1.3.2: Data Wrangling

(Asynchronous-Online)

Session Objectives

1. To read in data.
2. To subset data.
3. To select, create, and modify variables.
4. To filter the observations according to a certain condition.
5. To explore and summarize data.
6. To run descriptive statistics.

Key points to cover:

1. Import data.
2. Introduce to tidyverse.
3. Summarize data and run descriptive statistics.
4. Introduce other resources (e.g., books, blogs, or websites) trainees can refer to.

---

0. Prework - Before You Begin

Install Packages

Before you begin, please go ahead and install the following packages - these are all on CRAN, so you can install them using the RStudio Menu Tools/Install Packages interface:

- `haven`
- `readr`
- `readxl`
- `dplyr`

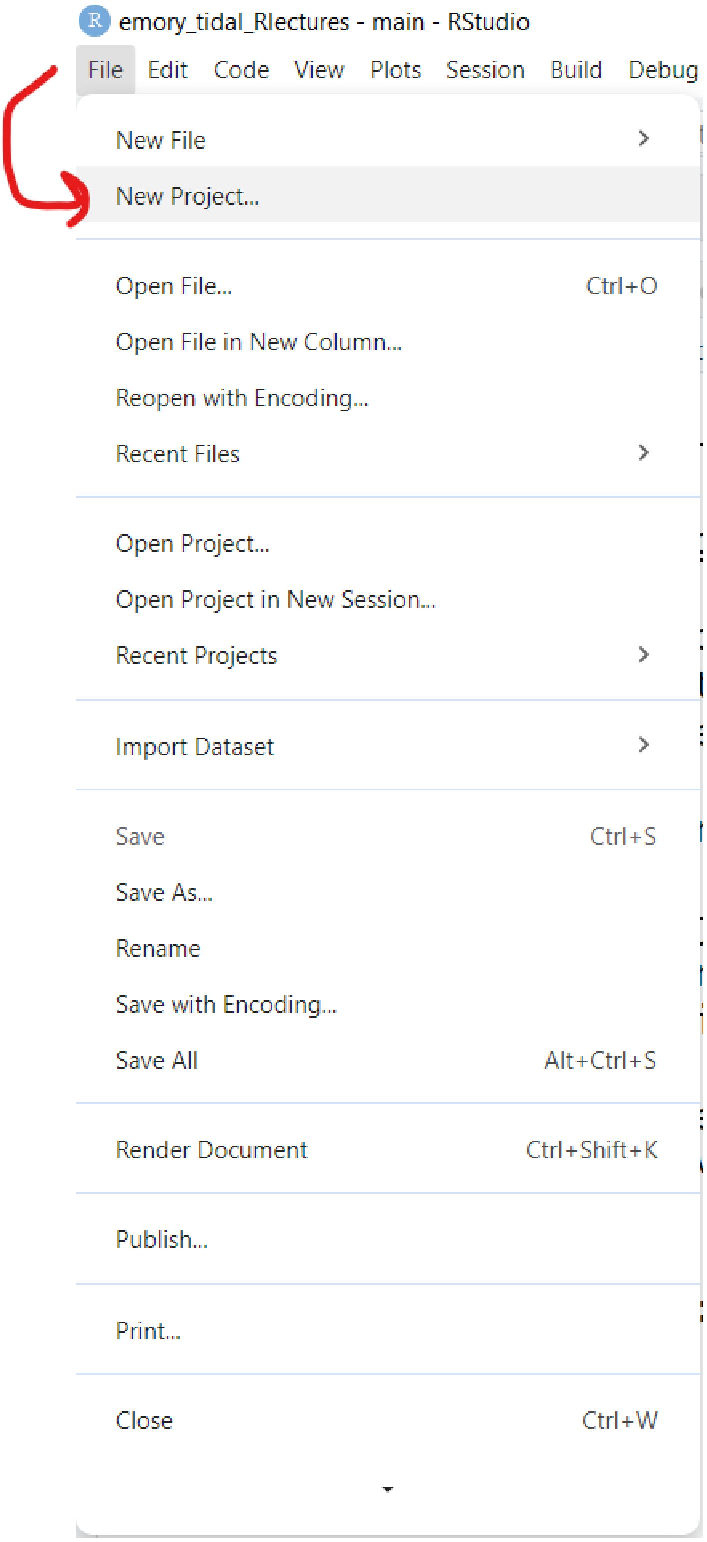
See Module 1.3.1 on Installing Packages

1. To read in data.

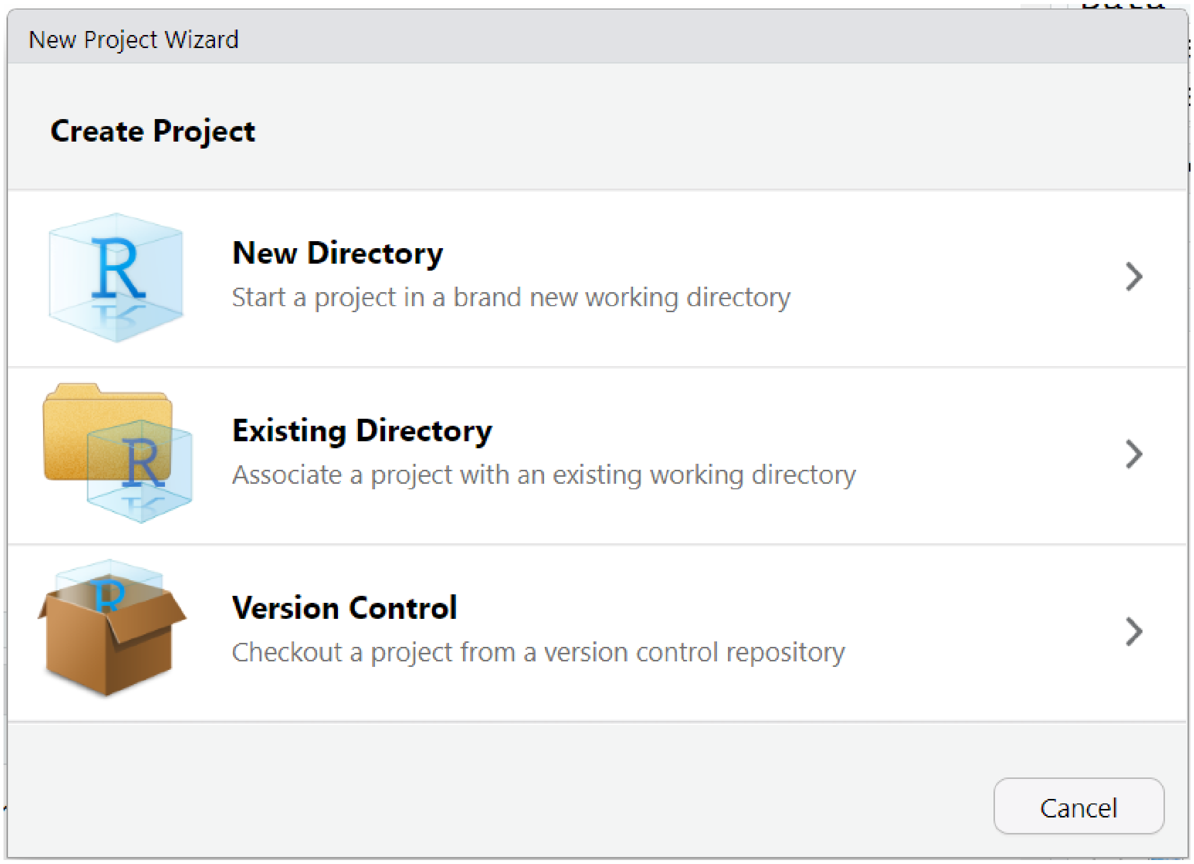
Begin with a NEW RStudio Project

Let's begin with a new RStudio Project.

1. First click on the menu at top for File/New Project:


emory_tidal_Rlectures - main - RStudio
File
Edit
Code
View
Plots
Session
Build
Debug
New File
New Project...
Open File... Ctrl+O
Open File in New Column...
Reopen with Encoding...
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save Ctrl+S
Save As...
Rename
Save with Encoding...
Save All Alt+Ctrl+S
Render Document Ctrl+Shift+K
Publish...
Print...
Close Ctrl+W

2. Next choose either an "Existing Directory" or "New Directory" depending on whether you want to use a folder that already exists on your computer or you want to create a new folder.



New Project Wizard

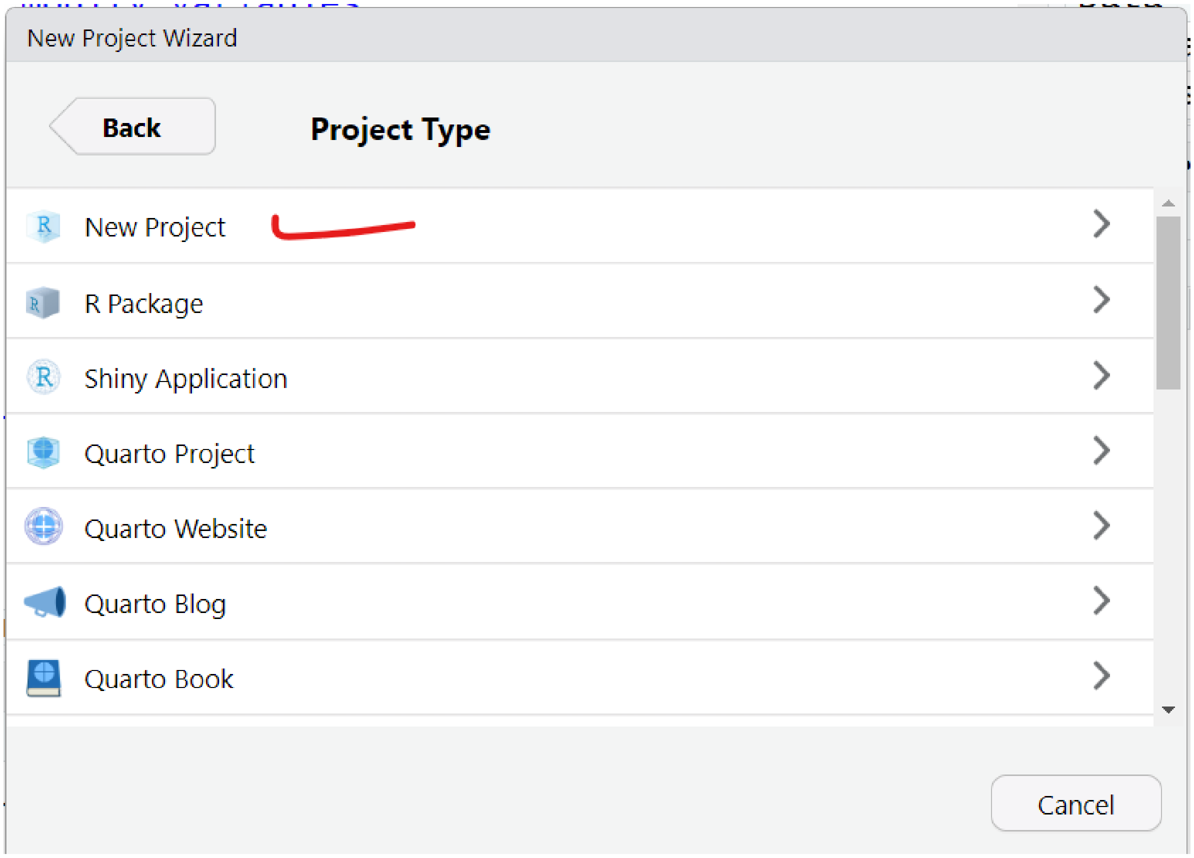
Create Project

New Directory
Start a project in a brand new working directory

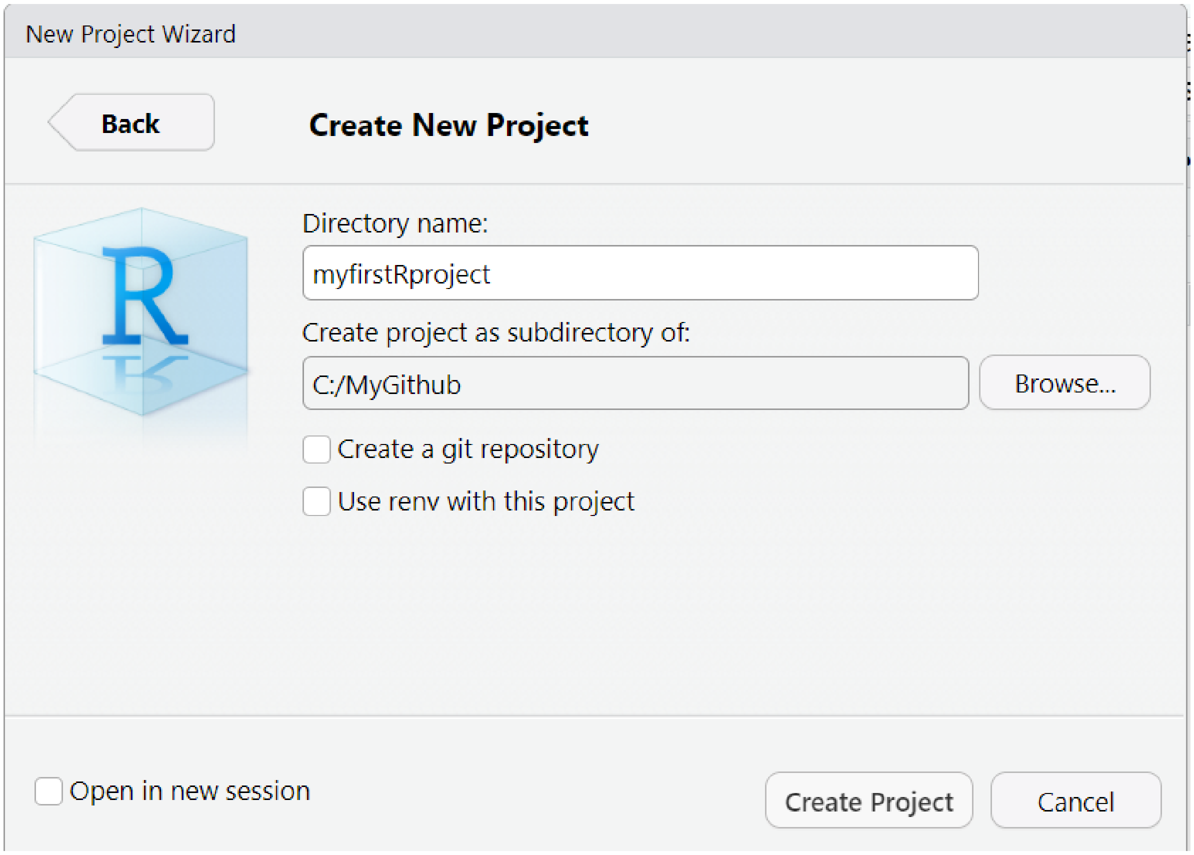
Existing Directory
Associate a project with an existing working directory

Version Control
Checkout a project from a version control repository

3. For now, let's choose a "New Directory" and then select "New Project"



4. When the next window opens, as an example, I'm creating a new folder called `myfirstRproject` for my RStudio project under the parent directory, `C:\MyGithub`.



Create New Project

Directory name:

myfirstRproject

Create project as subdirectory of:

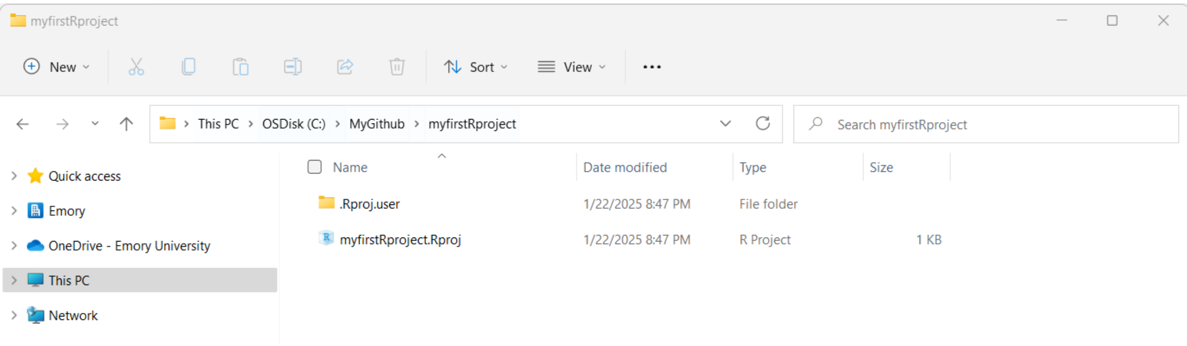
C:/MyGithub

□ Create a git repository

□ Use renv with this project

□ Open in new session

5. So, if I look back on my computer in my file manager (I'm on a computer with Windows 11 operating system) - I can now see this new folder on my computer for `C:\MyGithub\myfirstRproject`.

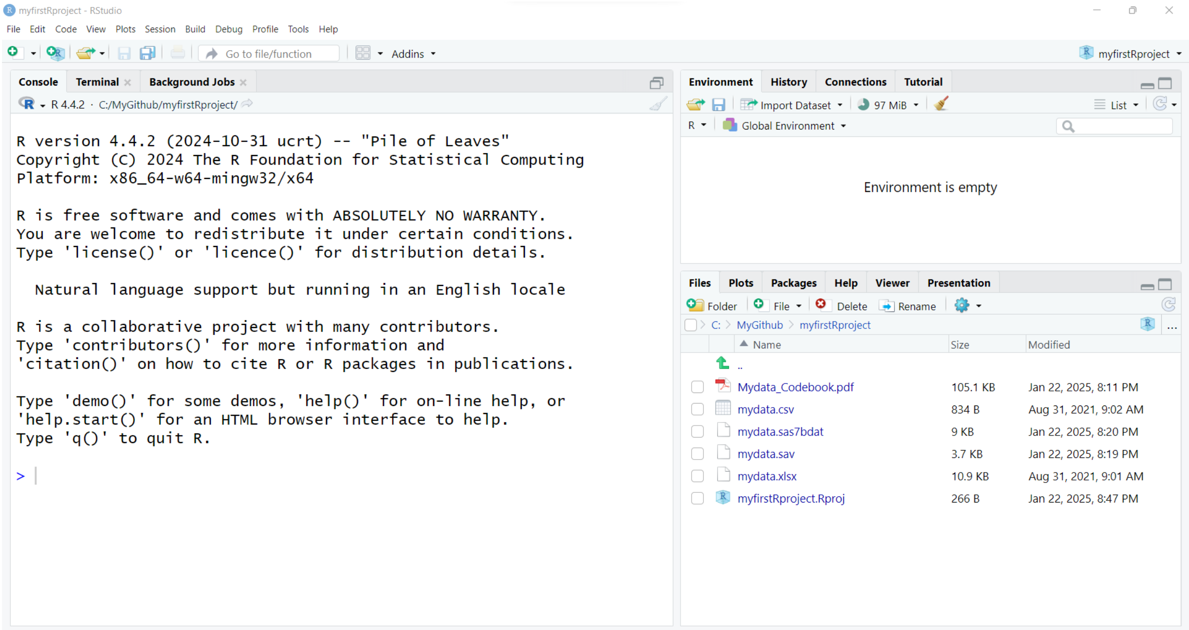


Name	Date modified	Type	Size
.Rproj.user	1/22/2025 8:47 PM	File folder	
myfirstRproject.Rproj	1/22/2025 8:47 PM	R Project	1 KB

6. Now let's put some data into this folder. Feel free to move datasets of your own into this new RStudio project directory. But here are some test datasets you can download and place into this new directory on your computer - choose at least one to try out - right click on each link and use "Save As" to save the file on your computer.

- `mydata.csv` - CSV (comma separated value) formatted data
- `mydata.xlsx` - EXCEL file
- `mydata.sav` - SPSS Dataset
- `mydata.sas7bdat` - SAS Dataset
- `Mydata_Codebook.pdf` - Codebook Details on "mydata" dataset

7. After putting these files into your new RStudio project folder, you should see something like this now in your RStudio Files Listing (bottom right window pane):



```
R version 4.4.2 (2024-10-31 ucrt) -- "Pile of Leaves"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

Environment is empty

Name	Size	Modified
..		
Mydata_Codebook.pdf	105.1 KB	Jan 22, 2025, 8:11 PM
mydata.csv	834 B	Aug 31, 2021, 9:02 AM
mydata.sas7bdat	9 KB	Jan 22, 2025, 8:20 PM
mydata.sav	3.7 KB	Jan 22, 2025, 8:19 PM
mydata.xlsx	10.9 KB	Aug 31, 2021, 9:01 AM
myfirstRproject.Rproj	266 B	Jan 22, 2025, 8:47 PM

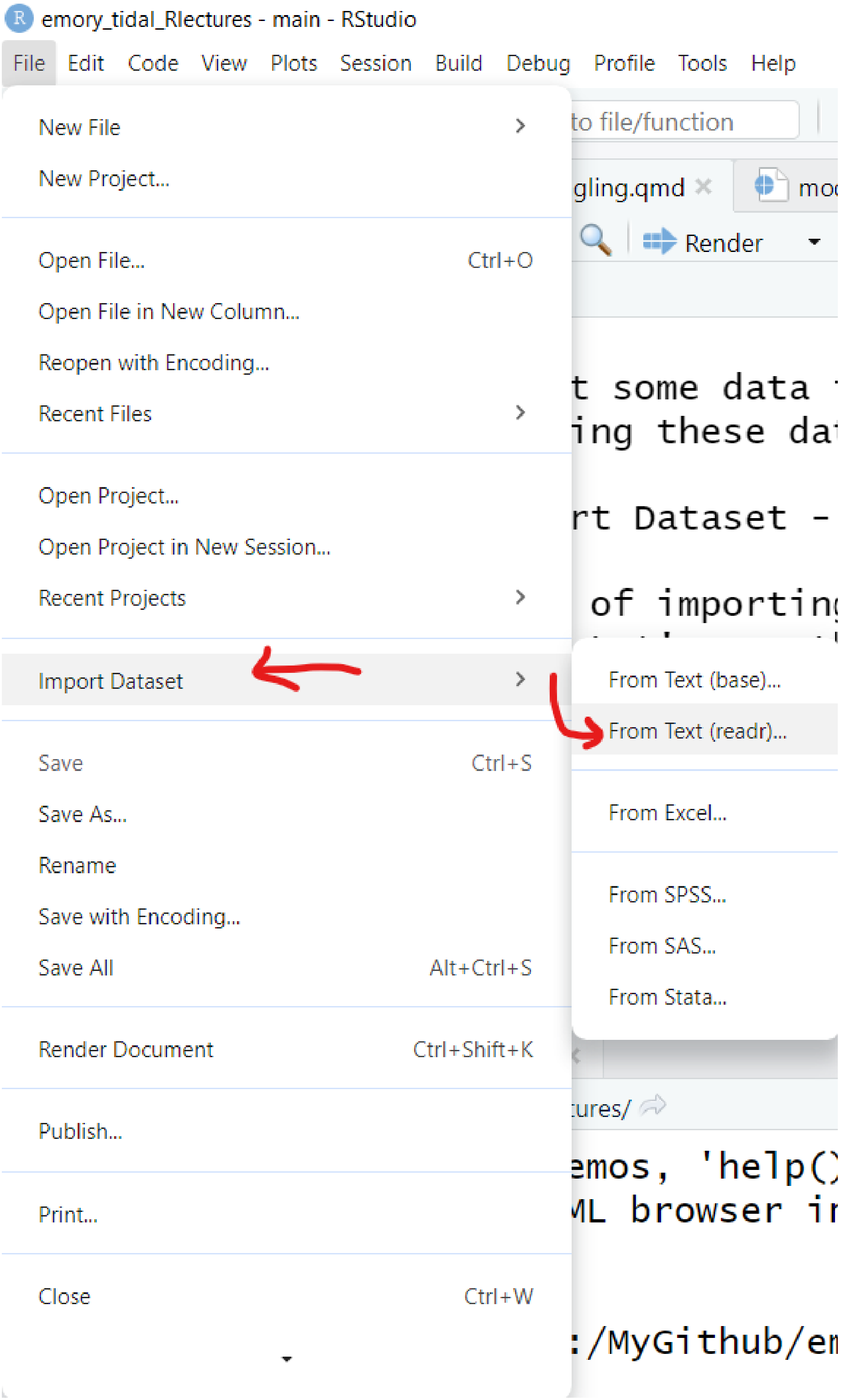
Importing Data

Now that you've got some data in your RStudio project folder, let's look at options for importing these datasets into your RStudio computing session.

Click on File/Import Dataset - and then choose the file format you want.

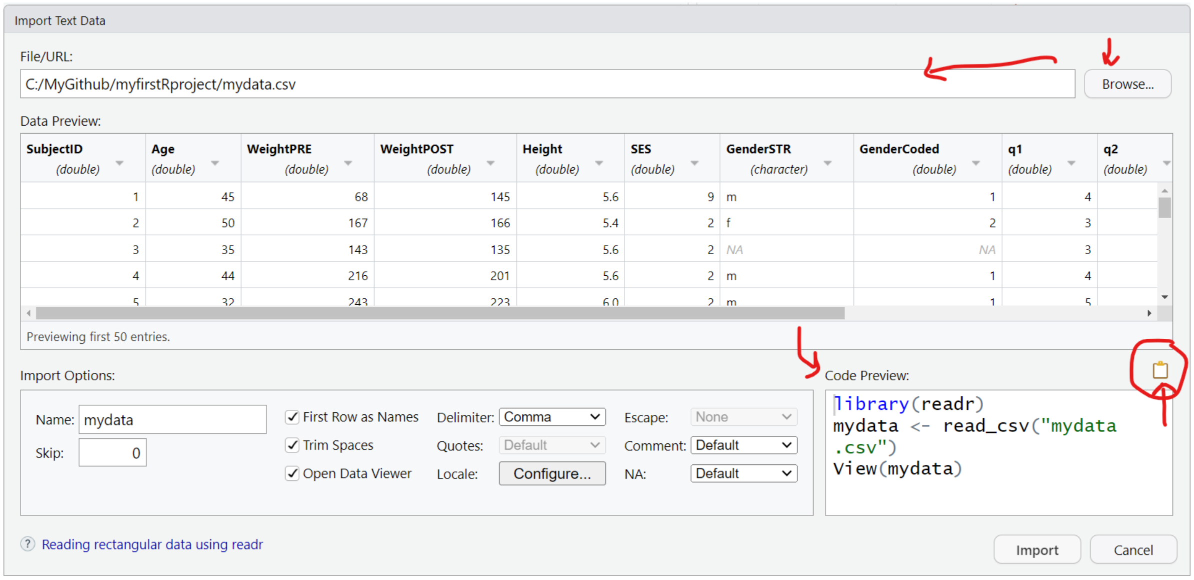
Import a CSV file

Here is an example of importing the `mydata.csv` - CSV formatted data. Let's use the `From Text (readr)` option.


emory_tidal_Rlectures - main - RStudio
File
Edit
Code
View
Plots
Session
Build
Debug
Profile
Tools
Help
New File
New Project...
Open File... Ctrl+O
Open File in New Column...
Reopen with Encoding...
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save Ctrl+S
Save As...
Rename
Save with Encoding...
Save All Alt+Ctrl+S
Render Document Ctrl+Shift+K
Publish...
Print...
Close Ctrl+W
From Text (base)...
From Text (readr)...
From Excel...
From SPSS...
From SAS...
From Stata...
to file/function
gling.qmd
mo
Render
t some data
ing these da
rt Dataset -
of importin
tures/
emos, 'help()
ML browser i
:/MyGithub/em

Once this opens, click on "Browse" and choose the mydata.csv file. Assuming all goes well, this window will read the top of the datafile and show you a quick preview to check that the import will work. And on the bottom right, the R code commands needed to import this dataset will be shown to you. You can then click on the little "clipboard" on the bottom right to copy this R code to your "clipboard", *(the R code option will be explained below)*.

OR You can also just click "Import" and the R code will be executed for you and the dataset brought into your R computing session.



File/URL: C:/MyGithub/myfirstRproject/mydata.csv

Data Preview:

SubjectID (double)	Age (double)	WeightPRE (double)	WeightPOST (double)	Height (double)	SES (double)	GenderSTR (character)	GenderCoded (double)	q1 (double)	q2 (double)
1	45	68	145	5.6	9	m	1	4	
2	50	167	166	5.4	2	f	2	3	
3	35	143	135	5.6	2	NA	NA	3	
4	44	216	201	5.6	2	m	1	4	
5	32	243	223	6.0	2	m	1	5	

Previewing first 50 entries.

Import Options:

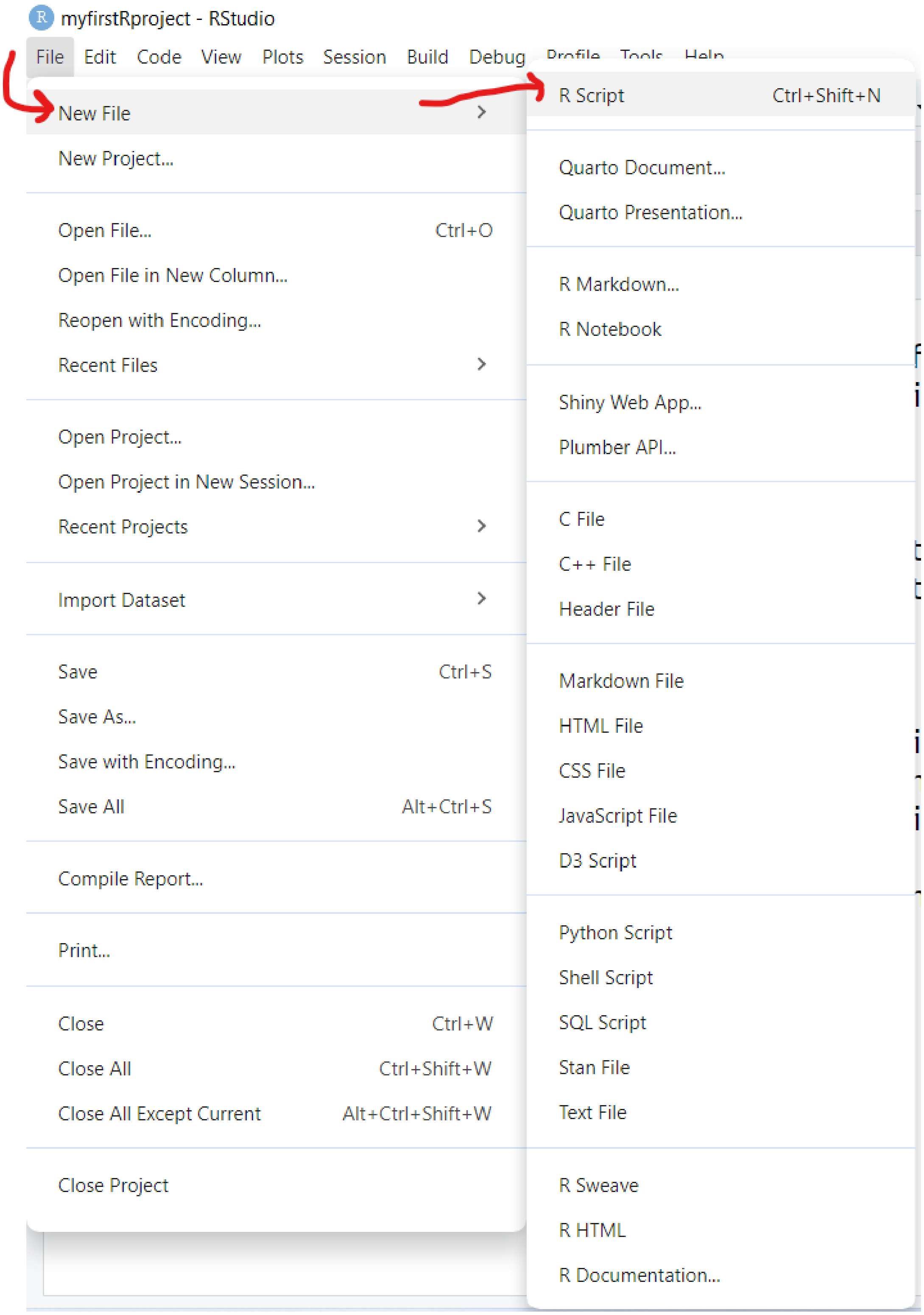
Name: mydata
Skip: 0
☑ First Row as Names
☑ Trim Spaces
☑ Open Data Viewer
Delimiter: Comma
Quotes: Default
Locale: Configure...
Escape: None
Comment: Default
NA: Default

Code Preview:

```
library(readr)
mydata <- read_csv("mydata
.csv")
View(mydata)
```

But the better way is to save the R code commands to import the data so you will be able to reproduce all steps in your data analysis workflow using code as opposed to non-reproducible point-and-click steps.

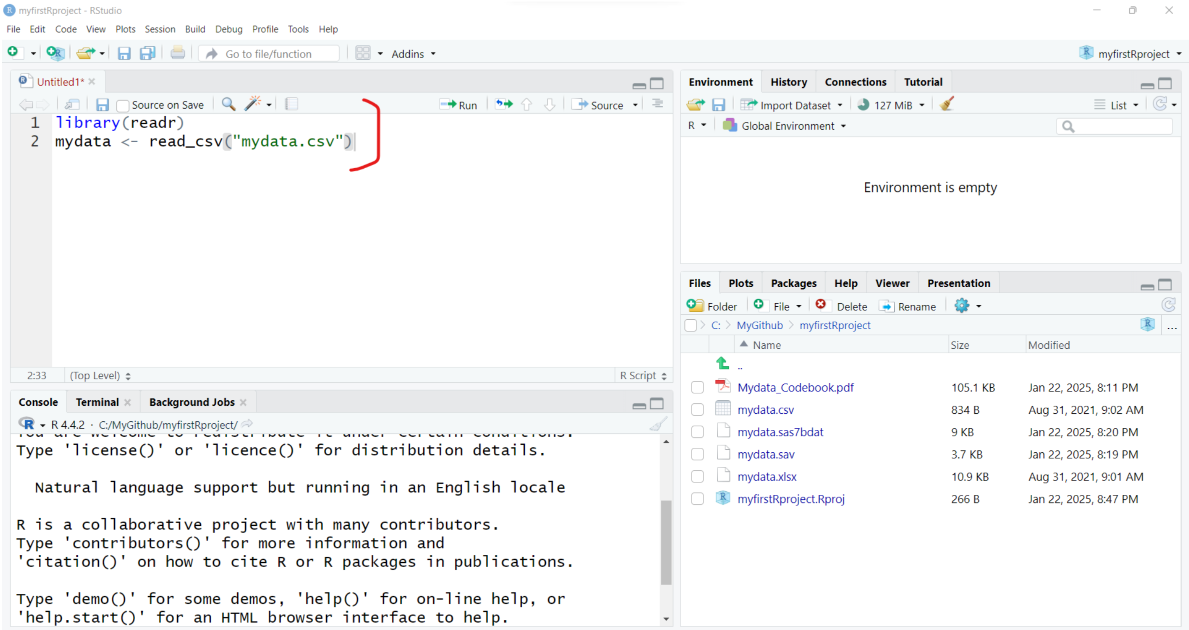
Once you copied the R code above to your clipboard, go to "File/New File/R Script" to open a script programming window:


myfirstRproject - RStudio
File
Edit
Code
View
Plots
Session
Build
Debug
Profile
Tools
Help
New File
New Project...
Open File... Ctrl+O
Open File in New Column...
Reopen with Encoding...
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save Ctrl+S
Save As...
Save with Encoding...
Save All Alt+Ctrl+S
Compile Report...
Print...
Close Ctrl+W
Close All Ctrl+Shift+W
Close All Except Current Alt+Ctrl+Shift+W
Close Project
R Script Ctrl+Shift+N
Quarto Document...
Quarto Presentation...
R Markdown...
R Notebook
Shiny Web App...
Plumber API...
C File
C++ File
Header File
Markdown File
HTML File
CSS File
JavaScript File
D3 Script
Python Script
Shell Script
SQL Script
Stan File
Text File
R Sweave
R HTML
R Documentation...

And then "paste" your R code into this window.

So as you can see importing the `mydata.csv` dataset, involves 2 steps:

1. Loading the `readr` package into your RStudio computing session, by running `library(readr)`
2. Running the `read_csv()` function from the `readr` package and then assigning `<-` this output into a new R data object called `mydata`.

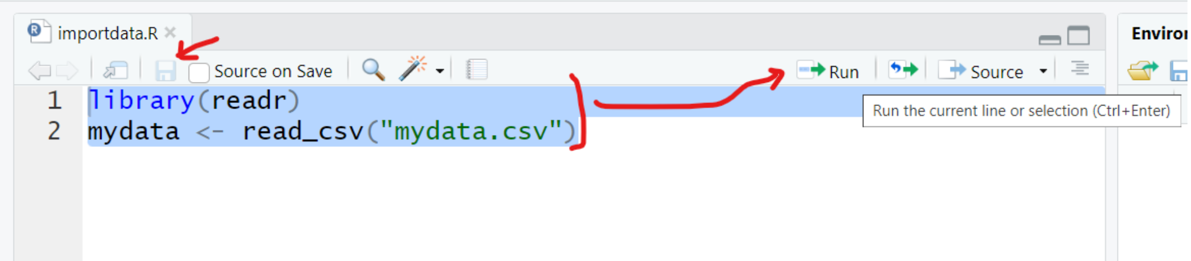

myfirstRproject - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
myfirstRproject
Untitled1*
Source on Save
Run
Source
2:33
(Top Level)
R Script
Environment History Connections Tutorial
Import Dataset
127 MiB
List
R
Global Environment
Environment is empty
Console Terminal Background Jobs
R 4.4.2 · C:/MyGithub/myfirstRproject/
Files Plots Packages Help Viewer Presentation
Folder File Delete Rename
C: > MyGithub > myfirstRproject

```r
library(readr)
mydata <- read_csv("mydata.csv")
```

```
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
```

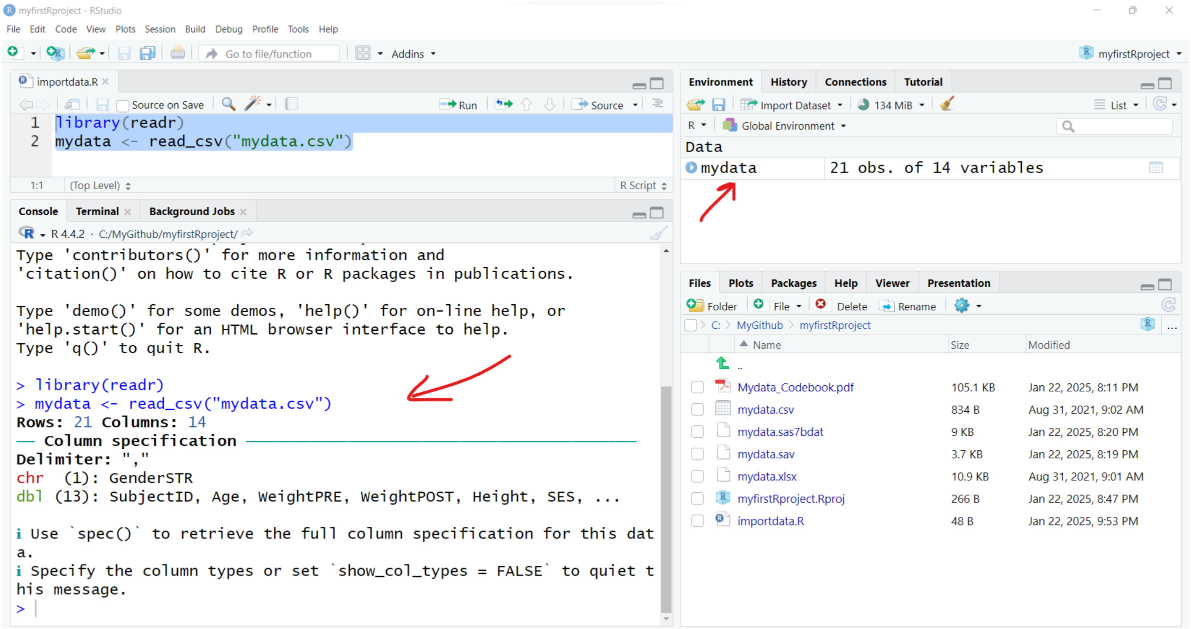
Name	Size	Modified
..		
Mydata_Codebook.pdf	105.1 KB	Jan 22, 2025, 8:11 PM
mydata.csv	834 B	Aug 31, 2021, 9:02 AM
mydata.sas7bdat	9 KB	Jan 22, 2025, 8:20 PM
mydata.sav	3.7 KB	Jan 22, 2025, 8:19 PM
mydata.xlsx	10.9 KB	Aug 31, 2021, 9:01 AM
myfirstRproject.Rproj	266 B	Jan 22, 2025, 8:47 PM

To import the dataset, select these 2 lines of code and then click "Run" to run the R code. And be sure to click "Save" to save your first R program - for example "importdata.R".



```r
library(readr)
mydata <- read_csv("mydata.csv")
```

After running these 2 lines of code, you should see something like this - the code messages in the bottom left "Console" window pane and a new R data object "mydata" in the top right "Global Environment" window pane.



```r
library(readr)
mydata <- read_csv("mydata.csv")
```

```
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(readr)
> mydata <- read_csv("mydata.csv")
Rows: 21 Columns: 14
── Column specification ──────────────────────────────────────────
Delimiter: ","
chr  (1): GenderSTR
dbl (13): SubjectID, Age, WeightPRE, WeightPOST, Height, SES, ...

ℹ Use `spec()` to retrieve the full column specification for this dat
a.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet t
his message.
> 
```

```
Data
mydata    21 obs. of 14 variables
```

Name	Size	Modified
..		
Mydata_Codebook.pdf	105.1 KB	Jan 22, 2025, 8:11 PM
mydata.csv	834 B	Aug 31, 2021, 9:02 AM
mydata.sas7bdat	9 KB	Jan 22, 2025, 8:20 PM
mydata.sav	3.7 KB	Jan 22, 2025, 8:19 PM
mydata.xlsx	10.9 KB	Aug 31, 2021, 9:01 AM
myfirstRproject.Rproj	266 B	Jan 22, 2025, 8:47 PM
importdata.R	48 B	Jan 22, 2025, 9:53 PM

Import an EXCEL file

aaaaaaaaaaa

exploring builtin datasets - see environment also - including data with packages

2. To subset data.

____

3. To select, create, and modify variables.

4. To filter the observations according to a certain condition.

____

5. To explore and summarize data.

base r - summary stats

making tables

gtsummary package

other packages - arsenal, tableone and Rmarkdown formatting of tables... get inspired

table competitions

6. To run descriptive statistics.

____

References

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Other Helpful Resources

Other Helpful Resources