Data Analysis Grab-Bag

percent change

$$\frac{\text{NEW - OLD}}{\text{OLD}}$$

If teacher salaries were $31,500 in 2017 and $32,000 in 2018, we can say:

- Teacher salaries increased by $500.
  - 32000-31500 = 500
- Teacher salaries increased by 1.6 PERCENT.
  - (32000-31500)/31500 = 0.01587

It works the same way with a decrease. If teacher salaries were $31,500 in 2017 and $30,000 in 2018, we can say:

- Teacher salaries decreased by $1,500.
  - 30000-31500 = -1500
- Teacher salaries decreased by 4.8 PERCENT.
  - (30000-31500)/31500 = -0.04762

percent change of a percent

$$\frac{\text{NEW - OLD}}{\text{OLD}}$$

What if we're dealing with changes to something that's ALREADY measured as a percentage?

If 25% of teachers had a masters degree in 2017 and 30% had a masters degree in 2018, we can say:

- The share of teachers with a masters degree increased by 5 PERCENTAGE POINTS.
  - 30-25 = 5
- The share of teachers with a masters degree increased by 20 PERCENT.
  - (30-25)/25 = 0.20

Or, if we're in a decrease situation: If 25% of teachers had a masters degree in 2017 and 18% had a masters degree in 2018, we can say:

- The share of teachers with a masters degree decreased by 7 PERCENTAGE POINTS.
  - 18-25 = -7
- The share of teachers with a masters degree decreased by 28 PERCENT.
  - (18-25)/25 = -0.28

per-capita

How many murders were there in New York City versus Austin, Texas?

To get a reasonable comparison, be sure to account for how many people live in each place!

city	homicide_rate_2017	population_2017	homicides_per_capita	
New York City	290	8,622,698	3.4	per 100,000
Austin, Texas	29	931,830	3.1	per 100,000
Detroit	267	672,795	39.7	per 100,000

** numbers not fact-checked!*

choosing your denominator wisely

How should we measure participation in an election in a particular county?

Some options:

- votes cast / registered voters in the county
- votes cast / eligible voters in the county
- votes cast / U.S. citizens who are at least 18 yrs old
- votes cast / people who live in the county

There’s not necessarily a RIGHT answer. You’re answering a different question with each option.

Analyzing two variables

Data

Let’s use the schools data again. But this time we’re interested in how two variables move *together*.

Types of questions:

As the rate of poor students at a school increases, what happens to test scores?

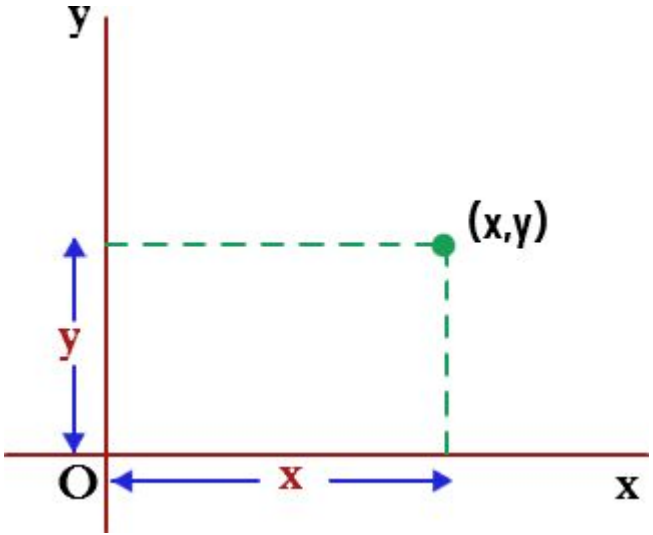
Is there a relationship between student teacher-ratio and math test scores?

School Reports data:
https://docs.google.com/spreadsheets/d/1HdzL31Sw0MUJxPrzJ1kWRUasCjtQakyFVJmq1XCRaK0/edit?usp=sharing

Cartesian coordinate review

In general, the x-axis is the thing we think is a predictor.

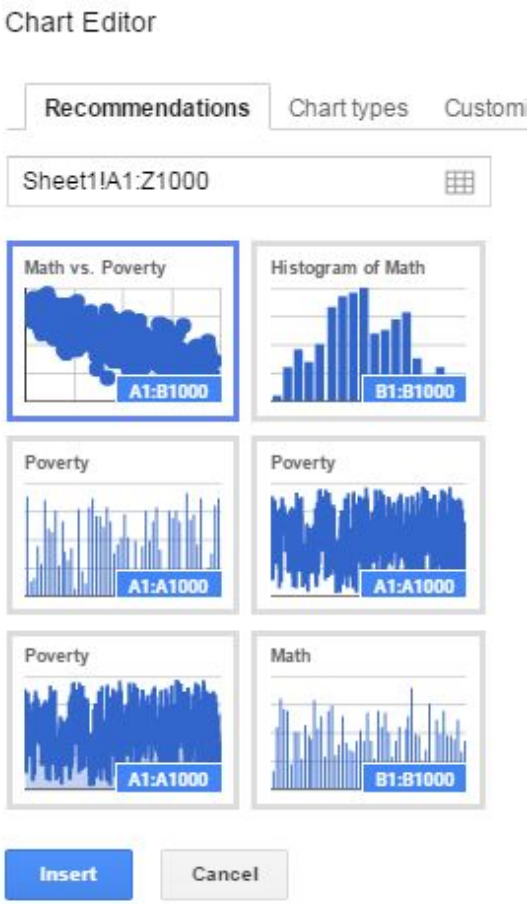
The y-axis is the thing we think might be related to the x-axis.



Are poverty and test scores related?

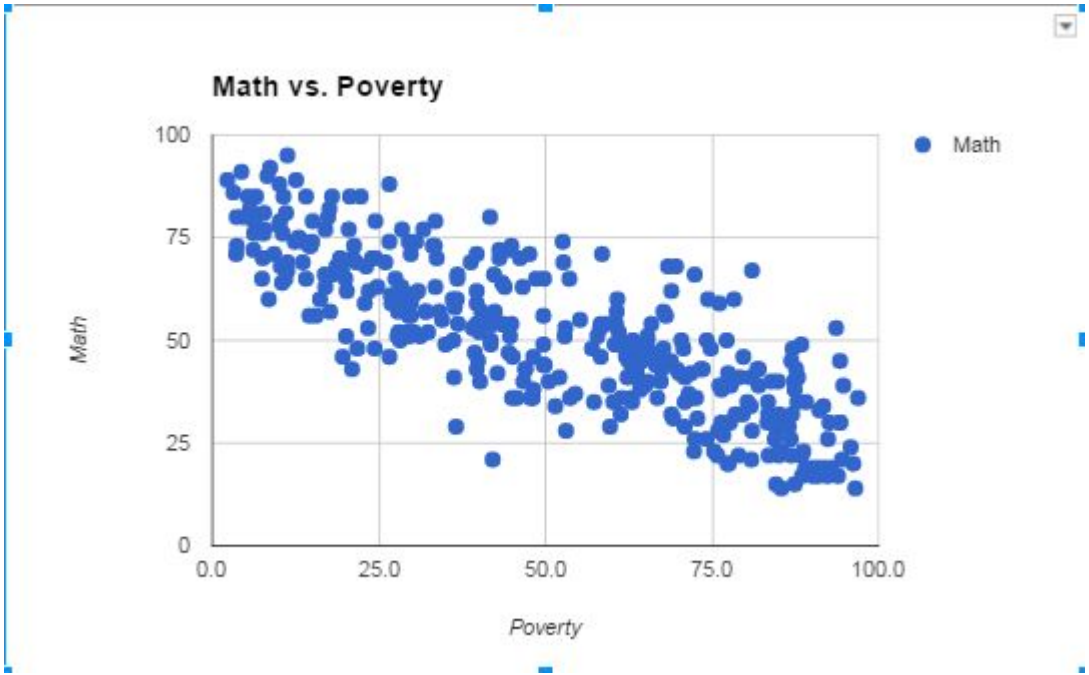
To assess this, we want to plot poverty rates and school average test scores on our x and y axis.

- Select the poverty and math score columns and paste them into a new tab.
- Choose Math vs. Poverty from the options



Are poverty and test scores related?

Your result should look like this:



Interpreting a scatterplot

If y goes up as X goes up, that is a positive correlation.

If y goes down as X goes up, that is a negative correlation.

There are two uses for a scatterplot:

1. What is the relationship between the variables overall?
2. Are there ‘outliers’, or observations that don’t fit the pattern. This can be the source of good reporting leads.

correlation

Calculate the correlation between two columns in Google Sheets with

```
=CORREL(A2:A,B2:B)
```

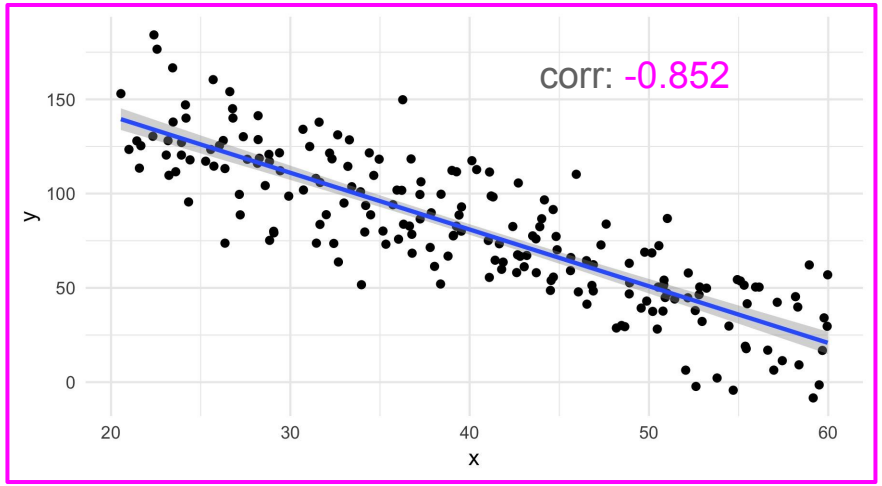
This will return a single value, a correlation coefficient. The value measures how close the two variables are to having a perfectly linear relationship with each other.

It will always be between -1 and 1.

- -1 : perfectly negatively correlated
- 0 : no correlation
- 1 : perfectly positively correlated

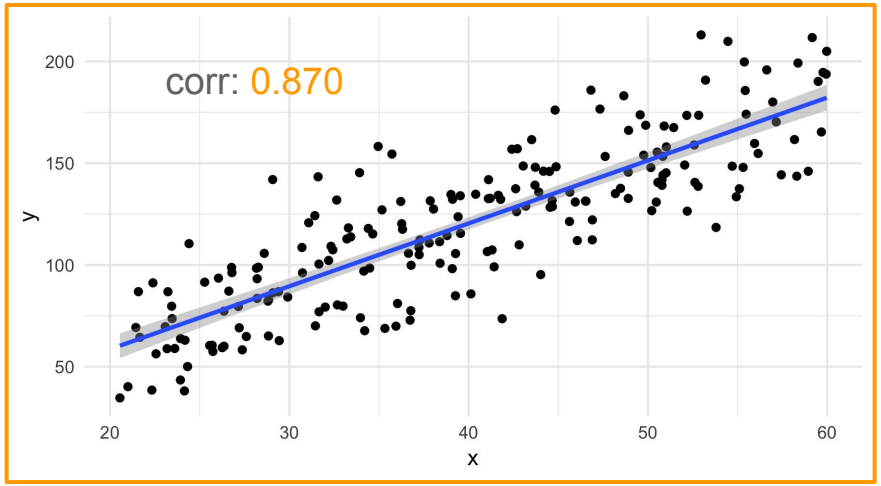
Negative correlation:

as one variable goes up, the other goes down



Positive correlation:

as one variable goes up, the other also goes up



Slope and correlation are different concepts

