You Have To Get Your Data From Somewhere

Andy Grogan-Kaylor

2020-12-06

Contents

You Have To Get Your Data From Somewhere

In learning R, as well as RMarkdown, one of the most difficult tasks seems to be understanding how to import data.

Read your data into a dataset using the *right function* for the *right format* from the *correct location*.

Note that while learning the correct syntax is *very helpful*, RStudio can generate much of this syntax for you using the **File | Import Dataset | . . .** menu.

Basic Idea

`mydata`[1] `<-` `function`[2]`("path/to/`[3]`file.extension`[4]`")`[5]

[1] How R will refer to your data. R's *nickname* for your data.

[2] The correct function for the type of data you are reading in, e.g. RData, CSV, Excel.

[3] Where is your data located? The *directory path* to your data.

[4] What is the *filename* of the file containing your data? Note that the extension often tells you *what kind* of data this is.

[5] Sometimes, especially on a Mac, it is necessary to refer to your directory with a `~`, e.g. `~/downloads/project1.RData` or `~/project1/project1.RData`.

Examples

Data Already in R Format

```
mydata <- load("/project1/mydata.RData") # load R format data
```

Data in CSV

CSV stands for *comma separated values*, and is essentially a raw text format for storing data. CSV is often an excellent format for exchanging data between programs. A few lines of *simulated data on clients* in CSV format are reproduced below.

```
"ID","age","gender","program","mental_health_T1","mental_health_T2","latitude","longitude"
2941,32,"Male","Program A",98.81,95.49,42.1108308238603,-83.6103627437424
2745,22,"Other Identity","Program B",86.28,104.84,42.0016810856589,-83.8064503632259
1320,28,"Female","Program B",89.17,107.48,42.0398163096771,-83.6793088312261
1211,20,"Male","Program D",94.15,95.71,42.2673004816002,-83.8247411126595
2293,20,"Female","Program D",85.38,105.09,42.300730845518,-83.7526918820329
```

```
library(readr) # library to read CSV

dataset <- read_csv("/project1/mydata.csv")
```

Data in Excel

```
library(readxl) # library to read Excel

dataset <- read_excel("/project1/mydata.xlsx")
```

Data in Stata

```
library(haven) # library to read other file formats

dataset <- read_stata("/project1/mydata.dta")
```

Data in SPSS

```
library(haven) # library to read other file formats

dataset <- read_sav("/project1/mydata.sav")
```

Working Directory

R uses the concept of a *working directory* to know where to find files, and where to save files.

If you do not specify a particular path to the data file that you are trying to import, R will assume that it is in your working directory.

It is often helpful to simply set your working directory to a particular location and by default, files will be accessed from, and saved to, that directory e.g.:

```
getwd() # "get", or find out, your working directory

setwd("/project1") # set your working directory
```

- Note that R uses a forward slash `/` to specify directory paths. R does not understand the use of a backward slash `\` to specify directories. R uses `~` to refer to the user's (usually your) home directory.
- If you are working in RStudio, you can use the menu option *Session | Set Working Directory | Choose Directory* to choose a particular working directory.
- If you double click on a `*.Rmd` file to start RStudio, R will assume that your working directory is the directory in which that `*.Rmd` file is located. Thus, it is often a good idea to have your data and RMarkdown document saved in the same directory.