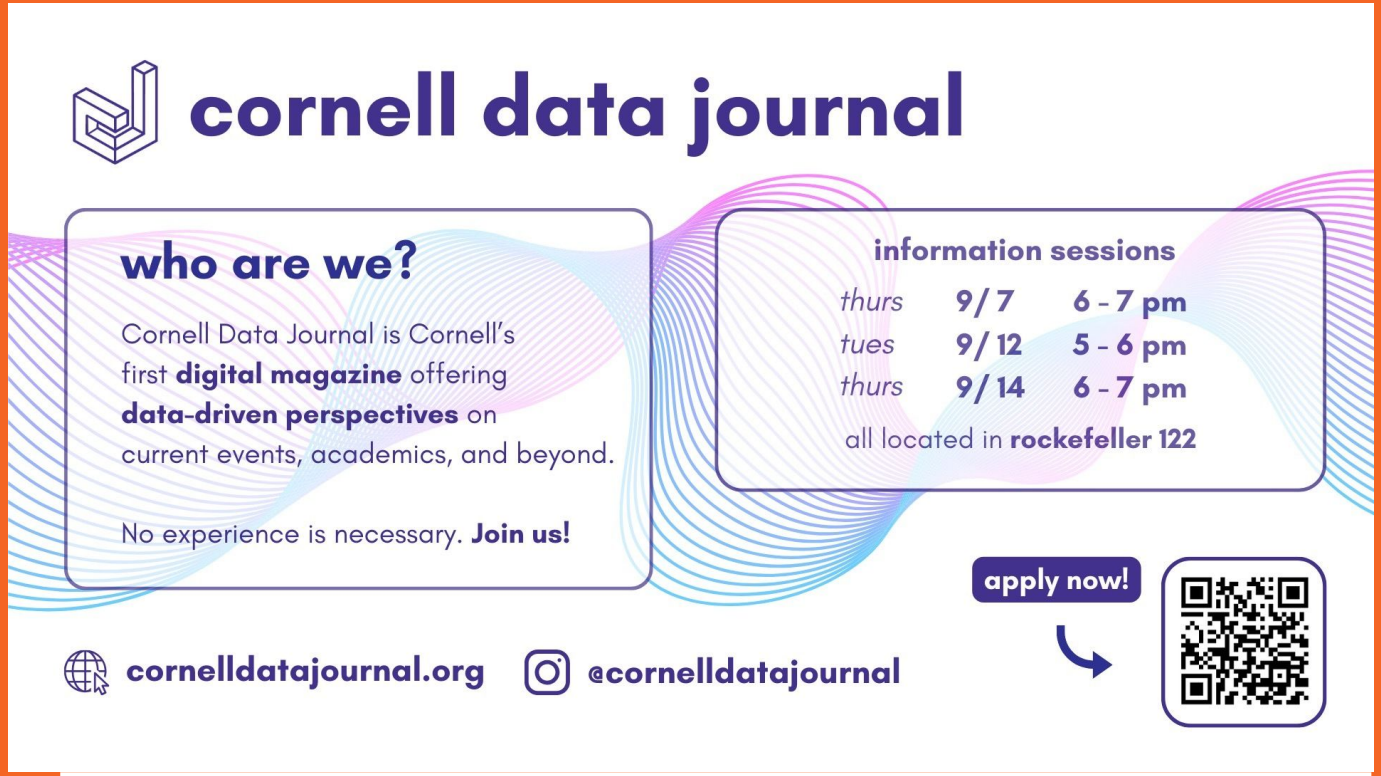
INFO 2950: Intro to Data Science

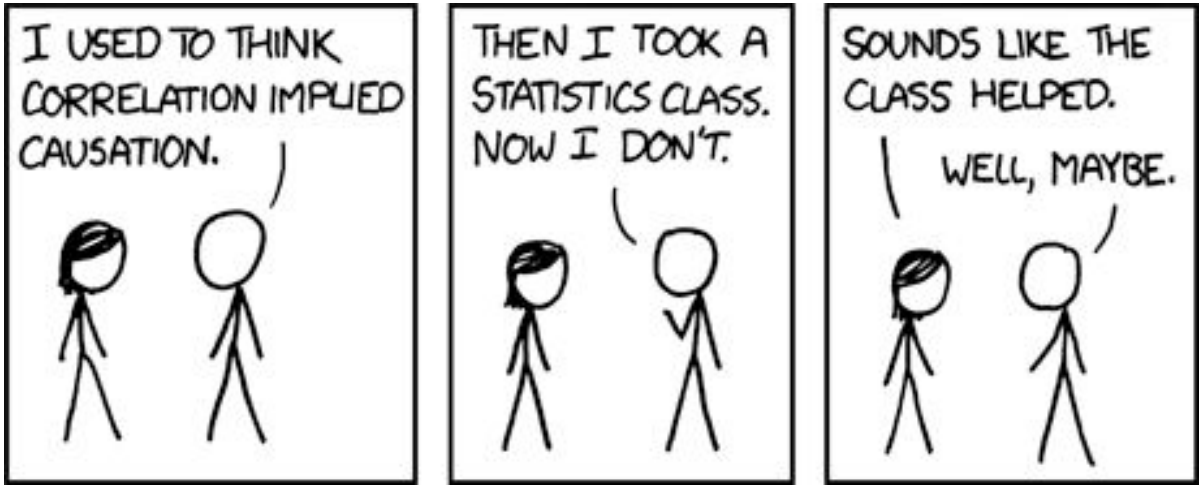


Lecture 5

2023-09-06

Today's agenda

- **Covariance**
  - SQL review
- **Correlation**
  - != Causation
- **HTML**
- **SQL Review**



https://imgs.xkcd.com/comics/correlation.png

Relationships between two variables

Temperature and ice cream sales

Relationships between two variables

Temperature and ice cream sales

Temperature and dental floss sales

Relationships between two variables

- ↑ Temperature and ice cream sales ↑
- ↑ Temperature and dental floss sales —
- ↑ Temperature and ski trips

Relationships between two variables

- Temperature and ice cream sales ↑
- Temperature and dental floss sales —
- Temperature and ski trips ↓
- Ice cream sales and temperature

Relationships between two variables

- ↑ Temperature and ice cream sales ↑
- ↑ Temperature and dental floss sales —
- ↑ Temperature and ski trips ↓
- ↑ Ice cream sales and temperature

Also tends to be higher, but not caused by ice cream

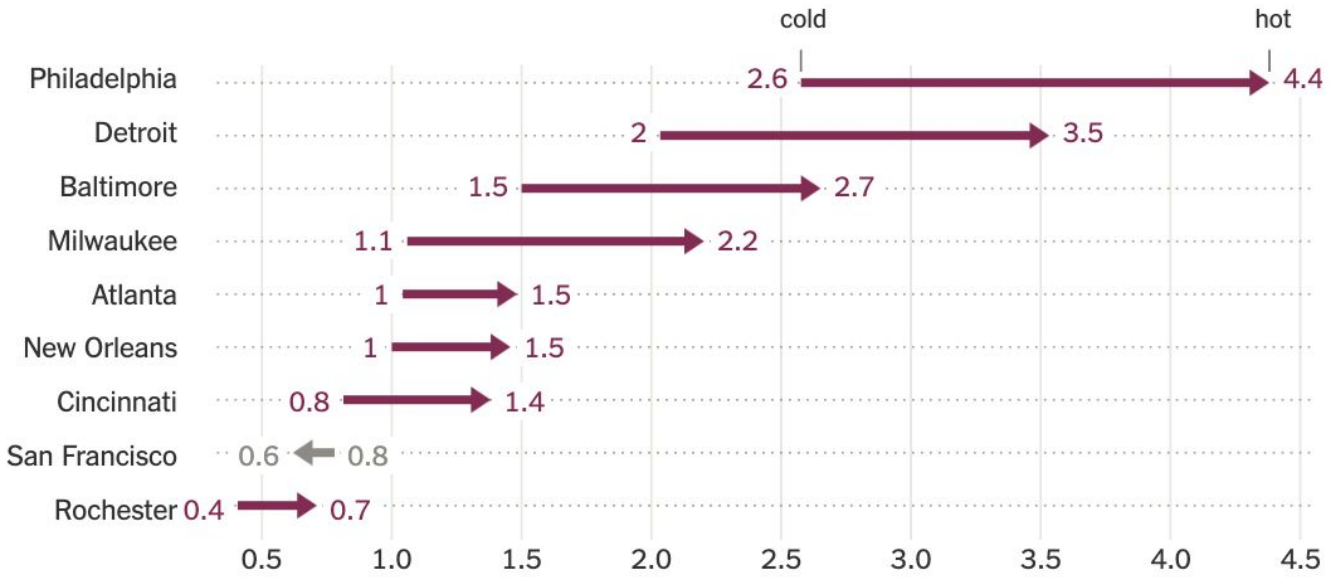
Relationships between two variables

- ↑ Temperature and ice cream sales ↑
- ↑ Temperature and dental floss sales ▬
- ↑ Temperature and ski trips ↓
- ↑ Ice cream sales and temperature — Also tends to be higher, but not caused by ice cream
- ↑ Ice cream sales and murder

Rise in Shooting Victims on Hot Days

The average number of shooting victims per cold day (defined here as under 50 degrees Fahrenheit) and per hot day (85 and up) in nine cities in recent years.

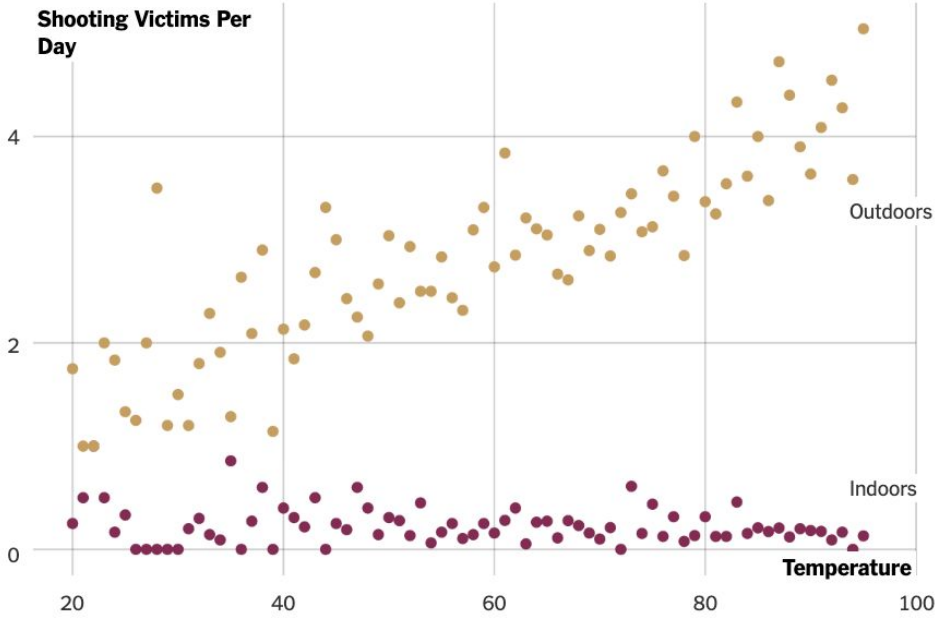
Shooting victims per day, based on weather



https://www.nytimes.com/2018/09/21/upshot/a-rise-in-murder-lets-talk-about-the-weather.html

Difference Between Violence Outdoors and Indoors in Philadelphia

Gun victimization increases outdoors in the city as temperatures rise; there's virtually no change indoors.



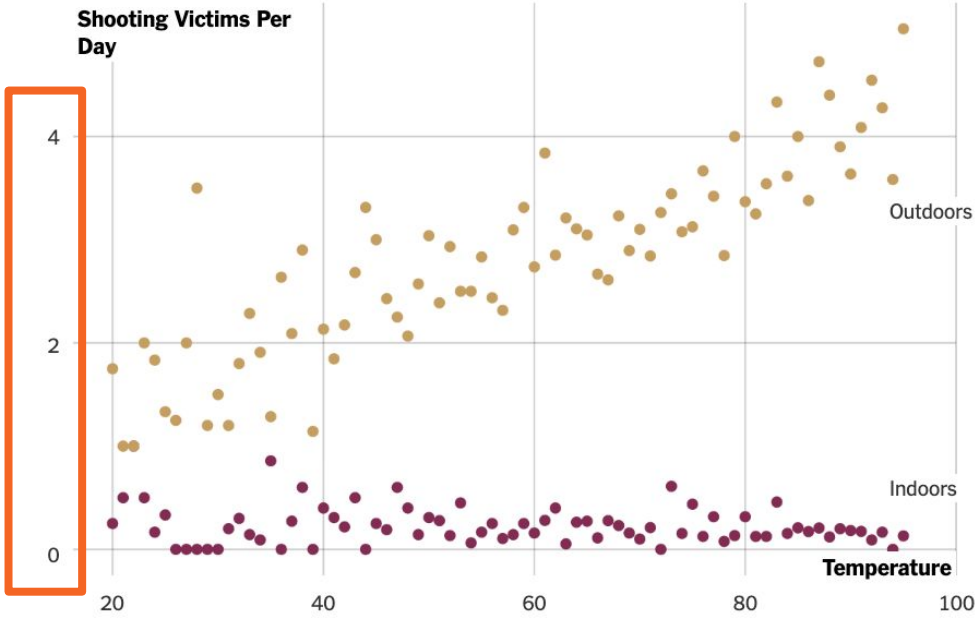
https://www.nytimes.com/2018/09/21/upshot/a-rise-in-murder-lets-talk-about-the-weather.html

Notice the y-axis!

"Murders are relatively rare, so even minor changes in any number of factors (including randomness) can have a major effect on a city's murder count."

Difference Between Violence Outdoors and Indoors in Philadelphia

Gun victimization increases outdoors in the city as temperatures rise; there's virtually no change indoors.



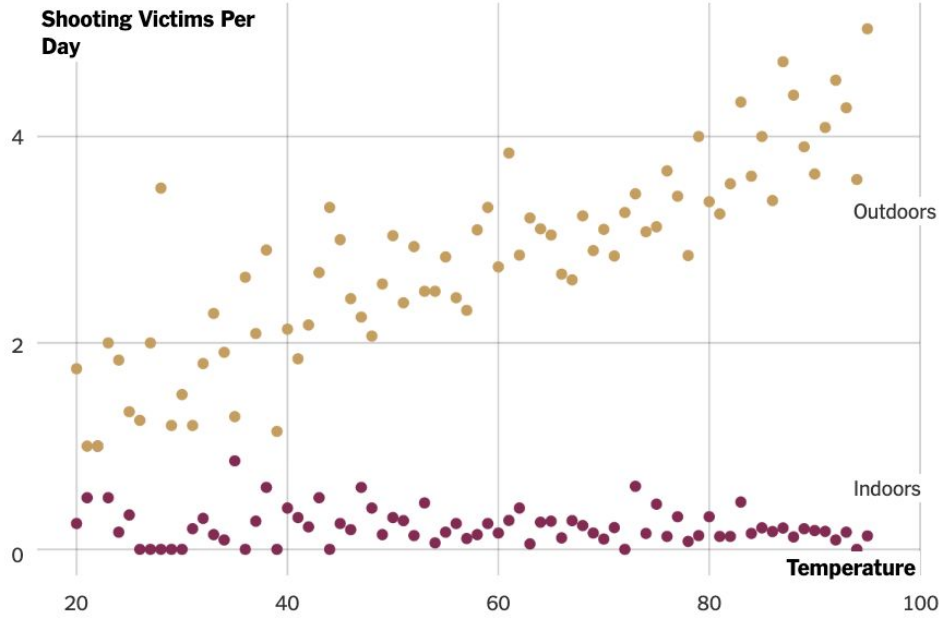
https://www.nytimes.com/2018/09/21/upshot/a-rise-in-murder-lets-talk-about-the-weather.html

Difference Between Violence Outdoors and Indoors in Philadelphia

Gun victimization increases outdoors in the city as temperatures rise; there's virtually no change indoors.

Other considerations:

- People are outside more since school is out
- Weather can't explain many big increases in murder rates
- High temperatures → turning on AC → staying inside (but, disparities in income)



https://www.nytimes.com/2018/09/21/upshot/a-rise-in-murder-lets-talk-about-the-weather.html

Relationships between two variables

Temperature and ice cream sales ↑

Temperature and dental floss sales —

Temperature and gas prices ↑

Ice cream sales and temperature

Also tends to be higher, but not caused by ice cream

Ice cream sales and murder

Maybe observed to be higher, but only in some cities / circumstances

What do summary stats tell us?

- **With mean, variance for one variable X:**
  - **is this specific value of X unusually high or low?**

What do summary stats tell us?

avg(temp), var(temp) temperature

- With mean, variance for one variable X:
  - is this specific value of X unusually high or low?

What do summary stats tell us?

avg(temp), var(temp) temperature

- With mean, variance for one variable X:
  - is this specific value of X unusually high or low?

Today it's 87° F. Historically, (in Ithaca in Sept) the average temperature is ~72° with a standard deviation of ~3°.

https://weatherspark.com/y/22130/Average-Weather-in-Ithaca-New-York-United-States-Year-Round

What do summary stats tell us?

- With mean, variance for one variable X:
  - is this specific value of X unusually high or low?

Today it's 87° F. Historically, (in Ithaca in Sept) the average temperature is ~72° with a standard deviation of ~3°.

Today's temp is way higher than the mean!

https://weatherspark.com/y/22130/Average-Weather-in-Ithaca-New-York-United-States-Year-Round

What do summary stats tell us?

- **With mean, variance for one variable X:**
  - **is this specific value of X unusually high or low?**

What do summary stats tell us?

- With mean, variance for one variable X:
  - is this specific value of X unusually high or low?
- With covariance for two variables X, Y:
  - if my value for X is unusually high, is my value for Y likely to be unusually high or low?

Covariance: do X, Y move together?

- Measures the direction of the relationship between two variables X, Y

Covariance: do X, Y move together?

- Measures the direction of the relationship between two variables X, Y
- If higher X mostly corresponds with higher Y, then X and Y have positive covariance (temperature, ice cream sales)

Covariance: do X, Y move together?

- Measures the direction of the relationship between two variables X, Y
- If higher X mostly corresponds with higher Y, then X and Y have positive covariance (temperature, ice cream sales)
- If higher X mostly corresponds with lower Y, then X and Y have negative covariance (temperature, ?)

nature human behaviour

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature human behaviour > articles > article

Article | Published: 05 October 2020

Learning is inhibited by heat exposure, both internationally and within the United States

R. Jisung Park, A. Patrick Behrer & Joshua Goodman

Heat and Learning

- **Data**: 10 million students retaking the PSATs
- **Findings**: hotter school days in the years before the test → reduce scores (hotter weekends/summers have little impact)
- "Without air conditioning, a 1°F hotter school year reduces that year's learning by 1 percent."
  - Disproportionately impacts minority students, (~5% of the racial achievement gap)

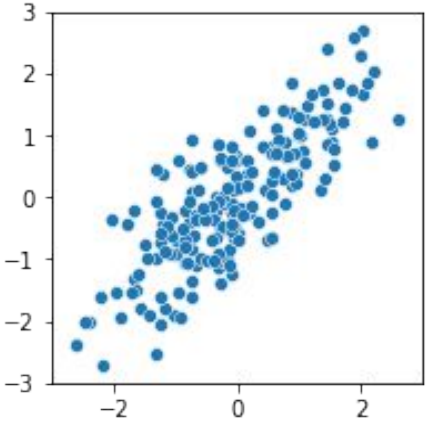
https://www.aeaweb.org/articles?id=10.1257/pol.20180612

Covariance: do X, Y move together?

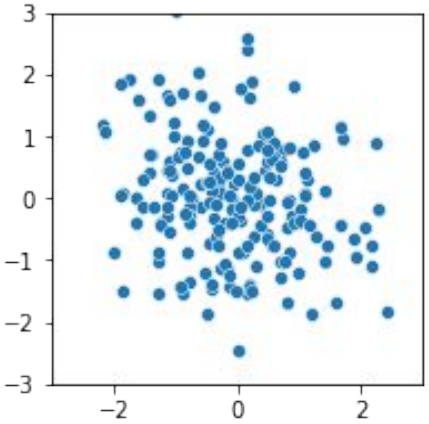
- Measures the direction of the relationship between two variables X, Y
- If higher X mostly corresponds with higher Y, then X and Y have positive covariance (temperature, ice cream sales)
- If higher X mostly corresponds with lower Y, then X and Y have negative covariance (temperature, test scores)

Sort these by increasing covariance

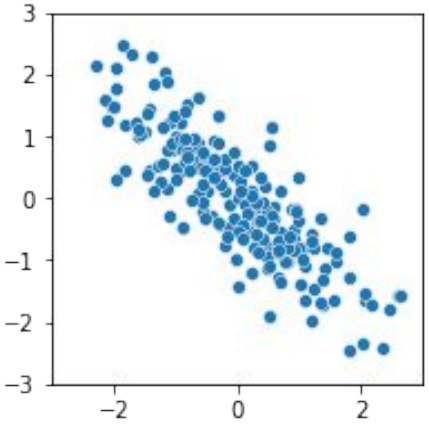
A



B

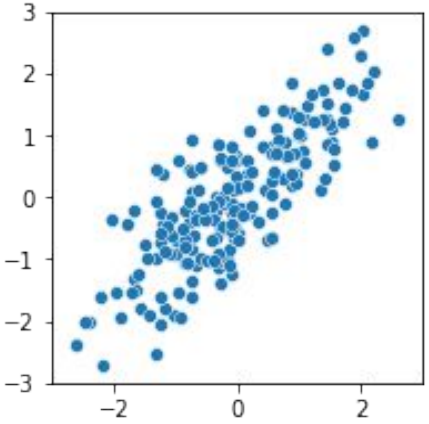


C

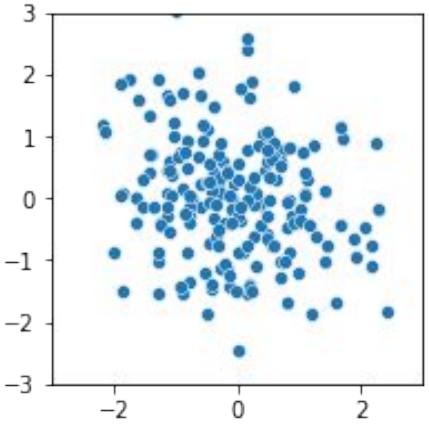


Sort these by increasing covariance

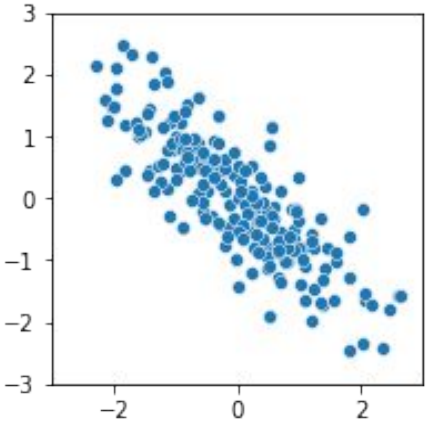
A



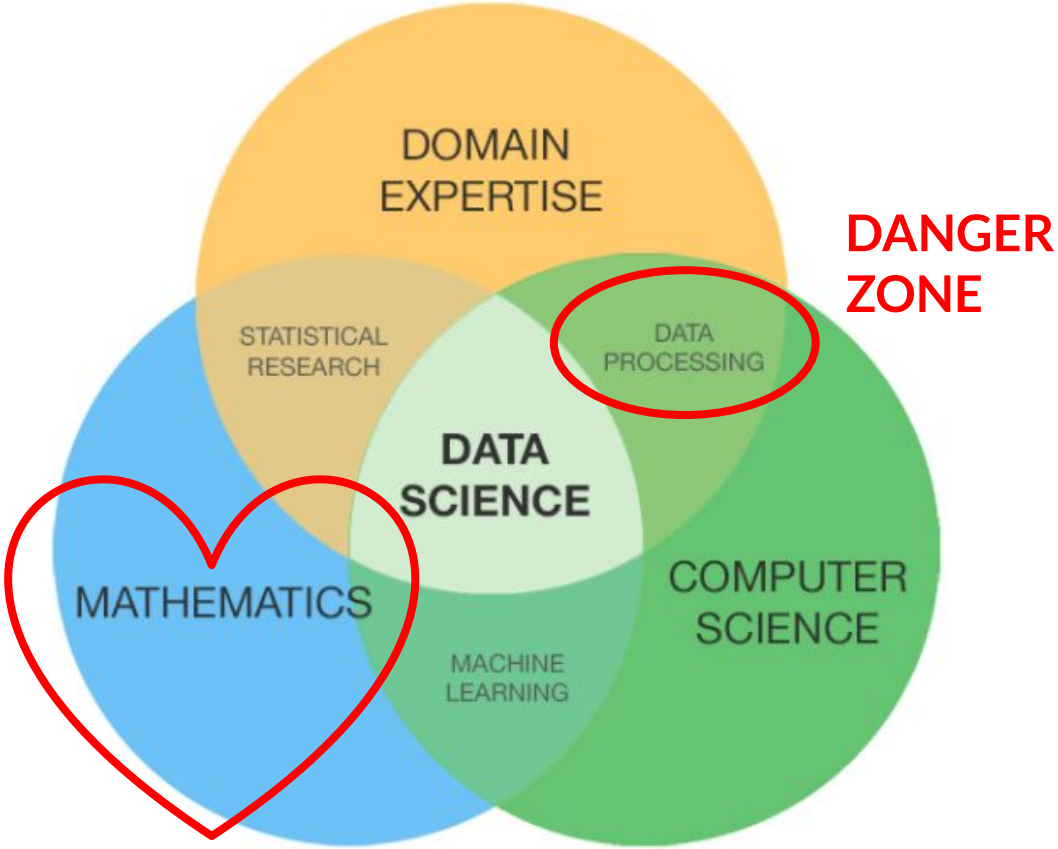
B



C



C (-0.8) < B (-0.2) < A (0.8)



Content source: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

—

Sample variance is the average squared distance to the sample mean

(using N to keep things simple)

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

Variance

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

i	Day	X
1	Sep 6	87°
2	Sep 7	90°
3	Sep 8	84°
…	…	…

Calculating variance manually

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

i	Day	X
1	Sep 6	87°
2	Sep 7	90°
3	Sep 8	84°
…	…	…

$$\bar{X} = (87+90+84)/3 = 87$$

SELECT _____ FROM Table

Calculating variance manually

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

i	Day	X
1	Sep 6	87°
2	Sep 7	90°
3	Sep 8	84°
…	…	…

$$\bar{X} = (87+90+84)/3 = 87$$

```
SELECT AVG(X) FROM Table
```

Calculating variance manually

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

i	Day	X	diff $X - \bar{X}$
1	Sep 6	87°	0
2	Sep 7	90°	3
3	Sep 8	84°	-3
…	…	…	

```
SELECT __, X - 87 __ diff FROM Table
```

Calculating variance manually

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

i	Day	X	diff $X - \bar{X}$
1	Sep 6	87°	0
2	Sep 7	90°	3
3	Sep 8	84°	-3
…	…	…	

```
SELECT *, X - 87 AS diff FROM Table
```

Calculating variance manually

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

i	Day	X	diff $X - \bar{X}$	sqdiff $(X_i - \bar{X})^2$
1	Sep 6	87°	0	0
2	Sep 7	90°	3	9
3	Sep 8	84°	-3	9
…	…	…		

```
SELECT *, _____ __ sqdiff FROM Table
```

Calculating variance manually

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

i	Day	X	diff $X - \bar{X}$	sqdiff $(X_i - \bar{X})^2$
1	Sep 6	87°	0	0
2	Sep 7	90°	3	9
3	Sep 8	84°	-3	9
…	…	…		

```
SELECT *, diff^2 AS sqdiff FROM Table
```

Calculating variance manually

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

i	Day	X	diff $X - \bar{X}$	sqdiff $(X_i - \bar{X})^2$
1	Sep 6	87°	0	0
2	Sep 7	90°	3	9
3	Sep 8	84°	-3	9
…	…	…		

Var = (0+9+9) / 3 = 6

```
SELECT SUM(____) / _____(_) FROM Table
```

Calculating variance manually

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

i	Day	X	diff $X - \bar{X}$	sqdiff $(X_i - \bar{X})^2$
1	Sep 6	87°	0	0
2	Sep 7	90°	3	9
3	Sep 8	84°	-3	9
…	…	…		

Var = (0+9+9) / 3 = 6

```
SELECT SUM(sqdiff) / COUNT(*) FROM Table
```

Calculating variance manually

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

i	Day	X	diff $X - \bar{X}$	sqdiff $(X_i - \bar{X})^2$
1	Sep 6	87°	0	0
2	Sep 7	90°	3	9
3	Sep 8	84°	-3	9
…	…	…		

Var = (0+9+9) / 3 = 6

(This also works)

```
SELECT SUM(sqdiff) / COUNT(sqdiff) FROM Table
```

Calculating variance ~~manually~~

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

i	Day	X
1	Sep 6	87°
2	Sep 7	90°
3	Sep 8	84°
…	…	…

In practice, you'd just do np.var(X)

What if we have another variable? (Y = ice cream sales)

i	Day	X	Y
1	Sep 6	87°	50
2	Sep 7	90°	60
3	Sep 8	80°	45
…	…	…	

—

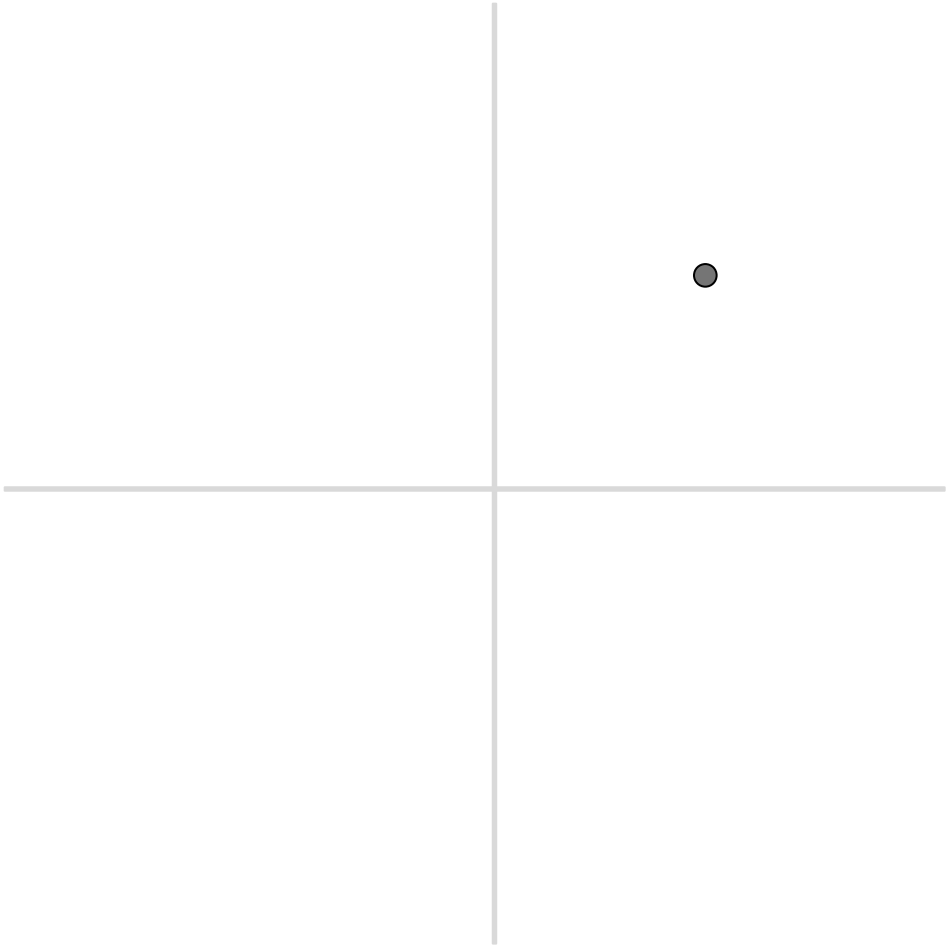
Sample covariance is the average product of distances to the sample means

$$\frac{\Sigma_i (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

—

Sample covariance is the average product of distances to the sample means

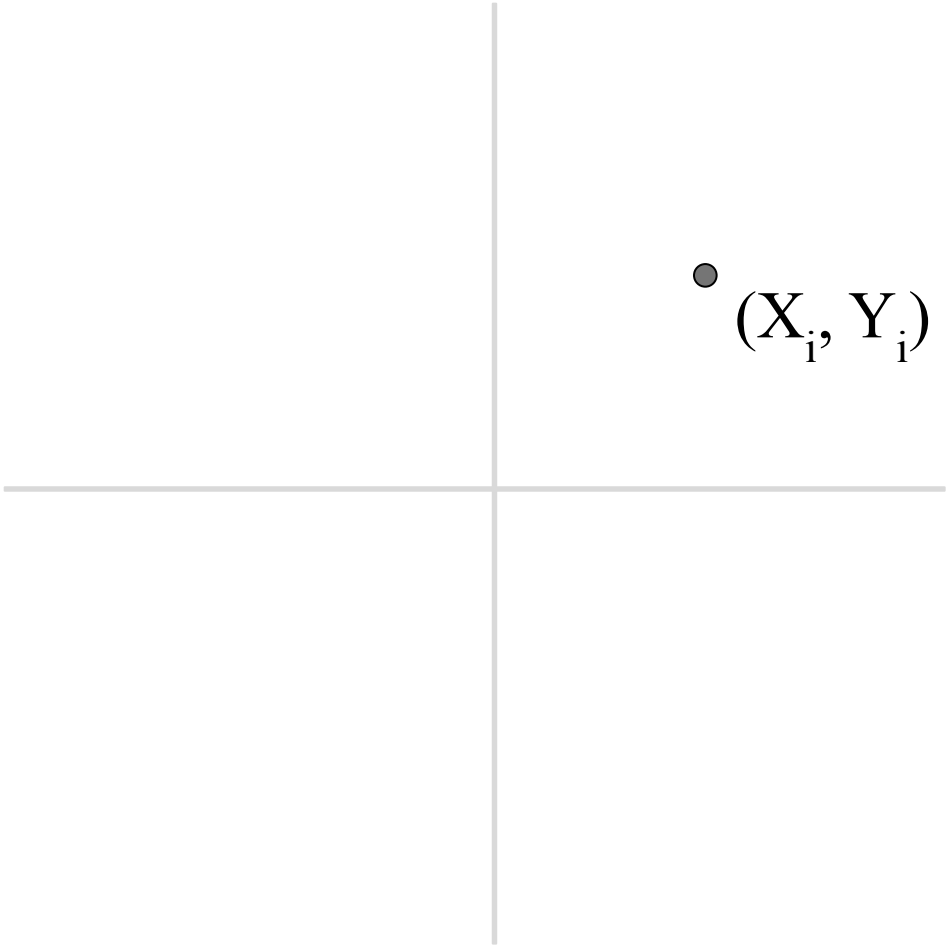
$$\frac{\Sigma_i (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$



—

Sample covariance is the average product of distances to the sample means

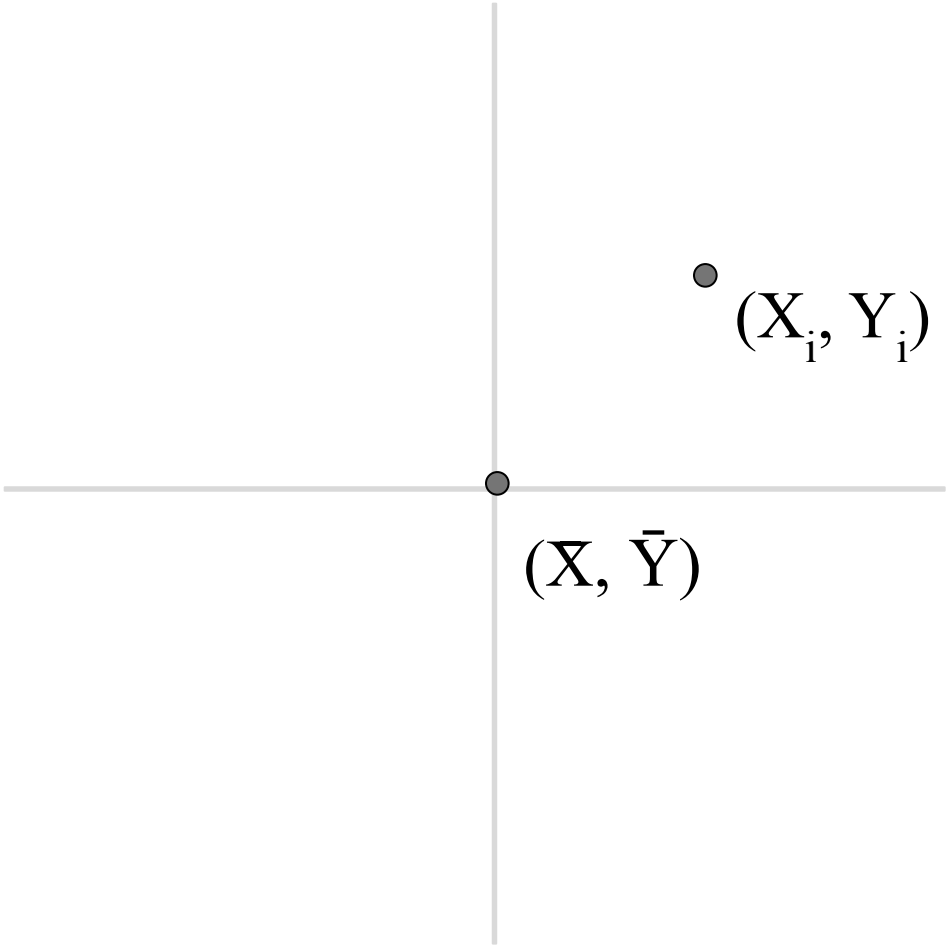
$$\frac{\Sigma_i (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$



—

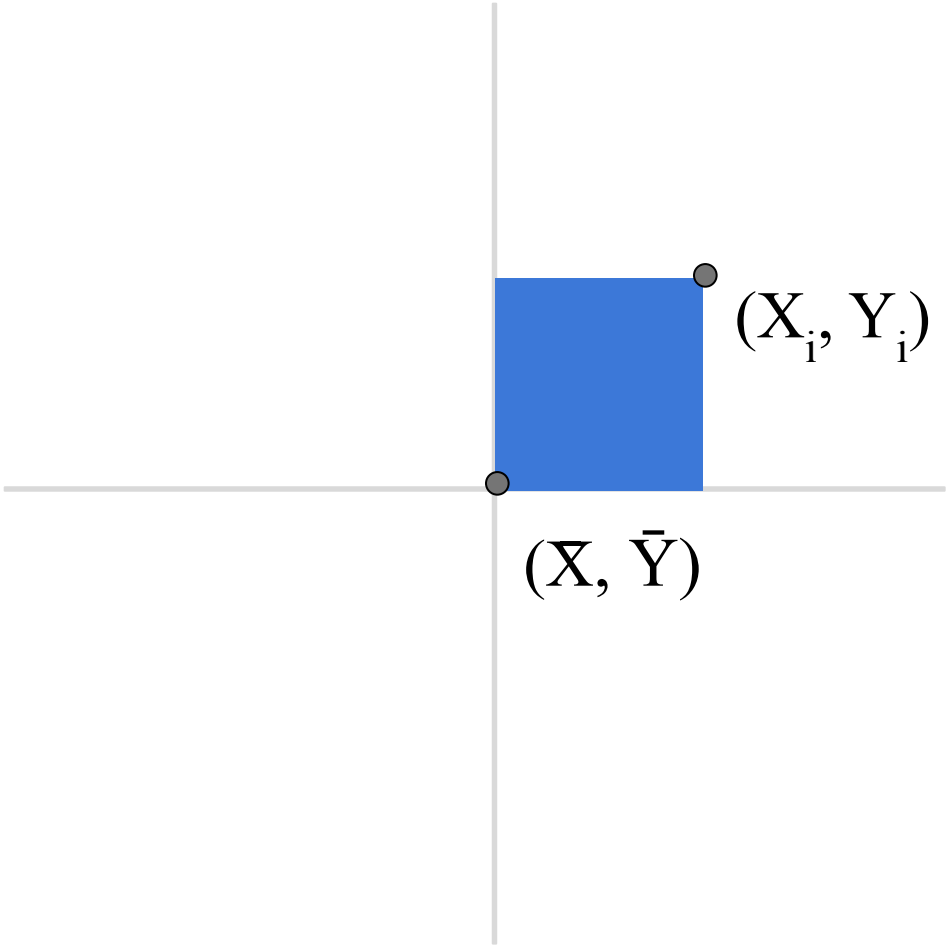
Sample covariance is the average product of distances to the sample means

$$\frac{\Sigma_i (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$



—

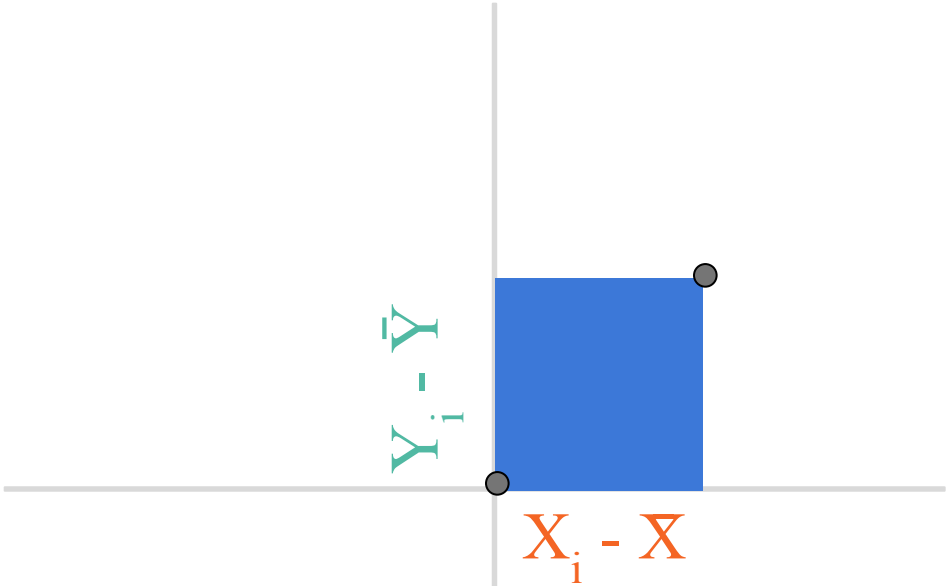
Sample covariance is the average product of distances to the sample means



—

Sample covariance is the average product of distances to the sample means

$$\frac{\Sigma_i (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

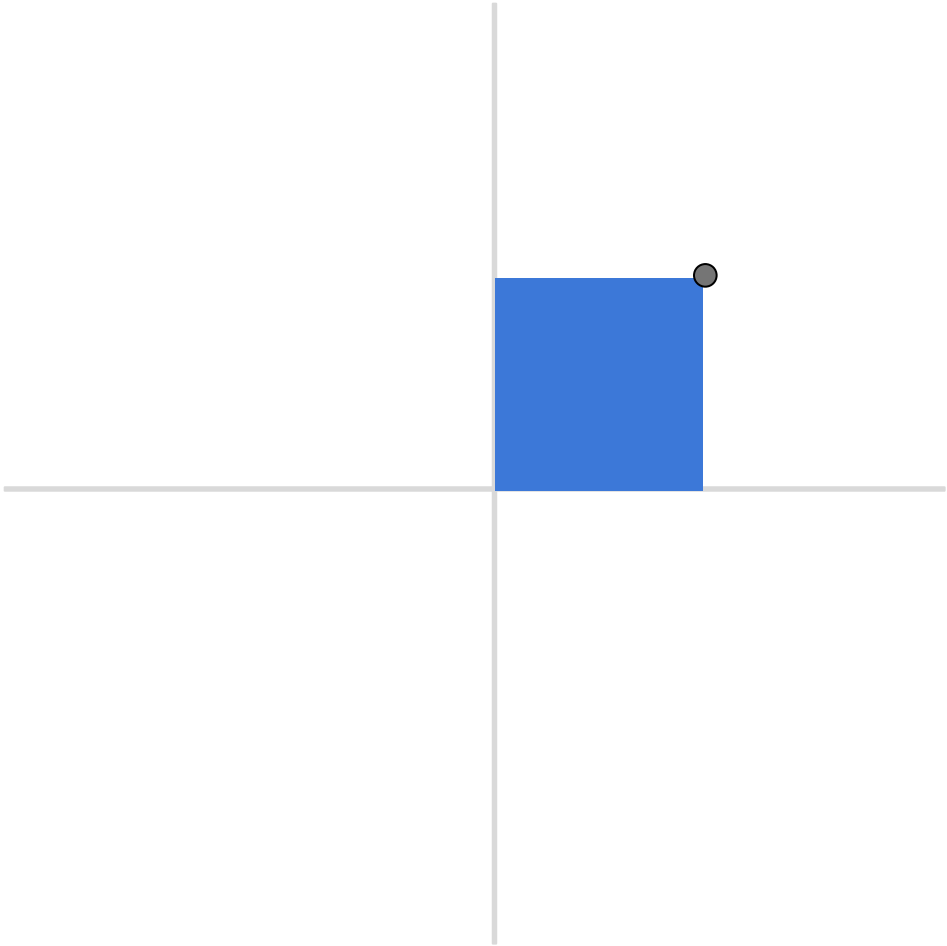


Area of square:

$$(X_i - \bar{X})(Y_i - \bar{Y})$$

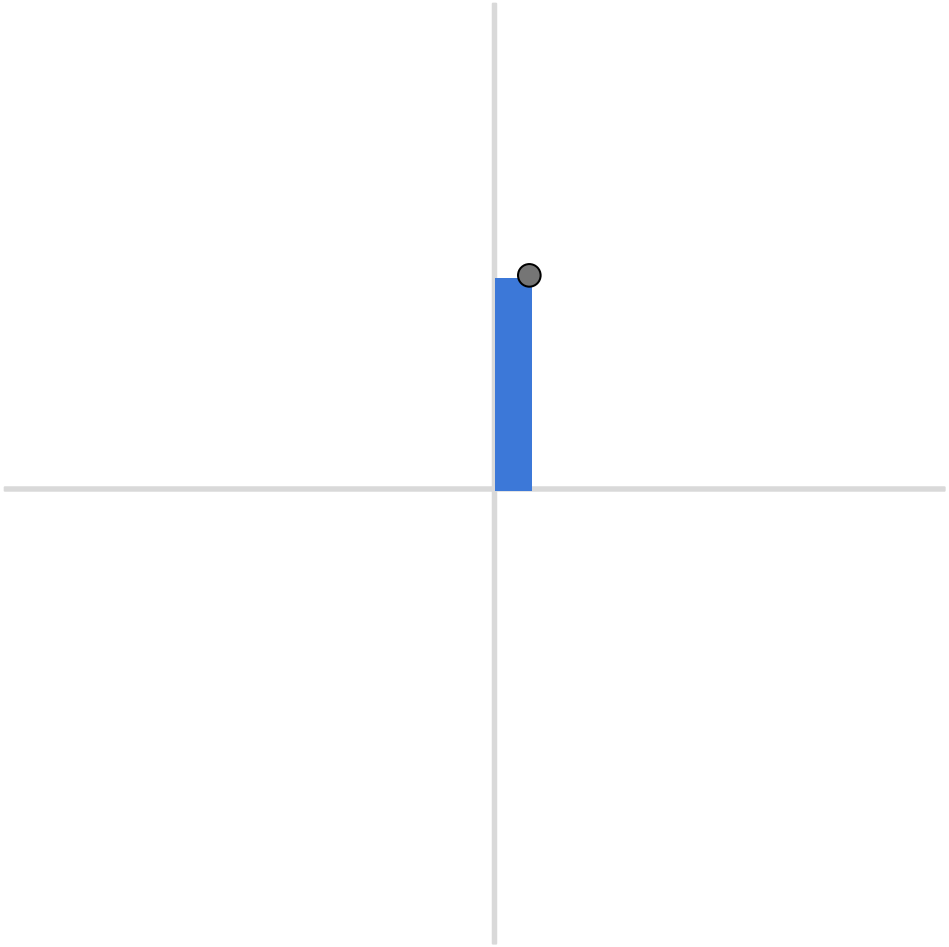
—

When is this product large?



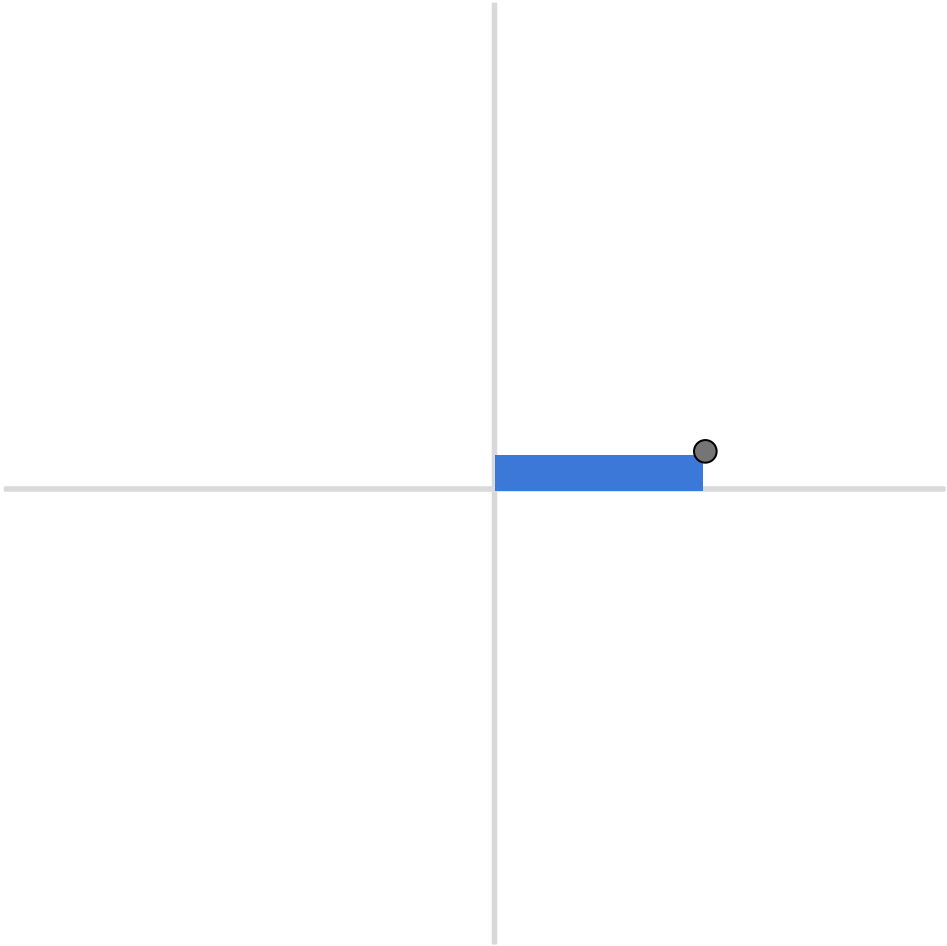
—

When is this product large?



—

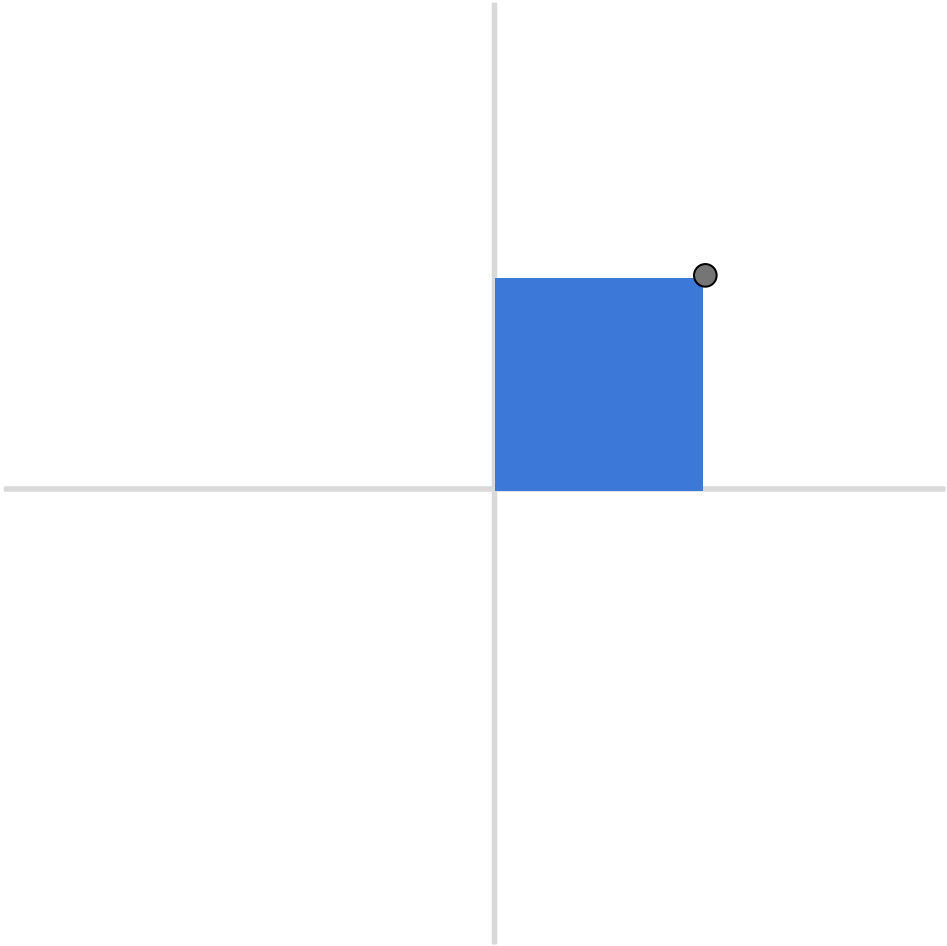
When is this product large?



—

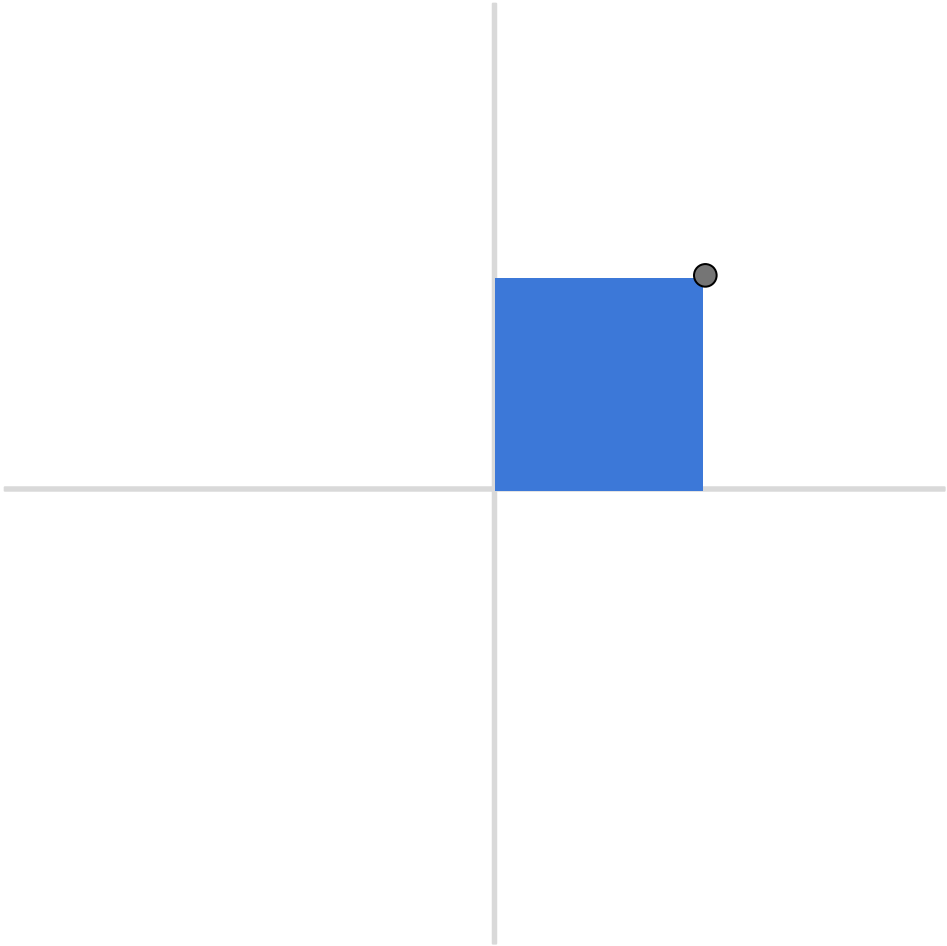
When is this product large?

When both X and Y are far above their means



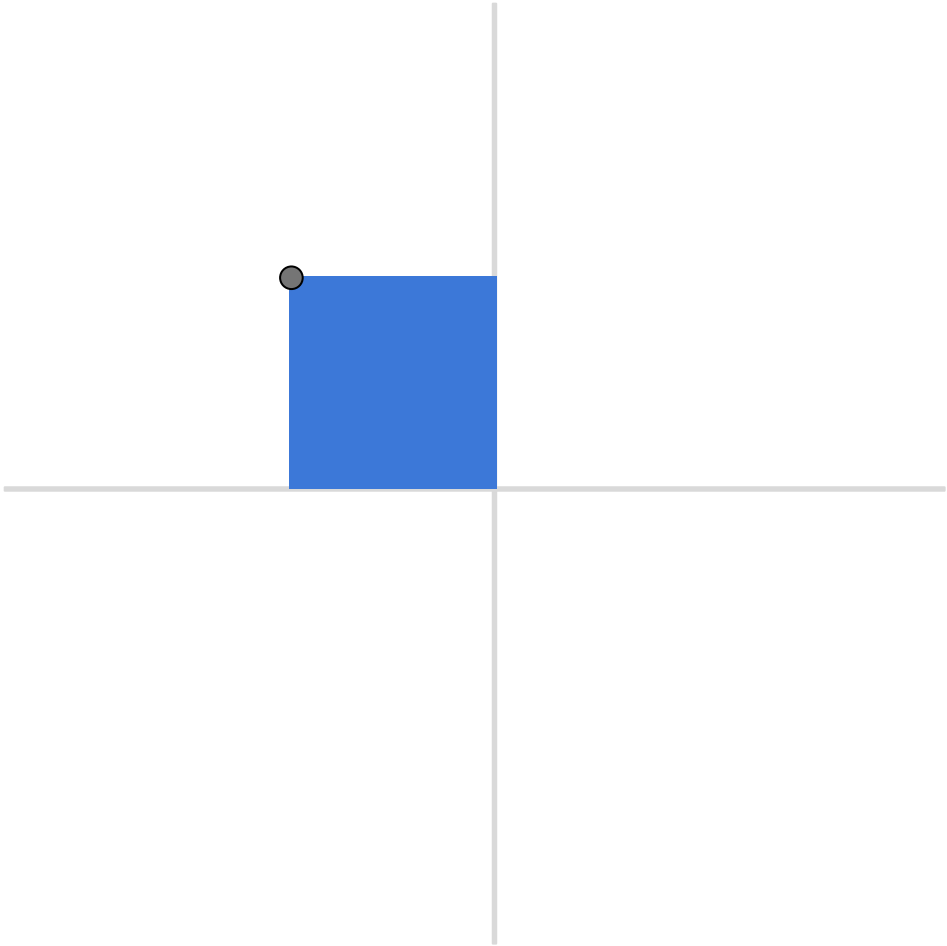
—

What is the sign of the product of the two sides?



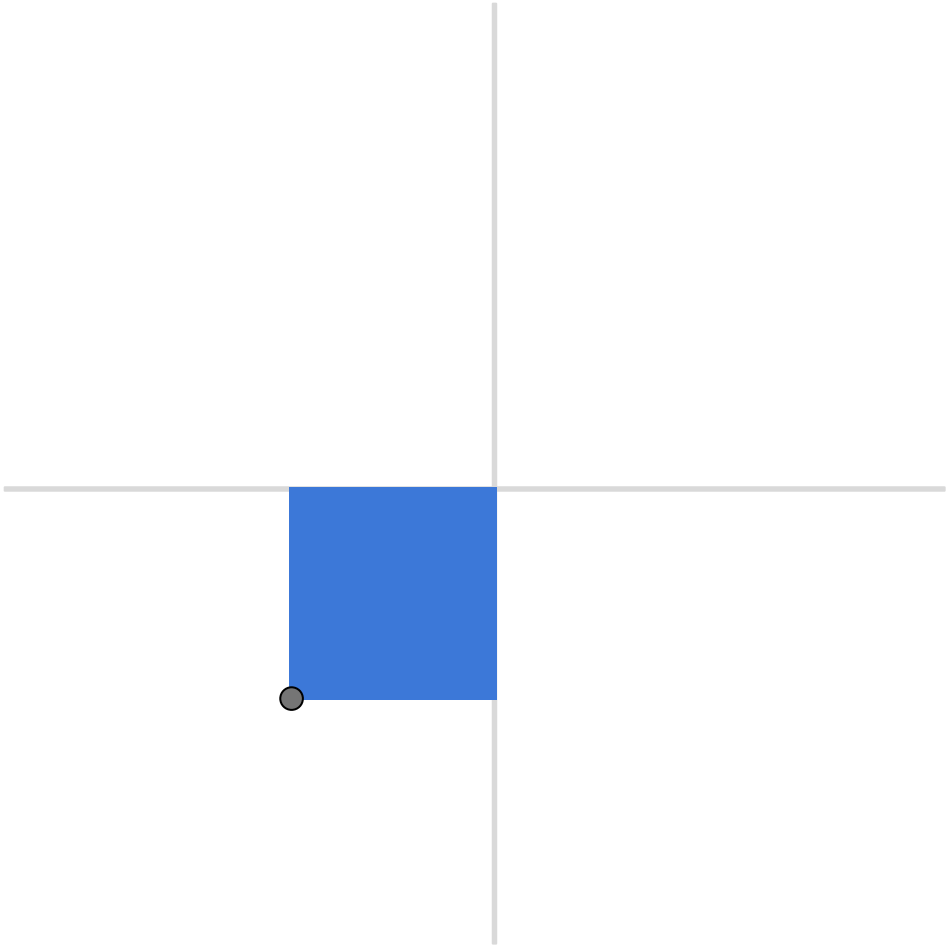
—

What is the sign of the product of the two sides?



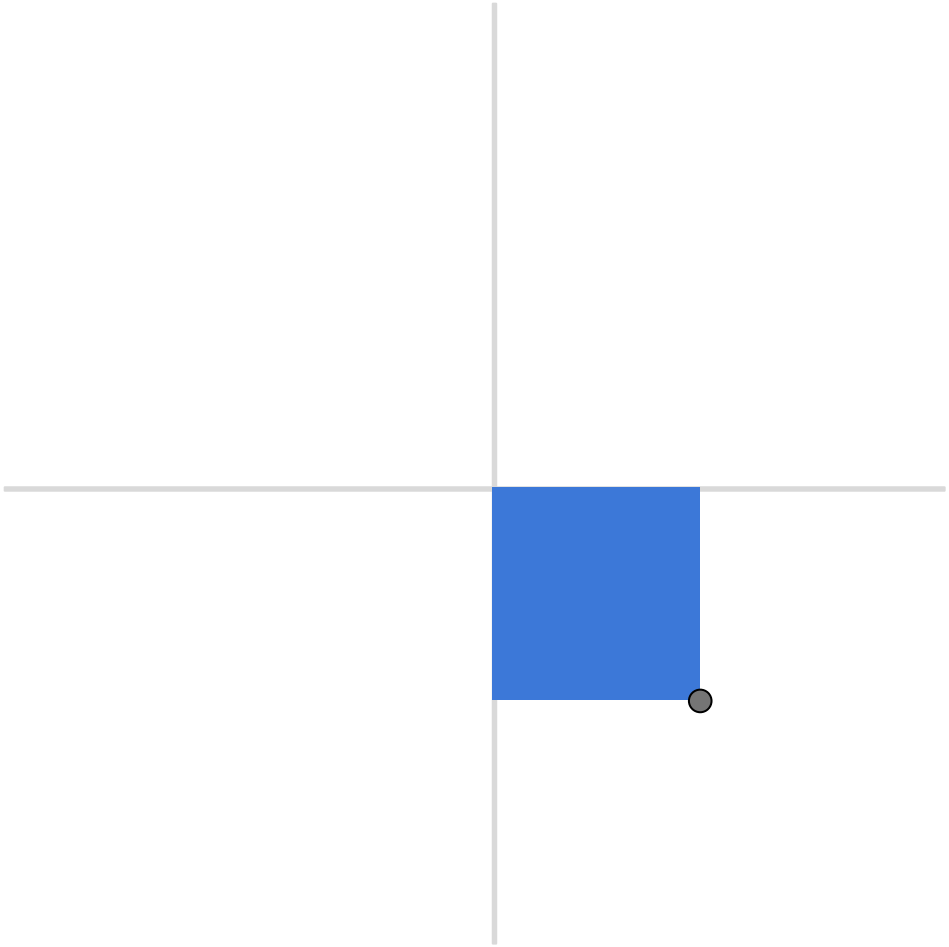
—

What is the sign of the product of the two sides?



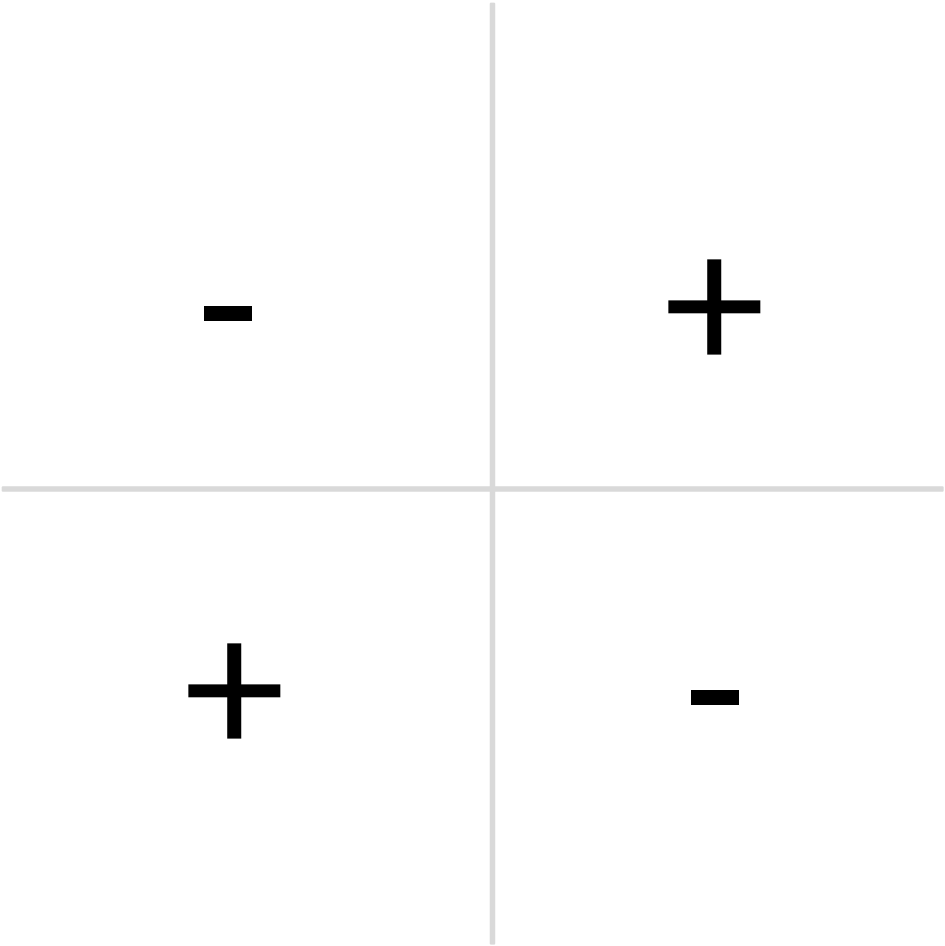
—

What is the sign of the product of the two sides?



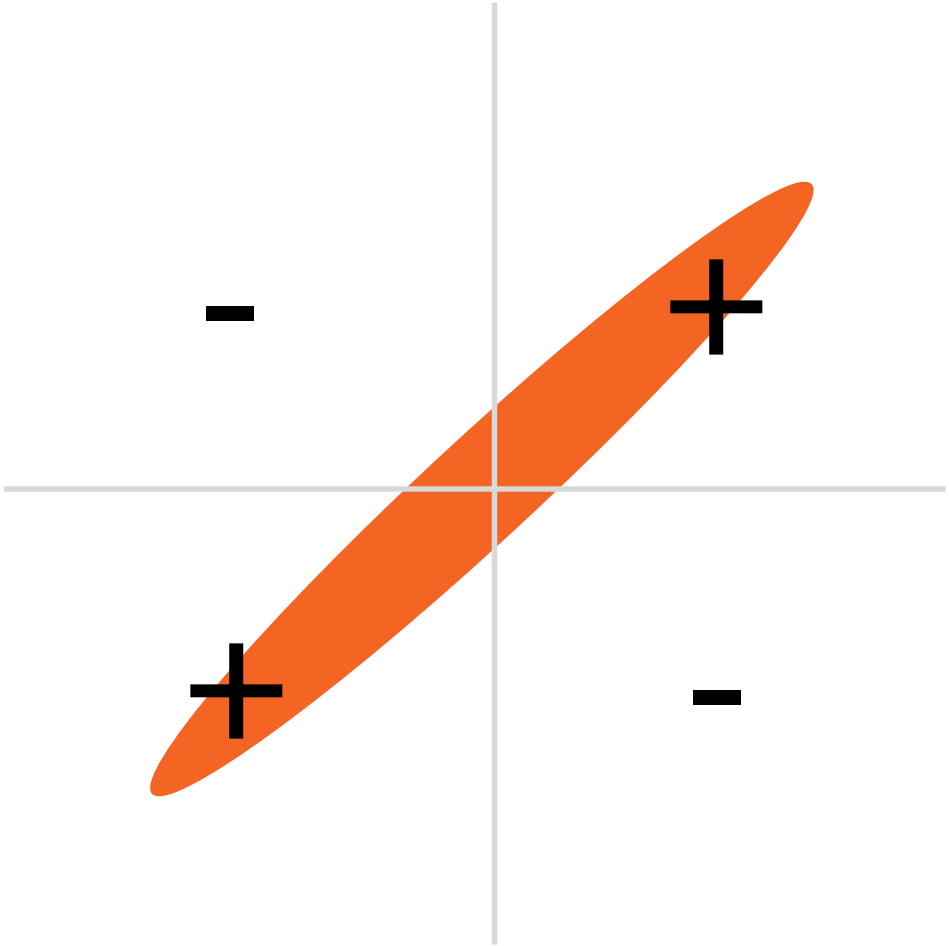
—

The sign measures how X and Y predict each other



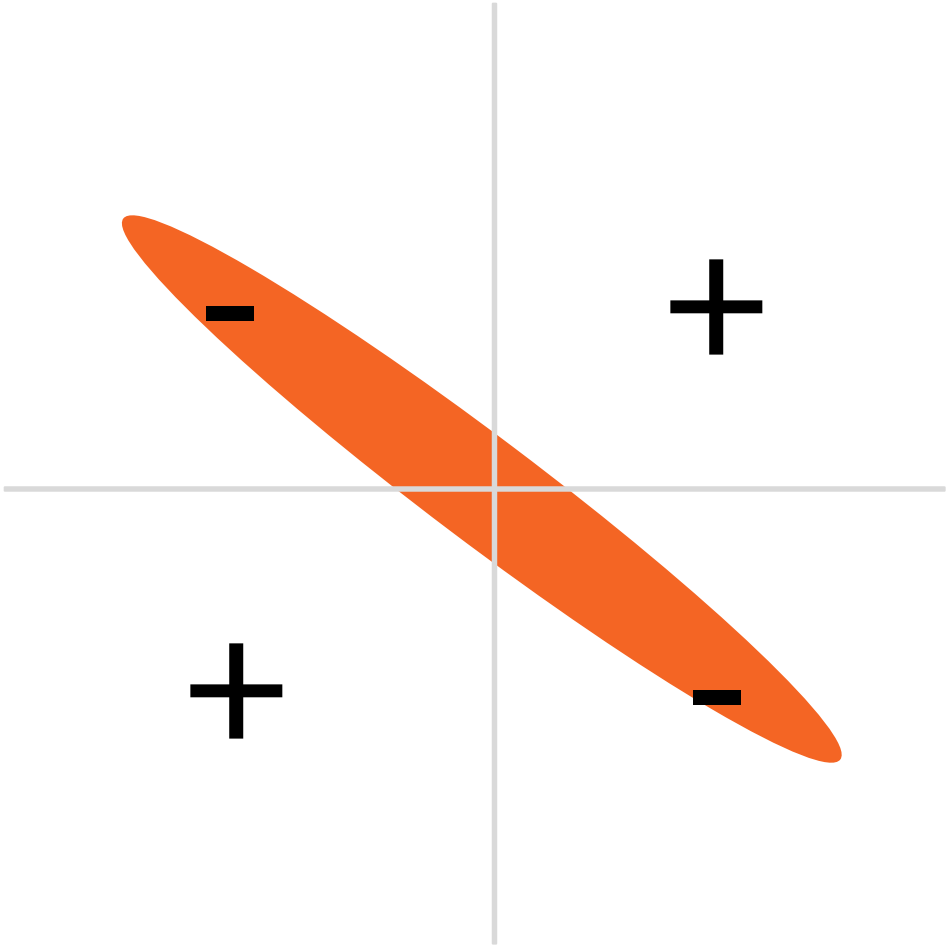
—

most positive covariance: close to the diagonal, in the positive quadrants



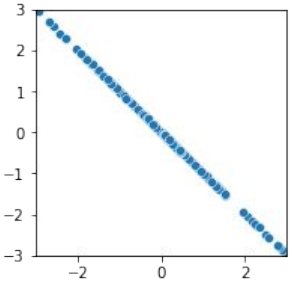
—

most negative covariance: close to the diagonal, in the positive quadrants

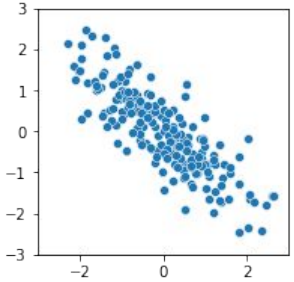


Match the plot to the value of covariance

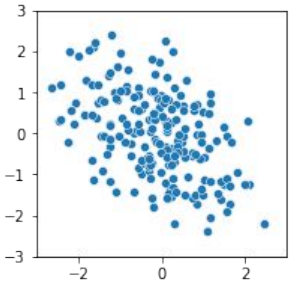
A



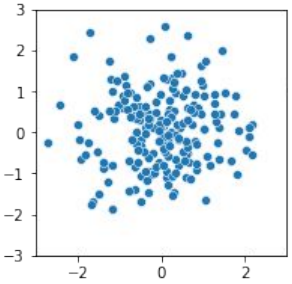
B



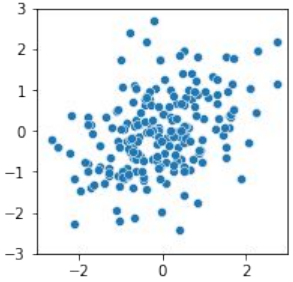
C



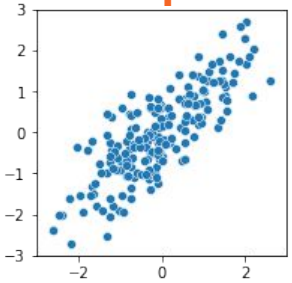
D



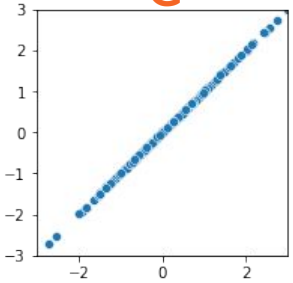
E



F



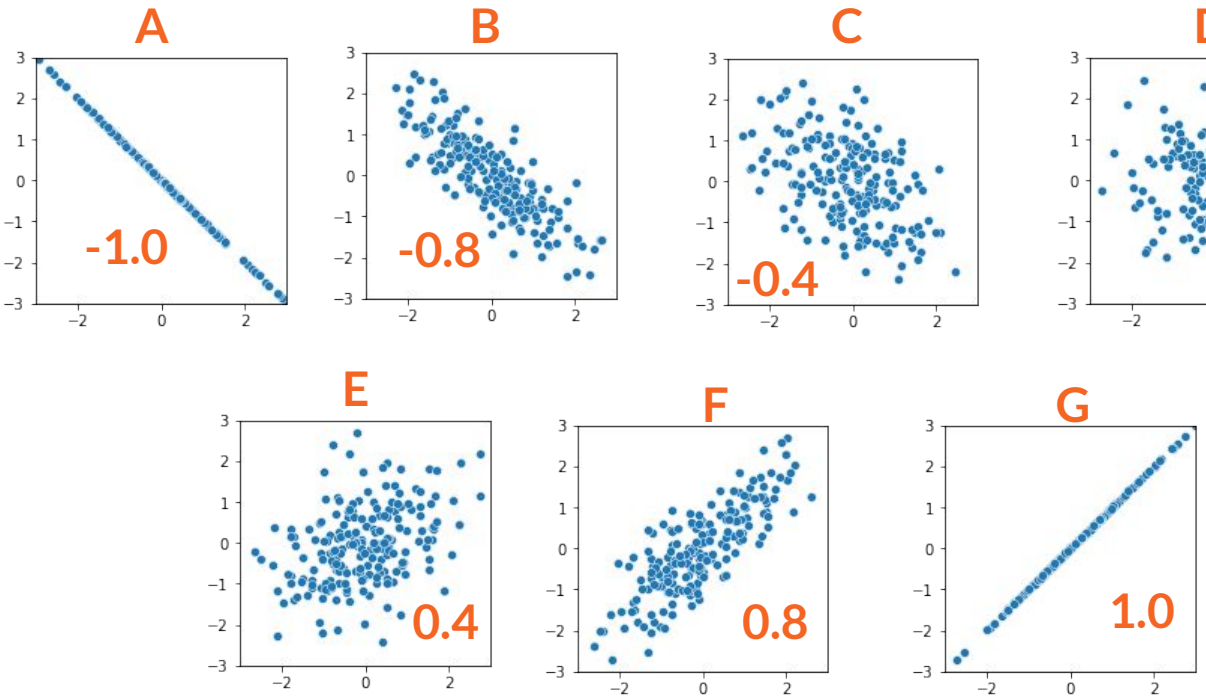
G



Number Bank:

1.0
0.8
0.4
0.0
-0.4
-0.8
-1.0

Match the plot to the value of covariance



Number Bank:

1.0
0.8
0.4
0.0
-0.4
-0.8
-1.0

—

What do we call the covariance of a variable X with itself?

$$\frac{\Sigma_i (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

—

What do we call the covariance of a variable X with itself?

Now Y = X

$$\frac{\Sigma_i (X_i - \bar{X})\overset{(X_i - \bar{X})}{\cancel{(Y_i - \bar{Y})}}}{N}$$

$$(X_i - \bar{X})$$

$$\frac{\Sigma_i (X_i - \bar{X})\cancel{(Y_i - \bar{Y})}}{N}$$

What do we call the covariance of a variable X with itself?

Now Y = X

Which gives us the variance!

$$\frac{\Sigma_i (X_i - \bar{X})^2}{N}$$

—

What happens to covariance if we add a constant to all the X values? Why?

$$\frac{\Sigma_i (X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

—

What happens to covariance if we add a constant to all the X values? Why?

$$\frac{\Sigma_i (X_i^{+c} - \bar{X}^{+c})(Y_i - \bar{Y})}{N}$$

The **mean** of our new X values is: $[(X_1+c) + (X_2+c) + (X_3+c) \ldots] / N$

$= [X_1 + X_2 + X_3 + \ldots + c + c + c \ldots] / N$

= our original mean(X) + c

—

What happens to covariance if we add a constant to all the X values? Why?

$$\frac{\Sigma_i (X_i \cancel{+c} - \bar{X})(Y_i \cancel{+c} - \bar{Y})}{N}$$

But because we're subtracting each of the updated values from the new mean, the +c's all cancel out!

So, **nothing happens to the covariance** since we just care about differences from the means.

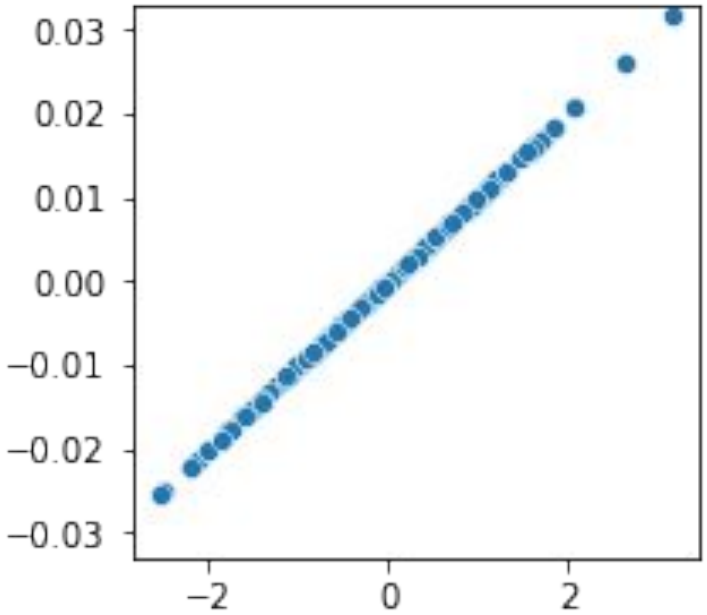
Attendance & 1 min break

tinyurl.com/2uf9uy26

"It's important to remember that correlation does not imply causation. Besides, we all know it was Brian."

https://andertoons.com/science/cartoon/7252/correlation-does-not-imply-causation-we-know-it-was-brian

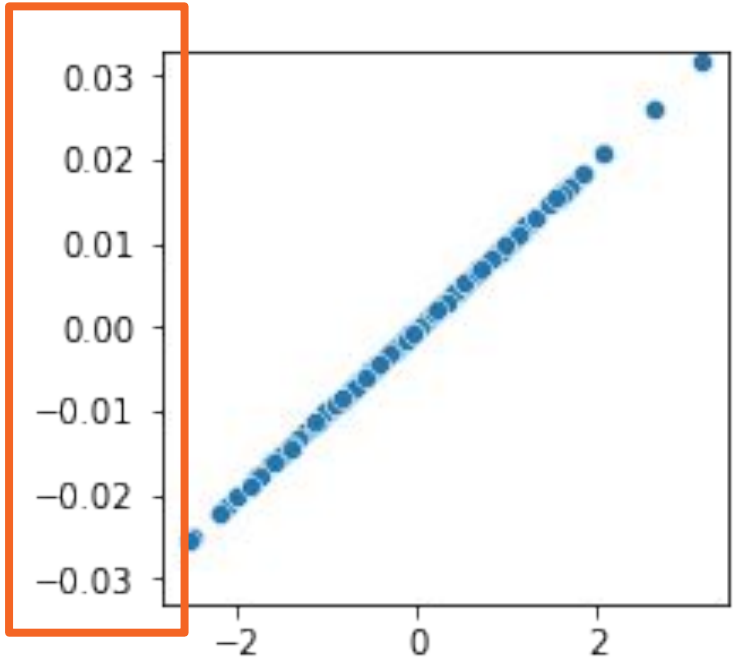
Refresher: what is the covariance here?



Refresher: what is the covariance here?

0.01 (!!)

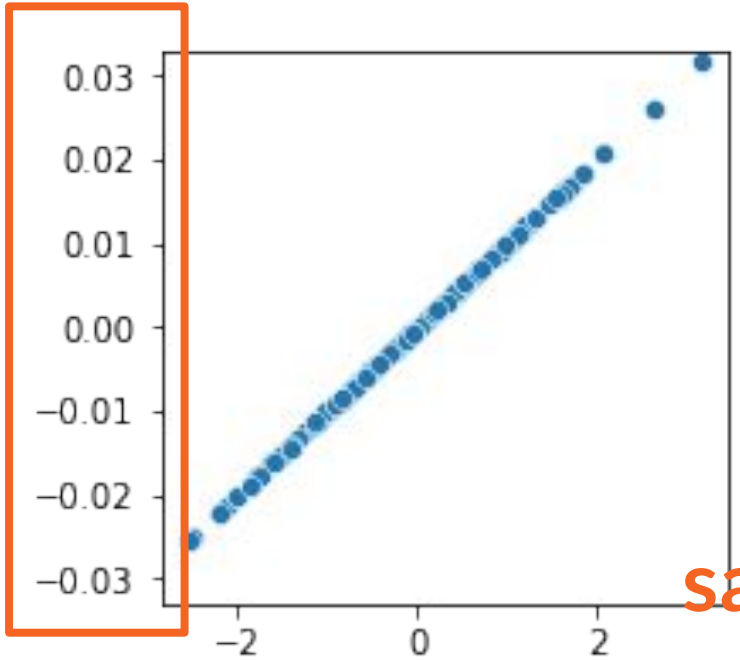
Trick question: look at the axes!!

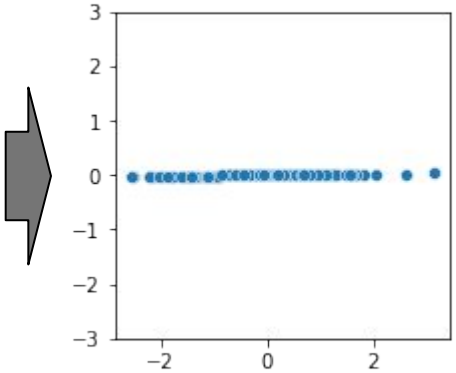


Refresher: what is the covariance here?

0.01 (!!)

Trick question: look at the axes!!



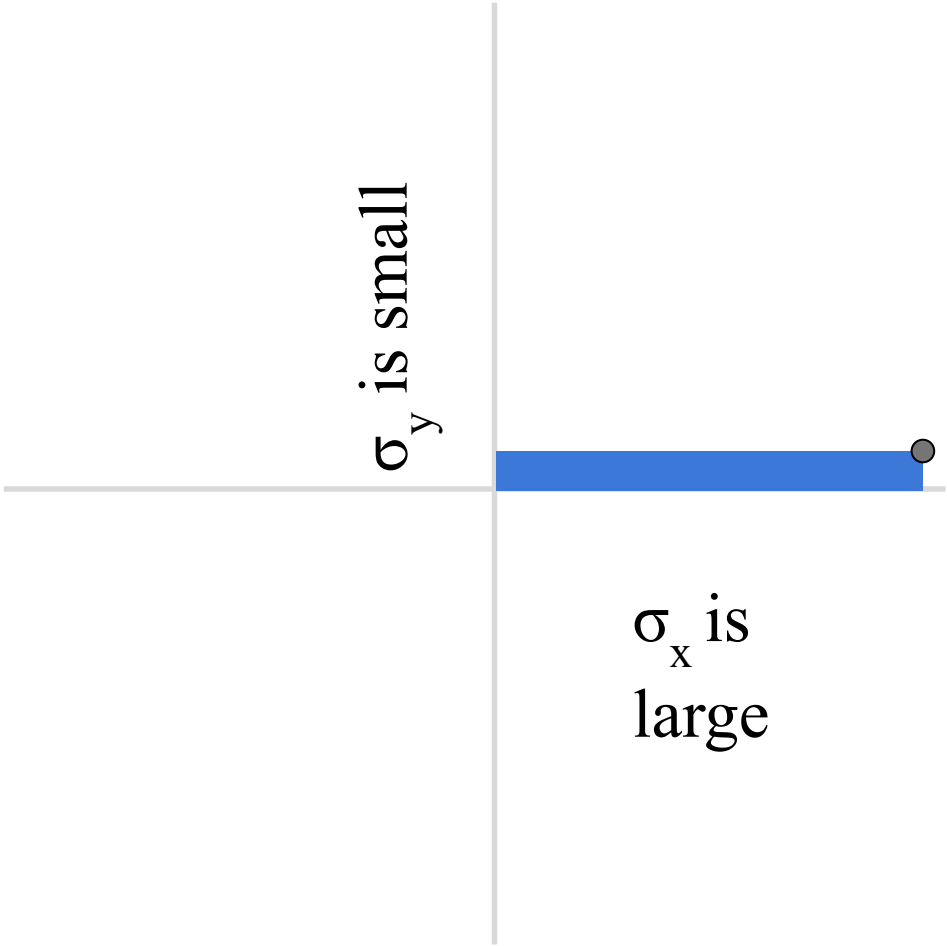


same Y, different ylim()

—

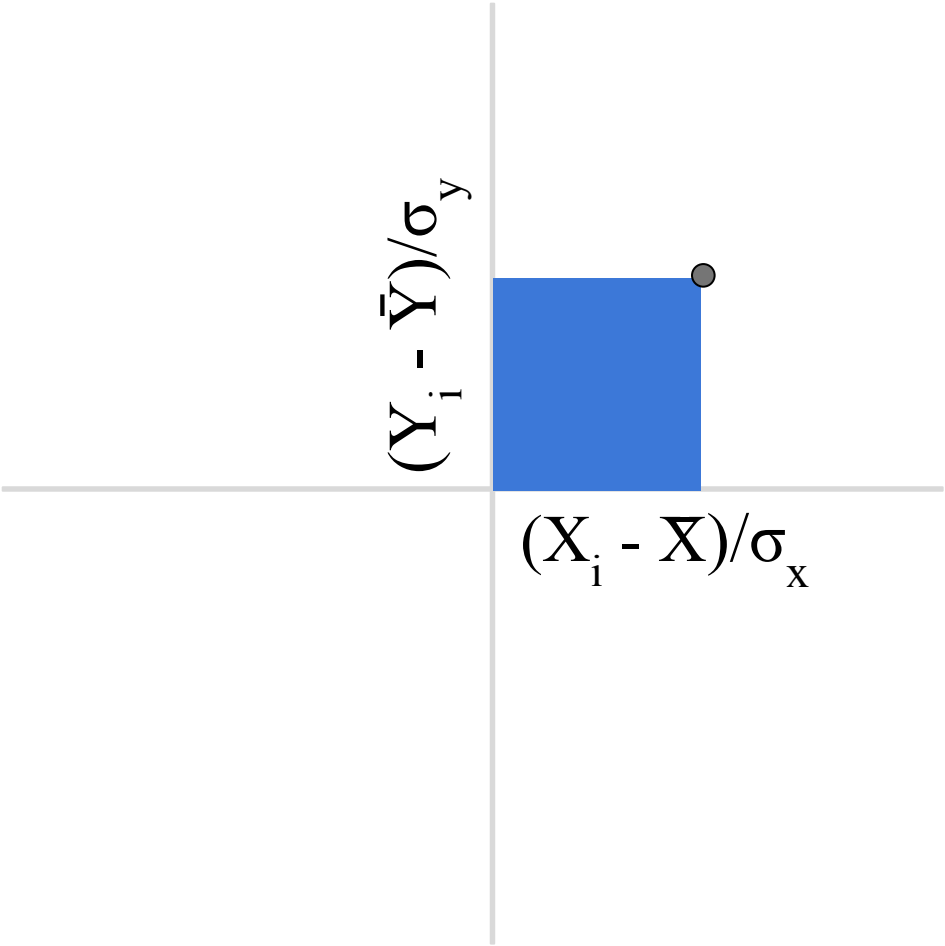
Covariance depends on how much X and Y vary individually

Example: country population vs average age



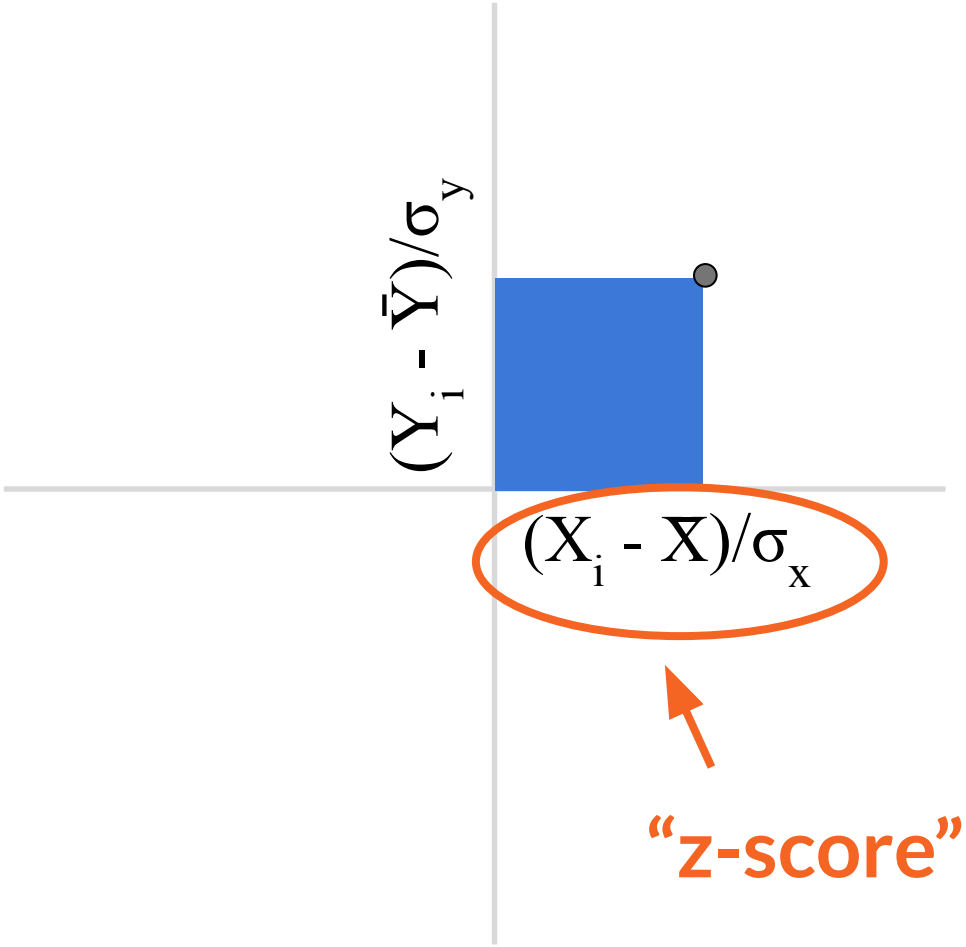
Subtracting the mean and dividing by the std. dev. *normalizes* the variables

Values are now comparable, regardless of units



Subtracting the mean and dividing by the std. dev. *normalizes* the variables

Values are now comparable, regardless of units



—

What if we calculate covariance with z-scores instead?

$$\frac{\Sigma_i ((X_i - \bar{X})/\sigma_x)((Y_i - \bar{Y})/\sigma_y)}{N}$$

—

$$\frac{\Sigma_i ((X_i - \bar{X})/\sigma_x)((Y_i - \bar{Y})/\sigma_y)}{N}$$

Std. dev. is the same for all Xs and the same for all Ys...

$$\frac{\Sigma_i (X_i - \bar{X})(Y_i - \bar{Y})/(\sigma_x \sigma_y)}{N}$$

$$\frac{\Sigma_i ((X_i - \bar{X})/\sigma_x)((Y_i - \bar{Y})/\sigma_y)}{N}$$

Familiar?

$$\frac{\Sigma_i (X_i - \bar{X})(Y_i - \bar{Y})/(\sigma_x \sigma_y)}{N}$$

—

Covariance divided by the product of standard deviations is correlation

$$\frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y}$$

Covariance vs Correlation

- Covariance measures the direction of the relationship between two variables X, Y
- Correlation measures the strength (and direction) of the relationship between two variables X, Y

—

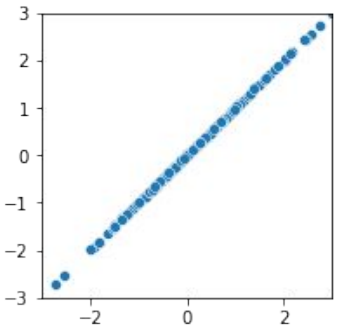
What is the correlation of a variable X with itself?

$$\frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y}$$

—

What is the correlation of a variable X with itself?

i	X	Y
1	87	87
2	90	90
3	84	84
…	…	…



Come up with toy examples for intuition

—

What is the correlation of a variable X with itself?

$$\frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y}$$

$\mathrm{corr}(X,X) = \mathrm{cov}(X,X) / \sigma_x \sigma_x = \mathrm{var}(X) / \sigma_x^2 = \boxed{}$

1

—

When is correlation equal to covariance?

$$\frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y}$$

—

When is correlation equal to covariance?

$$\frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y}$$

When the standard deviations multiply to 1

When is correlation *not* equal to covariance?

	Country	Continent	Region	Location	Area	Borders	Length	Coastline	Highest Point	Height	...	S
0	Indonesia	Asia	South-eastern Asia	5 00 S, 120 00 E	1904569.0	3	2958.0	54716.0	Puncak Jaya	4884	...	
1	Panama	Americas	Central America	9 00 N, 80 00 W	75420.0	2	687.0	1519.0	Volcán Barú	3475	...	
2	China	Asia	Eastern Asia	35 00 N, 105 00 E	9596960.0	14	22457.0	14500.0	Mount Everest	8848	...	
3	Japan	Asia	Eastern Asia	36 00 N, 138 00 E	377915.0	0	0.0	29751.0	Mount Fuji	3776	...	
4	United States	Americas	Northern America	38 00 N, 97 00 W	9833517.0	2	12048.0	20010.0	Denali	6191	...	

Covariance between pairs of variables

```
countries["Age"].cov(countries["Change"])
```

```
-7.50165454365079
```

covariance in Python

Covariance between pairs of variables

X Y

```
countries["Age"].cov(countries["Change"])
```

```
-7.50165454365079
```

Covariance between pairs of variables

average age of people in country

```
countries["Age"].cov(countries["Change"])
```

```
-7.50165454365079
```

```
countries["Population"].cov(countries["Change"])
```

```
2135775.8711537886
```

total population

% change in population

If A has higher Age than B, do you expect A to have higher or lower Change than B?

```
countries["Age"].cov(countries["Change"])
```

```
-7.50165454365079
```

```
countries["Population"].cov(countries["Change"])
```

```
2135775.8711537886
```

If A has higher Age than B, do you expect A to have higher or lower Change than B?

Covariance is negative, so higher age is associated with slower growth

```
countries["Age"].cov(countries["Change"])
```

```
-7.50165454365079
```

```
countries["Population"].cov(countries["Change"])
```

```
2135775.8711537886
```

If A has higher Pop. than B, do you expect A to have higher or lower Change than B?

```
countries["Age"].cov(countries["Change"])
```

```
-7.50165454365079
```

```
countries["Population"].cov(countries["Change"])
```

```
2135775.8711537886
```

If A has higher Pop. than B, do you expect A to have higher or lower Change than B?

```
countries["Age"].cov(countries["Change"])
```

```
-7.50165454365079
```

```
countries["Population"].cov(countries["Change"])
```

```
2135775.8711537886
```

Covariance is positive, so more populous countries tend to grow faster

Which of these covariances matters more?

```
countries["Age"].cov(countries["Change"])
```

```
-7.50165454365079
```

```
countries["Population"].cov(countries["Change"])
```

```
2135775.8711537886
```

Which of these covariances matters more?

One is bigger,
I guess?
Not sure...

```
countries["Age"].cov(countries["Change"])
```

```
-7.50165454365079
```

```
countries["Population"].cov(countries["Change"])
```

```
2135775.8711537886
```

Which of these covariances matters more?

Remember these numbers:
-7 and **2.1M**

```
countries["Age"].cov(countries["Change"])
```

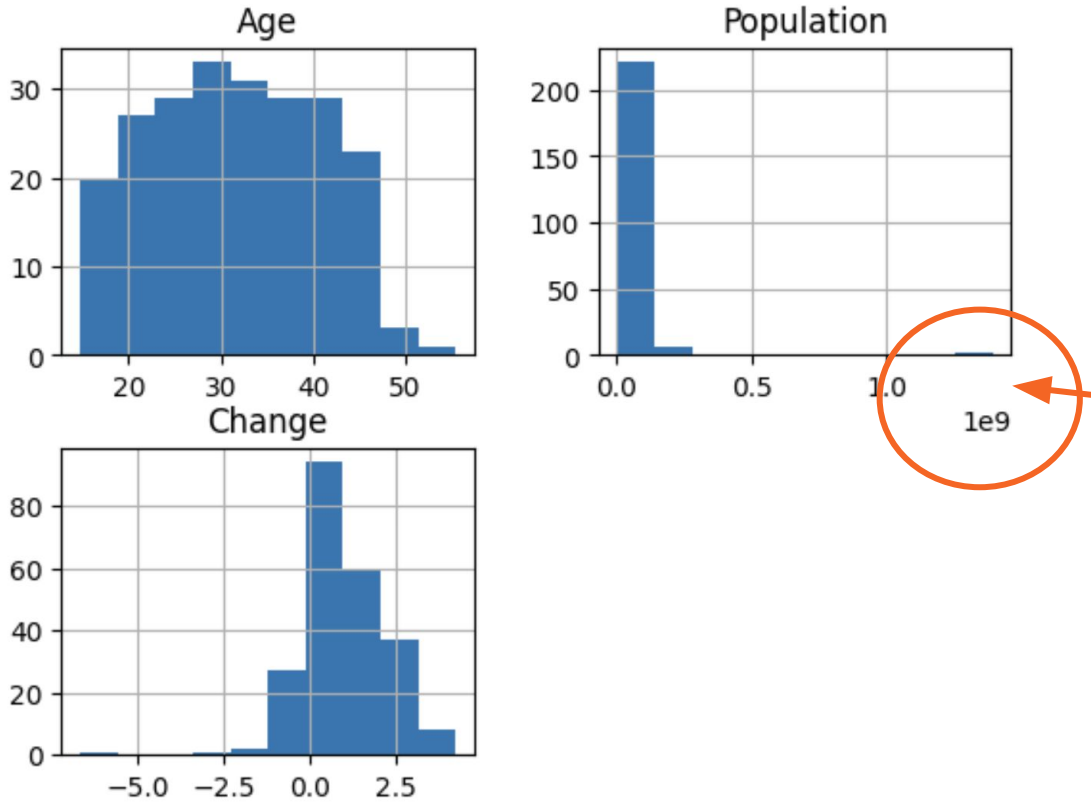
```
-7.50165454365079
```

```
countries["Population"].cov(countries["Change"])
```

```
2135775.8711537886
```

These variables are on vastly different scales

```
countries[ ["Age", "Population", "Change"] ].hist()
pyplot.show()
```



Most countries are much smaller than India and China

The covariance matrix shows cov and var

```
countries[ ["Age", "Population", "Change"] ].cov()
```

	Age	Population	Change
Age	8.152217e+01	-2.783326e+06	-7.501655e+00
Population	-2.783326e+06	1.748766e+16	2.135776e+06
Change	-7.501655e+00	2.135776e+06	1.411307e+00

The covariance matrix shows cov and var

Returns a matrix since doing pairwise covariance on each pair in this array

```
countries[ ["Age", "Population", "Change"] ].cov()
```

	Age	Population	Change
Age	8.152217e+01	-2.783326e+06	-7.501655e+00
Population	-2.783326e+06	1.748766e+16	2.135776e+06
Change	-7.501655e+00	2.135776e+06	1.411307e+00

The covariance matrix shows cov and var

```
countries[ ["Age", "Population", "Change"] ].cov()
```

	Age	Population	Change
Age	8.152217e+01	-2.783326e+06	-7.501655e+00
Population	-2.783326e+06	1.748766e+16	2.135776e+06
Change	-7.501655e+00	2.135776e+06	1.411307e+00

The covariance matrix shows cov and var

```
countries[ ["Age", "Population", "Change"] ].cov()
```

	Age	Population	Change
Age	8.152217e+01	-2.783326e+06	-7.501655e+00
Population	-2.783326e+06	1.748766e+16	2.135776e+06
Change	-7.501655e+00	2.135776e+06	1.411307e+00

kind of hard to read?

Can we turn off scientific notation?

```
# from https://re-thought.com/how-to-suppress-scientific-notation-in-pandas/

pd.options.display.float_format = '{:f}'.format
```

```
# format by itself without () is a value whose type is function

type("{:.2f}".format)
```

```
builtin_function_or_method
```

```
countries[ ["Age", "Population", "Change"] ].cov()
```

The covariance matrix shows cov and var

```
countries[ ["Age", "Population", "Change"] ].cov()
```

	Age	Population	Change
Age	81.522172	-2783325.974836	-7.501655
Population	-2783325.974836	17487655973542300.000000	2135775.871154
Change	-7.501655	2135775.871154	1.411307

whoa.

Take the square root of var to get the std. dev.

```
countries[ ["Age", "Population", "Change"] ].std()
```

```
Age                   9.028963
Population   132240901.288301
Change                1.187984
dtype: float64
```

Do you remember the covariance values?
How do they compare to the standard deviations?

Correlation accounts for the scale of variables

```
countries[ ["Age", "Population", "Change"] ].corr()
```

Toggle output scrolling

	Age	Population	Change
Age	1.000000	-0.002307	-0.698294
Population	-0.002307	1.000000	0.013595
Change	-0.698294	0.013595	1.000000

Correlation accounts for the scale of variables

```
countries[ ["Age", "Population", "Change"] ].corr()
```

Toggle output scrolling

	Age	Population	Change
Age	1.000000	-0.002307	-0.698294
Population	-0.002307	1.000000	0.013595
Change	-0.698294	0.013595	1.000000

Age is *much* more predictive of population change
The correlation between pop and change is ~0

Corr vs. Cov

- Correlation is more informative about relationships than covariance, ***especially*** when X and Y are on very different scales
  - Always check your axis scales to make sure you aren't tricking anyone with your plots!
- Covariance has units (sometimes meaningful) while correlation is unitless

—

Is correlation symmetric? If you swap X and Y, would you get the same value?

$$\frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y}$$

—

$$\frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y}$$

Is correlation symmetric? If you swap X and Y, would you get the same value?

Yes – it doesn't matter which order we multiply the two terms. (Covariance is also symmetric!)

Correlation and causation

Is growth higher *because* the population is younger?

Is the population younger *because* growth is higher?

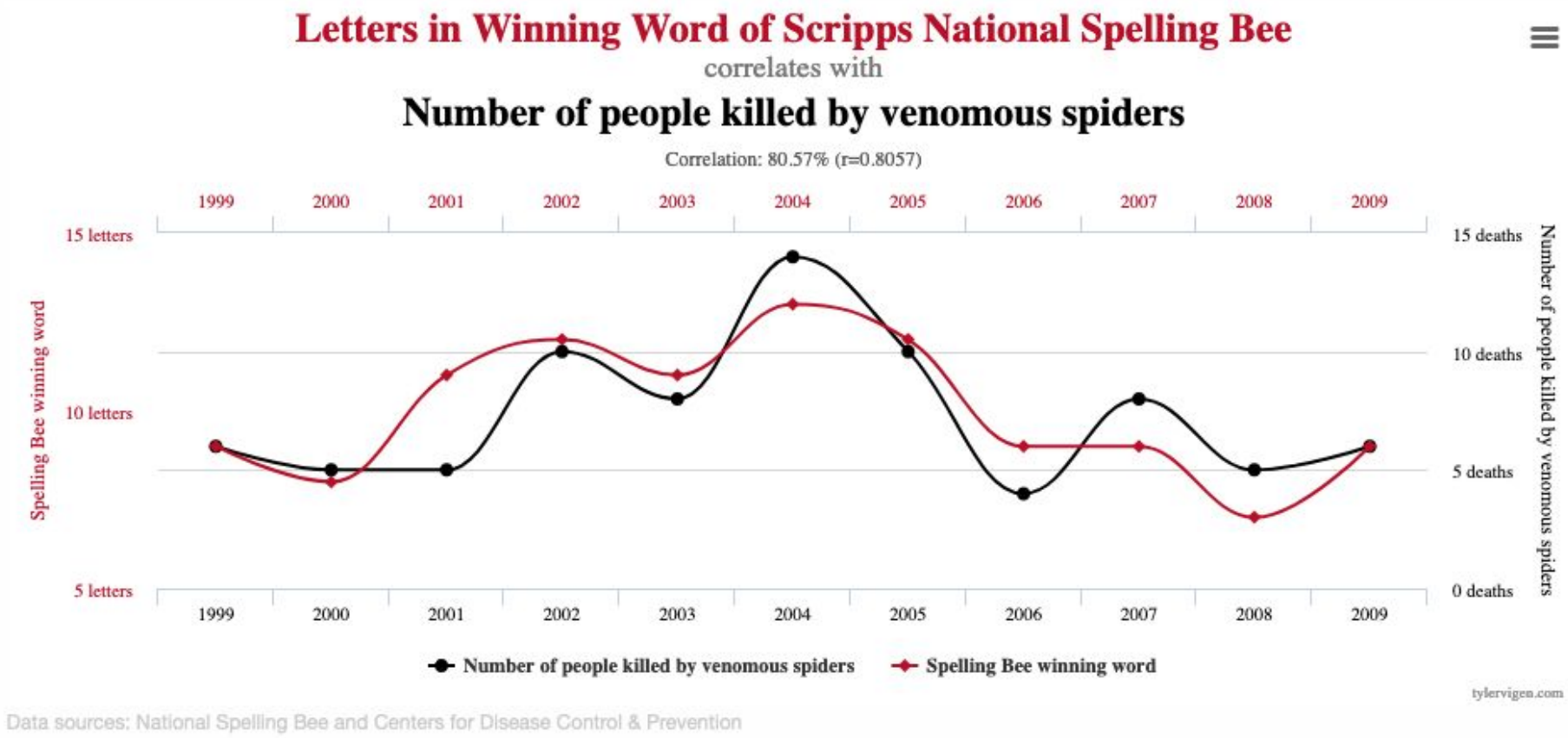
Or both? Or neither?

Correlation does not imply causation

Correlation is symmetric; causation is not! Must have a causal **direction**.

If X causes Y, they will often be correlated. Often there's a third underlying factor that causes both X and Y, so be careful in what you claim!

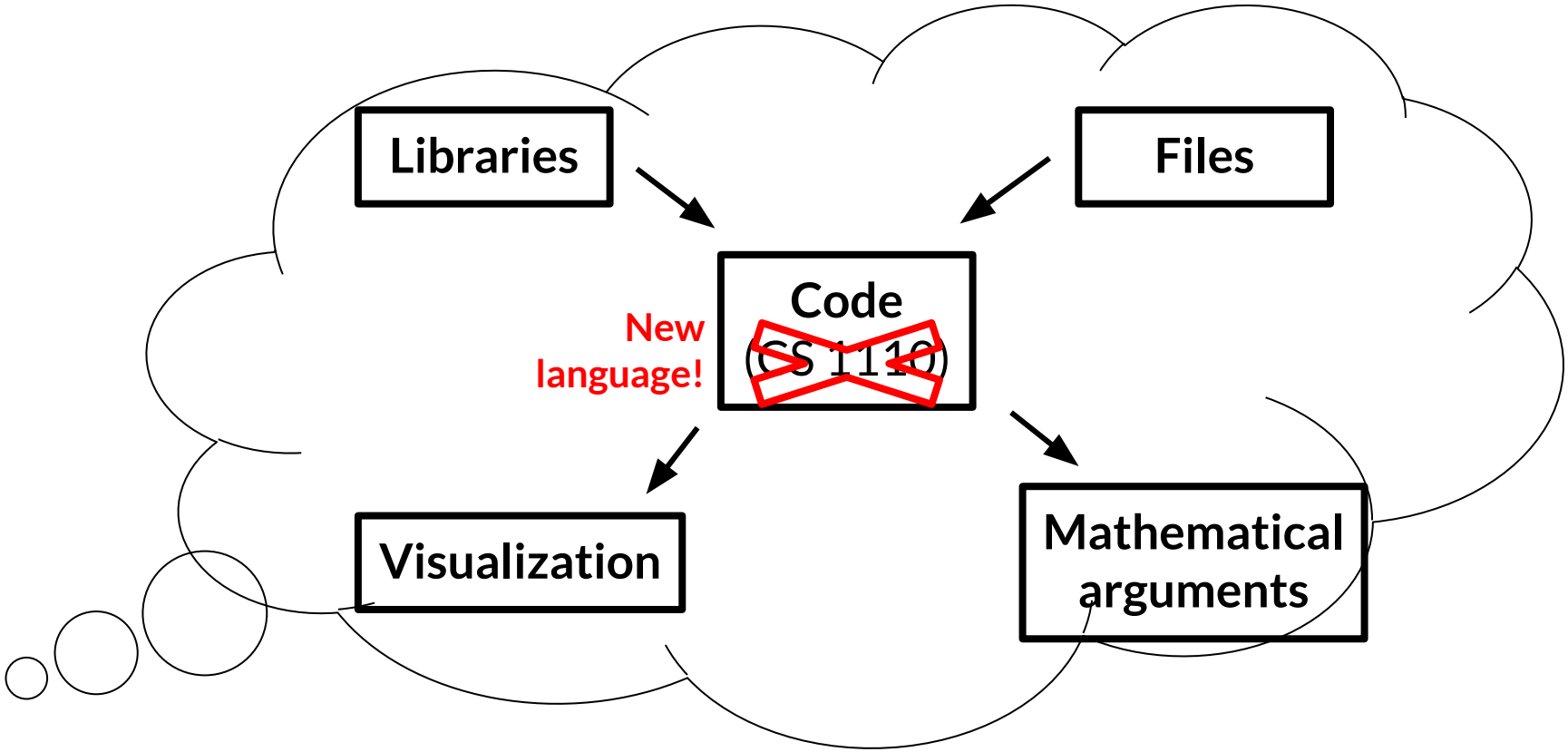
1 minute break



https://tylervigen.com/view_correlation?id=2941

Homework submission snafus (do not do these!!)

- Submitting the html instead of the pdf (or submitting a funky pdf)
- Tagging instructions instead of just solutions
- Tagging the correct question #s
- Putting full names instead of netids
- Not submitting the ipynb file


Libraries
Files
Code
(CS 1110)
New language!
Visualization
Mathematical arguments

HTML

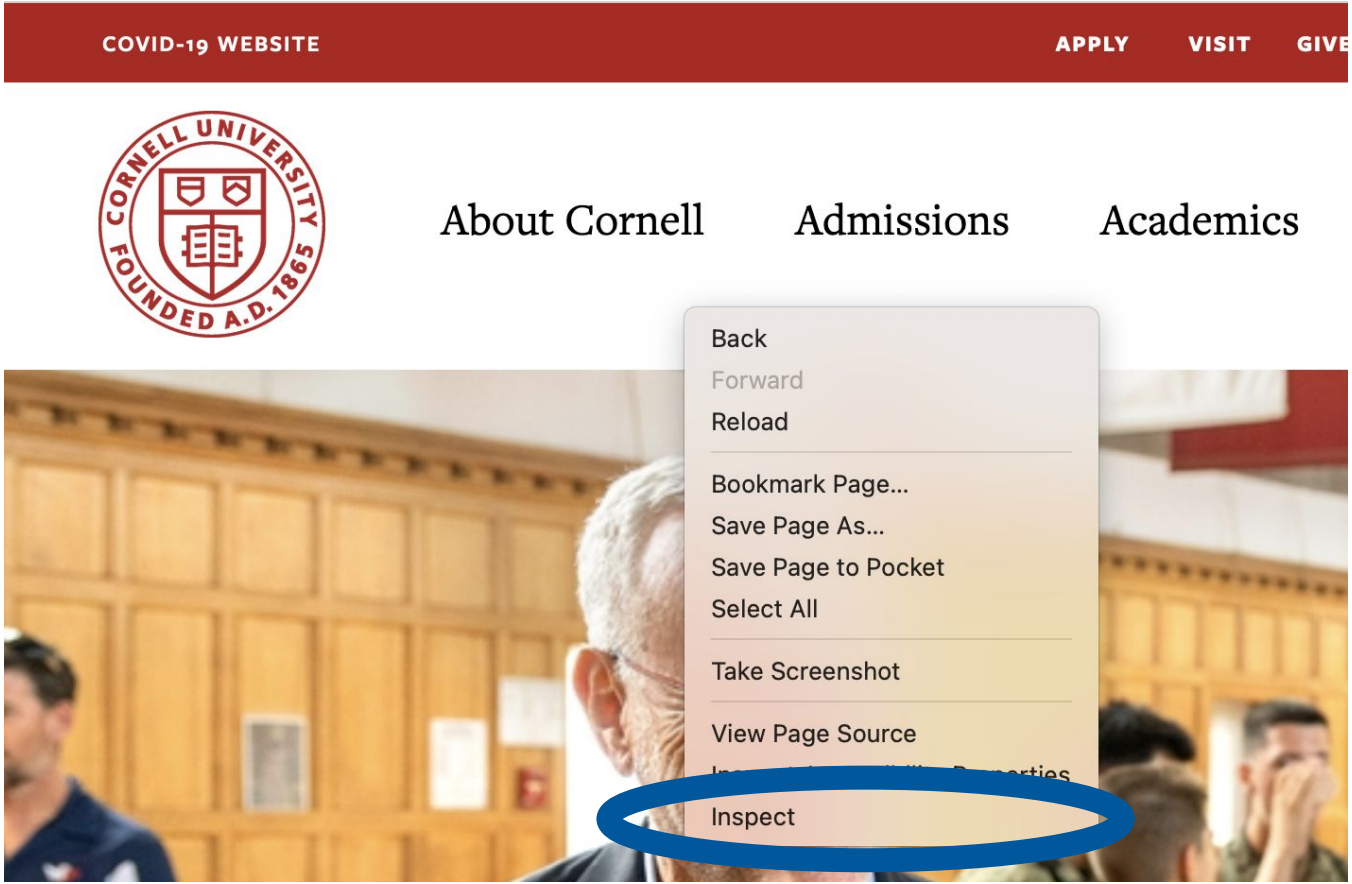
The HyperText Markup Language or HTML is the standard markup language for documents designed to be displayed in a web browser. It defines the meaning and structure of web content. It is often assisted by technologies such as Cascading Style Sheets and scripting languages such as JavaScript. Wikipedia

https://en.wikipedia.org/wiki/HTML

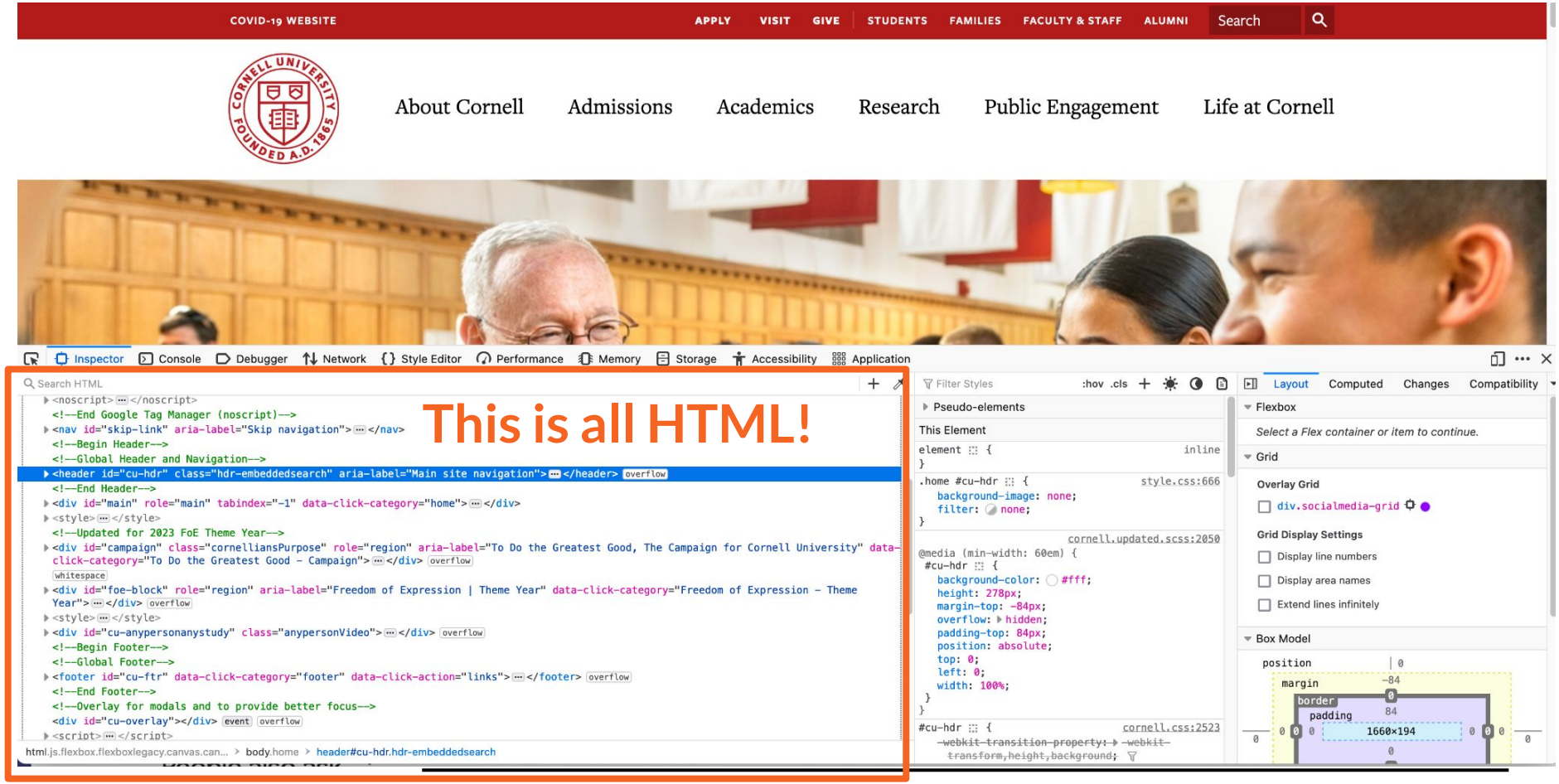
What is HTML?

- The markup language used to make websites!
- Why do we care about this in data science?
- Because we often get data directly *from websites*
    - Data scraping
    - Friday discussions = tutorial

Right click on any website…



https://www.cornell.edu/



https://www.cornell.edu/

Let's make a website that displays:

This is a Heading

This is a paragraph.

https://www.w3schools.com/html/

We need to use HTML

```
<!DOCTYPE html>
<html>
<head>
<title>Page Title</title>
</head>
<body>

<h1>This is a Heading</h1>

<p>This is a paragraph.</p>

</body>
</html>
```

https://www.w3schools.com/html/

Everything except for the actual text is in tags <>

```
<!DOCTYPE html>
<html>
<head>
<title>Page Title</title>
</head>
<body>

<h1>This is a Heading</h1>

<p>This is a paragraph.</p>

</body>
</html>
```

Every <> must be closed with </>

```
<!DOCTYPE html>
<html>
<head>
<title>Page Title</title>
</head>
<body>

<h1>This is a Heading</h1>

<p>This is a paragraph.</p>

</body>
</html>
```

https://www.w3schools.com/html/

Visible parts of the website are in the body

```
<!DOCTYPE html>
<html>
<head>
<title>Page Title</title>
</head>
<body>

<h1>This is a Heading</h1>

<p>This is a paragraph.</p>

</body>
</html>
```

What do the different elements do?

```
<!DOCTYPE html>
<html>
<head>
<title>Page Title</title>
</head>
<body>

<h1>This is a Heading</h1>

<p>This is a paragraph.</p>

</body>
</html>
```

https://www.w3schools.com/html/

Make different-sized headers

Markdown	HTML	Rendered Output
`# Heading level 1`	`<h1>Heading level 1</h1>`	Heading level 1
`## Heading level 2`	`<h2>Heading level 2</h2>`	Heading level 2
`### Heading level 3`	`<h3>Heading level 3</h3>`	Heading level 3
`#### Heading level 4`	`<h4>Heading level 4</h4>`	Heading level 4
`##### Heading level 5`	`<h5>Heading level 5</h5>`	Heading level 5
`###### Heading level 6`	`<h6>Heading level 6</h6>`	Heading level 6

https://www.markdownguide.org/basic-syntax/

Make different sections (organize headers and paragraphs)

Div Element

The `<div>` element is used as a container that divides an HTML document into sections and is short for “division”. `<div>` elements can contain *flow content* such as headings, paragraphs, links, images, etc.

```
<div>
  <h1>A section of grouped elements</h1>
  <p>Here’s some text for the section</p>
</div>
<div>
  <h1>Second section of grouped elements</h1>
  <p>Here’s some text</p>
</div>
```

https://www.codecademy.com/learn/learn-html/modules/learn-html-elements/cheatsheet

Make lists

 List Item Element

The `<li>` list item element create list items inside:

- Ordered lists `<ol>`
- Unordered lists `<ul>`

```
<ol>
  <li>Head east on Prince St</li>
  <li>Turn left on Elizabeth</li>
</ol>

<ul>
  <li>Cookies</li>
  <li>Milk</li>
</ul>
```

https://www.codecademy.com/learn/learn-html/modules/learn-html-elements/cheatsheet

Use hyperlinks

<a> Anchor Element

The `<a>` anchor element is used to create hyperlinks in an HTML document. The hyperlinks can point to other webpages, files on the same server, a location on the same page, or any other URL via the hyperlink reference attribute, `href`. The `href` determines the location the anchor element points to.

```
<!-- Creating text links -->
<a href="http://www.codecademy.com">Visit this site</

<!-- Creating image links -->
<a href="http://www.codecademy.com">
        <img src="logo.jpg">Click this image
</a>
```

https://www.codecademy.com/learn/learn-html/modules/learn-html-elements/cheatsheet

Embed images & videos

<img> Image Element

HTML image `<img>` elements embed images in documents. The `src` attribute contains the image URL and is mandatory. `<img>` is an *empty element* meaning it should not have a closing tag.

```
<img src="image.png">
```

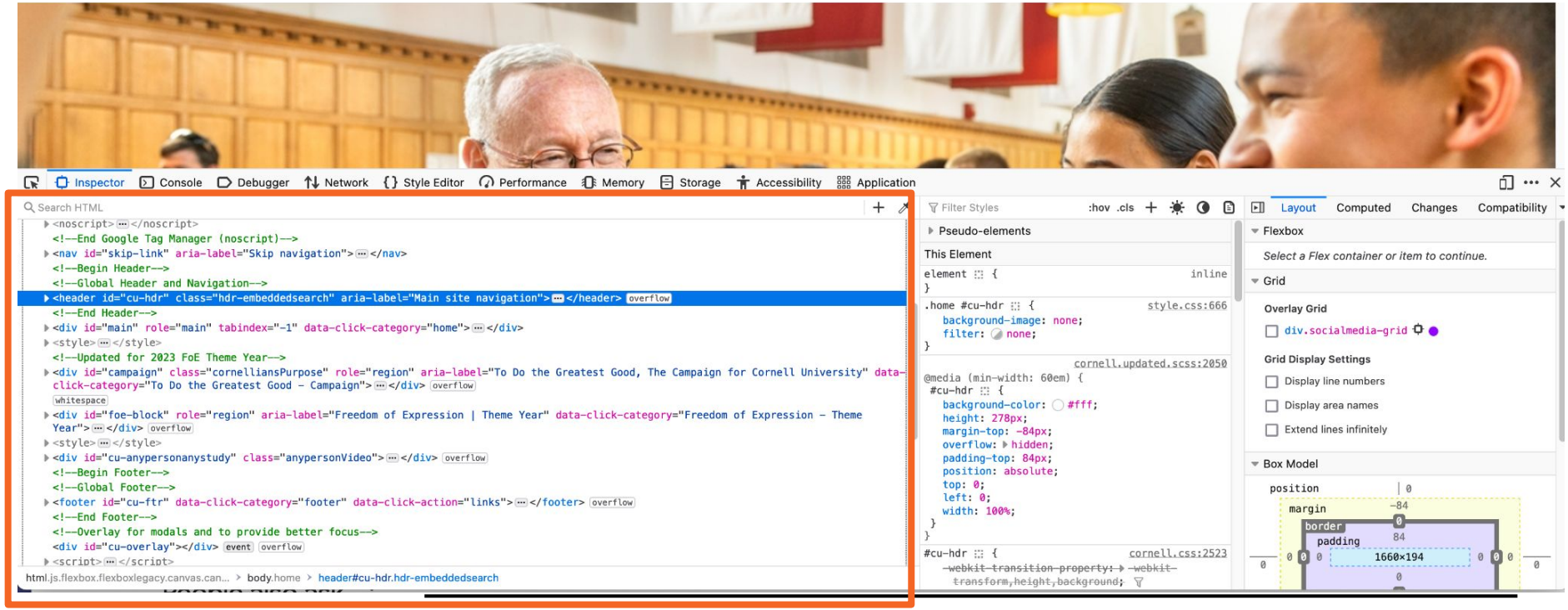
<video> Video Element

The `<video>` element embeds a media player for video playback. The `src` attribute will contain the URL to the video. Adding the `controls` attribute will display video controls in the media player.

```
<video src="test-video.mp4" controls>
  Video not supported
</video>
```

https://www.codecademy.com/learn/learn-html/modules/learn-html-elements/cheatsheet

Don’t panic! We’ll go through more during Friday discussions.



https://www.cornell.edu/

Before Friday Discussions

You must install two new modules: **BeautifulSoup** and **requests** (install them using the same method you used for duckdb).

If you run this code, you should not get errors:

```
import requests
from bs4 import BeautifulSoup
```

Group By's in SQL

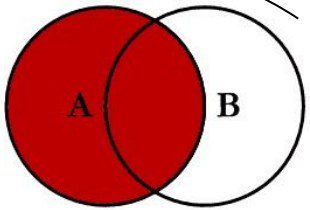
Restaurant	Food	Rating
Shi Miao Dao	Noodles	2
Pho Time	Noodles	2
De Tasty	Noodles	3
CTB	Sandwich	2
Gorgers	Sandwich	3
Dos Amigos	Tacos	2
Luna Inspired Street Food	Tacos	1

```
duckdb.sql("SELECT Food, AVG(Rating) FROM food_df GROUP BY Food;").df()
```

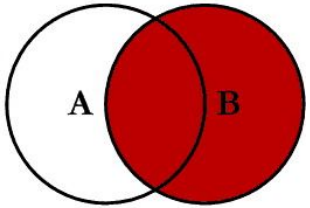
✓ 0.5s

	Food	avg(Rating)
0	Noodles	2.333333
1	Sandwich	2.500000
2	Tacos	1.500000

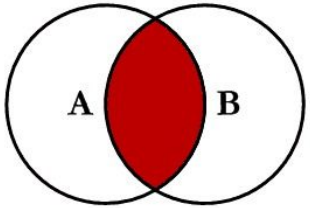
SQL JOINS



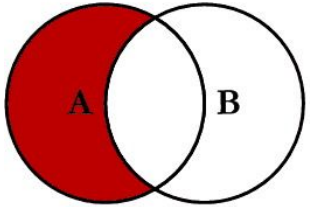

```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
```



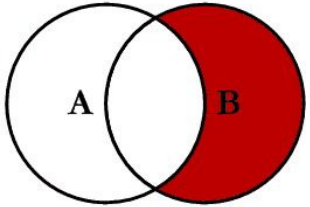

```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
```



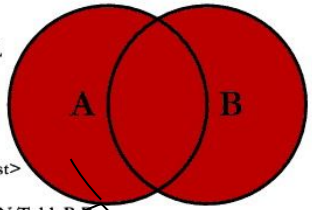

```
SELECT <select_list>
FROM TableA A
INNER JOIN TableB B
ON A.Key = B.Key
```



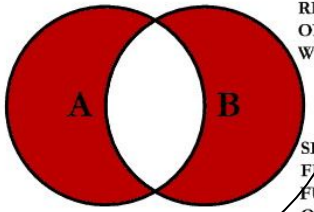

```
SELECT <select_list>
FROM TableA A
LEFT JOIN TableB B
ON A.Key = B.Key
WHERE B.Key IS NULL
```




```
SELECT <select_list>
FROM TableA A
RIGHT JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
```




```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
```




```
SELECT <select_list>
FROM TableA A
FULL OUTER JOIN TableB B
ON A.Key = B.Key
WHERE A.Key IS NULL
OR B.Key IS NULL
```

How do we inner join these tables?

SELECT _____, Food, Rating
FROM ratings_df INNER JOIN food_df
_____ _____ = _____

Restaurant	Rating
Shi Miao Dao	2
Pho Time	2
De Tasty	3
Dos Amigos	2
Luna Street Food	1

ratings_df

Restaurant	Food
Shi Miao Dao	Noodles
Pho Time	Noodles
De Tasty	Noodles
Dos Amigos	Tacos
Luna Street Food	Tacos

food_df

How do we inner join these tables?

```
SELECT food_df.Restaurant, Food, Rating
FROM ratings_df INNER JOIN food_df
ON ratings_df.Restaurant = food_df.Restaurant
```

Restaurant	Rating
Shi Miao Dao	2
Pho Time	2
De Tasty	3
Dos Amigos	2
Luna Street Food	1

ratings_df

Restaurant	Food
Shi Miao Dao	Noodles
Pho Time	Noodles
De Tasty	Noodles
Dos Amigos	Tacos
Luna Street Food	Tacos

food_df

Restaurant	Food	Rating
Shi Miao Dao	Noodles	2
Pho Time	Noodles	2
De Tasty	Noodles	3
Dos Amigos	Tacos	2
Luna Street Food	Tacos	1

How do we inner join these tables?

```
SELECT ratings_df.Restaurant, Food, Rating
FROM ratings_df INNER JOIN food_df
ON ratings_df.Restaurant = food_df.Restaurant
```

Restaurant	Rating
Shi Miao Dao	2
Pho Time	2
De Tasty	3
Dos Amigos	2
Luna Street Food	1

ratings_df

Restaurant	Food
Shi Miao Dao	Noodles
Pho Time	Noodles
De Tasty	Noodles
Dos Amigos	Tacos
Luna Street Food	Tacos

food_df

Restaurant	Food	Rating
Shi Miao Dao	Noodles	2
Pho Time	Noodles	2
De Tasty	Noodles	3
Dos Amigos	Tacos	2
Luna Street Food	Tacos	1

How do we inner join these tables?

```
SELECT ___.Restaurant, Food, Rating
FROM ratings_df a INNER JOIN food_df b
ON ___.Restaurant = ___.Restaurant
```

Restaurant	Rating
Shi Miao Dao	2
Pho Time	2
De Tasty	3
Dos Amigos	2
Luna Street Food	1

ratings_df

Restaurant	Food
Shi Miao Dao	Noodles
Pho Time	Noodles
De Tasty	Noodles
Dos Amigos	Tacos
Luna Street Food	Tacos

food_df

Restaurant	Food	Rating
Shi Miao Dao	Noodles	2
Pho Time	Noodles	2
De Tasty	Noodles	3
Dos Amigos	Tacos	2
Luna Street Food	Tacos	1

How do we inner join these tables?

```
SELECT a.Restaurant, Food, Rating
FROM ratings_df a INNER JOIN food_df b
ON a.Restaurant = b.Restaurant
```

Restaurant	Rating
Shi Miao Dao	2
Pho Time	2
De Tasty	3
Dos Amigos	2
Luna Street Food	1

ratings_df

Restaurant	Food
Shi Miao Dao	Noodles
Pho Time	Noodles
De Tasty	Noodles
Dos Amigos	Tacos
Luna Street Food	Tacos

food_df

Restaurant	Food	Rating
Shi Miao Dao	Noodles	2
Pho Time	Noodles	2
De Tasty	Noodles	3
Dos Amigos	Tacos	2
Luna Street Food	Tacos	1

Does left join yield the same as inner join?

```
SELECT a.Restaurant, Food, Rating
FROM ratings_df a LEFT JOIN food_df b
ON a.Restaurant = b.Restaurant
```

Restaurant	Rating
Shi Miao Dao	2
Pho Time	2
De Tasty	3
Dos Amigos	2
Luna Street Food	1

ratings_df

Restaurant	Food
Shi Miao Dao	Noodles
Pho Time	Noodles
De Tasty	Noodles
Dos Amigos	Tacos
Luna Street Food	Tacos

food_df

Does left join yield the same as inner join?

```
SELECT a.Restaurant, Food, Rating
FROM ratings_df a LEFT JOIN food_df b
ON a.Restaurant = b.Restaurant
```

Restaurant	Rating
Shi Miao Dao	2
Pho Time	2
De Tasty	3
Dos Amigos	2
Luna Street Food	1

ratings_df

Restaurant	Food
Shi Miao Dao	Noodles
Pho Time	Noodles
De Tasty	Noodles
Dos Amigos	Tacos
Luna Street Food	Tacos

food_df

Yes, because the # rows in BOTH ratings_df and food_df is 5

Which is the same as the total # rows in ratings_df

Which has more rows: INNER JOIN, or ratings_df2 LEFT JOIN food_df?

Restaurant	Rating
Shi Miao Dao	2
Pho Time	2
De Tasty	3
Dos Amigos	2
Luna Street Food	1
CTB	2

ratings_df2

Restaurant	Food
Shi Miao Dao	Noodles
Pho Time	Noodles
De Tasty	Noodles
Dos Amigos	Tacos
Luna Street Food	Tacos

food_df

LEFT JOIN has 6 rows, INNER JOIN only has 5

```
duckdb.sql('SELECT a.Restaurant, Food, Rating FROM ratings_df2 a LEFT JOIN food_df b
ON a.Restaurant=b.Restaurant').df()
```

Restaurant	Rating
Shi Miao Dao	2
Pho Time	2
De Tasty	3
Dos Amigos	2
Luna Street Food	1
CTB	2

ratings_df

Restaurant	Food
Shi Miao Dao	Noodles
Pho Time	Noodles
De Tasty	Noodles
Dos Amigos	Tacos
Luna Street Food	Tacos

food_df

Restaurant	Food	Rating
Shi Miao Dao	Noodles	2
Pho Time	Noodles	2
De Tasty	Noodles	3
Dos Amigos	Tacos	2
Luna Street Food	Tacos	1
CTB	NaN	2

Write the SQL query to get from the df on the left to the df on the right

merged_df

Restaurant	Food	Rating
Shi Miao Dao	Noodles	2
Pho Time	Noodles	2
De Tasty	Noodles	3
Dos Amigos	Tacos	2
Luna Street Food	Tacos	1
CTB	NaN	2

Food	Sum
Noodles	7.0
Tacos	3.0
NaN	2.0

```
SELECT _______, __________ AS _______
FROM merged_df
GROUP BY ____
```

Write the SQL query to get from the df on the left to the df on the right

merged_df

Restaurant	Food	Rating
Shi Miao Dao	Noodles	2
Pho Time	Noodles	2
De Tasty	Noodles	3
Dos Amigos	Tacos	2
Luna Street Food	Tacos	1
CTB	NaN	2

Food	Sum
Noodles	7.0
Tacos	3.0
NaN	2.0

```
SELECT Food, Sum(Rating) AS Sum
FROM merged_df
GROUP BY Food
```

Write the SQL query to restrict grouped_df to rows where Sum > 5.0

grouped_df

Food	Sum
Noodles	7.0
Tacos	3.0
NaN	2.0

Food	Sum
Noodles	7.0

SELECT _____
FROM grouped_df
_____

Write the SQL query to restrict grouped_df to rows where Sum > 5.0

grouped_df

Food	Sum
Noodles	7.0
Tacos	3.0
NaN	2.0

Food	Sum
Noodles	7.0

```
SELECT *
FROM grouped_df
WHERE Sum > 5.0
```

Write the SQL query to only show rows where food rating sums are > 5.0

merged_df

Restaurant	Food	Rating
Shi Miao Dao	Noodles	2
Pho Time	Noodles	2
De Tasty	Noodles	3
Dos Amigos	Tacos	2
Luna Street Food	Tacos	1
CTB	NaN	2

Food	Sum
Noodles	7.0

SELECT Food, Sum(Rating) AS Sum
FROM merged_df

_____ ___ _____
_____ _____ > 5.0

Write the SQL query to only show rows where food rating sums are > 5.0

merged_df

Restaurant	Food	Rating
Shi Miao Dao	Noodles	2
Pho Time	Noodles	2
De Tasty	Noodles	3
Dos Amigos	Tacos	2
Luna Street Food	Tacos	1
CTB	NaN	2

Food	Sum
Noodles	7.0

You must use HAVING instead of WHERE if doing aggregation

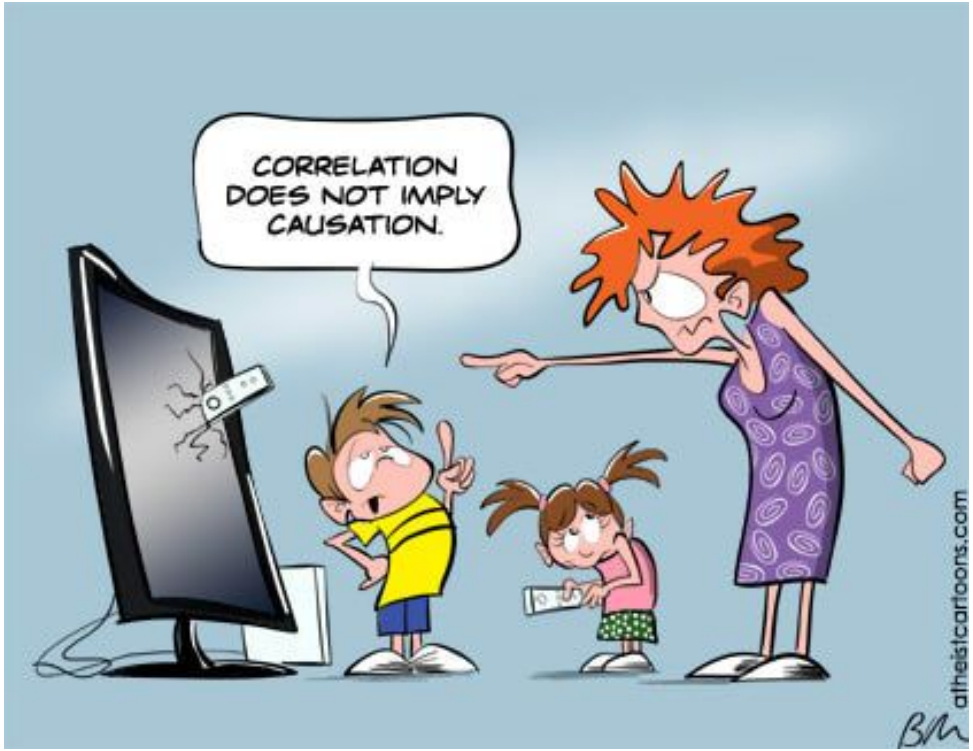
```
SELECT Food, Sum(Rating) AS Sum
FROM merged_df
GROUP BY Food
HAVING Sum > 5.0
```

Going to the SQL gym

These things come with practice!

Additional online resources:

- SQLBolt
- Leetcode
- W3 Schools



https://medium.com/@seema.singh/why-correlation-does-not-imply-causation-5b99790df07e