08

General Rules for Exploratory Data Analysis

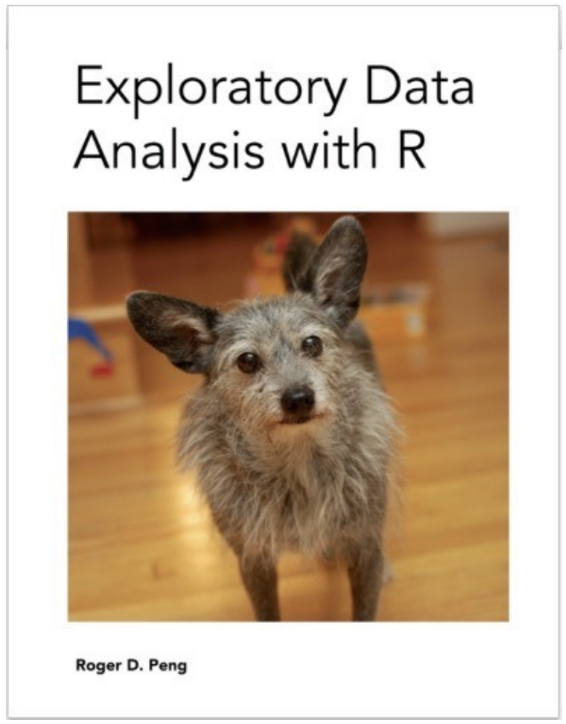
Notice

- **Author**
  - **João Moura Pires (jmp@fct.unl.pt)**

- **This material can be freely used for personal or academic purposes without any previous authorization from the author, provided that this notice is kept with.**

- **For commercial purposes the use of any part of this material requires the previous authorisation from the author.**

Bibliography….



Exploratory Data Analysis with R

Roger D. Peng

Table of Contents

General Rules for Exploratory Data Analysis

Principles of Analytic Graphics

- **Principle 1: Show comparisons**
  - **Evidence for a hypothesis is always relatives another competing hypothesis**
  - **Always ask “Compared to What?"**

Principles of Analytic Graphics

- **Principle 1: Show comparisons**
  - **Evidence for a hypothesis is always relatives another competing hypothesis**
  - **Always ask “Compared to What?"**

Testing whether an **air cleaner installed in a child’s home improves their asthma-related symptoms**.

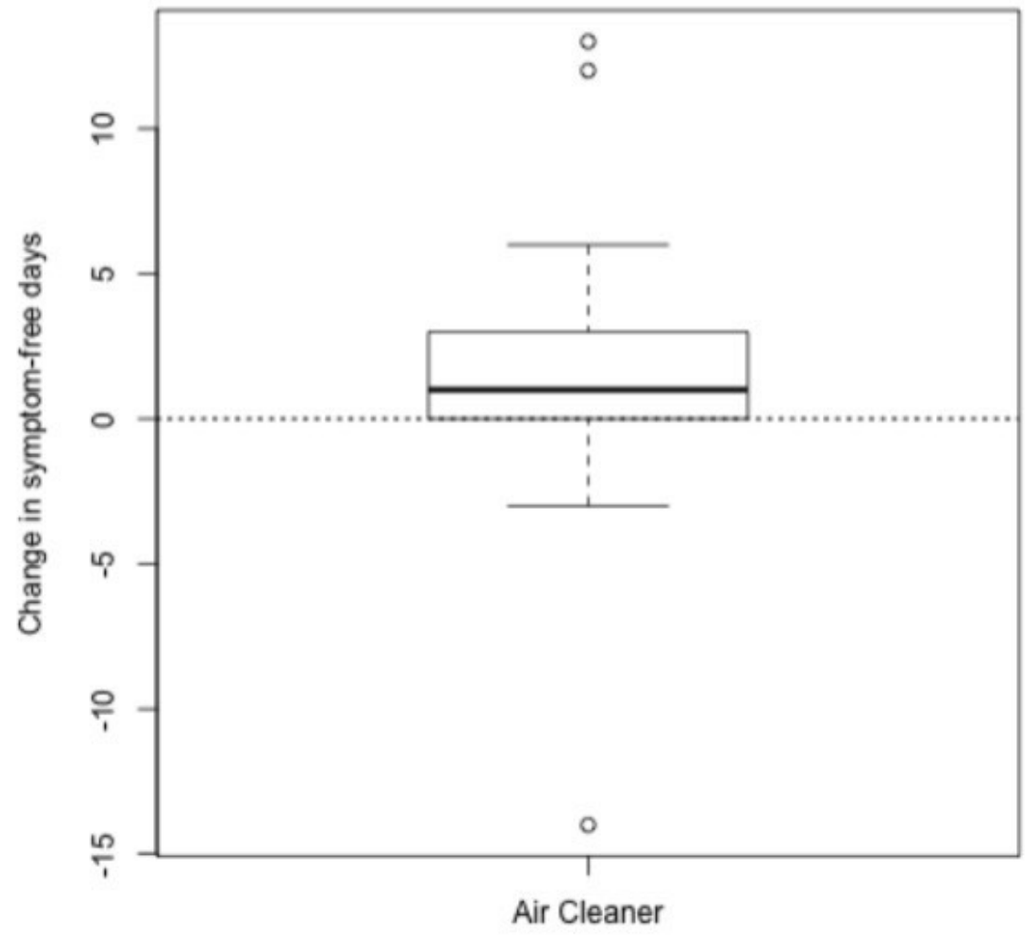
This study was conducted at the Johns Hopkins University School of Medicine and was conducted in homes where a smoker was living for at least 4 days a week.

Each child was assessed at baseline and then 6-months later at a second visit. The aim was to improve a child’s symptom-free days over the 6-month period. In this case, a higher number is better, indicating that they had *more* symptom-free days.

Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

Principles of Analytic Graphics

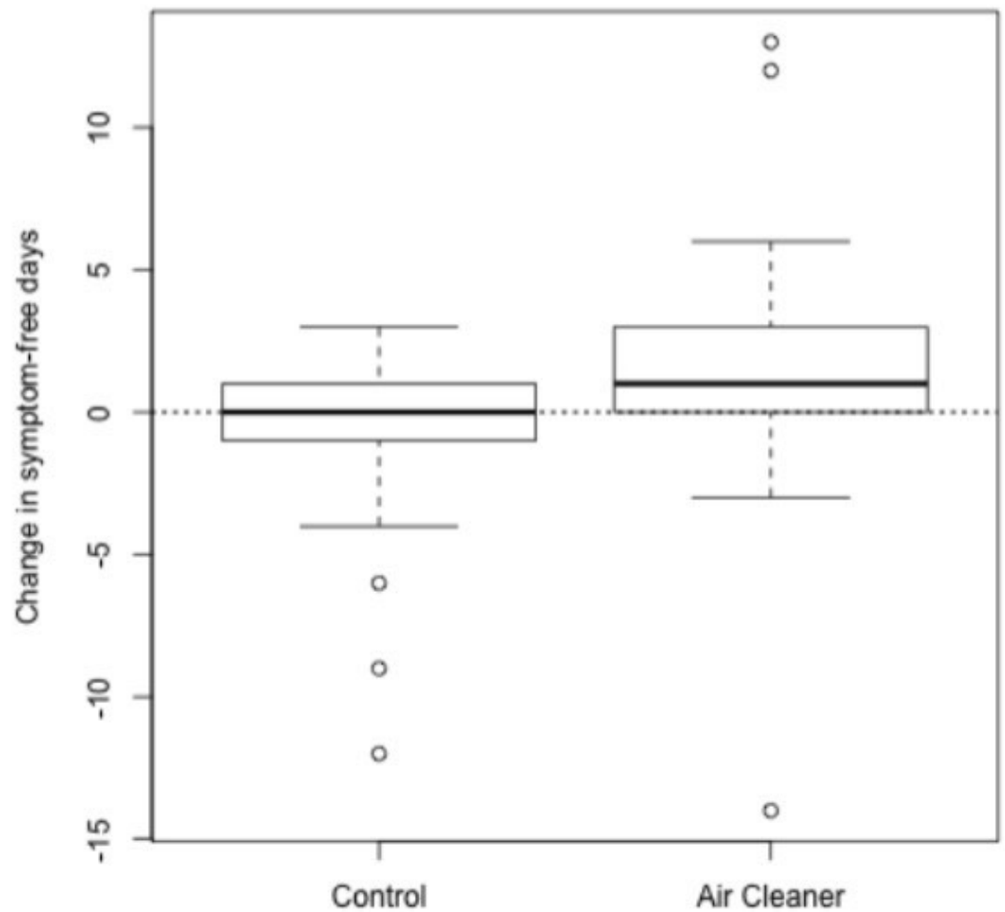
- **Principle 1: Show comparisons**
  - **Evidence for a hypothesis is always relatives another competing hypothesis**
  - **Always ask “Compared to What?"**



Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

Principles of Analytic Graphics

- **Principle 1: Show comparisons**
  - **Evidence for a hypothesis is always relatives another competing hypothesis**
  - **Always ask “Compared to What?"**



Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

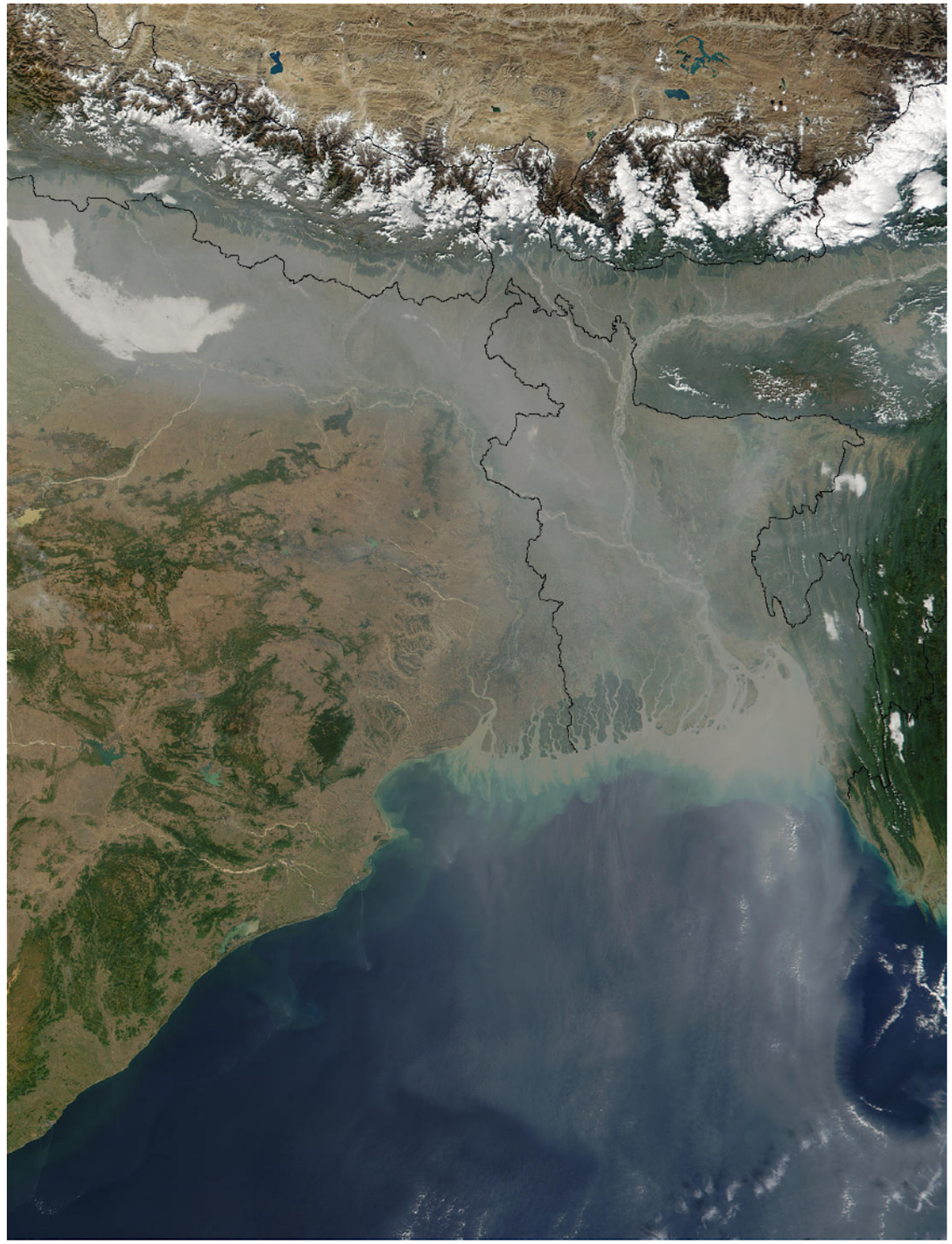
Principles of Analytic Graphics

- **Principle 1: Show comparisons**
  - **Evidence for a hypothesis is always relatives another competing hypothesis**
  - **Always ask “Compared to What?”**
- **Principle 2: Show causality, mechanism, explanation, systematic structure**
  - **What is your causal framework for thinking about a question?**

Principles of Analytic Graphics

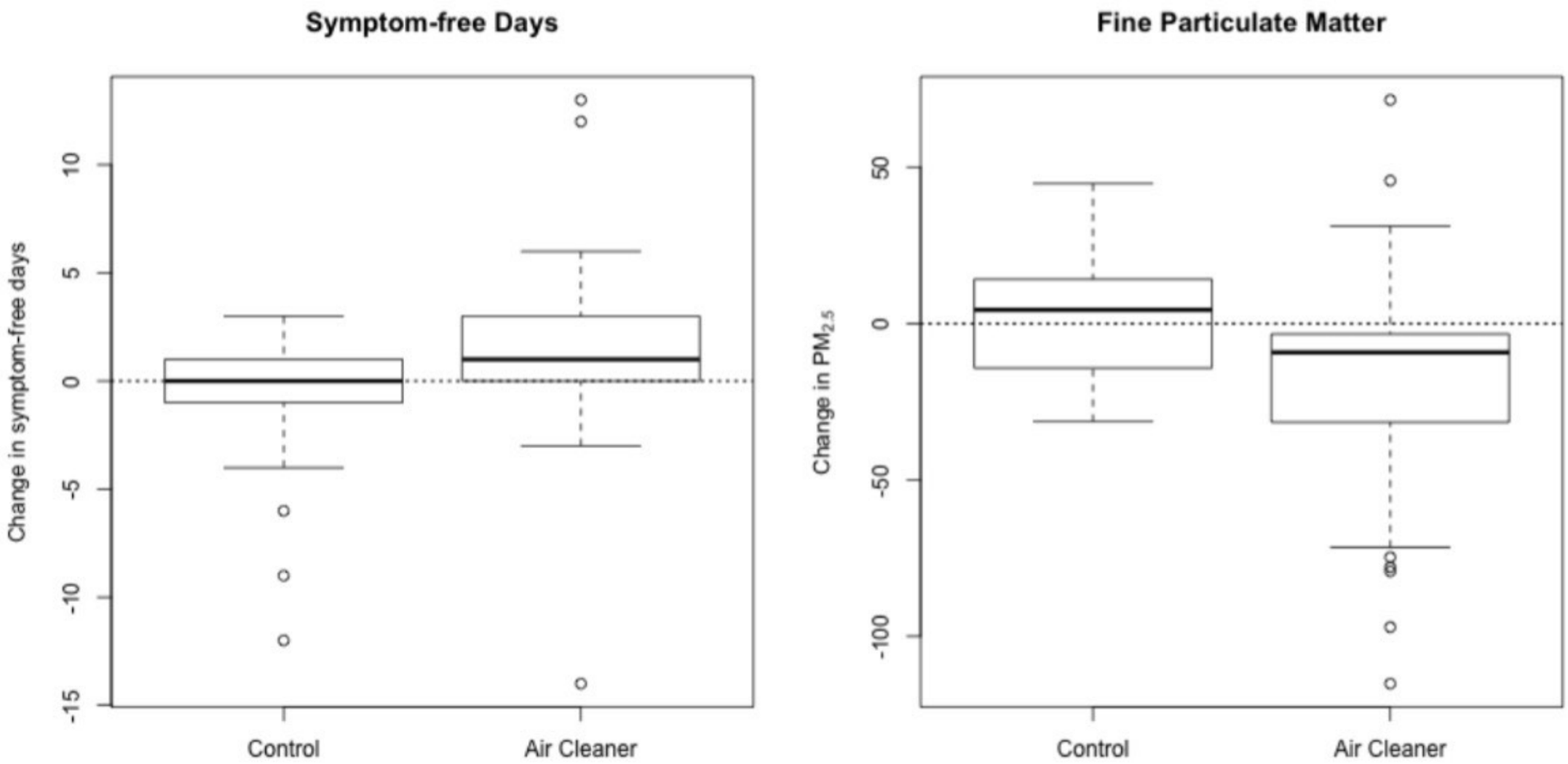
Air quality standards in the U.S. concerns the long-term average level of fine particle pollution, also referred to as PM2.5

The standard says that the “annual mean, averaged over 3 years” cannot exceed 12 micrograms per cubic meter.



Principles of Analytic Graphics

- **Principle 2: Show causality, mechanism, explanation, systematic structure**
  - **What is your causal framework for thinking about a question?**



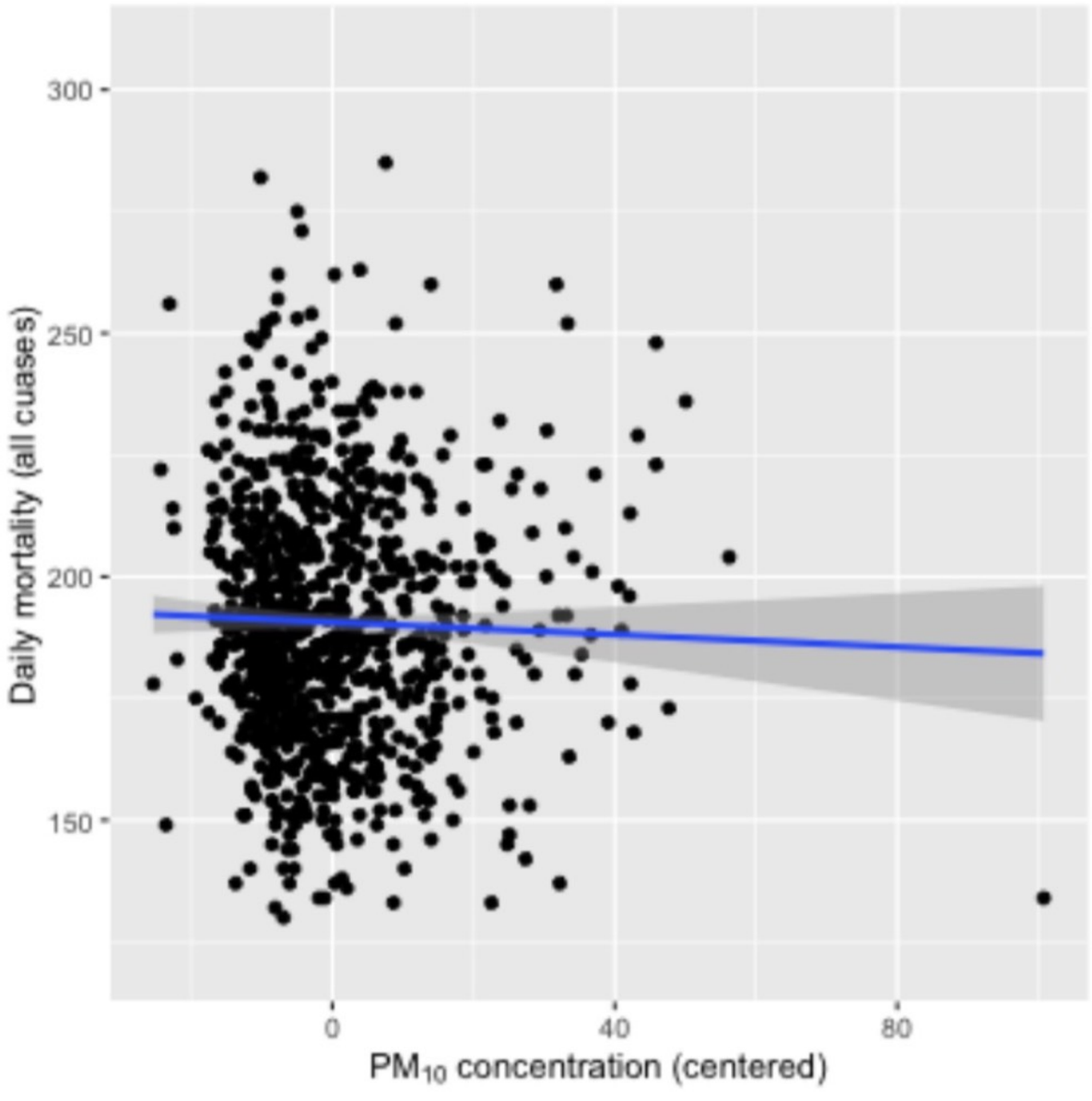
Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

Principles of Analytic Graphics

- **Principle 1: Show comparisons**
- **Principle 2: Show causality, mechanism, explanation, systematic structure**
- **Principle 3: Show multivariate data**
  - **Multivariate = more than 2 variables**
  - **The real world is multivariate**
  - **Need to “escape flatland"**

Principles of Analytic Graphics

- **Principle 3: Show multivariate data**



PM10 and mortality in New York City

it **seems** that there is a **slight negative relationship between the two variables**.

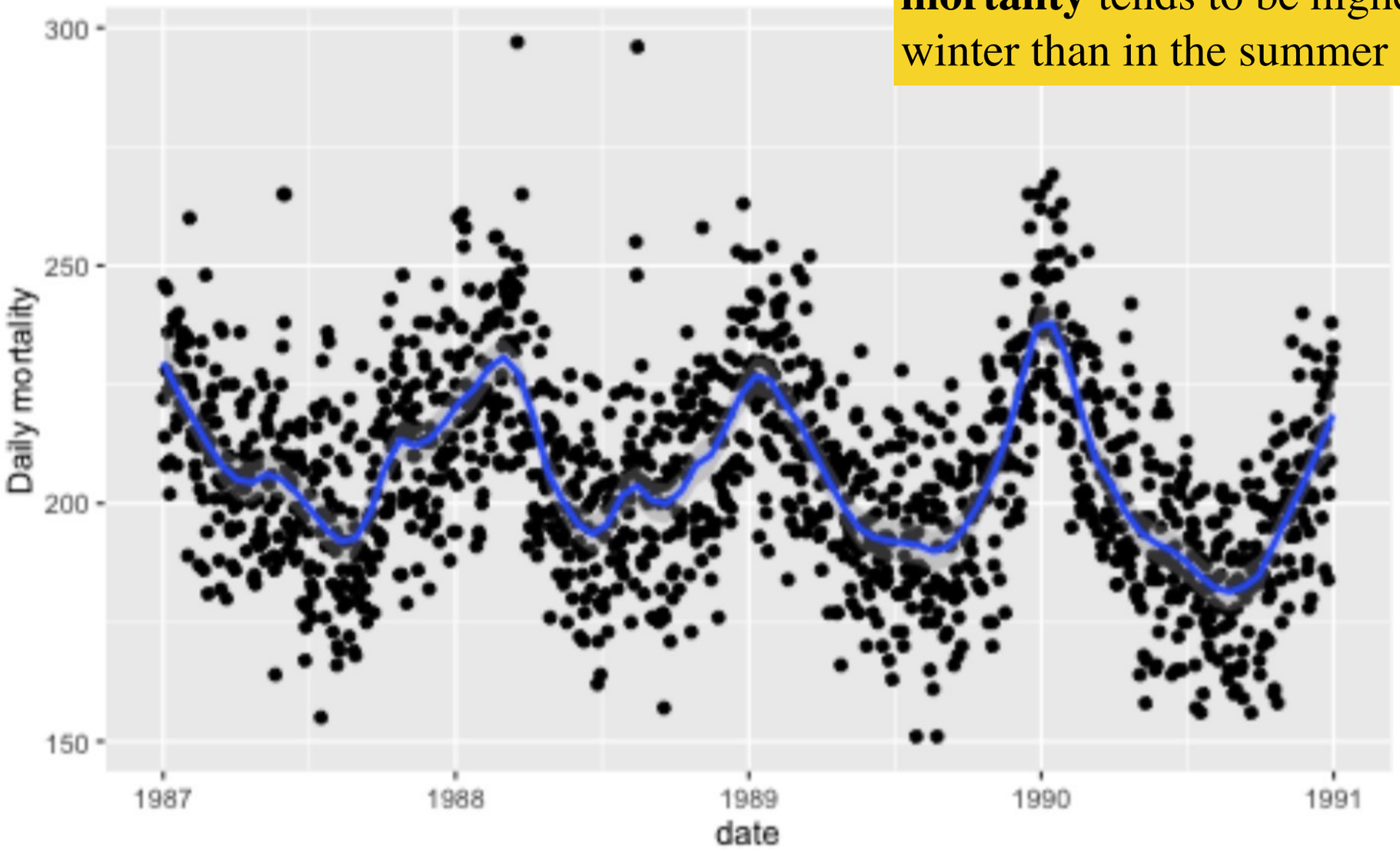
That is, higher daily average levels of PM10 appear to be associated with lower levels of mortality (fewer deaths per day).

Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

Principles of Analytic Graphics

- **Principle 3: Show multivariate data**

mortality tends to be higher in the winter than in the summer



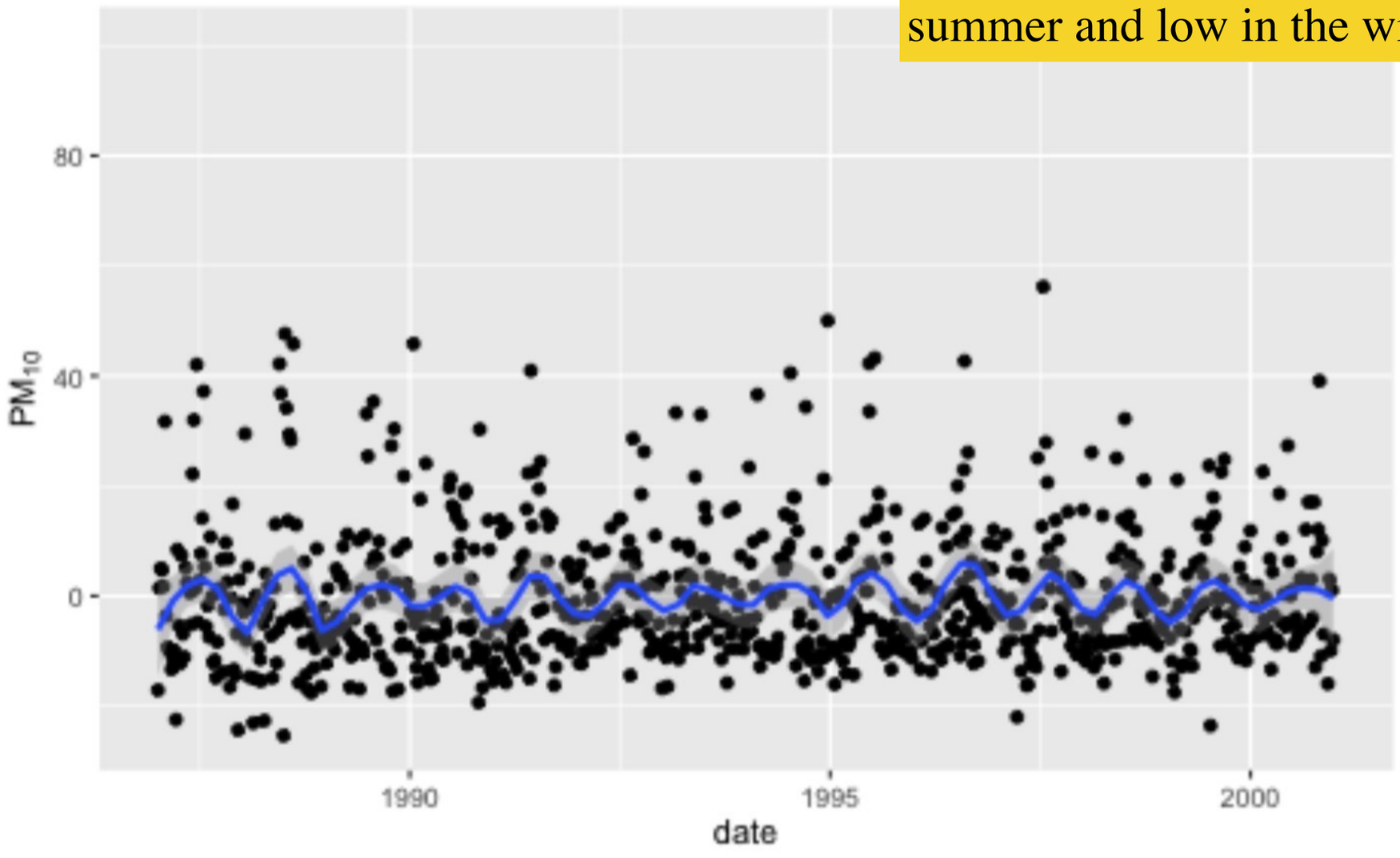
Daily mortality in New York City

Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

Principles of Analytic Graphics

- **Principle 3: Show multivariate data**

PM10 levels tend to be high in the summer and low in the winter.



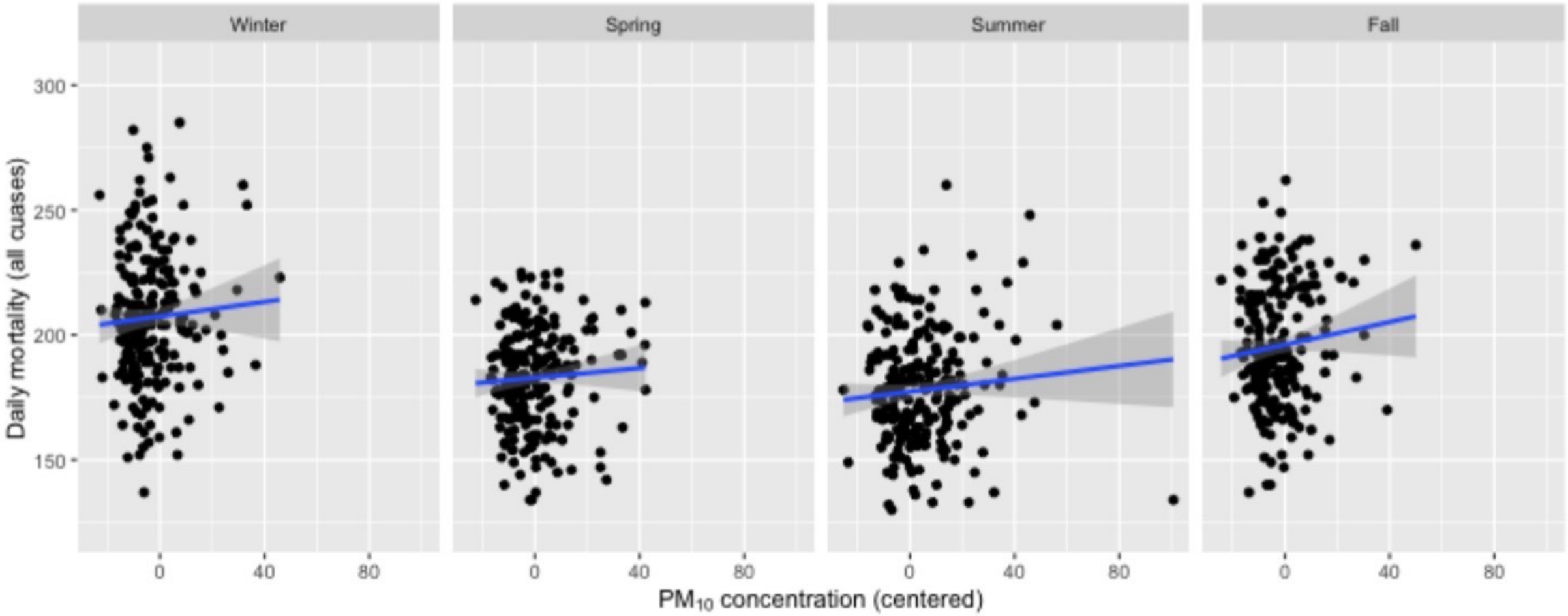
Daily PM10 in New York City

Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

Principles of Analytic Graphics

- **Principle 3: Show multivariate data**

There is a **slight positive** relationship between the two variables in each season



PM10 and mortality in New York City by season

Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

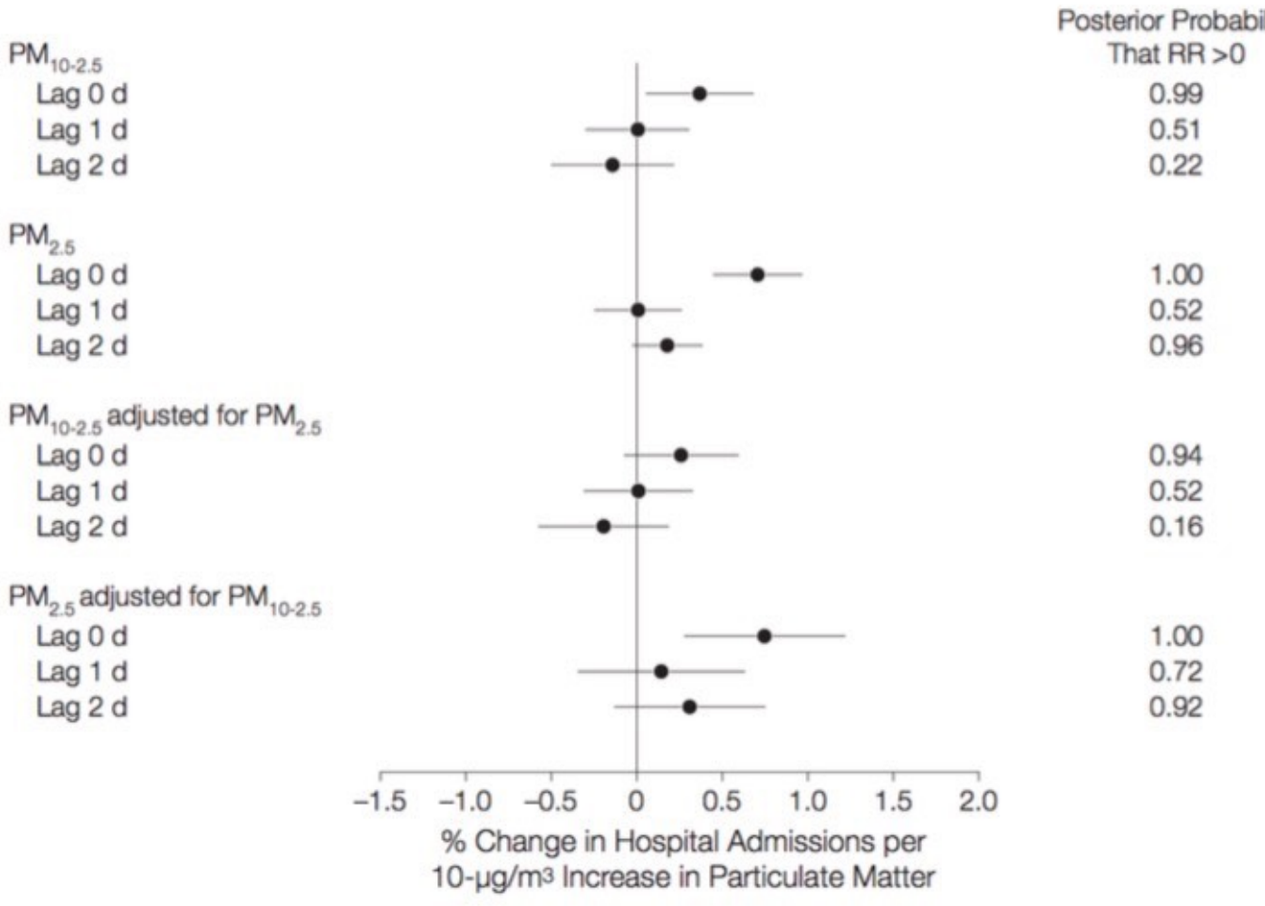
Principles of Analytic Graphics

- **Principle 1: Show comparisons**
- **Principle 2: Show causality, mechanism, explanation, systematic structure**
- **Principle 3: Show multivariate data**
- **Principle 4: Integration of evidence**
  - **Completely integrate words, numbers, images, diagrams**
  - **Data graphics should make use of many modes of data presentation**
  - **Don’t let the tool drive the analysis**

Principles of Analytic Graphics

- **Principle 4: Integration of evidence**

Figure 2. Percentage Change in Emergency Hospital Admissions Rate for Cardiovascular Diseases per a 10-μg/m³ Increase in Particulate Matter



Estimates are on average across 108 counties. $PM_{2.5}$ indicates particulate matter is 2.5 μm or less in aerodynamic diameter; PM_{10}, particulate matter is 10 μm or less in aerodynamic diameter; $PM_{10-2.5}$, particulate matter is greater than 2.5 μm and 10 μm or less in aerodynamic diameter; RR, relative risk. Error bars indicate 95% posterior intervals.

Ideally, a plot would have all of the necessary descriptions attached to it.

You might think that this level of documentation should be reserved for "final" plots as opposed to exploratory ones, but it's good to get in the habit of documenting your evidence sooner rather than later.

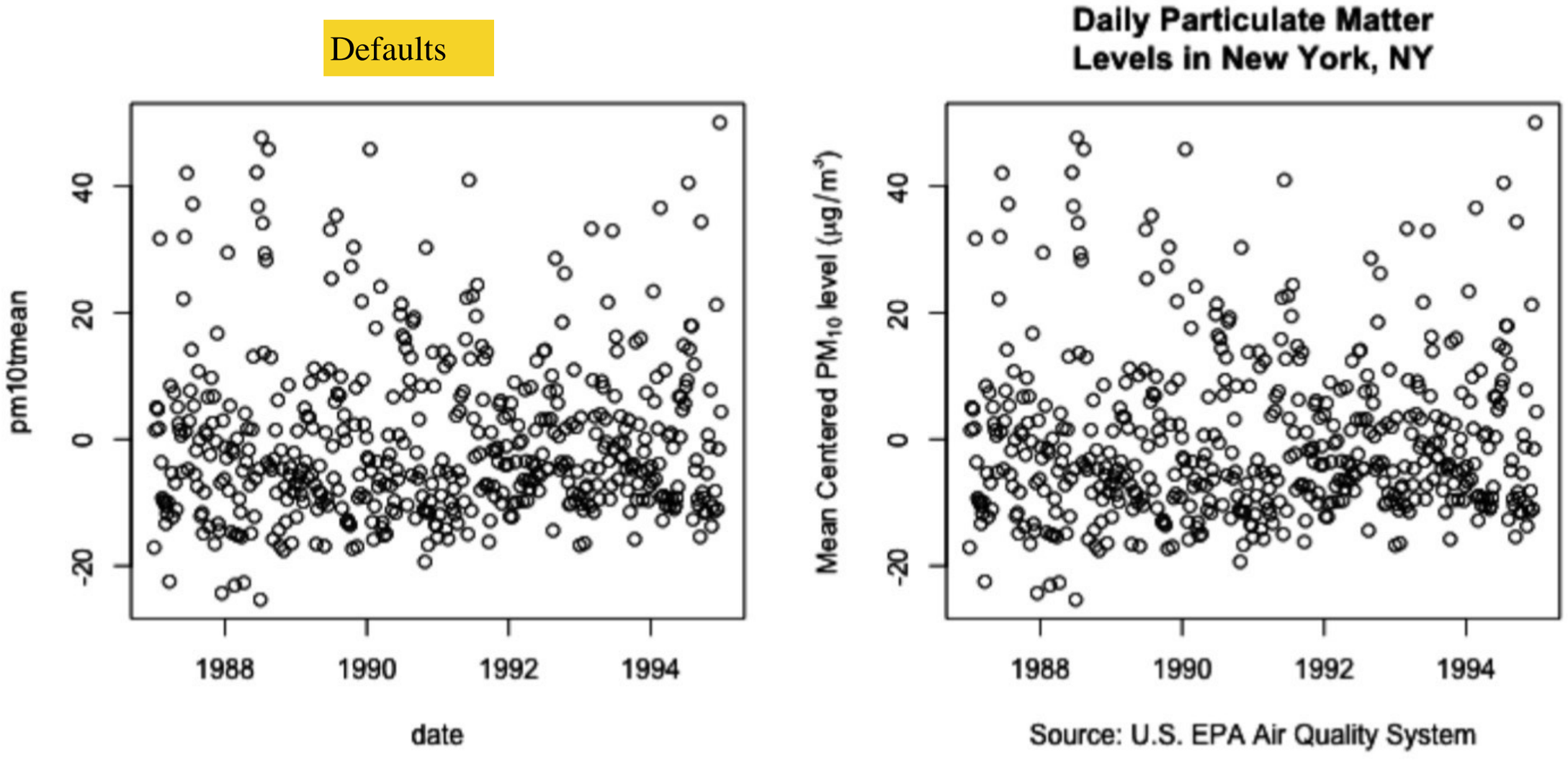
Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

Principles of Analytic Graphics

- **Principle 1: Show comparisons**
- **Principle 2: Show causality, mechanism, explanation, systematic structure**
- **Principle 3: Show multivariate data**
- **Principle 4: Integration of evidence**
- **Principle 5: Describe and document the evidence with appropriate labels, scales, sources**

Principles of Analytic Graphics

- **Principle 5: Describe and document the evidence with appropriate labels, scales, sources**



Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

Principles of Analytic Graphics

- **Principle 1: Show comparisons**
- **Principle 2: Show causality, mechanism, explanation, systematic structure**
- **Principle 3: Show multivariate data**
- **Principle 4: Integration of evidence**
- **Principle 5: Describe and document the evidence with appropriate labels, scales, sources**
- **Principle 6: Content is King**
  - **Analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content.**

Principles of Analytic Graphics

- **Principle 1: Show comparisons**
- **Principle 2: Show causality, mechanism, explanation, systematic structure**
- **Principle 3: Show multivariate data**
- **Principle 4: Integration of evidence**
- **Principle 5: Describe and document the evidence with appropriate labels, scales, sources**
- **Principle 6: Content is King**

Edward Tufte (2006). *Beautiful Evidence*, Graphics Press LLC. www.edwardtufte.com

Reference: Butz AM, *et al.*, *JAMA Pediatrics*, 2011.

Further Reading and Summary

Further Reading

- Exploratory Data Analysis with R, by Roger D. Peng
  - Chapters 5 - Exploratory Data Analysis Checklist
  - Chapter 6 - Principles of Analytic Graphics
- Optionally
  - Chapter 7 - Exploratory Graphs