



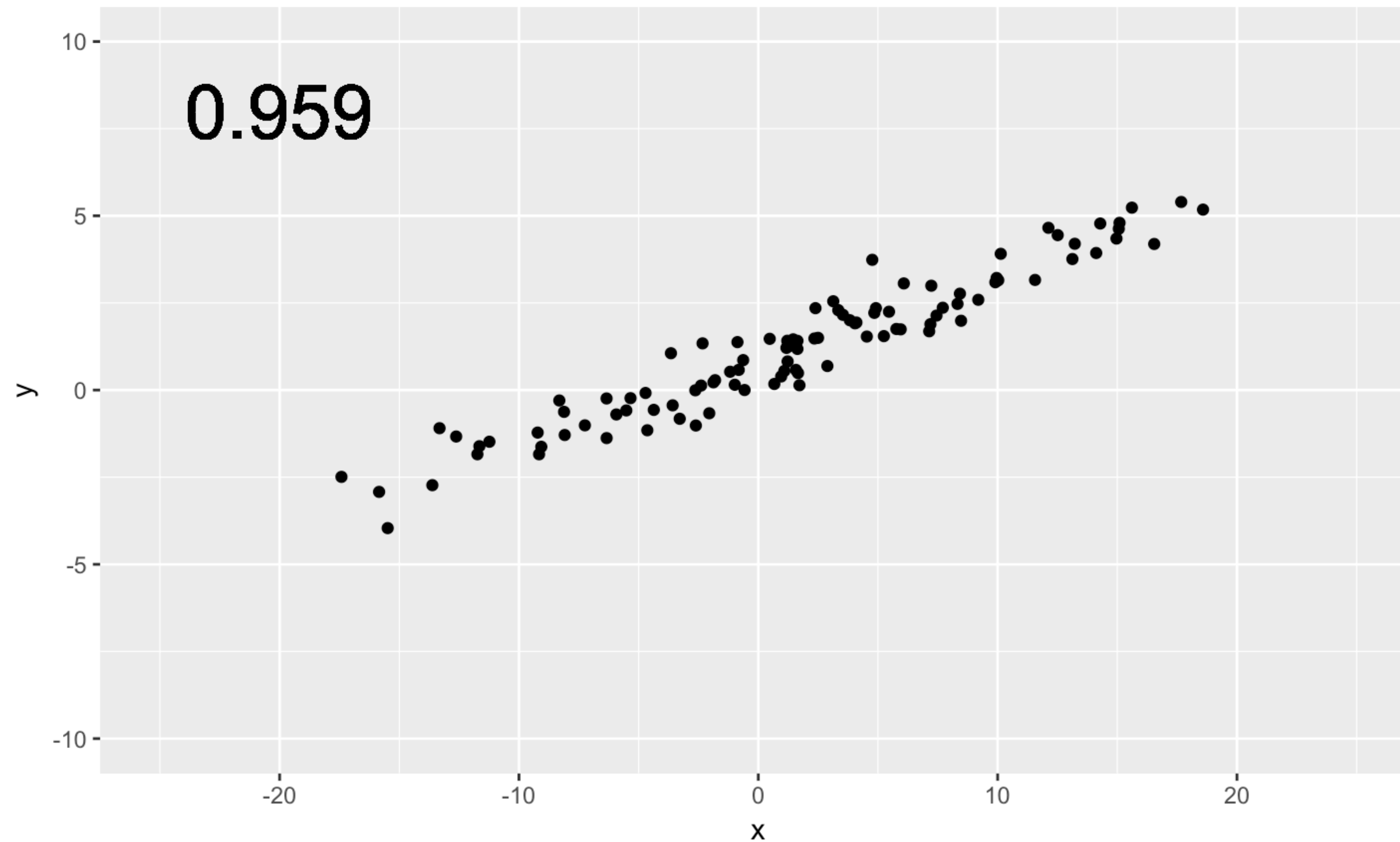
CORRELATION AND REGRESSION

Correlation

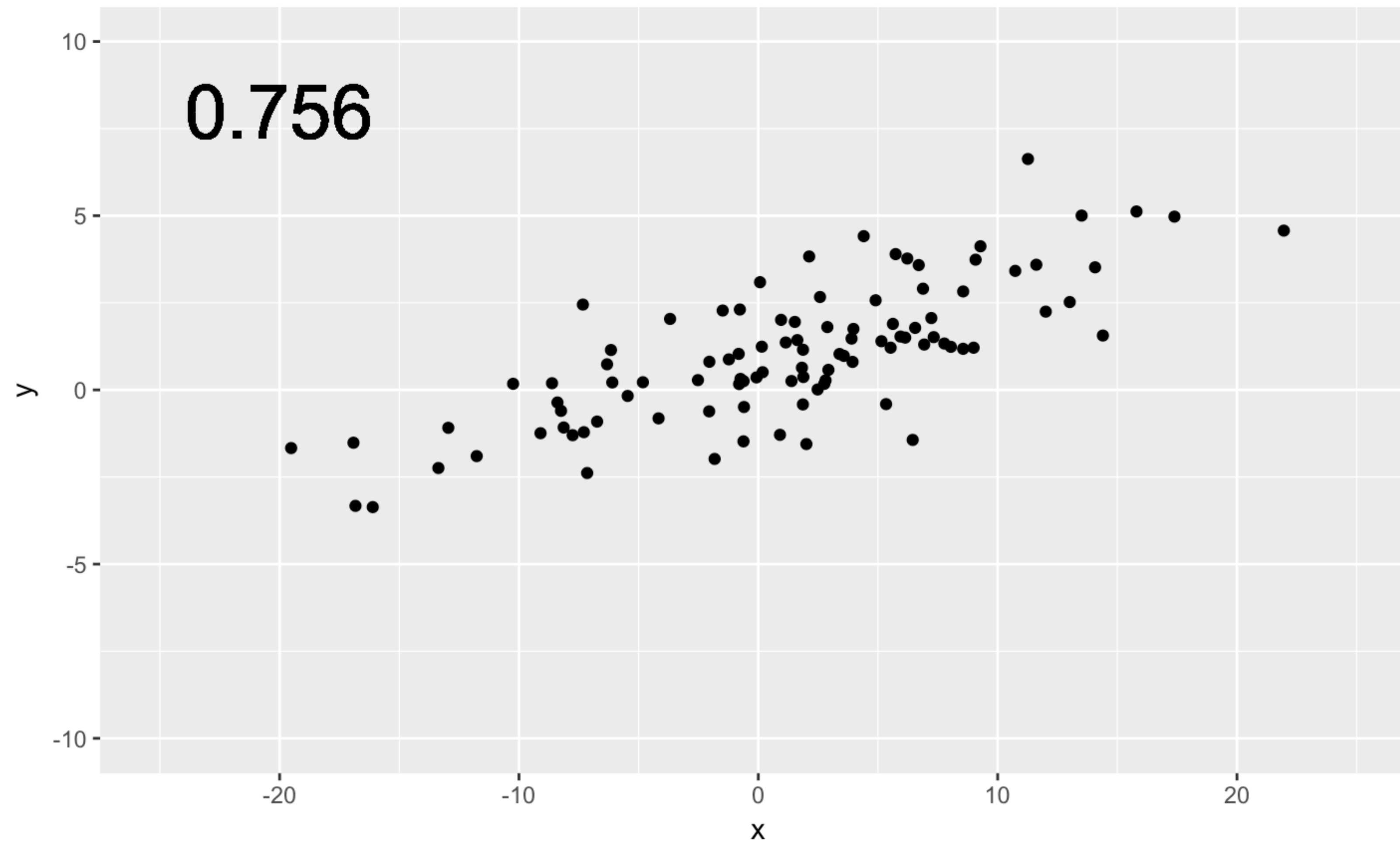
Correlation

- Correlation coefficient between -1 and 1
- Sign \longrightarrow direction
- Magnitude \longrightarrow strength

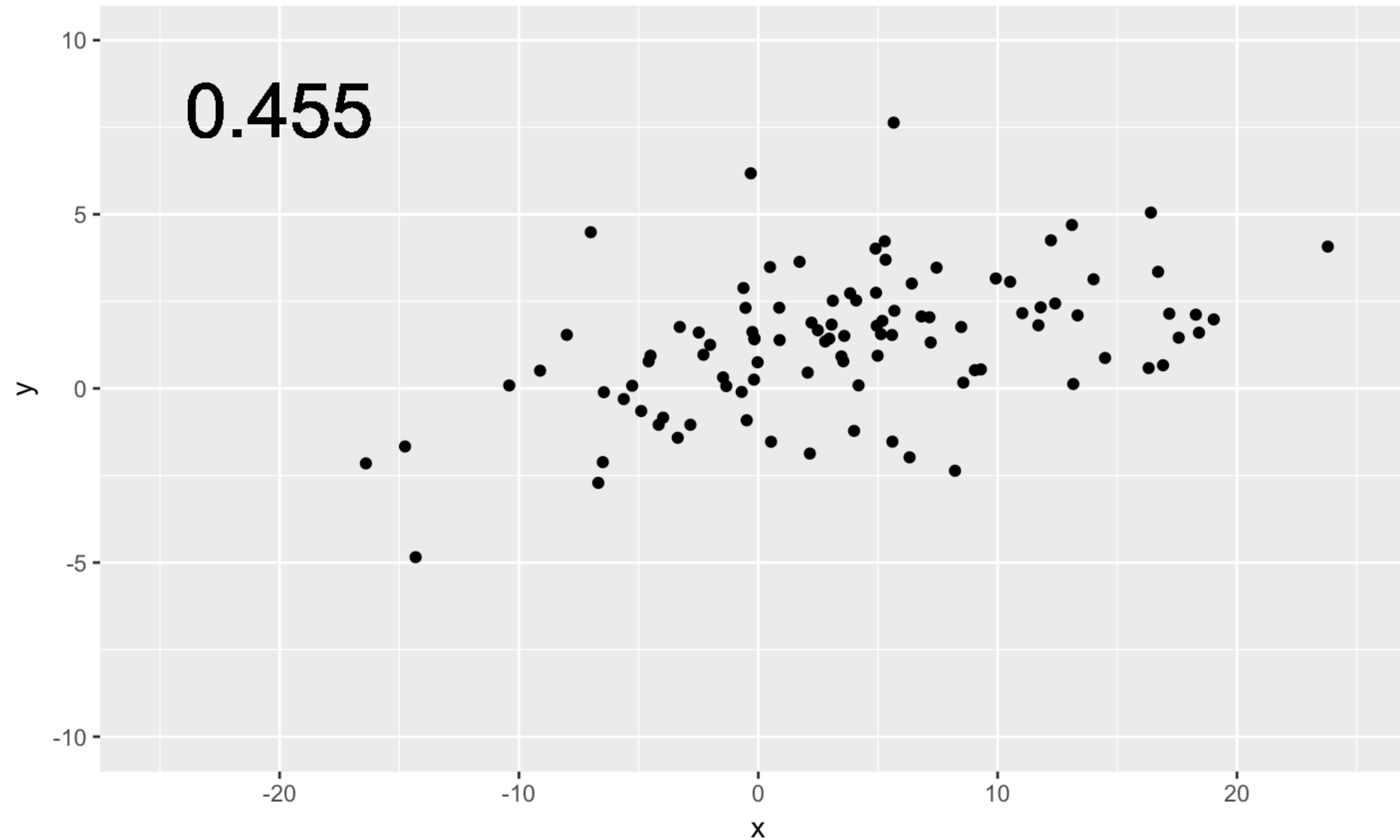
Near perfect correlation



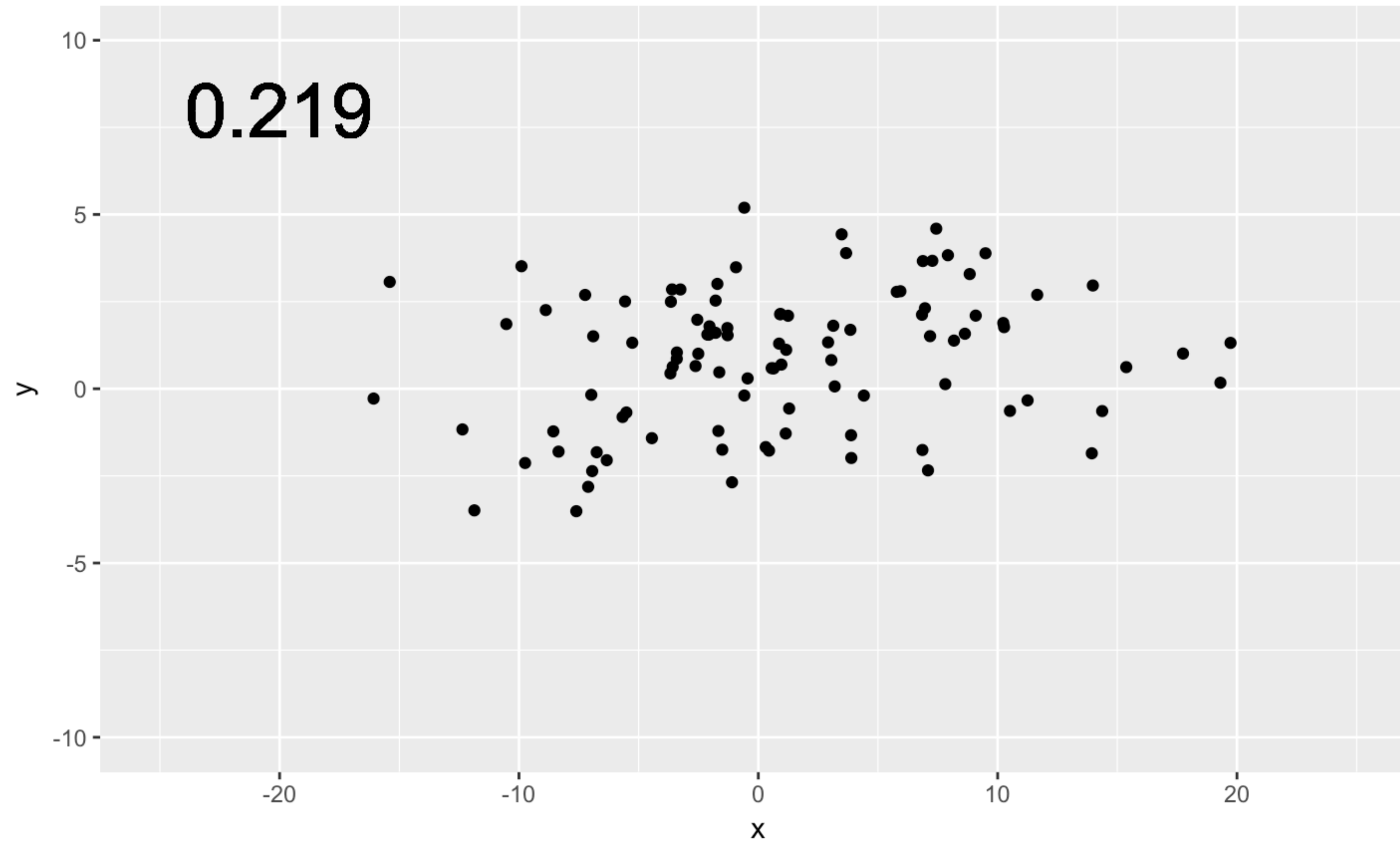
Strong



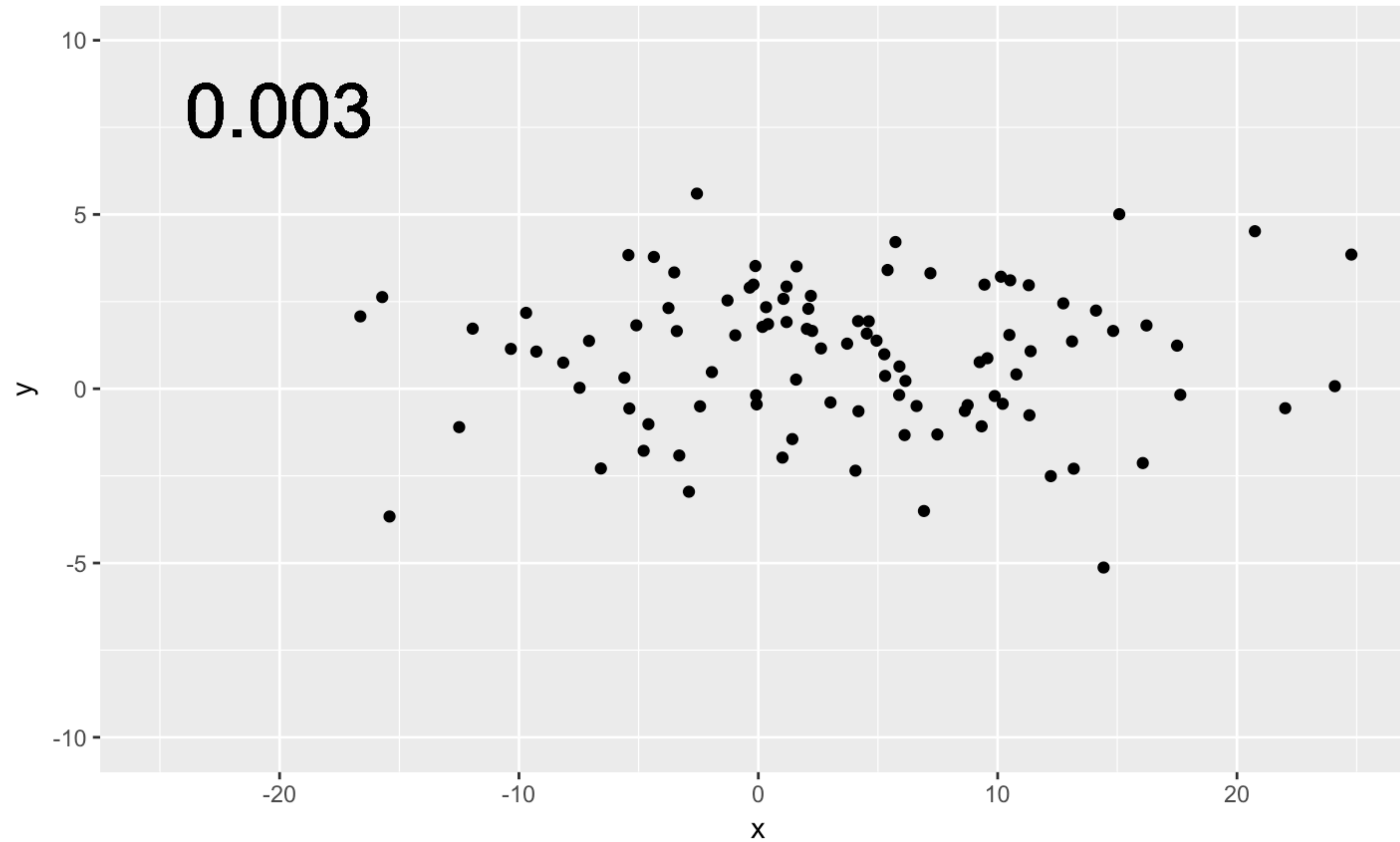
Moderate



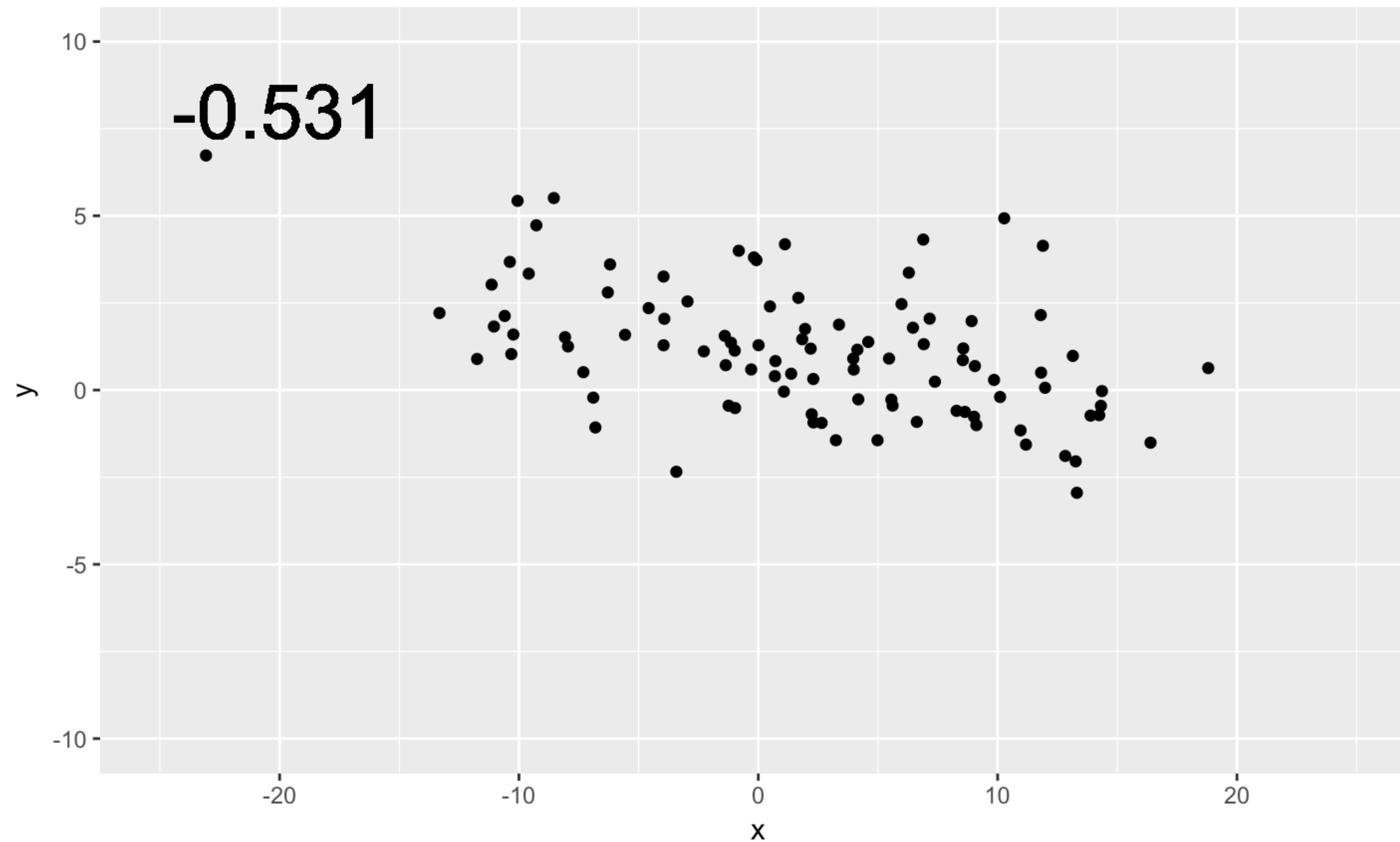
Weak



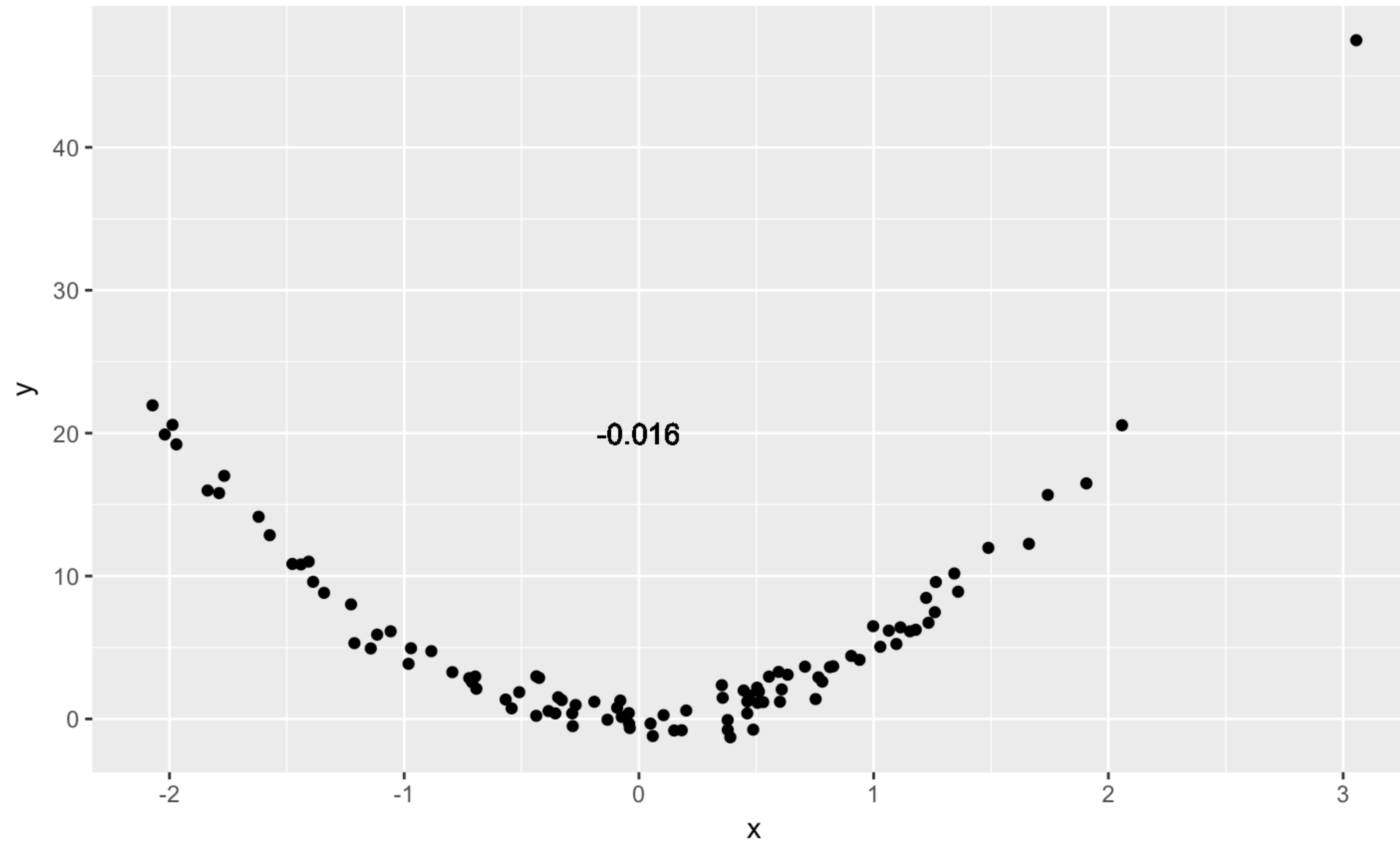
Zero



Negative

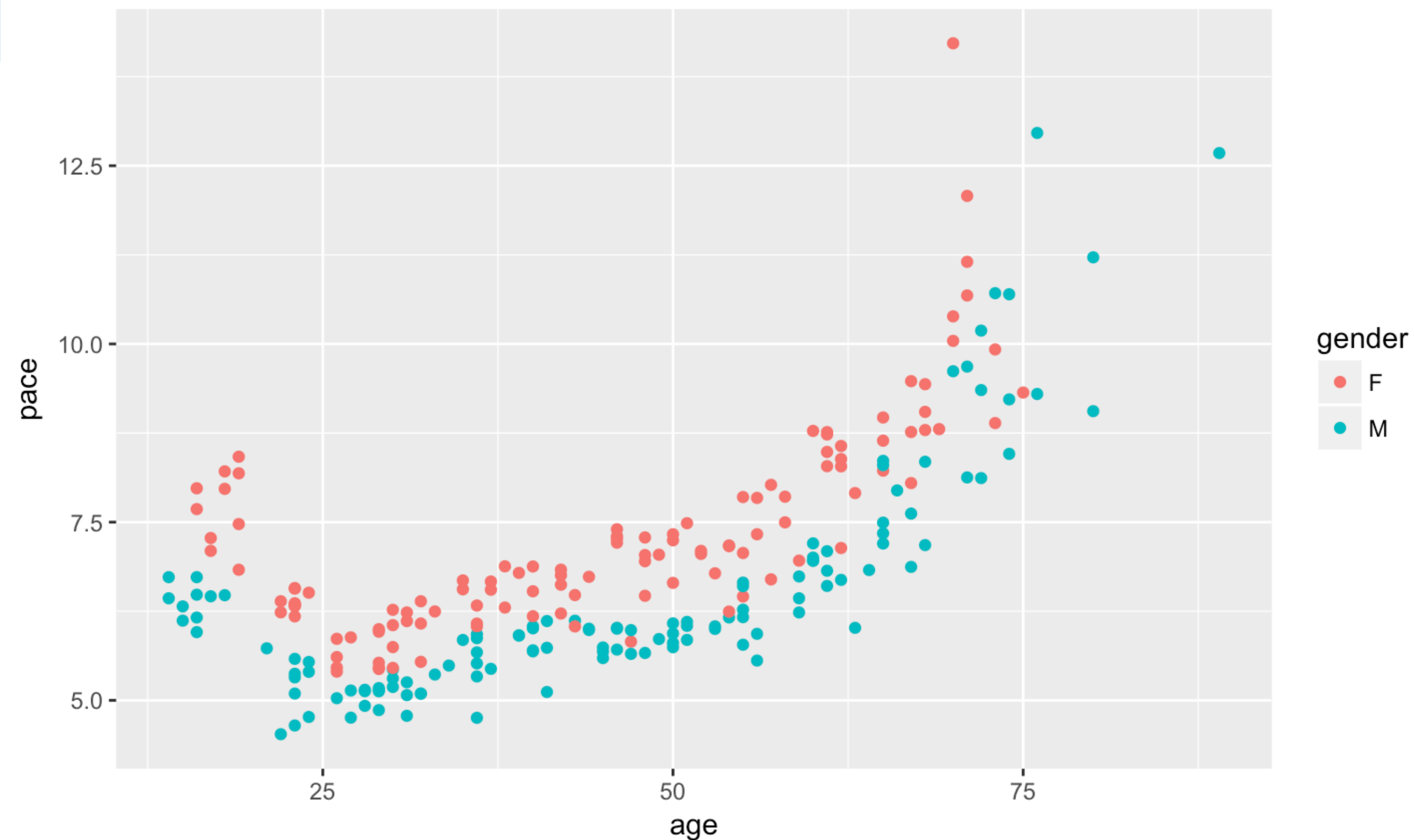


Non-linear



Non-linear correlation

```
> run10 %>%  
  filter(divPlace <= 10) %>%  
  ggplot(aes(x = age, y = pace, color = gender)) +  
  geom_point()
```



Pearson product-moment correlation

$$r(x, y) = \frac{Cov(x, y)}{\sqrt{SXX \cdot SY Y}}$$

Pearson product-moment correlation

$$r(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}$$



CORRELATION AND REGRESSION

Let's practice!

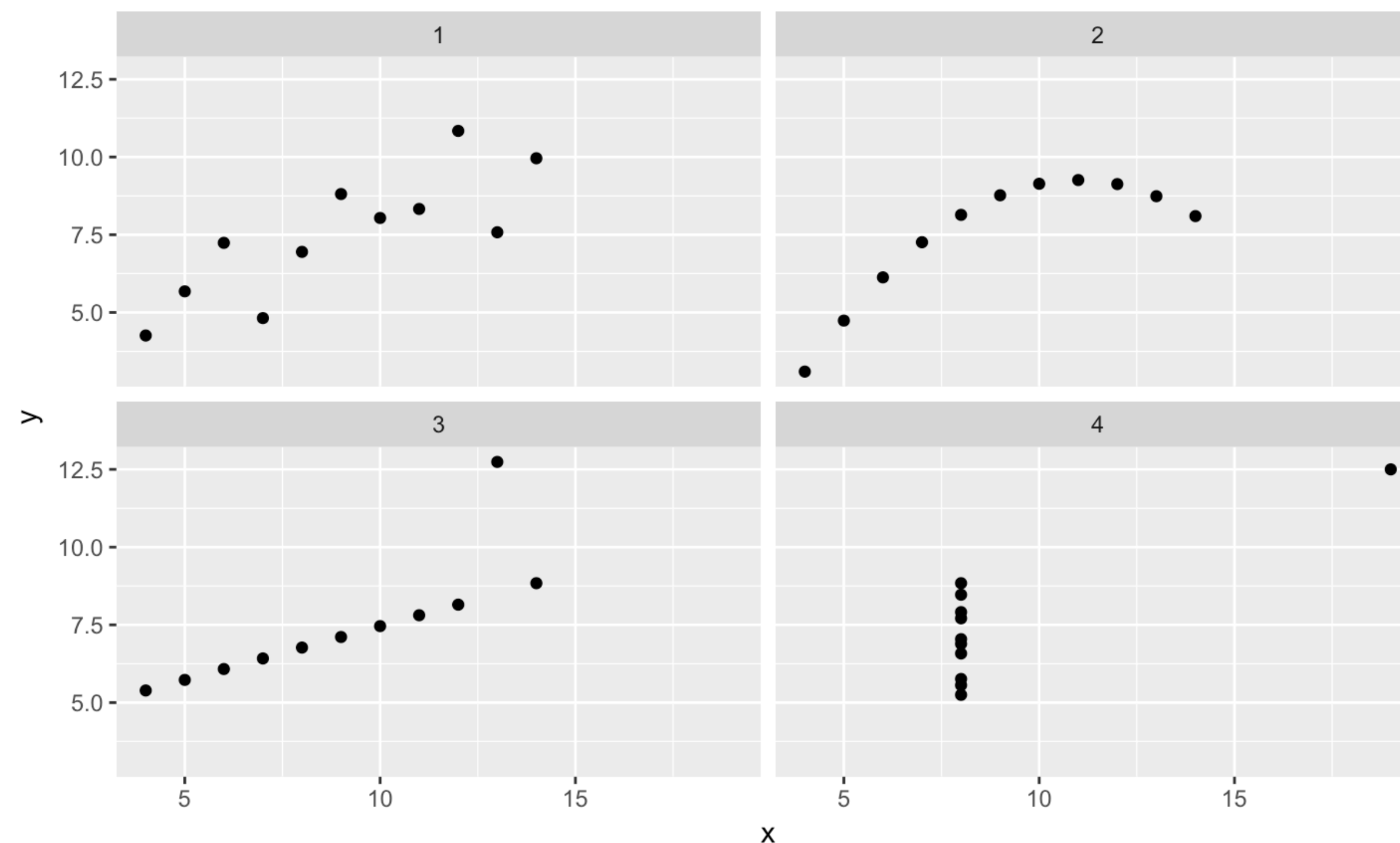


CORRELATION AND REGRESSION

Correlation

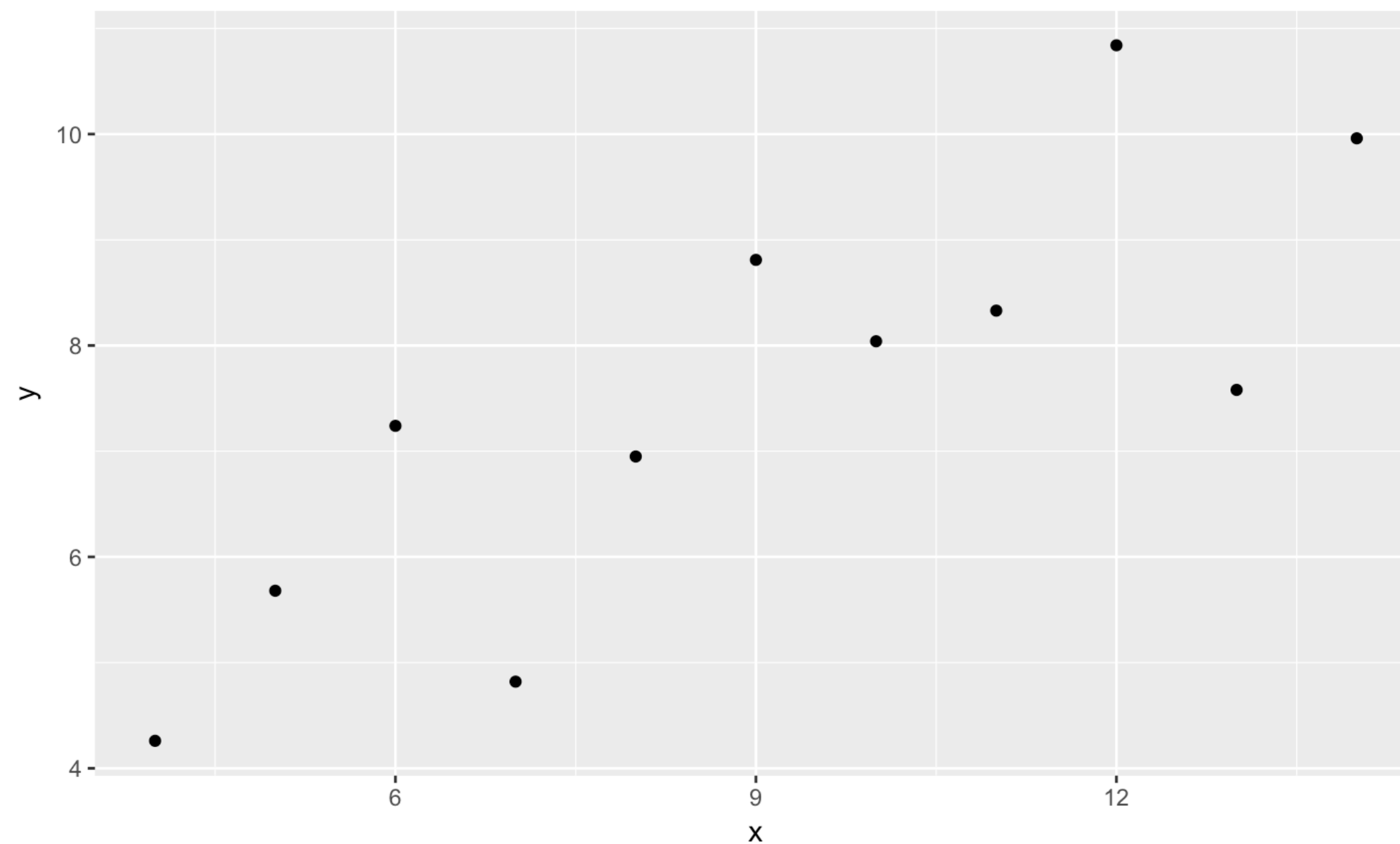
Anscombe

```
> ggplot(data = Anscombe, aes(x = x, y = y)) +  
  geom_point() +  
  facet_wrap(~ set)
```



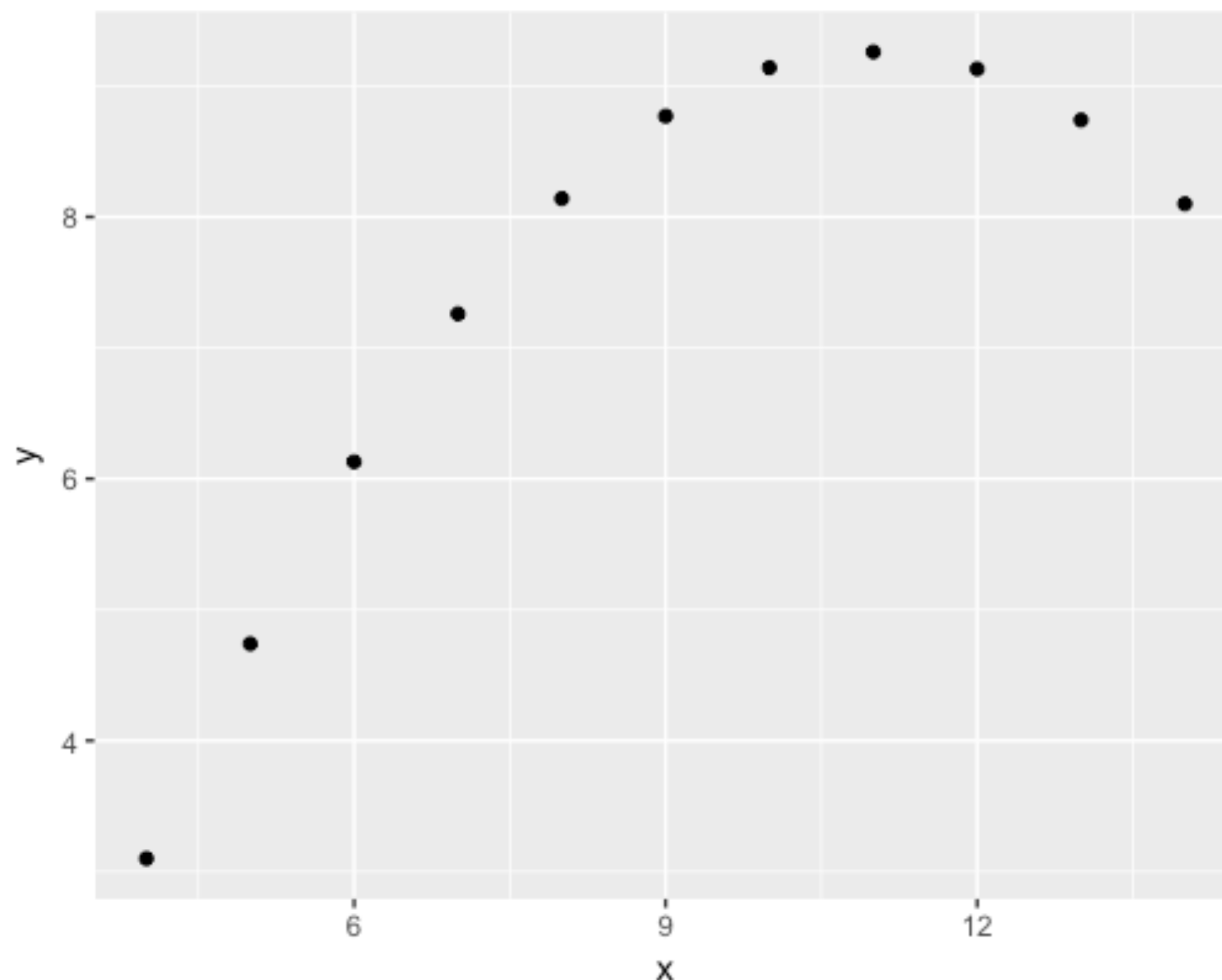
Anscombe 1

```
> Anscombe %>%  
  filter(set == 1) %>%  
  ggplot(aes(x = x, y = y)) +  
  geom_point()
```



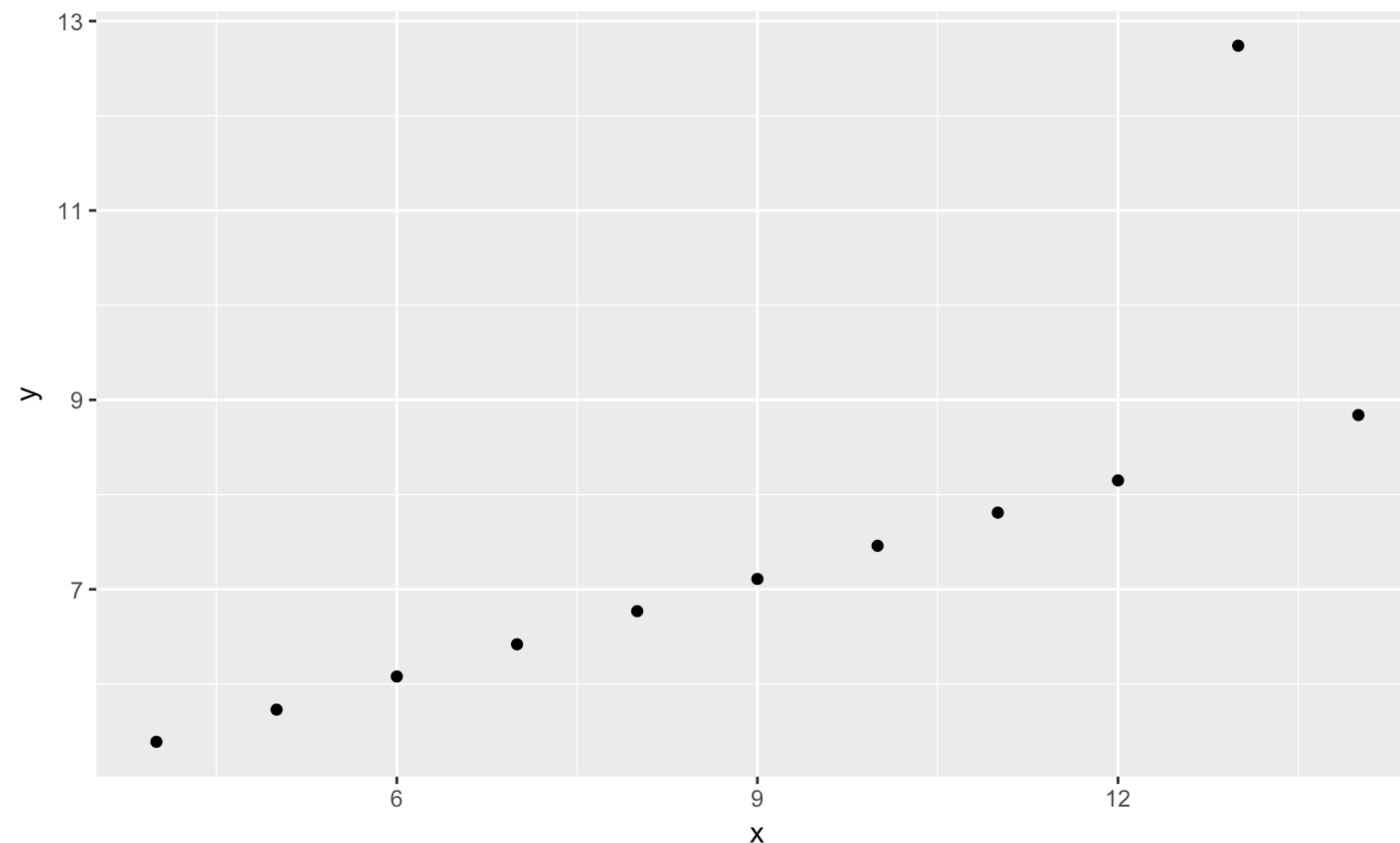
Anscombe 2

```
> Anscombe %>%  
  filter(set == 2) %>%  
  ggplot(aes(x = x, y = y)) +  
  geom_point()
```



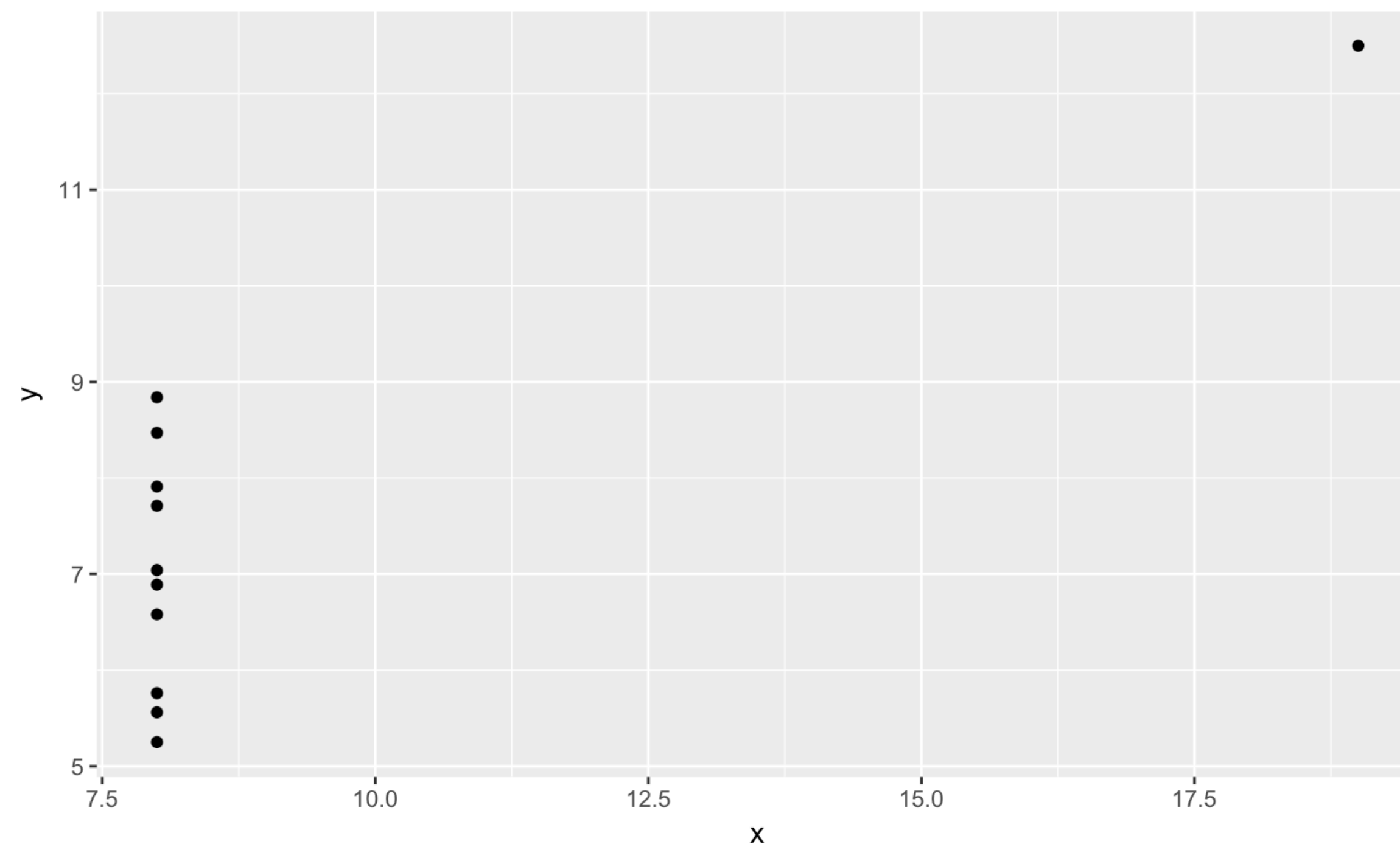
Anscombe 3

```
> Anscombe %>%  
  filter(set == 3) %>%  
  ggplot(aes(x = x, y = y)) +  
  geom_point()
```



Anscombe 4

```
> Anscombe %>%  
  filter(set == 4) %>%  
  ggplot(aes(x = x, y = y)) +  
  geom_point()
```





CORRELATION AND REGRESSION

Let's practice!



CORRELATION AND REGRESSION

Interpretation of Correlation

Exercise and beer



Getty Images

Email

Share

Tweet

Save

More

For many people, working out and alcohol are closely linked. Sports teams and training partners celebrate victories, bemoan defeats or mark the end of training sessions with a beer or three. Beer, in fact, provides a substantial portion of some exercisers' fluid intake after workouts.

But whether exercise encourages people to drink and, likewise, whether drinking encourages people to exercise has been in dispute.

Now two new studies suggest that exercise may well influence when and how much people drink. Drinking may even affect whether people exercise, and, the findings suggest, the interplay between

Exercise and beer

Past epidemiological studies have shown that people who exercise tend numerically also to be people who drink, and vice versa. In a [typical study from 2001](#), for example, researchers found that men and women who qualified as moderate drinkers, meaning they downed about a drink a day, were twice as likely to exercise regularly as teetotalers.

PHYS ED

Gretchen Reynolds on the science of fitness.



But most of these previous studies had limitations. They relied, for instance, on people's ability to recall their exercise and drinking habits over the course of, say, the past year, which can be notoriously unreliable. They also rarely took into account participants' ages and gender, which affect how much people exercise and drink.

And perhaps most problematic, these past studies rarely

Exercise and beer

write in their study, which was published recently in Health Psychology, “people drank more than usual on the same days that they engaged in more physical activity than usual.”

This relationship held true throughout all seasons of the year and whether someone was a man or a woman, a collegian or a retiree. Age and gender did not affect the results.

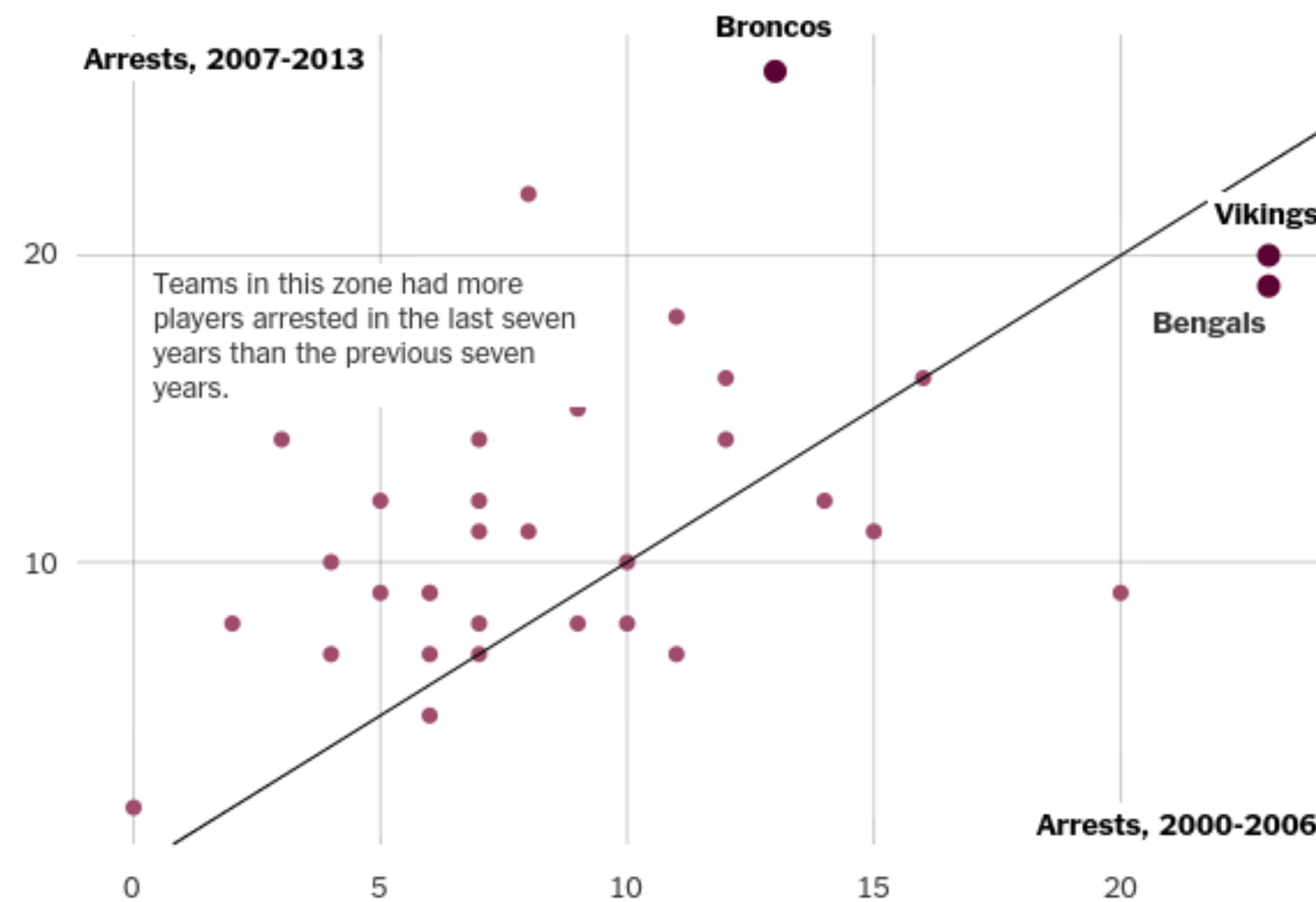
Thankfully, the data did not show that exercise incited or exacerbated problem drinking. Only very rarely during the study did anyone report drinking heavily, which the researchers defined as downing more than four drinks in succession for a woman and five for a man.

But of course this kind of epidemiological study cannot determine why working out and drinking should be associated at all, which makes the second study, a newly published review of past, related experiments, especially those involving animals, so compelling.

NFL arrests

Arrest Rates for an N.F.L. Franchise Tend to Persist Over Time

There was a 53 percent correlation between the number of a team's players arrested between 2000 and 2006 and the number arrested with the same team from 2007 and 2013.



Source: New York Times analysis of data from USA Today

NFL arrests

But there's a simple way to test that. If the results were random, you would expect there to be no correlation between the number of player arrests in one time period with a subsequent time period. You could even imagine a negative correlation, if teams that had a run of players getting in trouble took extra care not to sign players reputed to have character issues.

But that is not what happened over the last 14 years. If you chart the number of arrests of players from each franchise in the first seven years of the data, 2000 to 2006, versus the number of arrests that franchise experienced from 2007 to 2013, the correlation is a pretty solid 53 percent. A scatter plot shows a clear pattern in which those franchises with high numbers of arrests in the early years also tended to have high number arrests in later years and vice versa.

The data don't tell us anything about why these patterns are so persistent,

Correlation vs. regression

By JONATHAN WEISMAN NOV. 1, 2012



WASHINGTON — The Congressional Research Service has withdrawn [an economic report](#) that found **no correlation between top tax rates and economic growth**, a central tenet of conservative economic theory, after Senate Republicans raised concerns about the paper's findings and wording.

The decision, made in late September against the advice of the agency's economic team leadership, drew almost no notice at the time. Senator Charles E. Schumer, Democrat of New York, cited the study a week and a half after it was withdrawn in a speech on tax policy at the National Press Club.

Can you plot a correlation?

They plotted a correlation between various types of abuse – including racism, misogyny and homophobia – and political positions including voting to Leave:

Correlation Overview (UK)

LONG TERM UNEMPLOYMENT

	RACISM	TRANSPHOBIA	HOMOPHOBIA	MASCULINITY	MISOGYNY	VOTED YES TO LEAVE EU	CON 2015	UKIP	TURNOUT 2015	LONG TERM UNEMPLOYMENT
RACISM		-0.295	0.478	0.279	0.632	0.0139	-0.0658	-0.047	0.086	0.145
TRANSPHOBIA	-0.295		-0.315	-0.443	-0.334	0.334	-0.0813	0.095	-0.197	0.152
HOMOPHOBIA	0.478	-0.315		0.157	0.752	0.138	-0.0691	0.200	-0.142	0.163
MASCULINITY	0.279	-0.443	0.157		0.376	-0.077	-0.251	-0.054	0.128	0.101
MISOGYNY	0.632	-0.334	0.752	0.376		0.114	-0.111	0.273	-0.307	0.405



CORRELATION AND REGRESSION

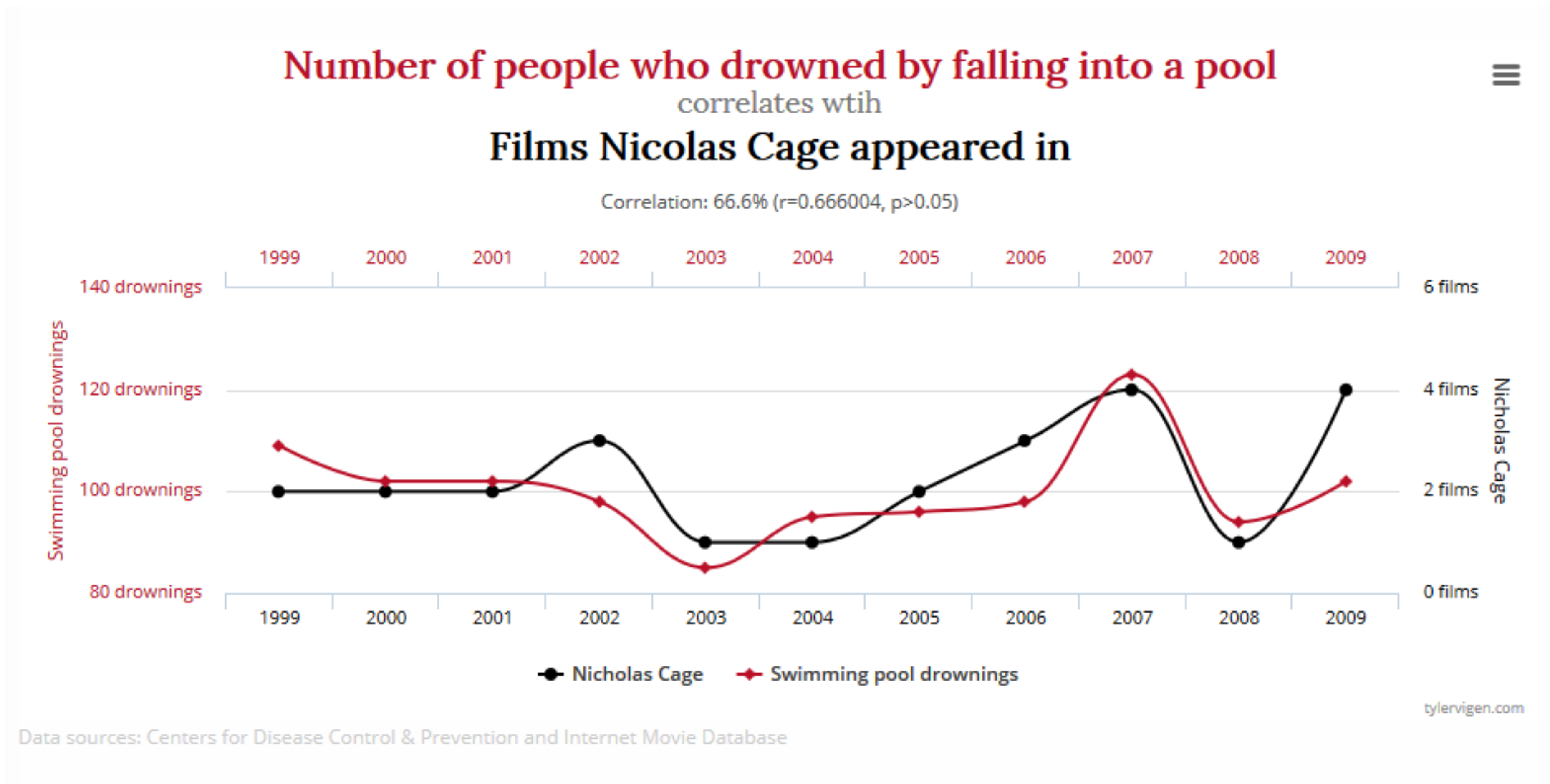
Let's practice!



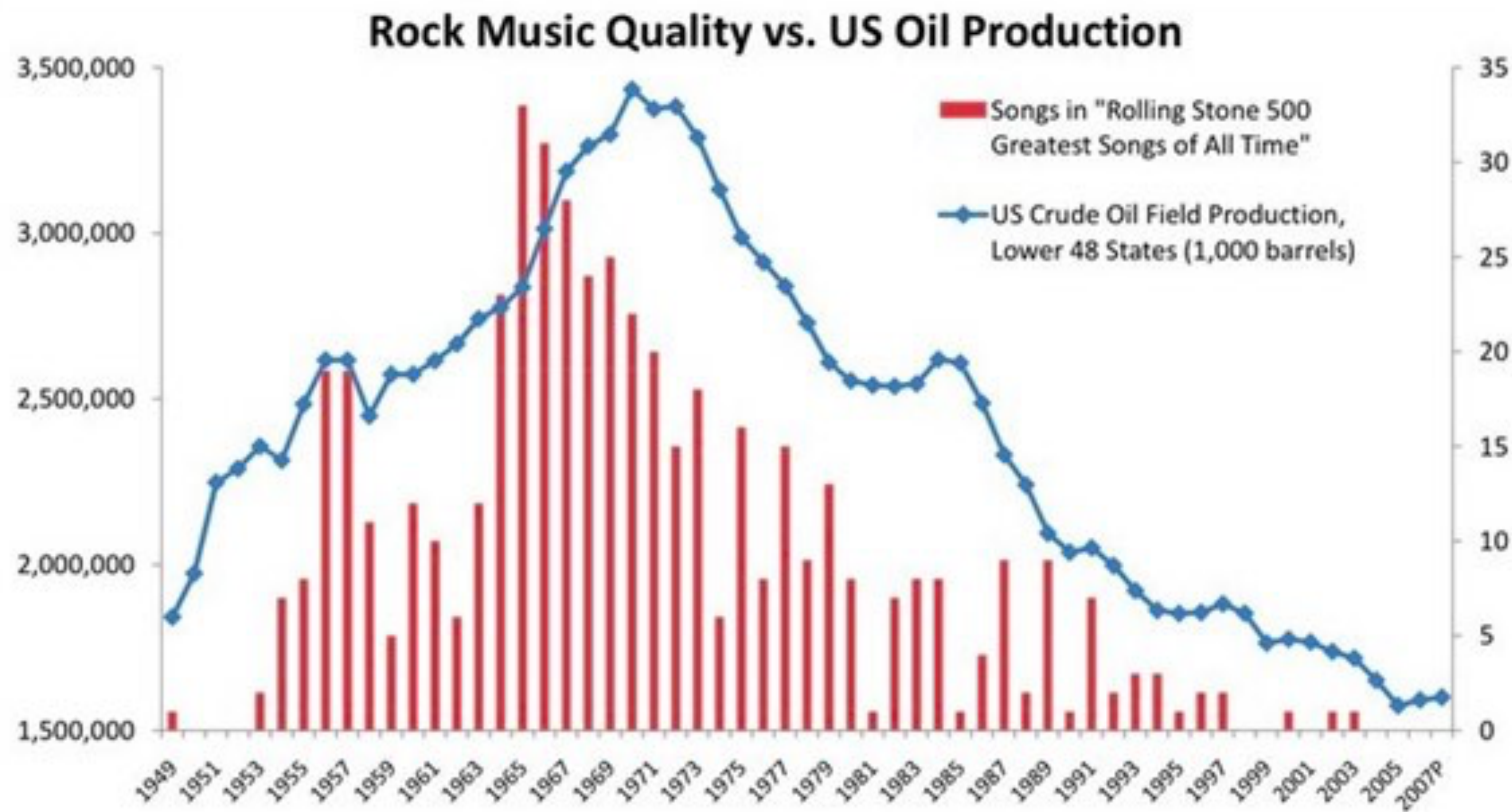
CORRELATION AND REGRESSION

Spurious correlations

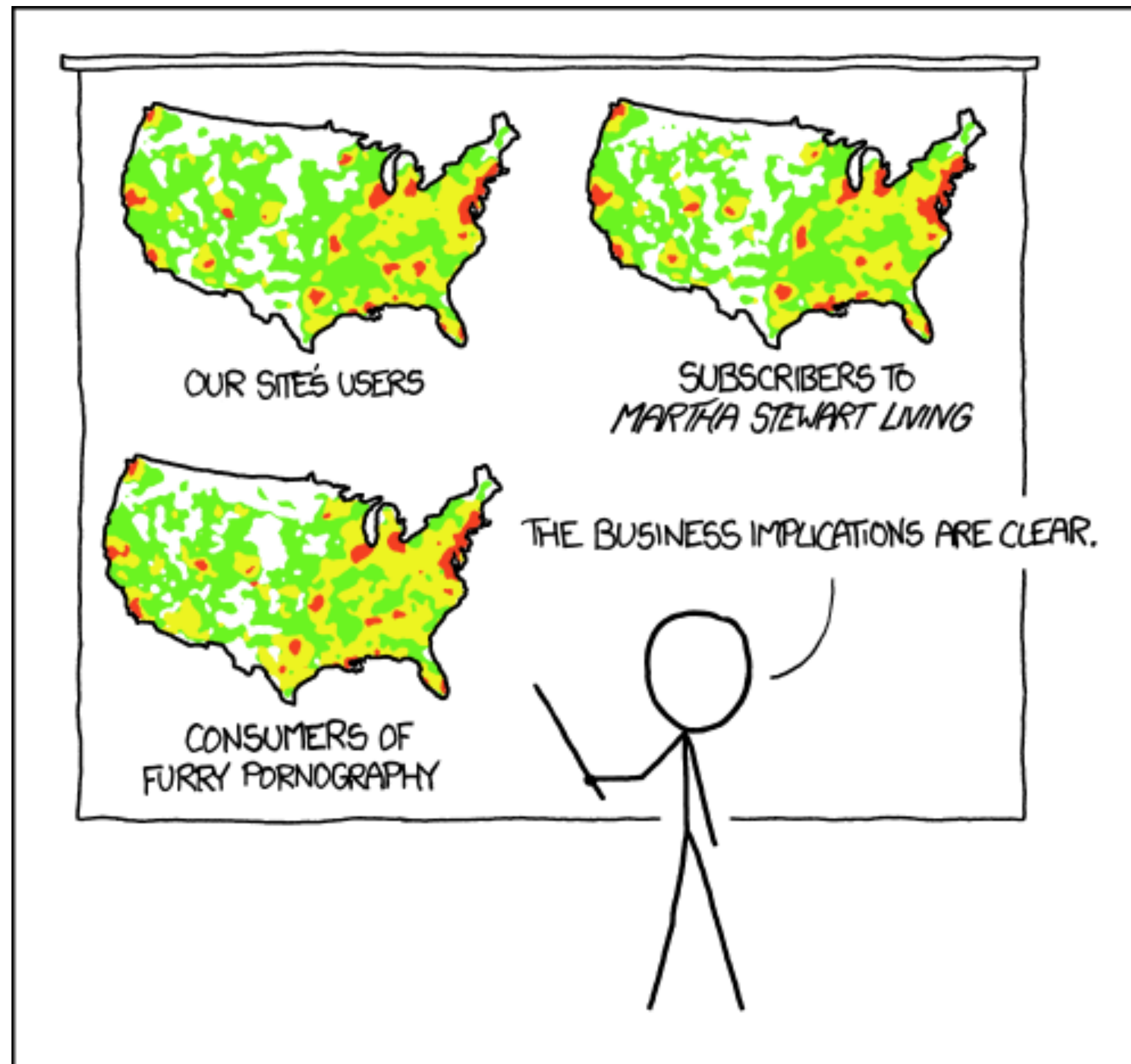
Spurious over time



Spurious over time

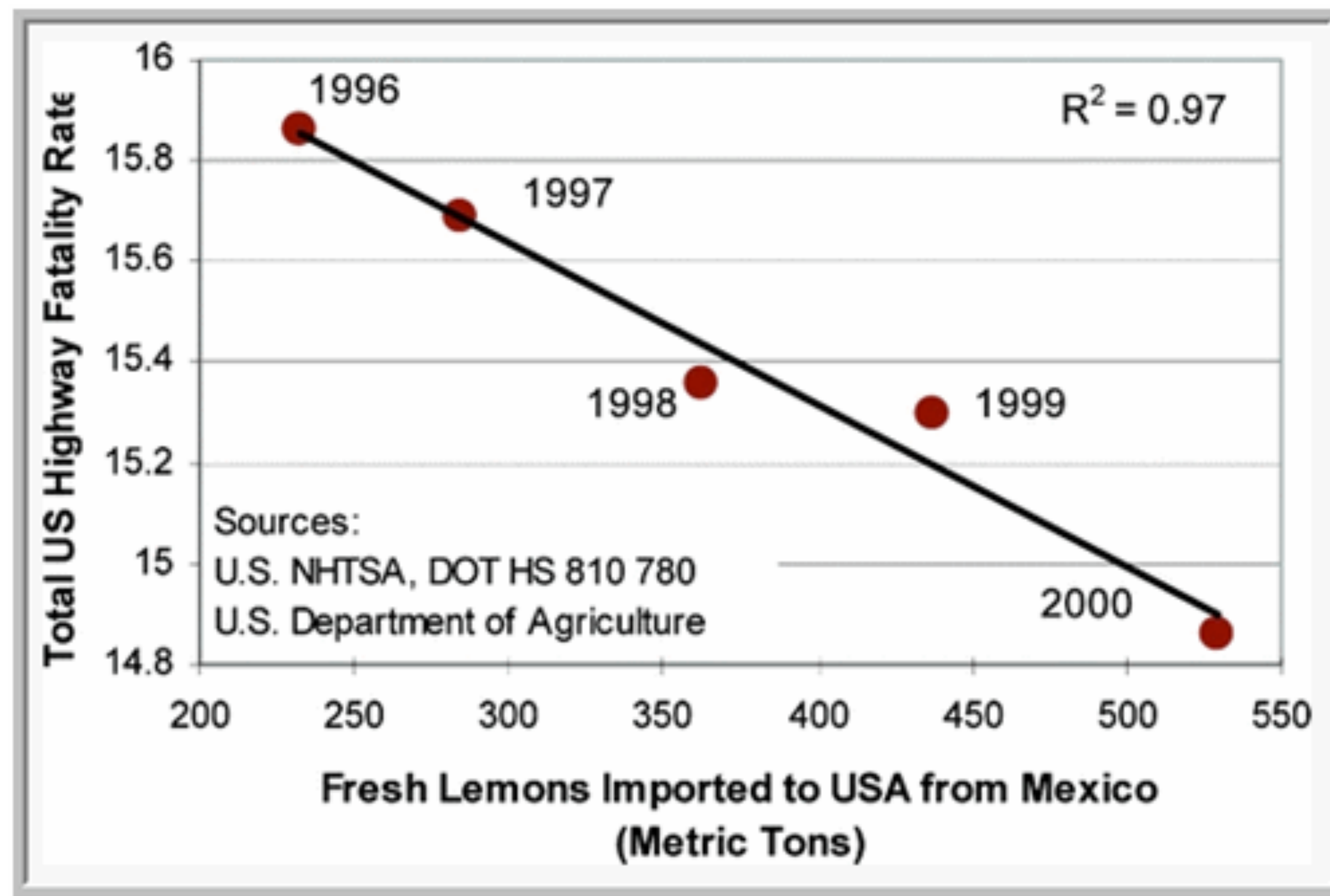


Spurious over space



PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE
BASICALLY JUST POPULATION MAPS

Spurious for whatever reason





CORRELATION AND REGRESSION

Let's practice!