Introduction to Telling Stories with Data

Author: Nicholas G Reich

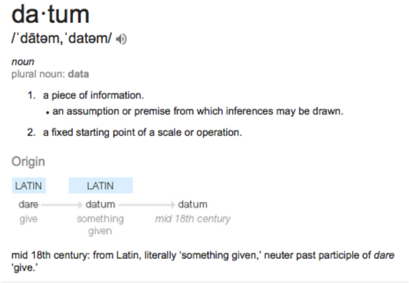
This material is part of the ***statsTeachR*** *project*

Communicating ideas with evidence

What is a narrative? [From the OED]

An account of a series of events, facts, etc., given in order and with the establishing of connections between them; a narration, a story, an account.

What is data? [From Google: literally, “what is data”]



da·tum

/ˈdātəm,ˈdatəm/

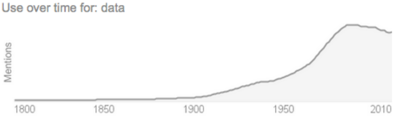
noun

plural noun: **data**

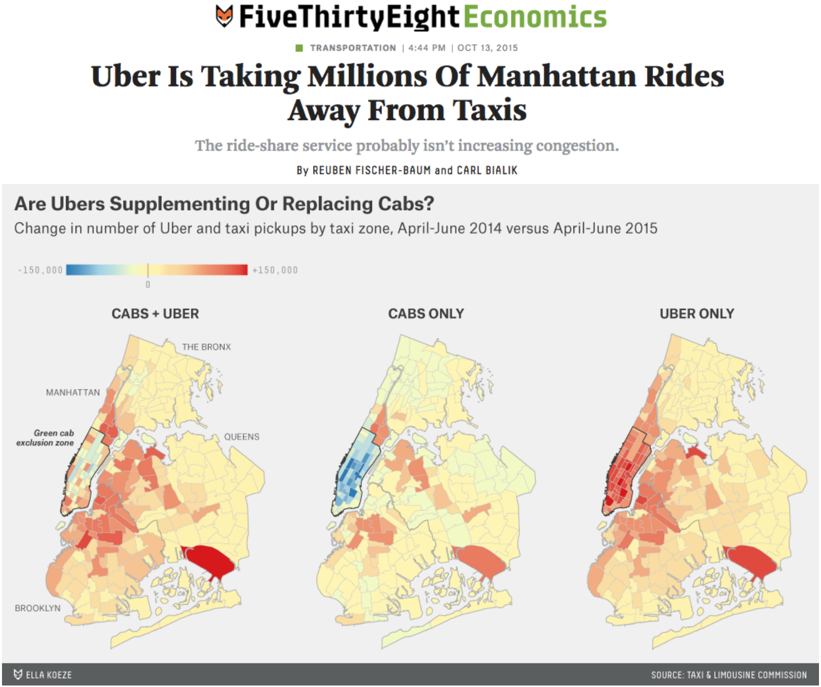
1. a piece of information.
   - an assumption or premise from which inferences may be drawn.
2. a fixed starting point of a scale or operation.

Origin

mid 18th century: from Latin, literally ‘something given,’ neuter past participle of *dare* ‘give.’



Use over time for: data



FiveThirtyEight Economics

TRANSPORTATION | 4:44 PM | OCT 13, 2015

Uber Is Taking Millions Of Manhattan Rides Away From Taxis

The ride-share service probably isn't increasing congestion.

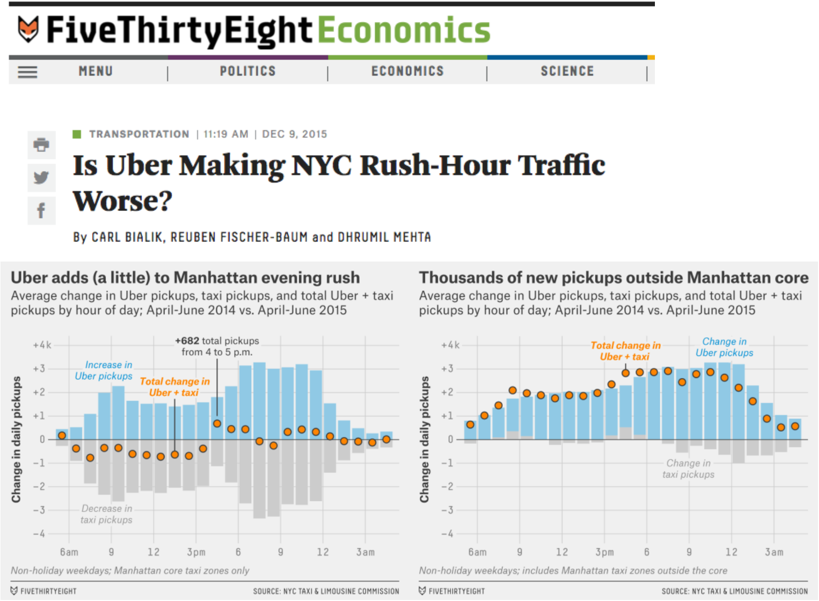
By REUBEN FISCHER-BAUM and CARL BIALIK

Are Ubers Supplementing Or Replacing Cabs?

Change in number of Uber and taxi pickups by taxi zone, April-June 2014 versus April-June 2015

ELLA KOEZE

SOURCE: TAXI & LIMOUSINE COMMISSION



TRANSPORTATION | 11:19 AM | DEC 9, 2015

Is Uber Making NYC Rush-Hour Traffic Worse?

By CARL BIALIK, REUBEN FISCHER-BAUM and DHRUMIL MEHTA

Uber adds (a little) to Manhattan evening rush

Average change in Uber pickups, taxi pickups, and total Uber + taxi pickups by hour of day; April-June 2014 vs. April-June 2015

Non-holiday weekdays; Manhattan core taxi zones only

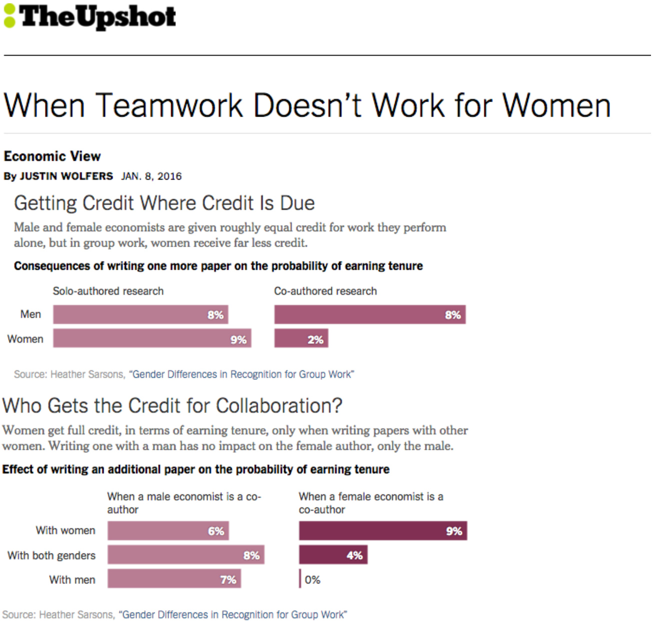
FIVETHIRTYEIGHT SOURCE: NYC TAXI & LIMOUSINE COMMISSION

Thousands of new pickups outside Manhattan core

Average change in Uber pickups, taxi pickups, and total Uber + taxi pickups by hour of day; April-June 2014 vs. April-June 2015

Non-holiday weekdays; includes Manhattan taxi zones outside the core

FIVETHIRTYEIGHT SOURCE: NYC TAXI & LIMOUSINE COMMISSION



TheUpshot

When Teamwork Doesn’t Work for Women

Economic View

By JUSTIN WOLFERS JAN. 8, 2016

Getting Credit Where Credit Is Due

Male and female economists are given roughly equal credit for work they perform alone, but in group work, women receive far less credit.

Consequences of writing one more paper on the probability of earning tenure

Source: Heather Sarsons, “Gender Differences in Recognition for Group Work”

Who Gets the Credit for Collaboration?

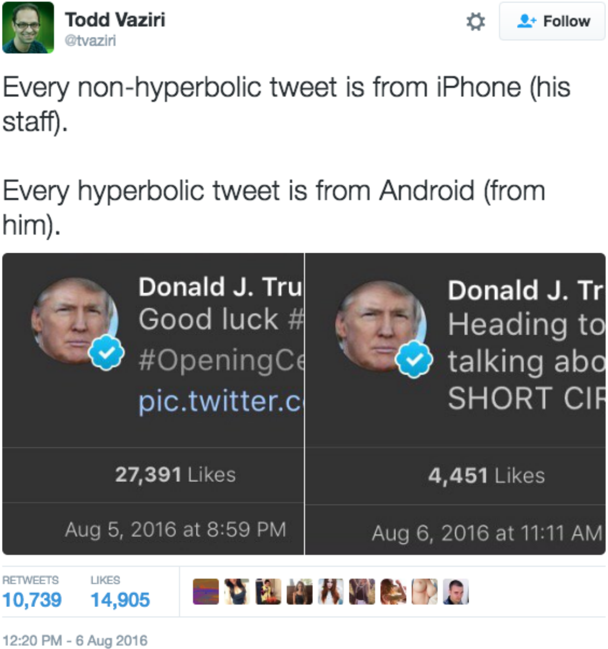
Women get full credit, in terms of earning tenure, only when writing papers with other women. Writing one with a man has no impact on the female author, only the male.

Effect of writing an additional paper on the probability of earning tenure

When a male economist is a co-author
When a female economist is a co-author
With women
6%
9%
With both genders
8%
4%
With men
7%
0%

Source: Heather Sarsons, “Gender Differences in Recognition for Group Work”

Trump's tweets: from varianceexplained.org



Todd Vaziri
@tvaziri

Every non-hyperbolic tweet is from iPhone (his staff).

Every hyperbolic tweet is from Android (from him).

Donald J. Tru
Good luck #
#OpeningCe
pic.twitter.c

27,391 Likes

Aug 5, 2016 at 8:59 PM

Donald J. Tr
Heading to
talking abo
SHORT CIR

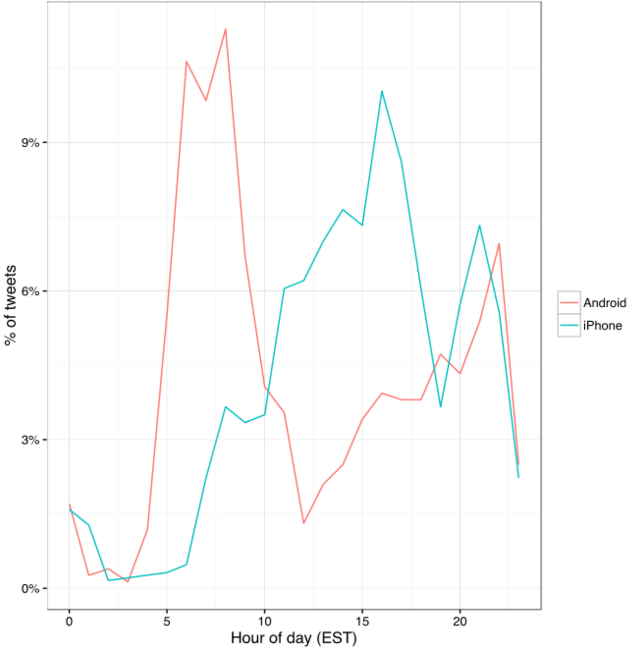
4,451 Likes

Aug 6, 2016 at 11:11 AM

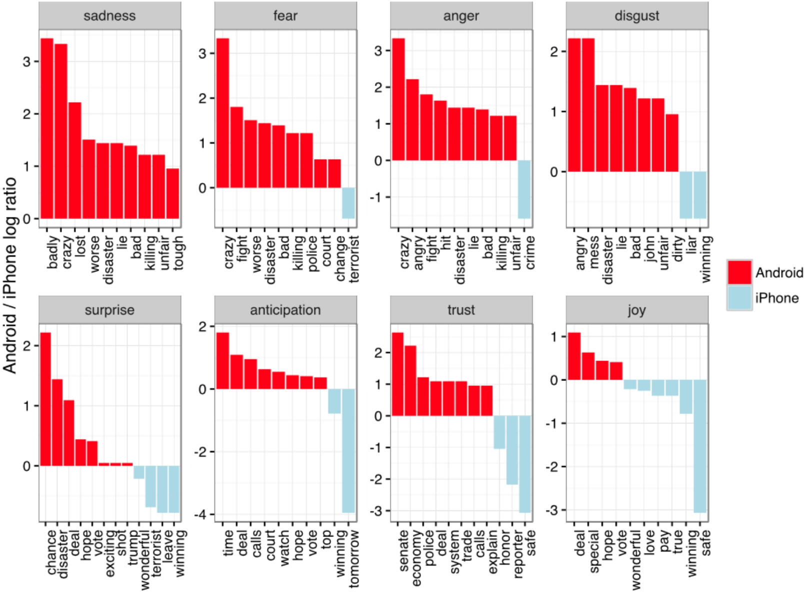
RETWEETS 10,739 LIKES 14,905

12:20 PM - 6 Aug 2016

Trump’s tweets: from varianceexplained.org



Trump's tweets: from varianceexplained.org

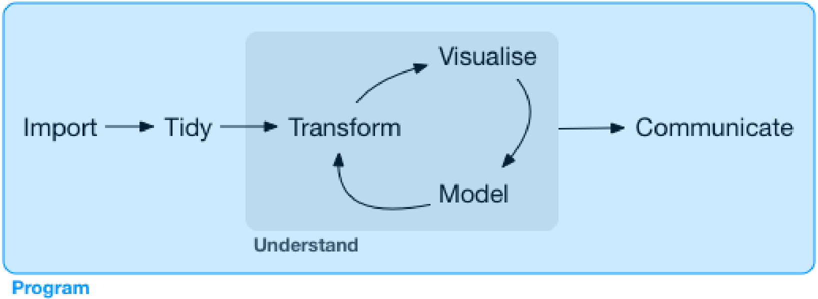


How to tell a story using data

Telling stories with data requires

- detective work
- creativity, both scientific and artistic
- experimentation with different storylines
- good data, (good data does not nescessarily equal “big data”)

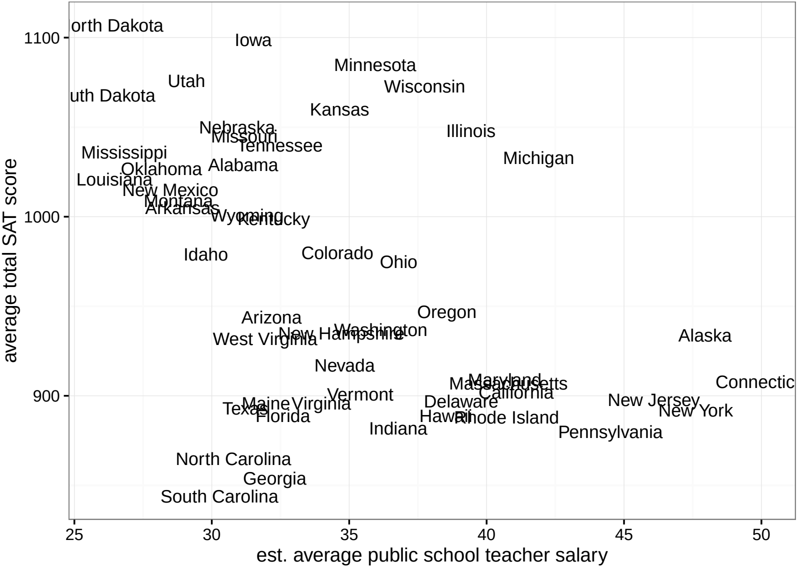
The tidy-verse: a process for data analysis



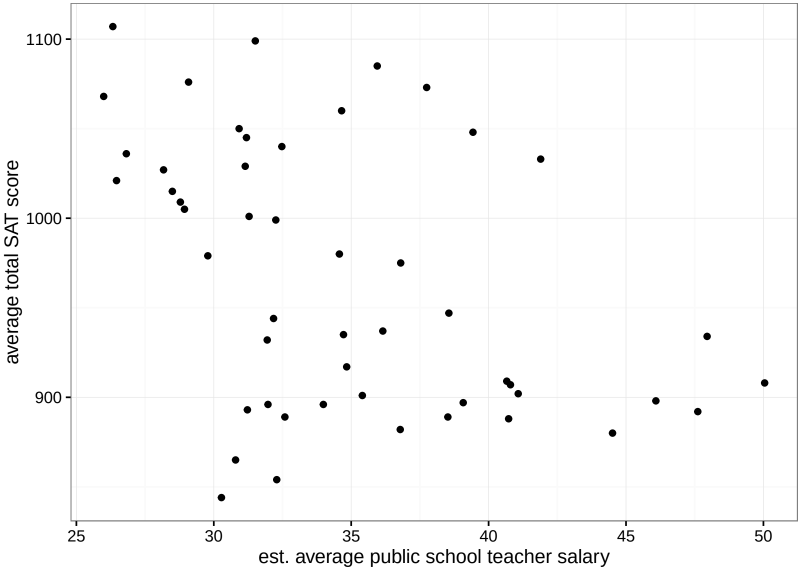
A common modeling tool: regression

- The goal is to learn about the relationship between two variables: a “covariate” (or “predictor”) of interest and an “outcome” of interest.
  - Some models focus on prediction.
  - Other models focus on description.
- Regression is an exercise in inferential statistics: we are drawing evidence and conclusions from data about “complex aspects of reality”, i.e. “noisy” systems.

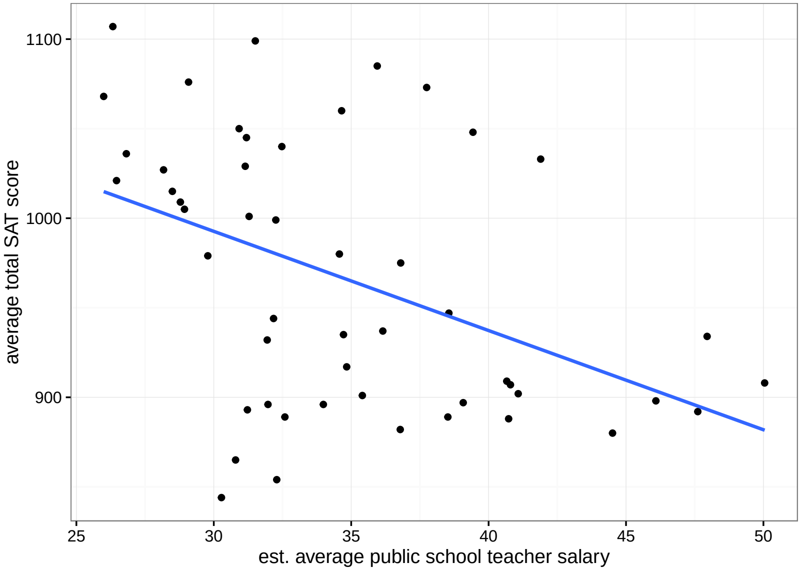
State-level SAT score data (1994-95)



State-level SAT score data (1994-95)



State-level SAT score data (1994-95)



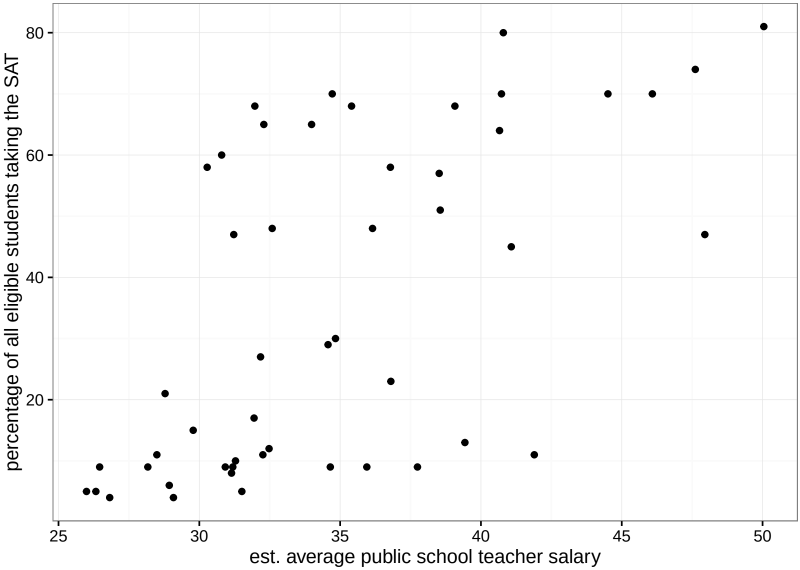
The SAT example

What is the outcome variable?

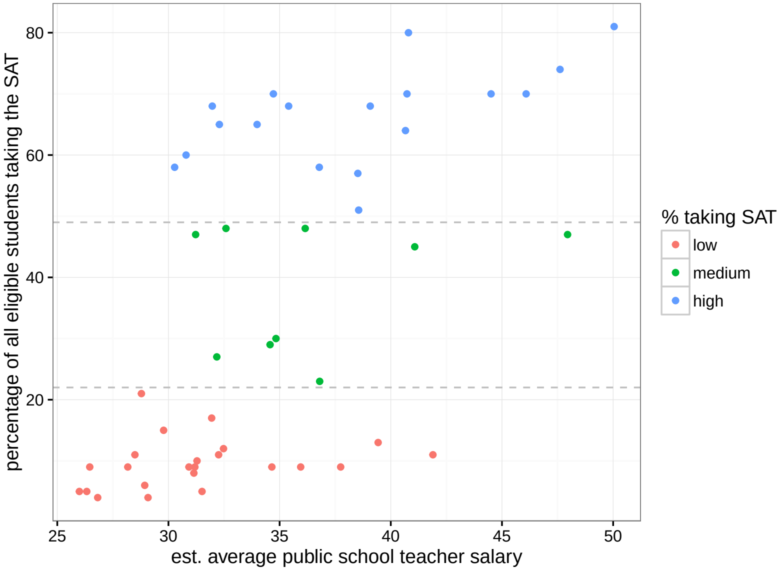
What is the covariate or predictor variable?

What other data might be part of this story?

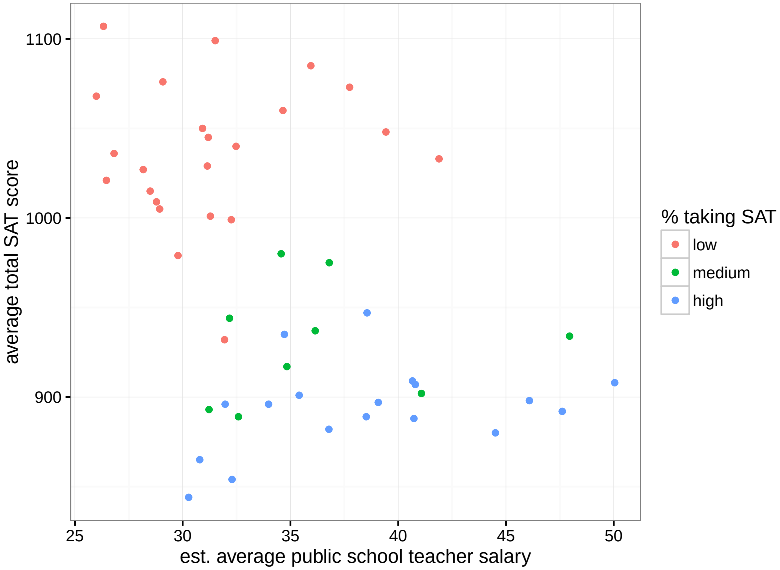
State-level SAT score data (1994-95)



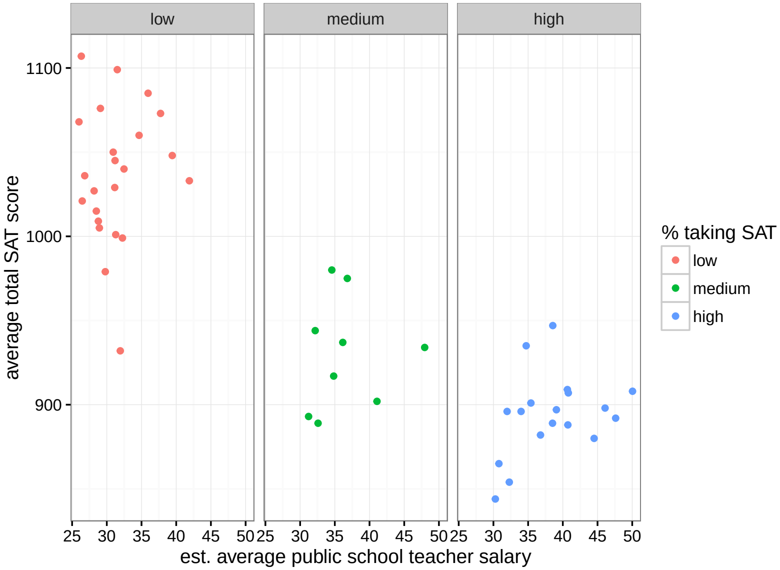
State-level SAT score data (1994-95)



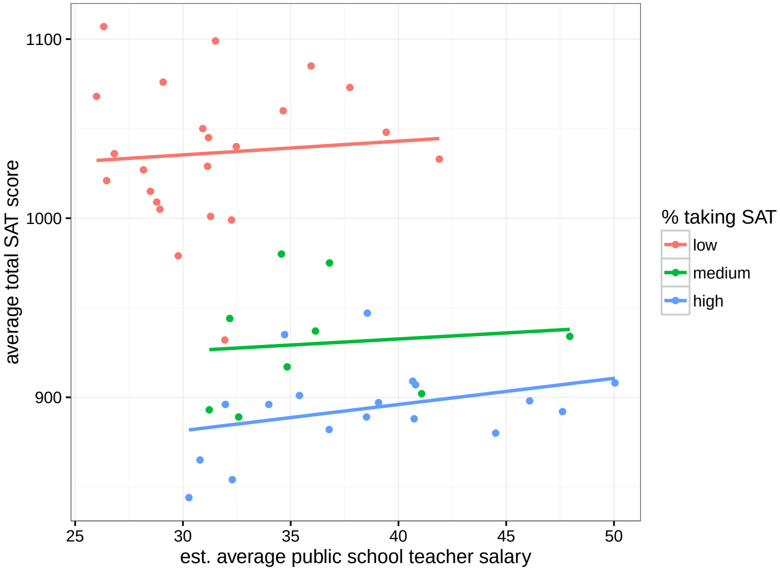
State-level SAT score data (1994-95)



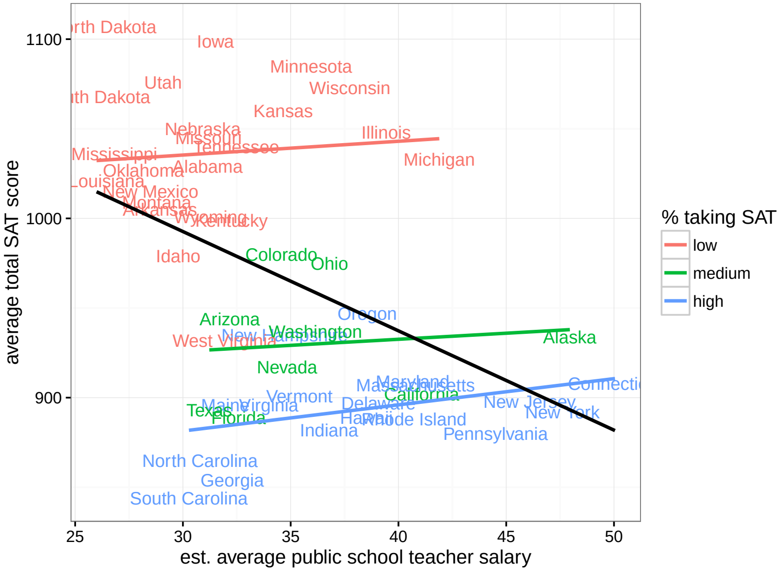
State-level SAT score data (1994-95)



State-level SAT score data (1994-95)



State-level SAT score data (1994-95)



State-level SAT score data (1994-95)

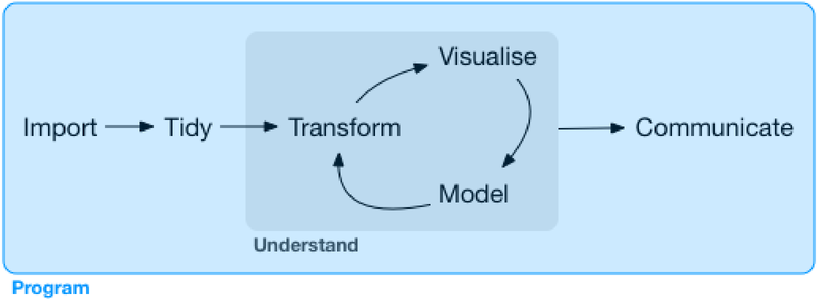
What can we conclude from all of this? (BTW, this is an example of "Simpson's Paradox".)

Regression modeling

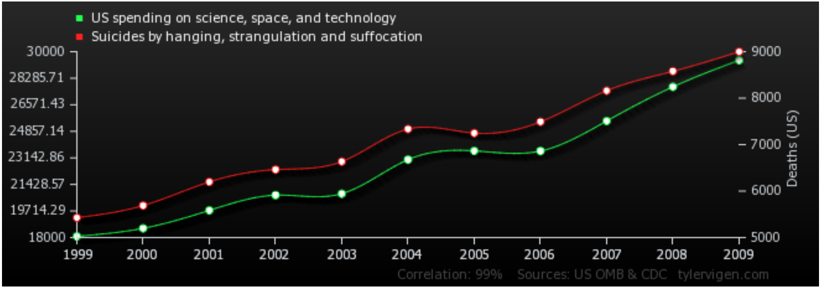
The process of using data to describe the relationship between outcomes and predictors is called modeling.

- 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”.

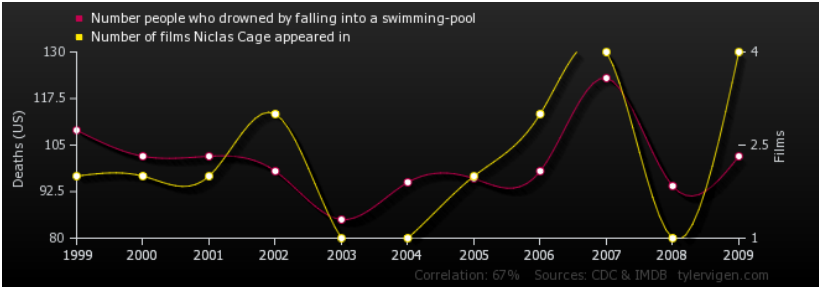
Things to come



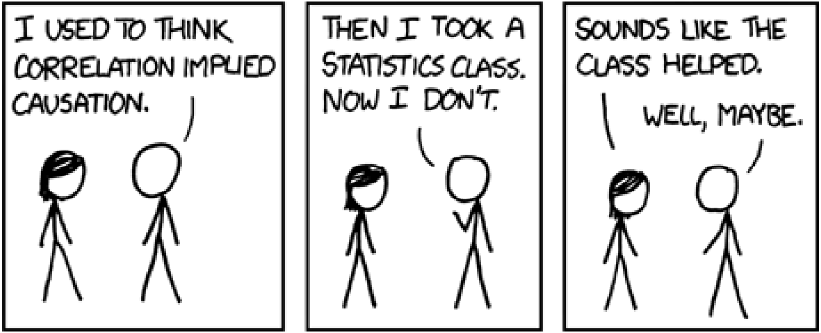
Beware of correlation!



Beware of correlation!



Hat tip to www.tylervigen.com


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

https://xkcd.com/552/