Data Analytics

Lecture Series: Part 1

Peter J. Mattingly

pjm407@nyu.edu

Welcome!

Welcome!

In the course, we will:

Welcome!

In the course, we will:

- Apply data analytics to real estate using R

Welcome!

In the course, we will:

- Apply data analytics to real estate using R
- Create reproducible working environments

Welcome!

In the course, we will:

- Apply data analytics to real estate using R
- Create reproducible working environments
- Visualize insight and information

Overview

In this section, we will:

Overview

In this section, we will:

- Learn R Studio and R Markdown basics

Overview

In this section, we will:

- Learn R Studio and R Markdown basics
- Packages and API Keys

Overview

In this section, we will:

- Learn R Studio and R Markdown basics
- Packages and API Keys
- Data wrangling concepts

@tladeras

Data

- What is it?

@tladeras

Data

- What is it?
- What does it look like?

@tladeras

Data

	GeoFips	GeoName	2007	2010	2015
1	998	United States (Metropolitan Portion)	12821473605	13265807959	16264292775
2	10180	Abilene, TX (Metropolitan Statistical Area)	5406560	5861966	6836185
3	10420	Akron, OH (Metropolitan Statistical Area)	28749370	28922844	34122942
4	10500	Albany, GA (Metropolitan Statistical Area)	5101790	5109457	5434079
5	10540	Albany-Lebanon, OR (Metropolitan Statistical Area)	3643627	3460440	4116193
6	10580	Albany-Schenectady-Troy, NY (Metropolitan Statistical...	42731610	47044326	54546375
7	10740	Albuquerque, NM (Metropolitan Statistical Area)	35108392	36388647	39419488
8	10780	Alexandria, LA (Metropolitan Statistical Area)	4636424	5369983	5862965
9	10900	Allentown-Bethlehem-Easton, PA-NJ (Metropolitan Sta...	33647001	35746169	41542046
10	11020	Altoona, PA (Metropolitan Statistical Area)	4643209	4962165	5547413
11	11100	Amarillo, TX (Metropolitan Statistical Area)	9788225	10722023	12784064

@tladeras

Data

	GeoFips	GeoName	2007	2010	2015
1	998	United States (Metropolitan Portion)	12821473605	13265807959	16264292775
2	10180	Abilene, TX (Metropolitan Statistical Area)	5406560	5861966	6836185
3	10420	Akron, OH (Metropolitan Statistical Area)	28749370	28922844	34122942
4	10500	Albany, GA (Metropolitan Statistical Area)	5101790	5109457	5434079
5	10540	Albany-Lebanon, OR (Metropolitan Statistical Area)	3643627	3460440	4116193
6	10580	Albany-Schenectady-Troy, NY (Metropolitan Statistical...	42731610	47044326	54546375
7	10740	Albuquerque, NM (Metropolitan Statistical Area)	35108392	36388647	39419488
8	10780	Alexandria, LA (Metropolitan Statistical Area)	4636424	5369983	5862965
9	10900	Allentown-Bethlehem-Easton, PA-NJ (Metropolitan Sta...	33647001	35746169	41542046
10	11020	Altoona, PA (Metropolitan Statistical Area)	4643209	4962165	5547413
11	11100	Amarillo, TX (Metropolitan Statistical Area)	9788225	10722023	12784064

@tladeras

Data

	GeoFips	GeoName	2007	2010	2015
1	998	United States (Metropolitan Portion)	12821473605	13265807959	16264292775
2	10180	Abilene, TX (Metropolitan Statistical Area)	5406560	5861966	6836185
3	10420	Akron, OH (Metropolitan Statistical Area)	28749370	28922844	34122942
4	10500	Albany, GA (Metropolitan Statistical Area)	5101790	5109457	5434079
5	10540	Albany-Lebanon, OR (Metropolitan Statistical Area)	3643627	3460440	4116193
6	10580	Albany-Schenectady-Troy, NY (Metropolitan Statistical...	42731610	47044326	54546375
7	10740	Albuquerque, NM (Metropolitan Statistical Area)	35108392	36388647	39419488
8	10780	Alexandria, LA (Metropolitan Statistical Area)	4636424	5369983	5862965
9	10900	Allentown-Bethlehem-Easton, PA-NJ (Metropolitan Sta...	33647001	35746169	41542046
10	11020	Altoona, PA (Metropolitan Statistical Area)	4643209	4962165	5547413
11	11100	Amarillo, TX (Metropolitan Statistical Area)	9788225	10722023	12784064

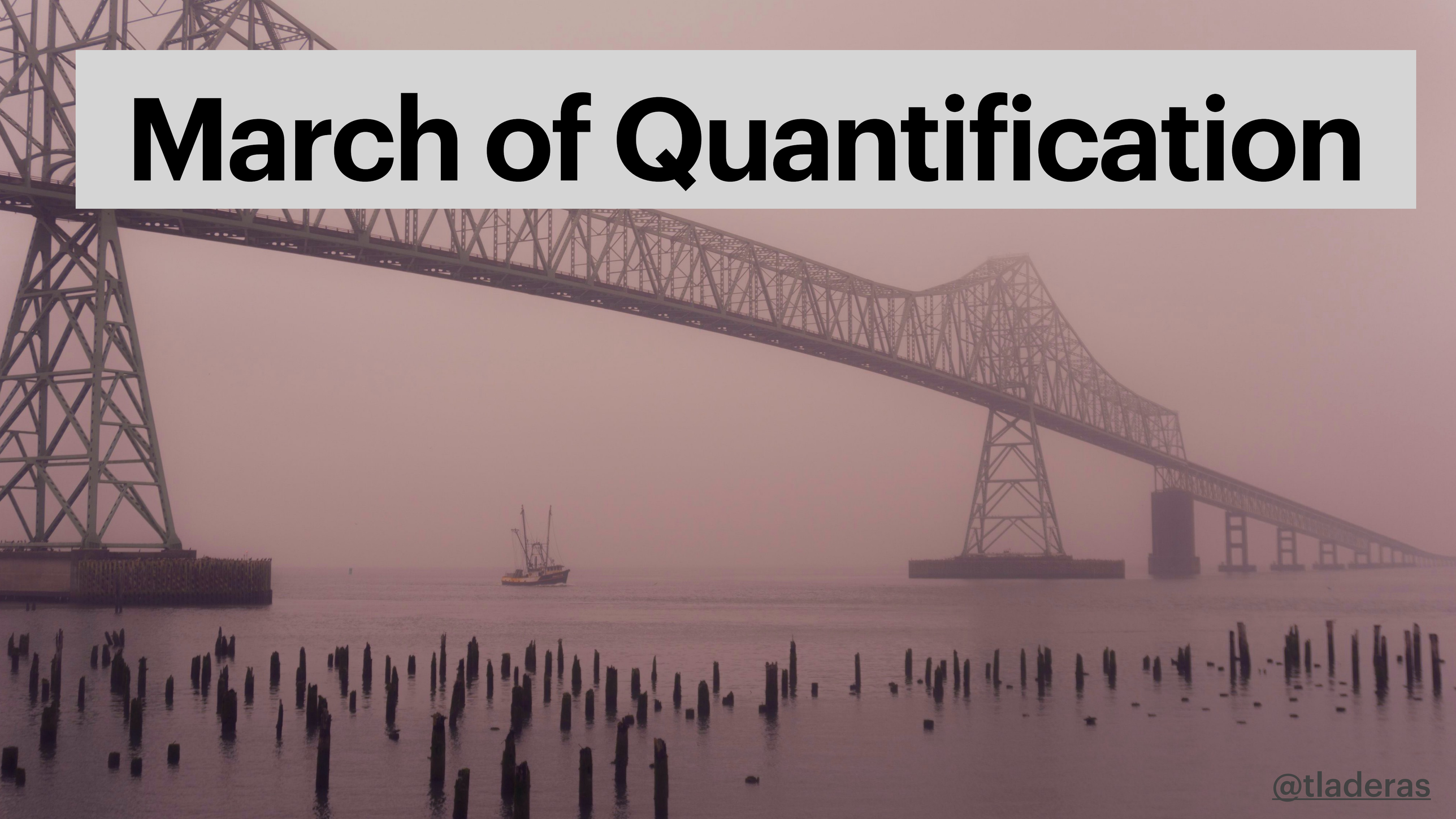
@tladeras

Data

	GeoFips	GeoName	2007	2010	2015
1	998	United States (Metropolitan Portion)	12821473605	13265807959	16264292775
2	10180	Abilene, TX (Metropolitan Statistical Area)	5406560	5861966	6836185
3	10420	Akron, OH (Metropolitan Statistical Area)	28749370	28922844	34122942
4	10500	Albany, GA (Metropolitan Statistical Area)	5101790	5109457	5434079
5	10540	Albany-Lebanon, OR (Metropolitan Statistical Area)	3643627	3460440	4116193
6	10580	Albany-Schenectady-Troy, NY (Metropolitan Statistical...	42731610	47044326	54546375
7	10740	Albuquerque, NM (Metropolitan Statistical Area)	35108392	36388647	39419488
8	10780	Alexandria, LA (Metropolitan Statistical Area)	4636424	5369983	5862965
9	10900	Allentown-Bethlehem-Easton, PA-NJ (Metropolitan Sta...	33647001	35746169	41542046
10	11020	Altoona, PA (Metropolitan Statistical Area)	4643209	4962165	5547413
11	11100	Amarillo, TX (Metropolitan Statistical Area)	9788225	10722023	12784064

@tladeras

March of Quantification



@tladeras

March of Quantification

Across all fields, disciplines, and industries :

@tladeras

March of Quantification

Across all fields, disciplines, and industries :

- Data more accessible via Internet & APIs

@tladeras

March of Quantification

Across all fields, disciplines, and industries :

- Data more accessible via Internet & APIs
- Methodologies more sophisticated & idiosyncratic

March of Quantification

Across all fields, disciplines, and industries :

- Data more accessible via Internet & APIs
- Methodologies more sophisticated & idiosyncratic
- Software more advanced (e.g., R, Python, Stata)

March of Quantification

Across all fields, disciplines, and industries :

- Data more accessible via Internet & APIs
- Methodologies more sophisticated & idiosyncratic
- Software more advanced (e.g., R, Python, Stata)
- Still need to uncover more

@tladeras

Software

@tladeras

Excel
Software
@tladeras

Excel

STATA®

Software

@tladeras

Software

Excel

STATA®

python

@tladeras

RStudio®
@tladeras

RStudio®

Pros :

@tladeras

RStudio®

Pros :

- Open source
- Array of packages
- Statistics
- Compatibility
- Graphics

@tladeras

RStudio

Pros :

- Open source
- Array of packages
- Statistics
- Compatibility
- Graphics

Cons :

@tladeras

RStudio

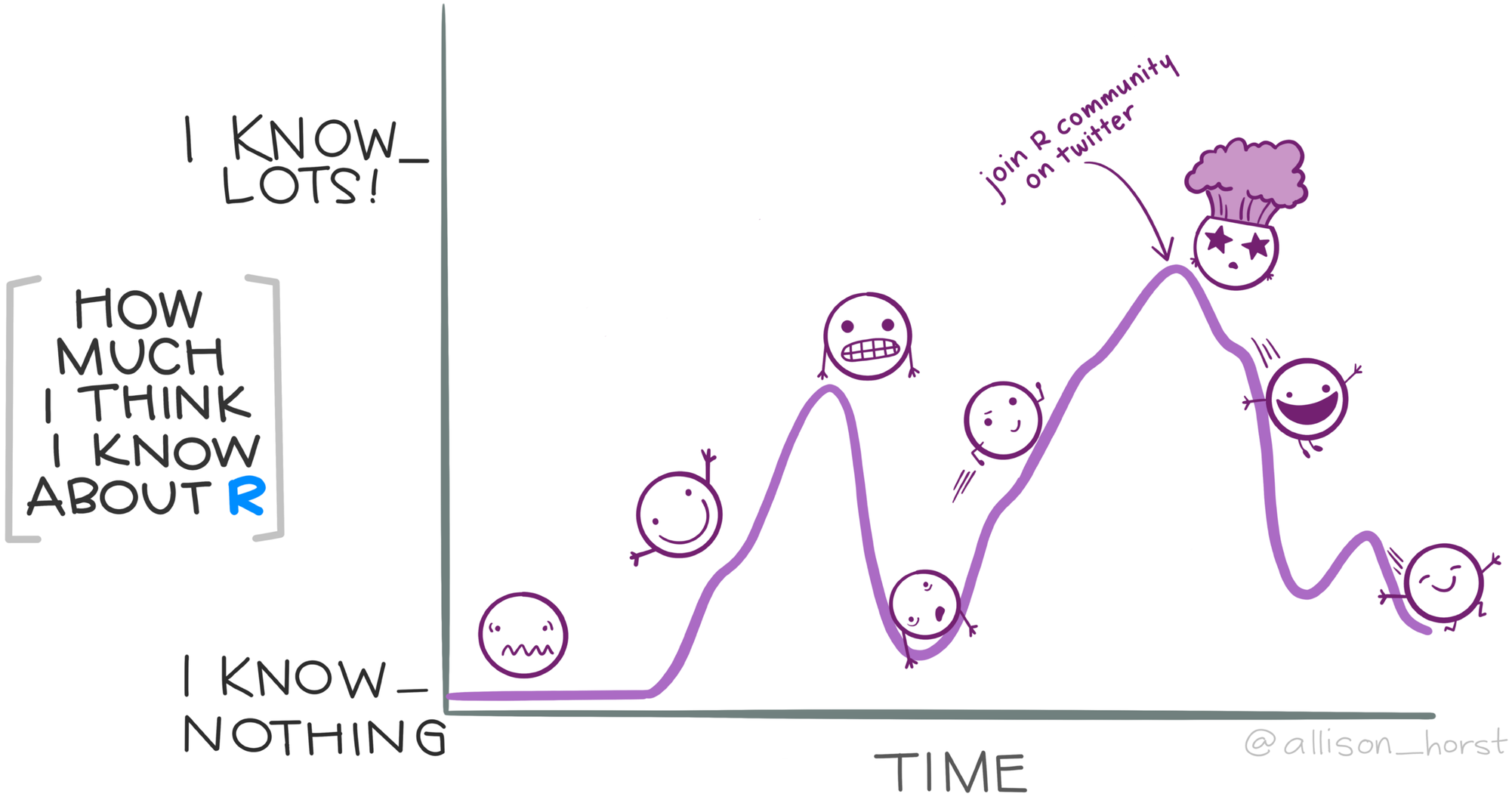
Pros :

- Open source
- Array of packages
- Statistics
- Compatibility
- Graphics

Cons :

- Memory
- Array of packages
- Security
- Learning curve

@tladeras


I KNOW LOTS!
HOW MUCH I THINK I KNOW ABOUT R
I KNOW NOTHING
join R community on twitter
TIME
@allison_horst

@tladeras

Terms

@tladeras

Terms

Packages

@tladeras

Terms

Packages :

- Imported and reusable

@tladeras

Terms

Packages :

- Bundled functions for cleaning, wrangling, and visualizing data

@tladeras

Packages

tidyverse

@tladeras

Packages

tidyverse

devtools

@tladeras

Packages

tidyverse

devtools

fredr

@tladeras

Packages

tidyverse

devtools

fredr

tidycensus

@tladeras

Terms

API Keys

@tladeras

Terms

API Keys :

- fredr

@tladeras

Register User Account - St. Lo

research.stlouisfed.org/useraccount/login/secure/

Apps

ECONOMIC RESEARCH
FEDERAL RESERVE BANK of ST. LOUIS

Economists | Research and Publications | The Research Division

Already have an account?
Sign In
Forgot your password?

Want to create a new account?
Register

Why Register?

- Subscribe to email updates for economic data series.
- Create personalized lists of economic data series.
- Save customized graphs and maps for later use.
- Build and share personalized dashboards with series that interest you.
- Access the FRED API to integrate data with your favorite software packages.
- Play FREDcast™.

Learn more about user accounts

API Key - St. Louis Fed

research.stlouisfed.org/useraccount/apikey

Apps

ECONOMIC RESEARCH
FEDERAL RESERVE BANK of ST. LOUIS

Economists ▾ Research and Publications ▾ The Research Division ▾

Your registered API key is:

YOUR API KEY HERE

Documentation is available on the St. Louis Fed web services website.

API Key

Describe the application or program you intend to write:

Utilizing the resources and information through FRED and import into R for analysis and visualization.

☑ **I have read and agree to the St. Louis Fed's Terms of Use, Privacy Notice & Policy, and Legal Notices,**

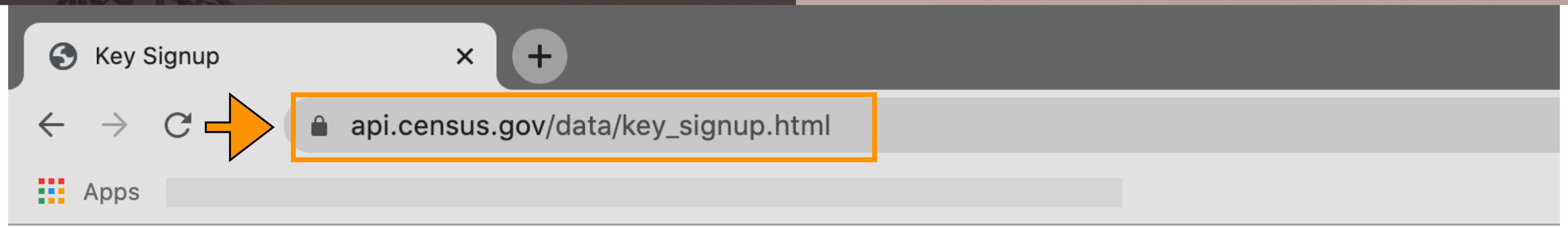
Request API Key

Terms

API Keys :

- tidycensus

@tladeras



Request A Key

Organization Name:

Email Address:

I agree to the terms of service

Submit Key Request

Terms

R Studio

@tladeras

Go to file/funct | Addins | Project: (None)

GreatRecession.Rmd | MEC_0412.Rmd | floodzone_censustract.Rmd | R Tutorial.Rmd | outlook_housing.Rmd | R Tutorial Par

Knit | Insert | Run

```
---
title: "R Tutorial"
author: "Mattingly"
date: "2/10/2020"
output: pdf_document
---

getwd()
setwd("/Users/petermattingly/Desktop/")

### creating a notebook chunk
'control' + 'option', then 'i'

```{r}

```

running individual lines of code
mac: 'command' then 'return'
pc: 'control' then 'enter'

assignment operator <-

creating pipe operator %>%
'command' 'shift' 'm' =

libraries and packages

```{r}
install.packages('data.table', 'tidyverse')
library(data.table)
library(tidyverse)
```
```

11:30 | creating a notebook chunk | R Markdown

Console | Terminal | R Markdown

```
+      xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+      main="Daily Interest Rates Since 2000", pch=16, col='blue')
Error in (function (formula, data = NULL, subset = NULL, na.action = na.fail,  :
  invalid type (list) for variable 'strptime(threemonth$value, "%Y-%m-%d")'
> plot(strptime(threemonth$value,"%Y-%m-%d"), strptime(tenyear$value,"%Y-%m-%d"),
+      xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+      main="Daily Interest Rates Since 2000", pch=16, col='blue')
Error in plot.window(...) : need finite 'xlim' values
In addition: Warning messages:
1: In min(x) : no non-missing arguments to min; returning Inf
2: In max(x) : no non-missing arguments to max; returning -Inf
3: In min(x) : no non-missing arguments to min; returning Inf
4: In max(x) : no non-missing arguments to max; returning -Inf
> plot(threemonth$value, tenyear$value,
+      xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+      main="Daily Interest Rates Since 2000", pch=16, col='blue')
> cor(tenyear$value ~ threemonth$value)
Error in cor(tenyear$value ~ threemonth$value) :
  supply both 'x' and 'y' or a matrix-like 'x'
> cor(tenyear$value, threemonth$value)
[1] 0.7608
> threemonth = drop_na(fredr(series_id = "DGS3MO", observation_start = as.Date("2000-01-01")))
> tenyear = drop_na(fredr(series_id = "DGS10", observation_start = as.Date("2000-01-01")))
> plot(threemonth$value, tenyear$value,
+      xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+      main="Daily Interest Rates Since 2000", pch=16, col='blue')
```

Environment | History | Connections

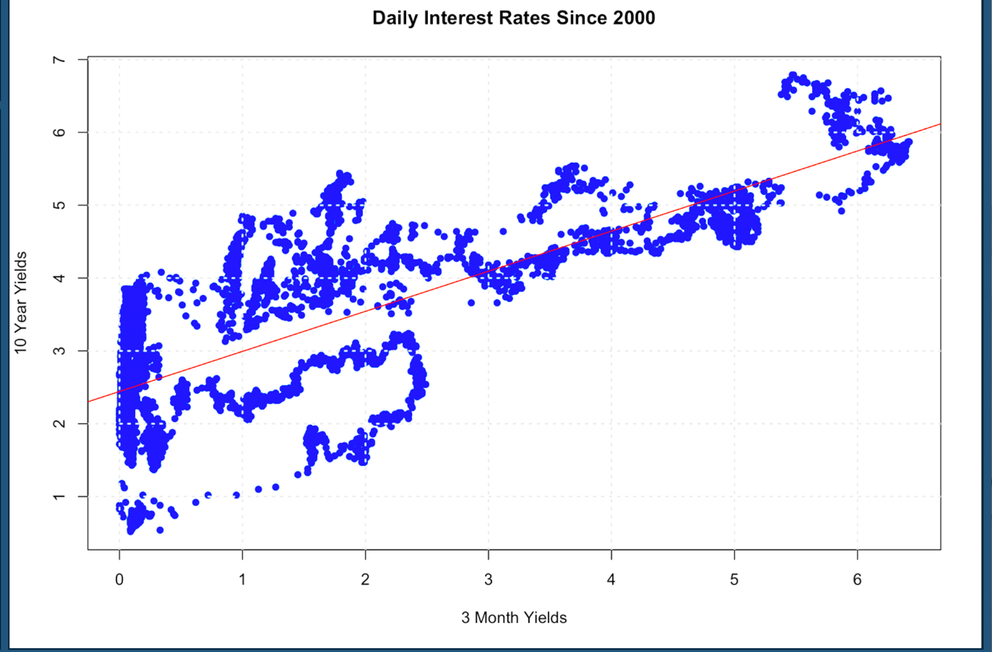
Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
dailyavg_table	tbl_df	7	2 KB	3 obs. of 7 variables
dailyavg_wtmeans	grouped_df	4	66.4 KB	1095 obs. of 4 variables
data1990	tbl_df	6	22 KB	373 obs. of 6 variables
data1990_2018_race_total	data.frame	5	8.7 KB	174 obs. of 5 variables
data1990_hisp	tbl_df	6	7 KB	62 obs. of 6 variables
data1990_main	tbl_df	6	19.1 KB	311 obs. of 6 variables
data1999_2000	grouped_df	5	4.4 KB	12 obs. of 5 variables
data1999_2000_total	data.frame	5	4.3 KB	66 obs. of 5 variables
data1999_2018_race_total	matrix	10	7.9 KB	List of 10
data1999_2018_total	data.frame	5	8.6 KB	174 obs. of 5 variables
f1	function	1	10.1 KB	function (x, y, p = 0)
geo_northern	data.table	9	30.6 KB	97 obs. of 9 variables
geospatial	data.table	9	73.7 KB	246 obs. of 9 variables
il	sf	6	1.4 MB	408 obs. of 6 variables
labTheme	function	1	18 KB	function (base_size = 48)
logo	rastergrob	12	1.8 MB	Large rastergrob (12 elements, 1.8 Mb)
model1	lm	12	1.3 MB	Large lm (12 elements, 1.3 Mb)
monthlyavg_countries	grouped_df	7	47 KB	730 obs. of 7 variables
name_region	data.table	5	38.5 KB	246 obs. of 5 variables
numbers	integer	10	96 B	int [1:10] 1 2 3 4 5 6 7 8 9 10
numlist	numeric	10	176 B	num [1:10] 1 2 3 4 5 6 7 8 9 10

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



@tladeras

Script

GreatRecession.Rmd | MEC_0412.Rmd | floodzone_censustract.Rmd | R Tutorial.Rmd | outlook_housing.Rmd | R Tutorial Par

```
---
title: "R Tutorial"
author: "Mattingly"
date: "2/10/2020"
output: pdf_document
---

getwd()
setwd("/Users/petermattingly/Desktop/")

### creating a notebook chu
'control' + 'option', then

```{r}

```

running individual lines
mac: 'command' then 'return'
pc: 'control' then 'enter'

assignment operator <-

creating pipe operator %>%
'command' 'shift' 'm' =

libraries and packages

```{r}
install.packages('data.table', 'tidyverse')
library(data.table)
library(tidyverse)
```
```

11:30 creating a notebook chunk — R Markdown

Console | Terminal | R Markdown

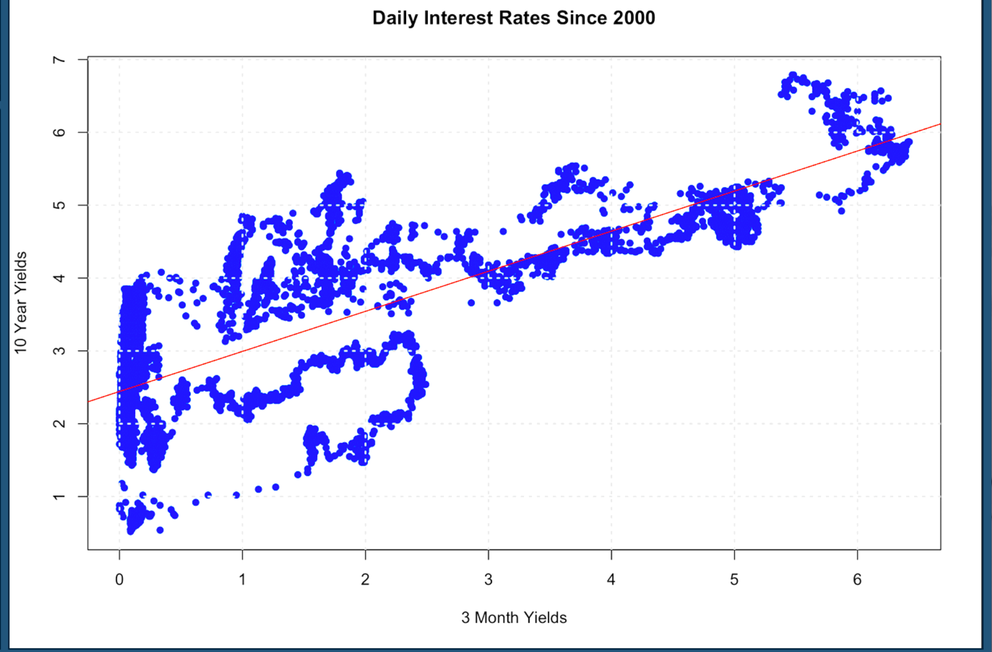
```
+      xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+      main="Daily Interest Rates Since 2000", pch=16, col='blue')
Error in (function (formula, data = NULL, subset = NULL, na.action = na.fail,  :
  invalid type (list) for variable 'strptime(threemonth$value, "%Y-%m-%d")'
> plot(strptime(threemonth$value,"%Y-%m-%d"), strptime(tenyear$value,"%Y-%m-%d"),
+      xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+      main="Daily Interest Rates Since 2000", pch=16, col='blue')
Error in plot.window(...) : need finite 'xlim' values
In addition: Warning messages:
1: In min(x) : no non-missing arguments to min; returning Inf
2: In max(x) : no non-missing arguments to max; returning -Inf
3: In min(x) : no non-missing arguments to min; returning Inf
4: In max(x) : no non-missing arguments to max; returning -Inf
> plot(threemonth$value, tenyear$value,
+      xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+      main="Daily Interest Rates Since 2000", pch=16, col='blue')
> cor(tenyear$value ~ threemonth$value)
Error in cor(tenyear$value ~ threemonth$value) :
  supply both 'x' and 'y' or a matrix-like 'x'
> cor(tenyear$value, threemonth$value)
[1] 0.7608
> threemonth = drop_na(fredr(series_id = "DGS3MO", observation_start = as.Date("2000-01-01")))
> tenyear = drop_na(fredr(series_id = "DGS10", observation_start = as.Date("2000-01-01")))
> plot(threemonth$value, tenyear$value,
+      xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+      main="Daily Interest Rates Since 2000", pch=16, col='blue')
```

Environment | History | Connections

Global Environment

Name	Type	Length	Size	Value
dailyavg_table	tbl_df	7	2 KB	3 obs. of 7 variables
dailyavg_wtmeans	grouped_df	4	66.4 KB	1095 obs. of 4 variables
data1990	tbl_df	6	22 KB	373 obs. of 6 variables
data1990_2018_race_total	data.frame	5	8.7 KB	174 obs. of 5 variables
data1990_hisp	tbl_df	6	7 KB	62 obs. of 6 variables
data1990_main	tbl_df	6	19.1 KB	311 obs. of 6 variables
data1999_2000	grouped_df	5	4.4 KB	12 obs. of 5 variables
data1999_2000_total	data.frame	5	4.3 KB	66 obs. of 5 variables
data1999_2018_race_total	matrix	10	7.9 KB	List of 10
data1999_2018_total	data.frame	5	8.6 KB	174 obs. of 5 variables
f1	function	1	10.1 KB	function (x, y, p = 0)
geo_northern	data.table	9	30.6 KB	97 obs. of 9 variables
geospatial	data.table	9	73.7 KB	246 obs. of 9 variables
il	sf	6	1.4 MB	408 obs. of 6 variables
labTheme	function	1	18 KB	function (base_size = 48)
logo	rastergrob	12	1.8 MB	Large rastergrob (12 elements, 1.8 Mb)
model1	lm	12	1.3 MB	Large lm (12 elements, 1.3 Mb)
monthlyavg_countries	grouped_df	7	47 KB	730 obs. of 7 variables
name_region	data.table	5	38.5 KB	246 obs. of 5 variables
numbers	integer	10	96 B	int [1:10] 1 2 3 4 5 6 7 8 9 10
numlist	numeric	10	176 B	num [1:10] 1 2 3 4 5 6 7 8 9 10

Files | Plots | Packages | Help | Viewer



@tladeras

Script

Console

GreatRecession.Rmd | MEC_0412.Rmd | floodzone_censustract.Rmd | R Tutorial.Rmd | outlook_housing.Rmd | R Tutorial Par

```
---
title: "R Tutorial"
author: "Mattingly"
date: "2/10/2020"
output: pdf_document
---

getwd()
setwd("/Users/petermattingly/Desktop/")

### creating a notebook chu
'control' + 'option', then

```{r}

```

running individual lines
mac: 'command' then 'return'
pc: 'control' then 'enter'

assignment operator <-

creating pipe operator %>%
'command' 'shift' 'm' =

libraries and packages

```{r}
install.packages('data.table', 'tidyverse')
library(data.table)
library(tidyverse)
```
```

11:30 creating a notebook chunk — R Markdown

Console | Terminal | R Markdown

```
+      xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+      main="Daily Interest Rates Since 2000", pch=16, col='blue')
Error in (function (formula, data = NULL, subset = NULL, na.action = na.fail,  :
  invalid type (list) for variable 'strptime(threemonth$value, "%Y-%m-%d")'
> plot(strptime(threemonth$value,"%Y-%m-%d"), strptime(tenyear$value,"%Y-%m-%d"),
+      xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+      main="Daily Inter  Since 2000", pch=16, col='blue')
Error in plot.window(  ite 'xlim' values
In addition: Warning  ages:
1: In min(x) : no no  ssing arguments  rn
2: In max(x) : no no  ssing argument  max;  rn  -Inf
3: In min(x) : no no  sing arg  min;  rn  Inf
4: In max(x) : no non  ts  max;  rn  -Inf
> plot(threemonth$value,  alue,
+      xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+      main="Daily Interest Rates Since 2000", pch=16, col='blue')
> cor(tenyear$value ~ threemonth$value)
Error in cor(tenyear$value ~ threemonth$value) :
  supply both 'x' and 'y' or a matrix-like 'x'
> cor(tenyear$value, threemonth$value)
[1] 0.7608
> threemonth = drop_na(fredr(series_id = "DGS3MO", observation_start = as.Date("2000-01-01")))
> tenyear = drop_na(fredr(series_id = "DGS10", observation_start = as.Date("2000-01-01")))
> plot(threemonth$value, tenyear$value,
+      xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+      main="Daily Interest Rates Since 2000", pch=16, col='blue')
```

Project: (None)

Environment | History | Connections

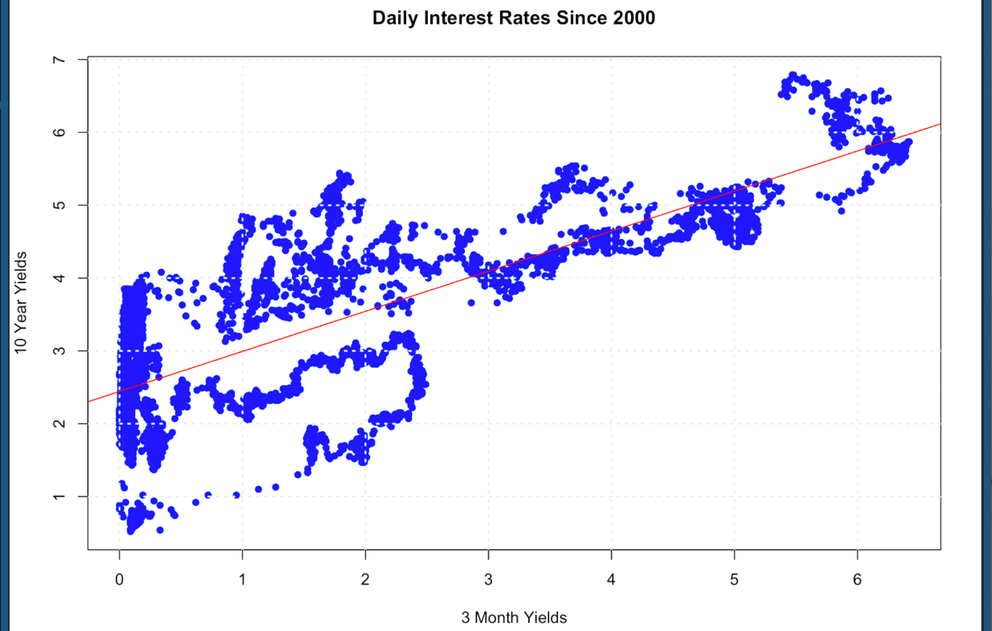
Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
dailyavg_table	tbl_df	7	2 KB	3 obs. of 7 variables
dailyavg_wtmeans	grouped_df	4	66.4 KB	1095 obs. of 4 variables
data1990	tbl_df	6	22 KB	373 obs. of 6 variables
data1990_2018_race_total	data.frame	5	8.7 KB	174 obs. of 5 variables
data1990_hisp	tbl_df	6	7 KB	62 obs. of 6 variables
data1990_main	tbl_df	6	19.1 KB	311 obs. of 6 variables
data1999_2000	grouped_df	5	4.4 KB	12 obs. of 5 variables
data1999_2000_total	data.frame	5	4.3 KB	66 obs. of 5 variables
data1999_2018_race_total	matrix	10	7.9 KB	List of 10
data1999_2018_total	data.frame	5	8.6 KB	174 obs. of 5 variables
f1	function	1	10.1 KB	function (x, y, p = 0)
geo_northern	data.table	9	30.6 KB	97 obs. of 9 variables
geospatial	data.table	9	73.7 KB	246 obs. of 9 variables
il	sf	6	1.4 MB	408 obs. of 6 variables
labTheme	function	1	18 KB	function (base_size = 48)
logo	rastergrob	12	1.8 MB	Large rastergrob (12 elements, 1.8 Mb)
model1	lm	12	1.3 MB	Large lm (12 elements, 1.3 Mb)
monthlyavg_countries	grouped_df	7	47 KB	730 obs. of 7 variables
name_region	data.table	5	38.5 KB	246 obs. of 5 variables
numbers	integer	10	96 B	int [1:10] 1 2 3 4 5 6 7 8 9 10
numlist	numeric	10	176 B	num [1:10] 1 2 3 4 5 6 7 8 9 10

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish

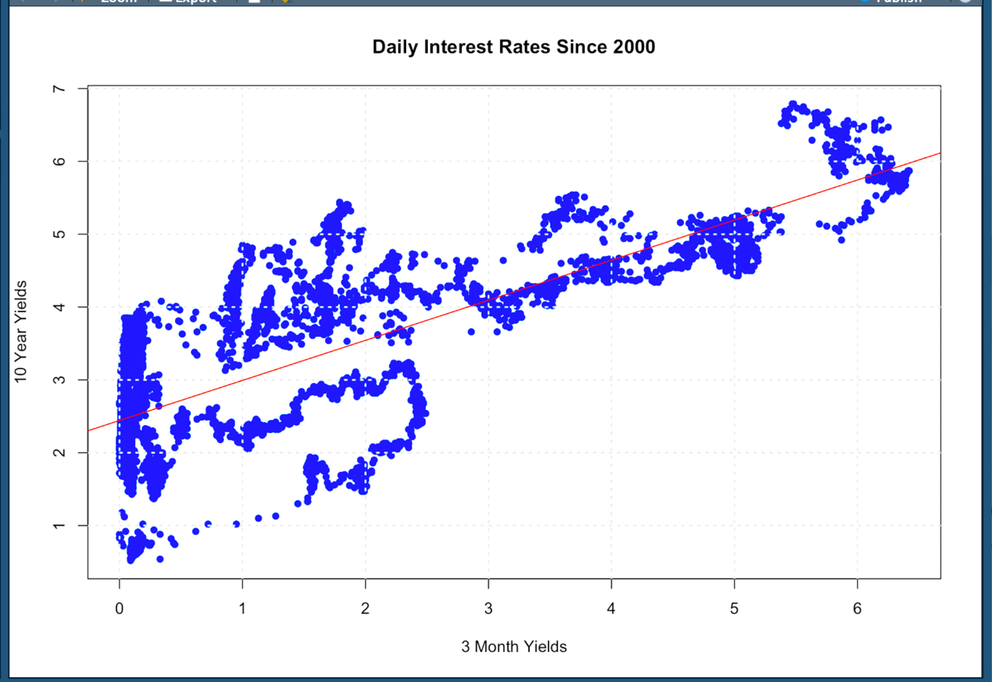


@tladeras

Script

Environ-ment

Console



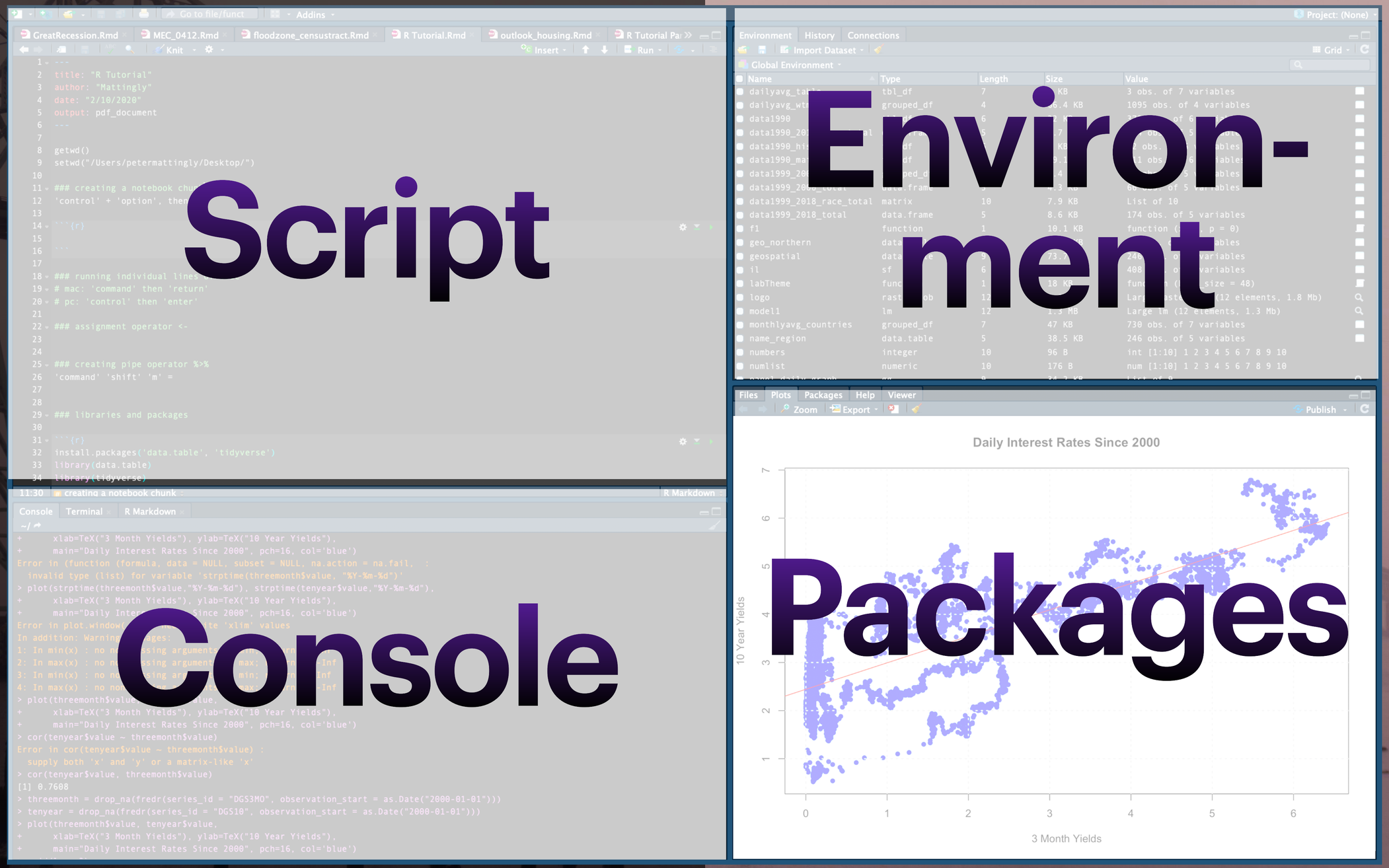
@tladeras

Script

Environ-ment

Console

Packages



@tladeras

Terms

R Studio :

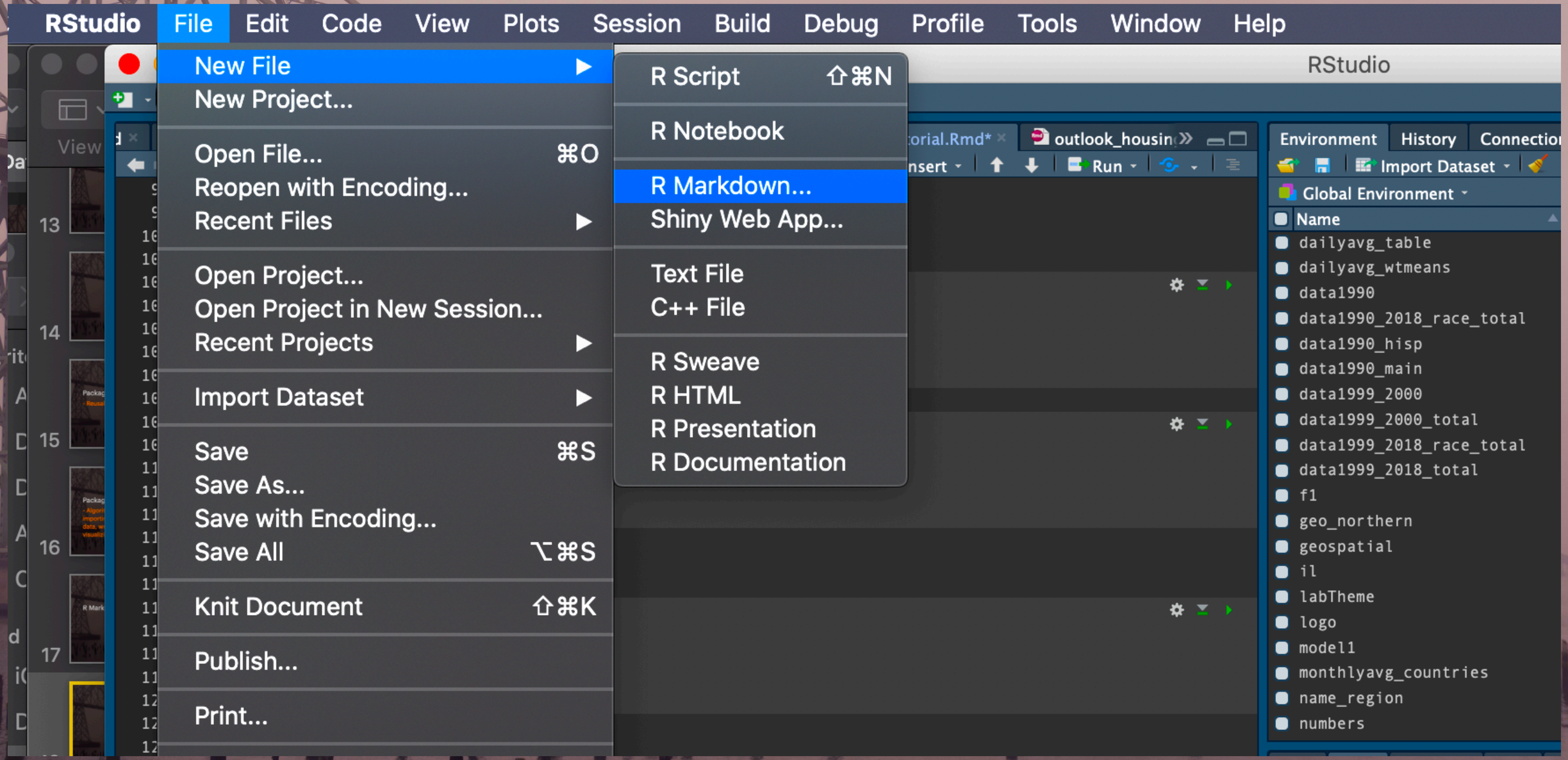
- R Markdown

@tladeras

Rmarkdown

TEXT. CODE. OUTPUT.

(GET IT TOGETHER, PEOPLE.)


RStudio
File
Edit
Code
View
Plots
Session
Build
Debug
Profile
Tools
Window
Help
New File
New Project...
Open File... ⌘O
Reopen with Encoding...
Recent Files
Open Project...
Open Project in New Session...
Recent Projects
Import Dataset
Save ⌘S
Save As...
Save with Encoding...
Save All ⌥⌘S
Knit Document ⇧⌘K
Publish...
Print...
R Script ⇧⌘N
R Notebook
R Markdown...
Shiny Web App...
Text File
C++ File
R Sweave
R HTML
R Presentation
R Documentation
Environment
History
Import Dataset
Global Environment
Name
dailyavg_table
dailyavg_wtmeans
data1990
data1990_2018_race_total
data1990_hisp
data1990_main
data1999_2000
data1999_2000_total
data1999_2018_race_total
data1999_2018_total
f1
geo_northern
geospatial
il
labTheme
logo
model1
monthlyavg_countries
name_region
numbers

@tladeras

Terms

R Markdown :

- Code chunks

@tladeras

```
### subsetting

##### subsetting by value

```{r}
### base r
setosa <- iris[iris$Species == "setosa",]
glimpse(setosa)
```

```{r}
### dplyr
setosa_tidy <- iris %>% filter(Species = "setosa")
glimpse(setosa_tidy)
```

subsetting by columns

```{r}
### base r
iris_length <- iris[, c(1,3,5,9)]
glimpse(iris_length)
```


```{r}
### dplyr
iris_length_dplyr <- iris %>% dplyr::select(matches("(Length|Species)"))
glimpse(iris_length_dplyr)
```

```

@tladeras

# Terms

R Studio :

- Working directory

@tladeras

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help

RStudio

Go to file/funct | Addins

GreatRecession.Rmd | MEC_0412.Rmd | floodzone_censu | ing.Rmd | R Tutorial

Knit | Run

```
1 ---
2 title: "R Tutorial"
3 author: "Mattingly"
4 date: "2/10/2020"
5 output: pdf_document
6 ---
7
8 getwd()
9 setwd("/Users/petermattingly/Desktop/")
10
11 ### creating a notebook chunk
12 'control' + 'option', then 'i'
13
14 ```{r}
15
16 ```
17
18 ### running individual lines of code
19 # mac: 'command' then 'return'
20 # pc: 'control' then 'enter'
21
22 ### assignment operator <-
23
24
25 ### creating pipe operator %>%
26 'command' 'shift' 'm' =
27
28
29 ### libraries and packages
30
31 ```{r}
32 install.packages('data.table', 'tidyverse')
```

Session

- New Session
- Interrupt R
- Terminate R...
- Restart R ⇧⌘F10
- Restart R and Clear Output
- Restart R and Run All Chunks
- Set Working Directory ▶
  - To Source File Location
  - To Files Pane Location
  - Choose Directory... ^⇧H
- Load Workspace...
- Save Workspace As...
- Clear Workspace...
- Quit Session...

Environment | History | Conn

Import Dataset

Global Environment

Name

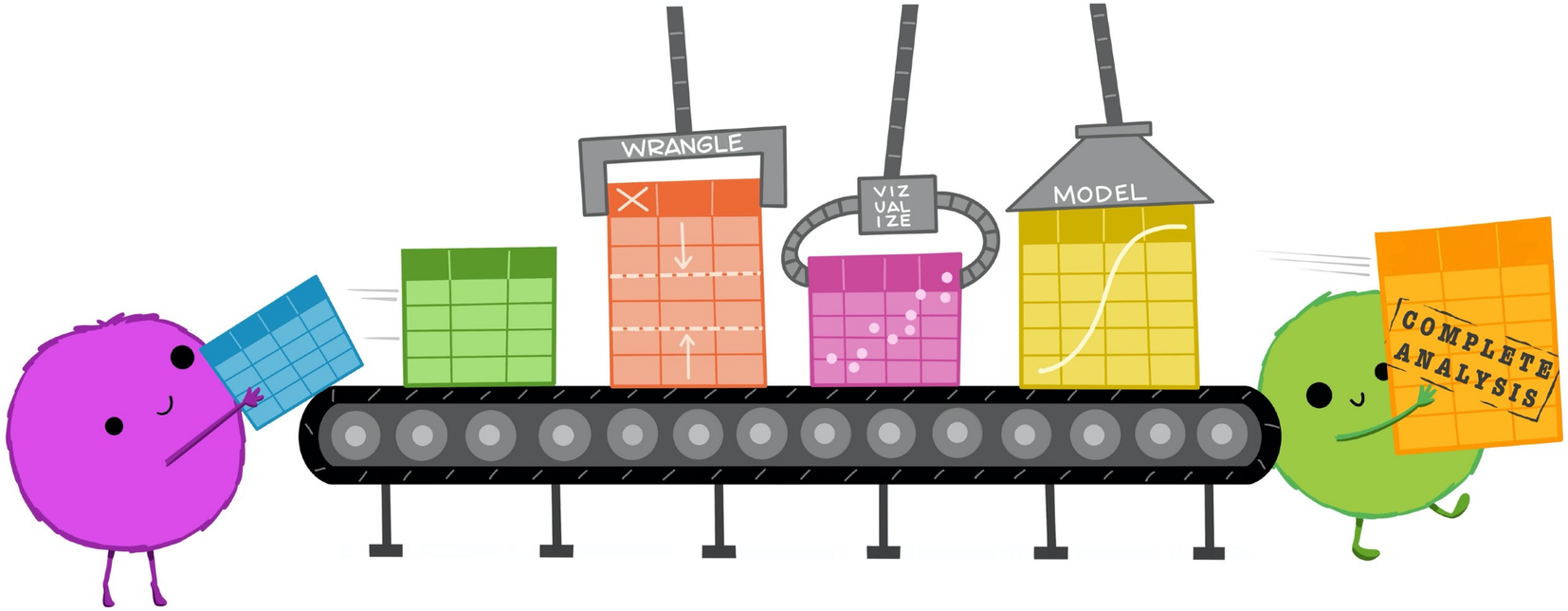
- dailyavg_table
- dailyavg_wtmeans
- data1990
- 1990_2018_race_total
- 1990_hisp
- 1990_main
- 1999_2000
- 1999_2000_total
- data1999_2018_race_total
- data1999_2018_total
- f1
- geo_northern
- geospatial
- il
- labTheme
- logo
- model1
- monthlyavg_countries
- name_region
- numbers

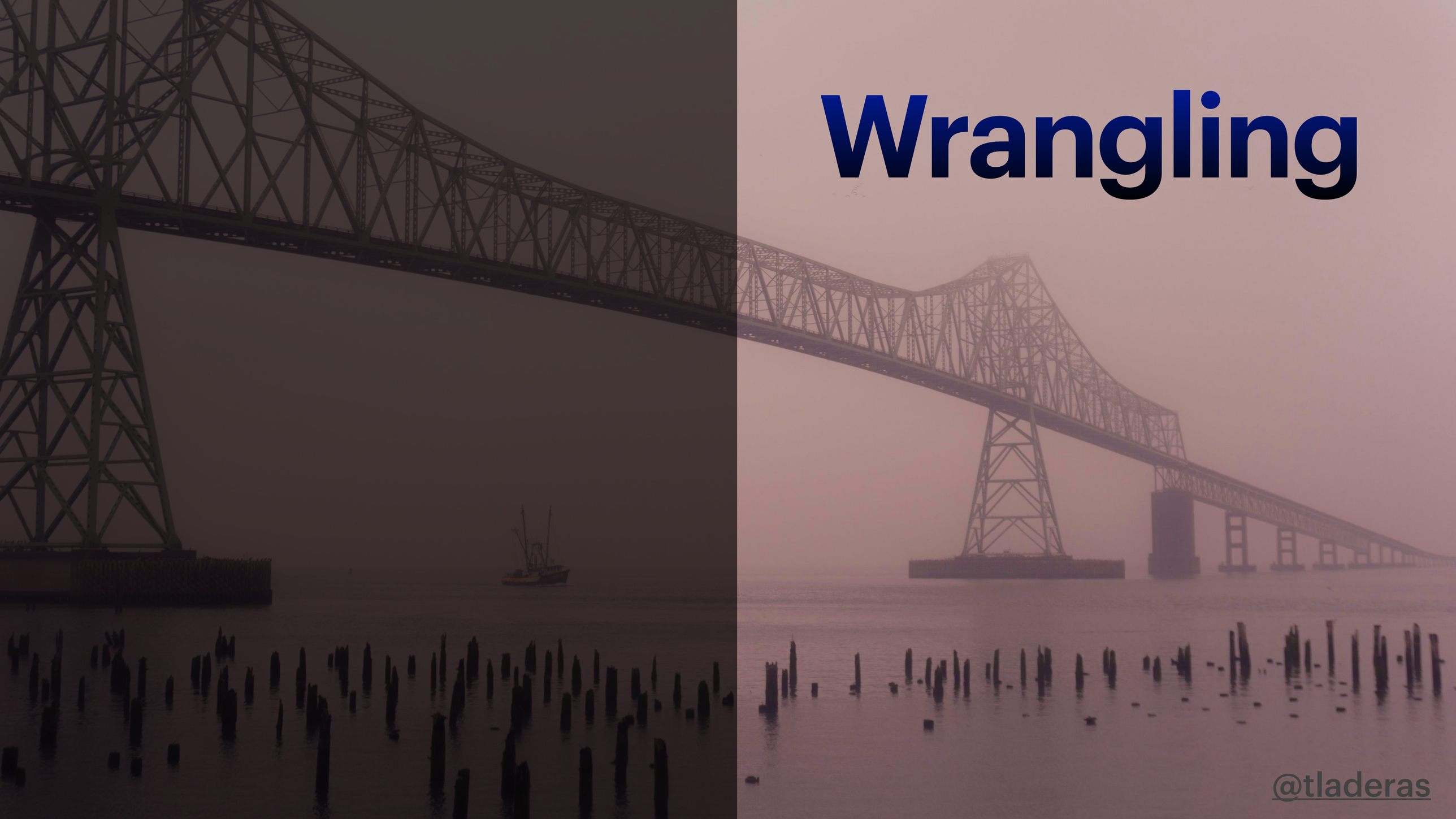
Files | Plots | Packages | Help

Zoom | Export

@tladeras

@tladeras

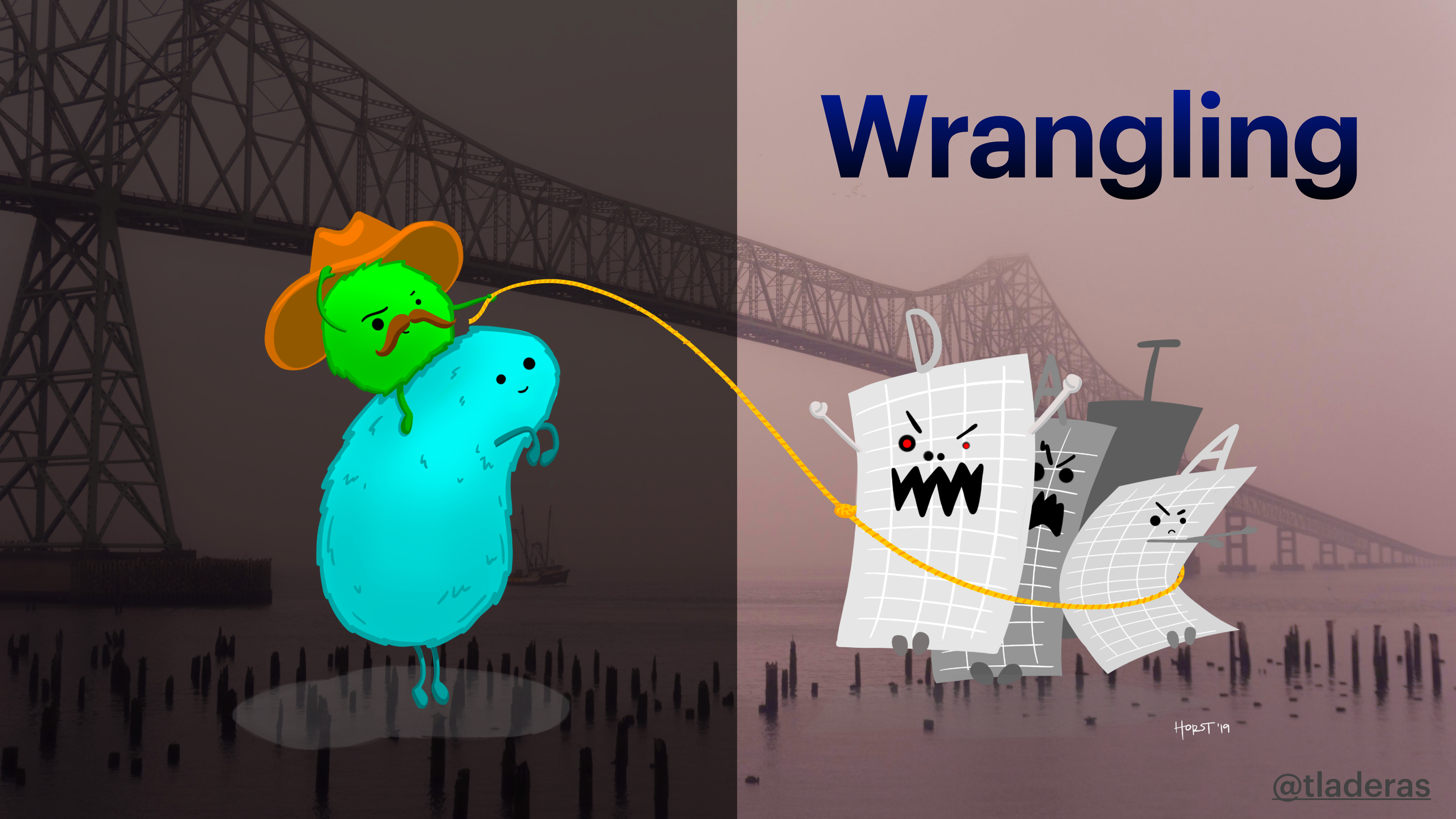

WRANGLE
VIZ
UAL
IZE
MODEL
COMPLETE
ANALYSIS



Wrangling

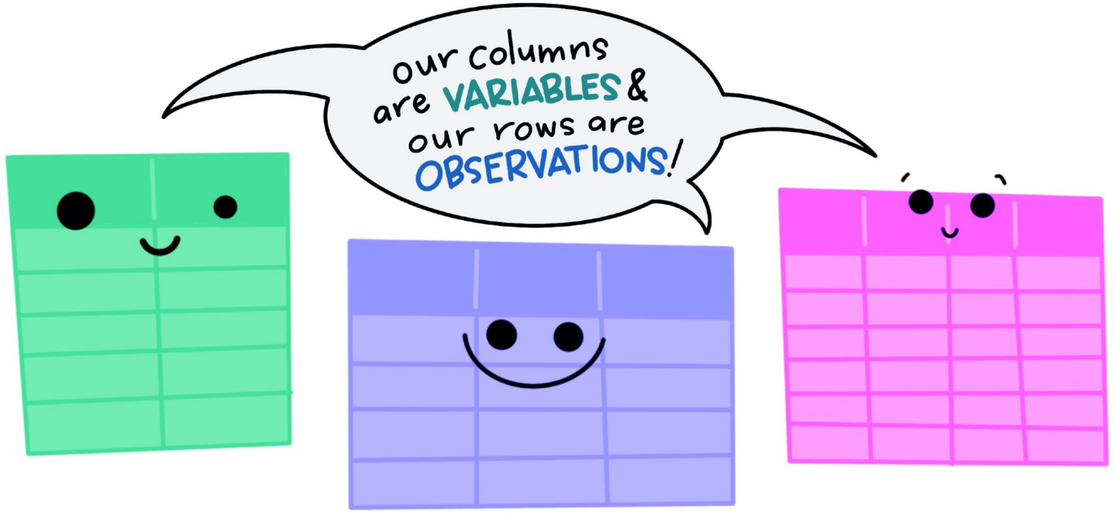
@tladeras

Wrangling



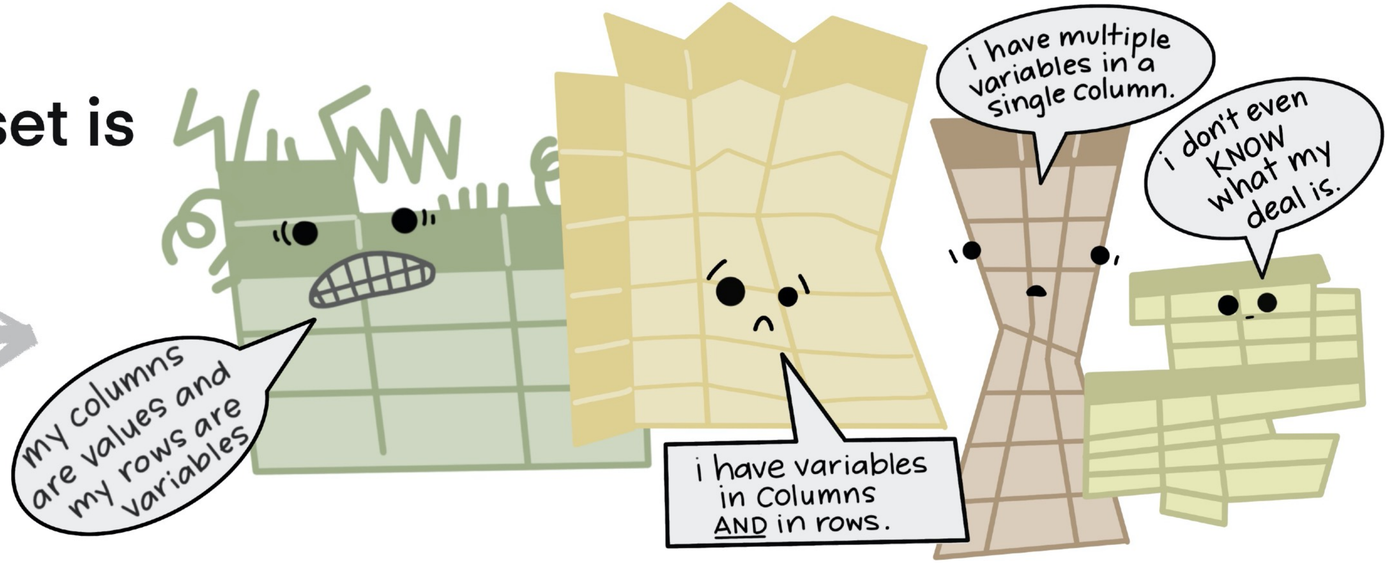
@tladeras

The standard structure of tidy data means that "tidy datasets are all alike..."



"...but every messy dataset is messy in its own way."

—HADLEY WICKHAM



Terms

Data wrangling

@tladeras

Terms

Data wrangling :

- Reshaping by lengthening or widening data

@tladeras

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

—HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

each column a variable

id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

each row an observation

Wickham, H. (2014). Tidy Data. Journal of Statistical Software *59* (10). DOI: 10.18637/jss.v059.i10

Data

@tladeras

Data

VARIABLE 1	VARIABLE 2	VARIABLE 3	VARIABLE 4	VARIABLE 5

@tladeras

Columns

VARIABLE 1	VARIABLE 2	VARIABLE 3	VARIABLE 4	VARIABLE 5

@tladeras

Rows

VARIABLE 1	VARIABLE 2	VARIABLE 3	VARIABLE 4	VARIABLE 5

@tladeras

Terms

Data wrangling :

- “Gathering” or lengthening with more rows/ observations

@tladeras

Original

Var 1	Var 2	Date 1	Date 2	Date 3

@tladeras

Original

Var 1	Var 2	Date 1	Date 2	Date 3

Gathering

Var 1	Var 2	Date	Value
		1	
		2	
		3	

@tladeras

Terms

Data wrangling :

- “Spreading” or widening with more columns/variables

@tladeras

Original

Var 1	Var 2	Date	Value
		1	
		2	
		3	

@tladeras

Original

Var 1	Var 2	Date	Value
		1	
		2	
		3	

Spreading

Var 1	Var 2	Date 1	Date 2	Date 3

@tladeras

Terms

Data wrangling :

- Variable creation or “mutation”
- Descriptive statistics
- Formulas

@tladeras

dplyr::mutate
add column(s),
keep existing.
A
B
C
1
5
10
0
2
-3
4
9
6
3
2
14
9
8
THE PLAN:
A
B
C
+
=
HORST '18

Terms

Data wrangling :

- Working with variables like dates

- Dates in R:

“YYYY-MM-DD”

@tladeras

@tladeras

Terms

Data wrangling :

- Subsetting or filtering

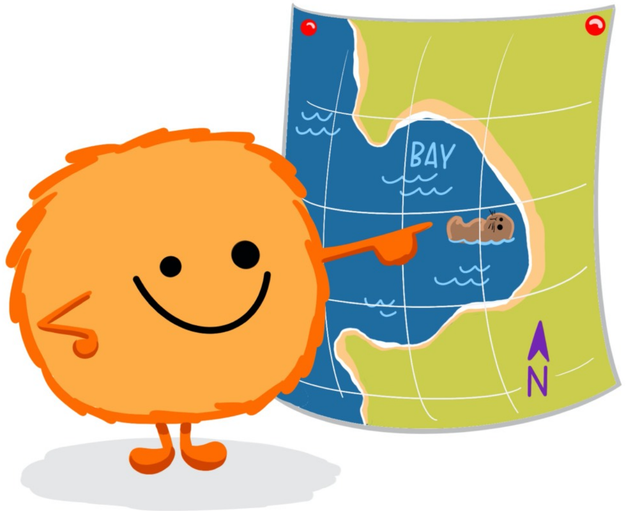
@tladeras

dplyr::filter()

KEEP ROWS THAT s·a·t·i·s·f·y **your CONDITIONS**

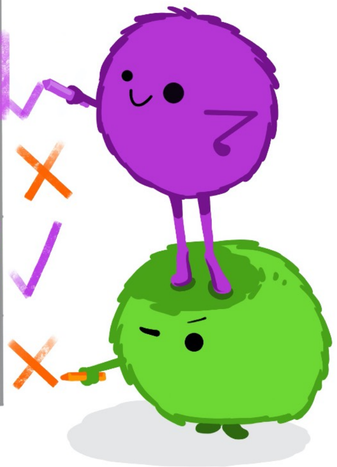
keep rows from… | this data… | ONLY IF… | type is "otter" | AND | site is "bay"

```
filter(df, type == “otter” & site == “bay”)
```



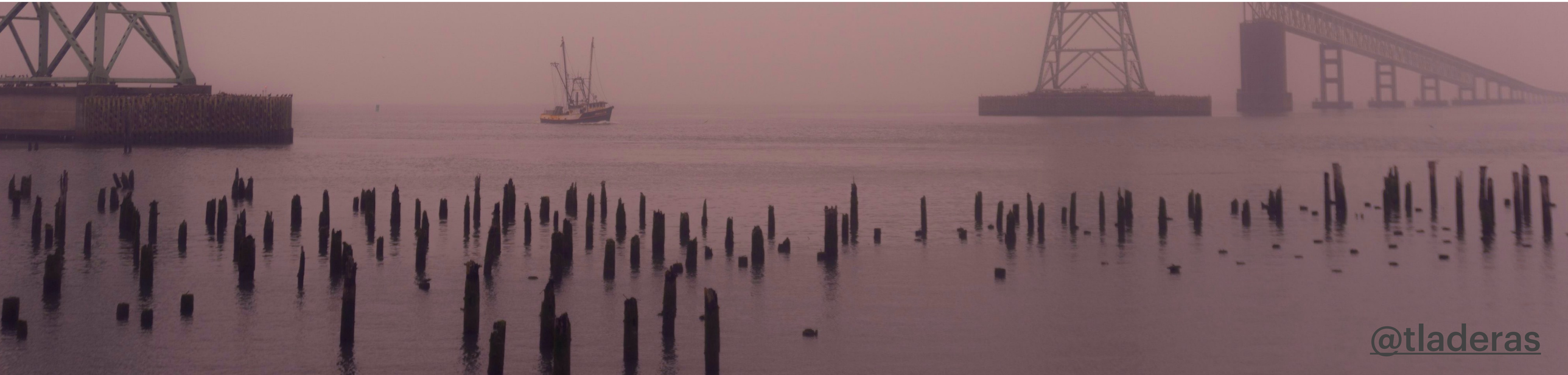
type	food	site
otter	urchin	bay
shark	seal	channel
otter	abalone	bay
otter	crab	wharf

@allison_horst



@tladeras

RStudio®



@tladeras

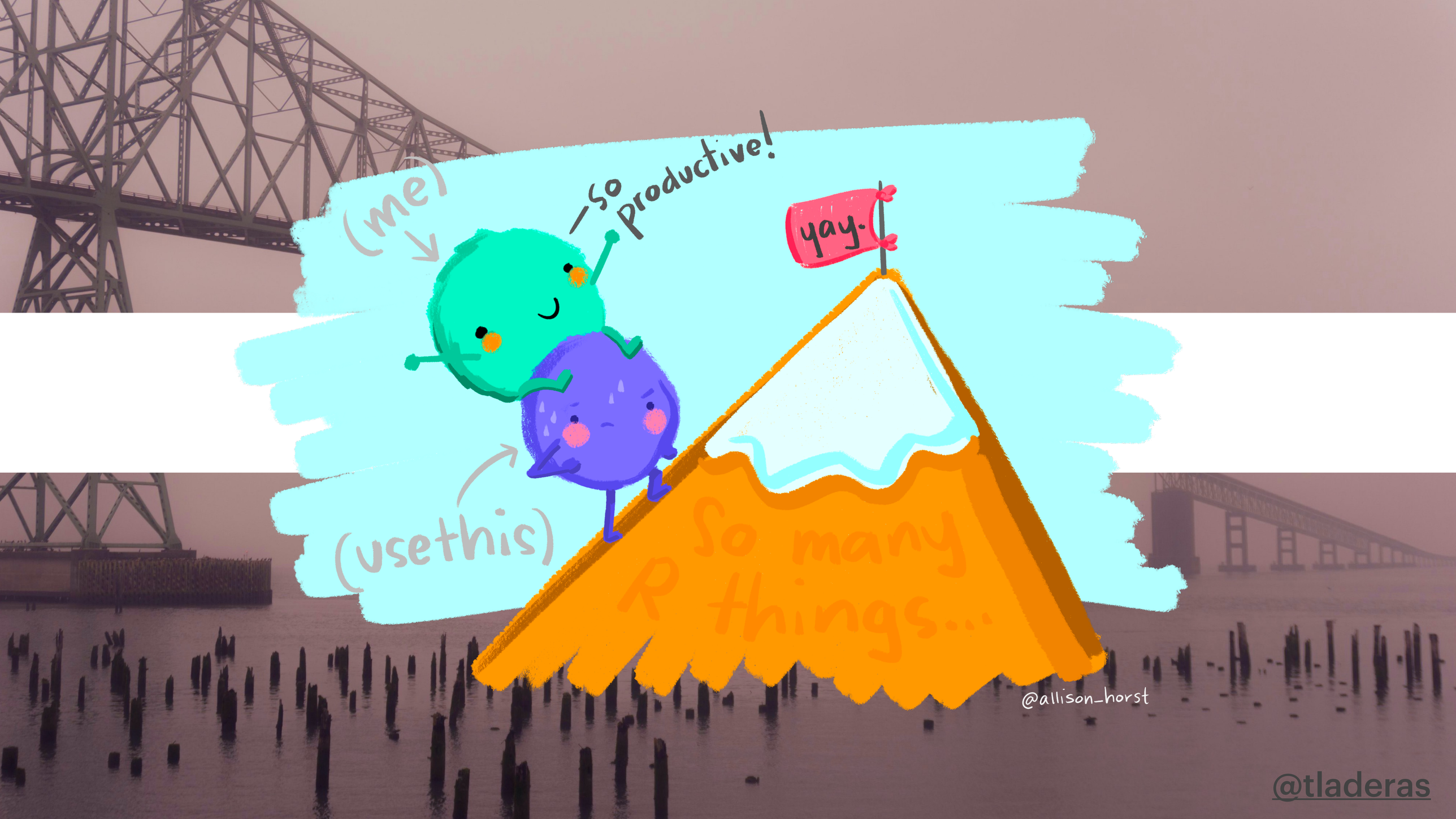
Download RStudio - RStudio

rstudio.com/products/rstudio/download/

Apps

	RStudio Desktop	RStudio Desktop Pro	RStudio Server	RStudio Server Pro
	Open Source License	Commercial License	Open Source License	Commercial License
	Free	$995 /year	Free	$4,975 /year (5 Named Users)
	DOWNLOAD	BUY	DOWNLOAD	BUY
	Learn more	Learn more	Learn more	Evaluation | Learn more
Integrated Tools for R	✓	✓	✓	✓
Priority Support		✓		✓
Access via Web Browser			✓	✓
RStudio Professional Drivers		✓		✓
Connect to RStudio Server Pro remotely		✓		

@tladeras



@tladeras