Correlations

Students continue to interpret scatter plots, and think about direction and strength of linear relationships.

Prerequisites	None
Relevant Standards OK K12CS CSTA NGSS CC-Math	*Select one or more standards from the menu on the left (⌘-click on Mac, Ctrl-click elsewhere).*
Lesson Goals	Students will be able to... • Confirm if a scatter plot appears linear • Understand how correlation measures direction in a linear relationship • Understand how correlation measures strength in a linear relationship
Student-facing Lesson Goals	• Let's explore scatter plots and what they can tell us about data relationships.
Materials	• Lesson Slides (Google Slides) • Computer for each student (or pair), with access to the internet • Student workbook, and something to write with
Preparation	• Make sure all materials have been gathered • Decide how students will be grouped in pairs • All students should log into CPO and open the "Animals Starter File" they saved from the prior lesson. If they don't have the file, they can open a new one
Supplemental Resources	Spurious Correlations
Language Table	

Types	Functions	Values
Number	`num-sqrt, num-sqr, mean, median, modes`	`4, -1.2, 2/3`
String	`string-repeat, string-contains`	`"hello", "91"`
Boolean	`==, <, <=, >=, string-equal`	`true, false`
Image	`triangle, circle, star, rectangle, ellipse, square, text, overlay, bar-chart, pie-chart, bar-chart-summarized, pie-chart-summarized, histogram, scatter-plot`	
Table	`count, .row-n, .order-by, .filter, .build-column`	

Glossary

form :: of a relationship between two quantitative variables: whether the two variables together vary linearly or in some other way

r :: a number between −1 and 1 that measures the direction and strength of a linear relationship between two quantitative variables (also known as correlation value)

Correlations have *Form*

10 minutes

Overview

Students identify and make use of patterns in scatter plots, learning to characterize them as being linear, curved, or showing no clear pattern. This builds intuition for determining if the ***form*** is linear, in which case we can proceed to correlation and linear regression

Launch

By now we have learned ways to summarize a single quantitative variable, like the `age` of an animal in our dataset: report the center, spread, and shape of the distribution. Together, those numbers tell us what age is typical, how much the ages vary, and what kind of age values are usual or unusual. We could do the same for `pounds`, `weeks`, or any other quantitative column.

But those individual summaries tell us nothing about the *relationship* between animals' ages and weights. In order to understand such relationships, we have to expand our view from a single dimension (along one axis) to two dimensions. This goes hand in hand with expanding our display from a one-dimensional histogram to a two-dimensional scatter plot.

Rather than summarizing each distribution in one dimension, we can summarize a *linear relationship* between two quantitative variables. But this only makes sense if the scatter plot follows a *straight-line pattern*, as opposed to being curved. So the very first assessment we have to make is to identify the ***form*** of the relationship as being linear or not.

Form: whether a relationship is linear or not

Investigate

The relationship between two quantitative variables can take many forms - some patterns are *linear*, and appear as a straight line sloping up or down. Some patterns are *non-linear*, and may look like a curve or an arc. And sometimes there is no pattern or relationship at all!

Have students turn to Identifying Form, Direction and Strength (Page 73) in their student workbooks. For each scatter plot, identify whether the relationship is linear, non-linear or if there's no relationship at all.

Synthesize

Data Scientists use their eyes all the time! It doesn't make sense to search for correlations when there's no pattern at all, and only linear relationships make sense if we want to summarize with a correlation.

Going Deeper

In an AP Statistics class or full-year Data Science class, it's appropriate to discuss non-linear relationships here. In a dedicated computer science class, it may also be appropriate to talk about *transforming* the x- or y-axis (using `.build-column` !) via a quadratic, exponential, or logarithmic function and then looking for a linear pattern in the resulting scatter plot. All of these are **extensions** to the materials presented here.

Correlations have *Direction & Strength* 20 minutes

Overview

Once students have learned to identify a possible linear relationship, they can turn their attention to other qualities of that relationship: its *direction* and *strength*. Each of these is expressed in the r-value, which students learn to read.

Launch

Assuming a relationship is linear, data scientists calculate a single number called "correlation" - or r-value - that reports both the direction and strength.

Direction: whether a linear relationship is positive or negative

Direction: whether a linear relationship is positive or negative.

A linear relationship between two quantitative variables is *positive* if, in general, the scatter plot points are sloping up: smaller x values tend to go with smaller y values, and larger x values tend to go with larger y values. The relationship is *negative* if points slope down: smaller x values tend to go with *larger* y values, and larger x values tend to go with *smaller* y values.

- **Positive** directions are by far more common because of natural tendencies for variables to increase in tandem. For example, "the older the animal, the more it tends to weigh". This is usually true for human animals, too!
- **Negative** relationships can also occur. For example, "the older a child gets, the fewer new words he or she learns each day."

Strength: how closely the two variables are correlated.

A relationship between two quantitative variables is strong if the scatter plot points are tightly clustered together. In this case, knowing the x-value of a data point gives us a very good idea of what its y-value will be. In other words, if the relationship is linear and strong, the scatter plot points are clumped together in a thin cloud.

- A **strong** linear relationship means that the points in the scatter plot are all clustered closely around an invisible line. If the cloud point is very tightly packed around the line, the relationship is said to be strong.
- A **weak** linear relationship means that the cloud of points is scattered very loosely around the line.

Investigate

Have students turn to Identifying Form, Direction and Strength (Page 73) in their student workbooks. For each scatter plot, identify whether the relationship is positive or negative, and whether it is strong or weak.

The correlation ***r*** is a number (between -1 and 1) that tells us the direction and strength of a linear relationship between two variables. r is positive or negative depending on whether the correlation is positive or negative. **The strength of a correlation is the distance from zero** : an r-value of zero means there is no correlation at all, and stronger correlations will be closer to −1 or 1.

An r-value of about ±0.65 or ±0.70 or more is typically considered a strong correlation, and anything between ±0.35 and ±0.65 is "moderately correlated". Anything less than about ±0.25 or ±0.35 may be considered weak. However, these cutoffs are not an exact science! In some contexts an r-value of ±0.50 might be considered impressively strong!

Calculating r from a data set only tells us the direction and strength of the relationship in *that particular sample* . If the correlation between adoption time and age for a representative sample of about 30 shelter animals turns out to be +0.44, the correlation for the larger population of animals will probably be *close* to that, but certainly not the same.

Have students turn to Identifying Form and r-Values (Page 74) in their student workbooks. For each scatter plot, identify whether the relationship linear, and use r to summarize direction and strength.

- In the Interactions Area, create a scatter plot for the Animals Dataset, using `"pounds"` as the xs and `"weeks"` as the ys.
- **Form:** Does the point cloud appear linear or non-linear?
- **Direction:** If it's linear, does it appear to go up or down as you move from left to right?
- **Strength:** Is the point cloud tightly packed, or loosely dispersed?
- Would you predict that the r-value is positive or negative? Will it be closer to zero, closer to ±1, or in between?
- Have Pyret compute the r-value, by typing `r-value(animals-table, "pounds", "weeks")` . Does this match your prediction?
- Repeat this process using `"age"` as the xs. Is this correlation stronger or weaker than the correlation for `"pounds"` ? What does that *mean*?

Common Misconceptions

- Students often conflate strength and direction, thinking that a strong correlation *must* be positive and a weak one *must* be negative.
- Students may also falsely believe that there is ALWAYS a correlation between any two variables in their dataset.
- Students often believe that strength and sample size are interchangeable, leading to mistaken assumptions like "any

- Students often believe that strength and sample size are interchangeable, leading to mistaken assumptions like "any correlation found in a million data points *must* be strong!"

Synthesize

It is useful to ask students probing questions, to help address the misconceptions listed above. Some examples:

- What is the difference between a *weak* relationship and a *negative* relationship?
- What is the difference between a *strong* relationship and a *positive* relationship?
- If we find a strong relationship in a sample, can we always infer that relationship holds for the whole population?
- Suppose we have two correlations, one drawn from 10 data points and one drawn from 50. If both correlations are identical in direction and strength, should we trust them equally when making an inference about the larger population?

Correlation does NOT imply causation.

It's easy to be seduced by large r-values, and believe that we're really onto something that will help us make predictions! But Data Scientists know better than that...
Here are some real-life correlations that have absolutely no causal relationship; they come about either by chance or because both of them are tied in with another variable that's (often) lurking in the background.

- "Number of people who drowned after falling out of a fishing boat" v. "Marriage rate in Kentucky" (r = 0.98)
- "Average per-person consumption of chicken" v. "U.S. crude oil imports" (r = 0.95)
- "Marriage rate in Wyoming" v. "Domestic production of cars" (r = 0.99)
- "Number of people who get tangled in their own bedsheets" v. "Amount of cheese consumed that year" (r = 0.95)

All of these correlations come from the Spurious Correlations website. If time allows, have your students explore the site to see more!

Your Analysis

flexible

Overview

Students repeat the previous activity, this time applying it to their own dataset and interpreting their own results. **Note: this activity can be done as a homework assignment, but we recommend giving students an *additional class period* to work on this.**

Launch

What correlations do you think there are in your dataset? Would you like to investigate a subset of your data to find those correlations?

Investigate

- Brainstorm a few possible correlations that you might expect to find in your dataset, and make some scatter plots to investigate.
- Turn to Correlations in My Dataset (Page 75), and list three correlations you'd like to search for.
- Investigate these correlations. If you need blank Design Recipes, you can find them at the back of your workbook, just before the Contracts.
- What correlations did you find?
- Did you need to filter out certain rows in order to get those correlations?

Synthesize

Have students share back their correlations, and why they expect to find them.
After looking at the scatter plot for our animal shelter, do students still agree with the claim on (Dis)Proving a Claim (Page 71)? (Perhaps they need more information, or to see the analysis broken down separately by animal!)

Additional Exercises:

- Identifying Form, Direction and Strength (Matching)