



UNIVERSITY OF SAN FRANCISCO
CHANGE THE WORLD FROM HERE

Multivariate Data Visualization

Outline

- Bubble Chart
- Heatmap
- Scatterplot Matrix
- Small Multiples
- Parallel Coordinates
- Spider Plots
- *Encoding Data*
- *Examples from Books*
- *Examples from D3*
- *Examples from R*
- *Examples from Web*
- *Discussion*

DATA TYPES

Data Types

- **Numerical**

- Continuous (-11.2, -1.3, 0.4, 4.8, 14.9, ...)
- Discrete (-2, -1, 0, 1, 2, ...)

- **Categorical**

- Ordered (January, February, March, April, ...)
- Unordered (Red, Blue, Green, Purple, ...)

Data Types

- **Special/Structured**

- Time (1998-01-27 12:00, 1999-04-13 22:10, ...)
- GPS (37.77679 latitude, -122.45117 longitude, ...)
- Social Security Numbers (555-55-5555, ...)

- **Unstructured/Semi-Structured**

- Free-Form Text (Twitter Feed, Screenplay, ...)

BUBBLE CHART

Encoding Data

- **Horizontal position**
 - Continuous data
- **Vertical position**
 - Continuous data
- **Circle area**
 - Numerical data
- **Circle color**
 - Numerical or categorical

"Visualize This: The FlowingData Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.

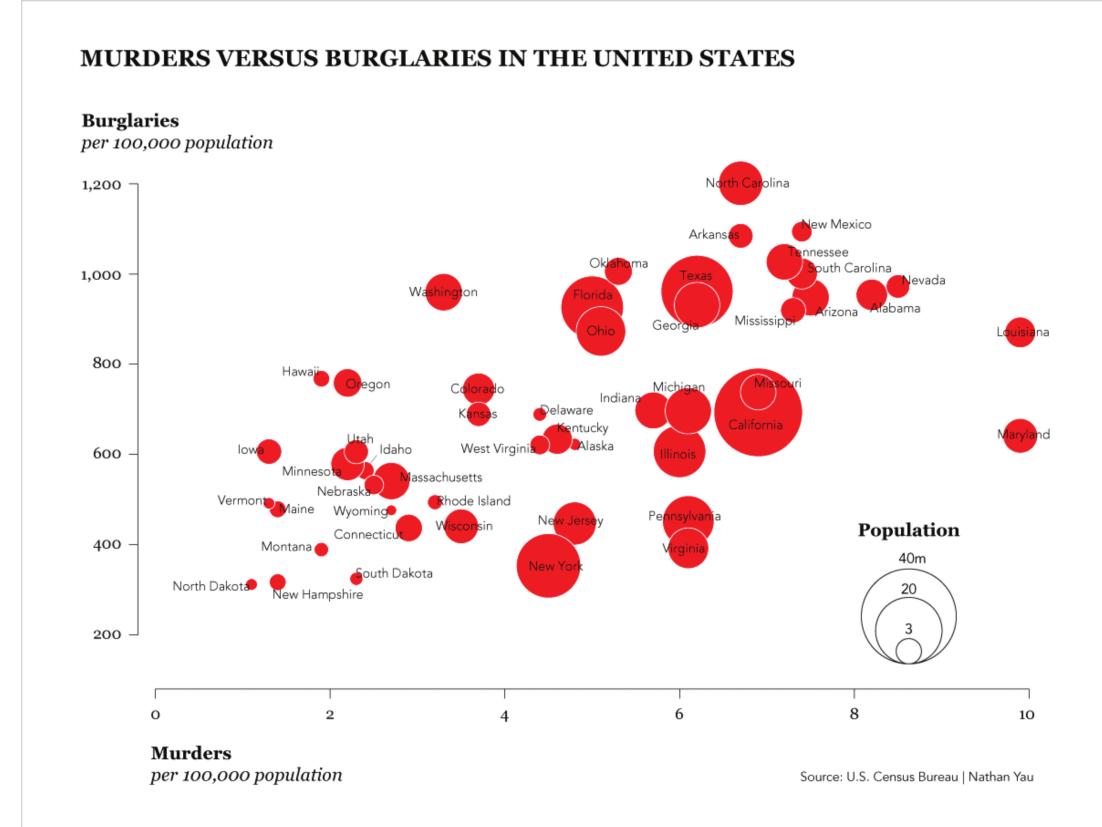


FIGURE 6-15 Bubble plot showing crime in the United States

"Visualize This: The FlowingData Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.

The Wealth & Health of Nations

March 13, 2012

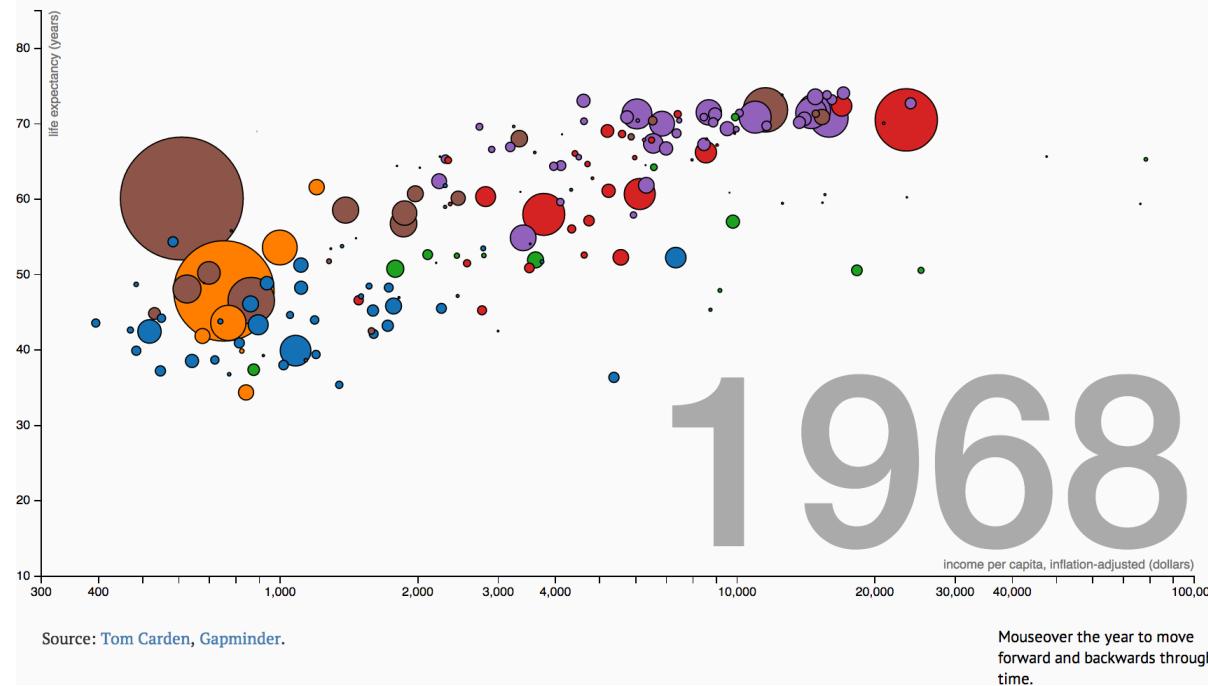
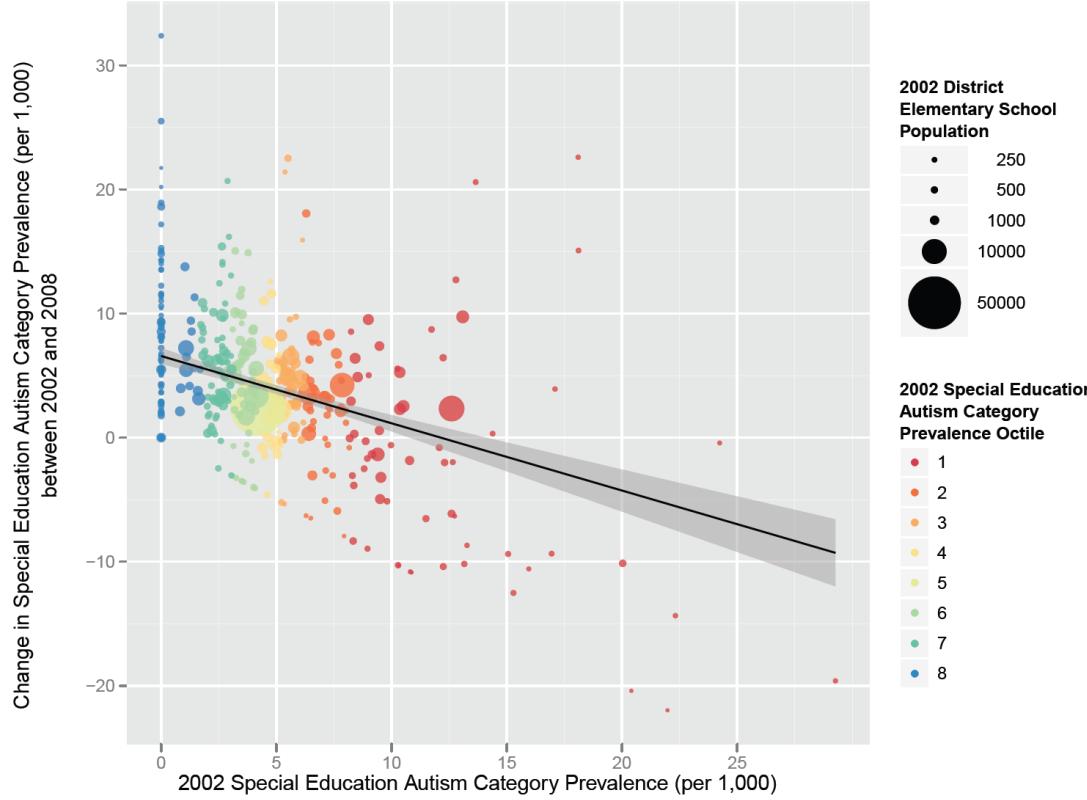
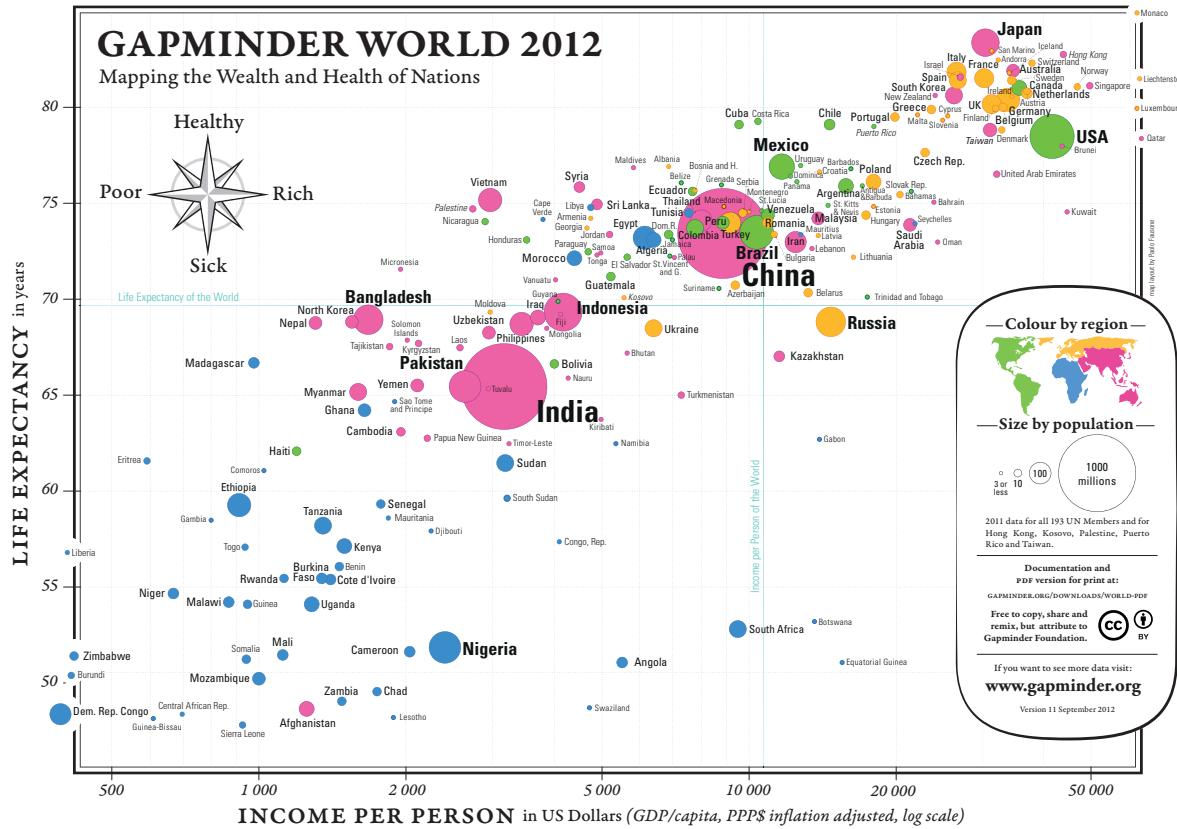


Figure 3. Change in Special Education Autism Category Prevalence between 2002 and 2008 vs Baseline (2002) Prevalence, Wisconsin Elementary School Districts (with weighted linear best-fit line and 95% confidence band)



<http://www.matthewmaenner.com/blog/?p=150>



<http://www.gapminder.org/downloads/world-pdf/>

Discussion

- Able to encode four dimensions of data
 - Ideal if one dimension is categorical (color)
- Rough comparison possible
 - Beware comparing circle areas
- Obscuring data may be an issue
 - Large circles should be behind smaller ones
 - Issues increases with density

HEATMAP

Encode Data

- **Horizontal position**
 - Column from dataset
- **Vertical position**
 - Row from dataset
- **Box color**
 - Value from row, column in dataset
 - Numerical or categorical

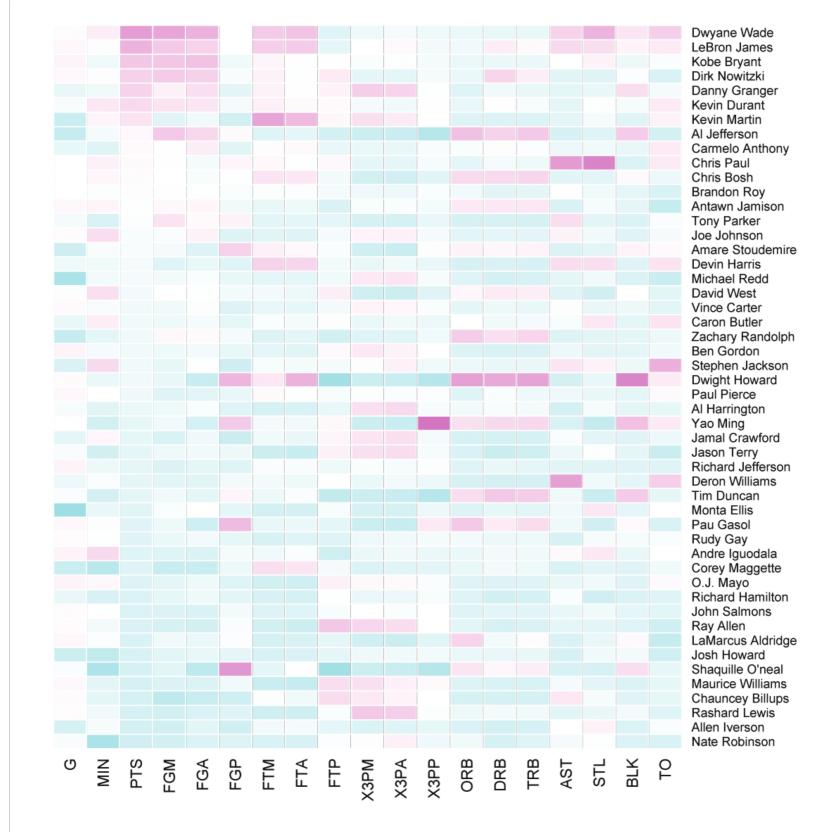
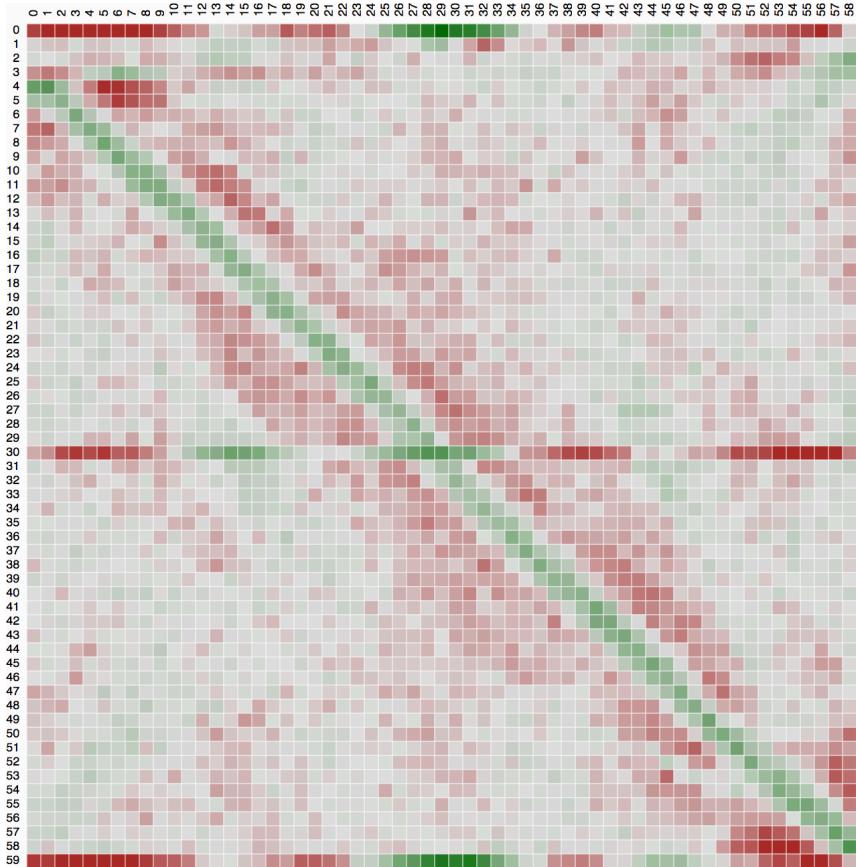
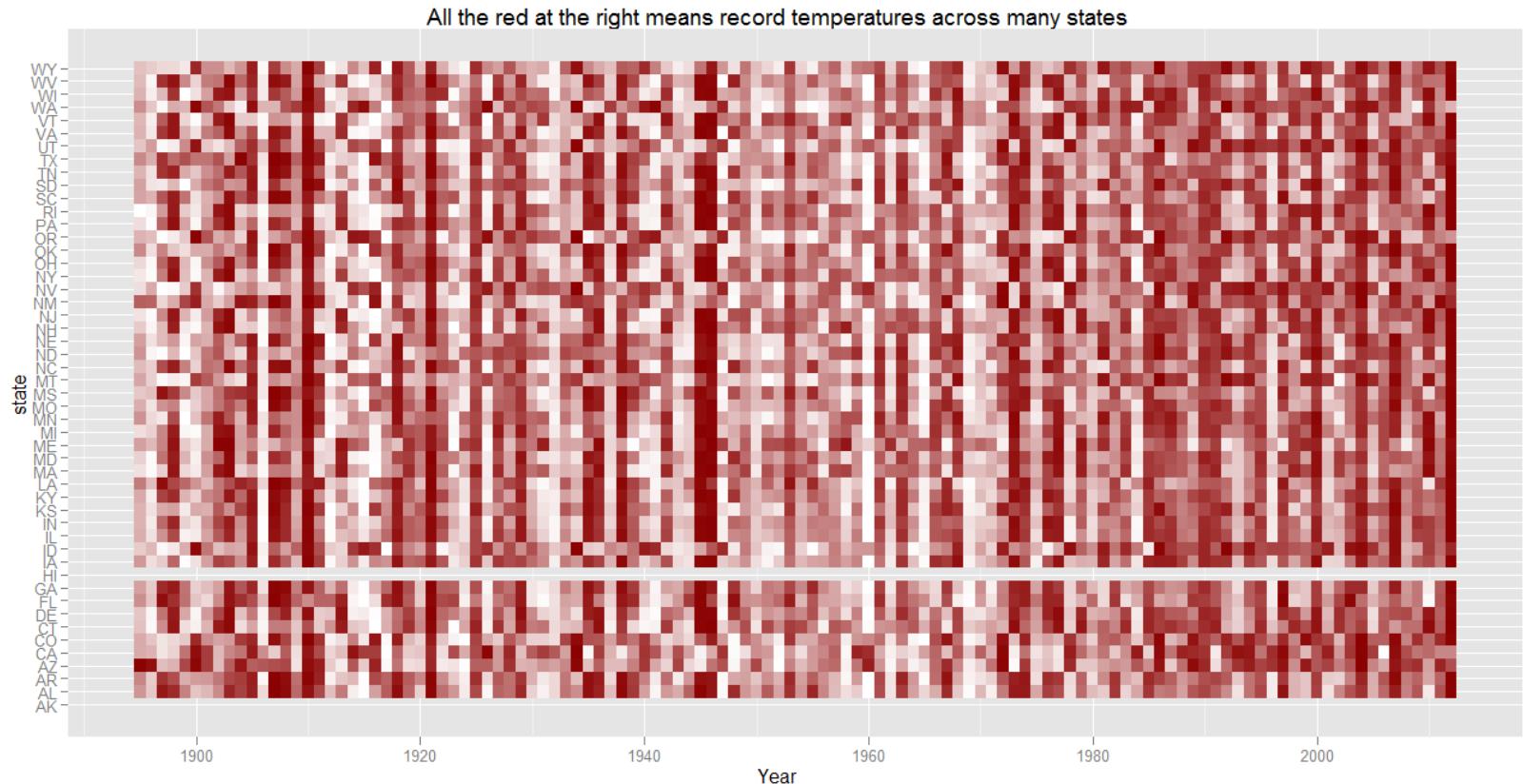


FIGURE 7-3 Default heatmap ordered by points per game

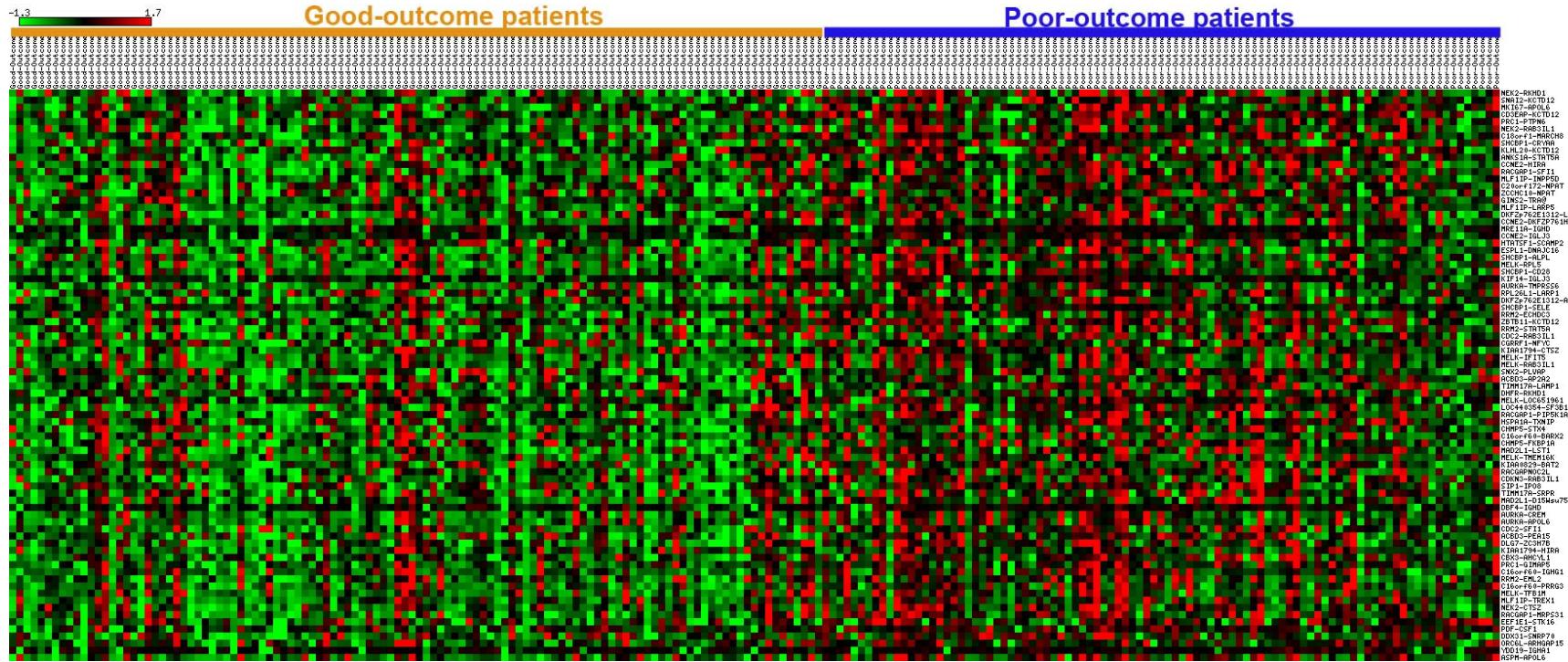
"Visualize This: The FlowingData Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.



<http://bost.ocks.org/mike/shuffle/compare.html>



<http://www.r-bloggers.com/118-years-of-us-state-weather-data/>



<http://www.biomedcentral.com/1471-2105/9/125>

Discussion

- Essentially showing the data set, but replaces numbers by color values
- Good for certain types of data
 - Continuous data well-suited
 - Unordered categorical limited to 7 categories
- Encourages comparison and pattern finding
 - Sorting changes patterns!

SCATTERPLOT MATRIX

Encode Data

- **Scatterplot**
 - Horizontal position maps to one variable
 - Vertical position maps to another variable
- **Matrix of scatterplots**
 - Each scatterplot focuses on one pair
 - Which pair is determined by row and column

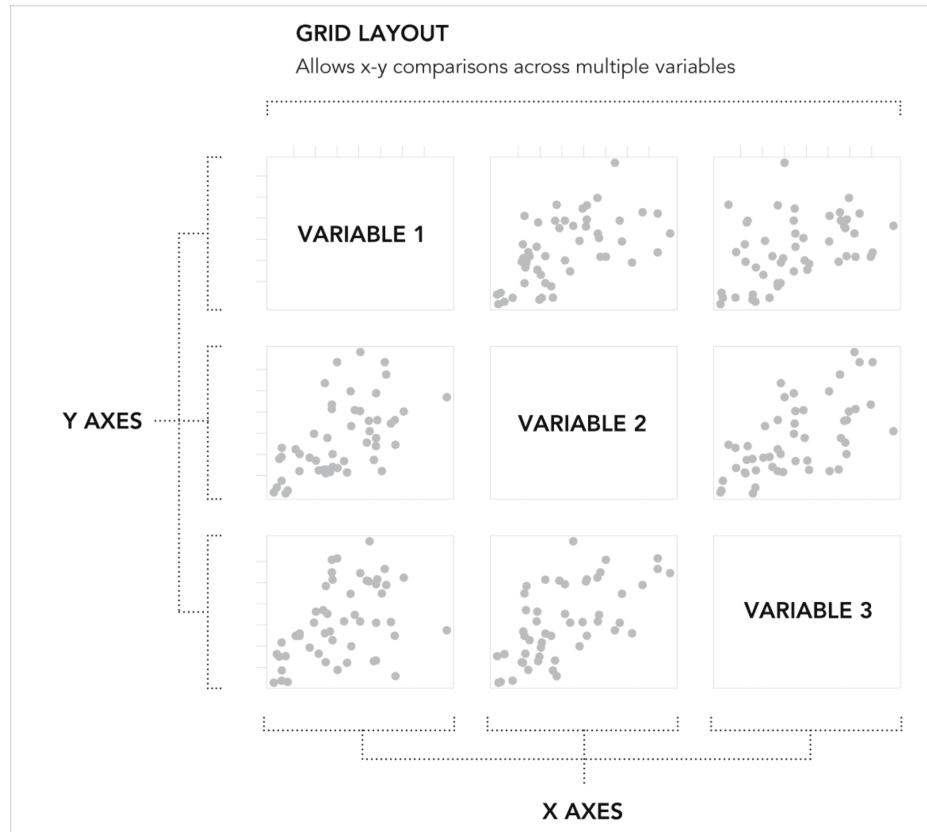


FIGURE 6-8 Scatterplot matrix framework

"Visualize This: The FlowingData Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.

MURDERS VERSUS BURGLARIES IN THE UNITED STATES

States with higher murder rates tend to have higher burglary rates.

Burglaries
per 100,000 population

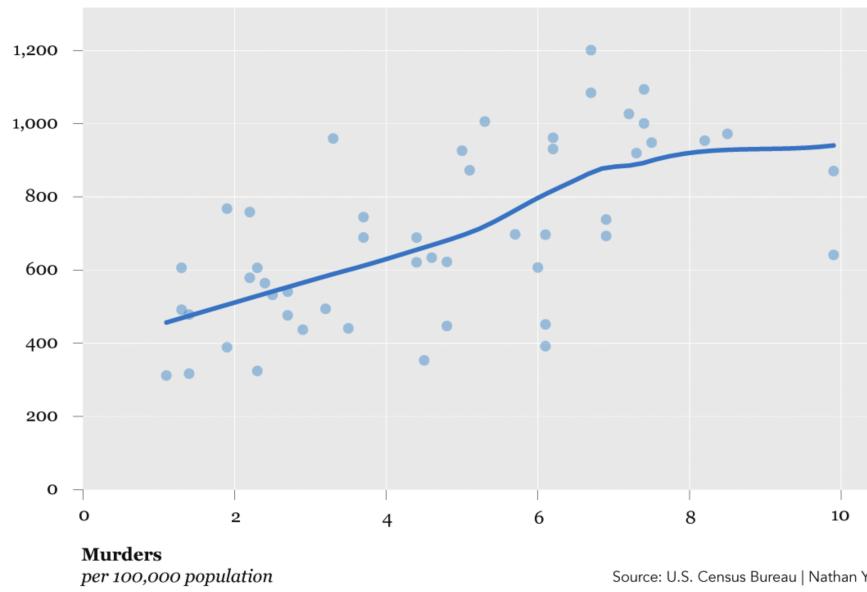


FIGURE 6-7 Revised scatterplot on murder versus burglary

"Visualize This: The FlowingData Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.

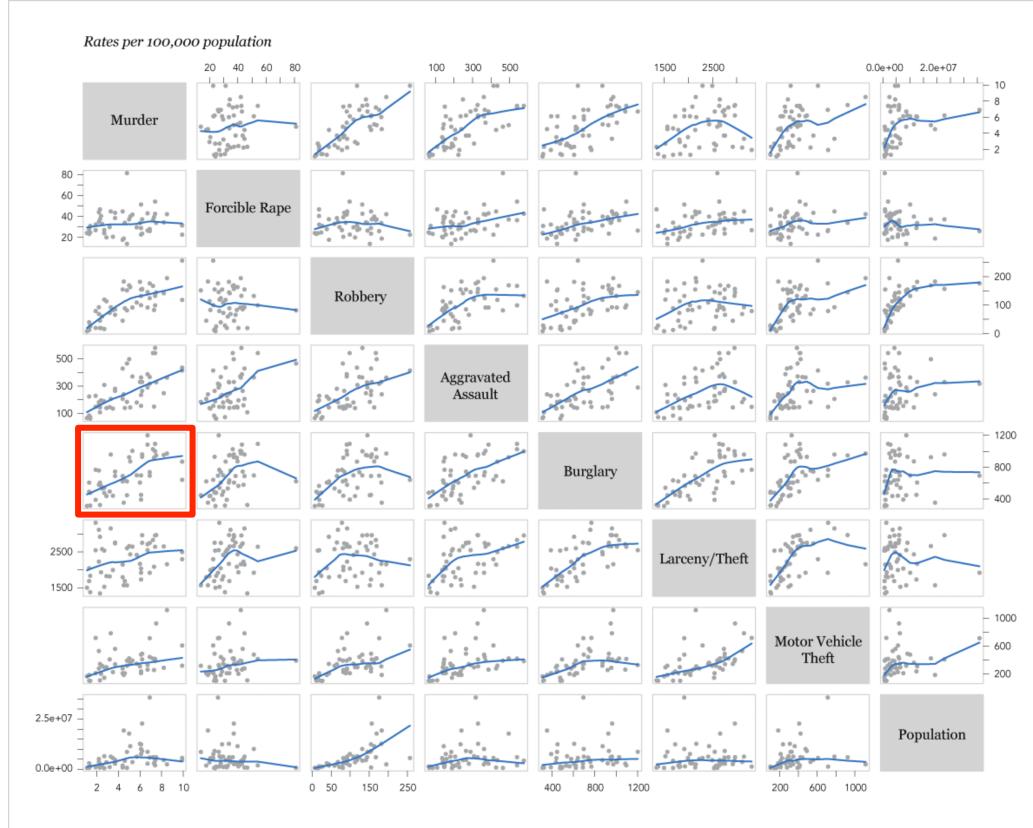
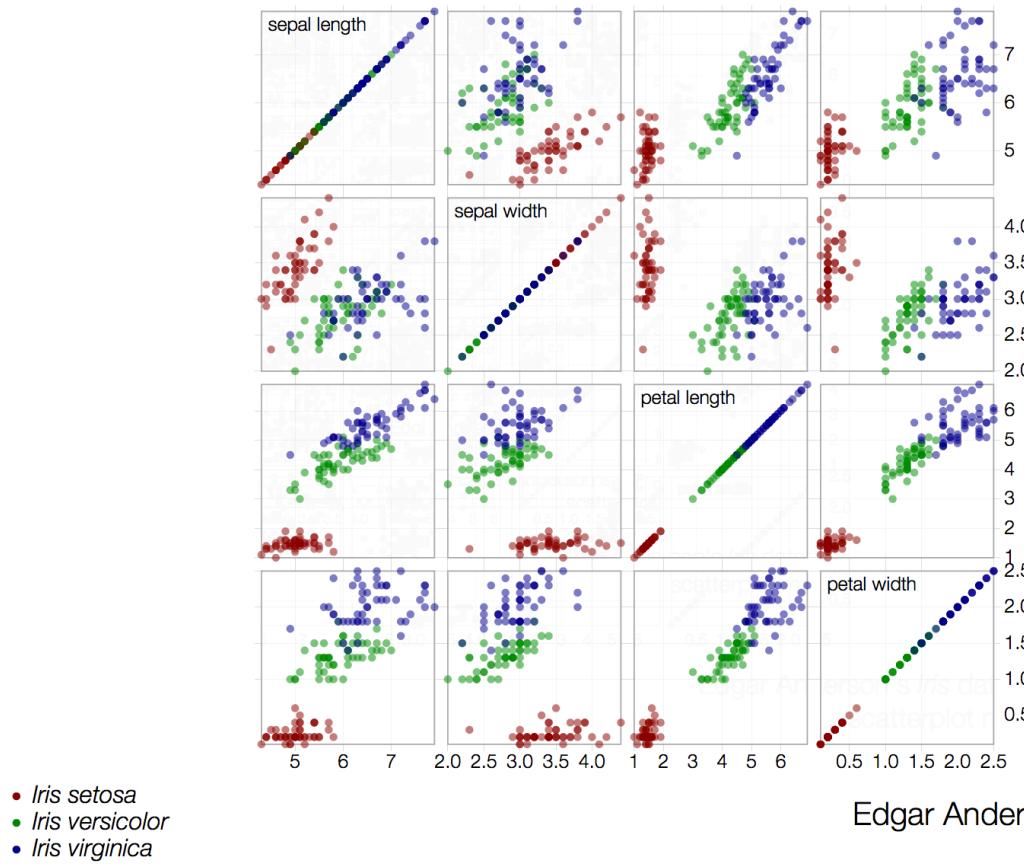


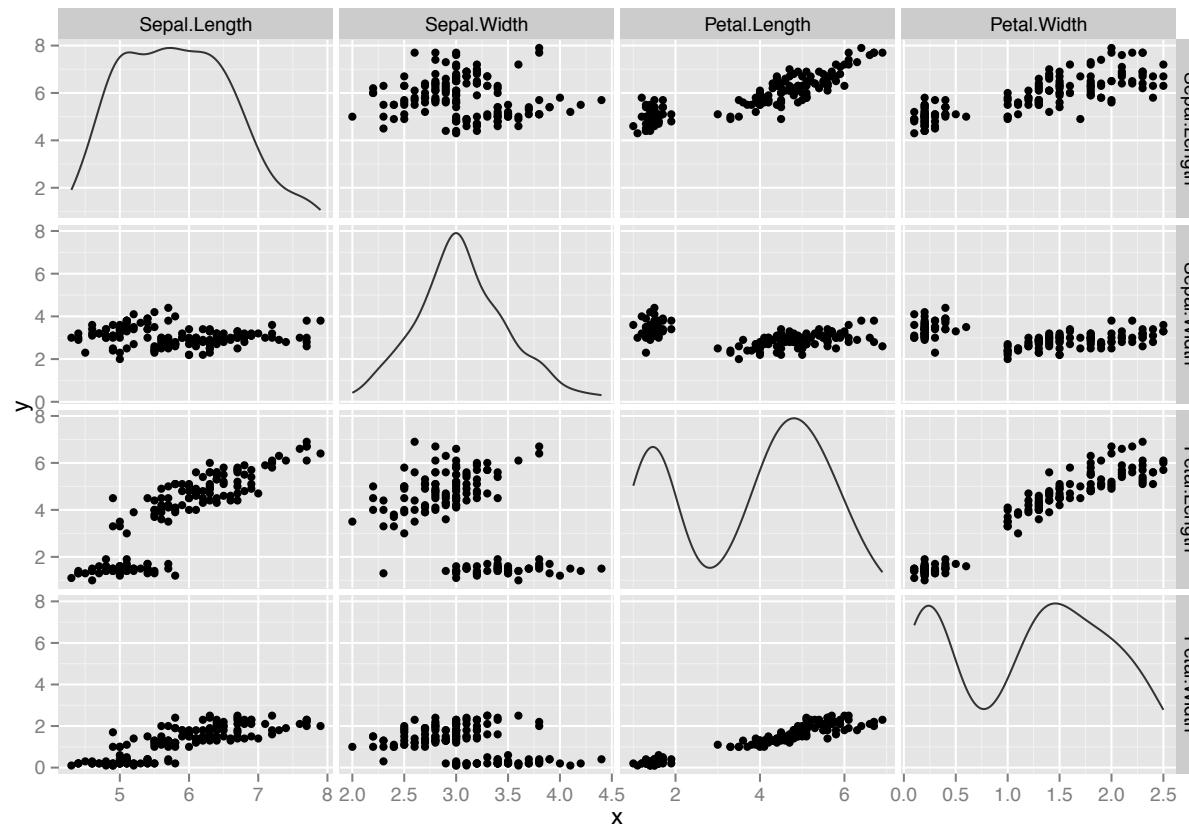
FIGURE 6-9 Scatterplot matrix of crime rates

"Visualize This: The FlowingData Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.

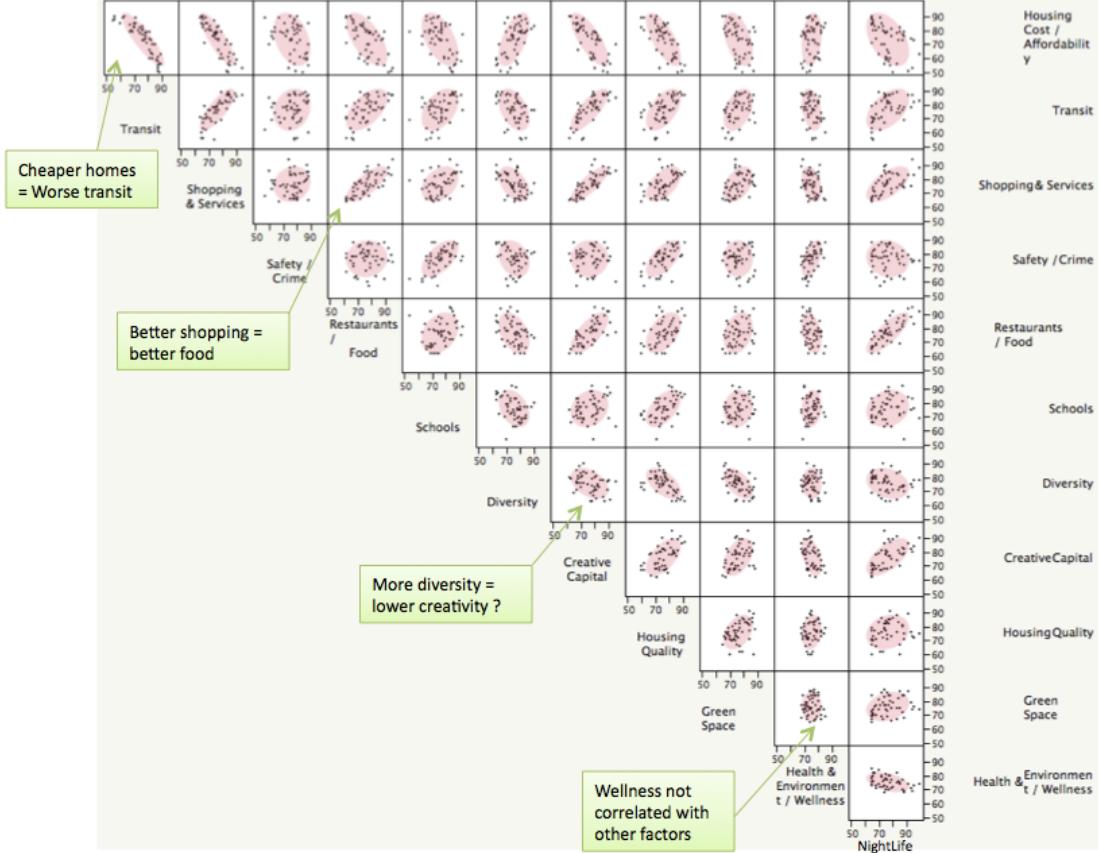


Edgar Anderson's *Iris* data set
scatterplot matrix

<http://mbostock.github.com/d3/talk/20111116/iris-splom.html>



<http://gettinggeneticsdone.blogspot.com/2011/07/scatterplot-matrices-in-r.html>



http://junkcharts.typepad.com/junk_charts/2010/06/the-scatterplot-matrix-a-great-tool.html

Discussion

- Good for exploration and comparison
 - Can be a little overwhelming at first
- Works with numerical or ordered data
- Form of small-multiples plot
 - Multiple scatterplots

SMALL MULTIPLES

Encode Data

- Group data by a variable to divide it into subsets
 - Usually a categorical variable
- Create a small plot for each subset
- Show all subset plots on same page
 - Encourages comparison

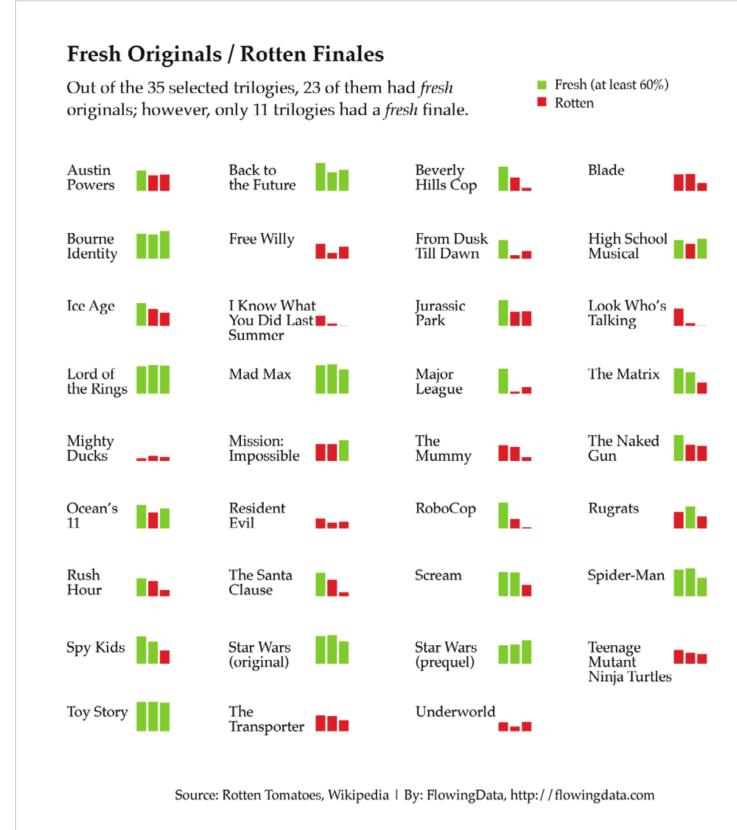
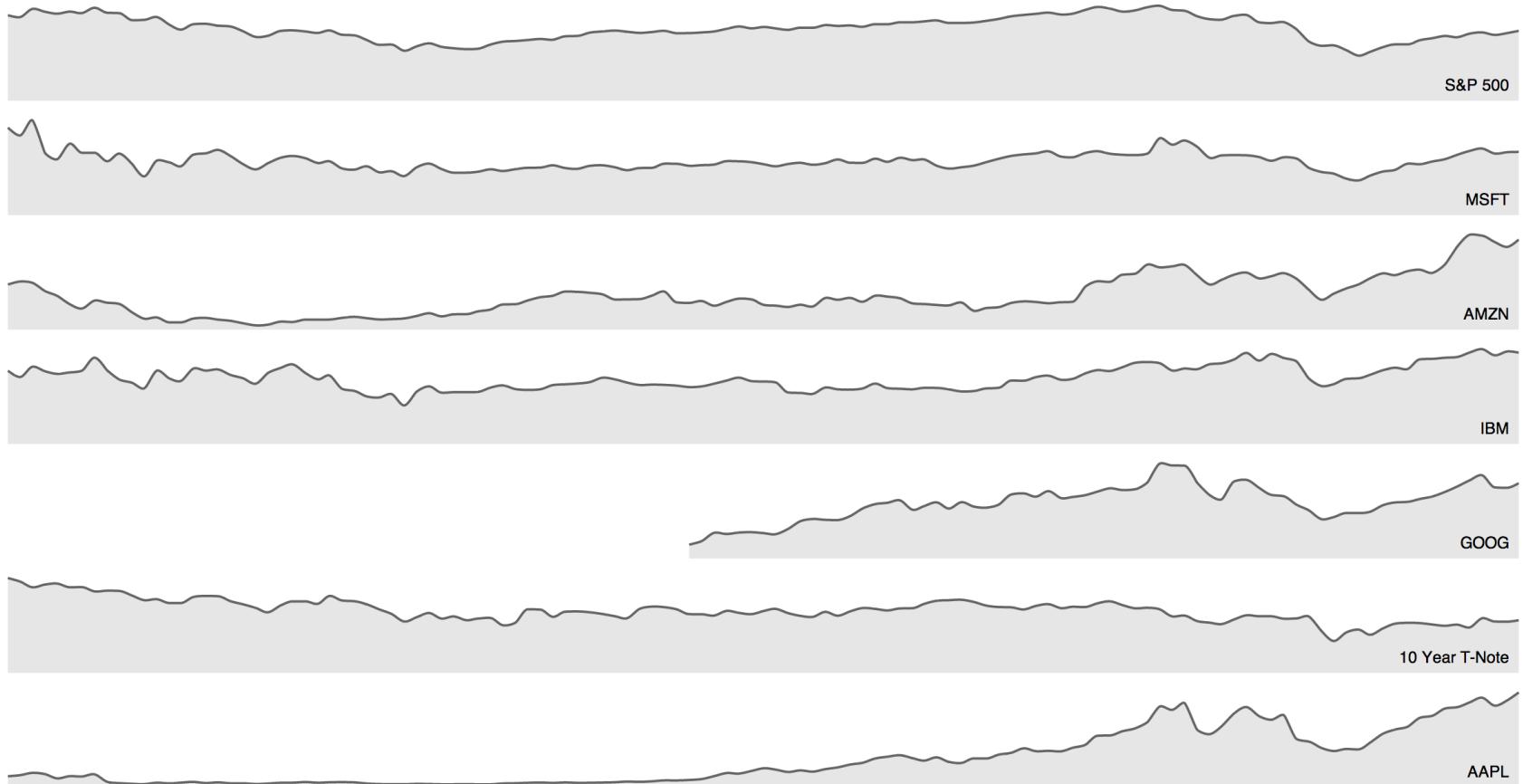
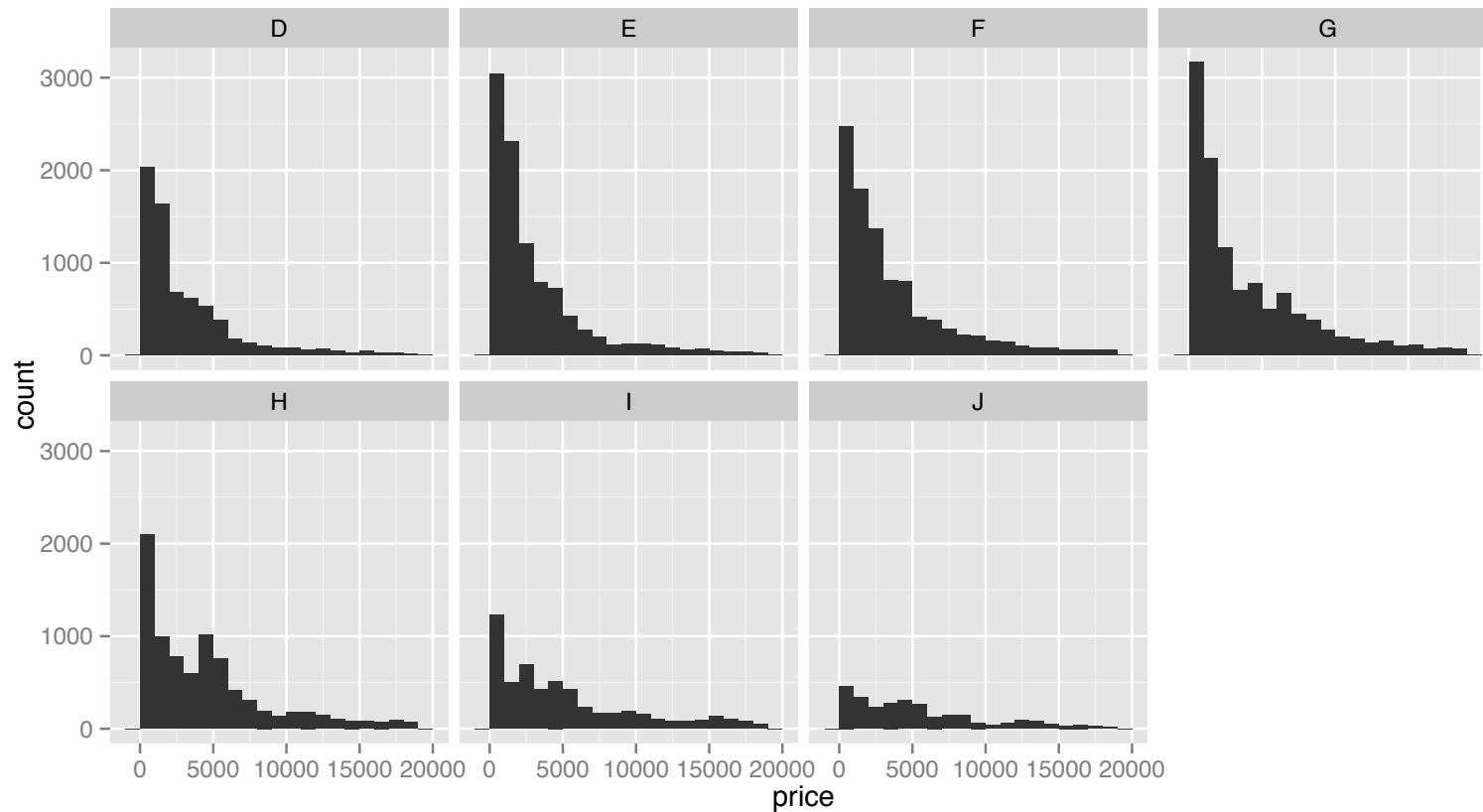


FIGURE 6-40 Small multiples for ratings of trilogies

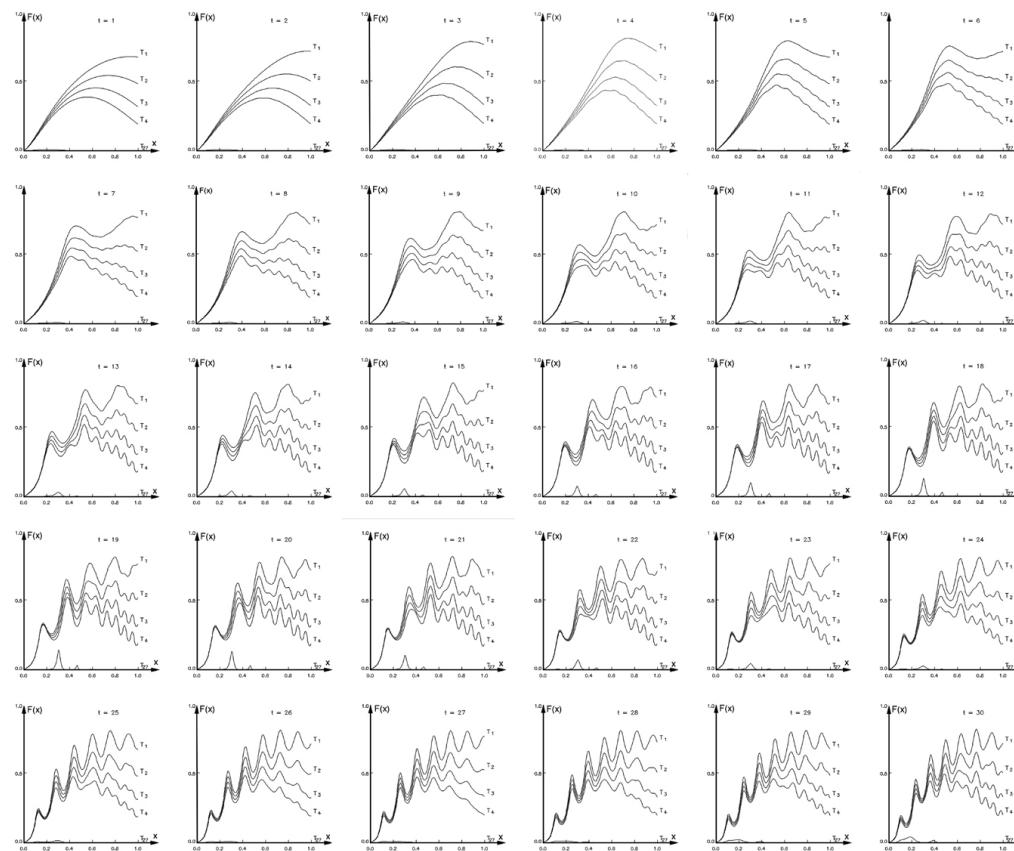
"Visualize This: The FlowingData Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.



<http://bl.ocks.org/1157787>



http://docs.ggplot2.org/current/facet_wrap.html



<http://www.sv.vt.edu/classes/ESM4714/methods/CogVizCmp.html>

Discussion

- Excellent for comparison
 - Depends on how to place the multiples!
- Requires a variable to use for grouping
 - Discrete or categorical data
- Can use with any type of plot
- Harder to tell exact values

PARALLEL COORDINATES

Encode Data

- Create one vertical line for every column
 - Numerical or ordered data
- Plot every row
 - x position is column
 - y position is value for that column
 - line connects values for a single row
- Picture an xy scatterplot, but putting both axis lines vertically

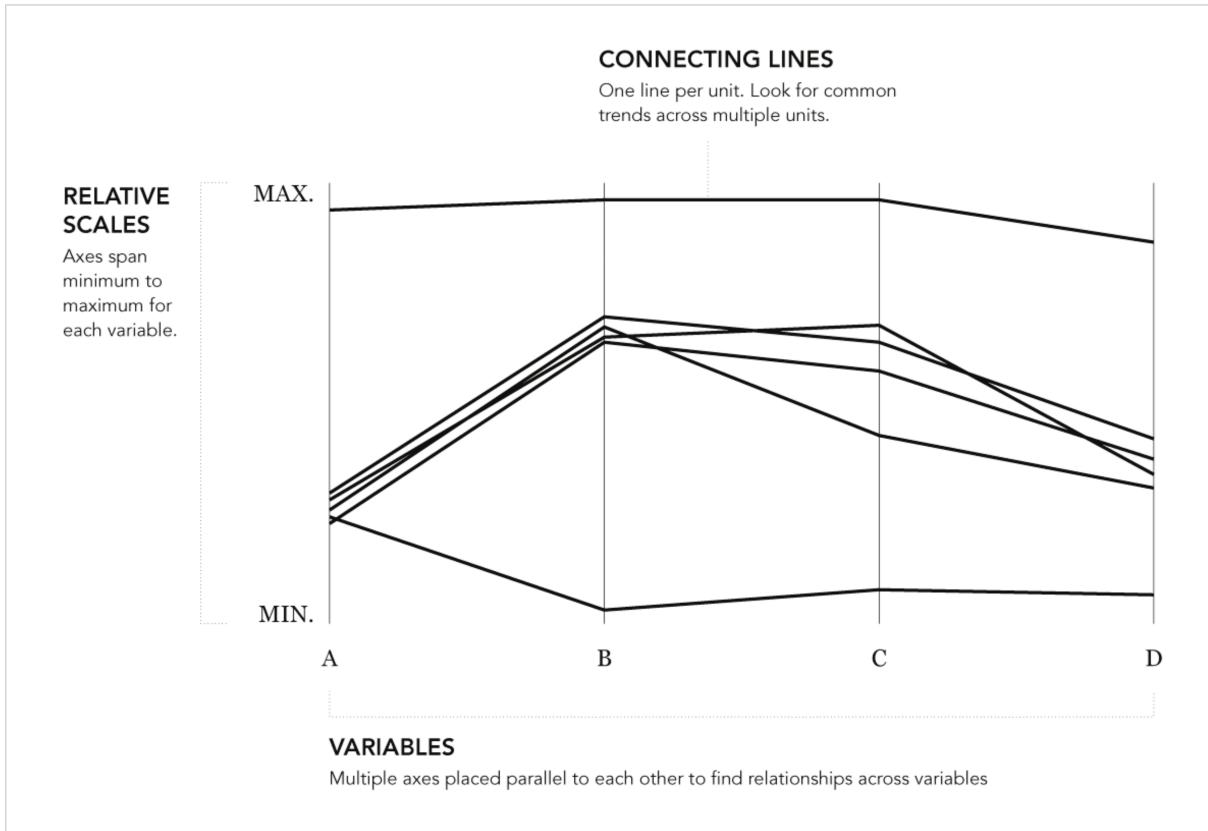


FIGURE 7-20 Parallel coordinates framework

"Visualize This: The FlowingData Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.

EDUCATION IN THE UNITED STATES

States with higher SAT reading scores predictably also have relatively higher math and writing scores. However, this does not necessarily mean that students are better educated in these states. It's more likely an indicator for what percentage of graduates actually took the test.

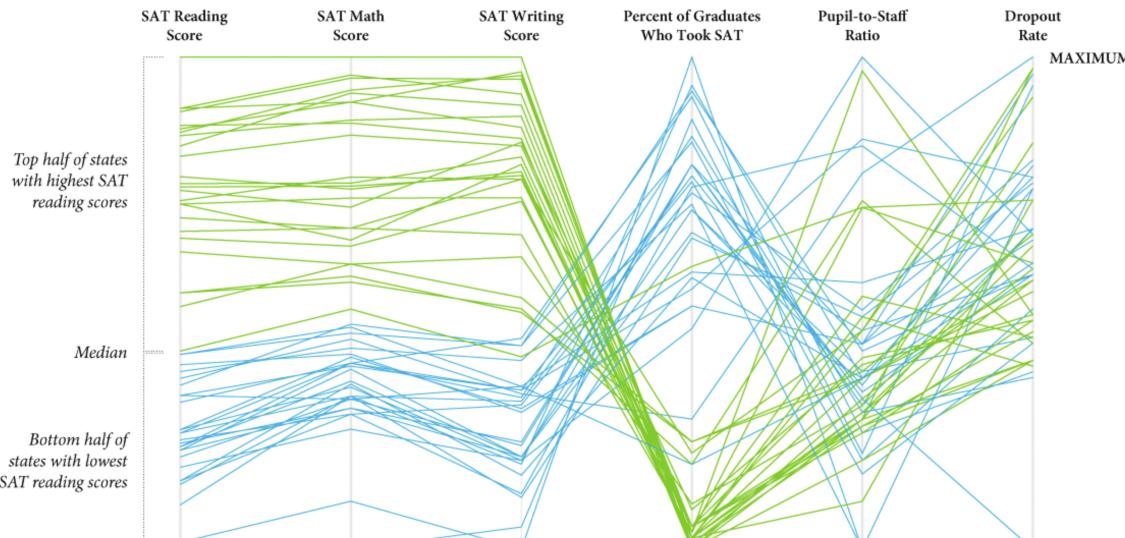
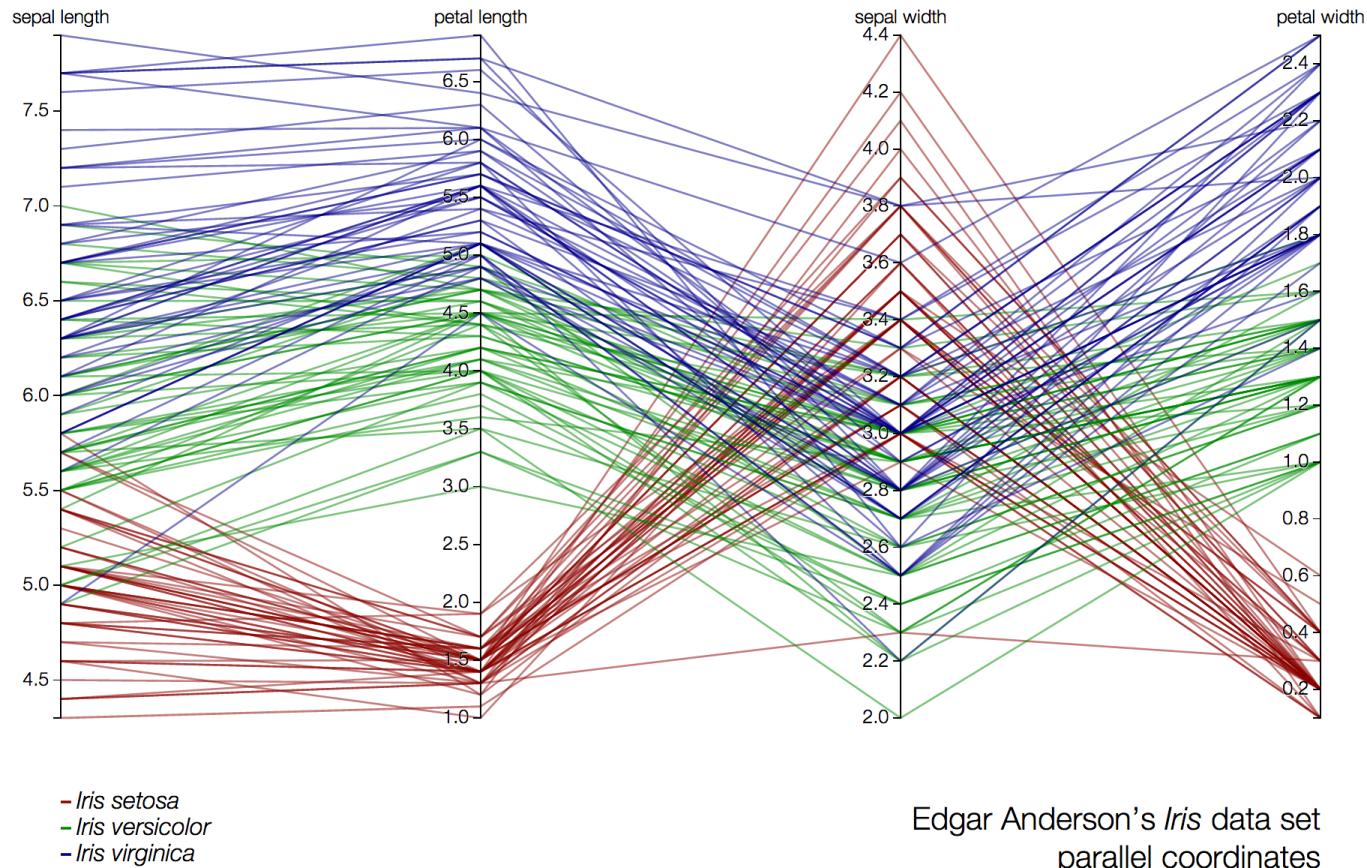


FIGURE 7-26 Standalone parallel coordinates plot on SAT scores

"Visualize This: The FlowingData Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.



<http://mbostock.github.com/d3/talk/20111116/iris-parallel.html>

Nutrient Contents – Parallel Coordinates

An interactive visualization of the [USDA Nutrient Database](#). For information on parallel coordinates, read [this tutorial](#).

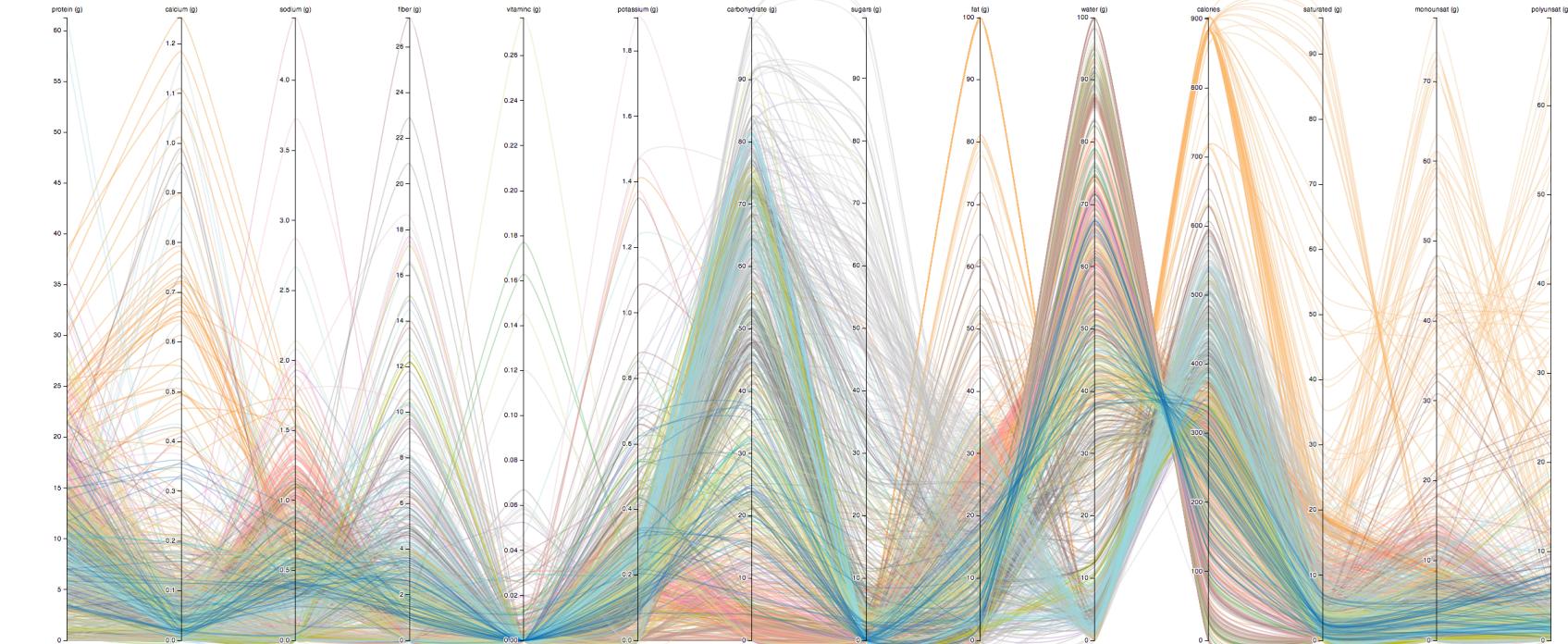
Hide Ticks | Dark | Shadows | Opacity: 34%

Per 100g of Food

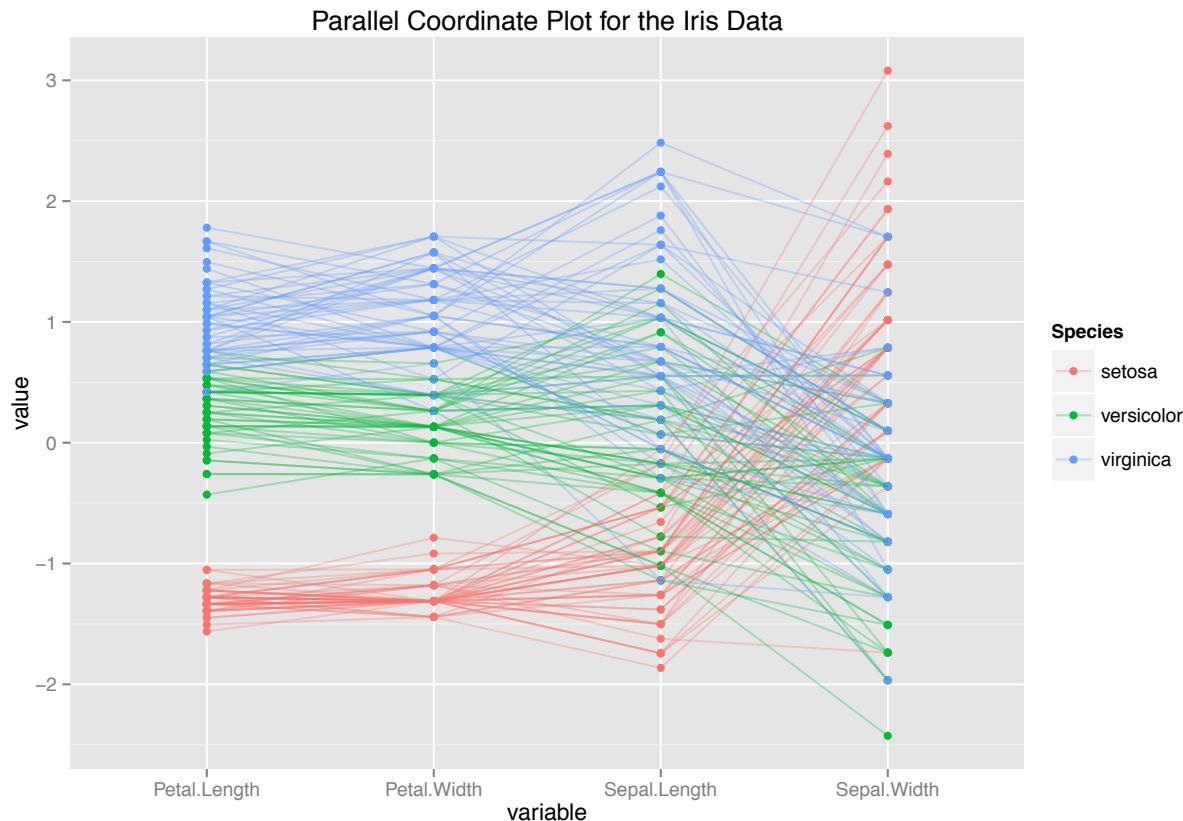
Dairy and Egg Products Fats and Oils Poultry Products Soups, Sauces, and Gravies Vegetables and Vegetable Products Sausages and Luncheon Meats Breakfast Cereals Fruits and Fruit Juices Nut and Seed Products Beverages Finfish and Shellfish Products Legumes and Legume Products Baked Products Sweets
 Cereal Grains and Pasta Fast Foods Meals, Entrees, and Sidedishes Snacks Restaurant Foods

Selected 1153 rows | Keep | Remove | Export

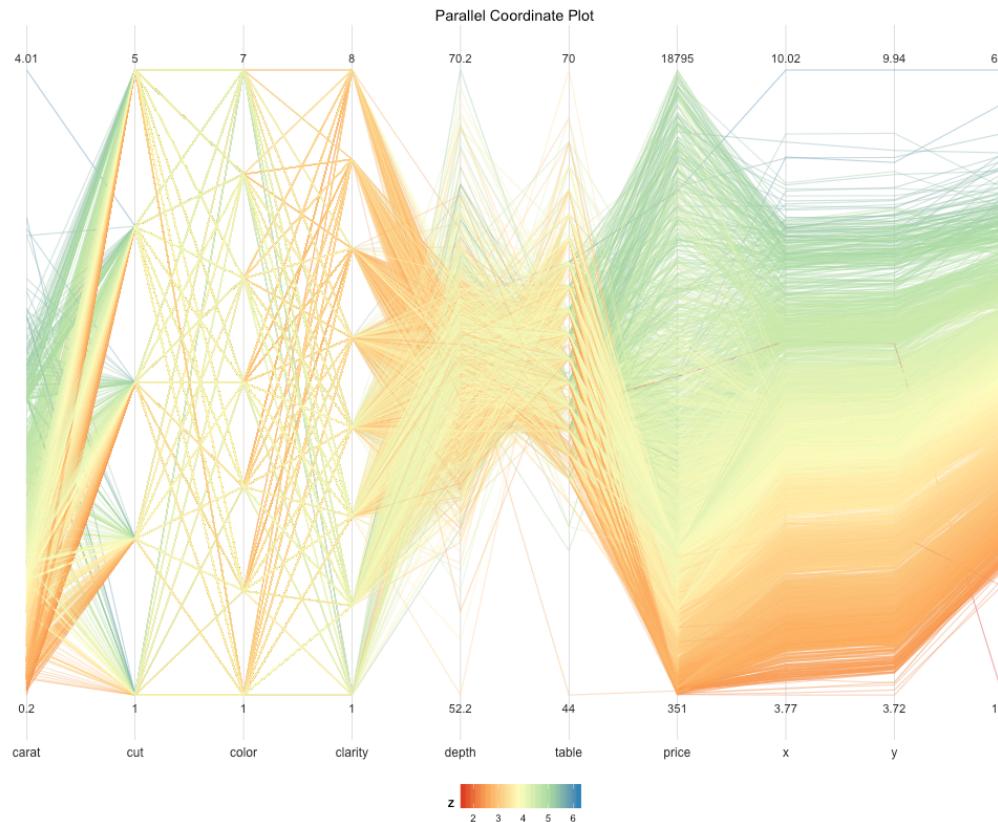
Group Breakdown | Total Selected



<http://exposedata.com/parallel/>



<http://www.inside-r.org/packages/cran/GGally/docs/ggparcoord>



Custom R and ggplot2 Implementation

Discussion

- Long startup time
- Almost always requires interactivity
 - Choose column to color by
 - Choose how to sort
 - Highlighting (brushing)
 - Clustering
- Very high density and data ink ratio, low lie factor

RADAR/STAR PLOT

Encode Data

- Same as a parallel coordinate plot,
but drawn radially

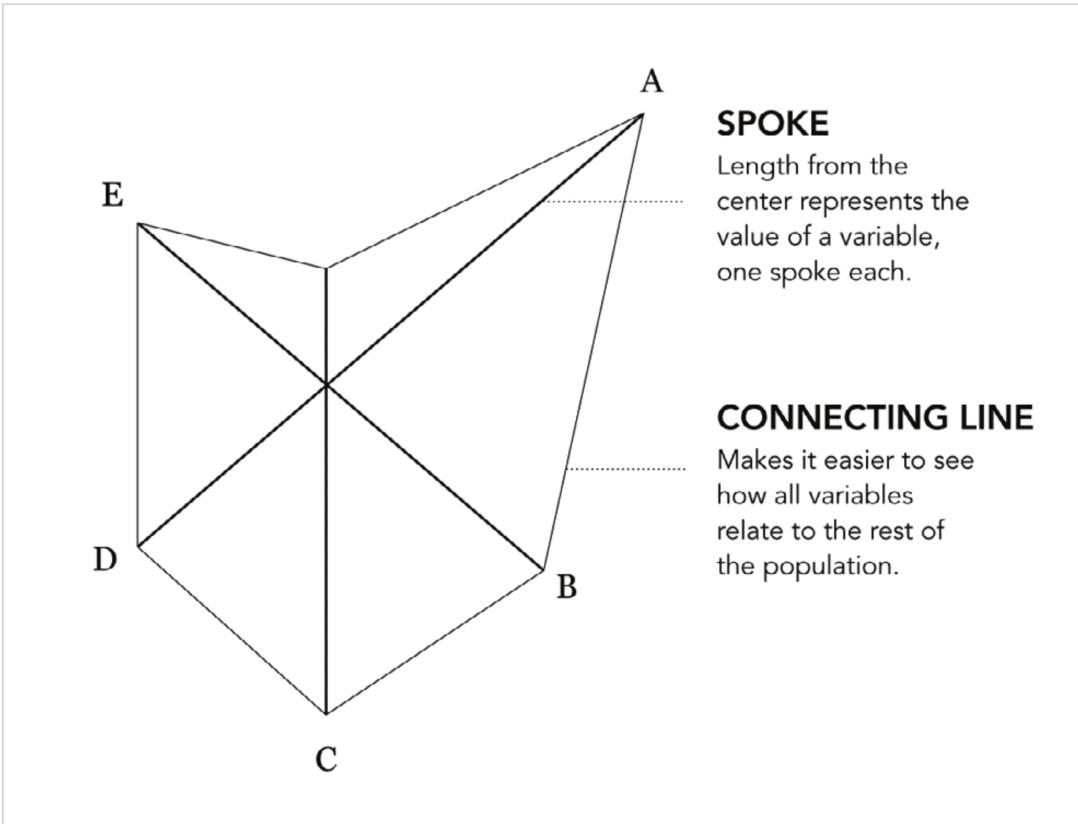


FIGURE 7-14 Star chart framework

"Visualize This: The FlowingData Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.

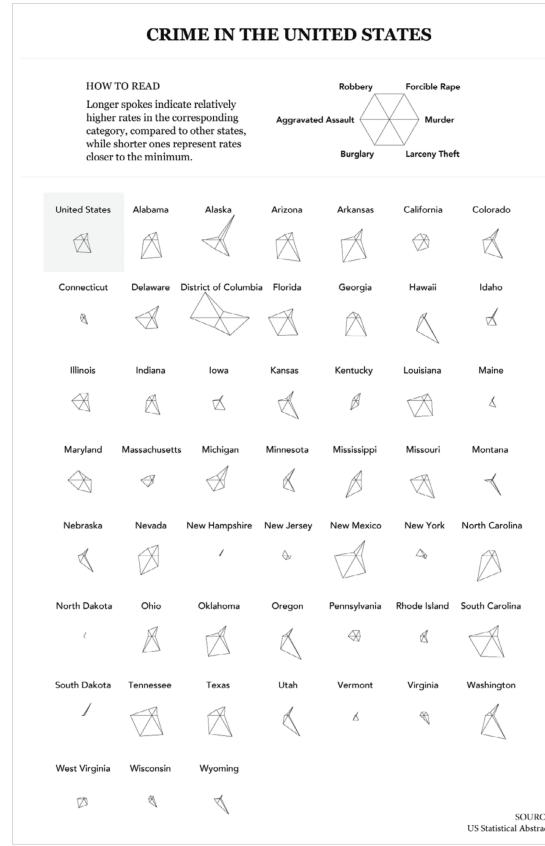
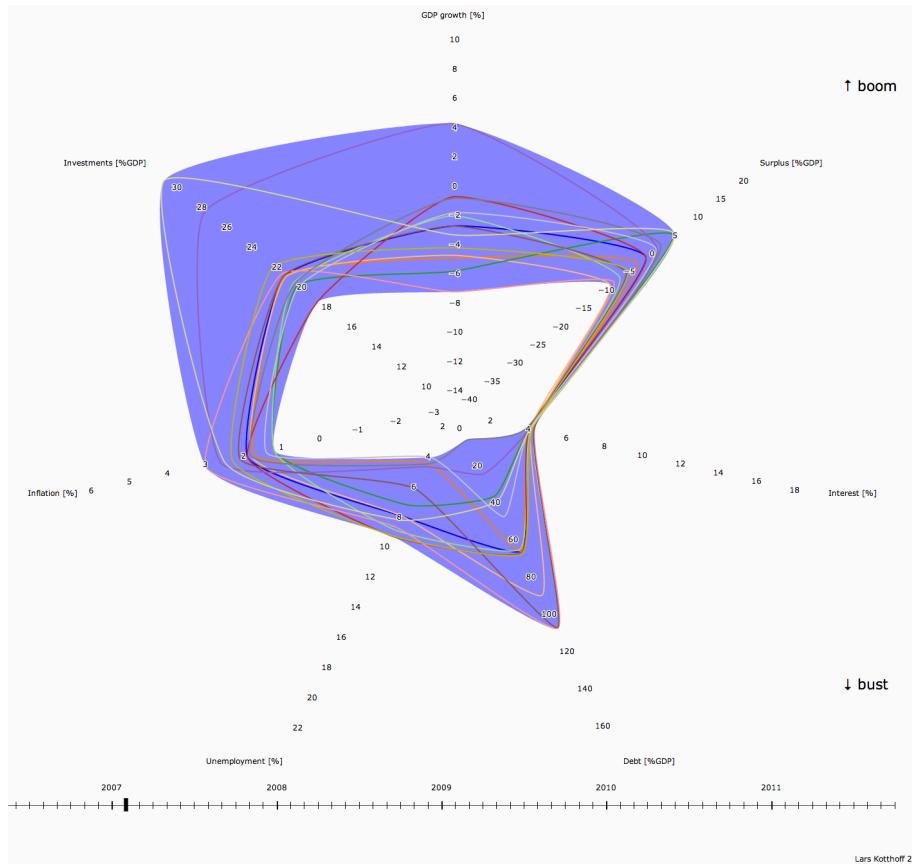
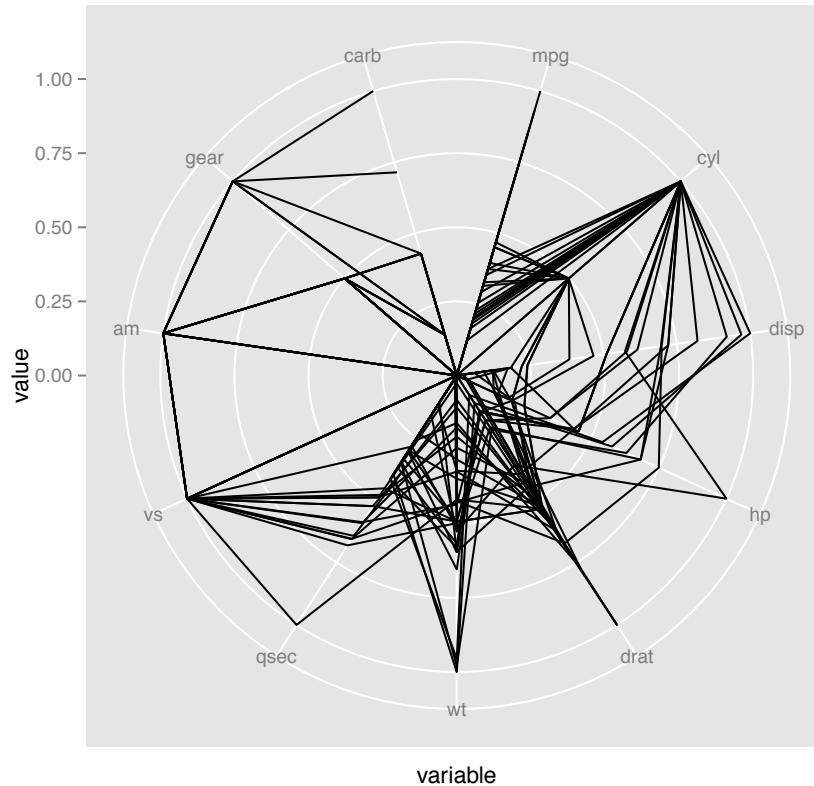


FIGURE 7-19 Series of star charts showing crime by state

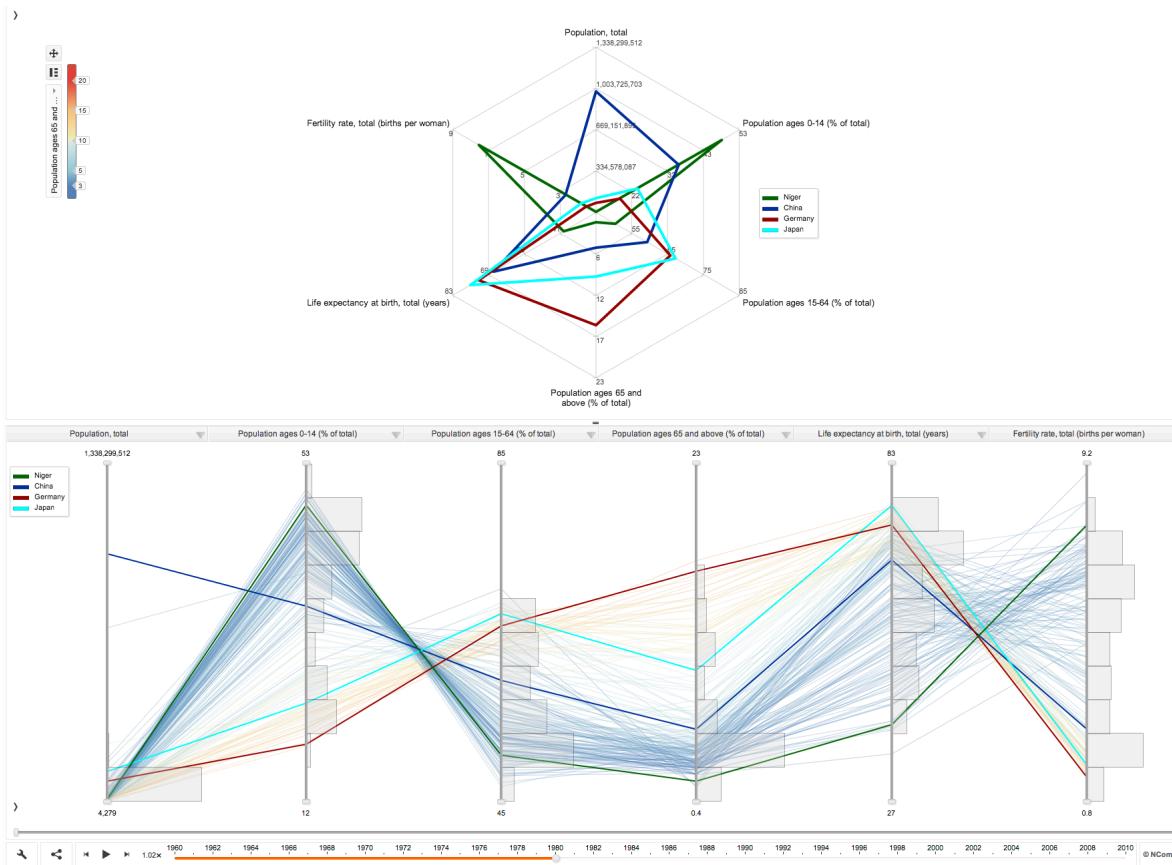
"Visualize This: The FlowingData Guide to Design, Visualization, and Statistics" by Nathan Yau, 2011.



<http://www.lasko.org/v/euc/>



<https://github.com/hadley/ggplot2/issues/516>



[http://www.ncomva.se/html5/dynamic/index.html?layout=\(radarchart,pcp\)](http://www.ncomva.se/html5/dynamic/index.html?layout=(radarchart,pcp))

Discussion

- Gives a shape to the data
- When plotted on top of each other, must have some ability to highlight and filter
- Works decently for small multiples?
 - Requires small number of rows
- Better for data with a circular aspect?
 - Monthly temperature time series

CONCLUSION

Other Techniques

- **Parallel Sets**
 - <http://eagereyes.org/parallel-sets>
- **Rose Diagrams**
 - <http://dd.dynamicdiagrams.com/2008/01/nightingales-rose/>

References



- Nathan Yau, **Visualize This: The FlowingData Guide to Design, Visualization, and Statistics**, Wiley Publishing, 2011.
- Mike Bostock, **Data Driven Documents** (D3.js)
 - <https://github.com/mbostock/d3/wiki/Gallery>
 - <http://bl.ocks.org/mbostock>
- Other references sourced on slides

QUESTIONS

<http://sjengle.cs.usfca.edu/>