Start with Data Science

as an Introduction to Statistical Thinking

Mine Çetinkaya-Rundel

Duke University + RStudio

@minebocek

mine-cetinkaya-rundel

mine@stat.duke.edu

audience

goal

a course that provides
a common (gateway) experience
to students wanting to get started with stats,
and that is

modern	place data front and center	quantitative (but without math prereqs)	different than HS stats	challenging, but not Intimidating

this course should...

emphasize modern and multivariate EDA + data visualization

start at the beginning of data analysis cycle with data collection and cleaning

encourage + enforce working collaboratively (think, code, write, present)

teach (not just expect) reproducible computing

approach statistics from a model based perspective

underscore effective communication of findings

...and more importantly

ask questions that students want to answer

equip students with the tools to answer questions of their own choosing


Sure!
Uh huh...


12345

1

rethink , don’t just add

don't start with this

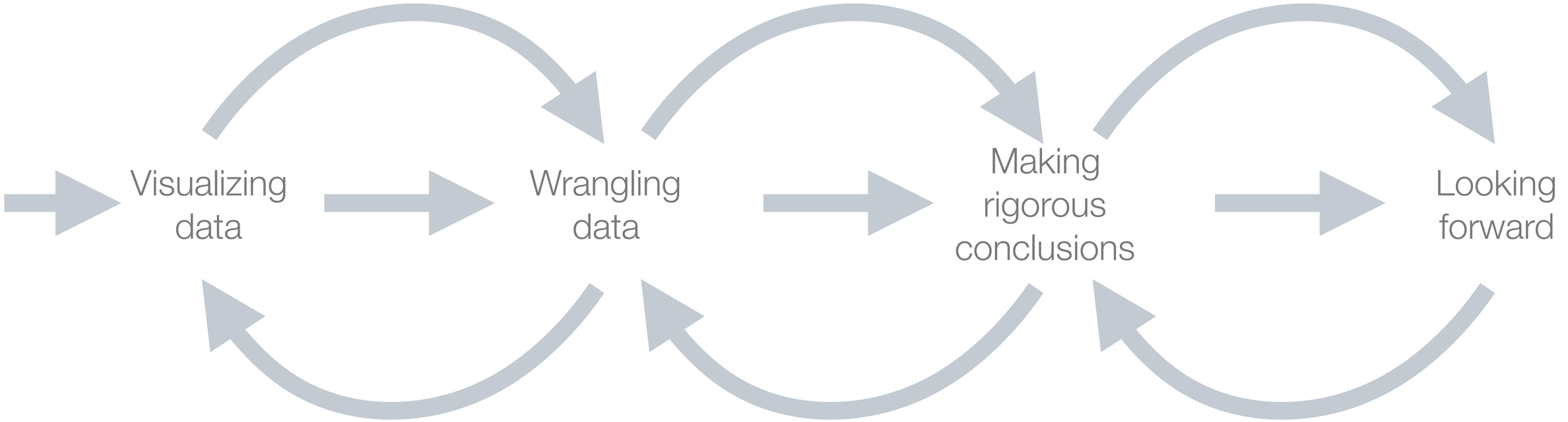
- Exploratory data analysis
- Study design
- Probability
- Random variables
- Central Limit Theorem
- One sample mean HT and CI
- One sample proportion HT and CI
- Two sample mean HT and CI
- Two sample proportion HT and CI
- Chi-square test
- ANOVA
- Simple linear regression

and add all this

+ R
+ R Markdown
+ git / GitHub
+ data scraping
+ iteration
+ working with non-rectangular data
+ interactive visualization

...

curriculum



Fundamentals of data & data viz, revision exercises, confounding variables and Simpson's paradox (and git/GitHub)

Tidy data, data frames vs. summary tables, recoding and transforming variables, web scraping and iteration

Building and selecting models, visualizing interactions, prediction and model validation, inference via simulation & discussion of CLT

Interactive visualization and reporting with Shiny, Bayesian inference, text analysis, ???

2

cherish

day

one

don't start like this

- Install R
- Install RStudio
- Install the following packages:
  - rmarkdown
  - tidyverse
  - ...
- Load these packages
- Install git

instead do this

- Go to rstudio.cloud (or some other server based solution)
- Log in with your ID & pass

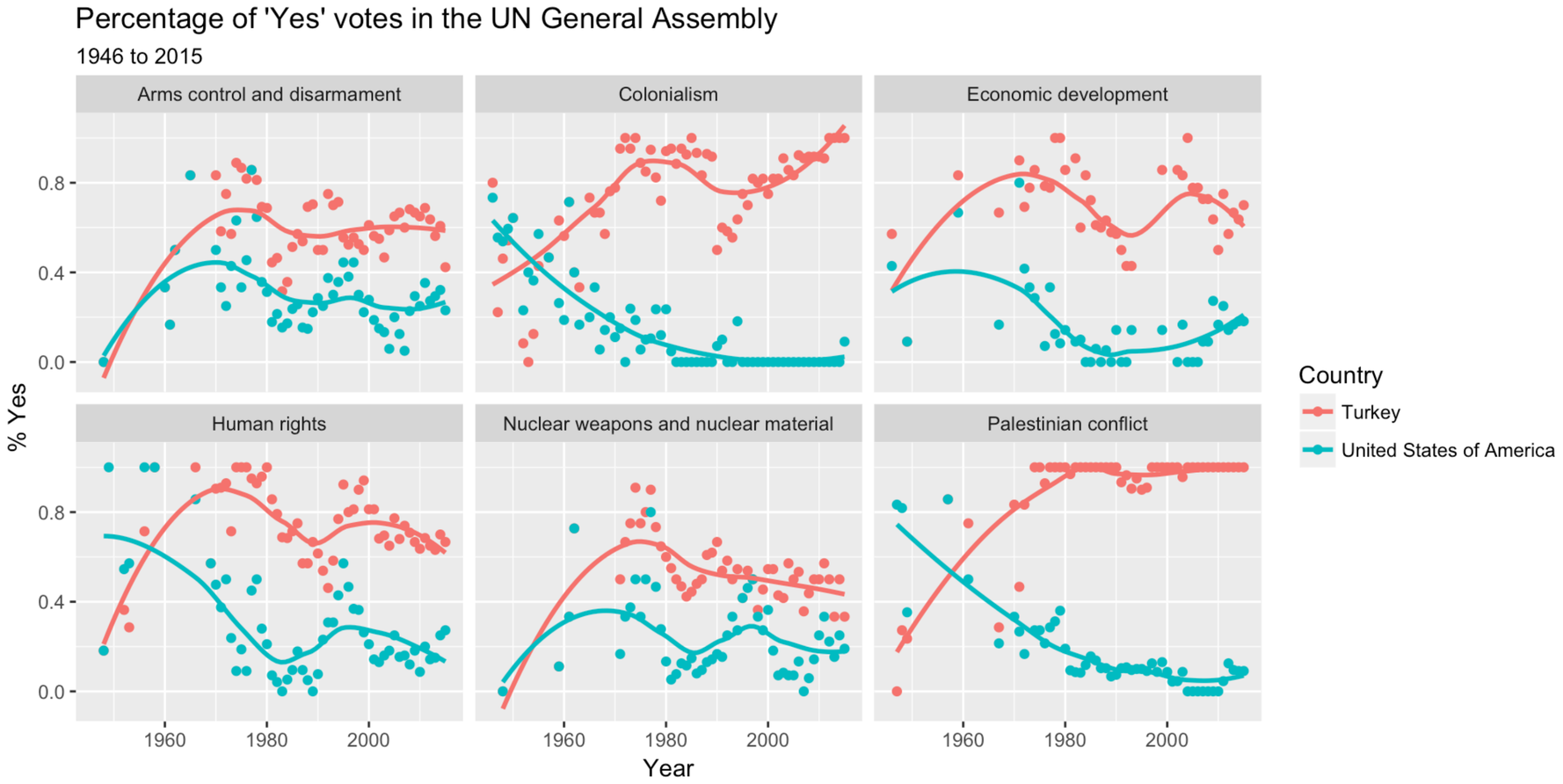
```
> hello R!
```

don’t start with this

```
class(mtcars$mpg)
#> [1] "numeric"
mean(mtcars$mpg)
#> [1] 20.09062
median(mtcars$mpg)
#> [1] 19.2
sd(mtcars$mpg)
#> [1] 6.026948
```

instead do this

- Open today's example project
- Knit the document and discuss the results with your neighbor



- Then, change countries plotted and knit again…

3

stick with a consistent grammar

base R

```r
# recode binary variable

mtcars$transmission <-
  ifelse(
    mtcars$am == 0,
    "automatic",
    "manual"
  )
```

tidyverse

```r
# recode binary variable

mtcars <- mtcars %>%
  mutate(
    transmission =
      case_when(
        am == 0 ~ "automatic",
        am == 1 ~ "manual"
      )
  )
```

base R

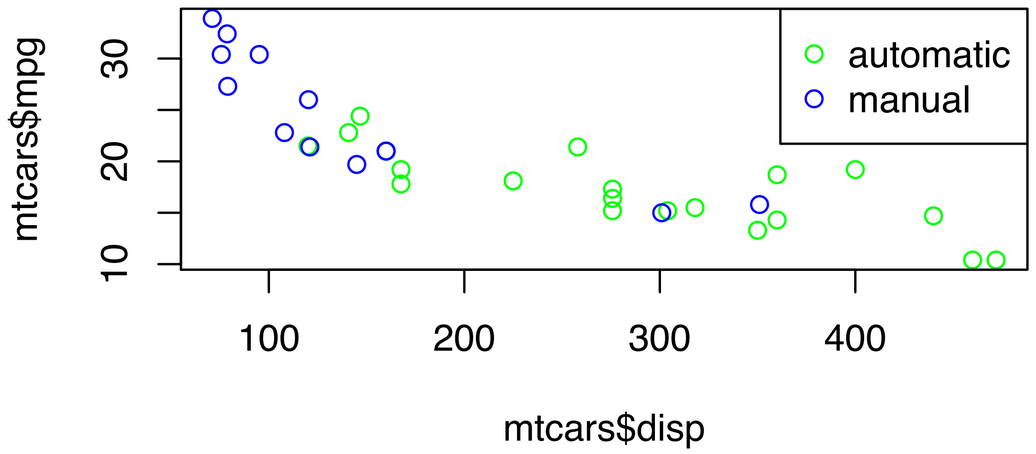
```r
# recode multi-level variable

mtcars$gear_char <-
  ifelse(
    mtcars$gear == 3,
    "three",
    ifelse(
      mtcars$gear == 4,
      "four",
      "five")
  )
```

tidyverse

```r
# recode multi-level variable

mtcars <- mtcars %>%
  mutate(
    gear_char =
      case_when(
        gear == 3 ~ "three",
        gear == 4 ~ "four",
        gear == 5 ~ "five"
      )
  )
```

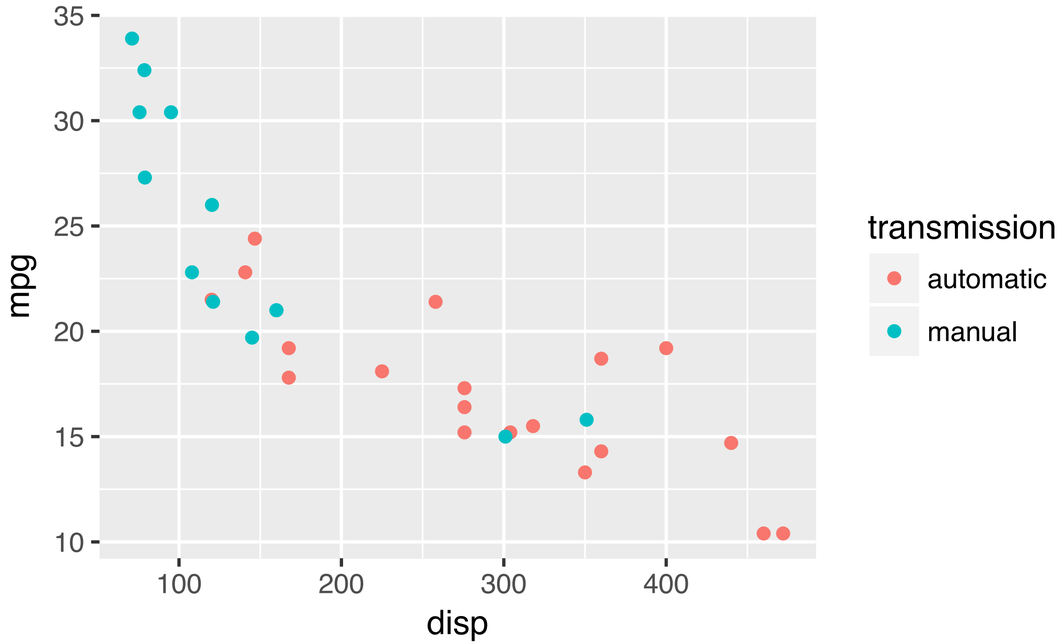
base R

```
# visualize three variables

mtcars$trans_color <-
  ifelse(mtcars$transmission == "automatic",
         "green",
         "blue")

plot(mtcars$mpg ~ mtcars$disp,
     col = mtcars$trans_color)
legend("topright",
       legend = c("automatic", "manual"),
       pch = 1, col = c("green", "blue"))
```



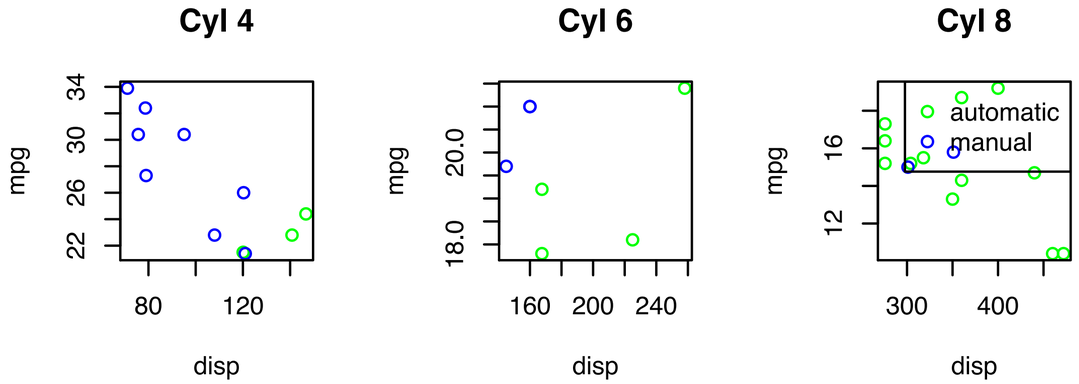
tidyverse

```
# visualize three variables

ggplot(mtcars,
       mapping = aes(
         x = disp, y = mpg,
         color = transmission
         )) +
  geom_point()
```



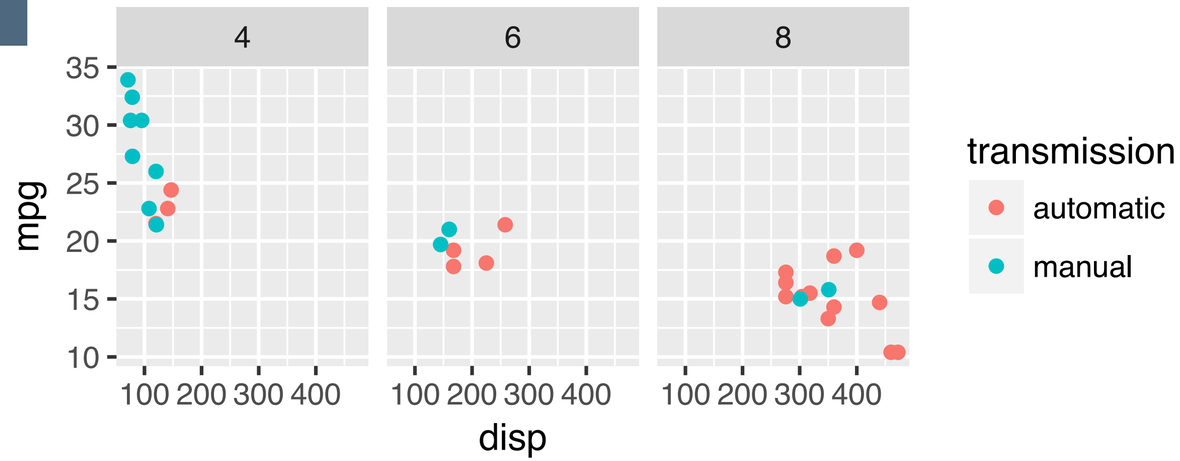
base R

```
# visualize four variables

mtcars_cyl4 <- mtcars[mtcars$cyl == 4, ]
mtcars_cyl6 <- mtcars[mtcars$cyl == 6, ]
mtcars_cyl8 <- mtcars[mtcars$cyl == 8, ]

par(mfrow = c(1, 3))
plot(mpg ~ disp, data = mtcars_cyl4,
     col = trans_color, main = "Cyl 4")
plot(mpg ~ disp, data = mtcars_cyl6,
     col = trans_color, main = "Cyl 6")
plot(mpg ~ disp, data = mtcars_cyl8,
     col = trans_color, main = "Cyl 8")
legend("topright",
       legend = c("automatic", "manual"),
       pch = 1, col = c("green", "blue"))
```



tidyverse

```
# visualize four variables

ggplot(mtcars,
       mapping = aes(
         x = disp, y = mpg,
         color = transmission
         )) +
  geom_point()
  facet_wrap(~ cyl)
```

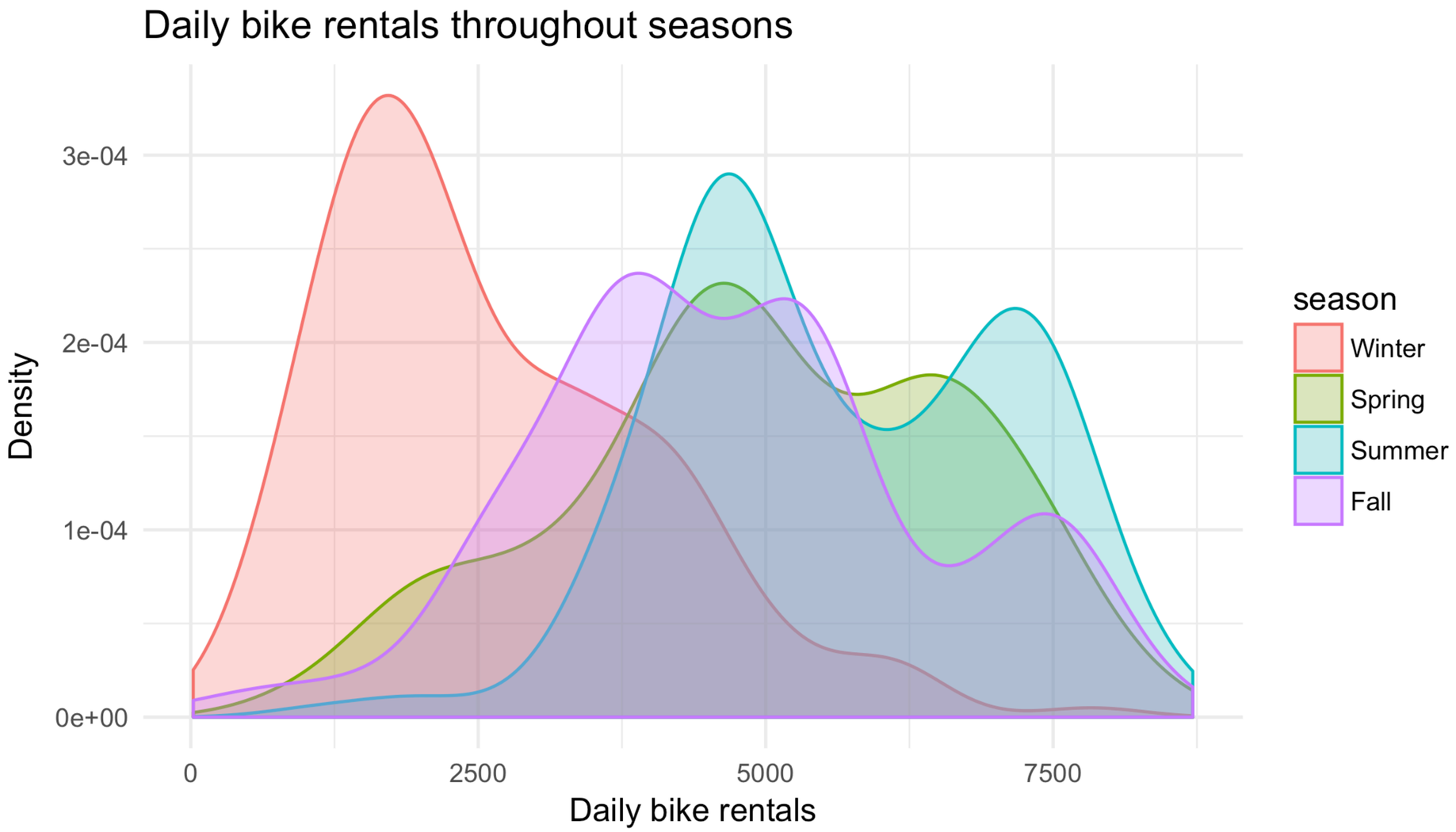


4

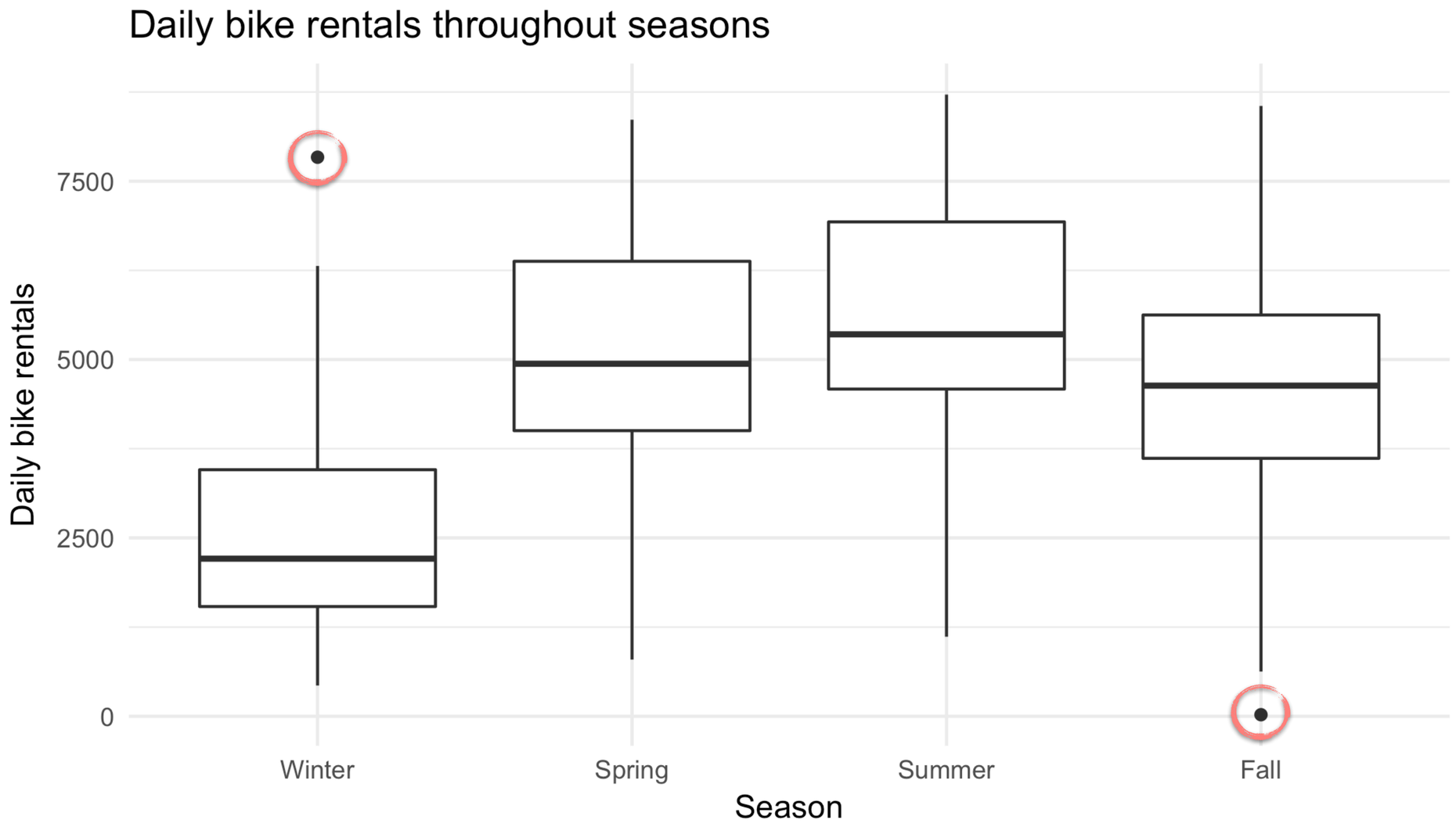
use real and relatable examples

citibikes in DC

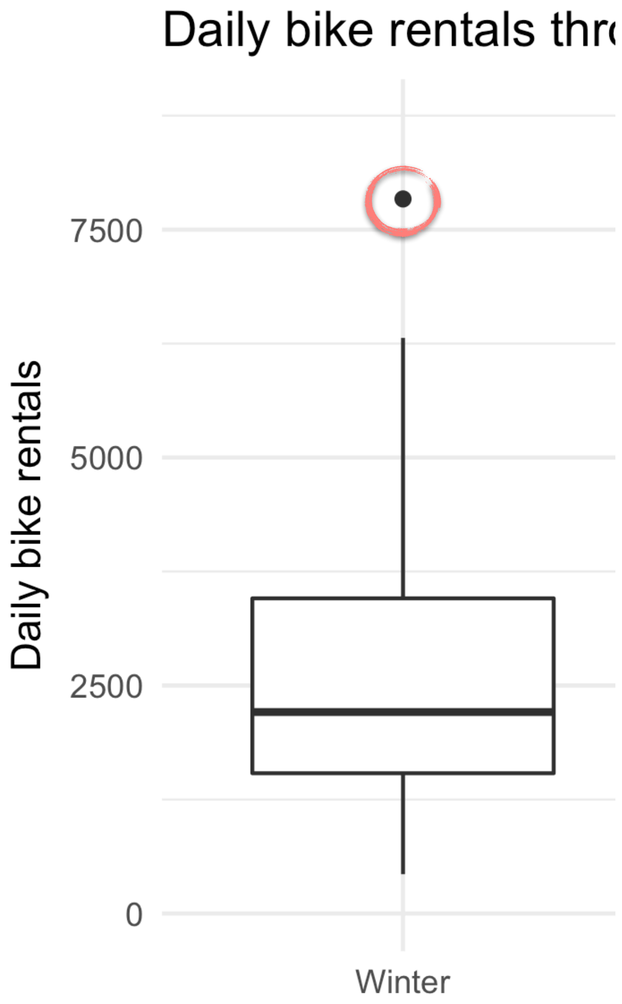
Question 8. Create a visualization displaying the relationship between bike rentals and season. Interpret the plot in context of the data.



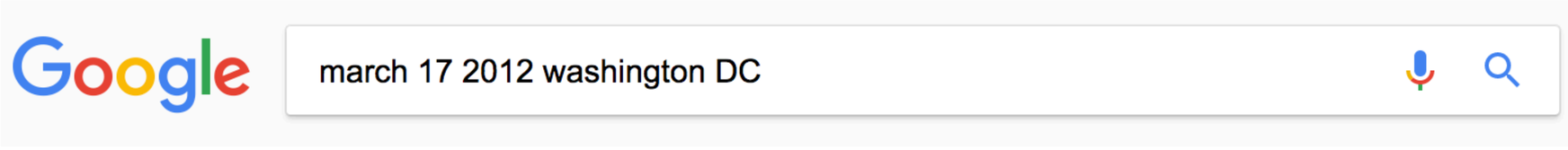
Question 8. Create a visualization displaying the relationship between bike rentals and season. Interpret the plot in context of the data.



Question 8. Create a visualization displaying the relationship between bike rentals and season. Interpret the plot in context of the data.

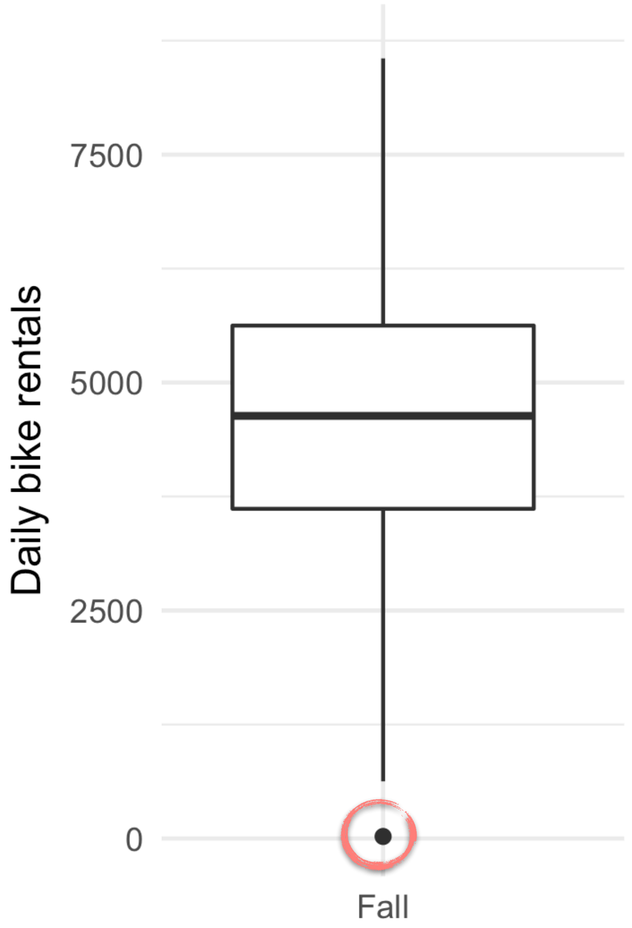



```
bike %>%
  filter(season == "Winter") %>%
  summarise(min = max(cnt), day_min = dteday[which.max(cnt)])
## # A tibble: 1 x 2
##     min day_min
##   <dbl> <date>
## 1  7836 2012-03-17
```

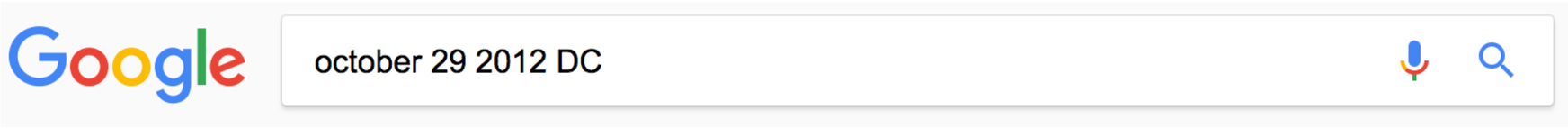


President Obama at the Dubliner on St. Patrick's Day | whitehouse.gov
https://obamawhitehouse.archives.gov/.../2012/.../17/president-obama-dubliner-st-patr...
17 Mar 2012 - President Barack Obama is reflected in a mirror at the Dubliner, an Irish pub in **Washington, D.C.**, with his Irish cousin, Henry Healy, and Ollie Hayes, a pub owner in Moneygall, Ireland, on St. Patrick's Day, **Saturday, March 17, 2012**. (Official White House Photo by Pete Souza). President Obama Greets the ...

Question 8. Create a visualization displaying the relationship between bike rentals and season. Interpret the plot in context of the data.




```
bike %>%
  filter(season == "Fall") %>%
  summarise(min = min(cnt), day_min = dteday[which.min(cnt)])
## # A tibble: 1 x 2
##     min day_min
##   <dbl> <date>
## 1  22.0 2012-10-29
```



Washington DC shuts down in preparation for hurricane Sandy | US ...
https://www.theguardian.com/.../2012/oct/29/washington-dc-shutdown-hurricane-sand...
28 Oct 2012 - Panic buying and fears over capability of power companies as first light rains reach city.

Hurricane Sandy: high winds and flooding hit US east coast – Monday ...
https://www.theguardian.com/.../us.../2012/oct/29/hurricane-sandy-new-york-live-blo...
30 Oct 2012 - Announces plans for a Storm Recovery Support Call Center operated by @ Delaware_DHSS will open tomorrow at noon. Gov. Markell: The message for tonight is to hunker down. Weather Underground (@wunderground). 470,000 without power in 13 states and **DC**. **October 29**,

learning goals

main
prediction and
model selection

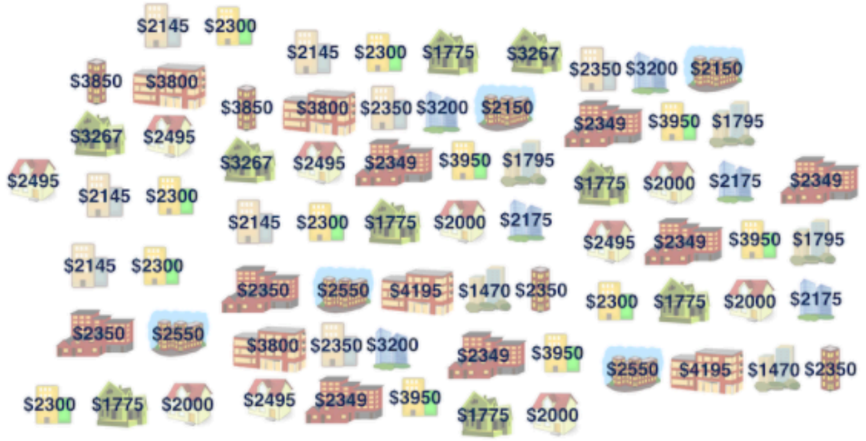
get for free
use of
outside data


manhattan apartments

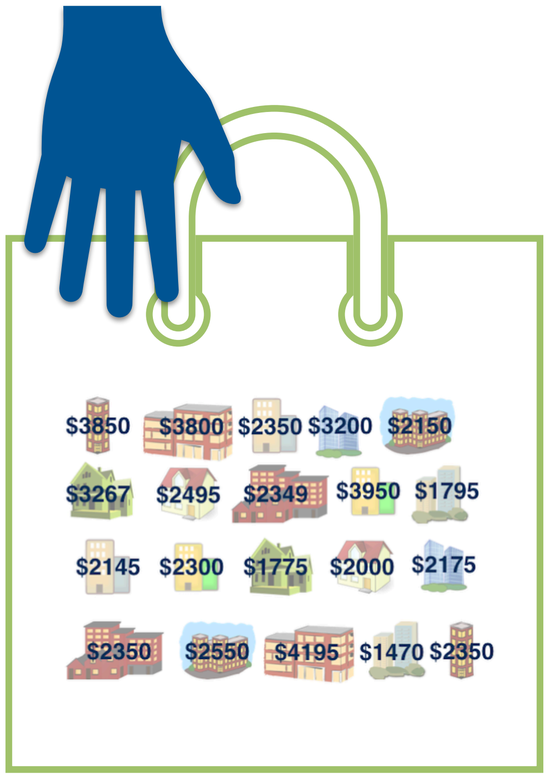
observed sample

Sample median = $2350 😱

population



Population median = ❓



Sample:

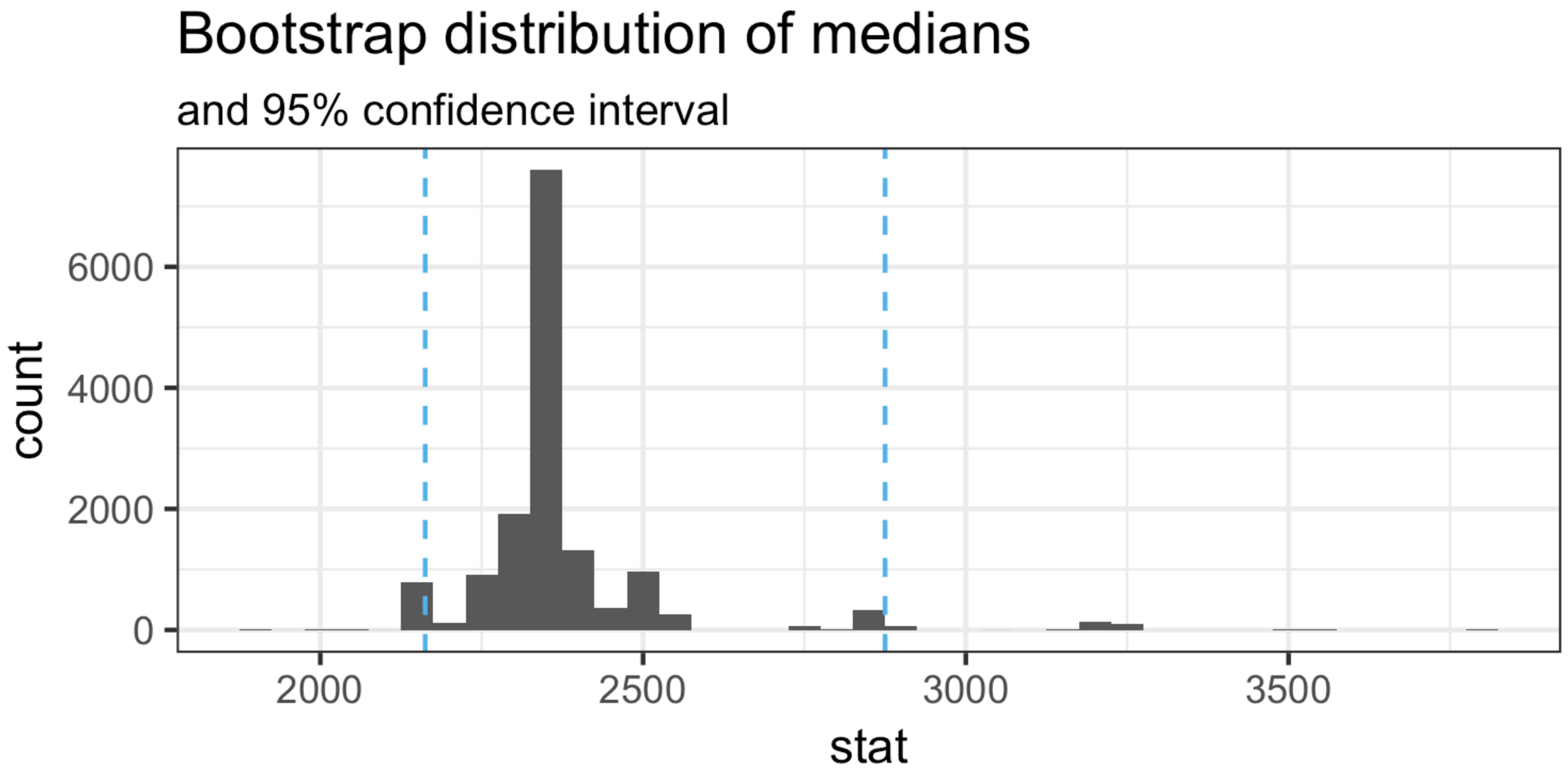
1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20

Ordered sample:

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20

Bootstrap median:

```r
library(infer)

manhattan %>%

  # specify the variable of interest
  specify(response = rent) %>%

  # generate 15000 bootstrap samples
  generate(reps = 15000, type = "bootstrap") %>%

  # calculate the median of each bootstrap sample
  calculate(stat = "median")
```


Bootstrap distribution of medians
and 95% confidence interval
count
6000
4000
2000
0
2000
2500
3000
3500
stat

learning goals

main
estimation
via
bootstrapping

get for free
discussion on
representativeness
of samples

5

teach tools for good science

literate programming

version control

+ transparent commit history

1 rethink, don't just add

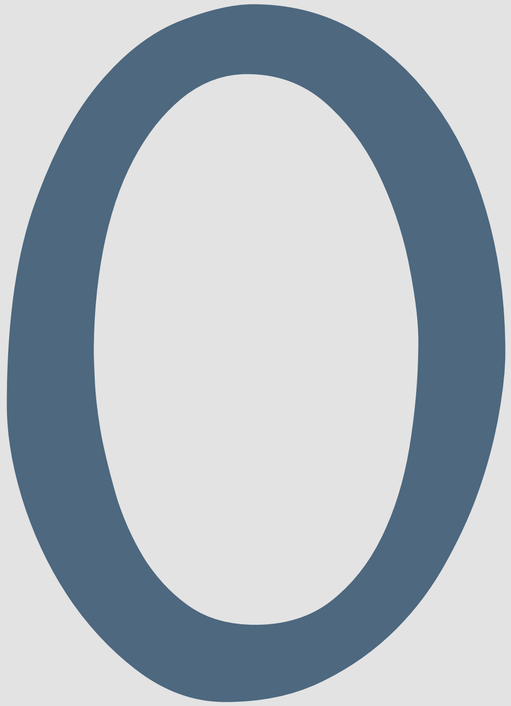
2 cherish day one

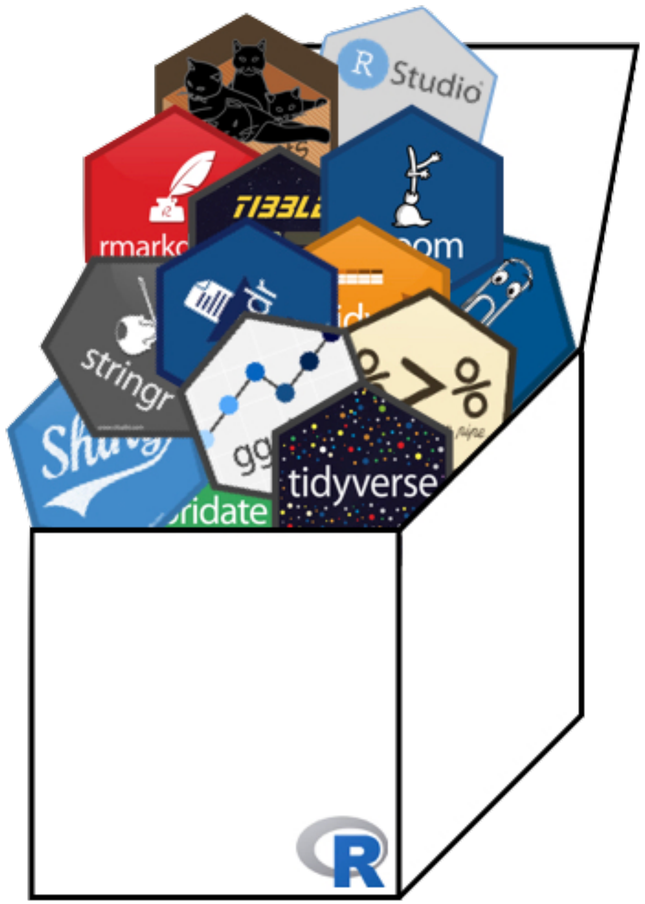
3 stick with a consistent grammar

4 use real and relatable examples

5 teach tools for good science

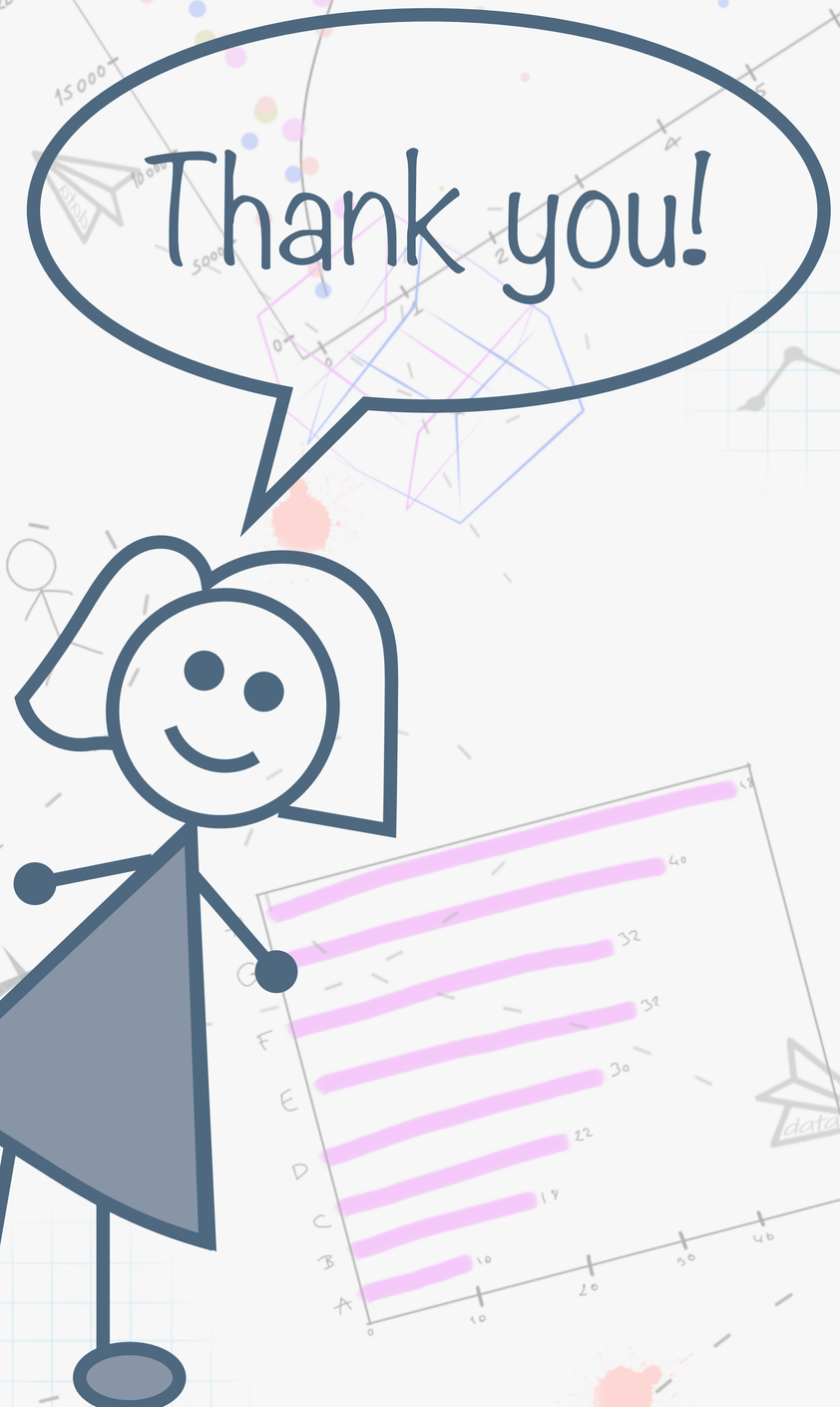
you're not alone





bit.ly/ds-box

```
Data Science Course in a Box
├── assignments
├── exams
├── labs
├── exams
├── slides
├── tutorials (WIP)
├── book (WIP)
│   ├── overview of materials
│   ├── computing infrastructure
├── dsbox-package (WIP)
```



bit.ly/ds-box

@minebocek

mine-cetinkaya-rundel

mine@stat.duke.edu