

Chapter 6 Solutions

Making Sense of Social Data

Mark Paradis, Ph.D.

September 2, 2026

These are the worked solutions to the *end-of-chapter* exercises for Chapter 6 of *Making Sense of Social Data*. Solutions to the exercises inside each section are in the book itself, printed immediately after the exercises they answer.

Do the exercise before you read the solution. A solution read first is just more text to read; a solution read second is a check on your reasoning, and only the second one teaches you anything.

The complete set for every chapter is at <https://www.markparadispolitics.com/books/making-sense-of-social-data/online/solutions/>.

Chapter 6 Organizing and Visualizing Data

Solution to Exercise 6.E1. (a) The four patterns.

I is what the summary statistics suggest: a roughly linear relationship with ordinary scatter around it. This is the only one for which the correlation is a fair description.

II is a clean curve. The points follow a smooth arc, rising and then falling. A straight-line summary is simply the wrong shape, and the correlation understates a relationship that is nearly perfect — just not linear.

III is a perfect straight line with one outlier far off it. Without that single point the correlation would be almost exactly 1; with it, the fitted line is pulled away from every other observation.

IV has all but one point at the same x value, with a single case far to the right. That one case creates the entire correlation. Remove it and there is no relationship at all, because x barely varies.

(b) That summary statistics describe particular features of a distribution and are silent about everything else. The mean, variance, and correlation are identical across all four, so those three numbers cannot distinguish a linear relationship from a curve, from a line with an outlier, from a pattern produced entirely by one case.

Note what this does *not* show. The statistics are not wrong — each is computed correctly and reports exactly what it reports. The error would be in treating them as a complete description.

(c) Plot the data before computing anything from it, and plot it again before believing anything you computed. Specifically: a correlation coefficient should never be reported without the scatterplot having been examined, because the same coefficient is produced by shapes with entirely different meanings and entirely different implications for what to do next.

Solution to Exercise 6.E2. Your answer depends on the table you find. Four things make it complete.

Build the frequency table correctly. Absolute frequencies summing to the stated total, relative frequencies as percentages, and a note if rounding prevents them summing to exactly 100. Include a cumulative column only if the variable is ordinal.

Let the variable choose the chart, not your preference. Categorical variable, bar chart; numerical, histogram. If you choose a pie chart, defend it against §6.3.5 — and that defence is available only for two or three categories of clearly different size.

Include the required elements in the sketch. Labelled axes with units, a descriptive title, a source note, and an axis starting at zero if the chart uses bars.

Make the misleading version specific. “You could make it look bigger” is not yet an answer. “Starting the vertical axis at 40 rather than 0 would make a

six-point difference fill the panel” is: it names a technique from §6.7 and explains the mechanism that produces the impression.

Solution to Exercise 6.E3. (A) *Annual household income*. Right-skewed. Bounded below at zero, unbounded above, with a small number of very high incomes stretching the upper tail. *Median*, because the tail drags the mean above the typical household. *Check*: read the codebook before anything else, and find out how the file stores a refusal. Some datasets use a sentinel code such as 999999, which enters every calculation as though it were an income; others top-code, which caps the variable instead; socsurvey top-codes. Then look at the zeros — a household reporting exactly zero income may be a genuine case or a missing value coded as 0.

(B) *Self-reported health, 1–5, strong public health system*. Left-skewed. Most respondents will report good or very good health, piling the data near the top of the scale and leaving the tail to stretch downward — a ceiling effect. *Median*, since the variable is ordinal and the distribution is skewed. *Check*: confirm the coding direction. Whether 1 means excellent or poor determines whether the distribution is left- or right-skewed, and codebooks differ.

(C) *Number of children under 18*. Right-skewed and discrete, with a large spike at zero, since most households contain no children under 18. *Median*, though the mean is also reported for this variable because policy questions concern totals. *Check*: confirm whether households with no children are included in the data at all — if the variable was only asked of parents, the spike at zero disappears and the distribution means something different.

(D) *Age at first marriage*. Roughly symmetric, or mildly right-skewed. Bounded below by legal minimums and clustered in the late twenties, with a modest tail of later marriages. *Mean* is defensible if the distribution is near-symmetric; the median is safer.

Watch for bimodality here. If men and women marry at different average ages and both are in the same histogram, two peaks may appear — and the fix is to separate the groups rather than to describe the combined shape (§6.4.3). *Check*: whether the variable includes only those who have married, which excludes everyone who has not and makes the sample a selected one.

Solution to Exercise 6.E4. (a) Two problems.

Area is not population. Northern census divisions are enormous and hold very few people. Shaded on a map they dominate the visual field, so a reader’s impression of “how much of Canada lags” is driven by territory rather than by households. The map overstates the geographic reach of the finding.

Median household income by division conceals variation within divisions. A division with a low median may contain both high- and low-income communities, and the map cannot show that. Reading the map as a statement about households in

the north is the ecological fallacy (§6.6.3).

A third defensible answer: class breaks. Where the shading boundaries fall determines which divisions appear dark, and equal-interval, quantile, and natural-break schemes would produce visibly different maps from identical data.

(b) A *sequential* scale — light to dark in one hue, darker meaning higher income. Median income runs from low to high with no meaningful midpoint dividing it into two kinds of value, which is what a diverging scale would imply. A diverging scale would be appropriate for a related but different map: income relative to the national median, where the national figure is a genuine midpoint separating above from below.

(c) A *proportional symbol* map. The revised question asks for a count — how many households — and counts scale with population. On a choropleth, a densely populated urban division containing tens of thousands of low-income households would occupy a small patch, while a northern division containing a few hundred would occupy a large one.

A proportional symbol map detaches the quantity from the area, so each division carries a circle sized by its count. The general rule: choropleths for rates, proportional symbols for counts.

Solution to Exercise 6.E5. 1. *Bar chart*, sorted by frequency. A pie chart is worse because nine categories cannot be compared as wedges, and several will be close in size.

2. *Histogram*. A bar chart is worse — it treats each distinct income as its own category, which for a continuous variable produces hundreds of one-case bars showing nothing.

3. *Frequency polygons*, one per group, plotted as percentages within each group. Two overlapping histograms are worse because the bars hide one another, and plotting counts would be worse still if the two groups differ in size.

4. *Small multiples*, twelve panels sharing one axis range. A single twelve-line chart is worse because no encoding keeps twelve series distinguishable, and the reader must trace each line behind eleven others.

5. *Dot plot*. A histogram is worse: with fifteen cases the bins hold one or two each, so the chart shows the binning rather than the data.

6. *Scatterplot*. Anything else is worse, because both variables are numerical and the question is about their relationship — which only a scatterplot displays.

The pattern the exercise is built around: the chart follows from the variable type and the question, in that order. Two numerical variables and a relationship question always give a scatterplot; one categorical variable and a distribution question always give a bar chart.

Solution to Exercise 6.E6. Your answer depends on the variable you choose. Four

things make it complete.

Classify the variable before choosing the chart. A histogram of a nominal variable stored as integers — region coded 1 to 7, say — is the failure §6.2.3 exists to prevent. Ask what kind of variable it is first, and let the answer choose the chart.

Record your choices honestly. Bin width, category order, axis range and missing-value treatment are all choices, including when you did not notice making them. Say what you chose and why — and “I accepted the software’s default bin width” is a legitimate entry, provided you write it down. The point of the record is that a default you accepted knowingly can be defended, and one you never saw cannot.

Evaluate your own chart properly. It is easy to produce a chart and declare it compliant. Go looking for something to change: a missing source note, an unlabelled axis, gridlines heavier than they need to be. If you find nothing at all, you have probably not looked at the checklist in §6.8 item by item.

Make the misleading version deliberately, and label it. This is the most instructive part of the exercise. Take your own honest chart, truncate its axis, name the technique you have just used, and say why you would submit the honest version. If you cannot find a way to make your chart mislead, reread §6.7 — almost any chart can be made to.

One thing you may notice, and it is worth noticing: the misleading version often looks more like the charts published in newspapers than the honest one does.

Solution to Exercise 6.E7. Seven problems.

3D effect. Distorts perception — perspective makes rear bars appear smaller and makes each bar’s height ambiguous. Prohibited by APA. Correction: flat 2D bars.

Colour-only encoding, including red and green. Fails for readers with the most common colour vision deficiency and in any greyscale reproduction. Correction: if the bars are five levels of one variable, one fill is correct and the colours are unnecessary entirely.

Truncated axis (15 to 40 per cent). A bar encodes quantity as length, so this makes the lengths misstate the vote shares. Correction: start at zero, without exception for bar charts.

Unlabelled vertical axis. The reader cannot tell what is measured or in what units. Correction: label it, with units.

Grey background and heavy gridlines. Non-data ink competing with the data. Correction: white background, gridlines removed or reduced to faint horizontal lines.

Alphabetical ordering. Party is nominal, so any order is a choice, and alphabetical arranges the data by an accident of spelling. Correction: sort by vote share, descending.

“Figure 1” as the caption. That is a number, not a title. Note that its *position* is

not the problem: in APA 7 the figure number and its italic title both sit above the figure, exactly as for a table (§6.8.4). What is missing is the title. Correction: keep the number where it is, put a descriptive italic title on the line beneath it, and add the source in an italicised *Note*. below the figure.

Older advice will tell you to move the whole caption below. That was APA 6, and a great deal of guidance still online has not caught up.

Which single change matters most? The truncated axis. The others make the chart uglier, harder to read, or non-compliant; the truncated axis makes it *wrong* — a reader who looks at it carefully and reads the axis will still come away with a distorted impression of the differences between parties. The 3D effect and the colours are real problems of a lesser kind, and worth naming as such — but they are not what makes the figure misleading.

Solution to Exercise 6.E8. Your answer will use different metropolitan areas from the ones below; the reasoning is the same whichever three you picked.

(a) Three series on one chart, distinguished by line type and marker shape rather than by colour alone (§6.3.4). Three colours would produce a chart that fails in this book's own printing.

(b) Twelve lines cross repeatedly and no encoding keeps them distinguishable. The fix is small multiples — twelve panels sharing one axis range — or highlighting two or three series in black against the remaining nine in light grey. Either is acceptable; the shared axis range is not optional.

(c) Probably yes, and this is the point of the question. A definitional break usually shows as a step: the series jumps at one wave and continues at a different level, in a way that no gradual social process would produce. Look at waves 11 to 13 and you should see it.

What it suggests: an unexplained discontinuity in a time series you did not collect is more likely a change in measurement than a change in the world, and the codebook is where to check. This is the same lesson as §4.5.3 — definitions drift, and a break can look exactly like a finding.

(d) The simulated series will not match the real levels, and should not. What you should find is that the general shape and the ordering across CMAs are broadly plausible while the numbers themselves are wrong — which is exactly what the Preface promises. The resemblance is in trend and ordering, not in level, and that is the design working as intended rather than a defect in it.