

Data Analytics

Lecture Series: Part 2

Data Wrangling

Peter J. Mattingly

pjm407@nyu.edu

Overview



Overview

In this section, we will:



Overview

In this section, we will:

- Learn variable types



Overview

In this section, we will:

- Learn variable types
- Source real estate and population data

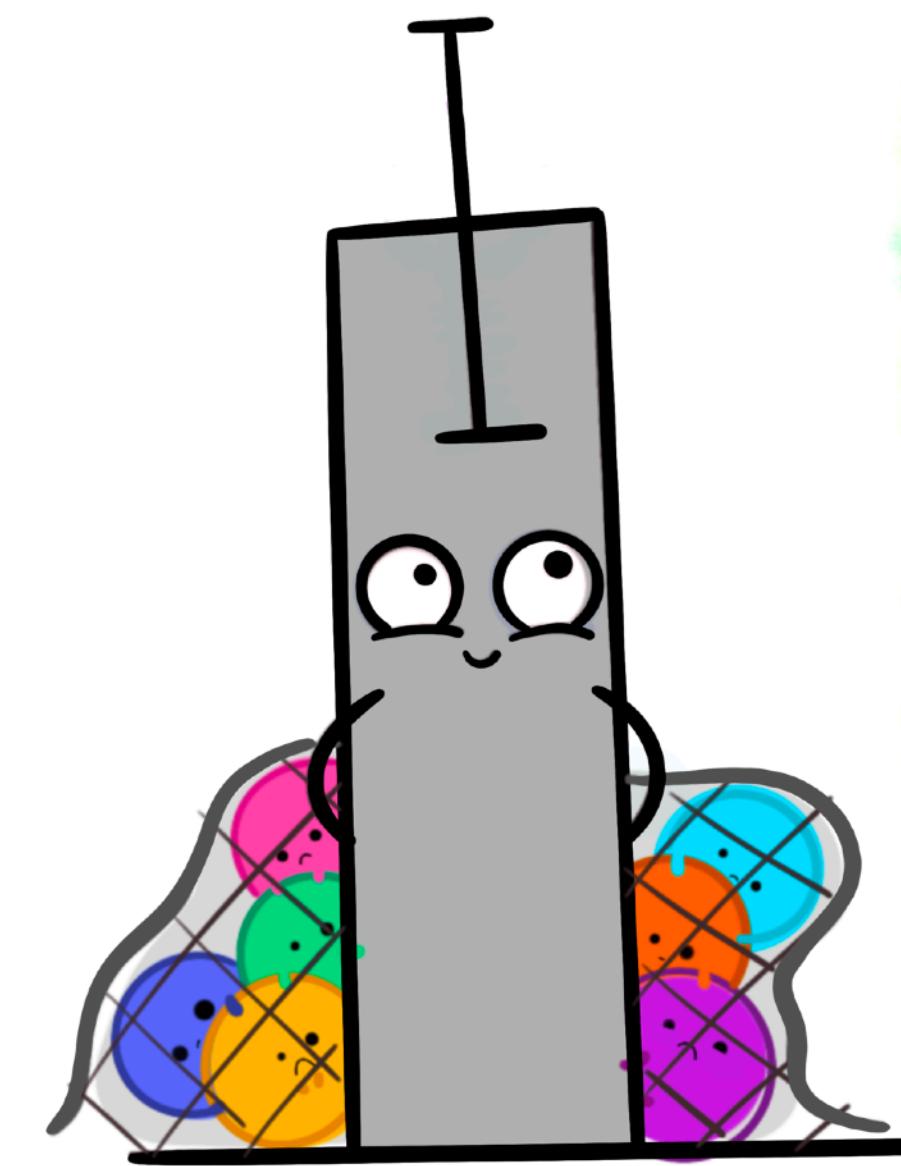


Overview

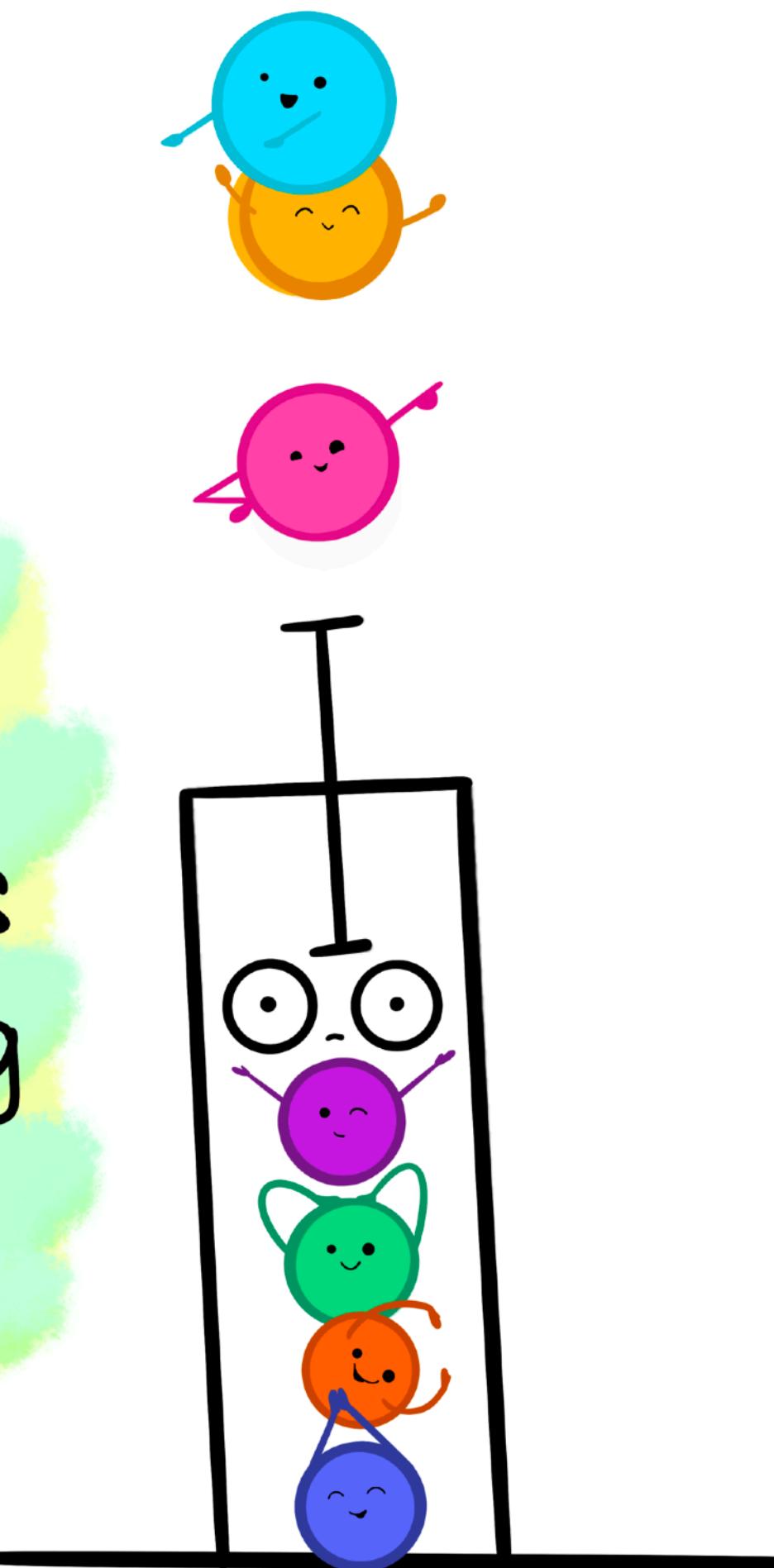
In this section, we will:

- Learn variable types
- Source real estate and population data
- Clean and prepare the data for analysis



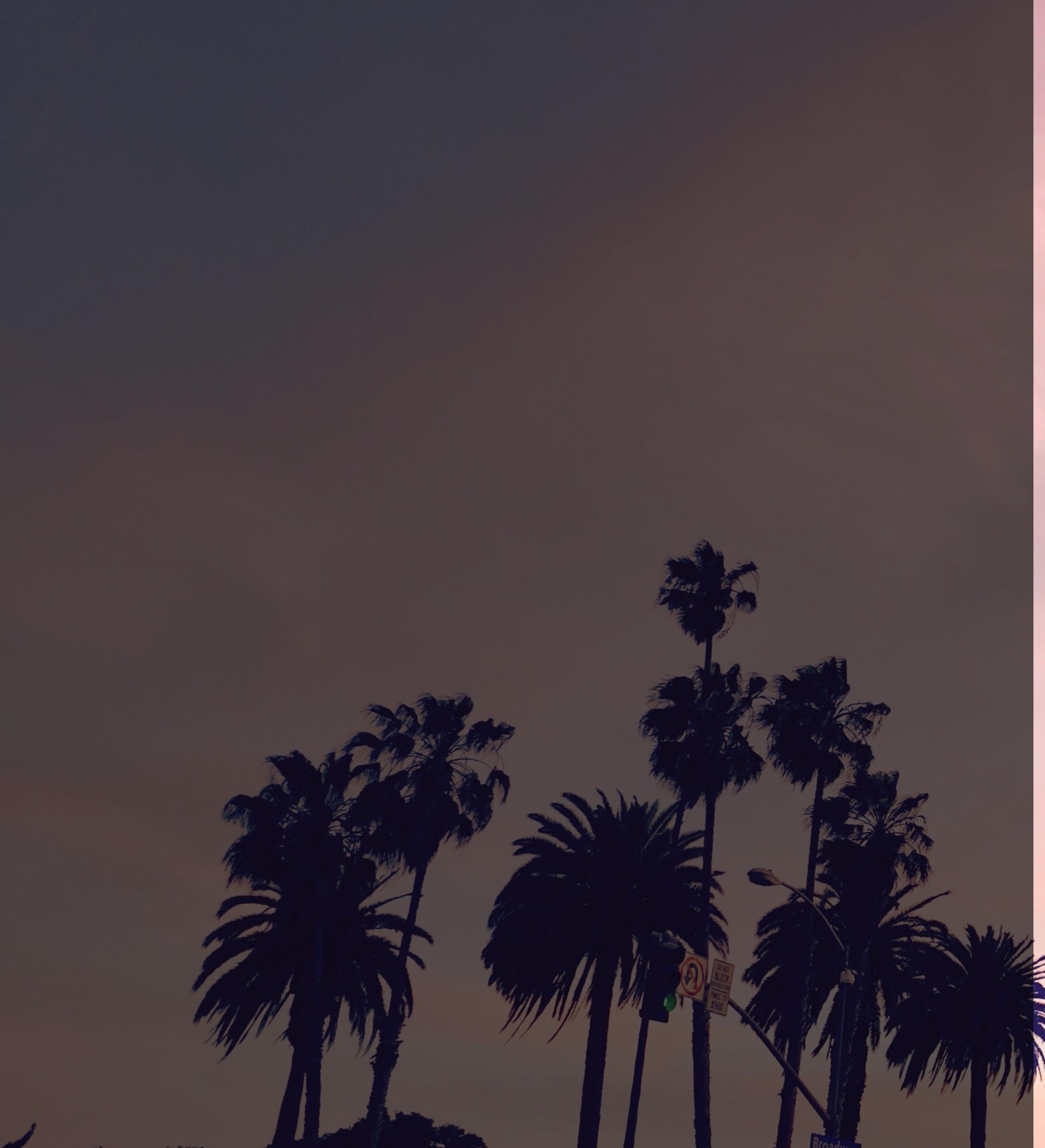


are your
summary statistics
hiding something
interesting?



@allison_horst

Variables



Variables :

- Categorical

Variables



NOMINAL

UNORDERED DESCRIPTIONS



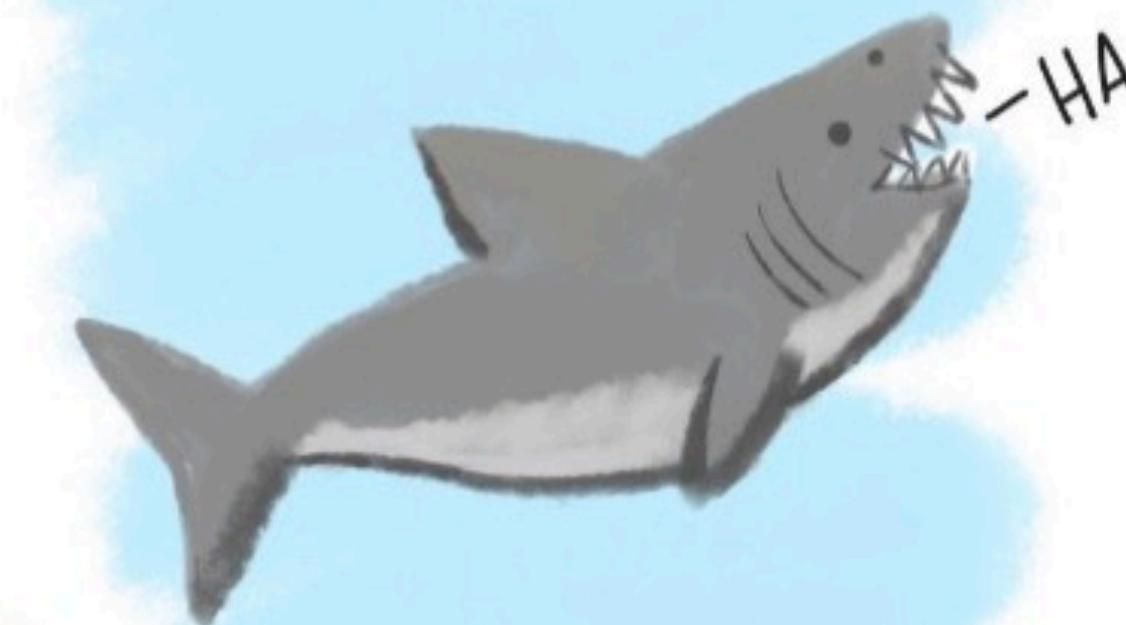
ORDINAL

ORDERED DESCRIPTIONS



BINARY

ONLY 2 MUTUALLY
exclusive OUTCOMES



@allison_horst

Variables :

- Categorical
- Numeric

Variables



CONTINUOUS

measured data, can have ∞ values within possible range.



I AM 3.1" TALL

I WEIGH 34.16 grams

DISCRETE

OBSERVATIONS can only exist at LIMITED VALUES, OFTEN COUNTS.



I HAVE 8 ARMS
and
4 SPOTS!

@allison_horst

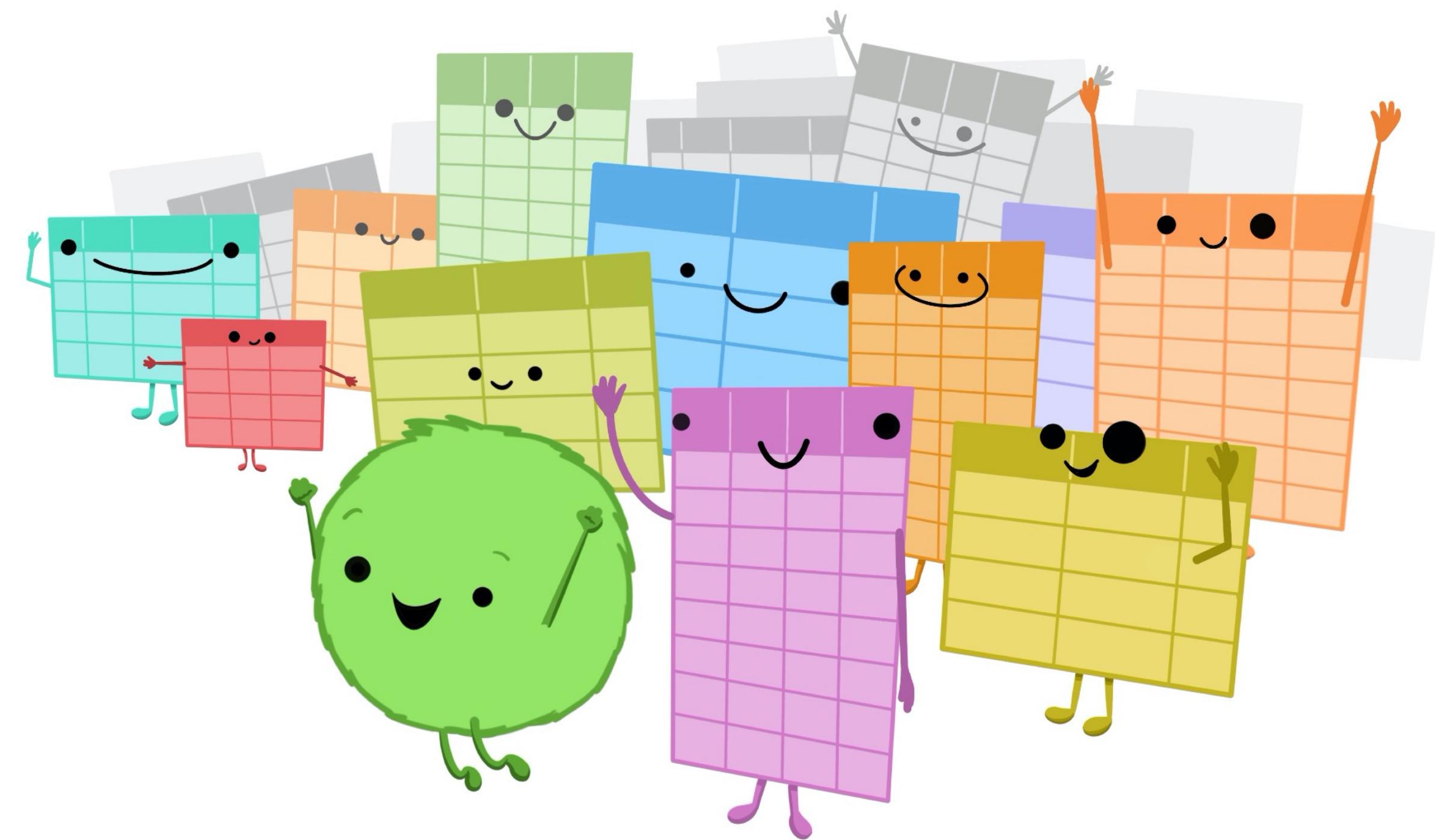
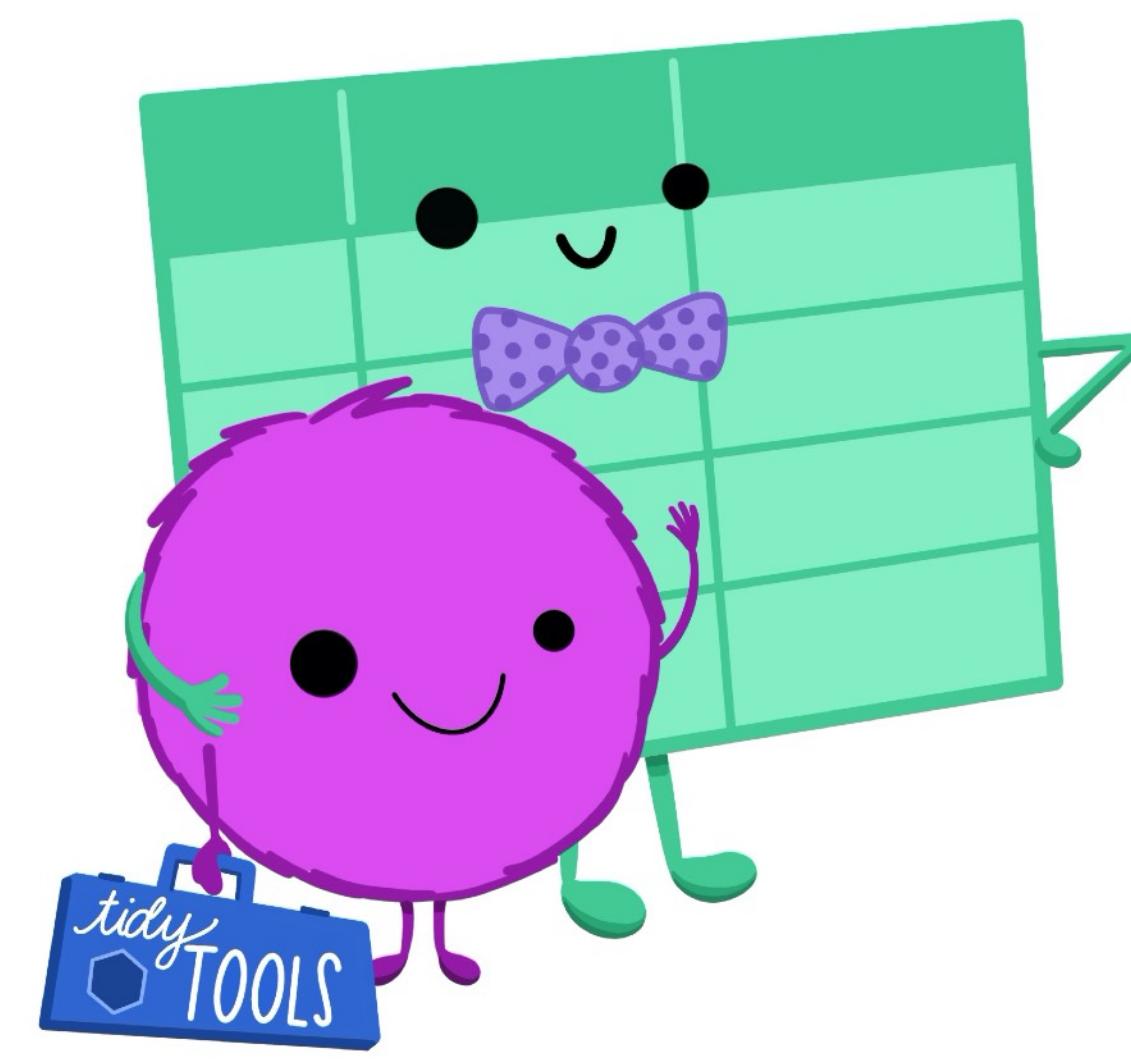


Variables :

- Categorical
- Numeric
- Dates

Variables







Script

```

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
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')
33 library(data.table)
34 library(tidyverse)

```

11:30 # creating a notebook chunk

```

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$value)
[1] 0.7608
> threemonth = drop_na(fredr(series_id = "DGS3M0", observation_start = as.Date("2000-01-01")))
> tenyear = drop_na(fredr(series_id = "DGS10", observation_start = as.Date("2000-01-01")))
> plot(threemonth$value, tenyear$value,
+   xlab=TeX("3 Month Yields"), ylab=TeX("10 Year Yields"),
+   main="Daily Interest Rates Since 2000", pch=16, col='blue')

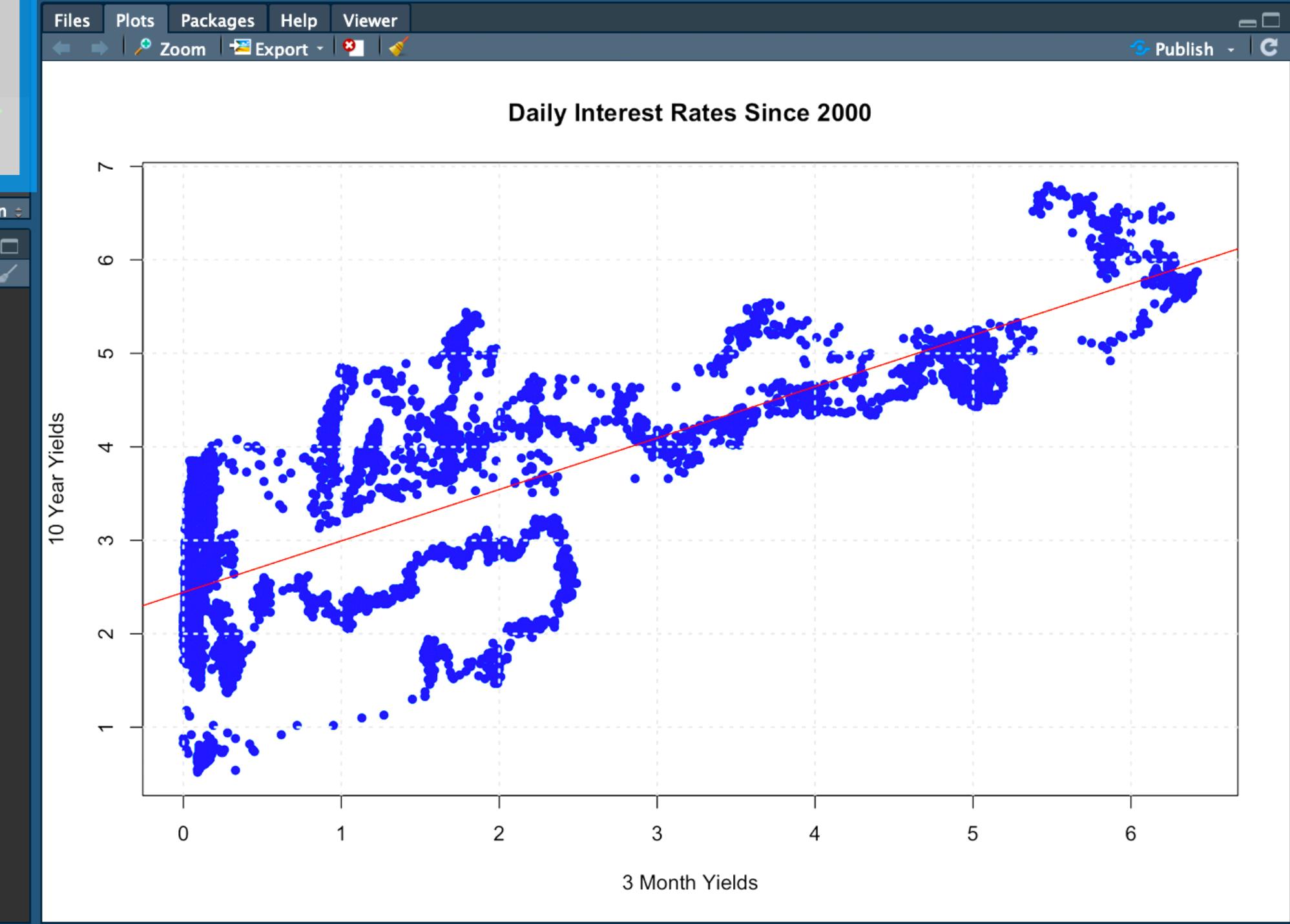
```

Environment History Connections

Import Dataset Grid C

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
open_daily_graph	gg	9	24.7 KB	List of 9



RStudio

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

RStudio

Go to file/function Addins

R Tutorial.Rmd x R Tutorial Part 2.Rmd x REDA.Rmd x GreatRecession.Rmd x

ABC Knit Insert Run

6

7

8 `### creating a notebook chunk`

9 `# on a mac: 'control' + 'option', then 'i'`

10 `# on a pc: 'control' + 'alt', then 'i'`

11

12 ````{r}`

13 `install.packages(c("tidyverse", "devtools", "tidycensus"))`

14 `````

Error in install.packages : Updating loaded packages

15

16 ````{r}`

17 `library(tidyverse)`

18 `library(devtools)`

19 `library(tidycensus)`

20 `````

21

22 ````{r}`

23 `devtools::install_github("sboysel/fredr")`

24 `library(fredr)`

25 `````

26

27

25:1 C Chunk 34 R Markdown

Console Terminal R Markdown

~/

> `library(tidyverse)`

> `library(devtools)`

> `library(tidycensus)`

> `devtools::install_github("sboysel/fredr")`

Skipping install of 'fredr' from a github remote, the SHA1 (97b244ed) has not changed since last install.

Use `force = TRUE` to force installation

> `library(fredr)`

>

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

RStudio

+ | Go to file/function | Addins

R Tutorial.Rmd R Tutorial Part 2.Rmd REDA.Rmd GreatRecession.Rmd

ABC Knit Insert Run

```
18 library(devtools)
19 library(tidyCensus)
20 ``
21 ``
22 ``{r}
23 devtools::install_github("sboysel/fredR")
24 library(fredR)
25 ``
26 
27 **FRED API KEY**
28 
29 ``{r}
30 fredr_set_key('YOUR API KEY HERE')
31 ``
32 
33 **CENSUS API KEY**
34 
35 ``{r}
36 census_api_key('YOUR API KEY HERE')
37 ````
```

1:1 # REDA R Markdown



Categories > Production & Business Activity > Housing

☆ Housing Inventory: Active Listing Count in Boston-Cambridge-Newton, MA-NH (CBSA) (ACTLISCOU14460)

Observation:

Sep 2020: **7,771** (+ more)

Updated: Oct 1, 2020

Units:

Level,
Not Seasonally Adjusted

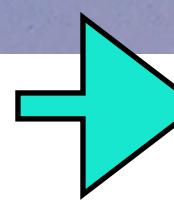
Frequency:

Monthly

1Y | 5Y | 10Y | Max

2016-07-01 to 2020-09-01





fred.stlouisfed.org/series/ACTLISCOU14460



ECONOMIC RESEARCH
FEDERAL RESERVE BANK OF ST. LOUIS

Search FRED

FRED® Economic Data Information Services Publications Working Papers Economists About

Categories > Production & Business Activity > Housing

★ Housing Inventory: Active Listing Count in Boston-Cambridge-Newton, MA-NH (CBSA)

(ACTLISCOU14460)

Observation:

Sep 2020: **7,771** (+ more)

Updated: Oct 1, 2020

Units:

Level,
Not Seasonally Adjusted

Frequency:

Monthly

1Y | 5Y | 10Y | Max

2016-07-01

to 2020-09-01



Housing Inventory: Active Listing Count in Boston-Cambridge-Newton, MA-NH (CBSA)



“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

each row an observation

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

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

REDA.Rmd x pulse_household_survey.Rmd x outlook_housing.Rmd x R Tutorial.Rmd x R Tutorial Part 2. >

47 ****HOUSING SUPPLY****
48
49 ````{r}`
50 `# Active listing count in Boston-Cambridge-Newton, MA-NH`
51 `# Earliest start date: July 2016`
52 `boston_supply <- fredr('ACTLISCOU14460',`
53 `observation_start = as.Date("2016-07-01"),`
54 `frequency = "m")`
55 `head(boston_supply)`
56 `````

<code>date</code>	<code>series_id</code>	<code>value</code>
<code><date></code>	<code><chr></code>	<code><dbl></code>
2016-07-01	ACTLISCOU14460	12198
2016-08-01	ACTLISCOU14460	11895
2016-09-01	ACTLISCOU14460	12274
2016-10-01	ACTLISCOU14460	12266
2016-11-01	ACTLISCOU14460	10879
2016-12-01	ACTLISCOU14460	8573

6 rows

REDA.Rmd x pulse_household_survey.Rmd x outlook_housing.Rmd x R Tutorial.Rmd x R Tutorial Part 2. >

47 ****HOUSING SUPPLY****
48
49 ````{r}`
50 `# Active listing count in Boston-Cambridge-Newton, MA-NH`
51 `# Earliest start date: July 2016`
52 `boston_supply <- fredr('ACTLISCOU14460',`
53 `observation_start = as.Date("2016-07-01"),`
54 `frequency = "m")`
55 `head(boston_supply)`
56 `````

<code>date</code>	<code>series_id</code>	<code>value</code>
<code><date></code>	<code><chr></code>	<code><dbl></code>
2016-07-01	ACTLISCOU14460	12198
2016-08-01	ACTLISCOU14460	11895
2016-09-01	ACTLISCOU14460	12274
2016-10-01	ACTLISCOU14460	12266
2016-11-01	ACTLISCOU14460	10879
2016-12-01	ACTLISCOU14460	8573

6 rows

REDA.Rmd x pulse_household_survey.Rmd x outlook_housing.Rmd x R Tutorial.Rmd x R Tutorial Part 2. >

47 ****HOUSING SUPPLY****
48
49 ````{r}`
50 `# Active listing count in Boston-Cambridge-Newton, MA-NH`
51 `# Earliest start date: July 2016`
52 `boston_supply <- fredr('ACTLISCOU14460',`
53 `observation_start = as.Date("2016-07-01"),`
54 `frequency = "m")`
55 `head(boston_supply)`
56 `````

date series_id value
<date> <chr> <dbl>
2016-07-01 ACTLISCOU14460 12198
2016-08-01 ACTLISCOU14460 11895
2016-09-01 ACTLISCOU14460 12274
2016-10-01 ACTLISCOU14460 12266
2016-11-01 ACTLISCOU14460 10879
2016-12-01 ACTLISCOU14460 8573

6 rows

Variables :

- Check the format

Variables



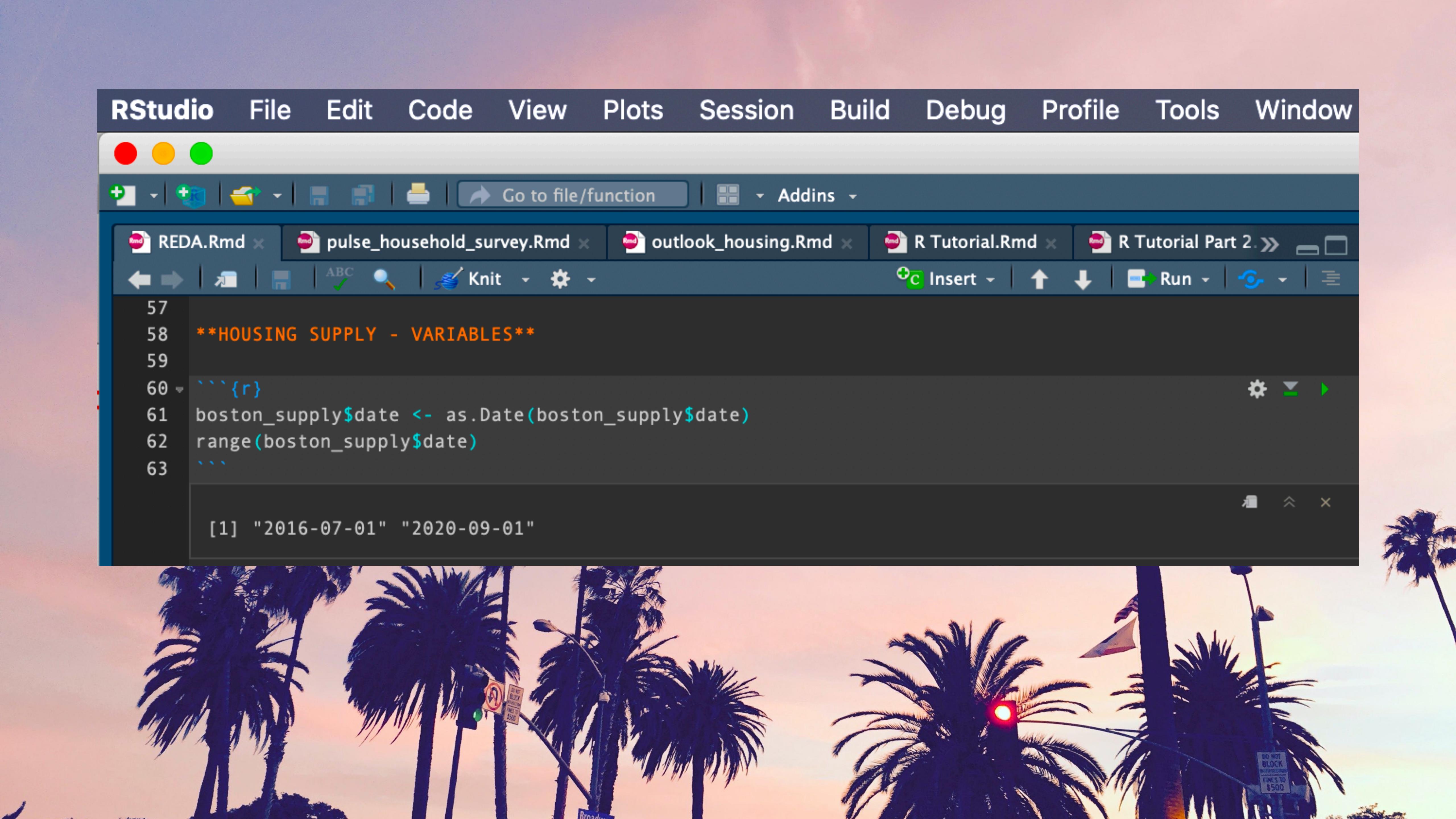
RStudio File Edit Code View Plots Session Build Debug Profile Tools Window

REDA.Rmd x pulse_household_survey.Rmd x outlook_housing.Rmd x R Tutorial.Rmd x R Tutorial Part 2. >

Go to file/function Addins

57
58 **HOUSING SUPPLY - VARIABLES**
59
60 ` `` {r}
61 boston_supply\$date <- as.Date(boston_supply\$date)
62 range(boston_supply\$date)
63 ` ``

[1] "2016-07-01" "2020-09-01"



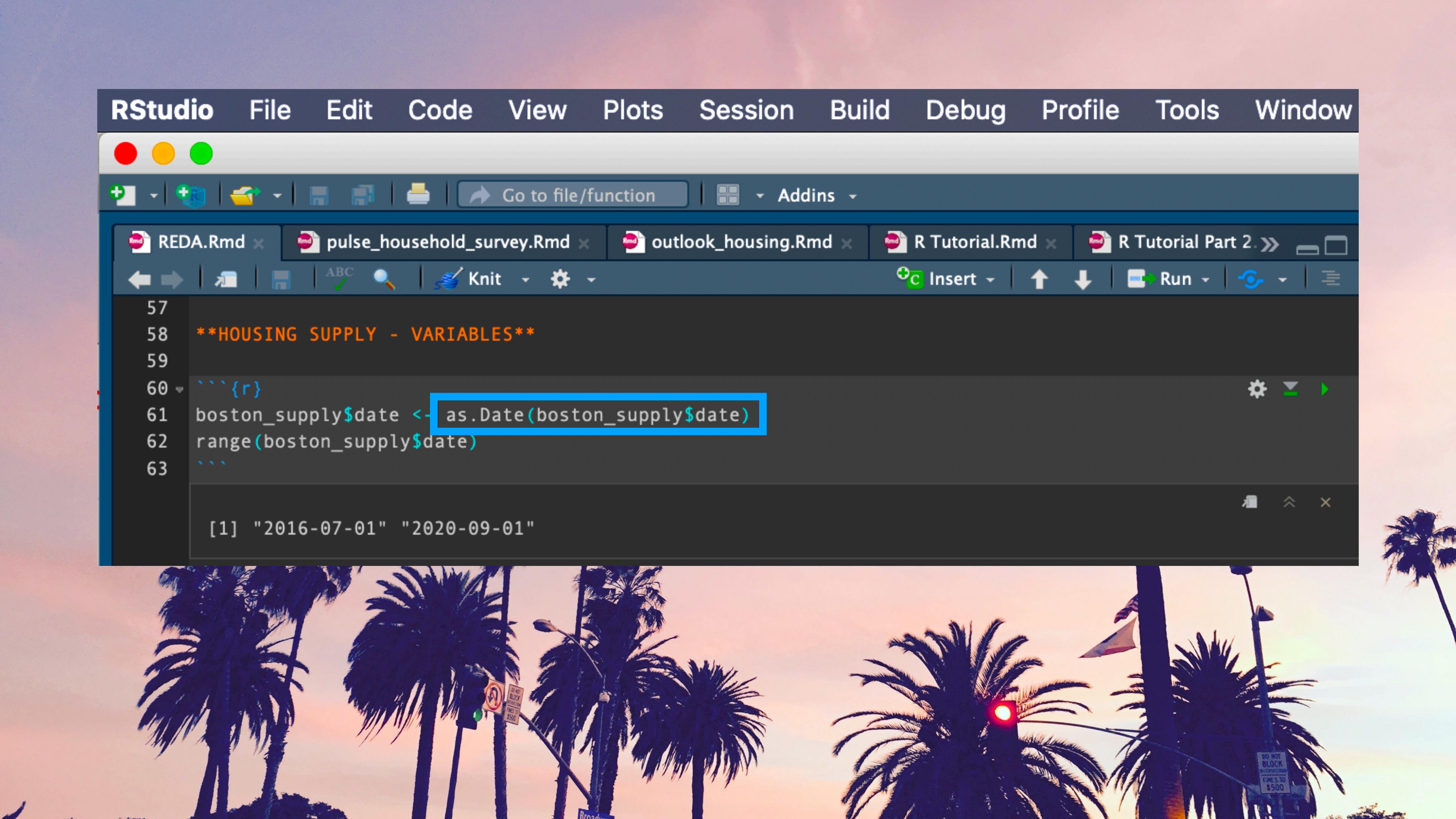
RStudio File Edit Code View Plots Session Build Debug Profile Tools Window

REDA.Rmd x pulse_household_survey.Rmd x outlook_housing.Rmd x R Tutorial.Rmd x R Tutorial Part 2. >

Go to file/function Addins

57
58 **HOUSING SUPPLY - VARIABLES**
59
60 ` `` {r}
61 boston_supply\$date <- as.Date(boston_supply\$date)
62 range(boston_supply\$date)
63 ` ``

[1] "2016-07-01" "2020-09-01"



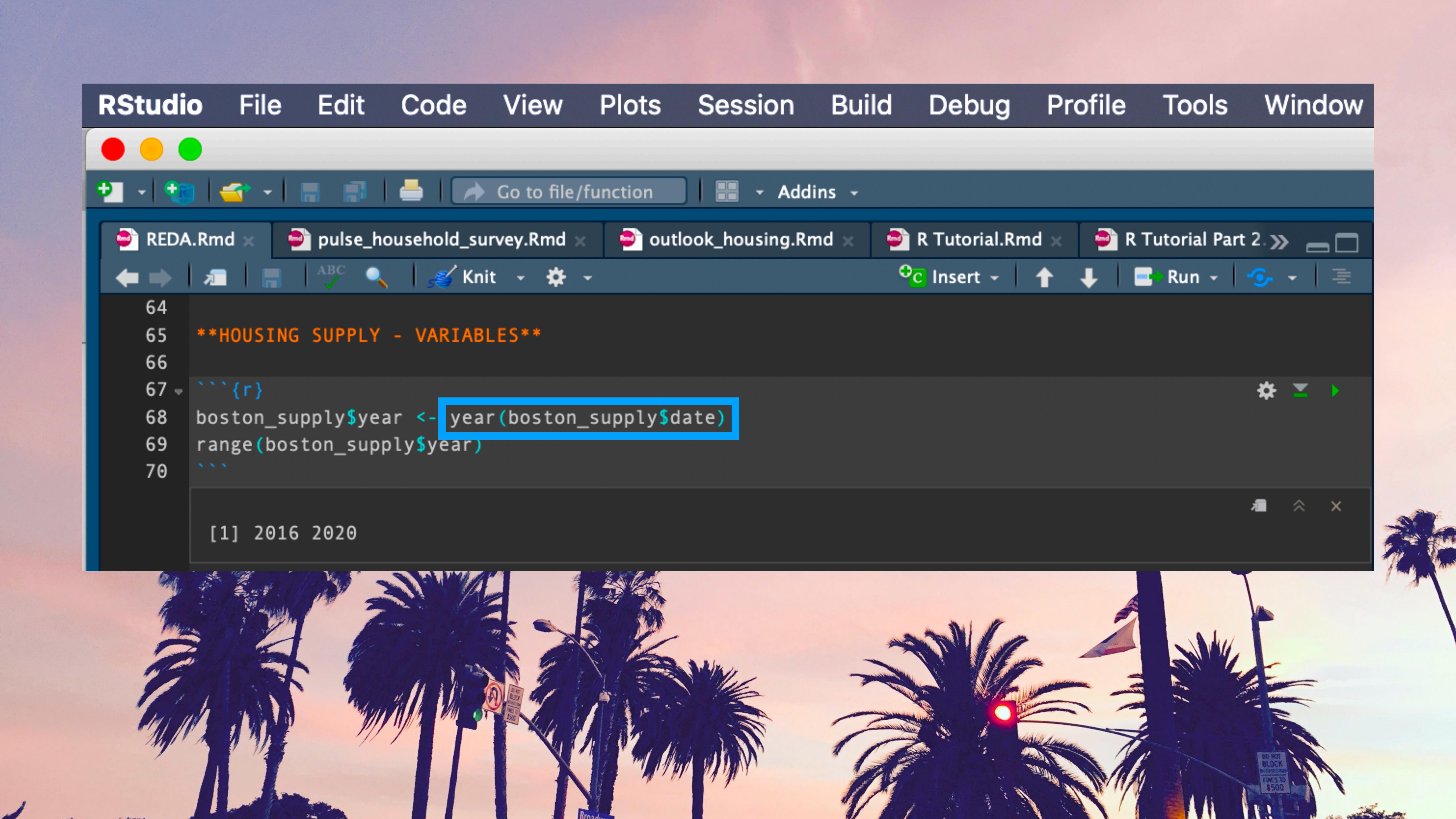
Variables :

- Check the format
- Create alternate version

Variables



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



A screenshot of the RStudio interface showing an R Markdown file named "pulse_household_survey.Rmd" open. The code editor displays the following R code:

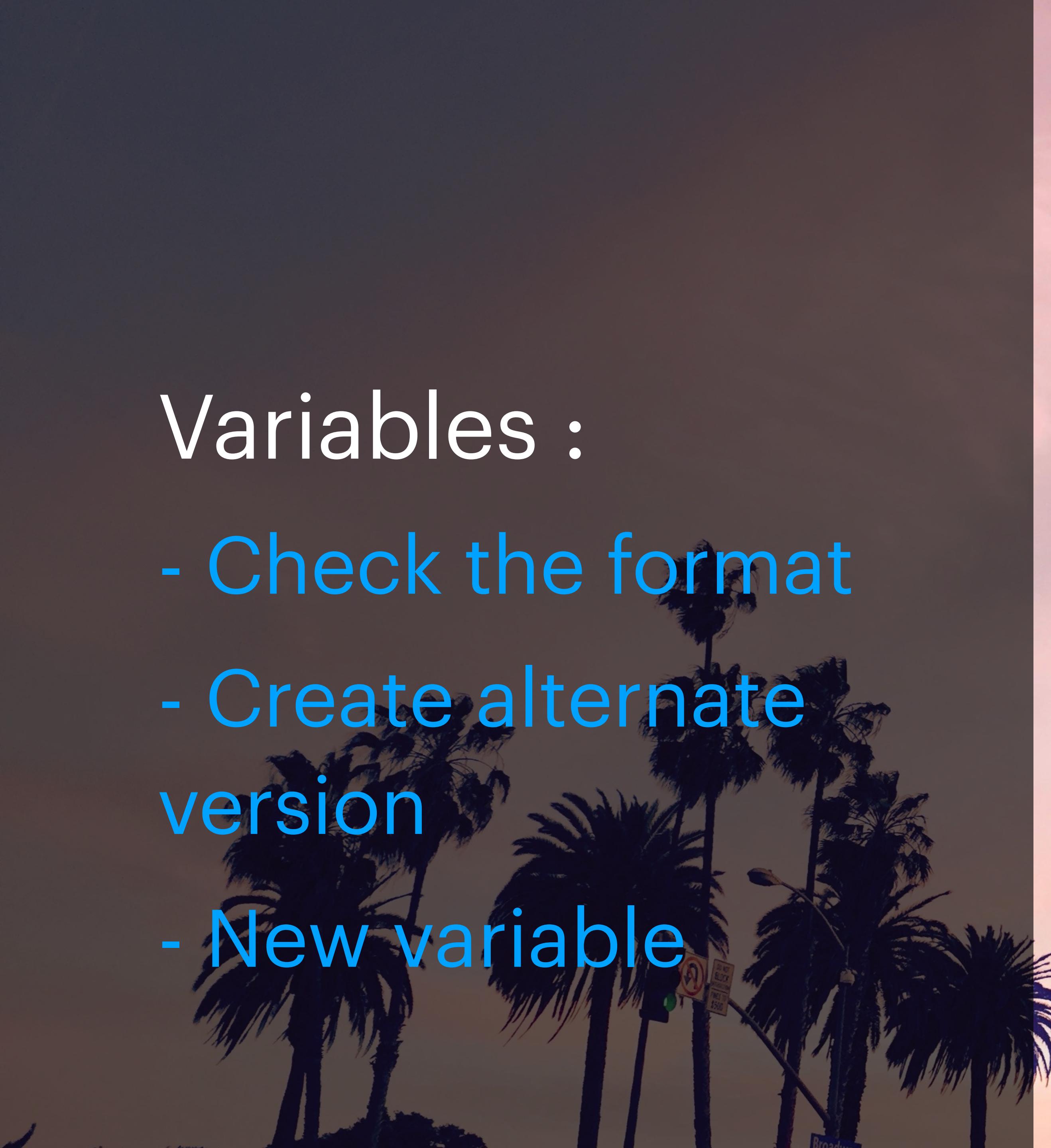
```
64  
65  **HOUSING SUPPLY - VARIABLES**  
66  
67  ```{r}  
68  boston_supply$year <- year(boston_supply$date)  
69  range(boston_supply$year)  
70  ````  
  
[1] 2016 2020
```

The line `boston_supply\$year <- year(boston_supply\$date)` is highlighted with a blue selection bar. The RStudio toolbar at the top includes icons for file operations, search, and navigation.

Variables

Variables :

- Check the format
- Create alternate version
- New variable



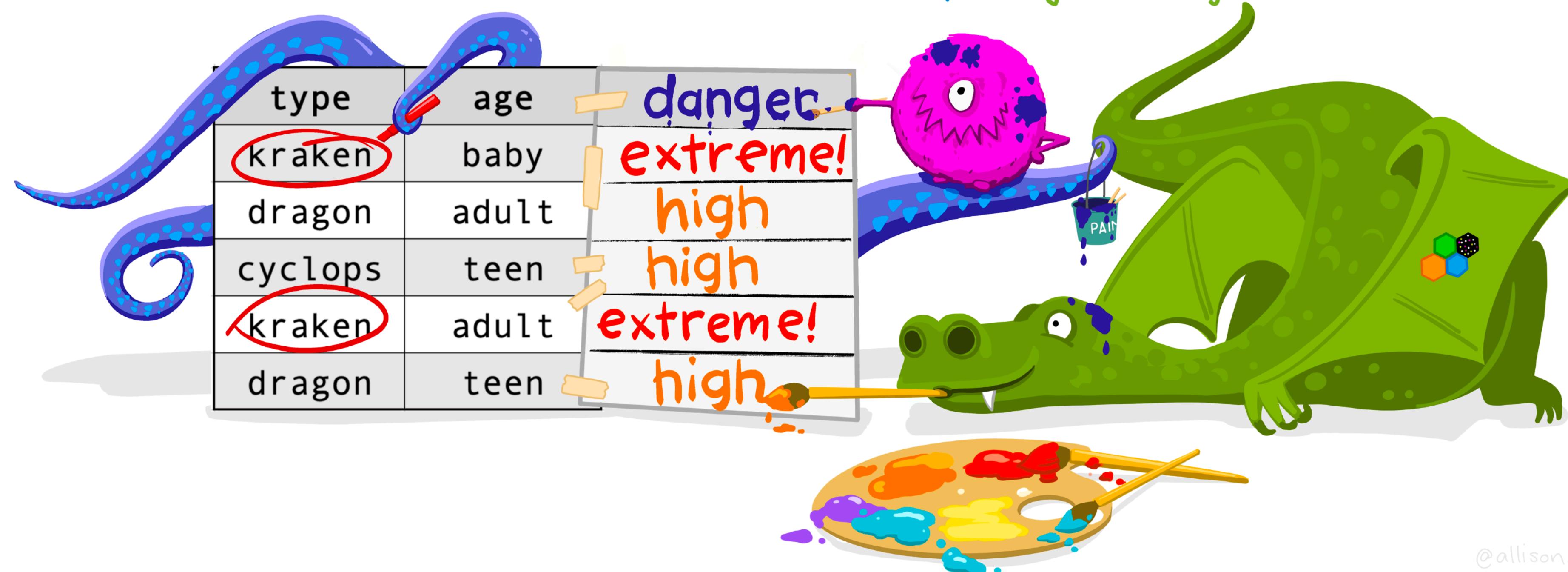


dplyr::case_when()

IF ELSE... but feels
(but you love it?)

df %>% ADD COLUMN
mutate(danger)

IF type is kraken THEN
TRUE ~ "high")
OTHERWISE, danger is high.
danger is extreme!



@allison_horst

@tladeras

```
73 **HOUSING SUPPLY - VARIABLES**
74
75 ````{r}
76 # Create dummy variable for pre/during pandemic
77 # 1 represents during pandemic
78 # 0 represents before
79 boston_supply <- boston_supply %>%
80   mutate(during_pandemic = case_when(date >= "2020-02-01" ~ 1,
81                                         date < "2020-02-01" ~ 0))
82 tail(boston_supply, n = 10)
83 ````
```

	date	series_id	value	year	during_pandemic
	<date>	<chr>	<dbl>	<int>	<dbl>
1	2019-12-01	ACTLISCOU14460	7353	2019	0
2	2020-01-01	ACTLISCOU14460	6058	2020	0
3	2020-02-01	ACTLISCOU14460	5969	2020	1
4	2020-03-01	ACTLISCOU14460	6871	2020	1
5	2020-04-01	ACTLISCOU14460	6721	2020	1
6	2020-05-01	ACTLISCOU14460	7487	2020	1
7	2020-06-01	ACTLISCOU14460	8073	2020	1
8	2020-07-01	ACTLISCOU14460	7664	2020	1
9	2020-08-01	ACTLISCOU14460	7396	2020	1
10	2020-09-01	ACTLISCOU14460	7771	2020	1

1-10 of 10 rows

The screenshot shows the RStudio interface with a dark theme. The top menu bar includes RStudio, File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Window. Below the menu is a toolbar with various icons. The main workspace shows a script file with R code. A specific line of code, `boston_supply %>%`, is highlighted with a blue rectangle. The code is as follows:

```
73 **HOUSING SUPPLY - VARIABLES**
74
75 ````{r}
76 # Create dummy variable for pre/during pandemic
77 # 1 represents during pandemic
78 # 0 represents before
79 boston_supply <- boston_supply %>%
80   mutate(during_pandemic = case_when(date >= "2020-02-01" ~ 1,
81                                         date < "2020-02-01" ~ 0))
82 tail(boston_supply, n = 10)
83 ````
```

A data preview pane below the code shows a subset of the `boston_supply` dataset. The table has columns: date, series_id, value, year, and during_pandemic.

	date	series_id	value	year	during_pandemic
	<date>	<chr>	<dbl>	<int>	<dbl>
1	2019-12-01	ACTLISCOU14460	7353	2019	0
2	2020-01-01	ACTLISCOU14460	6058	2020	0
3	2020-02-01	ACTLISCOU14460	5969	2020	1
4	2020-03-01	ACTLISCOU14460	6871	2020	1
5	2020-04-01	ACTLISCOU14460	6721	2020	1
6	2020-05-01	ACTLISCOU14460	7487	2020	1
7	2020-06-01	ACTLISCOU14460	8073	2020	1
8	2020-07-01	ACTLISCOU14460	7664	2020	1
9	2020-08-01	ACTLISCOU14460	7396	2020	1
10	2020-09-01	ACTLISCOU14460	7771	2020	1

At the bottom of the preview pane, it says "1-10 of 10 rows".

The screenshot shows the RStudio interface with several tabs open at the top: REDA.Rmd*, boston_supply, pulse_household_survey.Rmd, outlook_housing.Rmd, and R Tutorial.Rmd. The main area displays R code for creating a dummy variable 'during_pandemic' based on a date threshold. A data preview window is overlaid on the code, showing the first 10 rows of a dataset with columns: date, series_id, value, year, and during_pandemic.

```
73 **HOUSING SUPPLY - VARIABLES**
74
75 ````{r}
76 # Create dummy variable for pre/during pandemic
77 # 1 represents during pandemic
78 # 0 represents before
79 boston_supply <- boston_supply %>%
80   mutate(during_pandemic = case_when(date >= "2020-02-01" ~ 1,
81                                     date < "2020-02-01" ~ 0))
82 tail(boston_supply, n = 10)
83 ````
```

	date	series_id	value	year	during_pandemic
	<date>	<chr>	<dbl>	<int>	<dbl>
1	2019-12-01	ACTLISCOU14460	7353	2019	0
2	2020-01-01	ACTLISCOU14460	6058	2020	0
3	2020-02-01	ACTLISCOU14460	5969	2020	1
4	2020-03-01	ACTLISCOU14460	6871	2020	1
5	2020-04-01	ACTLISCOU14460	6721	2020	1
6	2020-05-01	ACTLISCOU14460	7487	2020	1
7	2020-06-01	ACTLISCOU14460	8073	2020	1
8	2020-07-01	ACTLISCOU14460	7664	2020	1
9	2020-08-01	ACTLISCOU14460	7396	2020	1
10	2020-09-01	ACTLISCOU14460	7771	2020	1

1-10 of 10 rows

REDA.Rmd* boston_supply pulse_household_survey.Rmd outlook_housing.Rmd R Tutorial.Rmd

ABC Knit Insert Run

```
73 **HOUSING SUPPLY - VARIABLES**
74
75 ````{r}
76 # Create dummy variable for pre/during pandemic
77 # 1 represents during pandemic
78 # 0 represents before
79 boston_supply <- boston_supply %>%
80   mutate(during_pandemic = case_when(date >= "2020-02-01" ~ 1,
81                                         date < "2020-02-01" ~ 0))
82 tail(boston_supply, n = 10)
83 ````
```

	date	series_id	value	year	during_pandemic
	<date>	<chr>	<dbl>	<int>	<dbl>
1	2019-12-01	ACTLISCOU14460	7353	2019	0
2	2020-01-01	ACTLISCOU14460	6058	2020	0
3	2020-02-01	ACTLISCOU14460	5969	2020	1
4	2020-03-01	ACTLISCOU14460	6871	2020	1
5	2020-04-01	ACTLISCOU14460	6721	2020	1
6	2020-05-01	ACTLISCOU14460	7487	2020	1
7	2020-06-01	ACTLISCOU14460	8073	2020	1
8	2020-07-01	ACTLISCOU14460	7664	2020	1
9	2020-08-01	ACTLISCOU14460	7396	2020	1
10	2020-09-01	ACTLISCOU14460	7771	2020	1

1-10 of 10 rows

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

The screenshot shows the RStudio interface with a dark theme. The top menu bar includes RStudio, File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Window. Below the menu is a toolbar with various icons. The main workspace shows a script file with R code. A specific line of code is highlighted with a blue box:

```
73 **HOUSING SUPPLY - VARIABLES**  
74  
75 ````{r}  
76 # Create dummy variable for pre/during pandemic  
77 # 1 represents during pandemic  
78 # 0 represents before  
79 boston_supply <- boston_supply %>%  
80   mutate(during_pandemic = case_when(date >= "2020-02-01" ~ 1,  
81                                         date < "2020-02-01" ~ 0))  
82 tail(boston_supply, n = 10)  
83 ````
```

The code creates a new column 'during_pandemic' in the 'boston_supply' dataset, which is 1 for dates on or after February 1, 2020, and 0 for dates before that. A data preview pane below the code shows the first 10 rows of the 'boston_supply' dataset:

date	series_id	value	year	during_pandemic
<date>	<chr>	<dbl>	<int>	<dbl>
2019-12-01	ACTLISCOU14460	7353	2019	0
2020-01-01	ACTLISCOU14460	6058	2020	0
2020-02-01	ACTLISCOU14460	5969	2020	1
2020-03-01	ACTLISCOU14460	6871	2020	1
2020-04-01	ACTLISCOU14460	6721	2020	1
2020-05-01	ACTLISCOU14460	7487	2020	1
2020-06-01	ACTLISCOU14460	8073	2020	1
2020-07-01	ACTLISCOU14460	7664	2020	1
2020-08-01	ACTLISCOU14460	7396	2020	1
2020-09-01	ACTLISCOU14460	7771	2020	1

At the bottom of the preview pane, it says "1-10 of 10 rows".

The screenshot shows an RStudio interface with a dark theme. The top menu bar includes 'File', 'Edit', 'Code', 'View', 'Plots', 'Session', 'Build', 'Debug', 'Profile', 'Tools', and 'Window'. Below the menu is a toolbar with various icons. The main area displays an R script and a data frame visualization.

The R script code is as follows:

```
73 **HOUSING SUPPLY - VARIABLES**
74
75 ````{r}
76 # Create dummy variable for pre/during pandemic
77 # 1 represents during pandemic
78 # 0 represents before
79 boston_supply <- boston_supply %>%
80   mutate(during_pandemic = case_when(
81     date >= "2020-02-01" ~ 1,
82     date < "2020-02-01" ~ 0))
83 tail(boston_supply, n = 10)
````
```

The data frame visualization shows the 'boston\_supply' dataset with columns: date, series\_id, value, year, and during\_pandemic. The期间变量(during\_pandemic)在2020年2月1日及之后的行中值为1，之前为0。行数从1到10。

| date       | series_id      | value | year  | during_pandemic |
|------------|----------------|-------|-------|-----------------|
| <date>     | <chr>          | <dbl> | <int> | <dbl>           |
| 2019-12-01 | ACTLISCOU14460 | 7353  | 2019  | 0               |
| 2020-01-01 | ACTLISCOU14460 | 6058  | 2020  | 0               |
| 2020-02-01 | ACTLISCOU14460 | 5969  | 2020  | 1               |
| 2020-03-01 | ACTLISCOU14460 | 6871  | 2020  | 1               |
| 2020-04-01 | ACTLISCOU14460 | 6721  | 2020  | 1               |
| 2020-05-01 | ACTLISCOU14460 | 7487  | 2020  | 1               |
| 2020-06-01 | ACTLISCOU14460 | 8073  | 2020  | 1               |
| 2020-07-01 | ACTLISCOU14460 | 7664  | 2020  | 1               |
| 2020-08-01 | ACTLISCOU14460 | 7396  | 2020  | 1               |
| 2020-09-01 | ACTLISCOU14460 | 7771  | 2020  | 1               |

1-10 of 10 rows

Variables :

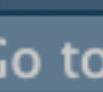
- Count



# Variables



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



Go to file/function Addins

REDA.Rmd x

boston\_supply x

pulse\_household\_survey.Rmd x

outlook\_housing.Rmd x

R Tutorial.Rmd x



85

86 \*\*HOUSING SUPPLY - COUNT\*\*

87

88 ``{r}

```
89 # Create 'count' dataframe
90 boston_count <- boston_supply %>%
91 count(during_pandemic) %>%
92 rename(count = n) %>%
93 mutate(total = sum(count),
94 share = (count/total)*100)
95 head(boston_count)
```

96



c

Insert



Run



| during_pandemic<br><dbl> | count<br><int> | total<br><int> | share<br><dbl> |
|--------------------------|----------------|----------------|----------------|
| 0                        | 43             | 51             | 84.31373       |
| 1                        | 8              | 51             | 15.68627       |

2 rows

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



REDA.Rmd x boston\_supply x pulse\_household\_survey.Rmd x outlook\_housing.Rmd x R Tutorial.Rmd x»



```
85
86 **HOUSING SUPPLY - COUNT**
87
88 ````{r}
89 # Create 'count' dataframe
90 boston_count <- boston_supply %>%
91 count(during_pandemic) %>%
92 rename(count = n) %>%
93 mutate(total = sum(count),
94 share = (count/total)*100)
95 head(boston_count)
96 ````
```



| during_pandemic | count | total | share |
|-----------------|-------|-------|-------|
| <dbl>           | <int> | <int> | <dbl> |

|   |    |    |          |
|---|----|----|----------|
| 0 | 43 | 51 | 84.31373 |
| 1 | 8  | 51 | 15.68627 |

2 rows

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



Go to file/function



Addins

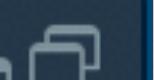
REDA.Rmd x

boston\_supply x

pulse\_household\_survey.Rmd x

outlook\_housing.Rmd x

R Tutorial.Rmd x



85

86 \*\*HOUSING SUPPLY - COUNT\*\*

87

88 ``{r}

89 # Create 'count' dataframe

90 boston\_count <- boston\_supply %>%

91 count(during pandemic) %>%

92 rename(count = n) %>%

93 mutate(total = sum(count),

94 share = (count/total)\*100)

95 head(boston\_count)

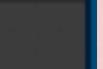
96 ````

+c

Insert



Run



| during_pandemic<br><dbl> | count<br><int> | total<br><int> | share<br><dbl> |
|--------------------------|----------------|----------------|----------------|
| 0                        | 43             | 51             | 84.31373       |
| 1                        | 8              | 51             | 15.68627       |

2 rows

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



Go to file/function



Addins

REDA.Rmd x

boston\_supply x

pulse\_household\_survey.Rmd x

outlook\_housing.Rmd x

R Tutorial.Rmd x



85

86 \*\*HOUSING SUPPLY - COUNT\*\*

87

88 ``{r}

89 # Create 'count' dataframe

90 boston\_count <- boston\_supply %>%

91 count(during\_pandemic) %>%

92 rename(count = n) %>%

93 mutate(total = sum(count),

94 share = (count/total)\*100)

95 head(boston\_count)

96 ````

+c

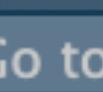
Insert



| during_pandemic<br><dbl> | count<br><int> | total<br><int> | share<br><dbl> |
|--------------------------|----------------|----------------|----------------|
| 0                        | 43             | 51             | 84.31373       |
| 1                        | 8              | 51             | 15.68627       |

2 rows

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



Addins

REDA.Rmd x

boston\_supply x

pulse\_household\_survey.Rmd x

outlook\_housing.Rmd x

R Tutorial.Rmd x



85

86 \*\*HOUSING SUPPLY - COUNT\*\*

87

88 ``{r}

89 # Create 'count' dataframe

90 boston\_count <- boston\_supply %>%

91 count(during\_pandemic) %>%

92 rename(count = n) %>%

93 mutate(total = sum(count),

94 share <- count/total)\*100)

95 head(boston\_count)

96

| during_pandemic<br><dbl> | count<br><int> | total<br><int> | share<br><dbl> |
|--------------------------|----------------|----------------|----------------|
| 0                        | 43             | 51             | 84.31373       |
| 1                        | 8              | 51             | 15.68627       |

2 rows

Variables :

- Descriptive  
statistics

# Variables



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

REDA.Rmd   boston\_supply   pulse\_household\_survey.Rmd   outlook\_housing.Rmd   R Tutorial.Rmd

Go to file/function   Addins

```
85
86 **HOUSING SUPPLY - SUMMARY**
87
88 ````{r}
89 # Create 'summary' dataframe
90 # includes mean, median, standard deviation, and coefficient of variation
91 boston_summary <- boston_supply %>%
92 group_by(during_pandemic) %>%
93 summarise(mean_value = mean(value),
94 median_value = median(value),
95 sd_value = sd(value),
96 cv_value = (sd_value/mean_value)*100)
97 head(boston_summary)
98 ````
```

R Console

tbl\_df  
2 x 5

| during_pandemic | mean_value | median_value | sd_value  | cv_value  |
|-----------------|------------|--------------|-----------|-----------|
| <dbl>           | <dbl>      | <dbl>        | <dbl>     | <dbl>     |
| 0               | 9657.791   | 10261.0      | 1876.3054 | 19.427895 |
| 1               | 7244.000   | 7441.5       | 682.6216  | 9.423269  |

2 rows

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

REDA.Rmd   boston\_supply   pulse\_household\_survey.Rmd   outlook\_housing.Rmd   R Tutorial.Rmd

Go to file/function   Addins

```
85
86 **HOUSING SUPPLY - SUMMARY**
87
88 ````{r}
89 # Create 'summary' dataframe
90 # includes mean, median, standard deviation, and coefficient of variation
91 boston_summary <- boston_supply %>%
92 group_by(during_pandemic) %>%
93 summarise(mean_value = mean(value),
94 median_value = median(value),
95 sd_value = sd(value),
96 cv_value = (sd_value/mean_value)*100)
97 head(boston_summary)
98 ````
```

R Console   tbl\_df   2 x 5

| during_pandemic | mean_value | median_value | sd_value  | cv_value  |
|-----------------|------------|--------------|-----------|-----------|
| <dbl>           | <dbl>      | <dbl>        | <dbl>     | <dbl>     |
| 0               | 9657.791   | 10261.0      | 1876.3054 | 19.427895 |
| 1               | 7244.000   | 7441.5       | 682.6216  | 9.423269  |

2 rows

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

REDA.Rmd   boston\_supply   pulse\_household\_survey.Rmd   outlook\_housing.Rmd   R Tutorial.Rmd

85  
86   \*\*HOUSING SUPPLY - SUMMARY\*\*  
87  
88   ```{r}  
89   # Create 'summary' dataframe  
90   # includes mean, median, standard deviation, and coefficient of variation  
91   boston\_summary <- boston\_supply %>%  
92   group\_by(during\_pandemic) %>%  
93   summarise(mean\_value = mean(value),  
94   median\_value = median(value),  
95   sd\_value = sd(value),  
96   cv\_value = (sd\_value / mean\_value) \* 100)  
97   head(boston\_summary)  
98   ````

```{r}  
Create 'summary' dataframe
includes mean, median, standard deviation, and coefficient of variation
boston_summary <- boston_supply %>%
 group_by(during_pandemic) %>%
 summarise(mean_value = mean(value),
 median_value = median(value),
 sd_value = sd(value),
 cv_value = (sd_value / mean_value) * 100)
head(boston_summary)
````

R Console   tbl\_df   2 x 5

during_pandemic	mean_value	median_value	sd_value	cv_value
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0	9657.791	10261.0	1876.3054	19.427895
1	7244.000	7441.5	682.6216	9.423269

2 rows

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

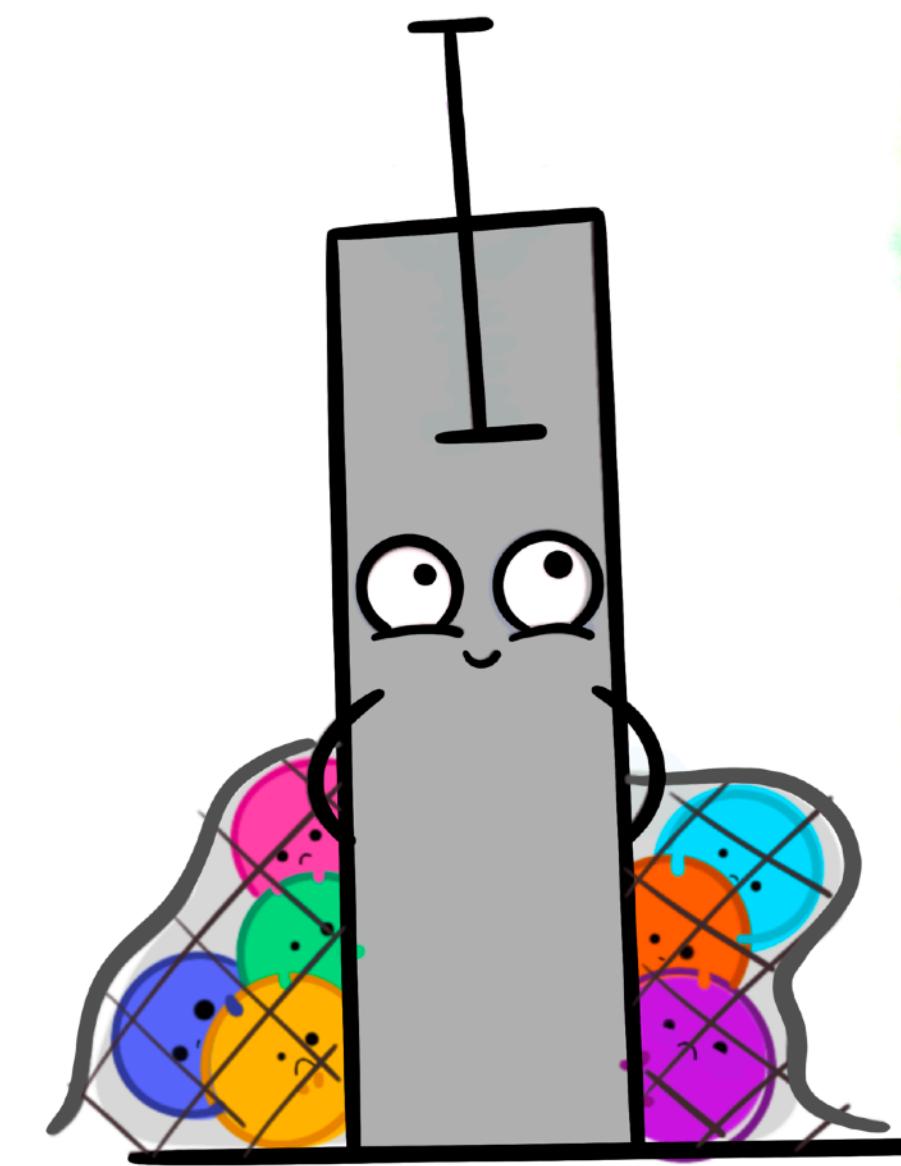
REDA.Rmd   boston\_supply   pulse\_household\_survey.Rmd   outlook\_housing.Rmd   R Tutorial.Rmd

85  
86   \*\*HOUSING SUPPLY - SUMMARY\*\*  
87  
88   ```{r}  
89   # Create 'summary' dataframe  
90   # includes mean, median, standard deviation, and coefficient of variation  
91   boston\_summary <- boston\_supply %>%  
92   group\_by(during\_pandemic) %>%  
93   summarise(mean\_value = mean(value),  
94   median\_value = median(value),  
95   sd\_value = sd(value),  
96   cv\_value = (sd\_value/mean\_value)\*100)  
97   head(boston\_summary)  
98   ````

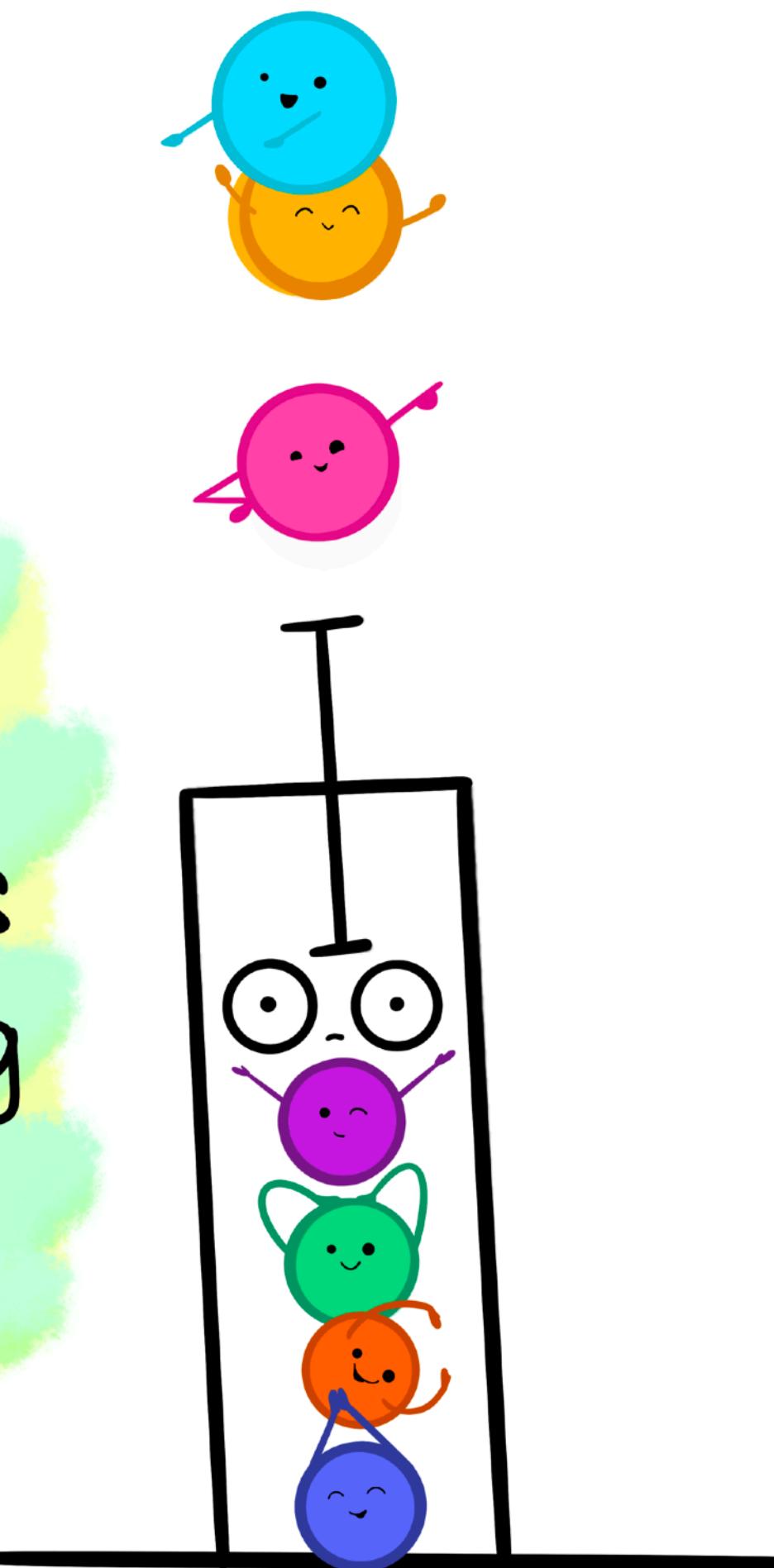
R Console   tbl\_df   2 x 5

during_pandemic	mean_value	median_value	sd_value	cv_value
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
0	9657.791	10261.0	1876.3054	19.427895
1	7244.000	7441.5	682.6216	9.423269

2 rows



are your  
summary statistics  
hiding something  
interesting?



@allison\_horst

