Unit 1 · Exploring Data Cheat Sheet

Unit 1 focus: displays, shape, center, spread, box plots, outliers, normal models and z-scores

1 KEY VOCABULARY

categorical variable	Places each individual in a group.	eye color, grade level
quantitative variable	A number you can average.	height, test score
distribution	The values a variable takes and how often.	a histogram of scores
outlier	A value far from the rest (1.5 · IQR rule).	30 in 3, 5, 7, …, 14
resistant	Not pulled much by outliers.	median and IQR

2 DESCRIBING A DISTRIBUTION

SOCS	What to say
Shape	symmetric, skewed left / right, unimodal / bimodal
Outliers	name any unusual values (or say none)
Center	mean (symmetric) or median (skewed)
Spread	SD (symmetric) or IQR (skewed); range

Always answer **in context**. Skewed right (long tail right): mean > median.

3 MEAN AND STANDARD DEVIATION

$\bar{x} = \frac{\sum x}{n}$, $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$: the typical distance from the mean.

Example: Data: 4, 6, 7, 9, 14
Mean: $\frac{40}{5} = 8$
Squared deviations: 16, 4, 1, 1, 36 (sum 58)
s: $\sqrt{58/4} = \sqrt{14.5}$

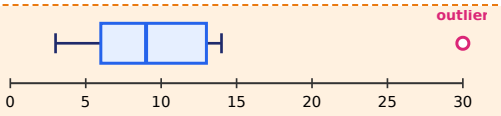
Answer: $\bar{x} = 8$, $s \approx 3.81$

4 BOX PLOTS AND THE OUTLIER RULE

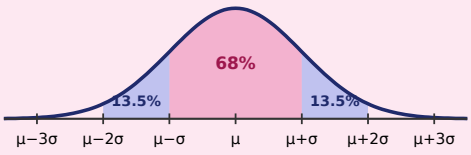
Outlier if below $Q_1 - 1.5\,\text{IQR}$ or above $Q_3 + 1.5\,\text{IQR}$.

Example: 3, 5, 7, 8, 9, 11, 12, 14, 30
Summary: min 3, Q_1 6, median 9, Q_3 13, max 30
Fences: IQR = 7, so $6 - 10.5 = -4.5$ and $13 + 10.5 = 23.5$

Answer: 30 is an outlier



5 NORMAL MODELS AND Z-SCORES



Empirical rule: **68%** within 1σ, **95%** within 2σ, **99.7%** within 3σ. $z = \frac{x-\mu}{\sigma}$ counts SDs from the mean. SAT $N(500, 100)$: a 650 has $z = 1.5$; $P(Z > 1.5) \approx 0.067$.

6 QUICK REMINDERS & COMMON MISTAKES

✓ Do remember

- Label displays and describe them in context.
- Use median & IQR when data are skewed or have outliers.
- A histogram's bars touch; a bar graph's don't.

✗ Common mistakes

- Dividing by n for s: divide by $n - 1$
- "The data is normal": say "approximately normal"
- Comparing only centers: compare SOCS for both groups

Unit 1 · Exploring Data Cheat Sheet

More practice: two-way tables, scatterplots, correlation, least-squares regression and residuals

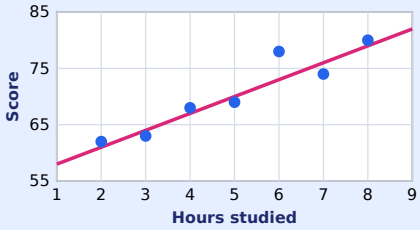
1 TWO-WAY TABLES

	Plays sport	No sport	Total
Grade 10	45	55	100
Grade 11	60	40	100
Total	105	95	200

Marginal: $P(\text{sport}) = \frac{105}{200} = 52.5\%$. Conditional: $\frac{45}{100} = 45\%$ of 10th graders vs 60% of 11th graders. Different conditional percents $\Longrightarrow$ an **association**.

2 SCATTERPLOTS AND CORRELATION

Describe **direction, form, strength** and unusual points. Correlation $-1 \le r \le 1$ measures the strength of a **linear** relationship; it has no units and is not resistant.



3 LEAST-SQUARES REGRESSION LINE

$\hat{y} = a + bx$ with $b = r\,\frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$. The line passes through $(\bar{x}, \bar{y})$.

Example: $\bar{x} = 5$, $s_x = 2$, $\bar{y} = 70$, $s_y = 8$, $r = 0.75$
Slope: $b = 0.75 \cdot \frac{8}{2} = 3$
Intercept: $a = 70 - 3(5) = 55$
Predict at 6 h: $55 + 3(6) = 73$

Answer: $\hat{y} = 55 + 3x$; predicted score 73

4 INTERPRETING THE MODEL

Value	Interpretation in context
Slope $b = 3$	predicted score rises 3 points per extra hour
Intercept $a = 55$	predicted score for 0 hours (only if 0 makes sense)
$r^2 = 0.75^2$	56.25% of the variation in score is explained by the line
s (residual SD)	typical prediction error, in y units

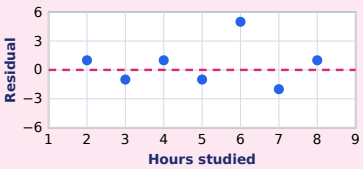
Template: "The predicted *y* increases/decreases by *b* for each 1-unit increase in *x*."

5 RESIDUALS

Residual = actual − predicted = $y - \hat{y}$. A residual plot with **no pattern** means a linear model fits.

Example: A student who studied 6 h scored 78.
Predicted: 73
Residual: $78 - 73 = 5$

Answer: +5: the line under-predicted by 5



6 CAUTIONS WITH REGRESSION

✓ Do remember

- Plot the data and the residuals first.
- Use $\hat{y}$ ("predicted") in the equation.
- Influential points (far right or left) can change the slope a lot.

✗ Common mistakes

- Correlation $\Longrightarrow$ causation: association only
- Predicting far outside the data: that is extrapolation
- Using r for a curved pattern: r is for linear form