Intro to R

Data Input

Day 1 Review

- the RStudio Editor (top) is for static code like scripts or R Markdown documents
- The console is for testing code (bottom) - best to save your code though!
- R code goes within what is called a chunk (the gray box with a green play button)
- **Objects** (like nouns) are data or variables.

Day 1 Review

- R functions as a calculator
- Use `<-` to save (assign) values to objects
- **Functions** (like verbs) perform specific tasks in R and are found within packages
- Use `c()` to **combine** vectors
- `length()`, `class()`, and `str()` tell you information about an object
- Install packages with `install.packages()`
- Load packages with `library()`
- Get help with `?` or help pane

Day 1 Review

- Make sure we have installed and loaded the `tidyverse` package!

Outline

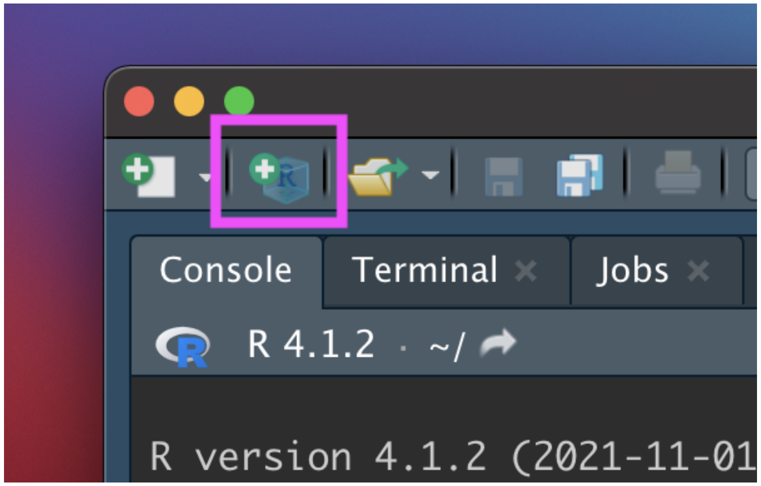
- Part 0: A little bit of set up!
- Part 1: reading in manually (point and click) (.csv)
- Part 2: checking data & multiple file formats (.xlsx)

Part 0: Setup - R Project

New R Project

Let’s make an R Project so we can stay organized in the next steps.

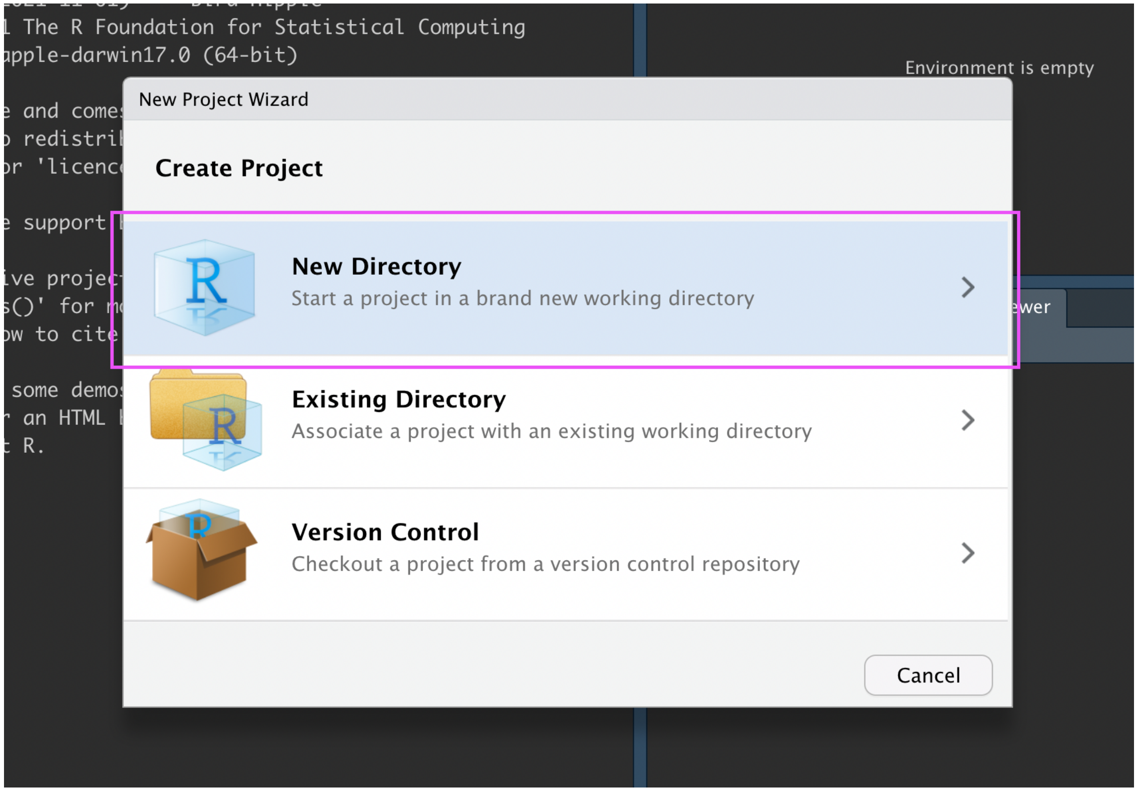
Click the new R Project button at the top left of RStudio:



```
R version 4.1.2 (2021-11-01
```

New R Project

In the New Project Wizard, click “New Directory”:



```
l The R Foundation for Statistical Computing
apple-darwin17.0 (64-bit)

e and comes
o redistrib
or 'licence

e support

ive projec
s()' for m
ow to cite

 some demos
r an HTML b
t R.
```

Environment is empty

New Project Wizard

Create Project

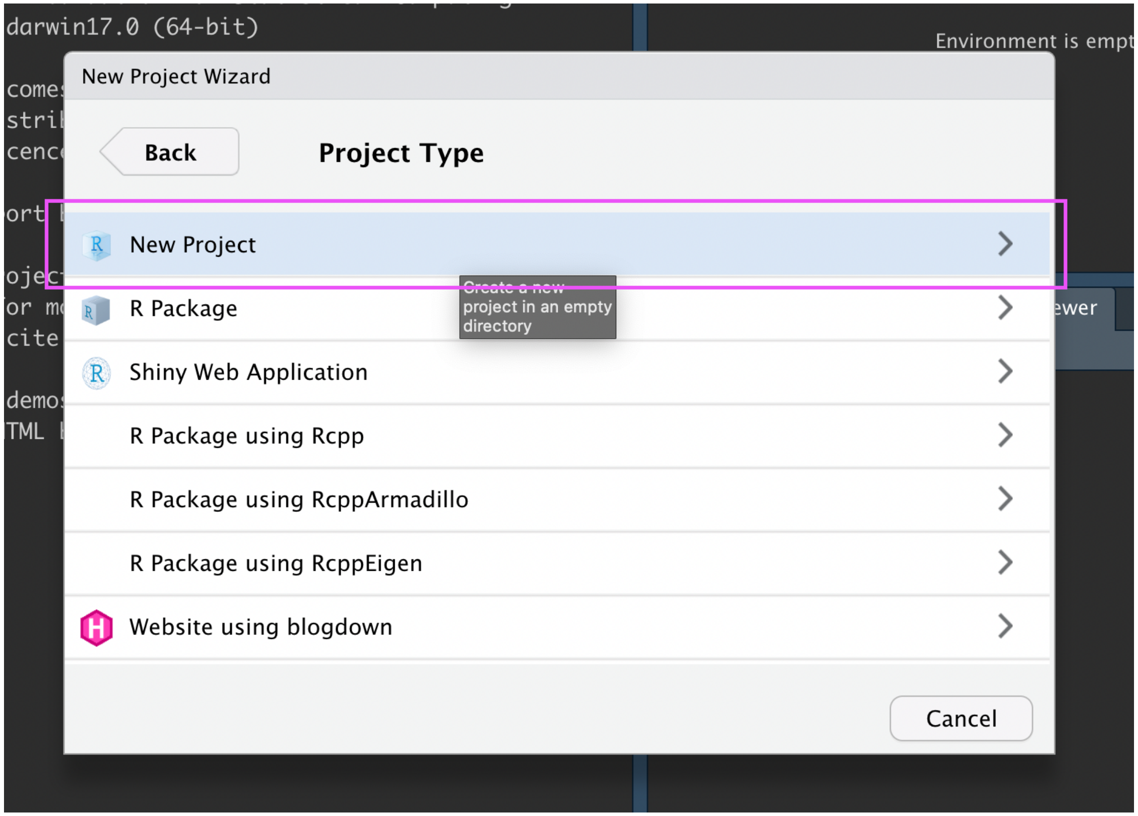
New Directory
Start a project in a brand new working directory

Existing Directory
Associate a project with an existing working directory

Version Control
Checkout a project from a version control repository

New R Project

Click “New Project”:

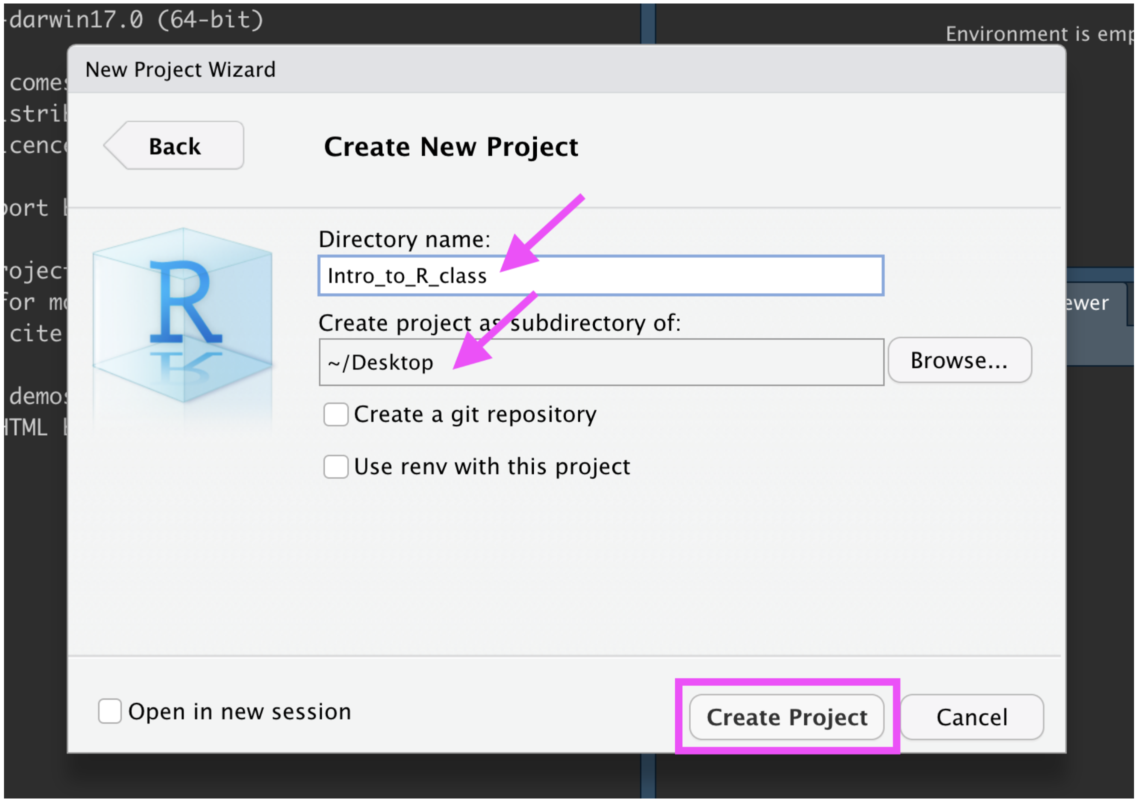


Create a new project in an empty directory

New R Project

Type in a name for your new folder.

Store it somewhere easy to find, such as your Desktop:



Directory name: Intro_to_R_class

Create project as subdirectory of: ~/Desktop

□ Create a git repository

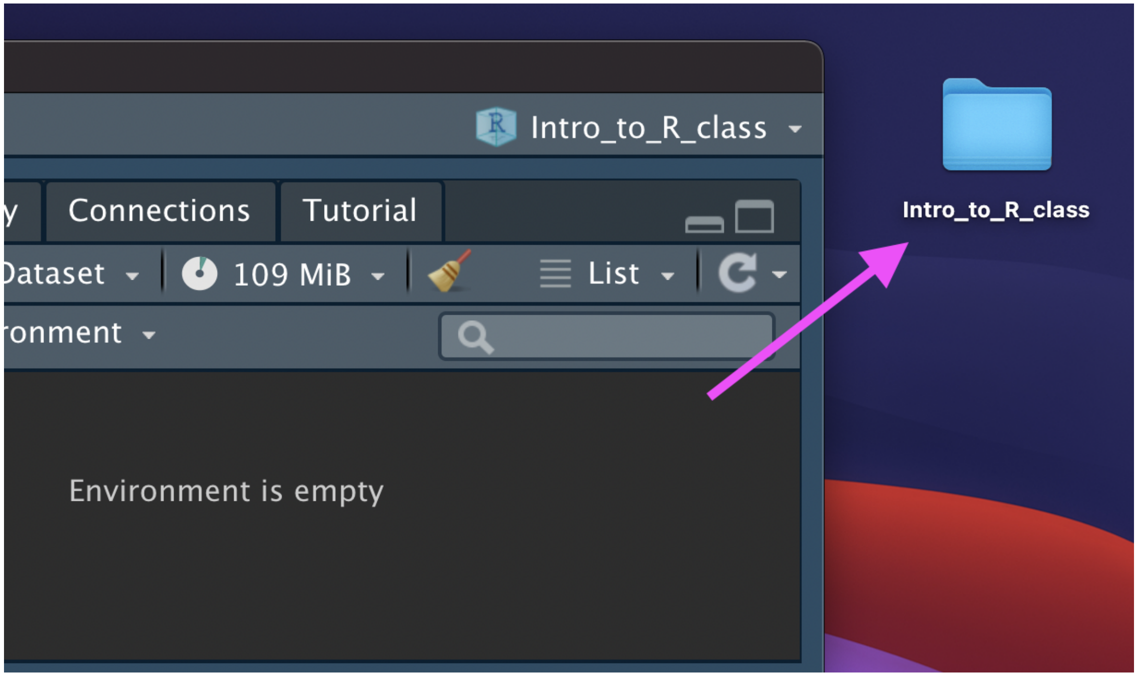
□ Use renv with this project

□ Open in new session

New R Project

You now have a new R Project folder on your Desktop!

Make sure you add any scripts or data files to this folder as we go through today's lesson. This will make sure R is able to "find" your files.



New R Project

Pause here! Confirm the Project is set up.

Part 1: Getting data into R (manual/point and click, .csv)

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’)

Data Input

Youth Tobacco Survey (YTS) dataset:

“The YTS was developed to provide states with comprehensive data on both middle school and high school students regarding tobacco use, exposure to environmental tobacco smoke, smoking cessation, school curriculum, minors’ ability to purchase or otherwise obtain tobacco products, knowledge and attitudes about tobacco, and familiarity with pro-tobacco and anti-tobacco media messages.”

- Check out the data at: https://catalog.data.gov/dataset/youth-tobacco-survey-yts-data

Data Input: Dataset Location

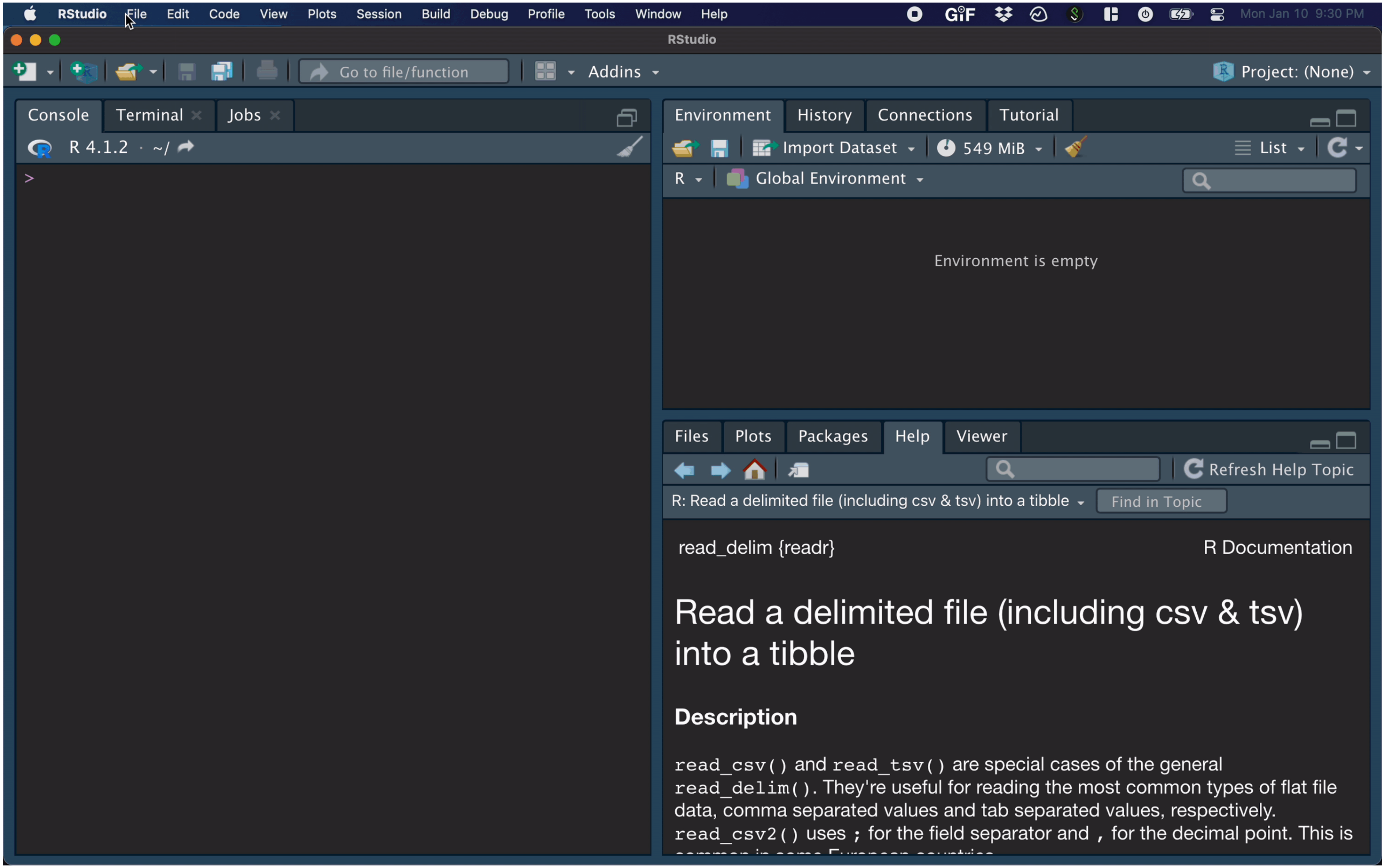
Dataset is located at https://hutchdatascience.org/SeattleStatSummer_R/data/Youth_Tobacco_Survey_YTS_D

- Download data by clicking the above link
    - Safari - if a file loads in your browser, choose File –> Save As, select, Format “Page Source” and save

Import Dataset

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

Import Dataset



read_delim {readr}   R Documentation

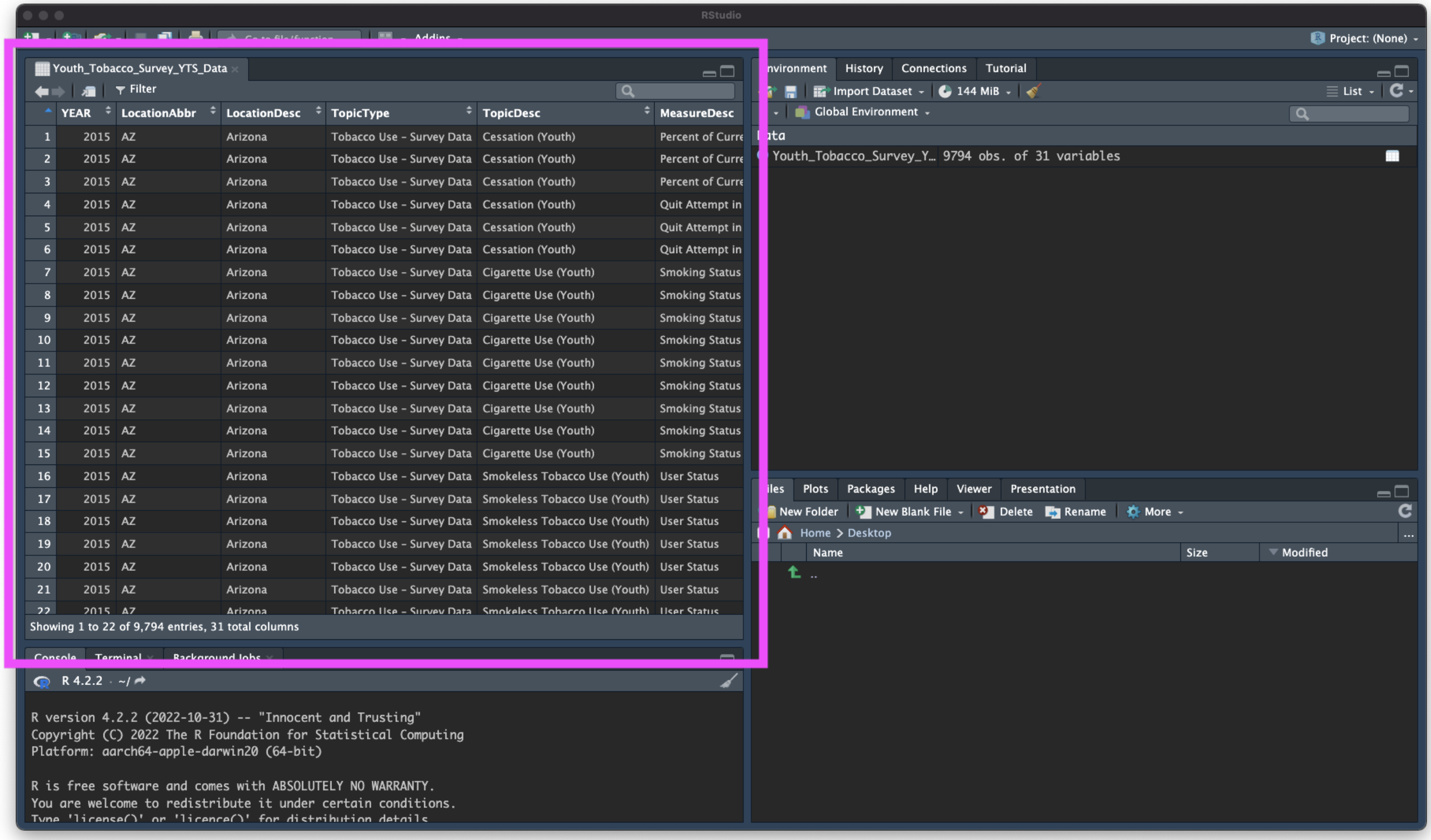
Read a delimited file (including csv & tsv) into a tibble

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

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

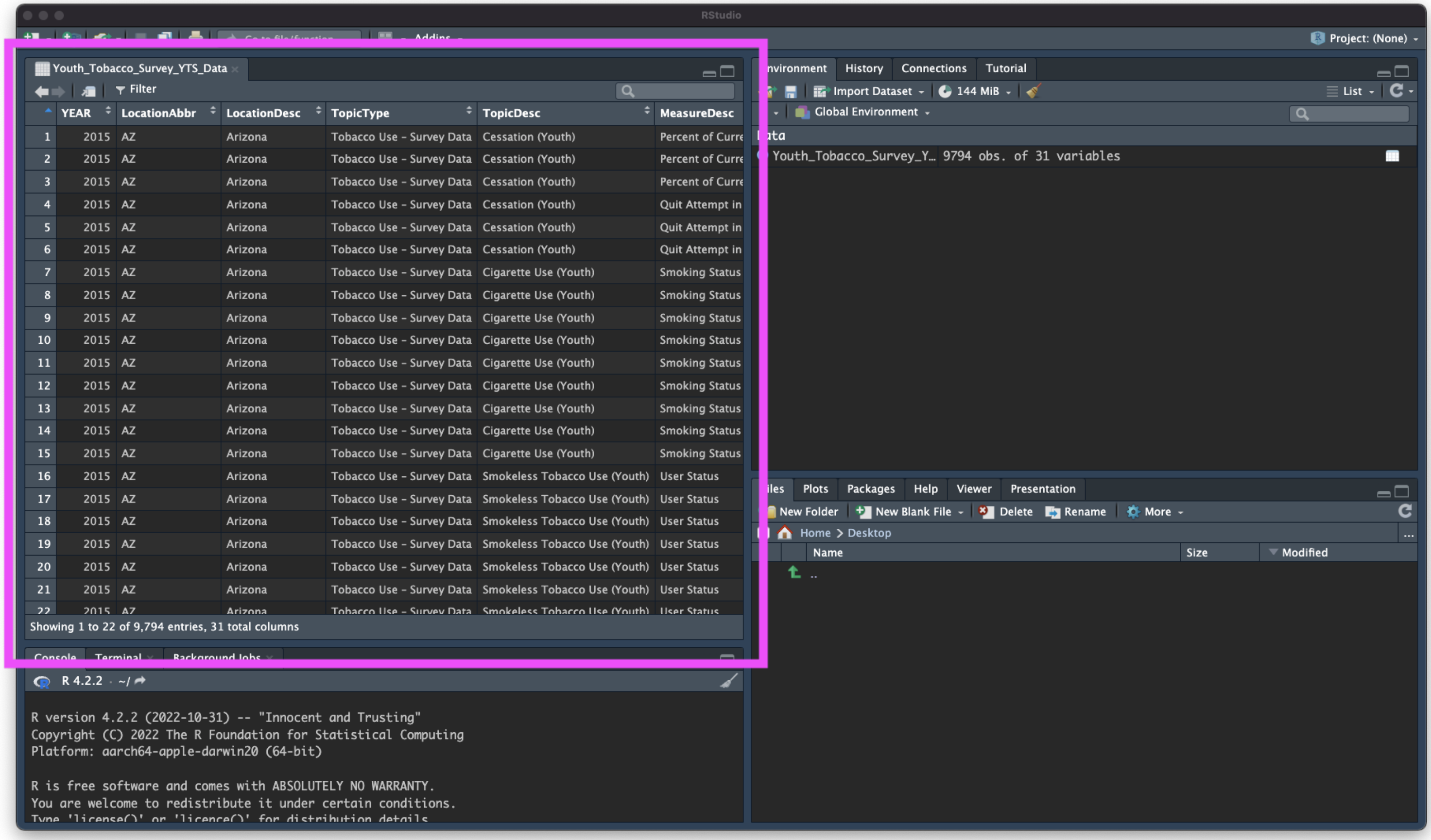
Youth_Tobacco_Survey_Y… 9794 obs. of 31 variables

```
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?

You see a new object called `Youth_Tobacco_Survey_YTS_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

Data

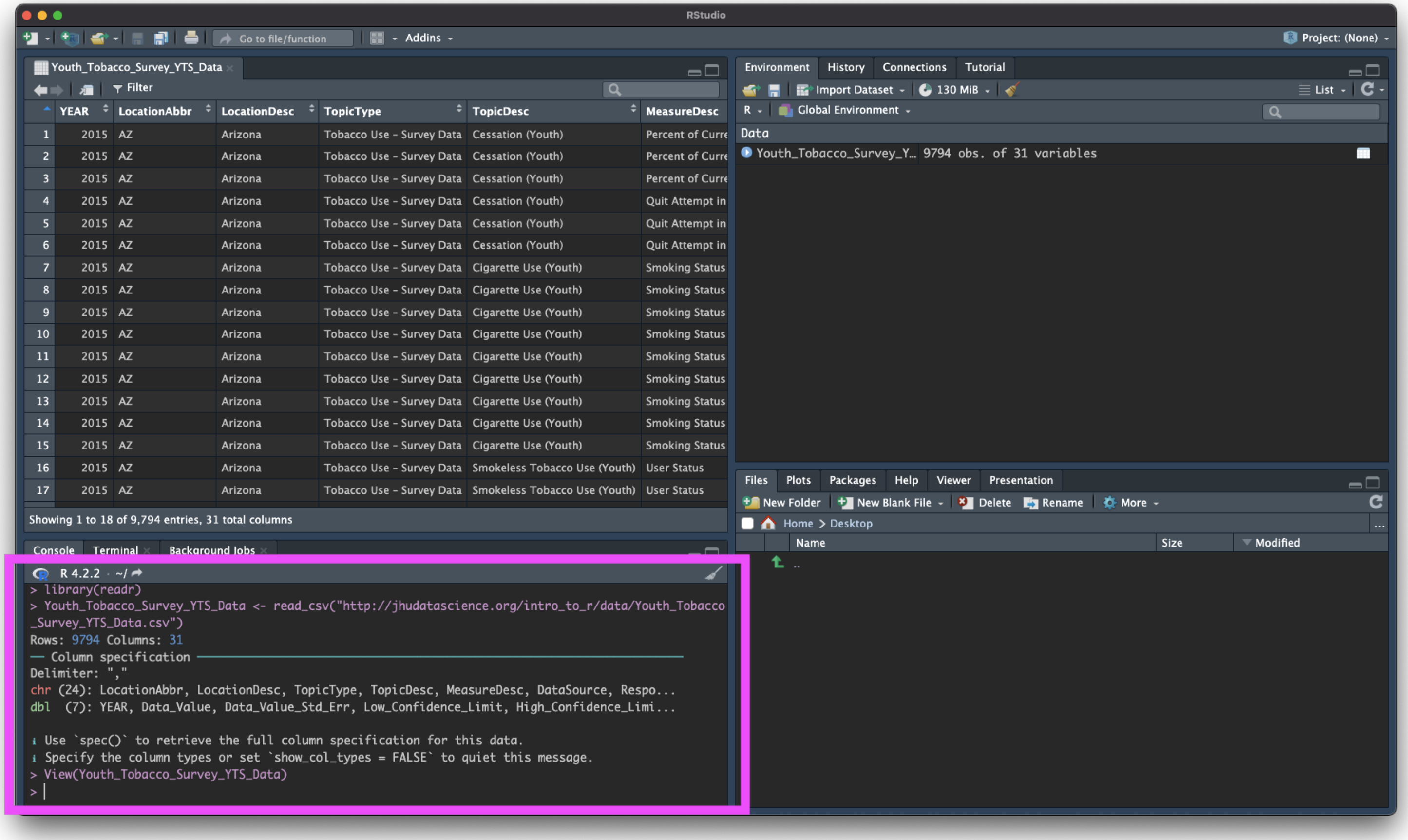
Youth_Tobacco_Survey_Y… 9794 obs. of 31 variables

```
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

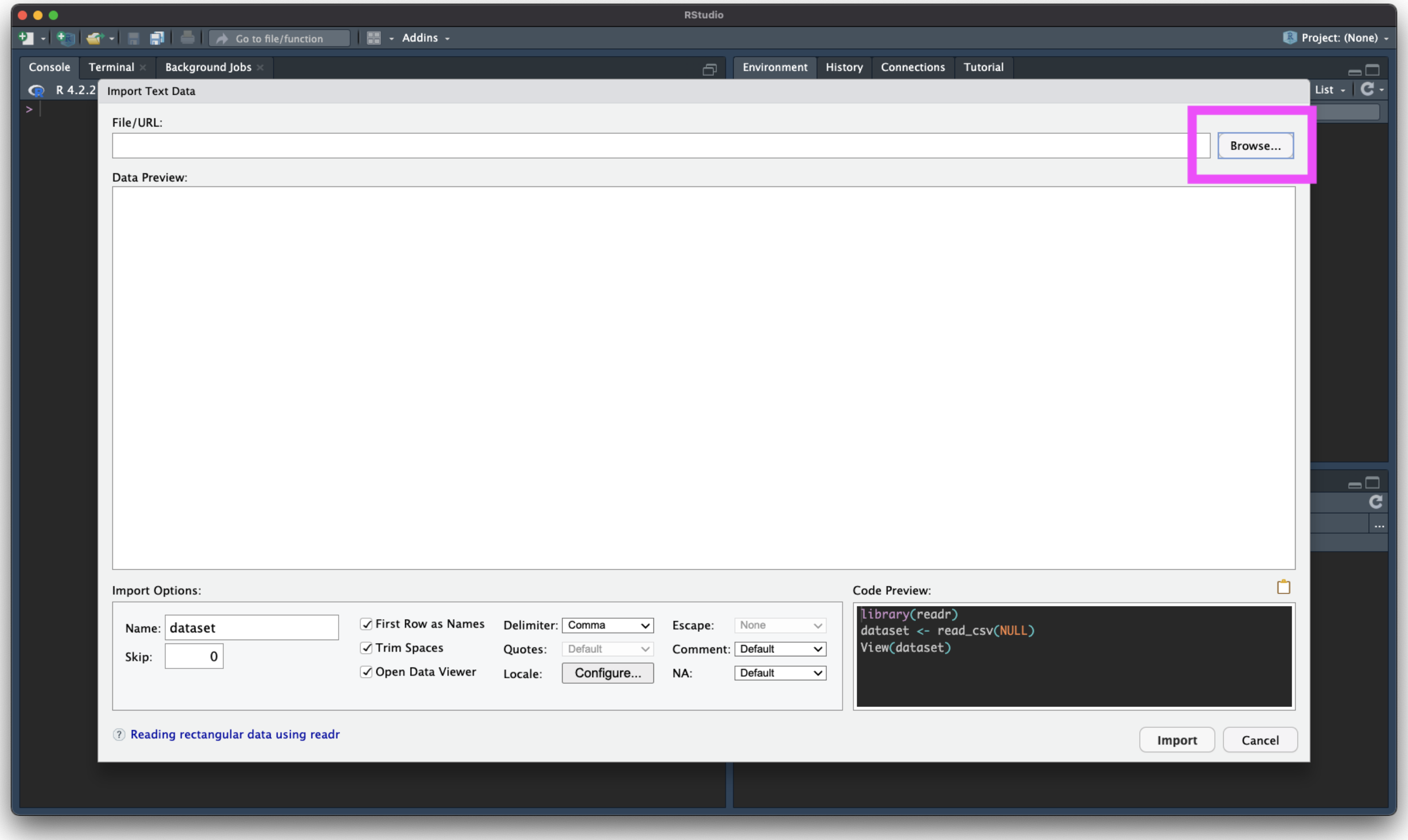
Data

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



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)
```

Summary

Review the process: `https://youtu.be/LEkNfJgpunQ`

- > File
- > Import Dataset
- > From Text (`readr`)
- > paste the url (https://hutchdatascience.org/SeattleStatSummer_R/data/Youth_Tobacco_Survey_YT
- > click "Update" and "Import"

Let’s practice!

Importing “states” data

- Try downloading the dataset located here: https://hutchdatascience.org/SeattleStatSummer_R/data/states.csv
- Use the File > Import Dataset > from Text (`readr`)
- Browse for the downloaded file on your machine
- Inspect the code that was run. Copy this code into your R Markdown document for later!

Looking at the code

```
library(readr)
states <- read_csv("~/Downloads/states.csv")
View(states)
```

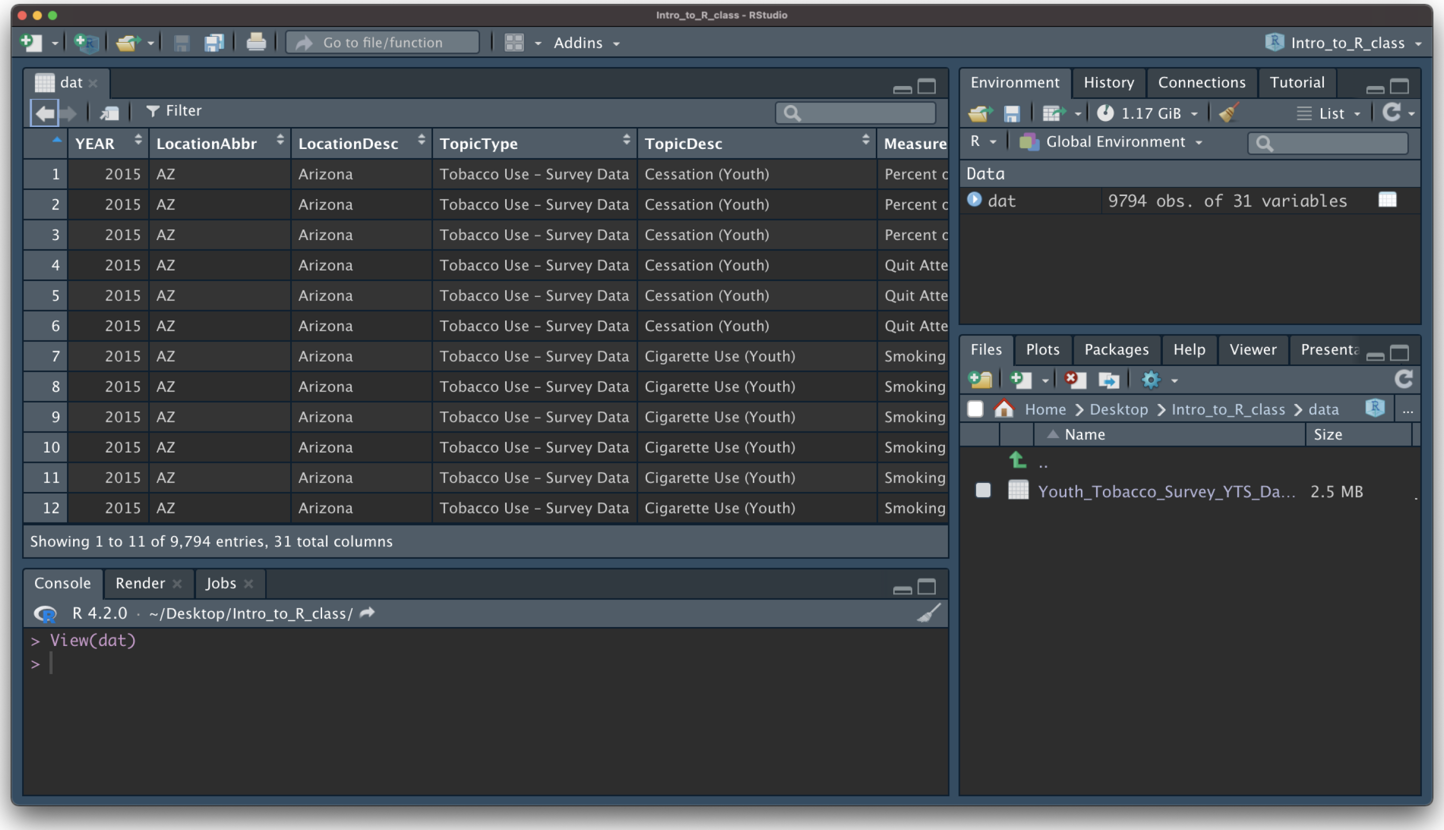
Notice that the part in quotes is a location on my computer. If I move the file, I won’t be able to use the same code again. Better to move that file to my project folder (instead of my downloads for example) for longer-term storage.

Part 2: 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(states)
```



	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

dat 9794 obs. of 31 variables

Youth_Tobacco_Survey_YTS_Da... 2.5 MB

```
R 4.2.0 · ~/Desktop/Intro_to_R_class/
> View(dat)
>
```

Data Input: Checking the data

The `str()` function can tell you about data/objects(different variables and their classes - more on this later).

```
str(states)
```

```
spec_tbl_df [52 × 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ entity                   : chr [1:52] "Alabama" "Alaska" "Arizona" "Arkansas" ...
 $ state_abb                : chr [1:52] "AL" "AK" "AZ" "AR" ...
 $ state_area_sq_miles      : num [1:52] 51609 589757 113909 53104 158693 ...
 $ state_division           : chr [1:52] "East South Central" "Pacific" "Mountain" "West South
 $ state_region             : chr [1:52] "South" "West" "West" "South" ...
 $ population               : num [1:52] 4903185 731545 7278717 3017804 39512223 ...
 $ births_in_2021           : num [1:52] 58054 9367 77916 35965 420608 ...
 $ fertility_rate_per_1000  : num [1:52] 59.5 64.9 55.5 61.7 52.8 52.5 52.1 56.5 54.9 55.9 ...
 $ cesarean_percent         : num [1:52] 35.1 24.2 28.7 34.3 30.8 27.3 35.4 31.9 35.8 35.1 ...
 $ life_expect              : num [1:52] 73.2 76.6 76.3 73.8 79 78.3 78.4 76.7 77.5 75.6 ...
 $ cancer_rate_per_100000   : num [1:52] 160 156 135 168 132 ...
 $ cancer_mortality         : num [1:52] 10429 1093 12813 6516 59503 ...
 $ Administered_Dose1_Pop_Pct: num [1:52] 64.8 72.8 77.1 69.6 84.3 83.3 95 87.7 82.1 68.1 ...
 $ Series_Complete_Pop_Pct  : num [1:52] 53 64.9 65.8 56.7 74.4 73.2 82.8 72.9 69.2 57.1 ...
 - attr(*, "spec")=
  .. cols(
  ..   entity = col_character(),
  ..   state_abb = col_character(),
  ..   state_area_sq_miles = col_double(),
  ..   state_division = col_character(),
  ..   state_region = col_character(),
  ..   population = col_double(),
  ..   births_in_2021 = col_double(),
       fertility_rate_per_1000 = col_double()
```

Data Input: Excel files

- Getting data from Excel is a bit more complicated. You have to download the file, either through R or manually.
- R does not know how to read excel files by default. We will use a package called `readxl` to do that.

Let's bring the following into R:

https://hutchdatascience.org/SeattleStatSummer_R/data/asthma.xlsx

- > File
- > Import Dataset
- > From Excel …
- > paste the url (https://hutchdatascience.org/SeattleStatSummer_R/data/asthma.xlsx)
- > click "Update" and "Import"

Looking at the code:

```
library(readxl)
url <- "https://hutchdatascience.org/SeattleStatSummer_R/data/asthma.xlsx"
destfile <- "asthma.xlsx"
curl::curl_download(url, destfile)
asthma <- read_excel(destfile)
View(asthma)
```

Let’s practice!

Importing tuberculosis data

- Try downloading the dataset located here: https://hutchdatascience.org/SeattleStatSummer_R/data/tb_incidence.xlsx
- Use the File > Import Dataset > from Excel
- Browse for the downloaded file on your machine
- Inspect the code that was run. Copy this code into your R Markdown document for later!

Looking at the code

```r
library(readxl)
tb_incidence <- read_excel("tb_incidence.xlsx")
View(tb_incidence)
```

Modifying the code

You can name the dataset whatever you want, it’s an object in your Environment now.

```
library(readxl)
my_data <- read_excel("tb_incidence.xlsx")
View(my_data)
```

Summary

- > File
- > Import Dataset
- > From Text (`readr`) **OR** From Excel
- > paste the url or Browse for the file
- > click “Update” and “Import”
- > save the code for later!

Workshop Website