

# Data Science

as a gateway to Statistics

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

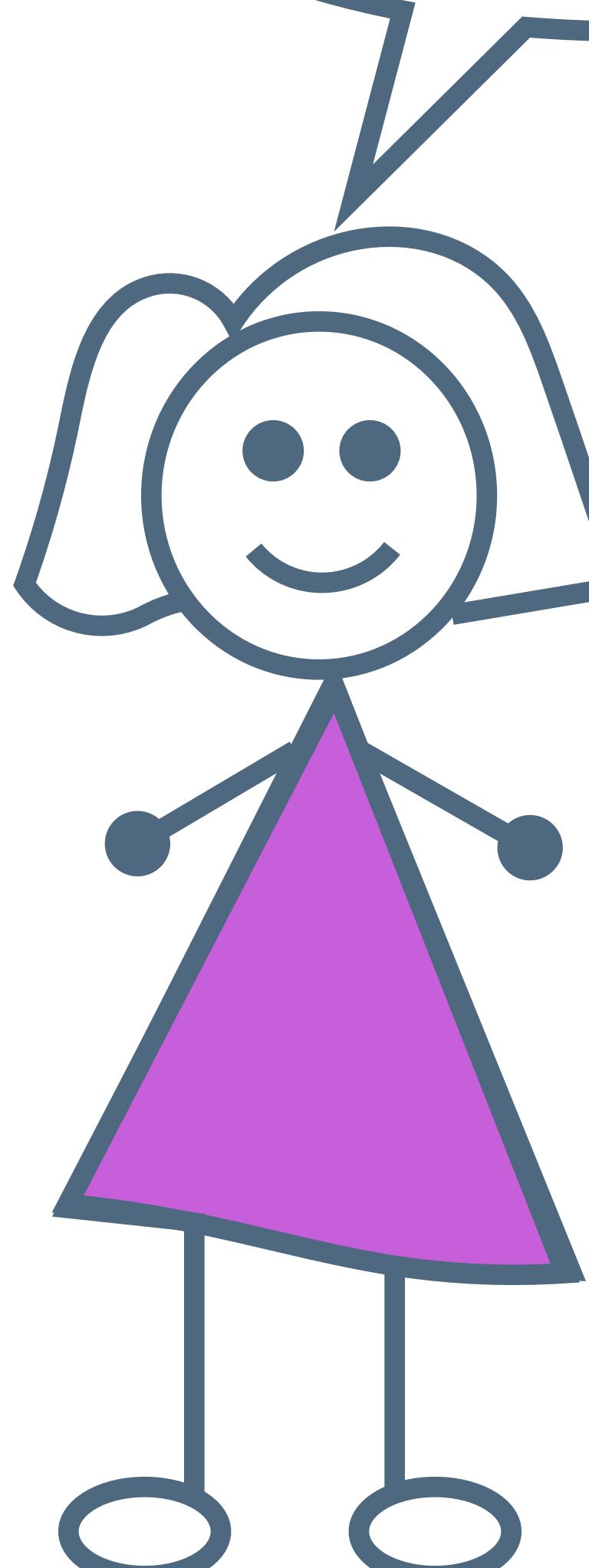
approach statistics  
from a model  
based perspective

underscore  
effective  
communication  
of findings

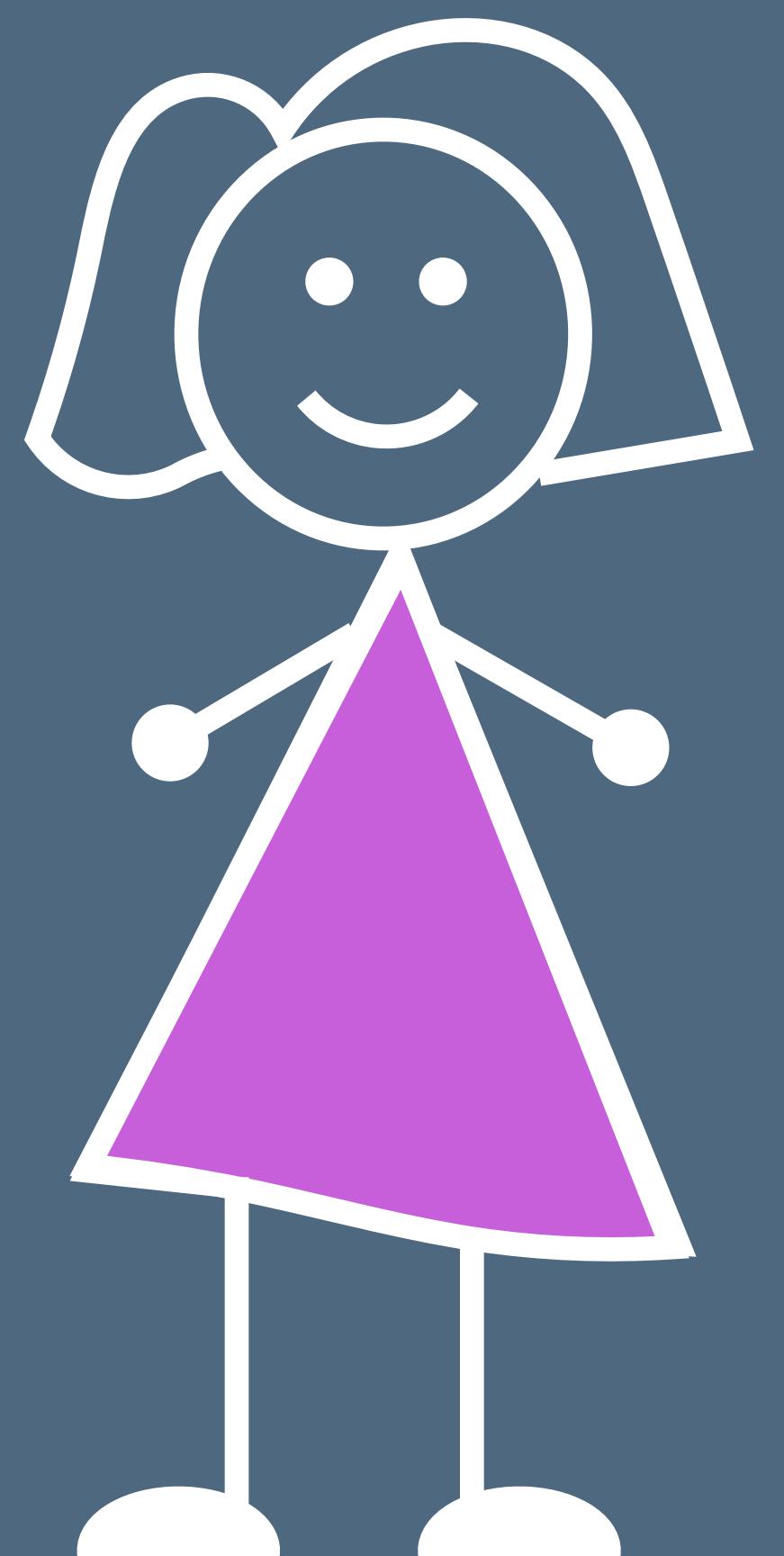
## ...and more importantly

ask questions that  
students want to  
answer

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



Sure!



Uh huh...

1

2

3

4

5

1

**rethink ,  
don't just  
add**

# don't start with this

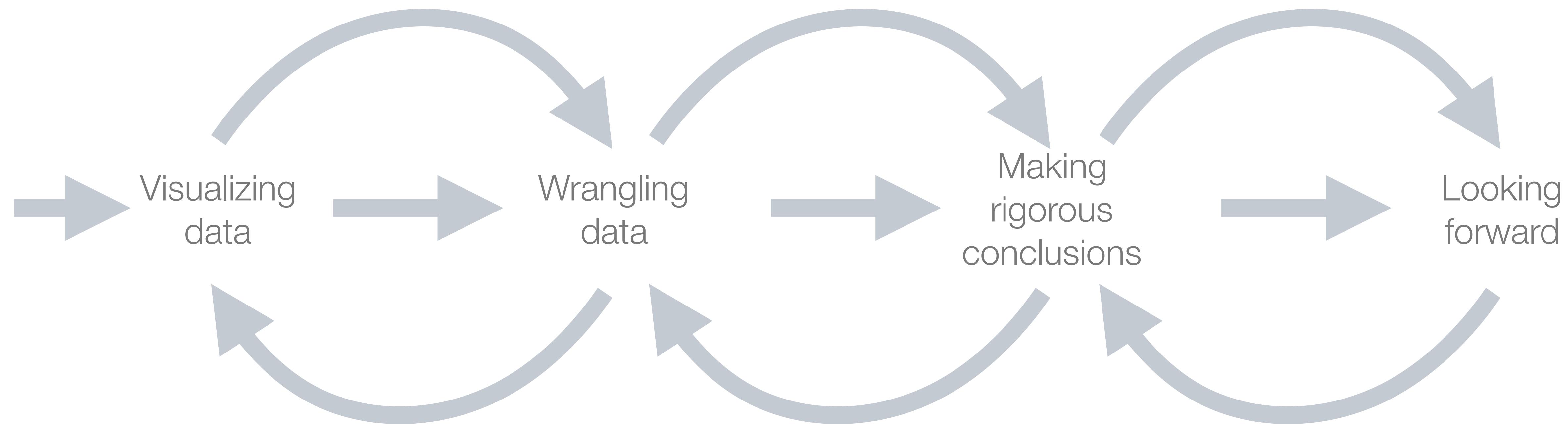
- Exploratory data analysis
- Study design
- Probability
- Random variables
- Central Limit Theorem
- One sample mean HT and CI
- One sample proportion HT and CI
- Two sample mean HT and CI
- Two sample proportion HT and CI
- Chi-square test
- ANOVA
- Simple linear regression

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, ???

## **structure:**

teams: in class  
exercises + projects  
individual: HW +  
two take home  
midterms

## **assessment:**

not just final work but  
also the process,  
peer evaluations and  
contribution  
diagnostics

2

cherish  
day  
one

## don't Start like this

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

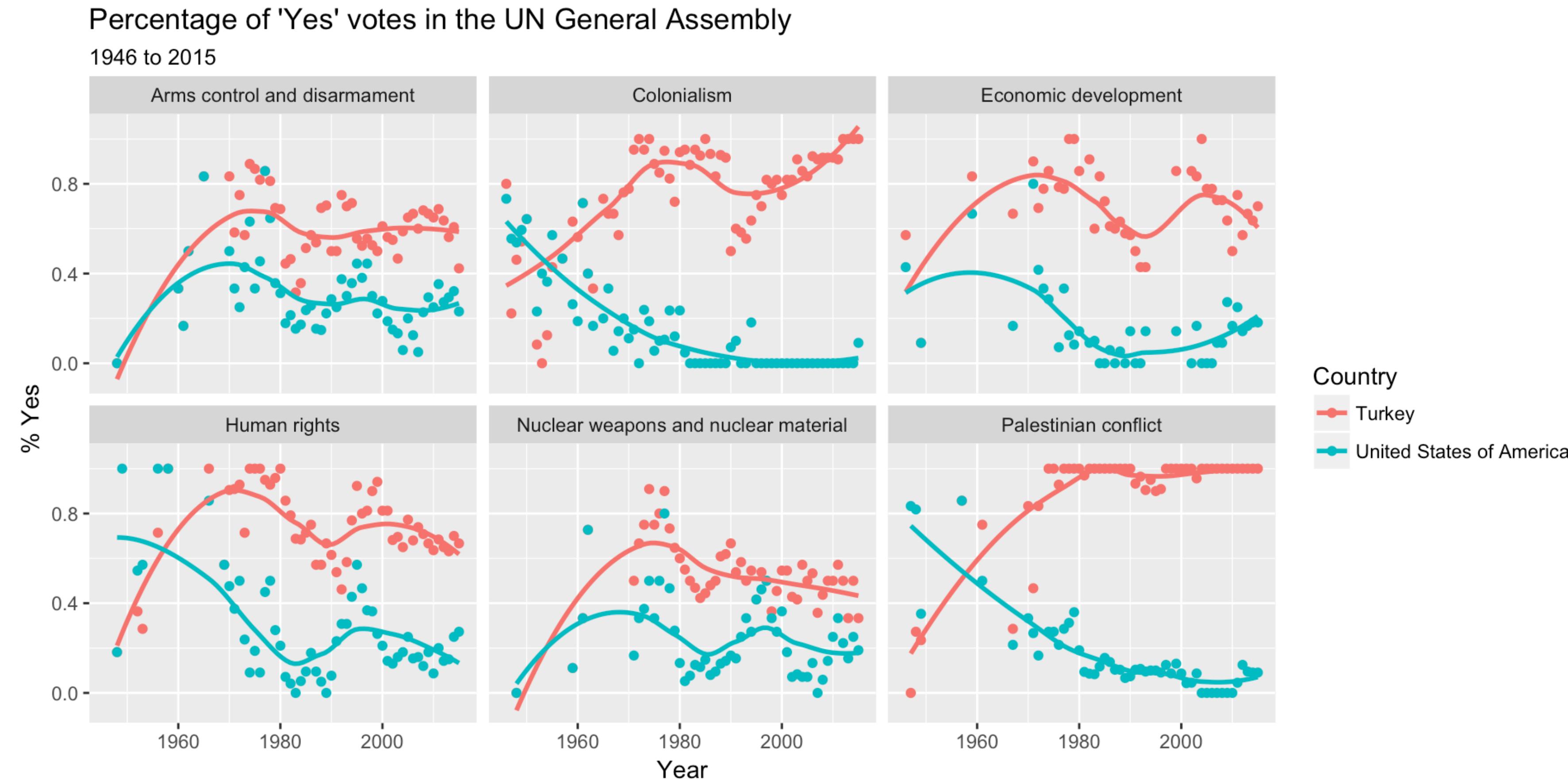
instead do this

- Go to [rstudio.cloud](https://rstudio.cloud) (or some other server based solution)
  - Log in with your ID & pass
- > hello R!

```
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

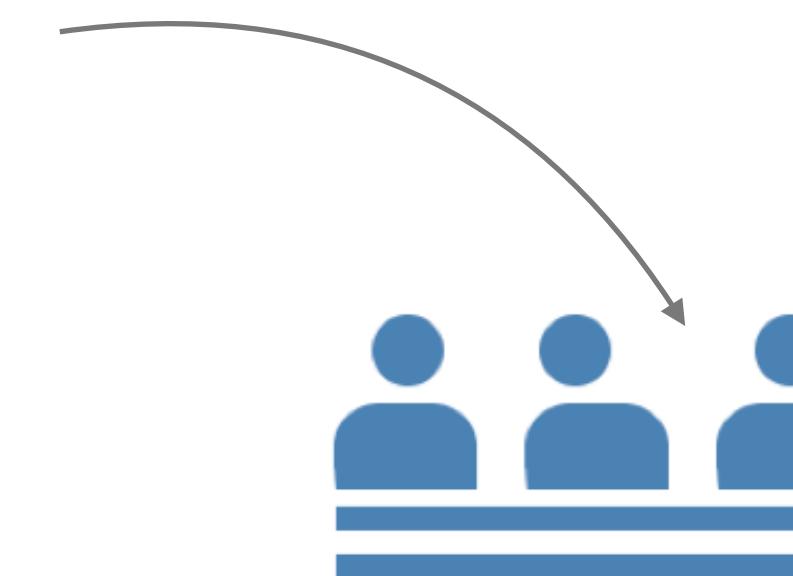


- Then, change countries plotted and knit again...

# logistics

~min 10 of first class  
give students a short link  
to join the workspace

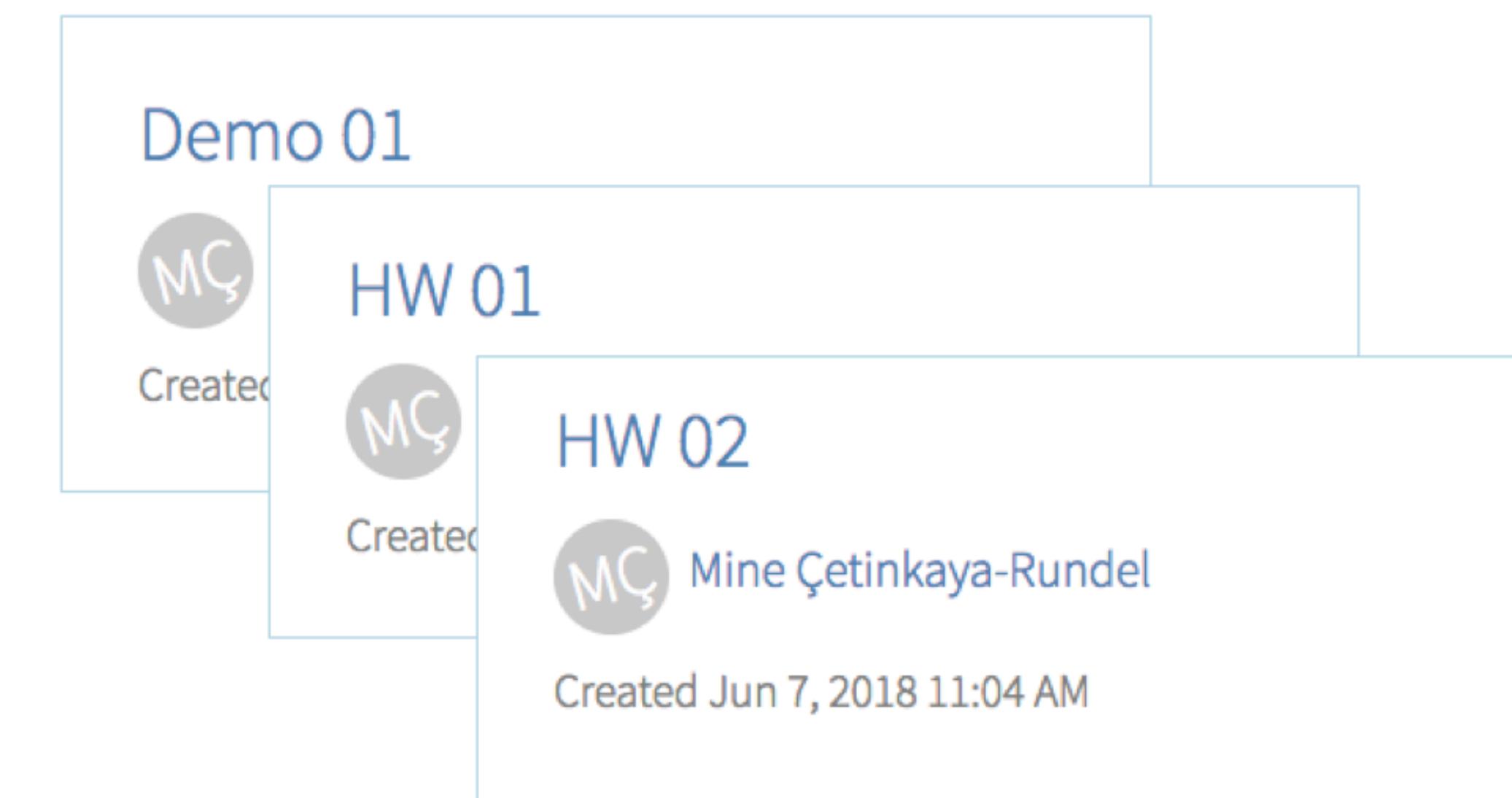
## WORKSPACE



## CLASS

students make a copy of project,  
which comes with all  
required packages installed  
+ Rmd of demo analysis

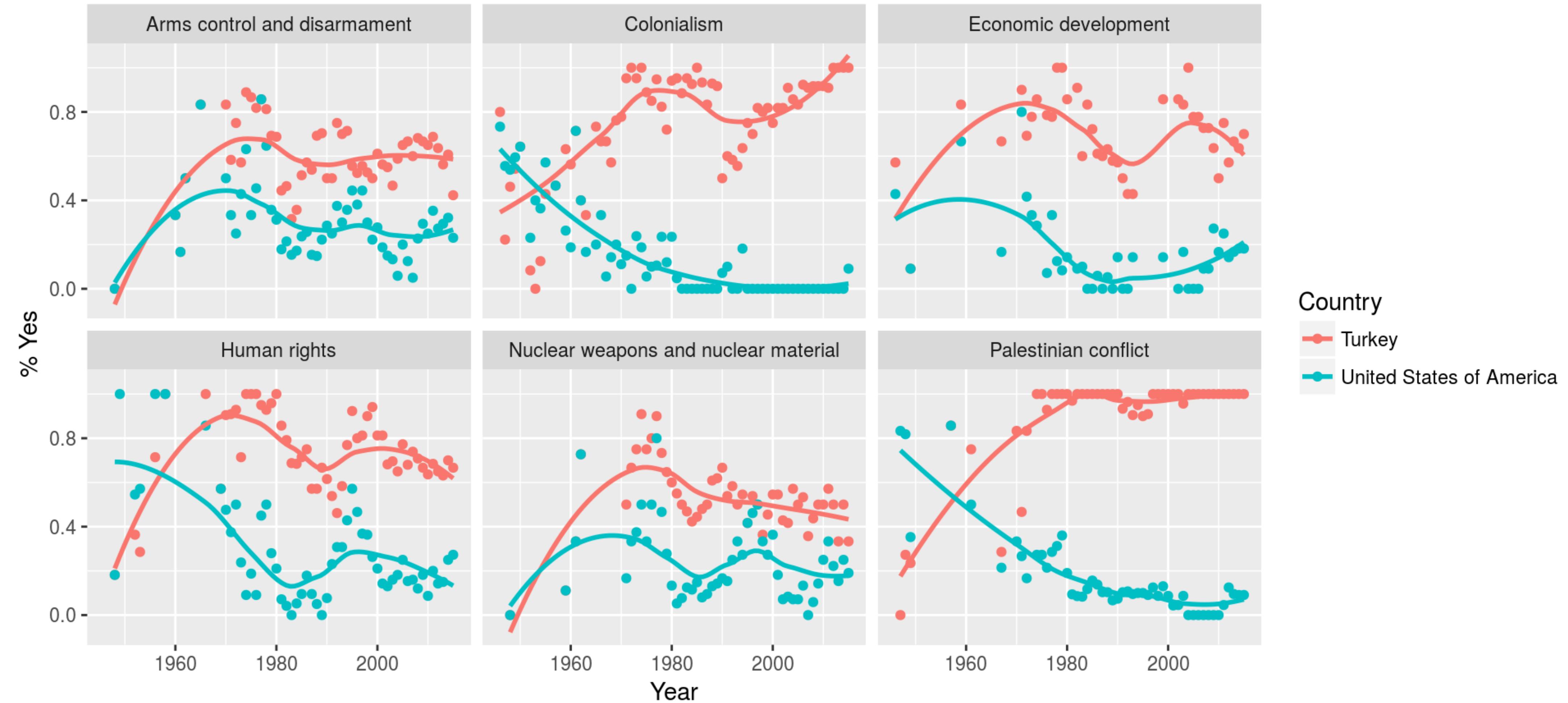
## PROJECT



## ASSIGNMENT

## Percentage of 'Yes' votes in the UN General Assembly

1946 to 2015



```
un_votes %>%
  filter(country %in% c("United States of America", "Turkey")) %>%
  inner_join(un_roll_calls, by = "rcid") %>%
  inner_join(un_roll_call_issues, by = "rcid") %>%
  group_by(country, year = year(date), issue) %>%
  summarize(
    votes = n(),
    percent_yes = mean(vote == "yes")
  ) %>%
  filter(votes > 5) %>% # only use records where there are more than 5 votes
  ggplot(mapping = aes(x = year, y = percent_yes, color = country)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE) +
  facet_wrap(~ issue) +
  labs(
    title = "Percentage of 'Yes' votes in the UN General Assembly",
    subtitle = "1946 to 2015",
    y = "% Yes",
    x = "Year",
    color = "Country"
  )
```



**stick with a  
consistent  
grammar**

## base R

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

## tidyverse

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

## base R

```
# recode multi-level variable  
  
mtcars$gear_char <-  
  ifelse(  
    mtcars$gear == 3,  
    "three",  
    ifelse(  
      mtcars$gear == 4,  
      "four",  
      "five"))
```

## tidyverse

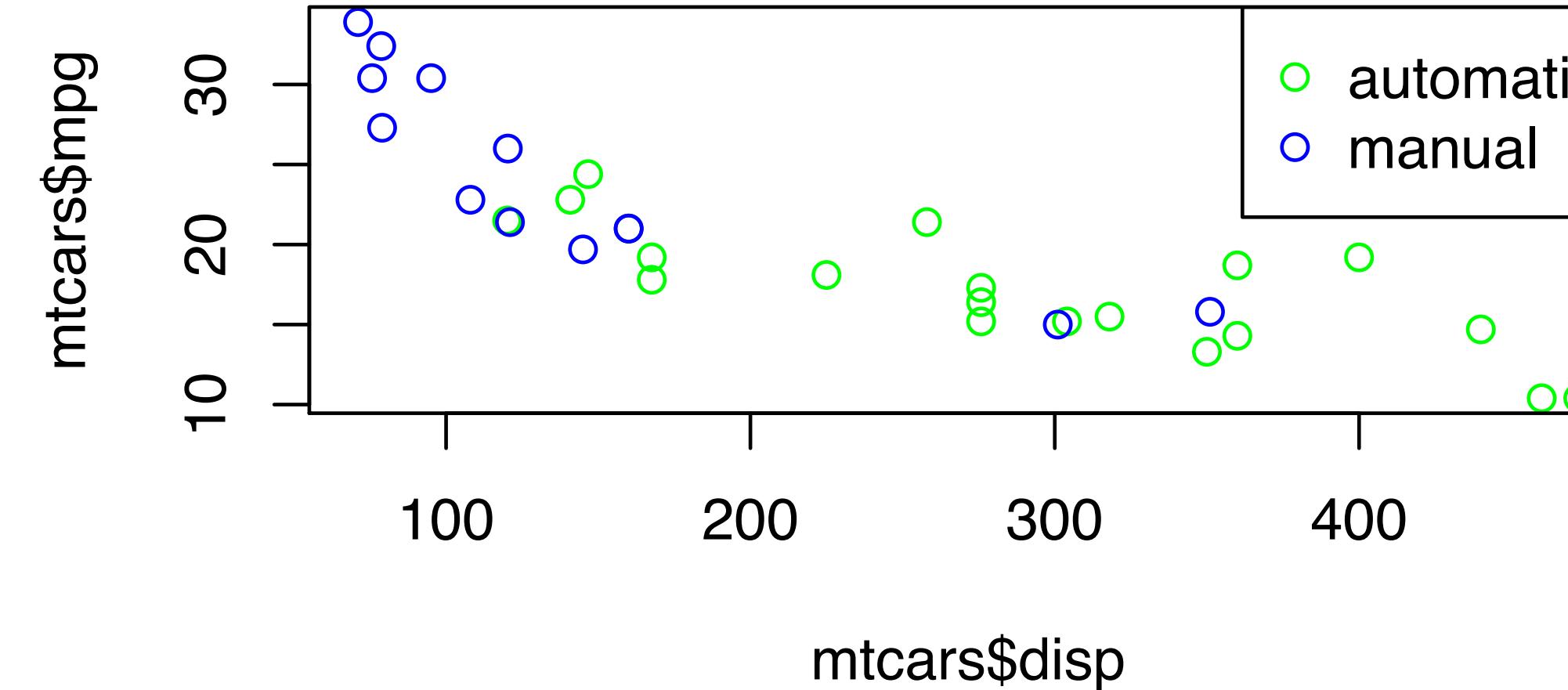
```
# recode multi-level variable  
  
mtcars <- mtcars %>%  
  mutate(  
    gear_char =  
      case_when(  
        gear == 3 ~ "three",  
        gear == 4 ~ "four",  
        gear == 5 ~ "five"))
```

# base R

```
# visualize three variables

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

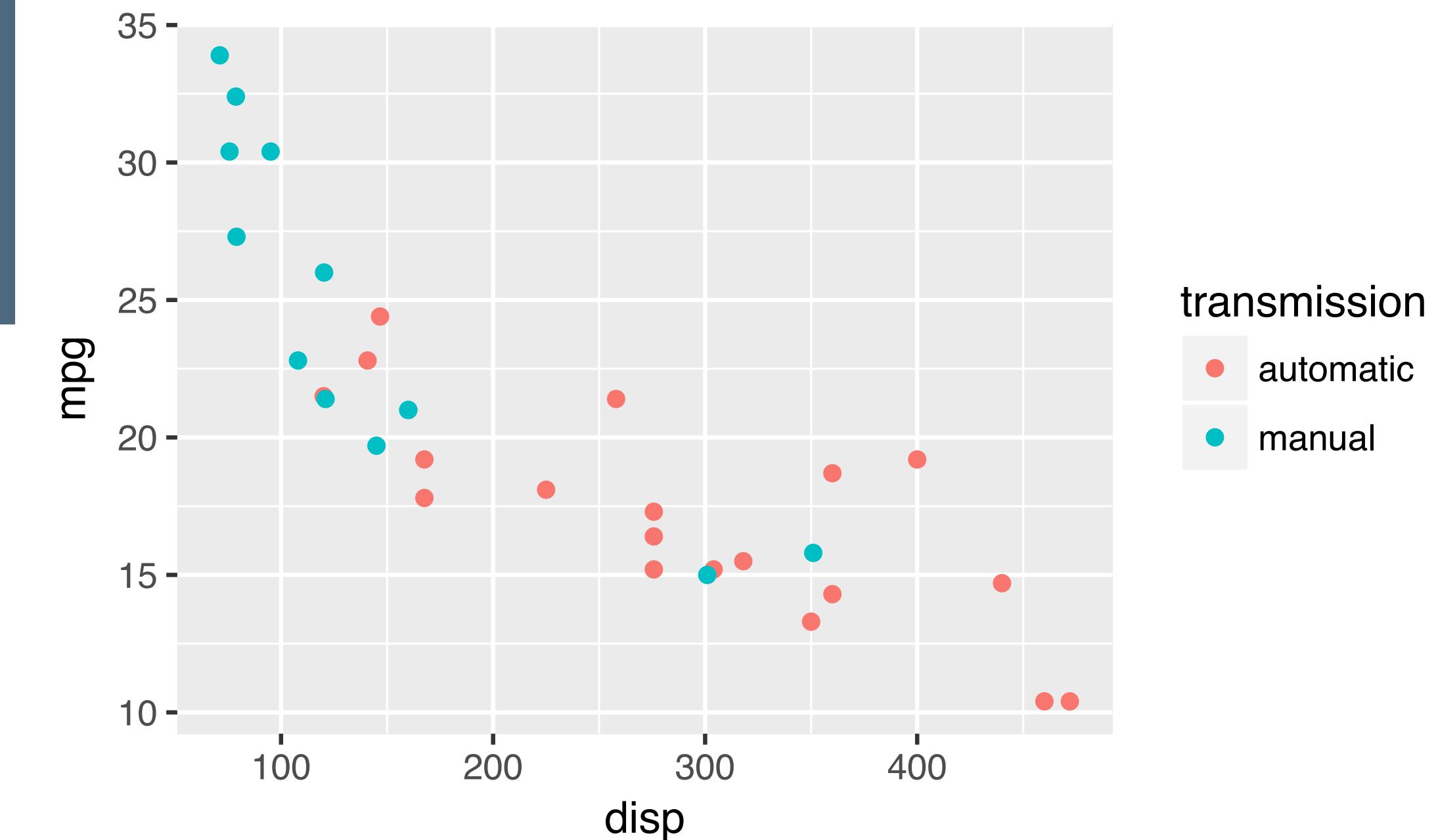
plot(mtcars$mpg ~ mtcars$disp,
     col = mtcars$trans_color)
legend("topright",
       legend = c("automatic", "manual"),
       pch = 1, col = c("green", "blue"))
```



# tidyverse

```
# visualize three variables

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

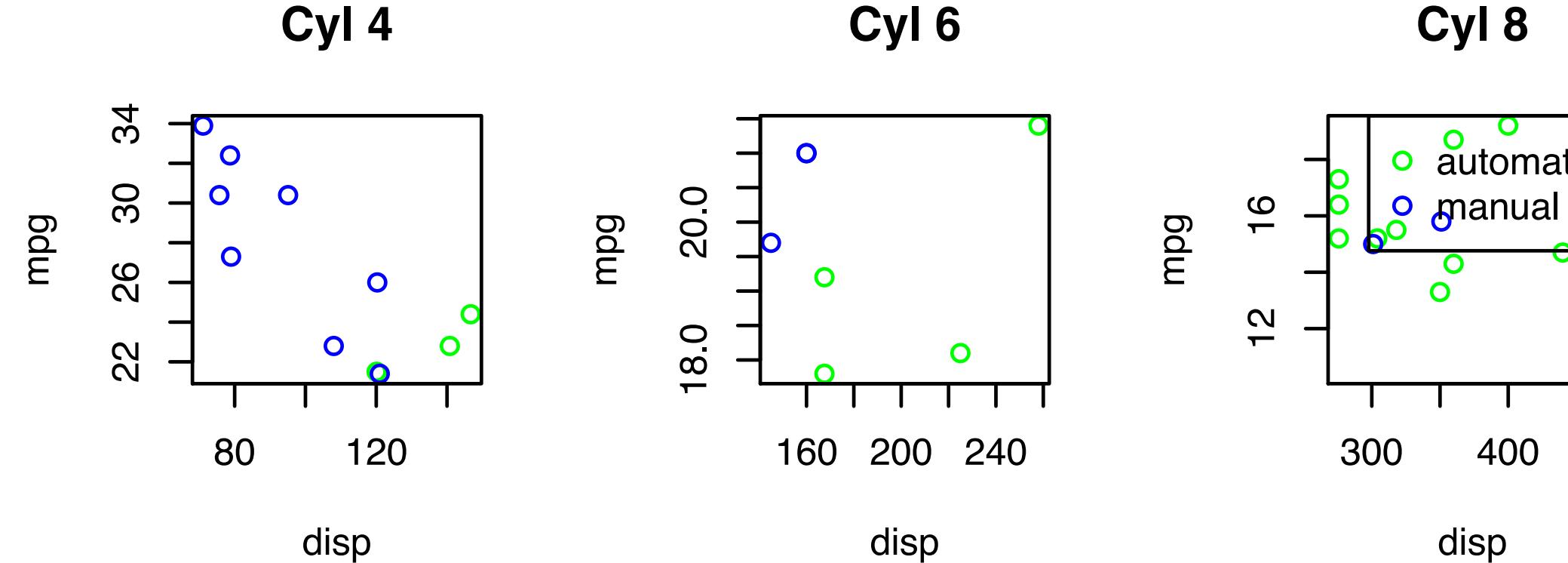


# base R

```
# visualize four variables

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

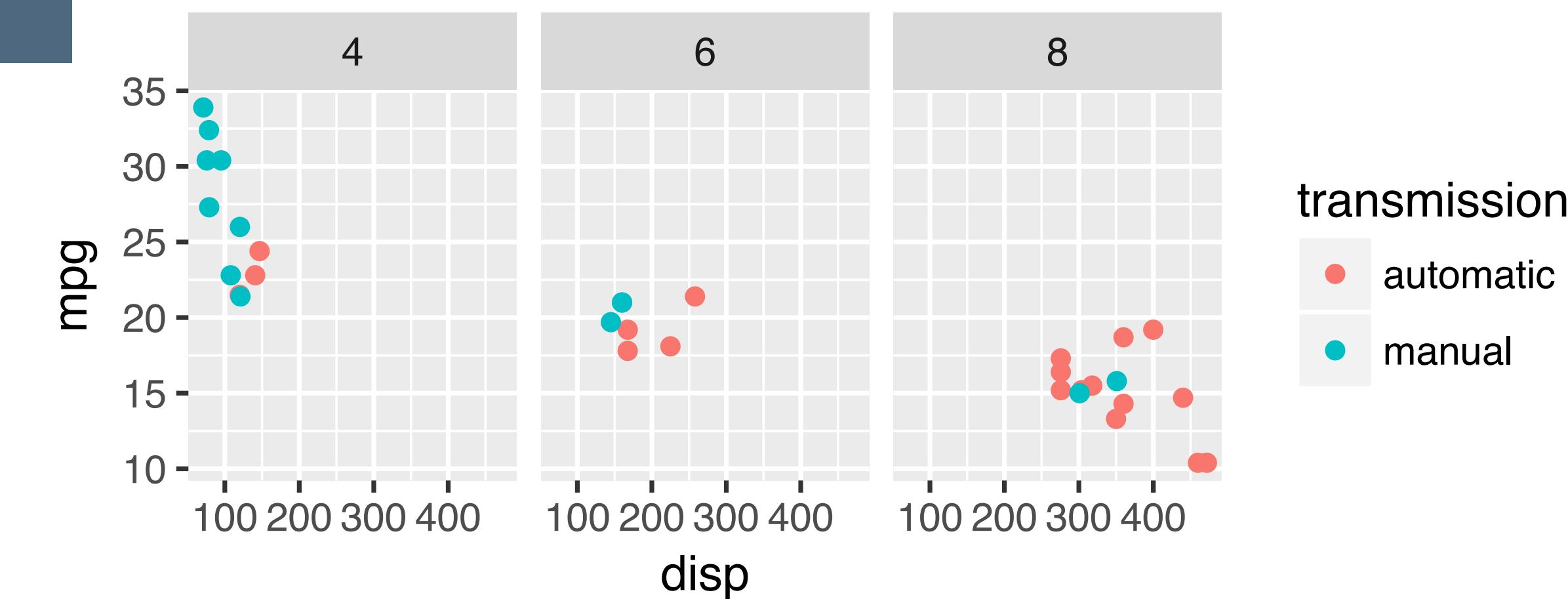
par(mfrow = c(1, 3))
plot(mpg ~ disp, data = mtcars_cyl4,
     col = trans_color, main = "Cyl 4")
plot(mpg ~ disp, data = mtcars_cyl6,
     col = trans_color, main = "Cyl 6")
plot(mpg ~ disp, data = mtcars_cyl8,
     col = trans_color, main = "Cyl 8")
legend("topright",
       legend = c("automatic",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



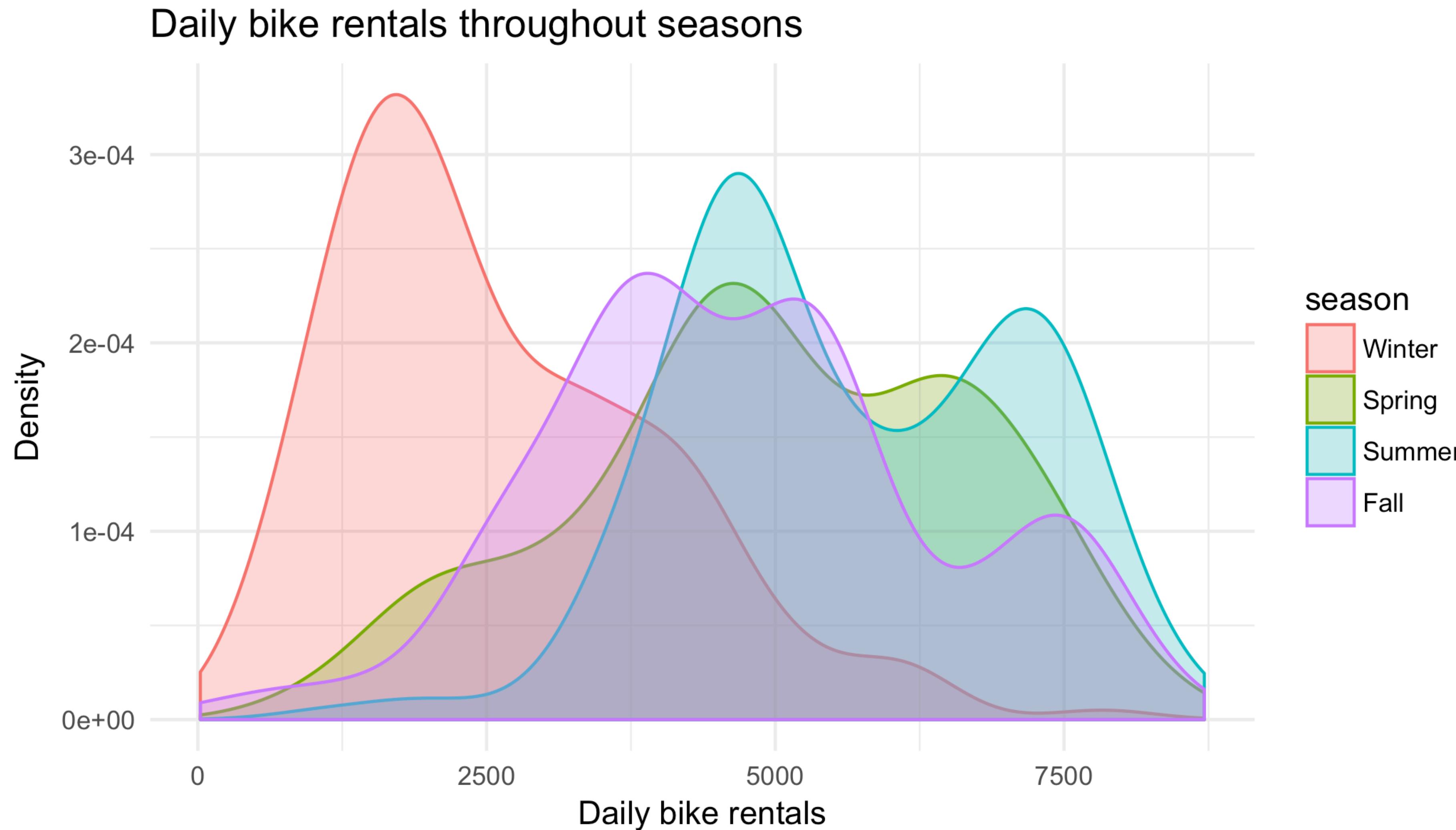


**use real and  
relatable  
examples**

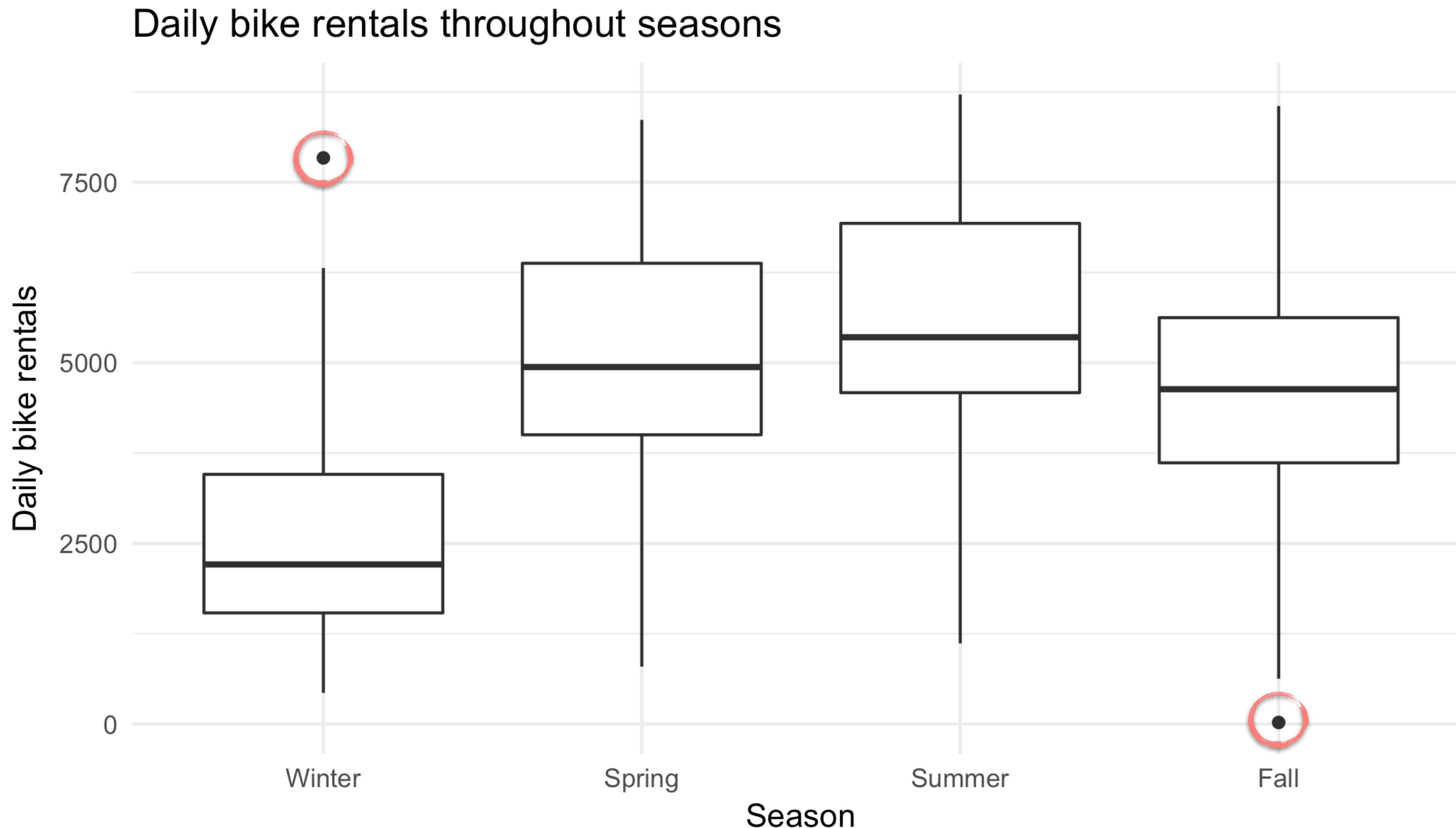
**citibikes** in DC



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



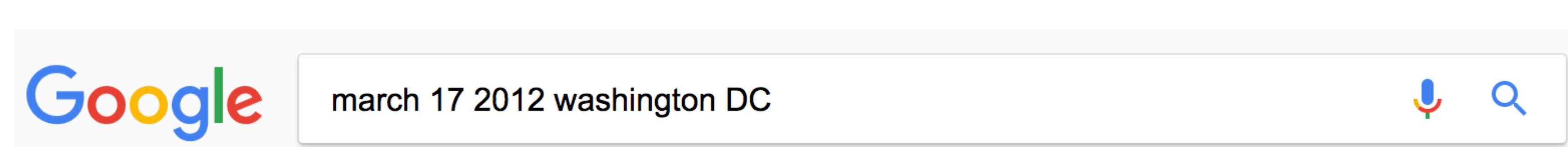
**Question 8.** Create a visualization displaying the relationship between bike rentals and season.  
Interpret the plot in context of the data.



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

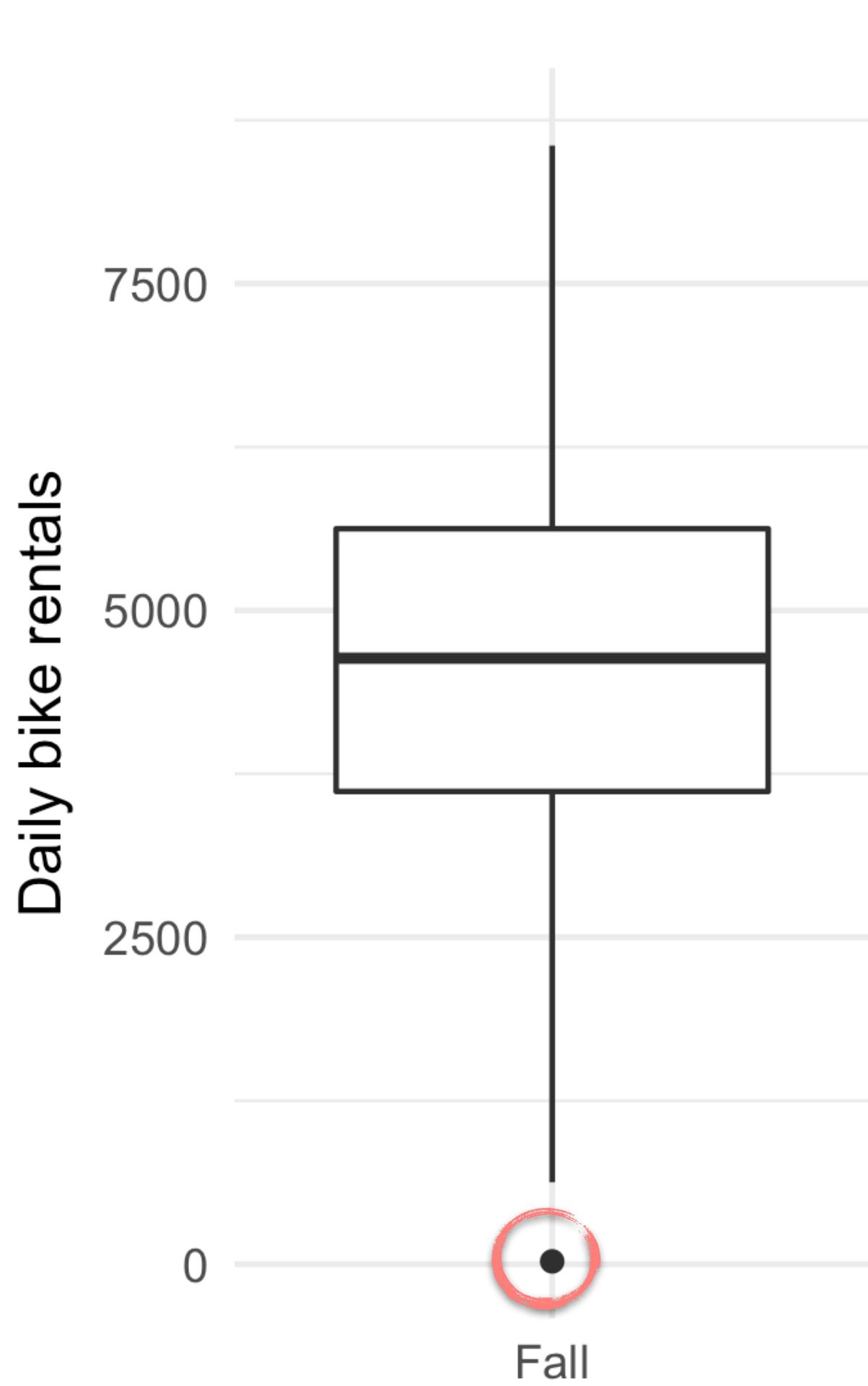


```
bike %>%
  filter(season == "Winter") %>%
  summarise(min = max(cnt), day_min = dteday[which.max(cnt)])
## # A tibble: 1 x 2
##       min day_min
##   <dbl> <date>
## 1 7836 2012-03-17
```



[President Obama at the Dubliner on St. Patrick's Day | whitehouse.gov](https://obamawhitehouse.archives.gov/.../2012/.../17/president-obama-dubliner-st-patr...)  
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
```

Google october 29 2012 DC

[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,

**main**  
prediction and  
model selection

**get for free**  
use of  
outside data

paris paintings





*Two paintings very rich in composition, of a beautiful execution, and whose merit is very remarkable, each 17 inches 3 lines high, 23 inches wide; the first, painted on wood, comes from the Cabinet of Madame la Comtesse de Verrue; it represents a departure for the hunt: it shows in the front a child on a white horse, a man who gives the horn to gather the dogs, a falconer and other figures nicely distributed across the width of the painting; two horses drinking from a fountain; on the right in the corner a lovely country house topped by a terrace, on which people are at the table, others who play instruments; trees and fabriques pleasantly enrich the background.*

## data transcription

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	name	sale	lot	dealer	year	origin_author	origin_cat	school_pntg	diff_origin	price	count	subject	authorstandard	artistliving	authorstyle	author	winnir
2517	R1777-86	R1777	86	R	1777	D/FL	D/FL	D/FL	0	620.0	1	2 femmes, enfants, paysage vu à travers une arcade	Bega, Cornelis Pieterszoon	0	n/a	Corneille Bega	Lebrun
2518	R1777-87	R1777	87	R	1777	D/FL	D/FL	D/FL	0	12,000.0	1	Course du hareng	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Donjeu
2519	R1777-88	R1777	88	R	1777	D/FL	D/FL	D/FL	0	8,000.0	1	Paysage sablonneux	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Lambe
2520	R1777-89a	R1777	89	R	1777	D/FL	D/FL	D/FL	0	5,300.0	1	Départ pour la chasse	Wouwerman, Philips	0	n/a	Philippe Wouwerman	Langlie

	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	winningbidder	winningbiddertype	endbuyer	Interm	type_intermed	Height_in	Width_in	Surface_Rect	Diam_in	Surface_Rnd	Shape	Surface	material	mat	quantity	nfigures	engraved
2516	Feuillet	D	D	0		16	20	320			squ_rect	320	toile	t	1	0	0
2517	Lebrun, Jean-Baptiste-Pierre	D	D	0		13.25	11	145.75			squ_rect	145.75	bois	b	1	0	0
2518	Donjeux, Vincent	D	D	0		23	29.25	672.75			squ_rect	672.75	toile	t	1	50	0
2519	Lambert, John (Chevalier Lambert)	C	C	0		23	30	690			squ_rect	690	toile	t	1	0	1
2520	Langlier, Jacques for Poullain, Antoine	DC	C	1	D	17.25	23	396.75			squ_rect	396.75	bois	b	1	0	0

- **mat** - category of material (a=silver, al=alabaster, ar=slate, b=wood, bc=wood and copper, br=bronze frames, bt=canvas on wood, c=copper, ca=cardboard, co=cloth, e=wax, g=grisaille technique, h=oil technique, m=marble, mi=miniature technique, o=other, p=paper, pa=pastel, t=canvas, ta=canvas?, v=glass, n/a=NA, (blanks)=NA)
- **Shape** - shape of painting

```
pp <- pp %>%
  mutate(
    Shape = fct_collapse(Shape, oval = c("oval", "ovale"),
                          round = c("round", "ronde"),
                          squ_rect = "squ_rect",
                          other = c("octogon", "octagon", "miniature")),
    mat = fct_collapse(mat, metal = c("a", "br", "c"),
                        canvas = c("co", "t", "ta"),
                        paper = c("p", "ca"),
                        wood = "b",
                        other = c("e", "g", "h", "mi", "o", "pa", "v", "al", "ar", "m"))
  )
```

## main

data provenance

modelling

diagnostic, log  
transformation

## get for free

iterative

data

cleanup, i.e.

working with other  
people's data

manhattan apartments



# observed sample



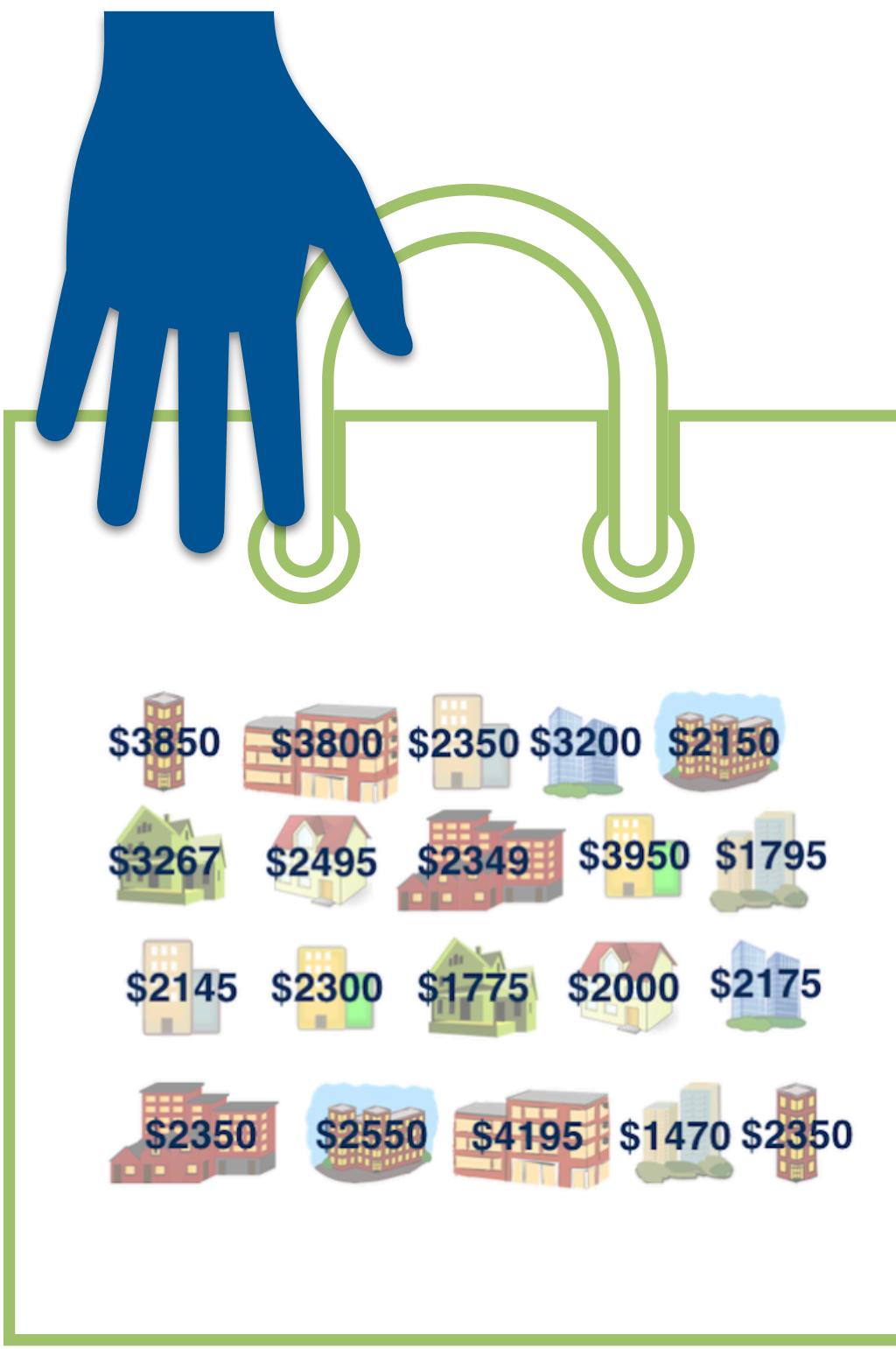
Sample median = \$2350 😱

# population



# Population median = ?

# first, tactile simulation



**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:**

```
library(infer)

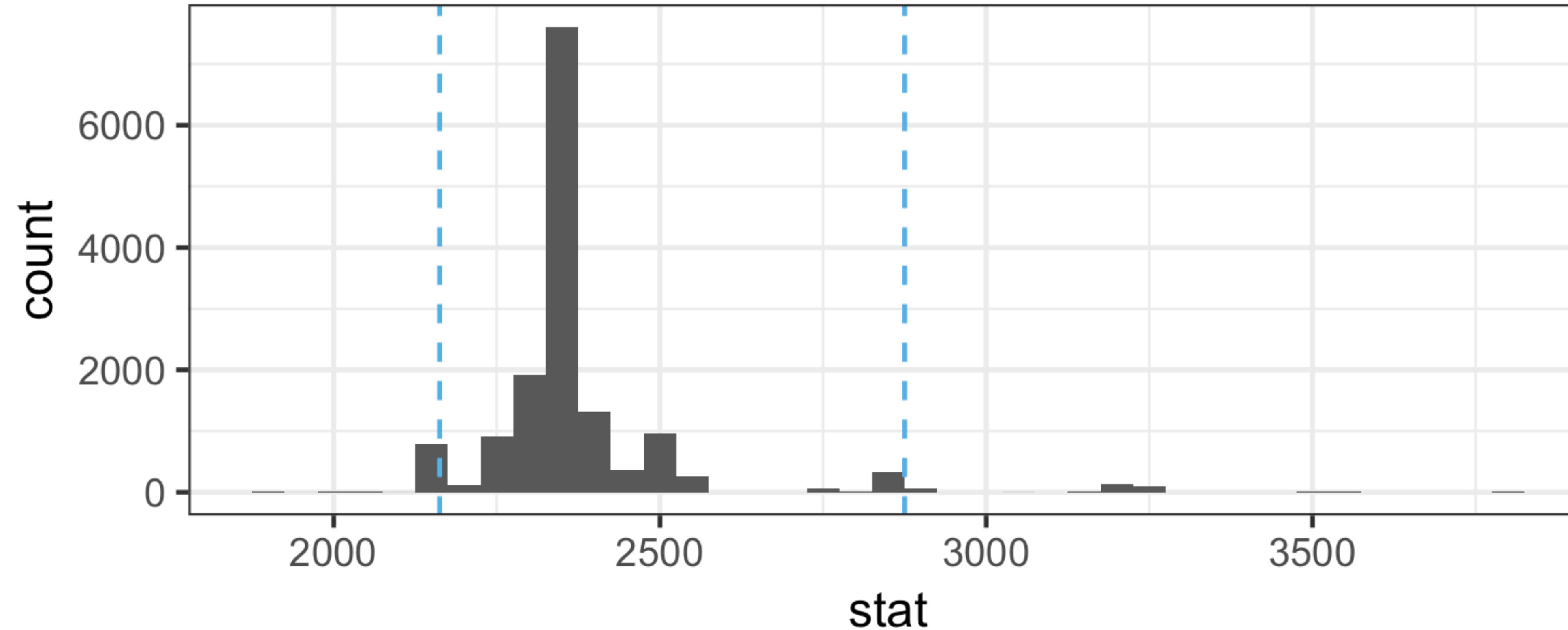
manhattan %>%

  # 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



**main**  
estimation  
via  
bootstrapping

**get for free**  
discussion on  
representativeness  
of samples



ST

HA

AN

F3

# North Carolina Breweries

241 active

25 closed

Name	Type	Beer Count	My Count	Est.
1718 Brewing Ocracoke Ocracoke	Brewpub	<input type="checkbox"/> 3	-	2018
217 Brew Works Wilson	Microbrewery	<input type="checkbox"/> 10	-	2017
3rd Rock Brewing Company Trenton	Microbrewery	<input type="checkbox"/> 13	-	2016
7 Clans Brewing Cherokee	Client Brewer	<input type="checkbox"/> 1	-	2018
Andrews Brewing Company Andrews	Microbrewery	<input type="checkbox"/> 18	-	2014
Angry Troll Brewing Elkin	Microbrewery	<input type="checkbox"/> 8	-	2017
Appalachian Mountain Brewery Boone	Microbrewery	<input type="checkbox"/> 79	-	2013
Archetype Brewing Asheville	Microbrewery	<input type="checkbox"/> 24	-	2017

copy-paste

	A	B	C	D	E
1	Name	Type	Beer Count	My Count	Est.
2	<a href="#">1718 Brewing Ocracoke</a>				
3	<a href="#">Ocracoke</a>	Brewpub	3	-	2018
4	<a href="#">217 Brew Works</a>				
5	<a href="#">Wilson</a>			-	2017
6	<a href="#">3rd Rock Brewing Company</a>				
7	<a href="#">Trenton</a>			-	2016
8	<a href="#">7 Clans Brewing</a>				
9	<a href="#">Cherokee</a>			-	2018
10	<a href="#">Andrews Brewing Company</a>				
11	<a href="#">Andrews</a>			-	2014
12	<a href="#">Angry Troll Brewing</a>				
13	<a href="#">Elkin</a>			-	2017
14	<a href="#">Appalachian Mountain Brewery</a>				
15	<a href="#">Boone</a>	Microbrewery	79	-	2013
16	<a href="#">Archetype Brewing</a>				
17	<a href="#">Asheville</a>	Microbrewery	24	-	2017



```
library(tidyverse)
library(rvest)

page ← read_html("https://www.ratebeer.com/breweries/north%20carolina/33/213/")

names ← page %>%
  html_nodes("#brewerTable a:nth-child(1)") %>%
  html_text() %>%
  str_trim()

active_cities ← page %>%
  html_nodes(".filter") %>%
  html_text()

closed_cities ← page %>%
  html_nodes("#brewerTable span") %>%
  html_text()

cities ← c(active_cities, closed_cities)

...

ncbreweries ← tibble(
  name = names,
  city = cities,
  ...
)

write_csv(ncbreweries, path = "data/ncbreweries.csv")
```

	name	city	type	beercount	est	status	url
1	1718 Brewing Ocracoke	Ocracoke	Brewpub	3	2018	Active	<a href="https://www.ratebeer.com//brewers/1718-brewing-ocracoke">https://www.ratebeer.com//brewers/1718-brewing-ocracoke</a>
2	217 Brew Works	Wilson	Microbrewery	10	2017	Active	<a href="https://www.ratebeer.com//brewers/217-brew-works">https://www.ratebeer.com//brewers/217-brew-works</a>
3	3rd Rock Brewing Company	Trenton	Microbrewery	13	2016	Active	<a href="https://www.ratebeer.com//brewers/3rd-rock-brewing-company">https://www.ratebeer.com//brewers/3rd-rock-brewing-company</a>
4	7 Clans Brewing	Cherokee	Client Brewer	1	2018	Active	<a href="https://www.ratebeer.com//brewers/7-clans-brewing">https://www.ratebeer.com//brewers/7-clans-brewing</a>
5	Andrews Brewing Company	Andrews	Microbrewery	18	2014	Active	<a href="https://www.ratebeer.com//brewers/andrews-brewing-company">https://www.ratebeer.com//brewers/andrews-brewing-company</a>
6	Angry Troll Brewing	Elkin	Microbrewery	8	2017	Active	<a href="https://www.ratebeer.com//brewers/angry-troll-brewing">https://www.ratebeer.com//brewers/angry-troll-brewing</a>
7	Appalachian Mountain Brewery	Boone	Microbrewery	79	2013	Active	<a href="https://www.ratebeer.com//brewers/appalachian-mountain-brewery">https://www.ratebeer.com//brewers/appalachian-mountain-brewery</a>
8	Archetype Brewing	Asheville	Microbrewery	24	2017	Active	<a href="https://www.ratebeer.com//brewers/archetype-brewing">https://www.ratebeer.com//brewers/archetype-brewing</a>
9	Asheville Brewing Company	Asheville	Brewpub	88	2003	Active	<a href="https://www.ratebeer.com//brewers/asheville-brewing-company">https://www.ratebeer.com//brewers/asheville-brewing-company</a>
10	Ass Clown Brewing Company	Cornelius	Microbrewery	108	2011	Active	<a href="https://www.ratebeer.com//brewers/ass-clown-brewing-company">https://www.ratebeer.com//brewers/ass-clown-brewing-company</a>
11	Aviator Brewing Company	Fuquay Varina	Microbrewery	60	2008	Active	<a href="https://www.ratebeer.com//brewers/aviator-brewing-company">https://www.ratebeer.com//brewers/aviator-brewing-company</a>
12	Balsam Falls Brewing Company	Sylva	Microbrewery	25	2018	Active	<a href="https://www.ratebeer.com//brewers/balsam-falls-brewing-company">https://www.ratebeer.com//brewers/balsam-falls-brewing-company</a>
13	Barking Duck Brewing Company	Mint Hill	Microbrewery	16	2014	Active	<a href="https://www.ratebeer.com//brewers/barking-duck-brewing-company">https://www.ratebeer.com//brewers/barking-duck-brewing-company</a>
14	Barrel Culture Brewing and Blending	Durham	Microbrewery	40	2017	Active	<a href="https://www.ratebeer.com//brewers/barrel-culture-brewing-and-blending">https://www.ratebeer.com//brewers/barrel-culture-brewing-and-blending</a>

Secure | https://www.ratebeer.com/brewers/1718-brewing-ocracoke/35300/

# 1718 Brewing Ocracoke

Brew Pub

📍 1129 Irvin Garrish HWY, Ocracoke, North Carolina, USA 27960

📞 (252) 928-2337

🕒 Mon-Sat: 12 ot 9 pm

ocracokebrewing.com

Associated place: **1718 Brewing Ocracoke**

BREWERS: [Discover our brewer resources here.](#)  
*Are you affiliated with this brewery?*

[★ Follow](#)

[Q Beers](#) [Reviews](#) [📍 Distribution Map](#) [🛒 Distributors](#)

▲ Name	▲ ABV	▲ Added
<b>1718 Ocracoke Coffee Kölsch</b> Kölsch	5.0	4/20/2018
<b>1718 Ocracoke Good Bones IPA</b> India Pale Ale (IPA)	-	4/20/2018
<b>1718 Ocracoke Pepperberry Saison</b> Saison	7.1	4/20/2018

**main**  
data  
harvesting

**get for free**

parsing  
text strings

iteration



**teach  
tools for  
good science**

# literate programming

## **reproducibility:**

train new analysts  
whose only  
workflow is a  
reproducible one

## **pedagogy:**

code + output +  
prose together  
  
syntax highlighting  
+ notebooks FTW!

## **efficiency:**

consistent  
formatting + built in  
“show your work”  
= easier grading

## **key to success:**

iterative  
development:  
knit early,  
and often



# version control

## version control:

lots of mistakes  
along the way,  
need ability keep  
track of history  
(revert)

## collaboration:

platform and  
interface designed  
to enable  
collaboration

## accountability:

transparent  
commit history

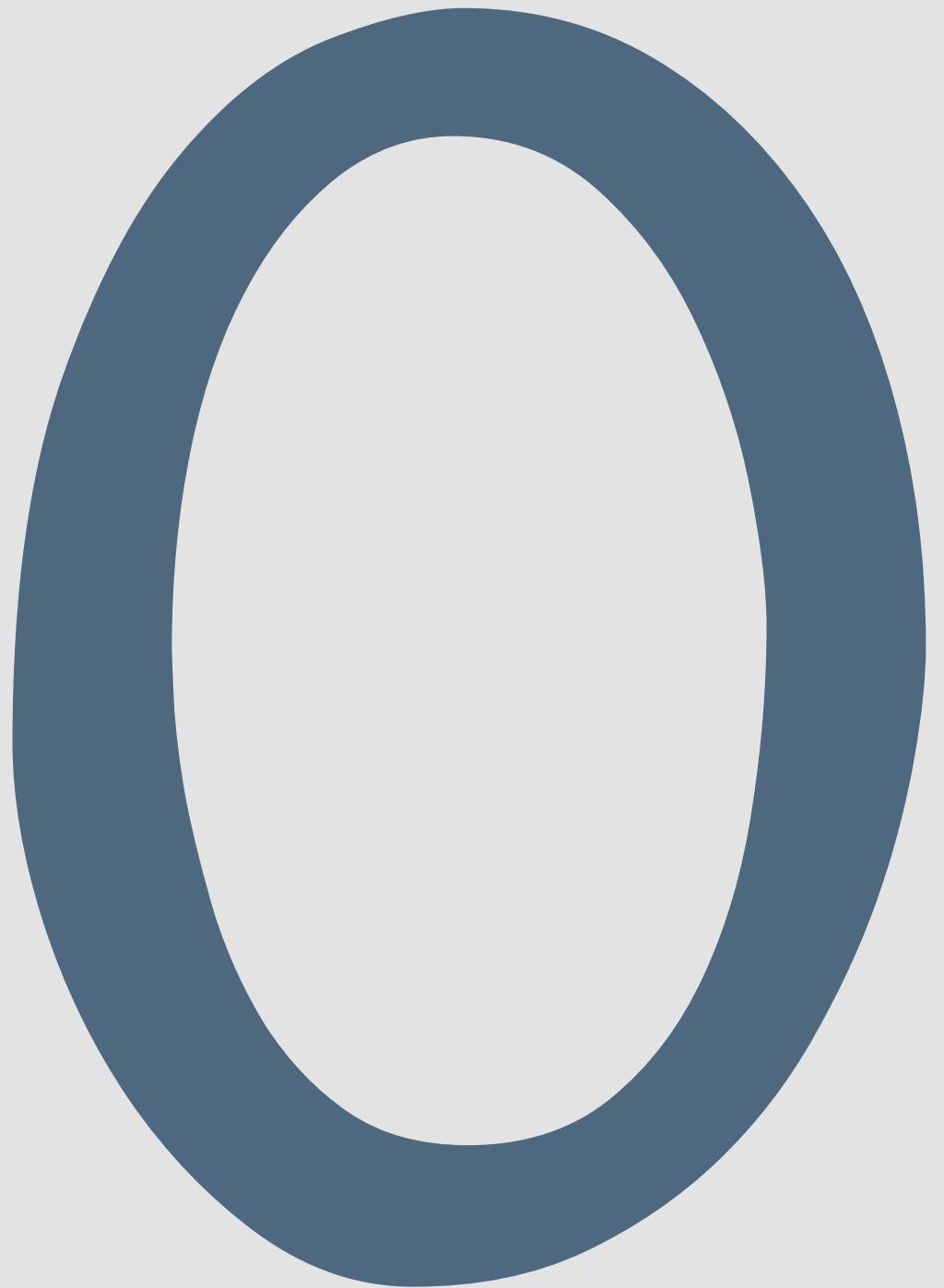
## early intro:

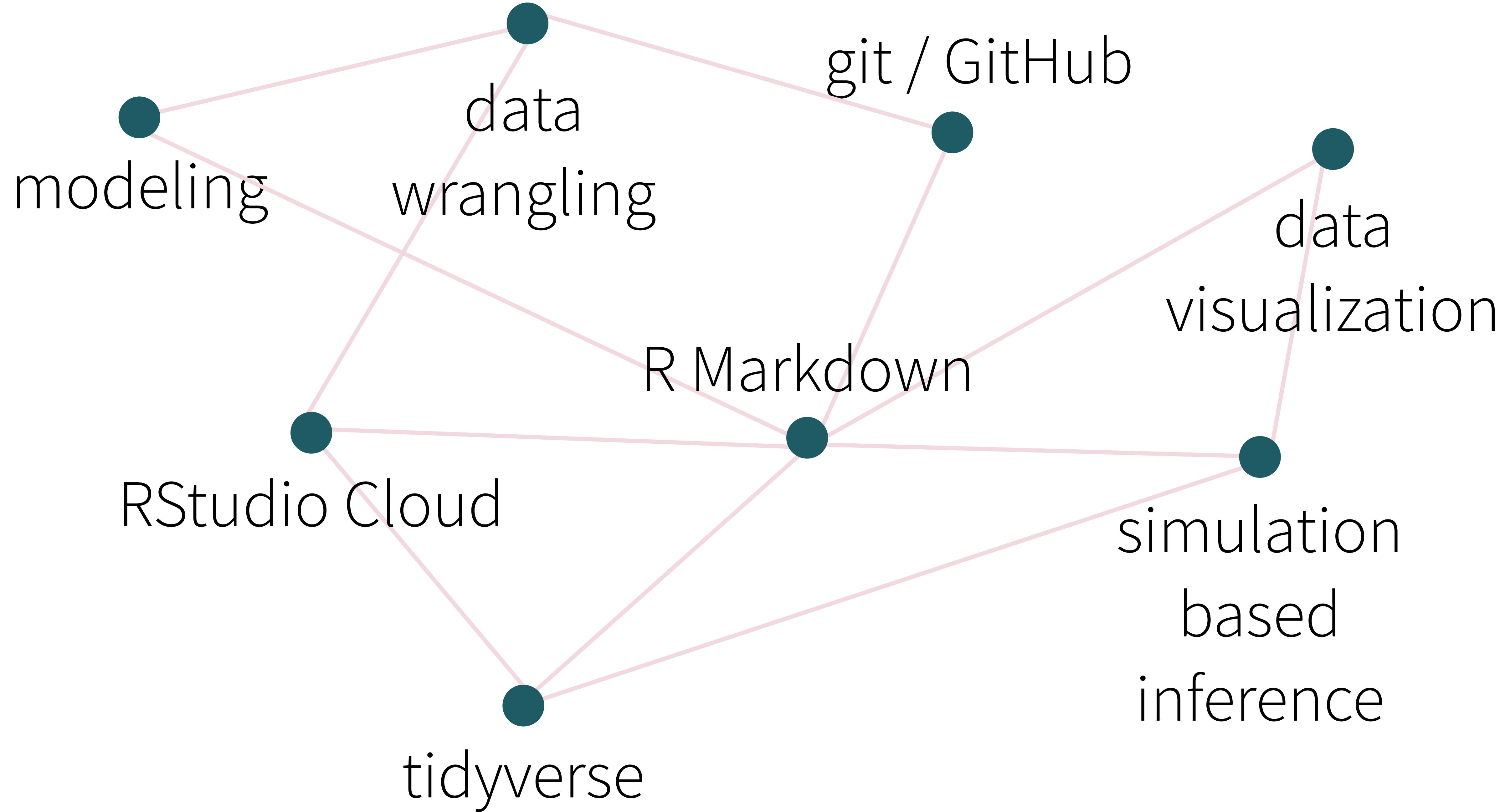
mastery takes time,  
start early (day 1)  
  
marketability +  
discoverability



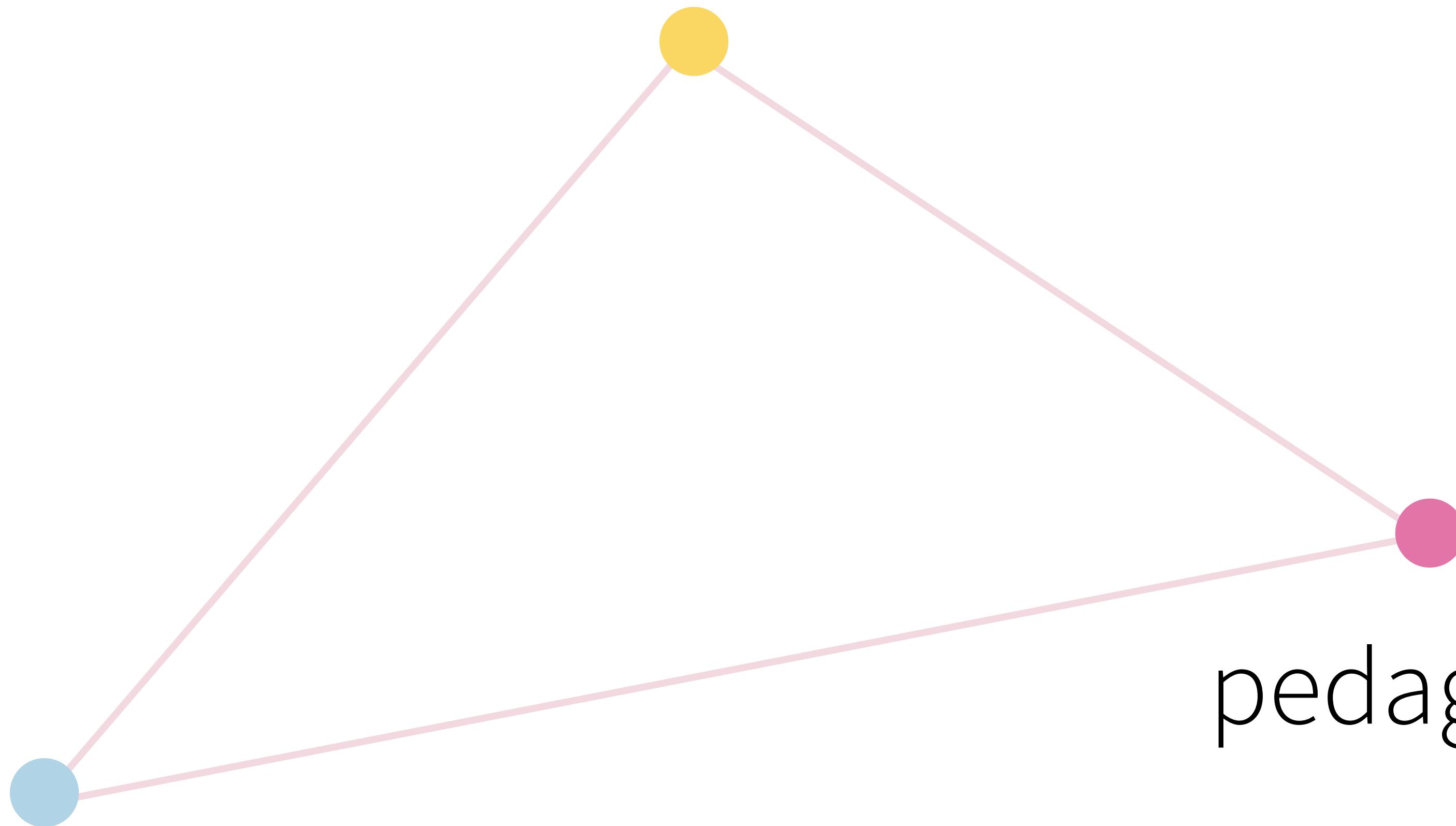
- 1 rethink, don't just add
- 2 cherish day one
- 3 stick with a consistent grammar
- 4 use real and relatable examples
- 5 teach tools for good science

**you're  
not  
alone**





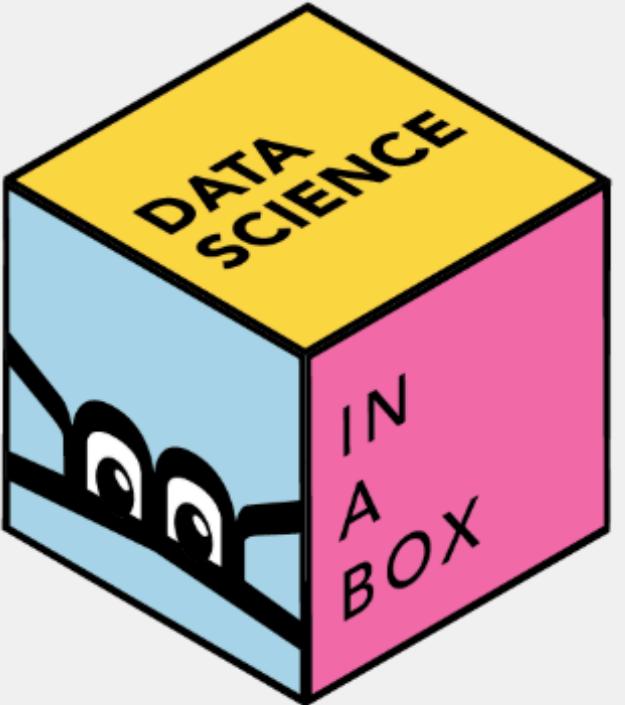
curriculum



tooling

pedagogy

[bit.ly/dsbox-web](https://bit.ly/dsbox-web)



Hello #dsbox

Course content

Technology stack

Pedagogy

Built with ❤️ and blogdown

[bit.ly/dsbox-repo](https://bit.ly/dsbox-repo)

# Data Science in a Box

How can we effectively and efficiently teach data science to students with little to no background in computing and statistical thinking? How can we equip them with the skills and tools for reasoning with various types of data and leave them wanting to learn more?

This introductory data science course that is our (working) answer to this question. The core content of the course focuses on data acquisition and wrangling, exploratory data analysis, data visualization, inference, modeling, and effective communication of results. Time permitting, the course also introduces additional concepts and tools like interactive visualization and reporting Bayesian inference. A heavy emphasis is placed on a consistent syntax (with tools from the [tidyverse](#)), reproducibility (with [R Markdown](#)) and version control and collaboration (with git/GitHub). In addition, out-of-class learning is supplemented with interactive [tutorials](#). The goal of the course is to bring students from zero to being able to work in a team to complete a fully reproducible data analysis project on a dataset of their choice and answering questions they care about.

Data Science in a Box contains the materials required to teach (or learn from) the course described above, all of which are [freely-available](#) and [open-source](#). They include course materials such as slide decks, homework assignments, guided labs, sample exams, a final project assignment, as well as pedagogical tips, computing infrastructure, technology stack, and course logistics.



**x 26**

**slides**

**x 10**

**labs**

**x 6**

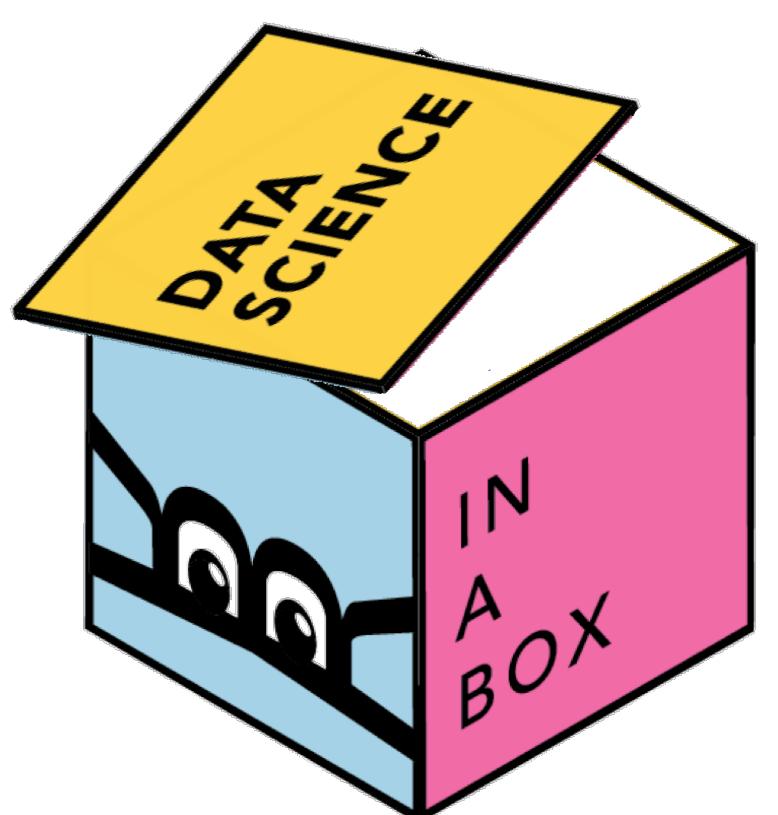
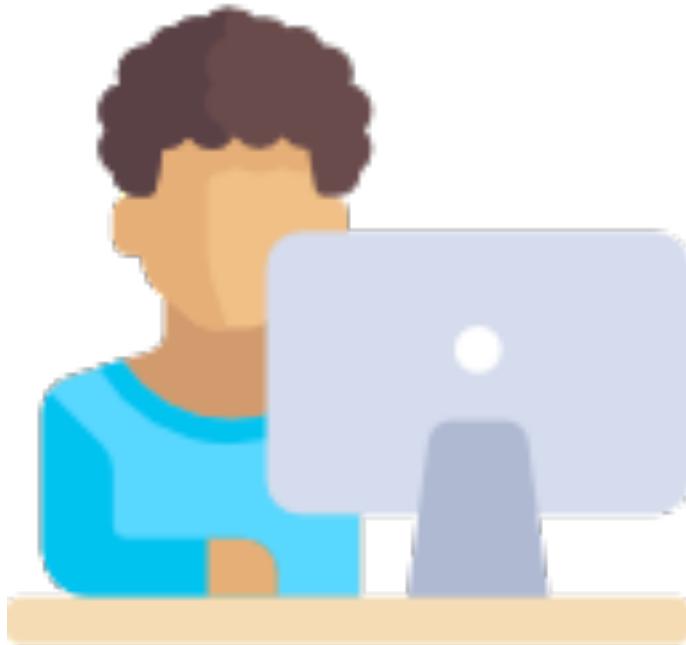
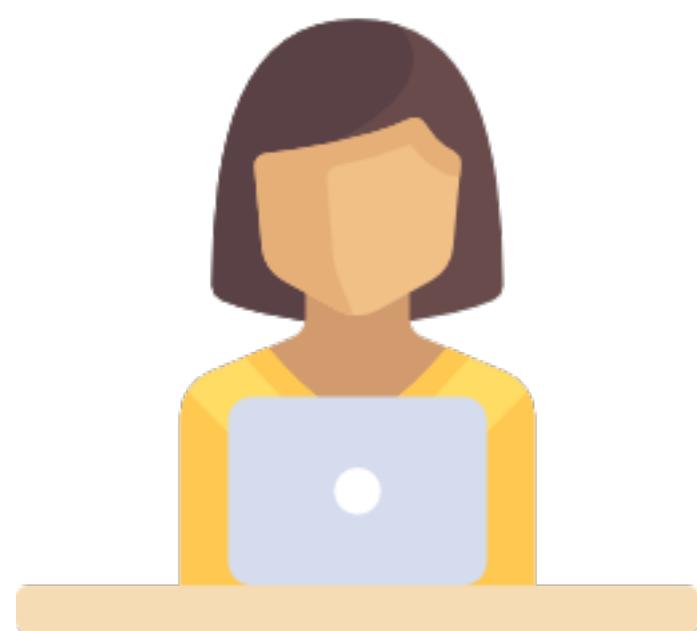
**assignments**

**x 1**

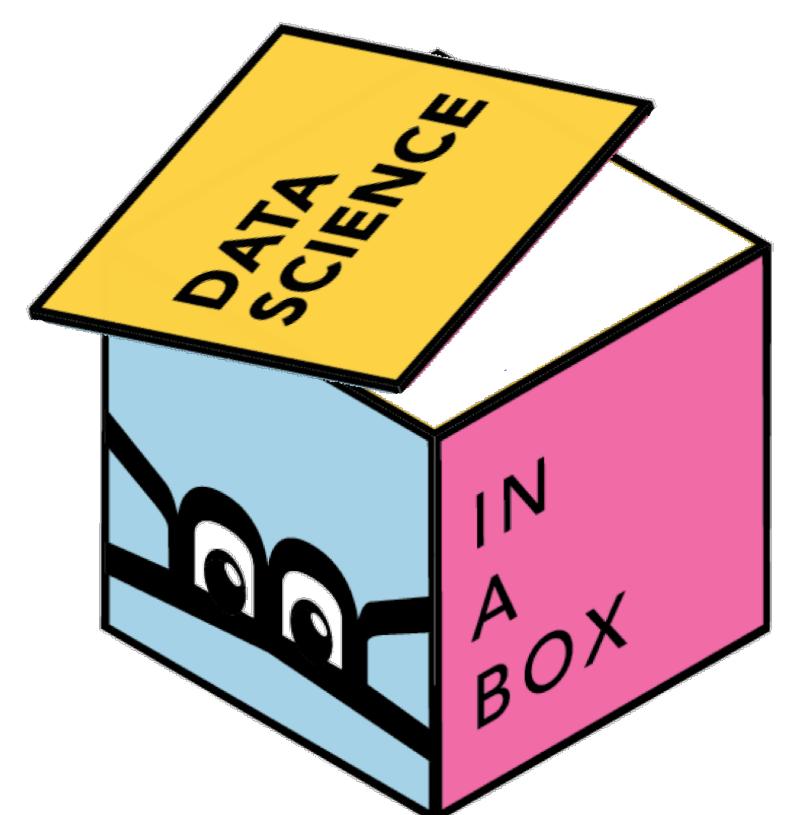
**project**

**x 2**

**exams**



**course infrastructure  
using the tech stack  
lesson plans  
pedagogical tips**

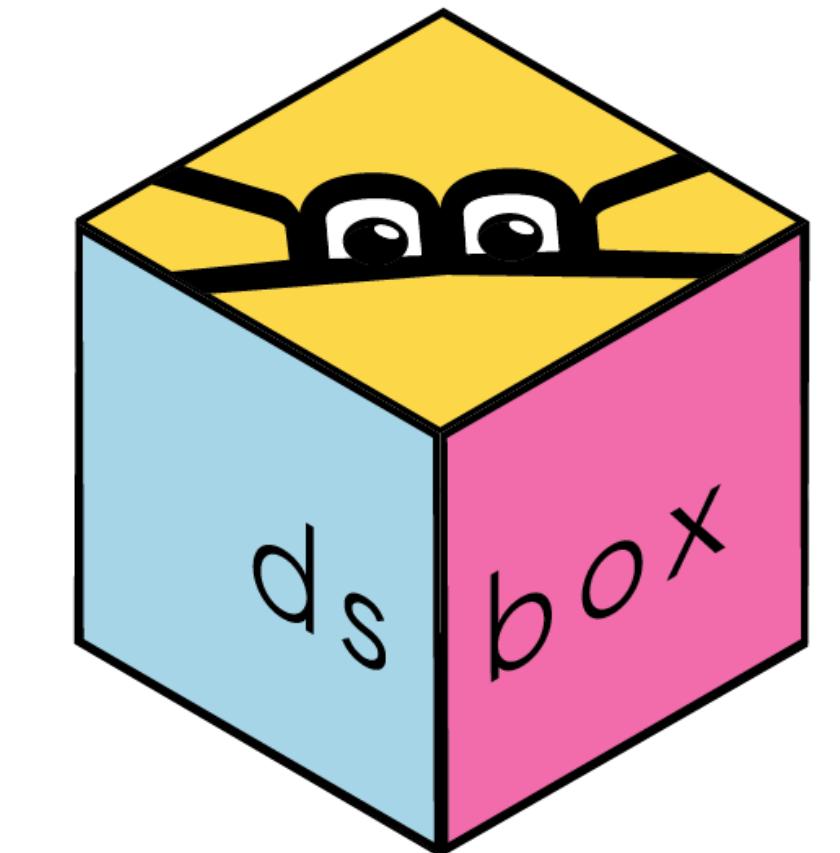




# dsbox

---

Datasets for the Data Science Course in a Box



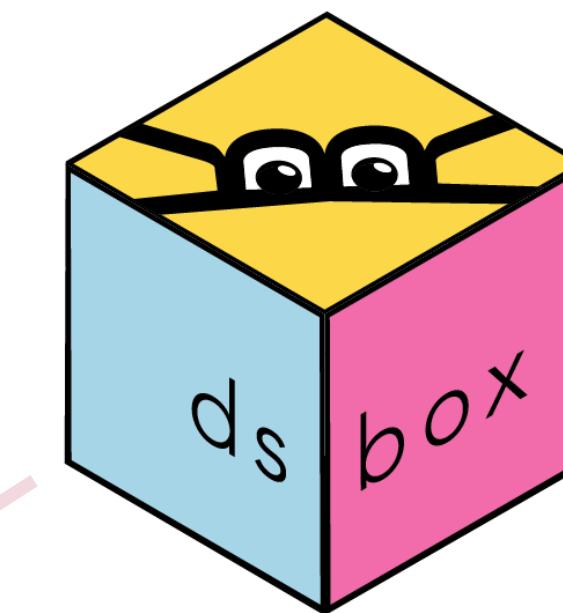
```
install_github("rstudio-education/dsbox")
```



R Studio Cloud



Teaching ▶



Thank you!



[bit.ly/dsbox-web](https://bit.ly/dsbox-web)



[bit.ly/ds-gateway](https://bit.ly/ds-gateway)

@minebocek

mine-cetinkaya-rundel

mine@stat.duke.edu

