What is Exploratory Data analysis?

Exploratory Data
Analysis with R

Roger D. Peng

Exploratory Data Analysis with R
Roger D. Peng

This book is for sale at http://leanpub.com/exdata
This version was published on 2015-11-12

"In the editing room"

▶ The process of making the "rough cut" for a data analysis.

▶ The goals：

1. identifying relationships between variables that are particularly interesting or unexpected;

2. checking to see if there is any evidence for or against a stated hypothesis;

3. checking for problems with the collected data, such as missing data or measurement error;

4. identifying certain areas where more data need to be collected

# Principles of Analytic Graphics

Edward Tufte <Beautiful Evidence>

Show comparisons

▶ The basis of all good scientific investigation

▶ Evidence for a hypothesis is always *relative* to another competing hypothesis.

▶ "the evidence favors hypothesis A"

▶ "the evidence favors hypothesis A versus hypothesis B"

▶ A good scientist is always asking "Compared to What?" when confronted with a scientific claim or statement.

▶ Data graphics should generally follow this same principle. You should always be comparing at least two things.

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.

Huang, F., & Kim, J. S. (2012). A Randomized Trial of Air Cleaners and a Health Coach to Improve Indoor Air Quality for Inner-City Children With Asthma and Secondhand Smoke Exposure. *Pediatrics*, *130*(Supplement 1), S33-S34.
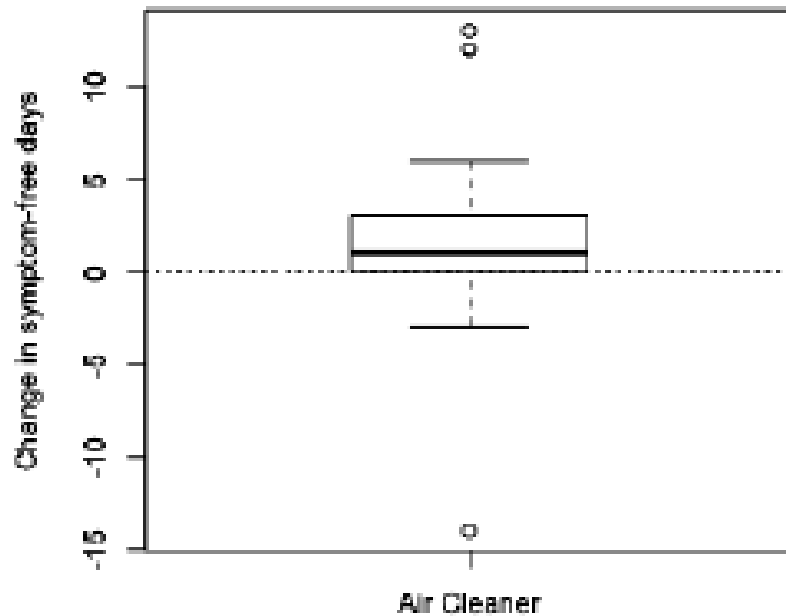
Change in symptom-free days with air cleaner
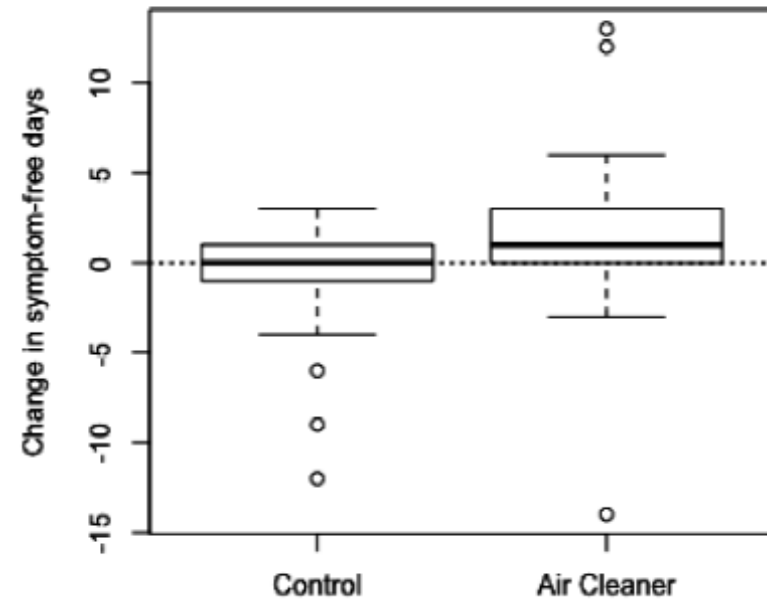
Change in symptom-free days by treatment group

Huang, F., & Kim, J. S. (2012). A Randomized Trial of Air Cleaners and a Health Coach to Improve Indoor Air Quality for Inner-City Children With Asthma and Secondhand Smoke Exposure. *Pediatrics, 130*(Supplement 1), S33-S34.

Show causality, mechanism, explanation, systematic structure

▶ If possible, it's always useful to show your causal framework for thinking about a question.

▶ Generally, it's 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.

"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.

A decrease in airborne particles?

Symptom-free Days

Fine Particulate Matter
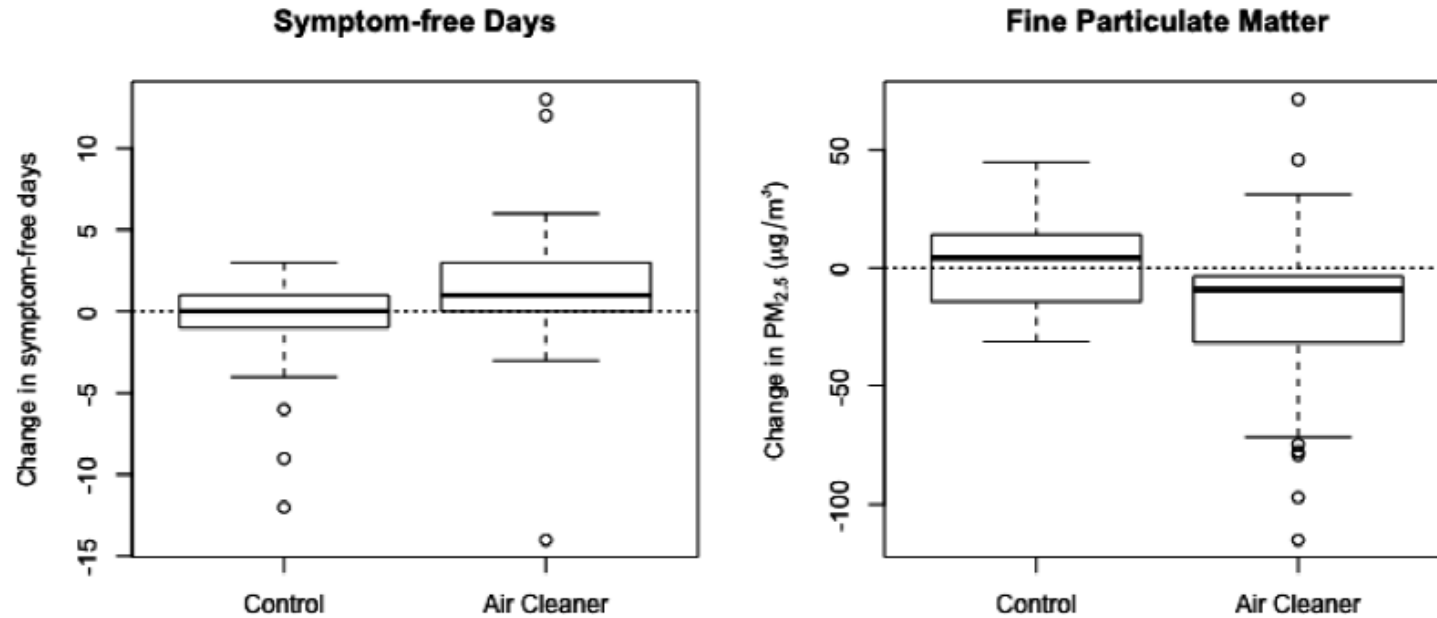
Change in symptom-free days and change in PM2.5 levels in-home

📢 However, it is not conclusive proof of this idea because there may be other unmeasured confounding factors that can lower levels of PM2.5 and improve symptom-free days.

Show multivariate data

▶ The real world is multivariate.

▶ For anything that you might study, there are usually many attributes that you can measure.

▶ The point is that data graphics should attempt to show this information as much as possible, rather than reduce things down to one or two features that we can plot on a page.

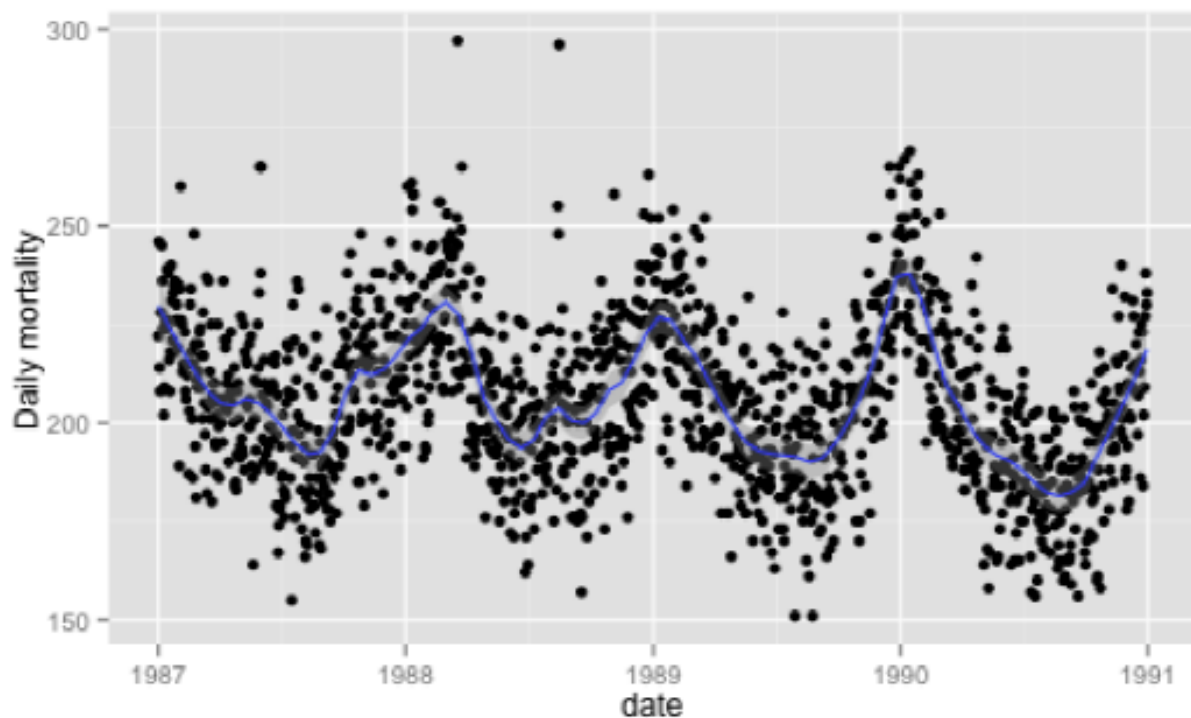▶ There are a variety of ways that you can show multivariate data.

Daily airborne particulate matter ("PM10") and mortality

▶ Daily PM10 in New York City and mortality from 1987 to 2000

▶ The PM10 data come from the U.S. Environmental Protection Agency

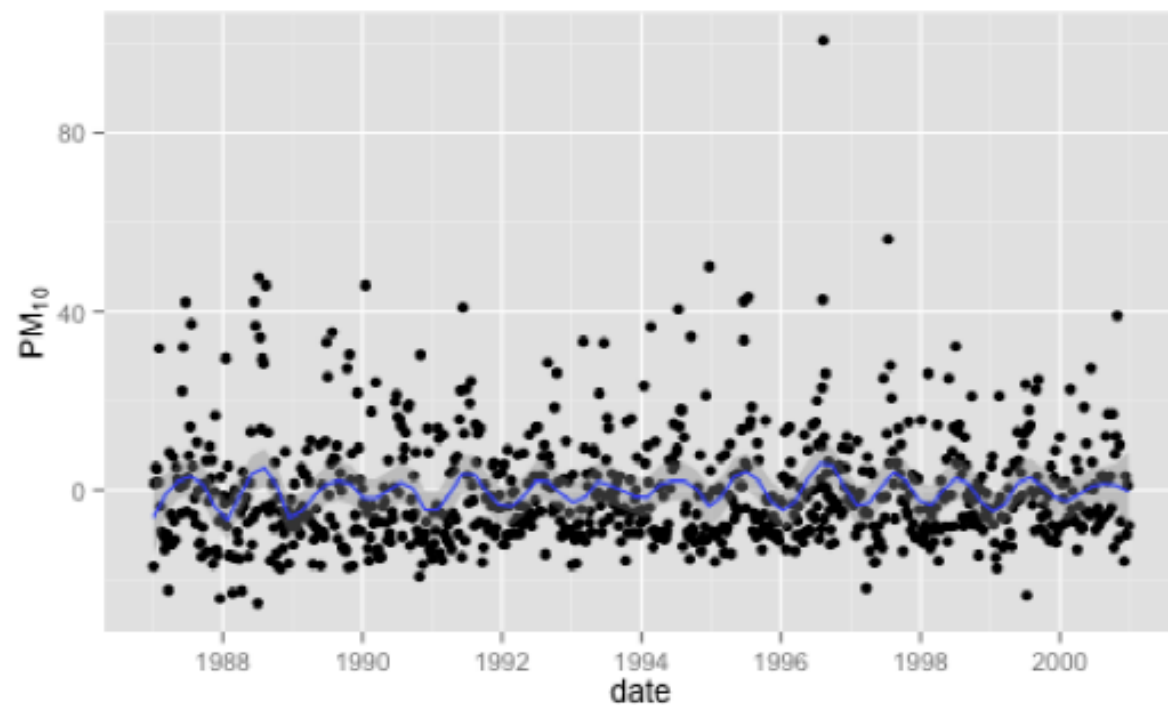▶ The mortality data come from the U.S. National Center for Health Statistics.
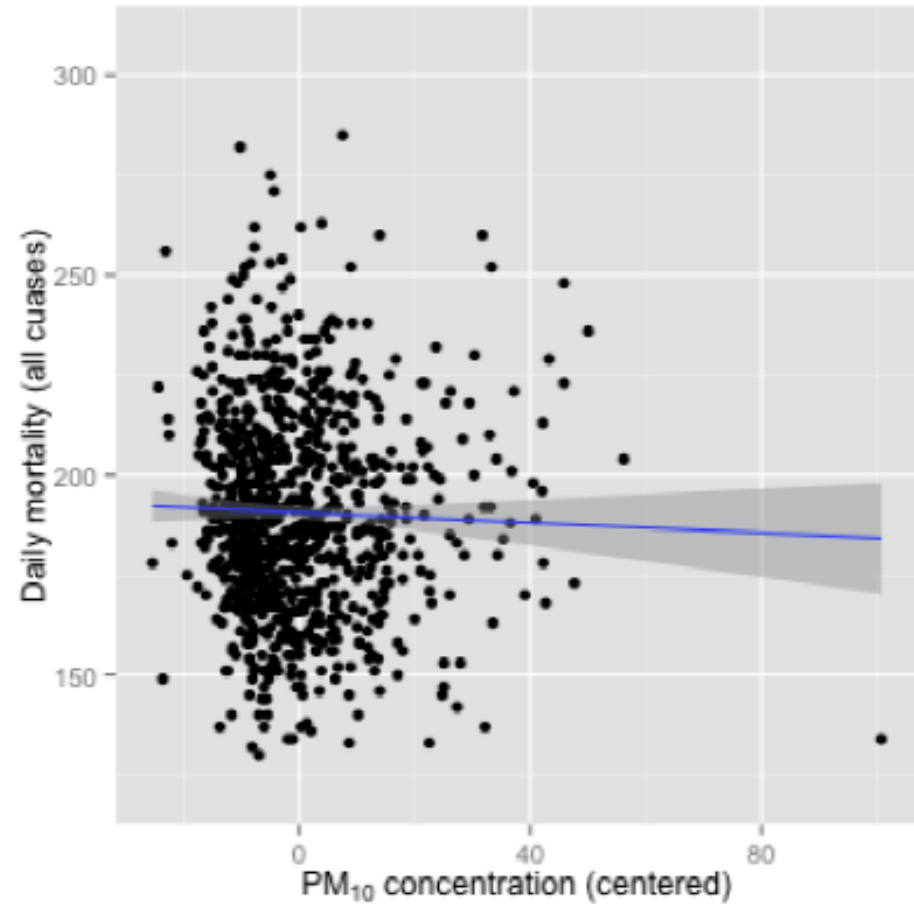
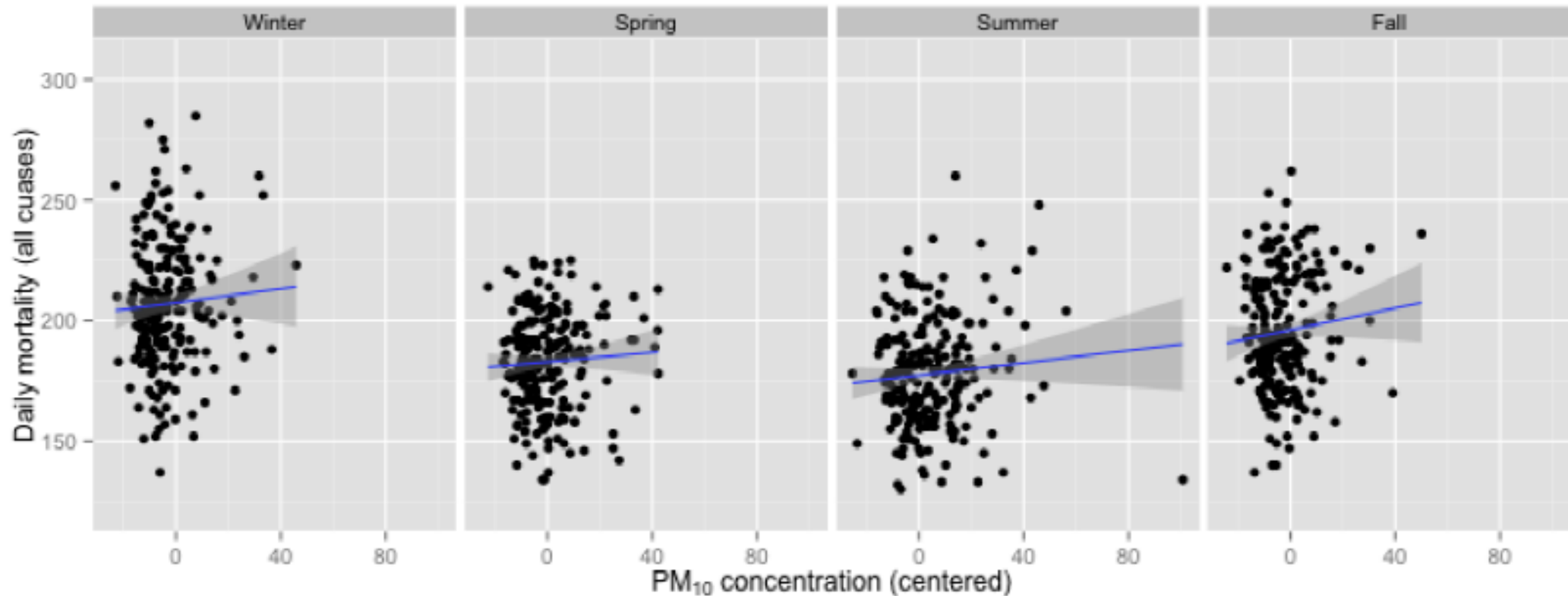**PM10 and mortality in New York City**

Season?

Daily mortality in New York City

Daily PM10 in New York City

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

PM10 and mortality in New York City by season

Describe and document the evidence

▶ Data graphics should be appropriately documented with labels, scales, and sources.

▶ A general rule for me is that a data graphic should tell a complete story all by itself.

Labelling and annotation of data graphics
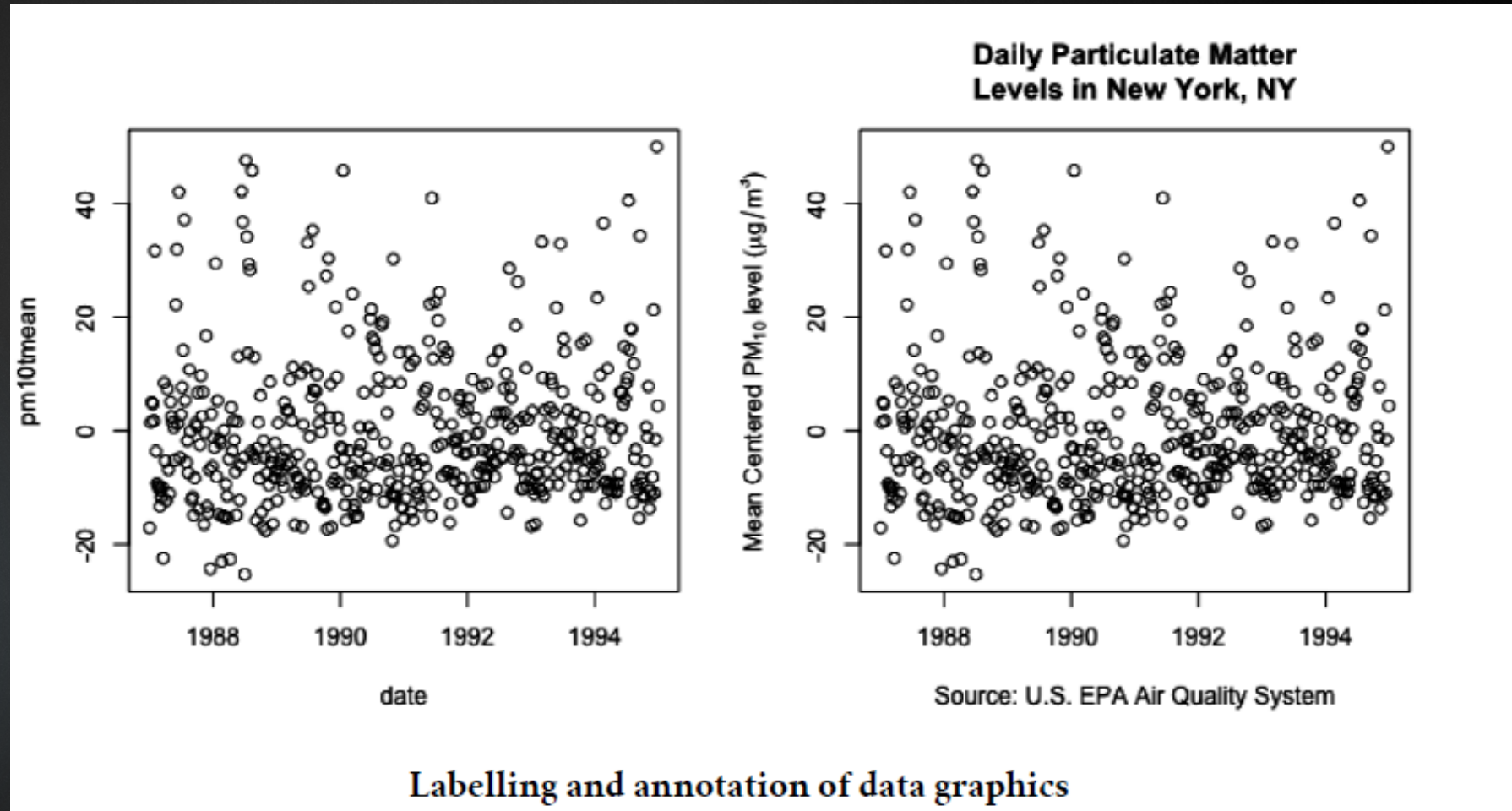
Content, content, content!

► Analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content.

► This includes the question being asked and the evidence presented in favor of certain hypotheses.

► 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.

No amount of visualization magic or bells and whistles can make poor data, or more importantly, a poorly formed question, shine with clarity.