

# WELCOME TO NDACAN MONTHLY OFFICE HOURS!

*NATIONAL DATA ARCHIVE ON CHILD ABUSE AND NEGLECT  
DUKE UNIVERSITY, CORNELL UNIVERSITY, & UNIVERSITY OF CALIFORNIA: SAN FRANCISCO*



- The session will begin at 11am EST
  - 11:00 - 11:30am – LeaRn with NDACAN (Introduction to R)
  - 11:30 - 12:00pm – Office hours breakout sessions
- Please submit LeaRn questions to the Q&A box
- This session is being recorded.
- See ZOOM Help Center for connection issues:  
<https://support.zoom.us/hc/en-us>
  - If issues persist and solutions cannot be found through Zoom, contact Andres Arroyo at [aa17@cornell.edu](mailto:aa17@cornell.edu).

# LEARN WITH NDACAN

Presented by Frank Edwards

# MATERIALS FOR THIS COURSE

- Course Box folder (<https://cornell.box.com/v/LeaRn-with-R-NDACAN-2024-2025>) contains
  - Data (will be released as used in the lessons)
    - Census state-level data, 2015-2019
    - AFCARS state-aggregate data, 2015-2019
    - AFCARS (FAKE) individual-level data, 2016-2019
    - NYTD (FAKE) individual-level data, 2017 Cohort
  - Documentation/codebooks for the provided datasets
  - Slides used in each week's lesson
  - Exercises as that correspond to each week's lesson
  - An .R file that will have example, usable R code for each lesson – will be updated and appended with code from each lesson

# WEEK 3: BASIC DATA MANAGEMENT

November 11, 2024

## DATA USED IN THIS WEEK'S EXAMPLE CODE

- Census aggregate data from 2015-2019 (census\_2015\_2019.csv)
  - Population counts by state, year, sex, race, and ethnicity
  - Publicly available from CDC Wonder:
    - <https://wonder.cdc.gov/single-race-population.html>
- AFCARS aggregate data from 2015-2019 (afcars\_aggreg\_suppressed.csv)
  - Counts by state, year, sex, race/ethnicity of children in foster care; number of children removed due to physical or sexual abuse, or neglect; the number of children who entered or exited foster care in that year
  - Can order full data from NDACAN:
    - <https://www.ndacan.acf.hhs.gov/datasets/request-dataset.cfm>

# HOW R WORKS WITH DATA

## READING, WRITING AND THE ENVIRONMENT

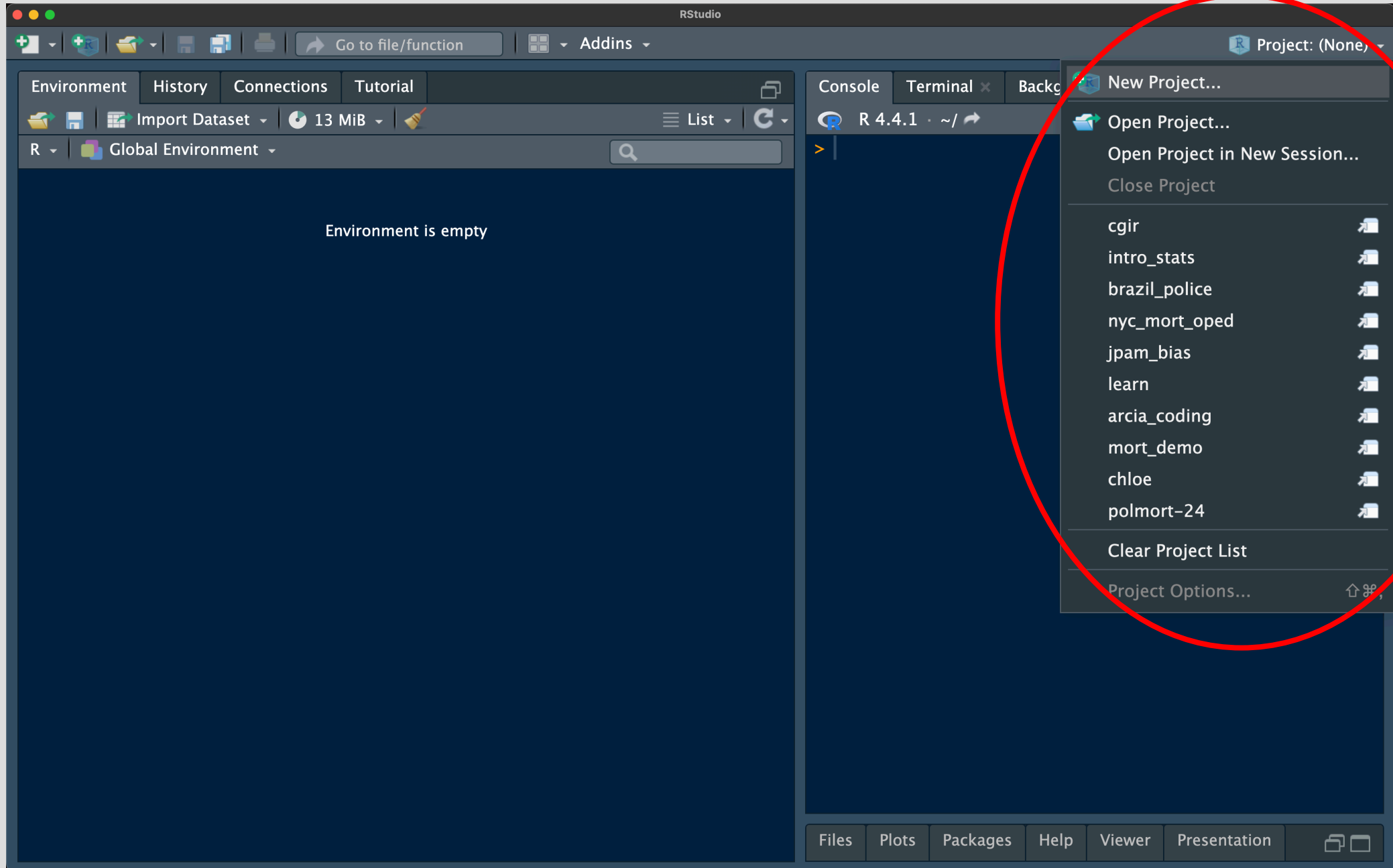
- The 'read' family of functions load data
  - `read_csv`, `read_tsv`, `read_fwf` (tidyverse version of `read`.)
- Data must be loaded into your environment (RAM)
- The environment is temporary
- The 'write' family of functions store data on disk
  - `write_csv` is most common, `saveRDS` has some uses

## YOUR LAPTOP (OR SERVER'S) DISK AND THE ENVIRONMENT

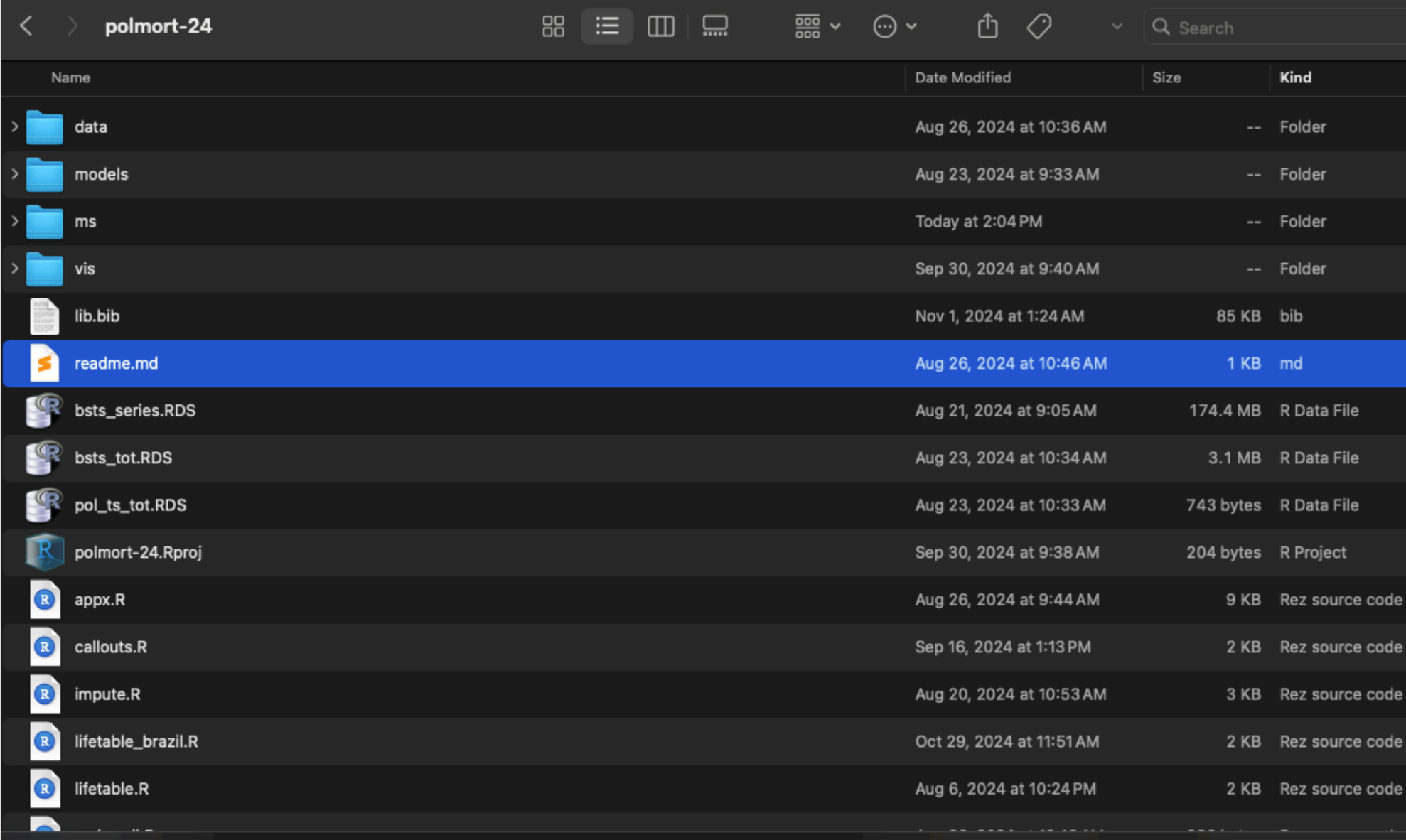
- setwd points R to a location
  - But this is tedious, and bad form with Git / collaboration
- RStudio Projects automatically orient your environment to a particular disk location
  - Get in the habit of using projects and developing a consistent directory structure



# RSTUDIO PROJECT MENU



# A BASIC PROJECT DIRECTORY LAYOUT



The image shows a file explorer window for a project named 'polmort-24'. The interface includes a navigation bar with back/forward arrows, a search bar, and various view icons. The main area displays a list of files and folders with columns for Name, Date Modified, Size, and Kind. The 'readme.md' file is highlighted in blue.

Name	Date Modified	Size	Kind
> data	Aug 26, 2024 at 10:36 AM	--	Folder
> models	Aug 23, 2024 at 9:33 AM	--	Folder
> ms	Today at 2:04 PM	--	Folder
> vis	Sep 30, 2024 at 9:40 AM	--	Folder
lib.bib	Nov 1, 2024 at 1:24 AM	85 KB	bib
readme.md	Aug 26, 2024 at 10:46 AM	1 KB	md
bsts_series.RDS	Aug 21, 2024 at 9:05 AM	174.4 MB	R Data File
bsts_tot.RDS	Aug 23, 2024 at 10:34 AM	3.1 MB	R Data File
pol_ts_tot.RDS	Aug 23, 2024 at 10:33 AM	743 bytes	R Data File
polmort-24.Rproj	Sep 30, 2024 at 9:38 AM	204 bytes	R Project
appx.R	Aug 26, 2024 at 9:44 AM	9 KB	Rez source code
callouts.R	Sep 16, 2024 at 1:13 PM	2 KB	Rez source code
impute.R	Aug 20, 2024 at 10:53 AM	3 KB	Rez source code
lifetable_brazil.R	Oct 29, 2024 at 11:51 AM	2 KB	Rez source code
lifetable.R	Aug 6, 2024 at 10:24 PM	2 KB	Rez source code

# LOADING DATA USING AN RSTUDIO PROJECT

The screenshot displays the RStudio interface with a script editor on the left and a file explorer on the right. The script editor shows the following code:

```
1 library(tidyverse)
2
3 afcars_demo<-read_csv("../data/afcars_aggreg_suppressed.csv")
4
5 census_demo<-read_csv("../data/census_2015_2019.csv")
6
7 #####
```

The file explorer on the right shows the current directory structure:

Name	Size	Modified
..		
data		
learn.Rproj	204 B	Nov 11, 2024, 2:27
Untitled.html	691 KB	Oct 18, 2024, 11:22
Untitled.Rmd	843 B	Oct 18, 2024, 11:22
week2solutions.Rmd	3.1 KB	Oct 18, 2024, 1:03
week3.R	140 B	Nov 11, 2024, 2:32



```
learn - RStudio
Go to file/function
Addins
learn

week3.R
Source on Save
Run
Source

1 library(tidyverse)
2
3 afcars_demo<-read_csv("../data/afcars_aggreg_suppressed.csv")
4
5 census_demo<-read_csv("../data/census_2015_2019.csv")
6
7 ### change census variable names
8 ### to match afcars names for join
9
10 census_demo<-census_demo %>%
11   rename(fy = cy,
12         stname = state,
13         state = stfips)
14
15

Console
Terminal
Background Jobs
R 4.4.1 · ~/docs/ndacan/learn/
quiet this message.
>
> ### change census variable names
> ### to match afcars names for join
> head(afcars_demo)
# A tibble: 6 × 10
  fy state sex raceethn numchild phyabuse sexabuse
  <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>
1  2015     1 1         1     2180     352     88
2  2015     1 1         2     1245     198     46
3  2015     1 1         4         10      NA      0
4  2015     1 1         5      NA      0      0
5  2015     1 1         6      245     30     NA
6  2015     1 1         7      204     56     22
# i 3 more variables: neglect <dbl>, entered <dbl>,
#   exited <dbl>
> head(census_demo)
# A tibble: 6 × 8
  cy stfips state st sex race6 hisp pop
  <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl>
1  2015     1 Alabama AL     1     1     0 331305
2  2015     1 Alabama AL     1     1     1  33105
3  2015     1 Alabama AL     1     2     0 164122
4  2015     1 Alabama AL     1     2     1   2763
5  2015     1 Alabama AL     1     3     0   2540
6  2015     1 Alabama AL     1     3     1   1195
>
```

# HARMONIZING THE ROWS (UNIT OF ANALYSIS)

The screenshot shows the RStudio interface with a script editor on the left and a console on the right. The script editor contains R code for reading two CSV files, renaming variables, and collapsing data. The console shows the output of the `head()` functions for both datasets.

```
1 library(tidyverse)
2
3 afcars_demo<-read_csv("../data/afcars_aggreg_suppressed.csv")
4
5 census_demo<-read_csv("../data/census_2015_2019.csv")
6
7 ### change census variable names
8 ### to match afcars names for join
9
10 census_demo<-census_demo %>%
11   rename(fy = cy,
12         stname = state,
13         state = stfips)
14
15 ## Collapse race / ethnicity / sex in both tables
16 ## to create total pop, total entries
17
18 census_collapse<-census_demo %>%
19   group_by(fy, state) %>%
20   summarize(pop = sum(pop))
21
22 afcars_collapse<-afcars_demo %>%
23   group_by(fy, state) %>%
24   summarize(entered = sum(entered, na.rm = T))
25
26
```

Console output:

```
> head(census_collapse)
# A tibble: 6 × 3
# Groups:   fy [1]
  fy state    pop
<dbl> <dbl> <dbl>
1  2015     1 1103159
2  2015     2  184134
3  2015     4 1629765
4  2015     5  706879
5  2015     6  9118819
6  2015     8 1258312

> head(afcars_collapse)
# A tibble: 6 × 3
# Groups:   fy [1]
  fy state entered
<dbl> <dbl> <dbl>
1  2015     1    3536
2  2015     2    1483
3  2015     4   12553
4  2015     5    4009
5  2015     6   31258
6  2015     8    4733
>
```

# JOINING THE DATA

```
learn - RStudio
Go to file/function
Addins
learn

week3.R x LeaN_module_week2.R x
Source on Save Run Source

7 ### change census variable names
8 ### to match afcars names for join
9
10 census_demo<-census_demo %>%
11   rename(fy = cy,
12         stname = state,
13         state = stfips)
14
15 ## Collapse race / ethnicity / sex in both tables
16 ## to create total pop, total entries
17
18 census_collapse<-census_demo %>%
19   group_by(fy, state) %>%
20   summarize(pop = sum(pop))
21
22 afcars_collapse<-afcars_demo %>%
23   group_by(fy, state) %>%
24   summarize(entered = sum(entered, na.rm = T))
25
26 ### join the data.frames, and compute a per capita entry rate
27
28 dat_join<-afcars_collapse %>%
29   left_join(census_collapse)
30
31 dat_join<-dat_join %>%
32   mutate(entries_per1000 = entered / pop * 1000)
33
```

```
R 4.4.1 · ~/docs/ndacan/learn/
+ left_join(census_collapse)
Joining with `by = join_by(fy, state)`
> head(dat_join)
# A tibble: 6 × 4
# Groups:   fy [1]
  fy state entered      pop
<dbl> <dbl> <dbl> <dbl>
1  2015     1   3536 1103159
2  2015     2   1483  184134
3  2015     4  12553 1629765
4  2015     5   4009  706879
5  2015     6  31258 9118819
6  2015     8   4733 1258312
> dat_join<-dat_join %>%
+   mutate(entries_per1000 = entered / pop * 1000)
> head(dat_join)
# A tibble: 6 × 5
# Groups:   fy [1]
  fy state entered      pop entries_per1000
<dbl> <dbl> <dbl> <dbl> <dbl>
1  2015     1   3536 1103159         3.21
2  2015     2   1483  184134         8.05
3  2015     4  12553 1629765         7.70
4  2015     5   4009  706879         5.67
5  2015     6  31258 9118819         3.43
6  2015     8   4733 1258312         3.76
> |
```

33:1 (Top Level) R Script

# VISUALIZING THE DATA: DISTRIBUTIONS OVER TIME

```
learn - RStudio
Go to file/function
Addins
learn

week3.R* x LeaRn_module_week2.R x
Source on Save Run Source
14
15 ## Collapse race / ethnicity / sex in both tables
16 ## to create total pop, total entries
17
18 census_collapse<-census_demo %>%
19   group_by(fy, state) %>%
20   summarize(pop = sum(pop))
21
22 afcars_collapse<-afcars_demo %>%
23   group_by(fy, state) %>%
24   summarize(entered = sum(entered, na.rm = T))
25
26 ### join the data.frames, and compute a per capita entry rate
27
28 dat_join<-afcars_collapse %>%
29   left_join(census_collapse)
30
31 dat_join<-dat_join %>%
32   mutate(entries_per1000 = entered / pop * 1000)
33
34 ### visualize the data
35
36 dat_join %>%
37   ggplot(aes(x = entries_per1000)) +
38   geom_histogram() +
39   facet_wrap(~fy)
40
```

Console

```
R 4.4.1 ~ /docs/ndacan/learn/
+ ggplot(aes(x = entries_per1000)) +
+ geom_histogram() +
+ facet_wrap(~fy)
`stat_bin()` using `bins = 30`. Pick better value
with `binwidth`.
Warning message:
Removed 5 rows containing non-finite outside the
scale range (`stat_bin()`).
> |
```

Files Plots Packages Help Viewer Presentation

Zoom Export

count

2015 2016 2017

2018 2019

entries\_per1000

40:1 (Top Level) R Script

Environment History Connections Tutorial

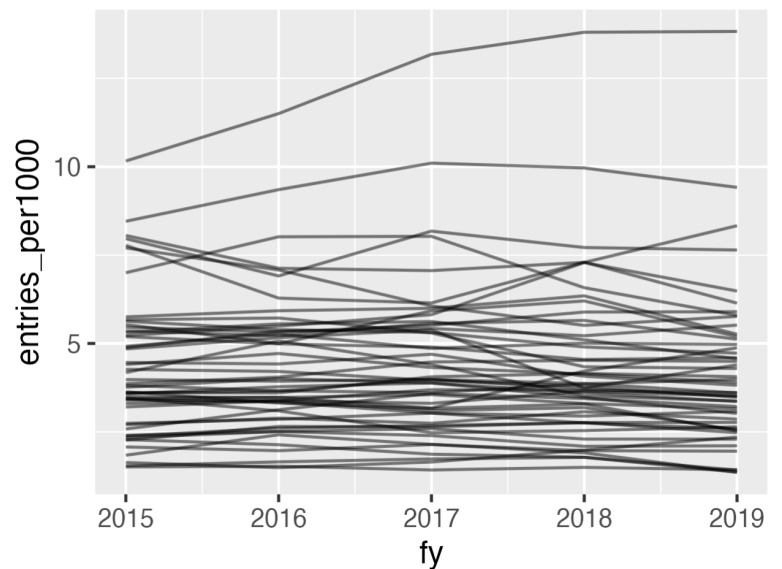
# VISUALIZING THE DATA: BY STATE

```
learn - RStudio
Go to file/function
Addins
week3.R
Source on Save
Run
Source
20 summarize(pop = sum(pop))
21
22 afcars_collapse<-afcars_demo %>%
23   group_by(fy, state) %>%
24   summarize(entered = sum(entered, na.rm = T))
25
26 ### join the data.frames, and compute a per capita entry rate
27
28 dat_join<-afcars_collapse %>%
29   left_join(census_collapse)
30
31 dat_join<-dat_join %>%
32   mutate(entries_per1000 = entered / pop * 1000)
33
34 ### visualize the data
35
36 dat_join %>%
37   ggplot(aes(x = entries_per1000)) +
38   geom_histogram() +
39   facet_wrap(~fy)
40
41 ### visualize state time series
42 dat_join %>%
43   ggplot(aes(x = fy, y = entries_per1000,
44             group = state)) +
45   geom_line(alpha = 0.5)
46
```

```
R 4.4.1 · ~/docs/ndacan/learn/
> ### visualize state time series
> dat_join %>%
+   ggplot(aes(x = fy, y = entries_per1000,
+             group = state)) +
+   geom_line(alpha = 0.5)
Warning message:
Removed 5 rows containing missing values or
values outside the scale range
(`geom_line()`).
> |
```

Files | Plots | Packages | Help | Viewer | Presentations

Zoom | Export



Environment | History | Connections | Tutorial



# STORING OUTPUT ON DISK

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for data manipulation and visualization. The code includes comments in blue and function calls in white. Line numbers 23-49 are visible.
- Console:** Shows the execution of the code. It displays a warning message: "Warning message: Removed 5 rows containing missing values or values outside the scale range (geom\_line()).".
- Files Panel:** Shows the file explorer for the project directory. It lists files and folders including 'data', 'learn.Rproj', 'Untitled.html', 'Untitled.Rmd', 'week2solutions.Rmd', 'week3.R', and 'vis'.

```
23 group_by(fy, state) %>%
24   summarize(entered = sum(entered, na.rm = T))
25
26 ### join the data.frames, and compute a per capita entry rate
27
28 dat_join<-afcars_collapse %>%
29   left_join(census_collapse)
30
31 dat_join<-dat_join %>%
32   mutate(entries_per1000 = entered / pop * 1000)
33
34 ### visualize the data
35
36 dat_join %>% |
37   ggplot(aes(x = entries_per1000)) +
38     geom_histogram() +
39     facet_wrap(~fy)
40
41 ### visualize state time series
42 dat_join %>%
43   ggplot(aes(x = fy, y = entries_per1000,
44             group = state)) +
45     geom_line(alpha = 0.5)
46
47 ggsave("./vis/state_ts.png", width = 5, height = 5)
48
49 write_csv(dat_join, "./data/join_demo.csv")
```

Console output:

```
R 4.4.1 · ~/docs/ndacan/learn/
> ggsave("./vis/state_ts.png", width = 5, height = 5)
Warning message:
Removed 5 rows containing missing values or
values outside the scale range
(`geom_line()`).
>
> write_csv(dat_join, "./data/join_demo.csv")
> |
```

Files Panel:

Name	Size	Mod
..		
data		
learn.Rproj	204 B	Nov
Untitled.html	691 KB	Oct
Untitled.Rmd	843 B	Oct
week2solutions.Rmd	3.1 KB	Oct
week3.R	1.1 KB	Nov
vis		