Chapter 4

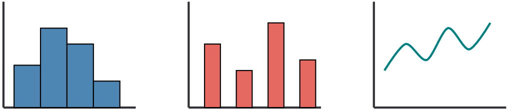
Scatterplots and Correlation

Intended Learning Outcomes

- Create and interpret scatterplots
- Identify explanatory and response variables
- Describe form, direction, and strength
- Recognise outliers
- Calculate and interpret correlation
- Apply properties of correlation
- Distinguish correlation from causation

Why Study Relationships Between Variables?

- So far, we have focused on **univariate** data analysis (one variable at a time).
- To investigate variables, we use tools like **histograms**, **bar graphs**, **stem plots**, and **time series**.



Histogram Bar Graph Time Series

- Many real-world questions involve more than one variable.

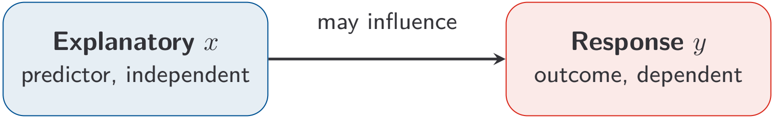
Examples of Bivariate Questions

1. Does **study time** affect **exam scores**?
2. Does **height** predict **earnings**?
3. Does **temperature** influence **ice cream sales**?
4. Does **exercise** relate to **resting heart rate**?

Response and Explanatory Variables

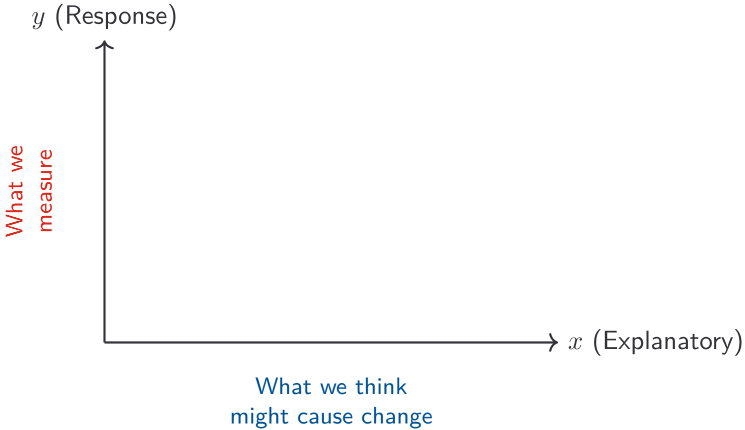
Response and Explanatory Variables

- The **response variable** y is the outcome of interest.
- The **explanatory variable** x may help explain or predict y.



Axis Placement Rule

Key Point: Place the **explanatory variable on the x-axis** and the **response variable on the y-axis**.



Example (`ex:ch4-identifying-variables`): Identifying Variables

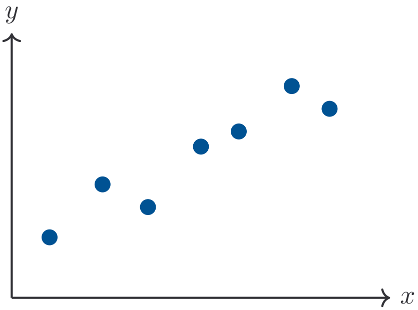
For each scenario, identify the explanatory and response variables:

1. (Study time, Exam scores)
2. (Ice cream sales, Temperature)
3. (Education level, Income)

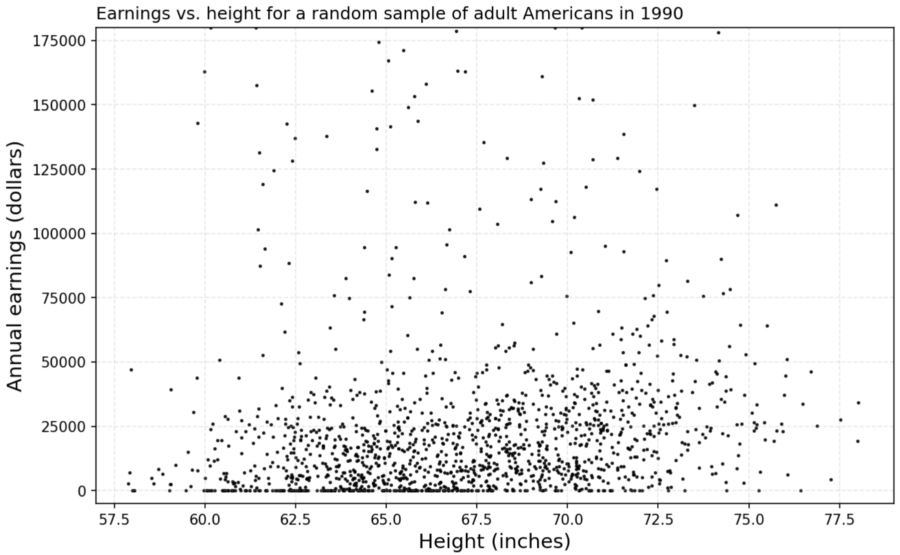
Scatterplot: Definition

Scatterplot

A **scatterplot** displays the relationship between two quantitative variables. Each individual appears as a **point** at coordinates (x_i, y_i).

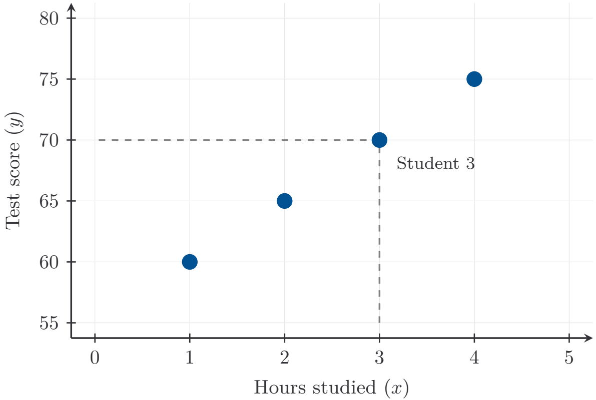


Heights and earnings



Study Time and Test Score

Hours studied (x)	1	2	3	4
Test score (y)	60	65	70	75



Constructing a Scatterplot

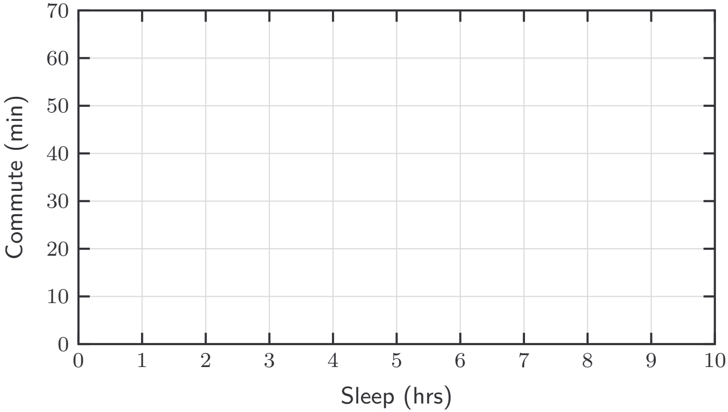
How to Create a Scatterplot

1. Identify the two quantitative variables
2. Decide which variable goes on each axis
3. Label the x-axis (horizontal)
4. Label the y-axis (vertical)
5. Plot each individual as a point at (x_i, y_i)
6. Add a descriptive title

Example (`ex:ch4-plotting-by-hand`): Plotting Data by Hand

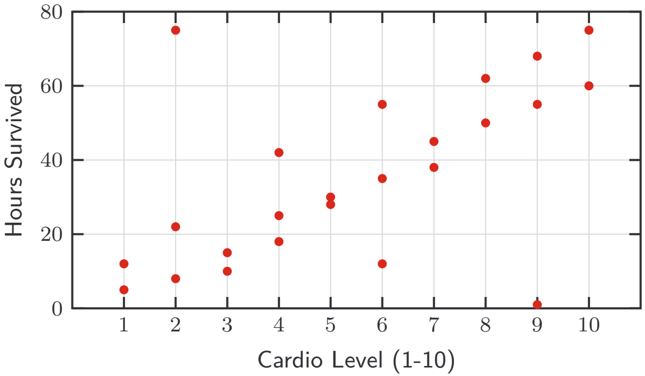
Plot the scatterplot for the following data:

Commute (min)	60	45	30	15	0	20	40
Sleep (hrs)	5.5	6.0	7.0	8.0	8.5	7.5	6.5



Example (`ex:zombie-scatter`): Reading a Scatterplot

Context: A researcher analyzes data from a Zombie Outbreak Simulation to see if "Rule #1: Cardio" actually helps.



1. What is the **longest time** anyone survived?
2. Look at the person with **Cardio Level 10**. What was their lowest survival time?
3. Are there any outlying points? If so, circle one and explain why it is an outlier.
4. How many people with **Cardio Level 4** survived longer than 40 hours?

Plotting several variables simultaneously

We can encode a third (categorical) variable using:

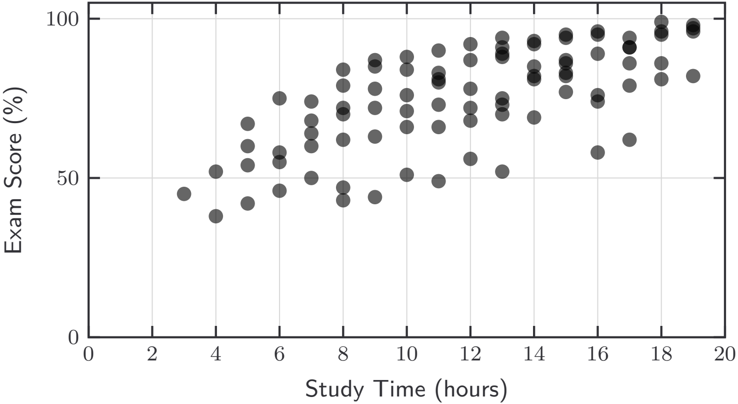
- **Color**: different groups shown in different colors
- **Shape**: different groups shown with different marker shapes
- **Size**: continuous third variable shown by point size

This allows us to explore relationships between three or more variables simultaneously.

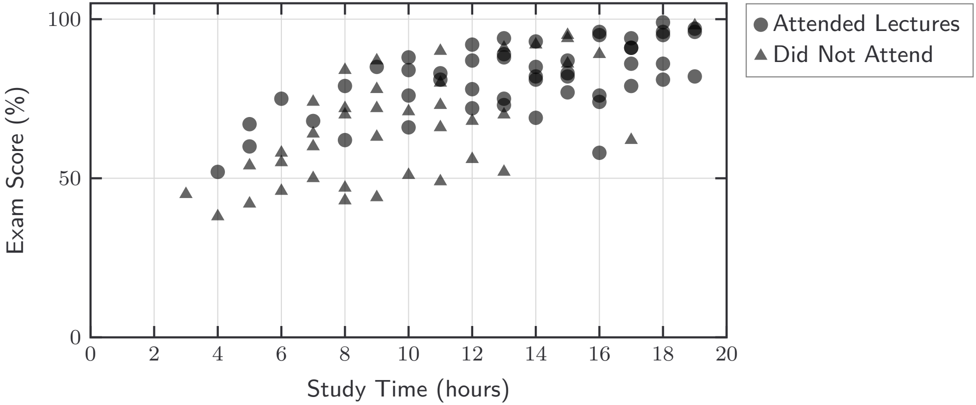
student_id	study_time	exam_score	attended_lectures	sleep_quantity
482910334	5	42	No	4
299401855	12	78	Yes	4.5
850223190	8	84	No	6
110594823	13	94	Yes	9
673820019	18	99	Yes	10
559201182	9	72	No	9

Example (`ex:ch4-exam-study-basic`): Exam Scores vs. Study Time

Context: Study time vs. exam score, with different colors indicating sleep level (low, medium, high).

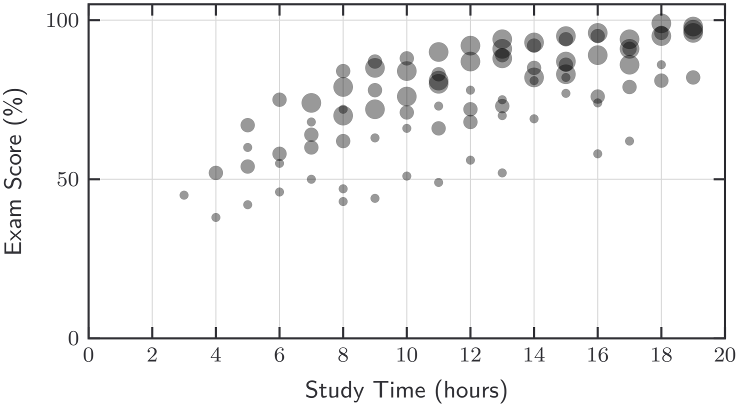


Example (`ex:ch4-third-var-shape`): Adding a Third Variable with Shape

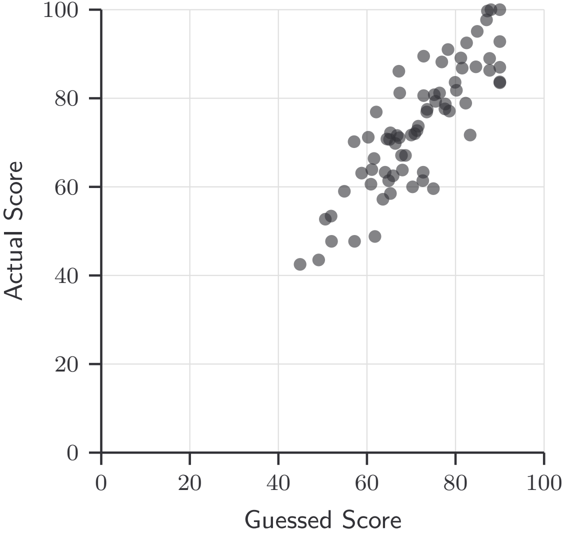


Example (`ex:ch4-third-var-size`): Adding a Third Variable with Size

Context: Study time vs. exam score, with point size indicating hours of sleep (larger = more sleep).

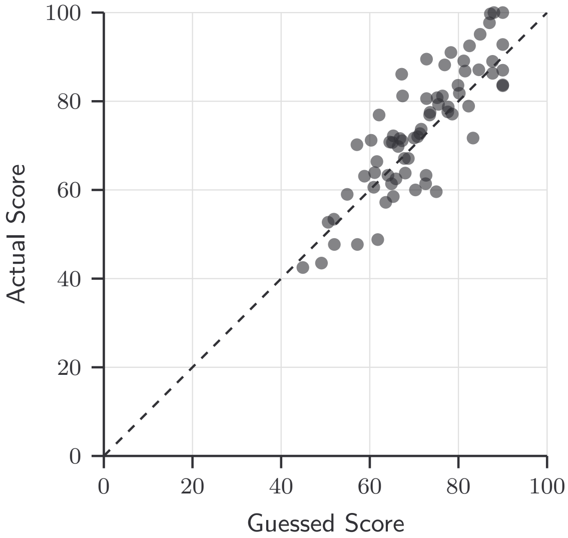


How good were UWaterloo students at predicting their performance?



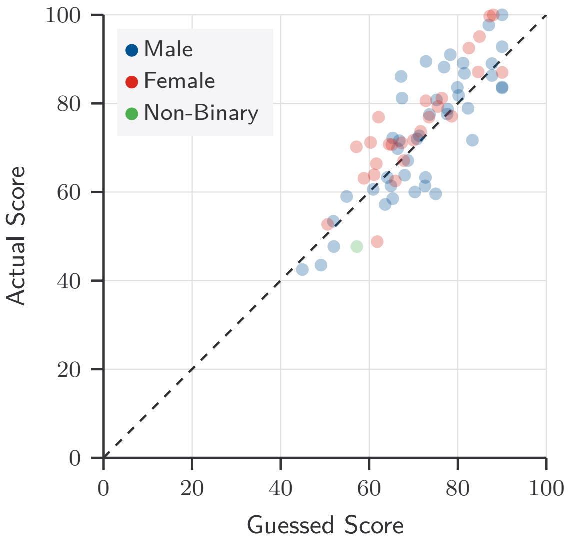
How good were UWaterloo students at predicting their performance?

Adding a reference line

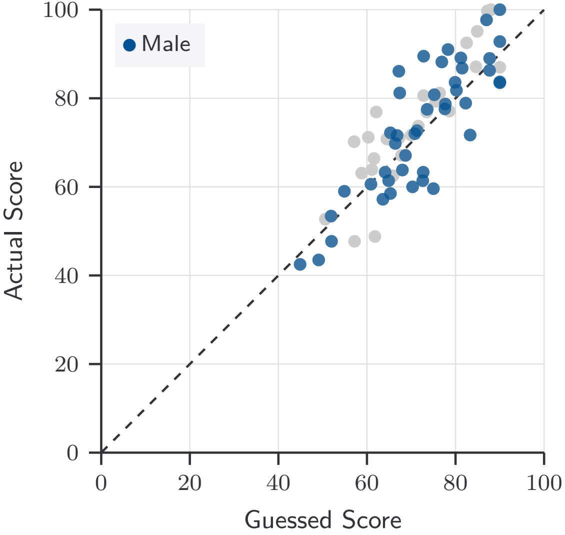


How good were UWaterloo students at predicting their performance?

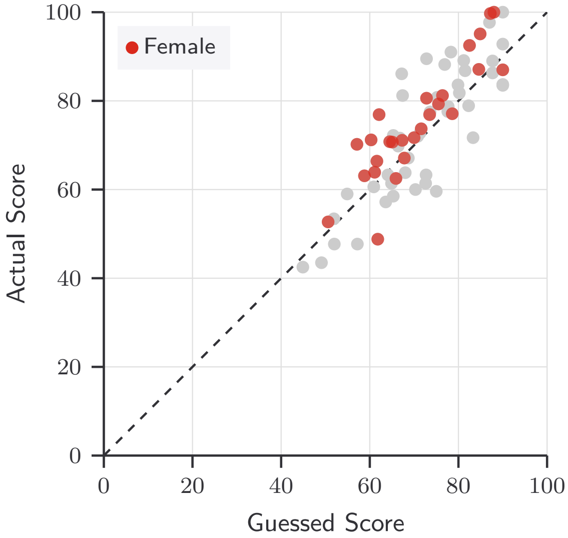
Broken down by gender



How good were men at predicting their performance?



How good were women at predicting their performance?



Depth of interpretation

This is a bit harder than the level of interpretation we will focus on in this course, but it shows how scatterplots can be used to explore complex relationships between multiple variables.

Interpreting Scatterplots at the DS1000 level: Four Key Features

Outliers

Points far from the pattern

Form

Linear or nonlinear?

Direction

Positive or negative?

Strength

Strong, moderate, or weak?

Outliers in Scatterplots

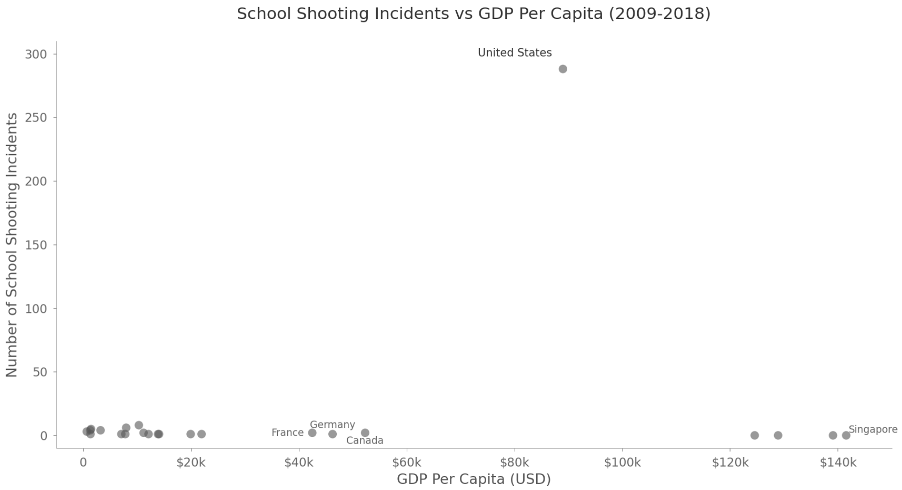
Outlier

A point that falls far from the overall pattern of the data.

Types of unusual points:

- **Vertically unusual**: far above or below the pattern
- **Horizontally unusual**: extreme x-value
- **Both**: unusual in both directions

Example (`ex:ch4-gun-violence`): Outlier Example

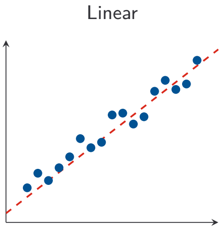


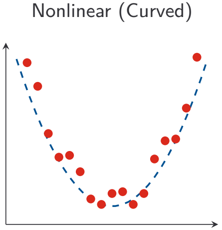
Form: Linear vs. Nonlinear

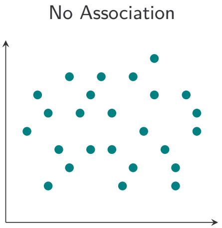
Form

The **form** of a scatterplot describes the overall shape of the relationship

- **Linear**: Points cluster around a straight line.
- **Nonlinear**: Points follow a curved pattern (quadratic, exponential, logarithmic, etc.).
- **No association**: Points appear scattered with no discernible pattern.

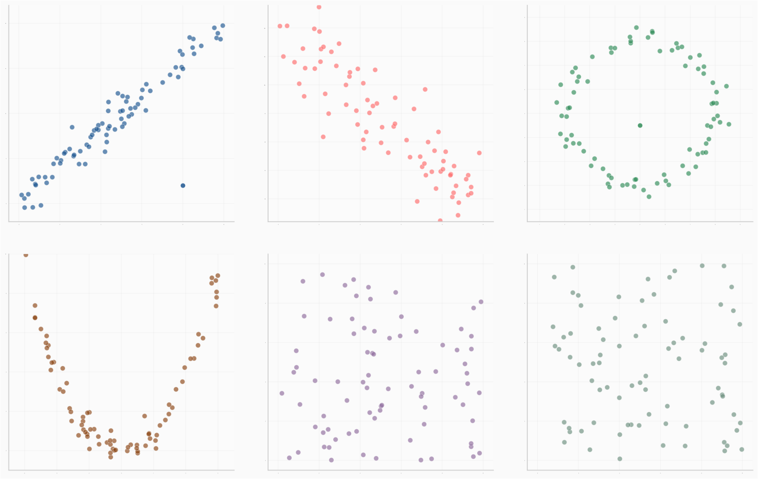






Example (`ex:ch4-finding-form`): Finding Form

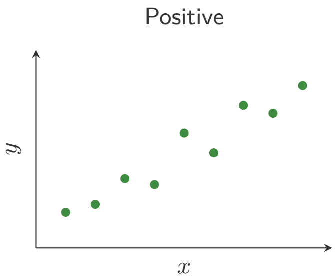
For each plot below, identify the form of the relationship:

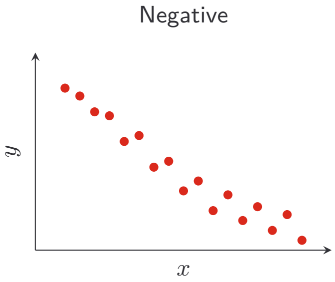


Direction

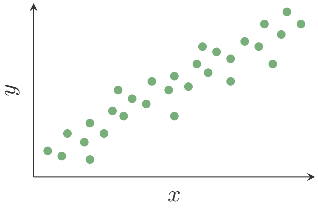
Direction

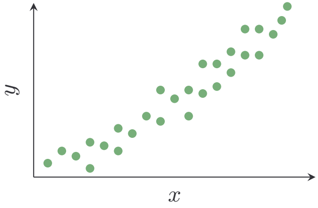
- **Positive**: as x increases, y tends to **increase**.
- **Negative**: as x increases, y tends to **decrease**.

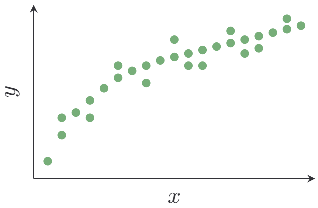


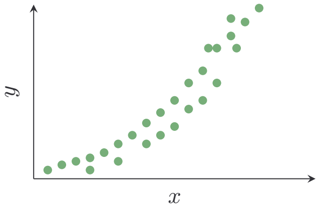


Positive Association

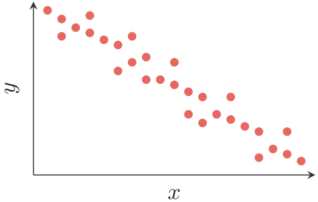


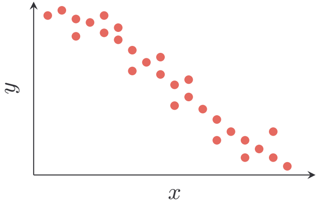


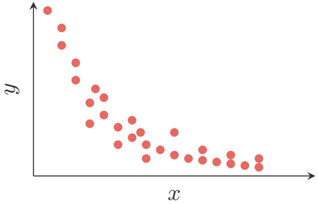


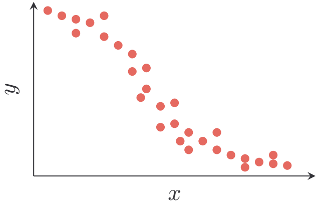


Negative Association









Strength: Strong, Moderate, or Weak

Strength

The **strength** of a relationship describes how closely the data points follow the overall pattern:

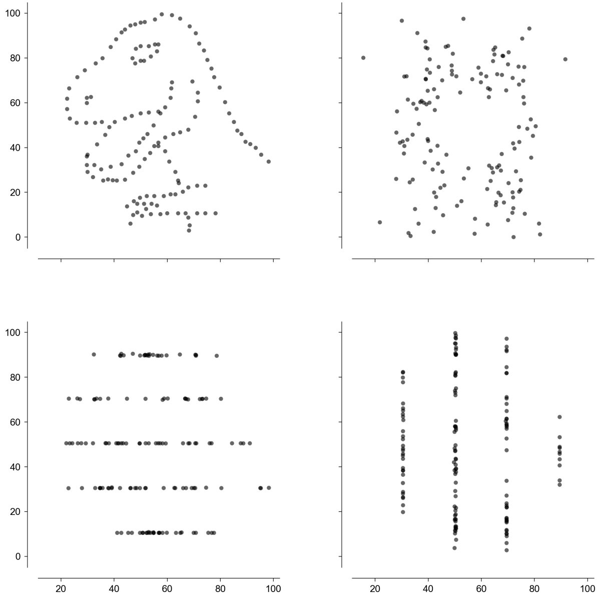
- **Strong**: Points lie close to the pattern with minimal scatter.
- **Moderate**: Points show noticeable scatter but the pattern remains clear.
- **Weak**: Points are widely scattered and the pattern is difficult to discern.

Describing a Scatterplot: Template

How to Describe a Scatterplot

1. Note any **outliers** or unusual points
2. State the **form**: linear, curved, or no clear form
3. State the **direction**: positive, negative, or none
4. State the **strength**: strong, moderate, or weak

Which plot below has the strongest linear relationship?



The Problem with "Eye-balling" Data

- **Subjectivity:** What looks "strong" to me might look "moderate" to you.
- **Scale:** Changing the axis scales can make the same data look steeper or flatter.
- **Precision:** We cannot compare relationships across different datasets (e.g., Height vs. Weight) using just adjectives.

We need an objective, numerical measure: **Correlation** (r).

What is a z-score?

z-score (Standardized Score)

The z-**score** of a value tells us how many standard deviations it is from the mean:

$$z = \frac{x - \bar{x}}{s_x}$$

- x: the value of interest
- $\bar{x}$: the mean of all values
- s_x: the standard deviation

- A positive z-score means the value is above the mean.
- A negative z-score means the value is below the mean.
- The larger the absolute value of z, the farther from the mean.

Example (`ex:ch4-z-score-practice`): Calculating z-scores

- **Example 1:** Suppose the mean height in a class is 170 cm with a standard deviation of 8 cm. What is the z-score for a student who is 178 cm tall?

- **Example 2:** If the mean test score is 75 with a standard deviation of 10, what is the z-score for a score of 60?

- **Example 3:** The average commute time is 30 minutes, with a standard deviation of 5 minutes. What is the z-score for a 35-minute commute?

Correlation Coefficient: Definition

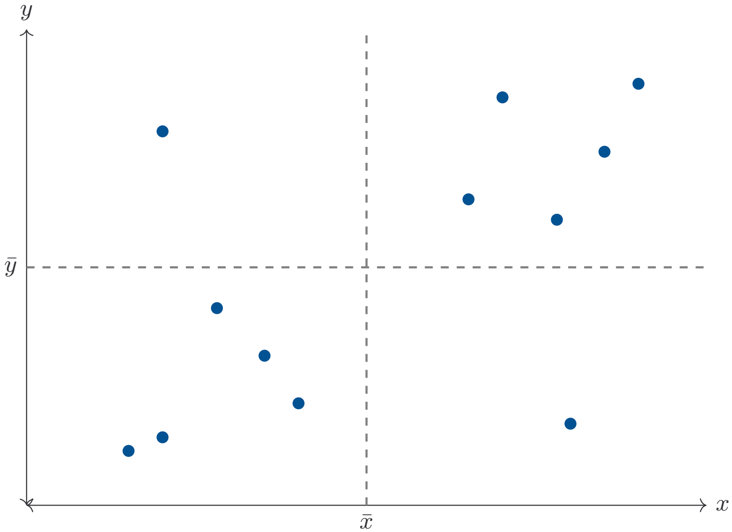
Correlation Coefficient r

$$r = \text{corr}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

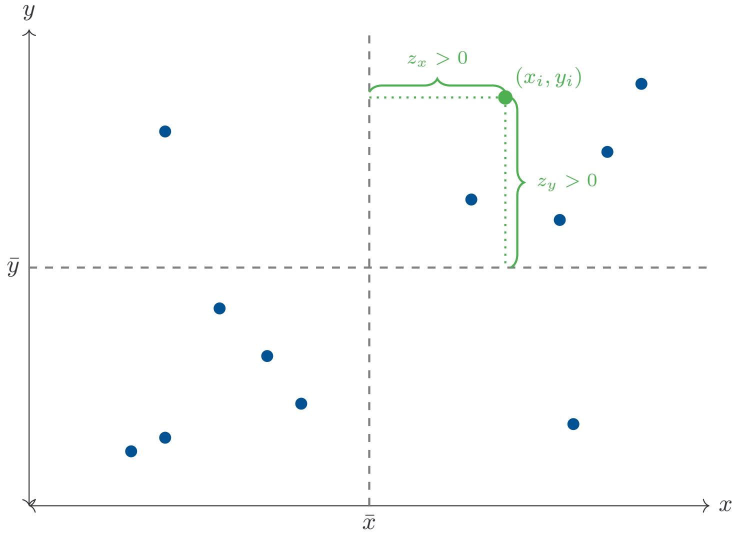
The correlation r is the "average" of the products of the z-**scores** for x and y.

When we need to emphasize the variables involved, we will use the notation $\text{corr}(x, y)$.

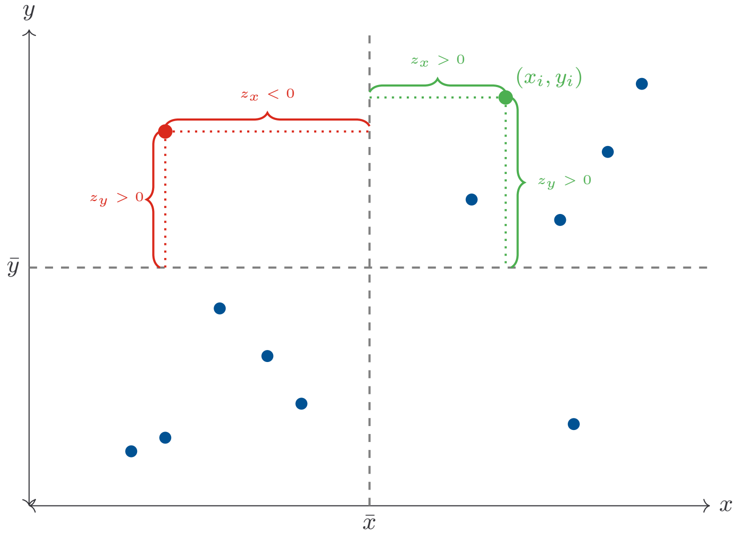
The Idea Behind Correlation



The Idea Behind Correlation



The Idea Behind Correlation



Correlation Formula Breakdown

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \times \frac{y_i - \bar{y}}{s_y}$$

$\frac{x_i - \bar{x}}{s_x}$: z-score for x

$\frac{y_i - \bar{y}}{s_y}$: z-score for y

Example (`ex:ch4-correlation-step`): Computing Correlation

The Data:

x	-1	0	1
y	1	3	2

Step 1: Compute Means and Standard Deviations

$\bar{x} = 0,\ \bar{y} = 2,\ s_x = 1,\ s_y = 1.$

Step 2: Compute the z-scores and their products for each data point.

x_i	y_i	z_x	z_y	
-1	1			
0	3			
1	2			
			Sum of Products:	

From the table, we calculated $\sum z_x z_y = 1$.

Step 3: Divide the sum of the product of the z-scores by $n-1$

Example (`ex:ch4-correlation-practice`): Computing Correlation

Find the correlation

x	8	5	5	6
y	3	3	6	8

Step 1: Compute Means and Standard Deviations

Step 2: Compute the z-scores and their products.

x_i	y_i	z_x	z_y	
8	3			
5	3			
5	6			
6	8			

Step 3: Divide by $n-1$ to get r.

Alternative Formulas for Correlation

Alternative Correlation Formulas

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Using alternative correlation formulas

Problem: Calculate r using all three formulas for: $(2,3),(4,7),(6,5)$

Step 1: Setting up the table

	Data points			Sum
x_i	2	4	6	$\sum x_i = 12$
y_i	3	7	5	$\sum y_i = 15$
x_i^2	4	16	36	$\sum x_i^2 = 56$
y_i^2	9	49	25	$\sum y_i^2 = 83$
$x_i y_i$	6	28	30	$\sum x_i y_i = 64$

Summary:

- $n = 3$
- $\bar{x} = 4$, $\bar{y} = 5$
- $s_x = 2$, $s_y = 2$

Using alternative correlation formulas (continued)

Method 1:

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

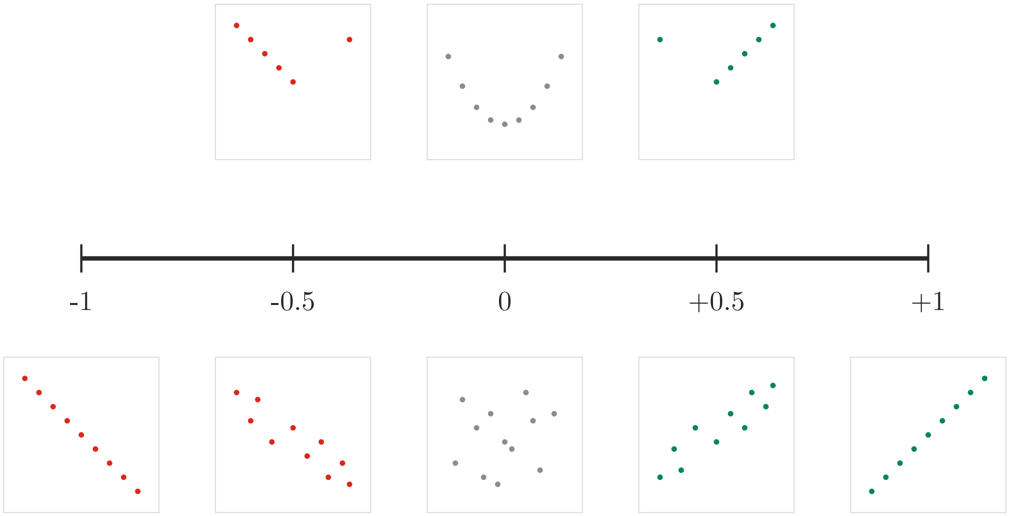
Another approach

Method 2: Deviation Formula

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}}$$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
2	3			
4	7			
6	5			
			Sum:	

Visualizing Different Values of r



Property 1: Boundedness

Bounded Range

The correlation coefficient is always between -1 and $+1$:

$$\boxed{-1 \leq r \leq +1}$$

- $r = +1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship

You will **never** compute a correlation outside this range.

Property 2: Symmetry

Symmetry

The correlation between x and y equals the correlation between y and x:

$$\boxed{\text{corr}(x, y) = \text{corr}(y, x)}$$

- There is no distinction between explanatory and response variables when computing correlation.
- The correlation between “Height” and “Weight” is the same as between “Weight” and “Height.”

Example (`ex:ch4-switching-vars`): Switching Variables

Problem: A researcher reports that the correlation between “hours of exercise per week” and “resting heart rate” is $r = -0.45$.

Find the correlation between “resting heart rate” and “hours of exercise per week.”

Property 3: Unit Invariance

Scale Invariance

Multiplying either variable by a positive constant does **not** change correlation:

$$\boxed{\text{corr}(Ax, By) = \text{corr}(x, y) \quad \text{for } A, B > 0}$$

Why? Correlation is based on **standardized scores** (z-scores), which have no units.

Example:

- Convert height from inches to centimeters (multiply by 2.54)
- Convert weight from lbs to kilograms (divide by 2.2)
- The correlation remains exactly the same!

Example (`ex:ch4-currency`): Currency Conversion

Problem: The correlation between "advertising spending (in USD)" and "sales revenue (in USD)" is $r = 0.72$.
A European branch wants to analyze the same data but in Euros (1 USD = 0.85 EUR).
What is $\text{corr}(\text{spending in EUR}, \text{revenue in EUR})$?

Property 4: Translation Invariance

Translation Invariance

Adding a constant to either variable does **not** change correlation:

$$\boxed{\text{corr}(x + a, y + b) = \text{corr}(x, y)}$$

Why? Shifting data up/down or left/right doesn't change the **shape** of the point cloud.

Example: Midterm and final exam scores

Problem: The correlation between midterm scores and final exam scores is $r = 0.68$.
The professor decides to add 5 bonus points to everyone's midterm and 10 bonus points to everyone's final exam.
What is the new correlation?

Combining Invariance Properties

Key Point: Correlation is **invariant under linear transformations**:

$$\text{corr}(Ax + a, By + b) = \text{corr}(x, y) \quad \text{for } A, B > 0$$

This means you can:

- Change units (Fahrenheit to Celsius, miles to kilometers)
- Shift baselines (add/subtract constants)
- Scale data (multiply by positive constants)

...and correlation **will not change**.

Using Properties to Simplify Calculations

The properties of correlation let us solve problems **without computation**.

Strategy: If you know corr(x, y), you can immediately deduce correlations for any linear transformation (function) of x or y.

Known: corr$(x, y) = r$ → Transformed: corr$(Ax + a, By + b) =$

Example (`ex:ch4-temperature-scales`): Temperature Conversion

Problem: A researcher finds that the correlation between daily temperature (in Celsius) and ice cream sales is $r = 0.82$.
What is the correlation if temperature is measured in Fahrenheit instead?

Recall: $F = \frac{9}{5}C + 32$

Example (`ex:ch4-standardized-test`): Standardized Test Scores

Problem: The correlation between SAT Math scores and first-year GPA is $r = 0.65$.
A new "adjusted SAT" is created: $\text{SAT}_{\text{adj}} = 2 \cdot \text{SAT} - 800$
What is $\text{corr}(\text{SAT}_{\text{adj}}, \text{GPA})$?

Example (`ex:ch4-symmetry-action`): Symmetry in Action

Problem: A study reports that the correlation between years of education and annual income is $r = 0.58$.
A journalist writes: "The correlation between income and education is different from the correlation between education and income."
Is this correct?

Sign Changes Under Negative Scaling

Effect of Negative Multipliers

For $A, B \neq 0$:

$$\text{corr}(Ax, By) = \text{sign}(A) \cdot \text{sign}(B) \cdot \text{corr}(x, y)$$

$\text{sign}(A)$	$\text{sign}(B)$	Effect on r	Example
$+$	$+$	Same sign	$\text{corr}(2x, 3y) = r$
$+$	$-$	Flips sign	$\text{corr}(2x, -3y) = -r$
$-$	$+$	Flips sign	$\text{corr}(-2x, 3y) = -r$
$-$	$-$	Same sign	$\text{corr}(-2x, -3y) = r$

Example (`ex:ch4-negative-scaling`): Negative Scaling Application

Problem: The correlation between "hours of sleep" and "number of errors on a task" is $r = -0.60$.
A researcher redefines the variables as:

- "Sleep deficit" $= 8 -$ hours of sleep
- "Accuracy" $= 100 -$ errors

Find corr(sleep deficit, accuracy).

Property 5: Perfect Correlation

Perfect Correlation

- $r = +1$: All points fall **exactly** on a line with **positive slope**.
- $r = -1$: All points fall **exactly** on a line with **negative slope**.

Example: If $y = 3x + 7$, then $r = 1$ because every point lies exactly on the line.

Example (`ex:ch4-slope-correlation`): Does Slope Affect Correlation?

Problem: Three datasets have the following exact relationships:

- Dataset A: $y = 10x$ (steep line)
- Dataset B: $y = 2x$ (moderate line)
- Dataset C: $y = 0.5x$ (shallow line)

Which dataset has the highest correlation?

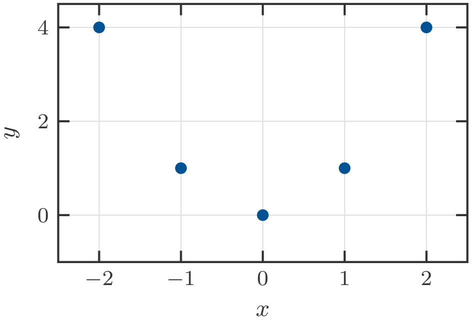
Example (`ex:ch4-zero-slope`): What About Zero Slope?

Problem: Consider the data points $(1,5),(2,5),(3,5),(4,5),(5,5)$.
All points lie exactly on the horizontal line $y=5$. What is the correlation?

Property 6: Correlation Measures Linear Relationships Only

⚠ **Caution:** Correlation (r) only measures the strength of **linear** relationships.

Check that the correlation of the following points is $r = 0$:



Example (`ex:ch4-mixed-practice-1`): Test Score Transformations

Problem: The correlation between raw test scores (x) and project grades (y) is $r = 0.75$. The instructor applies these transformations:

- Curved scores: $x_{\text{curved}} = 1.2x + 10$
- Weighted projects: $y_{\text{weighted}} = 0.8y$

(a) What is $\text{corr}(x_{\text{curved}}, y_{\text{weighted}})$?
(b) What is $\text{corr}(y, x)$?

Example (`ex:ch4-mixed-practice-2`): Health Metrics

Problem: A health study finds $\text{corr}(\text{BMI}, \text{blood pressure}) = 0.42$.
Define new variables:

- "BMI deficit" $= 25 - \text{BMI}$ (how far below "healthy" BMI)
- "Blood pressure in kPa" $= \text{BP in mmHg} \times 0.133$

Find $\text{corr}(\text{BMI deficit}, \text{BP in kPa})$.

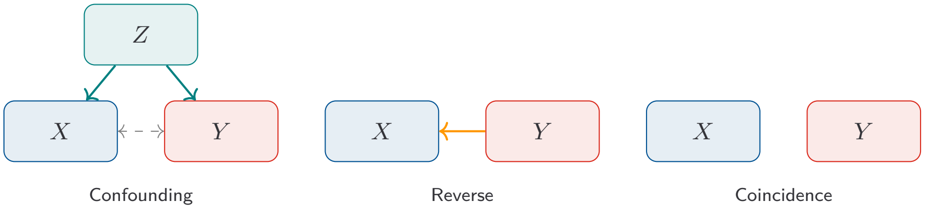
Example (`ex:ch4-mixed-practice-3`): True or False?

Problem: Identify which claims are **correct** or **incorrect**.

1. "If $r = 0$, then x and y are unrelated."
2. "The correlation between age and height is different if I measure height in feet vs. meters."
3. "If all points lie on the line $y = -5x + 100$, then $r = -1$."
4. "A steeper regression line means a higher correlation."

Correlation Does Not Imply Causation

Caution: Just because two variables are correlated does **not** mean one causes the other.

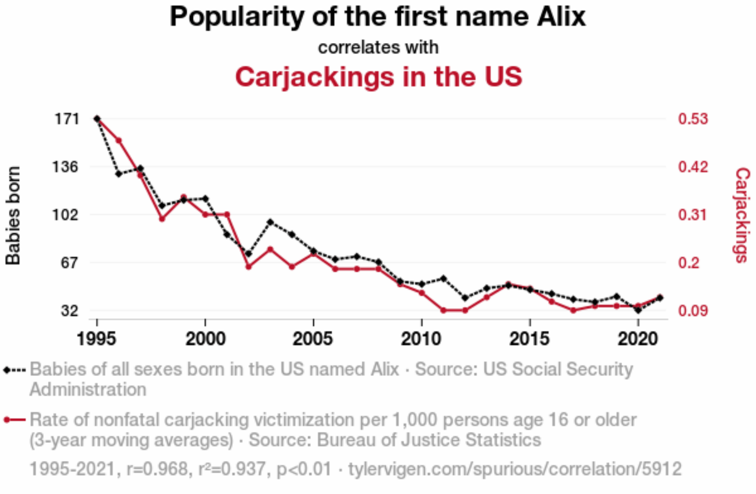


Example (`ex:ch4-confounding`): Confounding Example

Observation: Ice cream sales and drowning deaths are positively correlated.

Does ice cream cause drowning?

Example (`ex:ch4-spurious-correlation`): Spurious Correlations



Properties of Correlation: Summary

Property	What It Means
1. Boundedness	$-1 \leq r \leq +1$
2. Symmetry	$\text{corr}(x, y) = \text{corr}(y, x)$
3. Unit Invariance	$\text{corr}(Ax, By) = \lvert A \rvert \, \lvert B \rvert \, \text{corr}(x, y)$ for $A, B > 0$
4. Translation Invariance	Adding constants doesn't change r
5. Perfect Correlation	$r = \pm 1$ means points lie exactly on a line
6. Linearity Only	r only detects linear patterns

Chapter 4 Summary

Scatterplots

- Plot two or more quantitative variables on a scatterplot
- Explanatory on x; Response on y
- Identify: outliers, form, direction, strength

Correlation Formula

z-score:

$$z = \frac{x - \bar{x}}{s_x}$$

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Properties of r

- Bounded: $-1 \leq r \leq +1$
- Symmetric: order of variables doesn't matter
- Unit-free: invariant to (positive) scaling and shifting
- Measures **linear** relationships only

Caution

- Correlation $\neq$ Causation