Correlations

(Also available in Pyret)

Students deepen their understanding of scatter plots, learning to describe and interpret direction and strength of linear relationships.

Lesson Goals	Students will be able to… • Confirm if a scatter plot appears linear • Understand how correlation assesses 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.
Prerequisites	• Introduction to Data Science • Exploring CODAP • Scatter Plots
Materials	• PDF of all Handouts and Page • Animals Starter File • Data Exploration Project Slide Template • Lesson Slides • Printable Lesson Plan (a PDF of this web page)
Supplemental Materials	• Additional Printable Pages for Scaffolding and Practice • *Identifying Strength (Desmos)*
Supplemental Resources	• Spurious Correlations • Guess the Correlation

Key Points For The Facilitator	• Students frequently confuse correlation for causation! It can be tempting for focus on the *computational* element of the lesson alone, getting kids to think about R-value and identify patterns in graphcs. But this ignores the critical point about correlation and causation, and students who over-focus on obtaining R-values are likely to develop this misconception!

Glossary

correlation :: the degree to which knowing the value of an *explanatory* variable helps us predict the value of another, *response*, variable

direction :: the aspect of a linear relationship that tells if the line relating the two variables is sloping up or down

explanatory variable :: When modeling a possible relationship between an input and an output (e.g. - height and age), we are curious about how a change in the input (typically graphed on the x-axis of a scatter plot) might "explain" the output (y). When the behavior of the output may be explained by the input, we refer to the input as the "explanatory variable".

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

linear regression :: a type of analysis that models the relationship between two quantitative variables. The result is known as a regression line, or line of best fit.

linear relationship :: a mathematical relation between two quantitative variables x and y such that y changes by a constant amount (the slope) for every unit increase in x. When graphed, a linear relationship appears as a straight line (sloping up or down).

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)

response variable :: the variable in a relationship, generally plotted on the y-axis of a scatter plot, that is presumed to be affected by the explanatory variable; in some contexts the response variable is referred to as the "dependent variable" or the "output"

strength :: of a relationship between two quantitative variables: how much do the values of one variable tells us about the values of the other

Correlations have *Form* *5 minutes*

Overview

Students identify and make use of patterns in scatter plots, learning to characterize them as appearing to be linear, curved, or showing no clear pattern. Determining that a ***form*** appears to be linear is a prerequisite for proceeding to ***correlation*** and ***linear regression***.

Note: We can't make any definitive assertions about correlations without computation, which students will learn about in our Linear Regression lesson, but it doesn't make sense to search for correlations when there's no pattern at all, and summarizing with a correlation only makes sense for linear relationships, so it is important for students to develop an intuitive sense of what form means before engaging with the abstraction of computation.

Launch

We can analyze a single quantitative variable, such as `age` or `pounds` to identify a value that is **typical**, how much the values **vary**, and what kind of values are **usual or unusual**.

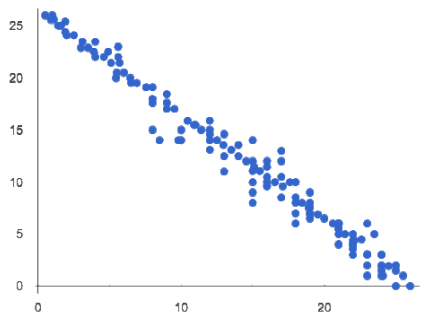
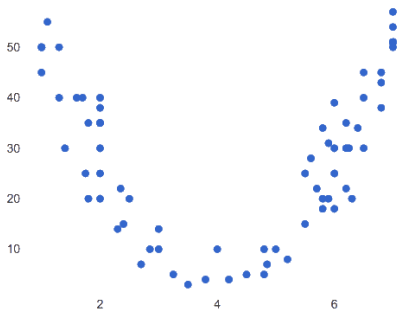
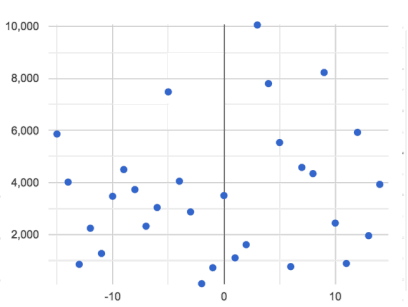
(The bolded words above all deal with notions of what it means for a value to be "normal" or "abnormal". These words have loaded meaning in the context of variability, and should be used carefully!)

But those analyses tell us nothing about the *relationship* between animals' ages and weights. In order to understand such relationships, we have to expand our view from one column to two. This goes hand-in-hand with expanding our display from a 1-dimensional histogram or box plot to a 2-dimensional scatter plot.

Rather than summarizing each distribution in one dimension, we can search for a *linear relationship* between two quantitative variables. Linear relationships only make sense if the scatter plot follows a *straight-line pattern*, so the first thing we need to ask is whether the ***form*** of the relationship appears to be linear or not.

Investigate

Form indicates whether a relationship is linear, nonlinear or undefined:

		
Linear	Nonlinear	No Relationship

Some patterns are ***linear***, and cluster around a straight line sloping up or down.

Some patterns are **nonlinear**, and may look like a curve, a hockey-stick, or some other shape!

And sometimes there is **no relationship** or pattern at all! That means there's no predictable change in the y-axis as we go from one side of the x-axis to the other.

Turn to Identifying Form, Direction and Strength, and complete *just the first question* for each scatter plot, identifying whether the relationship appears to be linear, nonlinear or if there's no relationship at all.

Synthesize

- Which scatter plots seem to have *linear* relationships?
  - *Students should feel very confident that A and C seem to have linear relationships.*
  - *Students will likely also identify D and F as seeming to have linear relationships.*
- Which scatter plots seem to have *nonlinear* relationships?
  - *Scatter plot E seems to have a non-linaer relationship.*
- Which scatter plots seem to have *no relationships*?
  - *Scatter plot B seems to have no relationship.*

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 summarizing with a correlation only makes sense for linear relationships!

Going Deeper

In an AP Statistics class or full-year Data Science class, it's appropriate to discuss nonlinear 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.

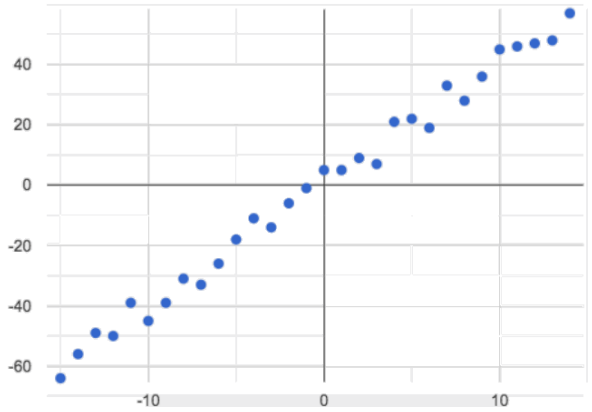
Correlations have *Direction* *10 minutes*

Overview

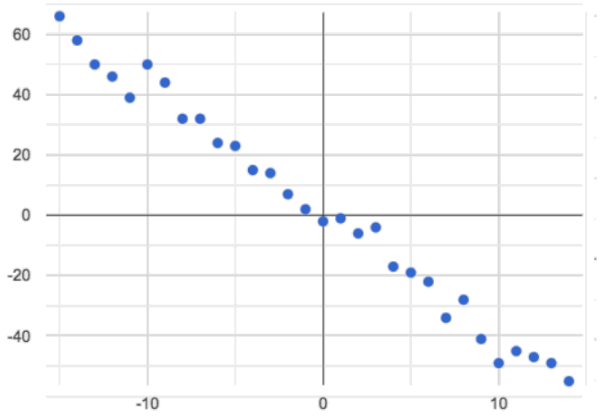
Once students have learned to identify a possible linear relationship, they can turn their attention to other qualities of that relationship, like its ***direction***.

Launch

We can also examine the direction of a linear relationship.



Positive Direction



Negative Direction

A **positive** direction means that the line slopes up as we look from left-to-right. Positive relationships are by far most 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!

A **negative** direction means that the line slopes *down* as we look from left-to-right. Negative relationships can also occur. For example, “the older a child gets, the fewer new words he or she learns each day.”

If the form is nonlinear or non-existent, "direction" doesn’t apply: A parabola might look like it has both a positive *and* negative correlation, and if there’s no form at all then there certainly can’t be a direction!

Investigate

Complete Identifying Form, Direction and Strength and focus *just on the second question*, determining whether each of the possible linear relationships you previously identified appears to have a positive or negative correlation.

Synthesize

- It only makes sense to look for direction in linear relationships!
- Which data sets appear to have a positive correlation between the variables?

Correlations have *Strength* *10 minutes*

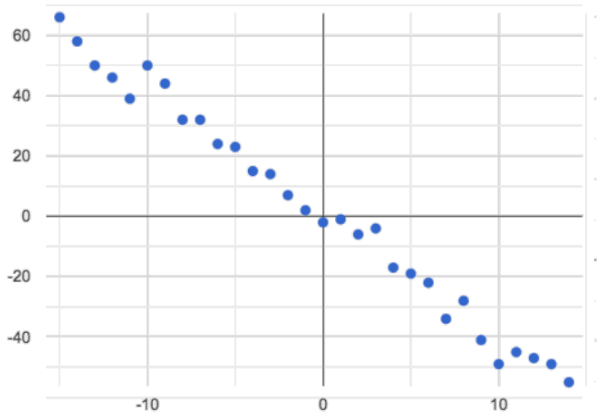
Overview

We'll explore another quality of a possible linear relationship: its ***strength***.

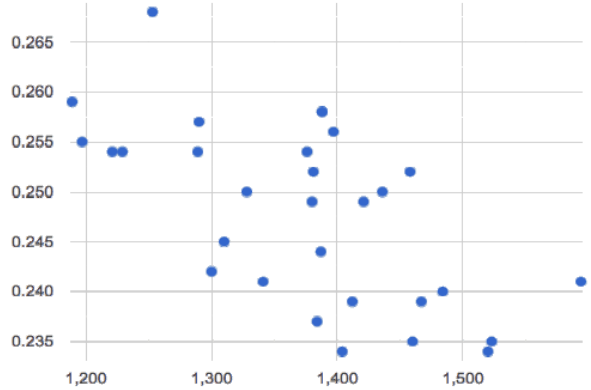
Launch

Strength indicates how closely the two variables are correlated.

How well does knowing the x-value allow us to predict what the y-value will be?



Strong Relationship



Weak Relationship

A relationship is strong if knowing the x-value of a data point gives us a very good idea of what its y-value will be (knowing a student's age gives us a very good idea of what grade they're in). A strong linear relationship means that the points in the scatter plot are all clustered *tightly* around an invisible line.

A relationship is weak if x tells us little about y (a student's age doesn't tell us much about their number of siblings). A weak linear relationship means that the cloud of points is scattered very *loosely* around the line.

If the form is non-existent, "strength" doesn't apply: without any form at all, there's nothing for data points to be tightly or loosely clustered around!

Investigate

- Complete Identifying Form, Direction and Strength, and focus on the third question for each scatter plot, identifying whether the relationship appears to be strong or weak.

- *Optional:* Complete the card sort on Identifying Strength (Desmos).

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 correlation found in a million data points *must* be strong!"

Synthesize

- Complete Reflection on Form, Direction and Strength.
- Be ready to discuss your answers with the class!

This page includes a series of probing questions that get at the common misconceptions listed above. Discuss the answers as a class.

Optional: If time permits, have students complete Identifying Form, Direction and Strength (Matching).

Summarizing Correlations using r-values *20 minutes*

Overview

Now that students know how to identify *direction* and *strength* for linear relationships, they'll learn to read how these are expressed in the *r*-value.

Launch

Students have learned that a correlation can be described by three pieces of information: *Form*, *Direction*, and *Strength*. Statisticians and Data Scientists have a shorter way of describing all three, called ***r-value***.

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!

If it works for you, give students five minutes to play a few rounds of the online game Guess the Correlation to develop intuition with r-values. (This will require creating an account.)

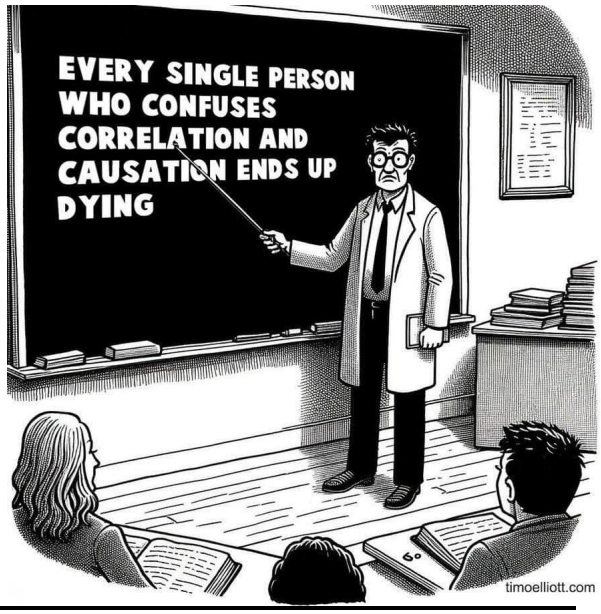
Investigate

- Complete Identifying Form and r-Values. For each scatter plot, identify whether the relationship appears to be linear, and, if so, use *r* to summarize direction and strength.
- Be prepared to discuss your answers with the class!

Calculating *r* from a dataset 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.

- Let's look for correlations in the Animals Dataset!
- Open your saved Animals Starter File, or make a new copy.
- Complete Correlations in the Animals Dataset.

It's easy to be seduced by large r-values, and believe that we're really onto something that will help us claim that one variable really impacts another! But Data Scientists know better than that…



Correlation does NOT imply causation.

Complete Correlation Does Not Imply Causation!

If time allows, you may want to emphasize the point that correlation does not imply causation by having students look at the nonsense claims that could be made from the graphs of real world data on the Spurious Correlations website.

Common Misconceptions

Students often giggle at some of the Spurious Correlations examples, but fail to internalize the point when it comes to the Animals dataset or their own analysis. Pay close attention to students' language when describing their correlations, and make sure they are not using causative wording!

Synthesize

Which corresponded more strongly with time to adoption, `"age"` or `"pounds"`? What does this *mean*?

The correlation with `"pounds"` *is higher, meaning that an animal's weight is a better predictor of the number of weeks an animal will live at the shelter before being adopted than its age.*

- People often confuse correlation with causation. What are some examples of this?
- Why is it a problem for society, that people confuse correlation and causation?

Exploration Project (Correlations) *flexible*

Overview

Students apply what they have learned about correlations to their chosen dataset. They will add two or more items to their Data Exploration Project Slide Template: (1) a correlation they think they see in the data set, and (2) the form, direction and strength of that correlation. To learn more about the sequence and scope of the Exploration Project, visit Project: Dataset Exploration. For teachers with time and interest, Project: Create a Research Project is an extension of the Dataset Exploration, where students select a single question to investigate via data analysis.

Launch

Let's review what we have learned about correlations.

- What kind of displays can we use to visualize a correlation?
  - *Scatter plots are used to visualize correlations.*
- When Data Scientists describe correlations to one another, what three properties do they talk about, and what do they mean?
  - *1) Form - describes the **shape** of a correlation. Correlations can be linear, nonlinear, or non existant (N/A).*
  - *2) Direction - linear correlations can be **positive** or **negative**, describing whether the point cloud seems to rise or fall as the explanatory variable gets larger.*
  - *3) Strength - describes how tightly the data is clustered around a line or curve.*

Investigate

Let's connect what we know about correlations to your chosen dataset.

- Open your chosen dataset starter file in CODAP.
  - *Teachers: Students have the opportunity to choose a dataset that interests them from our List of Datasets in the Choosing Your Dataset lesson.*
  - Turn to Correlations in My Dataset, and list three correlations you'd like to search for.
- Pick **one correlation** to explore. Which column do you think is the ***explanatory variable***? The ***response variable***?
- Make a scatter plot with the explanatory variable on the x-axis and the response variable on the y-axis.

- Do you see a correlation? What is its form? If it's linear, what is its direction and strength?
- Repeat this process for at least one more correlation.

Confirm that all students have created and understand how to interpret their correlations. Once you are confident that all students have made adequate progress, invite them to access their Data Exploration Project Slide Template from Google Drive.

- **It's time to add to your Data Exploration Project Slide Template.**
- Find the "Correlations I want to look into" section of the slide deck.
- For each correlation you wrote in Correlations in My Dataset, copy what you wrote into the slide.
- On the same slide, add your scatter plot and your description of the result.
- Repeat the process for each additional correlation you explored, making copies of the correlation slide as-needed.

Synthesize

Have students share their findings.

Did you discover anything surprising or interesting about their dataset?

Were any of the correlations especially strong? Were any of them surprising?

When students compared their your findings with those of their classmates, did they make any interesting discoveries? (For instance: Did everyone find a strong correlation? A linear one?)