

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 data



Overview

In this section, we will:

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



Variables



Variables :

- Categorical

Variables



NOMINAL

UNORDERED DESCRIPTIONS



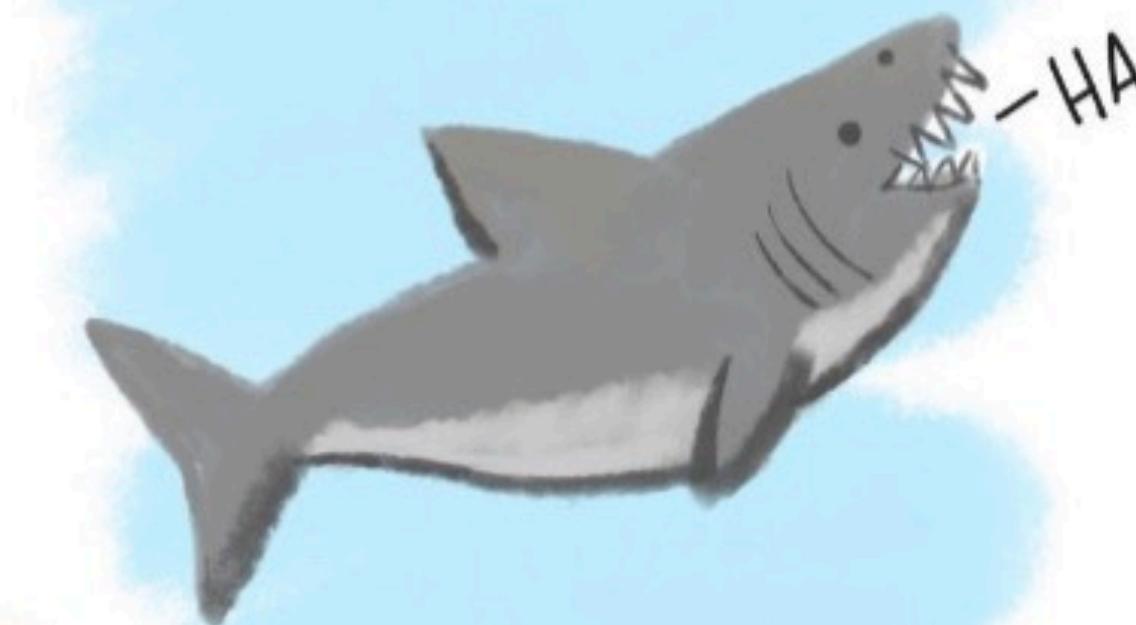
ORDINAL

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!

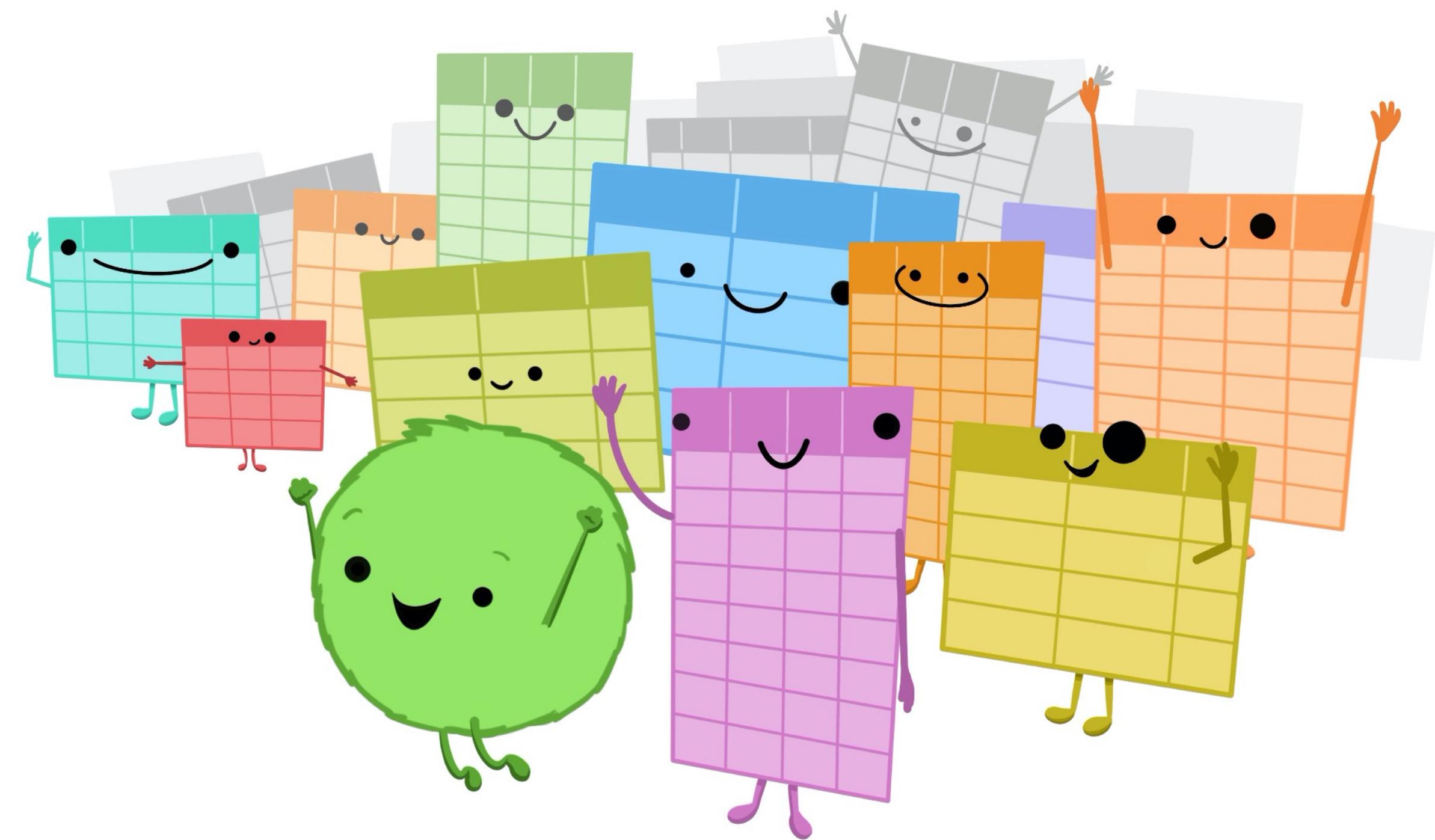
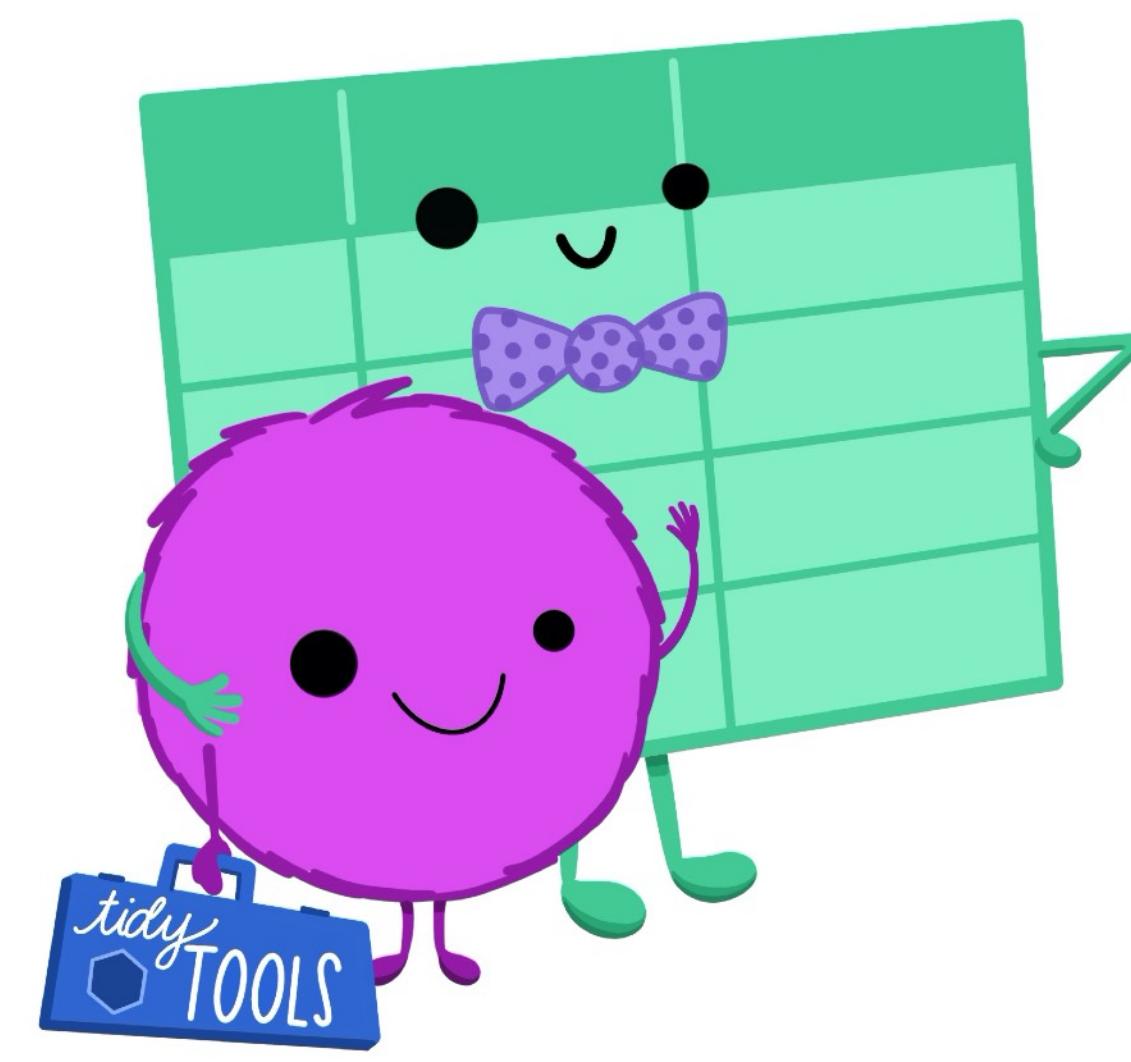
@allison_horst

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')

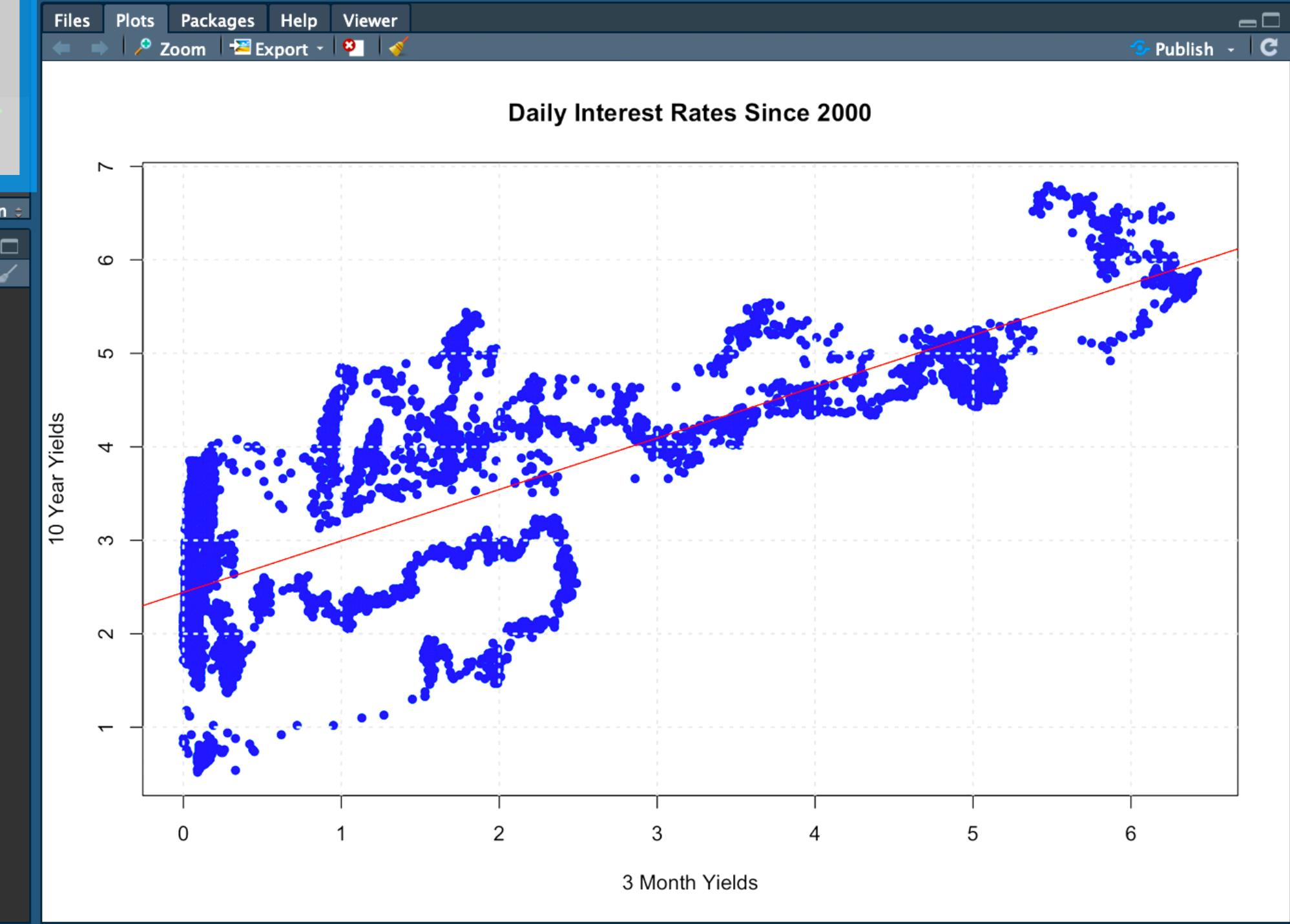
```

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:

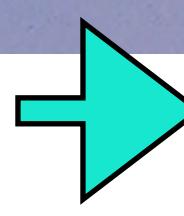
Monthly

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

The screenshot shows the RStudio desktop application. 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 for file operations like saving, opening, and printing, along with a 'Go to file/function' search bar and an 'Addins' dropdown.

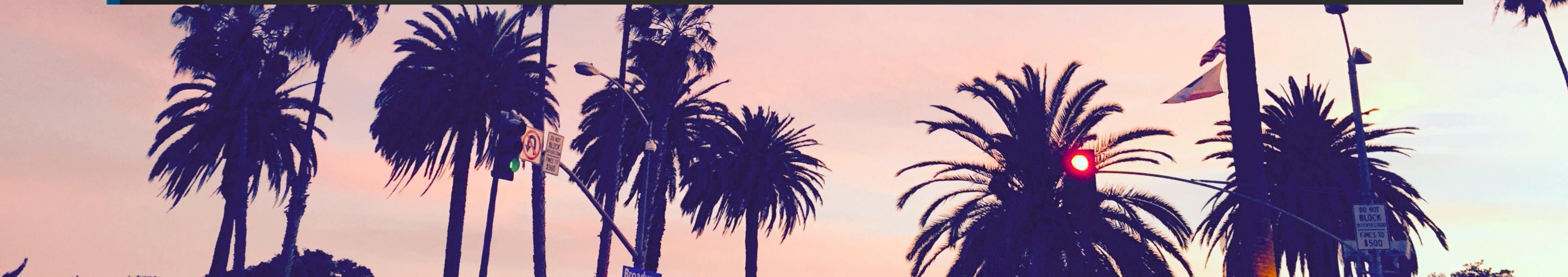
The main workspace displays several R Markdown files in tabs: 'REDA.Rmd', 'pulse_household_survey.Rmd' (which is currently active), 'outlook_housing.Rmd', 'R Tutorial.Rmd', and 'R Tutorial Part 2.'. Below the tabs are toolbars for navigation, code folding, and knit/run operations.

The code editor area contains the following R code:

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

The terminal output window at the bottom shows the result of the R code execution:

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



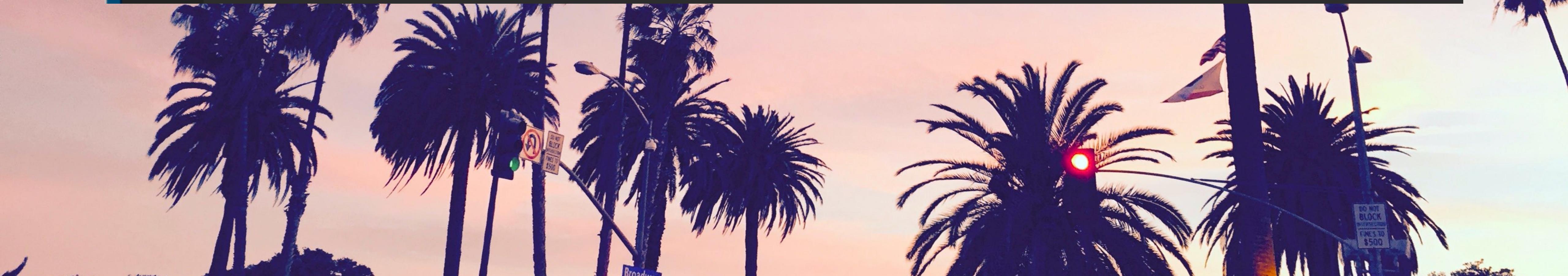
RStudio File Edit Code View Plots Session Build Debug Profile Tools Window

The screenshot shows the RStudio interface with a dark theme. The top menu bar includes 'RStudio' and various tabs like 'File', 'Edit', 'Code', etc. Below the menu is a toolbar with icons for file operations and a 'Go to file/function' search bar. A tab bar at the top lists several R Markdown files: 'REDA.Rmd', 'pulse_household_survey.Rmd' (which is currently selected), 'outlook_housing.Rmd', 'R Tutorial.Rmd', and 'R Tutorial Part 2.'. The main workspace displays R code and its output. The code is as follows:

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

The line 'as.Date(boston_supply\$date)' is highlighted with a blue selection box. The output pane below shows the result of the 'range' function:

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



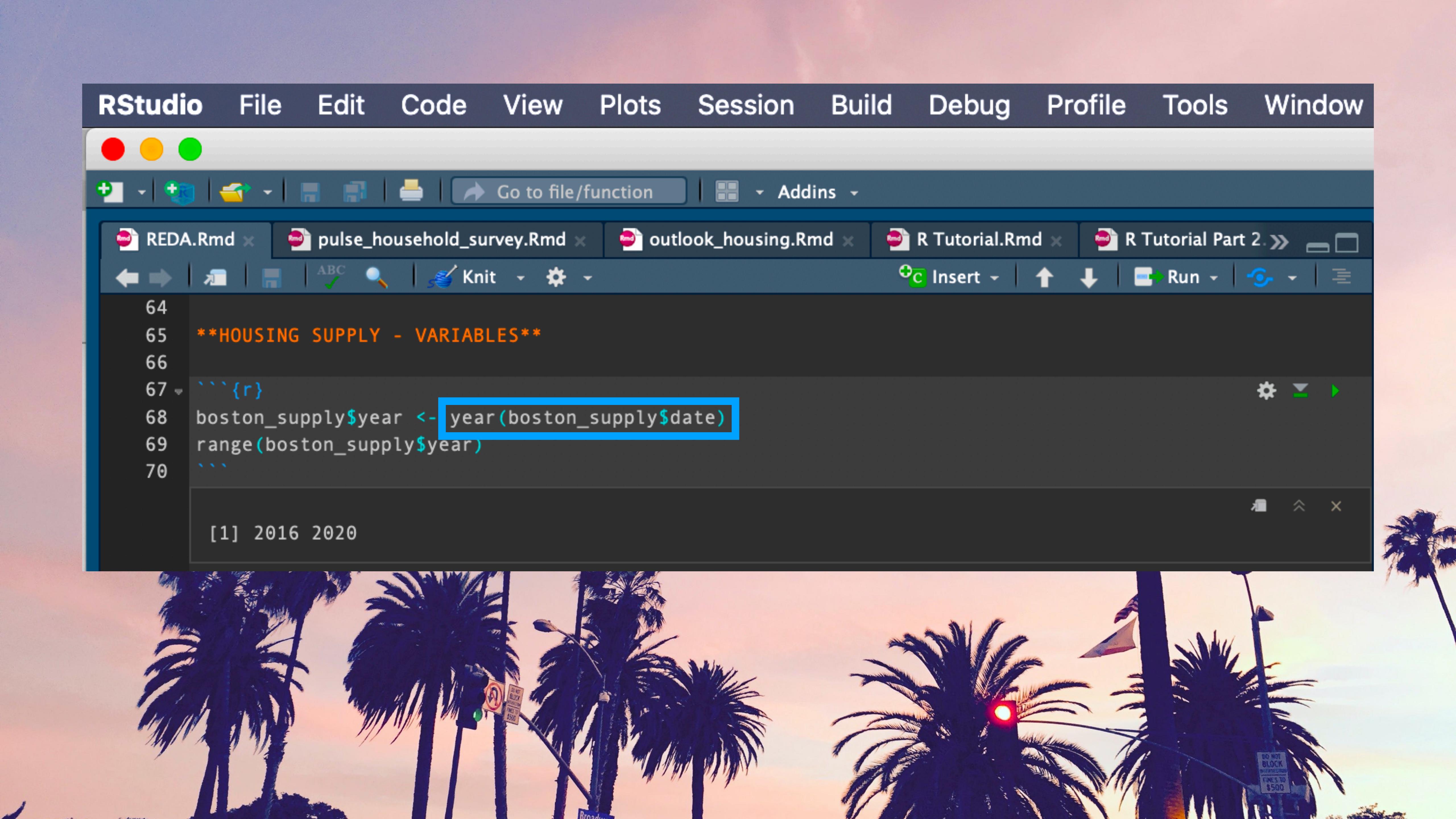
Variables :

- 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:

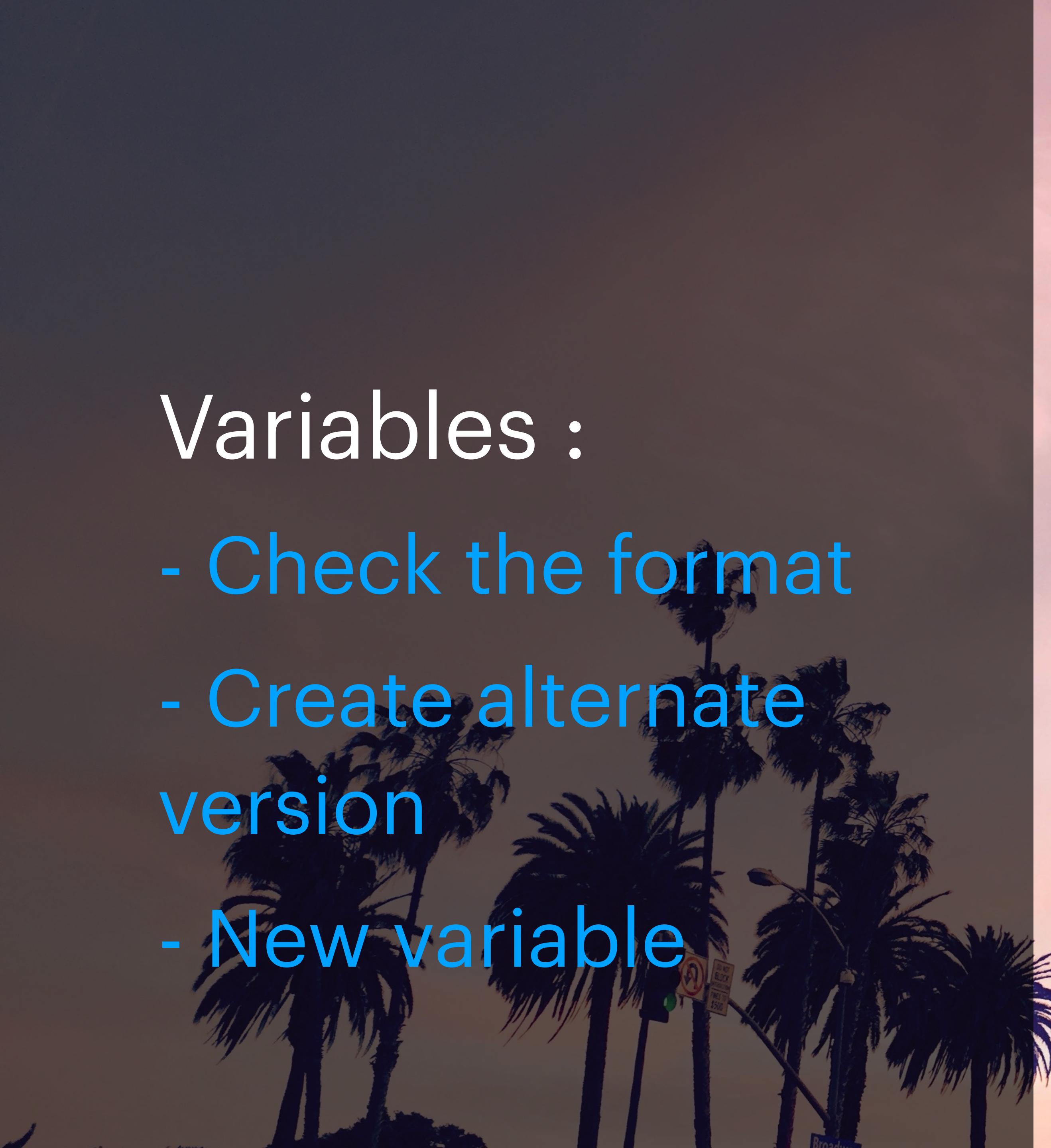
```
64  
65  **HOUSING SUPPLY - VARIABLES**  
66  
67  ```{r}  
68  boston_supply$year <- year(boston_supply$date)  
69  range(boston_supply$year)  
70  ````  
  
[1] 2016 2020
```

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

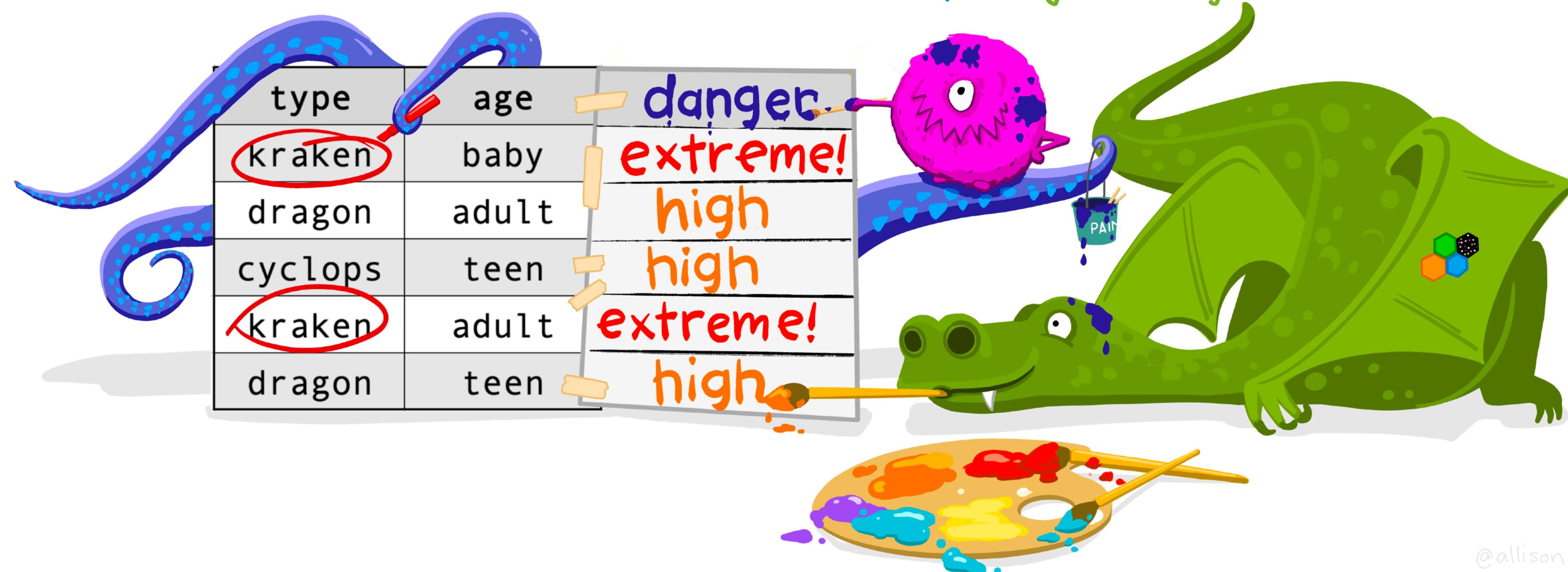


dplyr::case_when()

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

The screenshot shows the RStudio interface with a script editor containing R code and a data viewer pane displaying a subset of the data.

Script Editor Content:

```
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 ````
```

Data Preview:

	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. The toolbar below has icons for file operations like Open, Save, and Print, along with Go to file/function, Addins, and a search icon.

The main workspace shows a script titled "boston_supply.Rmd" with the following code:

```
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)
````
```

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>           |
| 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".

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(
81 date >= "2020-02-01" ~ 1,
82 date < "2020-02-01" ~ 0))
83 tail(boston_supply, n = 10)
````
```

| date
<date> | series_id
<chr> | value
<dbl> | year
<int> | during_pandemic
<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 :

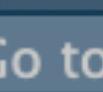
- Count



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
<dbl> | count
<int> | total
<int> | share
<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 ````
```

+c Insert ▾ | ↑ ↓ | Run ▾ | ⌂ | ⌂ | ⌂



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



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),

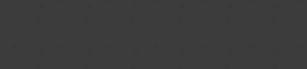
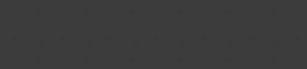
94 share = (count/total)*100)

95 head(boston_count)

96 ````

+c

Insert



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

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

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



Categories > Production & Business Activity > Housing

★ Housing Inventory: Average Listing Price in Boston-Cambridge-Newton, MA-NH (CBSA) (AVELISPRI14460)

Observation:

Sep 2020: **1,034,098** (+ more)

Updated: Oct 1, 2020

Units:

U.S. Dollars,
Not Seasonally Adjusted

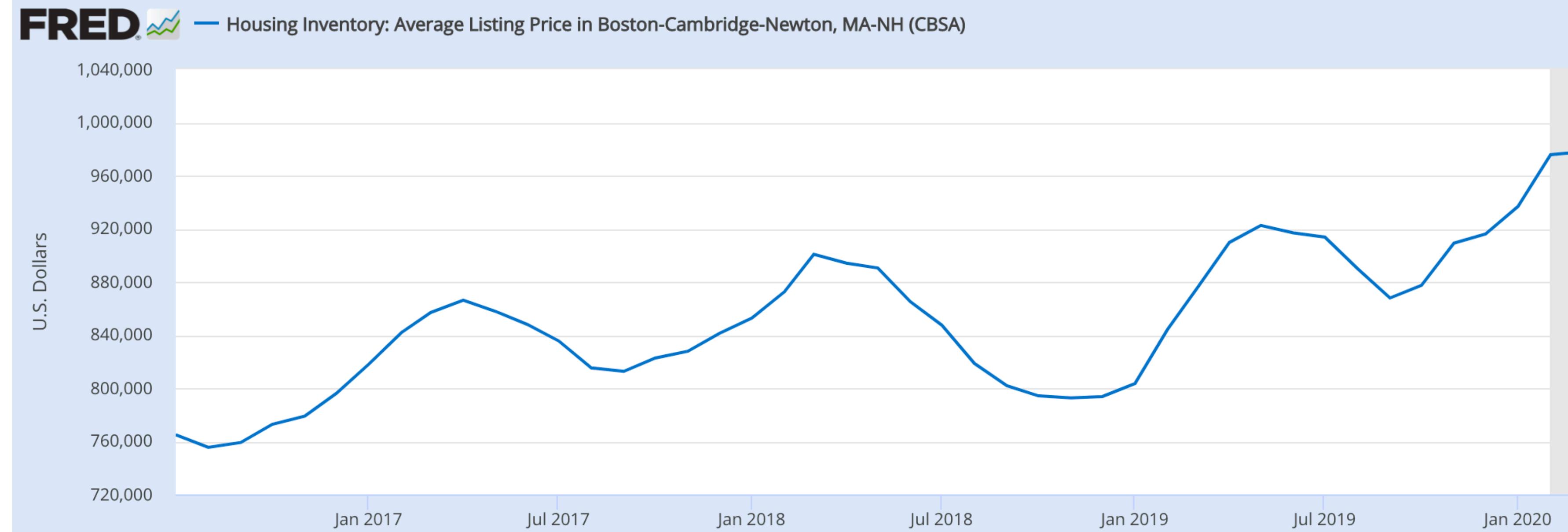
Frequency:

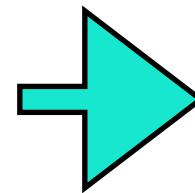
Monthly

1Y | 5Y | 10Y | Max

2016-07-01

to 2020-09-01





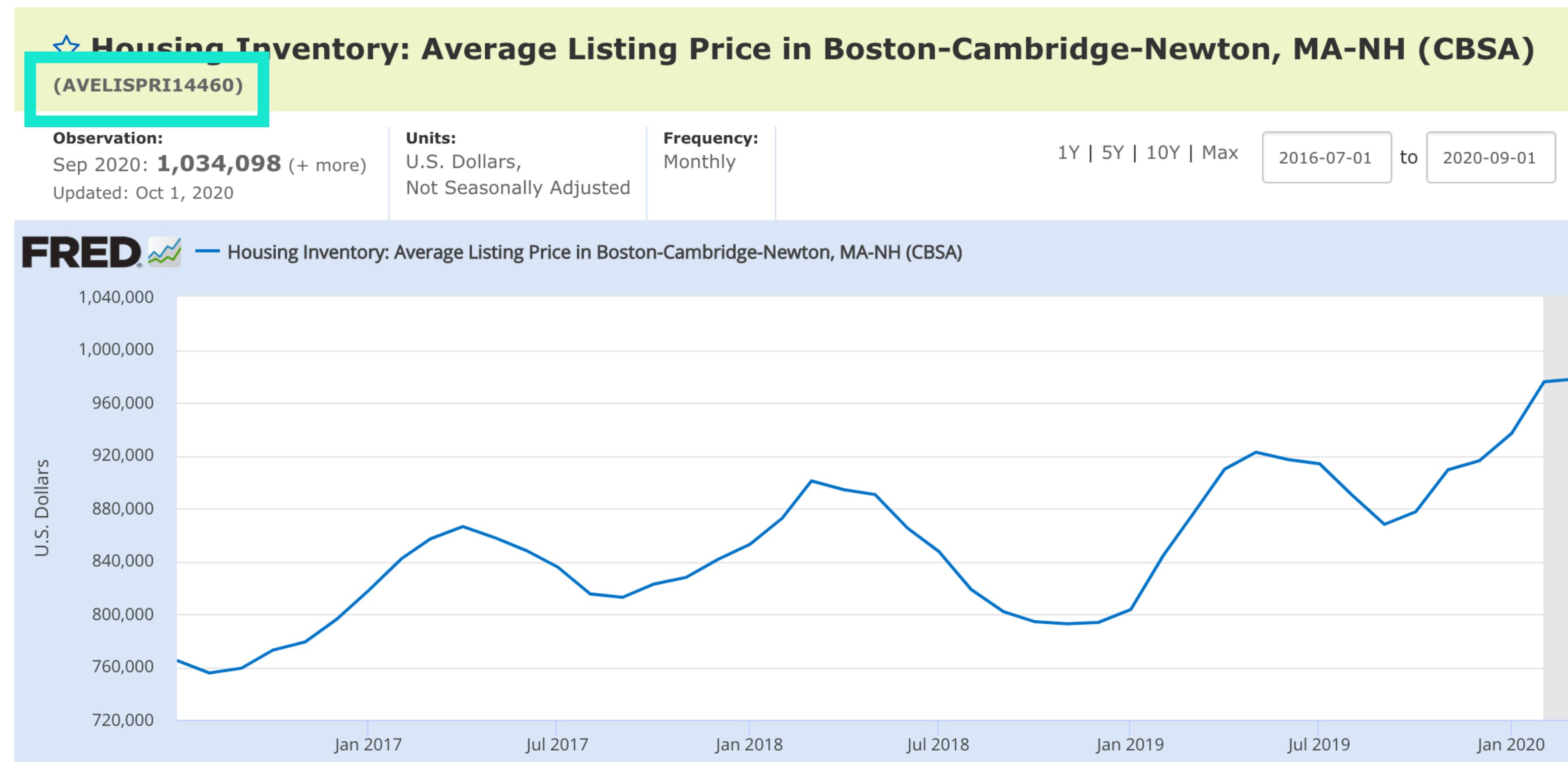
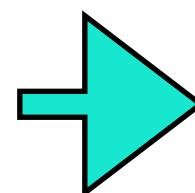
https://fred.stlouisfed.org/series/AVELISPRI14460

FRED | 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





Go to file/function Addins

REDA.Rmd x merged_price x Assignment 2 - CM.Rmd x boston_supply x pulse_household_survey.Rmd x»

ABC Knit Run Insert

120
127 **AVERAGE LISTING PRICE**
128
129 ``{r}
130 # Average housing listing price in Boston-Cambridge-Newton, MA-NH
131 # Earliest start date: July 2016
132 boston_price <- fredr('AVELISPRI14460',
133 observation_start = as.Date("2016-07-01"),
134 frequency = "m")
135
136 boston_price\$date <- as.Date(boston_price\$date)
137
138 head(boston_price)
139 ````

	date	series_id	value
	<date>	<chr>	<dbl>
	2016-07-01	AVELISPRI14460	764864
	2016-08-01	AVELISPRI14460	755404
	2016-09-01	AVELISPRI14460	759052
	2016-10-01	AVELISPRI14460	772755
	2016-11-01	AVELISPRI14460	778844
	2016-12-01	AVELISPRI14460	795947

6 rows



Go to file/function

Addins

REDA.Rmd x merged_price x Assignment 2 - CM.Rmd x boston_supply x pulse_household_survey.Rmd x

ABC Knit

Insert Run

```
127 **AVERAGE LISTING PRICE**  
128  
129 ``{r}  
130 # Average housing listing price in Boston-Cambridge-Newton, MA-NH  
131 # Earliest start date: July 2016  
132 boston_price <- freq('AVELISPRI14460',  
133   observation_start = as.Date("2016-07-01"),  
134   frequency = "m")  
135  
136 boston_price$date <- as.Date(boston_price$date)  
137  
138 head(boston_price)  
````
```

|  | date       | series_id      | value  |
|--|------------|----------------|--------|
|  | <date>     | <chr>          | <dbl>  |
|  | 2016-07-01 | AVELISPRI14460 | 764864 |
|  | 2016-08-01 | AVELISPRI14460 | 755404 |
|  | 2016-09-01 | AVELISPRI14460 | 759052 |
|  | 2016-10-01 | AVELISPRI14460 | 772755 |
|  | 2016-11-01 | AVELISPRI14460 | 778844 |
|  | 2016-12-01 | AVELISPRI14460 | 795947 |

6 rows



Go to file/function

Addins

REDA.Rmd x merged\_price x Assignment 2 - CM.Rmd x boston\_supply x pulse\_household\_survey.Rmd x

ABC Knit

Insert Run

```
127 **AVERAGE LISTING PRICE**
128
129 ``{r}
130 # Average housing listing price in Boston-Cambridge-Newton, MA-NH
131 # Earliest start date: July 2016
132 boston_price <- freq('AVELISPRI14460',
133 observation_start = as.Date("2016-07-01"),
134 frequency = "m")
135
136 boston_price$date <- as.Date(boston_price$date)
137
138 head(boston_price)
139 ````
```

|  | date       | series_id      | value  |
|--|------------|----------------|--------|
|  | <date>     | <chr>          | <dbl>  |
|  | 2016-07-01 | AVELISPRI14460 | 764864 |
|  | 2016-08-01 | AVELISPRI14460 | 755404 |
|  | 2016-09-01 | AVELISPRI14460 | 759052 |
|  | 2016-10-01 | AVELISPRI14460 | 772755 |
|  | 2016-11-01 | AVELISPRI14460 | 778844 |
|  | 2016-12-01 | AVELISPRI14460 | 795947 |

6 rows



Categories > Production & Business Activity > Housing

## ★ Housing Inventory: Average Listing Price in the United States (AVELISPRIUS)

**Observation:**

Sep 2020: **647,012** (+ more)

Updated: Oct 1, 2020

**Units:**

U.S. Dollars,  
Not Seasonally Adjusted

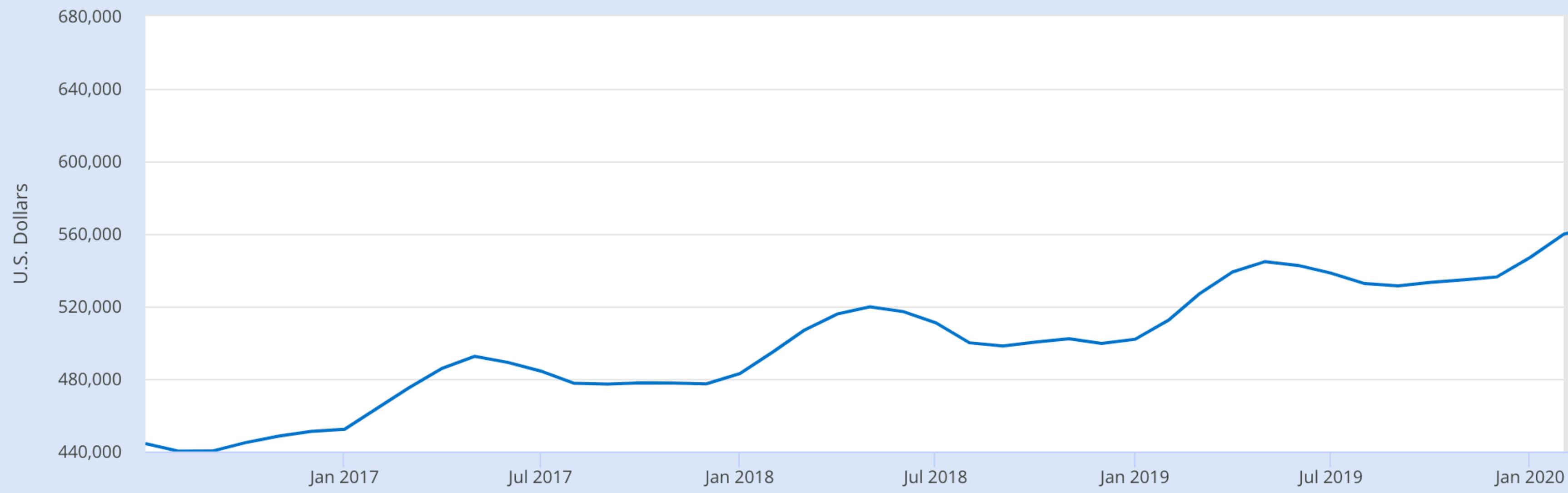
**Frequency:**

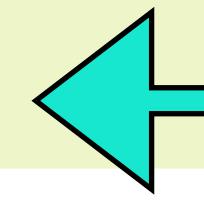
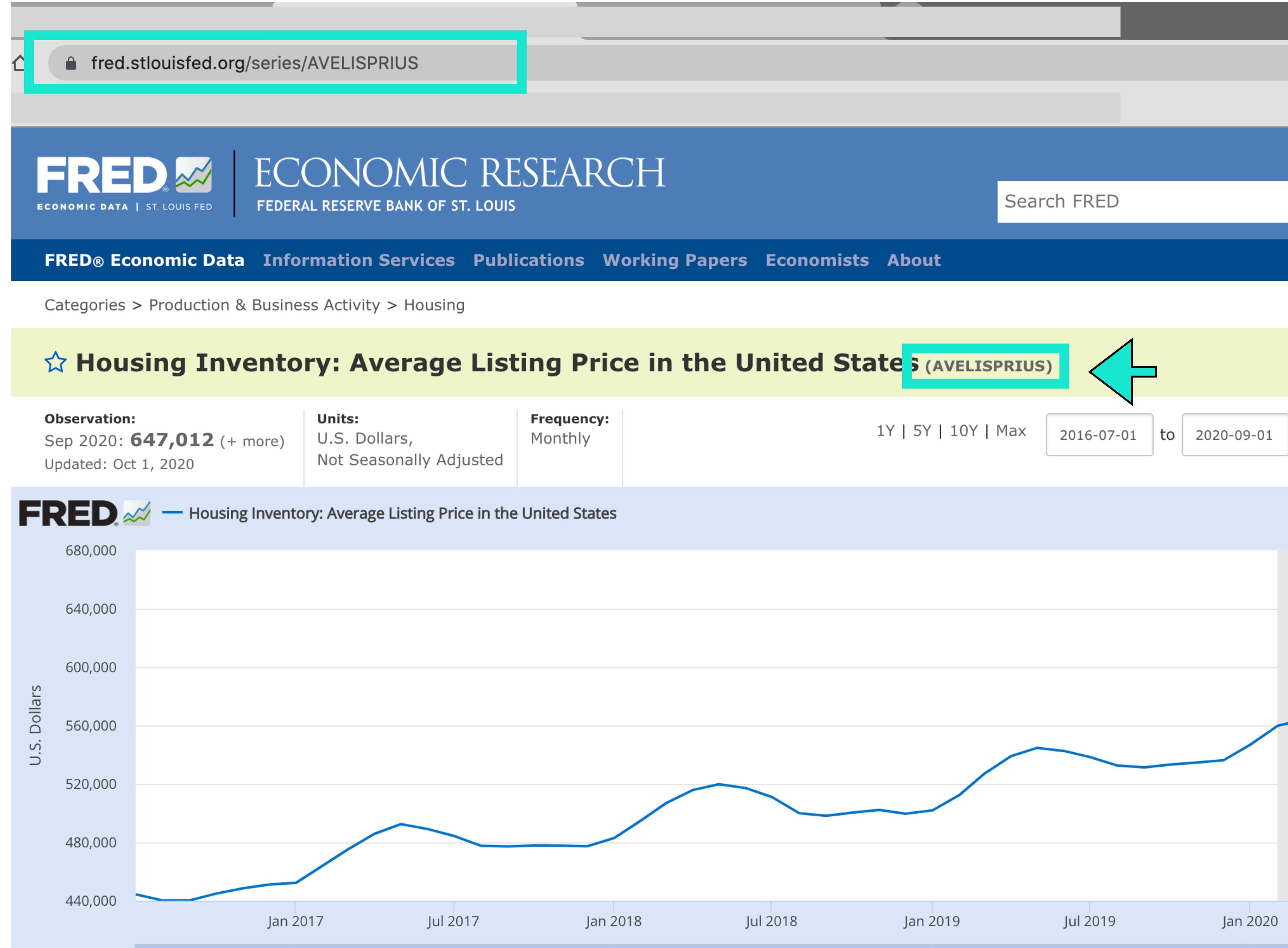
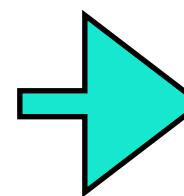
Monthly

1Y | 5Y | 10Y | Max

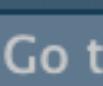
2016-07-01 to

2020-09-01





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



Go to file/function



REDA.Rmd x merged\_price x boston\_supply x pulse\_household\_survey.Rmd >>

ABC Knit Settings Insert Run

```
141 **AVERAGE LISTING PRICE - NATIONAL**
142
143 ``{r}
144 # Average housing listing price in U.S.
145 # Earliest start date: July 2016
146 national_price <- fredr('AVELISPRIUS',
147 observation_start = as.Date("2016-07-01"),
148 frequency = "m")
149
150 national_price$date <- as.Date(national_price$date)
151
152 head(national_price)
````
```

| | date | series_id | value |
|---|------------|-------------|--------|
| | <date> | <chr> | <dbl> |
| 1 | 2016-07-01 | AVELISPRIUS | 444350 |
| 2 | 2016-08-01 | AVELISPRIUS | 440166 |
| 3 | 2016-09-01 | AVELISPRIUS | 440295 |
| 4 | 2016-10-01 | AVELISPRIUS | 444839 |
| 5 | 2016-11-01 | AVELISPRIUS | 448456 |
| 6 | 2016-12-01 | AVELISPRIUS | 451075 |

6 rows

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



Go to file/function Addins

REDA.Rmd x

merged_price x

boston_supply x

pulse_household_survey.Rmd x



+c Insert



```
141 **AVERAGE LISTING PRICE - NATIONAL**
142
143 ``{r}
144 # Average housing listing price in U.S.
145 # Earliest start date: July 2016
146 national_price <- fredr('AVELISPRIUS',
147                         observation_start = as.Date("2016-07-01"),
148                         frequency = "m")
149
150 national_price$date <- as.Date(national_price$date)
151
152 head(national_price)
````
```

|  | date       | series_id   | value  |
|--|------------|-------------|--------|
|  | <date>     | <chr>       | <dbl>  |
|  | 2016-07-01 | AVELISPRIUS | 444350 |
|  | 2016-08-01 | AVELISPRIUS | 440166 |
|  | 2016-09-01 | AVELISPRIUS | 440295 |
|  | 2016-10-01 | AVELISPRIUS | 444839 |
|  | 2016-11-01 | AVELISPRIUS | 448456 |
|  | 2016-12-01 | AVELISPRIUS | 451075 |

6 rows

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



Go to file/function

Addins

REDA.Rmd x merged\_price x boston\_supply x pulse\_household\_survey.Rmd x

ABC Knit Run Insert

```
141 **AVERAGE LISTING PRICE - NATIONAL**
142
143 ``{r}
144 # Average housing listing price in U.S.
145 # Earliest start date: July 2016
146 national_price <- fredr('AVELISPRIUS',
147 observation_start = as.Date("2016-07-01"),
148 frequency = "m")
149
150 national_price$date <- as.Date(national_price$date)
151
152 head(national_price)
````
```

| | date | series_id | value |
|--|------------|-------------|--------|
| | <date> | <chr> | <dbl> |
| | 2016-07-01 | AVELISPRIUS | 444350 |
| | 2016-08-01 | AVELISPRIUS | 440166 |
| | 2016-09-01 | AVELISPRIUS | 440295 |
| | 2016-10-01 | AVELISPRIUS | 444839 |
| | 2016-11-01 | AVELISPRIUS | 448456 |
| | 2016-12-01 | AVELISPRIUS | 451075 |

6 rows

Joining :

- Merging the two data frames

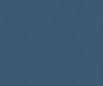
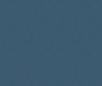
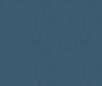
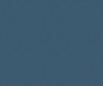
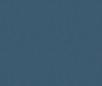


Joining





Go to file/function



REDA.Rmd

merged_price

boston_supply

pulse_household_survey.Rmd

```
154  
155 **AVERAGE LISTING PRICE - JOIN**  
156  
157 ```{r}  
158 merged_price <- full_join(boston_price, national_price)  
159 head(merged_price)  
160 ````
```

Joining, by = c("date", "series_id", "value")

R Consoletbl_df
6 x 3

date series_id value
2016-07-01 AVELISPRI14460 764864
2016-08-01 AVELISPRI14460 755404
2016-09-01 AVELISPRI14460 759052
2016-10-01 AVELISPRI14460 772755
2016-11-01 AVELISPRI14460 778844
2016-12-01 AVELISPRI14460 795947

| date | series_id | value |
|------------|----------------|--------|
| <date> | <chr> | <dbl> |
| 2016-07-01 | AVELISPRI14460 | 764864 |
| 2016-08-01 | AVELISPRI14460 | 755404 |
| 2016-09-01 | AVELISPRI14460 | 759052 |
| 2016-10-01 | AVELISPRI14460 | 772755 |
| 2016-11-01 | AVELISPRI14460 | 778844 |
| 2016-12-01 | AVELISPRI14460 | 795947 |

6 rows



REDA.Rmd

merged_price

boston_supply

pulse_household_survey.Rmd

```
154  
155 **AVERAGE LISTING PRICE - JOIN**  
156  
157 ````{r}  
158 merged_price <- full_join(boston_price, national_price)  
159 head(merged_price)  
160 ````
```

Joining, by = c("date", "series_id", "value")
R Console

date series_id value
2016-07-01 AVELISPRI14460 764864
2016-08-01 AVELISPRI14460 755404
2016-09-01 AVELISPRI14460 759052
2016-10-01 AVELISPRI14460 772755
2016-11-01 AVELISPRI14460 778844
2016-12-01 AVELISPRI14460 795947
tbl_df
6 x 3

| date | series_id | value |
|------------|----------------|--------|
| <date> | <chr> | <dbl> |
| 2016-07-01 | AVELISPRI14460 | 764864 |
| 2016-08-01 | AVELISPRI14460 | 755404 |
| 2016-09-01 | AVELISPRI14460 | 759052 |
| 2016-10-01 | AVELISPRI14460 | 772755 |
| 2016-11-01 | AVELISPRI14460 | 778844 |
| 2016-12-01 | AVELISPRI14460 | 795947 |

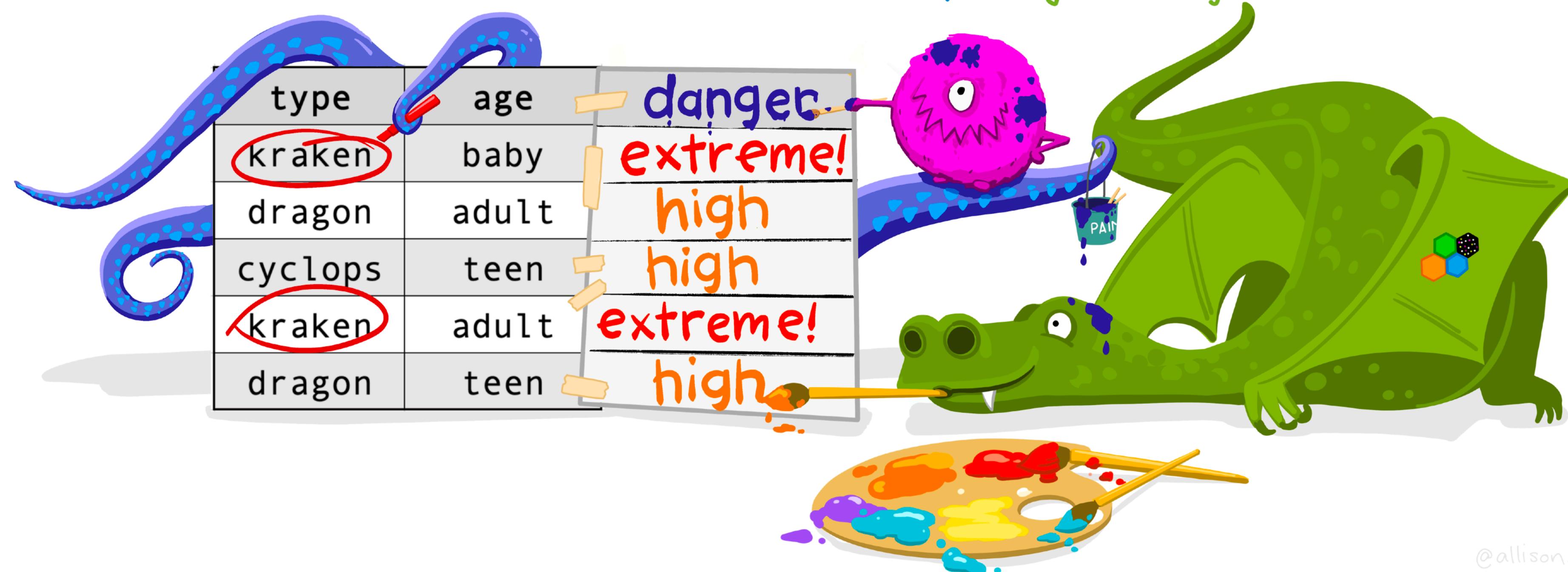
6 rows

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



Go to file/function



Addins

REDA.Rmd x

merged_price x

boston_supply x

pulse_household_survey.Rmd x



```
172  
173 **AVERAGE LISTING PRICE - VARIABLES**  
174  
175 ````{r}  
176 # Create dummy variable for pre/during pandemic  
177 # 1 represents during pandemic  
178 # 0 represents before  
179 merged_price <- merged_price %>%  
180   mutate(during_pandemic = case_when(date >= "2020-02-01" ~ 1,  
181                     date < "2020-02-01" ~ 0))  
182 head(merged_price)  
183 ````
```

| date | series_id | value | during_pandemic |
|------------|----------------|--------|-----------------|
| <date> | <chr> | <dbl> | <dbl> |
| 2016-07-01 | AVELISPRI14460 | 764864 | 0 |
| 2016-08-01 | AVELISPRI14460 | 755404 | 0 |
| 2016-09-01 | AVELISPRI14460 | 759052 | 0 |
| 2016-10-01 | AVELISPRI14460 | 772755 | 0 |
| 2016-11-01 | AVELISPRI14460 | 778844 | 0 |
| 2016-12-01 | AVELISPRI14460 | 795947 | 0 |

6 rows

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



REDA.Rmd x merged_price x boston_supply x pulse_household_survey.Rmd x



```
172  
173 **AVERAGE LISTING PRICE - VARIABLES**  
174  
175 ````{r}  
176 # Create dummy variable for pre/during pandemic  
177 # 1 represents during pandemic  
178 # 0 represents before  
179 merged_price <- merged_price %>%  
180 mutate(during_pandemic = case_when(date >= "2020-02-01" ~ 1,  
181 date < "2020-02-01" ~ 0))  
182 head(merged_price)  
183
```

| date | series_id | value | during_pandemic |
|------------|----------------|--------|-----------------|
| <date> | <chr> | <dbl> | <dbl> |
| 2016-07-01 | AVELISPRI14460 | 764864 | 0 |
| 2016-08-01 | AVELISPRI14460 | 755404 | 0 |
| 2016-09-01 | AVELISPRI14460 | 759052 | 0 |
| 2016-10-01 | AVELISPRI14460 | 772755 | 0 |
| 2016-11-01 | AVELISPRI14460 | 778844 | 0 |
| 2016-12-01 | AVELISPRI14460 | 795947 | 0 |

6 rows

REDA.Rmd | merged_price | boston_supply | pulse_household_survey.Rmd

Go to file/function | Addins | Insert | Run | Knit | ABC | Search | Settings | Help | Editor | View | Tools | Help

```
161
162 **AVERAGE LISTING PRICE - TREND VARIABLE**
163
164 ````{r}
165 # Create new variable indicating geography
166 merged_price <- merged_price %>%
167   mutate(trend = case_when(series_id == "AVELISPRI14460" ~ "Boston",
168                           series_id == "AVELISPRIUS" ~ "National")) %>%
169   arrange(date)
170 head(merged_price)
171 ````
```

| date | series_id | value | during_pandemic | trend |
|------------|----------------|--------|-----------------|----------|
| <date> | <chr> | <dbl> | <dbl> | <chr> |
| 2016-07-01 | AVELISPRI14460 | 764864 | 0 | Boston |
| 2016-07-01 | AVELISPRIUS | 444350 | 0 | National |
| 2016-08-01 | AVELISPRI14460 | 755404 | 0 | Boston |
| 2016-08-01 | AVELISPRIUS | 440166 | 0 | National |
| 2016-09-01 | AVELISPRI14460 | 759052 | 0 | Boston |
| 2016-09-01 | AVELISPRIUS | 440295 | 0 | National |

6 rows

REDA.Rmd | merged_price | boston_supply | pulse_household_survey.Rmd

Go to file/function | Addins | Insert | Run | Knit | ABC | Settings | Help

```
161
162 **AVERAGE LISTING PRICE - TREND VARIABLE**
163
164 ````{r}
165 # Create new variable indicating geography
166 merged_price <- merged_price %>%
167   mutate(trend = case_when(
168     series_id == "AVELISPRI14460" ~ "Boston",
169     series_id == "AVELISPRIUS" ~ "National")) %>%
170   arrange(date)
171 head(merged_price)
172 ````
```

| date | series_id | value | during_pandemic | trend |
|------------|----------------|--------|-----------------|----------|
| <date> | <chr> | <dbl> | <dbl> | <chr> |
| 2016-07-01 | AVELISPRI14460 | 764864 | 0 | Boston |
| 2016-07-01 | AVELISPRIUS | 444350 | 0 | National |
| 2016-08-01 | AVELISPRI14460 | 755404 | 0 | Boston |
| 2016-08-01 | AVELISPRIUS | 440166 | 0 | National |
| 2016-09-01 | AVELISPRI14460 | 759052 | 0 | Boston |
| 2016-09-01 | AVELISPRIUS | 440295 | 0 | National |

6 rows

Variables :

- Count



Variables



REDA.Rmd | merged_price | boston_supply | pulse_household_survey.Rmd

ABC Knit Insert Run

```
185 **AVERAGE LISTING PRICE - COUNT**
186
187 ````{r}
188 # Create 'count' dataframe
189 merged_count <- merged_price %>%
  count(trend, during_pandemic) %>%
  rename(count = n) %>%
  mutate(total = sum(count),
         share = (count/total)*100)
190 head(merged_count)
191
192 ````
```

| trend
<chr> | during_pandemic
<dbl> | count
<int> | total
<int> | share
<dbl> |
|----------------|--------------------------|----------------|----------------|----------------|
| Boston | 0 | 43 | 102 | 42.156863 |
| Boston | 1 | 8 | 102 | 7.843137 |
| National | 0 | 43 | 102 | 42.156863 |
| National | 1 | 8 | 102 | 7.843137 |

4 rows

REDA.Rmd | merged_price | boston_supply | pulse_household_survey.Rmd

ABC Knit Insert Run

```
185 **AVERAGE LISTING PRICE - COUNT**
186
187 ````{r}
188 # Create 'count' dataframe
189 merged_count <- merged_price %>%
190   count(trend, during_pandemic) %>%
191   rename(count = n) %>%
192   mutate(total = sum(count),
193         share = (count/total)*100)
194 head(merged_count)
195 ````
```

| trend
<chr> | during_pandemic
<dbl> | count
<int> | total
<int> | share
<dbl> |
|----------------|--------------------------|----------------|----------------|----------------|
| Boston | 0 | 43 | 102 | 42.156863 |
| Boston | 1 | 8 | 102 | 7.843137 |
| National | 0 | 43 | 102 | 42.156863 |
| National | 1 | 8 | 102 | 7.843137 |

4 rows

REDA.Rmd | merged_price | boston_supply | pulse_household_survey.Rmd

ABC Knit Insert Run

```
185 **AVERAGE LISTING PRICE - COUNT**
186
187 ````{r}
188 # Create 'count' dataframe
189 merged_count <- merged_price %>%
190   count(trend, during_pandemic) %>%
191   rename(count = n) %>%
192   mutate(total = sum(count),
193         share = (count/total)*100)
194 head(merged_count)
195 ````
```

| trend
<chr> | during_pandemic
<dbl> | count
<int> | total
<int> | share
<dbl> |
|----------------|--------------------------|----------------|----------------|----------------|
| Boston | 0 | 43 | 102 | 42.156863 |
| Boston | 1 | 8 | 102 | 7.843137 |
| National | 0 | 43 | 102 | 42.156863 |
| National | 1 | 8 | 102 | 7.843137 |

4 rows

REDA.Rmd | merged_price | boston_supply | pulse_household_survey.Rmd

ABC Knit Insert Run

```
185 **AVERAGE LISTING PRICE - COUNT**
186
187 ````{r}
188 # Create 'count' dataframe
189 merged_count <- merged_price %>%
  count(trend, during_pandemic) %>%
  rename(count = n) %>%
  mutate(total = sum(count),
         share = (count/total)*100)
194 head(merged_count)
195
```

| trend
<chr> | during_pandemic
<dbl> | count
<int> | total
<int> | share
<dbl> |
|----------------|--------------------------|----------------|----------------|----------------|
| Boston | 0 | 43 | 102 | 42.156863 |
| Boston | 1 | 8 | 102 | 7.843137 |
| National | 0 | 43 | 102 | 42.156863 |
| National | 1 | 8 | 102 | 7.843137 |

4 rows

Variables :

- Descriptive
statistics

Variables





Go to file/function

Addins

REDA.Rmd x merged_price x boston_supply x pulse_household_survey.Rmd >> □

◀ ▶ ABC Knit □ Insert □ Run □

```
198  
199 **AVERAGE LISTING PRICE - SUMMARY**  
200  
201 ``{r}  
202 # Create 'summary' dataframe  
203 # includes mean, median, standard deviation, and coefficient of variation  
204 merged_summary <- merged_price %>%  
205   group_by(trend, during_pandemic) %>%  
206   summarise(mean_value = mean(value),  
207             median_value = median(value),  
208             sd_value = sd(value),  
209             cv_value = (sd_value/mean_value)*100)  
210 head(merged_summary)  
````
```

'summarise()' regrouping output by 'trend'  
(override with '.groups' argument)

R Console

trend during\_pandemic mean\_value median\_value sd\_value cv\_value  
Boston 0 845951.5 847356.0 49106.71 5.804908  
Boston 1 992156.9 985821.5 38023.08 3.832366  
National 0 496961.9 499572.0 31711.00 6.380973  
National 1 593160.8 583615.0 35199.00 5.934142

grouped\_df  
4 x 6

| trend<br><chr> | during_pandemic<br><dbl> | mean_value<br><dbl> | median_value<br><dbl> | sd_value<br><dbl> | cv_value<br><dbl> |
|----------------|--------------------------|---------------------|-----------------------|-------------------|-------------------|
| Boston         | 0                        | 845951.5            | 847356.0              | 49106.71          | 5.804908          |
| Boston         | 1                        | 992156.9            | 985821.5              | 38023.08          | 3.832366          |
| National       | 0                        | 496961.9            | 499572.0              | 31711.00          | 6.380973          |
| National       | 1                        | 593160.8            | 583615.0              | 35199.00          | 5.934142          |

4 rows



Go to file/function

Addins

REDA.Rmd

merged\_price

boston\_supply

pulse\_household\_survey.Rmd

```
198
199 **AVERAGE LISTING PRICE - SUMMARY**
200
201 ```{r}
202 # Create 'summary' dataframe
203 # includes mean, median, standard deviation, and coefficient of variation
204 merged_summary <- merged_price %>%
205 group_by(trend, during_pandemic) %>%
206 summarise(mean_value = mean(value),
207 median_value = median(value),
208 sd_value = sd(value),
209 cv_value = (sd_value/mean_value)*100)
210 head(merged_summary)
211 ```


```

'summarise()' regrouping output by 'trend'  
(override with '.groups' argument)

R Console

trend during\_pandemic mean\_value median\_value sd\_value cv\_value  
Boston 0 845951.5 847356.0 49106.71 5.804908  
Boston 1 992156.9 985821.5 38023.08 3.832366  
National 0 496961.9 499572.0 31711.00 6.380973  
National 1 593160.8 583615.0 35199.00 5.934142

grouped\_df  
4 x 6

| trend<br><chr> | during_pandemic<br><dbl> | mean_value<br><dbl> | median_value<br><dbl> | sd_value<br><dbl> | cv_value<br><dbl> |
|----------------|--------------------------|---------------------|-----------------------|-------------------|-------------------|
| Boston         | 0                        | 845951.5            | 847356.0              | 49106.71          | 5.804908          |
| Boston         | 1                        | 992156.9            | 985821.5              | 38023.08          | 3.832366          |
| National       | 0                        | 496961.9            | 499572.0              | 31711.00          | 6.380973          |
| National       | 1                        | 593160.8            | 583615.0              | 35199.00          | 5.934142          |

4 rows



REDA.Rmd x merged\_price x boston\_supply x pulse\_household\_survey.Rmd >> □



```
198
199 **AVERAGE LISTING PRICE - SUMMARY**
200
201 ``{r}
202 # Create 'summary' dataframe
203 # includes mean, median, standard deviation, and coefficient of variation
204 merged_summary <- merged_price %>%
205 group_by(trend, during_pandemic) %>%
206 summarise(mean_value = mean(value),
207 median_value = median(value),
208 sd_value = sd(value),
209 cv_value = (sd_value/mean_value)*100)
210 }
```

```
summarise() regrouping output by 'trend'
(override with `...groups` argument)
```

R Console  
grouped\_df  
4 x 6

| trend    | during_pandemic | mean_value | median_value | sd_value | cv_value |
|----------|-----------------|------------|--------------|----------|----------|
| <chr>    | <dbl>           | <dbl>      | <dbl>        | <dbl>    | <dbl>    |
| Boston   | 0               | 845951.5   | 847356.0     | 49106.71 | 5.804908 |
| Boston   | 1               | 992156.9   | 985821.5     | 38023.08 | 3.832366 |
| National | 0               | 496961.9   | 499572.0     | 31711.00 | 6.380973 |
| National | 1               | 593160.8   | 583615.0     | 35199.00 | 5.934142 |

4 rows



Go to file/function

Addins

REDA.Rmd

merged\_price

boston\_supply

pulse\_household\_survey.Rmd

```
198
199 **AVERAGE LISTING PRICE - SUMMARY**
200
201 ````{r}
202 # Create 'summary' dataframe
203 # includes mean, median, standard deviation, and coefficient of variation
204 merged_summary <- merged_price %>%
205 group_by(trend, during_pandemic) %>%
206 summarise(mean_value = mean(value),
207 median_value = median(value),
208 sd_value = sd(value),
209 cv_value = (sd_value/mean_value)*100)
210 head(merged_summary)
211
```

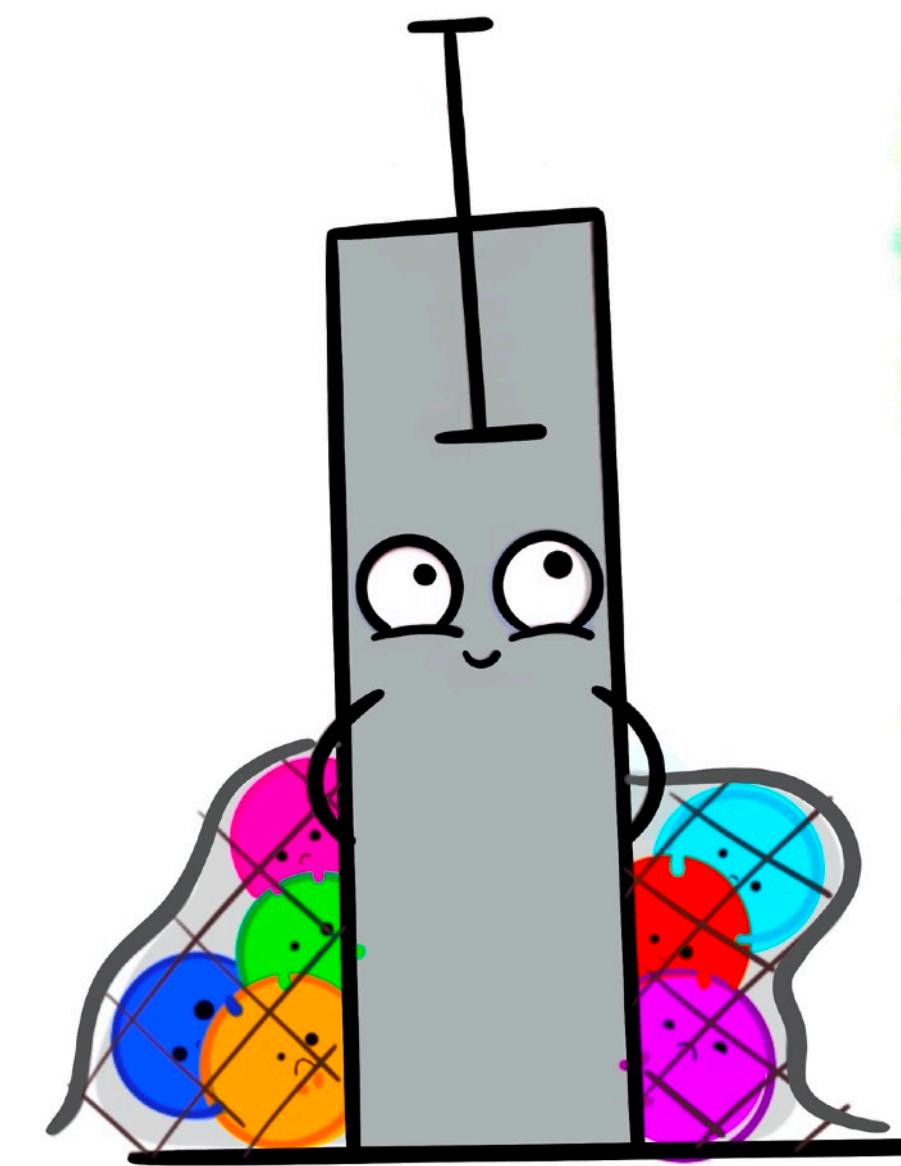
R Console

grouped\_df  
4 x 6

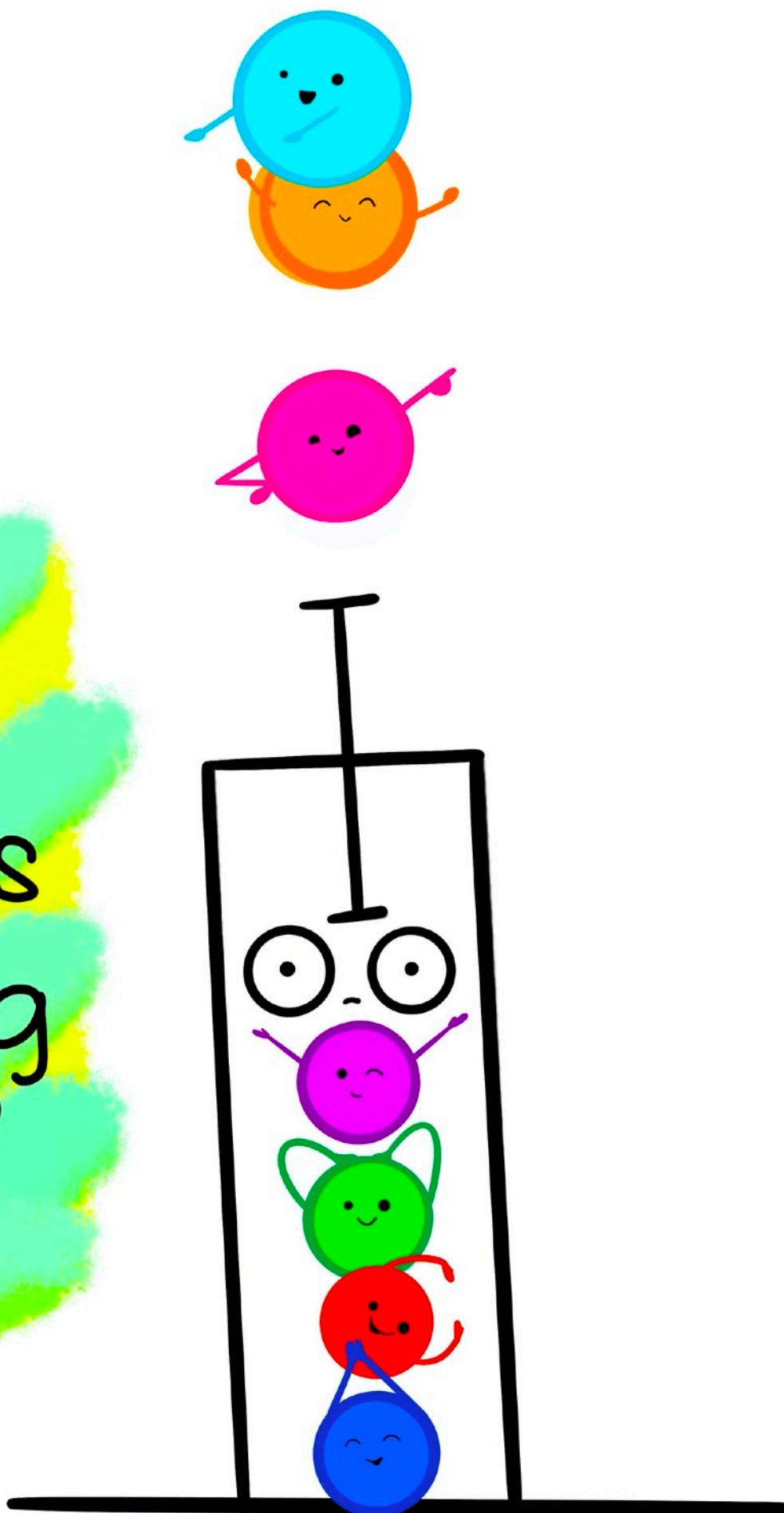
| trend<br><chr> | during_pandemic<br><dbl> | mean_value<br><dbl> | median_value<br><dbl> | sd_value<br><dbl> | cv_value<br><dbl> |
|----------------|--------------------------|---------------------|-----------------------|-------------------|-------------------|
| Boston         | 0                        | 845951.5            | 847356.0              | 49106.71          | 5.804908          |
| Boston         | 1                        | 992156.9            | 985821.5              | 38023.08          | 3.832366          |
| National       | 0                        | 496961.9            | 499572.0              | 31711.00          | 6.380973          |
| National       | 1                        | 593160.8            | 583615.0              | 35199.00          | 5.934142          |

4 rows





are your  
summary statistics  
hiding something  
interesting?



@allison\_horst





# ggplot2:

Build a data  
**MASTERpiece**

