Today: What Is Data? Graphics Principles Friday: Introduction to R and Reproducibility Monday: No class (Labor Day)

Sam Ventura
36-315

Department of Statistics
Carnegie Mellon University

August 31, 2016

What Is Data?

How Do We Describe Data?

Two measurements used to describe datasets:

Data is usually in matrix form:

Types of Data

Categorical:

Continuous:

Graphics and Their Goals (from Tufte)

Graphics: visually display measured quantities by combining points, lines, coordinate system, numbers, symbols, words, shading, color

Goals: show data!

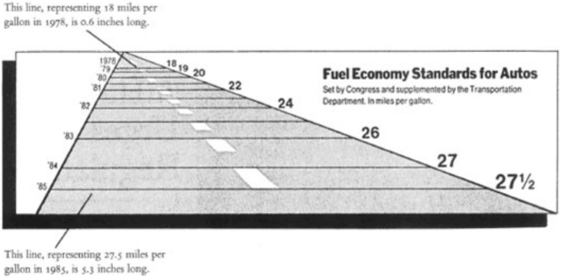
- induce viewer to think about substance, not graphical methodology
- avoid **distorting** the data
- present numbers in small space
- make large, complicated datasets more coherent
- encourage comparison of different pieces of data
- reveal data at several levels of detail
- describe, explore, tabulate, or decorate
- be closely integrated with statistical/verbal descriptions of dataset

Graphs that do not meet these goals are not successful

Graphs leading viewers to make misleading conclusions should be avoided

Distortion = “graph doesn’t match the data”

Visual representation of data is inconsistent with numerical representation



Tufte suggests optimizing the Lie Factor:

“Decorating” and Data-Ink

Graphics should not draw the viewer’s attention away from the data.
Extras get in the way.

Note: Decoration does not refer to appropriate graph labeling.
Labels should always be clear, detailed, and thorough.
Label key parts of the data. Add text explanations if necessary.

Data Ink should primarily present information about the data:
the non-erasable, non-redundant core of a graphic

Tufte suggests using the *data-ink ratio*:

“Decorating” / Data-Ink

Two ways to increase the proportion of data-ink:

Remove non-data-ink:

Remove redundant data-ink: