

DS100 Lecture 6: Visualization

Slides by:

Deborah Nolan

deborah.nolan@berkeley.edu

Why is Graphics a topic in this course?

- Visualization belongs in every stage of the data life cycle
- Plots can uncover structure in data that can't be detected from numerical summaries
- Visualization is an important communication skill

Goals of this lecture

- Guidelines and general philosophy
 - Reveal the data
 - Facilitate Comparisons
 - Add information
 - Iterate
- Techniques for following guidelines
 - Scale
 - Conditioning
 - Perception
 - Transformations
 - Adding context
 - Smoothing & other large data considerations

Good Starting Place – Know your data type

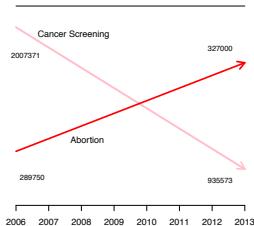
- Quantitative (Numeric)
 - Continuous (e.g., health care expenditure)
 - Discrete (e.g., number of siblings)
- Qualitative (Categorical)
 - Nominal (e.g., lane of traffic, country)
 - Ordinal (e.g., Yelp rating, education level)
- See table that maps data types to plot types at end of slides

4 Examples

2015 Congressional Hearing: Planned Parenthood

- Congressman Chaffetz (R-UT), chair US House Oversight Committee
- Investigation of federal funding of Planned Parenthood
- Chaffetz showed a plot which originally appeared in a report by Americans United for Life (<http://www.au.org/>).
- Report available at <https://oversight.house.gov/interactivepage/plannedparenthood/>

Planned Parenthood Procedures

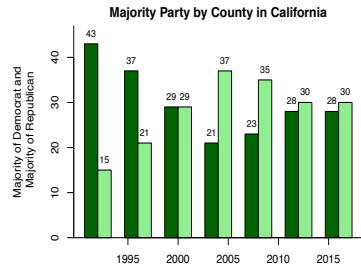


- Procedures:
 - Cancer screenings
 - Abortion
- Time: 2006 to 2013
- How many data points are in this plot?
- What's suspicious about this plot?

Voter Registration Trends in California

- State of California publishes voter registration summaries
http://www.sos.ca.gov/elections/ror/60day_presprim/histreg_stats.pdf
- Historical registration counts available for presidential election years

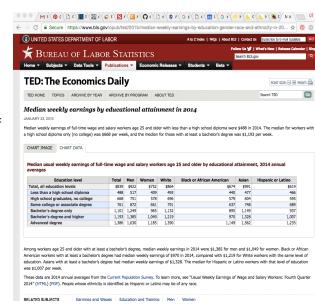
Voter Registration Trends in California



What's confusing or annoying about this plot?

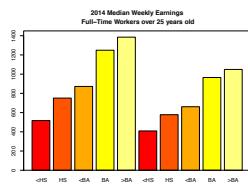
Earnings

- Bureau of Labor Statistics
 - Oversees scientific surveys related to economic health of the country
- Current Population Survey
 - Collects data on the earnings
- www.bls.gov - Web interface to a report generating app



Earnings

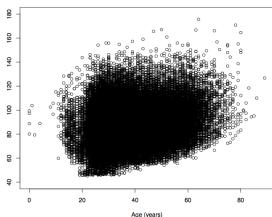
Which comparisons can be easily made with this plot?



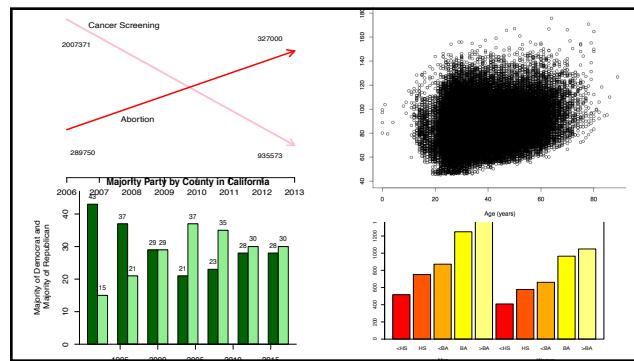
Cherry Blossom Run

- 10 mile run in Washington DC each April
- Race organizers make results available on Web
 - Runner name, age, gender, address, hometown, time
 - Race results from 1999 to 2016
 - In 2012 nearly 17,000 runners ranging in age from 9 to 89 participated
 - <http://www.cherryblossom.org/>

Cherry Blossom Run



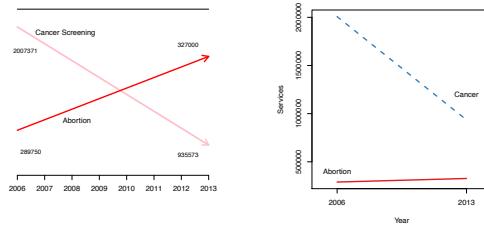
- Scatter plot of run time (minutes) by age (years)
- 70,000 points in this plot.
- What's the relationship between run time and age?



Techniques

Scale Conditioning Perception Transformation Context Smoothing Philosophy

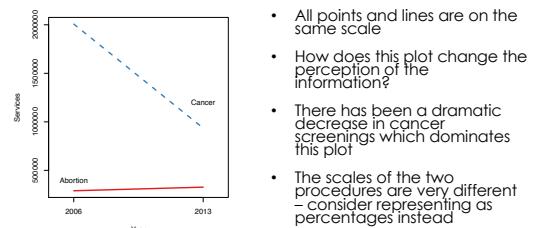
Planned Parenthood Procedures



Scale

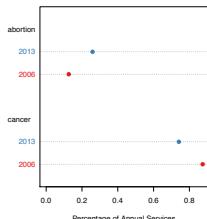
Scale Conditioning Perception Transformation Context Smoothing Philosophy

Planned Parenthood Procedures



- All points and lines are on the same scale
- How does this plot change the perception of the information?
- There has been a dramatic decrease in cancer screenings which dominates this plot
- The scales of the two procedures are very different – consider representing as percentages instead

Planned Parenthood Procedures

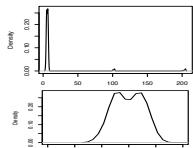


- Procedures in 2006 and 2013 as a percentage
- Abortions increased from 13% to 26% of total procedures
- May want to plot the percent change, screenings fell 50%

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

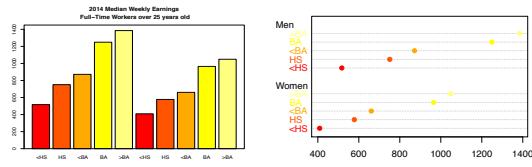
Choosing the Scale

- Choose axis limit to fill the plotting region
- In necessary,
 - Zoom in to focus on region with bulk of data
 - Make multiple plots of different regions
 - Transform data to improve resolution (TBC)
- Don't change scale mid-axis
- Don't use two different scales for the same axis



Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

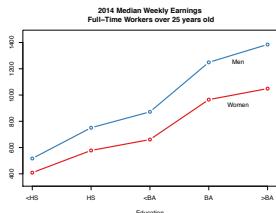
Earnings



Conditioning

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Earnings



- Emphasize the important difference –
- Lines make it easier to see growth in gap
- Placement of one point above the other makes it easier to compare males & females

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Conditioning – Distributions & Relationships in subgroups

- Superpose density curves, fitted curves and lines from different subgroups
- Juxtapose scatter plots, histograms & keep x and y scales the same across plots to facilitate comparison
- Use color and plotting symbols to represent additional variables

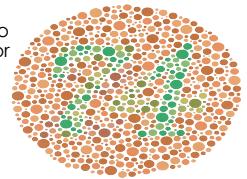
Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Perception - Color

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Color Guidelines

- Choosing a set of colors which work well together is a challenging task for anyone who does not have an intuitive gift for color
- 7-10% of males are red-green color blind.



Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Colorfulness

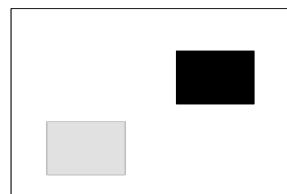
- Saturated/colorful colors are hard to look at for a long time.
- They tend to produce an after-image effect which can be distracting.



Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Luminance

- Areas should be rendered with colors of similar luminance (brightness).
- Lighter colors tend to make areas look larger than darker colors



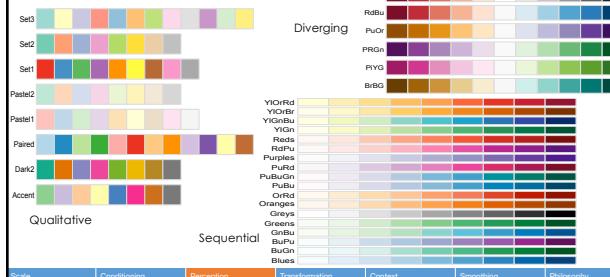
Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Data Type and Color

- Qualitative – Choose a **qualitative** scheme that makes it easy to distinguish between categories
- Quantitative – Choose a color scheme that implies magnitude.
 - Does the data progress from low to high? Use a **sequential** scheme where light colors are for low values
 - Do both low and high value deserve equal emphasis? Use a **diverging** scheme where light colors represent middle values

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Examples of Palettes



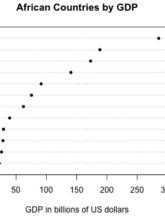
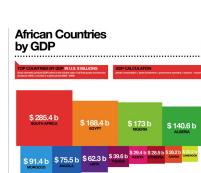
Perception - Length

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Bar plot, Pie chart, Dot chart

- Experiments found that angle judgments based on pie charts are less accurate than length judgments from bar charts
- Length is easier to compare than area or volume
- Lengths that fall on a line are easier to compare than lengths on parallel lines, i.e., judgments based on dot charts are easier to make than judgments based on bar plots

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy



Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Stacking and Jiggling

- Stacked bar plots and histograms are difficult to read because the base line moves from one bar to the next
- Line plots where the area between successive lines represent the measurement are very difficult to read because the base line jiggles up and down.

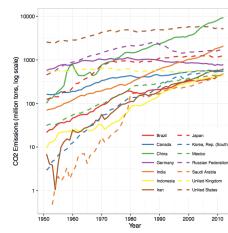
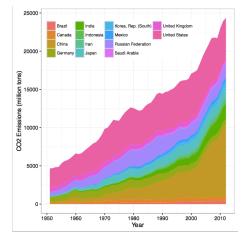
Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

CO2 Emissions from Fuel Consumption

- Data on historical carbon dioxide (CO2) emissions from fuel combustion (<http://cait.wri.org>)
- Country annual CO2 emissions date back to 1850
- Typical report on trends since 1950 for the 14 countries that emitted the greatest amount of CO2 in 2012
- World Resources Institute (<http://www.wri.org/>)

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

CO2 Emissions



Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Transformations

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Why Transform Variables?

- Reveal distribution of most of the observations (otherwise much of the data is squashed in a small region)
- Numerical summaries of a transformed data are better summaries of a symmetric distribution
- Choose a transformation that's simple and easily interpreted in the context of the problem, e.g., a power of 2, 3, $\frac{1}{2}$, 0 (log), -1

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Power Transformation

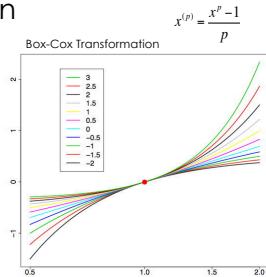
Preserve order of values

Effective when max / min > 5

Sometimes add a shift before transform

Ratio of hinges can help select a transformation

$$\frac{\text{Upper Quartile} - \text{Median}}{\text{Median} - \text{Lower Quartile}} = 1$$



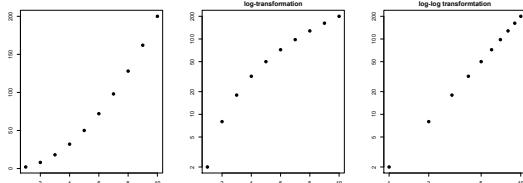
Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Why Straighten Relationships?

- Easier to uncover the form of the relationship if we can transform it to linear relationship; we see what transformation used to make it linear
- Linear relationships are particularly simple to interpret & fit
- Choose a transformation that's simple and easily interpreted in the context of the problem, e.g., a power of 2, 3, $\frac{1}{2}$, 0 (log), -1

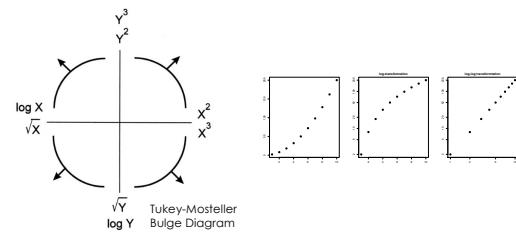
Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Straighten Relationships with Transformations



Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Straighten Relationships with Transformations



Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

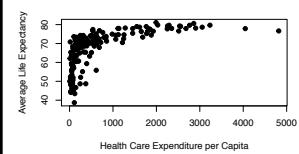
World Bank Country Statistics

- World Bank provides financial and technical assistance to developing countries.
- In 2010, the World Bank launched an Open Data Website that provides access to data from their reports on topics such as GDP, education, health, and the environment.
- We are interested in the relationship between life expectancy and health expenditures
- These variables are measured at the country level

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

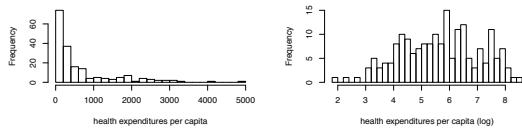
Healthcare Costs & Life Expectancy

What does the bulge diagram tell us?



Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

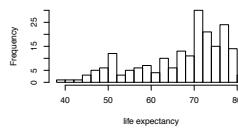
Healthcare Costs



Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

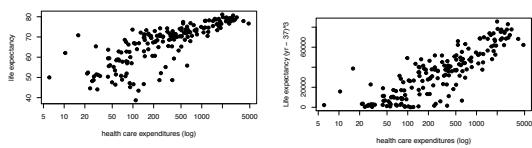
Life Expectancy

- Unusual feature – left skew. Upper bound on life expectancy
- What transformation?
 - Range is factor of 2 so shift first
 - Pull up the low end with a square or cube transformation



Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Healthcare Costs & Life Expectancy



Issues remaining:
3 unusual countries
Complexity of cube model

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Add Context

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Add Context

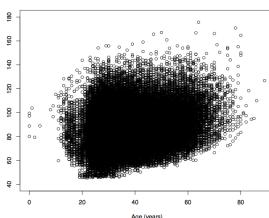
- Label axes, including units
- Add Reference lines and markers for important values
- Label points of unusual/interesting observations
- Include captions that describe data, how plotted, and describe important features

Scale Conditioning Perception Transformation Context Smoothing Philosophy

Large n (records)

Scale Conditioning Perception Transformation Context Smoothing Philosophy

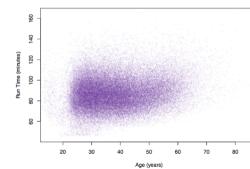
Cherry Blossom Run



Scale Conditioning Perception Transformation Context Smoothing Philosophy

- 3-dimensional histogram is needed, but hard to come by
- Use heat map or hexbin plot or transparency
- Add smooth curve that takes local averages to see the conditional center, i.e., average y in a neighborhood of x

Cherry Blossom Run



Slightly transparent points

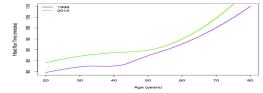
Scale Conditioning Perception Transformation Context Smoothing Philosophy

Local Smoothing helps us see the center

Control for year – race popularity

Observational data – snapshot in time

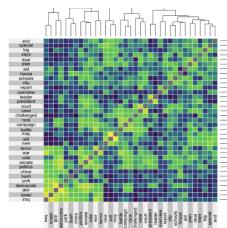
Not Longitudinal – follow same people in time



Large p (variables)

Scale Conditioning Perception Transformation Context Smoothing Philosophy

Heat map



Scale Conditioning Perception Transformation Context Smoothing Philosophy

Documents – records

Words counts – variables

Hierarchical clustering groups documents that have similar distributions of words

Heat Map - Color is used to denote closeness

Philosophy

Scale	Conditioning	Perception
-------	--------------	------------

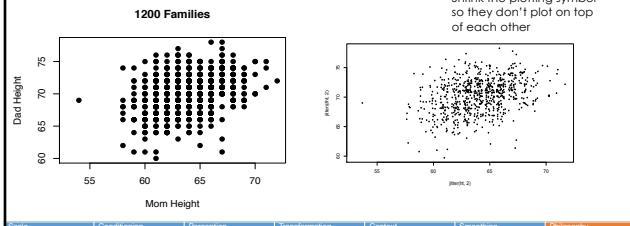
- ## Reveal the Data

- Choose scale appropriately
 - Avoid having other graph elements interfere with data
 - Use visually prominent symbols
 - Eliminate superfluous material, aka chart junk
 - Avoid over-plotting

Scale	Conditioning	Perception	Transformation	Context	Smoothing	Philosophy
-------	--------------	------------	----------------	---------	-----------	------------

Avoid over-plotting

Why are there so few data points?



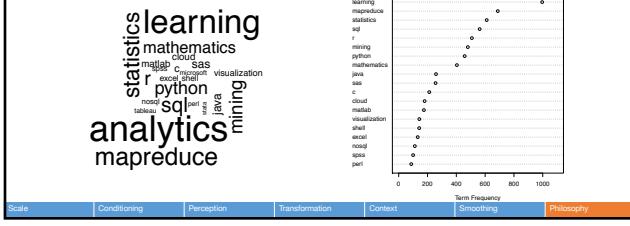
- ## Facilitate Comparisons

- Put Juxtaposed plots on same scale
 - Make it easy to distinguish elements of superposed plots, e.g. color, line type
 - Avoid Stacking and Jiggling the baseline
 - Avoid angles, extra dimensions (e.g., areas rather than lines)
 - Don't break the visual metaphor, i.e., if use rectangles, then area should correspond to value

Score	Conditioning	Perception	Transformation	Context	Smoothing	Philosophy
-------	--------------	------------	----------------	---------	-----------	------------

Comparison: area vs length

Broken Visual metaphor –
count is represented by
height of word, not area



- Make a plot information rich

- Describe what you see in the Caption
 - Add context with Reference Markers (lines and points) including text
 - Add Legends and Labels
 - Use color and plotting symbols to add more information
 - Plot the same thing more than once in different ways/scales
 - Reduce clutter

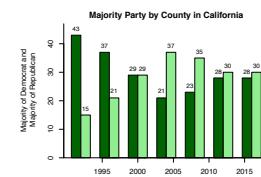
Scale Conditioned Perception Transformation Context Term Frequency Smoothing Philology

Captions

- Captions should be comprehensive
- Self-contained
- Captions should:
 - Describe what has been graphed
 - Draw attention to important features
 - Describe conclusions drawn from graph

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

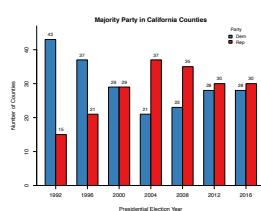
Iterate – Example Voter Registration



- Location of tick marks under bars
- Color of bars – indicate party
- Title confusing
- Y-axis label confusing
- X-axis label missing

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

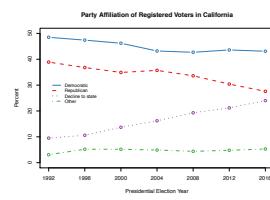
Voter Registration Trends



- Check data for understanding of how plot is made
- Observation? People vote, not counties
- Lurking variable? County size - small counties tend to be rural and conservative

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Voter Registration Trends

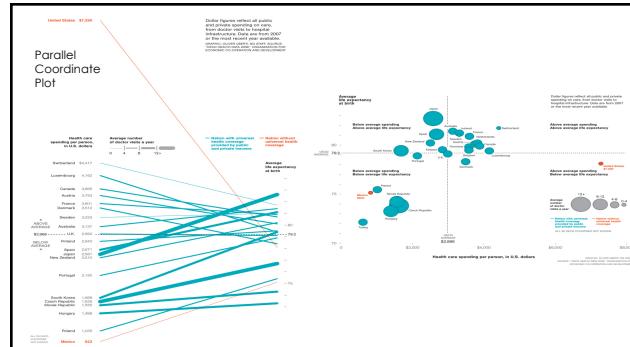


- What is the message?
- Can we improve it?
- Collect better/more data
 - Decline to state and other parties are missing
 - Voter registration totals may be more useful

Scale | Conditioning | Perception | Transformation | Context | Smoothing | Philosophy

Good Plot Making Practice

- Put major conclusions in graphical form
- Provide reference information
- Proof read for clarity and consistency
- Graphing is an iterative process
- Multiplicity is OK, i.e., two plots of the same variable may provide different messages
- Make plots data rich



Univariate Displays

Type	Plot
Numeric – few observations	Histogram, Density curve Box plot, Violin plot Normal quantile plot Few Observations - Rug plot, Dot plot Caution if discrete: density curves and box plots may be misleading
Categorical – Counts of categories	Dot chart Bar chart Pie chart (avoid) Caution if ordinal-order of bars, dots, etc. should reflect category order

Bivariate Displays

	Numeric	Categorical
Numeric	Scatter plot Smooth scatter Smooth lines and curves	Multiple histograms, density curves, Avoid jiggling!
Categorical		Side-by-side bar plot Overlaid Lines plot Side-by-side dot chart Mosaic plot Avoid stacking!