THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Correlation & Simple Regression 1

Lecture 9

STA 371G

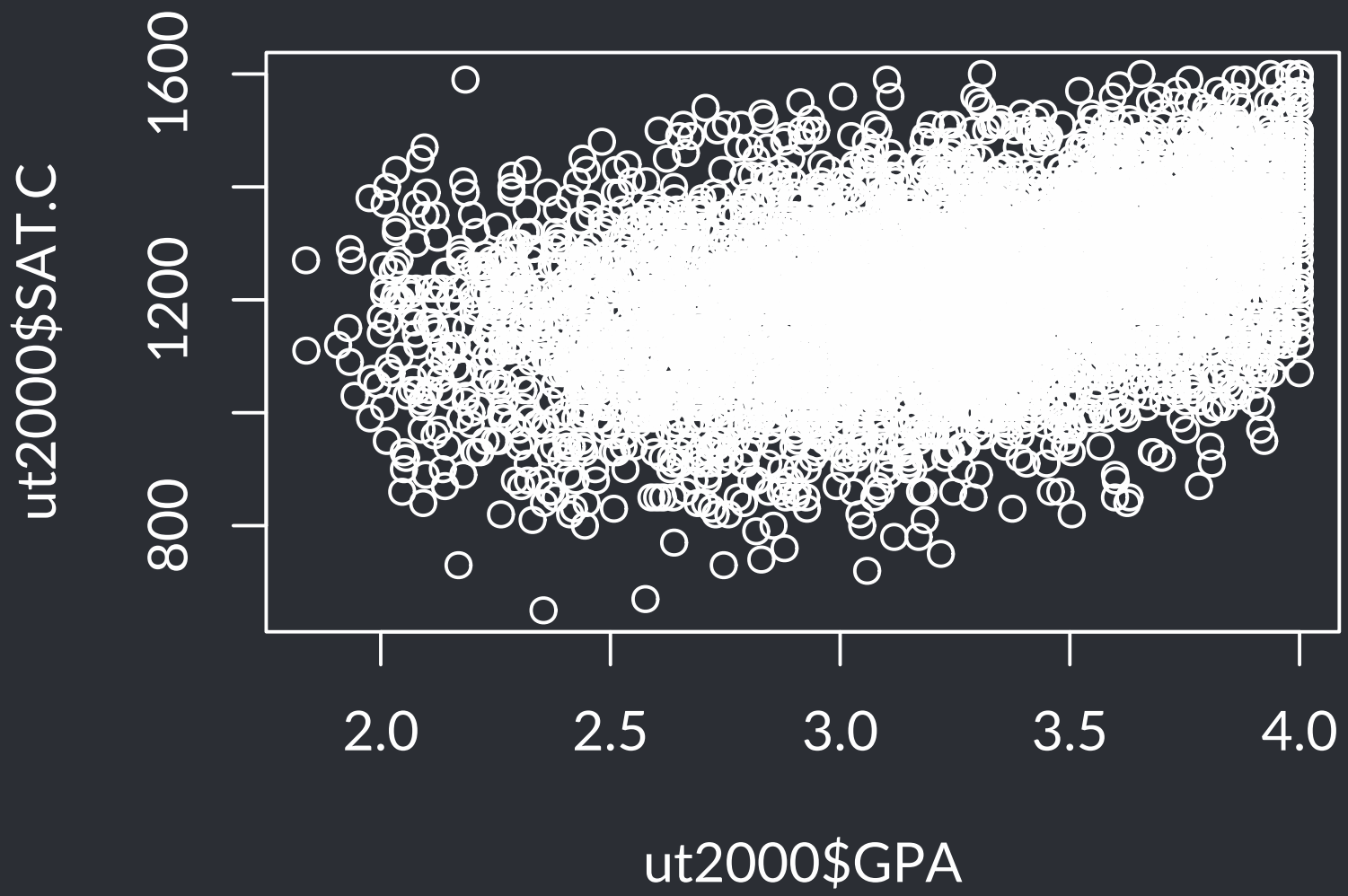
Scatterplots

- Scatterplots are the main graphical tool used to examine the relationship between two variables
- The explanatory (or independent) variable goes on the *X*-axis: it's the variable you want to use to explain observed outcomes
- The response (or dependent) variable goes on the *Y*-axis: it's the variable you want to explain

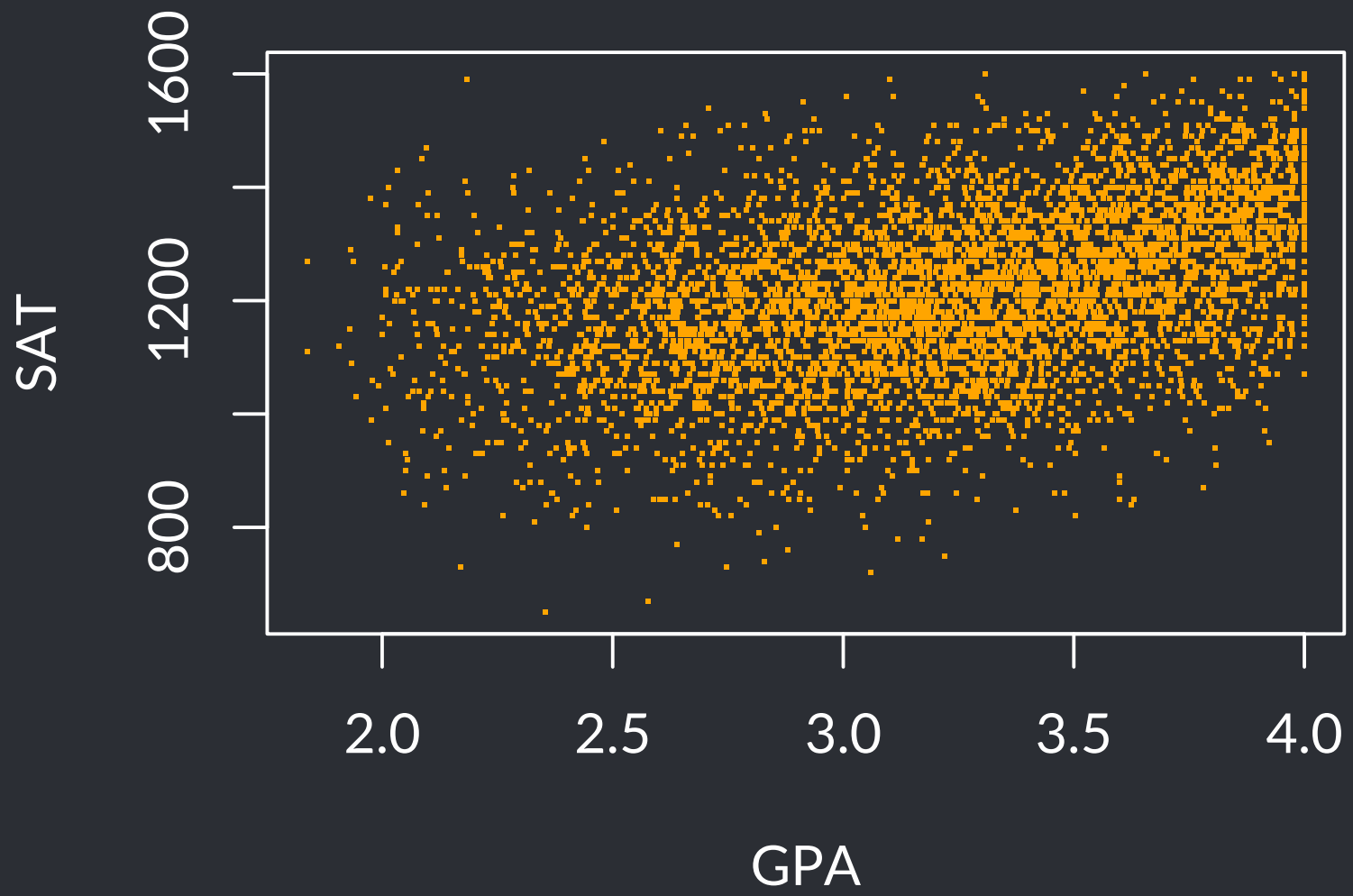
Predicting SAT scores from high school GPA

- Cases are individual students entering UT in Fall 2000
- SAT score is the response variable
- High school GPA is the explanatory variable

```
plot(ut2000$GPA, ut2000$SAT.C)
```



```
plot(ut2000$GPA, ut2000$SAT.C, col="orange", pch=".",
     xlab="GPA", ylab="SAT")
```



Correlation coefficient

Definition

The correlation r between two variables X and Y measures the strength of the linear relationship between them. Correlation ranges from —1 (perfect negative relationship) to 0 (no relationship) to 1 (perfect positive relationship).

Correlation coefficient

Definition

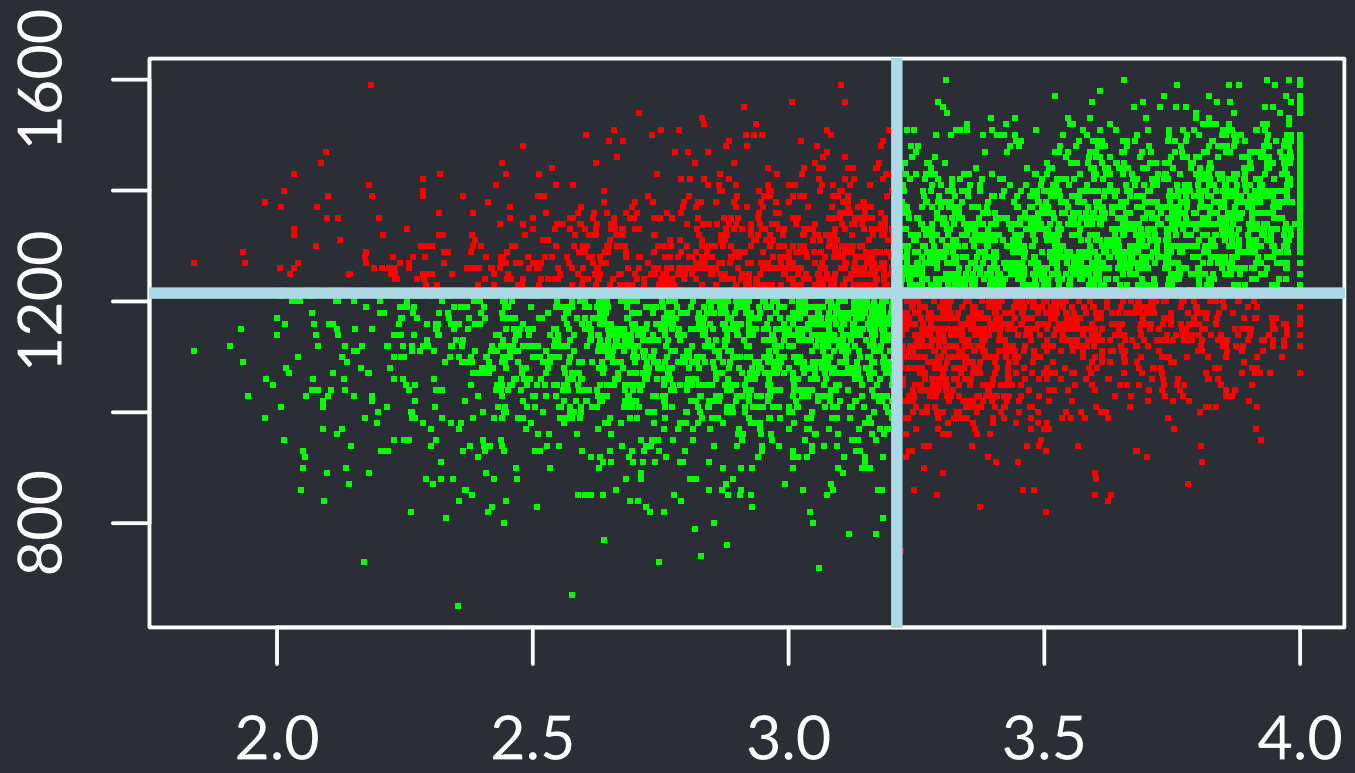
The correlation r between two variables X and Y measures the strength of the linear relationship between them. Correlation ranges from —1 (perfect negative relationship) to 0 (no relationship) to 1 (perfect positive relationship).

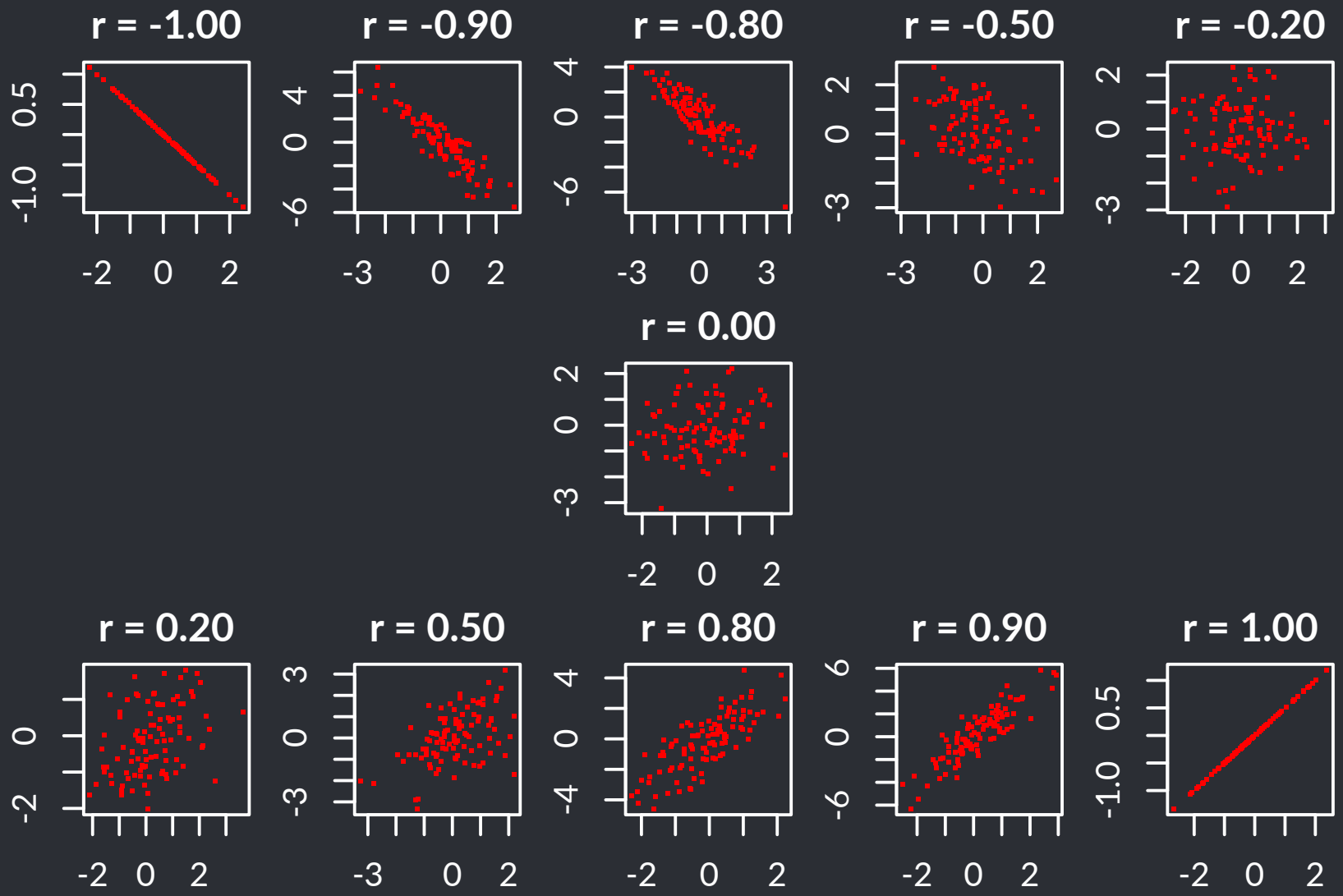
It is calculated as

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\mathrm{SD}(X)} \right) \left(\frac{Y_i - \overline{Y}}{\mathrm{SD}(Y)} \right)$$

How correlation is calculated

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \underbrace{\left(\frac{X_i - \bar{X}}{\mathrm{SD}(X)} \right)}_{z\text{-score for } i\text{th GPA}} \cdot \underbrace{\left(\frac{Y_i - \bar{Y}}{\mathrm{SD}(Y)} \right)}_{z\text{-score for } i\text{th SAT}}$$




r = -1.00
r = -0.90
r = -0.80
r = -0.50
r = -0.20
r = 0.00
r = 0.20
r = 0.50
r = 0.80
r = 0.90
r = 1.00

Calculating correlations in R

```
cor(ut2000$GPA, ut2000$SAT.C)

[1] 0.3904258
```

Calculating correlations in R

```
cor(ut2000$GPA, ut2000$SAT.C)
```

```
[1] 0.3904258
```

```
cor(ut2000$SAT.C, ut2000$GPA)
```

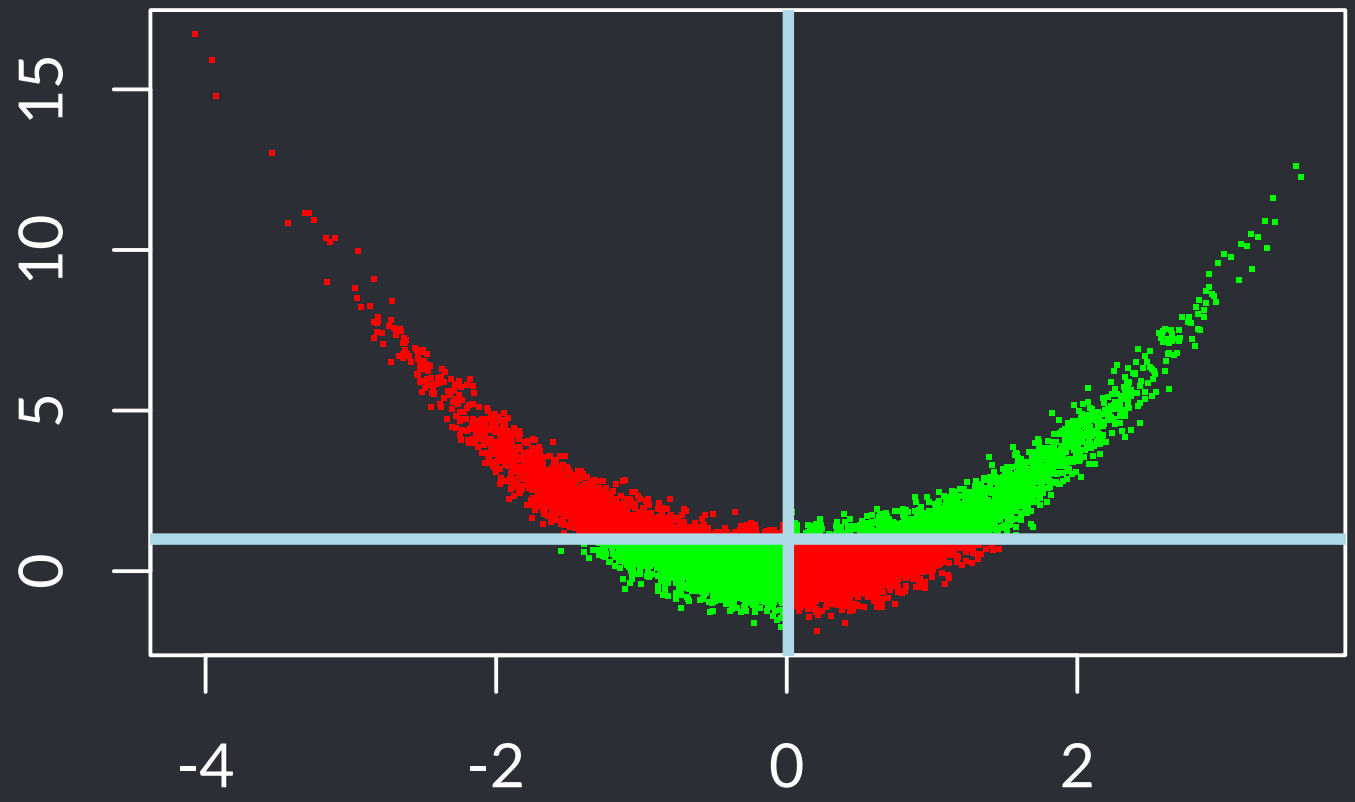
```
[1] 0.3904258
```

Conditions for correlation

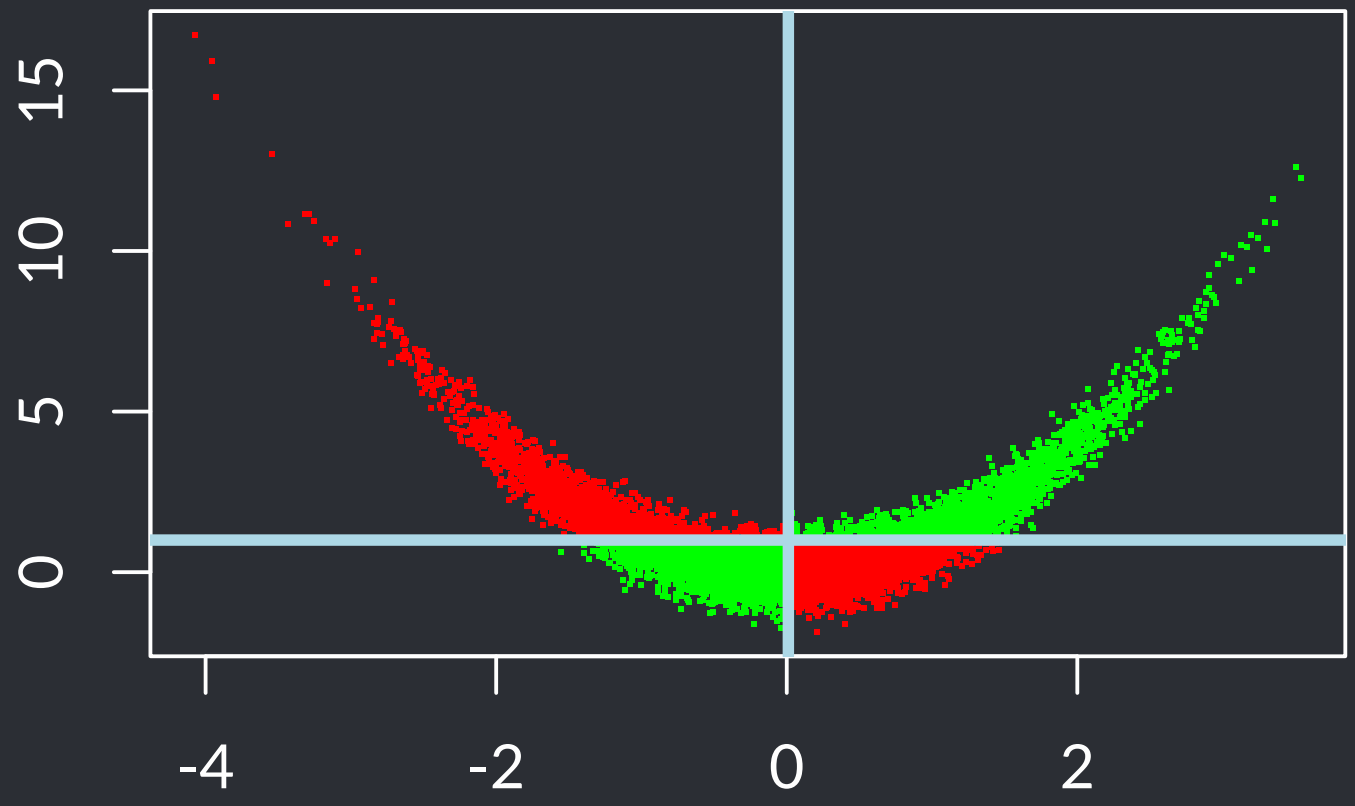
In order for correlation coefficients to make sense:

- Both variables must be quantitative
- The relationship between X and Y must be linear (i.e., not curved)
- There must not be extreme outliers

What if the relationship is not linear?

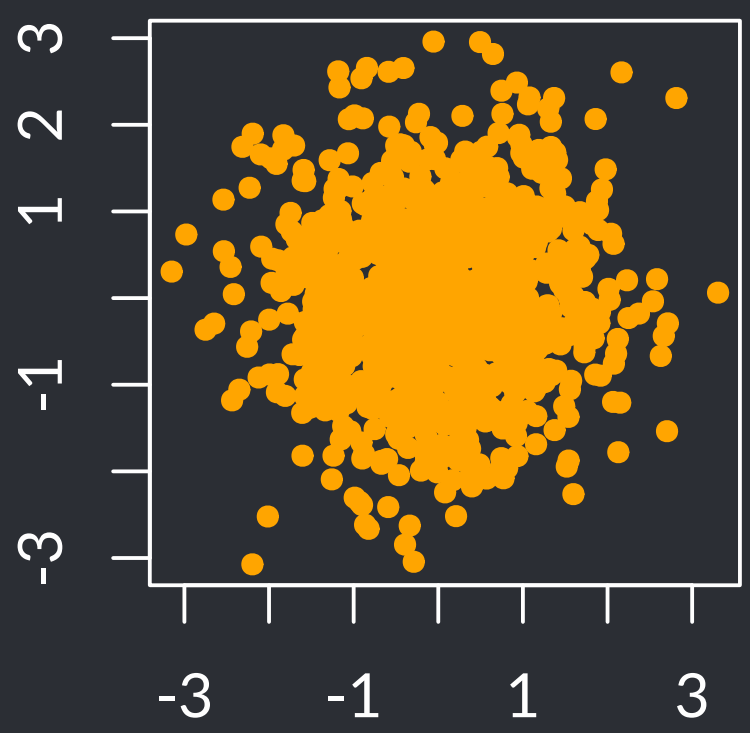


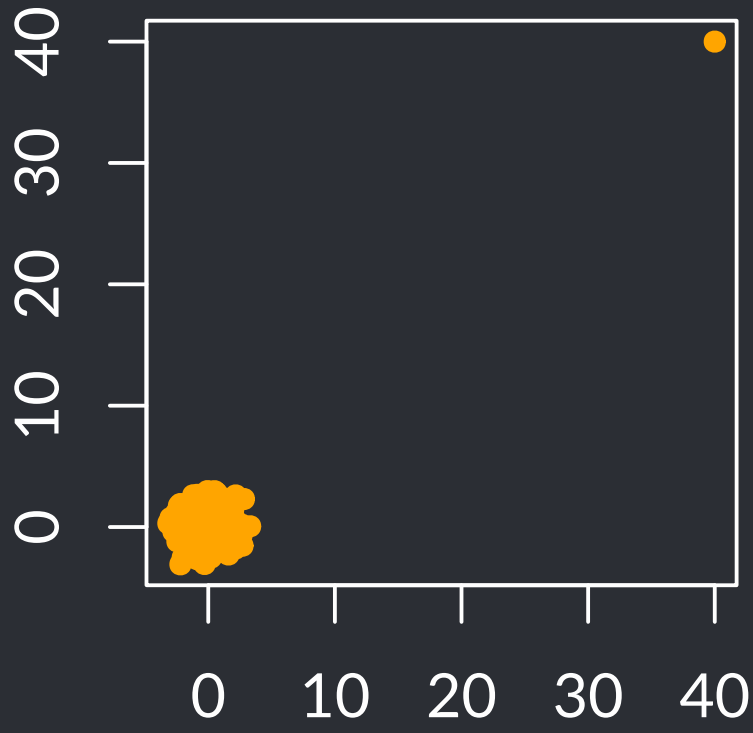
What if the relationship is not linear?



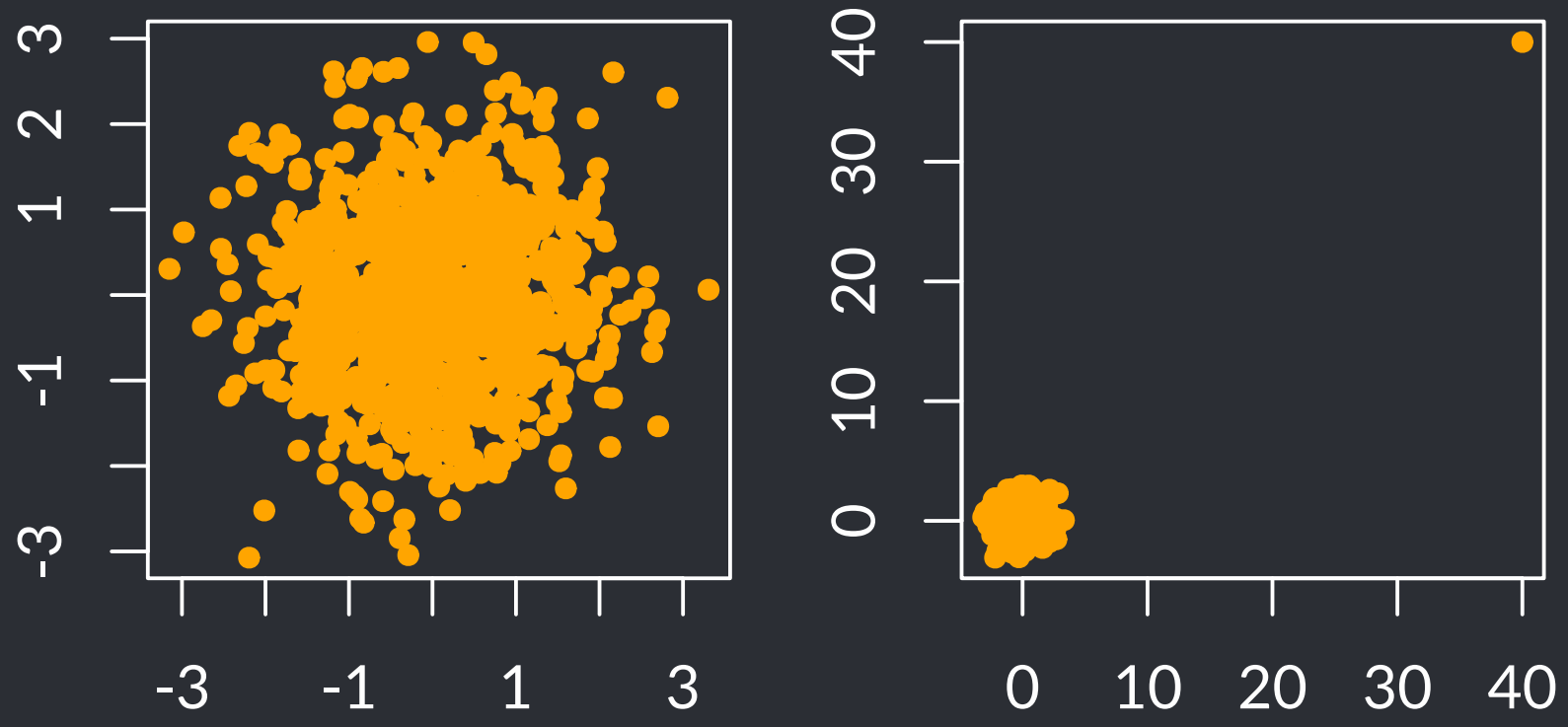
The correlation is 0 even though there is a strong relationship between X and Y!

What if there are extreme outliers?





What if there are extreme outliers?



Adding just one outlier makes the correlation jump from 0 to 0.6!

Correlation $\neq$ causation

- Just because the correlation between X and Y is large does not mean that X causes Y (or Y causes X)

Correlation $\neq$ causation

- Just because the correlation between X and Y is large does not mean that X causes Y (or Y causes X)
- The correlation could be caused by a lurking variable that is causing both X and Y

Correlation $\neq$ causation

- Just because the correlation between X and Y is large does not mean that X causes Y (or Y causes X)
- The correlation could be caused by a lurking variable that is causing both X and Y
- Or the correlation could be a coincidence!

Lurking variables

What could the lurking variable be?

- Students with higher GPAs tend to have higher SAT scores

Lurking variables

What could the lurking variable be?

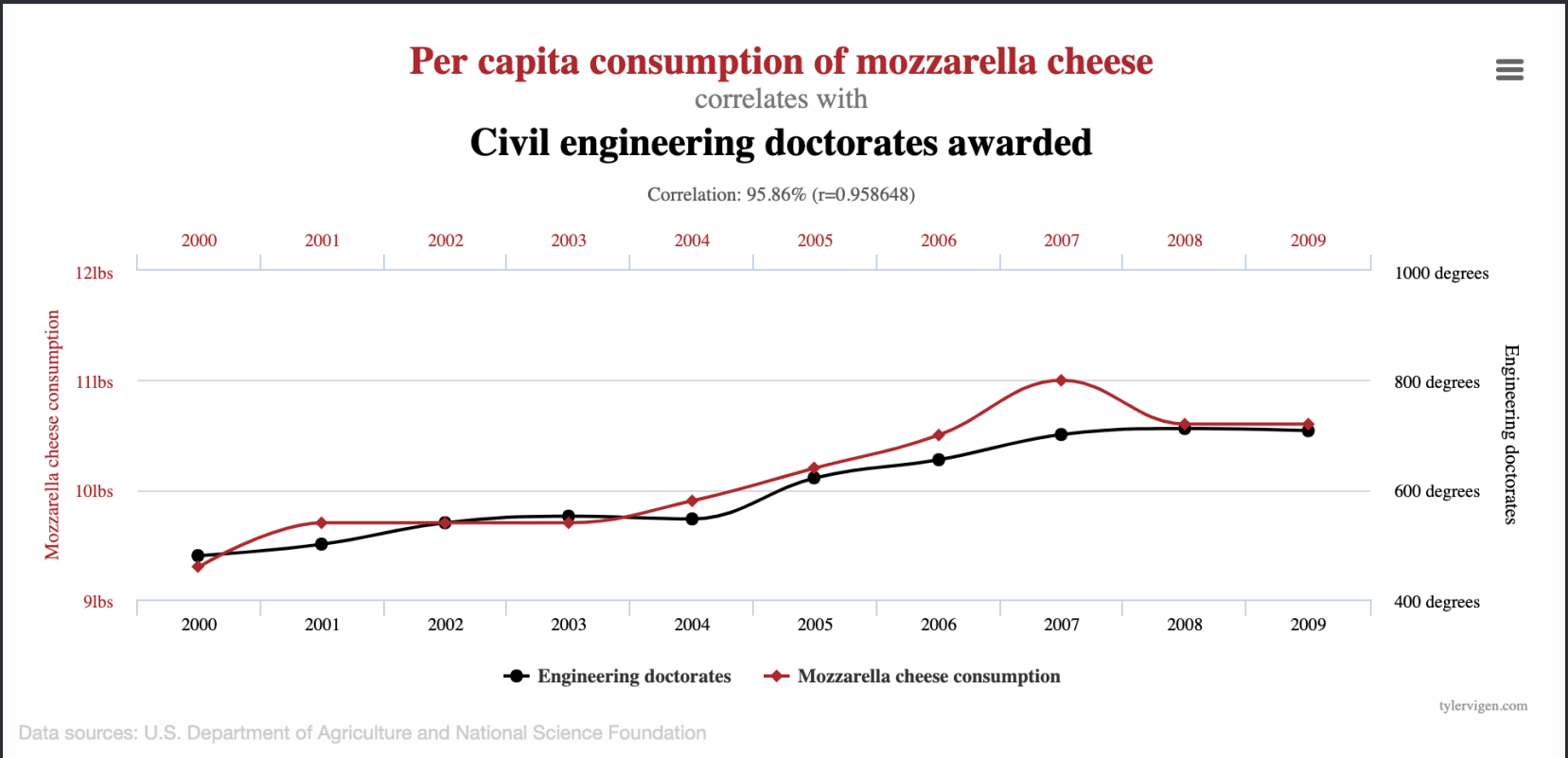
- Students with higher GPAs tend to have higher SAT scores
- More people tend to drown when ice cream sales are higher

Lurking variables

What could the lurking variable be?

- Students with higher GPAs tend to have higher SAT scores
- More people tend to drown when ice cream sales are higher
- In Europe, more babies tend to be born where there are more storks

Seriously, correlation ≠ causation, even if *r* is high



Correlation ≠ causation