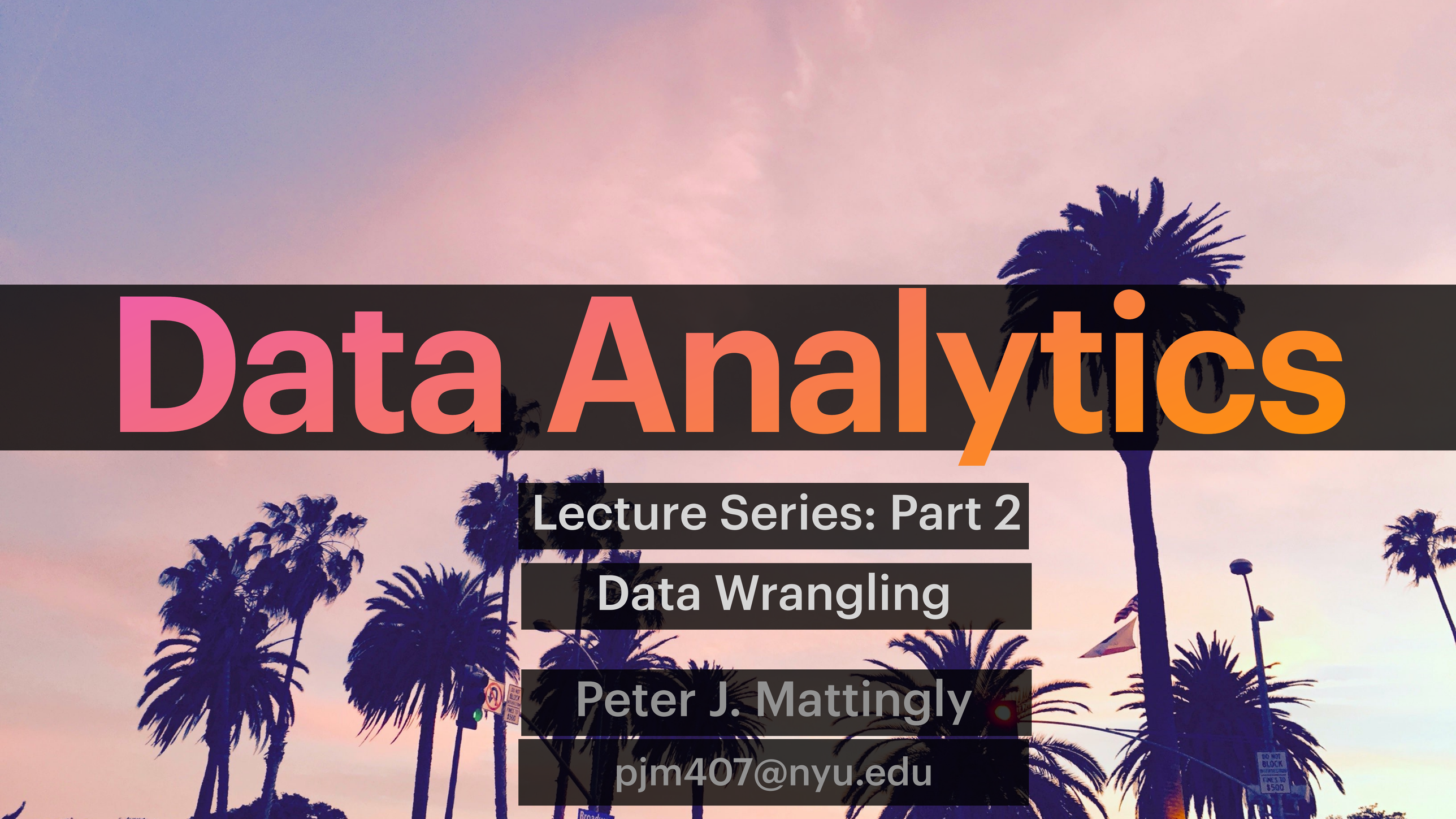


Data Analytics

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

Variables :

- Categorical


NOMINAL
UNORDERED DESCRIPTIONS
I'm a TURTLE!
i'm a snail!
I'm a butterfly!
ORDINAL
ORDERED DESCRIPTIONS
I am unhappy.
I am O.K.
I am Awesome!!!
BINARY
ONLY 2 MUTUALLY EXCLUSIVE OUTCOMES
I am EXTINCT!
HA.
@allison_horst

Variables

Variables :

- Categorical
- Numeric

CONTINUOUS

measured data, can have ∞ values within possible range.

I AM 3.1" TALL

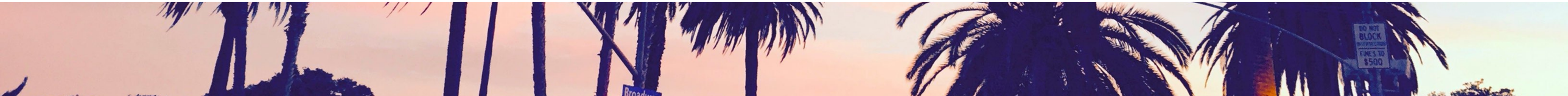
I WEIGH 34.16 grams

DISCRETE

OBSERVATIONS CAN ONLY EXIST AT LIMITED VALUES, OFTEN COUNTS.

I HAVE 8 ARMS and 4 SPOTS!

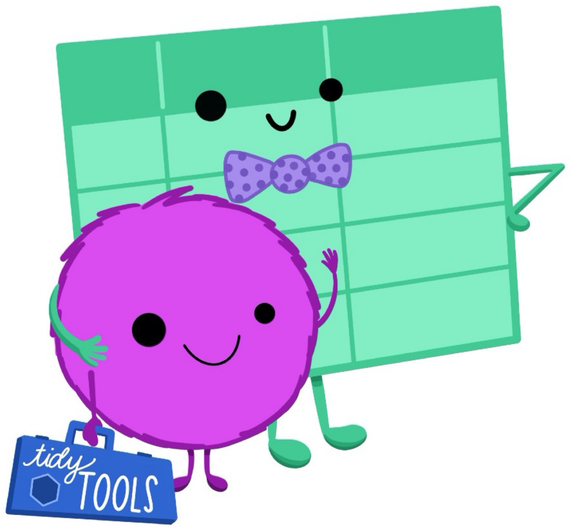
@allison_horst

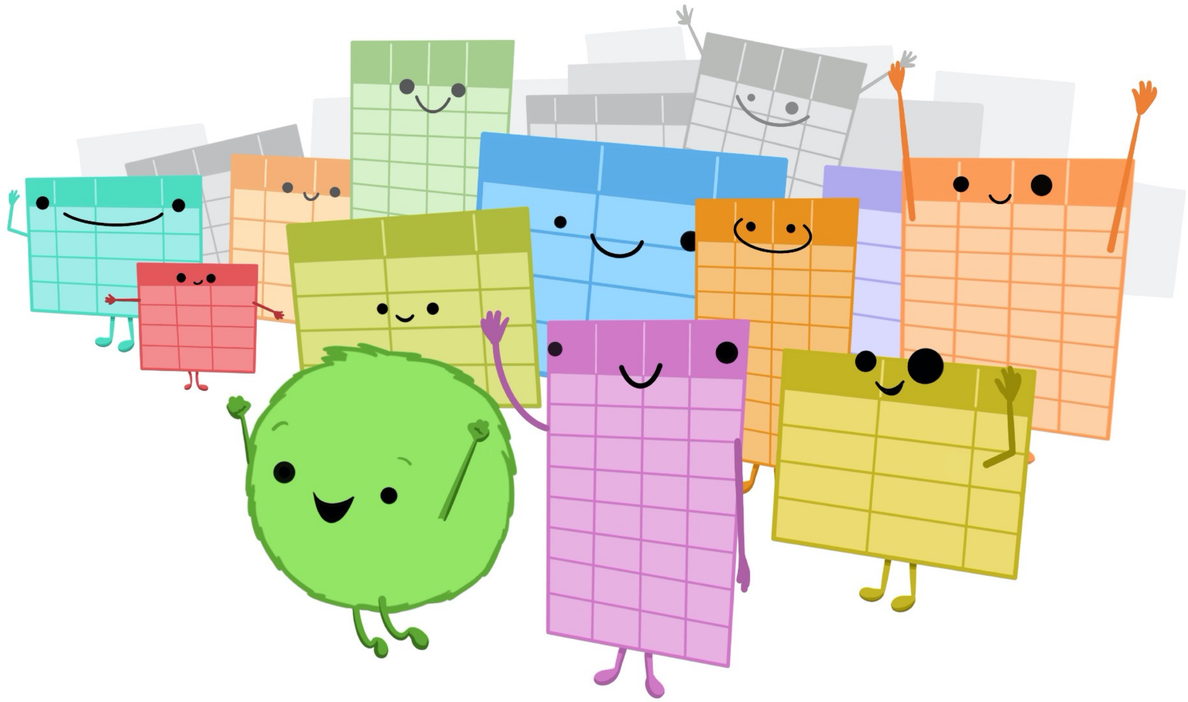


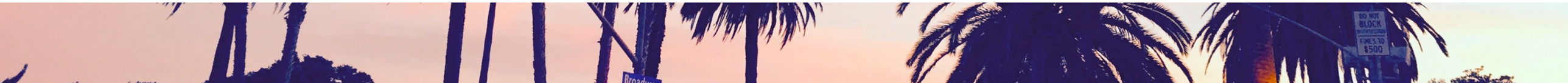
Variables

Variables :

- Categorical
- Numeric
- Dates


tidy
TOOLS




DO NOT
BLOCK
FINES TO
$500


R Studio®
DO NOT BLOCK
FINES TO $500

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', ther

```{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')
```

Project: (None)

Environment | History | Connections

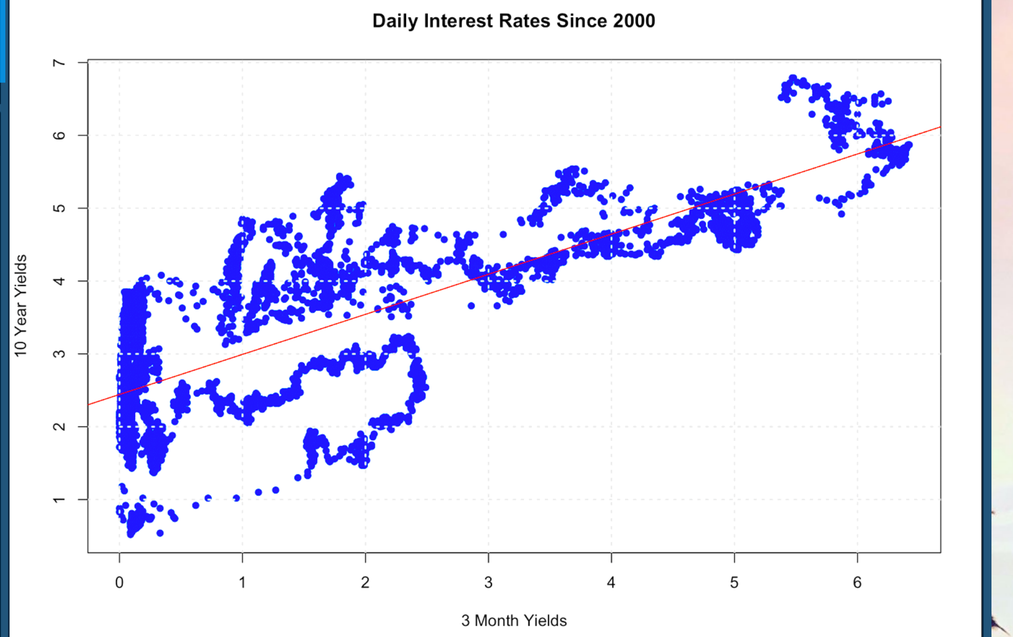
Import Dataset | Grid

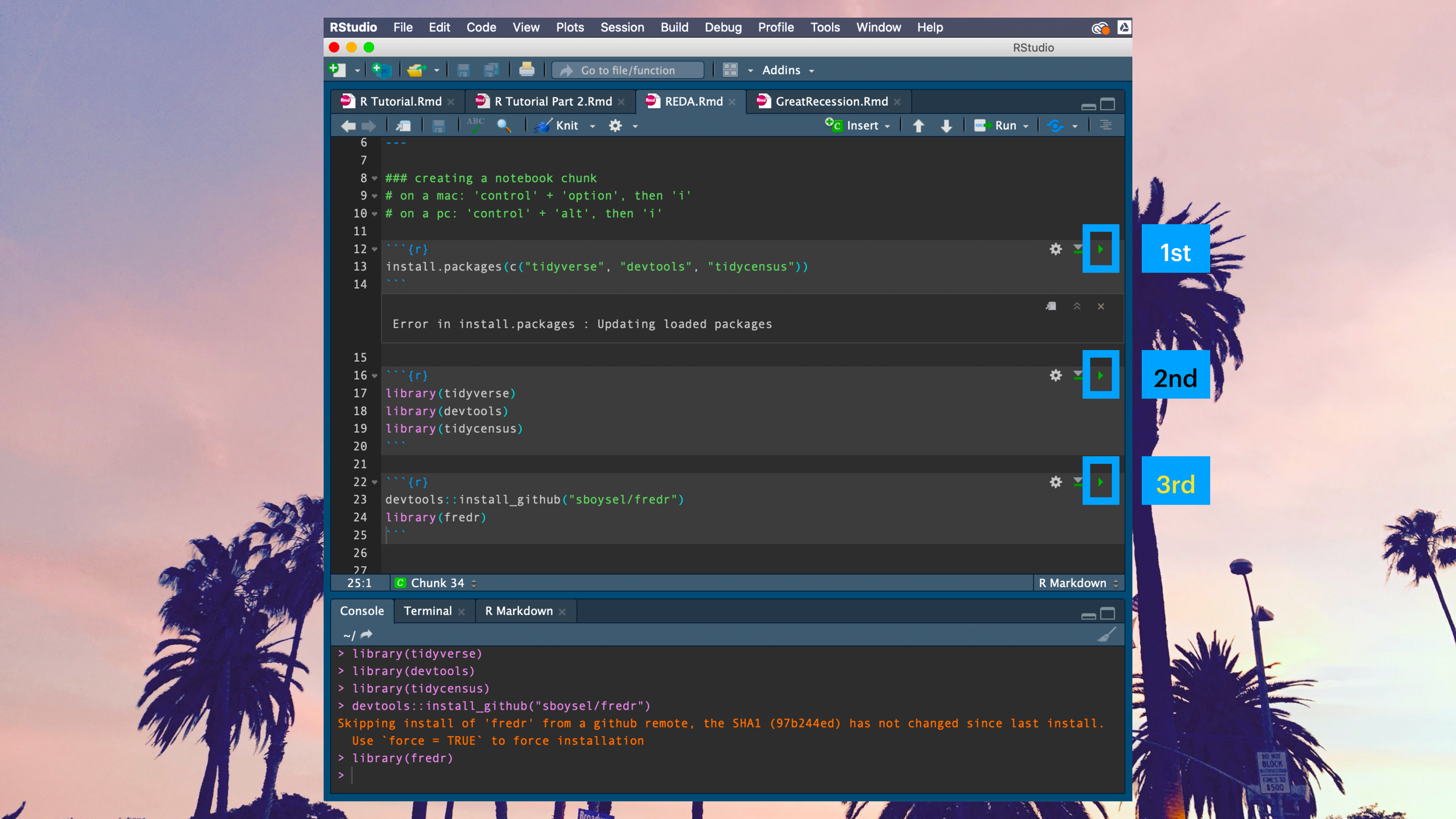
Global Environment

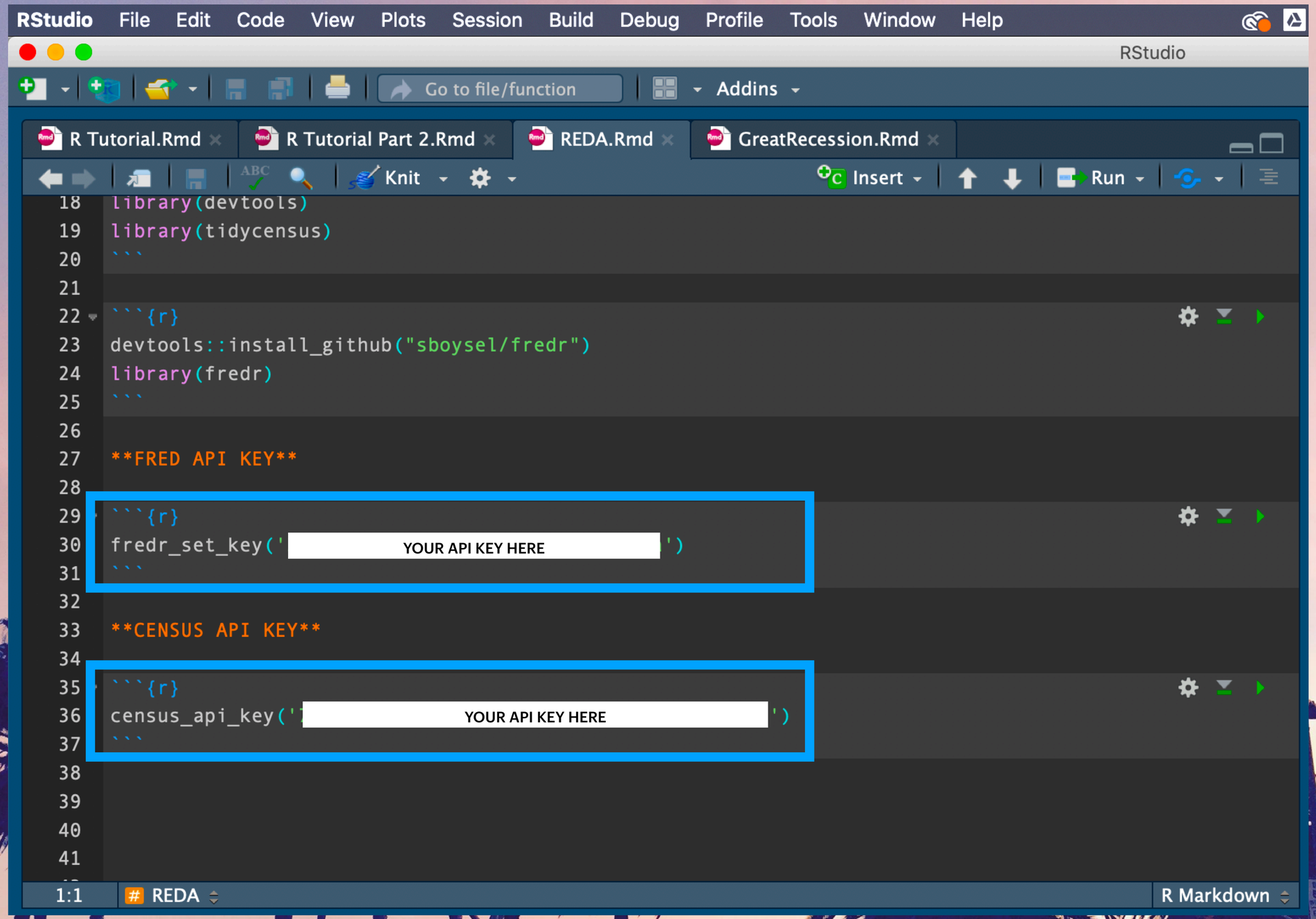
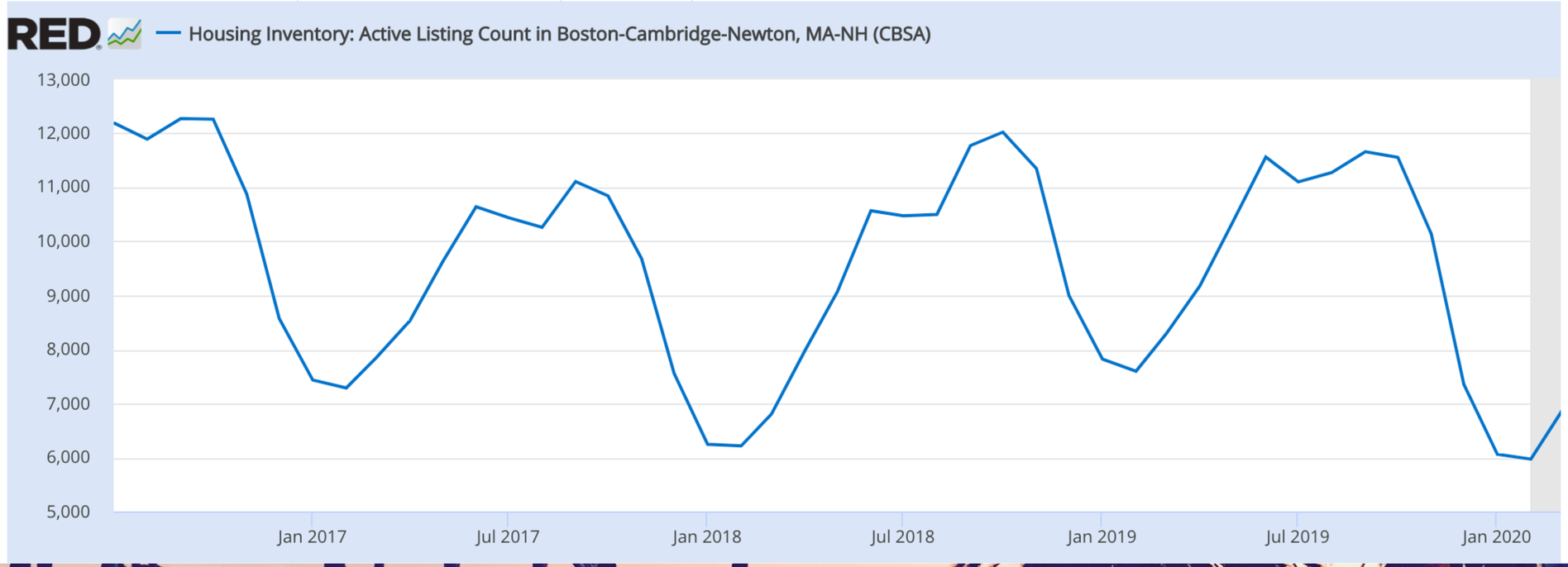
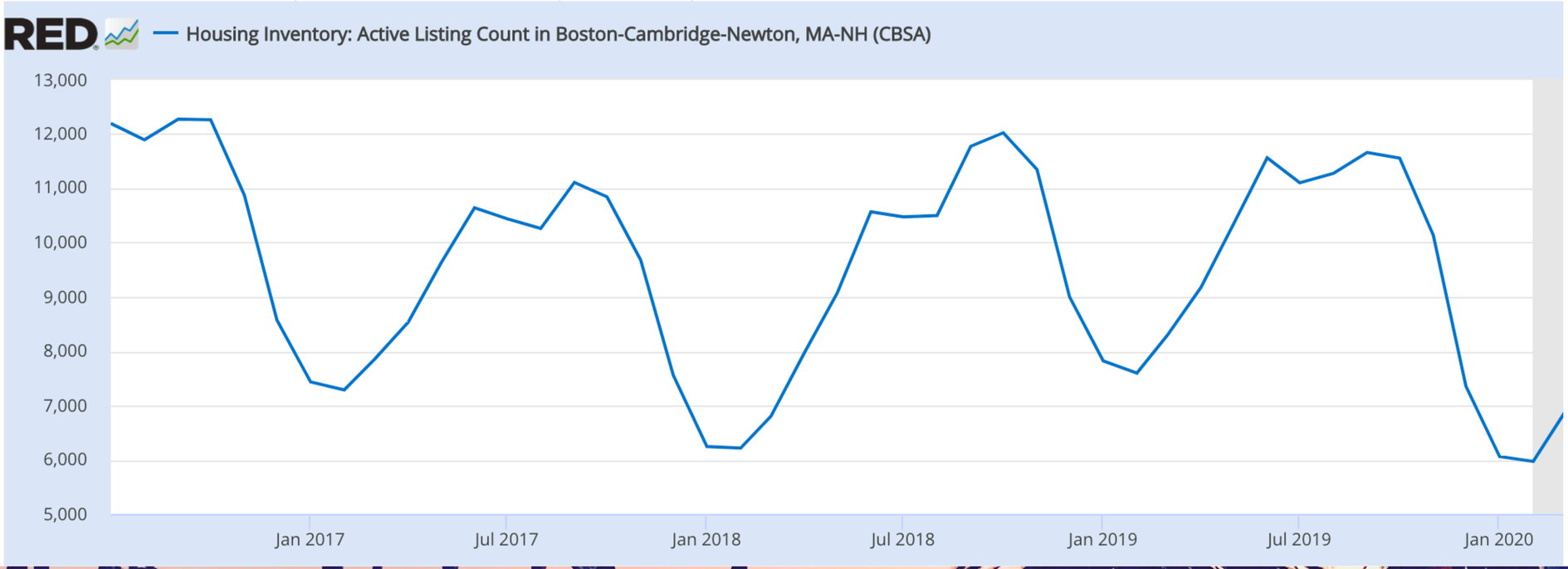
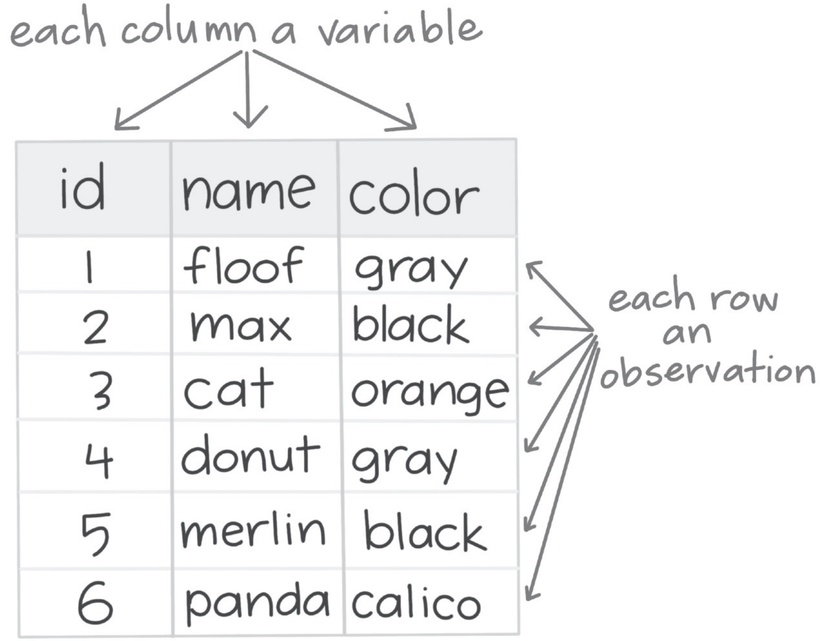
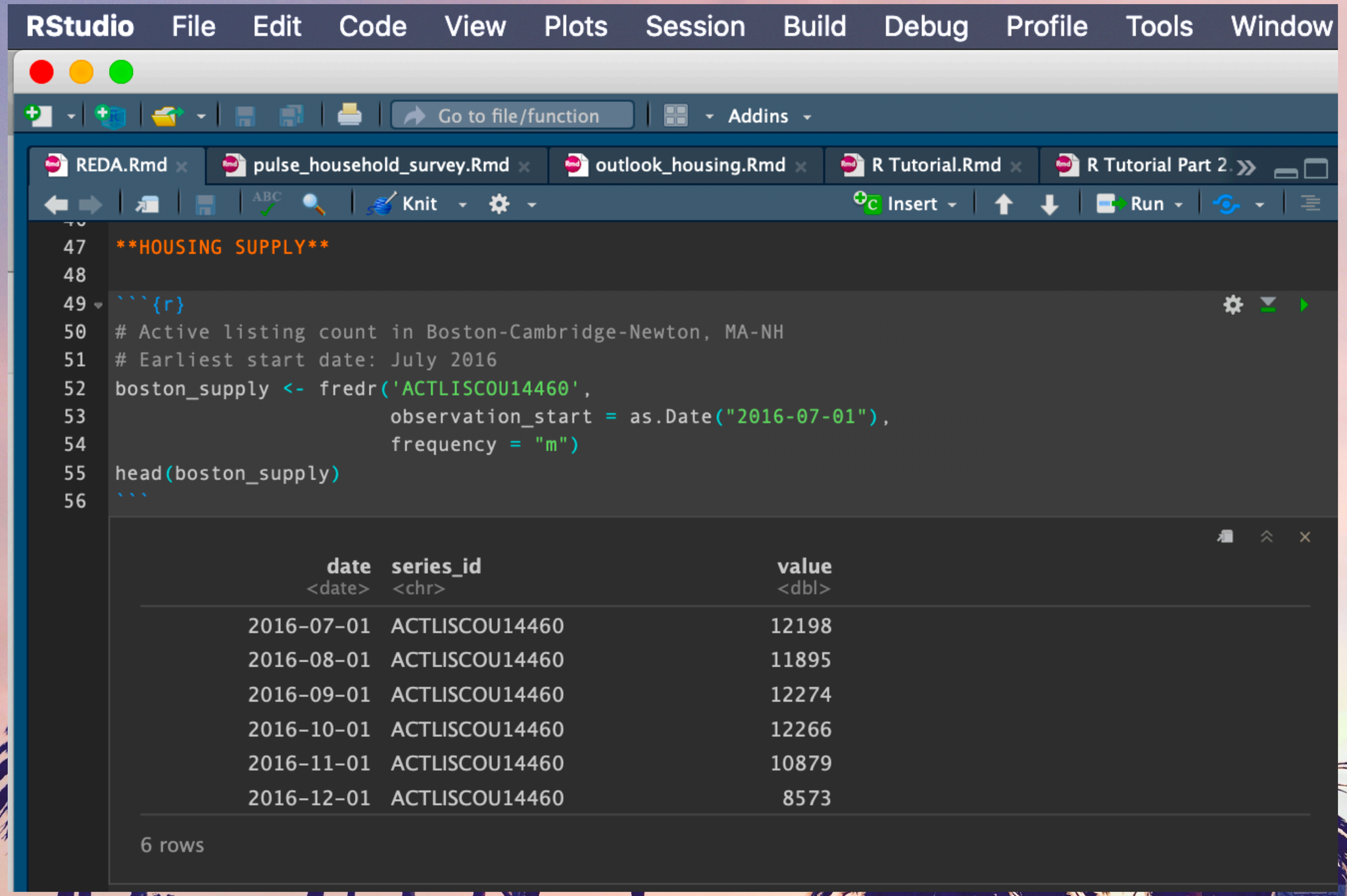
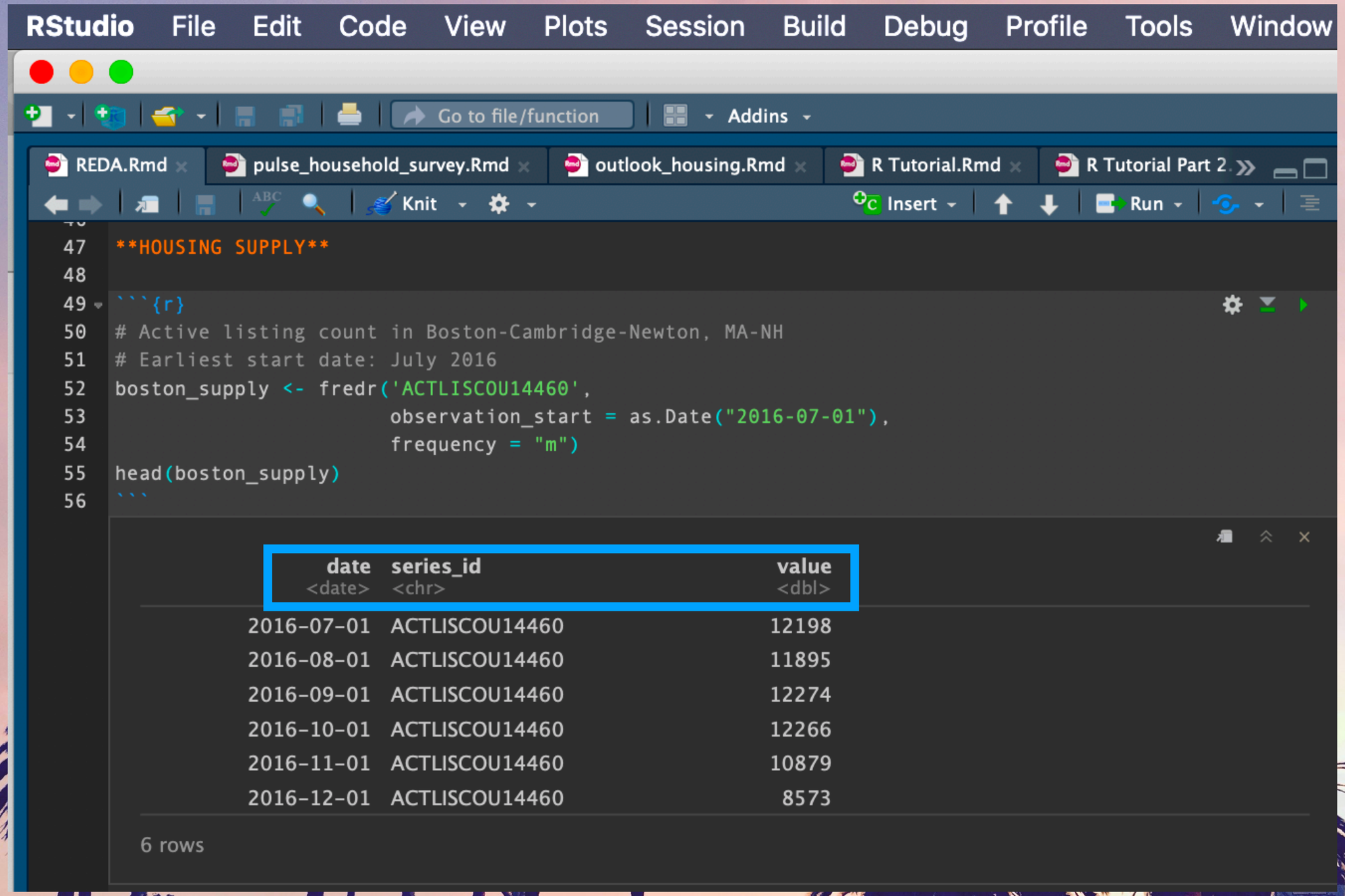
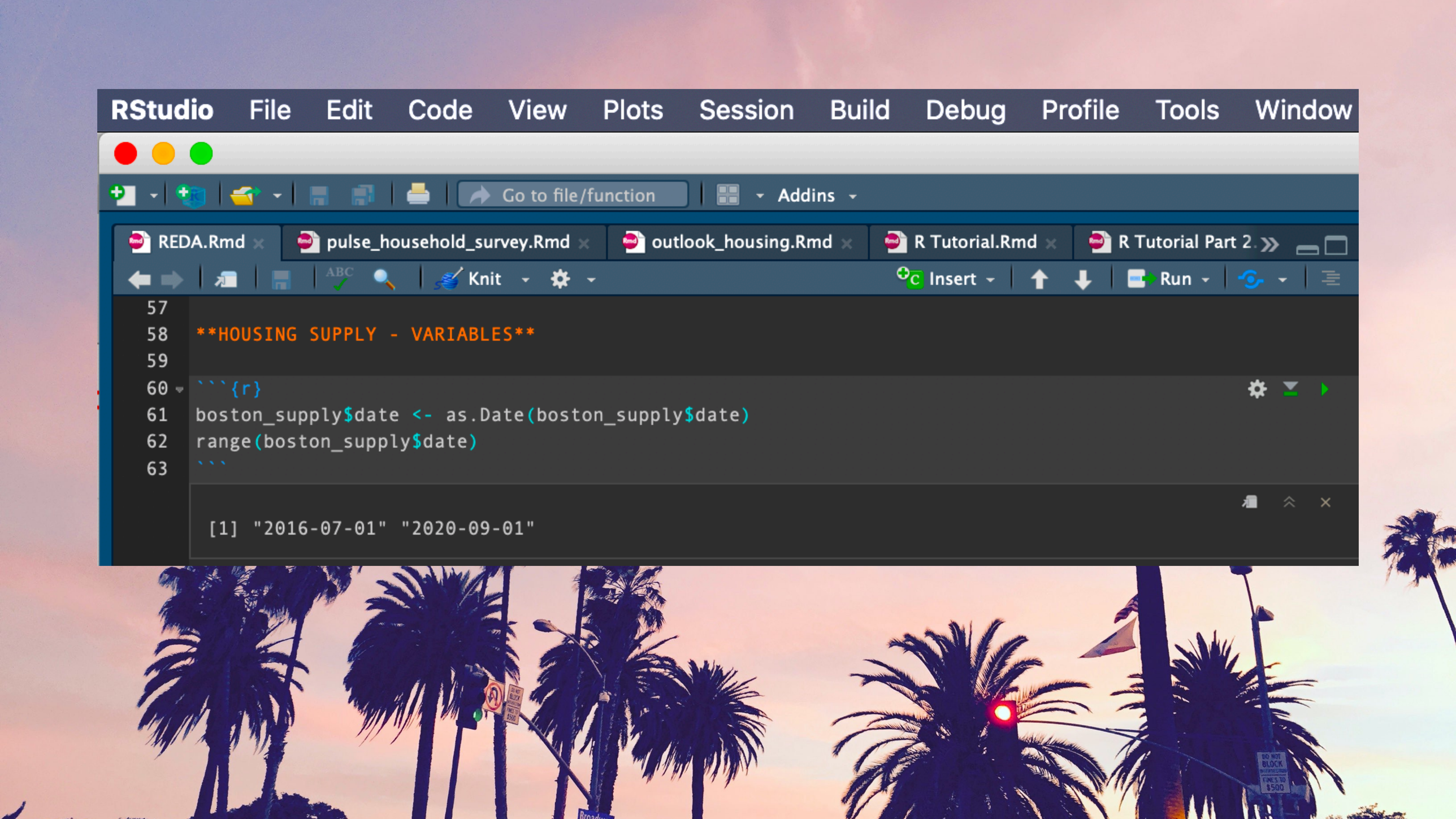
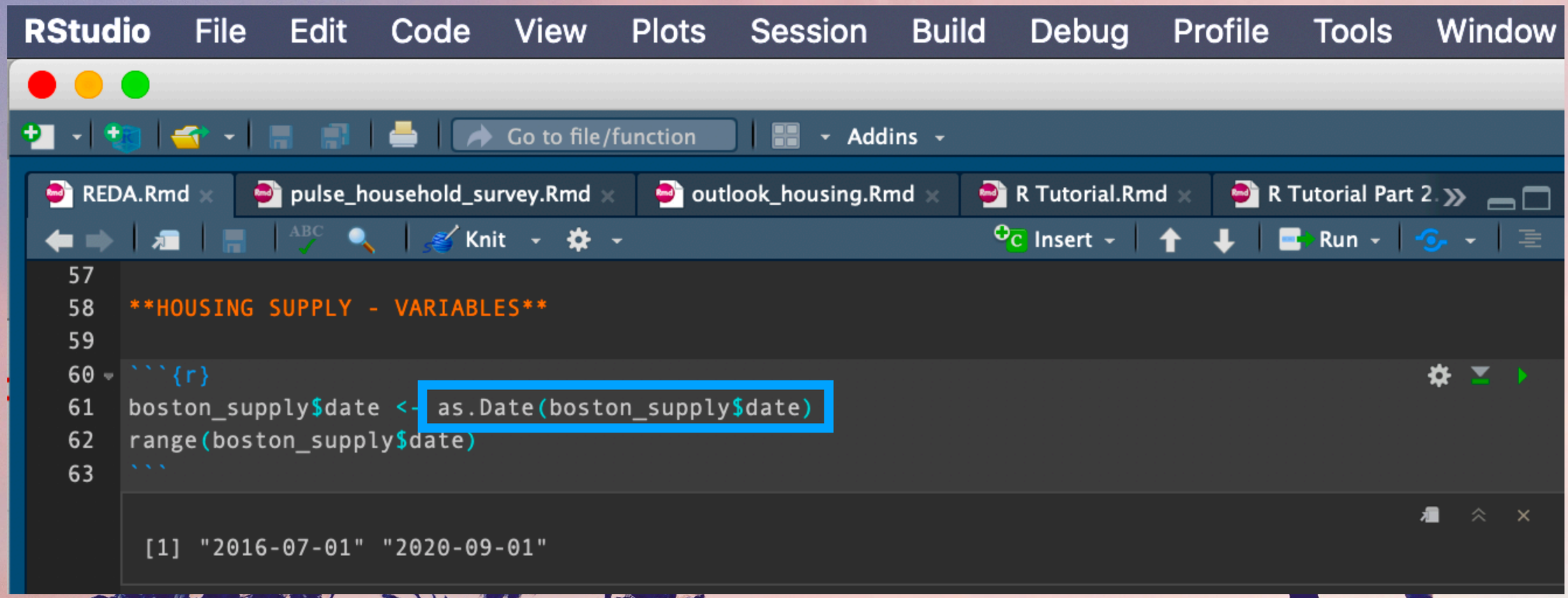
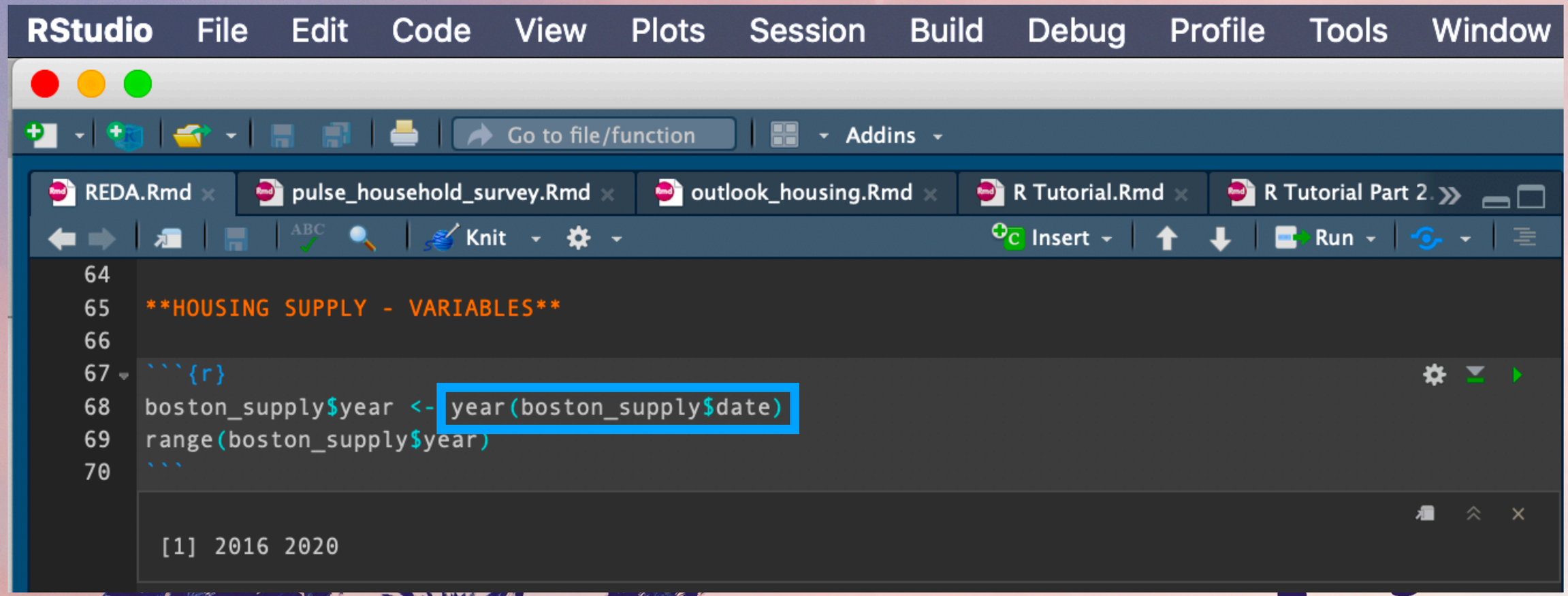
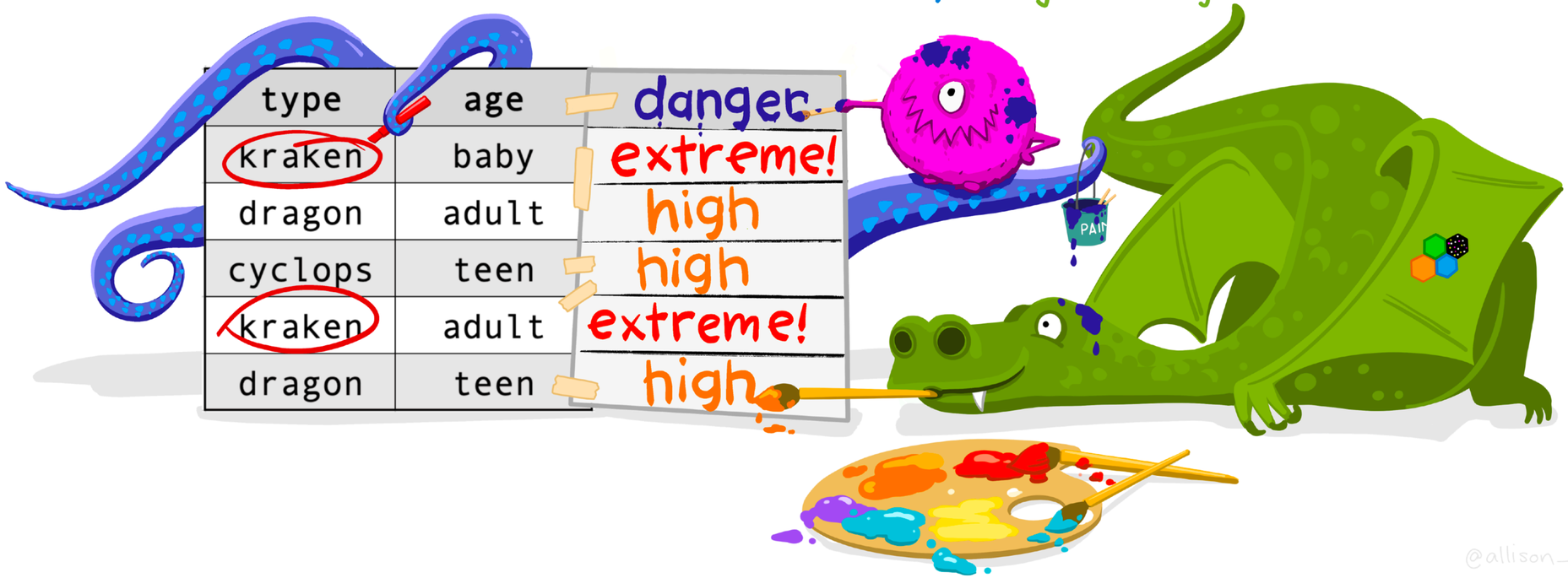
Name	Type	Length	Size	Value
dailyavg_table	tbl_df	7	2 KB	3 obs. of 7 variables
dailyavg_wtmeans	grouped_df	4	66.4 KB	1095 obs. of 4 variables
data1990	tbl_df	6	22 KB	373 obs. of 6 variables
data1990_2018_race_total	data.frame	5	8.7 KB	174 obs. of 5 variables
data1990_hisp	tbl_df	6	7 KB	62 obs. of 6 variables
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




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
Knit
Insert
Run
---
creating a notebook chunk
on a mac: 'control' + 'option', then 'i'
on a pc: 'control' + 'alt', then 'i'
```{r}
install.packages(c("tidyverse", "devtools", "tidycensus"))
```
Error in install.packages : Updating loaded packages
```{r}
library(tidyverse)
library(devtools)
library(tidycensus)
```
```{r}
devtools::install_github("sboysel/fredr")
library(fredr)
```
25:1
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)
>
1st
2nd
3rd

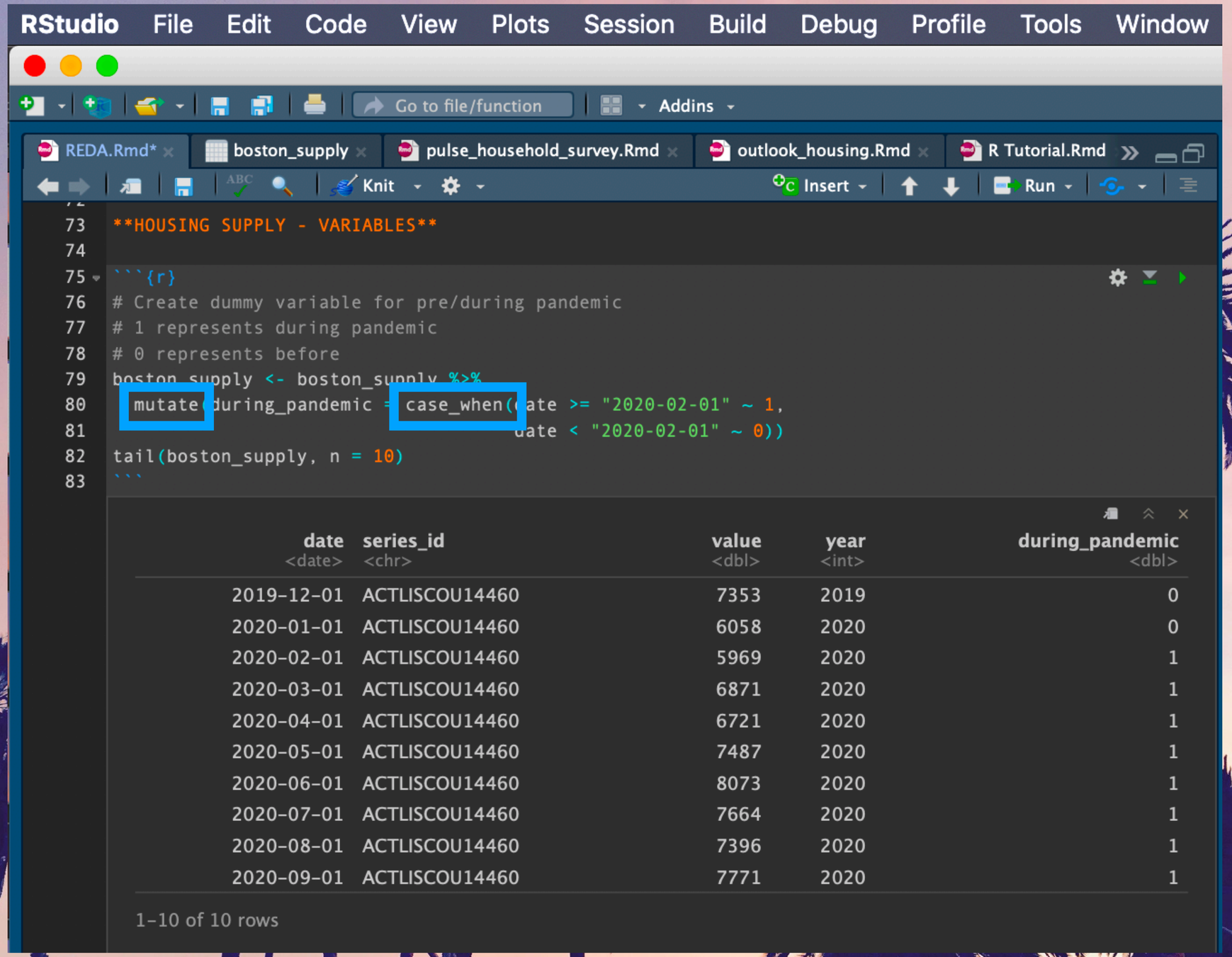

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
RStudio
Go to file/function
Addins
R Tutorial.Rmd
R Tutorial Part 2.Rmd
REDA.Rmd
GreatRecession.Rmd
Knit
Insert
Run
library(devtools)
library(tidycensus)
```
```{r}
devtools::install_github("sboysel/fredr")
library(fredr)
```
FRED API KEY
```{r}
fredr_set_key('YOUR API KEY HERE')
```
CENSUS API KEY
```{r}
census_api_key('YOUR API KEY HERE')
```
1:1
REDA
R Markdown
```

fred.stlouisfed.org/series/ACTLISCOU14460

FRED ECONOMIC DATA | ST. LOUIS FED | 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



fred.stlouisfed.org/series/ACTLISCOU14460

FRED ECONOMIC DATA | ST. LOUIS FED | 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



"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



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


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window
Go to file/function
Addins
REDA.Rmd
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
R Tutorial Part 2.
Knit
Insert
Run
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 <date> series_id <chr> value <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

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

Go to file/function | Addins

REDA.Rmd | pulse_household_survey.Rmd | outlook_housing.Rmd | R Tutorial.Rmd | R Tutorial Part 2

Knit | Insert | Run

```
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 <date>	series_id <chr>	value <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


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window
Go to file/function
Addins
REDA.Rmd
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
R Tutorial Part 2.
Knit
Insert
Run
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 <date> series_id <chr> value <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

Variables :

- Check the format


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window
Go to file/function
Addins
REDA.Rmd
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
R Tutorial Part 2
Knit
Insert
Run
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
Go to file/function
Addins
REDA.Rmd
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
R Tutorial Part 2
Knit
Insert
Run
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

Variables :

- Check the format
- Create alternate version


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window
Go to file/function
Addins
REDA.Rmd
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
R Tutorial Part 2
Knit
Insert
Run
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

Variables

Variables :

- Check the format
- Create alternate version
- New variable


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

dplyr::case_when()

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

```
df %>% ADD COLUMN 'danger'
  mutate(danger = case_when(type == "kraken" ~ "extreme!",
                            TRUE ~ "high"))
```

IF type is kraken THEN danger is extreme!

OTHERWISE, danger is high.



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

Go to file/function | Addins

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

Knit | Insert | Run

```
**HOUSING SUPPLY - VARIABLES**

```{r}
# Create dummy variable for pre/during pandemic
# 1 represents during pandemic
# 0 represents before
boston_supply <- boston_supply %>%
  mutate(during_pandemic = case_when(date >= "2020-02-01" ~ 1,
                                     date < "2020-02-01" ~ 0))
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

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

Go to file/function | Addins

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

Knit | Insert | Run

```
**HOUSING SUPPLY - VARIABLES**

```{r}
# Create dummy variable for pre/during pandemic
# 1 represents during pandemic
# 0 represents before
boston_supply <- boston_supply %>%
  mutate(during_pandemic = case_when(date >= "2020-02-01" ~ 1,
                                     date < "2020-02-01" ~ 0))
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


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window
Go to file/function
Addins
REDA.Rmd*
boston_supply
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
Knit
Insert
Run
HOUSING SUPPLY - VARIABLES
```{r}
# Create dummy variable for pre/during pandemic
# 1 represents during pandemic
# 0 represents before
boston_supply <- boston_supply %>%
mutate(during_pandemic = case_when(date >= "2020-02-01" ~ 1,
date < "2020-02-01" ~ 0))
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

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

Go to file/function | Addins

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

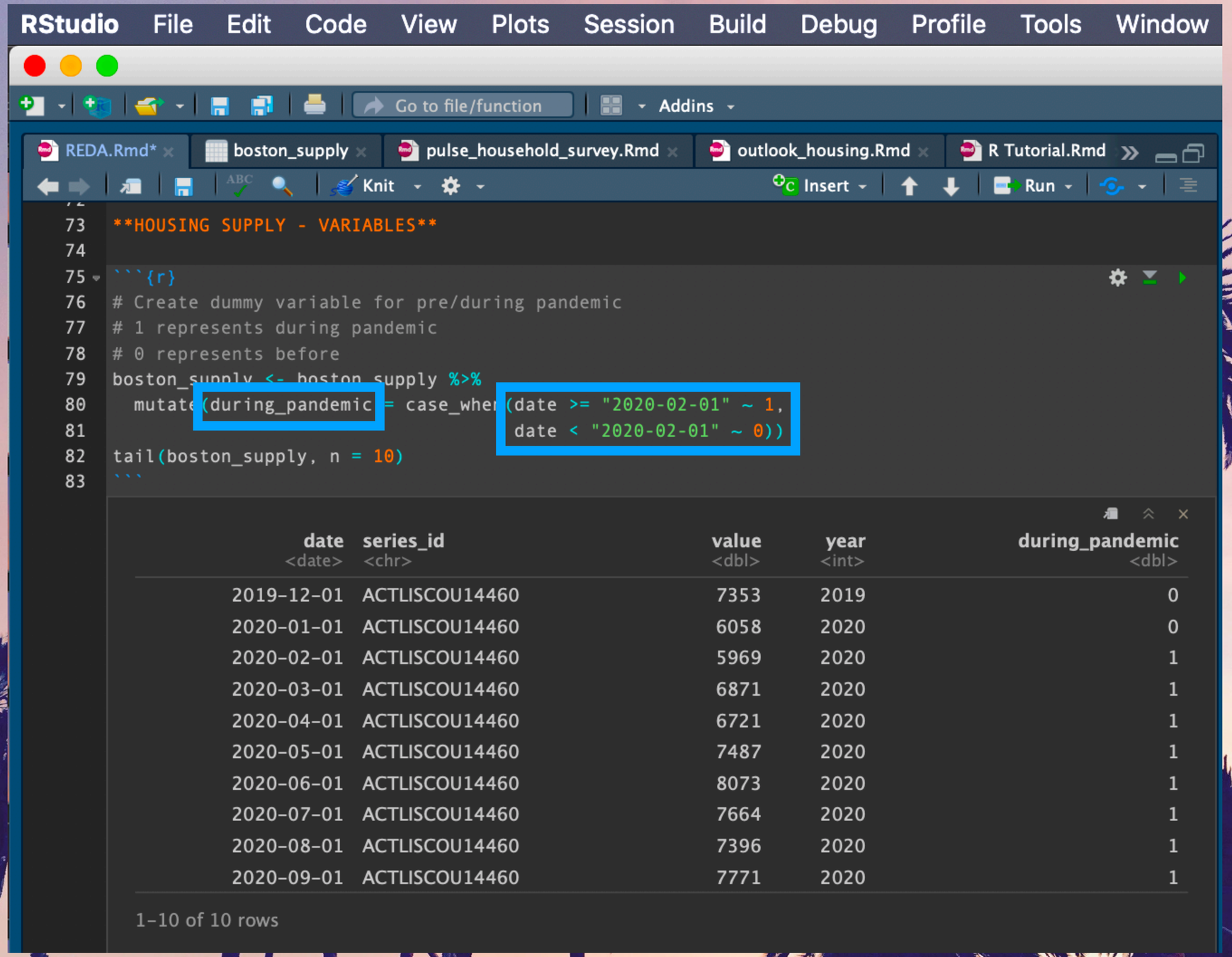
Knit | Insert | Run

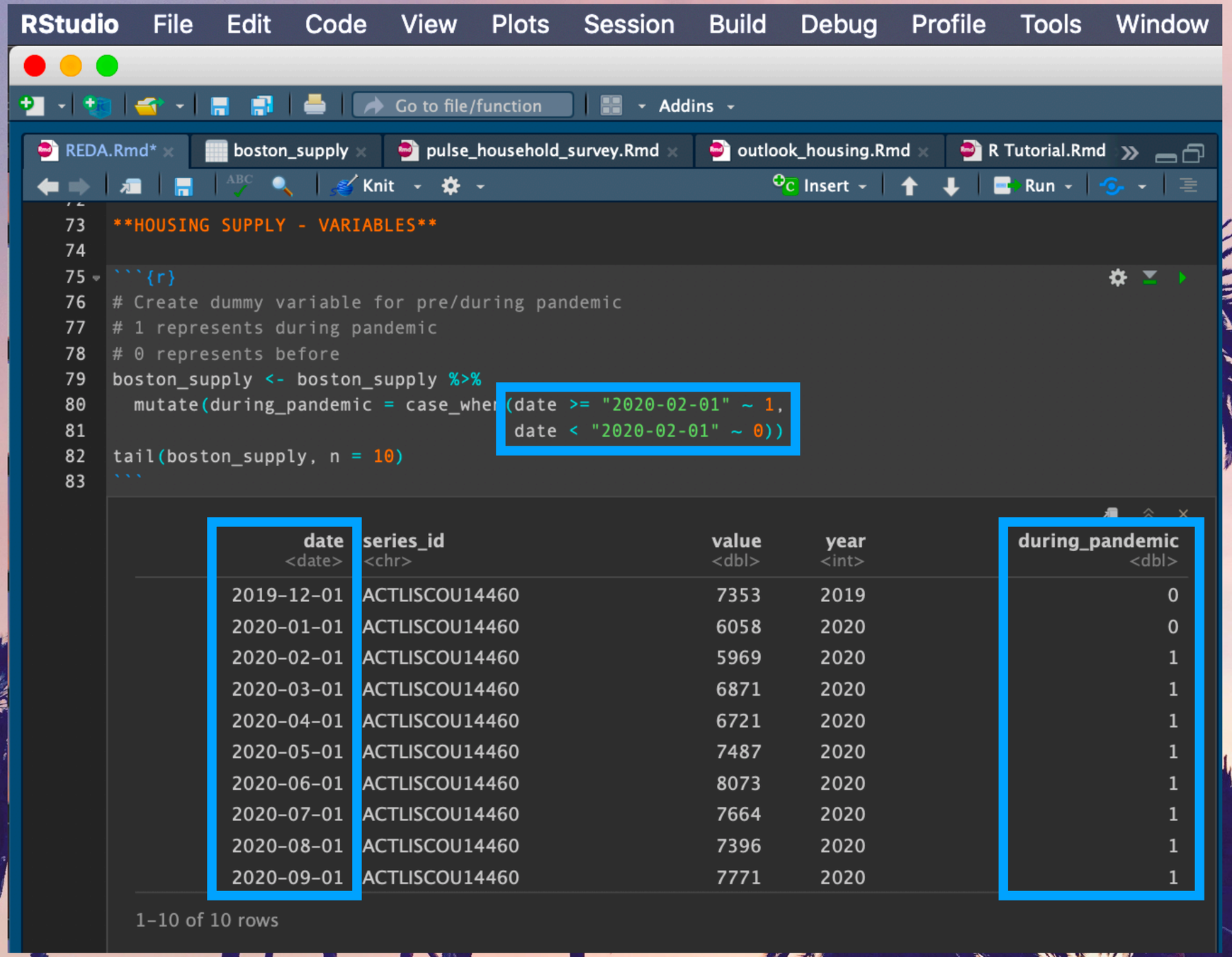
```
**HOUSING SUPPLY - VARIABLES**

```{r}
# Create dummy variable for pre/during pandemic
# 1 represents during pandemic
# 0 represents before
boston_supply <- boston_supply %>%
  mutate(during_pandemic = case_when(date >= "2020-02-01" ~ 1,
                                     date < "2020-02-01" ~ 0))
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

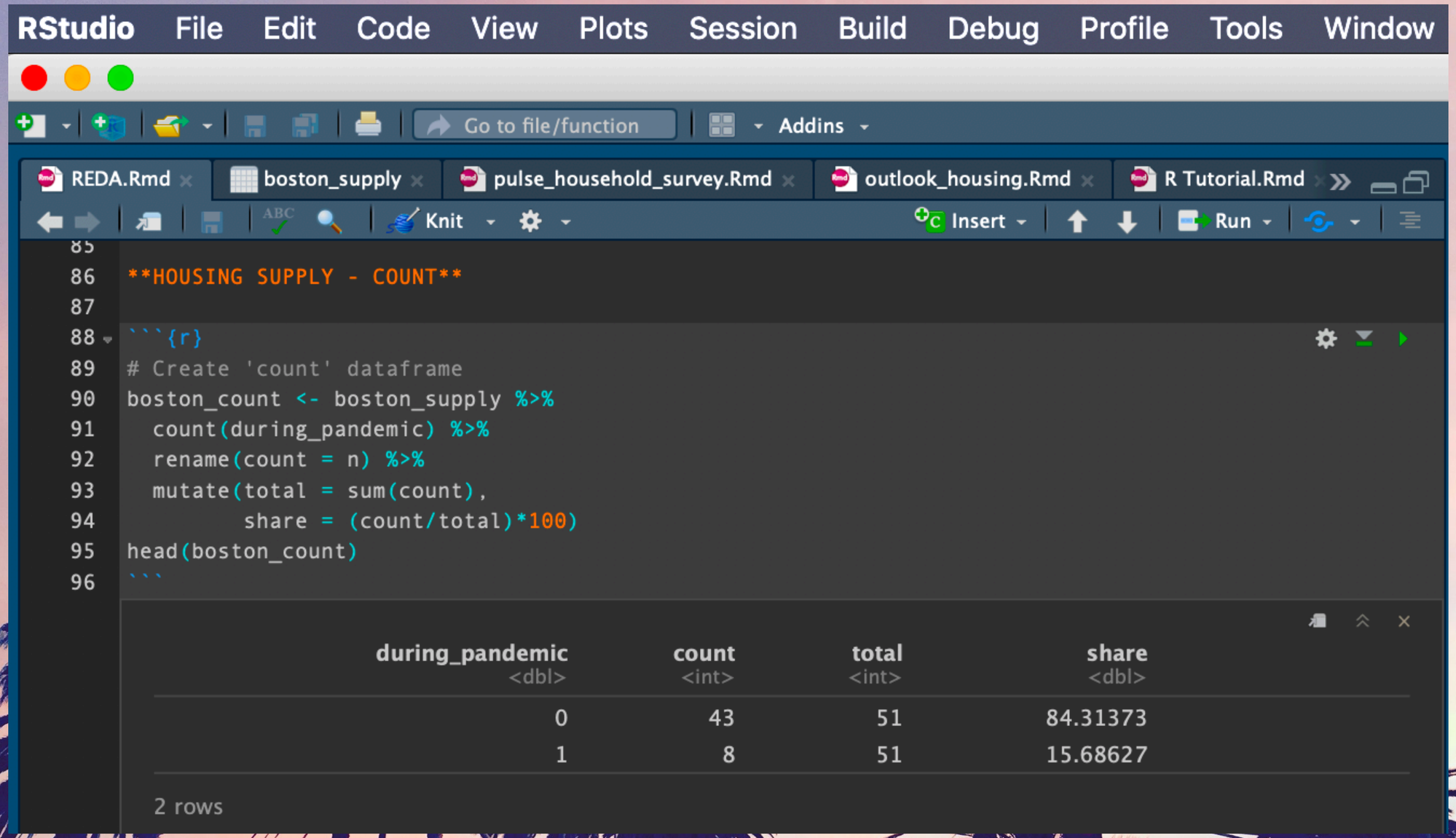

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window
Go to file/function
Addins
REDA.Rmd*
boston_supply
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
Knit
Insert
Run
73 **HOUSING SUPPLY - VARIABLES**
74
75 ```{r}
76 # Create dummy variable for pre/during pandemic
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 <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

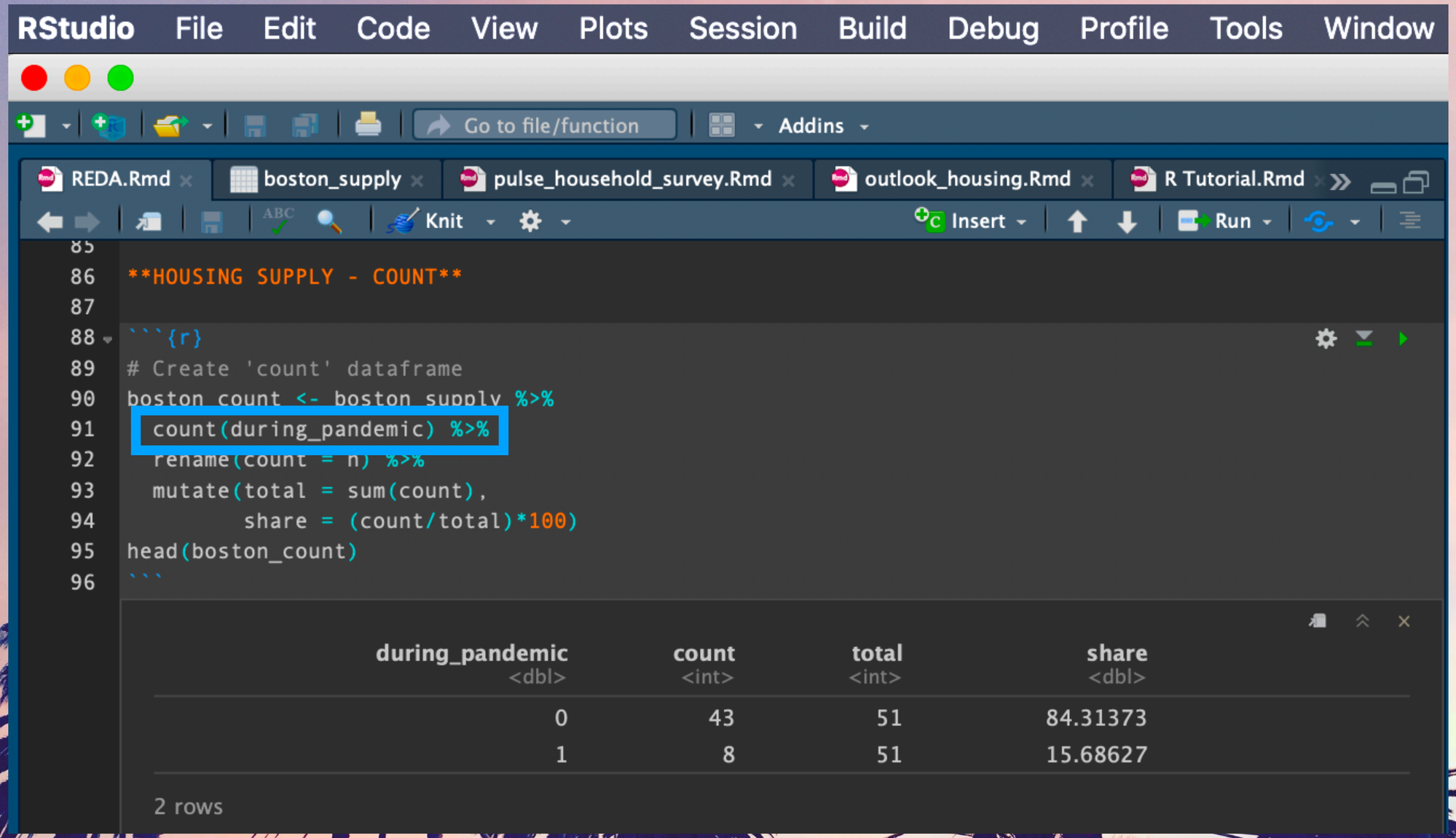

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window
Go to file/function
Addins
REDA.Rmd*
boston_supply
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
Knit
Insert
Run
73 **HOUSING SUPPLY - VARIABLES**
74
75 ```{r}
76 # Create dummy variable for pre/during pandemic
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 <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

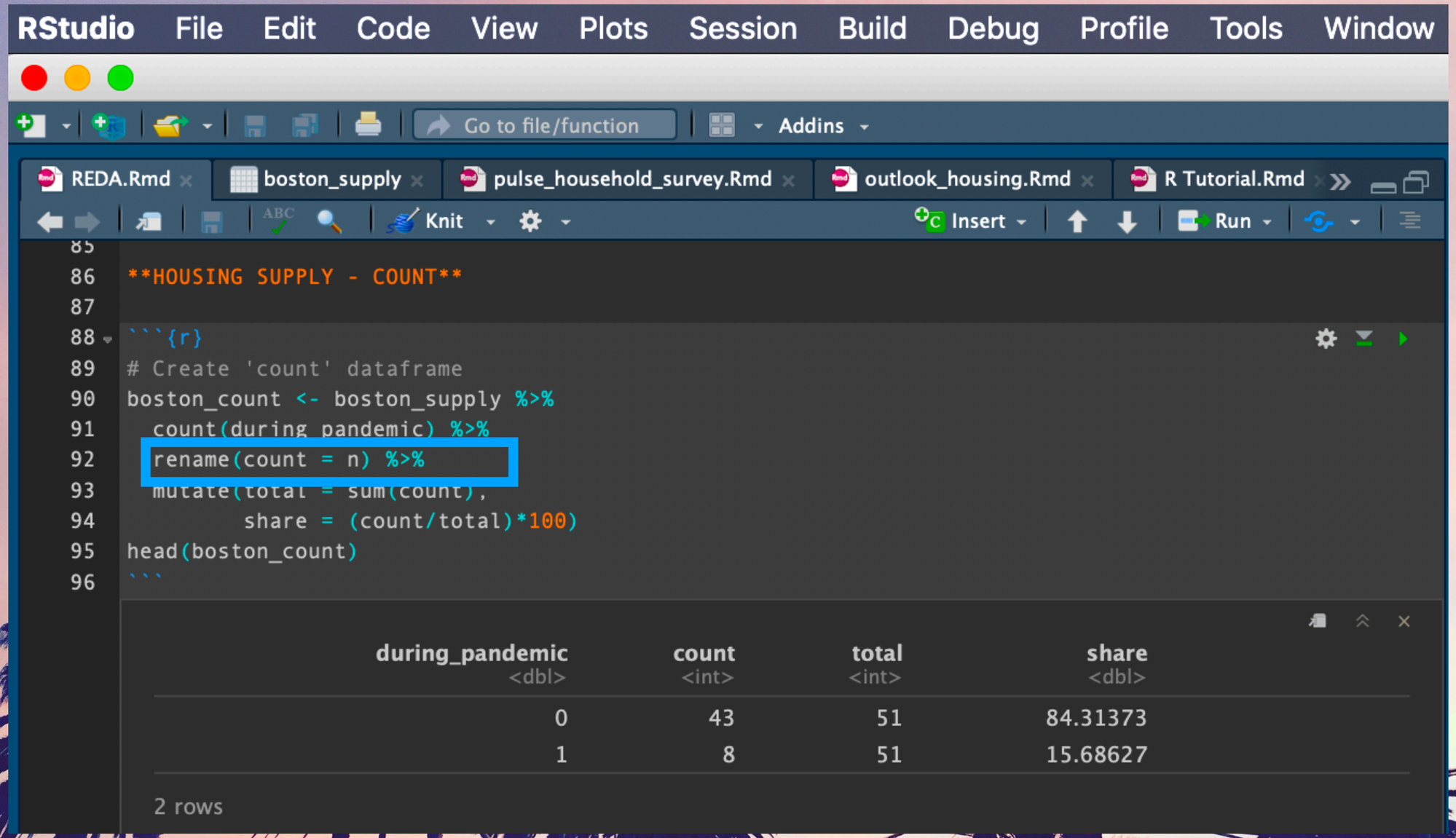
Variables

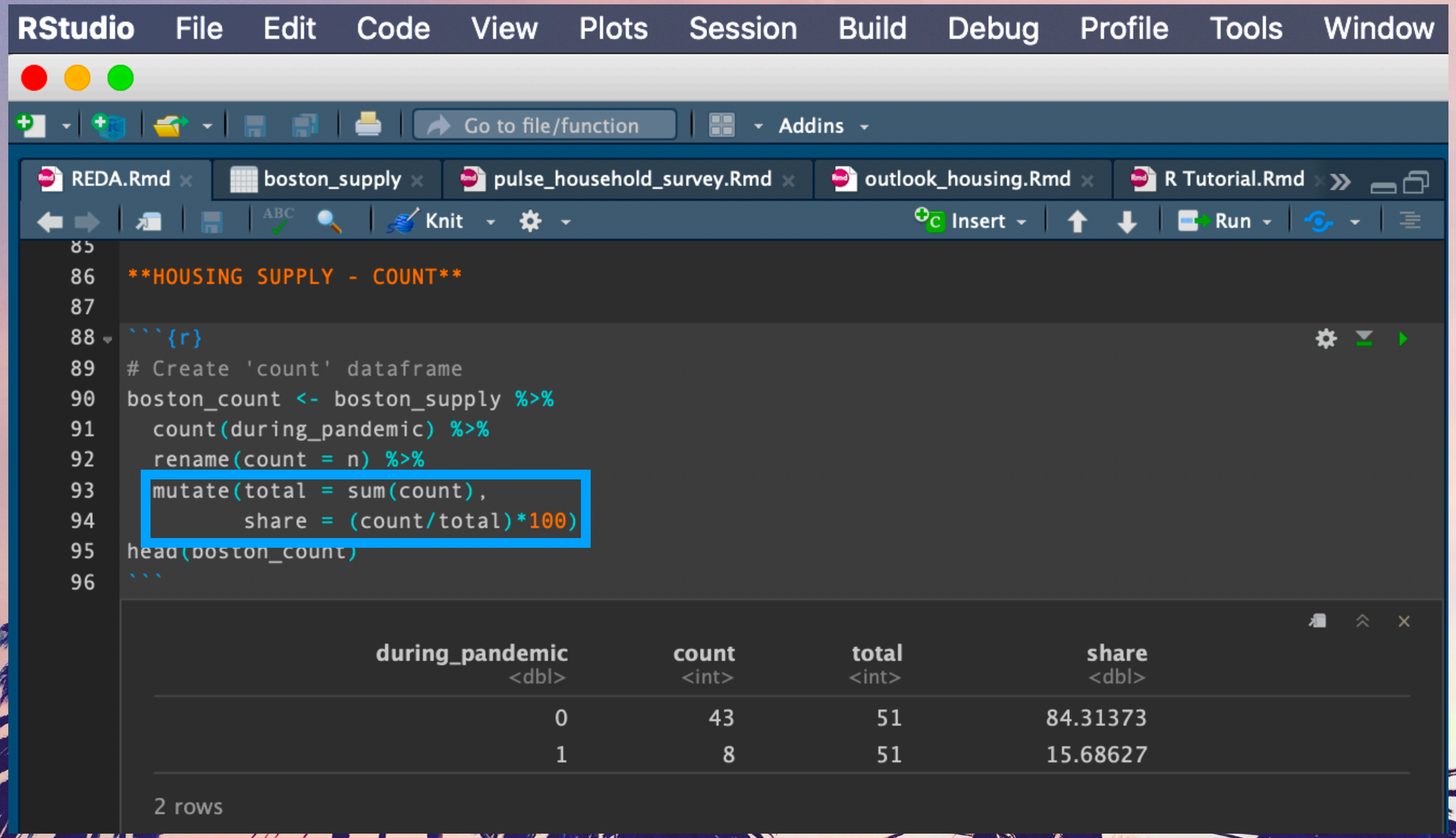
Variables :

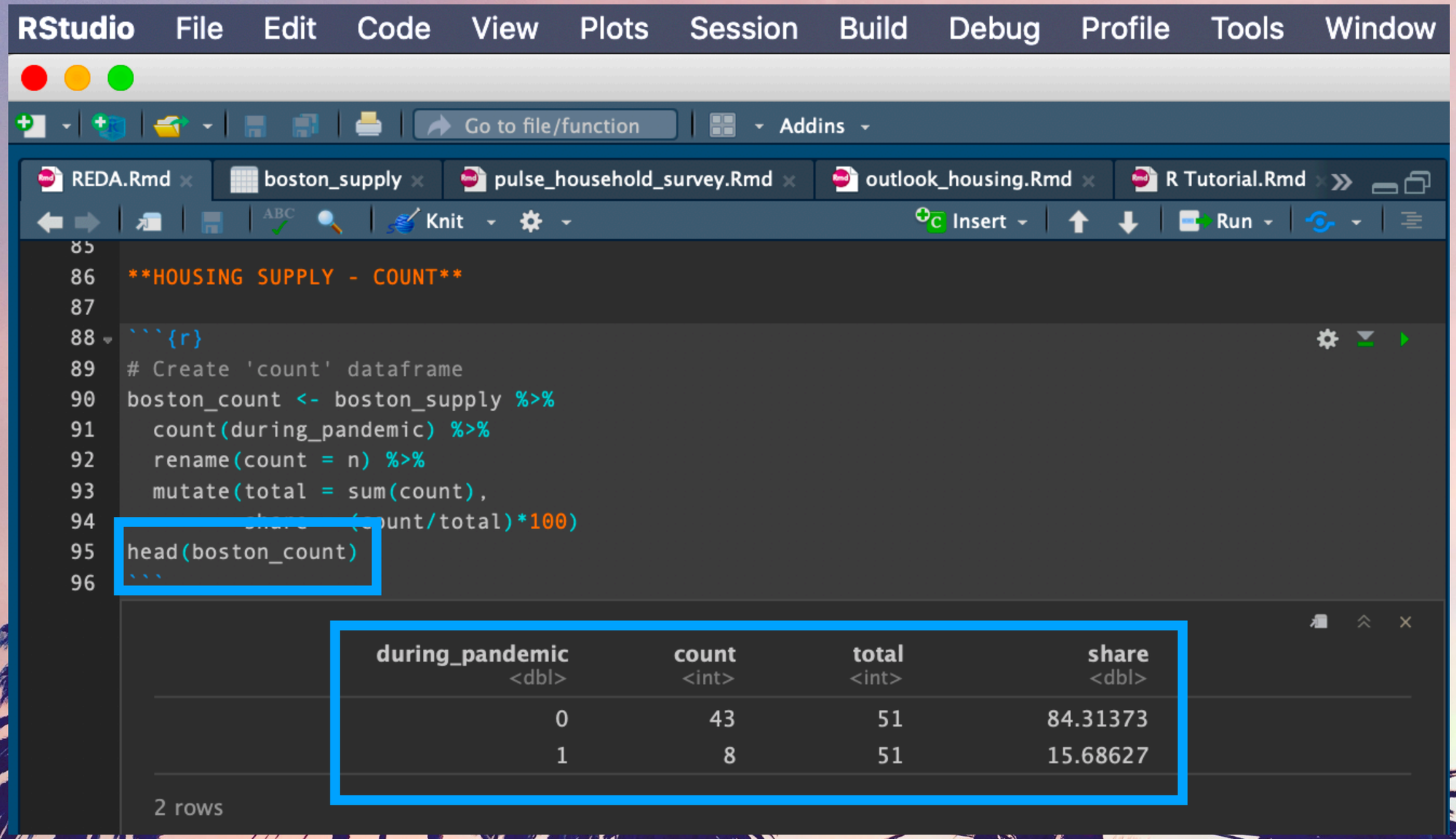
- Count


RStudio File Edit Code View Plots Session Build Debug Profile Tools Window
Go to file/function
Addins
REDA.Rmd
boston_supply
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
Knit
Insert
Run
HOUSING SUPPLY - COUNT
```{r}
# Create 'count' dataframe
boston_count <- boston_supply %>%
count(during_pandemic) %>%
rename(count = n) %>%
mutate(total = sum(count),
share = (count/total)*100)
head(boston_count)
```
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
boston_supply
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
Knit
Insert
Run
HOUSING SUPPLY - COUNT
```{r}
# Create 'count' dataframe
boston_count <- boston_supply %>%
count(during_pandemic) %>%
rename(count = n) %>%
mutate(total = sum(count),
share = (count/total)*100)
head(boston_count)
```
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
boston_supply
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
Knit
Insert
Run
HOUSING SUPPLY - COUNT
```{r}
# Create 'count' dataframe
boston_count <- boston_supply %>%
count(during_pandemic) %>%
rename(count = n) %>%
mutate(total = sum(count),
share = (count/total)*100)
head(boston_count)
```
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
boston_supply
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
Knit
Insert
Run
HOUSING SUPPLY - COUNT
```{r}
# Create 'count' dataframe
boston_count <- boston_supply %>%
count(during_pandemic) %>%
rename(count = n) %>%
mutate(total = sum(count),
share = (count/total)*100)
head(boston_count)
```
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
boston_supply
pulse_household_survey.Rmd
outlook_housing.Rmd
R Tutorial.Rmd
Knit
Insert
Run
HOUSING SUPPLY - COUNT
```{r}
# Create 'count' dataframe
boston_count <- boston_supply %>%
count(during_pandemic) %>%
rename(count = n) %>%
mutate(total = sum(count),
(count/total)*100)
head(boston_count)
```
during_pandemic <dbl> count <int> total <int> share <dbl>
0 43 51 84.31373
1 8 51 15.68627
2 rows

Variables

Variables :

- Descriptive statistics

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

Go to file/function | Addins

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

Knit | Insert | Run

```
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 <dbl>	mean_value <dbl>	median_value <dbl>	sd_value <dbl>	cv_value <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

Go to file/function | Addins

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

Knit | Insert | Run

```
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 <dbl>	mean_value <dbl>	median_value <dbl>	sd_value <dbl>	cv_value <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

Go to file/function | Addins

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

Knit | Insert | Run

```
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 <dbl>	mean_value <dbl>	median_value <dbl>	sd_value <dbl>	cv_value <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

Go to file/function | Addins

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

Knit | Insert | Run

```
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 <dbl>	mean_value <dbl>	median_value <dbl>	sd_value <dbl>	cv_value <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


DO NOT
BLOCK
FINES TO
$500