

# Introduction to Telling Stories with Data

Author: Nicholas G Reich

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

*Made available under the Creative Commons Attribution-ShareAlike 3.0 Unported  
License: [http://creativecommons.org/licenses/by-sa/3.0/deed.en\\_US](http://creativecommons.org/licenses/by-sa/3.0/deed.en_US)*

# 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

/ˈdætəm, ˈdætəm/ ⓘ

*noun*

plural noun: *data*

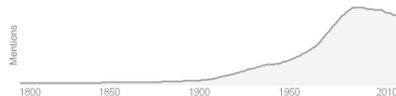
1. a piece of information.
  - an assumption or premise from which inferences may be drawn.
2. a fixed starting point of a scale or operation.

#### Origin



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

#### Use over time for: data



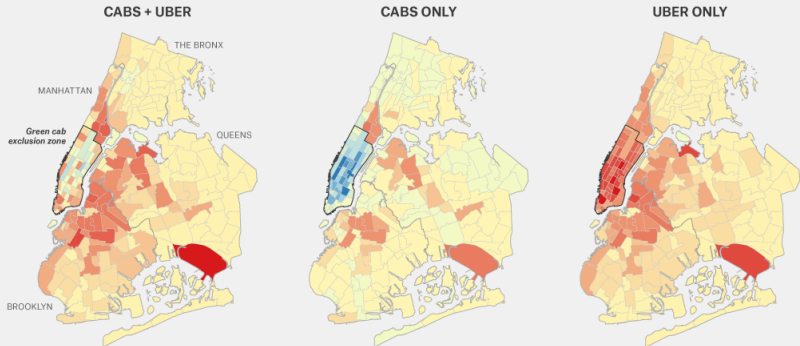
# Uber Is Taking Millions Of Manhattan Rides Away From Taxis

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





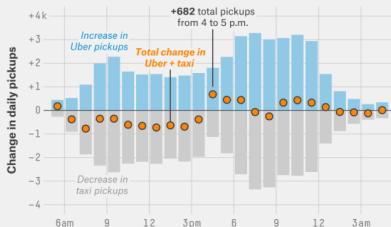
TRANSPORTATION | 11:19 AM | DEC 9, 2015

## 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



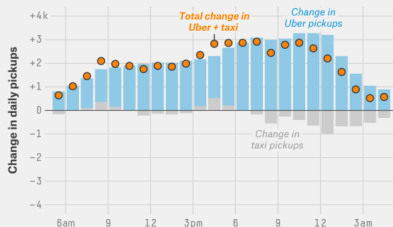
Non-holiday weekdays; Manhattan core taxi zones only

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

# How To Spot A Front-Runner On The 'Bachelor' Or 'Bachelorette'

What we learned from analyzing all 33 seasons.

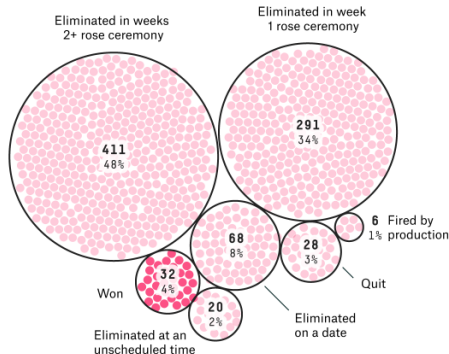
By [Ella Koeze](#) and [Walt Hickey](#)

Filed under [TV](#)

Get the data on [BitHub](#)

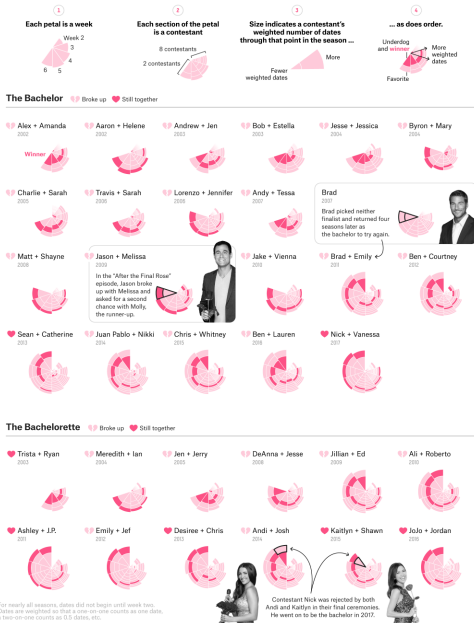
Published May 22, 2017

## The fate of every 'Bachelor' and 'Bachelorette' contestant



## A rose for every season

The path of every winner on every season of the "Bachelor" and "Bachelorette"

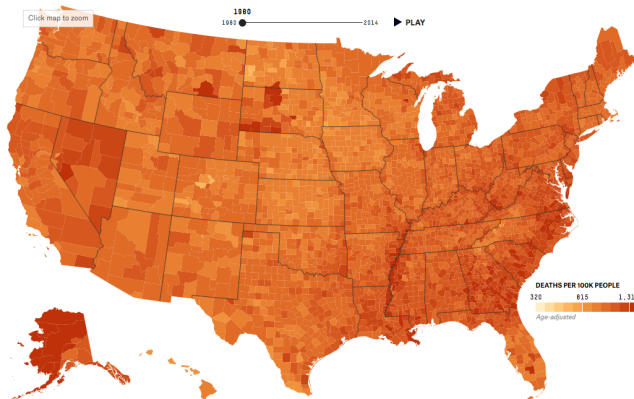


# 35 Years Of American Death

Mortality rates for leading causes of death in every U.S. county from 1980 to 2014.

By [Ella Koeze](#)

< All causes of death >



Mortality rates are age adjusted to account for higher mortality in older populations and geographic variations in the ages of county populations.

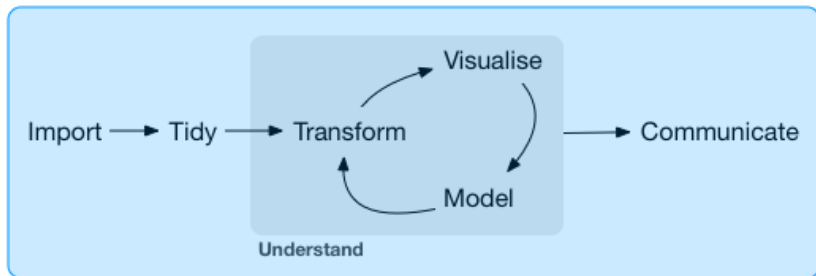
# How to tell a story using data

Telling stories with data requires

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



# The tidy-verse: a process for data analysis

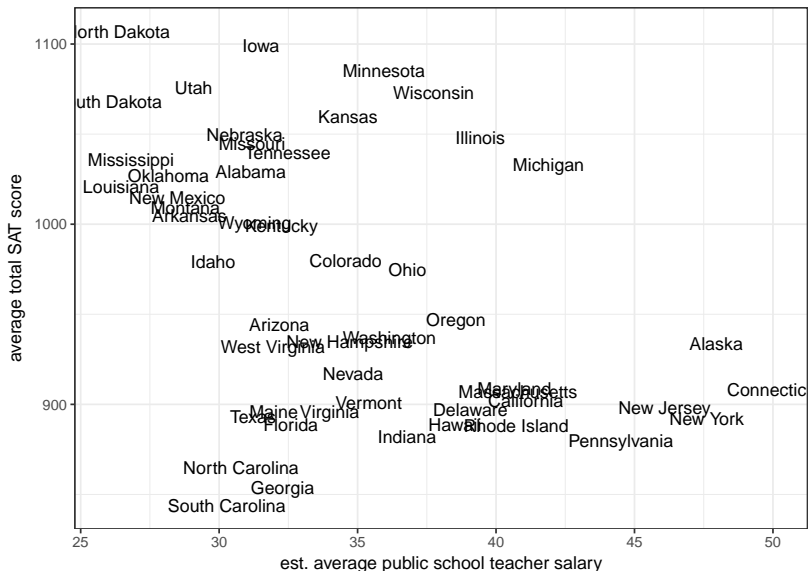


Program

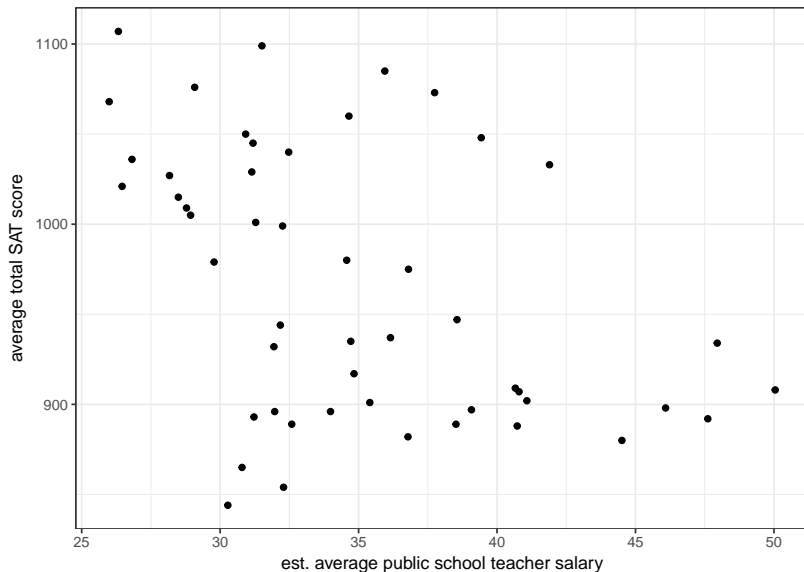
# A common modeling tool: regression

- 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.

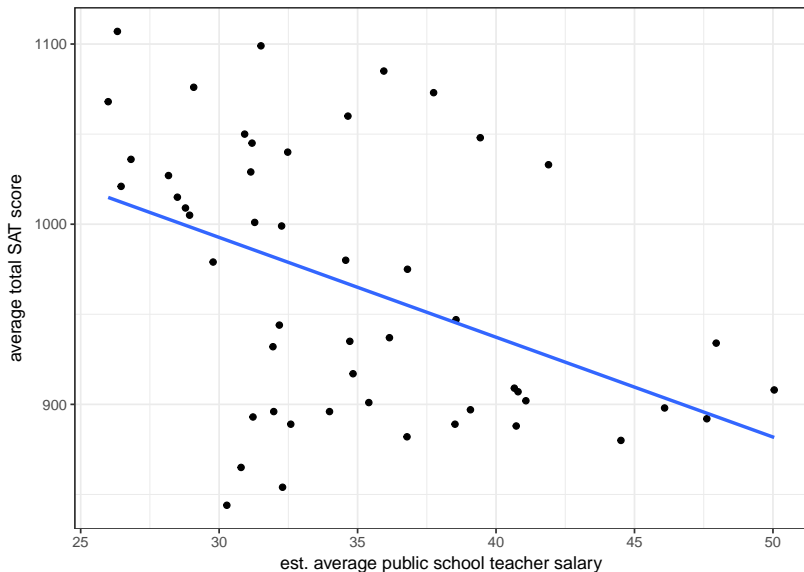
# State-level SAT score data (1994-95)



# State-level SAT score data (1994-95)



# State-level SAT score data (1994-95)



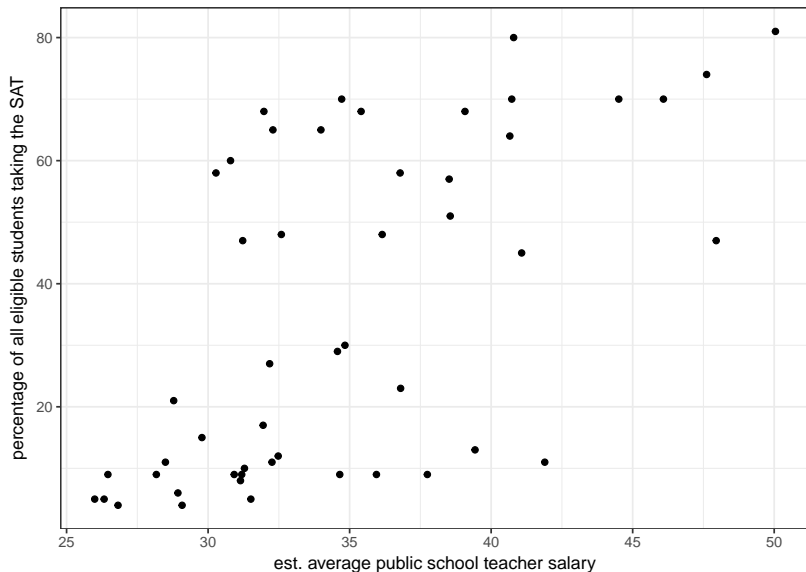
# The SAT example

What is the outcome variable?

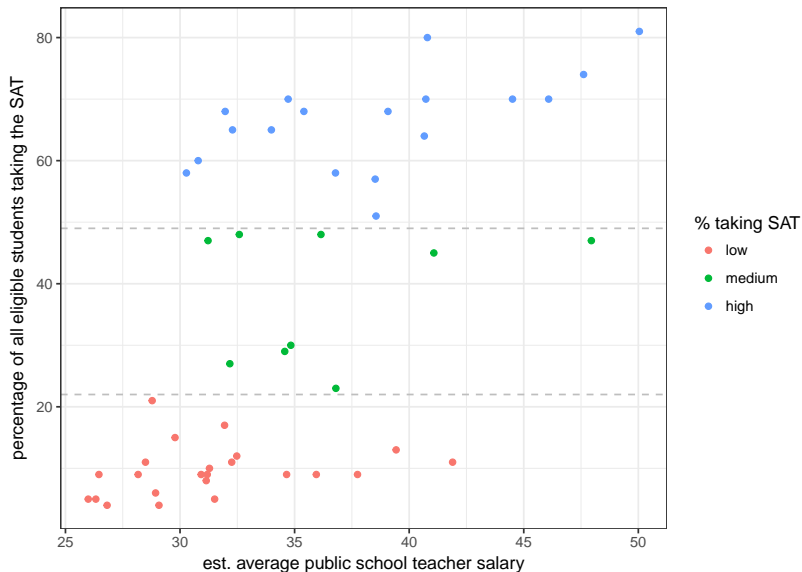
What is the covariate or predictor variable?

What other data might be part of this story?

# State-level SAT score data (1994-95)

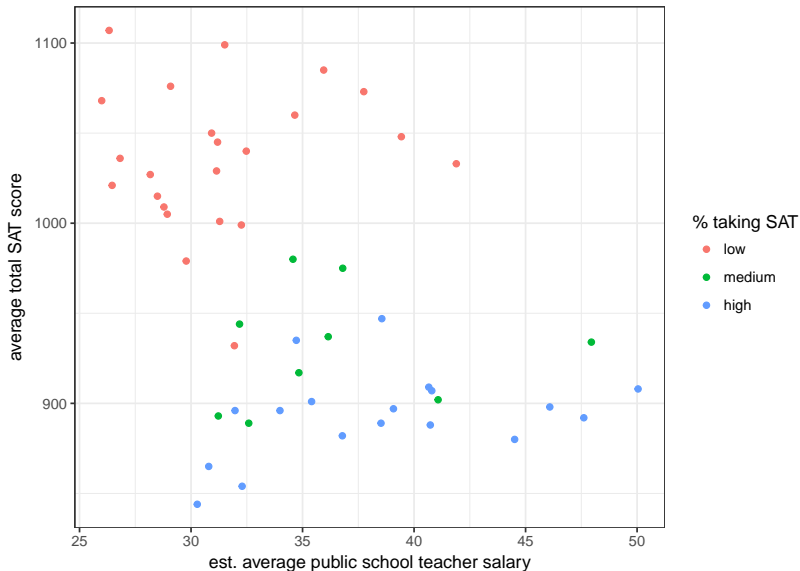


# State-level SAT score data (1994-95)

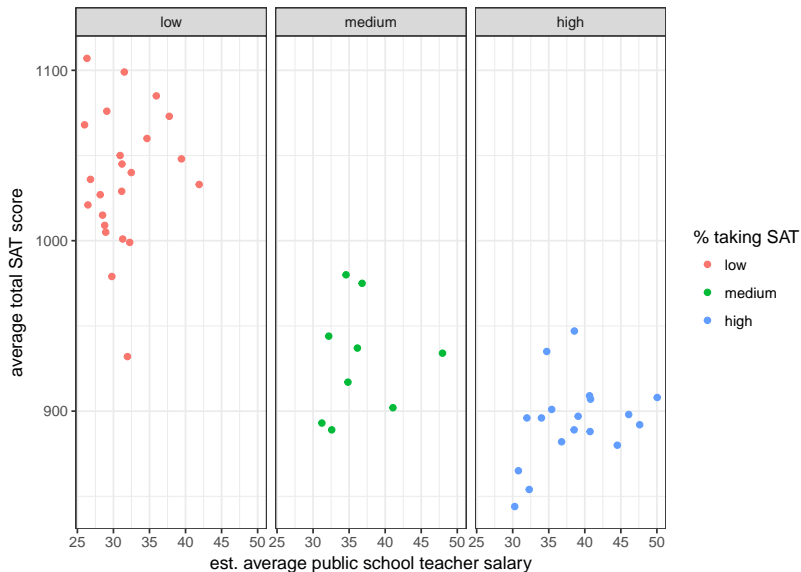




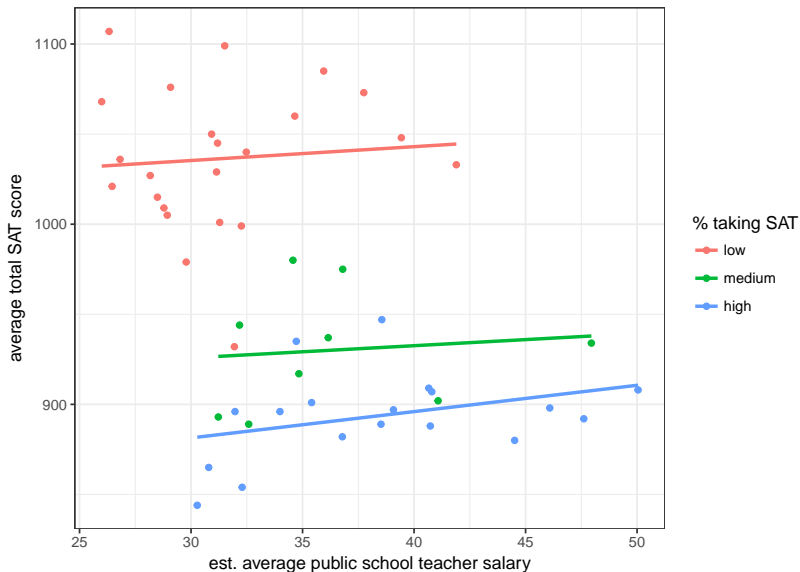
# State-level SAT score data (1994-95)



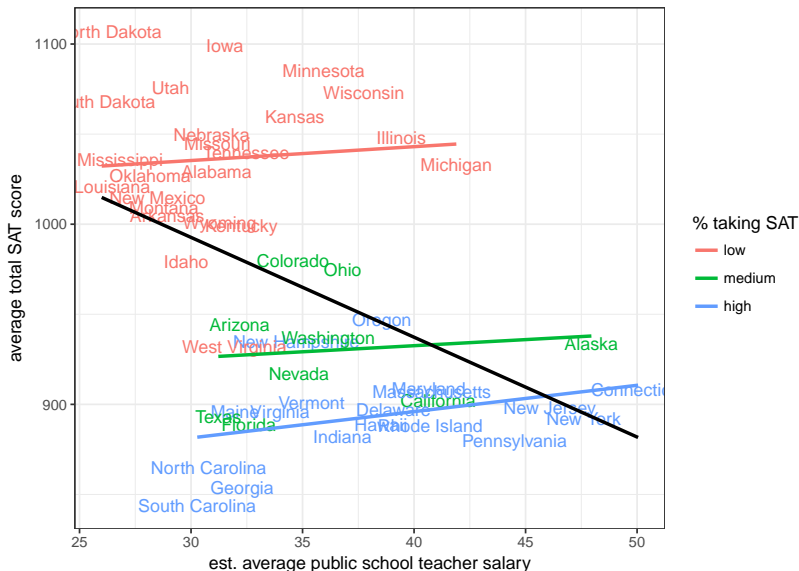
# State-level SAT score data (1994-95)



# State-level SAT score data (1994-95)



# State-level SAT score data (1994-95)



## State-level SAT score data (1994-95)

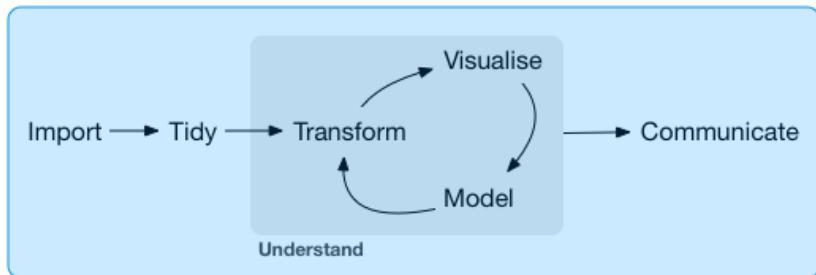
What can we conclude from all of this? (BTW, this is an example of "Simpson's Paradox".)

# Regression modeling

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

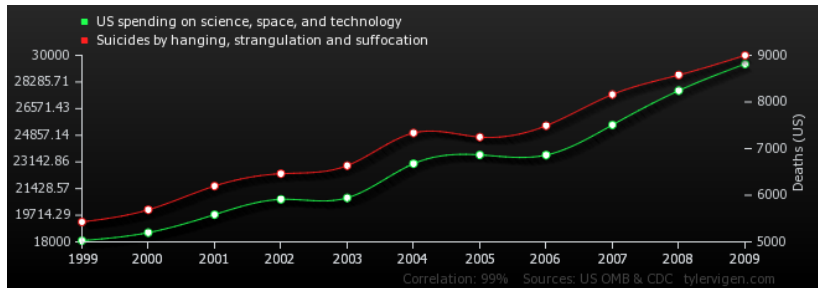
- Models are models, not reality.
- “All models are wrong, but some are useful.”
- Introduce structure to our model that balances realism with “goodness of fit” .

# Things to come



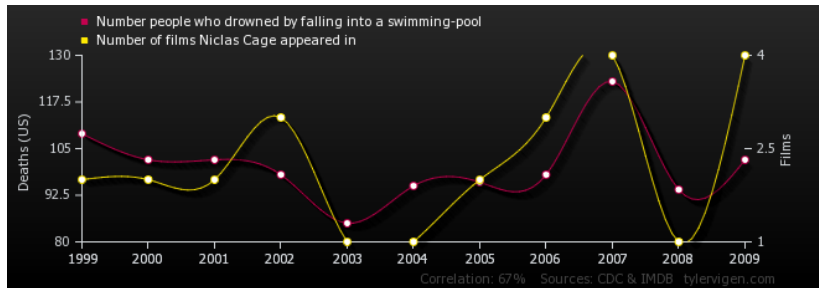
Program

# Beware of correlation!

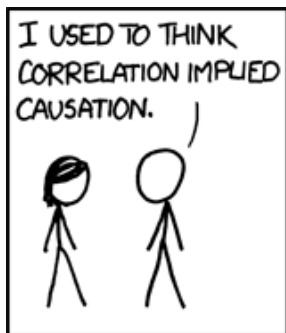




# Beware of correlation!



Hat tip to [www.tylervigen.com](http://www.tylervigen.com)



<https://xkcd.com/552/>