full_join

Using R to defend immigrant's rights at the ACLU

Brooke Watson, Senior Data Scientist, ACLU Analytics

Case	Package(s)	Use Case
1. **Damus v. Nielsen**	1. tabulizer + daff	1. Look at your data
2. **Hernandez v. Sessions**	2. readr + visdat	2. **Look at your data**
3. **Ms L. v. ICE**	3. ggplot2	3. ***Look at your data***

Functions

tabulizer

1. locate_areas()
2. get_n_pages()
3. extract_tables()

daff

4. diff_data()
5. render_daff()

visdat

6. vis_miss()

readr

7. read_file()
8. write_file()

ggplot2

9. stat_ecdf()
10. geom_density()

Damus v. Nielsen

Damus v. Nielsen

Legal Ruling

- The judge in this case issued an injunction ordering ICE to give parole interviews to all class members, to issue determinations in a timely manner, and to report those determinations monthly to our litigation team.

Data Question

- Do the statistics ICE reports match the raw PDF data they are required to share?
- Do the data indicate that people are being detained arbitrarily?

Tabulizer: extract data from PDFs

Daff: check changes between two datasets

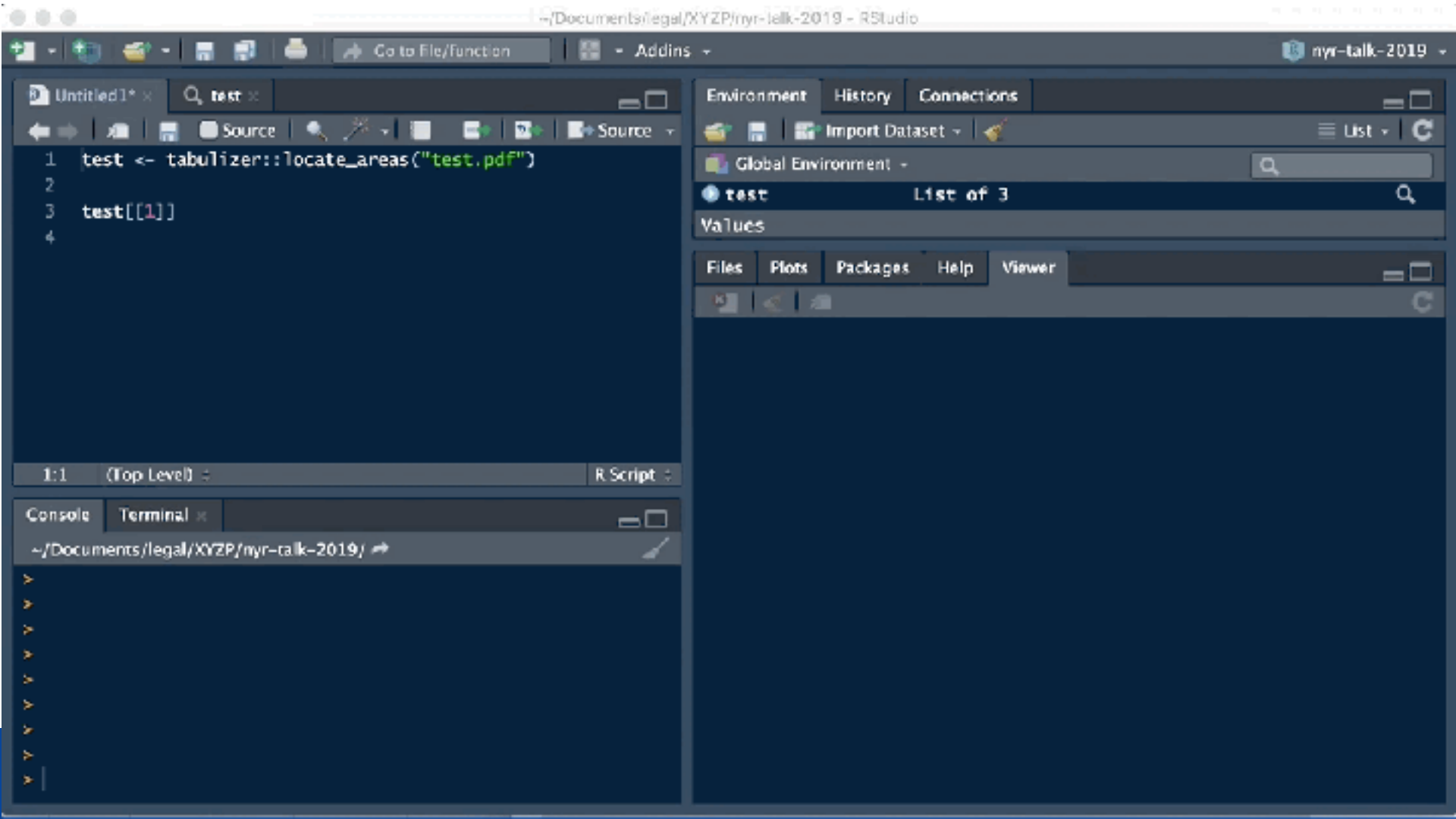
Tabulizer + daff in Damus v. Nielsen:
monitor changes government reporting over time

Tabulizer

1. **locate_areas()** # to find the positions of tables or columns

2. **get_n_pages()** # to determine the length of a PDF

3. **extract tables()** # to get tables from said PDF


~/Documents/legal/XYZP/nyr-talk-2019 - RStudio
Go to File/Function
Addins
nyr-talk-2019
Untitled1*
test
Source
Source
1 test <- tabulizer::locate_areas("test.pdf")
2
3 test[[1]]
4
1:1
(Top Level)
R Script
Console
Terminal
~/Documents/legal/XYZP/nyr-talk-2019/
Environment
History
Connections
Import Dataset
List
Global Environment
test
List of 3
Values
Files
Plots
Packages
Help
Viewer

```
multi_tbl_extract <- function(path, pass = pwd) {

    n <- get_n_pages(path, password = pass)

    d <- extract_tables(path,
                        password = pass,
                        pages = 1:n,
                        method = "lattice",
                        columns = list(c(38, 58, 94, 124.5)))
    names(d) <- 1:n

    map(d, as_tibble) %>%
        bind_rows()
}
```

The result of locate_areas()

daff

4. **diff_data()** # track changes between two datasets

5. **render_diff()** # view a color-coded HTML file of the changes

```r
library(daff)
y <- iris[1:10,]
x <- y

x <- head(x,9) # remove a row
x[1,1] <- 10 # change a value
xShello <- "world"  # add a column
xSSpecies <- NULL # remove a column


```

```
> 
> 
> 
> 
> 
> 
> 
> dif <- diff_data(y, x)
> render_diff(dif)
> 
```

daff: diff_data() + render_diff()

```
mar <- read_rds("march_cleaned.RDS")
apr <- multi_tbl_extract("raw/april/dir")

mar_apr_changes <- diff_data(mar, apr)

render_diff(mar_apr_changes)
```

daff

Identify changes over time

Identify corrupted data that may seem facially harmless

‘fake_data’ vs. ‘fake_data_corrupt’

2019-05-06 19:43:33

	#	Modified	Reordered	Deleted	Added
Rows	5 → 6	4	0	0	1
Columns	3	0	0	0	0

@@	id	date	status
⇒	1	2019-05-16 → 2019-05-06	parole granted
⇒	2	2019-05-17 → 2019-05-07	parole granted → parole denied
⇒	3	2019-05-18 → 2019-05-08	parole denied
⇒	4	2019-05-19 → 2019-05-09	parole denied
	5	2019-05-20	null
+++	6	2019-05-06	null

Hernandez v. Sessions

Hernandez v. Sessions

Legal Ruling

- If courts don't consider an individual's ability to pay a bond, the government "risks detention that accomplishes **'little more than punishing a person for his poverty.'"**

Data Question

- How can we use administrative court records (EOIR data) to find how many individuals are being detained because they **can't afford their bond**?

The “Beetlejuice Provision” of the Freedom of Information Act

“Each agency, in accordance with published rules, shall make available for public inspection in an electronic format […] copies of all records, regardless of form or format—that have been released to any person under paragraph (3); and […] that have been requested 3 or more times.”

- 5 U.S. Code §552(a)(2)

readr: read in and clean multiple file types

visdat: look at data types and missingness

readr + visdat with EOIR data:
Make public immigration data usable

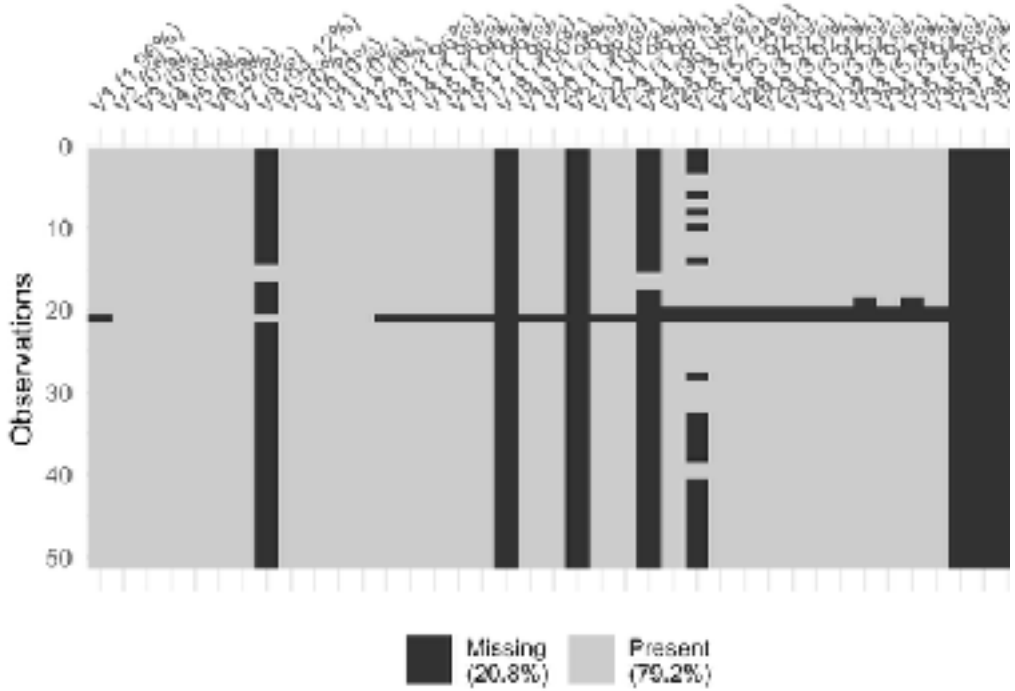
visdat + readr

6. **visdat::vis_miss():** visualize patterns of missingness in data

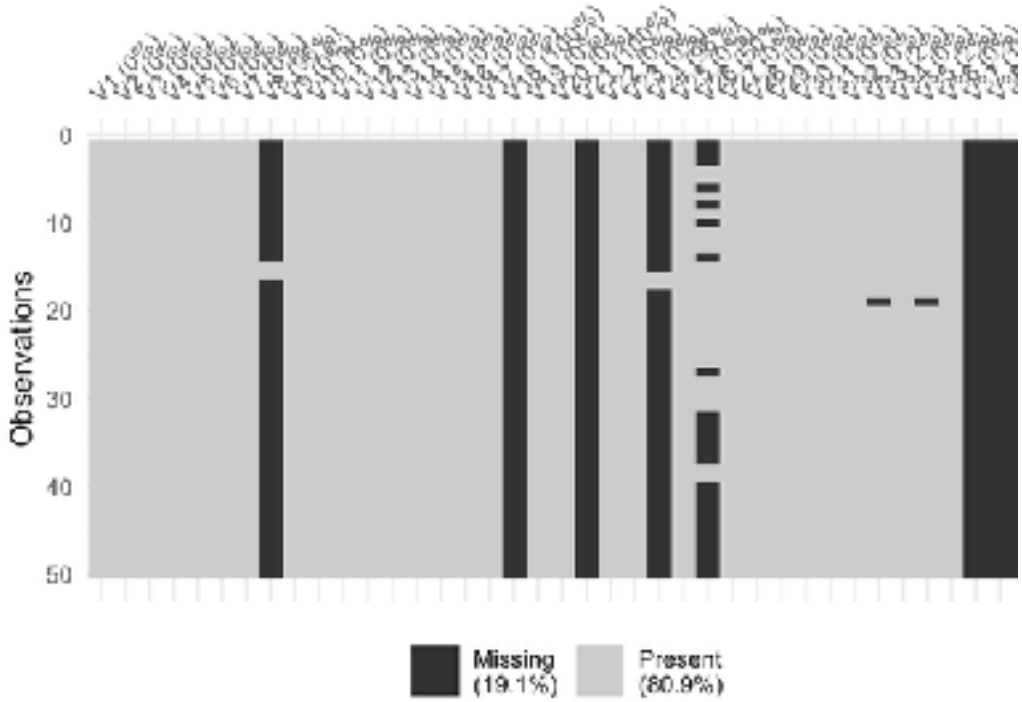
7. **readr::read_file()** # read **raw** text files to identify errors in tabs, spaces, & carriage returns

8. **readr::write_file()** # write raw text files (e.g. after cleaning)

```
df <- read.csv("df.csv",
               sep = "\t")

visdat::vis_miss(df)
```

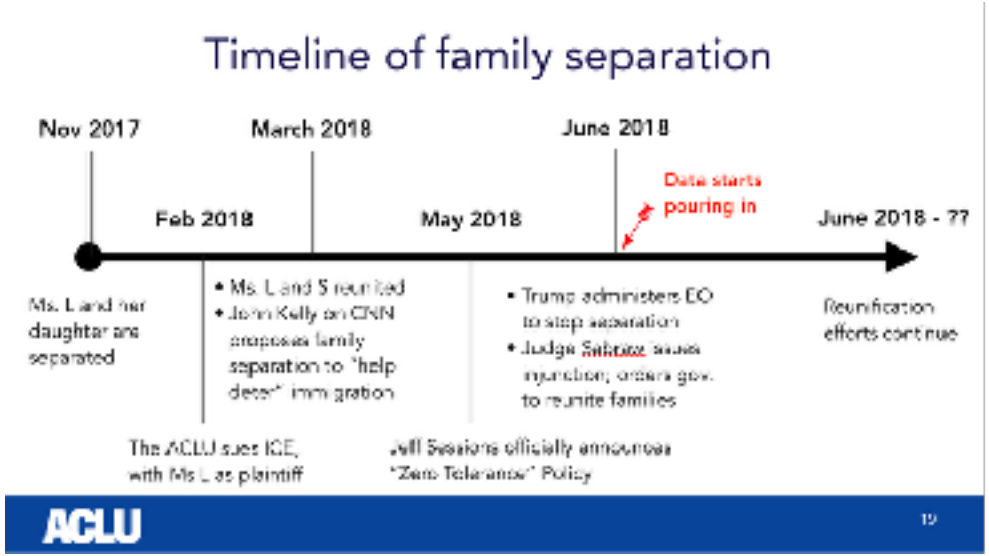


```
read_file("df.csv") %>%
    str_remove(., "\r") %
    write_file("df.csv")

df <- read.csv("df.csv",
           sep = "\t")

visdat::vis_miss(df)
```



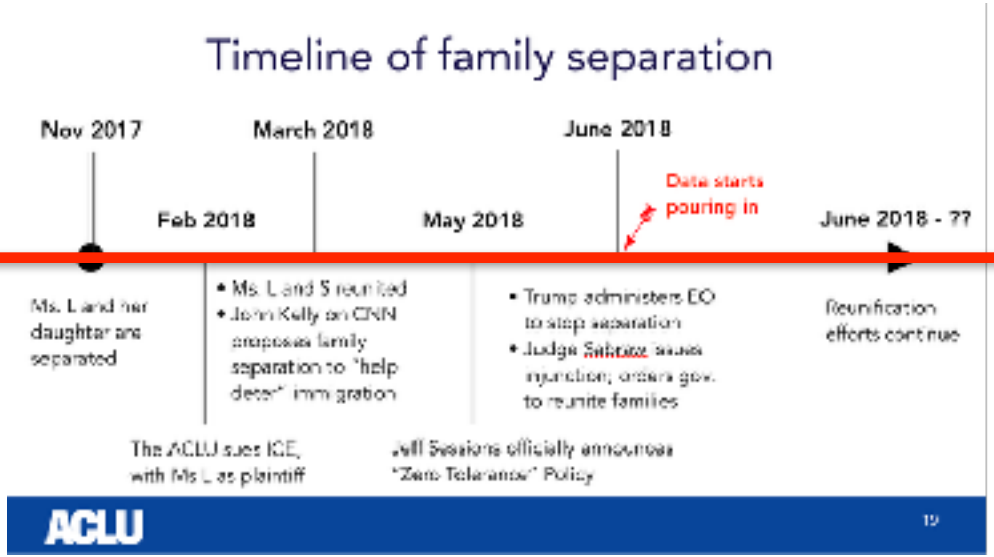
Ms. L
v.
ICE

In January, this is what we knew:



Over **2700 children** were separated from their parents before a federal injunction stopped the practice, most of them in May and June of 2018.

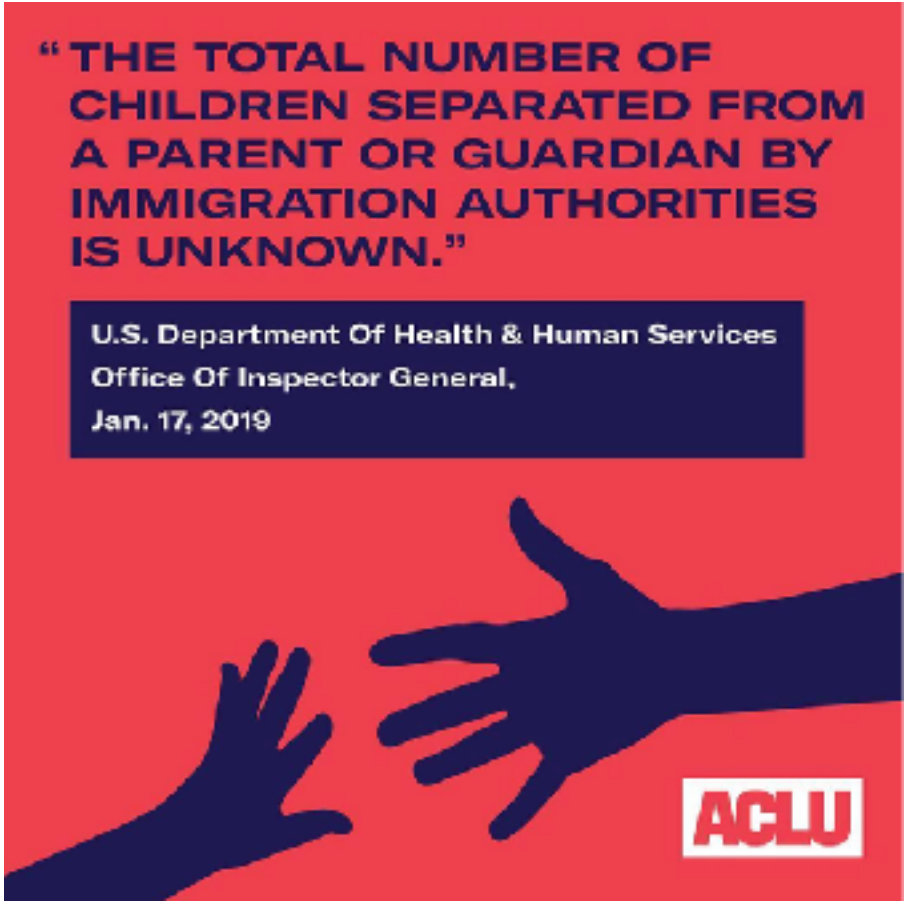
Separations were occurring as early as **July 1, 2017.** The government has now been ordered to ID and reunite these families.



At LEAST **389** separations have occurred **since the federal injunction** on June 26, 2018.

Over **2700 children** were separated from their parents before a federal injunction stopped the practice, most of them in May and June of 2018.

All the analyses following use government-reported data, and **certainly undercount** the true number of children separated from their families.



ggplot2: explore a variety of plots with the same base structure and small changes in syntax.

ggplot2 in Ms. L vs. ICE: Understand the shifting nature of family separation practices over time.

ggplot2: look at data in multiple ways

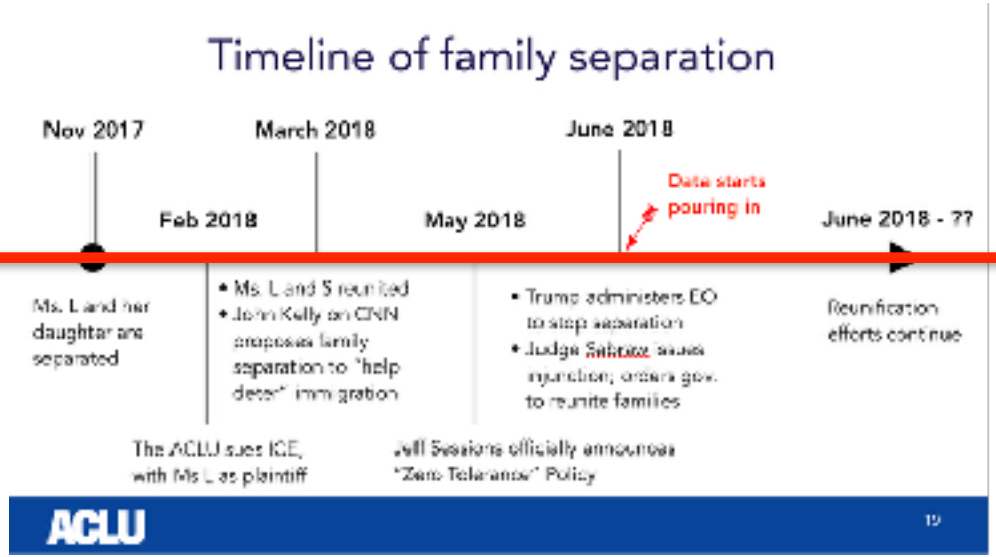
9. stat_ecdf()

- histogram complement
- Useful when data is particularly susceptible to bin effects (e.g. time series with varying numbers of days/weekdays in a month)

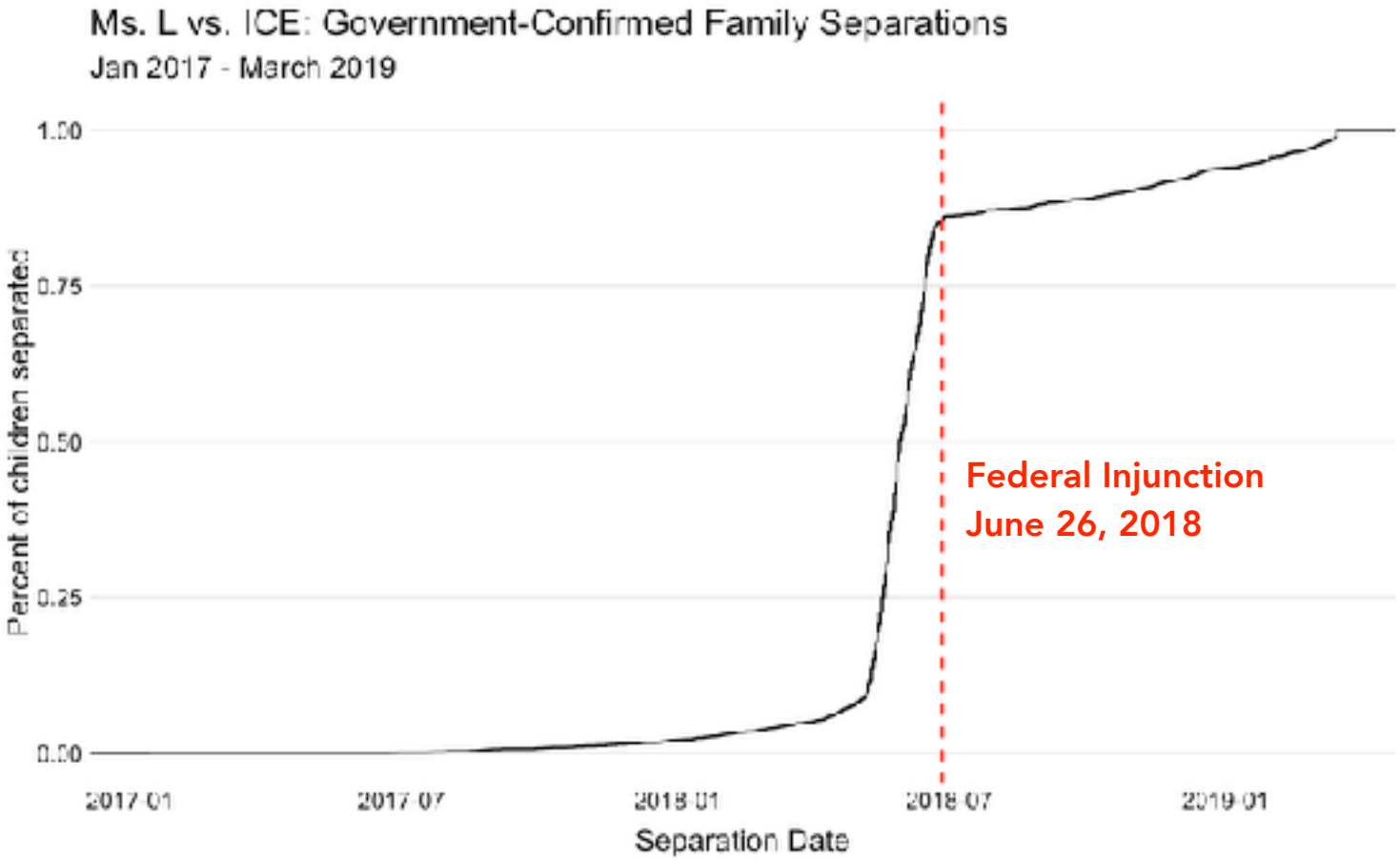
10. geom_density()

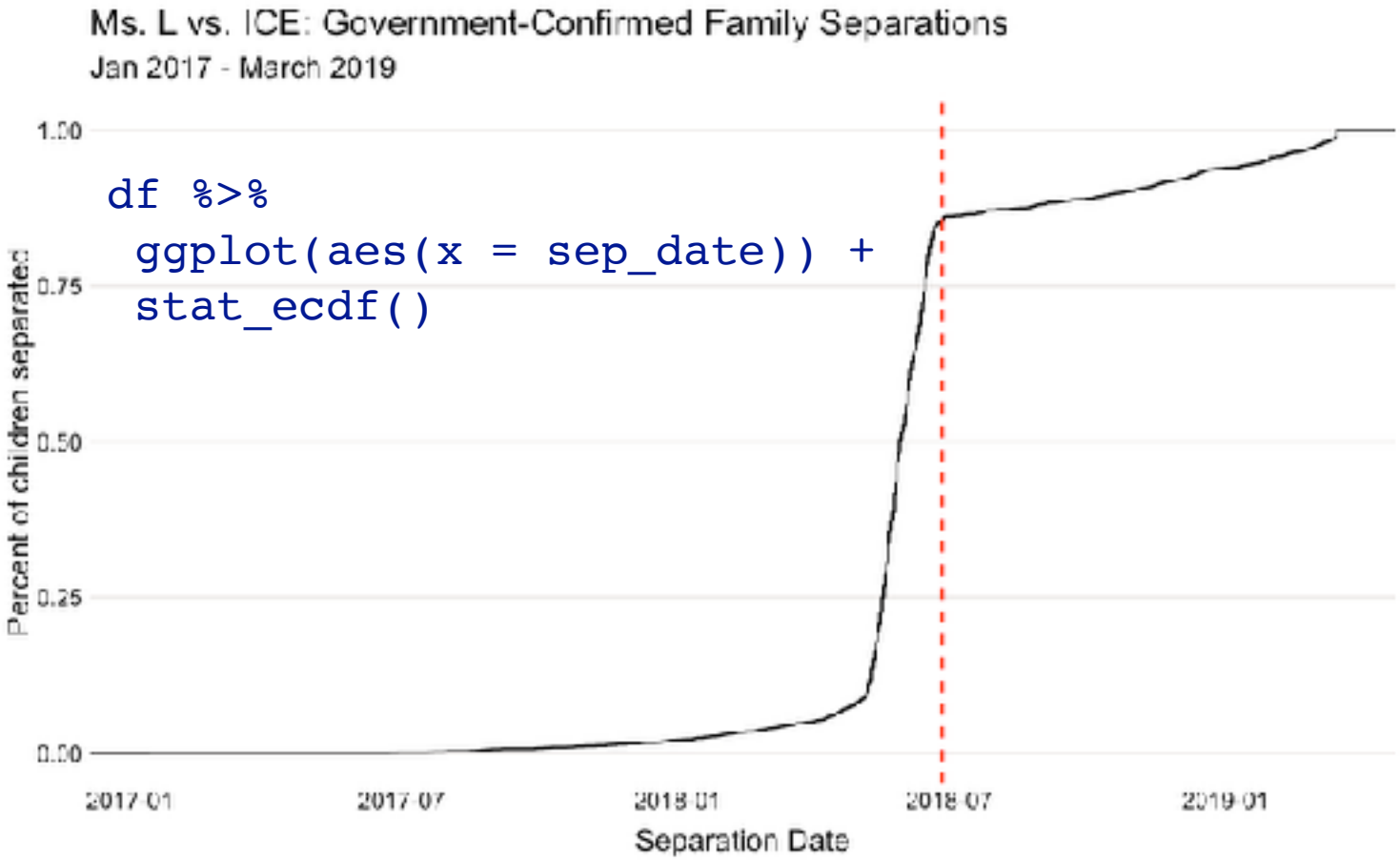
- Histo or bar chart complement
- Useful when comparing the proportional makeup of differently-sized groups

Separations were occurring as early as **July 1, 2017.** The government has now been ordered to ID and reunite these families.



At LEAST **389** separations have occurred since the federal injunction on June 26, 2018

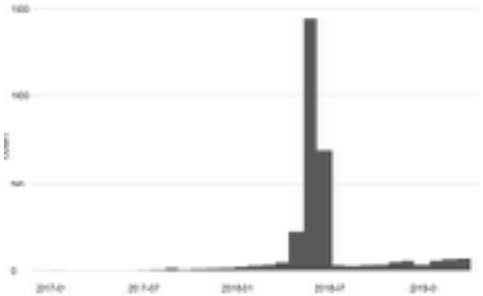

Ms. L vs. ICE: Government-Confirmed Family Separations
Jan 2017 - March 2019
Percent of children separated
1.00
0.75
0.50
0.25
0.00
2017 01
2017 07
2018 01
2018 07
2019 01
Separation Date
Federal Injunction
June 26, 2018


Ms. L vs. ICE: Government-Confirmed Family Separations
Jan 2017 - March 2019
df %>%
ggplot(aes(x = sep_date)) +
stat_ecdf()
1.00
0.75
0.50
0.25
0.00
Percent of children separated
2017 01
2017 07
2018 01
2018 07
2019 01
Separation Date

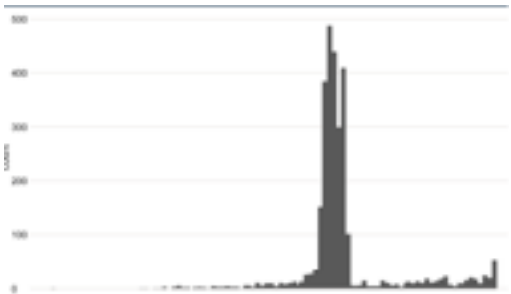
Histograms + ECDFs: three views of the same variable

```
ggplot(df, aes(x = date)) +
```

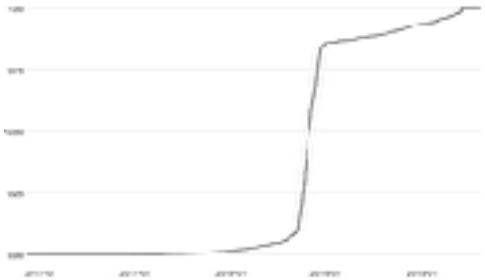
```
geom_histogram()
```



```
geom_histogram(
  bins = 100)
```



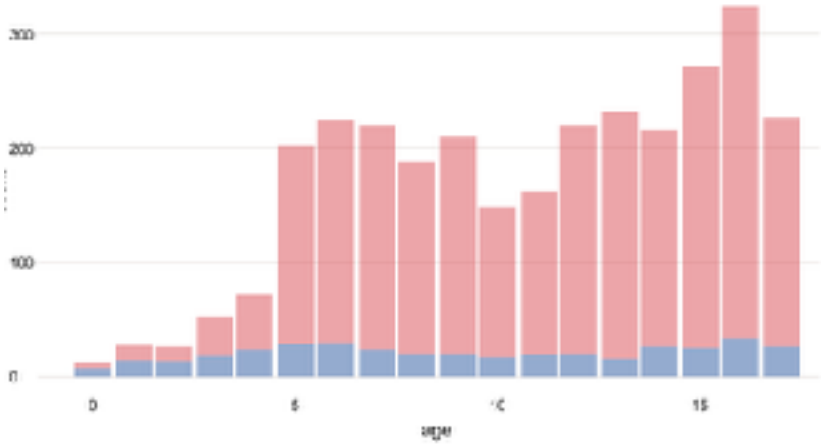
```
stat_ecdf()
```



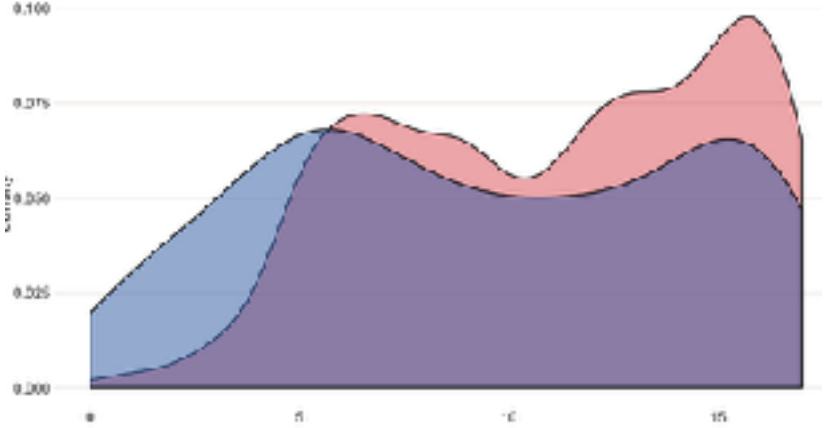
geom_bar + geom_density: two views of the same variable

```
ggplot(df, aes(x = age, fill = new)) +
```

geom_bar()

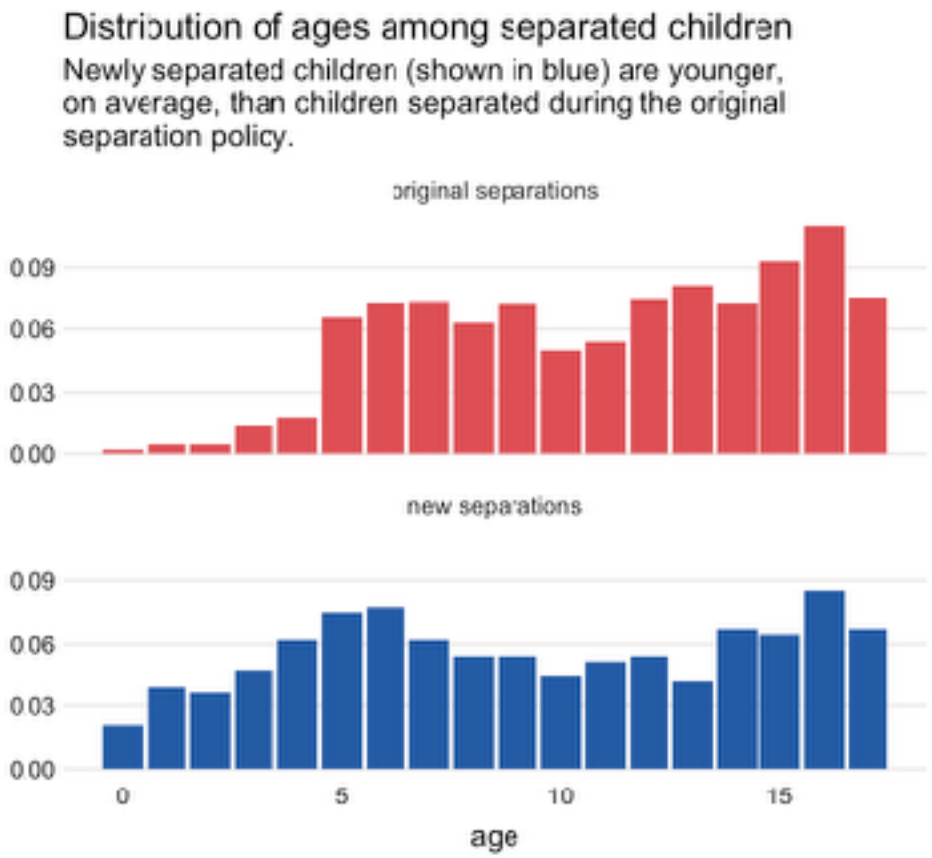


geom_density()

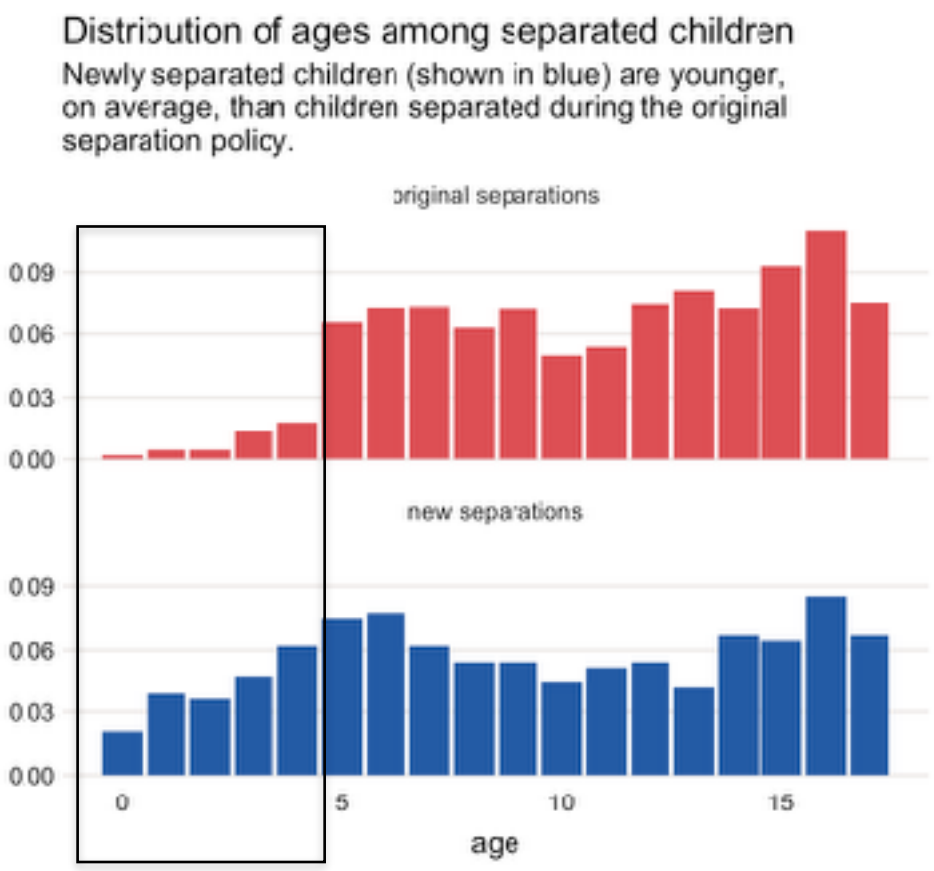


Different plot types lead to different understandings.

```
df %>%
  tabyl(age, new) %>%
  adorn_percentages("col") %>%
  gather(group, pct, -age) %>%
  ggplot(aes(x = age,
             y = pct,
             fill = group)) +
  geom_col() +
  facet_wrap(~group, ncol = 1)
```



- 3.9% of the originally identified separated children were under 5 years old.
- **More than 20% of children** separated since June 28, 2019 are under **5**.



Looking at data in different ways can deepen your understanding of a problem.

If you can look from many angles, you can look for many kinds of solutions.

Functions

tabulizer

1. locate_areas()
2. get_n_pages()
3. extract_tables()

daff

4. diff_data()
5. render_daff()

visdat

6. vis_miss()

readr

7. read_file()
8. write_file()

ggplot2

9. stat_ecdf()
10. geom_density()

Thank you

Github: brooke-watson

Twitter: @brooklynevery1

aclu.org

ACLU