When summary statistics lie

Four data sets. Eleven points each. Are they the same or different?
Open `data_detective.xlsx` in Google Sheets (File → Import → Upload) and work on the **Anscombe** tab.

1. Calculate the summary statistics

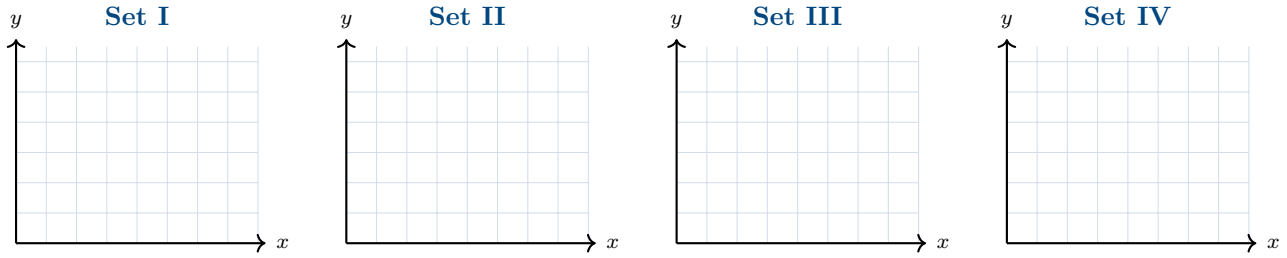
In the yellow cells of the **Anscombe** tab, enter formulas for the mean, standard deviation, and correlation of each set. Useful: `=AVERAGE(...)`, `=STDEVP(...)`, `=CORREL(..., ...)`. Copy your results below, to 2 d.p.

	Set I		Set II		Set III		Set IV	
Statistic	x	y	x	y	x	y	x	y
Mean								
Std. dev.								
Correlation r								

What do you notice about the four columns of statistics?

2. Predict the scatter plots

Based only on the statistics above, sketch what you *predict* each scatter plot will look like.



3. Plot them for real

In Google Sheets, select the x and y columns for Set I, then Insert → Chart → Scatter. Repeat for Sets II, III, IV.

How were your predictions? Describe how each actual plot differs from the others, even though their summary statistics are nearly identical.

Anscombe's quartet — published by Francis Anscombe in 1973 to demonstrate why plotting matters.

When the groups disagree with the whole

Two tutoring methods, A and B, were tested at Lakeside School. Students from each group sat one of two exam papers. Results are on the **Simpson** tab.

	Multiple-choice exam		Long-answer exam	
Method	Passed	Total	Passed	Total
A	81	87	192	263
B	234	270	55	80

4. Calculate the pass rates

In the yellow cells on the **Simpson** tab, calculate each pass rate. Format as percentages (Format → Number → Percent). Copy your results here:

Method	Multiple choice	Long answer	Overall passed / total	Overall rate
A			/	
B			/	

5. Which method is better?

Write an inequality each time — e.g. "93.1% > 86.7%, so Method A is better."

(a) Among **multiple-choice** students:

(b) Among **long-answer** students:

(c) **Overall**, across all 700 students:

6. What is going on, and what would you recommend?

Method A wins both subgroups but loses overall. Look back at the **Total** columns. **Why does this reversal happen? Refer to specific numbers from the table.**

If you were the principal, which method would you recommend, and why?

When the graph misleads

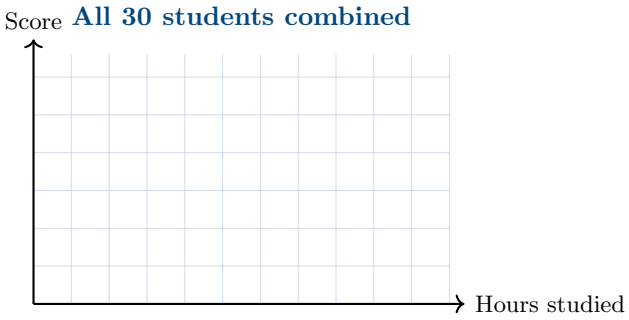
Thirty Year 10 students sat one of three maths papers — **Foundation**, **Standard**, or **Extension** — each with different content and difficulty. For each student we have the hours per week they studied and the score they earned.

Switch to the **Hidden Groups** tab of `data_detective.xlsx`.

7. Plot *all* the data together

Use columns `J` and `K` (all 30 students stacked). Select `J5:K35`, then Insert → Chart → Scatter. Add a linear trendline with Use Equation.

Sketch the scatter and trendline below:



Equation of the line of best fit: $y =$ _____ $x +$ _____

8. Find the correlation r

In the yellow cell next to "**r (All 30 combined)**", enter `=CORREL(J6:J35, K6:K35)`.

State the value: $r =$ _____

Is r positive or negative? Strong or weak?

Based on this plot alone, does studying more hours seem to lead to higher or lower scores?

Something seems wrong. More studying → lower scores? Turn over to find the hidden truth.

Group the same data — and watch the trend flip

The data is the same. The students are the same. But now we will look at **each paper separately**.

9. Find the correlation *inside* each group

In the yellow cells, enter:

- Foundation: `=CORREL(A6:A15, B6:B15)`
- Standard: `=CORREL(D6:D15, E6:E15)`
- Extension: `=CORREL(G6:G15, H6:H15)`

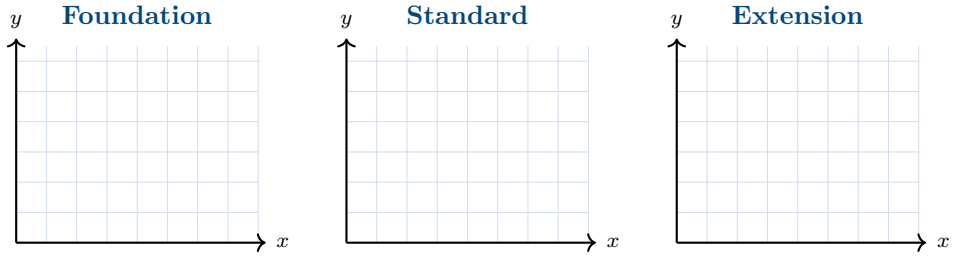
Record the values to 3 d.p.:

	Foundation	Standard	Extension
r			

What do all three group correlations have in common, and how does this compare to the combined r on side 3?

10. Plot each group — what does each trendline do?

Insert a Scatter chart on the **Hidden Groups** tab. Add three series (instructions are on the tab), one per group, and add a **linear trendline** to each. Sketch the trendline you see for each group below:



What direction does each trendline go? Compare with the combined trendline from Side 3.

11. Explain the paradox

What was the hidden variable that made the combined plot misleading? Refer to specific groups and numbers in your explanation.

Connect: across all three tasks (Anscombe, Simpson, Hidden Groups), what **general lesson** would you give to someone reading a statistic in a newspaper?