VISUALIZATION

Jeff Goldsmith, PhD

Department of Biostatistics

Exploratory data analysis

- Exploratory analysis is a loosely-defined process
- Roughly, the stuff between loading data and formal analysis is "exploratory"
- This includes
  - Visualization
  - Checks for data completeness and reliability
  - Quantification of centrality and variability
  - Initial evaluation of hypotheses
  - Hypothesis generation

- Current emphasis is visualization

A picture is worth 1000 words

- Looking at data is critical
  - True for you as an analyst
  - True for you as a communicator

- You should make dozens, maybe even hundreds, of graphics for each dataset
  - Most of these are for your eyes only
  - A small subset are for others

A _{good} picture is worth 1000 words
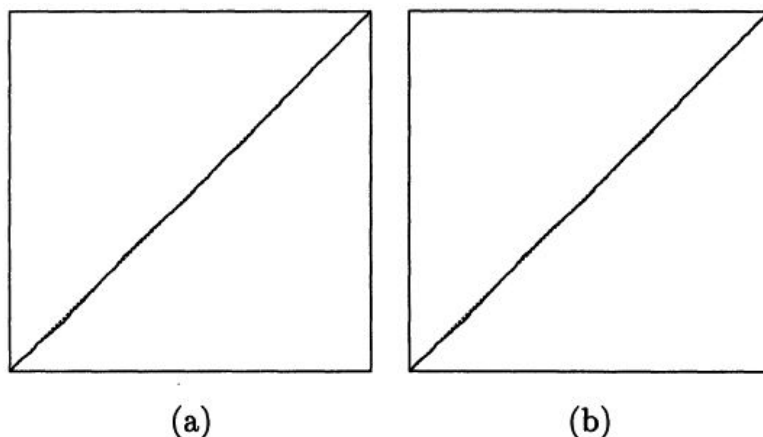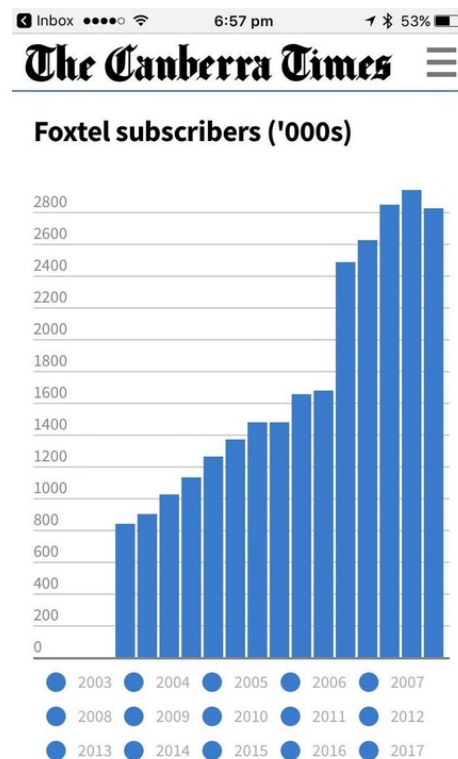
- Bad graphics are worth only a few words

Figure 1. SRQ Plots of T_i/T_n (Vertical Axes) Against i/n (Horizontal Axes) for the Gibbs Sampler (a) and an Alternating Gibbs/Independence Sampler (b) for the Pump Failure Data Based on Runs of Length 5,000. Lines through the origin with unit slope are shown dashed; axis ranges are from 0 to 1 for all axes.
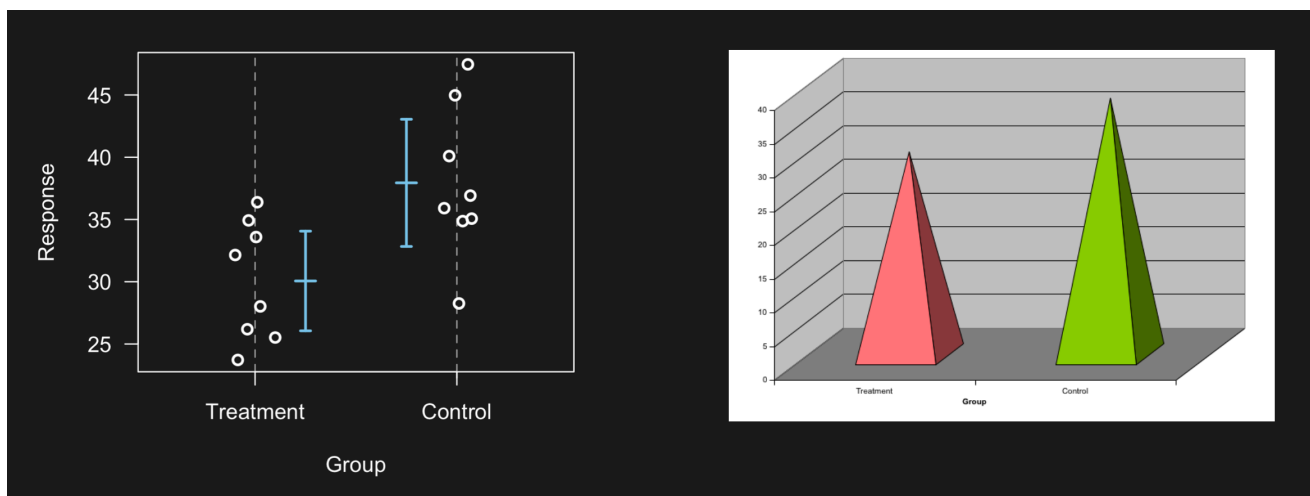
For more bad graphics, see Karl Broman's "Top Ten Worst Graphics"
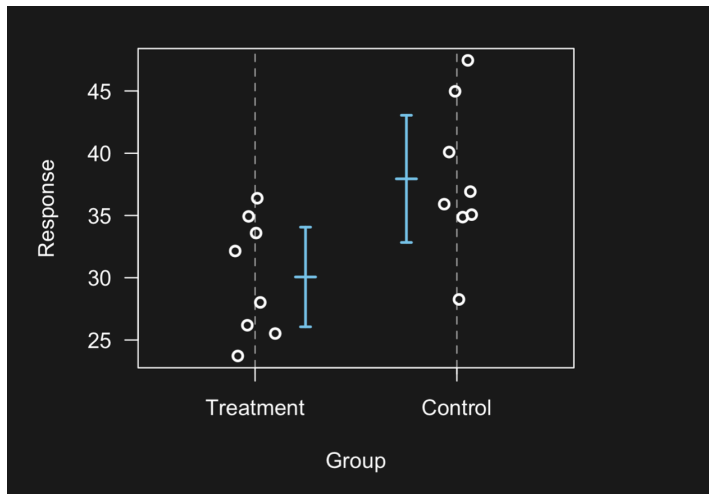
What makes a "good" picture?

- Show as much of the data as possible
- Avoid superfluous frills (e.g. 3D ...)
- Facilitate comparisons
  - Put groups in a sensible order
  - Use common axes
  - Use color to highlight groups
  - No pie charts

"Creating effective tables and figures" – talk by Karl Broman

What makes a "good" picture?

- Show as much of the data as possib
- Avoid superfluous frills (e.g. 3D ...)
- Facilitate comparisons
  - Put groups in a sensible order
  - Use common axes
  - Use color to highlight groups
  - No pie charts

.@earino: There is a time and a place for pie charts

That is when you are talking about pie

Otherwise, never

#PlotCon

12:58 PM - 18 Nov 2016

"Creating effective tables and figures" – talk by Karl Broman

What makes a "good" picture?

- "Good" figures aren't necessarily "publication quality" pictures
  - Most figures are for you, and even these should be good
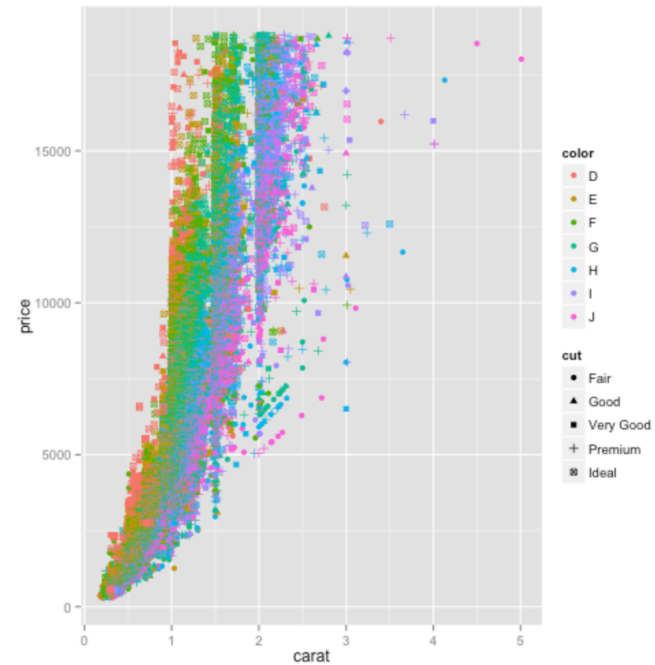 - Graphics for others require more fiddly detailing than is necessary for graphics for you

Why ggplot?

- Makes good graphics with relative ease
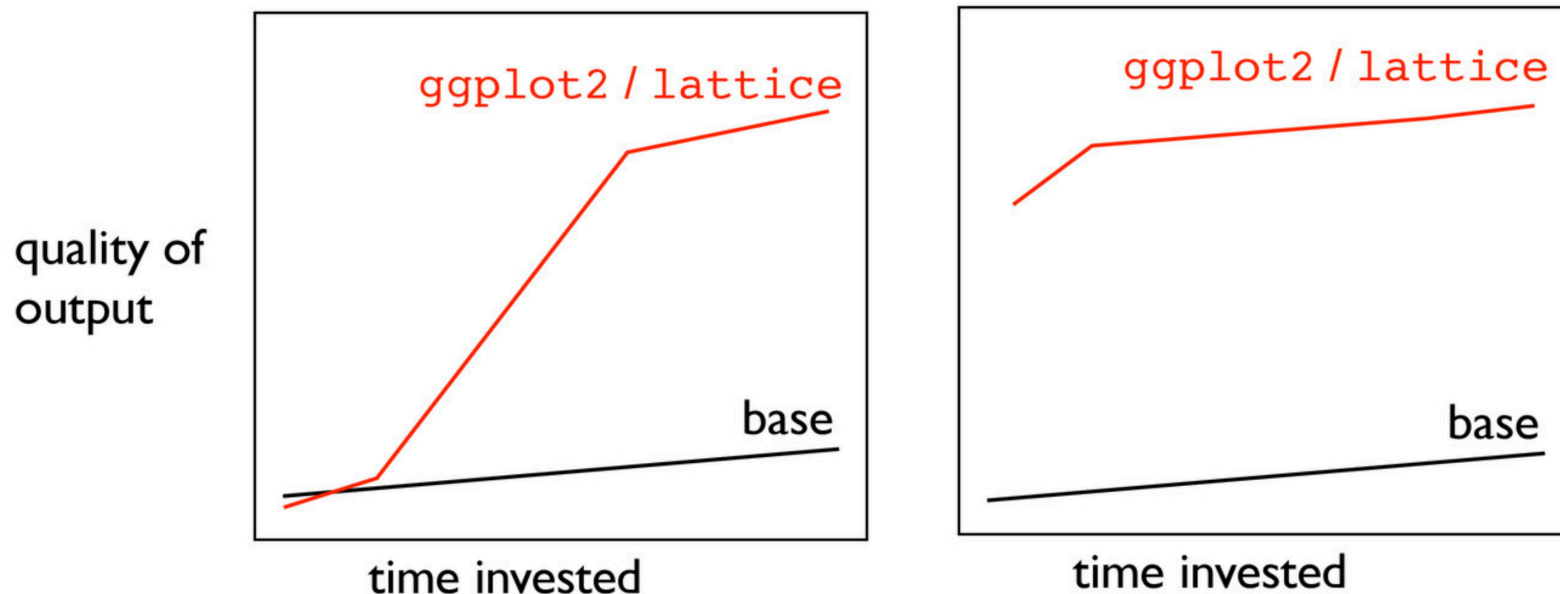 - "Relative" here is compared to base R graphics
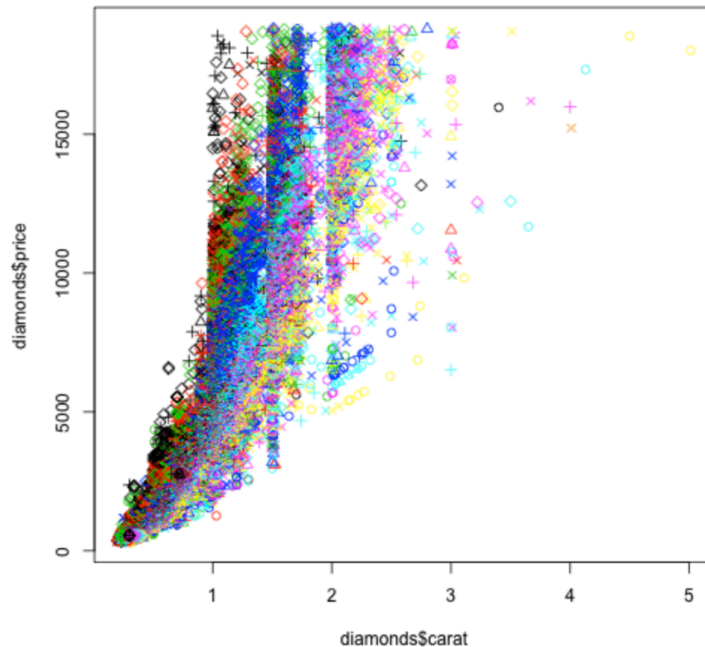
VS

"Don't teach built-in plotting to beginners (teach ggplot2)" – blog post by David Robinson

Why ggplot?

- Cohesiveness shortens the learning curve
  - Same principles underlie all graphic types

Learning `ggplot`

- Lots of materials
- google is your friend
  - Start searches with "ggplot"
  - StackOverflow has lots of questions and useful answers
  - Don't worry about googling stuff you "should know"

Using `ggplot`

- Based around the "tidy data" framework
- Trouble making a plot is often trouble with data tidiness in disguise
  – Think about how your data organization affects your ability to visualize
  – Factors can help with ordering
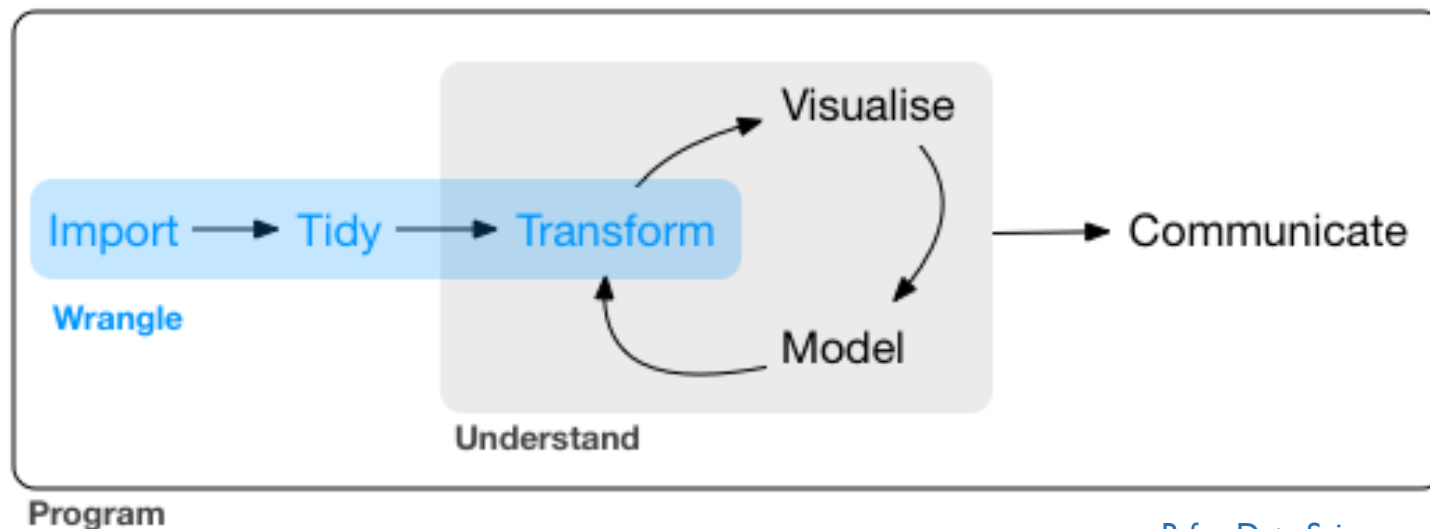
R for Data Science

Using ggplot

- Basic graph components
  - data
  - aesthetic mappings
  - geoms

- Advanced graph components
  - facets
  - scales
  - statistics

- A graph is built by combining these components
- Components are consistent across graph types
  - Scatterplots, bar graphs, density plots, ridge plots …