Start with **data** science

🔗 **bit.ly/start-w-ds**

Mine Cetinkaya-Rundel
Duke University + RStudio

@minebocek
mine-cetinkaya-rundel
cetinkaya.mine@gmail.com

RStudio®

Duke Statistical Science

Goal: Educate the new generation of data scientists

- working on ML and AI problems
- not intimidated by learning new computing technologies

Where do we start?

Where do we start?

- How early?
- How long?
- How inclusive?

Q How early?	as early as possible
Q How long?	10-15 weeks
Q How inclusive?	yes!

So, really,
where do we start?

case study

teacher salaries

average total SAT score, 1994-95

estimated average annual salary of teachers in public elementary and secondary schools

percentage of all eligible students taking the SAT

	state	salary	sat	frac
1	Alabama	31.1	1029	8
2	Alaska	48.0	934	47
3	Arizona	32.2	944	27
4	Arkansas	28.9	1005	6
5	California	41.1	902	45
6	Colorado	34.6	980	29
7	Connecticut	50.0	908	81
8	Delaware	39.1	897	68
9	Florida	32.6	889	48
10	Georgia	32.3	854	65

... with 40 more rows

mosaicData

Randall Pruim, Daniel Kaplan and Nicholas Horton (2018). mosaicData: Project MOSAIC Data Sets. R package version 0.17.0. https://CRAN.R-project.org/package=mosaicData

tidyverse

Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

broom

David Robinson and Alex Hayes (2018). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.0. https://CRAN.R-project.org/package=broom

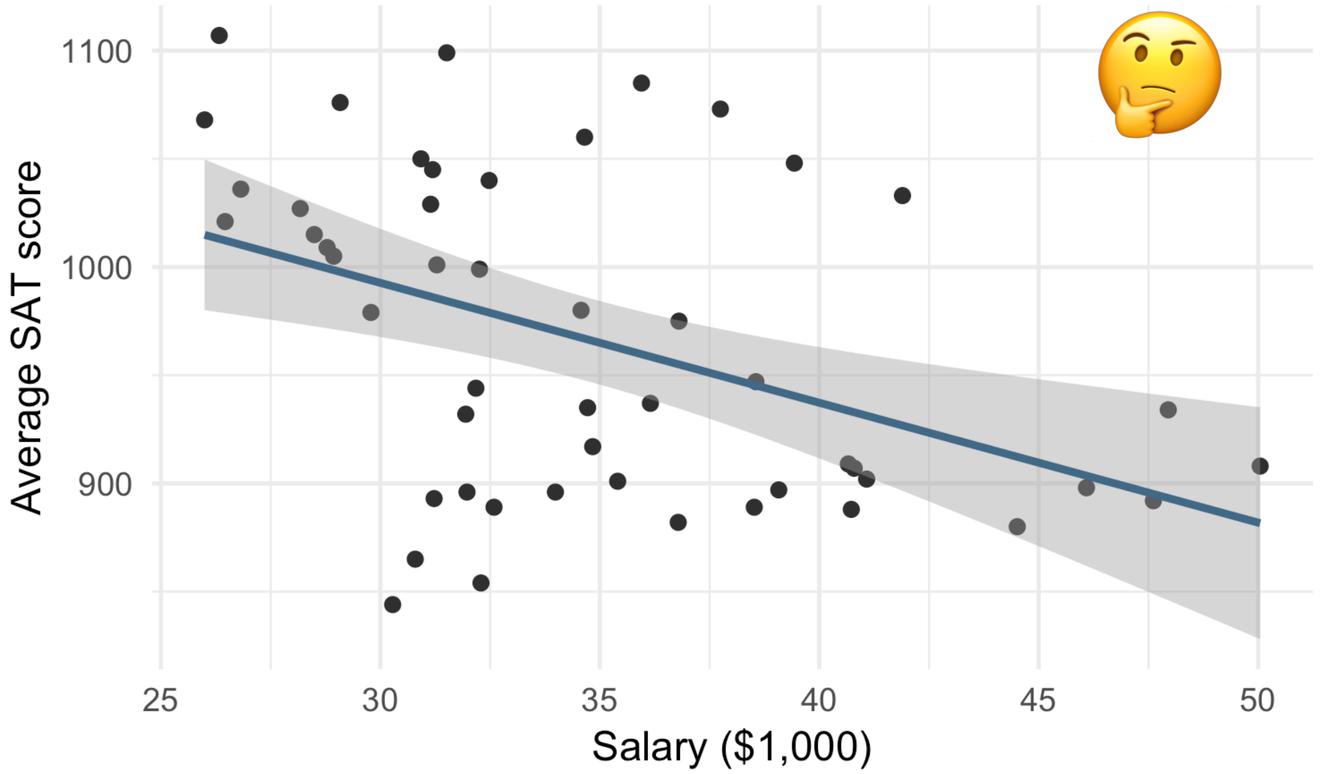
reprex

Jennifer Bryan, Jim Hester, David Robinson and Hadley Wickham (2018). reprex: Prepare Reproducible Example Code via the Clipboard. R package version 0.2.1. https://CRAN.R-project.org/package=reprex

option 1

prediction

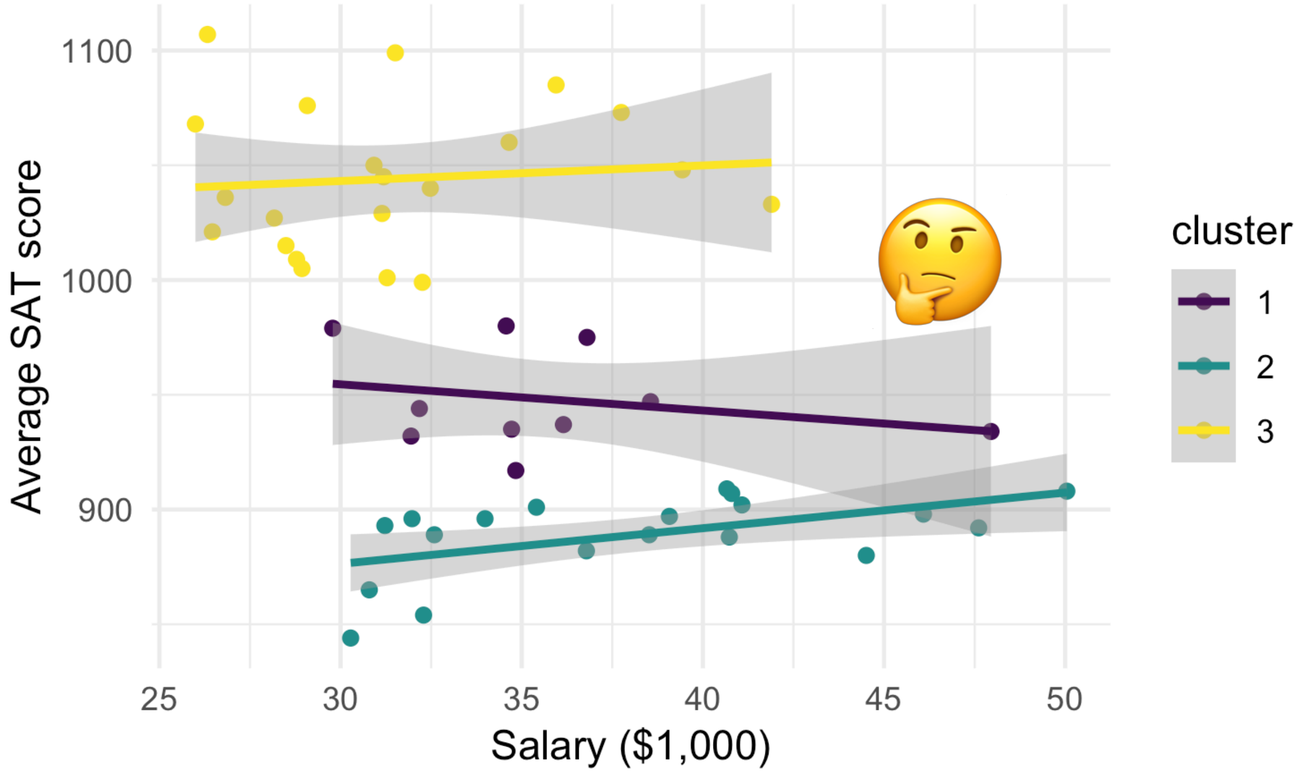
```r
mod_sat_sal ← lm(sat ~ salary, data = SAT)

new_teacher ← tibble(salary = 40)

predict(mod_sat_sal, new_teacher)
#>        1
#> 937.2742
```



option 2

clustering

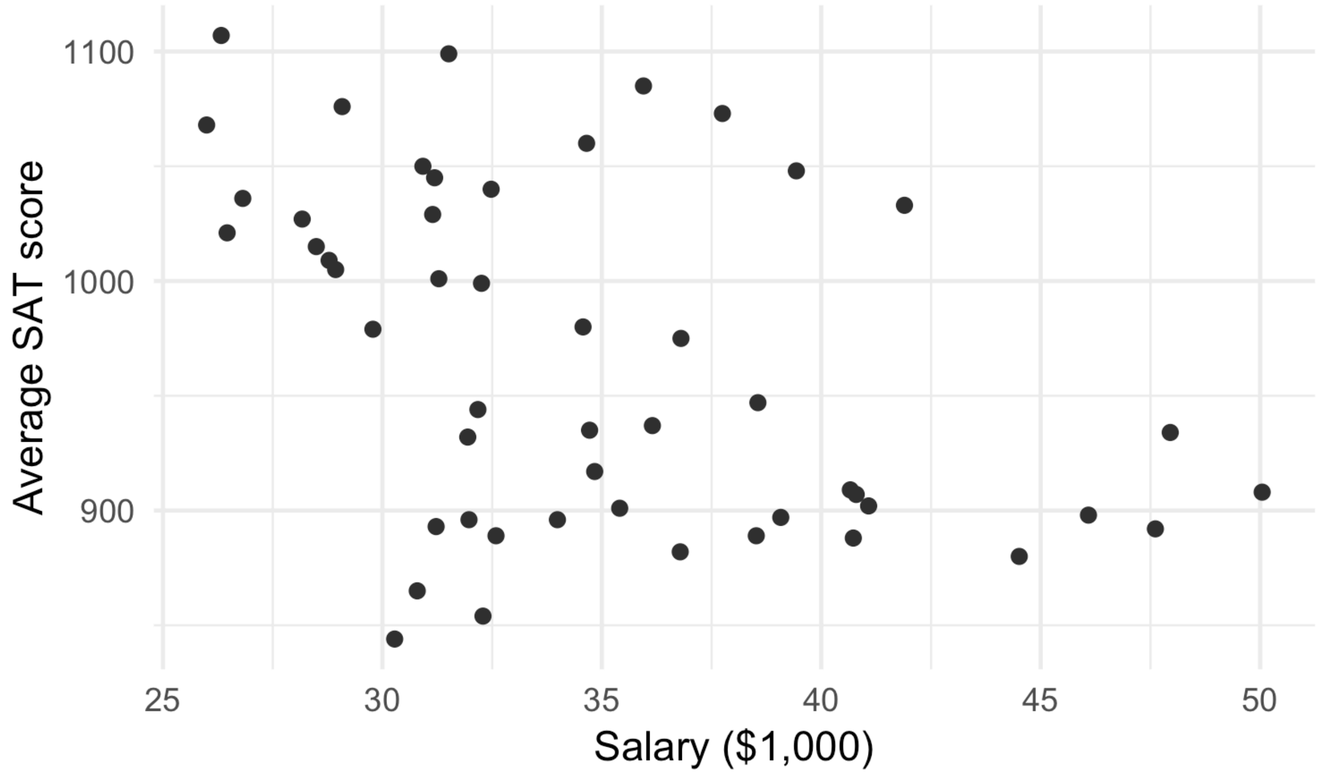
```r
clusters ← kmeans(SAT %>% select(salary, sat, frac), centers = 3)

SAT ← SAT %>%
  mutate(cluster = factor(clusters$cluster))
```



option 3

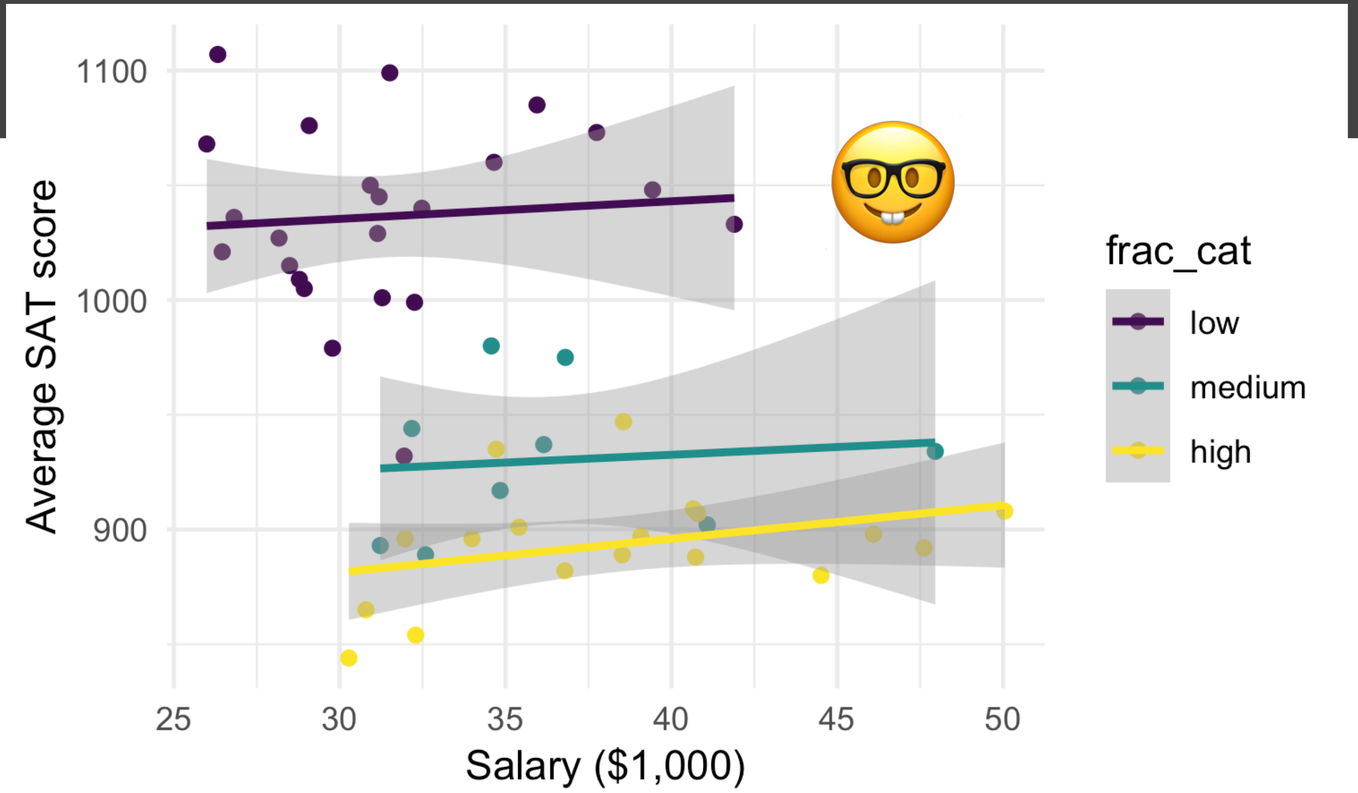
exploration

```r
ggplot(SAT, aes(x = salary, y = sat)) +
  geom_point() +
  labs(x = "Salary ($1,000)", y = "Average SAT score") +
  theme_minimal()
```



```r
ggplot(SAT, aes(x = salary, y = sat, color = frac)) +
  geom_point() +
  theme_minimal() +
  labs(x = "Salary ($1,000)", y = "Average SAT score") +
  scale_color_viridis_c()
```

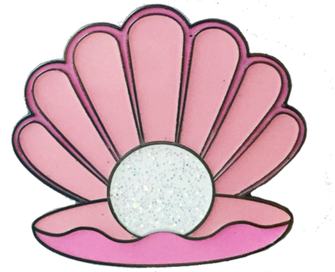
```r
SAT ← SAT %>%
  mutate(frac_cat = cut(frac, breaks = c(0, 22, 49, 81),
                        labels = c("low", "medium", "high")))

ggplot(SAT, aes(x = salary, y = sat, color = frac_cat)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Salary ($1,000)", y = "Average SAT score") +
  theme_minimal() +
  scale_color_viridis_d()
```



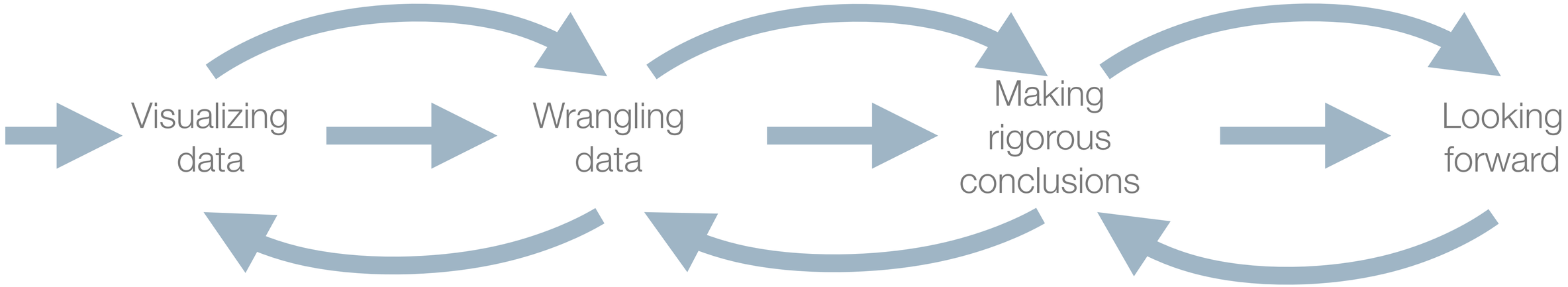
exploratory data analysis

descriptive models

predictive models



What does a semester long curriculum look like?

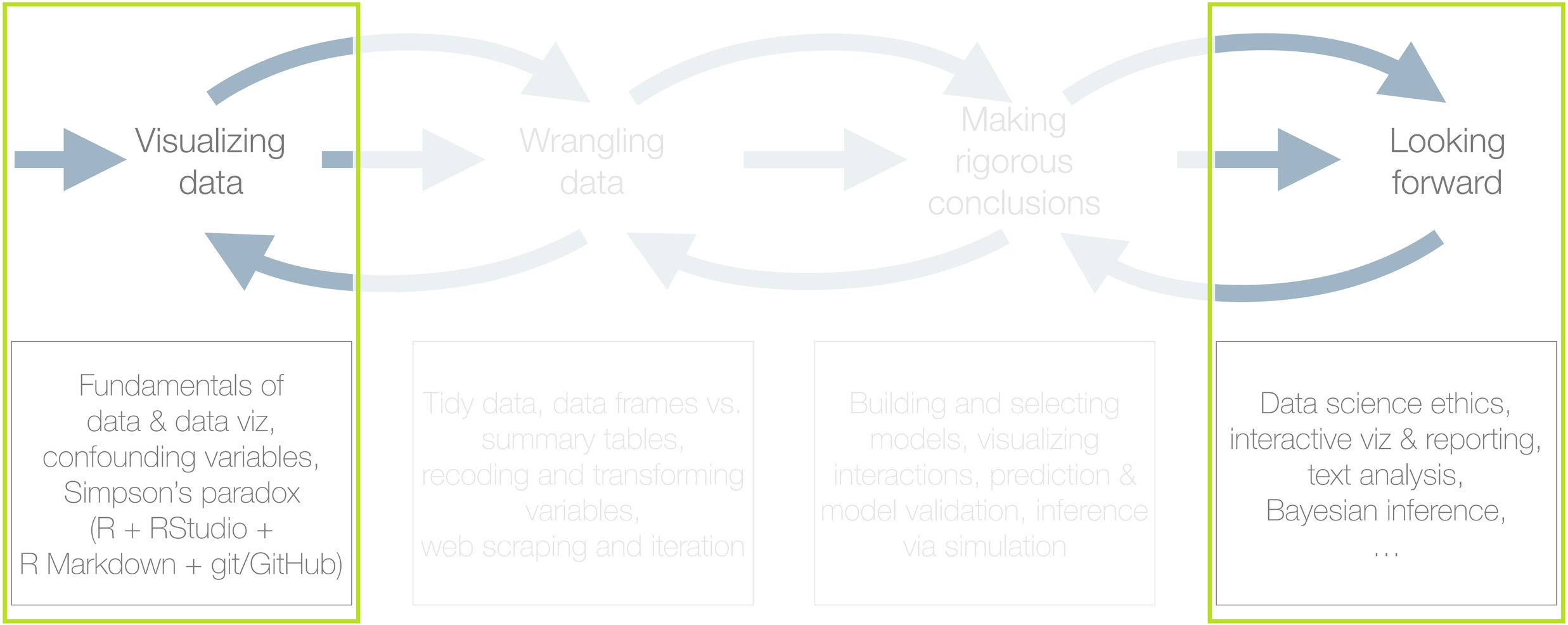


Fundamentals of data & data viz, confounding variables, Simpson's paradox (R + RStudio + R Markdown + git/GitHub)

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

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

Data science ethics, interactive viz & reporting, text analysis, Bayesian inference, …


Visualizing data
Wrangling data
Making rigorous conclusions
Looking forward
Fundamentals of data & data viz, confounding variables, Simpson's paradox (R + RStudio + R Markdown + git/GitHub)
Tidy data, data frames vs. summary tables, recoding and transforming variables, web scraping and iteration
Building and selecting models, visualizing interactions, prediction & model validation, inference via simulation
Data science ethics, interactive viz & reporting, text analysis, Bayesian inference, ...

Why start with visualization?

more likely for students to have intuition for interpretations coming in

easier for them to catch their own mistakes

great way to introduce programming

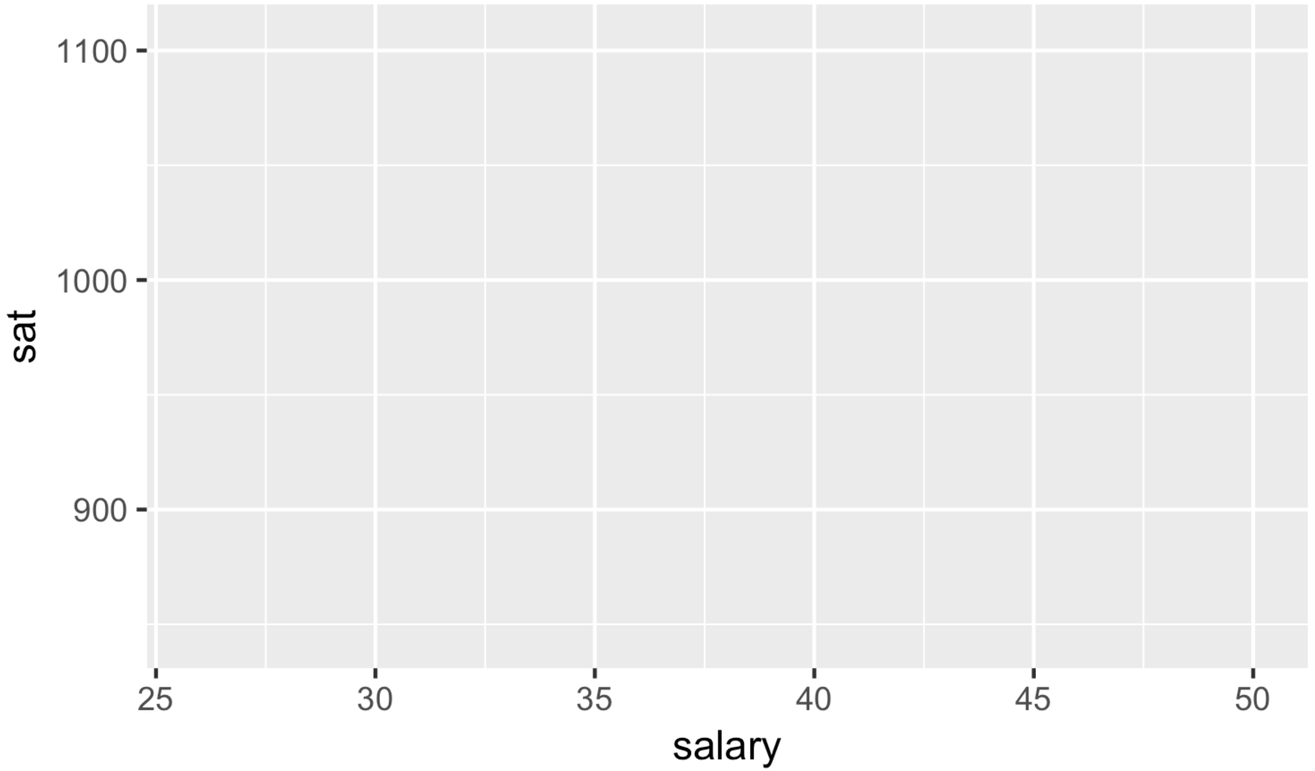
```
ggplot(SAT)
```

```
ggplot(SAT, aes(x = salary, y = sat))
```

```
function( arguments )
```

often a verb

what to apply that verb to

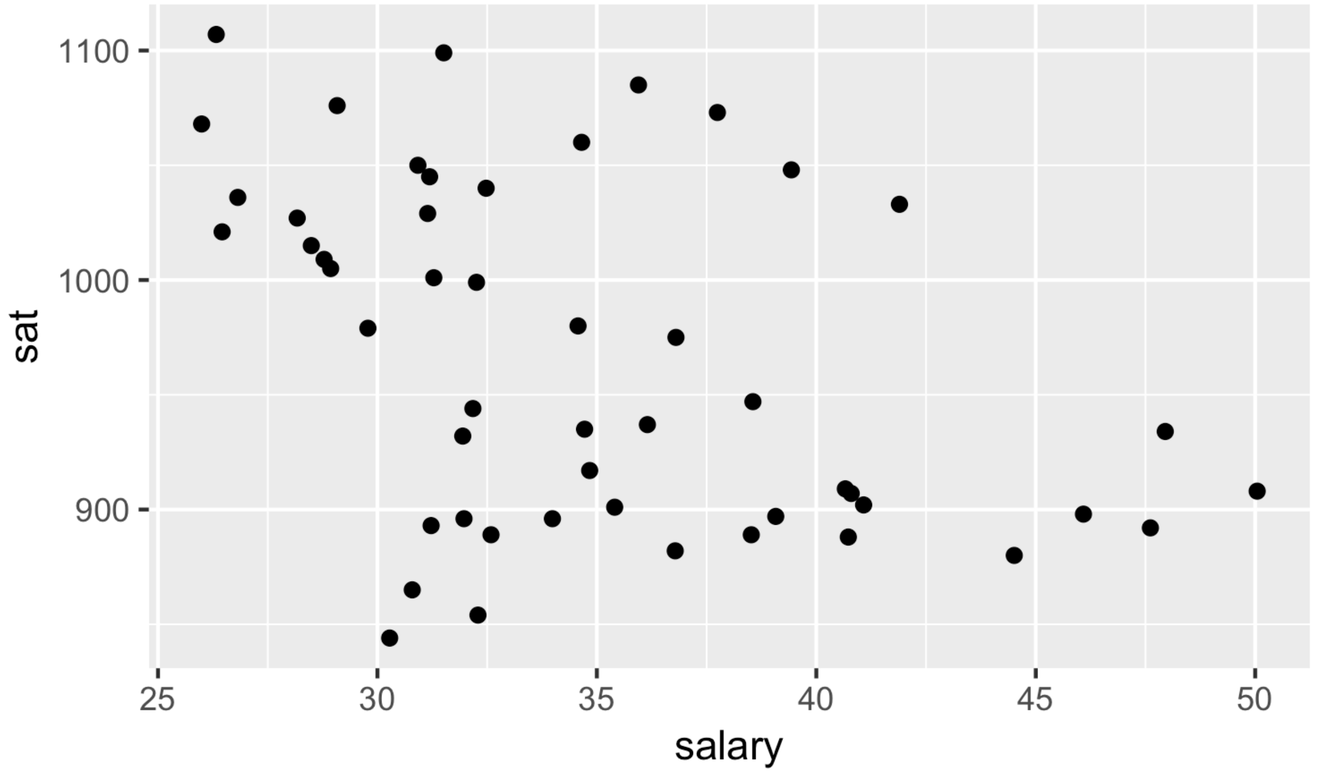


```
ggplot(SAT, aes(x = salary, y = sat)) +
  geom_point()
```

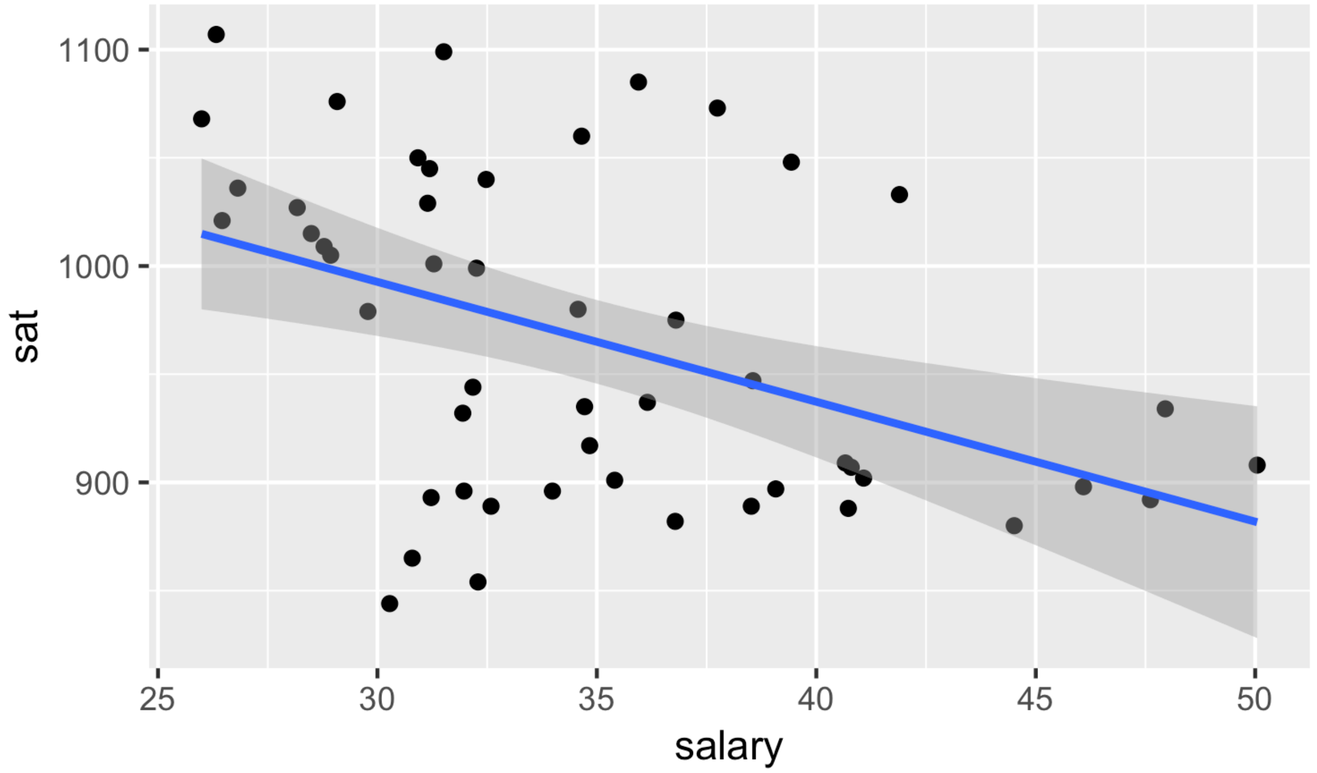
tidy data frame

	state	salary	sat	frac
1	Alabama	31.1	1029	8
2	Alaska	48.0	934	47
3	Arizona	32.2	944	27
4	Arkansas	28.9	1005	6
5	California	41.1	902	45
6	Colorado	34.6	980	29
7	Connecticut	50.0	908	81
8	Delaware	39.1	897	68
9	Florida	32.6	889	48
10	Georgia	32.3	854	65

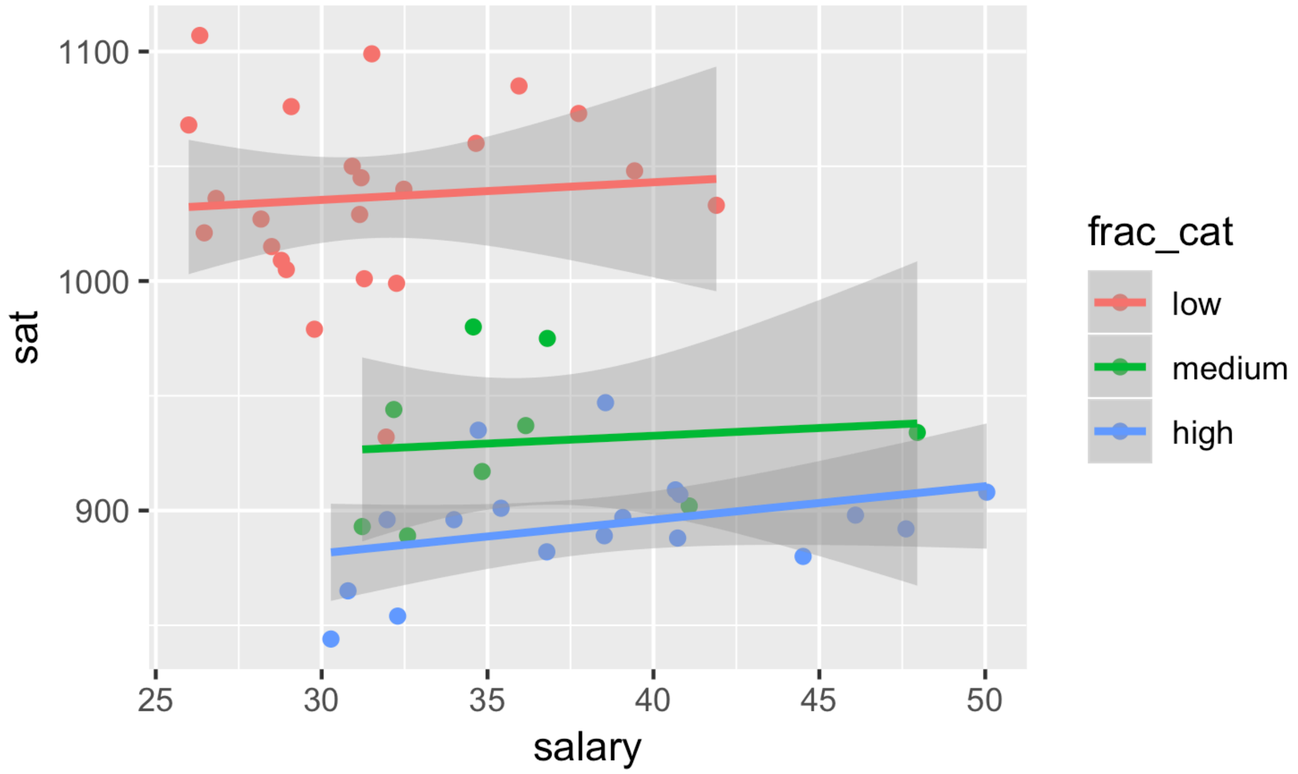
... with 40 more rows



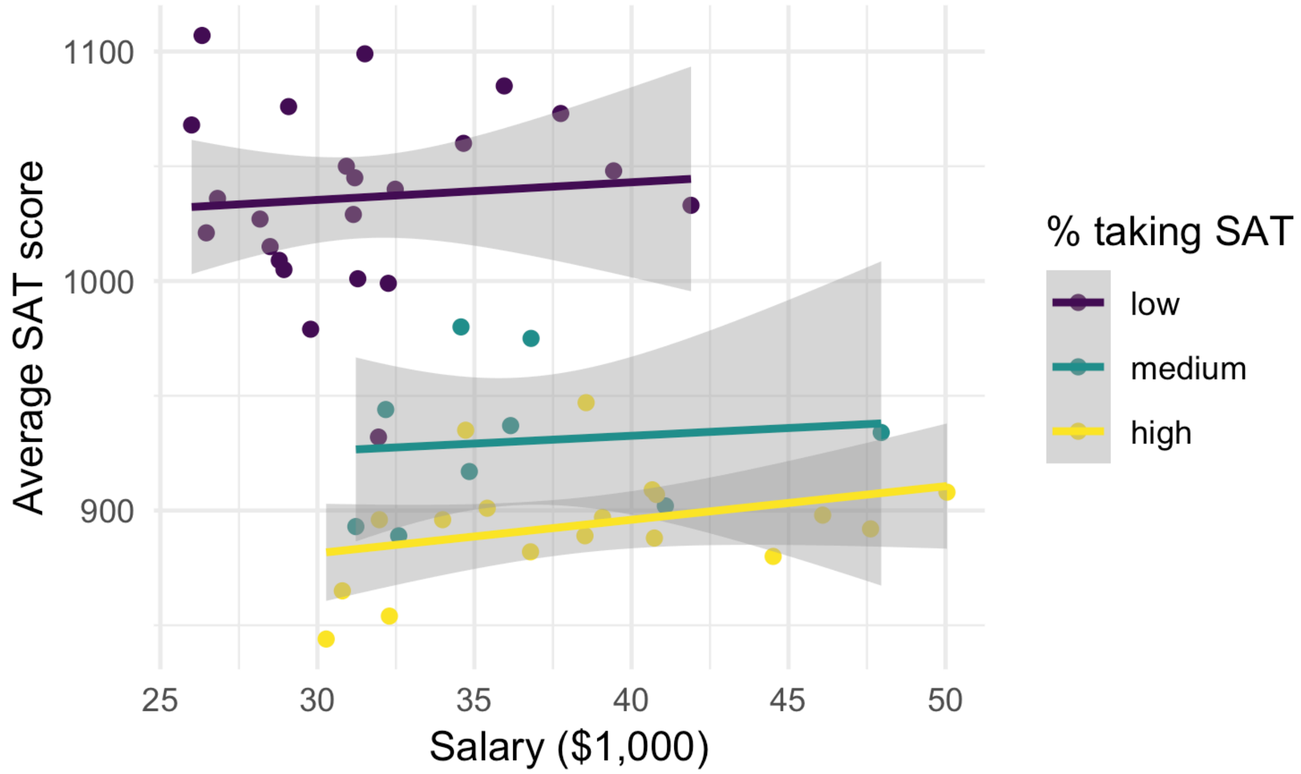
```
ggplot(SAT, aes(x = salary, y = sat)) +
  geom_point() +
  geom_smooth(method = "lm")
```



```
ggplot(SAT, aes(x = salary, y = sat, color = frac_cat)) +
  geom_point() +
  geom_smooth(method = "lm")
```



```
ggplot(SAT, aes(x = salary, y = sat, color = frac_cat)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Salary ($1,000)", y = "Average SAT score",
       color = "% taking SAT") +
  theme_minimal() +
  scale_color_viridis_d()
```



Why touch on ethics?
And how?

empower,
and warn,
at the same time

help students think beyond what the course curriculum can offer

do so using case studies they can relate to based on course curriculum

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

conditional probabilities

prediction

data available!

Source: propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

How to write a racist AI in R without really trying

5 min read

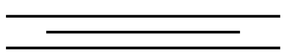
2018/09/27

Last year, Rob Speer wrote a really great post How to make a racist AI without really trying. Go read it.

The idea is to do sentiment analysis with obvious, off-the-shelf tools. As the post says

> So that's what we're going to do here, following the path of least resistance at every step, obtaining a classifier that should look very familiar to anyone involved in current NLP.

The original post used Python and I'm teaching an undergraduate data science course using R at the moment, so I wanted an R version. There were two issues in converting the code: my laptop doesn't really have



Things seem to be working. Now for the punch line

```
> sentiment("Let's go out for Italian food.")
[1] 1.387002
> sentiment("Let's go out for Chinese food.")
[1] 1.04452
> sentiment("Let's go out for Mexican food.")
[1] 0.6954334
```

training a model

sentiment analysis

implementation in R

Source: notstatschat.rbind.io/2018/09/27/how-to-write-a-racist-ai-in-r-without-really-trying/

Fine,
I'm intrigued,
but I need to see
the big picture

Topics :: Data Science in a Box

datasciencebox.org

Search...

Hello #dsbox

- Overview
- Philosophy
- Topics
- Tech stack
- Community

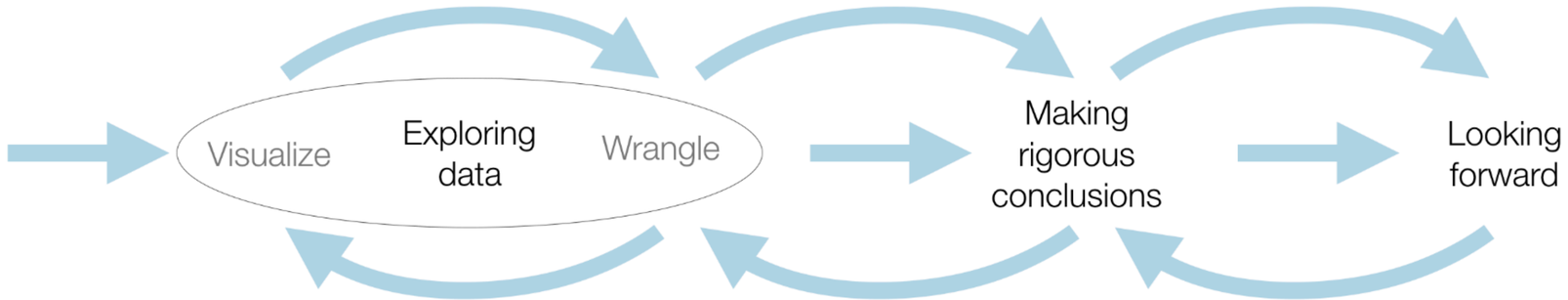
Course content

Infrastructure

Pedagogy

Topics

The course content is organized in three units:



Unit 1 - Exploring data: This unit focuses on data visualiation and data wranling. Specifically we cover fundamentals of data and data visualization, confounding variables, and Simpson's paradox as well as the concept of tidy data, data import, data cleaning, and data curation. We end the unit with web scraping and introduce the idea of iteration in preparation for the next unit. Also in this unit students are introduced to the toolkit: R, RStudio, R Markdown, Git, GitHub, etc.

Unit 2 - Making rigorous conclusions: In this part we introduce modeling and statistical inference for making data based conclusions. We discuss building, interpreting, and selecting models, visualizing interaction effects, and prediction and model validity. Statistical inference is introduced from a simulation based perspective, and the Central Limit Theorem is discussed very briefly to lay the foundation for future coursework in statistics.

Unit 3 - Looking forward: In the last unit we present a series of modules such as interactive reporting and visualizaiton with Shiny, text analysis, and Bayesian inference. These are independent modules that instructors can choose to include in their introductory data science curriculum depending on how much time they have left in the semester.

bit.ly/start-w-ds

@minebocek

mine-cetinkaya-rundel

cetinkaya.mine@gmail.com