

Chapter 3

PRINCIPLES OF ANALYTIC GRAPHICS FOR MEANINGFUL DATA VISUALIZATIONS

DATA ANALYSIS IN R

Outline

- Principle 1: Show comparisons
- Principle 2: Show causality, mechanism, explanation, systematic structure
- Principle 3: Show multivariate data
- Principle 4: Integrate evidence
- Principle 5: Describe and document the evidence
- Principle 6: Content is the King

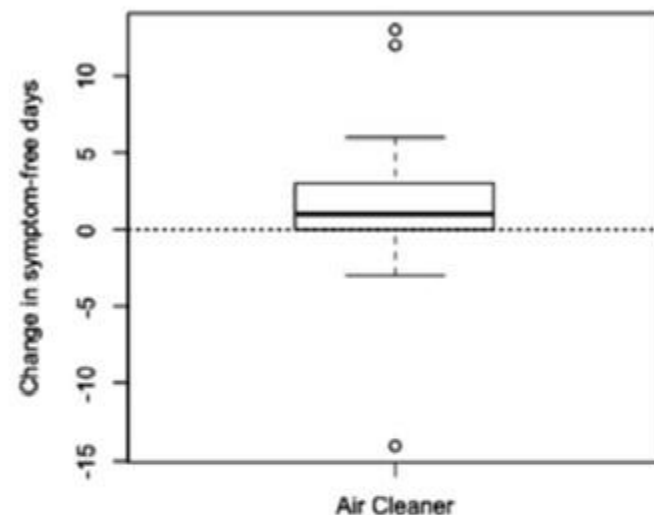
Principle1: Show comparisons

- Evidence for a hypothesis is always *relative* to another competing hypothesis.
- When you say, "the evidence favors hypothesis A", what you mean to say is that "the evidence favors hypothesis A versus hypothesis B".
- Good scientist is always asking "Compared to What?" when confronted with a scientific claim or statement.
- Data graphics should generally follow this same principle: always compare at least two things.

Show comparisons: Example

Boxplot which looks at the effect of an air cleaner on the asthma symptoms of children who is in a set of homes.

- ▶ Study conducted at the Johns Hopkins University School of Medicine and was conducted in homes where a smoker was living for at least 4 days a week.
- ▶ 47 children who received the air cleaner and was assessed at baseline and then 6-months later at a second visit
- ▶ The idea is that an air cleaner was introduced into a child's home to reduce indoor air pollution levels.
- ▶ we want to see if their asthma symptoms are improving. So the outcome is what's called symptom-free days: a higher number is better.



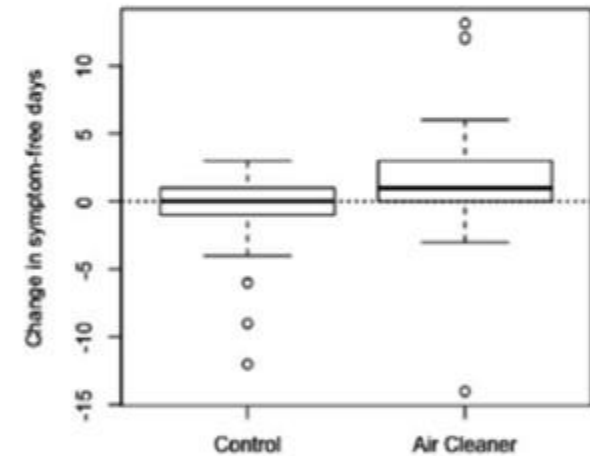
Is this a positive outcome?
Compared to what??

Show comparisons: Example

- **Boxplot which looks at the effect of an air cleaner on the asthma symptoms of children**

► Figure shows both the group that received the air cleaner and the control group that did not.

On average the number of symptom-free days increased by about 1 day (the solid line in the middle of the box is the median of the data → Improvement when introducing Air Cleaner



Principle2: Show causality, mechanism, explanation, systematic structure

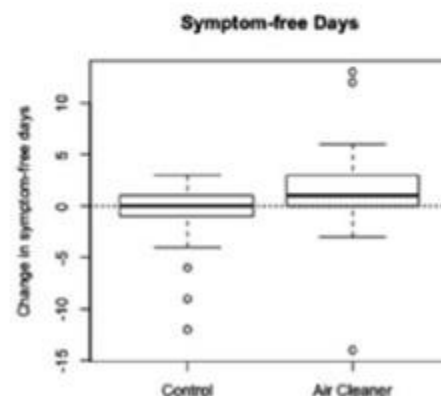
- How do you believe the system work
- What is the causal framework for thinking about a question?
 - It is difficult to prove that one thing causes another thing even with the most carefully collected data.
 - But it's still often useful for your data graphics to indicate what you are thinking about in terms of cause.

Such a display may suggest hypotheses or refute them, but most importantly, they will raise new questions that can be followed up with new data or analyses.

Example:

From the plot, it seems clear that on average, the group that received an air cleaner experienced improved asthma morbidity (more symptom-free days, a good thing).

An interesting question might be "Why do the children with the air cleaner improve?"

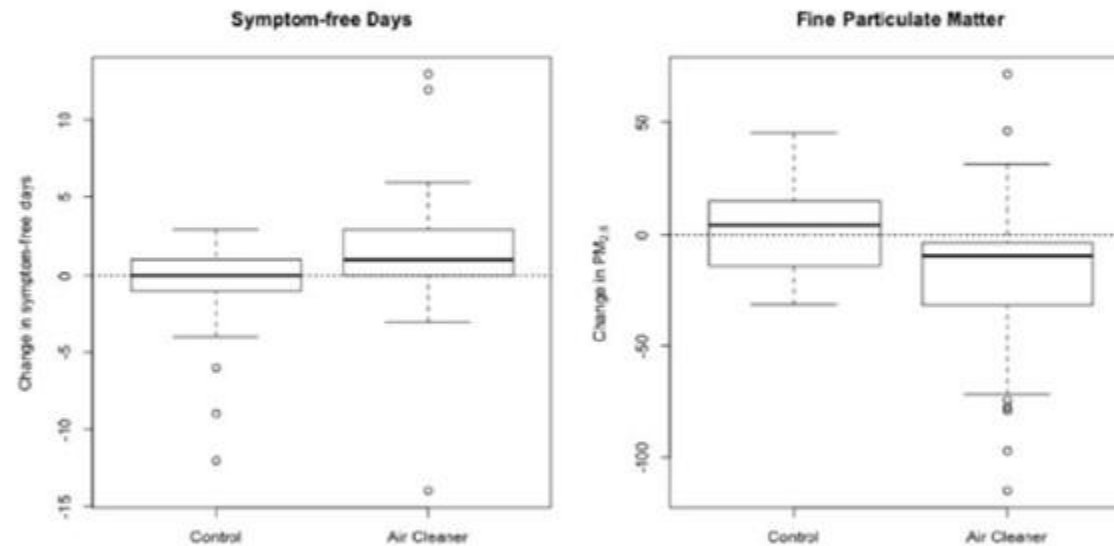


Principle2: Show causality, mechanism, explanation, systematic structure

- **Answering the question of "why?" might lead to improvements or new developments.**
 - Why do the children with the air cleaner improve?
- **The hypothesis behind air cleaners improving asthma morbidity in children is that the air cleaners remove airborne particles from the air.**
 - homes had smokers => high level of particles in the air, primarily from second-hand smoke=> inhaling fine particles can exacerbate asthma symptoms
 - Therefore, we would expect that the group receiving the air cleaners should on average see a decrease in airborne particles.
 - In this case we should track fine particulate matter, also called PM2.5 which stands for particulate matter less than or equal to 2.5 microns in aerodynamic diameter.

Principle2: Show causality, mechanism, explanation, systematic structure

- The plot shows both the change in symptom-free days for both groups (left) and the change in PM2.5 in both groups (right).



On average in the control group, the level of PM2.5 actually increased a little bit while in the air cleaner group the levels decreased on average. This pattern shown in the plot above is consistent with the idea that air cleaners improve health by reducing airborne particles

Principle3:Show multivariate data

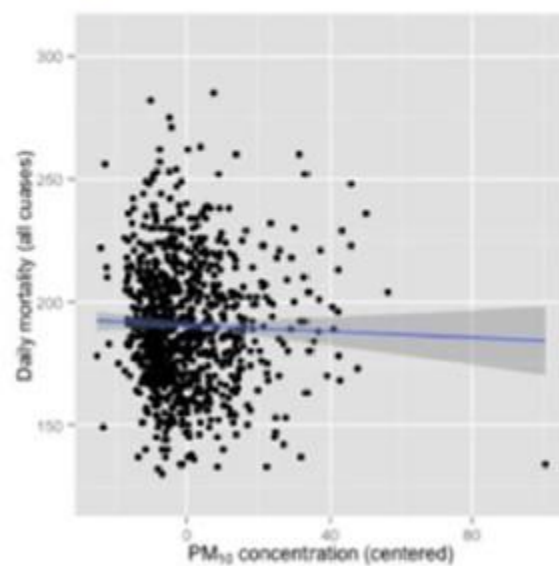
- Multivariate= more than 2 variables
- The real world is multivariate.
- For anything that you might study, there are usually many attributes that you can measure.
- Data graphics should attempt to show this information as much as possible.
- There are a variety of ways that you can show multivariate data

Principle3:Show multivariate data

- **Example:**

- data on daily airborne particulate matter ("PM10") in New York City and mortality from 1987 to 2000.
- Each point on the plot represents the average PM10 level for that day (measured in micrograins per cubic meter) and the number of deaths on that day.
- The PM10 data come from the U.S. Environmental Protection Agency and the mortality data come from the U.S. National Center for Health Statistics.

- ❑ This is a bivariate plot showing two variables in this dataset.
- ❑ From the plot it seems that there is a slight negative relationship between the two variables : higher daily average levels of PM10 appear to be associated with lower levels of mortality

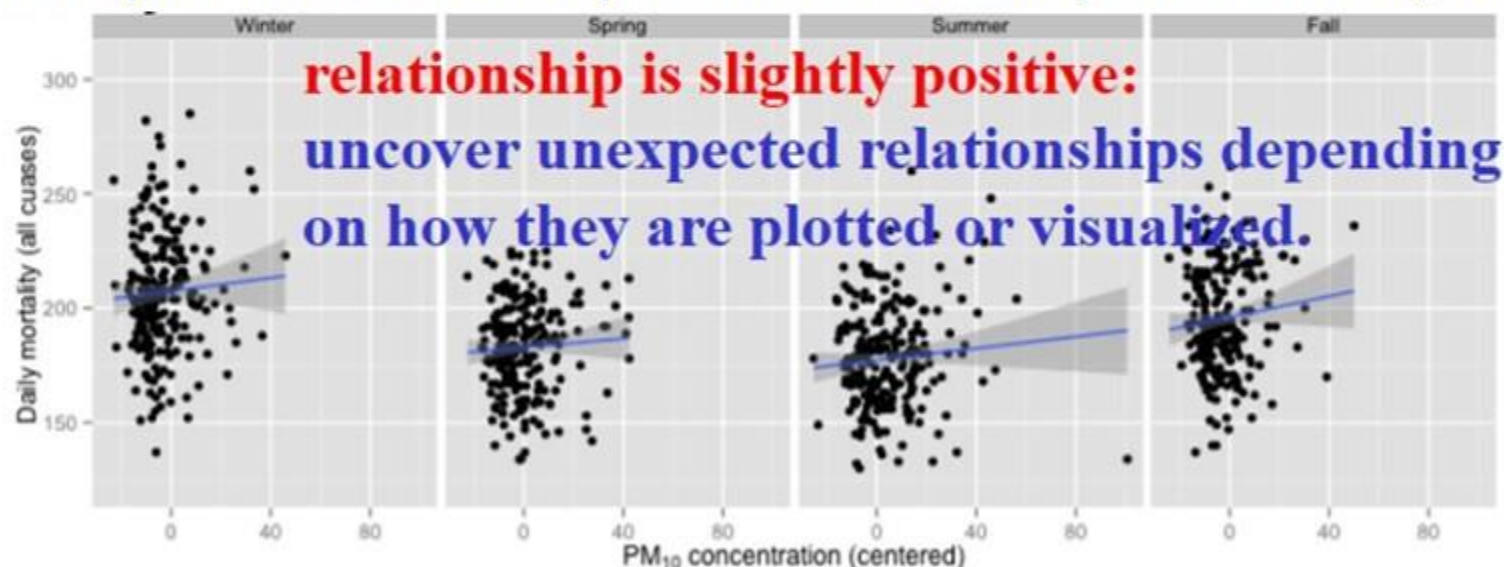


PM10 and mortality in New York City

Principle3:Show multivariate data

- However, there are other factors that are associated with both mortality and PM10 levels: the season.
 - It's well known that mortality tends to be higher in the winter than in the summer.
 - PM10 and mortality have opposite seasonality with mortality being high in the winter and PM10 being high in the summer.

What happens if we plot the relationship between mortality and PM10 *by season*?



PM10 and mortality in New York City by season

Principle 4: Integrate evidence

- Just because you may be making data graphics, doesn't mean you have to rely solely on circles and lines to make your point.
- Include printed numbers, words, images, and diagrams to tell your story.
- Data graphics should make use of many modes of data presentation simultaneously, not just the ones that are familiar to you or that the software can handle.
 - One should never let the available tools drive the analysis
 - One should integrate as much evidence as possible to a graphic.

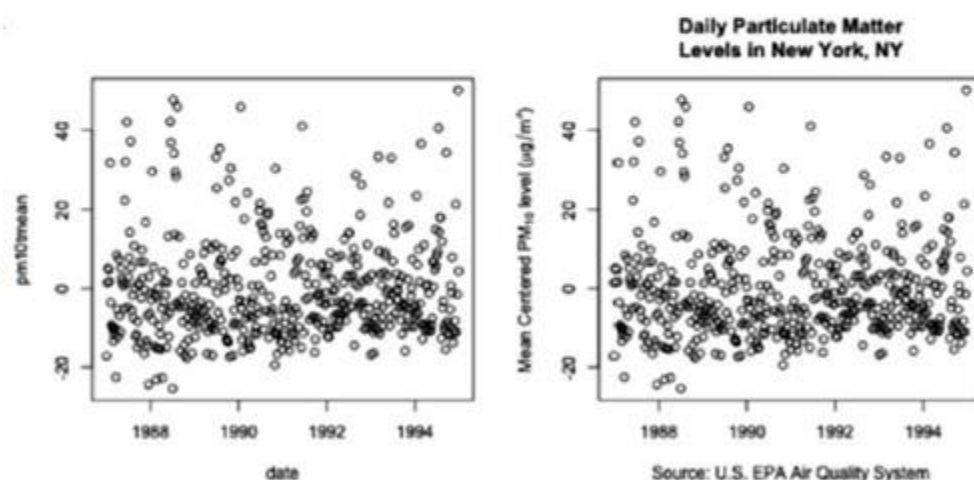
Principle 5: Describe and document the evidence

- Data graphics should be appropriately documented with labels, scales, and sources.
- Datagraphic should tell a complete story all by itself.
- Ask the question: Is there enough information on that graphic for the person to get the story?

- The plot on the left is a default plot generated by the plot function in R.
- The plot on the right uses the same plot function but adds annotations like a title, y-axis label, x-axis label.

Key information included is where the data were collected (New York), the units of measurement, the time scale of measurements (daily), and the source of the data (EPA).

Plot of the same data twice!



Labelling and annotation of data graphics

Principle 6: Content is the king

- **Analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content.**
- **Question being asked and the evidence presented in favor of certain hypotheses.**
 - No amount of visualization magic or bells and whistles can make poor data, or more importantly a poorly formed question, shine with clarity.
 - Starting with a good question, developing a sound approach and only presenting information that is necessary for answering that question, is essential to every data graphic.

Questions