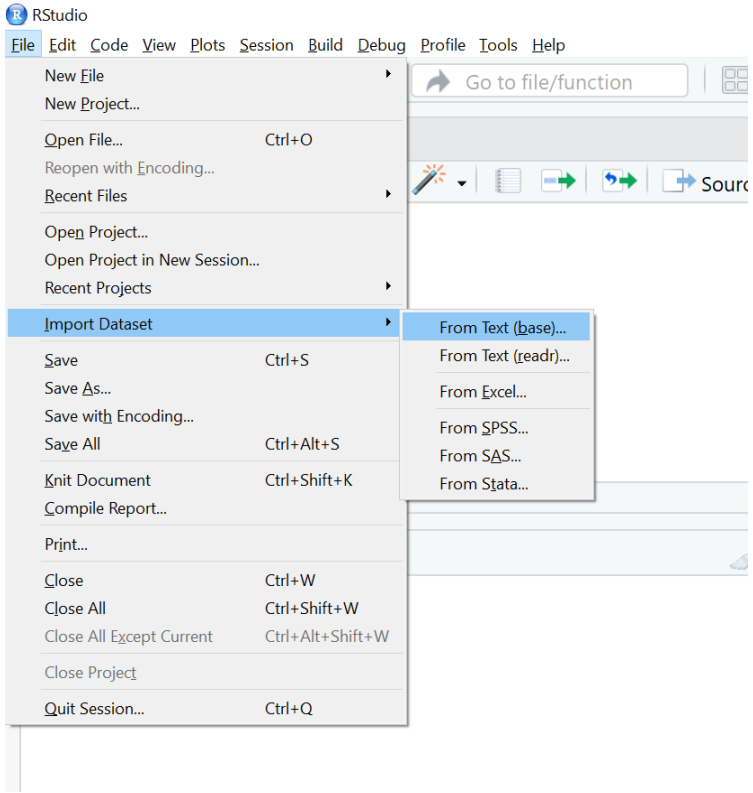
1 - Go to this page: https://osf.io/ds5w4 and download the file `sample_vat.csv`

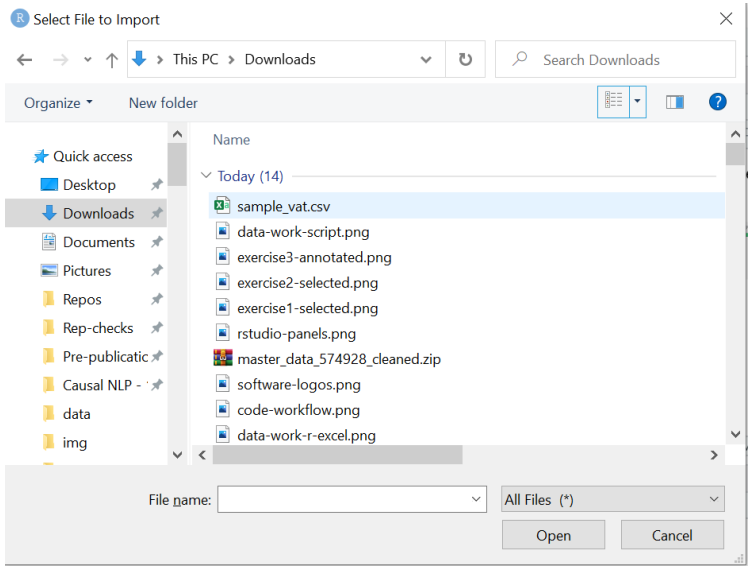


Exercise 4: Loading data into R

2 - In RStudio, go to `File` > `Import Dataset` > `From Text (base)` and select the file `sample_vat.csv`

- If you don't know where the file is, check in the `Downloads` folder

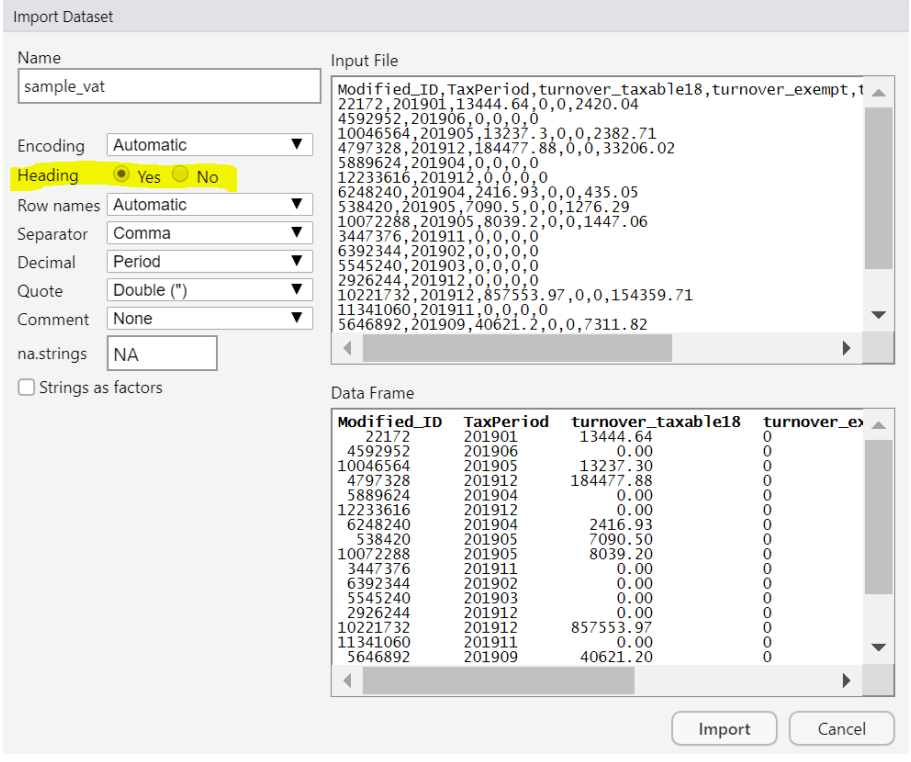




Exercise 4: Loading data into R

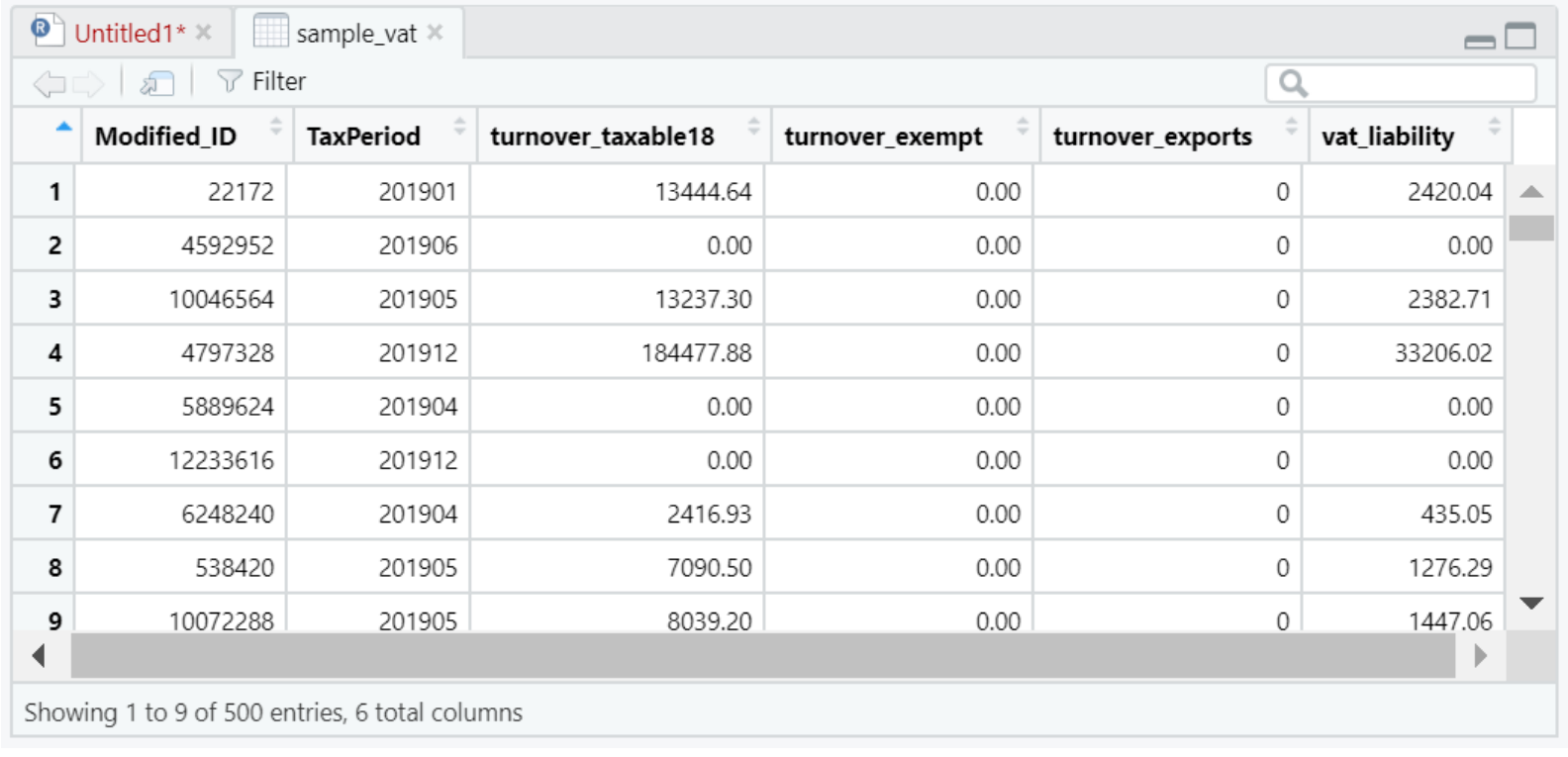
3 - Make sure to select `Heading` > `Yes` in the next window

4 - Select `Import`



Data in R // მონაცემები R

- If you did this correctly, you will note that a viewer of the data now appears in RStudio
- You can click on the `x` next to `sample_vat` to return to the script
- To open the viewer again, use the code: `View(sample_vat)` (notice the uppercase "V")

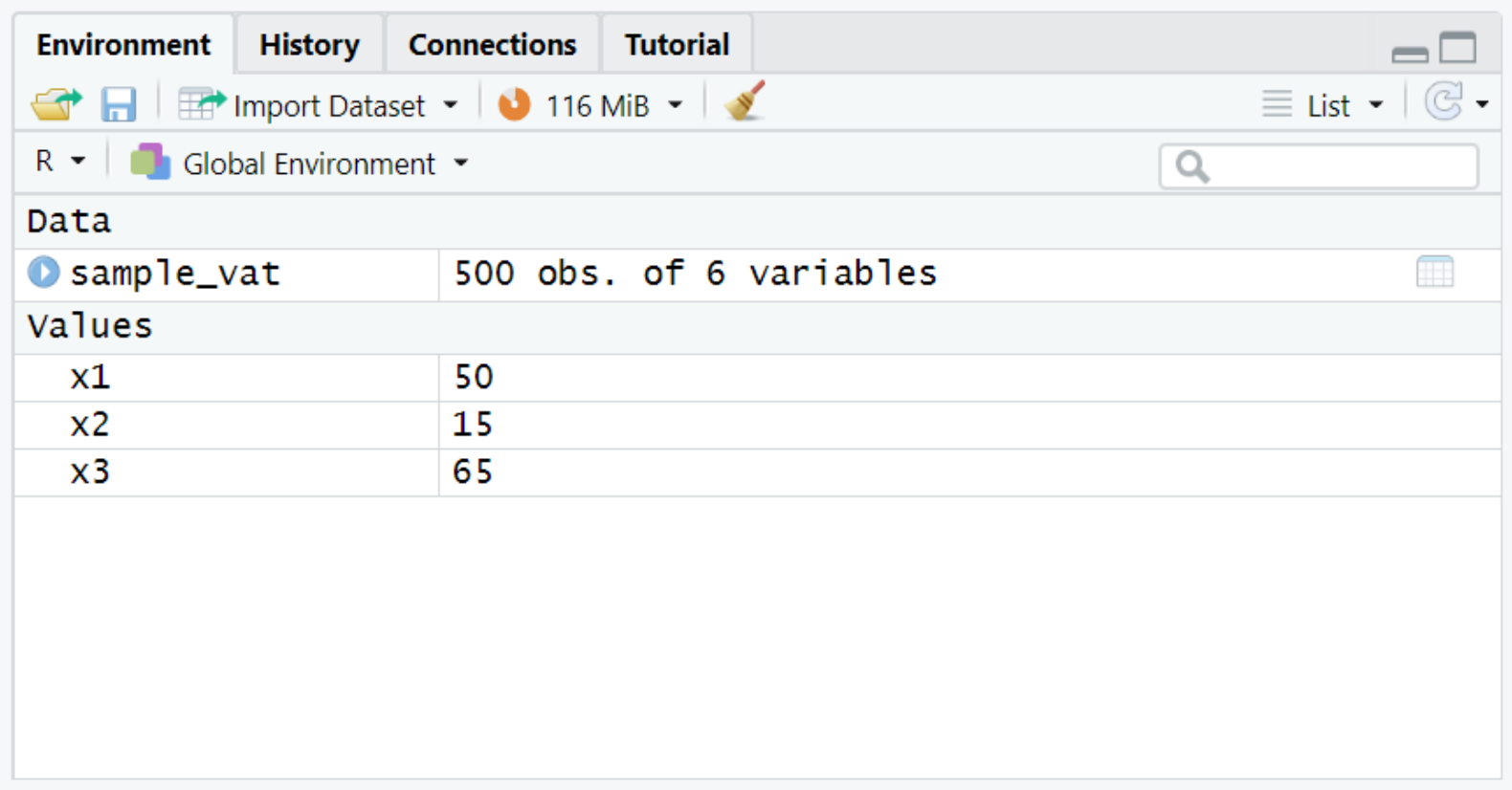


	Modified_ID	TaxPeriod	turnover_taxable18	turnover_exempt	turnover_exports	vat_liability
1	22172	201901	13444.64	0.00	0	2420.04
2	4592952	201906	0.00	0.00	0	0.00
3	10046564	201905	13237.30	0.00	0	2382.71
4	4797328	201912	184477.88	0.00	0	33206.02
5	5889624	201904	0.00	0.00	0	0.00
6	12233616	201912	0.00	0.00	0	0.00
7	6248240	201904	2416.93	0.00	0	435.05
8	538420	201905	7090.50	0.00	0	1276.29
9	10072288	201905	8039.20	0.00	0	1447.06

Showing 1 to 9 of 500 entries, 6 total columns

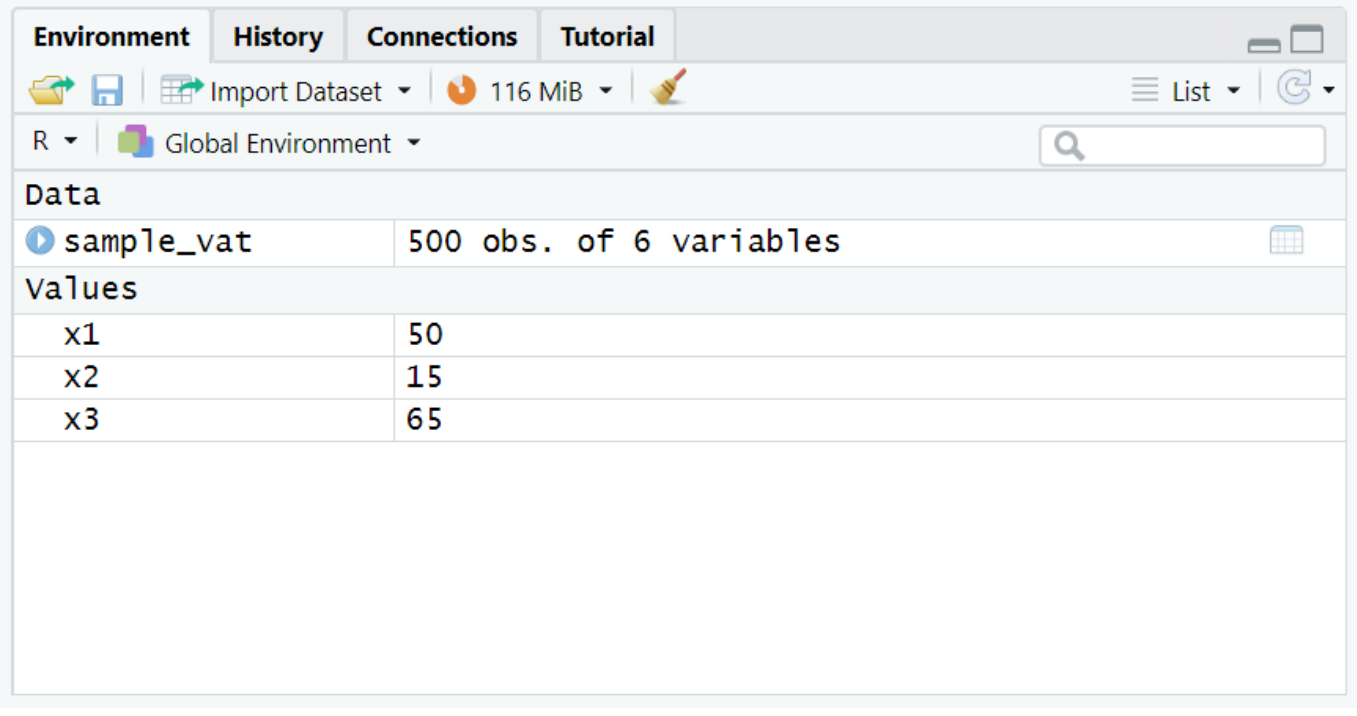
Data in R // მონაცემები R

- Additionally, you will now see an object named `sample_vat` in your environment



Data in R // მონაცემები R

- Remember we mentioned objects before? For R, `sample_vat` is an object just like `x1`, `x2`, or `x3`
- The difference is that `sample_vat` is not a single number like `x1`, but a collection of numeric values similar to an Excel spreadsheet. In R, this type of objects are called **dataframes**
- From now, we will refer to data loaded into R as **dataframes**



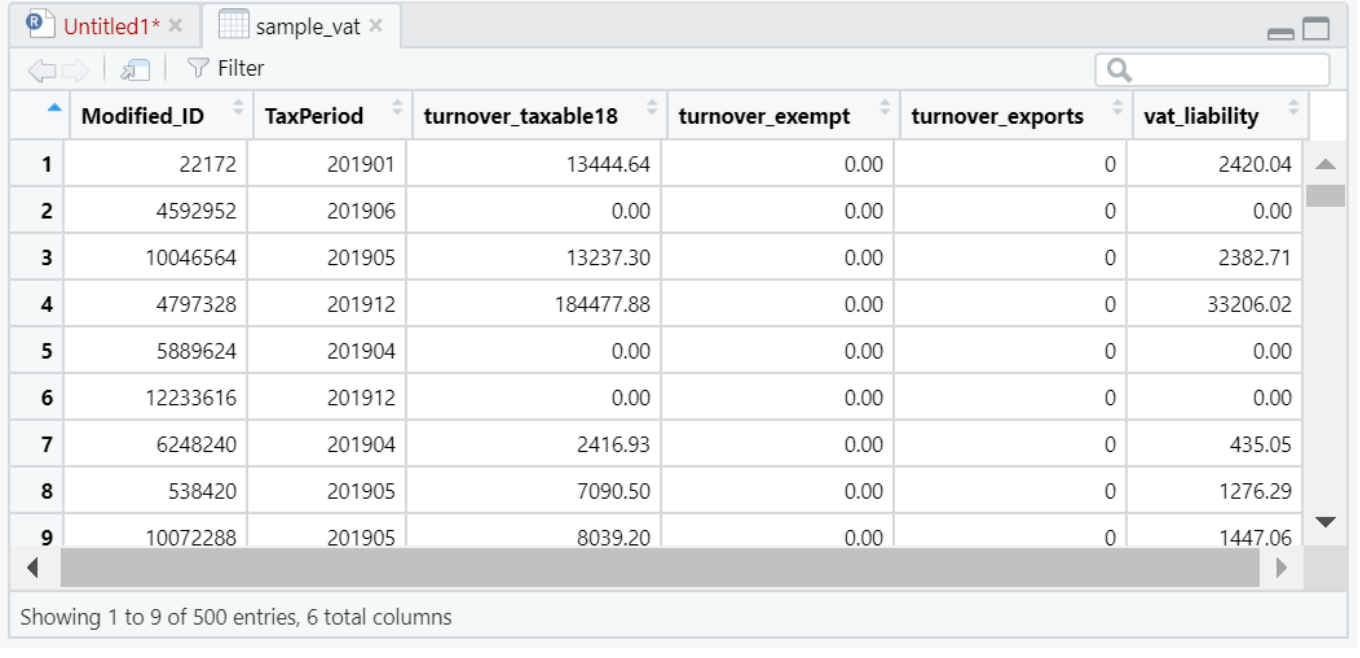
Data in R // მონაცემები R

- Since dataframes are also objects, we can refer to them with their names (exm: `sample_vat`)
- We'll see an example of that in the next exercise

A note about this dataframe

Understanding the data you use is very important. For this training, `sample_vat` is a "toy" dataframe that simulates tax data.

- `Modified_ID` is a taxpayer identifier
- `TaxPeriod` is a month variable (year + month)
- The rest are tax-related variables that we are not going to focus on in this session



Untitled1* | sample_vat

Filter

	Modified_ID	TaxPeriod	turnover_taxable18	turnover_exempt	turnover_exports	vat_liability
1	22172	201901	13444.64	0.00	0	2420.04
2	4592952	201906	0.00	0.00	0	0.00
3	10046564	201905	13237.30	0.00	0	2382.71
4	4797328	201912	184477.88	0.00	0	33206.02
5	5889624	201904	0.00	0.00	0	0.00
6	12233616	201912	0.00	0.00	0	0.00
7	6248240	201904	2416.93	0.00	0	435.05
8	538420	201905	7090.50	0.00	0	1276.29
9	10072288	201905	8039.20	0.00	0	1447.06

Showing 1 to 9 of 500 entries, 6 total columns

Exercise 5: Subset the data

1. Use the following code to subset `sample_vat` and leave only the observations in January 2019: `subset(sample_vat, TaxPeriod == 201901)`
    - Note that the "T" and "P" in `TaxPeriod` are **uppercase**
    - Note that there are **two equal signs** in the condition, not one
2. Use `View(sample_vat)` to visualize the dataframe again and see how it changed (note the uppercase "V")

Data in R // მონაცემები R

Does anything look strange?

Untitled1* | sample_vat

Filter

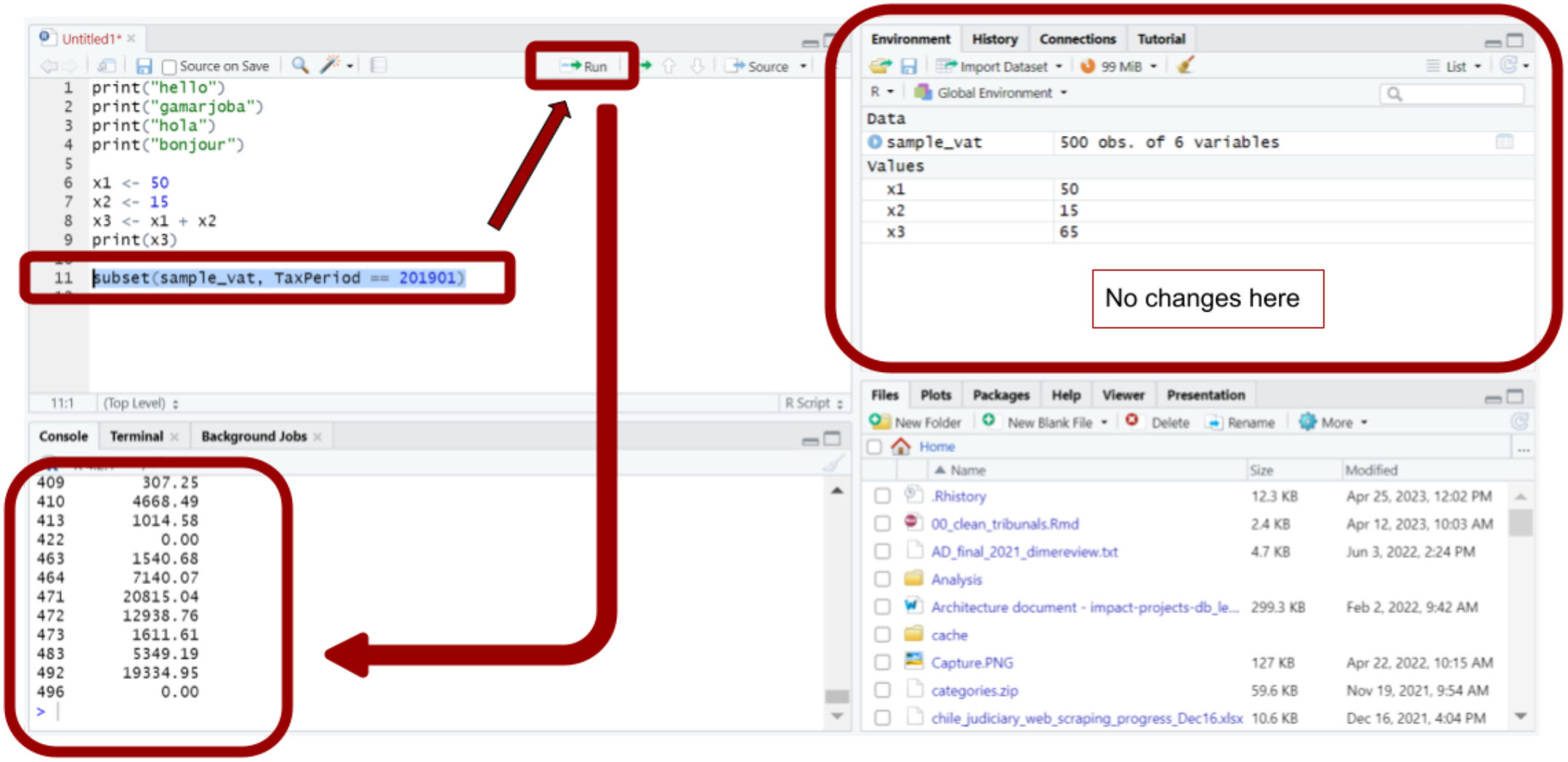
	Modified_ID	TaxPeriod	turnover_taxable18	turnover_exempt	turnover_exports	vat_liability
1	22172	201901	13444.64	0.00	0	2420.04
2	4592952	201906	0.00	0.00	0	0.00
3	10046564	201905	13237.30	0.00	0	2382.71
4	4797328	201912	184477.88	0.00	0	33206.02
5	5889624	201904	0.00	0.00	0	0.00
6	12233616	201912	0.00	0.00	0	0.00
7	6248240	201904	2416.93	0.00	0	435.05
8	538420	201905	7090.50	0.00	0	1276.29
9	10072288	201905	8039.20	0.00	0	1447.06

Showing 1 to 9 of 500 entries, 6 total columns

Data in R // მონაცემები R

- Indeed, the dataframe `sample_vat` didn't change
- That is because we didn't use the arrow operator (`<-`) to store the result in an object
- Instead, R only printed the result in the console for us (and nothing else)

Data in R // მონაცემები R



Exercise 6: Subsetting and storing the result

1. Subset the dataframe again and save the result into a new dataframe called `sample_vat_201901`. You can use this code for that: `sample_vat_201901 <- subset(sample_vat, TaxPeriod == 201901)`

2. Use `View(sample_vat_201901)` to visualize the result (notice the uppercase "V")

Data in R // მონაცემები R

Now the resulting dataframe looks correct!

Untitled1* | sample_vat_201901

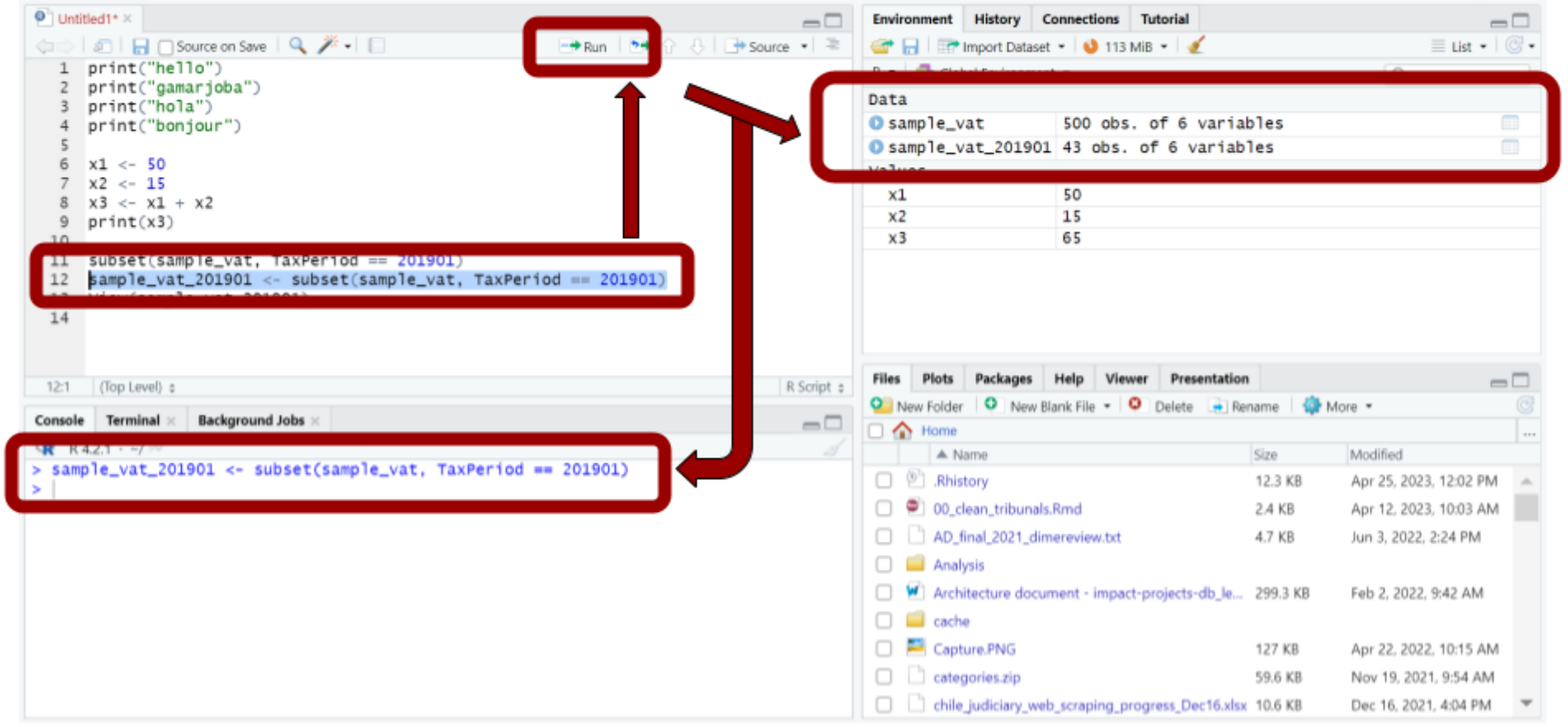
Filter

	Modified_ID	TaxPeriod	turnover_taxable18	turnover_exempt	turnover_exports	vat_liability
1	22172	201901	13444.64	0.00	0	2420.04
35	10829692	201901	0.00	0.00	0	0.00
37	11119000	201901	0.00	0.00	0	0.00
43	10074084	201901	25998.13	0.00	0	4679.66
55	1056200	201901	15369.18	0.00	0	2766.45
76	10174704	201901	0.00	0.00	0	0.00
108	4469128	201901	6695.25	0.00	0	1205.15
140	3594132	201901	397468.56	0.00	0	71544.34
144	10874928	201901	71489.97	0.00	0	12868.19

Showing 1 to 9 of 43 entries, 6 total columns

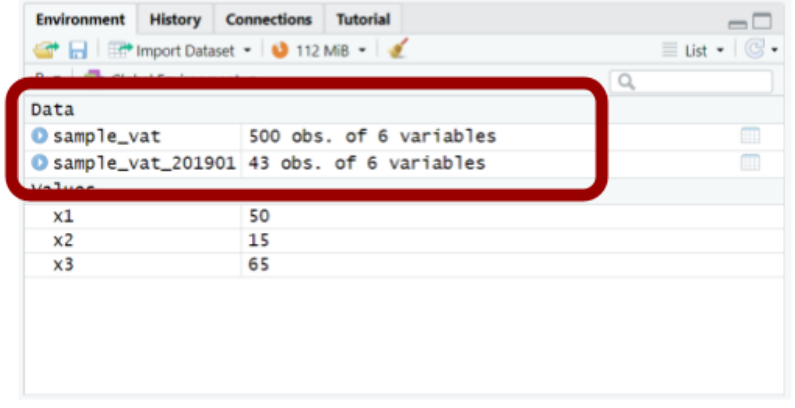
Data in R // მონაცემები R

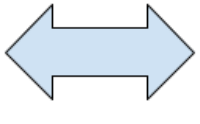
Note that this time R didn't print the resulting dataframe in the console, it only showed the code we were running. Also, now the new dataframe appears in the environment.

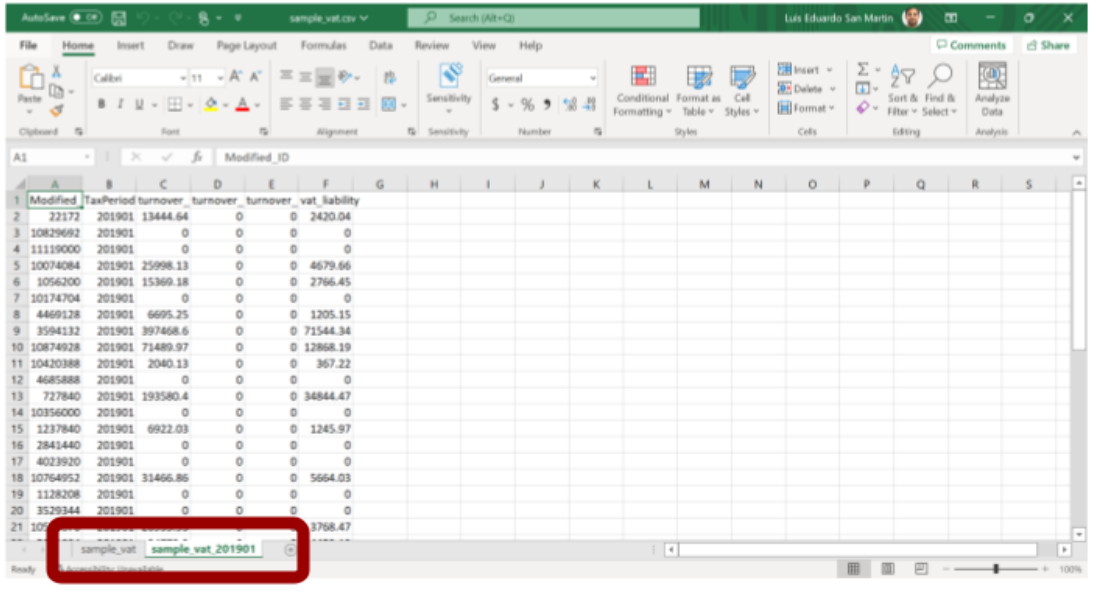


Data in R // მონაცემები R

- R can store multiple dataframes in the environment. This is analogous to having different spreadsheets in the same Excel window
- Always that dataframes are just objects in R. R differentiates which dataframe the code refers to with the dataframe name







Functions in R // ფუნქციები R

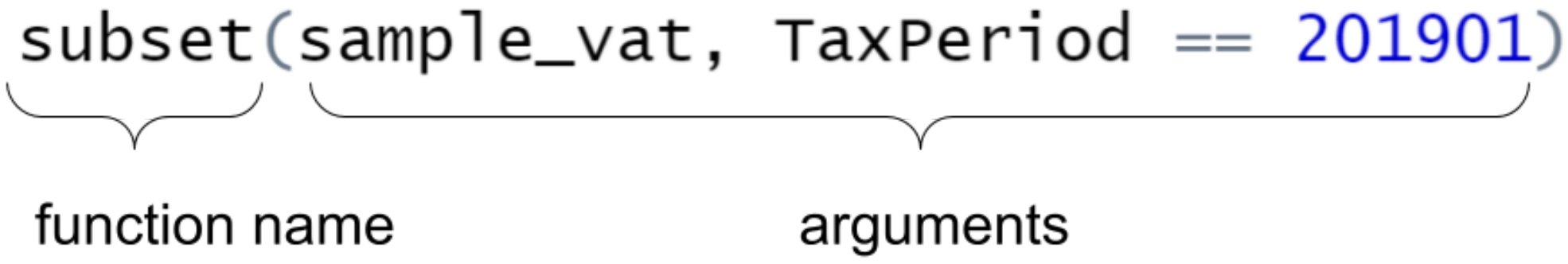
Functions in R // ფუნქციები R

- Functions are how we apply operations to objects in R
- We have used a few functions in the previous exercises. For example, `subset()` and `View()` are functions
- Everything that has a name plus parentheses is a function in R

```
subset(sample_vat, TaxPeriod == 201901)
```

Functions in R // ფუნქციები R

Functions have the following syntax:



- **Function name:** the name we use to call a function. It goes before the parentheses
- **Arguments:** inputs and specifications for the function to be applied.
  - Arguments go inside the parentheses
  - The first argument is the object you apply the function on

Functions in R // ფუნქციები R

- The results of a function can always be stored in an object with the arrow operator (`<-`)

```
sample_vat_201901 <- subset(sample_vat, TaxPeriod == 201901)
```

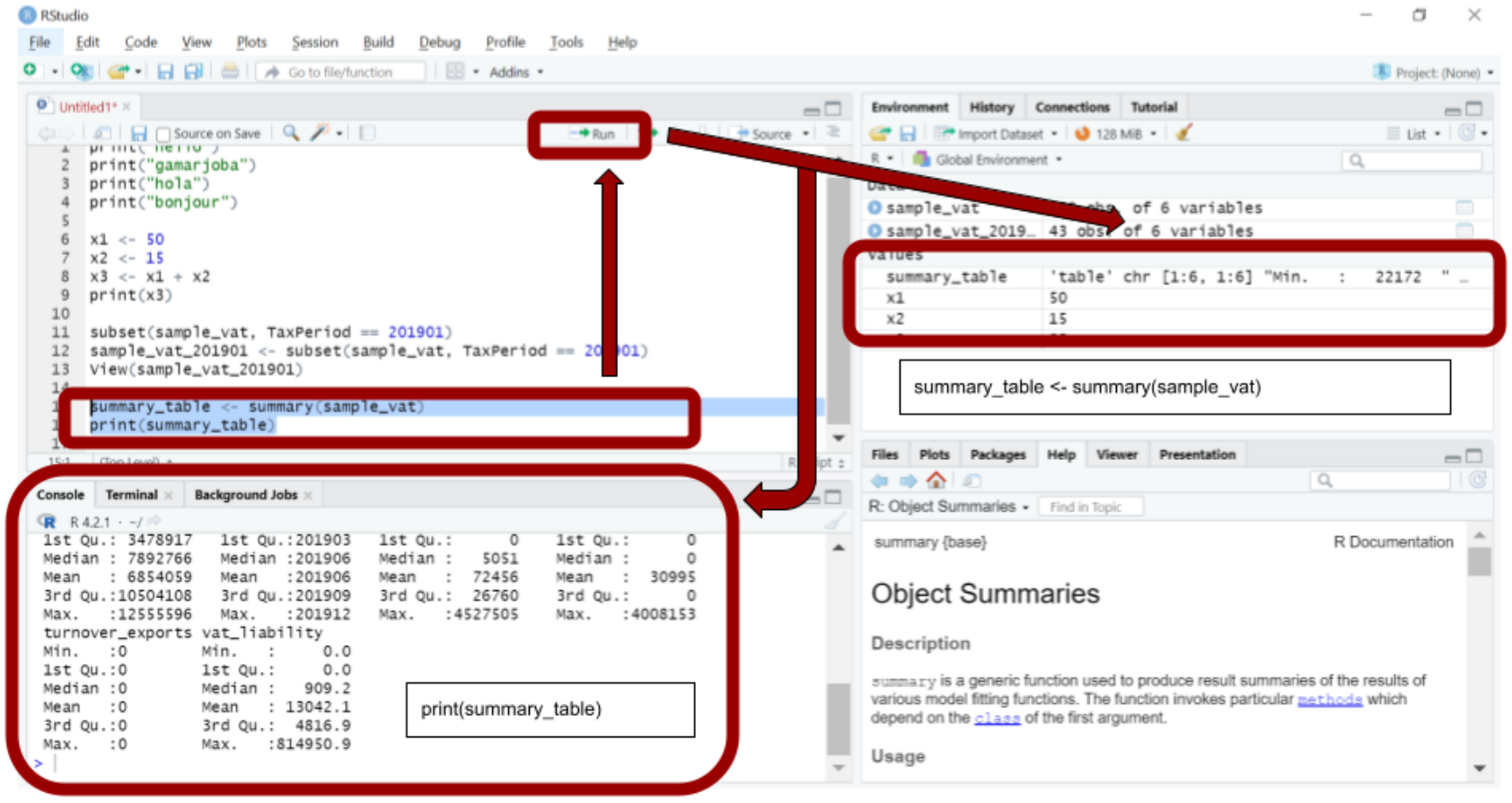
- As we saw in exercise 5, the results of a function will only be printed in the console if you don't store them

Exercise 7: Using the function `summary()`

1. Compute the summary statistics of the variables of `sample_vat` and save the result with `summary_table <- summary(sample_vat)`
2. Print the stored result with `print(summary_table)`

Functions in R // ფუნქციები R

Note that this code is both creating a new object (with `summary_table <- summary(sample_vat)`) and printing the result in the console (with `print(summary_table)`)



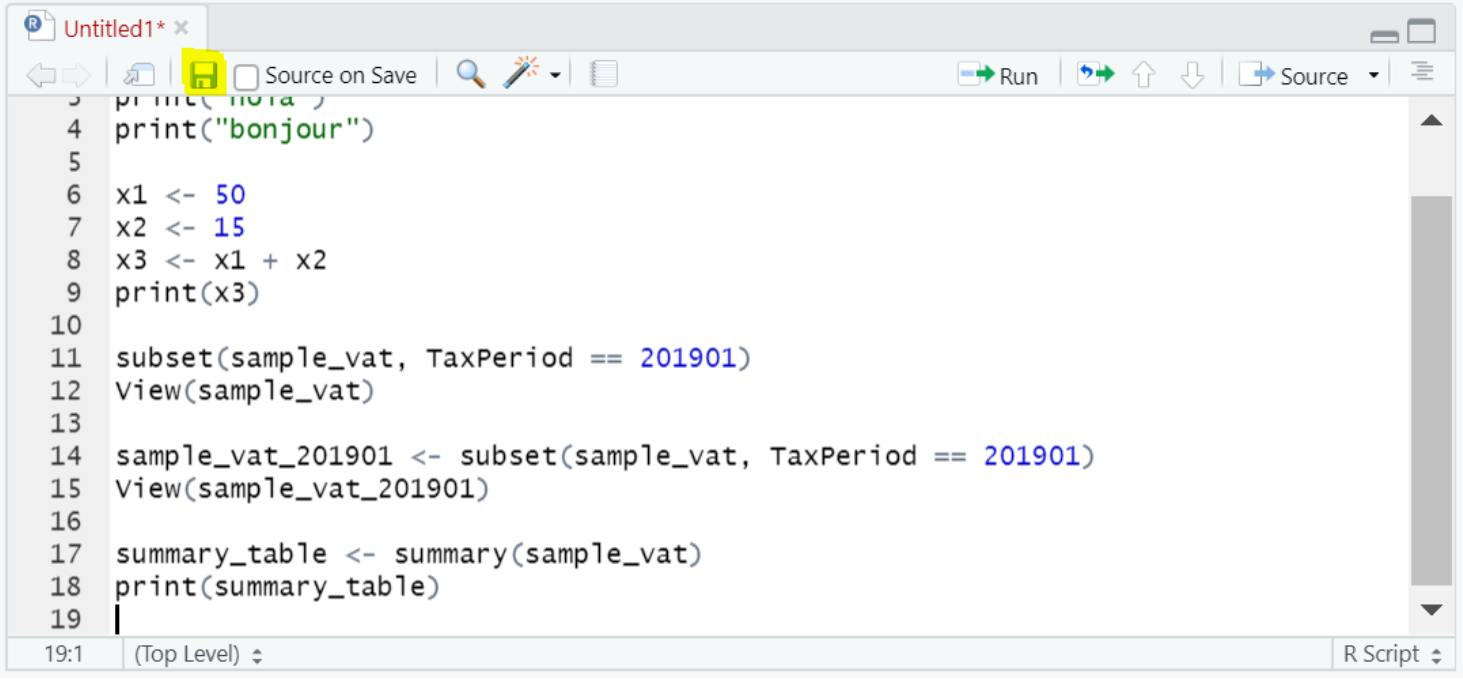
Questions? // კითხვები?

Wrapping up // შეფუთვა

Wrapping up // შეჯუმვა

Always save your work!

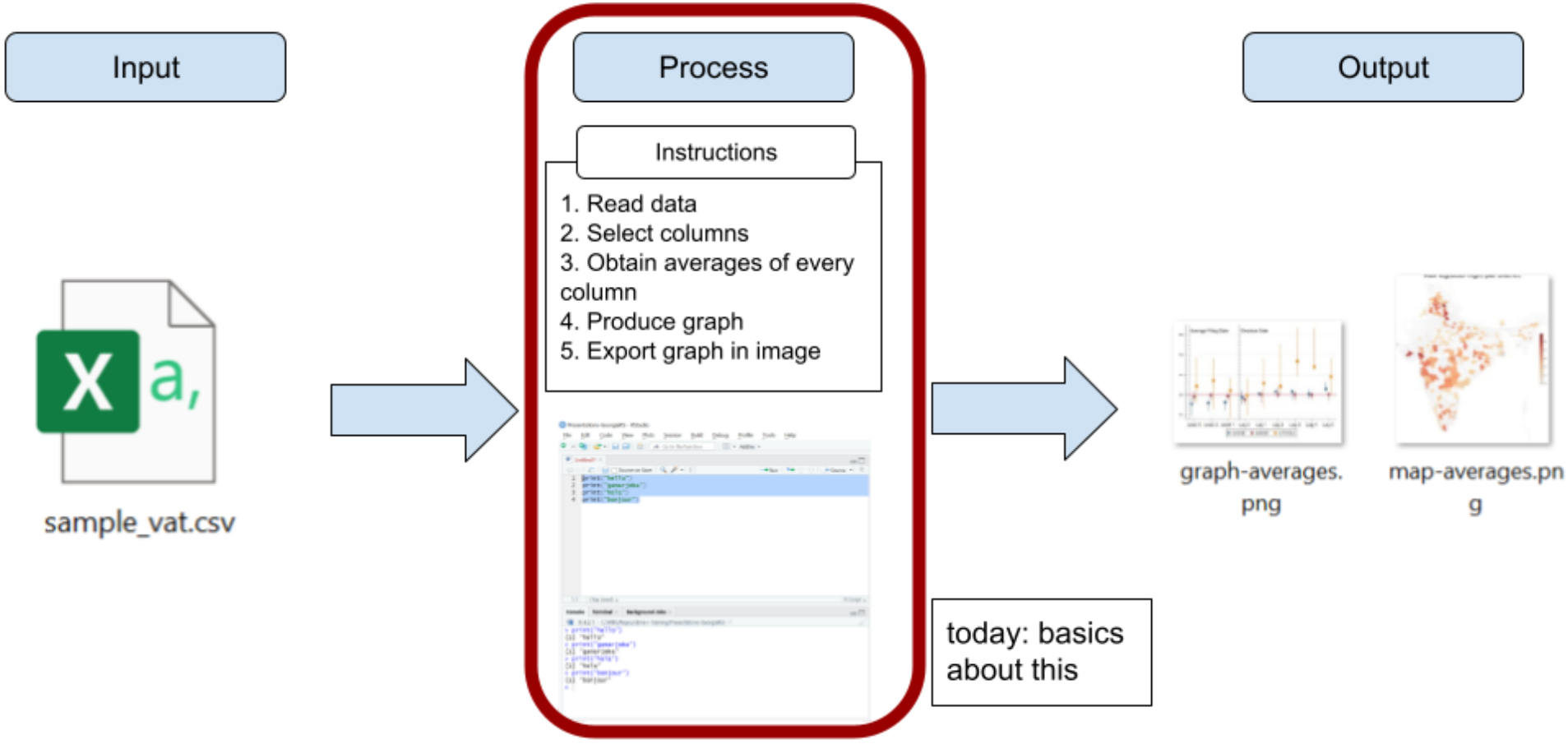
- Click the floppy disk icon to save your work
- Select a location for your file and remember where you're saving it



Wrapping up // შეფუთვა

Today

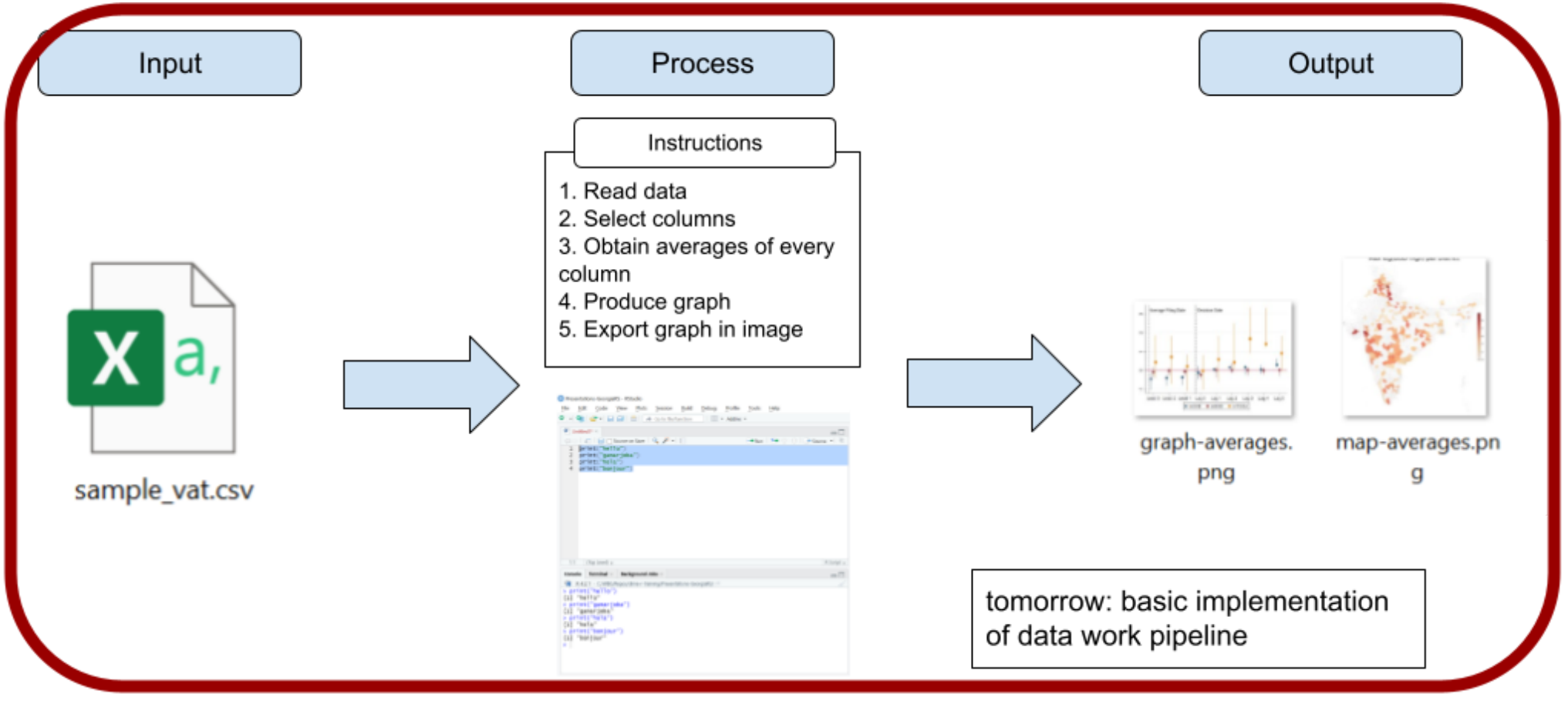
Today we focused on the basics for writing R code



Wrapping up // შეჯამება

Tomorrow

Tomorrow we'll review a few simple examples of the entire data work pipeline



Thanks! // მადლობა! // ¡Gracias!