

Reporting Statistics

T test

There was a significant difference in the change scores between X intervention ($M = 8.61, SD = 5.62$) and Y intervention ($M = 2.54, SD = 2.20$); $t(12.30) = 3.10, p = 0.009$. Since we see a greater change before and after X compared to Y, we can conclude that X is more effective than Y.

An independent samples t-test shows no significant difference between coffee and non-coffee drinkers' energy levels, $t(55) = .37, p = .567$.

Correlation

We found was a significant moderate, positive relationship between sleep duration and mood ($r = 0.53$, $p = < .01$).

Chi Square Tests

- We can reject the null hypothesis that the students are equally distributed across introduction classes, $X^2(2, N=1000) = 11.23$ $p = .003$. From looking at the observed frequencies compared to those expected, it looks like fewer students enrolled in introduction to biology (~20%), compared to introduction to statistics or psychology (~40%).

describing results

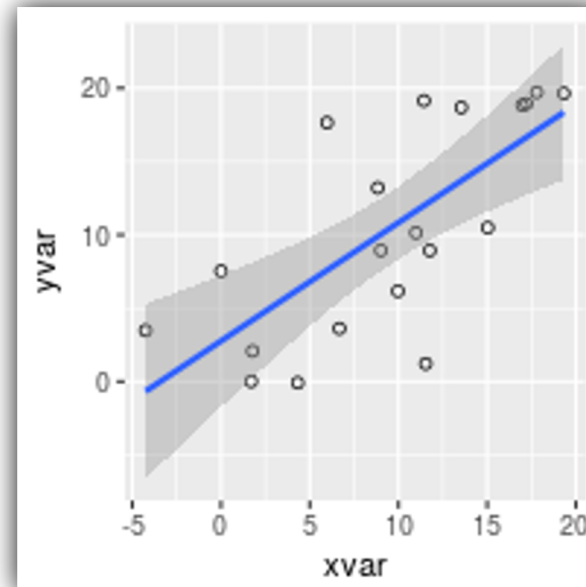
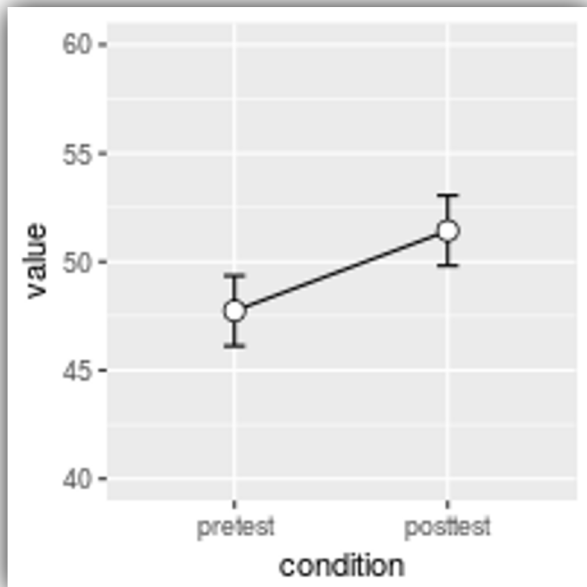
- See the American Psychological Association's guide to reporting results of statistical tests: <http://www.statisticssolutions.com/reporting-statistics-in-apa-format/>
- As predicted, results from an independent samples t test indicated that individuals diagnosed with schizophrenia ($M = .76$, $SD = .20$, $N = 10$) scored much higher on the sorting task than college students ($M = .17$, $SD = .13$, $N = 9$), $t(17) = 7.53$, $p < .001$, two-tailed. The difference of .59 scale points was large (scale range: 0 to 1; $d = 3.47$), and the 95% confidence interval around difference between the group means was relatively precise (.43 to .76).

describing results

- We found was a moderate positive relationship between sleep duration and mood ($r(112) = 0.53, p < .01$).
- By performing a linear regression, we can see there is a positive main effect of the number of hours studying on exam scores, $b = 8.2, t(67) = 5.21, p < .01$.
- We reject the null hypothesis that the students are equally distributed across introduction classes, $X^2(2, N=1000) = 11.23 p = .003$. A striking difference was that fewer students enrolled in introduction to biology (20%), compared to introduction to psychology (40%).

Visualizing results

- Guides: <http://www.cookbook-r.com/Graphs/>
- Examples and code: <http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html>
- `ggtitle()`, `ggsave()`, `theme(text = element_text(size = 20))`



Discussion

- Tie results back to your research question and hypotheses
 - Our results provide support for our hypothesis that...., or Our results did not provide evidence for our hypothesis that...
 - Discuss impact of findings and tie to motivation
- Discuss at least one limitation
 - *Some* examples:
 - didn't have the right variables to fully explore your research question- if this is the case, be comprehensive in naming the types of variables that would have been better to test
 - composition of sample
 - method of data collection
- Discuss future directions
 - "Future research should examine..."

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](#)

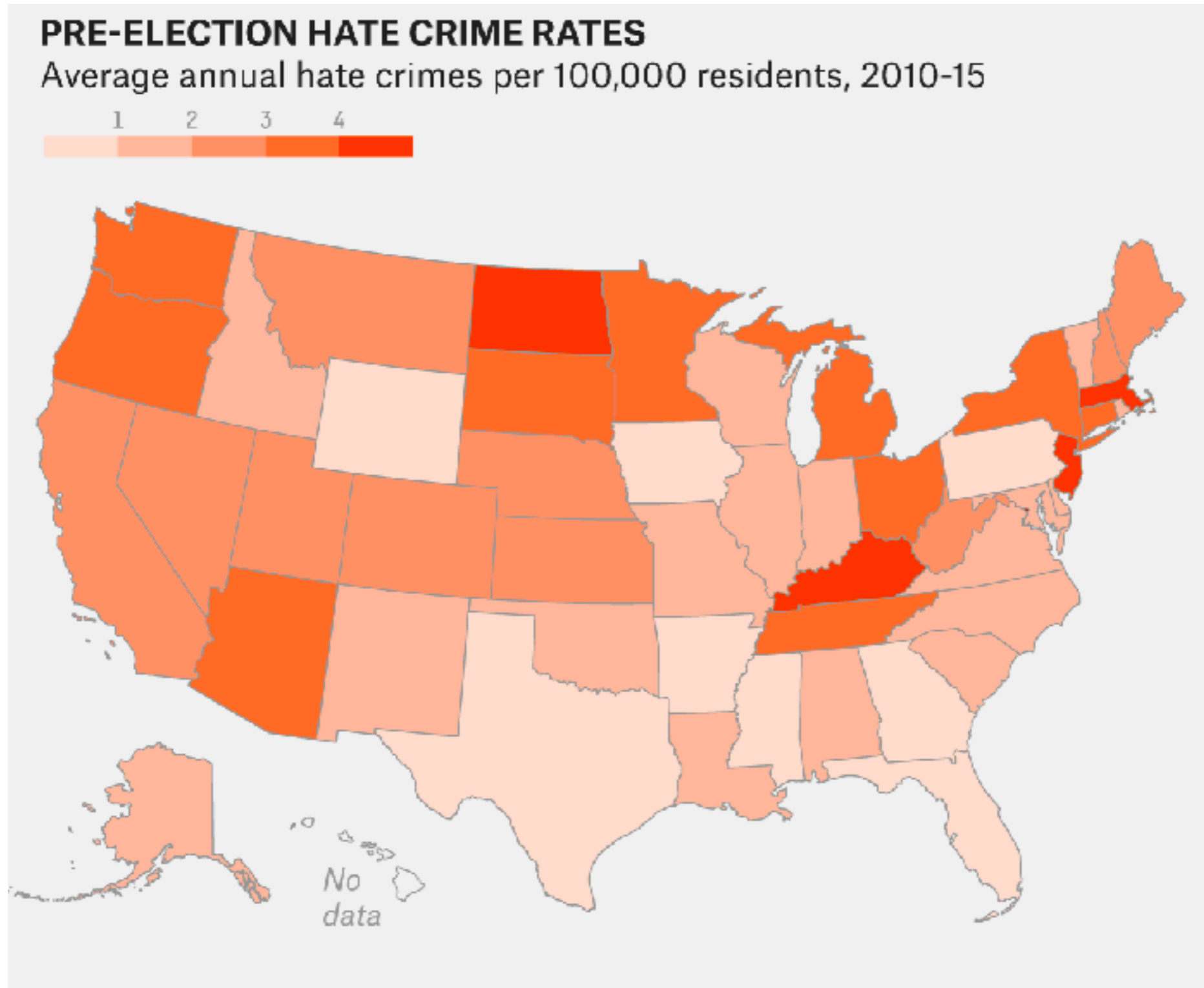
JAN. 23, 2017 AT 12:18 PM

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.

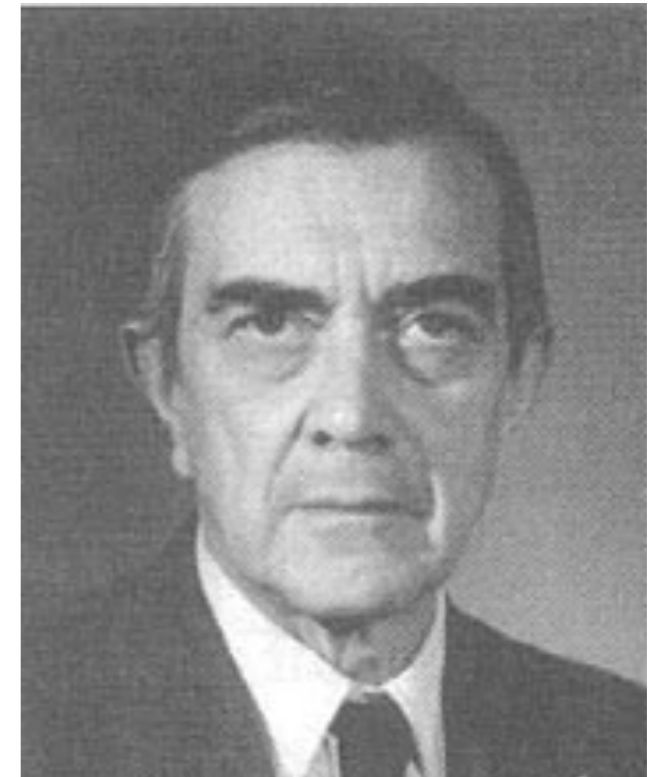
Hate crime rates differ across states



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”

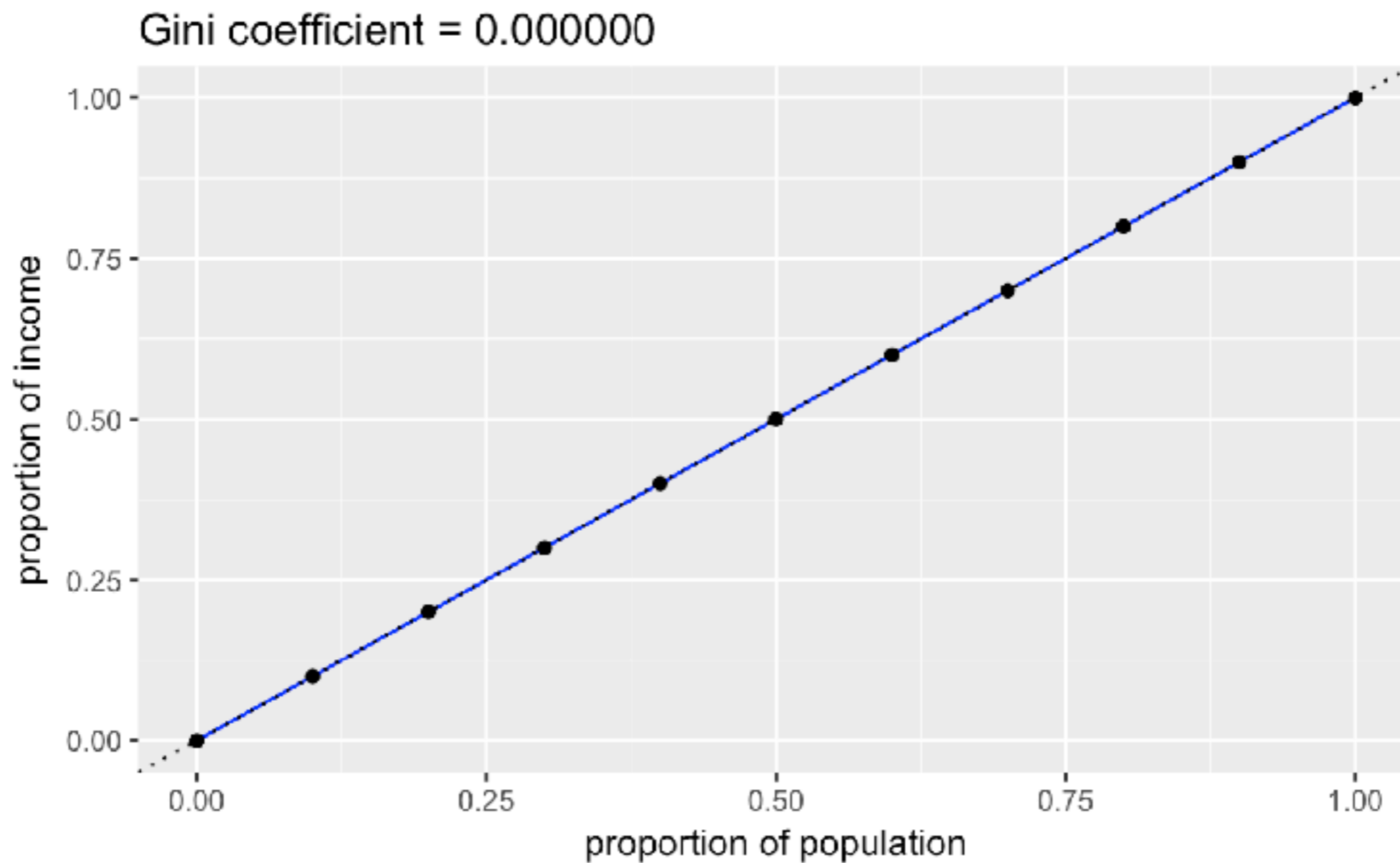
Corrado Gini



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

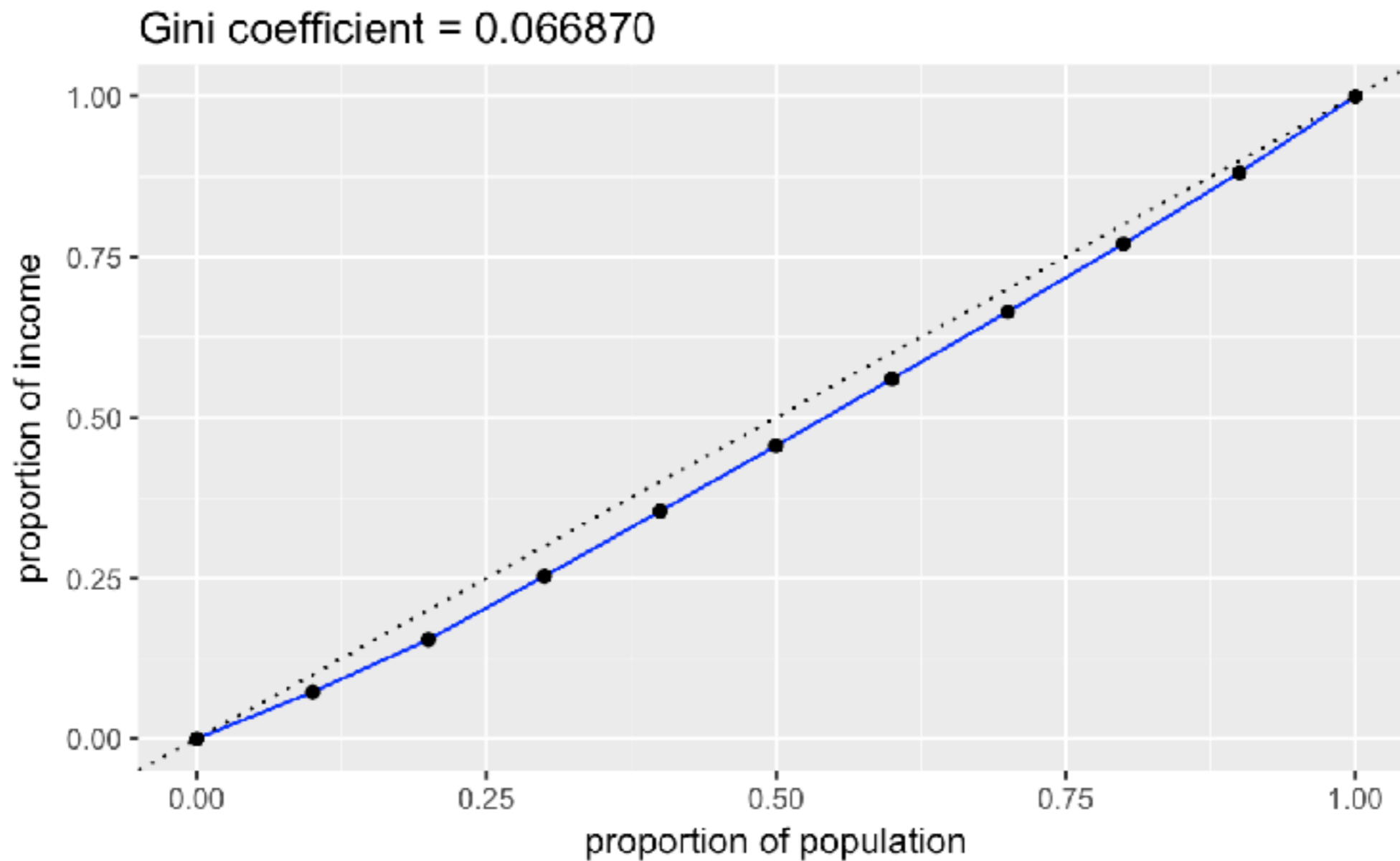
Example: perfect income equality

- 10 people, all incomes = \$40,000



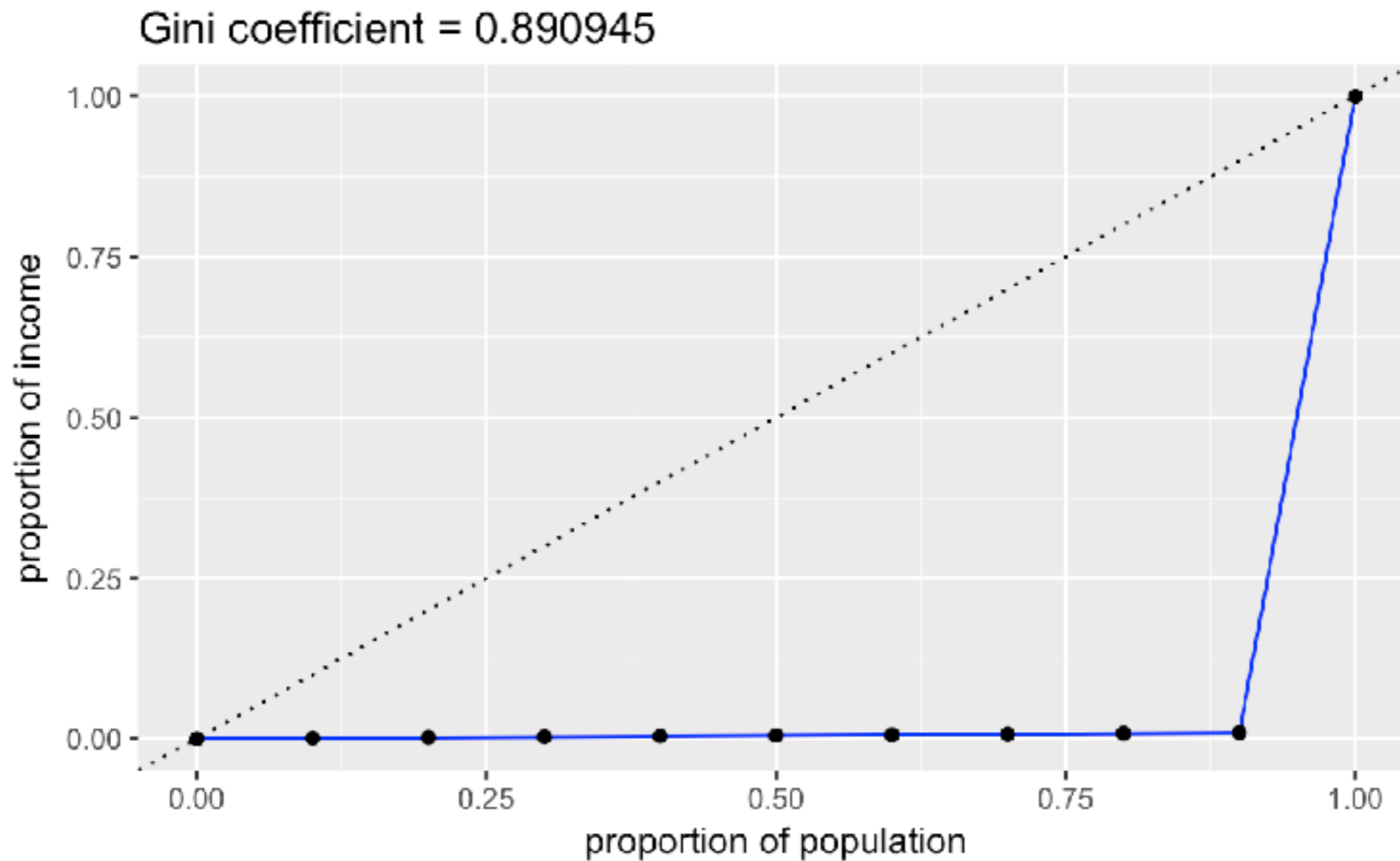
Example: mild inequality

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



Example: severe inequality

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



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

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

“cross product”



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

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

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

- 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

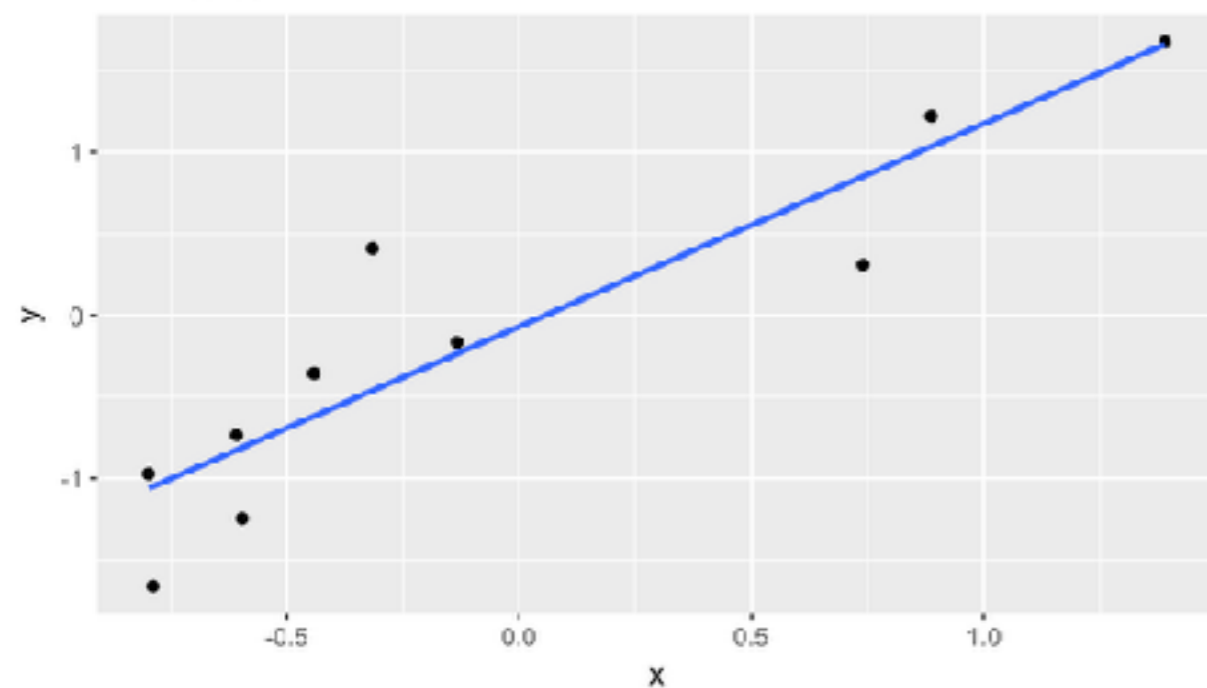
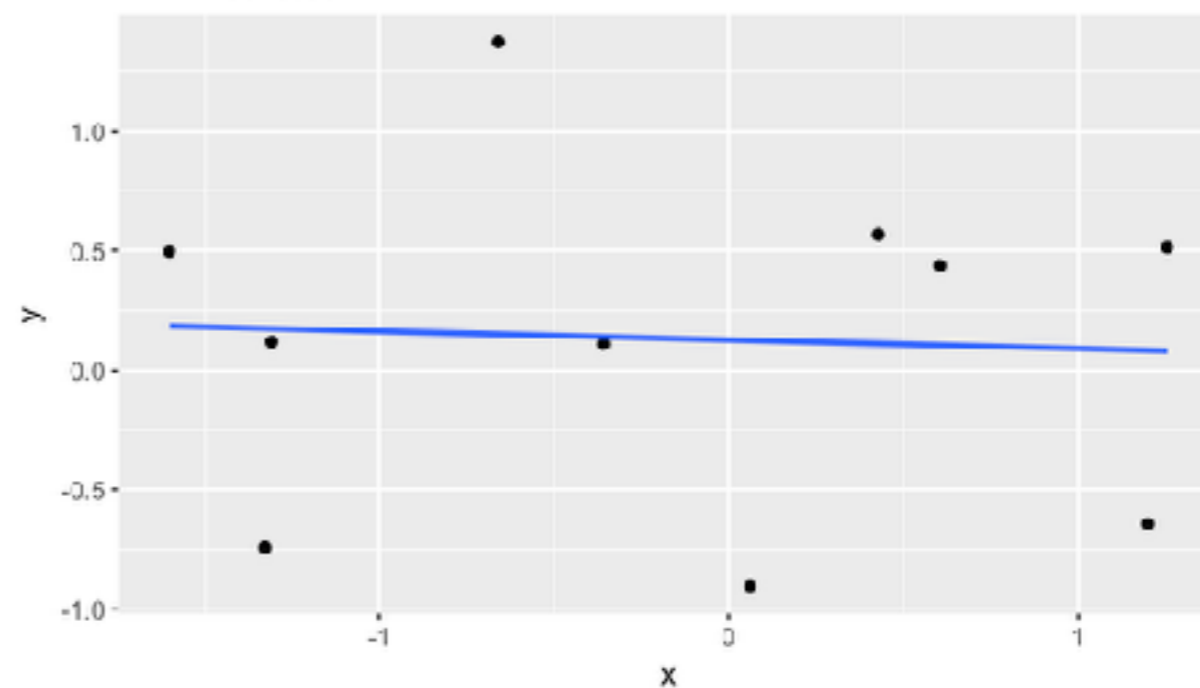
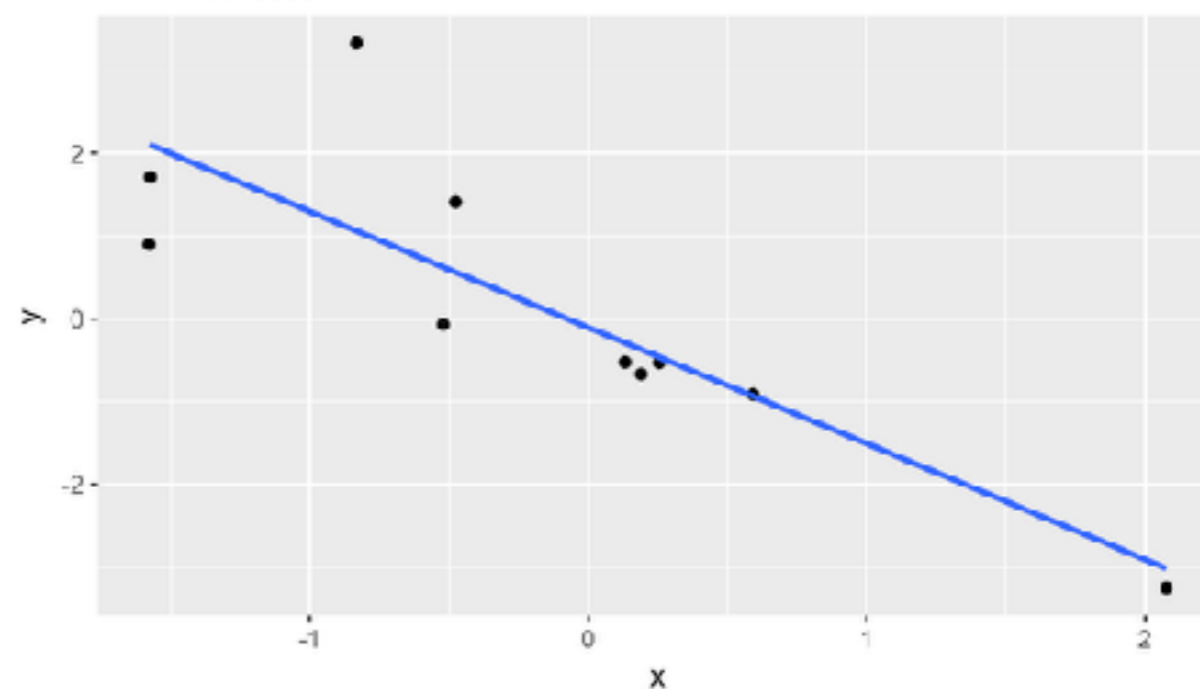
$$\text{sum} = 59$$

$$\text{covariance} = 59/4 = 14.85$$

$$\text{sd}(x) = 3.65$$

$$\text{sd}(y) = 4.42$$

$$r = 14.85 / (3.65 * 4.42) = 0.92$$

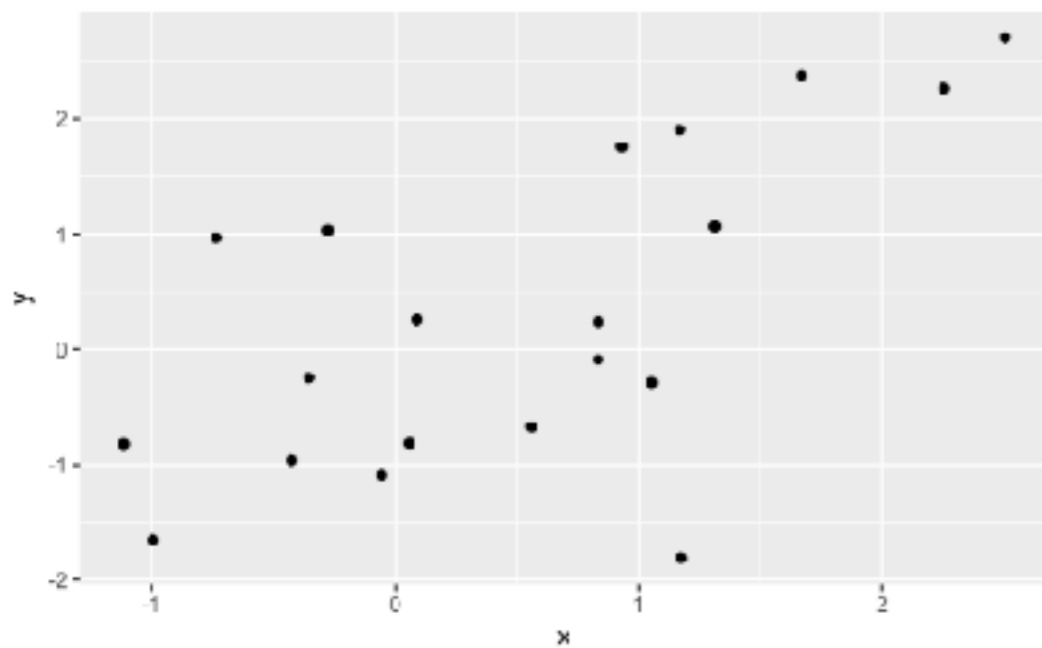
$r = 0.91$  **$r = -0.05$**  **$r = -0.85$** 

$r=1$: perfect positive relationship

$r=0$: no linear relationship

$r=-1$: perfect negative relationship

Guess the correlation coefficient for this dataset



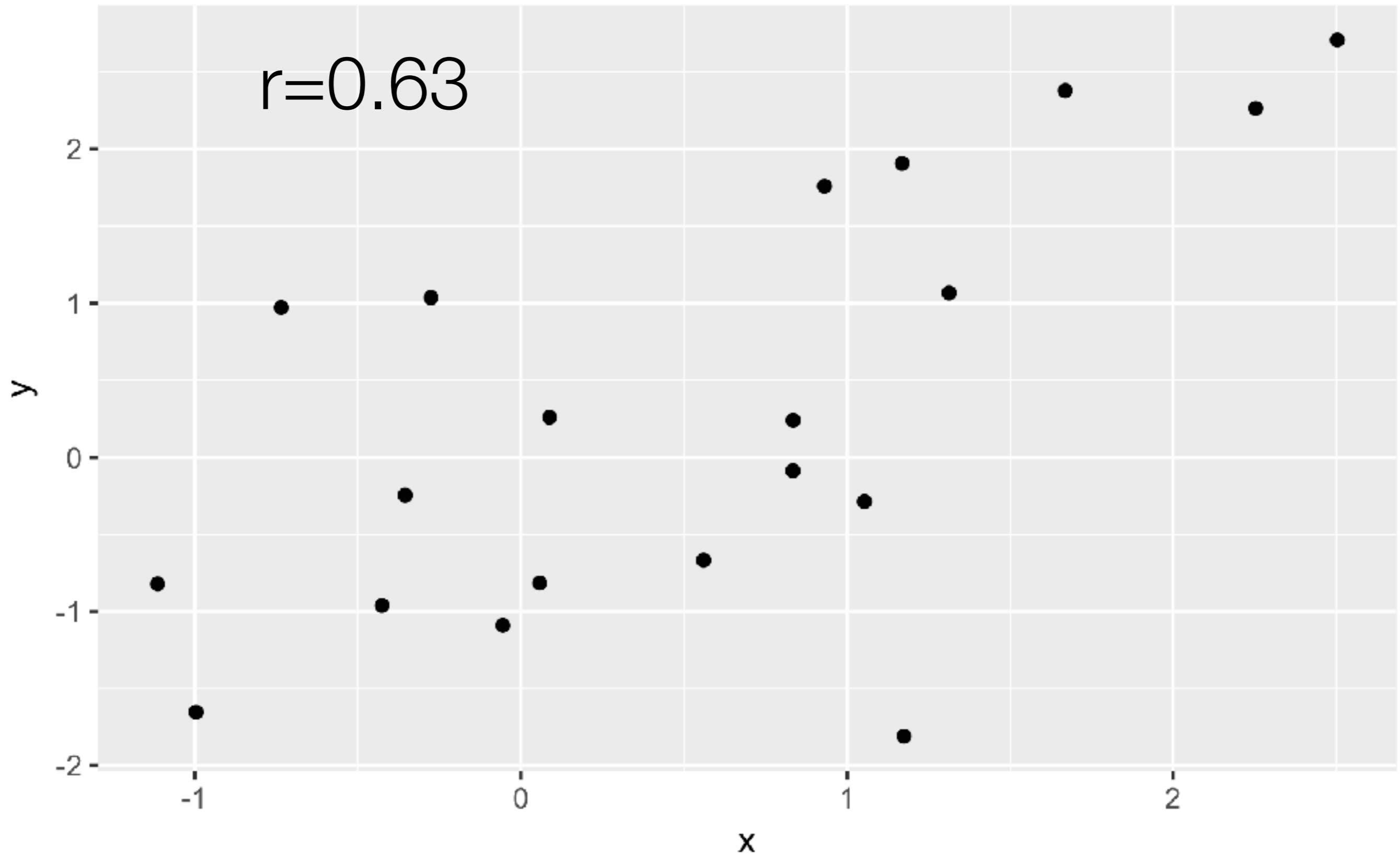
0.15

0.39

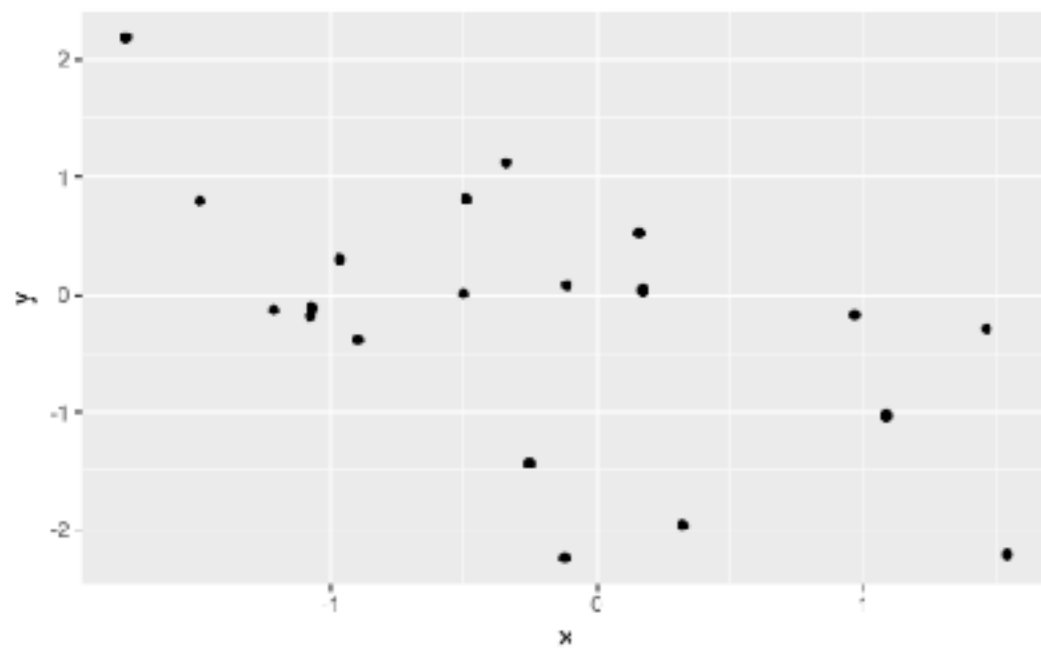
0.63

0.81

$r=0.63$



Guess the correlation coefficient for this dataset

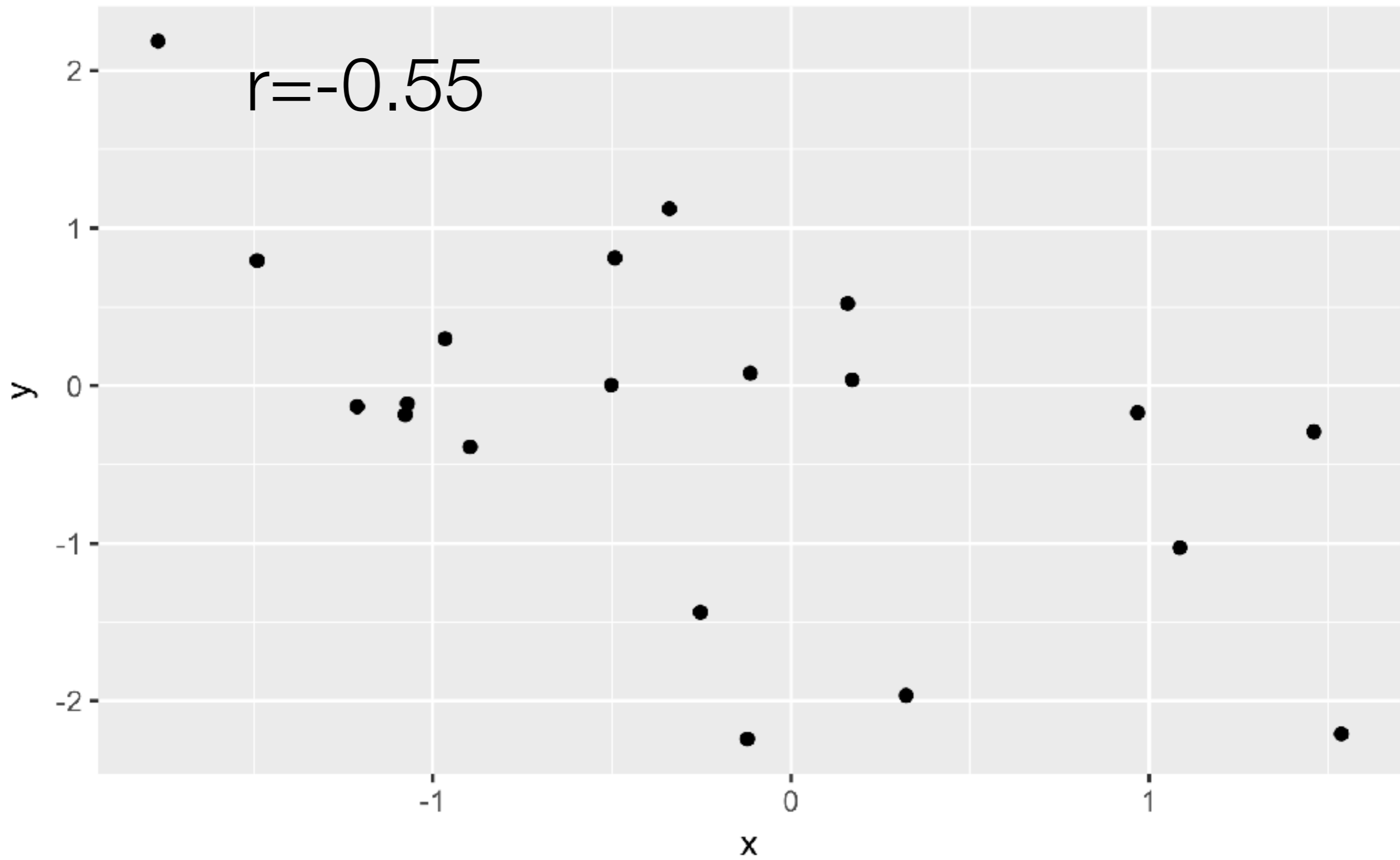


-0.71

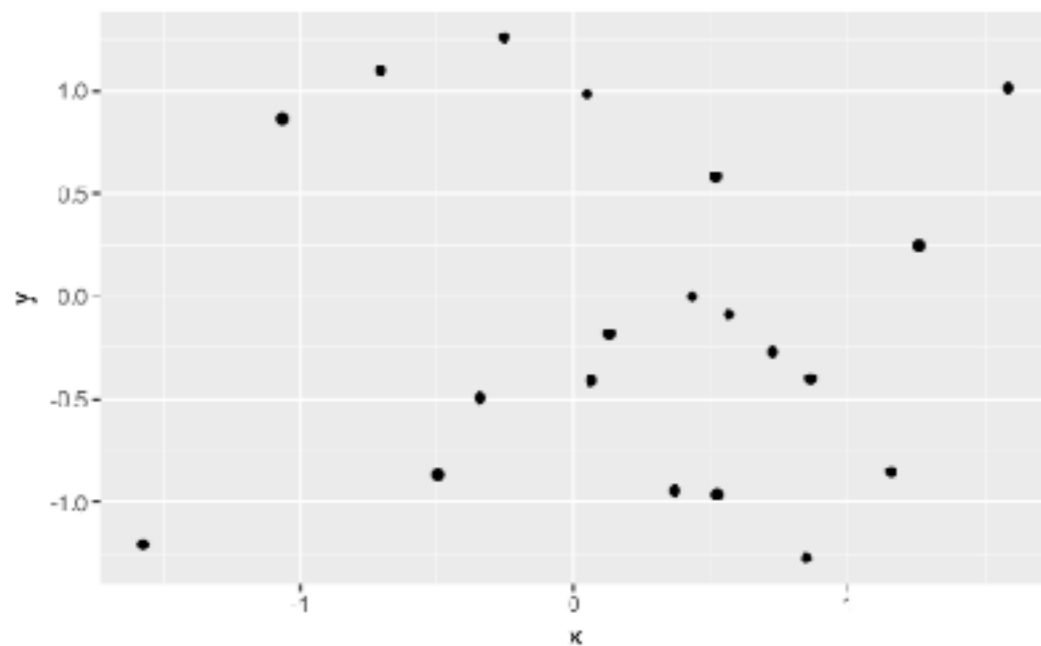
-0.55

-0.35

-0.12



Guess the correlation coefficient for this dataset

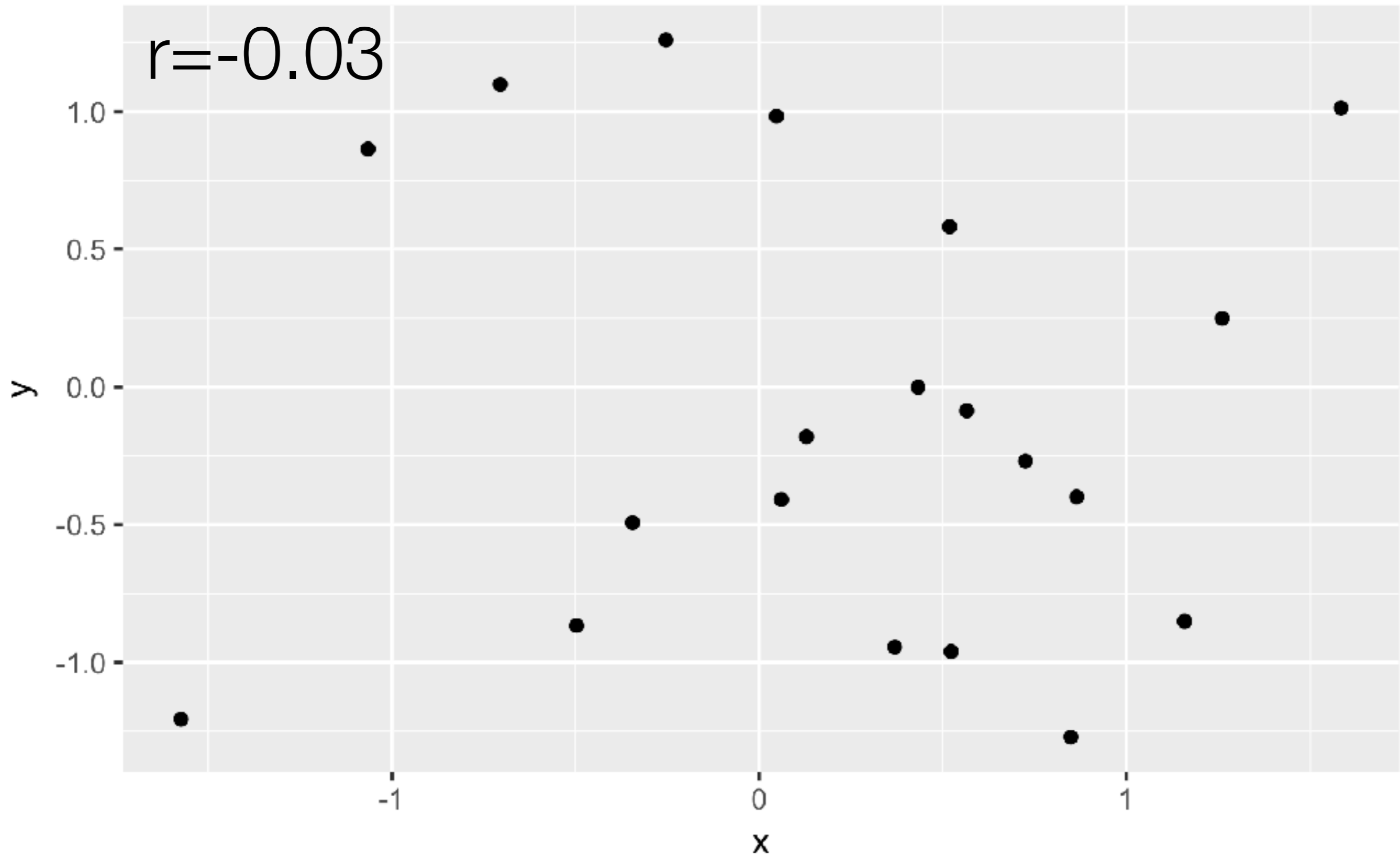


-0.28

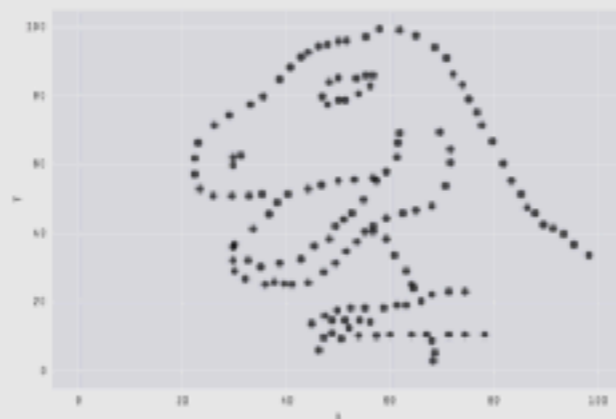
-0.03

0.25

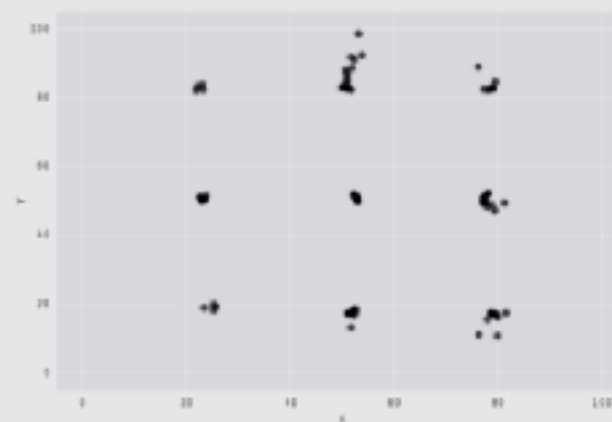
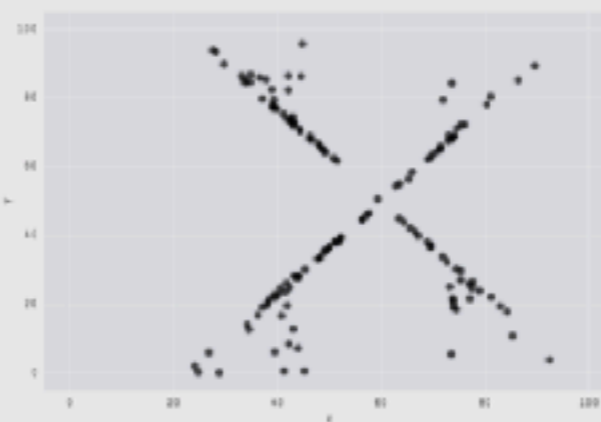
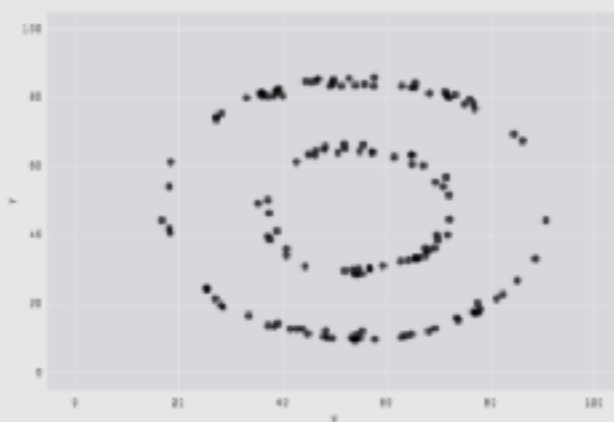
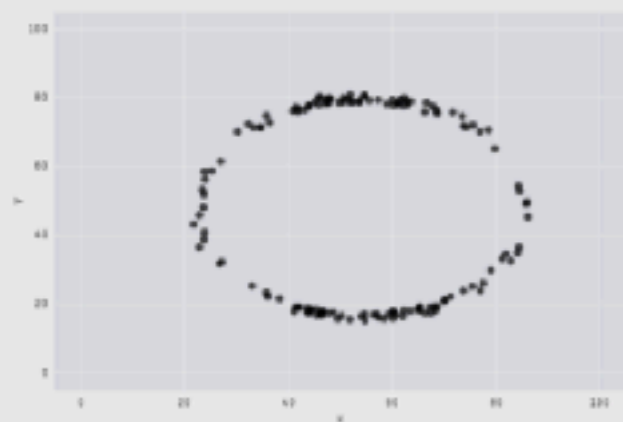
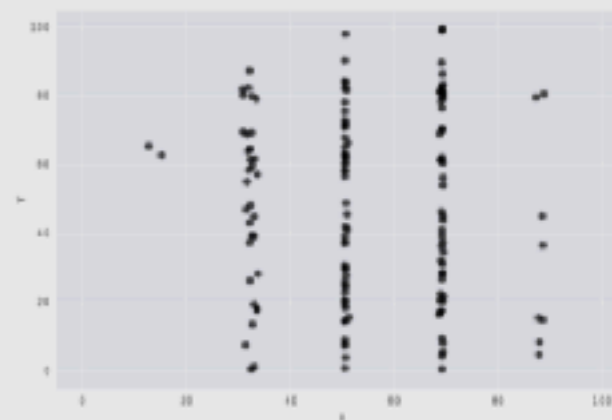
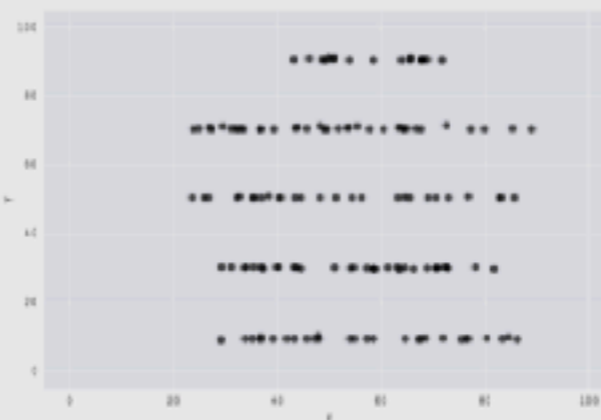
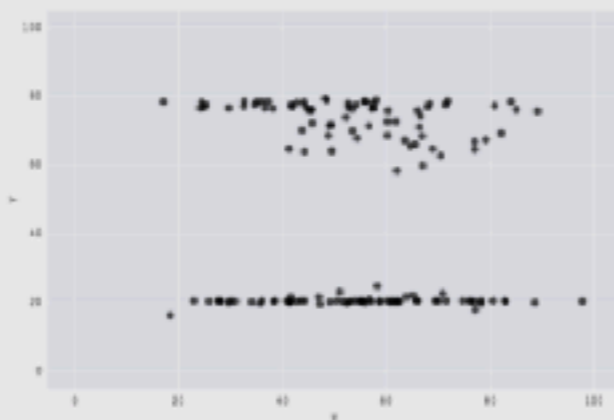
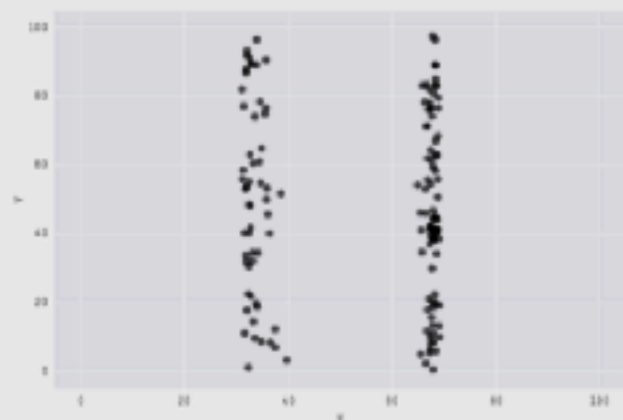
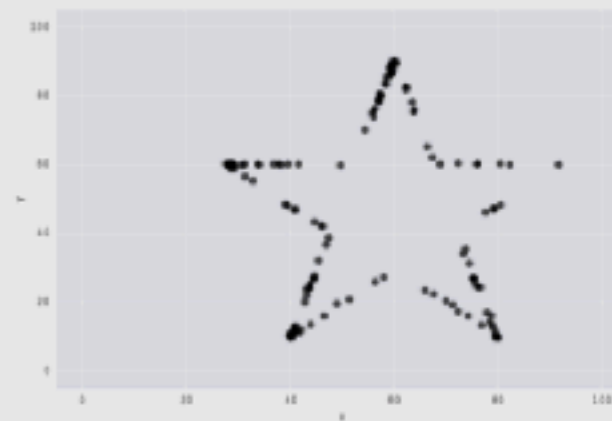
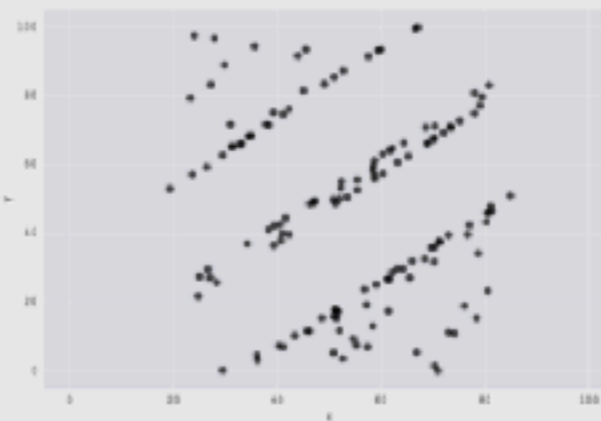
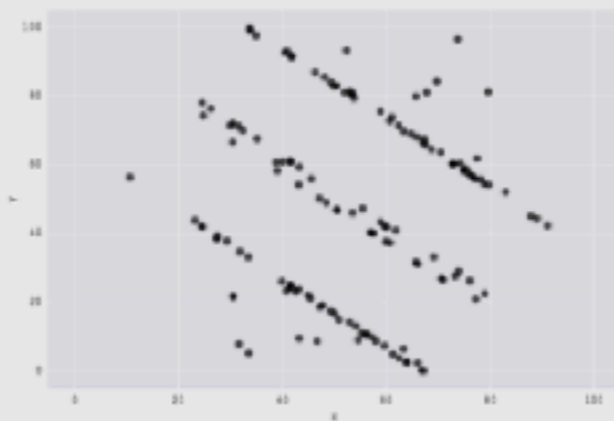
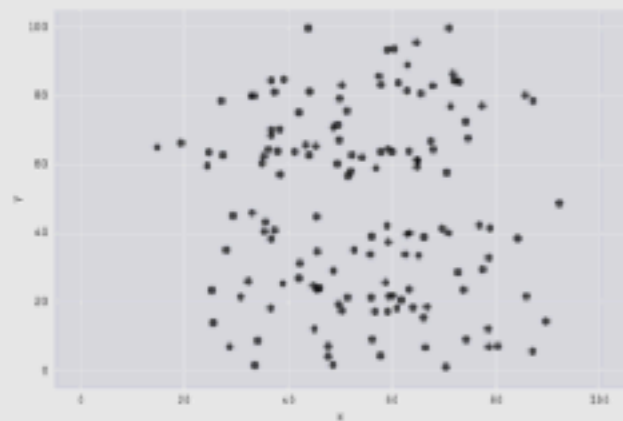
0.41



The “Datasaurus Dozen”



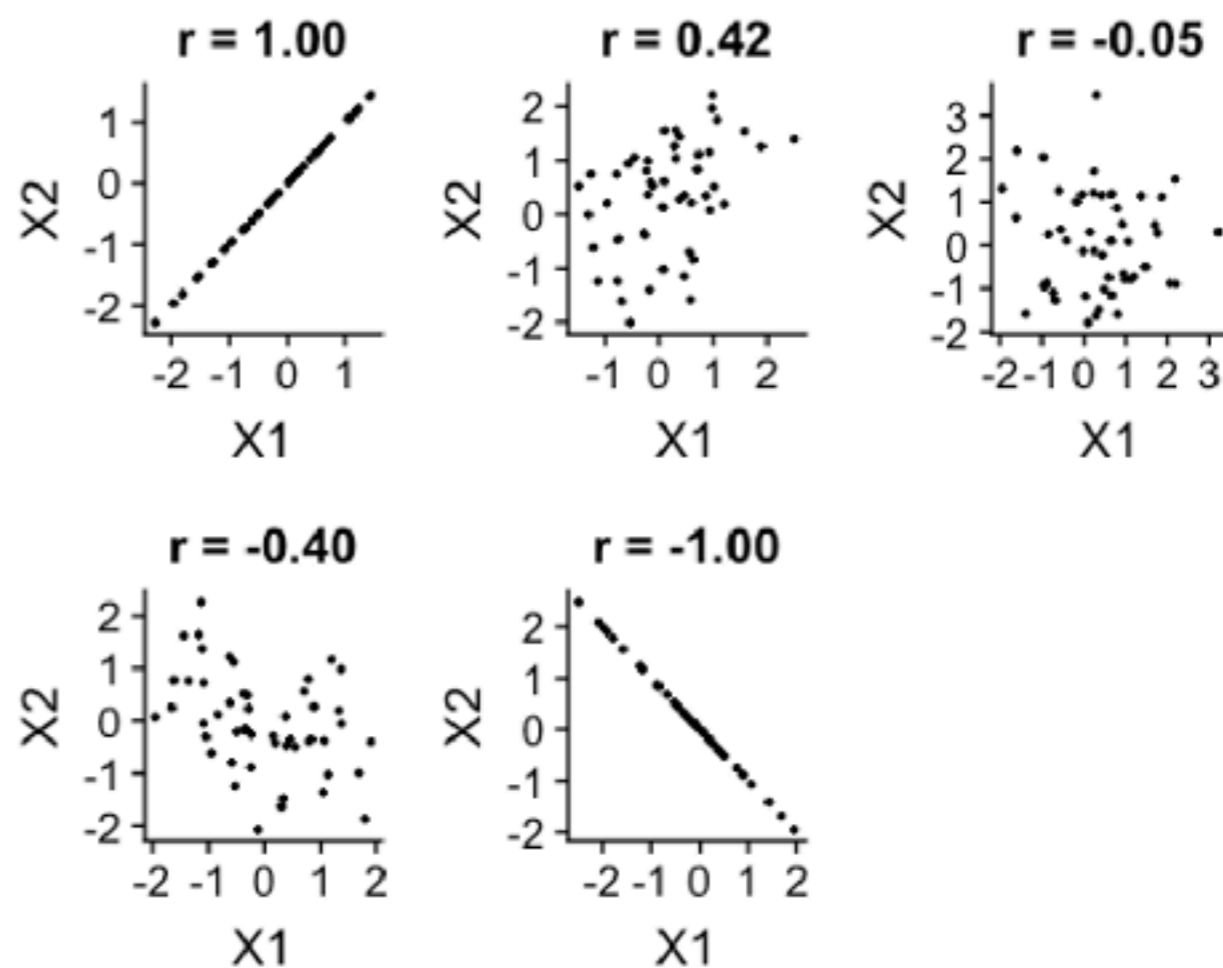
X Mean: 54.26
 Y Mean: 47.83
 X SD : 16.76
 Y SD : 26.93
 Corr. : -0.06



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.



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()`

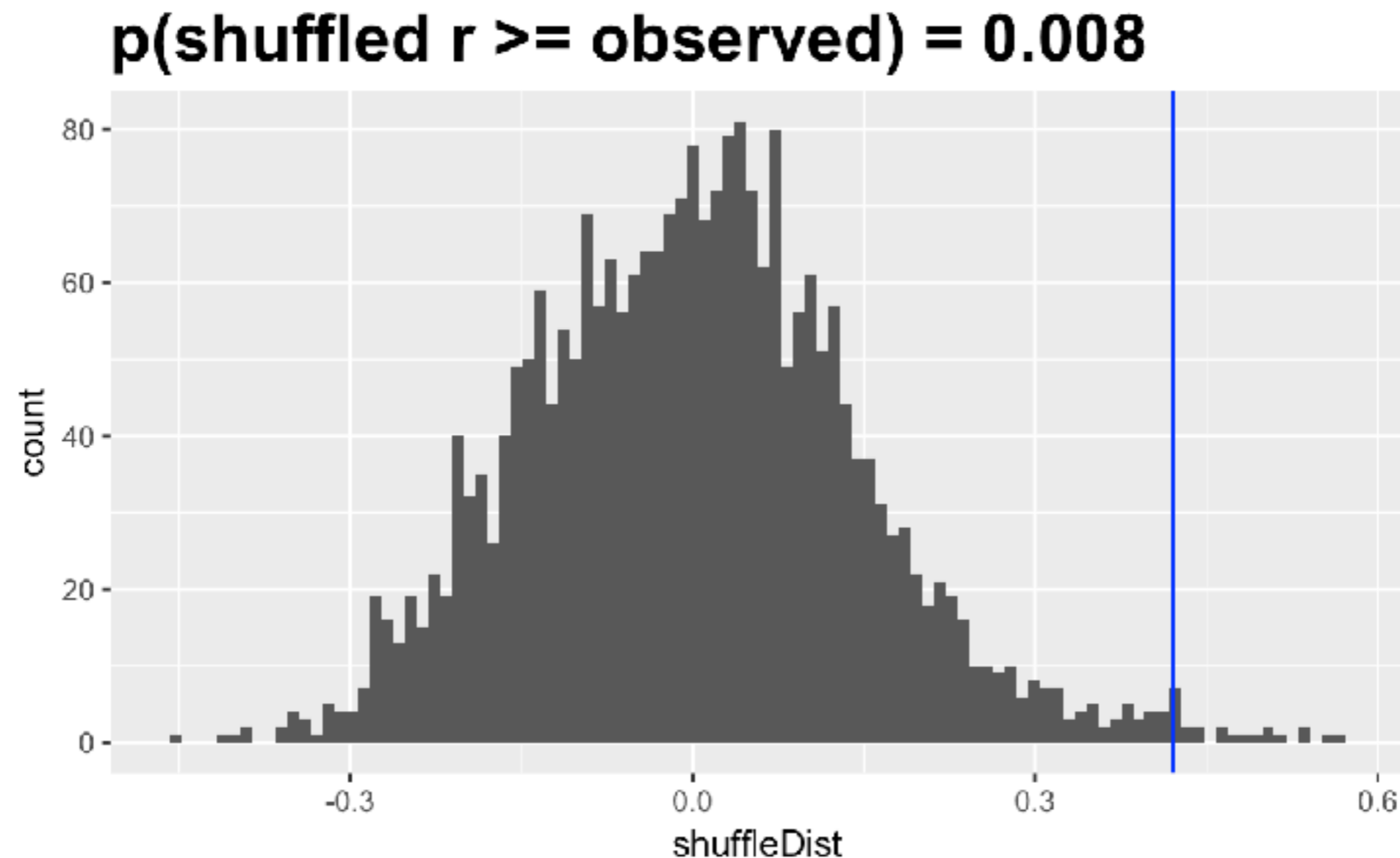
```
cor.test(hate_crimes$avg_hatecrimes_per_100k_fbi,  
         hate_crimes$gini_index, alternative='greater')
```

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
```

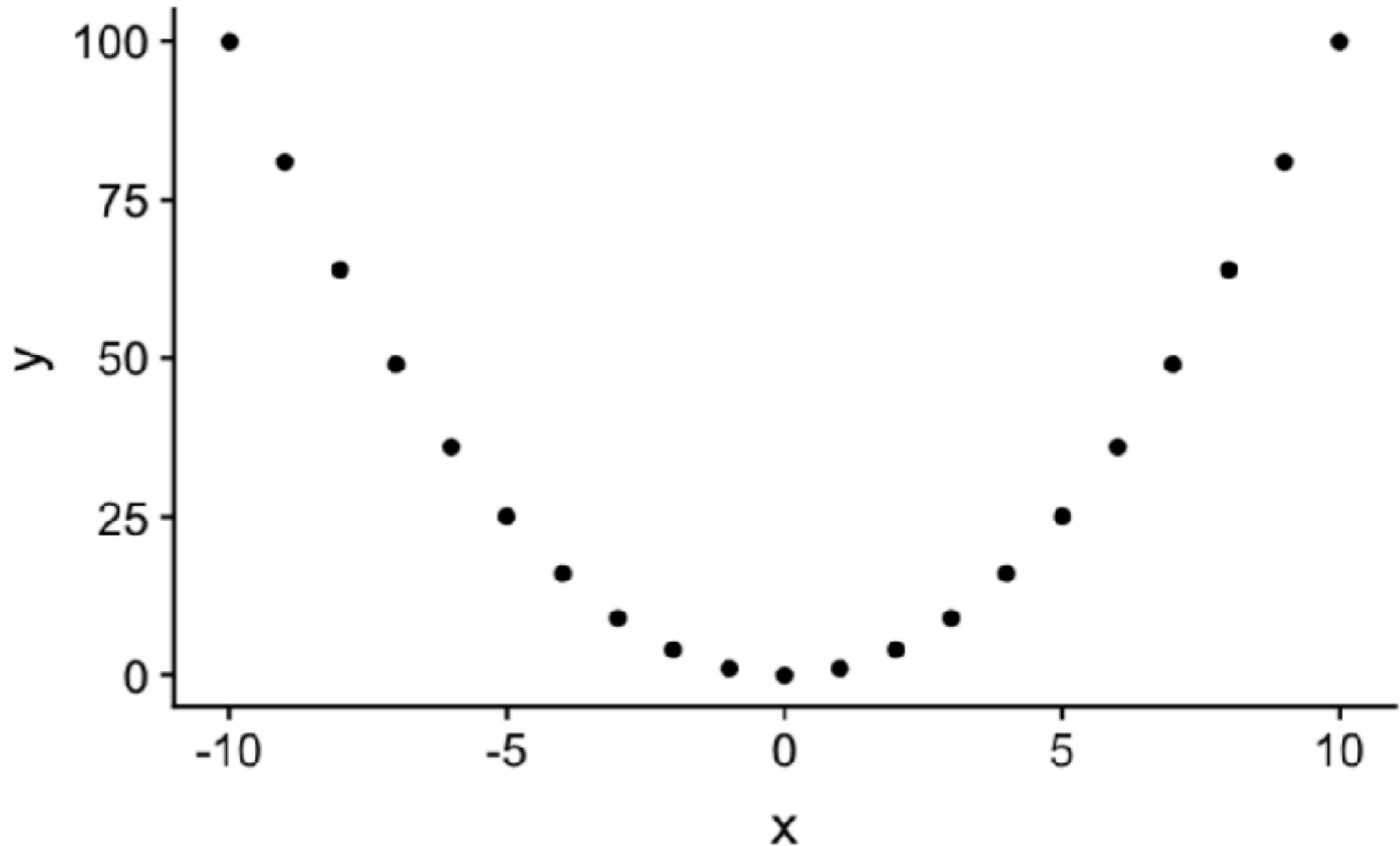
Randomization

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

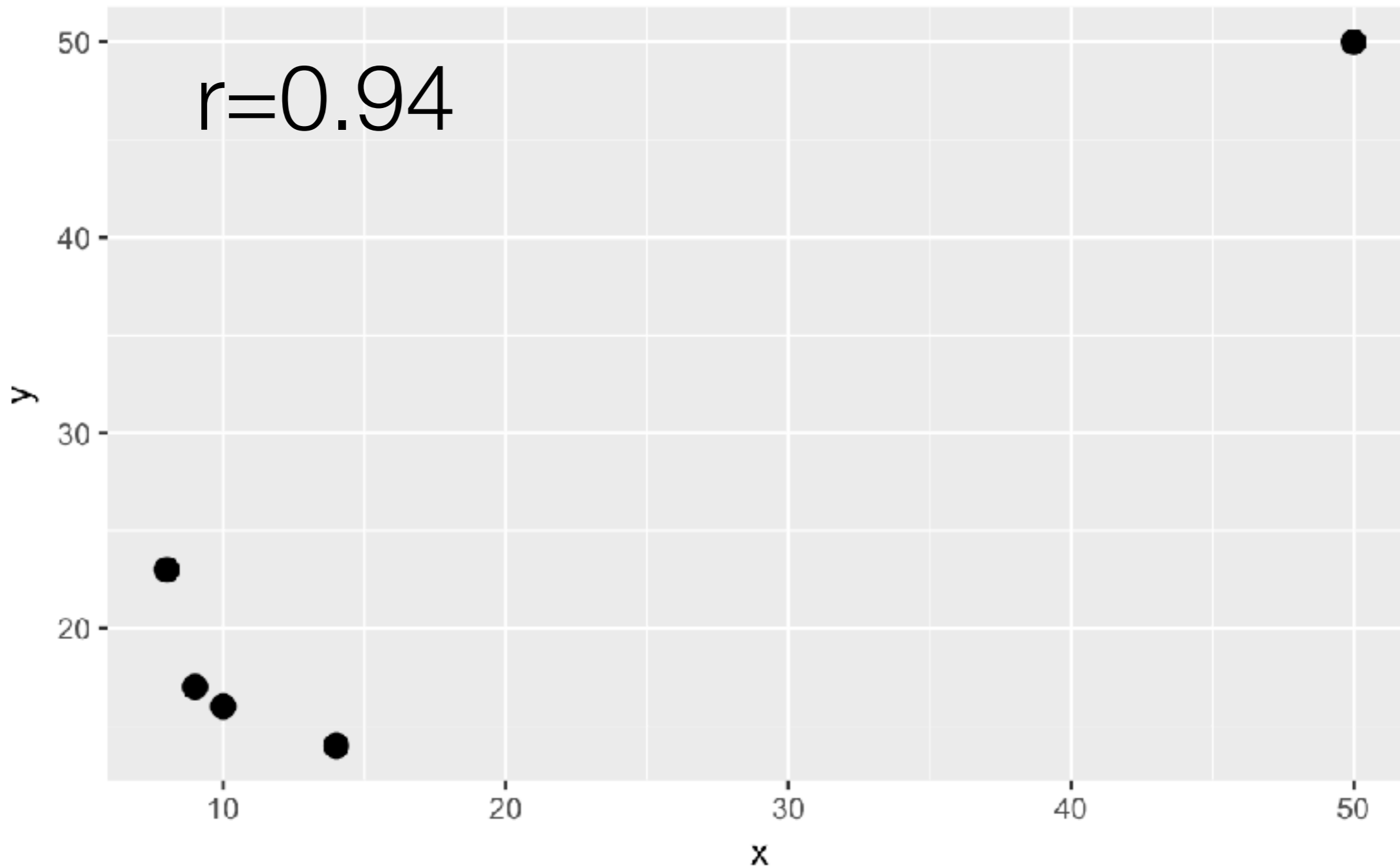


Correlation is only sensitive to linear relationships

$$y = x^2: \text{correlation} = 0$$



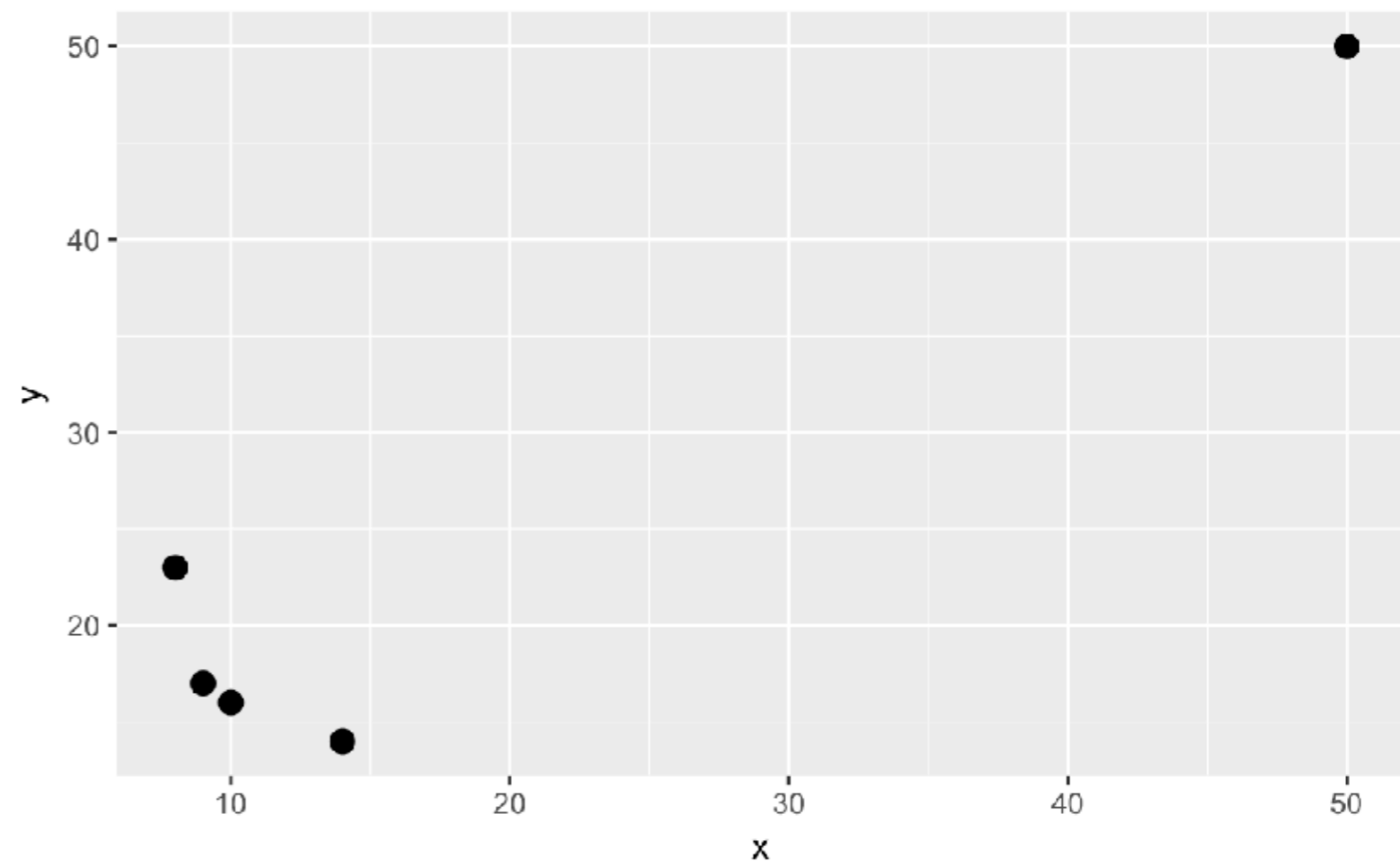
Correlation is very sensitive to outliers



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
```


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

FiveThirtyEight

Politics Sports Science & Health Economics Culture

JAN. 23, 2017 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.

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



Ed Cara

Wednesday 4:10pm • Filed to: ALCOHOL ▾

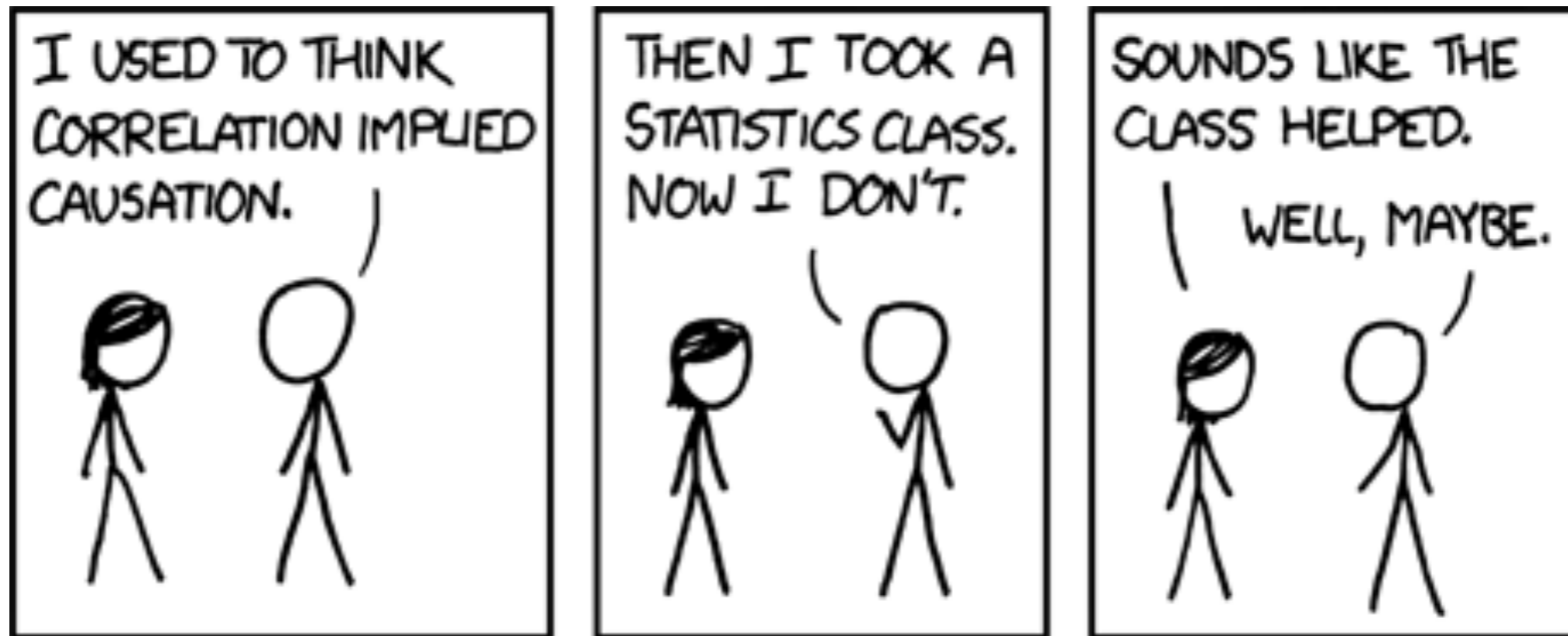
 45.3K  42  1

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.

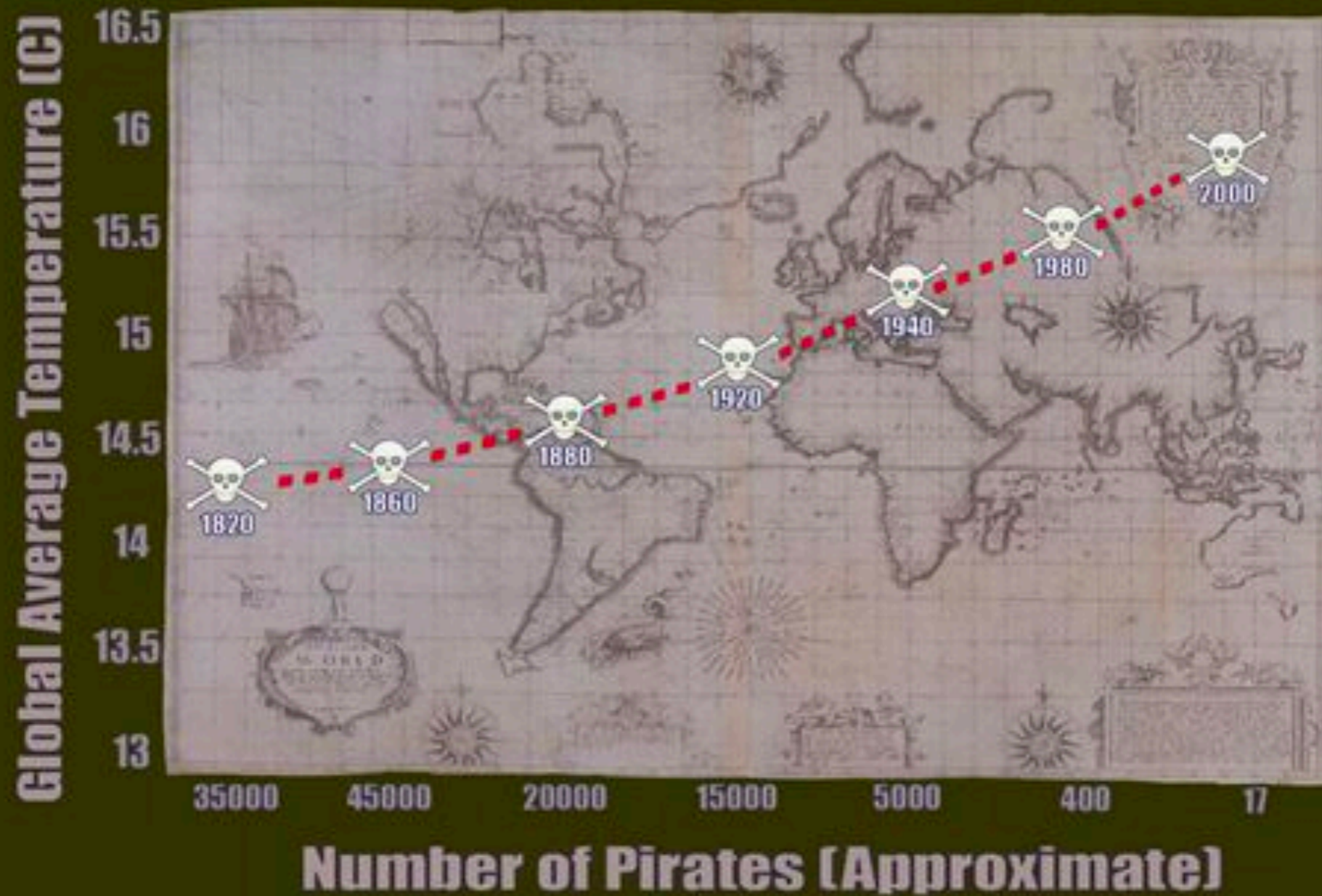
“If all these measures [increased alcohol taxes and advertising bans] are implemented widely, they could not only reduce dementia incidence or delay dementia onset, but also reduce all alcohol-attributable morbidity and mortality,” they wrote.

Correlation and causation



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

Global Temperature Vs. Number of Pirates



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

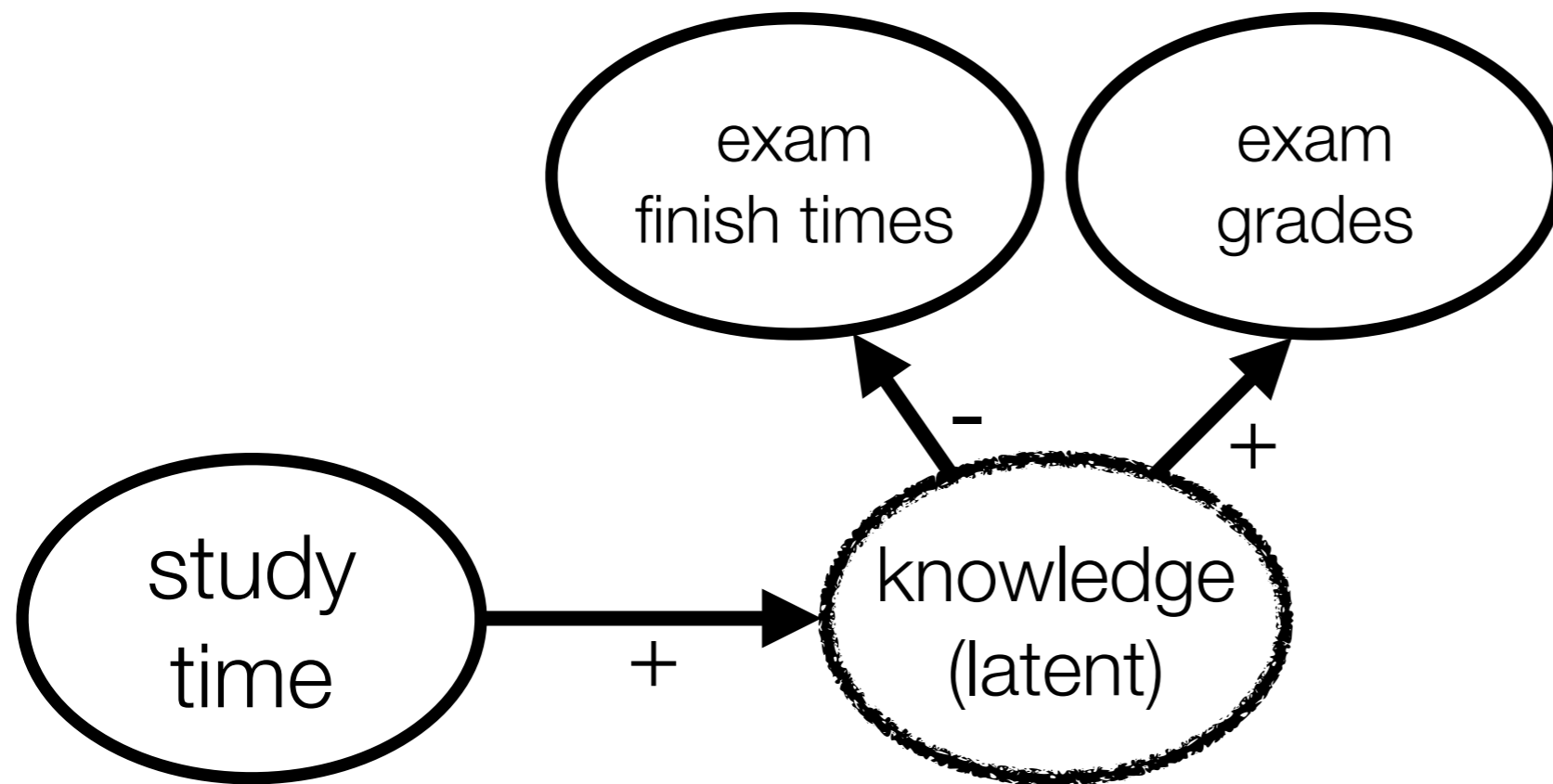
<http://www.tylervigen.com/spurious-correlations>

“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



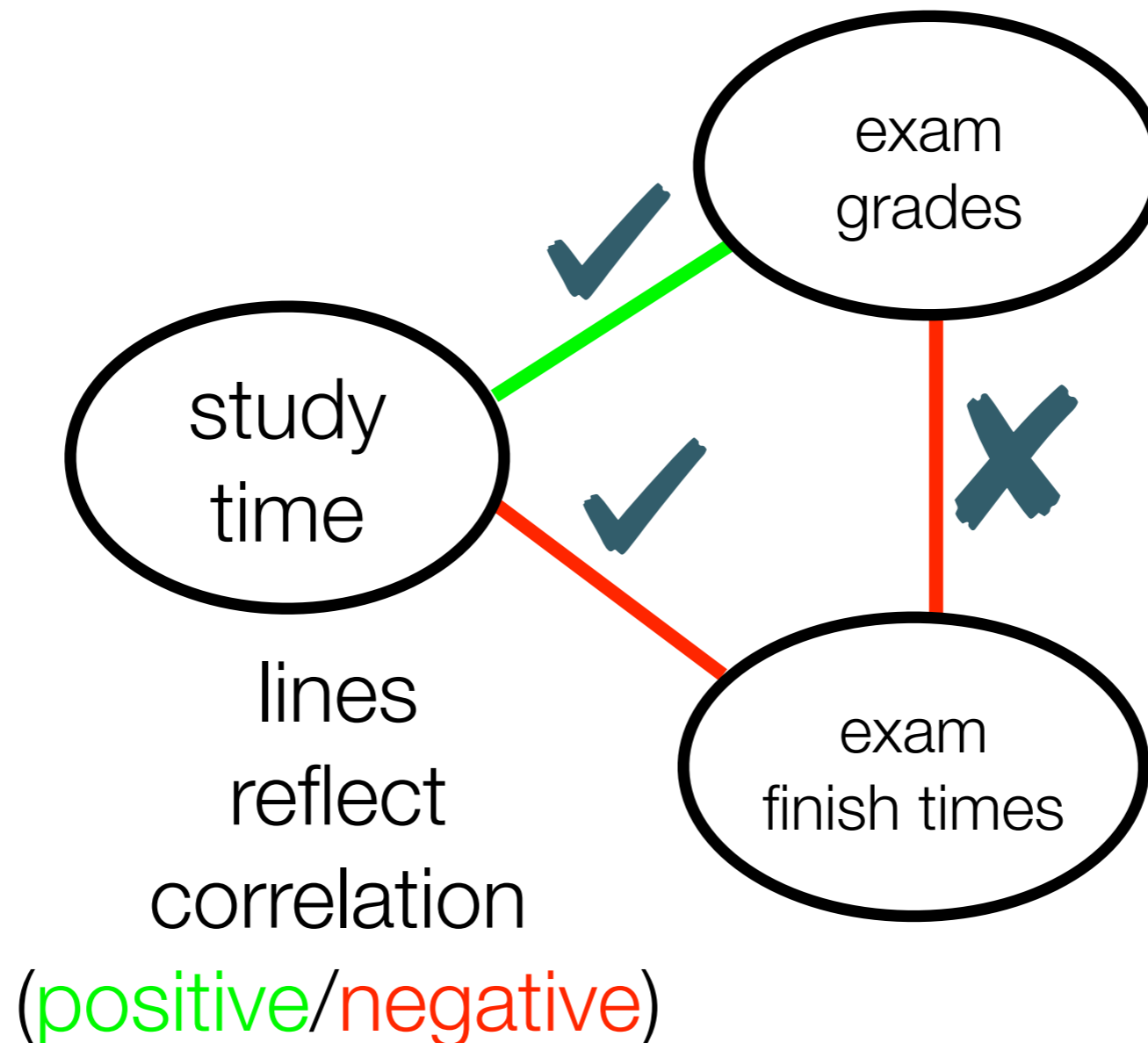
arrows reflect
causal relations

Causal relations mean
that manipulating one
variable will change
another

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

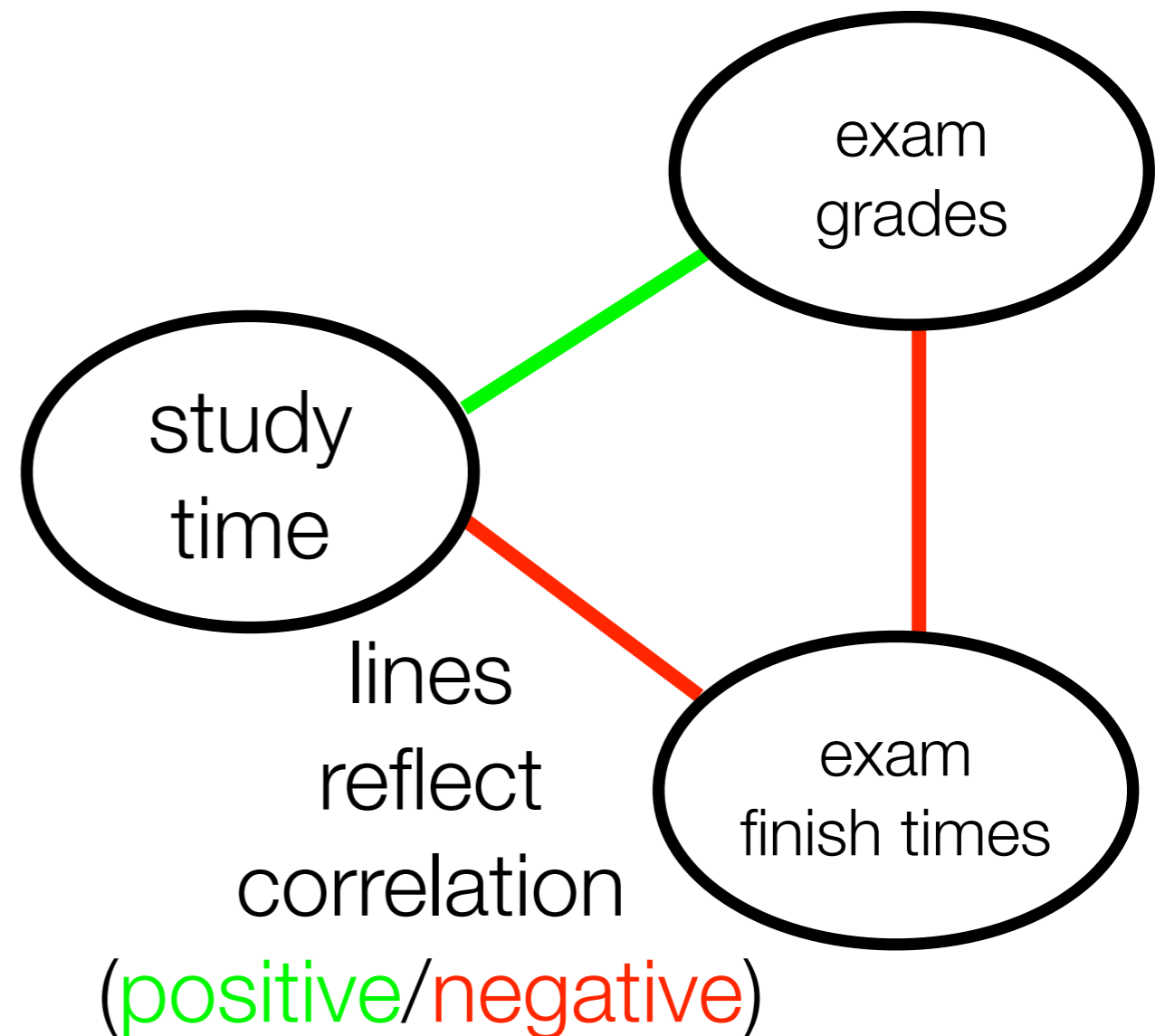
Correlation and causation

- 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!

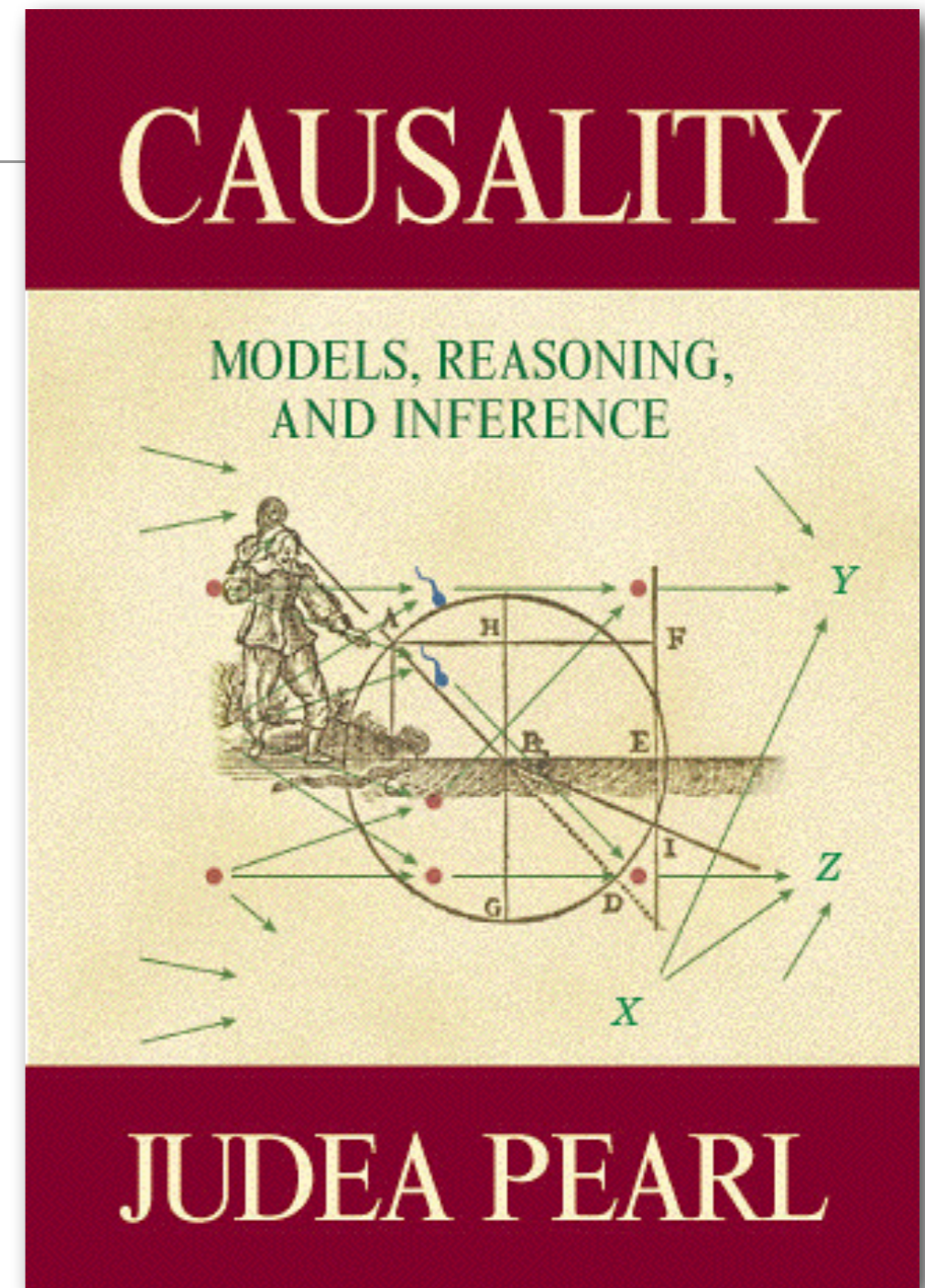


Group discussion

- 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