Data Input

R Projects

R Projects are a feature of RStudio that can help you stay organized. They are pretty straightforward to set up, but are not required. You can learn more about R Projects here:

https://daseh.org/resources/R_Projects.html

Getting data into R (manual/point and click)

Data Input

- ‘Reading in’ data is the first step of any real project/analysis
- R can read almost any file format, especially via add-on packages
- We are going to focus on simple delimited files first
    - comma separated (e.g. ‘.csv’)
    - tab delimited (e.g. ‘.txt’)
    - Microsoft Excel (e.g. ‘.xlsx’)

Note: data for demonstration

- We have added functionality to load some datasets directly in the `dasehr` package

Data Input

CalEnviroScreen Dataset:

CalEnviroScreen is a project that ranks census tracts in California based on potential exposures to pollutants, adverse environmental conditions, socioeconomic factors and the prevalence of certain health conditions. Data used in the CalEnviroScreen model come from national and state sources.

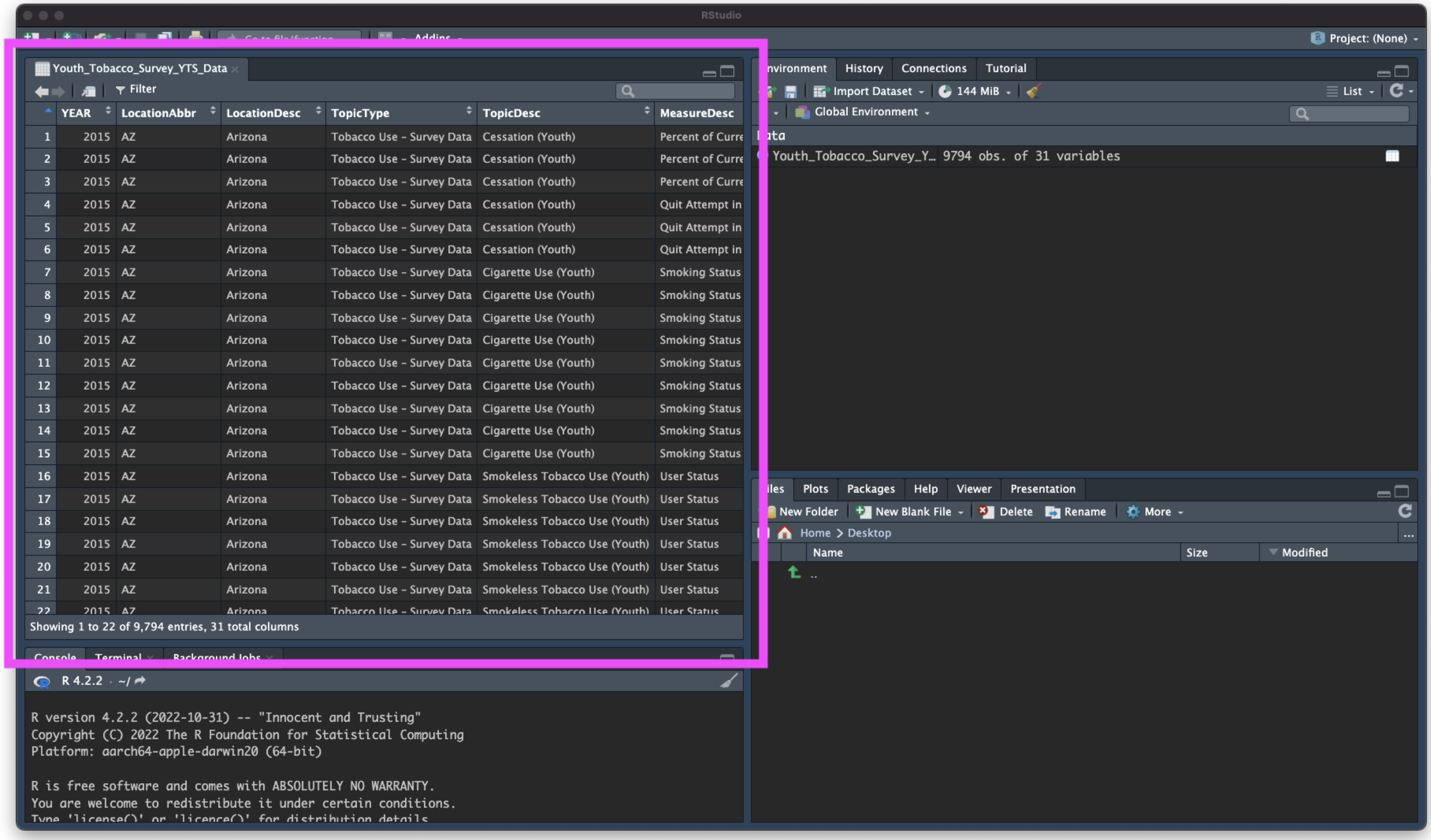
- Check out the data at: https://calenviroscreen-oehha.hub.arcgis.com/#Data

Import Dataset

- > File
- > Import Dataset
- > From Text (`readr`)
- > paste the url (https://daseh.org/data/CalEnviroScreen_data.csv)
- > click “Update” and “Import”

What Just Happened?

You see a preview of the data on the top left pane.



	YEAR	LocationAbbr	LocationDesc	TopicType	TopicDesc	MeasureDesc
1	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Percent of Curre
2	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Percent of Curre
3	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Percent of Curre
4	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Quit Attempt in
5	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Quit Attempt in
6	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Quit Attempt in
7	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
8	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
9	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
10	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
11	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
12	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
13	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
14	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
15	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
16	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
17	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
18	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
19	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
20	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
21	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
22	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status

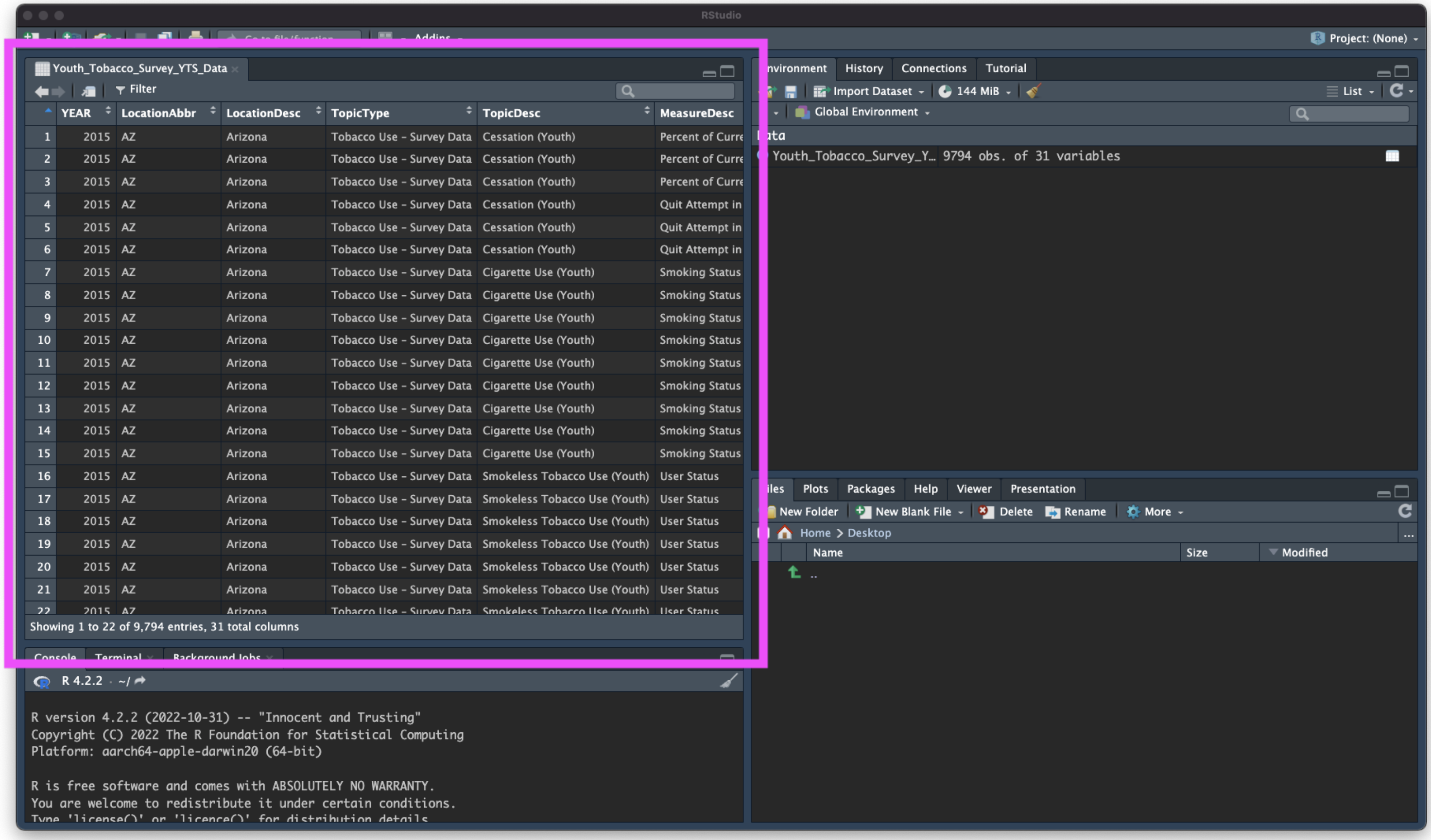
Showing 1 to 22 of 9,794 entries, 31 total columns

```
Youtth_Tobacco_Survey_Y… 9794 obs. of 31 variables
```

```
R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

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
```

What Just Happened?

You see a new object called `CalEnviroScreen_data` in your environment pane (top right). The table button opens the data for you to view.



	YEAR	LocationAbbr	LocationDesc	TopicType	TopicDesc	MeasureDesc
1	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Percent of Curre
2	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Percent of Curre
3	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Percent of Curre
4	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Quit Attempt in
5	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Quit Attempt in
6	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Quit Attempt in
7	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
8	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
9	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
10	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
11	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
12	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
13	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
14	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
15	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
16	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
17	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
18	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
19	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
20	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
21	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
22	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status

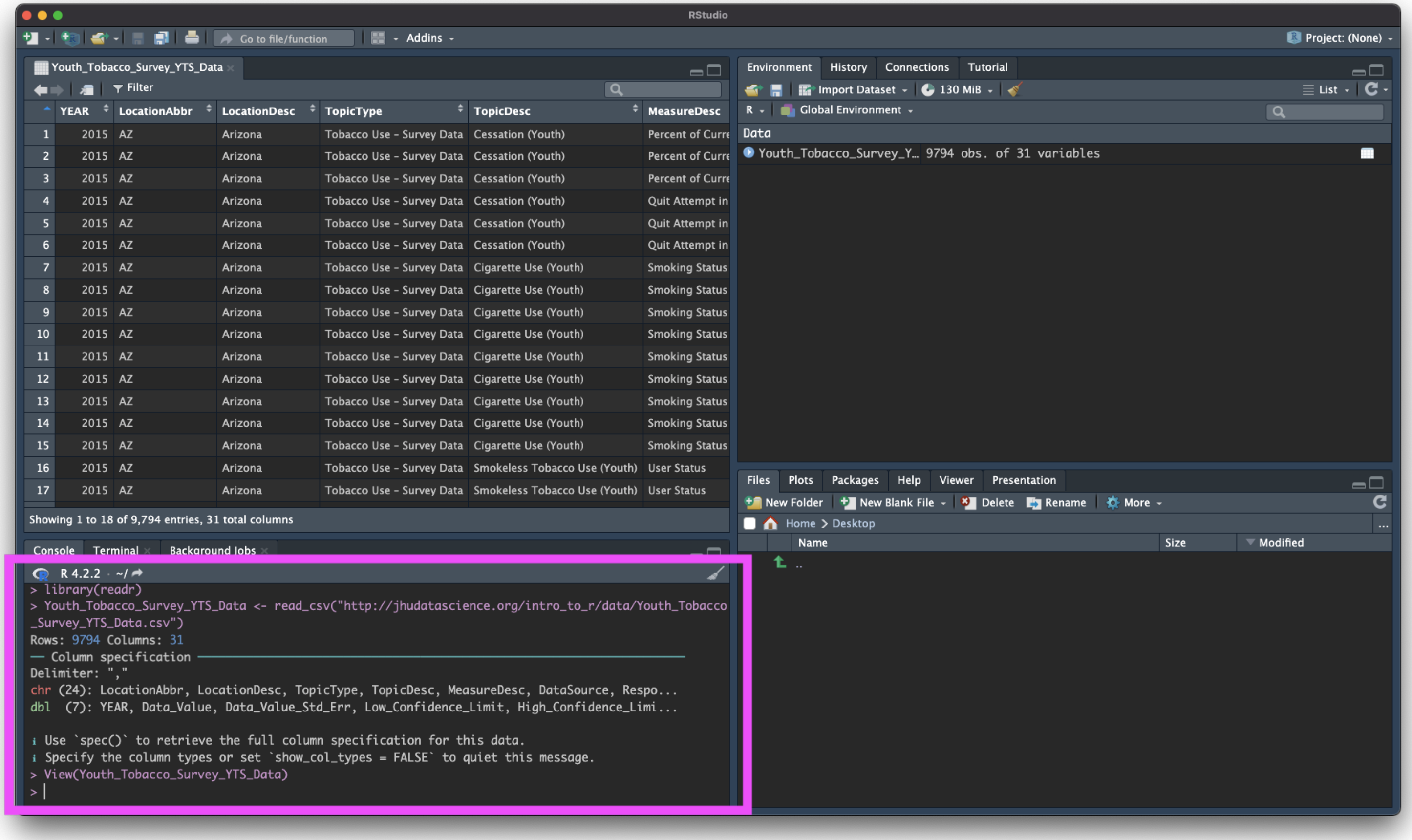
Showing 1 to 22 of 9,794 entries, 31 total columns

```
Youth_Tobacco_Survey_Y… 9794 obs. of 31 variables
```

```
R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

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.
```

What Just Happened?

R ran some code in the console (bottom left).



	YEAR	LocationAbbr	LocationDesc	TopicType	TopicDesc	MeasureDesc
1	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Percent of Curre
2	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Percent of Curre
3	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Percent of Curre
4	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Quit Attempt in
5	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Quit Attempt in
6	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Quit Attempt in
7	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
8	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
9	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
10	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
11	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
12	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
13	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
14	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
15	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking Status
16	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status
17	2015	AZ	Arizona	Tobacco Use – Survey Data	Smokeless Tobacco Use (Youth)	User Status

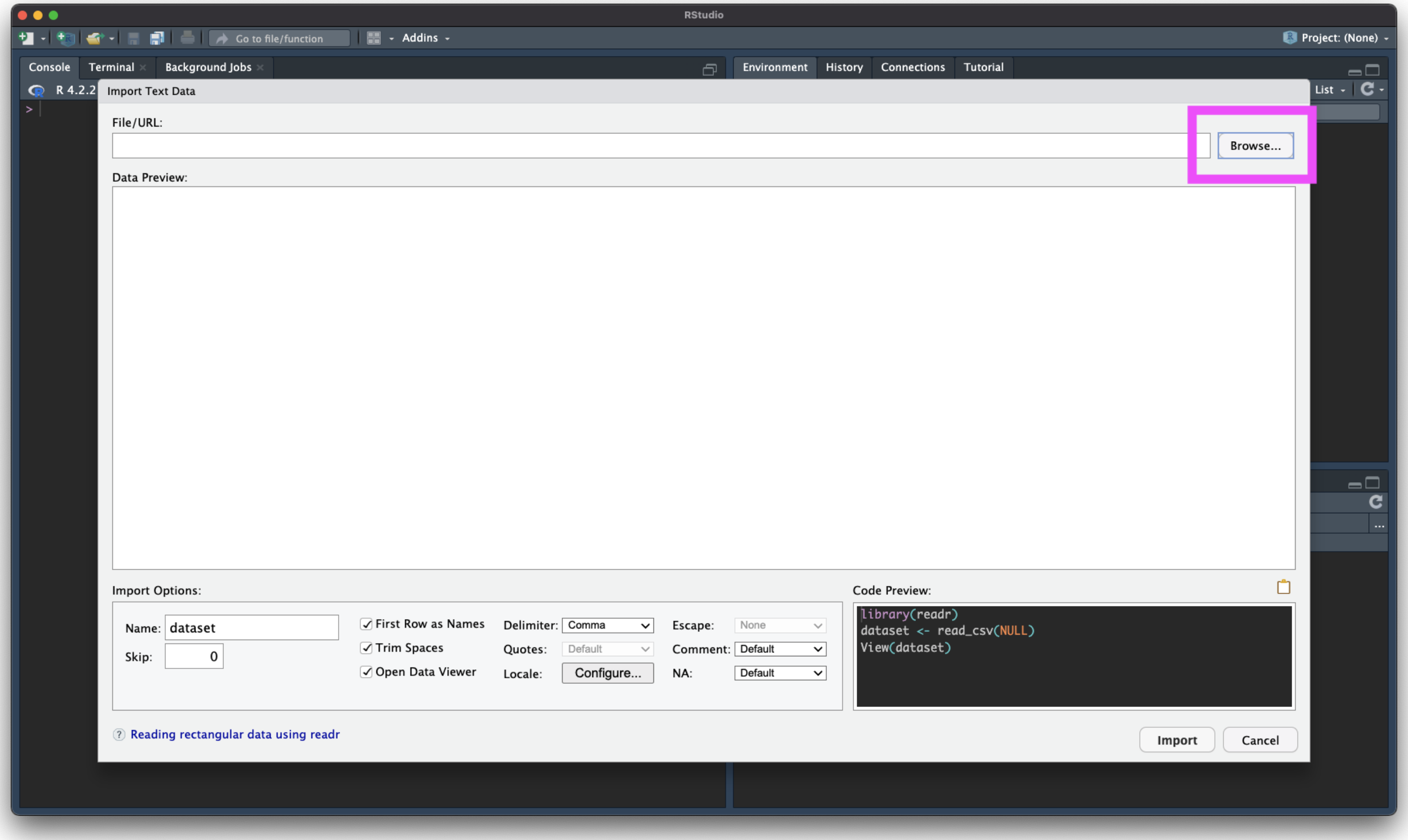
Showing 1 to 18 of 9,794 entries, 31 total columns

Data

Youth_Tobacco_Survey_Y… 9794 obs. of 31 variables

```
> library(readr)
> Youth_Tobacco_Survey_YTS_Data <- read_csv("http://jhudatascience.org/intro_to_r/data/Youth_Tobacco
_Survey_YTS_Data.csv")
Rows: 9794 Columns: 31
── Column specification ────────────────────────────────────────────────────
Delimiter: ","
chr (24): LocationAbbr, LocationDesc, TopicType, TopicDesc, MeasureDesc, DataSource, Respo...
dbl  (7): YEAR, Data_Value, Data_Value_Std_Err, Low_Confidence_Limit, High_Confidence_Limi...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
> View(Youth_Tobacco_Survey_YTS_Data)
> |
```

Browsing for Data on Your Machine



File/URL:

Name: dataset

Skip: 0

☑ First Row as Names

☑ Trim Spaces

☑ Open Data Viewer

Delimiter: Comma

Quotes: Default

Locale: Configure...

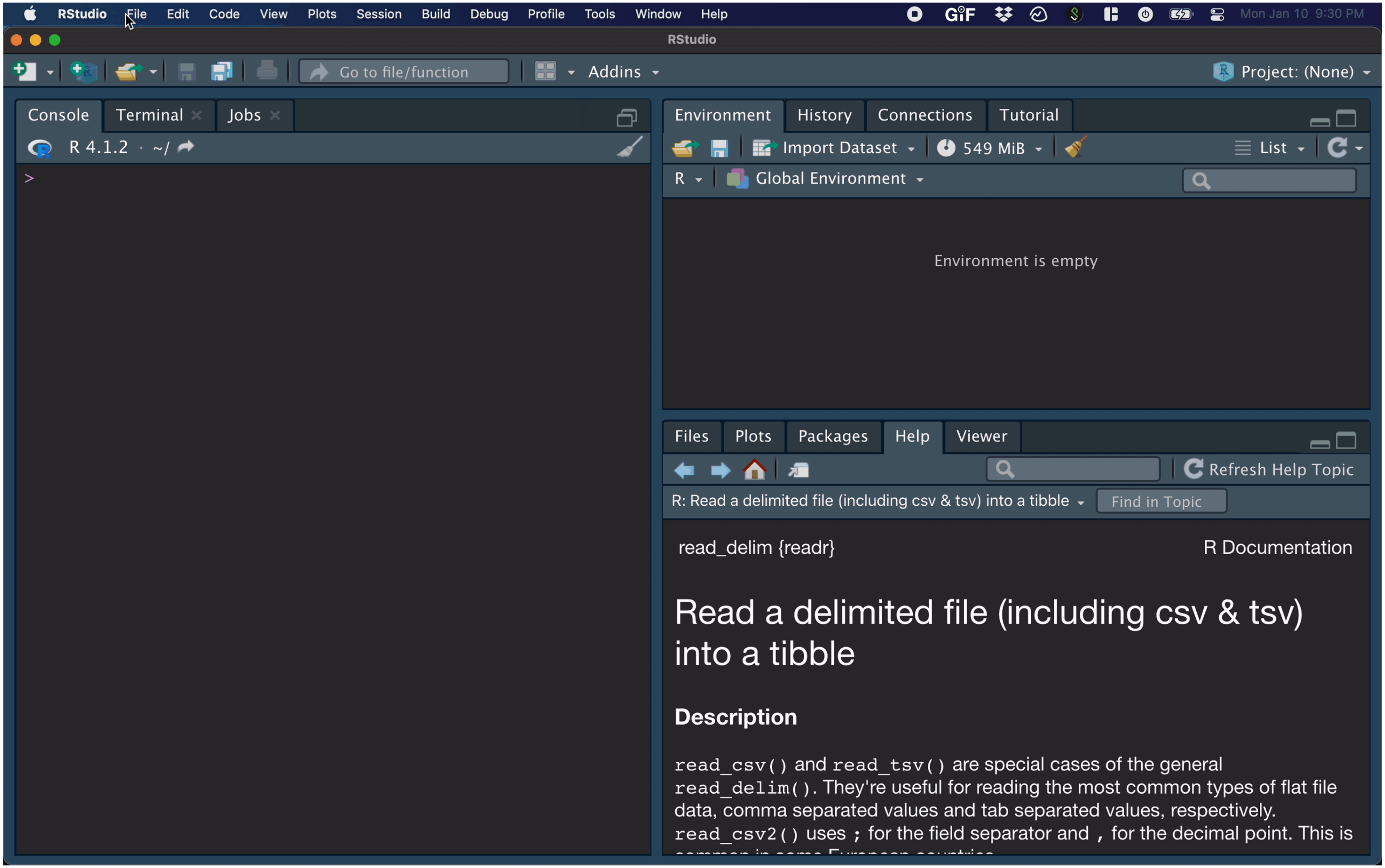
Escape: None

Comment: Default

NA: Default

```
library(readr)
dataset <- read_csv(NULL)
View(dataset)
```

Import Dataset



read_delim {readr} R Documentation

Read a delimited file (including csv & tsv) into a tibble

Description

`read_csv()` and `read_tsv()` are special cases of the general `read_delim()`. They're useful for reading the most common types of flat file data, comma separated values and tab separated values, respectively. `read_csv2()` uses `;` for the field separator and `,` for the decimal point. This is common in some European countries

Manual Import: Pros and Cons

Pros: easy!!

Cons: obscures some of what’s happening, others will have difficulty running your code

Getting data into R (directly)

Data Input: Read in Directly

```r
# load library `readr` that contains function `read_csv`
library(readr)
dat <- read_csv(
  file = "https://daseh.org/data/CalEnviroScreen_data.csv"
)

# `head` displays first few rows of a data frame. `tail()` works the same way.
head(dat, n = 5)
```

```
# A tibble: 5 × 68
   ...1 CensusTract CaliforniaCounty   ZIP Longitude Latitude ApproxLocation
  <dbl>       <dbl> <chr>            <dbl>     <dbl>    <dbl> <chr>
1     1  6001400100 Alameda          94704     -122.     37.9 Oakland
2     2  6001400200 Alameda          94618     -122.     37.8 Oakland
3     3  6001400300 Alameda          94618     -122.     37.8 Oakland
4     4  6001400400 Alameda          94609     -122.     37.8 Oakland
5     5  6001400500 Alameda          94609     -122.     37.8 Oakland
# ℹ 61 more variables: CES4.0Score <dbl>, CES4.0Percentile <dbl>,
#   CES4.0PercRange <chr>, Ozone <dbl>, OzonePctl <dbl>, PM2.5 <dbl>,
#   PM2.5.Pctl <dbl>, DieselPM <dbl>, DieselPMPctl <dbl>, DrinkingWater <dbl>,
#   DrinkingWaterPctl <dbl>, Lead <dbl>, LeadPctl <dbl>, Pesticides <dbl>,
#   PesticidesPctl <dbl>, ToxRelease <dbl>, ToxReleasePctl <dbl>,
#   Traffic <dbl>, TrafficPctl <dbl>, CleanupSites <dbl>,
#   CleanupSitesPctl <dbl>, GroundwaterThreats <dbl>, …
```

Data Input: Declaring Arguments

```
dat <- read_csv(
  file = "https://daseh.org/data/CalEnviroScreen_data.csv"
)
# EQUIVALENT TO
dat <- read_csv(
  "https://daseh.org/data/CalEnviroScreen_data.csv"
)
```

Data Input: Read in Directly

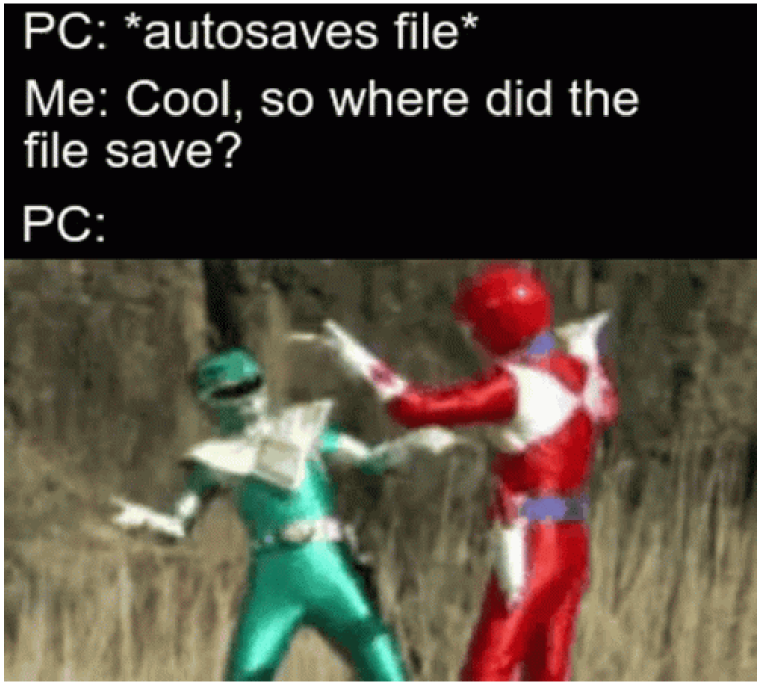
`read_csv()` needs an argument `file =`.

- `file` is the path to your file, **in quotation marks**
- can be path to a file on a website (URL)
- can be **path** in your local computer – absolute file path or relative file path

```
# Examples

dat <- read_csv(file = "www.someurl.com/table1.csv")

dat <- read_csv(file = "/Users/avahoffman/Downloads/CalEnviroScreen_data.csv")

dat <- read_csv(file = "CalEnviroScreen_data.csv")
```

Data Input: File paths

What is a file path ????

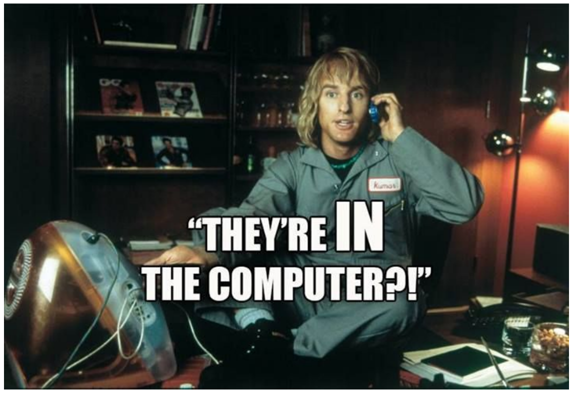


The working directory

When we work in R, we automatically have a **working directory**.

Working directory is a folder (directory) that RStudio assumes “you are working in”.

It’s where R looks for files.



Getting the working directory

Run the `getwd()` function to determine your working directory.

```
# Get the working directory
getwd()
```

Relative path

Let’s say my data is in a folder called “data” in my working directory.

`data/my_data.csv` would be the **relative path**. It’s relative to the working directory.

The whole address, for example `/Users/avahoffman/Downloads/data/my_data.csv` is the **absolute path**.

Setting the working directory

You can set the working directory manually with the `setwd()` function:

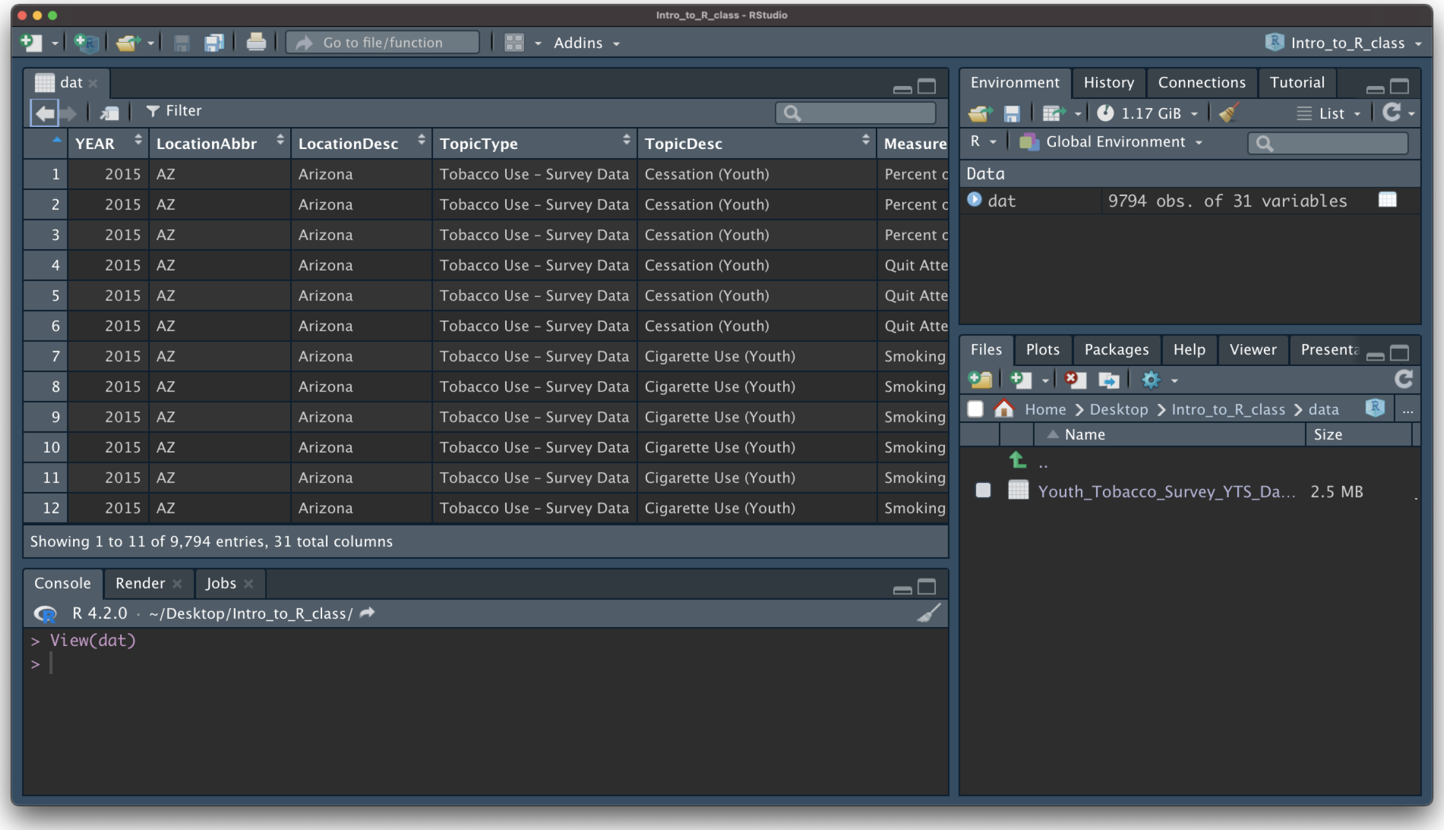
```
# set the working directory
setwd("/Users/avahoffman/Desktop")
```

Now what? Checking data & Other formats

Data Input: Checking the data

- the `View()` function shows your data in a new tab, in spreadsheet format
- be careful if your data is big!

```
View(dat)
```



	YEAR	LocationAbbr	LocationDesc	TopicType	TopicDesc	Measure
1	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Percent c
2	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Percent c
3	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Percent c
4	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Quit Atte
5	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Quit Atte
6	2015	AZ	Arizona	Tobacco Use – Survey Data	Cessation (Youth)	Quit Atte
7	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking
8	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking
9	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking
10	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking
11	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking
12	2015	AZ	Arizona	Tobacco Use – Survey Data	Cigarette Use (Youth)	Smoking

Showing 1 to 11 of 9,794 entries, 31 total columns

```
> View(dat)
> 
```

Data

dat 9794 obs. of 31 variables

Name	Size
..	
Youth_Tobacco_Survey_YTS_Da...	2.5 MB

Data Input: Other delimiters with `read_delim()`

`read_csv()` is a special case of `read_delim()` – a general function to read a delimited file into a data frame

`read_delim()` needs path to your file and **file’s delimiter**, will return a tibble

- `file` is the path to your file, in quotes
- `delim` is what separates the fields within a record

```
## Examples
dat <- read_delim(file = "www.someurl.com/table1.tsv", delim = "\t")

dat <- read_delim(file = "data.txt", delim = "|")
```

Data Input: Excel files

- You **cannot** read in an excel file from a URL.
- Need to load the `readxl` package with `library()`.
- The argument is `path` (not `file`).

```
library(readxl)
```

```
read_excel(path = "nitrate.xlsx")
```

Data input: other file types

- `haven` package has functions to read SAS, SPSS, Stata formats
- There are also resources for REDCap : `REDCapR`

WARNING! `read.csv` is * base R *

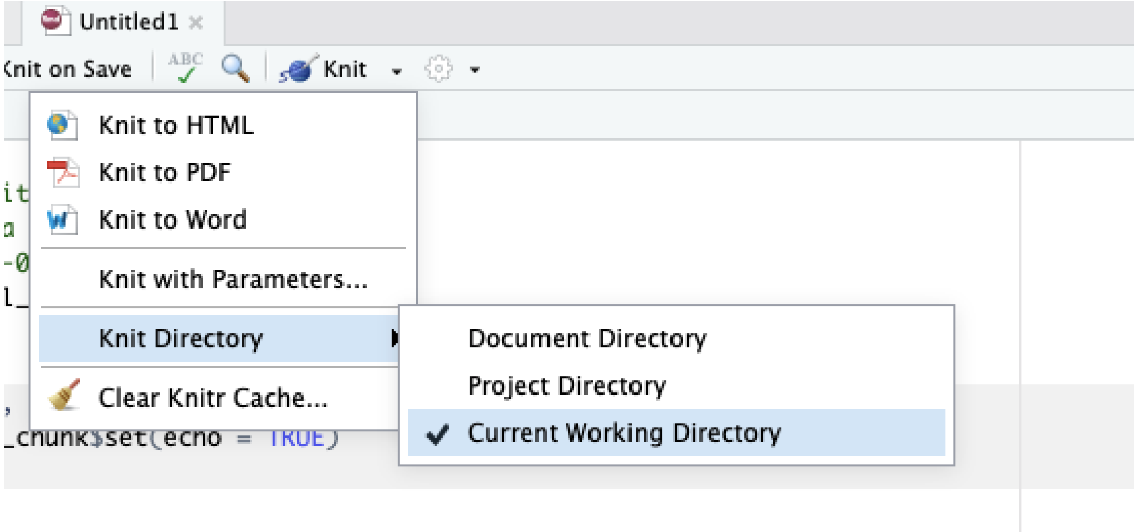
There are also data importing functions provided in base R (rather than the `readr` package), like `read.delim()` and `read.csv()`.

These functions have slightly different syntax for reading in data (e.g. `header` argument).

However, while many online resources use the base R tools, the latest version of RStudio switched to use these new `readr` data import tools, so we will use them in the class for slides. They are also up to two times faster for reading in large datasets, and have a progress bar which is nice.

TROUBLESHOOTING: Setting the working directory

If you are trying to knit your work, it might help to set the knit directory to the “Current Working Directory”:



Other Useful Functions

- The `str()` function can tell you about data/objects.
- We will also discuss the `glimpse()` function later, which does something very similar.
- `head()` shows first few rows
- `tail()` shows the last few rows

Summary

R Projects can make it easier to find files. Check out this resource.

Importing data manually:

- File > Import Dataset > From Text (`readr`)
- Paste the url
- Click “Update” and “Import”
- Review the process: `https://youtu.be/LEkNfJgpunQ`

Importing data programmatically:

- `read_csv()` function from `readr` package
- Use `getwd()` to check your working directory, where R looks for your data files

Summary - Part 2

Look at your data!

- Check the environment for a data object
- `View()` gives you a preview of the data in a new tab

Other file types

- `readr` package: `read_delim()` for general delimited files
- `readxl` package: `read_excel()` for Excel files

Don’t forget to use `<-` to assign your data to an object!

Lab

- Class Website
- Data Input Lab

Image by Gerd Altmann from Pixabay