Introduction

Meet the toolkit

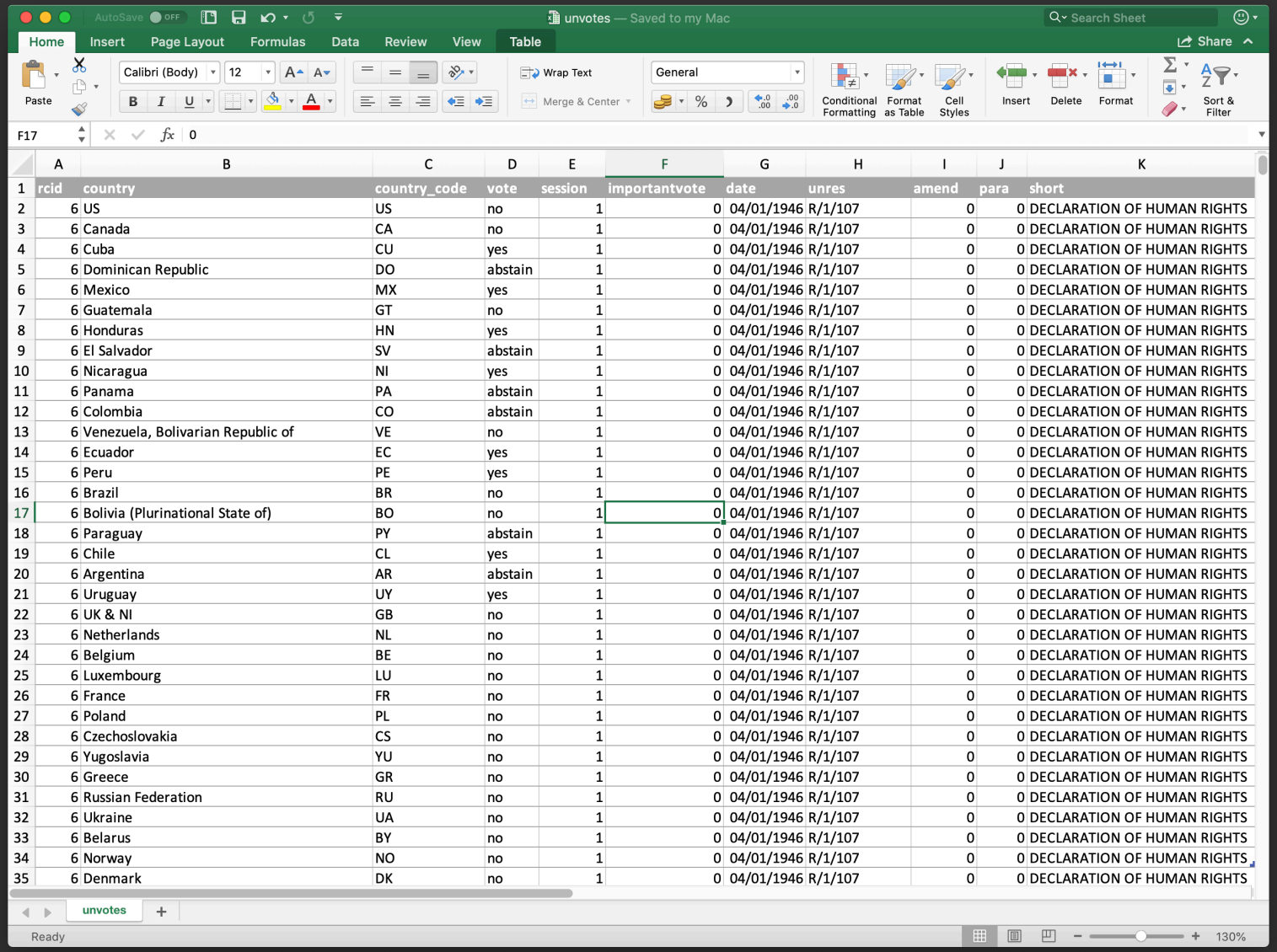
Termeh Shafie

These slides are adapted from Data Science in a Box
datasciencebox.org

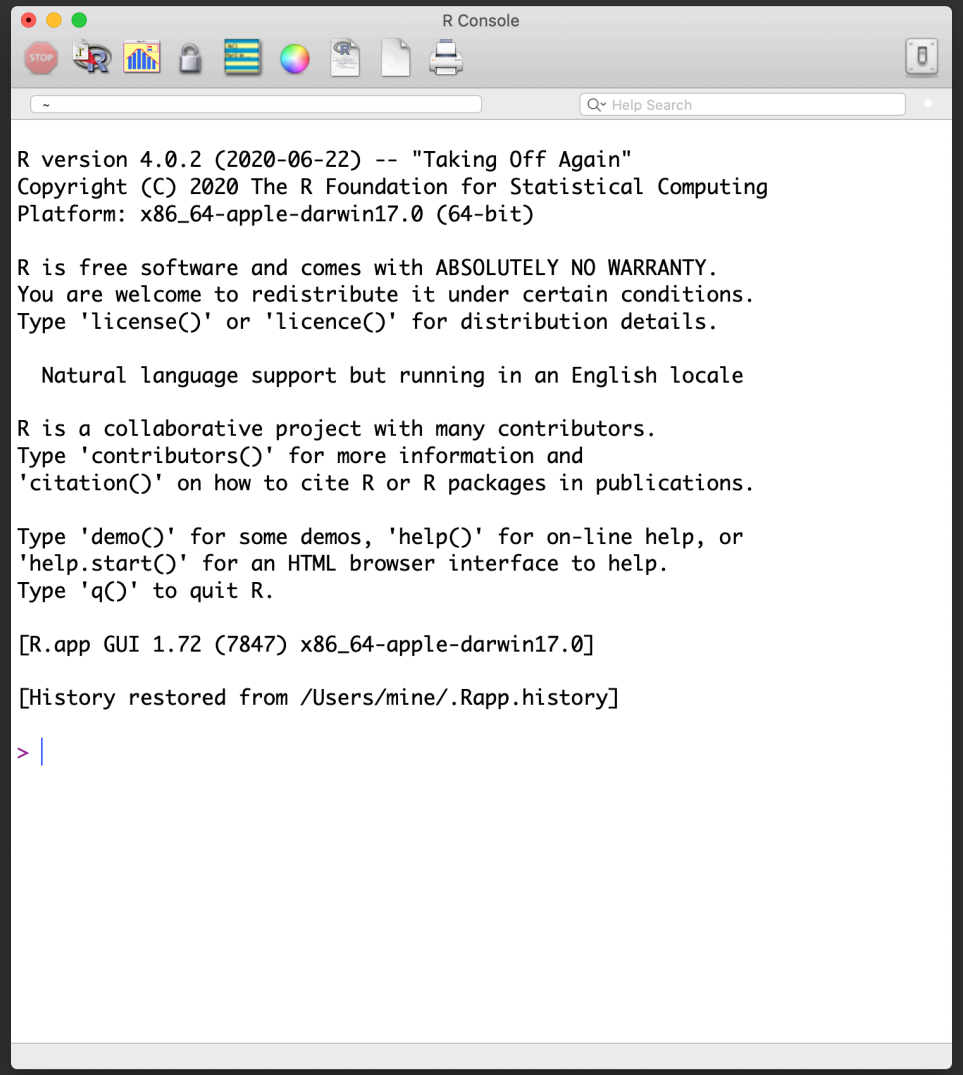
Data science

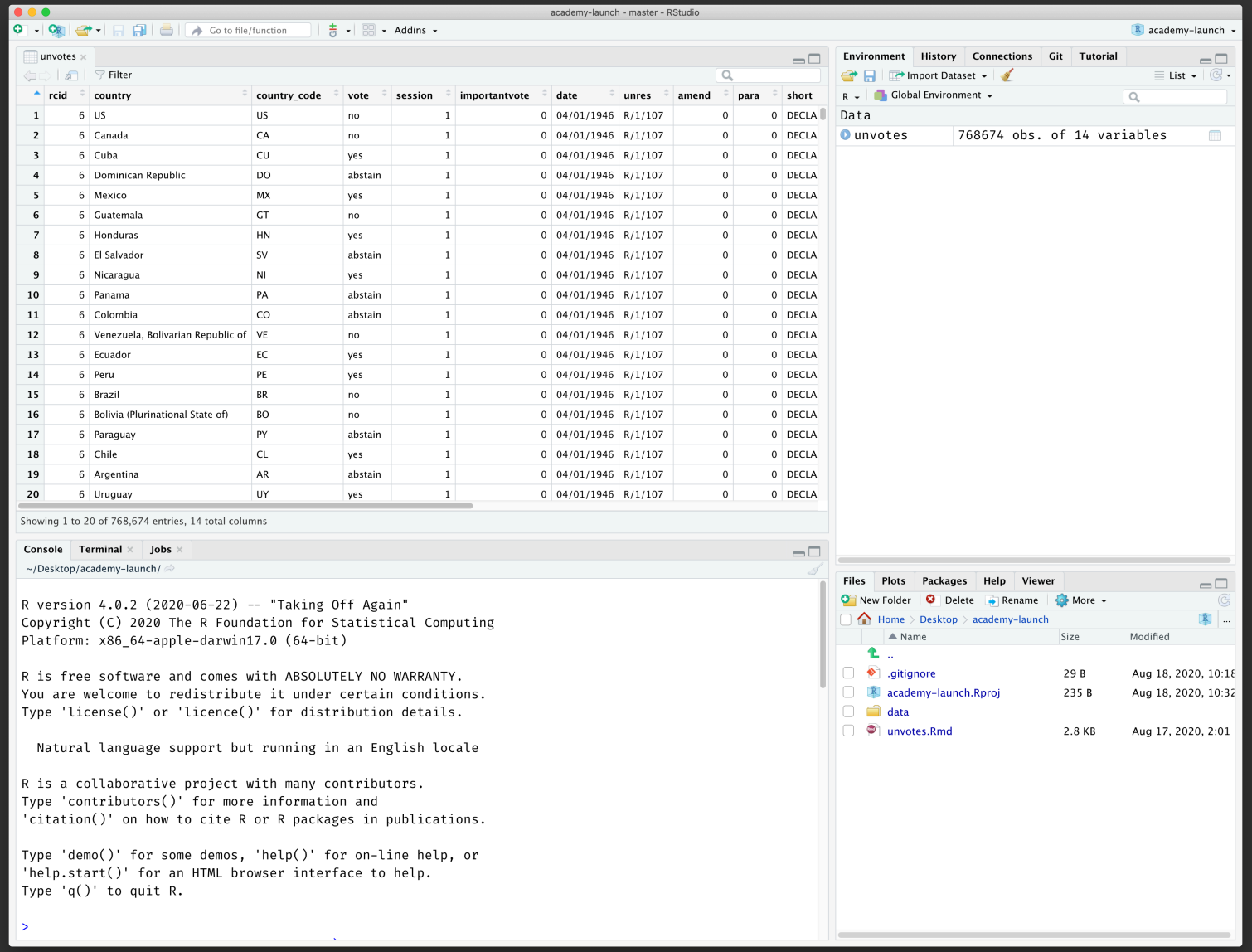
- Data science is an exciting discipline that allows you to turn raw data into understanding, insight, and knowledge.
- We're going to learn to do this in a `tidy` way -- more on that later!
- This is a course on introduction to data science, with an emphasis on statistical thinking.

Software

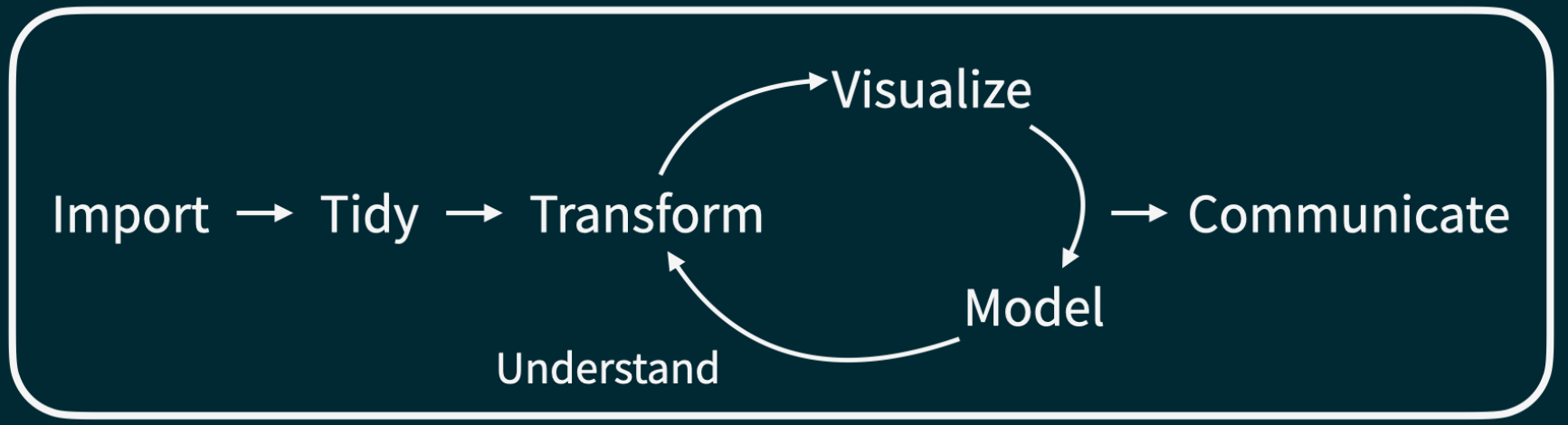


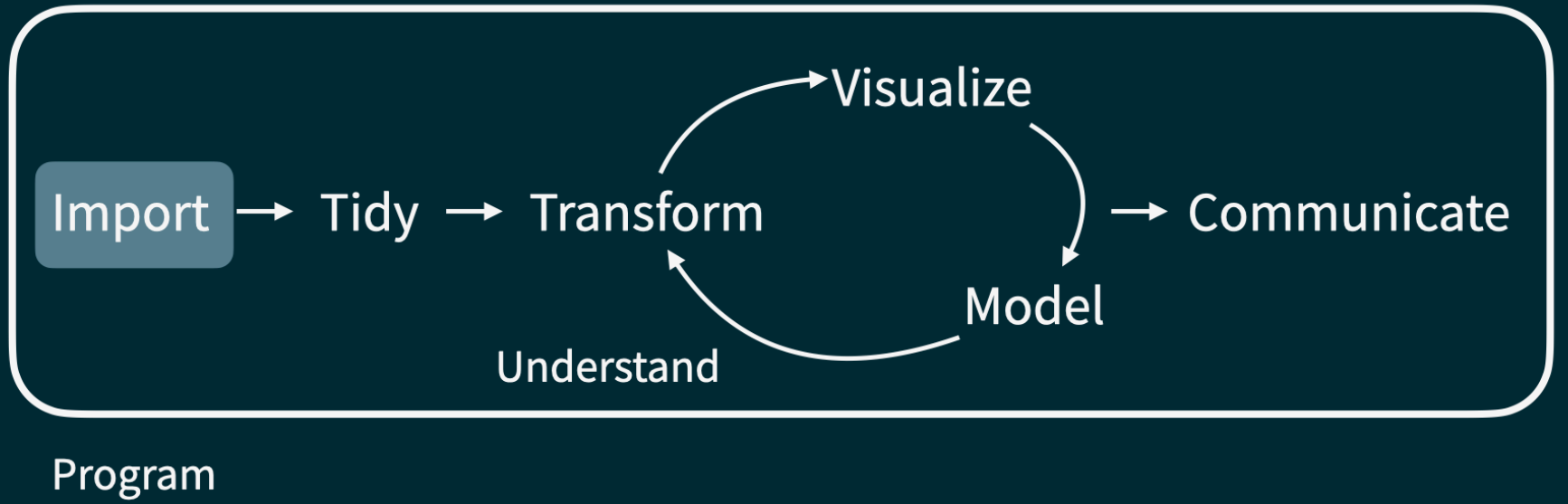
	A	B	C	D	E	F	G	H	I	J	K
1	rcid	country	country_code	vote	session	importantvote	date	unres	amend	para	short
2	6	US	US	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
3	6	Canada	CA	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
4	6	Cuba	CU	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
5	6	Dominican Republic	DO	abstain	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
6	6	Mexico	MX	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
7	6	Guatemala	GT	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
8	6	Honduras	HN	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
9	6	El Salvador	SV	abstain	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
10	6	Nicaragua	NI	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
11	6	Panama	PA	abstain	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
12	6	Colombia	CO	abstain	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
13	6	Venezuela, Bolivarian Republic of	VE	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
14	6	Ecuador	EC	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
15	6	Peru	PE	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
16	6	Brazil	BR	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
17	6	Bolivia (Plurinational State of)	BO	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
18	6	Paraguay	PY	abstain	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
19	6	Chile	CL	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
20	6	Argentina	AR	abstain	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
21	6	Uruguay	UY	yes	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
22	6	UK & NI	GB	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
23	6	Netherlands	NL	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
24	6	Belgium	BE	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
25	6	Luxembourg	LU	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
26	6	France	FR	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
27	6	Poland	PL	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
28	6	Czechoslovakia	CS	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
29	6	Yugoslavia	YU	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
30	6	Greece	GR	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
31	6	Russian Federation	RU	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
32	6	Ukraine	UA	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
33	6	Belarus	BY	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
34	6	Norway	NO	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS
35	6	Denmark	DK	no	1	0	04/01/1946	R/1/107	0	0	DECLARATION OF HUMAN RIGHTS

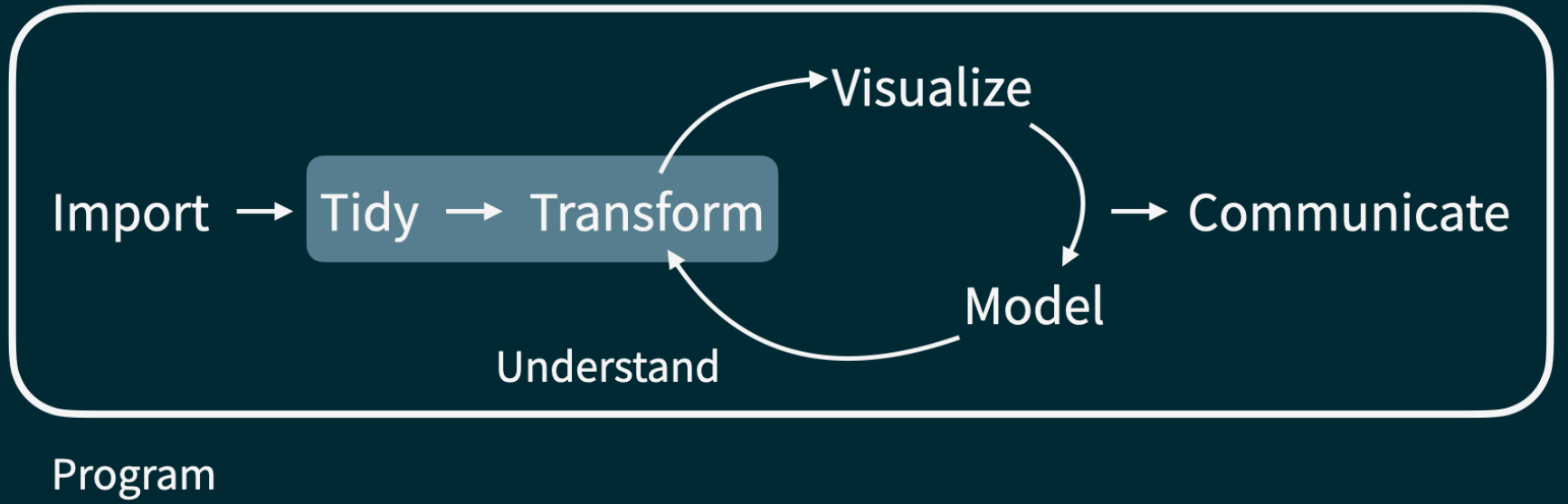

R Console
~
Help Search
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
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.
[R.app GUI 1.72 (7847) x86_64-apple-darwin17.0]
[History restored from /Users/mine/.Rapp.history]
>

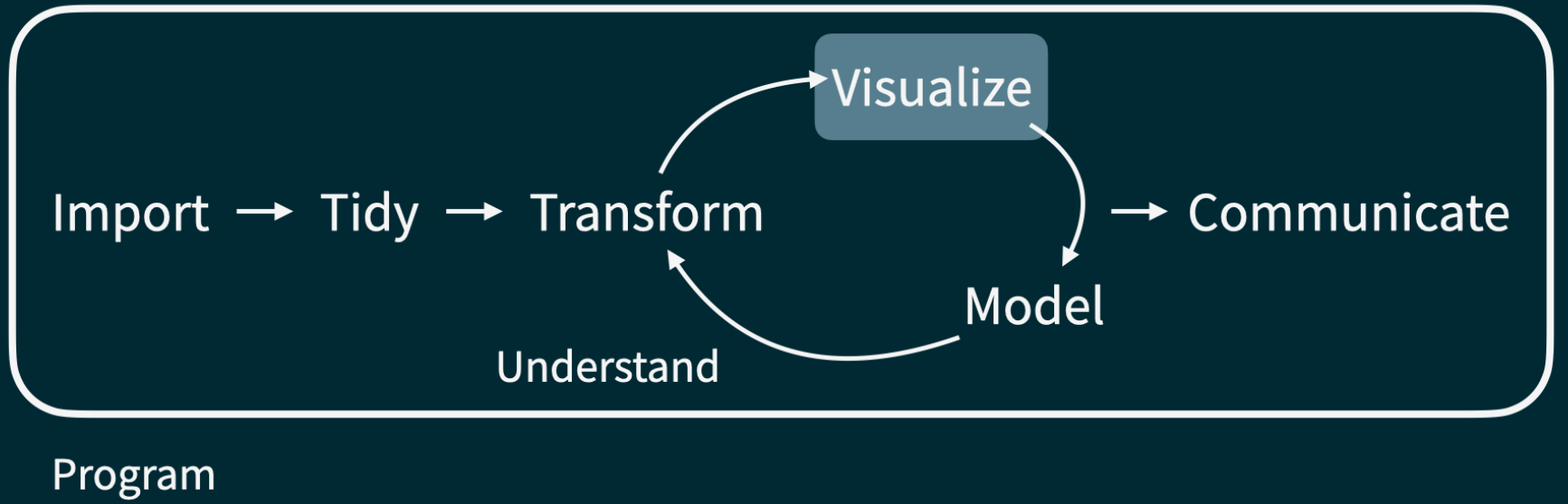

academy-launch - master - RStudio
Go to file/function
Addins
academy-launch
unvotes
Filter
rcid country country_code vote session importantvote date unres amend para short
1 6 US US no 1 0 04/01/1946 R/1/107 0 0 DECLA
2 6 Canada CA no 1 0 04/01/1946 R/1/107 0 0 DECLA
3 6 Cuba CU yes 1 0 04/01/1946 R/1/107 0 0 DECLA
4 6 Dominican Republic DO abstain 1 0 04/01/1946 R/1/107 0 0 DECLA
5 6 Mexico MX yes 1 0 04/01/1946 R/1/107 0 0 DECLA
6 6 Guatemala GT no 1 0 04/01/1946 R/1/107 0 0 DECLA
7 6 Honduras HN yes 1 0 04/01/1946 R/1/107 0 0 DECLA
8 6 El Salvador SV abstain 1 0 04/01/1946 R/1/107 0 0 DECLA
9 6 Nicaragua NI yes 1 0 04/01/1946 R/1/107 0 0 DECLA
10 6 Panama PA abstain 1 0 04/01/1946 R/1/107 0 0 DECLA
11 6 Colombia CO abstain 1 0 04/01/1946 R/1/107 0 0 DECLA
12 6 Venezuela, Bolivarian Republic of VE no 1 0 04/01/1946 R/1/107 0 0 DECLA
13 6 Ecuador EC yes 1 0 04/01/1946 R/1/107 0 0 DECLA
14 6 Peru PE yes 1 0 04/01/1946 R/1/107 0 0 DECLA
15 6 Brazil BR no 1 0 04/01/1946 R/1/107 0 0 DECLA
16 6 Bolivia (Plurinational State of) BO no 1 0 04/01/1946 R/1/107 0 0 DECLA
17 6 Paraguay PY abstain 1 0 04/01/1946 R/1/107 0 0 DECLA
18 6 Chile CL yes 1 0 04/01/1946 R/1/107 0 0 DECLA
19 6 Argentina AR abstain 1 0 04/01/1946 R/1/107 0 0 DECLA
20 6 Uruguay UY yes 1 0 04/01/1946 R/1/107 0 0 DECLA
Showing 1 to 20 of 768,674 entries, 14 total columns
Console
Terminal
Jobs
~/Desktop/academy-launch/
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (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.
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
History
Connections
Git
Tutorial
Import Dataset
List
R
Global Environment
Data
unvotes
768674 obs. of 14 variables
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home
Desktop
academy-launch
Name
Size
Modified
..
.gitignore
29 B
Aug 18, 2020, 10:18
academy-launch.Rproj
235 B
Aug 18, 2020, 10:32
data
unvotes.Rmd
2.8 KB
Aug 17, 2020, 2:01

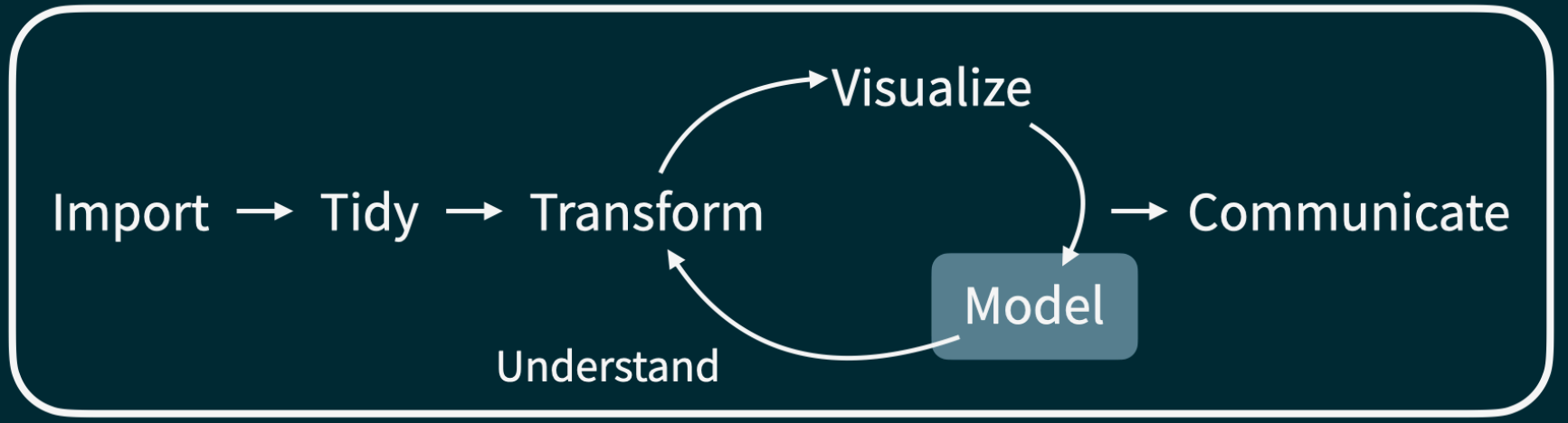
Data science life cycle

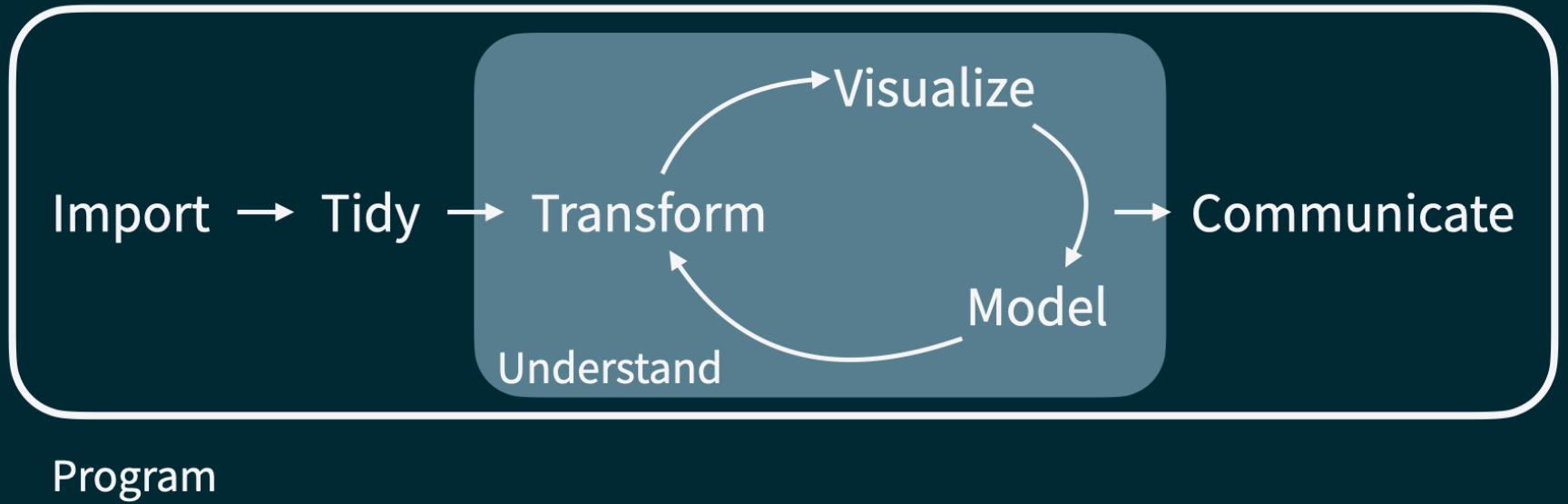

Visualize
Import
Tidy
Transform
Communicate
Model
Understand
Program


Visualize
Import
Tidy
Transform
Communicate
Model
Understand
Program


Visualize
Import
Tidy
Transform
Communicate
Model
Understand
Program


Visualize
Import
Tidy
Transform
Communicate
Model
Understand
Program


Visualize
Import
Tidy
Transform
Communicate
Model
Understand
Program


Import
Tidy
Transform
Visualize
Model
Understand
Communicate
Program

Google Trend Index

Trend index

Travel

Date range

2012-01-01 to 2017-07-31

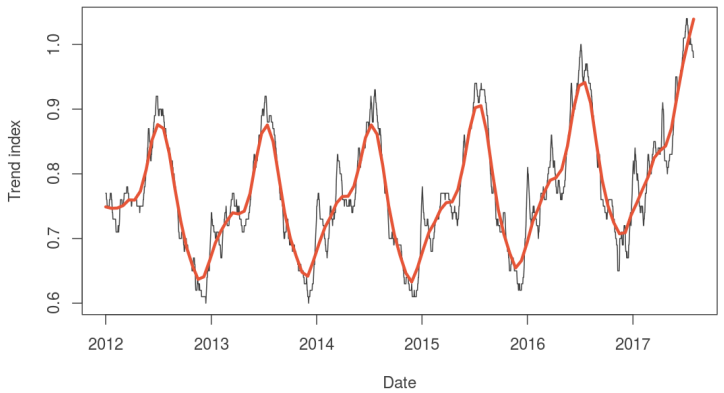
Overlay smooth trend line

Smoother span:

0.05 1

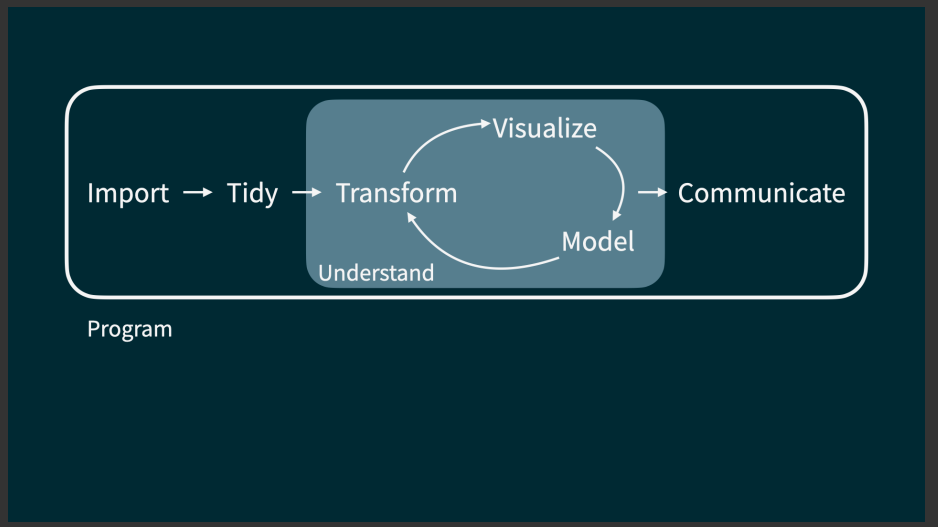
0.01 0.11 0.21 0.31 0.41 0.51 0.61 0.71 0.81 0.91 1

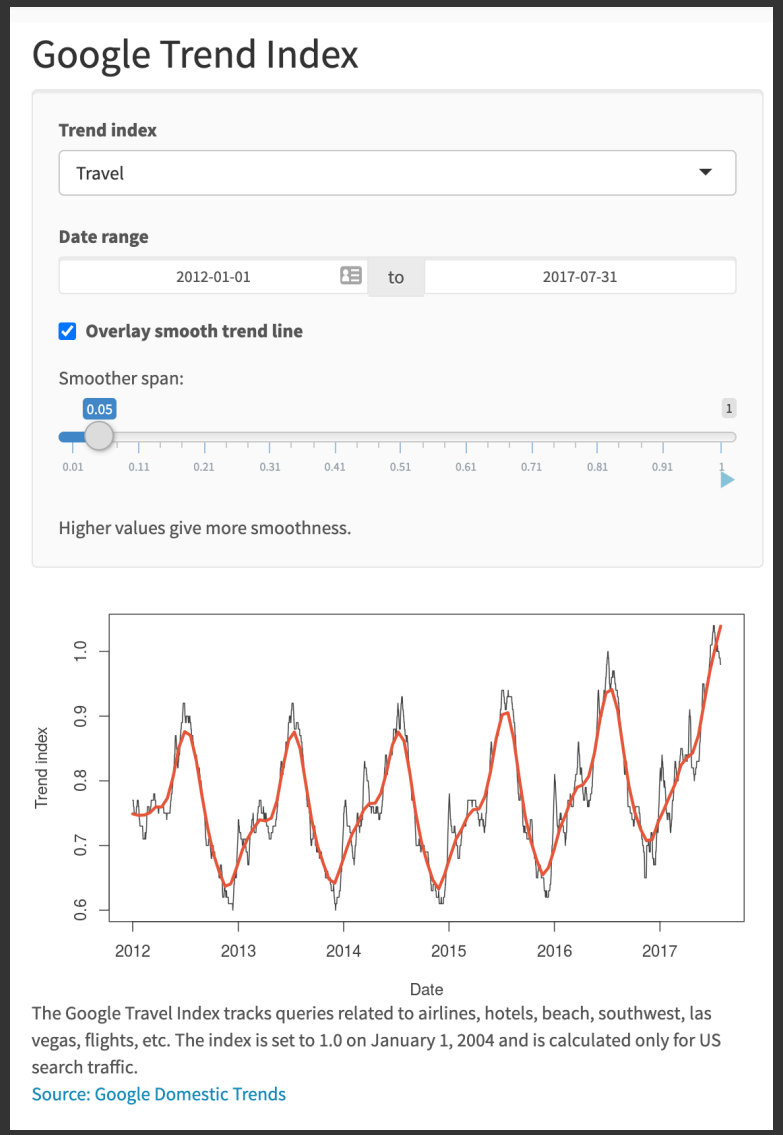
Higher values give more smoothness.

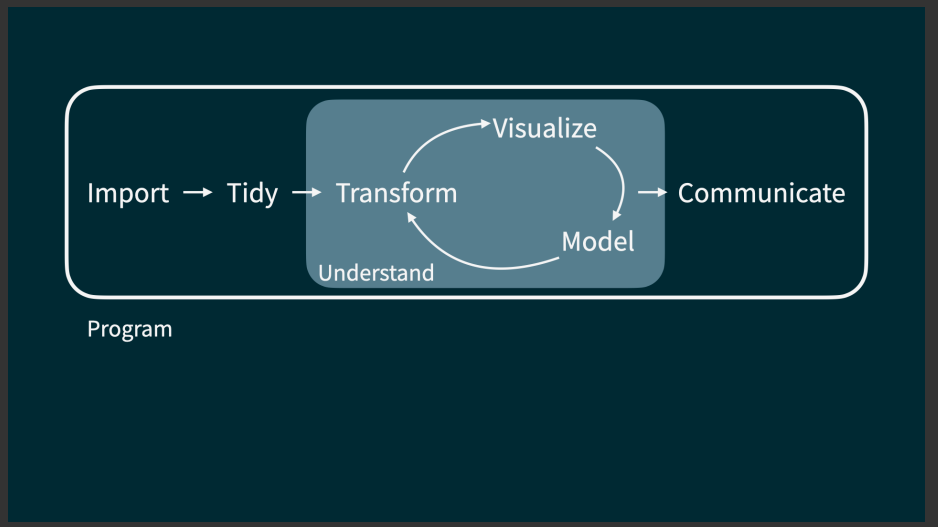


The Google Travel Index tracks queries related to airlines, hotels, beach, southwest, las vegas, flights, etc. The index is set to 1.0 on January 1, 2004 and is calculated only for US search traffic.

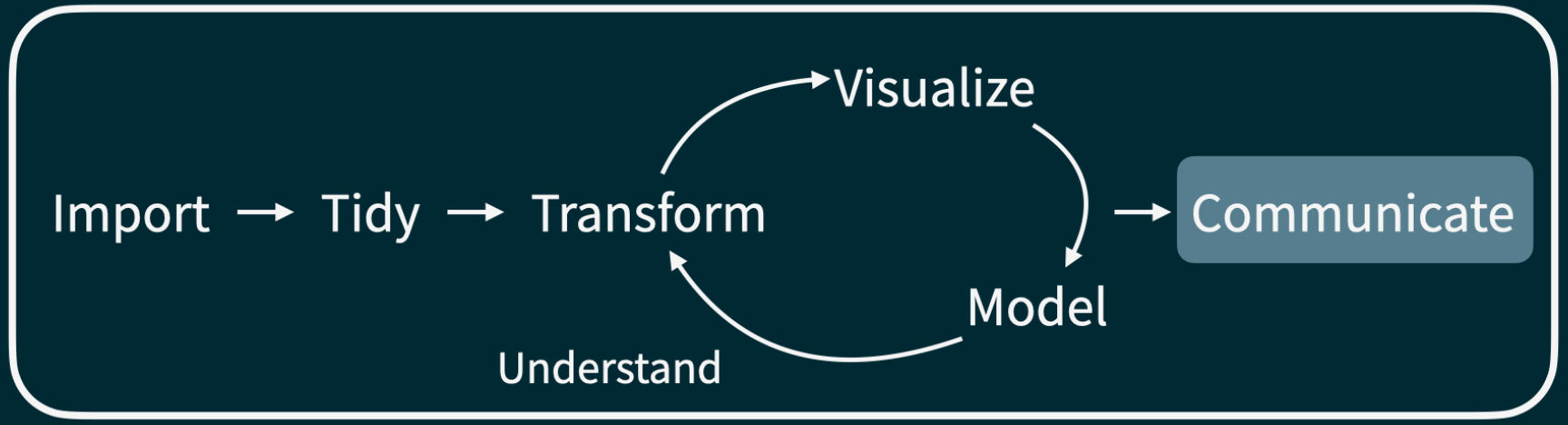
Source: Google Domestic Trends

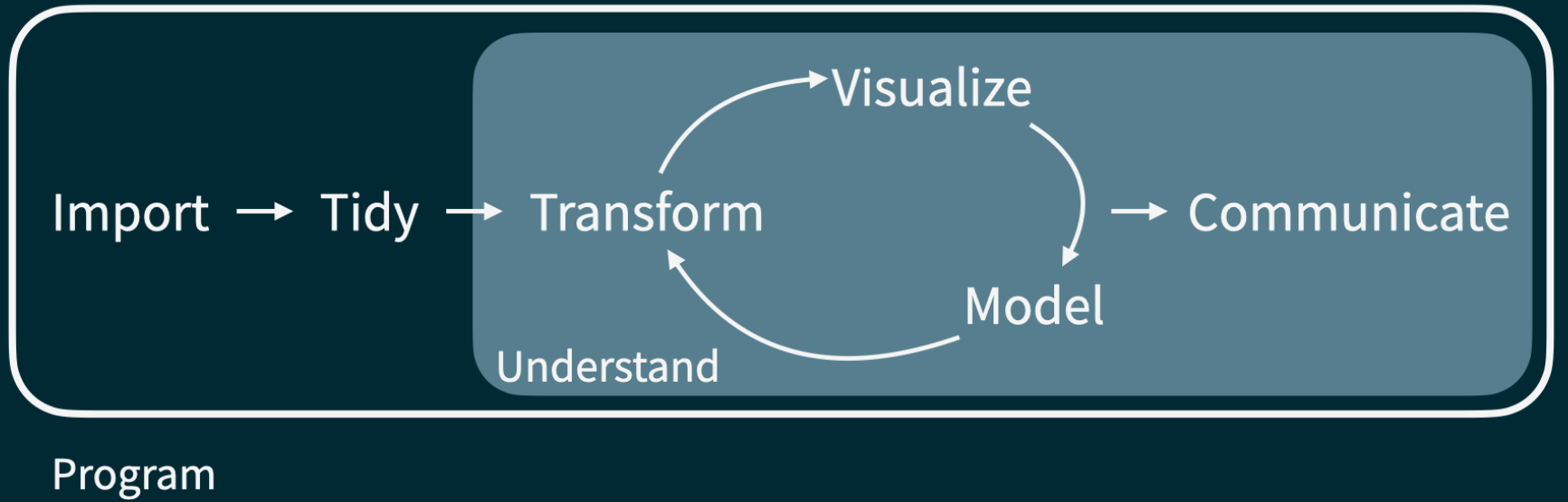


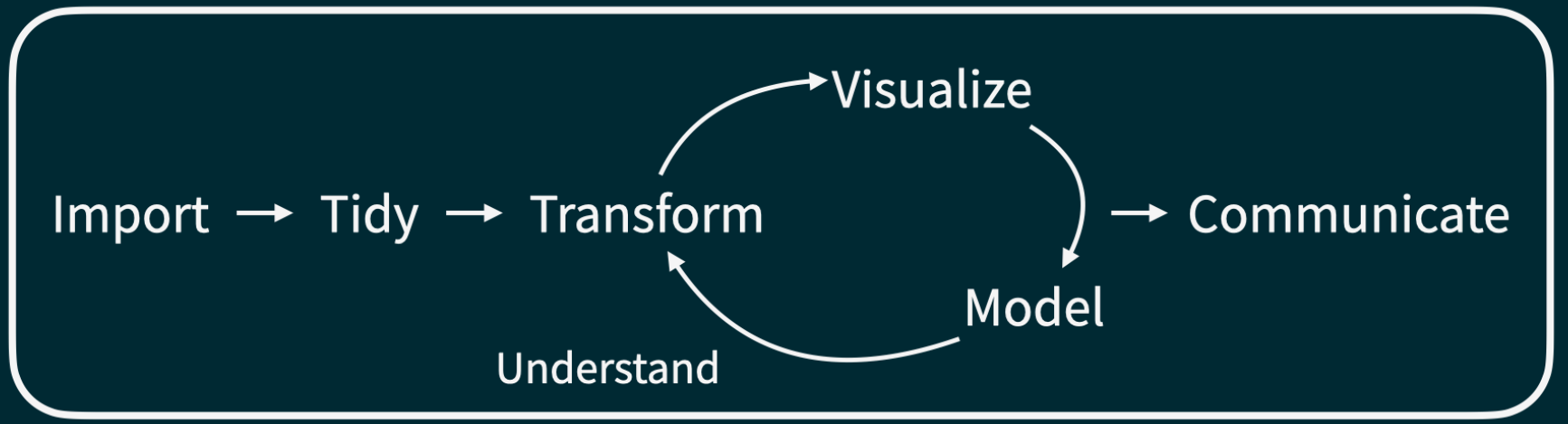





```
## # A tibble: 5 × 2
##   date              season
##   <chr>             <chr>
## 1 23 January 2017   winter
## 2 4 March 2017      spring
## 3 14 June 2017      summer
## 4 1 September 2017  fall
## 5 ...               ...
```


Visualize
Import
Tidy
Transform
Communicate
Model
Understand
Program


Import
Tidy
Transform
Visualize
Model
Understand
Communicate
Program


Visualize
Import
Tidy
Transform
Communicate
Model
Understand

Program

Course toolkit

Doing data science

- Programming:
  - R
  - RStudio
  - tidyverse
  - R Markdown

Learning goals

By the end of the course, you will be able to...

- gain insight from data
- gain insight from data, **reproducibly**
- gain insight from data, reproducibly, **using modern programming tools and techniques**
- gain insight from data, reproducibly **and collaboratively**, using modern programming tools and techniques
- gain insight from data, reproducibly **(with literate programming and version control)** and collaboratively, using modern programming tools and techniques

Reproducible data analysis

Reproducibility checklist

> What does it mean for a data analysis to be "reproducible"?

Near-term goals:

- Are the tables and figures reproducible from the code and data?
- Does the code actually do what you think it does?
- In addition to what was done, is it clear *why* it was done?

Long-term goals:

- Can the code be used for other data?
- Can you extend the code to do other things?

Toolkit for reproducibility

- Scriptability $\rightarrow$ R
- Literate programming (code, narrative, output in one place) $\rightarrow$ R Markdown
- Version control $\rightarrow$ Git / GitHub

R and RStudio

R and RStudio

- R is an open-source statistical **programming language**
- R is also an environment for statistical computing and graphics
- It's easily extensible with *packages*

- RStudio is a convenient interface for R called an **IDE** (integrated development environment), e.g. *"I write R code in the RStudio IDE"*
- RStudio is not a requirement for programming with R, but it's very commonly used by R programmers and data scientists

R packages

- **Packages** are the fundamental units of reproducible R code. They include reusable R functions, the documentation that describes how to use them, and sample data[1]
- As of September 2020, there are over 16,000 R packages available on **CRAN** (the Comprehensive R Archive Network)[2]
- We're going to work with a small (but important) subset of these!

[1] Wickham and Bryan, R Packages.

[2] CRAN contributed packages.

A short list (for now) of R essentials

- Functions are (most often) verbs, followed by what they will be applied to in parentheses:

```
do_this(to_this)
do_that(to_this, to_that, with_those)
```

- Packages are installed with the `install.packages` function and loaded with the `library` function, once per session:

```
install.packages("package_name")
library(package_name)
```

R essentials (continued)

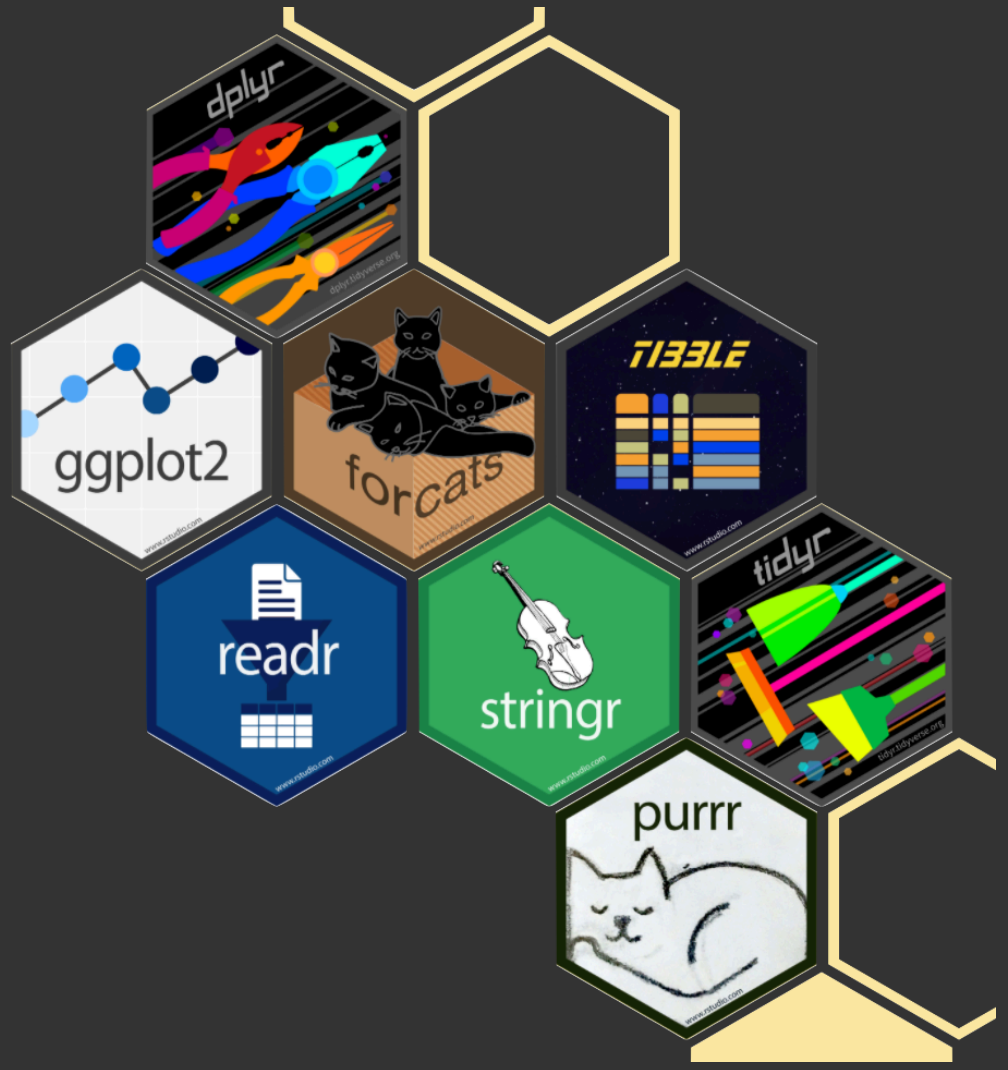
- Columns (variables) in data frames are accessed with `$`:

```
dataframe$var_name
```

- Object documentation can be accessed with `?`

```
?mean
```

tidyverse



tidyverse.org

- The **tidyverse** is an opinionated collection of R packages designed for data science
- All packages share an underlying philosophy and a common grammar

rmarkdown

rmarkdown.rstudio.com

- **rmarkdown** and the various packages that support it enable R users to write their code and prose in reproducible computational documents
- We will generally refer to R Markdown documents (with `.Rmd` extension), e.g. *"Do this in your R Markdown document"* and rarely discuss loading the rmarkdown package

Quarto

https://quarto.org/

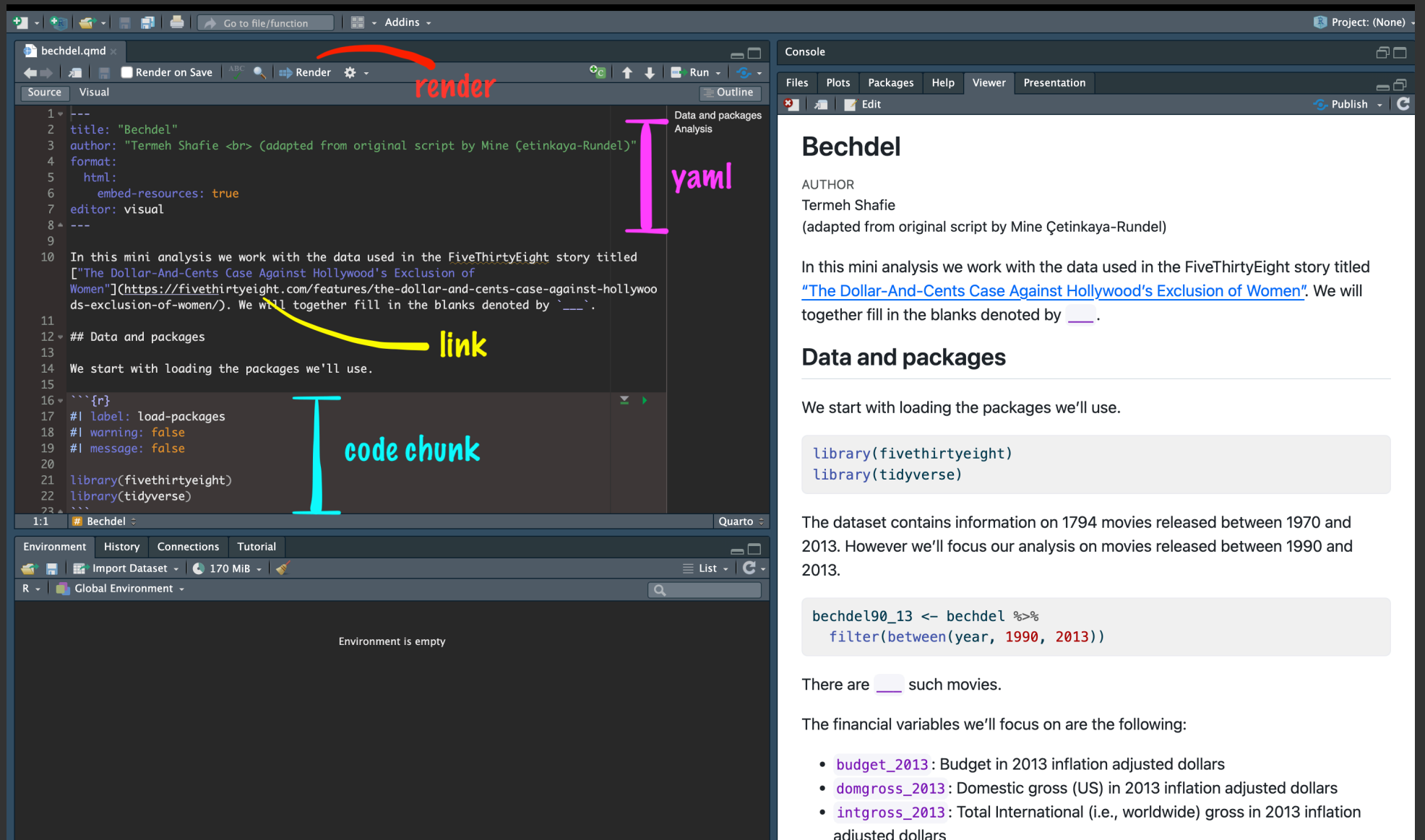
- An open-source scientific and technical publishing system
- Create dynamic content with Python, R, Julia, and Observable.
- Publish reproducible, production quality articles, presentations, dashboards, websites, blogs, and books in HTML, PDF, MS Word, ePub, and more.

R Markdown and Quarto

R Markdown and Quarto

- Fully reproducible reports -- each time you knit the analysis is ran from the beginning
- Simple markdown syntax for text
- Code goes in chunks, defined by three backticks, narrative goes outside of chunks

Tour: Quarto


bechdel.qmd
render
yaml
link
code chunk
Data and packages
Analysis
Environment is empty
Bechdel
AUTHOR
Termeh Shafie
(adapted from original script by Mine Çetinkaya-Rundel)
In this mini analysis we work with the data used in the FiveThirtyEight story titled "The Dollar-And-Cents Case Against Hollywood's Exclusion of Women". We will together fill in the blanks denoted by ___.
Data and packages
We start with loading the packages we'll use.
library(fivethirtyeight)
library(tidyverse)
The dataset contains information on 1794 movies released between 1970 and 2013. However we'll focus our analysis on movies released between 1990 and 2013.
bechdel90_13 <- bechdel %>%
filter(between(year, 1990, 2013))
There are ___ such movies.
The financial variables we'll focus on are the following:
budget_2013: Budget in 2013 inflation adjusted dollars
domgross_2013: Domestic gross (US) in 2013 inflation adjusted dollars
intgross_2013: Total International (i.e., worldwide) gross in 2013 inflation adjusted dollars

Environments

> The environment of your R Markdown/Quarto document is separate from the Console!

Remember this, and expect it to bite you a few times as you're learning to work with them!

How will we use R Markdown and Quarto?

- Every assignment / report / project / etc. is an R Markdown document
- You'll always have a template R Markdown document to start with

What's with all the hexes?

Mitchell O'Hara-Wild, useR! 2018 feature wall

Practical: Bechdel + Quarto

- The Bechdel test asks whether a work of fiction features at least two women who talk to each other about something other than a man, and there must be two women named characters.