Modeling continuous relationships

Stats 60/Psych 10
Ismael Lemhadri

This time

- Modeling continuous relationships
- Correlation
  - Pearson's coefficient
  - Statistical significance
- Correlation and causation

What does “correlation” mean to you?

FiveThirtyEight

Politics Sports Science & Health Economics Culture

JAN. 23, 2017 AT 12:18 PM

Higher Rates Of Hate Crimes Are Tied To Income Inequality

By Maimuna Majumder

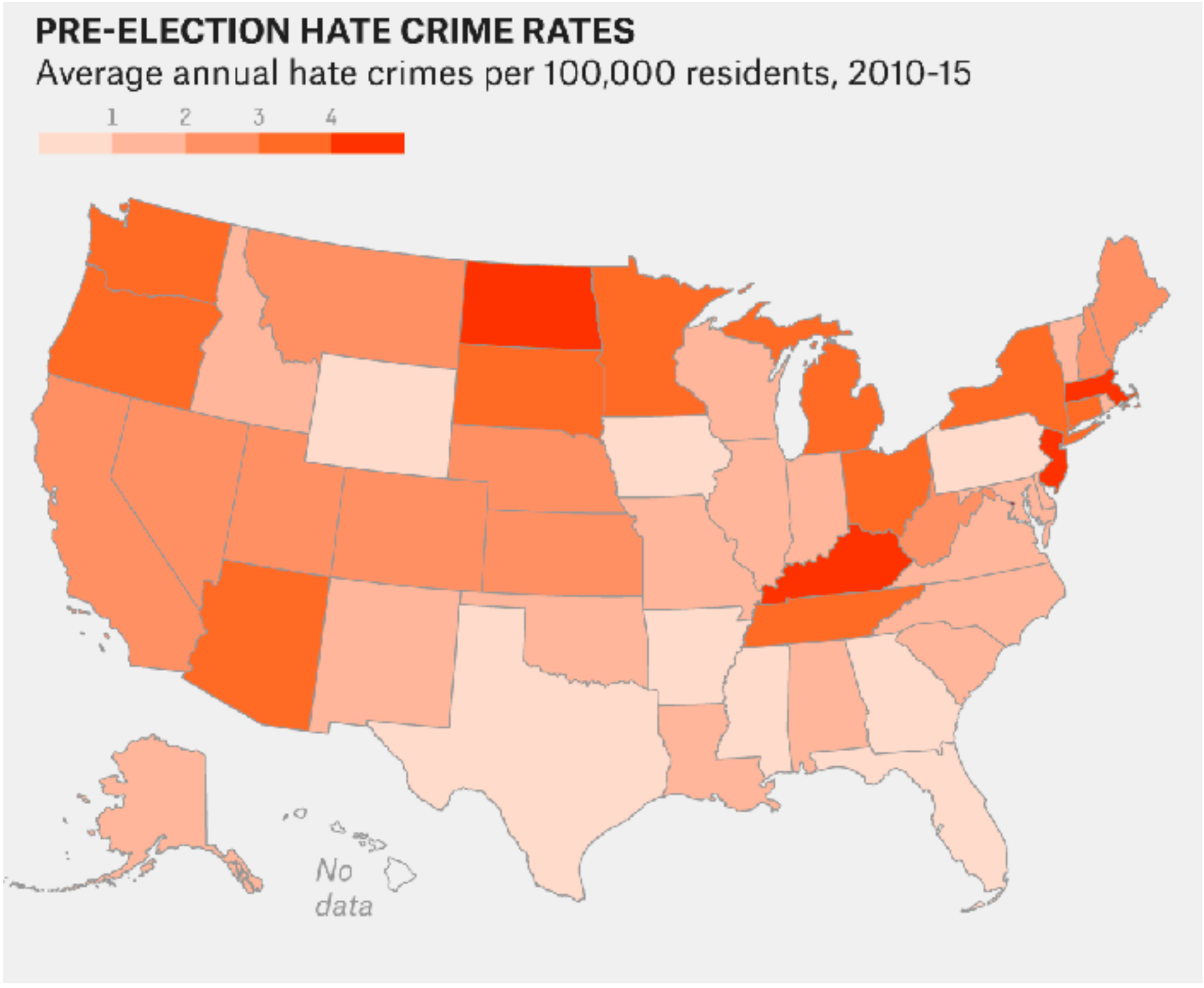
Filed under Hate Crimes

Get the data on GitHub

In the 10 days after the 2016 election, nearly 900 hate incidents were reported to the Southern Poverty Law Center, averaging out to 90 per day. By comparison, about 36,000 hate crimes were reported to the FBI from 2010 through 2015 — an average of 16 per day.

https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/

Hate crime rates differ across states



https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/

How can we define income inequality?

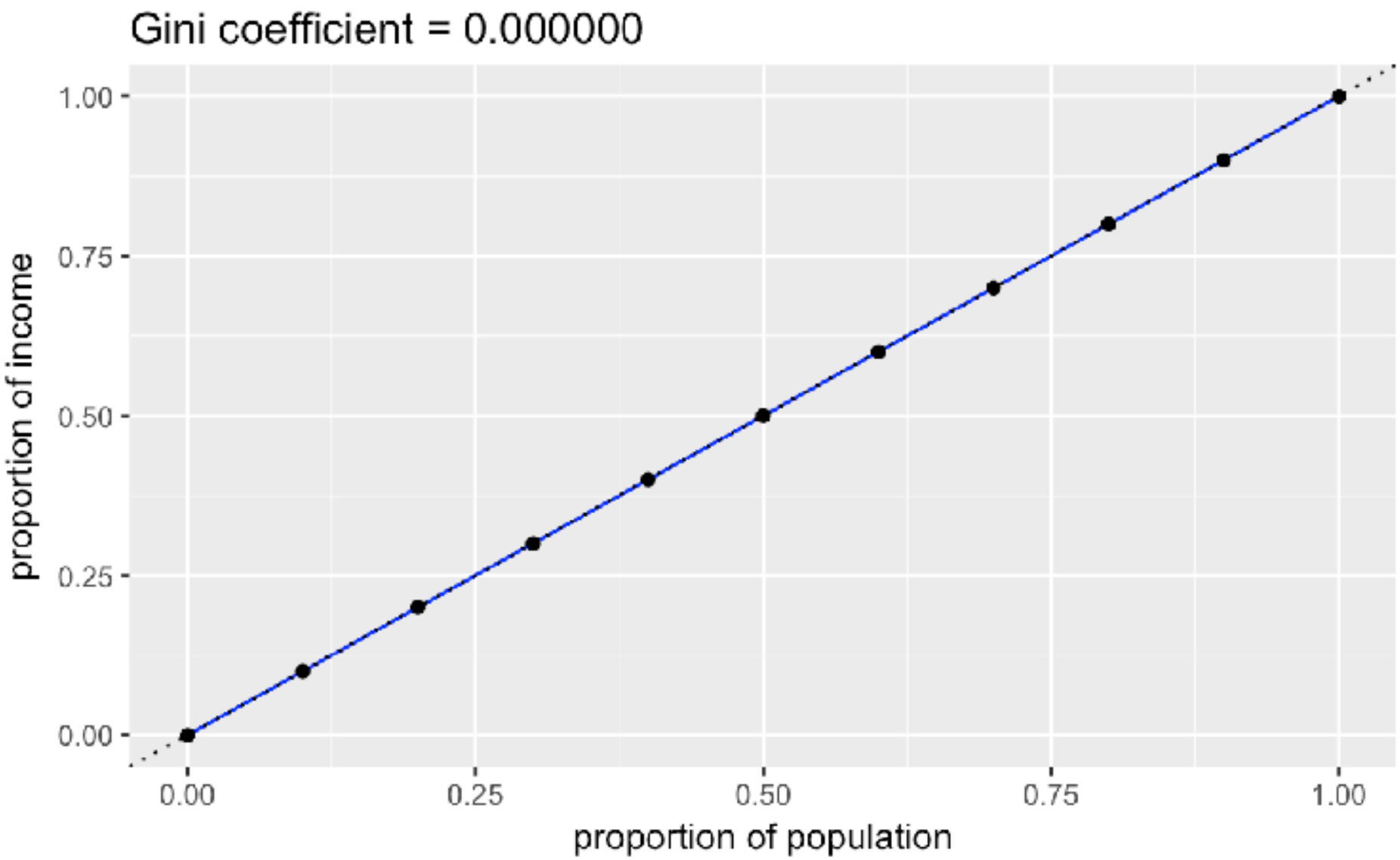
- Gini index
  - What is the mean relative absolute difference between incomes in the relevant population?
  - Usually defined in terms of a "Lorenz curve"

Corrado Gini

https://www.umass.edu/wsp/resources/tales/gini.html

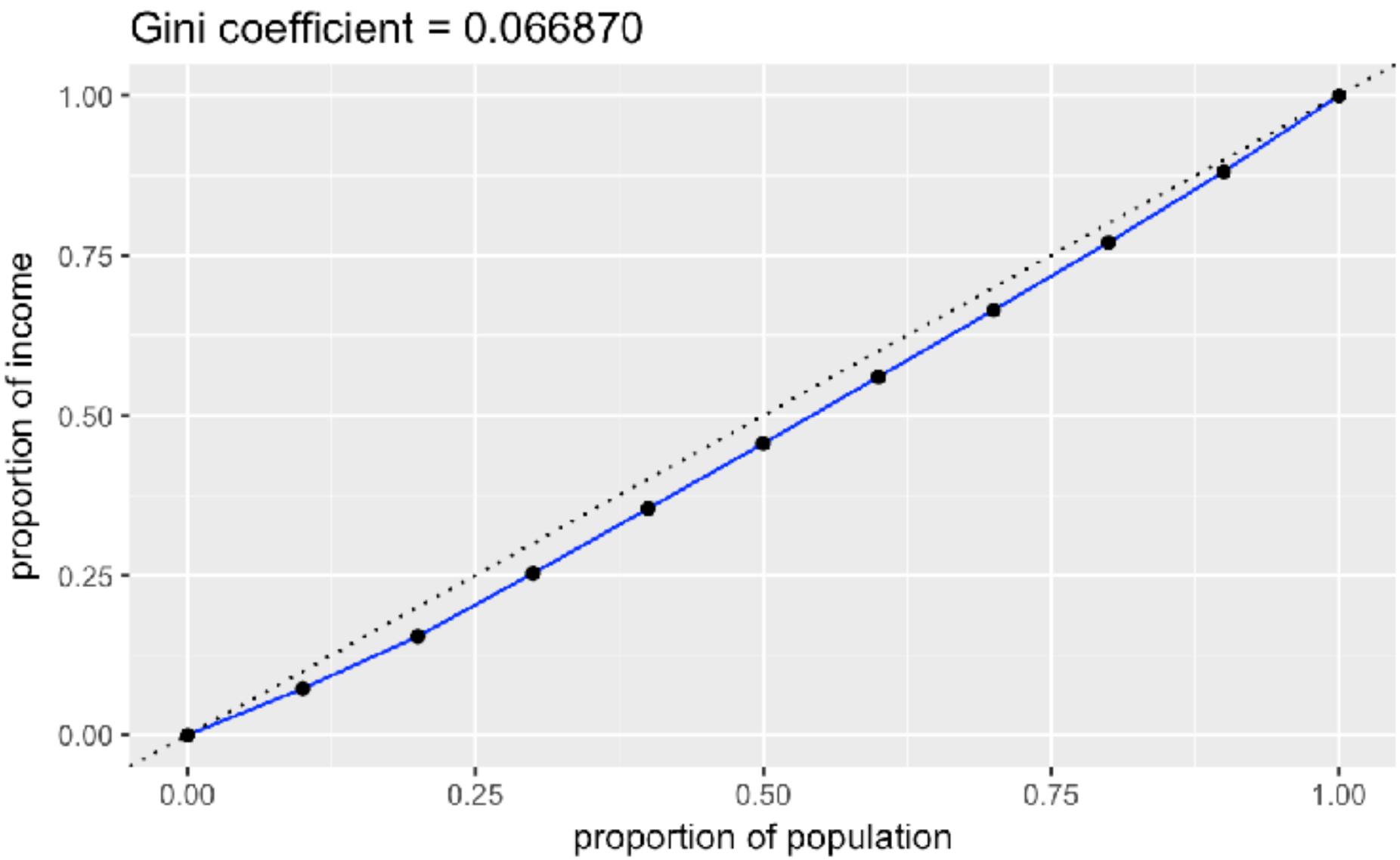
Example: perfect income equality

- 10 people, all incomes =$40,000



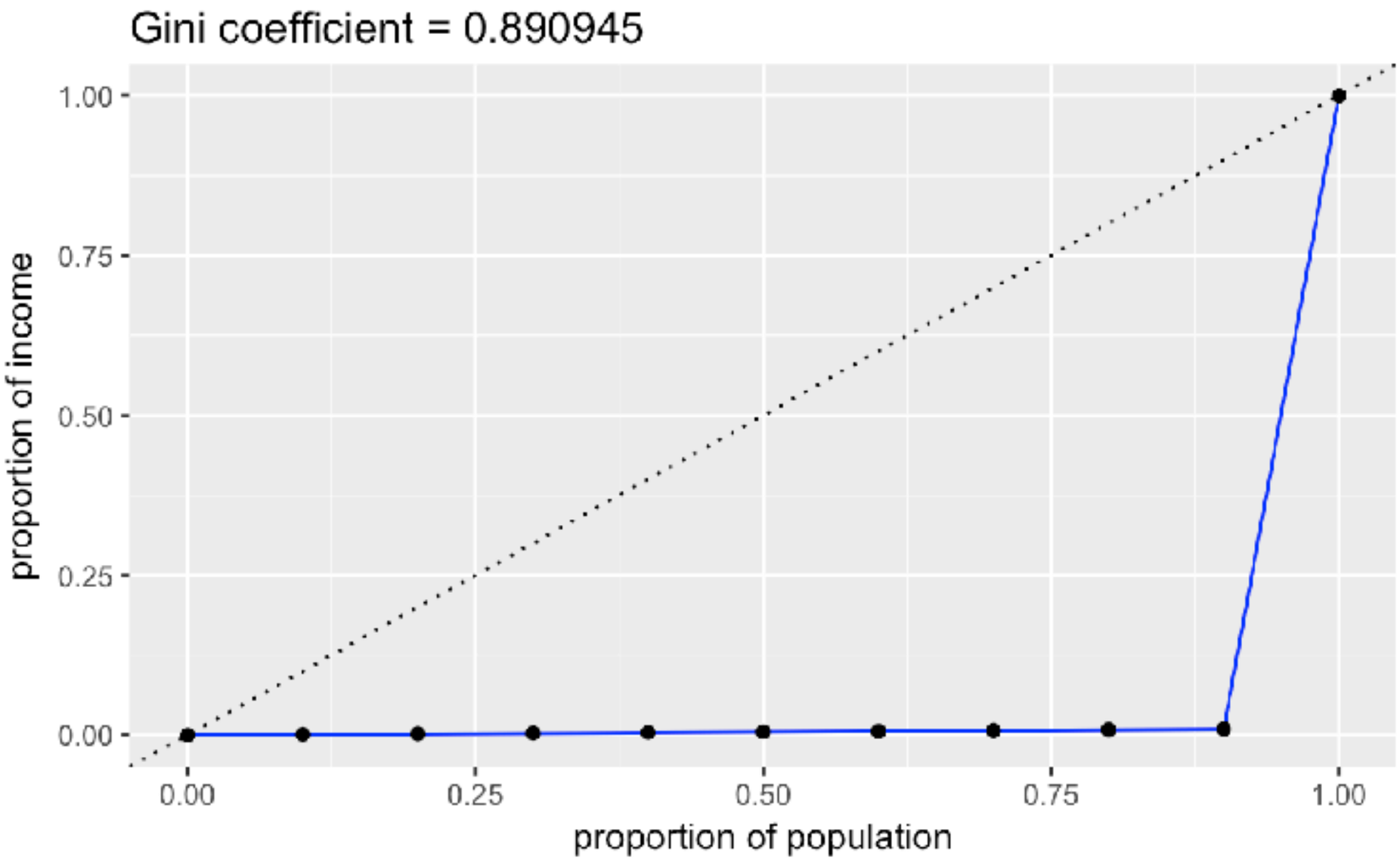
Example: mild inequality

- 10 people, incomes = rnorm(mean=40000,sd=10000)

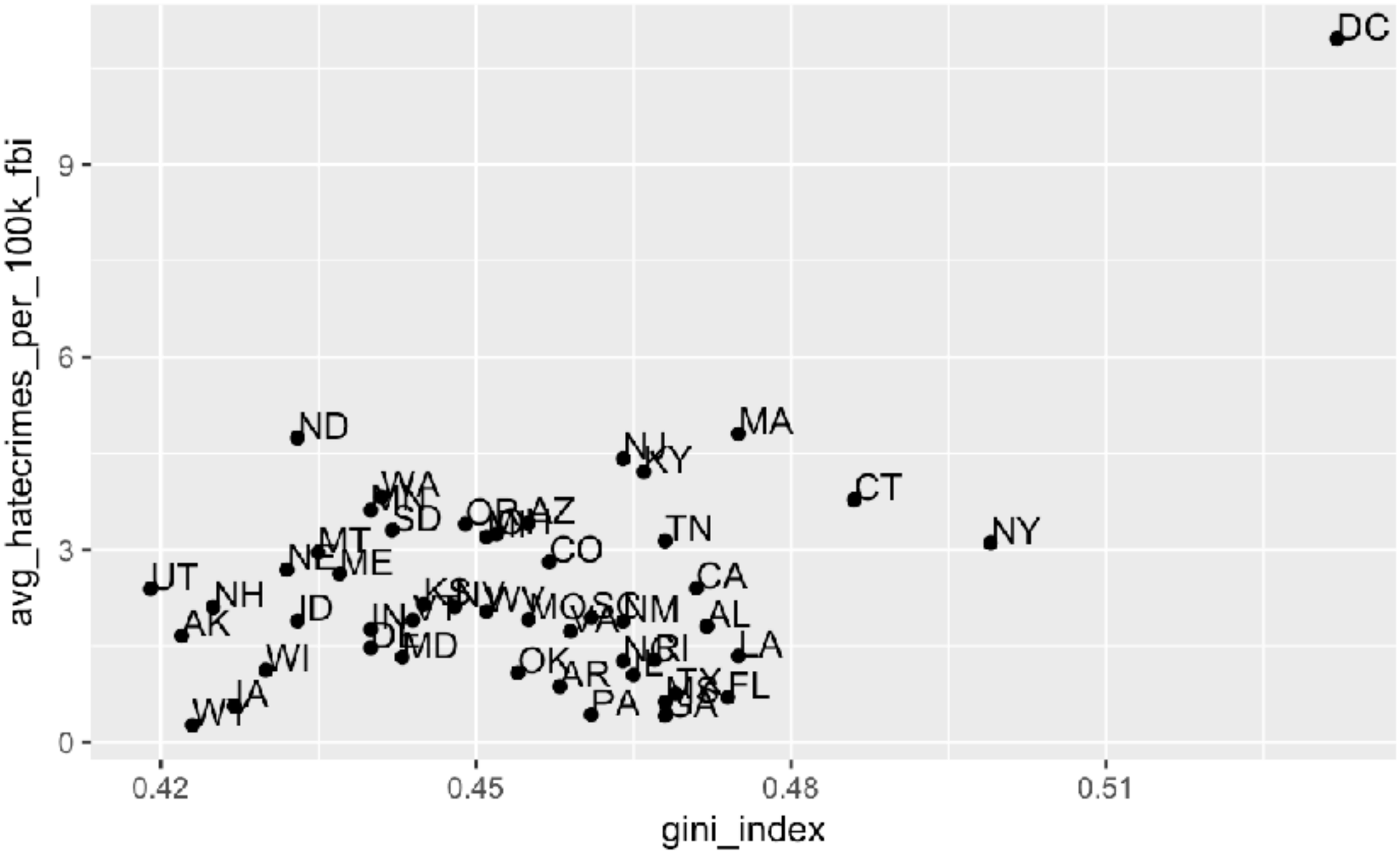


Example: severe inequality

- 10 people: 9 with $40,000, one with $40,000,000



How strong is the relationship between hate crimes and income inequality?



hate_crimes from fivethirtyeight R package

Quantifying continuous relationships

- Variance for a single variable

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}$$

- Covariance between two variables

"cross product"

$$covariance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$covariance = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

x	y	y_dev	x_dev	crossproduct
3	1	-7	-4.6	32.2
5	8	0	-2.6	0.0
8	8	0	0.4	0.0
10	10	2	2.4	4.8
12	13	5	4.4	22.0

sum = 59

covariance = 59/4 =14.85

Pearson's correlation coefficient

- The correlation coefficient (*r*) scales the covariance so that it has a standard scale

$$r = \frac{covariance}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

- This is exactly the same as the covariance between z-scored data (since the std deviation of z-scored data is 1)

x	y	y_dev	x_dev	crossproduct
3	1	-7	-4.6	32.2
5	8	0	-2.6	0.0
8	8	0	0.4	0.0
10	10	2	2.4	4.8
12	13	5	4.4	22.0

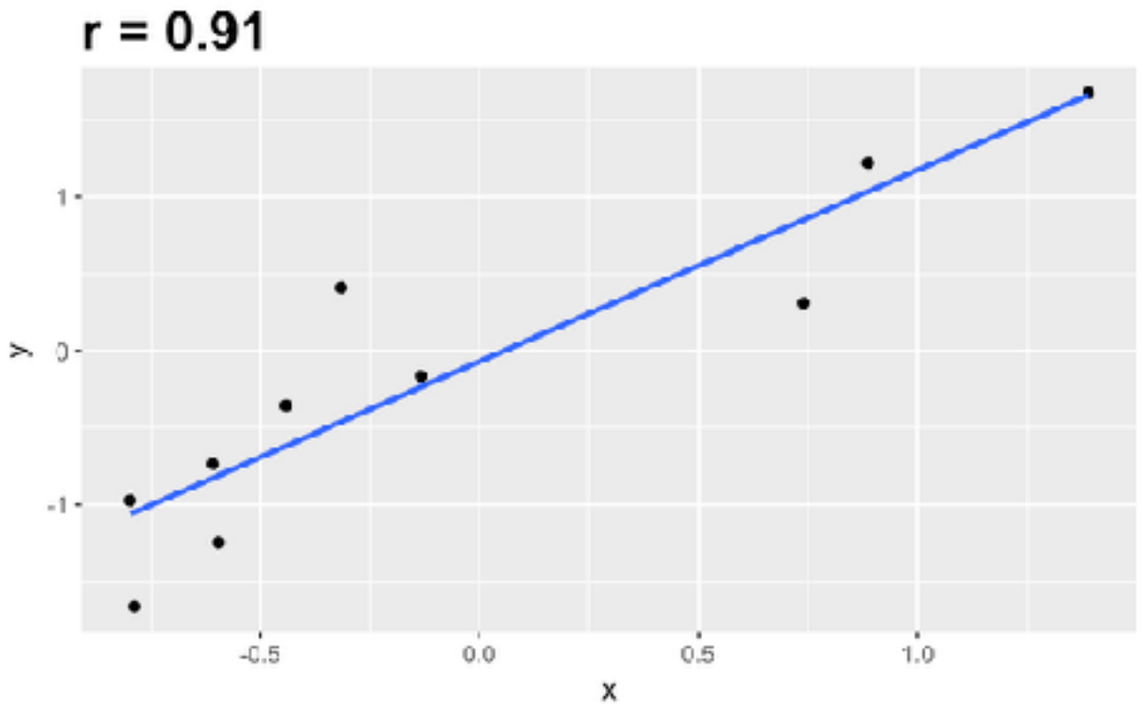
sum = 59

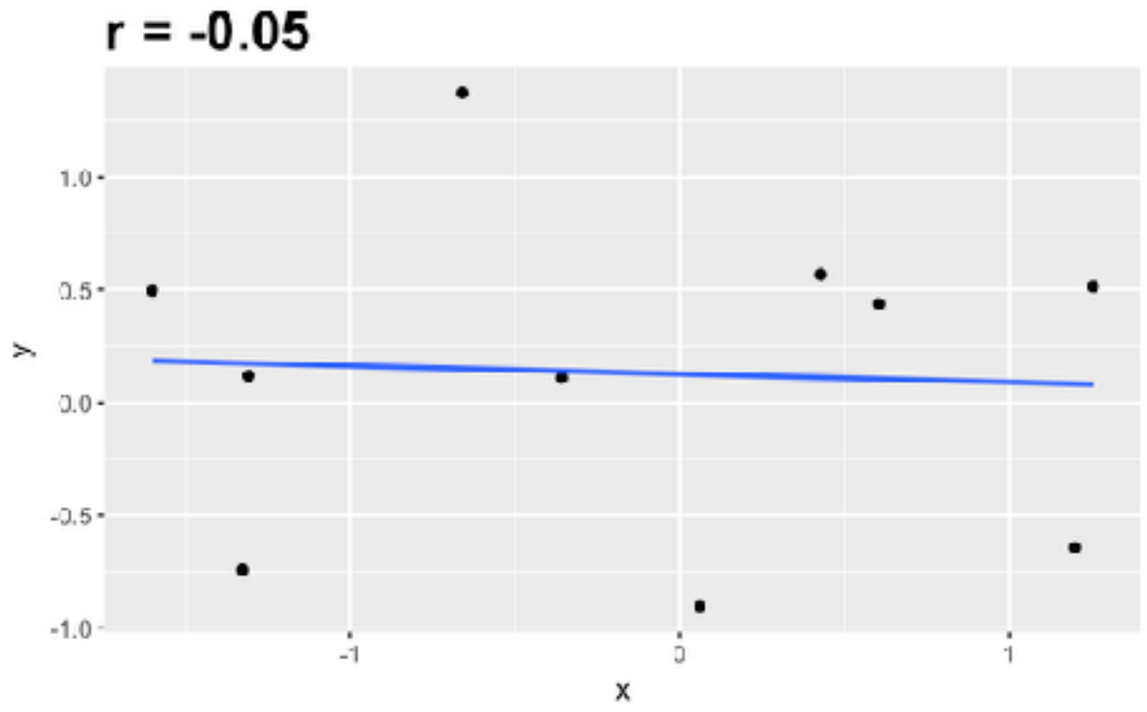
covariance = 59/4 =14.85

sd(x) = 3.65

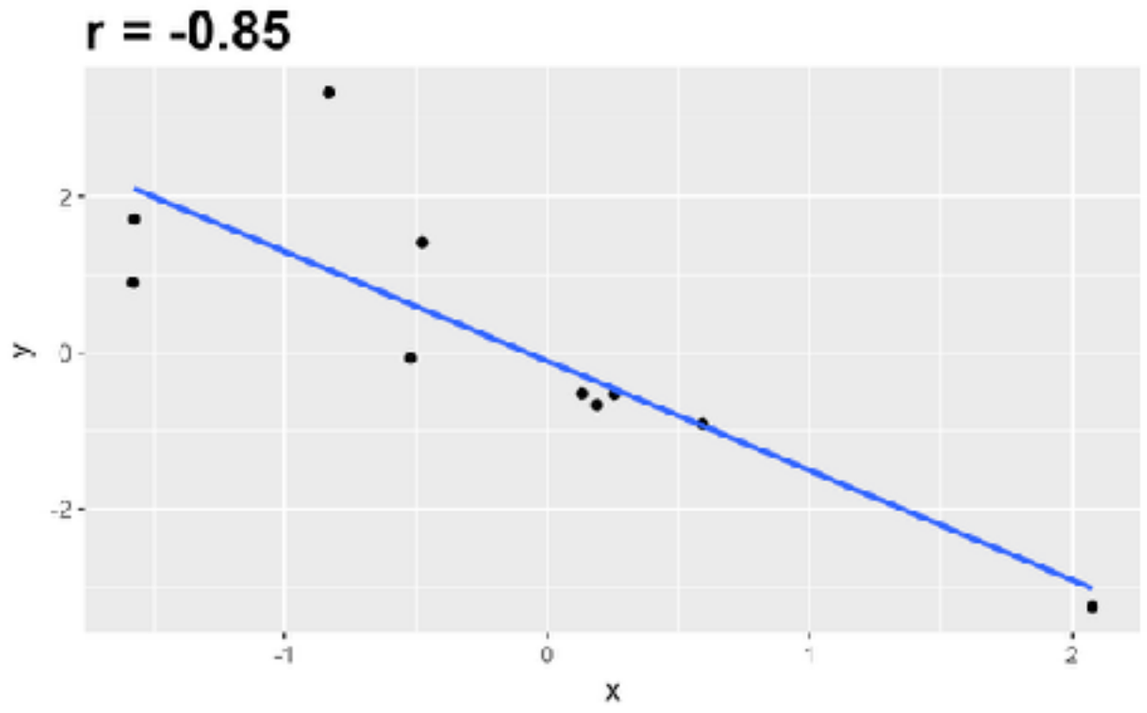
sd(y) = 4.42

r = 14.85/(3.65*4.42) = 0.92

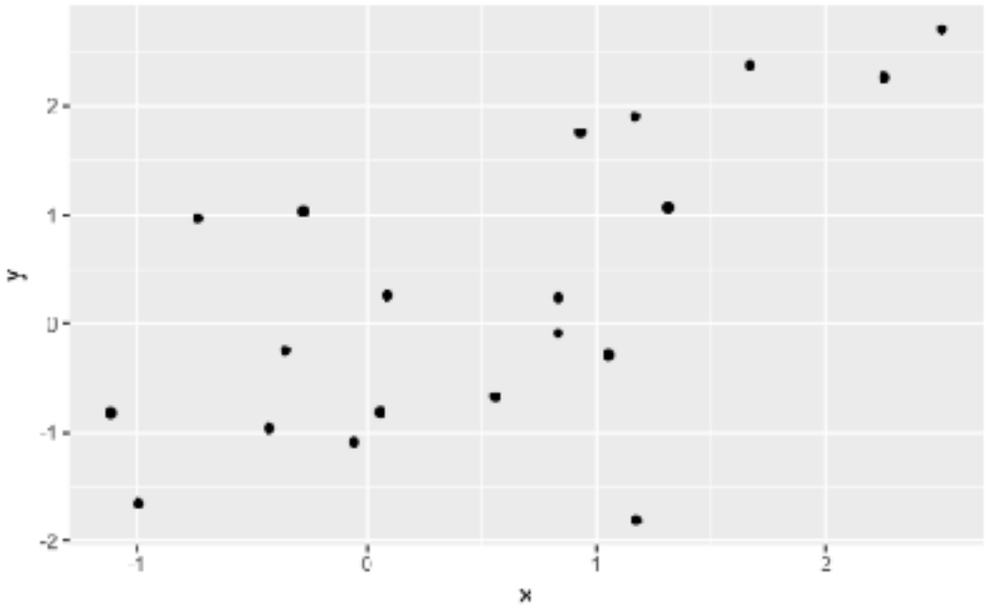




r=1: perfect positive relationship
r=0: no linear relationship
r=-1: perfect negative relationship



Guess the correlation coefficient for this dataset

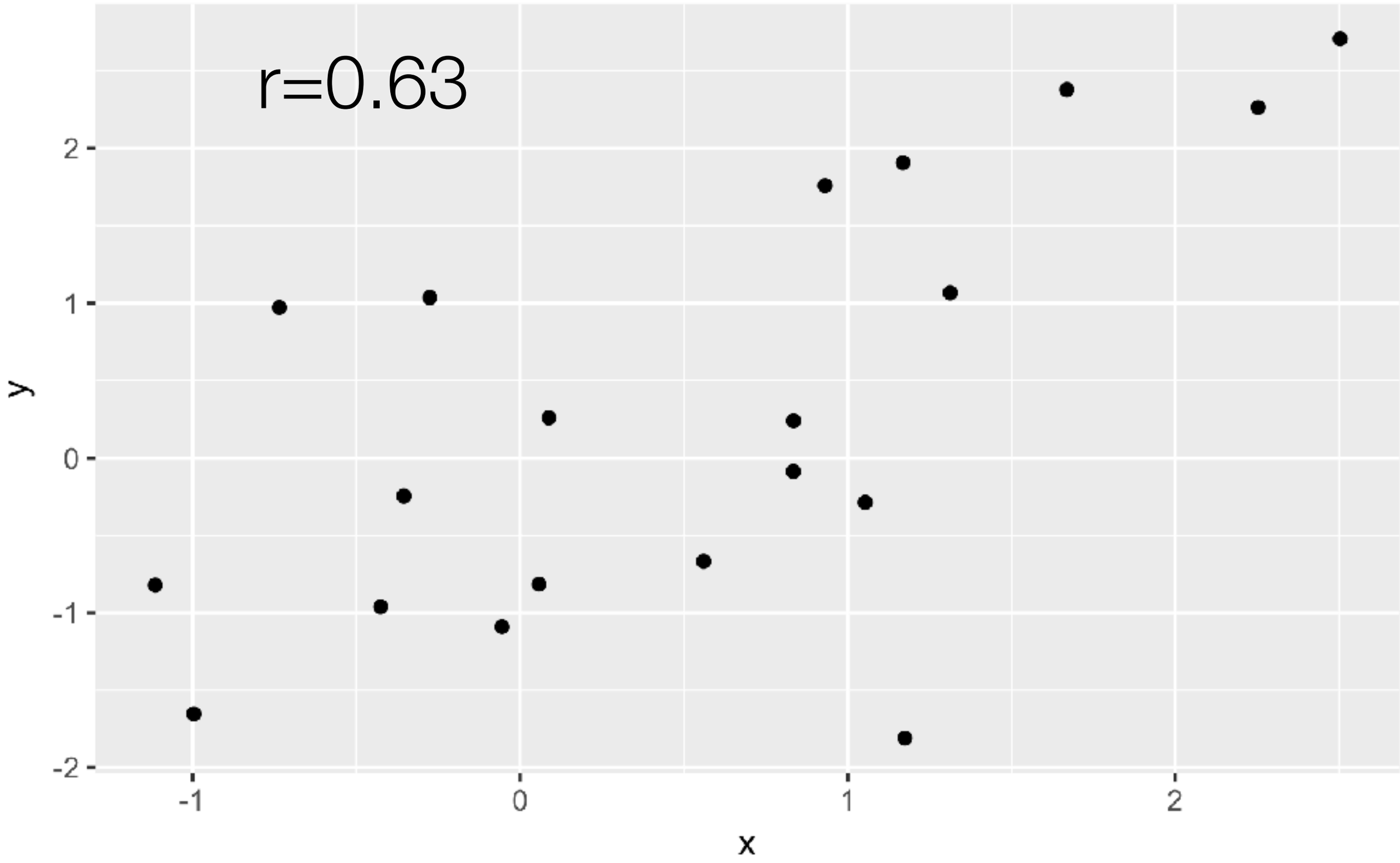


0.15

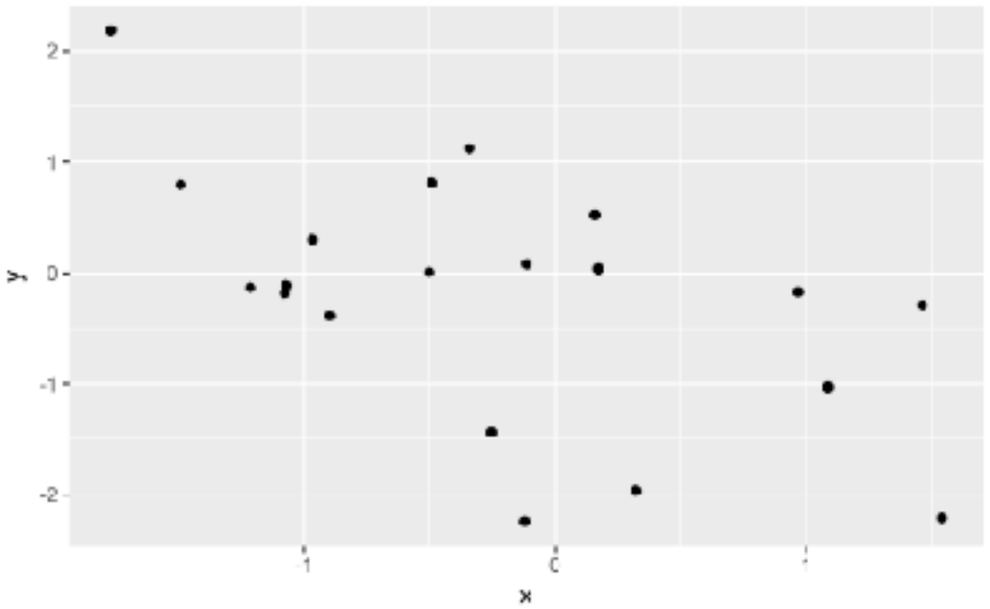
0.39

0.63

0.81


r=0.63
2
1
0
-1
-2
y
-1
0
1
2
x

Guess the correlation coefficient for this dataset

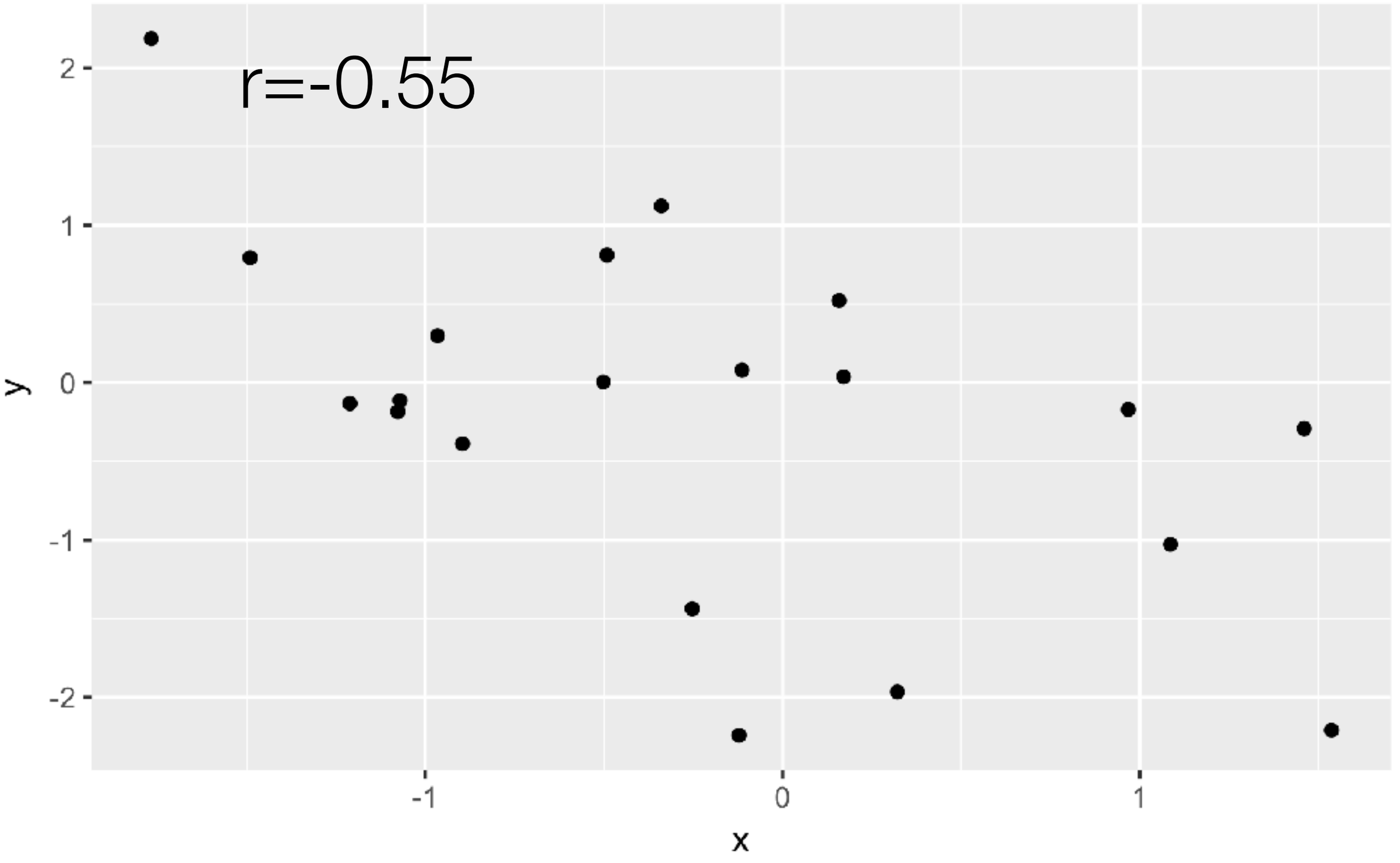


-0.71

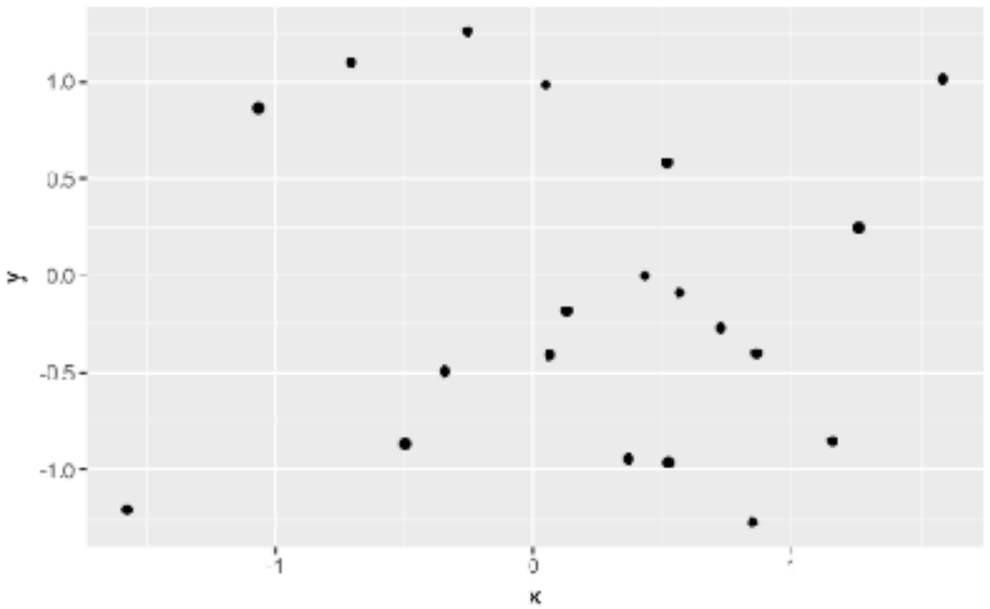
-0.55

-0.35

-0.12


r=-0.55
2
1
0
-1
-2
y
-1
0
1
x

Guess the correlation coefficient for this dataset

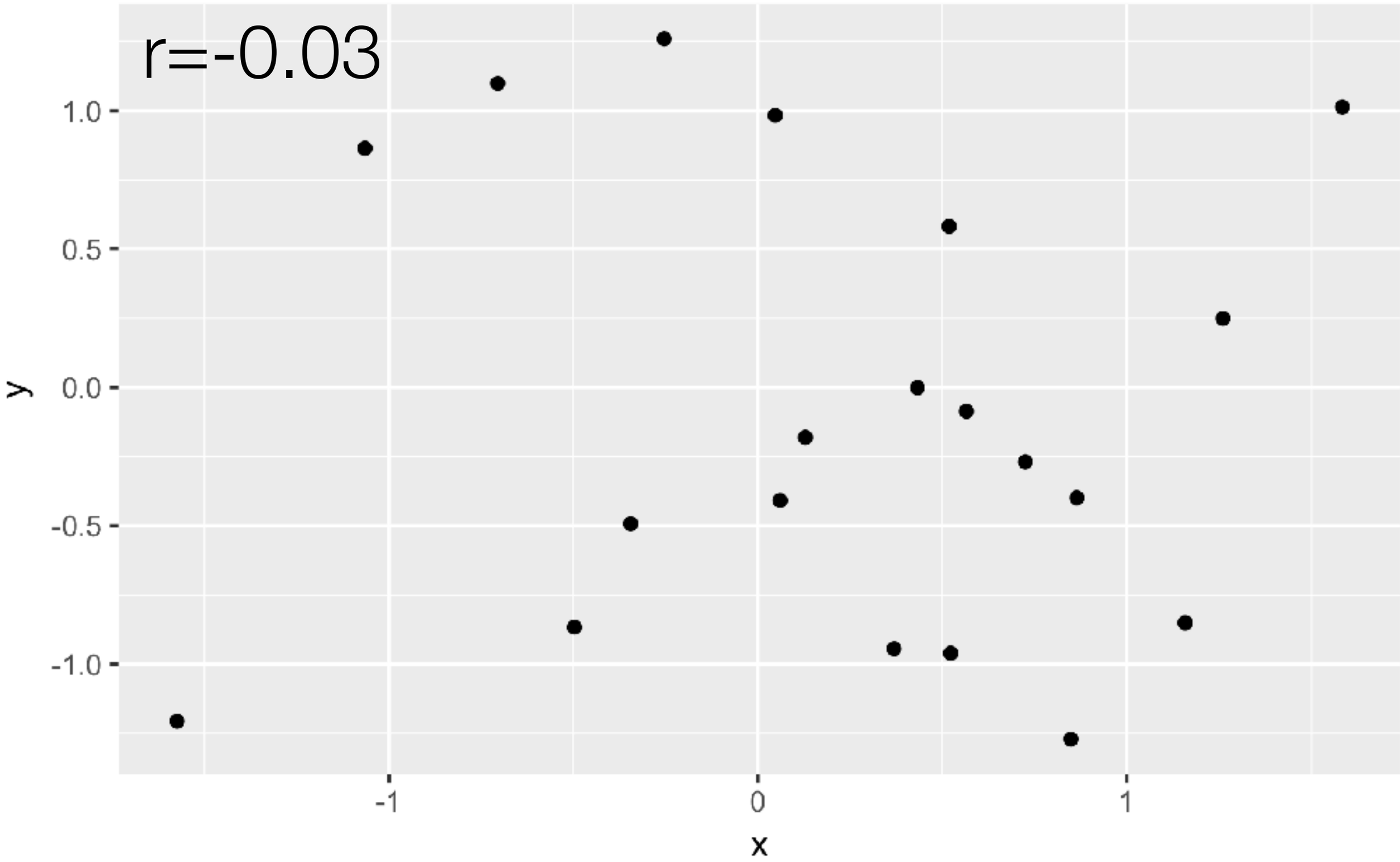


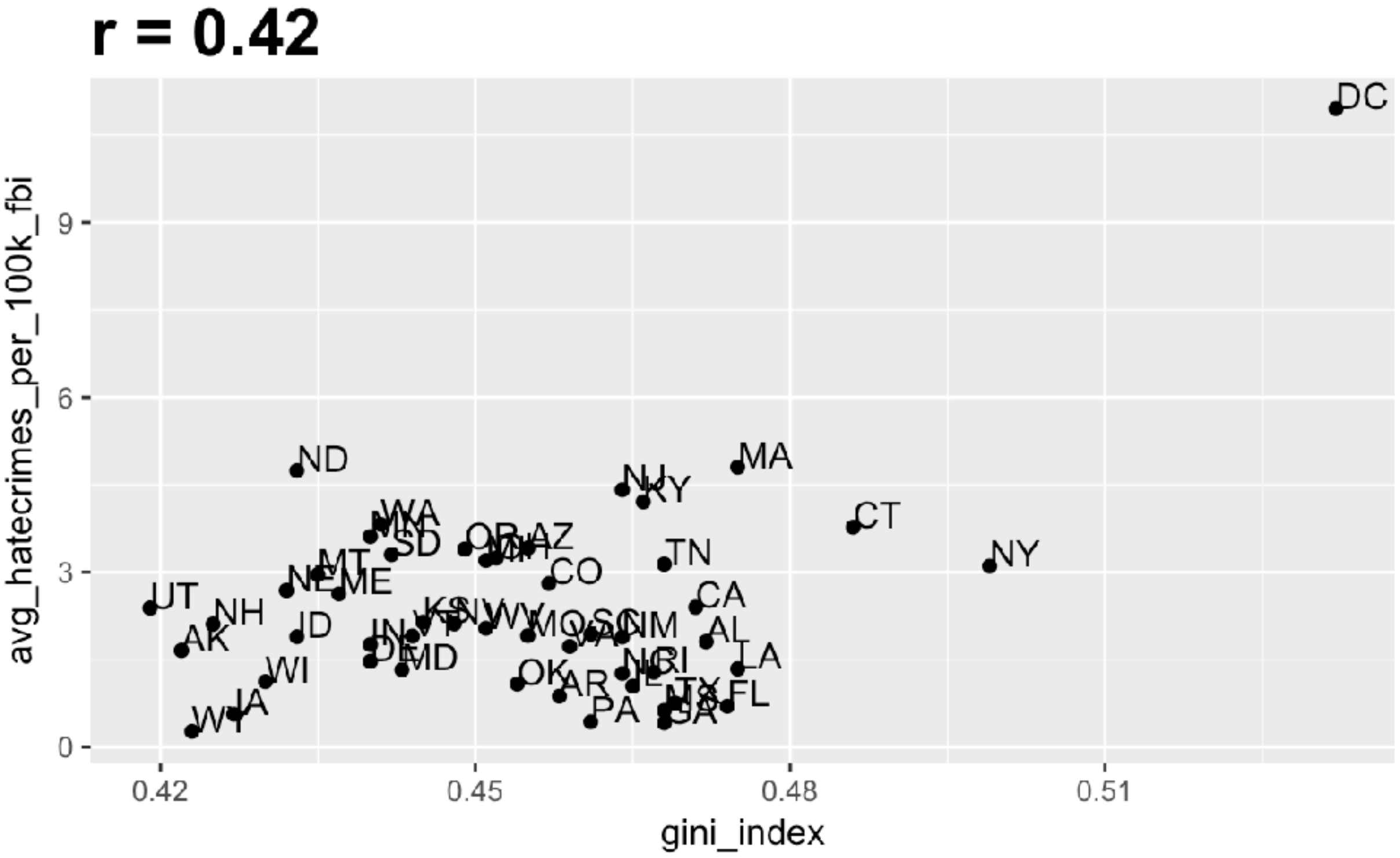
-0.28

-0.03

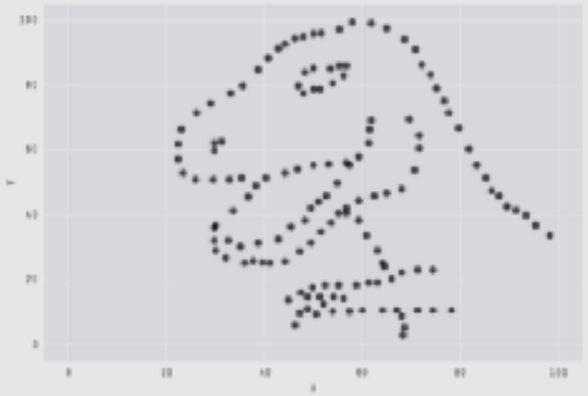
0.25

0.41

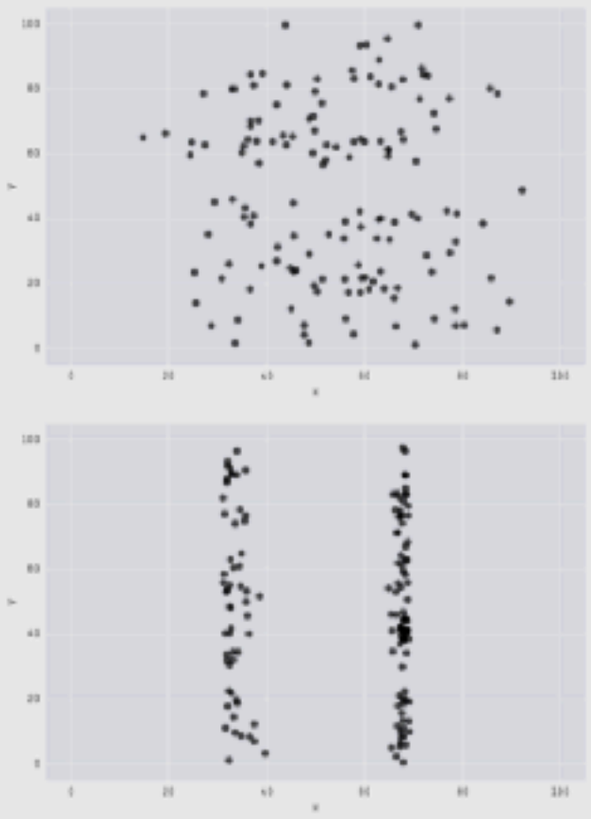

r=-0.03
1.0
0.5
0.0
-0.5
-1.0
y
-1
0
1
x

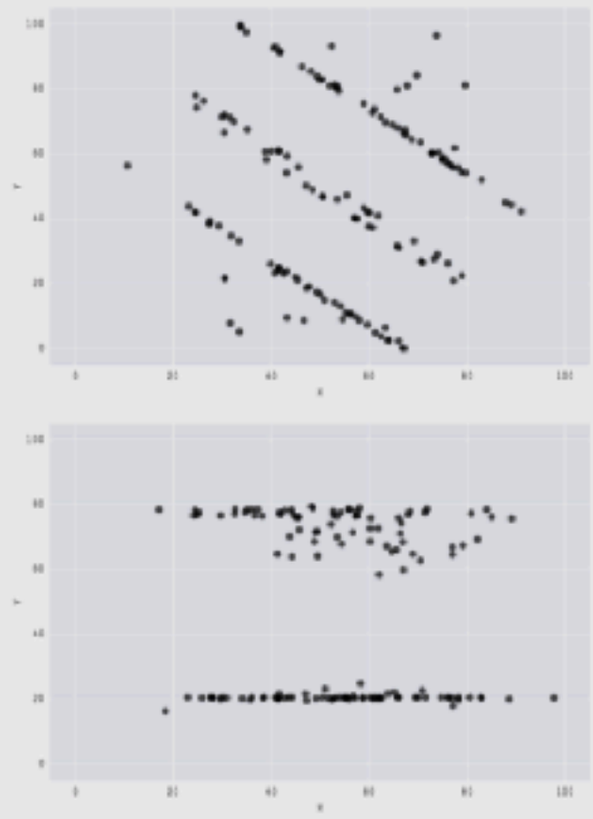

r = 0.42
avg_hatecrimes_per_100k_fbi
gini_index
0
3
6
9
0.42
0.45
0.48
0.51
DC
ND
MA
NJ
KY
CT
NY
TN
CO
CA
AL
LA
FL
ME
NH
AK
WI
MD
OK
PA
UT

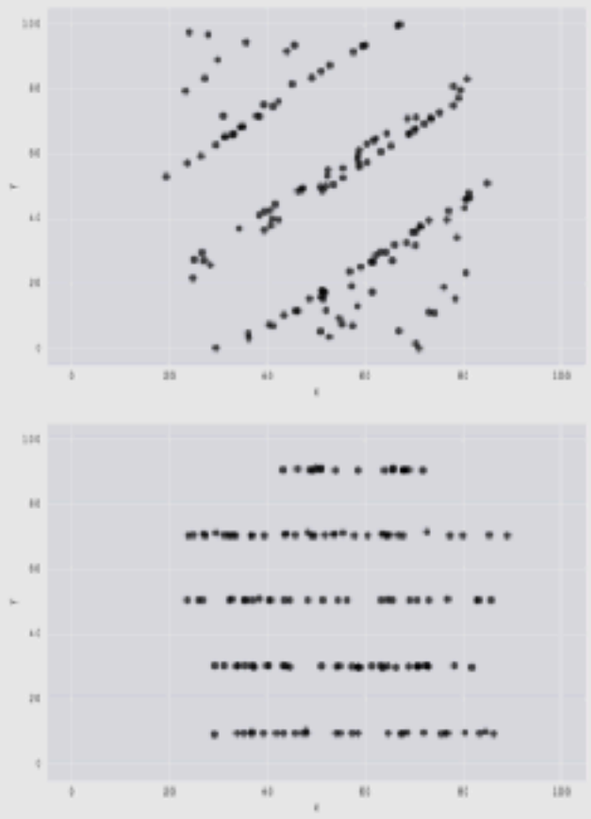
The "Datasaurus Dozen"

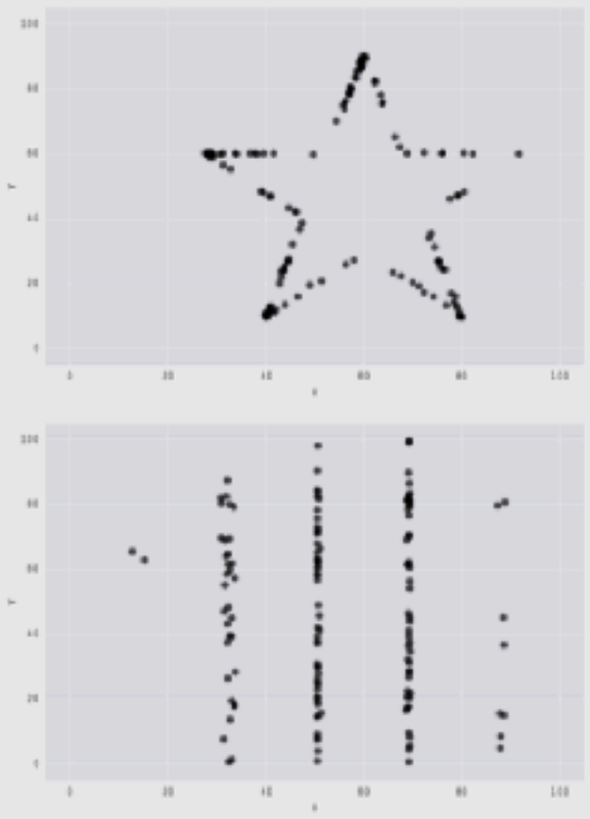


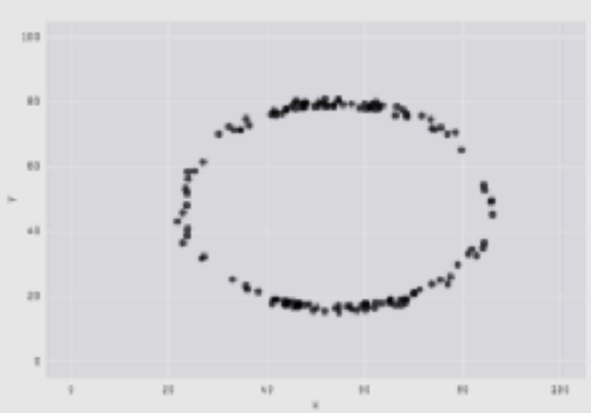
X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

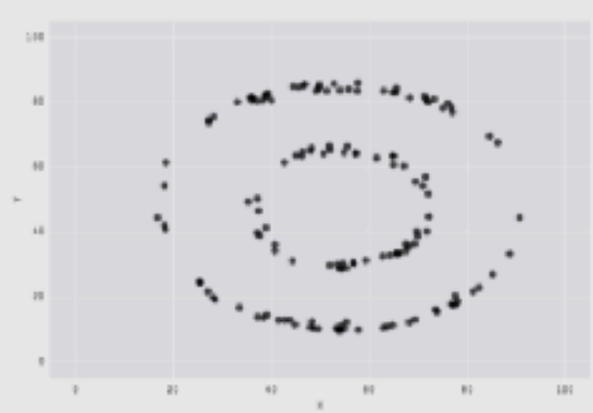


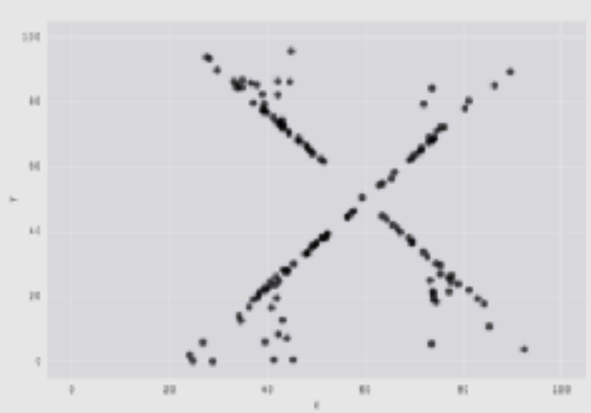


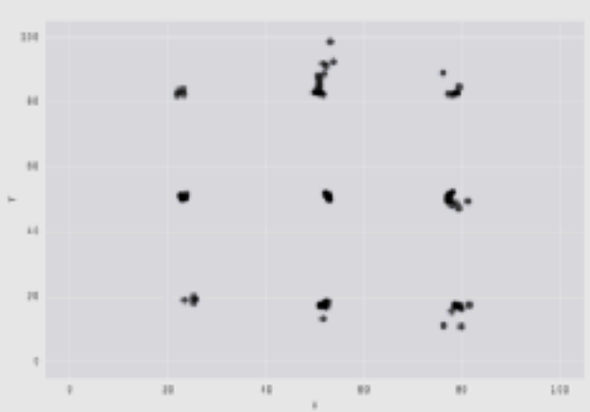










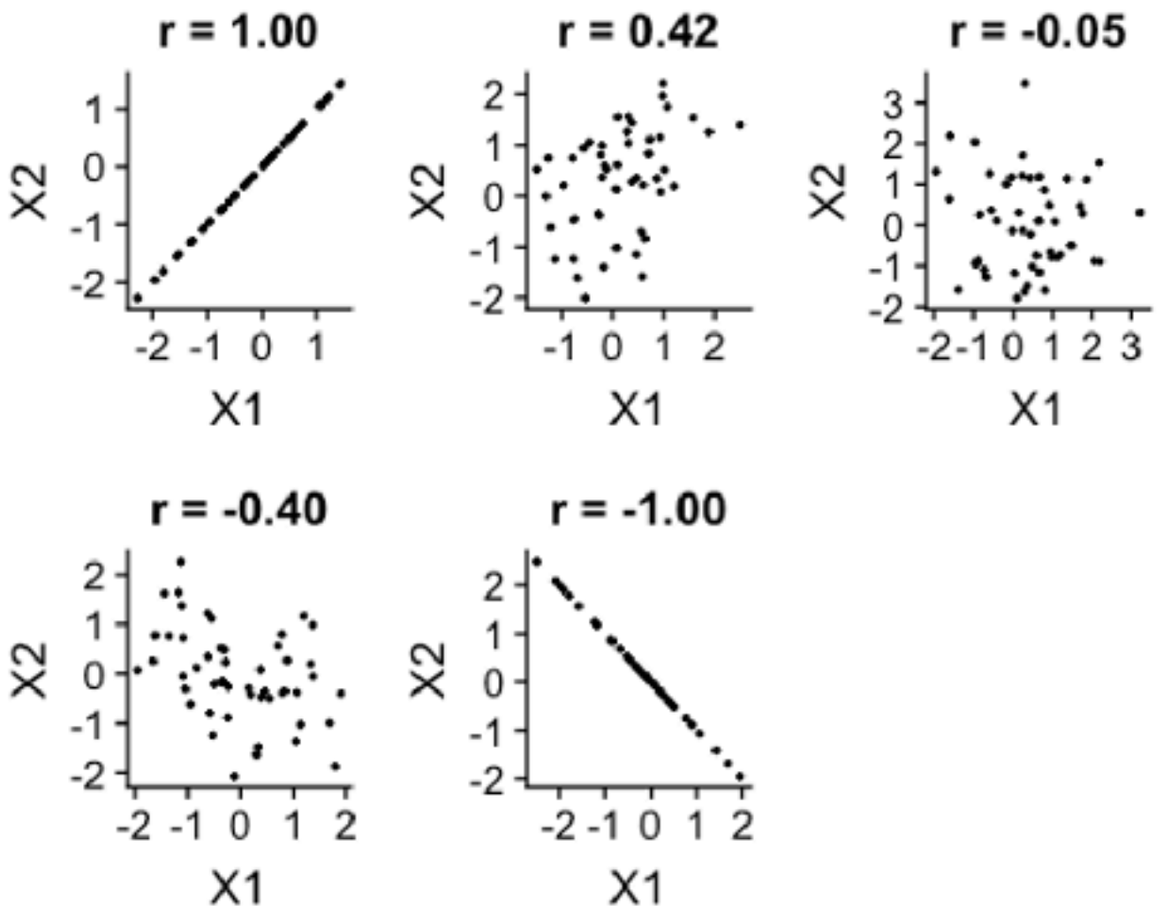


https://www.autodeskresearch.com/publications/samestats

Summary

PEARSON'S R

- also known as the *correlation coefficient*
- r is a measure of the strength of the linear relationship between two continuous variables.
- varies from -1 to 1
- 1 represents a perfect positive relationship between the variables
- 0 represents no relationship
- -1 represents a perfect negative relationship. F



Examples of various levels of Pearson's r.

Statistical significance of the correlation

- As usual, there are multiple ways…
-

Statistical significance of the correlation

- As usual, there are multiple ways…
- Simple approach: t-test

$$t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

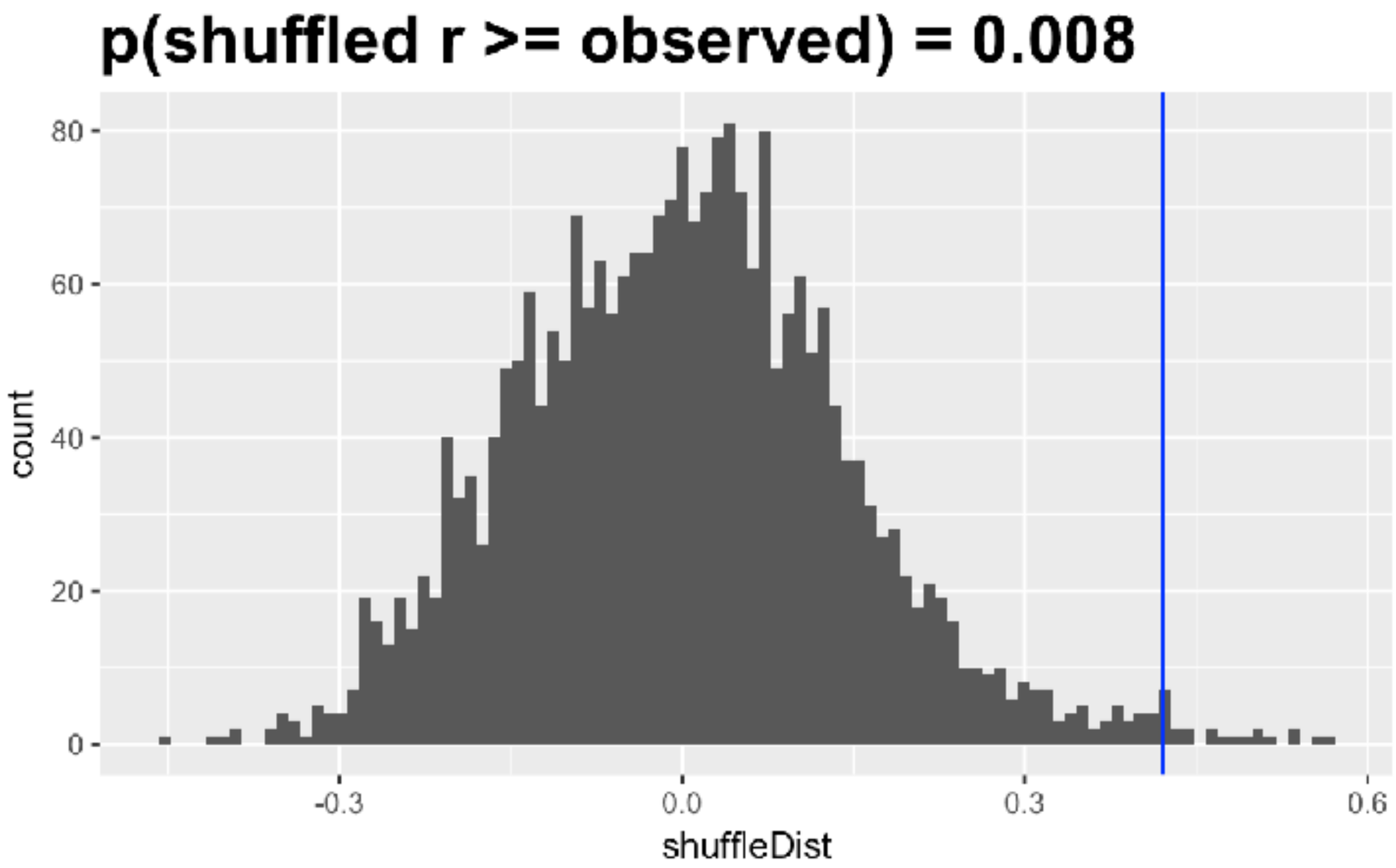
Distributed as t(N-2) under H_0: r=0
Assumes that underlying data are normally distributed

In R: `cor.test()`

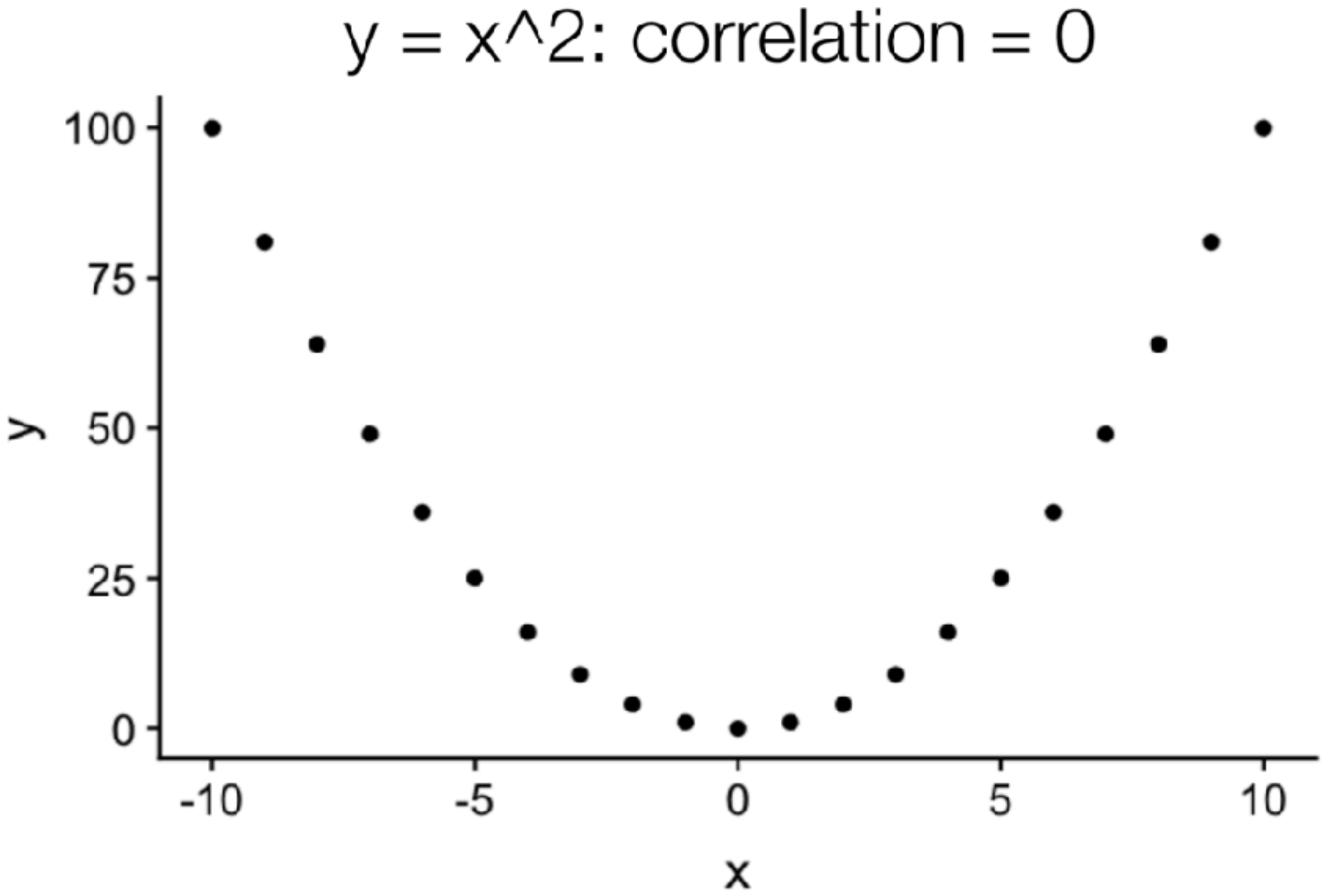
```
cor.test(hate_crimes$avg_hatecrimes_per_100k_fbi,
    hate_crimes$gini_index,alternative='greater')

  Pearson's product-moment correlation

data:  hate_crimes$avg_hatecrimes_per_100k_fbi and
hate_crimes$gini_index
t = 3.2182, df = 48, p-value = 0.001157
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.2063067 1.0000000
sample estimates:
      cor
0.4212719
```

Randomization

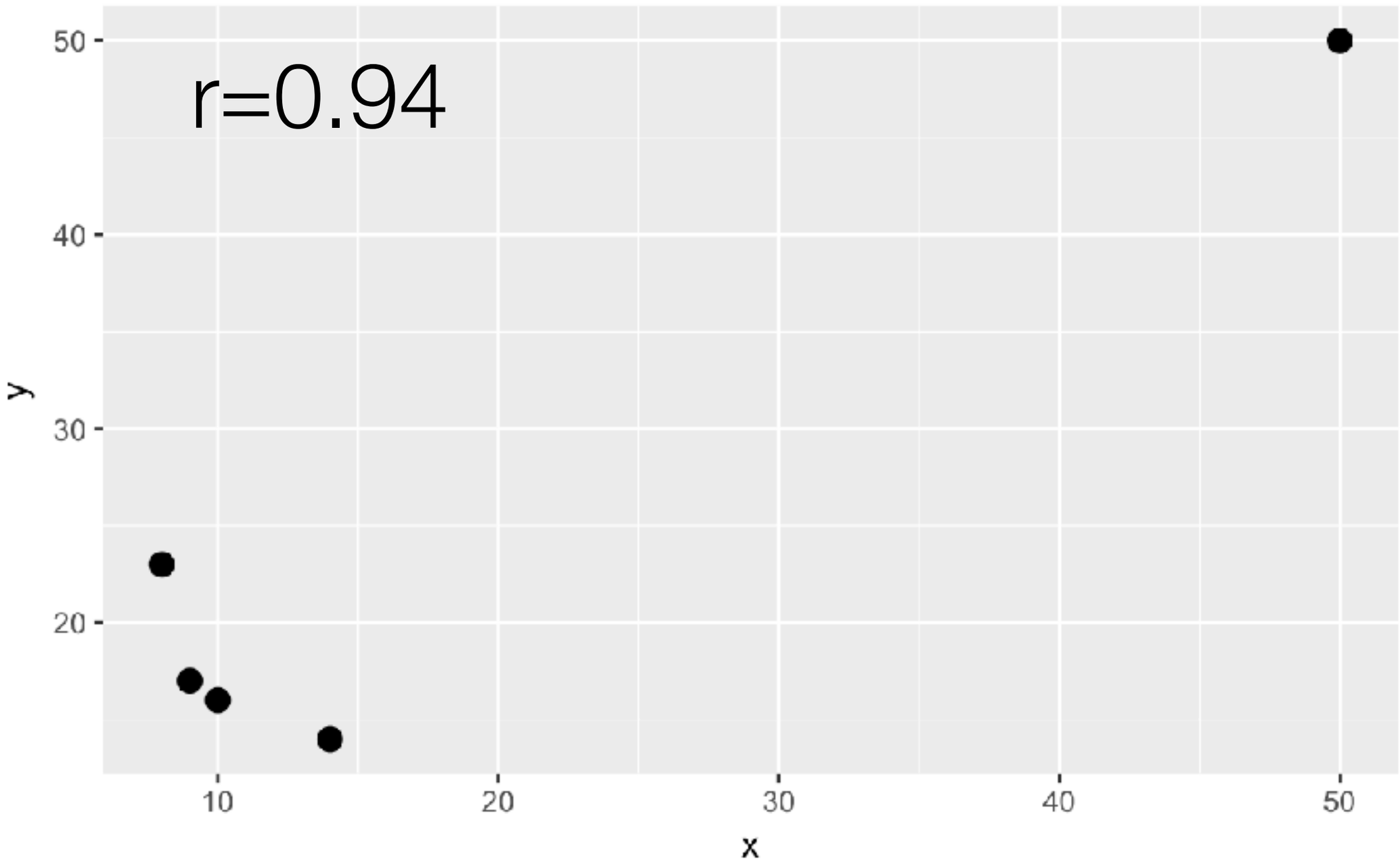
- Randomly shuffle values for one variable and compute correlation to obtain empirical null distribution



Correlation is only sensitive to linear relationships



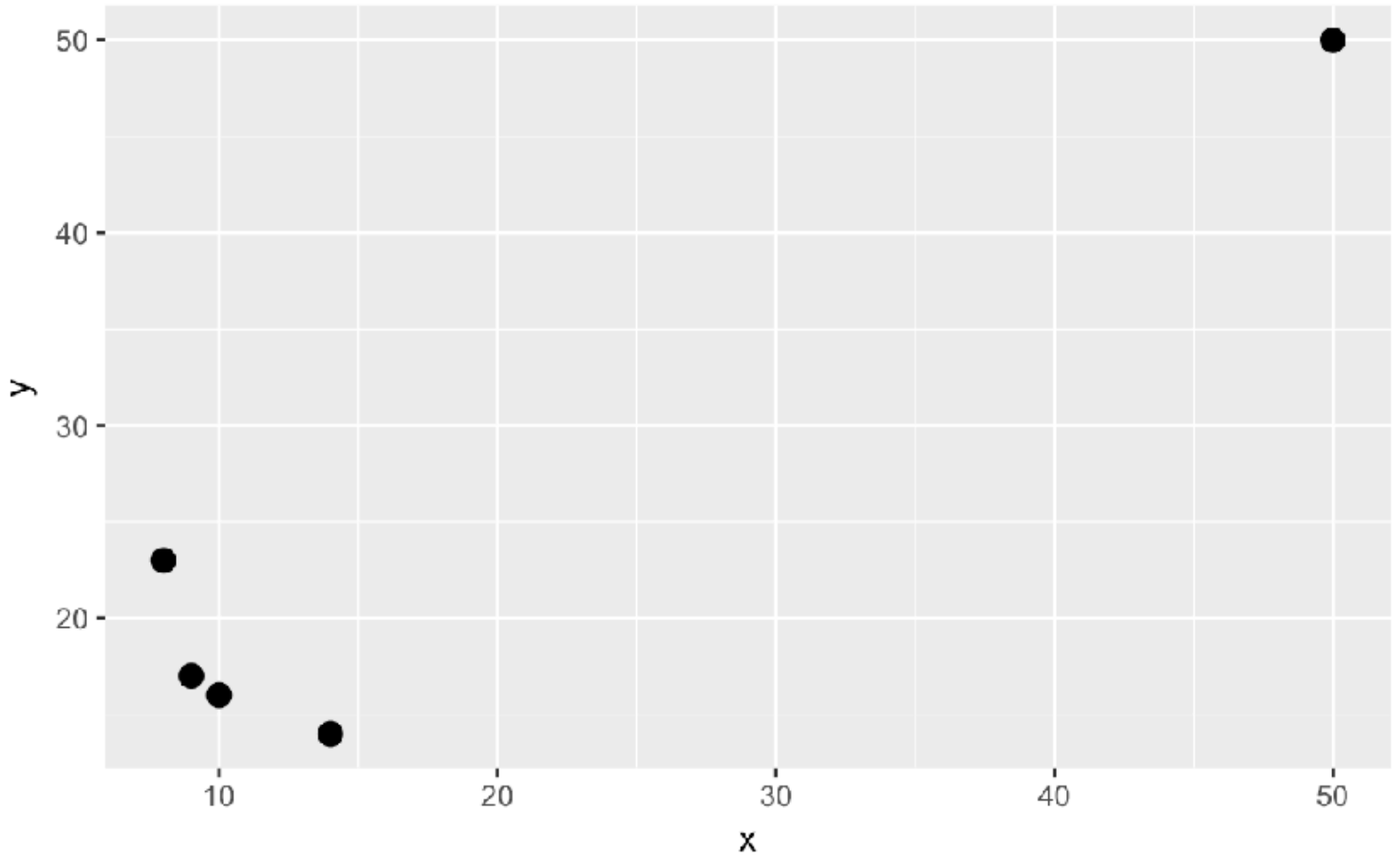
Correlation is very sensitive to outliers



Robust correlation: Spearman's rank correlation

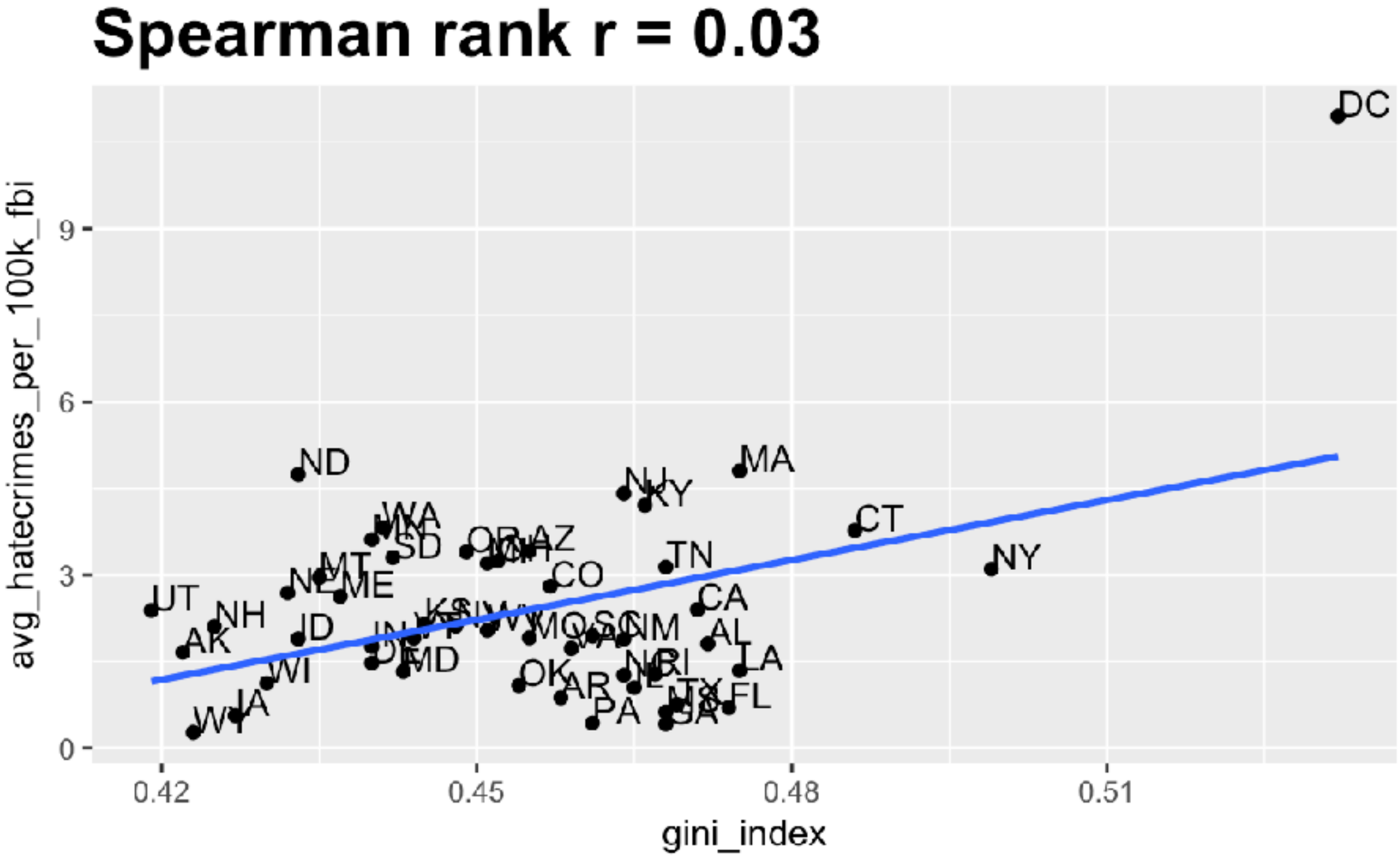
- Instead of computing correlation on raw values, compute correlation on ranks

x	y	rank(x)	rank(y)
8	23	1	4
9	17	2	3
10	16	3	2
14	14	4	1
50	50	5	5




```
> cor(df$x,df$y)
[1] 0.9435793
> cor(df$rankx,df$ranky)
[1] 0
```

Reducing the effects of outliers



Why it's always important to look at the data…

FiveThirtyEight

Polit Sports Science & Health Economics Culture

JAN. 23, 2017 A 12:18 PM

Higher Rates Of Hate Crimes Are Tied To Income Inequality

By Maimuna Majumder

Filed under Hate Crimes

Get the data on GitHub

In the 10 days after the 2016 election, nearly 900 hate incidents were reported to the Southern Poverty Law Center, averaging out to 90 per day. By comparison, about 36,000 hate crimes were reported to the FBI from 2010 through 2015 — an average of 16 per day.

https://fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/

Alcohol Plays a Much Bigger Role in Causing Dementia Than We Thought

Ed Cara

Wednesday 4:10pm • Filed to: ALCOHOL

45.3K 42 1

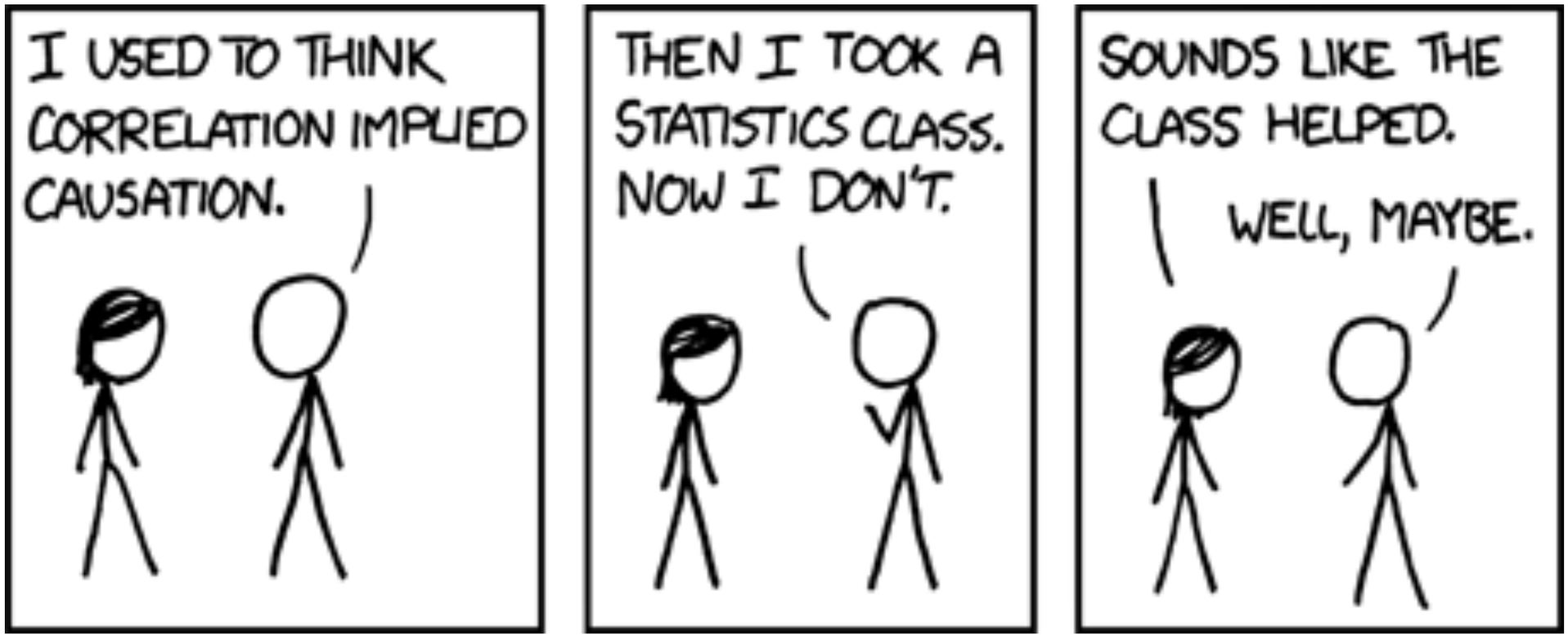
The researchers looked at a nationwide, anonymous database of more than 30 million adult French hospital patients who were discharged sometime between 2008 to 2013. ...

Narrowing in on the over 1 million patients newly diagnosed with dementia during that time, the researchers found that heavy alcohol use was a substantial risk factor for every common type of dementia, particularly early-onset cases caught before the age of 65. More than half of the 57,000 patients diagnosed with early-onset dementia—57 percent—showed signs of alcohol-related brain damage or were diagnosed with an alcohol use disorder at the same time.

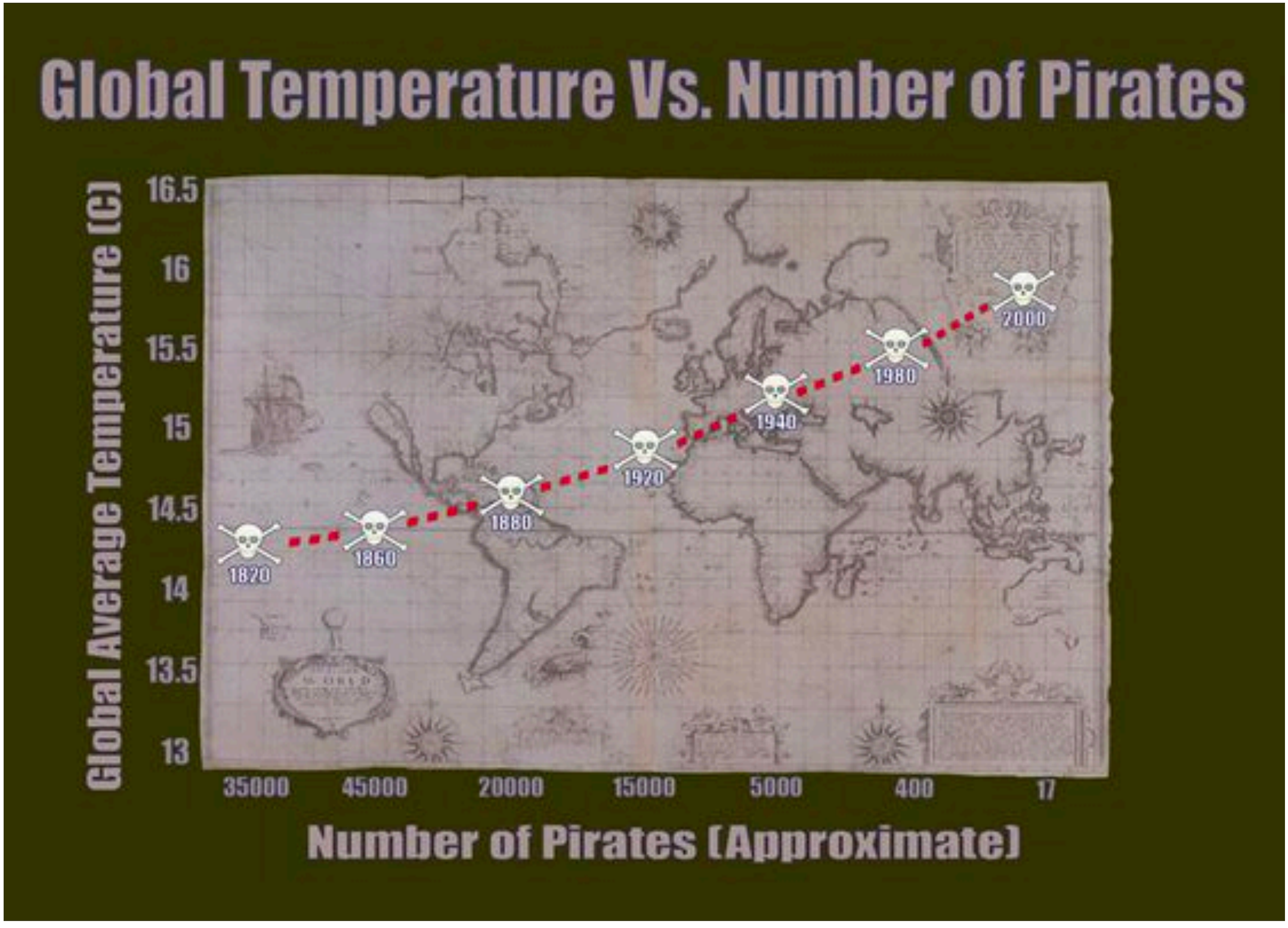
“If all these measures [increased alcohol taxes and advertising bans] are implemented widely, they could not only reduce dementia incidence or delay dementia onset, but also reduce all alcohol-attributable morbidity and mortality,” they wrote.

https://gizmodo.com/alcohol-plays-a-much-bigger-role-in-causing-dementia-th-1823198004

Correlation and causation



https://xkcd.com/552/



https://www.forbes.com/sites/erikaandersen/2012/03/23/true-fact-the-lack-of-pirates-is-causing-global-warming/

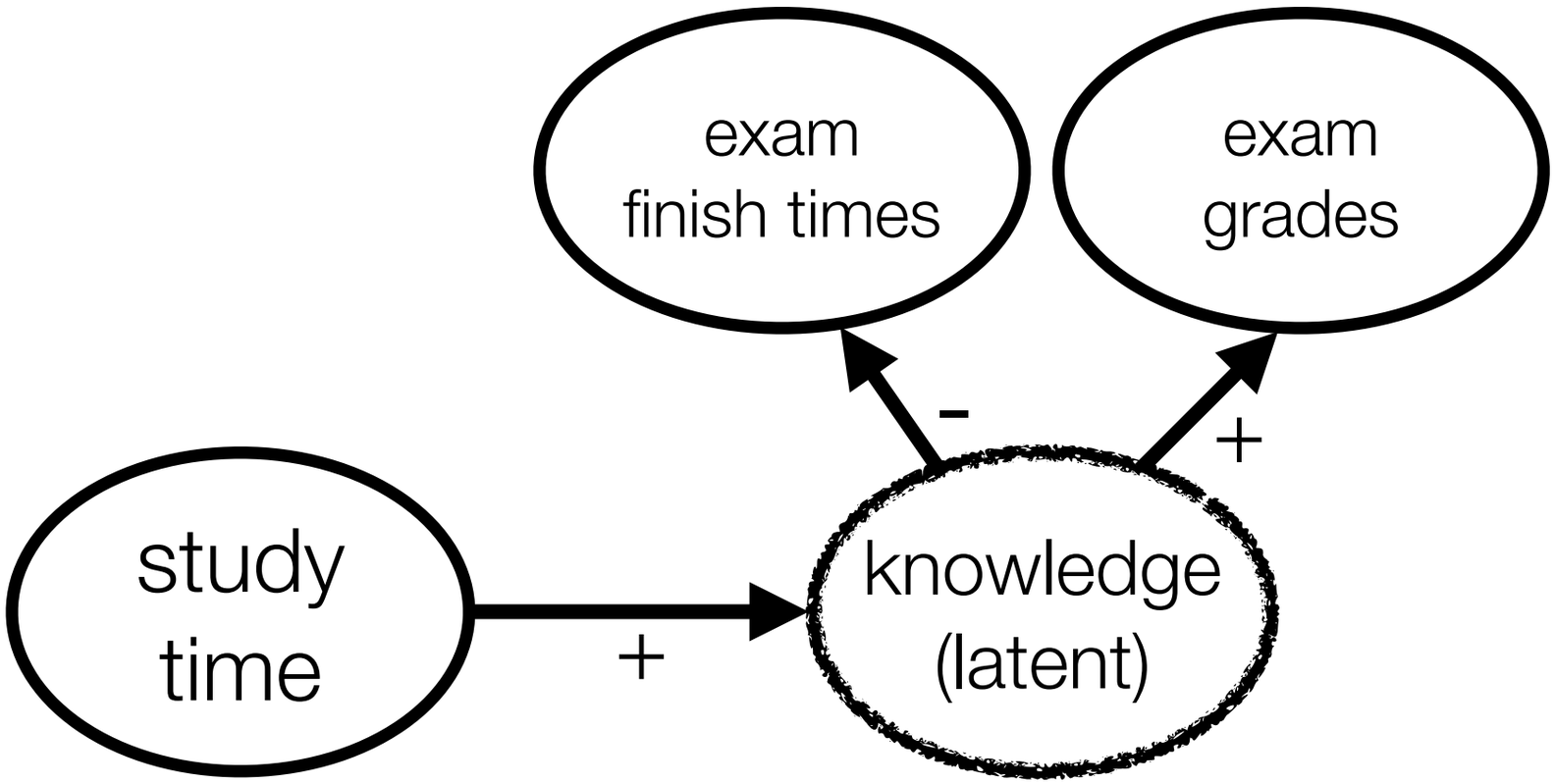
http://www.tylervigen.com/spurious-correlations

"Correlation does not imply causation, but it's a pretty good hint"

Edward Tufte

Understanding causation using causal graphs

- A causal graph describes the latent causal relations that give rise to the variables that we measure



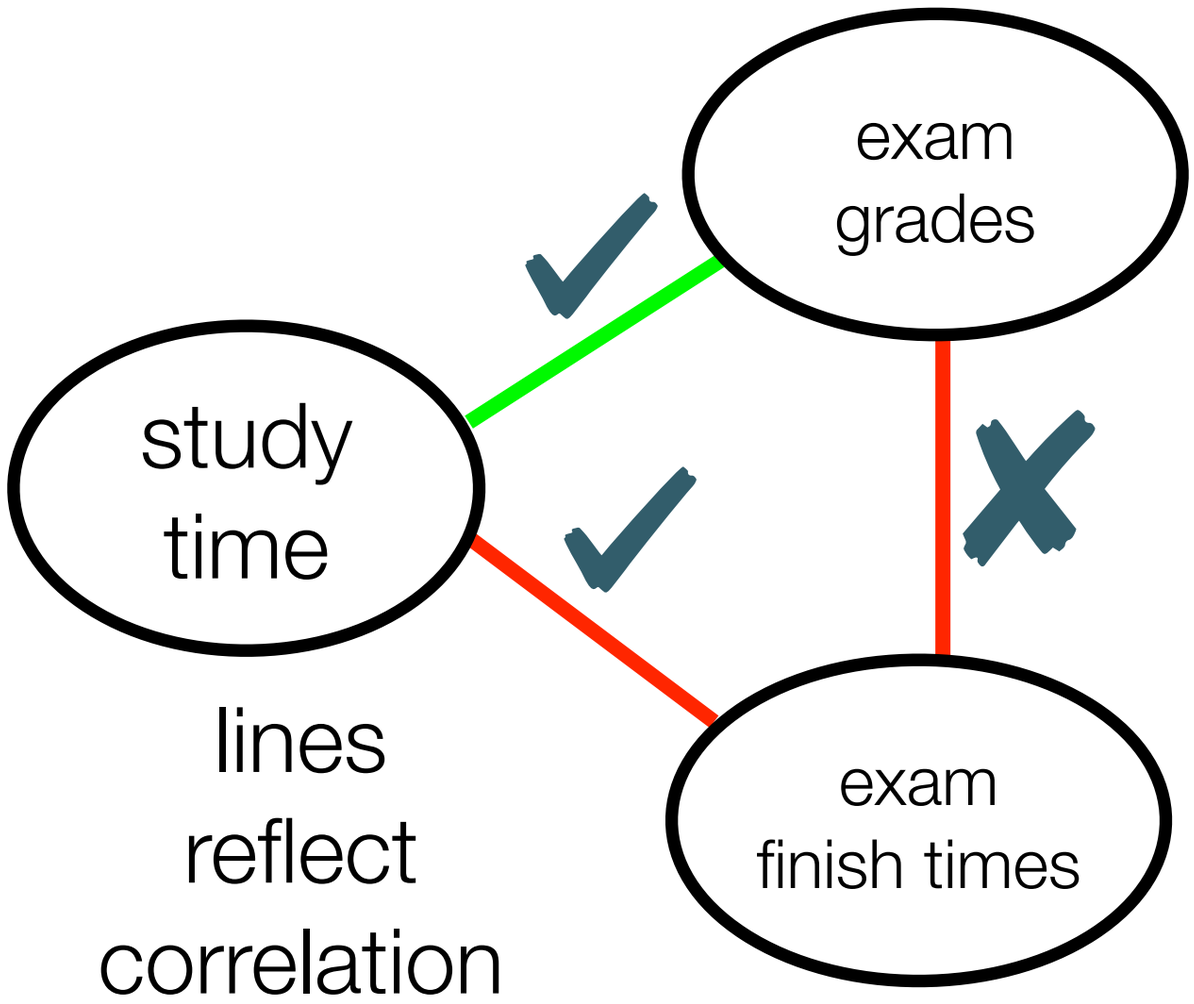
arrows reflect causal relations

Causal relations mean that manipulating one variable will change another

Increasing study time will increase knowledge, which increases grades and reduces exam finishing time

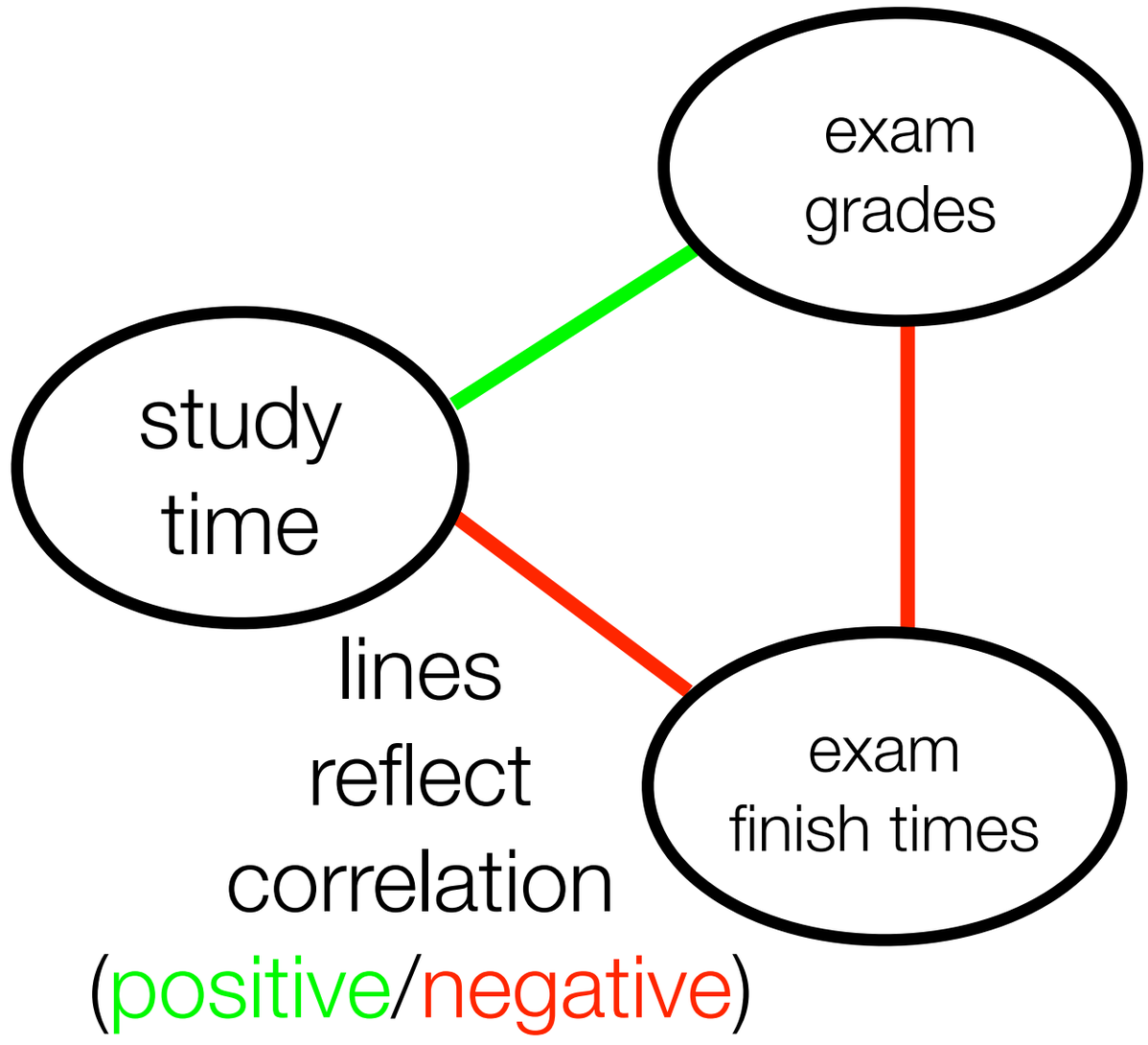
Correlation and causation

- Correlations can reflect causal relations or effects of common causes



Correlation and causation

- Correlations can sometimes imply the wrong causal relation
- Negative correlation between exam grades and exam finishing time
  - Implies that finishing the exam faster will improve grades!



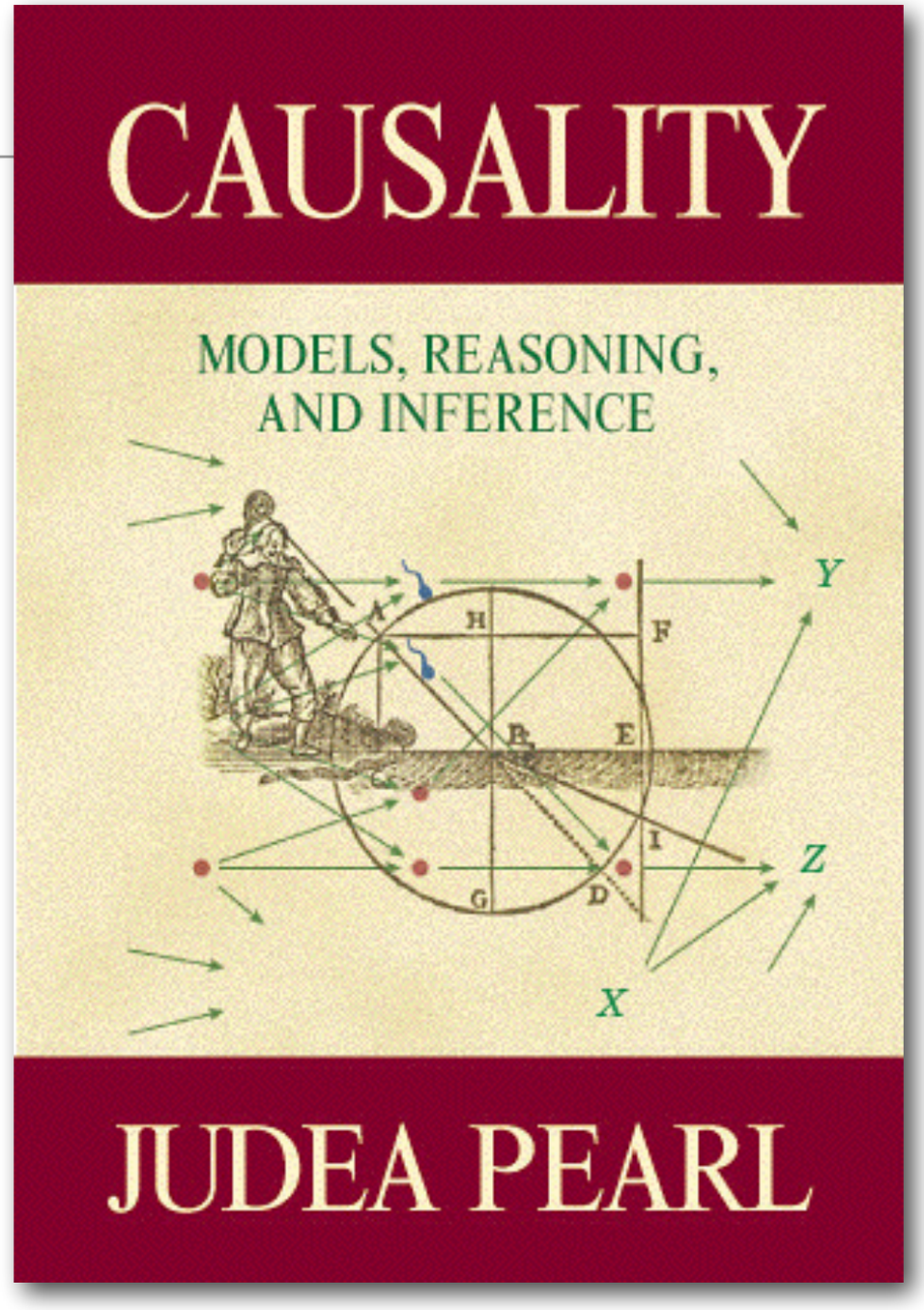
Group discussion

- Read this article:
  - https://www.washingtonpost.com/lifestyle/wellness/how-an-anti-inflammatory-diet-can-help-tame-an-autoimmune-condition/2019/02/14/21a52e24-2fcc-11e9-8ad3-9a5b113ecd3c_story.html
- Can you find any problematic causal claims?

Inferring causal relations

- With more than two variables, we can sometimes infer causal relations from correlational data

- This is a very active area in machine learning research



Recap

- Correlation quantifies the linear relationship between two variables
- Correlation is very sensitive to outliers
  - Always important to look at the data!
- Correlation does not imply causation, but it's often a pretty good hint