

# General data science overview

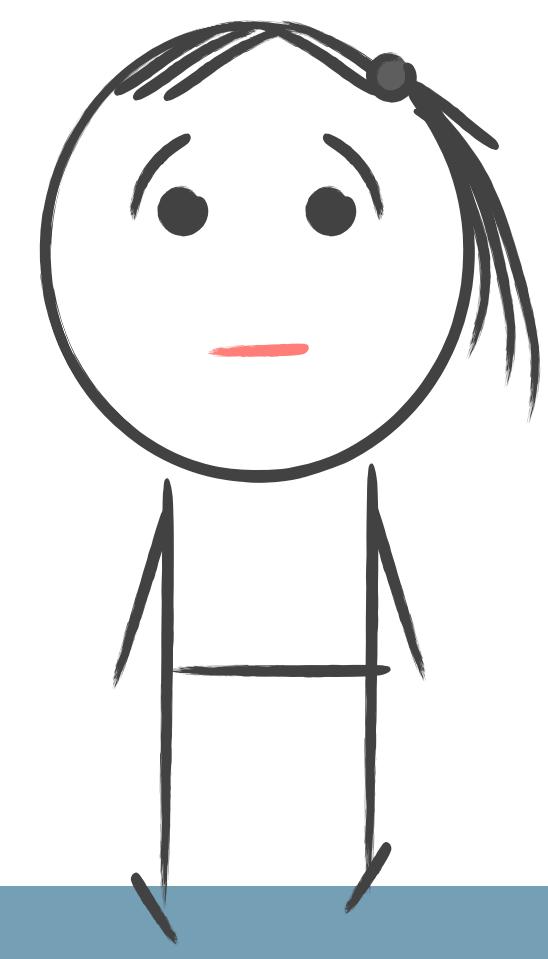
Mine Çetinkaya-Rundel

@minebocek

mine-cetinkaya-rundel

mine@rstudio.com

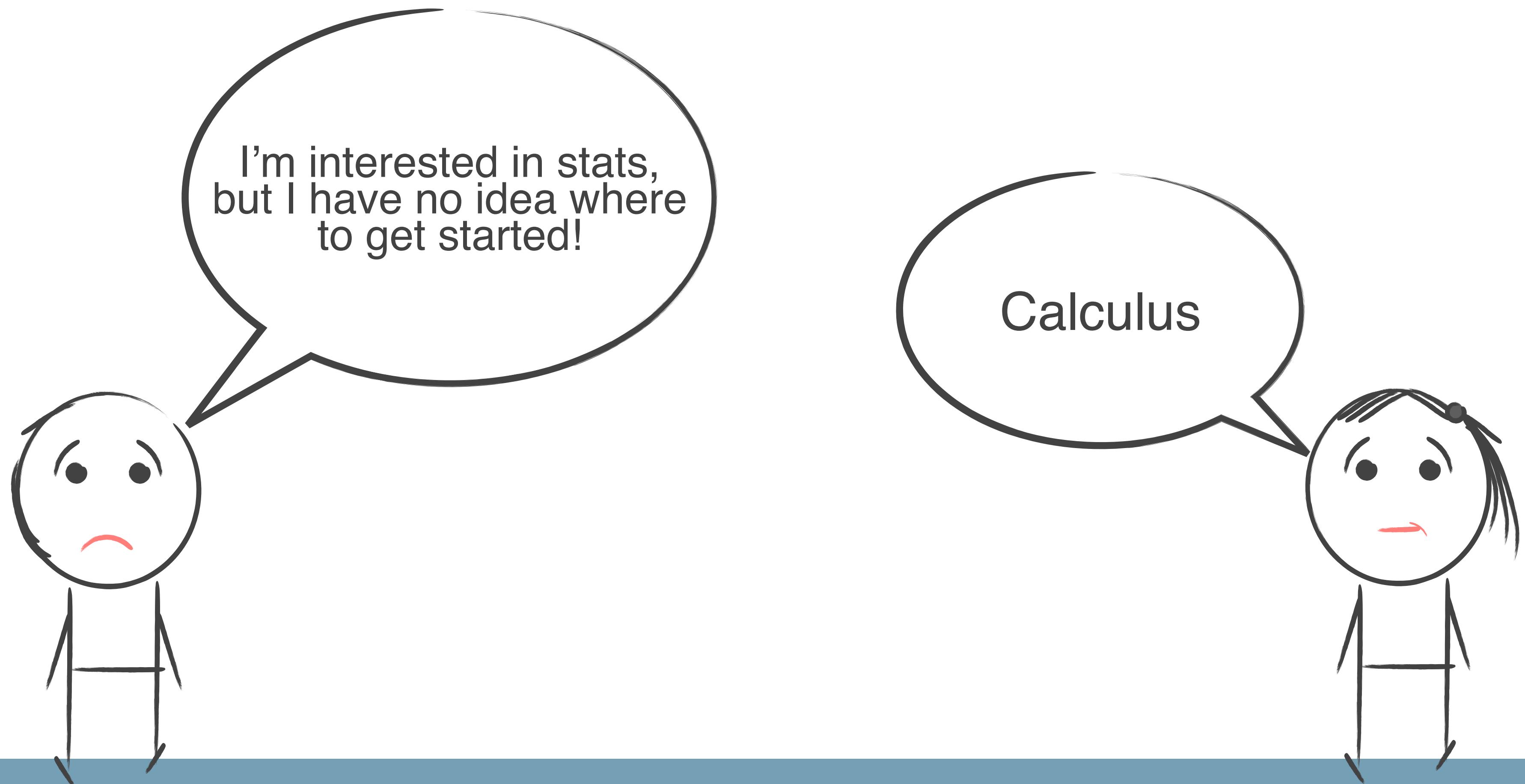


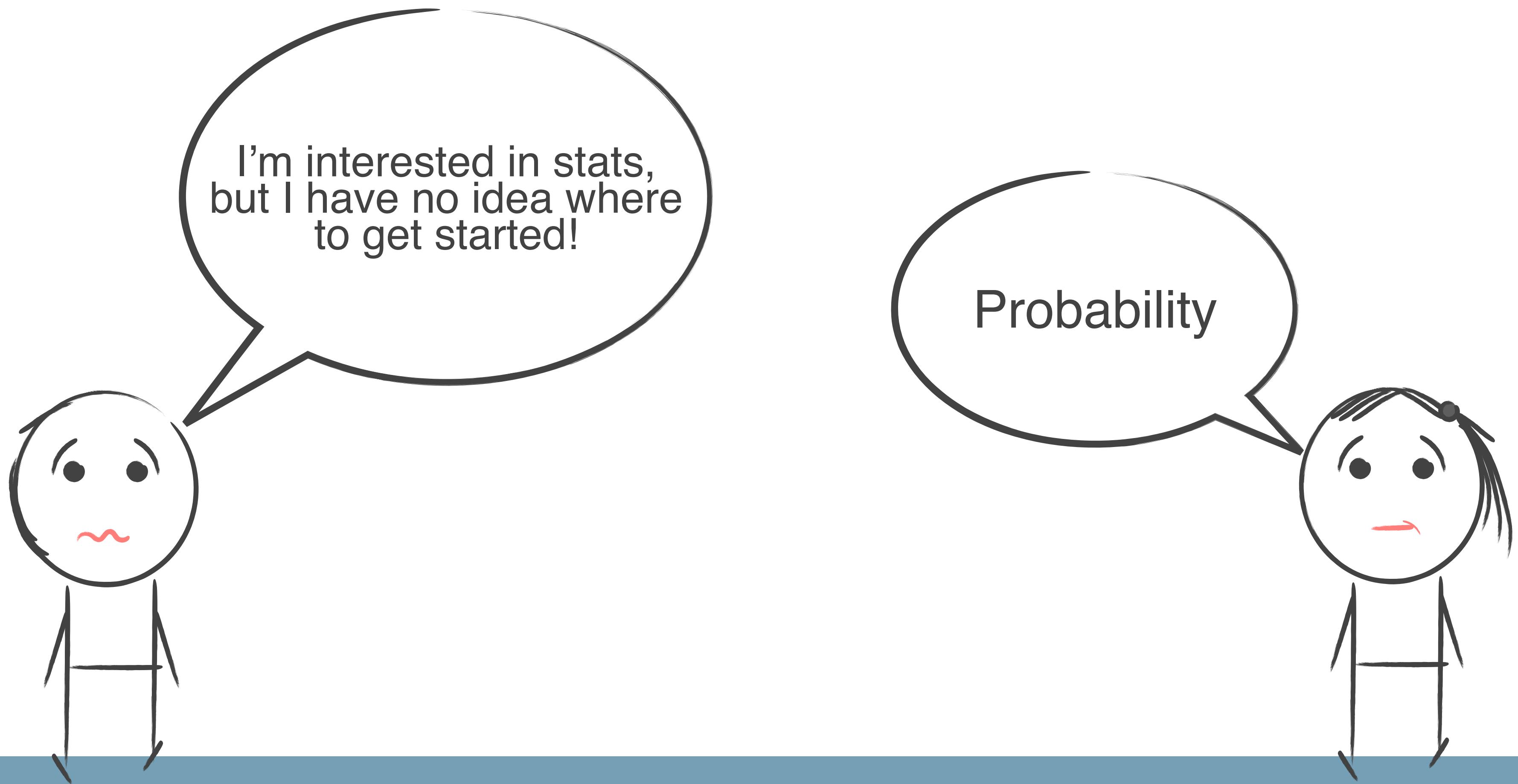




I'm interested in stats,  
but I have no idea where  
to get started!











motivation

computation

interest &  
impact

syllabus

data analysis  
examples

curricular  
considerations

**motivation**

computation

interest &  
impact

syllabus

data analysis  
examples

curricular  
considerations

# goal

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

- \* modern
- \* places data front and center
- \* quantitative (but not mathematical)
- \* 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  
computation

approach statistics  
from a model  
based perspective

underscore  
effective  
communication  
of findings

# and maybe more importantly...

ask questions that  
students want to  
answer

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

motivation

computation

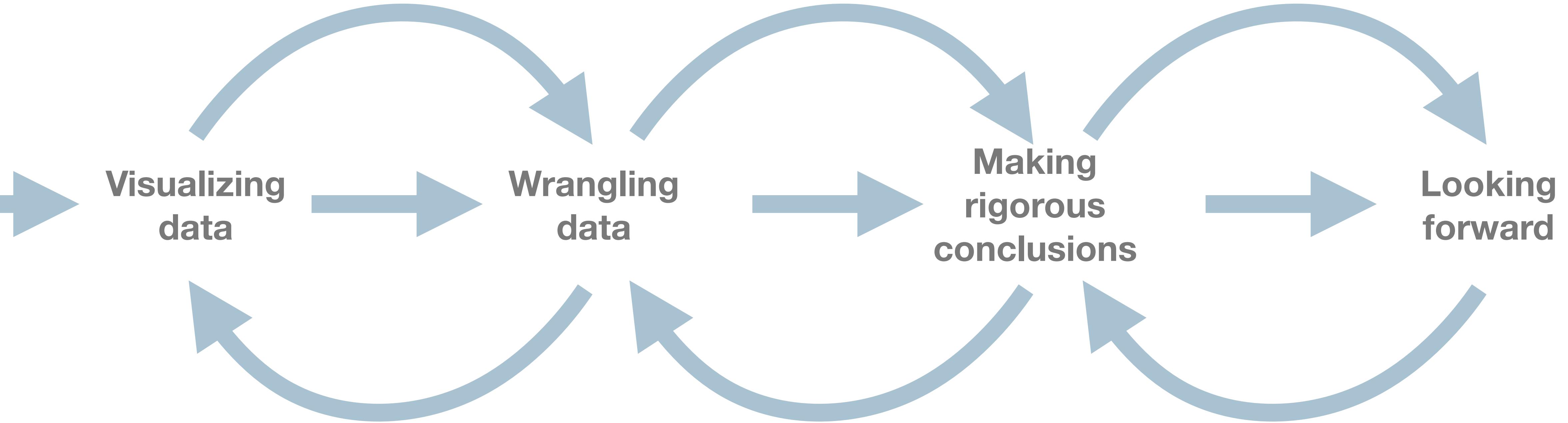
interest & impact

**syllabus**

data analysis  
examples

curricular  
considerations

# 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 (for fun, and for models)

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

Web scraping, Shiny, Bayesian inference, ???

# logistics

## **other goals:**

reproducible and  
collaborative workflow,  
effective  
communication

## **structure:**

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

## **assessment:**

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

# assignments

## **culminating:**

final open ended  
project on data of  
own choosing,  
team based

## **periodic:**

semi open ended  
homework,  
individual

## **in (every) class:**

semi open ended  
application  
exercises,  
team based

## **fast feedback:**

**learnr** modules,  
individual

Interactive Tutorials for R Mine

Secure <https://rstudio.github.io/learnr/>

learnr Home Exercises Questions Publishing Formats Examples

## Interactive Tutorials for R Overview

The **learnr** package makes it easy to turn any [R Markdown](#) document into an interactive tutorial. Tutorials consist of content along with interactive components for checking and reinforcing understanding. Tutorials can include any or all of the following:

1. Narrative, figures, illustrations, and equations.
2. Code exercises (R code chunks that users can edit and execute directly).
3. Quiz questions.
4. Videos (supported services include YouTube and Vimeo).
5. Interactive Shiny components.

Tutorials automatically preserve work done within them, so if a user works on a few exercises or questions and returns to the tutorial later they can pick up right where they left off.

### Examples

<https://rstudio.github.io/learnr/>

```
# on CRAN
> install.packages("learnr")
> library(learnr)
```

**earnr package:**

**Data Basics**

**Filtering Observations**

**Summarise Tables**

**Combining multiple operations**

**Summarizing Data**

Data Visualization Basics    Mine

Secure <https://tutorials.shinyapps.io/02-Vis-Basics/#section-geometric-objects>

# Data Visualization Basics

## Geometric objects

### Geoms

Welcome

A code template

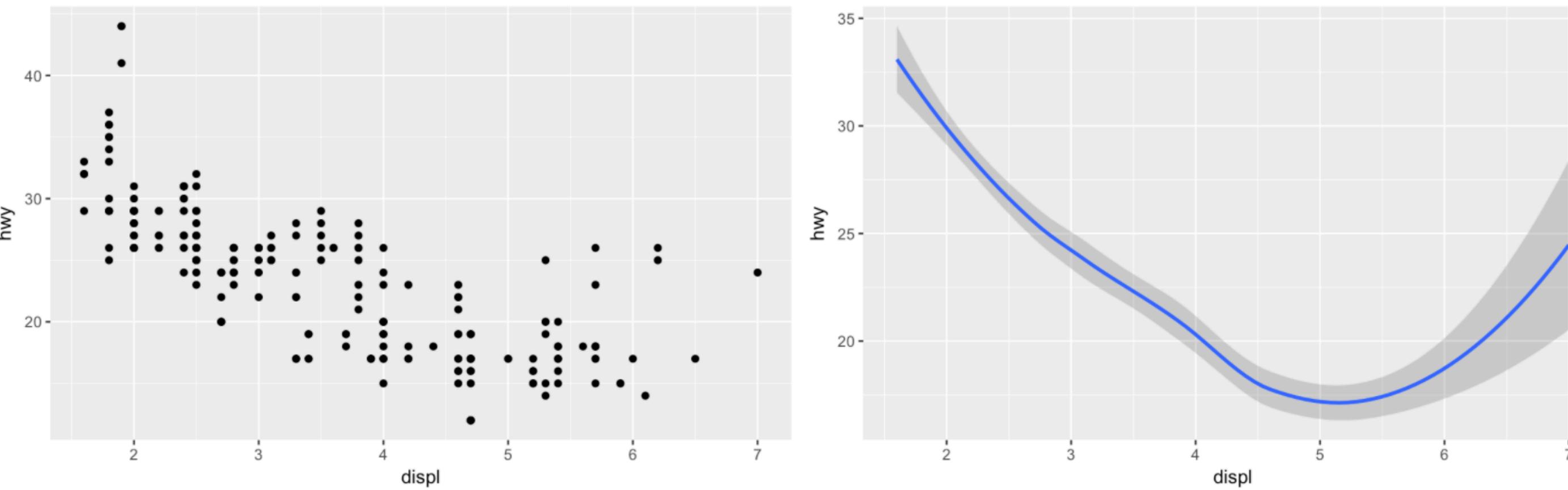
Aesthetic mappings

Geometric objects

The ggplot2 package

Start Over

How are these two plots similar?



Both plots contain the same x variable, the same y variable, and both describe the same data. But the plots are not identical. Each plot uses a different visual object to represent the data. In ggplot2 syntax, we say that they use different **geoms**.

A **geom** is the geometrical object that a plot uses to represent observations. People often describe plots by the type of geom that the plot uses. For example, bar charts use bar geoms, line charts use line geoms, boxplots use boxplot geoms, and so on. Scatterplots break the trend; they use the point geom.

As we see above, you can use different geoms to plot the same data. The plot on the left uses the point geom, and the plot on the right uses the smooth geom, a smooth line fitted to the data.

Continue

Data Visualization Basics

Secure https://tutorials.shinyapps.io/02-Vis-Basics/#section-aesthetic-mappings

✓ A strategy

# Data Visualization Basics

We can add the `class` variable to the plot by mapping the levels of an aesthetic (like color) to the values of `class`. For example, we can color a point green if it belongs to the compact class, blue if it belongs to the midsize class, and so on.

Let's give this a try. Fill in the blank piece of code below with `color = class`. What happens? Delete the commenting symbols (`#`) before running your code. (If you prefer British English, you can use `colour` instead of `color`.)

Welcome

A code template

Aesthetic mappings

Geometric objects

The ggplot2 package

Code Start Over Hint Run Code Submit Answer

```
1 ggplot(data = mpg) +  
2 | geom_point(mapping = aes(x = displ, y = hwy, color = class))  
3
```

Start Over

displ

hwy

class

- 2seater
- compact
- midsize
- minivan
- pickup
- subcompact
- suv

"Great Job! You can now tell which class of car each point represents by examining the color of the point."

This figure is a scatter plot titled "Aesthetic mappings". The x-axis is labeled "displ" (engine displacement) and ranges from approximately 1.6 to 7.0 liters. The y-axis is labeled "hwy" (highway fuel economy) and ranges from 10 to 45 mpg. The plot displays data points for various car classes, each represented by a different color: compact (yellow), midsize (green), minivan (teal), pickup (blue), subcompact (purple), and suv (pink). The points generally show a negative correlation, with fuel economy decreasing as engine size increases. There are several outliers, particularly in the compact and subcompact categories, which have relatively high fuel economy despite larger engines.

Data Visualization Basics    Mine

Secure <https://tutorials.shinyapps.io/02-Vis-Basics/#section-geometric-objects>

# Data Visualization Basics

## Exercise 2

What does the `se` argument to `geom_smooth()` do?

- Nothing. `se` is not an argument of `geom_smooth()`
- chooses a method for calculating the smooth line
- controls whether or not to show errors
- Adds or removes a standard error ribbon around the smooth line

[Submit Answer](#)

[Start Over](#)

[Continue](#)

motivation

computation

interest &  
impact

syllabus

data analysis  
examples

curricular  
considerations

# computation

## core:

R + RStudio server

## toolkit:

(mostly)

tidyverse

## reproducibility:

R Markdown +  
Git / GitHub

**core:**  
R + RStudio server

**why?**

# start up instructions

## # Local install

- Install R: <https://cran.r-project.org/>
- Install RStudio: <https://www.rstudio.com/products/rstudio/>
- Install the following packages:
  - rmarkdown
  - knitr
  - tidyverse
  - ...
- Load these packages

## vs. # RStudio Server

- Go to [smith.stat.duke.edu:8787](http://smith.stat.duke.edu:8787)
- Log in with your Net ID & pass

# core: R + RStudio Server

## **goal:**

minimize  
onboarding friction  
and time to 1st  
data viz

## **how:**

avoid local  
installation with  
RStudio Server  
(Pro)

## **at the end:**

provide  
instructions for +  
help with  
local install

**why?**

**toolkit:**  
(mostly)  
**tidyverse**

# recoding a binary variable

# base R

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

vs. # tidyverse

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

# recoding a multi-level variable

# base R

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

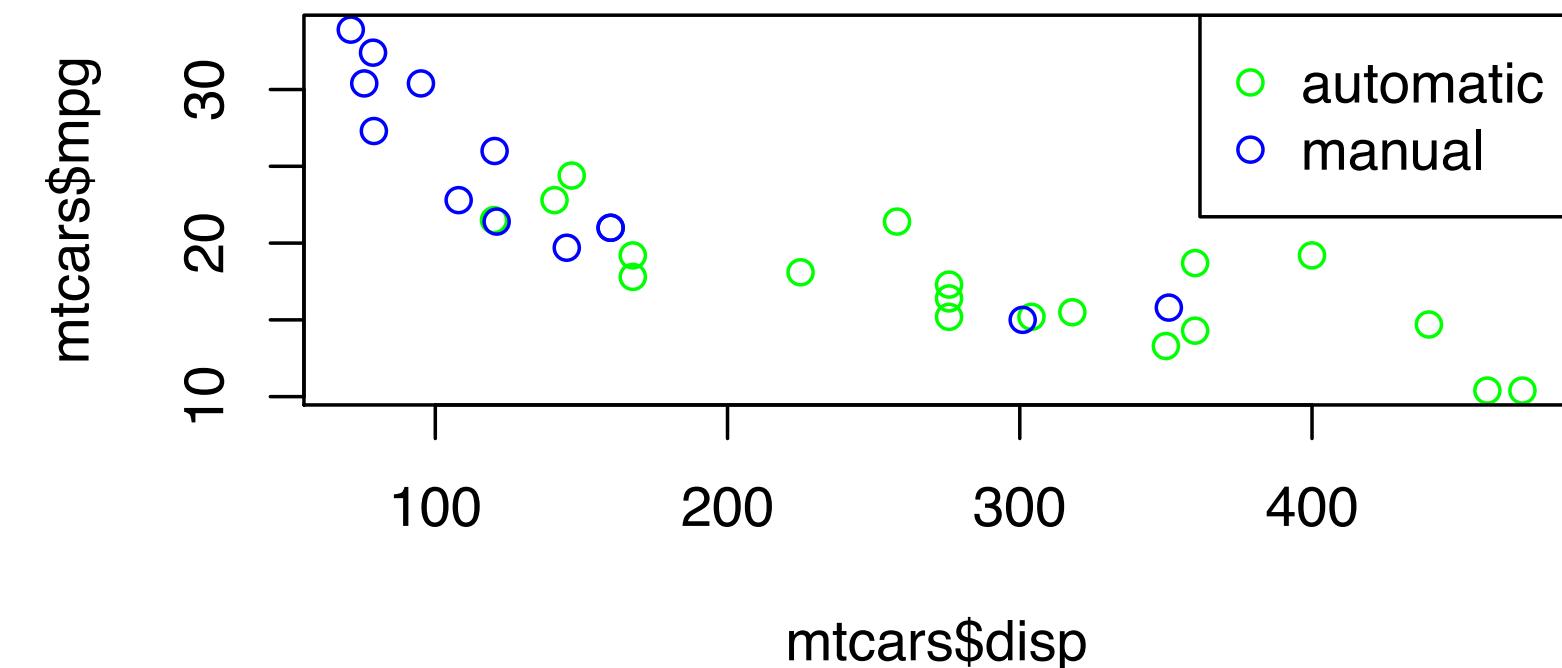
vs. # tidyverse

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

# visualizing multiple variables

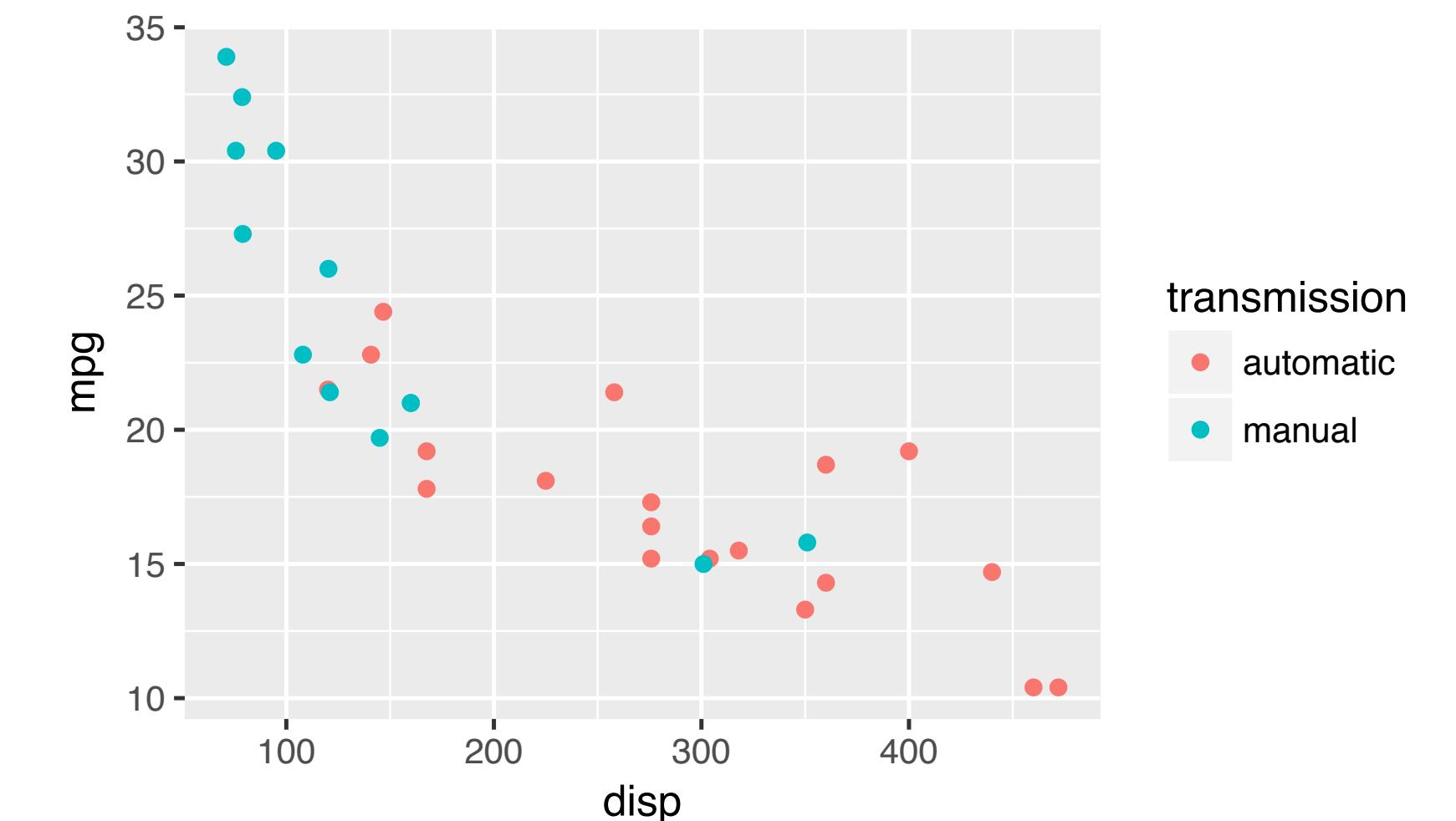
# base R

```
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"))
```



vs. # tidyverse

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

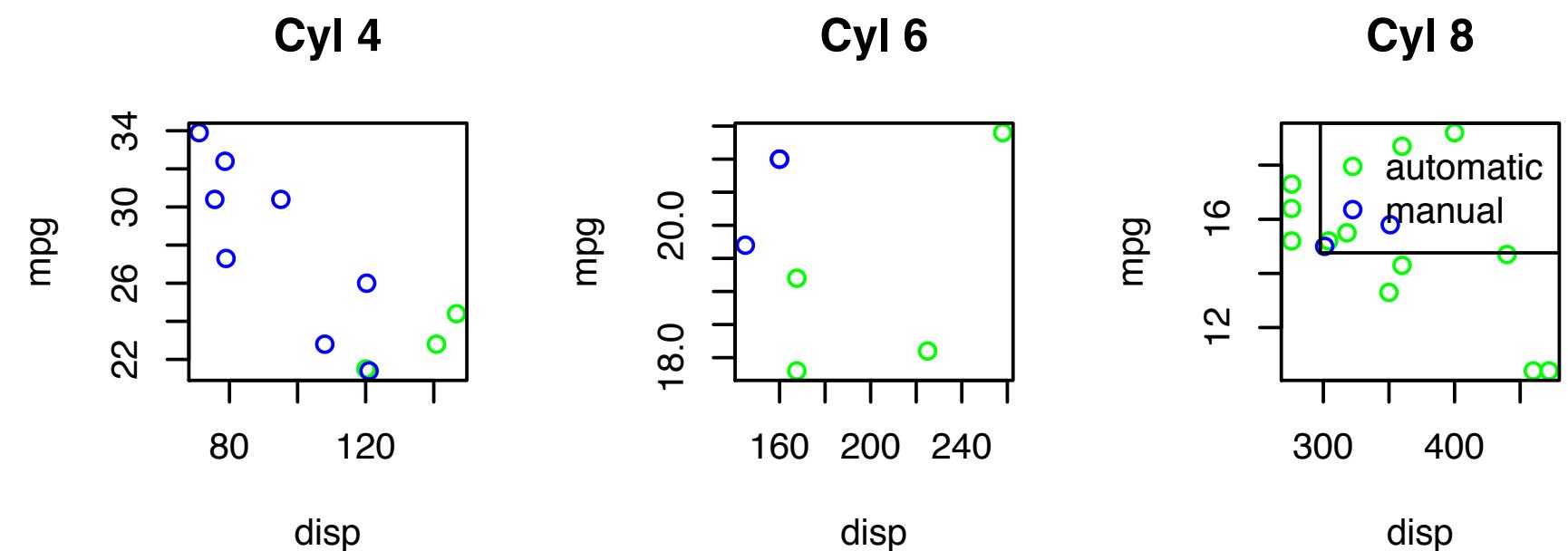


# visualizing even more variables

# base R

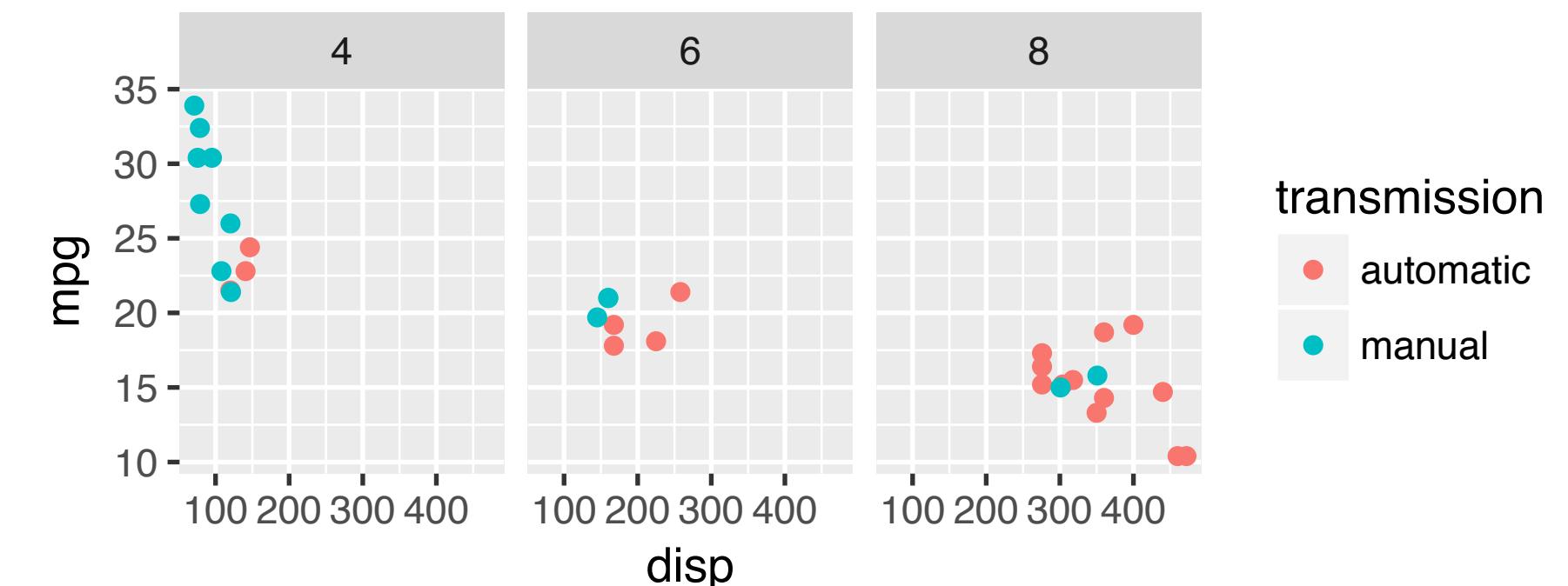
```
mtcars_cyl4 = mtcars[mtcars$cyl == 4, ]
mtcars_cyl6 = mtcars[mtcars$cyl == 6, ]
mtcars_cyl8 = mtcars[mtcars$cyl == 8, ]

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"))
```



vs. # tidyverse

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



# toolkit: (mostly) tidyverse

(closer to)  
human readable

consistent syntax

ease of  
multivariate  
visualizations

# why?

**reproducibility:**

R Markdown +  
Git / GitHub

# R Markdown

## **reproducibility:**

train new analysts  
whose only  
workflow is a  
reproducible one

## **efficiency:**

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

## **pedagogy:**

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

## **key to success:**

iterative  
development:  
knit early,  
and often

# Git + GitHub

## version control:

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

## accountability:

transparent  
commit history

## collaboration:

platform and  
interface designed  
to enable  
collaboration

## early intro:

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

motivation

computation

interest &  
impact

syllabus

**data analysis  
examples**

curricular  
considerations

# #1 paris paintings



# data expeditions



element of an  
undergrad course  
that introduces  
students to  
exploratory data  
analysis

pairs of grad  
students, work  
with course  
instructor to  
formulate question  
& pathway

graduate student  
participants  
receive  
a travel grant

# meet the experts



Hilary Coe Cronheim  
PhD, Art History

Sandra Van Ginhoven  
PhD, Art History



# data source: auction catalogs



*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

	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

# paris paintings

## **data:**

painting  
auction data  
1764 - 1780  
[3,393 x 57]

## **visualize:**

data visualization to  
explore patterns and  
possible interactions

## **clean:**

data cleaning and  
wrangling

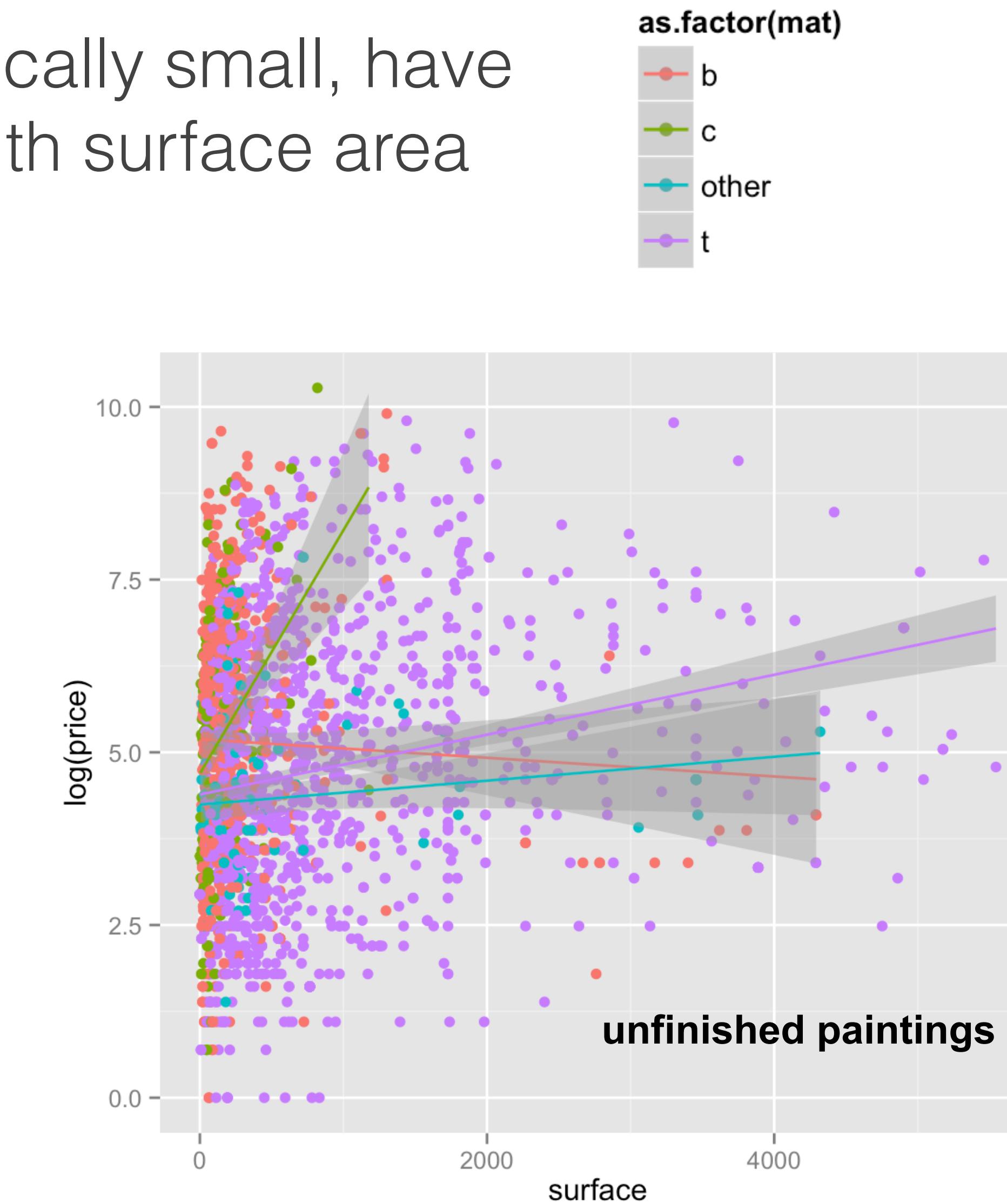
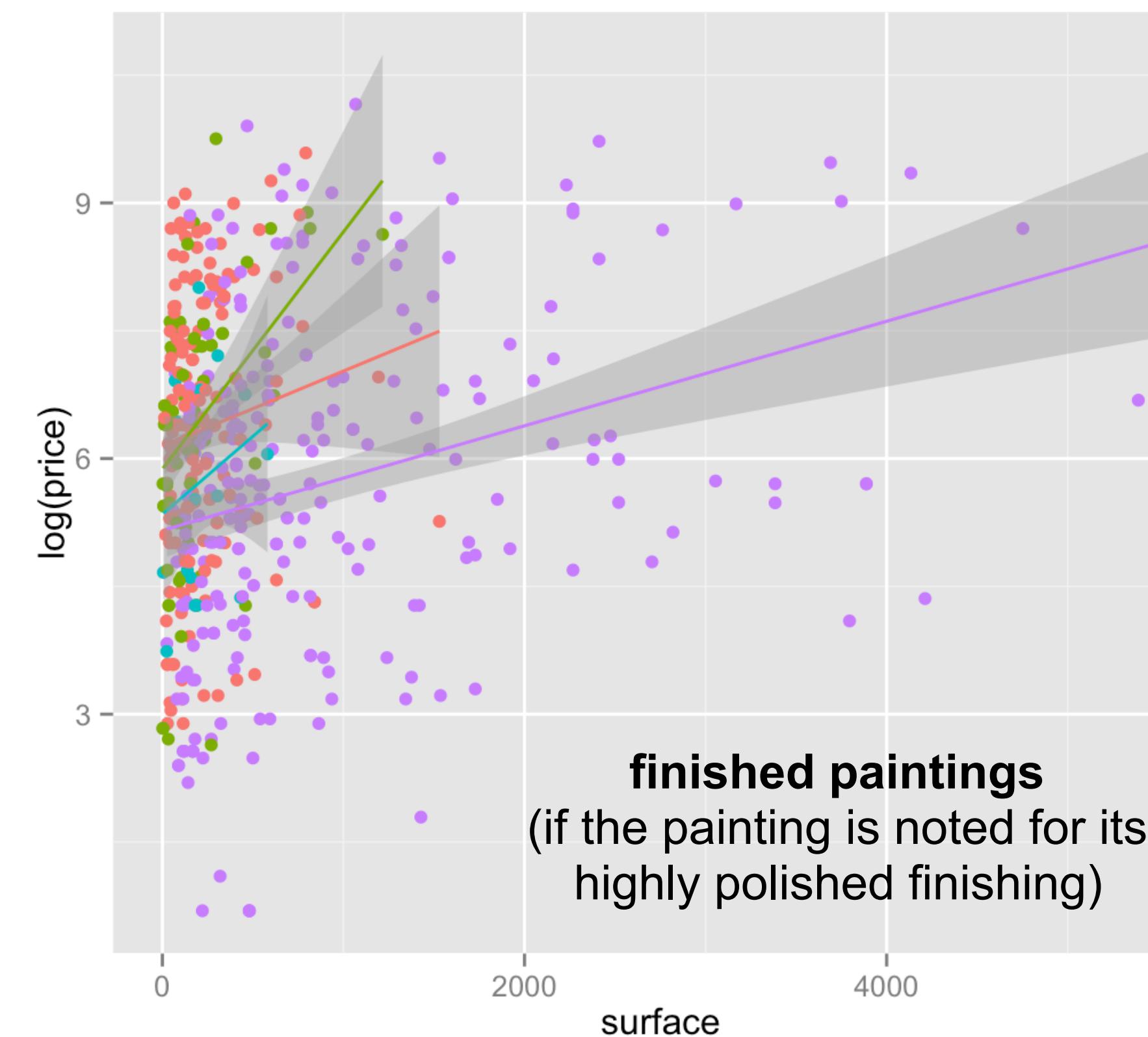
## **model:**

model  $\log(\text{price})$  and  
perform procedural  
and expert opinion  
based model  
selection

# sample exploration

# material and price

Copper paintings, though typically small, have a notably strong interaction with surface area



as.factor(mat)

- b
- c
- other
- t

# student experience

non-standard  
application piqued  
student interest

“massive” data  
overwhelming but  
expert input  
refreshing

unfamiliar  
variables made  
narrative  
challenging

novel  
application  
pushed  
creativity

#2 basketball



← → ⌂ i goduke.statsgeek.com/basketball-m/seasons/schedule.php?season=2014-15 ☆ RSS Print

2014-15 Schedule & Results							
Date	[Rk] Opponent	Duke Rank	Location (Venue)	Score (OT)	Att.	Tip Time	TV
11/14	~ <b>Presbyterian</b>	4	Durham, N.C. (Cameron Indoor Stadium)	W 113-44	9,314	6 p.m.	ESPNU
11/15	~ <b>Fairfield</b>	4	Durham, N.C. (Cameron Indoor Stadium)	W 109-59	9,314	8 p.m.	ESPN3
11/18	!! vs. [19] Michigan State	4	Indianapolis, Ind. (Bankers Life Fieldhouse)	W 81-71	19,306	7 p.m.	ESPN
11/21	~ vs. Temple	4	Brooklyn, N.Y. (Barclays Center)	W 74-54	10,135	9:30 p.m.	TruTV
11/22	~ vs. Stanford	4	Brooklyn, N.Y. (Barclays Center)	W 70-59	10,046	9:30 p.m.	TruTV
11/26	<b>Furman</b>	4	Durham, N.C. (Cameron Indoor Stadium)	W 93-54	9,314	5 p.m.	ESPNU
11/30	<b>Army</b>	4	Durham, N.C. (Cameron Indoor Stadium)	W 93-73	9,314	12 p.m.	ESPNU
12/3	# at [2] Wisconsin	4	Madison, Wisc. (Kohl Center)	W 80-70	17,279	9:30 p.m.	ESPN
12/15	<b>Elon</b>	2	Durham, N.C. (Cameron Indoor Stadium)	W 75-62	9,314	7 p.m.	ESPNU
12/18	vs. Connecticut	2	East Rutherford, N.J. (Izod Center)	W 66-56	16,541	8 p.m.	ESPN
12/29	<b>Toledo</b>	2	Durham, N.C. (Cameron Indoor Stadium)	W 86-69	9,314	7 p.m.	ESPN2
12/31	<b>Wofford</b>	2	Durham, N.C. (Cameron Indoor Stadium)	W 84-55	9,314	3 p.m.	RSN
1/3	* <b>Boston College</b>	2	Durham, N.C. (Cameron Indoor Stadium)	W 85-62	9,314	4 p.m.	RSN
1/7	* at Wake Forest	2	Winston-Salem, N.C. (Joel Coliseum)	W 73-65	12,651	9 p.m.	ACCN
1/11	* at N.C. State	2	Raleigh, N.C. (PNC Arena)	L 75-87	19,500	1:30 p.m.	CBS
1/13	* <b>Miami</b>	4	Durham, N.C. (Cameron Indoor Stadium)	L 74-90	9,314	9 p.m.	ESPNU
1/17	* at [6] Louisville	4	Louisville, Ky. (KFC Yum! Center)	W 63-52	22,791	12 p.m.	ESPN
1/19	* <b>Pittsburgh</b>	5	Durham, N.C. (Cameron Indoor Stadium)	W 79-65	9,314	7 p.m.	ESPN
1/25	at St. Johns	5	New York, N.Y. (Madison Square Garden)	W 77-68	19,812	2 p.m.	FOX
1/28	* at [8] Notre Dame	4	Notre Dame, Ind. (Joyce Center)	L 73-77	9,149	7:30 p.m.	ESPN2
1/31	* at [2] Virginia	4	Charlottesville, Va. (John Paul Jones Arena)	W 69-63	14,593	7 p.m.	ESPN
2/4	* <b>Georgia Tech</b>	4	Durham, N.C. (Cameron Indoor Stadium)	W 72-66	9,314	7 p.m.	ESPN2
2/7	* [10] <b>Notre Dame</b>	4	Durham, N.C. (Cameron Indoor Stadium)	W 90-60	9,314	1 p.m.	CBS
2/9	* at Florida State	4	Tallahassee, Fla. (Donald L. Tucker Center)	W 73-70	11,498	7 p.m.	ESPN
2/14	* at Syracuse	4	Syracuse, N.Y. (Carrier Dome)	W 80-72	35,446	6 p.m.	ESPN
2/18	* [15] <b>North Carolina</b>	4	Durham, N.C. (Cameron Indoor Stadium)	W 92-90 •	9,314	9 p.m.	ESPN/ACCN
2/21	* <b>Clemson</b>	4	Durham, N.C. (Cameron Indoor Stadium)	W 78-56	9,314	4 p.m.	ESPN
2/25	* at Virginia Tech	4	Blacksburg, Va. (Cassell Coliseum)	W 91-86 •	9,847	9 p.m.	ESPN2
2/28	* <b>Syracuse</b>	4	Durham, N.C. (Cameron Indoor Stadium)	W 73-54	9,314	7 p.m.	ESPN
3/4	* <b>Wake Forest</b>	3	Durham, N.C. (Cameron Indoor Stadium)	W 94-51	9,314	8 p.m.	ACCN
3/7	* at [19] North Carolina	3	Chapel Hill, N.C. (Dean Dome)	W 84-77	21,750	9 p.m.	ESPN
3/12	\$\$\$ vs. N.C. State	2	Greensboro, N.C. (Greensboro Coliseum)	W 77-53	22,026	7 p.m.	ESPN
3/13	\$\$\$\$ vs. [11] Notre Dame	2	Greensboro, N.C. (Greensboro Coliseum)	L 64-74	22,026	9 p.m.	ESPN
3/20	!! vs. Robert Morris	4	Charlotte, N.C. (Time Warner Cable Arena)	W 85-56	16,945	7 p.m.	CBS
3/22	!!! vs. San Diego State	4	Charlotte, N.C. (Time Warner Cable Arena)	W 68-49	18,482	2 p.m.	CBS
3/27	!!!! vs. [19] Utah	4	Houston, Texas (NRG Stadium)	W 63-57	21,168	7:45 p.m.	CBS
3/29	!!!!!! vs. [7] Gonzaga	4	Houston, Texas (NRG Stadium)	W 66-52	20,744	4 p.m.	CBS
4/4	!!!!!! vs. [23] Michigan State	4	Indianapolis, Ind. (Lucas Oil Stadium)	W 81-61	72,238	6 p.m.	TBS/TNT
4/6	!!!!!!! vs. [3] Wisconsin	4	Indianapolis, Ind. (Lucas Oil Stadium)	W 68-63	71,149	9:15 p.m.	CBS

# copy

2014-15 Schedule & Results							
Date	[Rk] Opponent	Duke Rank	Location (Venue)	Score (OT)	Att.	Tip Time	TV
11/14	~ Presbyterian	4	Durham, N.C. (Cameron Indoor Stadium)	W 113-44	9,314	6 p.m.	ESPNU
11/15	~ Fairfield	4	Durham, N.C. (Cameron Indoor Stadium)	W 109-59	9,314	8 p.m.	ESPN3
11/18	!! vs. [19] Michigan State	4	Indianapolis, Ind. (Bankers Life Fieldhouse)	W 81-71	19,306	7 p.m.	ESPN
11/21	~ vs. Temple	4	Brooklyn, N.Y. (Barclays Center)	W 74-54	10,135	9:30 p.m.	TruTV
11/22	~ vs. Stanford	4	Brooklyn, N.Y. (Barclays Center)	W 70-59	10,046	9:30 p.m.	TruTV
11/26	Furman	4	Durham, N.C. (Cameron Indoor Stadium)	W 93-54	9,314	5 p.m.	ESPNU
11/30	Army	4	Durham, N.C. (Cameron Indoor Stadium)	W 93-73	9,314	12 p.m.	ESPNU
12/3	# at [2] Wisconsin	4	Madison, Wisc. (Kohl Center)	W 80-70	17,279	9:30 p.m.	ESPN
12/15	Elon	2	Durham, N.C. (Cameron Indoor Stadium)	W 75-62	9,314	7 p.m.	ESPNU
12/18	vs. Connecticut	2	East Rutherford, N.J. (Izod Center)	W 66-56	16,541	8 p.m.	ESPN
12/29	Toledo	2	Durham, N.C. (Cameron Indoor Stadium)	W 86-69	9,314	7 p.m.	ESPN2
12/31	Wofford	2	Durham, N.C. (Cameron Indoor Stadium)	W 84-55	9,314	3 p.m.	RSN
1/3	* Boston College	2	Durham, N.C. (Cameron Indoor Stadium)	W 85-62	9,314	4 p.m.	RSN
1/7	* at Wake Forest	2	Winston-Salem, N.C. (Joel Coliseum)	W 73-65	12,651	9 p.m.	ACCN
1/11	* at N.C. State	2	Raleigh, N.C. (PNC Arena)	L 75-87	19,500	1:30 p.m.	CBS
1/13	* Miami	4	Durham, N.C. (Cameron Indoor Stadium)	L 74-90	9,314	9 p.m.	ESPNU
1/17	* at [6] Louisville	4	Louisville, Ky. (KFC Yum! Center)	W 63-52	22,791	12 p.m.	ESPN
1/19	* Pittsburgh	5	Durham, N.C. (Cameron Indoor Stadium)	W 79-65	9,314	7 p.m.	ESPN
1/25	at St. Johns	5	New York, N.Y. (Madison Square Garden)	W 77-68	19,812	2 p.m.	FOX
1/28	* at [8] Notre Dame	4	Notre Dame, Ind. (Joyce Center)	L 73-77	9,149	7:30 p.m.	ESPN2
1/31	* at [2] Virginia	4	Charlottesville, Va. (John Paul Jones Arena)	W 69-63	14,593	7 p.m.	ESPN
2/4	* Georgia Tech	4	Durham, N.C. (Cameron Indoor Stadium)	W 72-66	9,314	7 p.m.	ESPN2
2/7	* [10] Notre Dame	4	Durham, N.C. (Cameron Indoor Stadium)	W 90-60	9,314	1 p.m.	CBS
2/9	* at Florida State	4	Tallahassee, Fla. (Donald L. Tucker Center)	W 73-70	11,498	7 p.m.	ESPN
2/14	* at Syracuse	4	Syracuse, N.Y. (Carrier Dome)	W 80-72	35,446	6 p.m.	ESPN
2/18	* [15] North Carolina	4	Durham, N.C. (Cameron Indoor Stadium)	W 92-90 •	9,314	9 p.m.	ESPN/ACCN
2/21	* Clemson	4	Durham, N.C. (Cameron Indoor Stadium)	W 78-56	9,314	4 p.m.	ESPN
2/25	* at Virginia Tech	4	Blacksburg, Va. (Cassell Coliseum)	W 91-86 •	9,847	9 p.m.	ESPN2
2/28	* Syracuse	4	Durham, N.C. (Cameron Indoor Stadium)	W 73-54	9,314	7 p.m.	ESPN
3/4	* Wake Forest	3	Durham, N.C. (Cameron Indoor Stadium)	W 94-51	9,314	8 p.m.	ACCN
3/7	* at [19] North Carolina	3	Chapel Hill, N.C. (Dean Dome)	W 84-77	21,750	9 p.m.	ESPN
3/12	\$\$\$ vs. N.C. State	2	Greensboro, N.C. (Greensboro Coliseum)	W 77-53	22,026	7 p.m.	ESPN
3/13	\$\$\$\$ vs. [11] Notre Dame	2	Greensboro, N.C. (Greensboro Coliseum)	L 64-74	22,026	9 p.m.	ESPN
3/20	!! vs. Robert Morris	4	Charlotte, N.C. (Time Warner Cable Arena)	W 85-56	16,945	7 p.m.	CBS
3/22	!!! vs. San Diego State	4	Charlotte, N.C. (Time Warner Cable Arena)	W 68-49	18,482	2 p.m.	CBS
3/27	!!!! vs. [19] Utah	4	Houston, Texas (NRG Stadium)	W 63-57	21,168	7:45 p.m.	CBS
3/29	!!!!!! vs. [7] Gonzaga	4	Houston, Texas (NRG Stadium)	W 66-52	20,744	4 p.m.	CBS
4/4	!!!!!! vs. [23] Michigan State	4	Indianapolis, Ind. (Lucas Oil Stadium)	W 81-61	72,238	6 p.m.	TBS/TNT
4/6	!!!!!!! vs. [3] Wisconsin	4	Indianapolis, Ind. (Lucas Oil Stadium)	W 68-63	71,149	9:15 p.m.	CBS

# paste

A	B	C	D	E	F	G	H	I	J
3	Day	2013 Opponent	2013	2013 Opponent	2013	2013	2013	2013	2013
4	<a href="#">14-Nov</a>	~	<b>Presbyterian</b>		4 Durham, N.C. (W)	113-44	9,314	6 p.m.	ESPNU
5	<a href="#">15-Nov</a>	~	<b>Fairfield</b>		4 Durham, N.C. (W)	109-59	9,314	8 p.m.	ESPN3
6	<a href="#">18-Nov</a>	¶¶	vs. [19] Michigan		4 Indianapolis, Ind. W	81-71	19,306	7 p.m.	ESPN
7	<a href="#">21-Nov</a>	~	vs. Temple		4 Brooklyn, N.Y. W	74-54	10,135	9:30 p.m.	TruTV
8	<a href="#">22-Nov</a>	~	vs. Stanford		4 Brooklyn, N.Y. W	70-59	10,046	9:30 p.m.	TruTV
9	<a href="#">26-Nov</a>		<b>Furman</b>		4 Durham, N.C. (W)	93-54	9,314	5 p.m.	ESPNU
10	<a href="#">30-Nov</a>		<b>Army</b>			93-73	9,314	12 p.m.	ESPNU
11	<a href="#">3-Dec</a>	#	at [2] Wisconsin				17,279	9:30 p.m.	ESPN
12	<a href="#">15-Dec</a>		<b>Elon</b>				9,314	7 p.m.	ESPNU
13	<a href="#">18-Dec</a>		vs. Connecticut				16,541	8 p.m.	ESPN
14	<a href="#">29-Dec</a>		<b>Toledo</b>				9,314	7 p.m.	ESPN2
15	<a href="#">31-Dec</a>		<b>Wofford</b>				9,314	3 p.m.	RSN
16	<a href="#">3-Jan</a>	*	<b>Boston U</b>				9,314	4 p.m.	RSN
17	<a href="#">7-Jan</a>	*	at Wake Forest				9,314	9 p.m.	ACCN
18	<a href="#">11-Jan</a>	*	at N.C. State					1:30 p.m.	CBS
19	<a href="#">13-Jan</a>	*	<b>Miami (Fla.)</b>					9 p.m.	ESPNU
20	<a href="#">17-Jan</a>	*	at [6] Louisville					12 p.m.	ESPN
21	<a href="#">19-Jan</a>	*	<b>Pittsburgh</b>					7 p.m.	ESPN
22	<a href="#">25-Jan</a>		at St. John's				9,314	2 p.m.	FOX
23	<a href="#">28-Jan</a>	*	at [8] Notre Dame				9,314	7:30 p.m.	ESPN2
24	<a href="#">31-Jan</a>	*	at [2] Virginia				9,314	7 p.m.	ESPN
25	<a href="#">4-Feb</a>	*	<b>Georgia Tech</b>				9,314	7 p.m.	ESPN2
26	<a href="#">7-Feb</a>	*	[10] Notre Dame				9,314	1 p.m.	CBS
27	<a href="#">9-Feb</a>	*	at Florida State				11,498	7 p.m.	ESPN
28	<a href="#">14-Feb</a>	*	at Syracuse				35,446	6 p.m.	ESPN
29	<a href="#">18-Feb</a>	*	[15] North Carolina				9,314	9 p.m.	ESPN/ACCN
30	<a href="#">21-Feb</a>	*	<b>Clemson</b>		Durham, N.C. (W)	9,314	9,314	4 p.m.	ESPN
31	<a href="#">25-Feb</a>	*	at Virginia Tech		4 Blacksburg, Va. W	91-86	9,847	9 p.m.	ESPN2
32	<a href="#">28-Feb</a>	*	<b>Syracuse</b>		4 Durham, N.C. (W)	73-54	9,314	7 p.m.	ESPN
33	<a href="#">4-Mar</a>	*	<b>Wake Forest</b>		3 Durham, N.C. (W)	94-51	9,314	8 p.m.	ACCN
34	<a href="#">7-Mar</a>	*	at [19] North Carolina		3 Chapel Hill, N.C. W	84-77	21,750	9 p.m.	ESPN
35	<a href="#">12-Mar</a>	\$\$\$	vs. N.C. State		2 Greensboro, N.C. W	77-53	22,026	7 p.m.	ESPN
36	<a href="#">13-Mar</a>	\$\$\$\$	vs. [11] Notre Dame		2 Greensboro, N.C. L	64-74	22,026	9 p.m.	ESPN
37	<a href="#">20-Mar</a>	!!	vs. Robert Morris		4 Charlotte, N.C. W	85-56	16,945	7 p.m.	CBS
38	<a href="#">22-Mar</a>	!!!	vs. San Diego State		4 Charlotte, N.C. W	68-49	18,482	2 p.m.	CBS
39	<a href="#">27-Mar</a>	!!!!	vs. [19] Utah		4 Houston, Texas W	63-57	21,168	7:45 p.m.	CBS
40	<a href="#">29-Mar</a>	!!!!!	vs. [7] Gonzaga		4 Houston, Texas W	66-52	20,744	4 p.m.	CBS
41	<a href="#">4-Apr</a>	!!!!!!	vs. [23] Michigan		4 Indianapolis, Ind. W	81-61	72,238	6 p.m.	TBS/TNT
42	<a href="#">6-Apr</a>	!!!!!!	vs. [3] Wisconsin		4 Indianapolis, Ind. W	68-63	71,149	9:15 p.m.	CBS

# scrape

```
# Load packages -----
library(rvest)
library(stringr)
library(dplyr)

# Read page with season data -----
page <- read_html("http://goduke.statsgeek.com/basketball-m/seasons/schedule.php?season=2014-15")

# Harvest fields -----
date <- page %>%
  html_nodes(".stattextline b") %>%
  html_text()

opponent <- page %>%
  html_nodes(".stattextltgray2:nth-child(3)") %>%
  html_text() %>%
  str_trim()

venue <- page %>%
  html_nodes(".stattextltgray2:nth-child(5)") %>%
  html_text() %>%
  str_trim()

# Put fields into a tibble -----
blue_devils_1415 <- data_frame(date, opponent, venue)
```

# voila!

blue\_devils\_1415 \*

Filter

	date	opponent	venue
1	11/14	Presbyterian	Durham, N.C. (Cameron Indoor Stadium)
2	11/15	Fairfield	Durham, N.C. (Cameron Indoor Stadium)
3	11/18	vs. [19] Michigan State	Indianapolis, Ind. (Bankers Life Fieldhouse)
4	11/21	vs. Temple	Brooklyn, N.Y. (Barclays Center)
5	11/22	vs. Stanford	Brooklyn, N.Y. (Barclays Center)
6	11/26	Furman	Durham, N.C. (Cameron Indoor Stadium)
7	11/30	Army	Durham, N.C. (Cameron Indoor Stadium)
8	12/3	at [2] Wisconsin	Madison, Wisc. (Kohl Center)
9	12/15	Elon	Durham, N.C. (Cameron Indoor Stadium)
10	12/18	vs. Connecticut	East Rutherford, N.J. (Izod Center)
11	12/29	Toledo	Durham, N.C. (Cameron Indoor Stadium)
12	12/31	Wofford	Durham, N.C. (Cameron Indoor Stadium)

Showing 1 to 13 of 39 entries

# best laid plans...



**Mine CetinkayaRundel**

@minebocek

Students upset b/c website they need to scrape data from for hw assignment is down. Bad assignment or good lesson in working w/ real data?

RETWEET

1

LIKES

10



9:48 AM - 26 Nov 2015

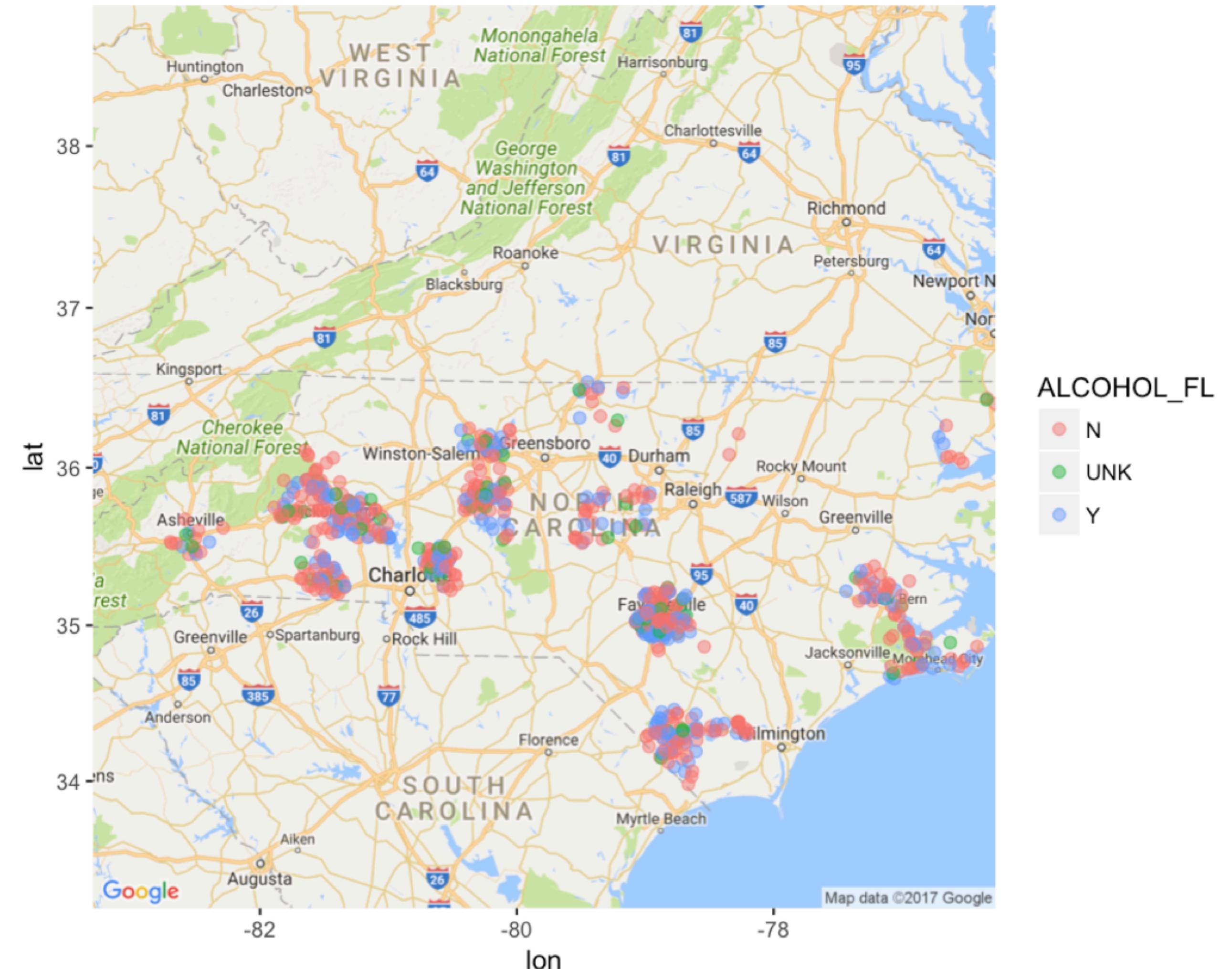
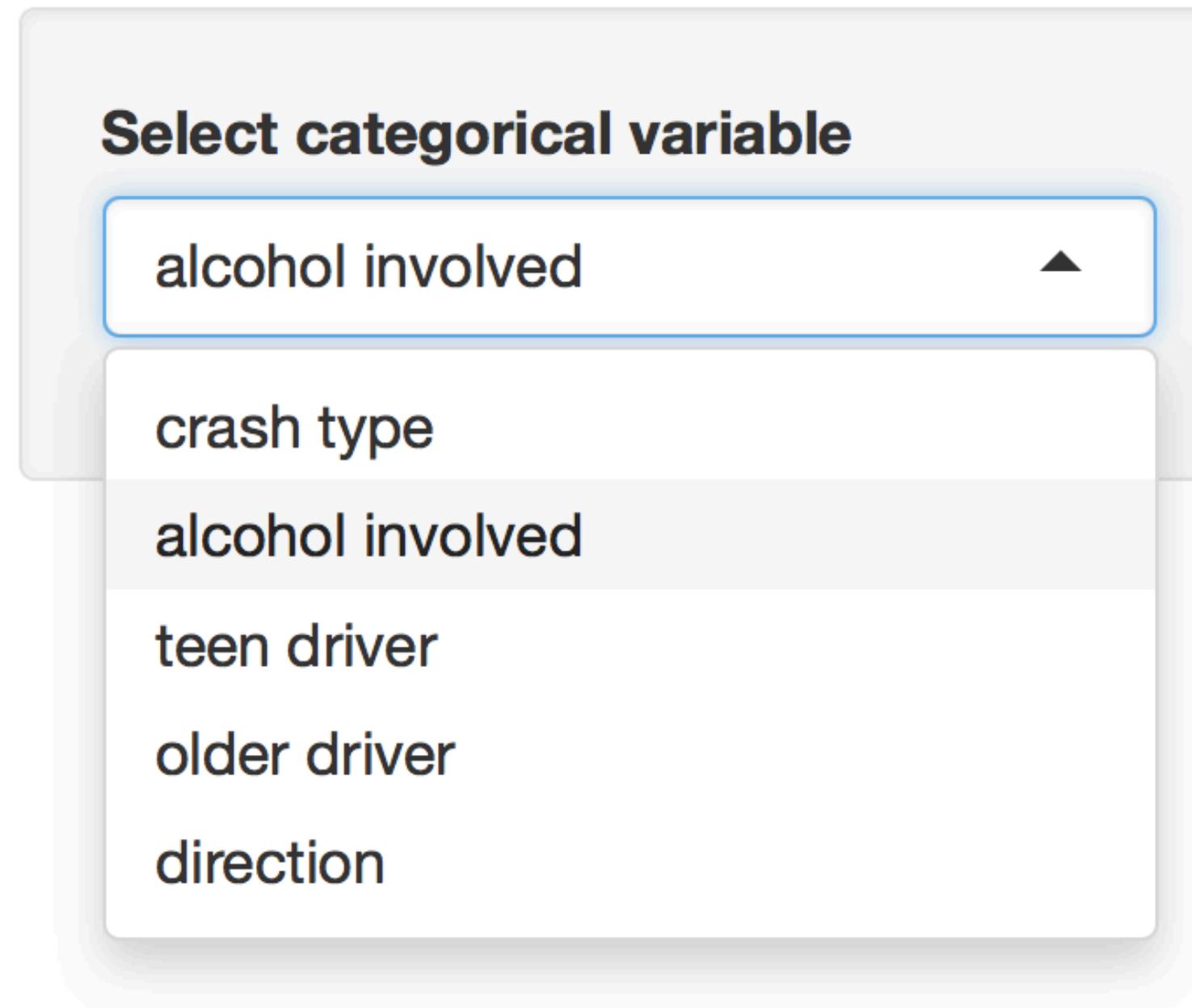
# #3 interactivity



```
> library(shiny)
```



# Modeling the Distributions of Fatal Car Crashes



motivation

computation

interest &  
impact

syllabus

data analysis  
examples

curricular  
considerations

# impact

## **pipeline for stats:**

2014: 19% declared  
2015: 31% declared  
2016: ~40%  
expressed interest

## **diversity:**

% female  
2014: 44%  
2015: 50%  
2016: 35%  
2017: 55%  
~25% in Probability

## **curricular:**

basis for  
gateway to stats  
major course  
to be offered in  
Spring 2018!

motivation

computation

interest &  
impact

syllabus

data analysis  
examples

**curricular  
considerations**

# curricular considerations

move away from  
ad-hoc computing  
education  
and/or  
expecting students  
to pick it up  
along the way

uniformity of tools is  
important: choose a  
toolkit that works for  
you and stick to it  
throughout the  
curriculum

teach computing  
early (without any  
prereqs) and often!



Thank you!

 @minebocek

 mine-cetinkaya-rundel

 mine@rstudio.com