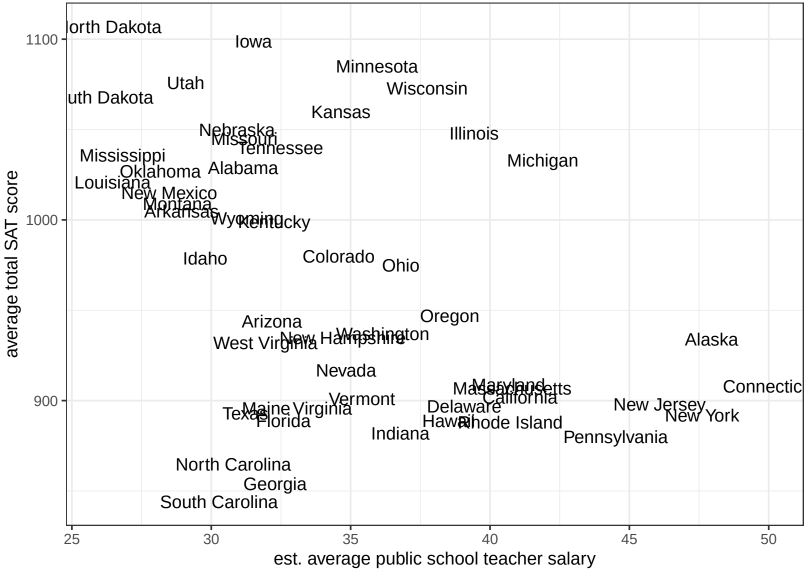
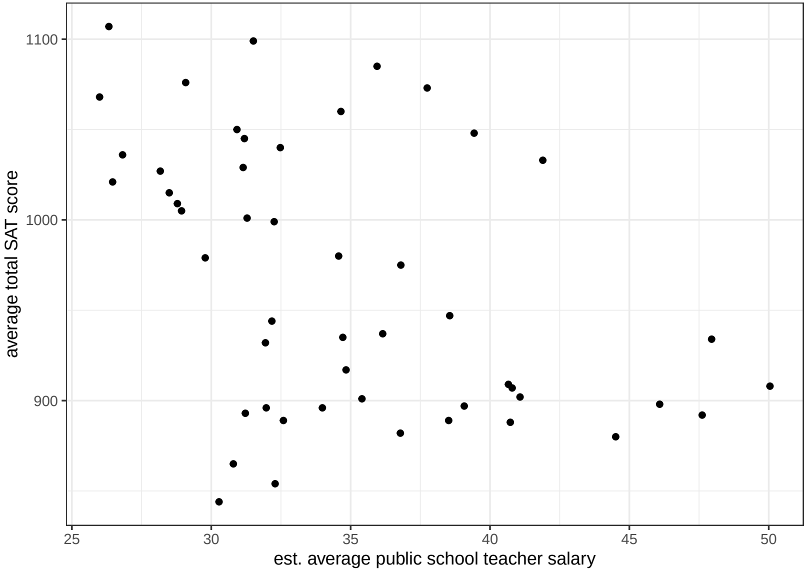
Introduction to Telling Stories with Data

Author: Nicholas G Reich

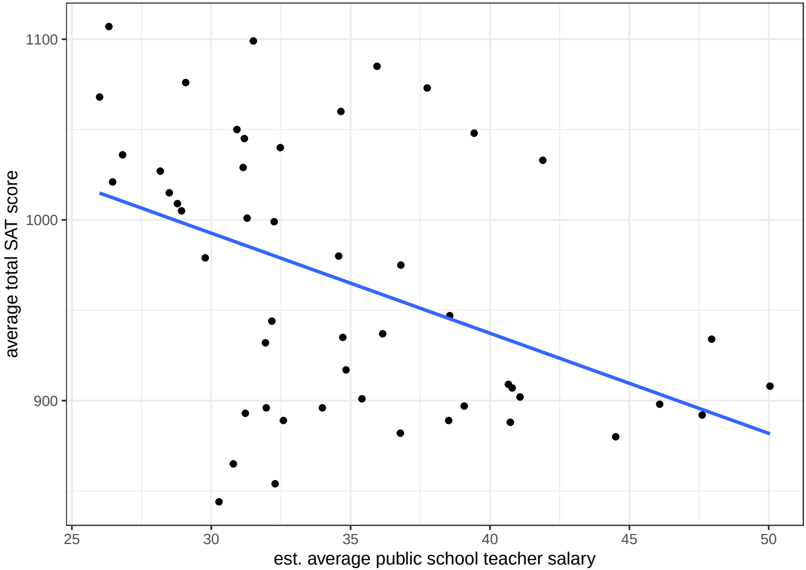
State-level SAT score data (1994-95)



State-level SAT score data (1994-95)



State-level SAT score data (1994-95)



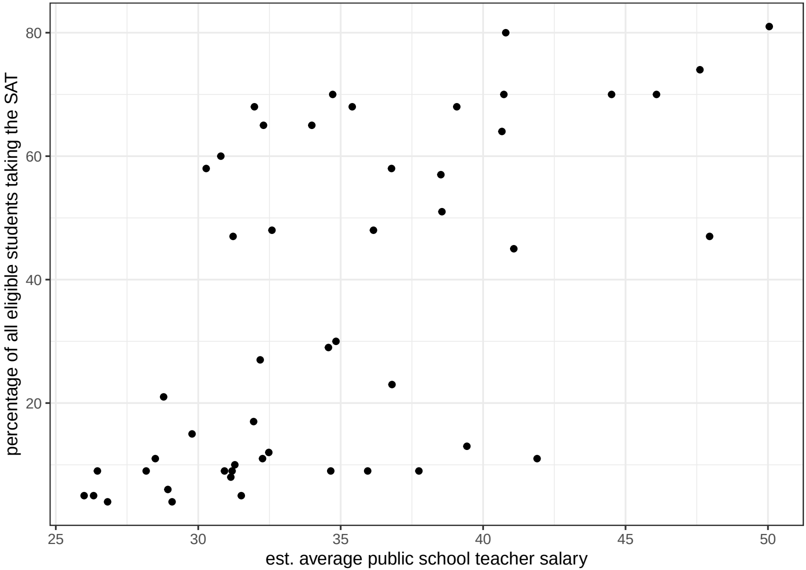
The SAT example

What is the outcome variable?

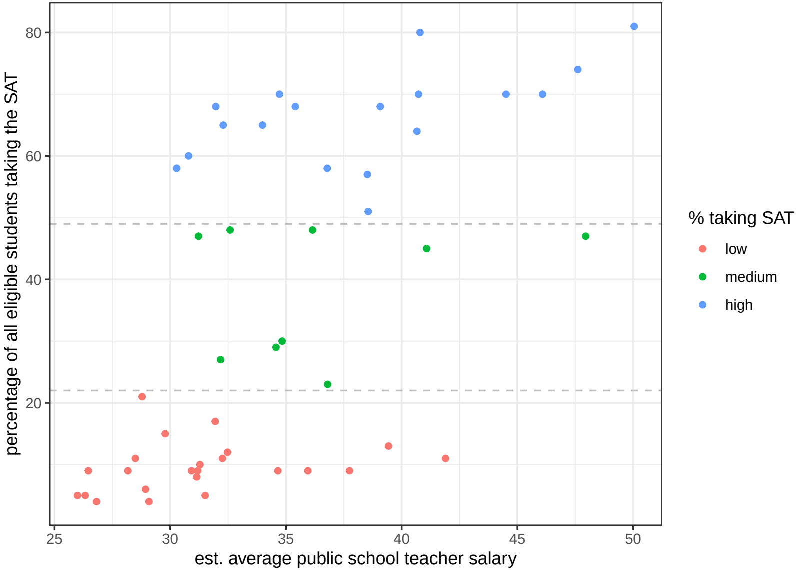
What is the covariate or predictor variable?

What other data might be part of this story?

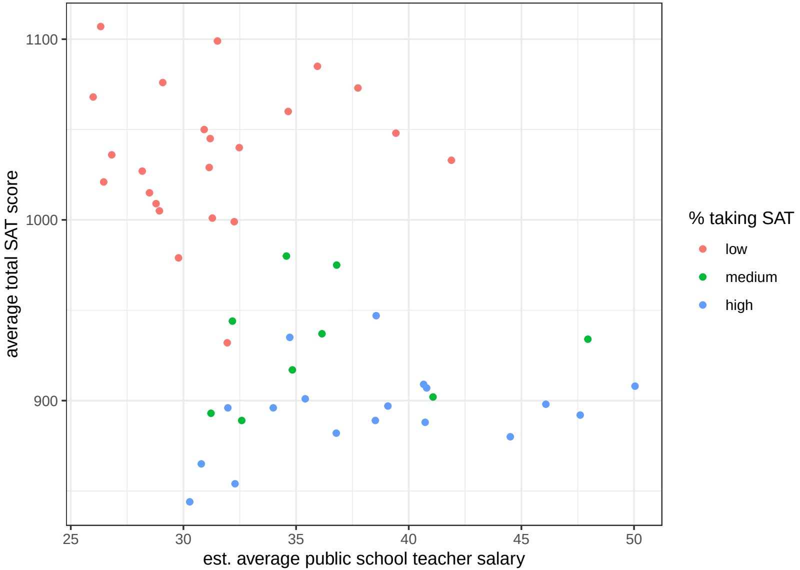
State-level SAT score data (1994-95)



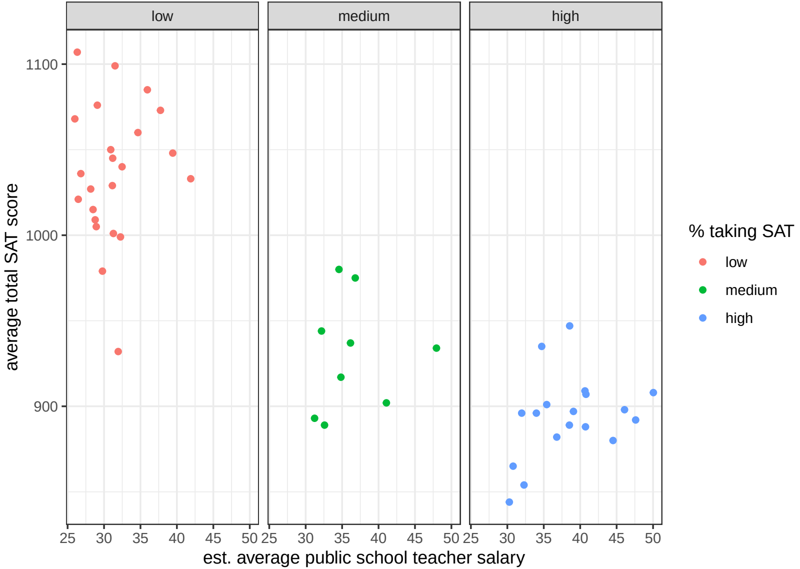
State-level SAT score data (1994-95)



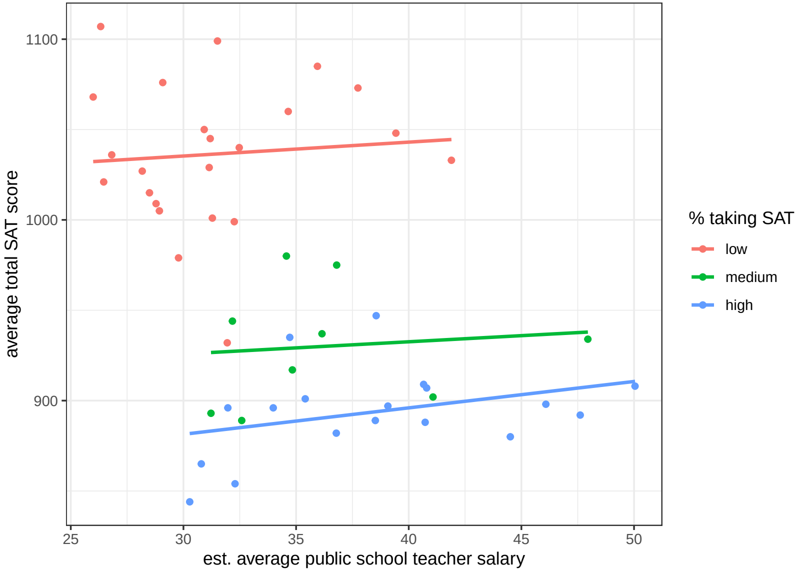
State-level SAT score data (1994-95)



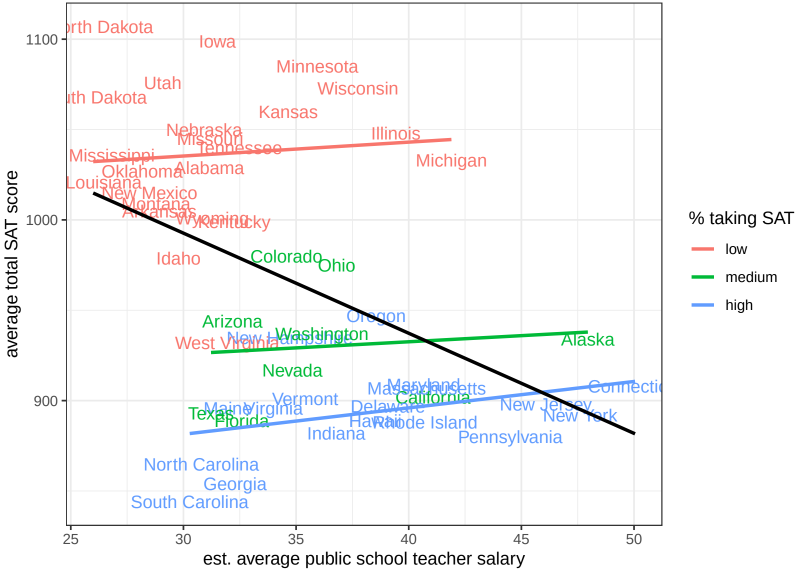
State-level SAT score data (1994-95)



State-level SAT score data (1994-95)



State-level SAT score data (1994-95)



State-level SAT score data (1994-95)

What can we conclude from all of this? (BTW, this is an example of "Simpson's Paradox".)

Regression modeling

The process of using data to describe the relationship between outcomes and predictors is called modeling.

- Models are models, not reality.
- “All models are wrong, but some are useful.”
- Introduce structure to our model that balances realism with “goodness of fit”.

Appendix: Code for plotting

```r
library(mosaicData)
library(ggplot2)
theme_set(theme_bw())
data(SAT)
SAT$fracgrp = cut(SAT$frac, breaks=c(0, 22, 49, 81),
                  labels=c("low", "medium", "high"))
ggplot(SAT) +
    geom_text(aes(x=salary, y=sat, label=state), size=4, show.legend=FALSE) +
    xlab("est. average public school teacher salary") +
    ylab("average total SAT score")
```

More plotting code available here.