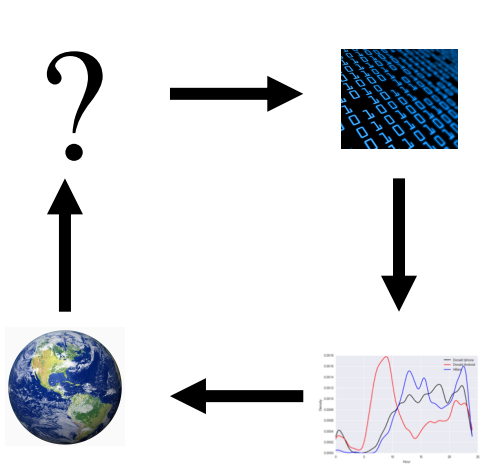
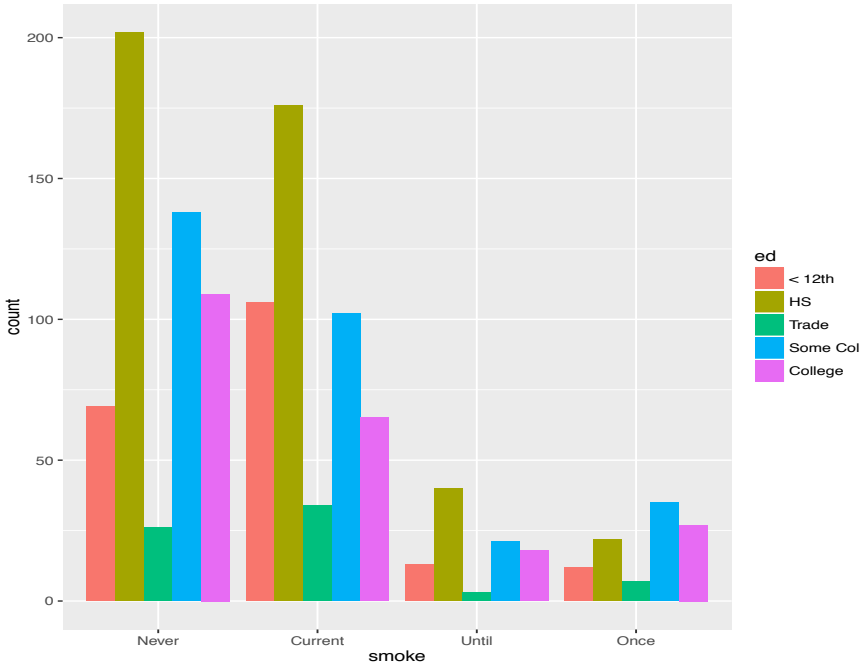
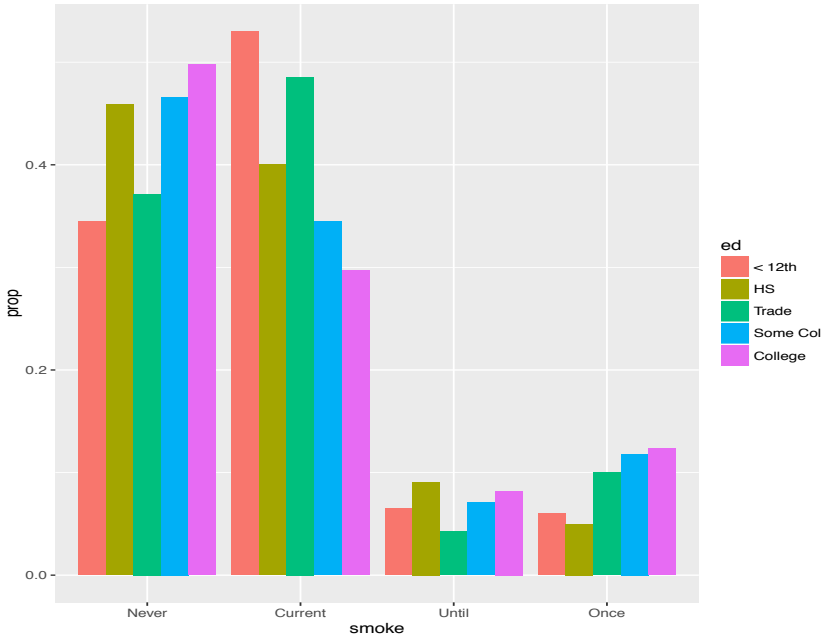


Data 100

Lecture 8: *Visualization*

Clarifications from last lecture





The y-axis for the plot on the left are counts, which can be misleading when comparing across groups

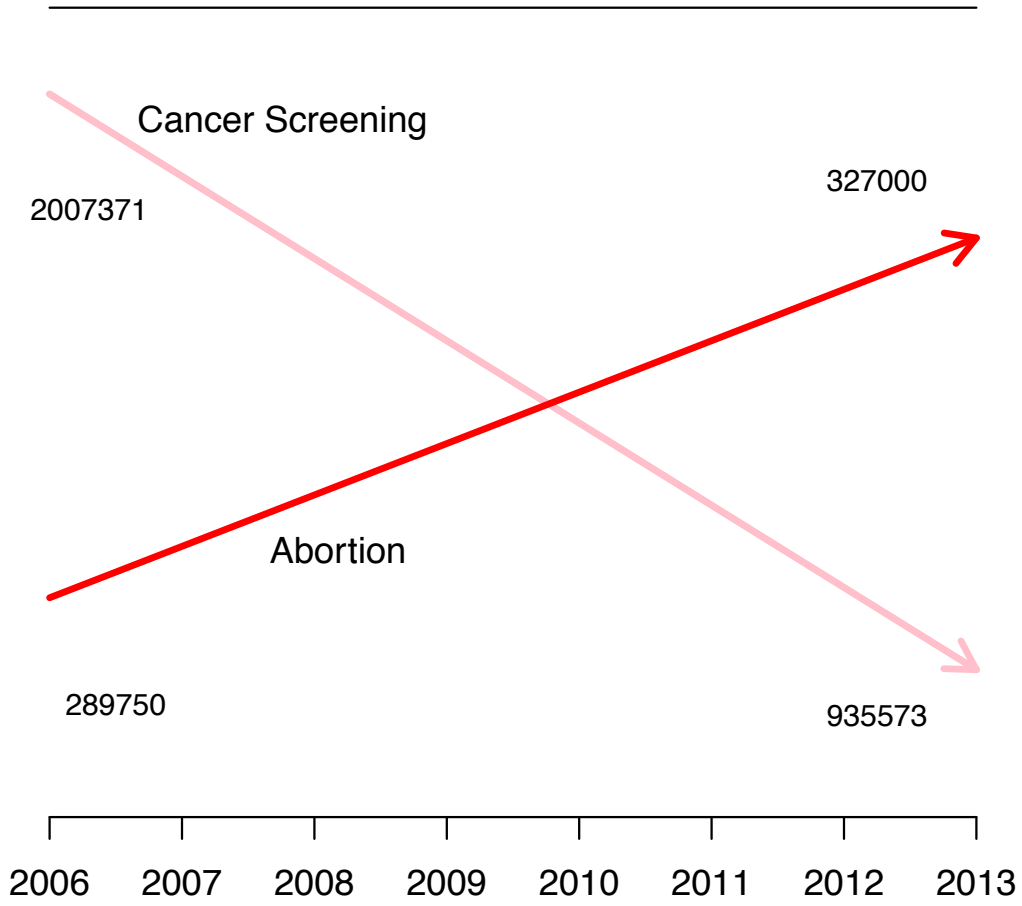
Grade in course –
Convert to grade points: Quan Discrete
Think of as letter grade: Qual Ordinal

Examples of Plots that Need Improvement

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.aul.org).
- Report available at: https://oversight.house.gov/interactivepage/plannedparenthood

Planned Parenthood Procedures

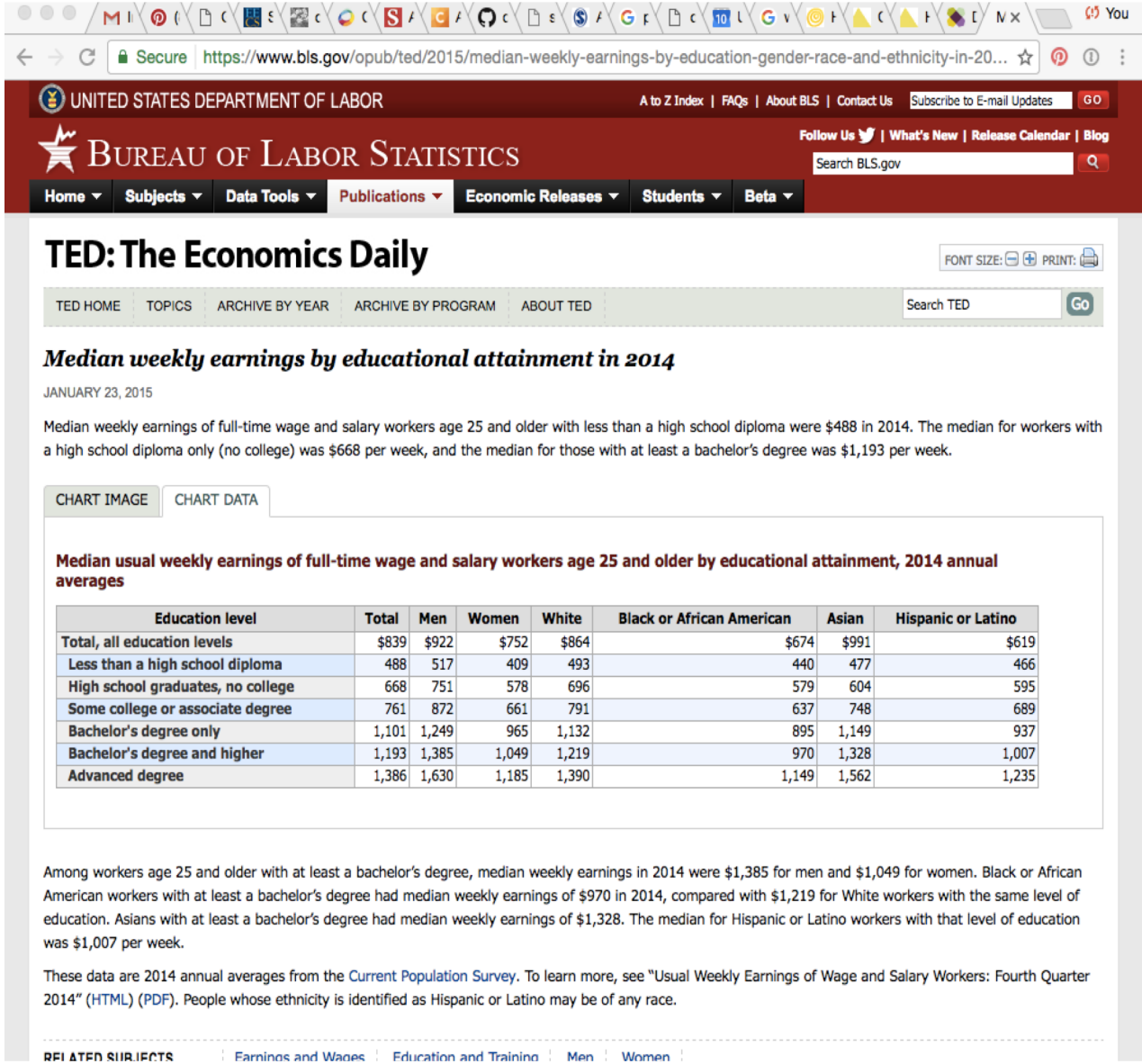


- Procedures:
  - Cancer screenings
  - Abortion
- Time: 2006 to 2013
- How many data points are in this plot?
- What's suspicious about this plot?

Note: This is an administrative dataset

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



Secure https://www.bls.gov/opub/ted/2015/median-weekly-earnings-by-education-gender-race-and-ethnicity-in-20...

UNITED STATES DEPARTMENT OF LABOR

A to Z Index | FAQs | About BLS | Contact Us | Subscribe to E-mail Updates | GO

BUREAU OF LABOR STATISTICS

Follow Us | What's New | Release Calendar | Blog

Search BLS.gov

Home | Subjects | Data Tools | Publications | Economic Releases | Students | Beta

TED: The Economics Daily

FONT SIZE: PRINT:

TED HOME | TOPICS | ARCHIVE BY YEAR | ARCHIVE BY PROGRAM | ABOUT TED | Search TED | Go

Median weekly earnings by educational attainment in 2014

JANUARY 23, 2015

Median weekly earnings of full-time wage and salary workers age 25 and older with less than a high school diploma were $488 in 2014. The median for workers with a high school diploma only (no college) was $668 per week, and the median for those with at least a bachelor's degree was $1,193 per week.

CHART IMAGE | CHART DATA

Median usual weekly earnings of full-time wage and salary workers age 25 and older by educational attainment, 2014 annual averages

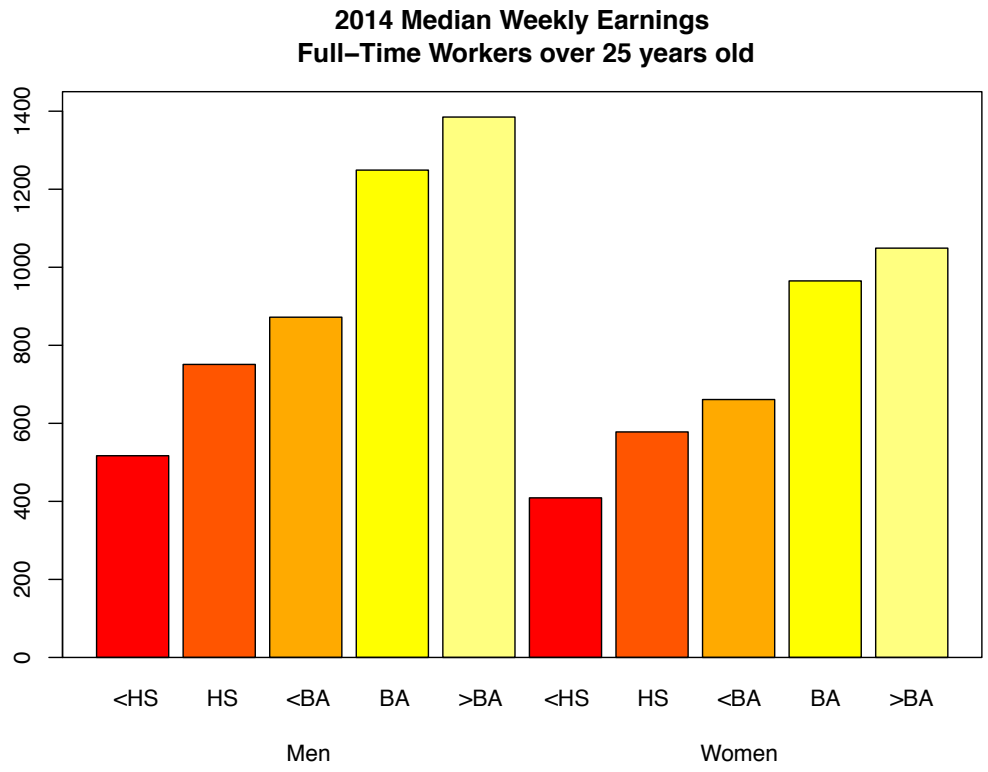
Education level	Total	Men	Women	White	Black or African American	Asian	Hispanic or Latino
Total, all education levels	$839	$922	$752	$864	$674	$991	$619
Less than a high school diploma	488	517	409	493	440	477	466
High school graduates, no college	668	751	578	696	579	604	595
Some college or associate degree	761	872	661	791	637	748	689
Bachelor's degree only	1,101	1,249	965	1,132	895	1,149	937
Bachelor's degree and higher	1,193	1,385	1,049	1,219	970	1,328	1,007
Advanced degree	1,386	1,630	1,185	1,390	1,149	1,562	1,235

Among workers age 25 and older with at least a bachelor's degree, median weekly earnings in 2014 were $1,385 for men and $1,049 for women. Black or African American workers with at least a bachelor's degree had median weekly earnings of $970 in 2014, compared with $1,219 for White workers with the same level of education. Asians with at least a bachelor's degree had median weekly earnings of $1,328. The median for Hispanic or Latino workers with that level of education was $1,007 per week.

These data are 2014 annual averages from the Current Population Survey. To learn more, see "Usual Weekly Earnings of Wage and Salary Workers: Fourth Quarter 2014" (HTML) (PDF). People whose ethnicity is identified as Hispanic or Latino may be of any race.

RELATED SUBJECTS | Earnings and Wages | Education and Training | Men | Women

Earnings



Which comparisons can be easily made with this plot?

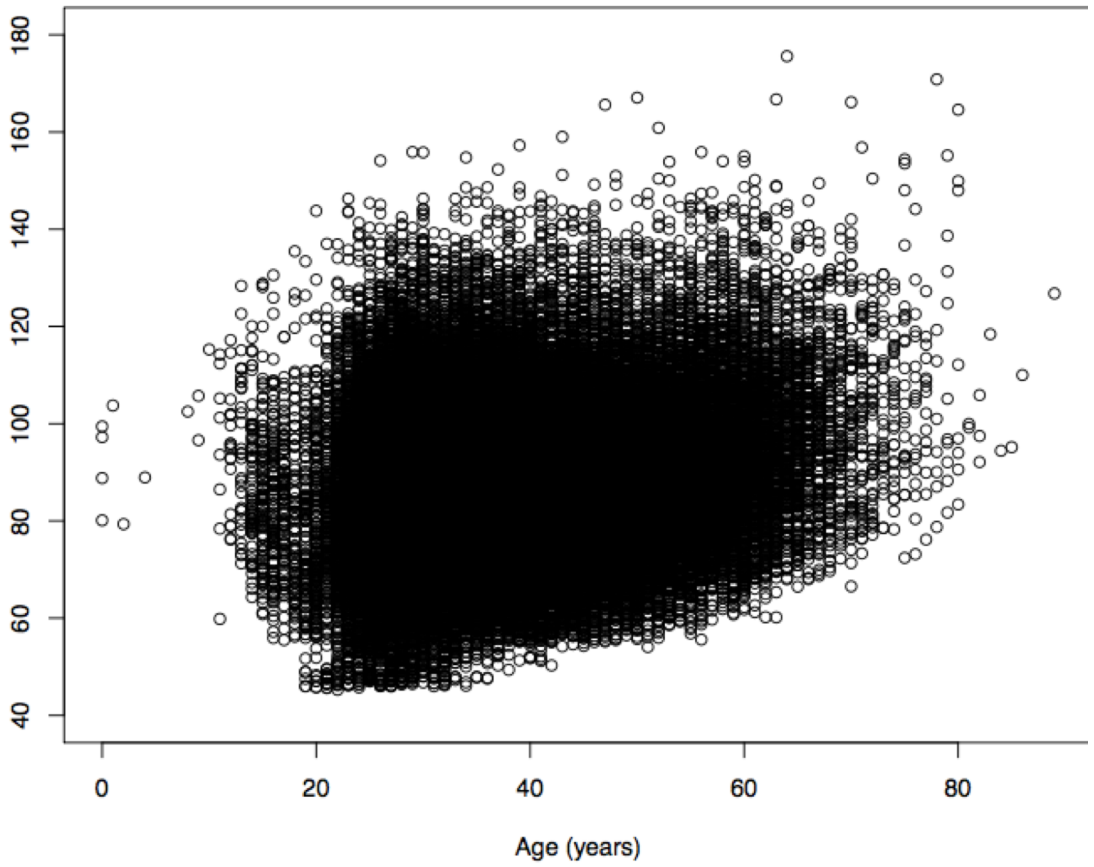
What comparisons are most interesting or important?

Note: This is a probability sample (AKA scientific sample)

Cherry Blossom Run

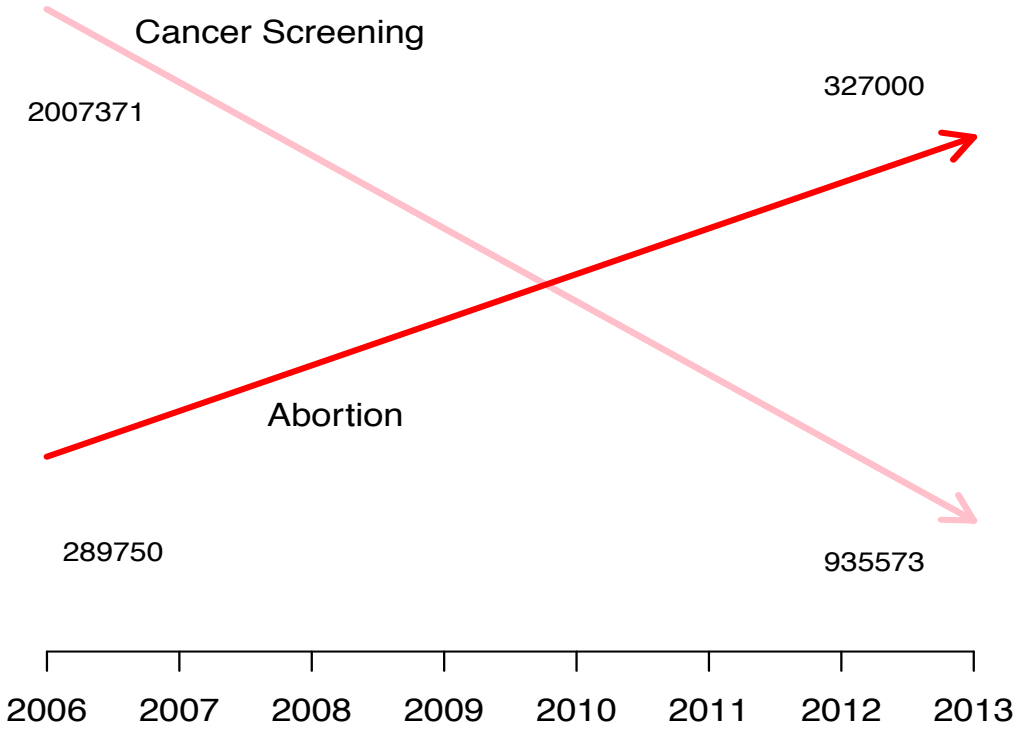
- 10 mile run in Washington DC each April
- Race organizers make results available on Web
  - Runner name, age, gender, address, hometown, time
  - Race results from 1999 to 2016
  - In 2012 nearly 17,000 runners ranging in age from 9 to 89 participated
  - http://www.cherryblossom.org/

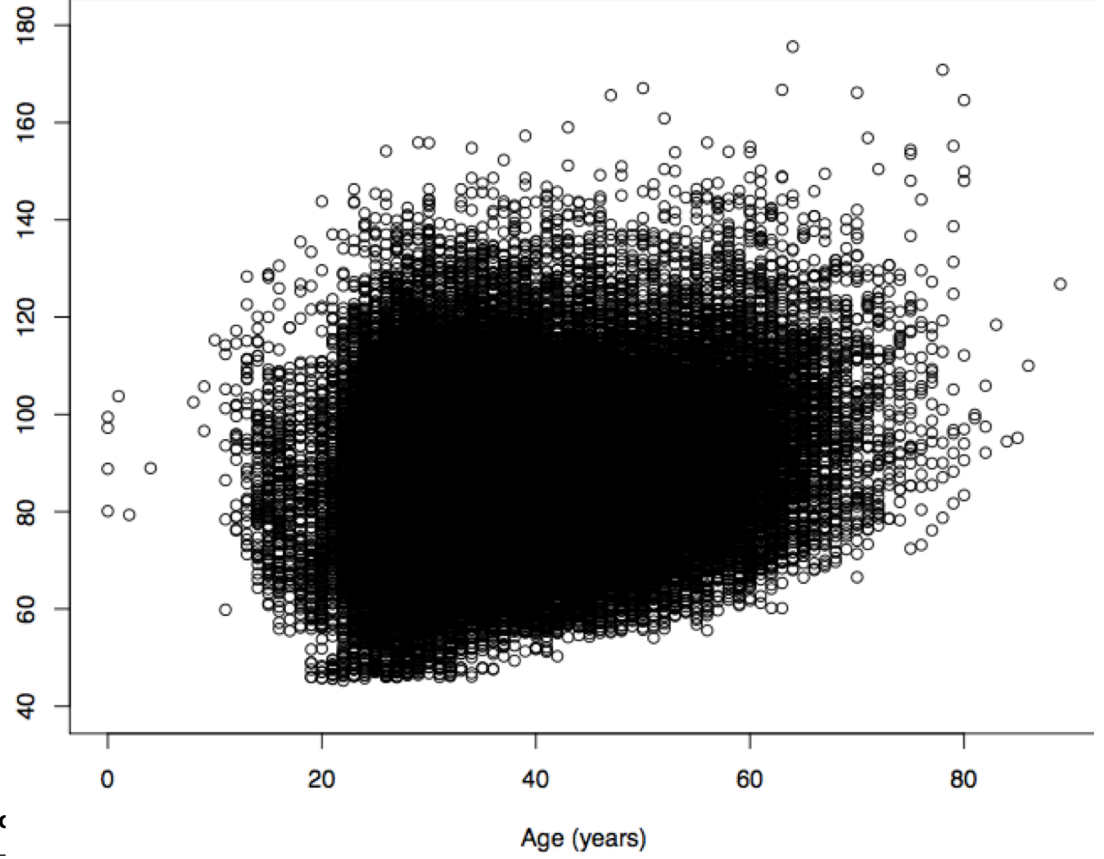
Cherry Blossom Run

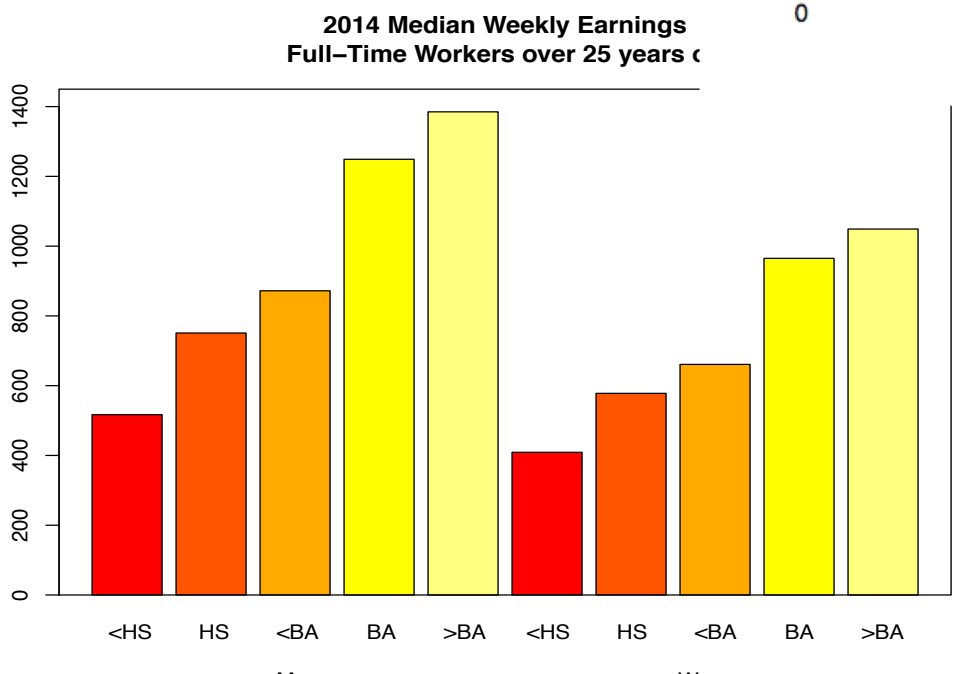


- Scatter plot of run time (minutes) by age (years)
- 70,000 points in this plot.
- What's the relationship between run time and age?

Note: This is a self-selected sample


Cancer Screening
2007371
327000
Abortion
289750
935573
2006
2007
2008
2009
2010
2011
2012
2013


40
60
80
100
120
140
160
180
0
20
40
60
80
Age (years)


2014 Median Weekly Earnings
Full–Time Workers over 25 years
0
200
400
600
800
1000
1200
1400
<HS
HS
<BA
BA
>BA
<HS
HS
<BA
BA
>BA

Techniques

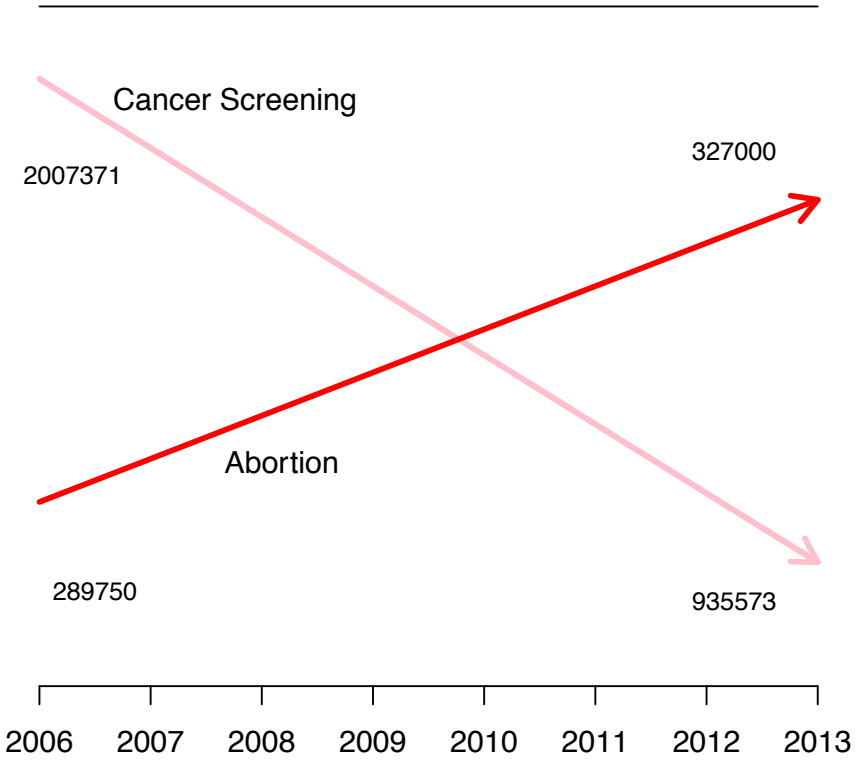
For Making Plots Informative and Effective

Scale	Conditioning	Perception	Transformation	Context	Smoothing & Reducing	Philosophy

Goals of this lecture

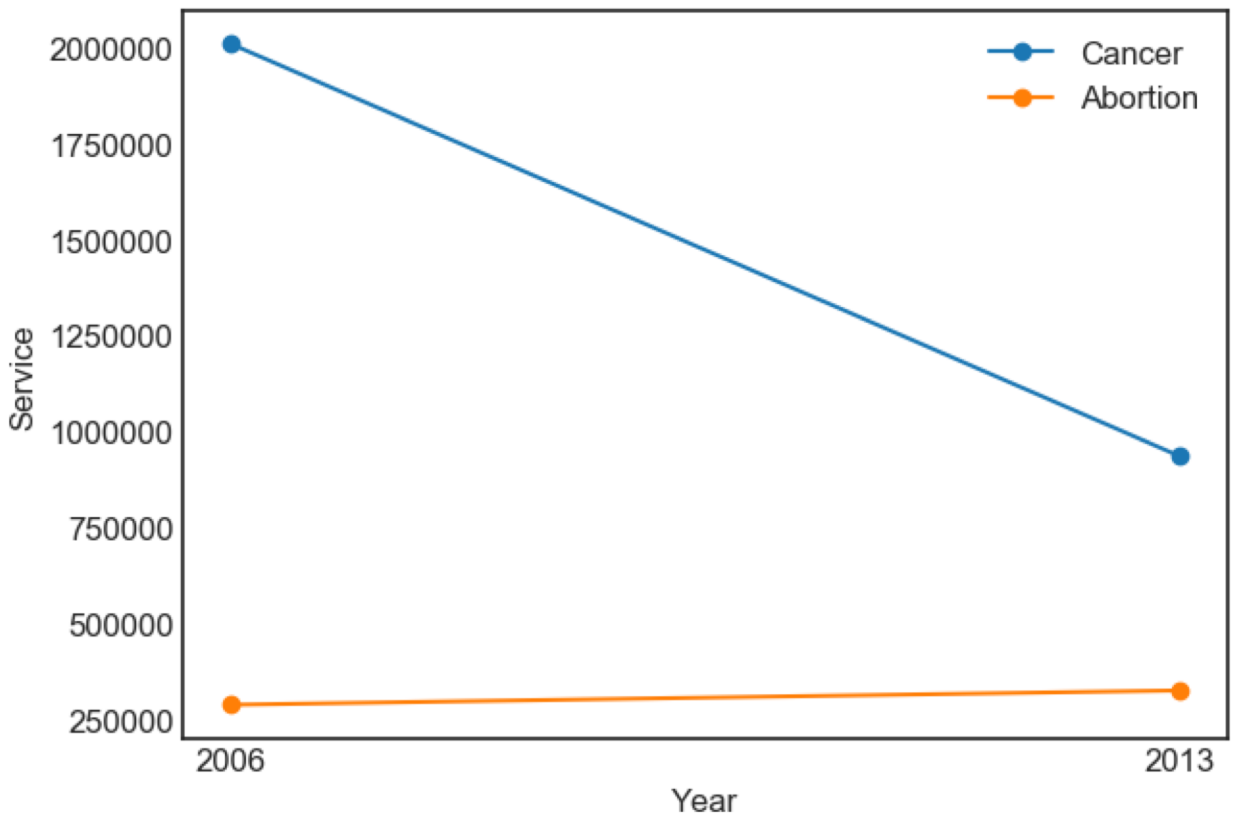
- Guidelines and general philosophy
  - Reveal the data
  - Facilitate Comparisons
  - Add information
- Techniques for following guidelines
  - Scale
  - Conditioning/faceting
  - Perception: color
  - Transformations
  - Adding context
  - Smoothing & other large data considerations

Planned Parenthood Procedures



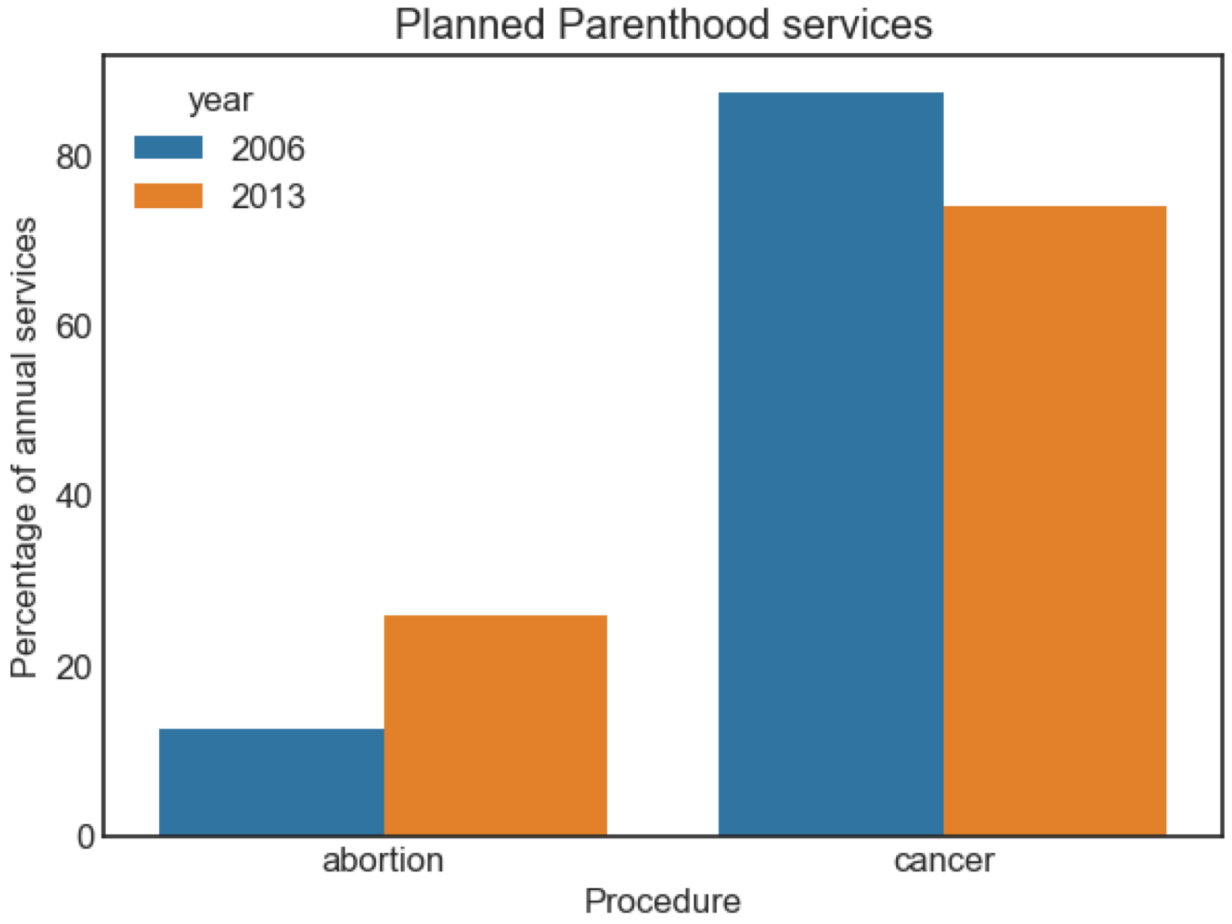
Scale

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

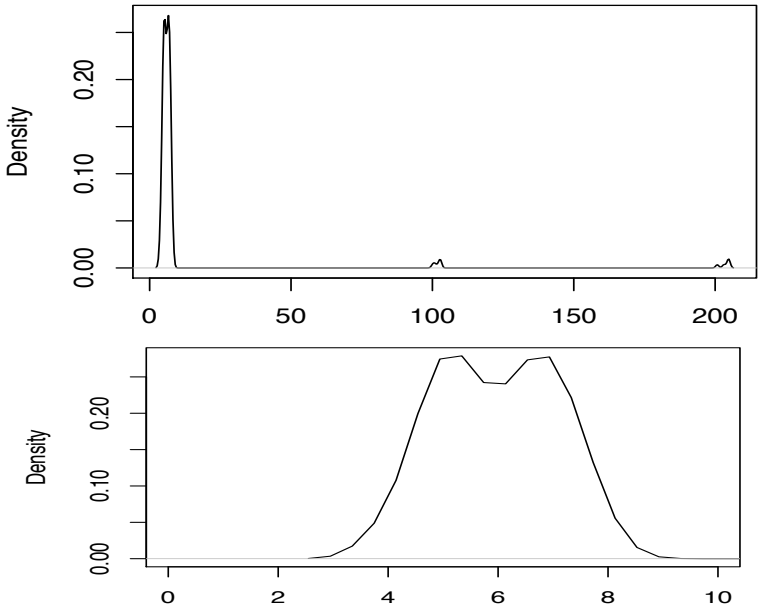
Planned Parenthood Procedures



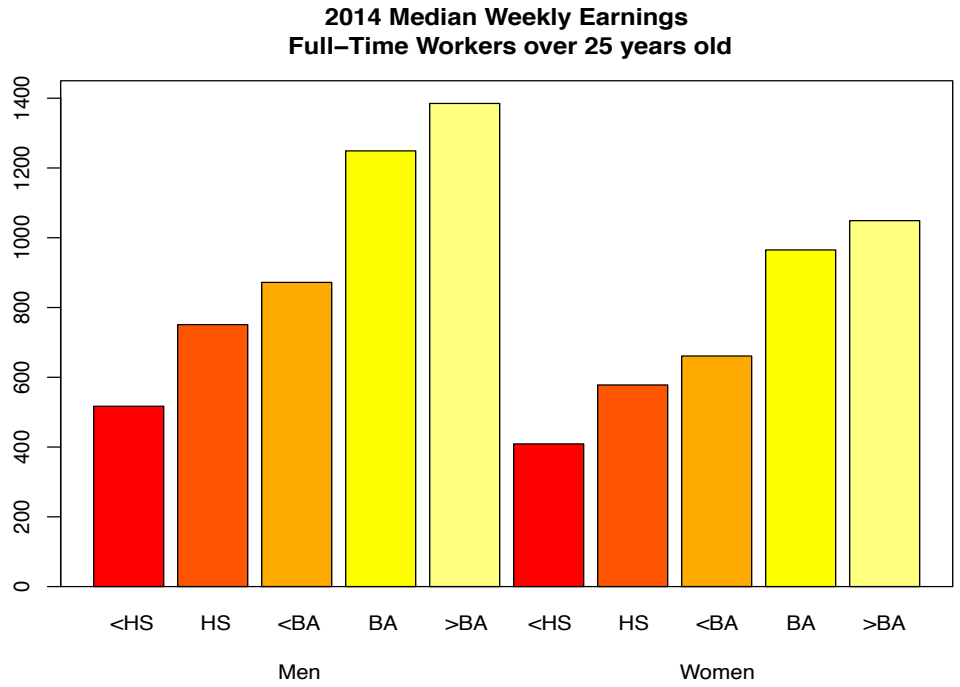
- Procedures in 2006 and 2013 as a percentage
- Abortions increased from 13% to 26% of total procedures
- May want to plot the percent change, screenings fell 50%

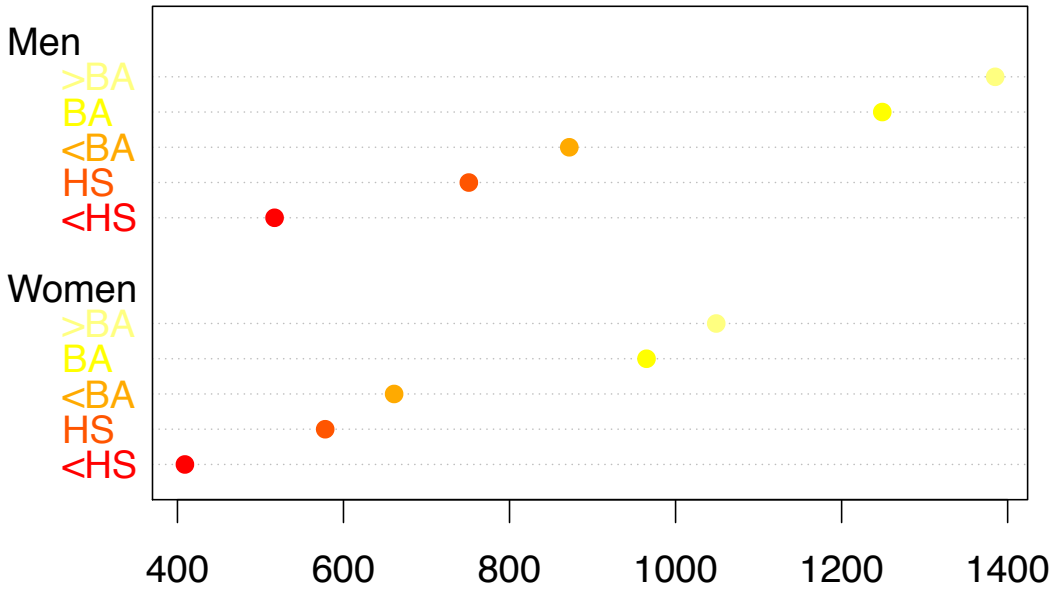
Choosing the Scale

- Choose axis limit to fill the plotting region
- If 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



Earnings

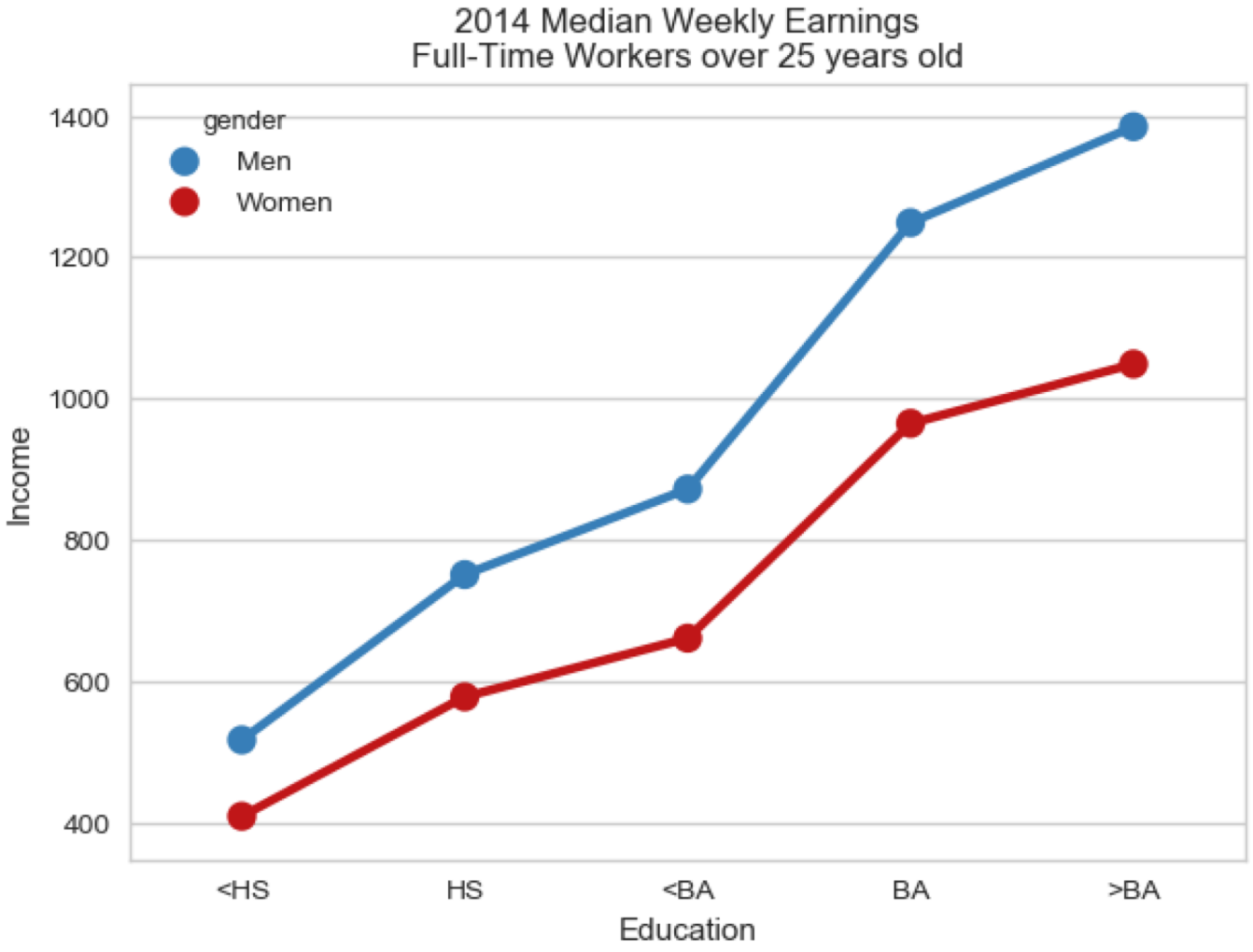

2014 Median Weekly Earnings
Full–Time Workers over 25 years old
0
200
400
600
800
1000
1200
1400
<HS
HS
<BA
BA
>BA
<HS
HS
<BA
BA
>BA
Men
Women


Men
>BA
BA
<BA
HS
<HS
Women
>BA
BA
<BA
HS
<HS
400
600
800
1000
1200
1400

Conditioning

Compare Distributions Across Sub-groups

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

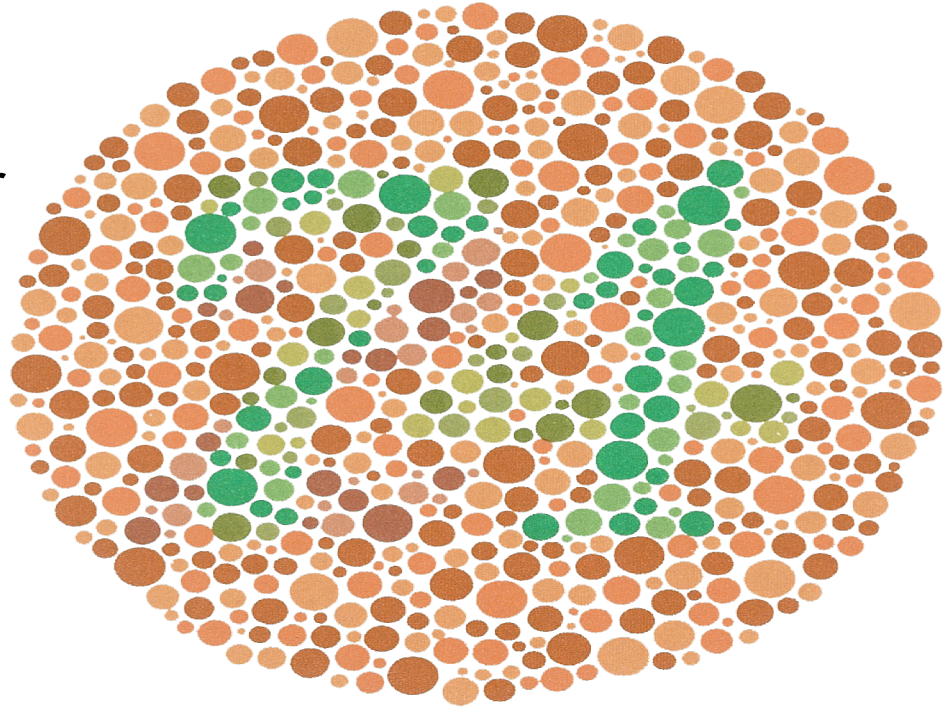
Conditioning – Distributions & Relationships in subgroups

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

Perception - Color

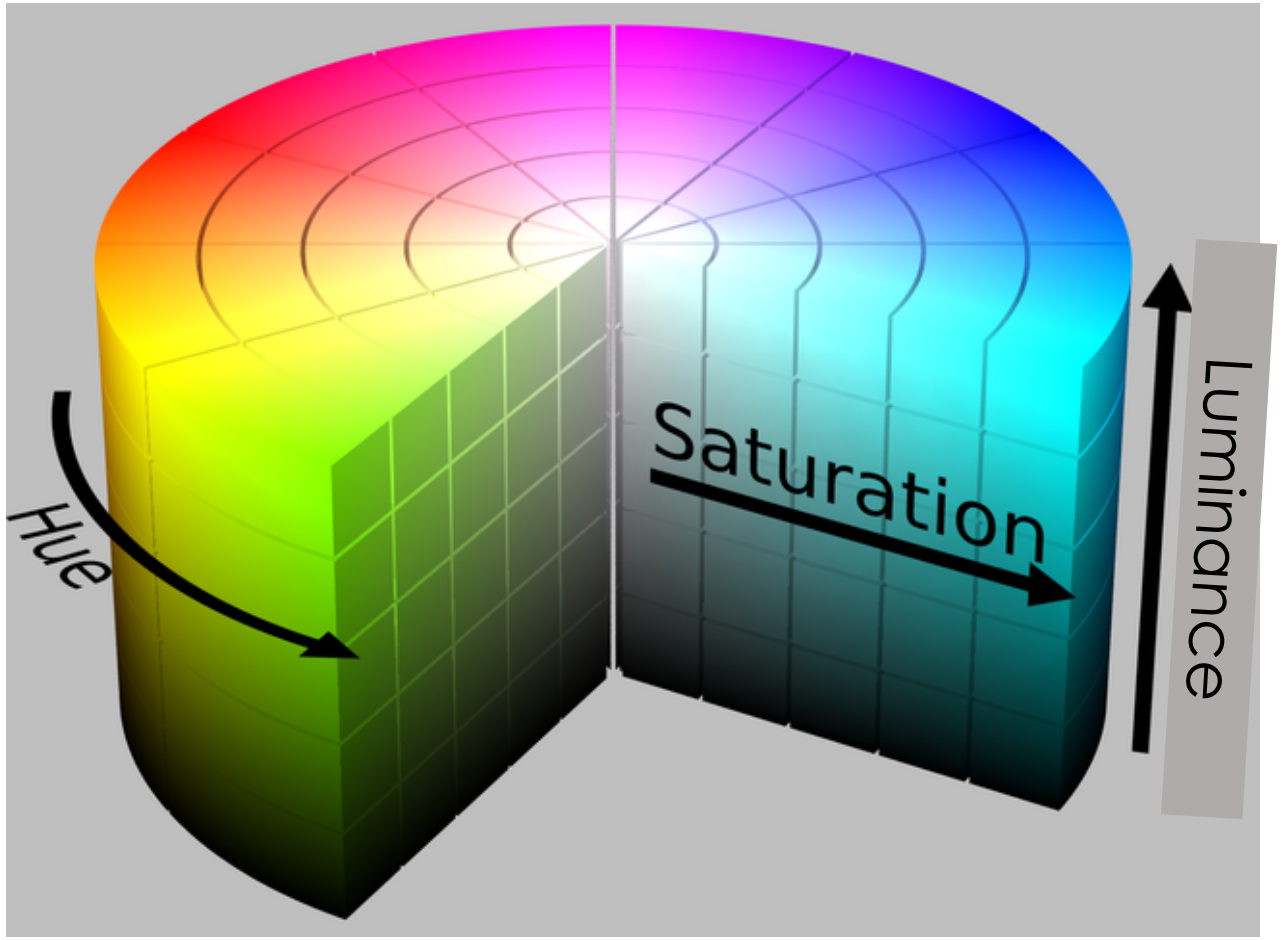
Color Guidelines

- 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.



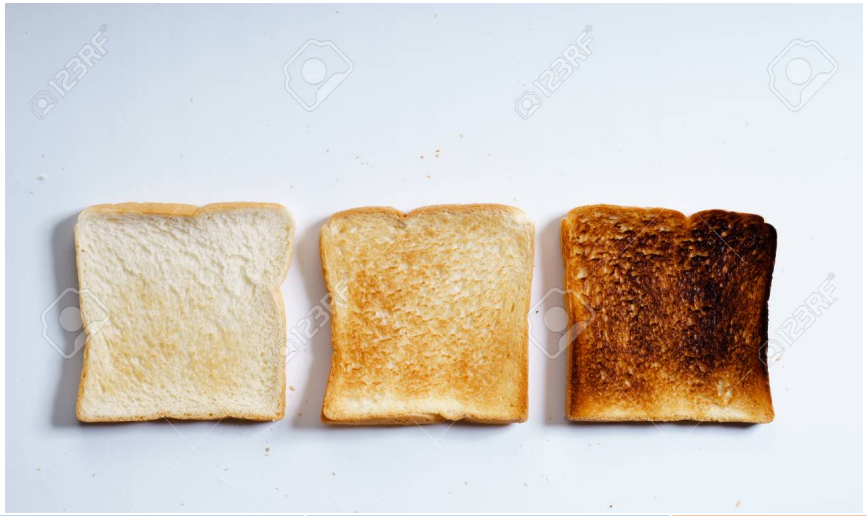
Hue – Saturation – Luminance

- This 3-d color cylinder helps describe the HSL components.



Hue – Unripe and Ripe bananas

Saturation – Chocolate milk strength



Luminance – White bread -> lightly toasted -> burnt

Hue

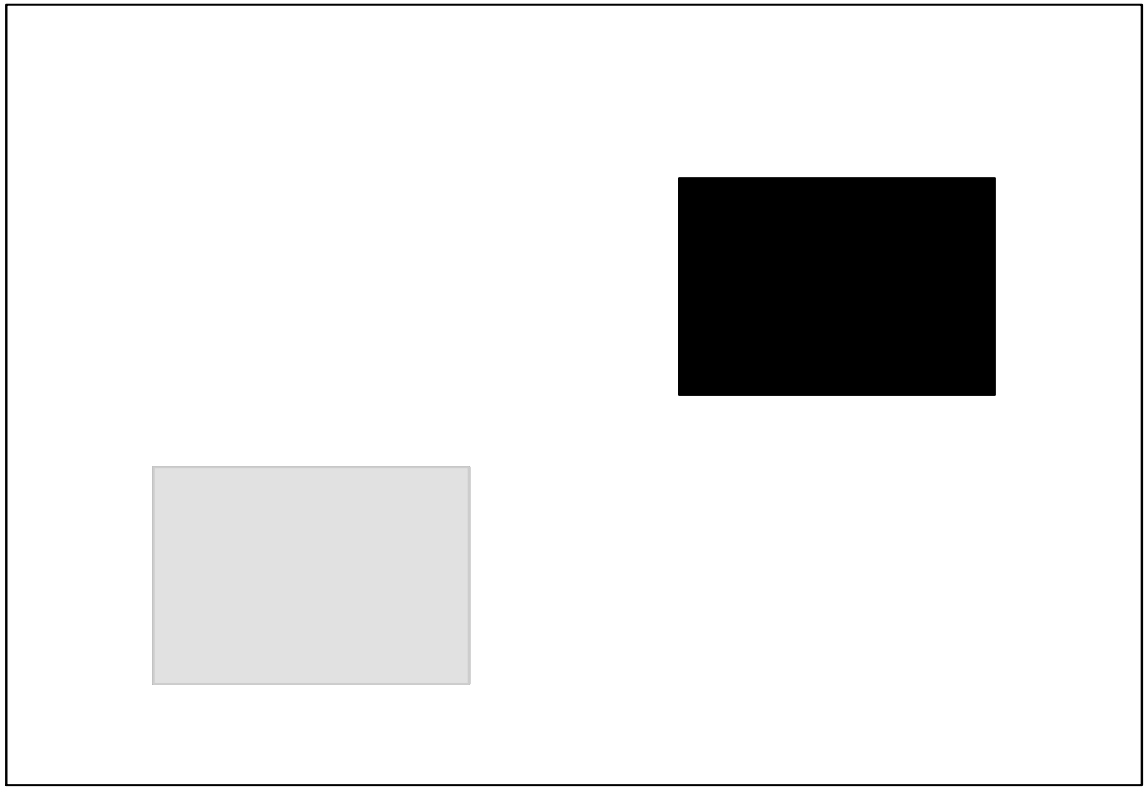
- We have difficulty in distinguishing between more than
- We also have trouble detecting combinations of colors accurately, e.g., 50% red and 50% blue looks more red to us than blue.

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.

Luminance

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

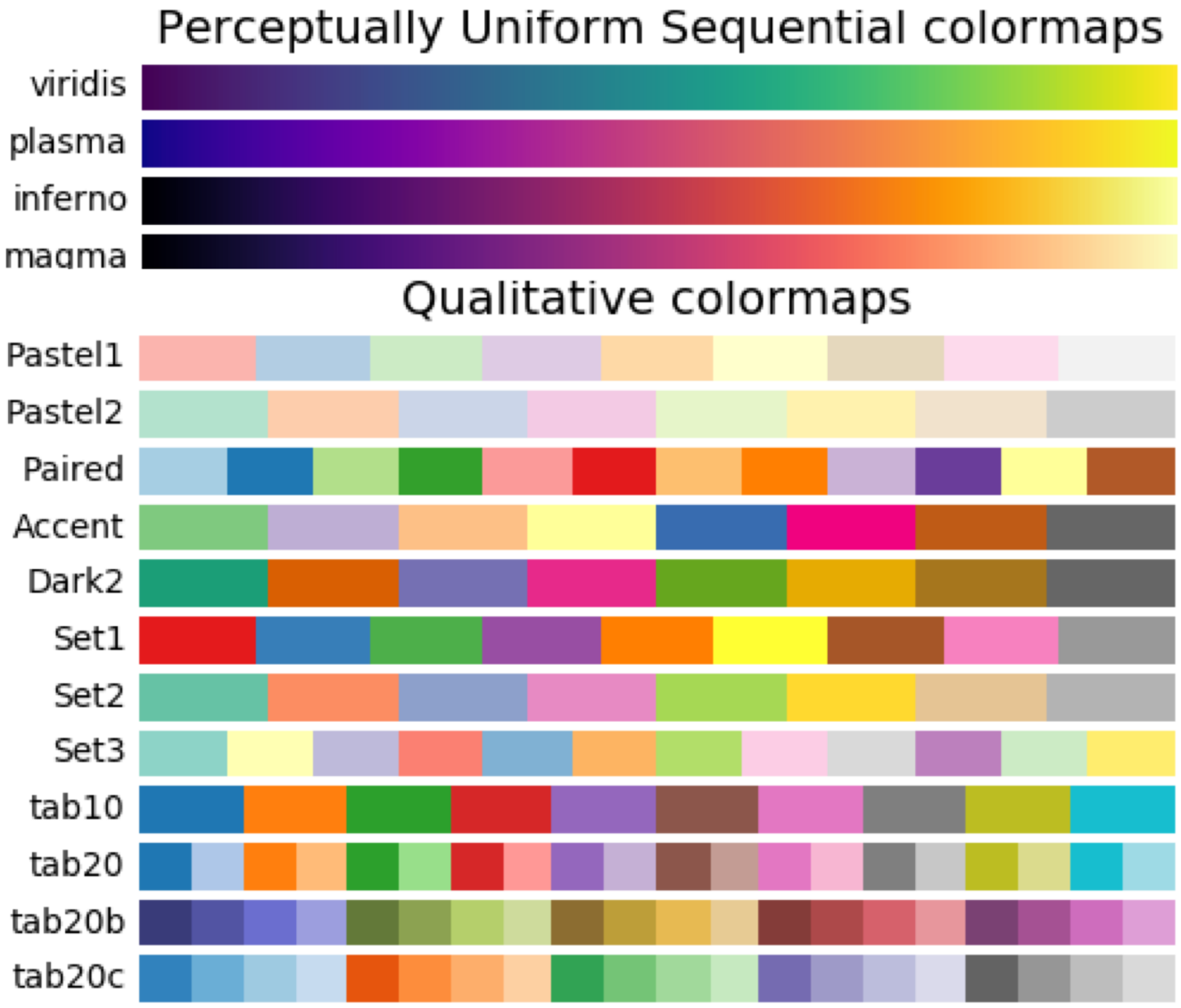


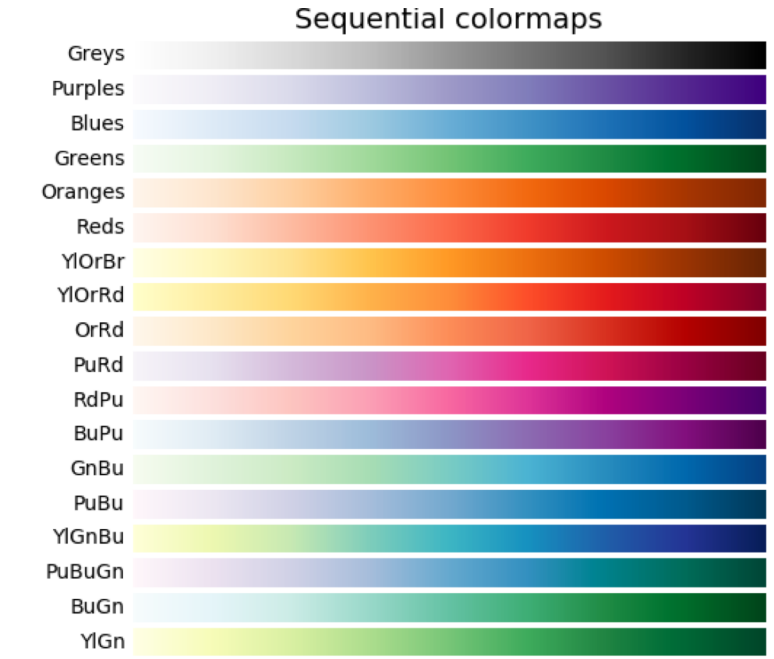
Data Type and Color

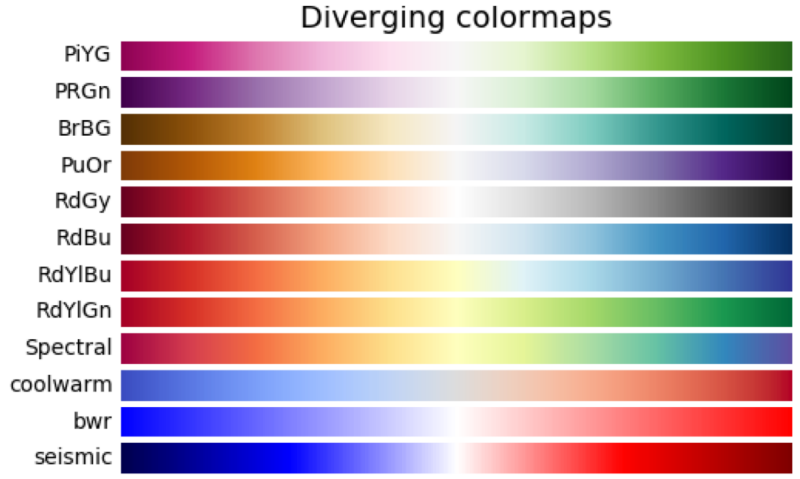
- Qualitative – Choose a **qualitative** scheme that makes it easy to distinguish between categories
- Quantitative – Choose a color scheme that implies magnitude.
  - Does the data progress from low to high? Use a **sequential** scheme where light colors are for low values
  - Do both low and high value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

Colormaps

https://matplotlib.org/examples/color/colormaps_reference.html







Perception Ranking

Based on Experiments by Cleveland and McGill

Most to Least Accurately Judged

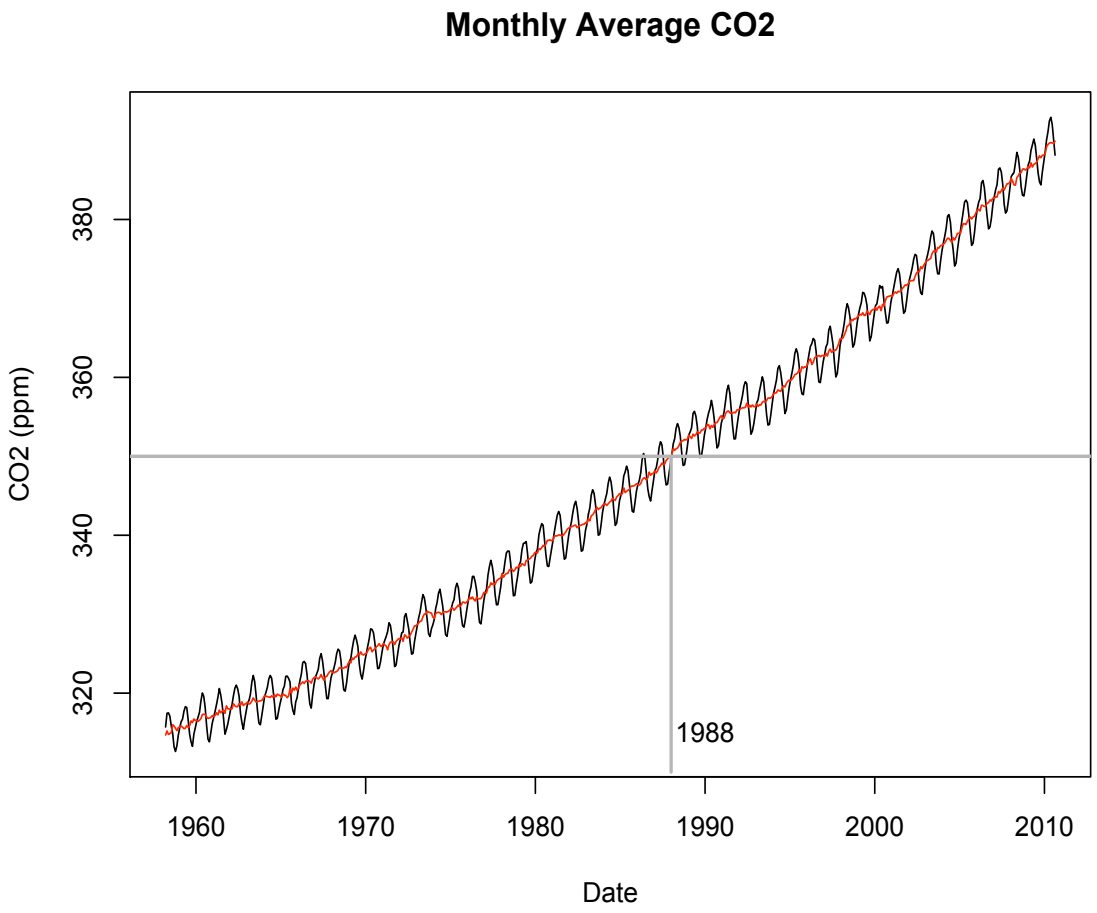
1. Positions along a common scale
2. Positions on identical, nonaligned scales
3. Length
4. Angle, slope
5. Area
6. Volume, density, color saturation
7. Color Hue

1. Strip plot, rug plot, dot plot
2. Bar plot
3. Segmented bar plot
4. Pie chart
5. Bubble chart
6. Scatter plot with too many points, 3-d bar plots
7. Purple election maps

Atmospheric Carbon Dioxide

- The increasing amount of CO_2 in the atmosphere from the burning of fossil fuels has become a serious environmental concern.
- Upper safety limit for atmospheric CO_2 is 350 parts per million
- Does a rise in CO_2 lead to a rise in world temperatures?

Aspect Ratio and Banking to 45°

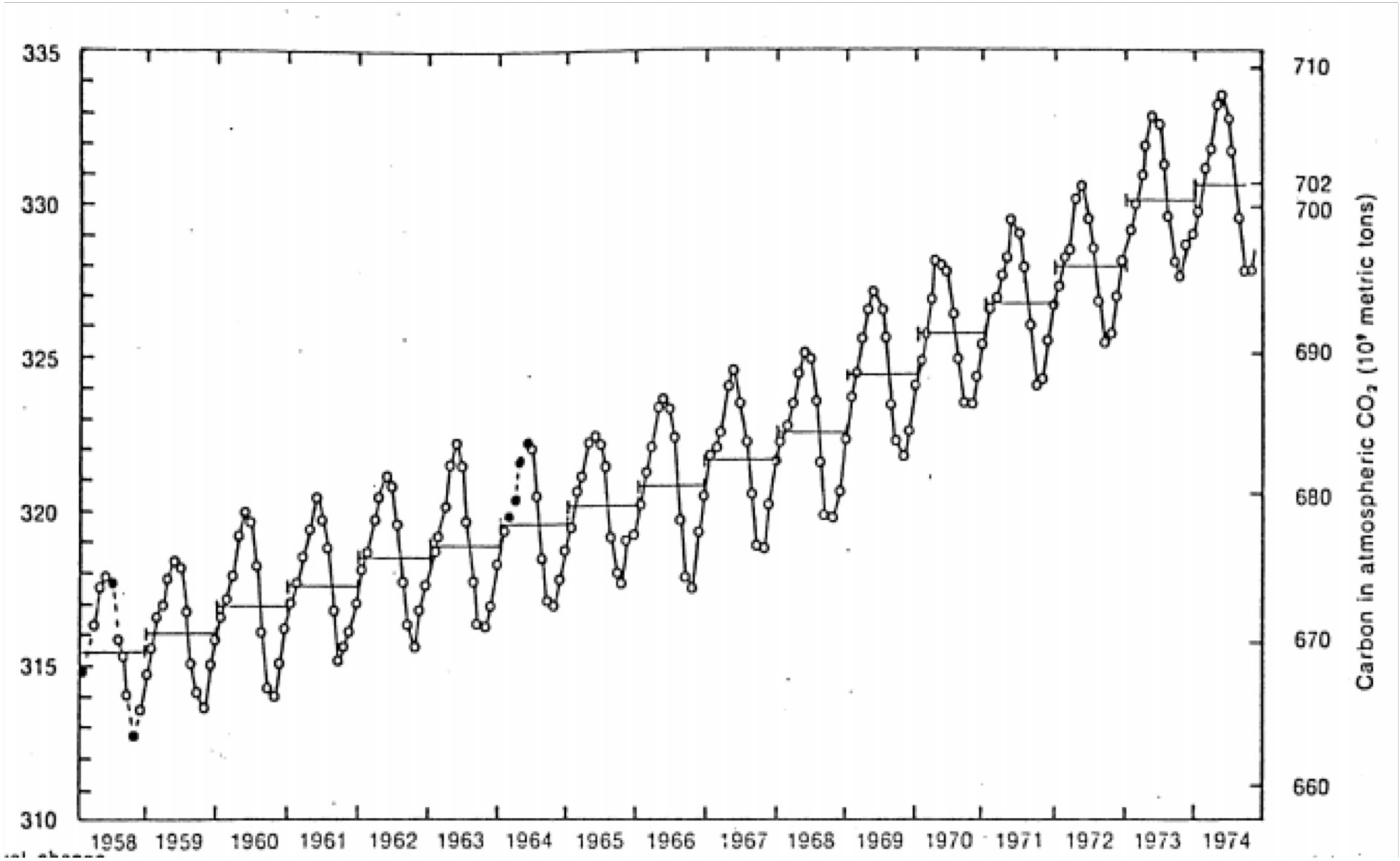


Aspect Ratio

- The height/width of the data region was selected to be about 1 so that the trend line is at about 45 degrees.
- The banking to 45 degrees let's us see that the curve is convex
- This means that the rate of increase of CO_2 is increasing through time

Global Warming

- 1981, Gore organized the first Congressional hearing on global warming
- Gore said that the Mauna Loa data clearly demonstrated increases in CO_2
- Pewitt (witness for the DOE) said that the graph was misleading because it doesn't include 0


335
330
325
320
315
310
710
702
700
690
680
670
660
Carbon in atmospheric CO_2 (10^9 metric tons)
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974

Chartology

Pewitt (p 90) took issue with the graph, saying

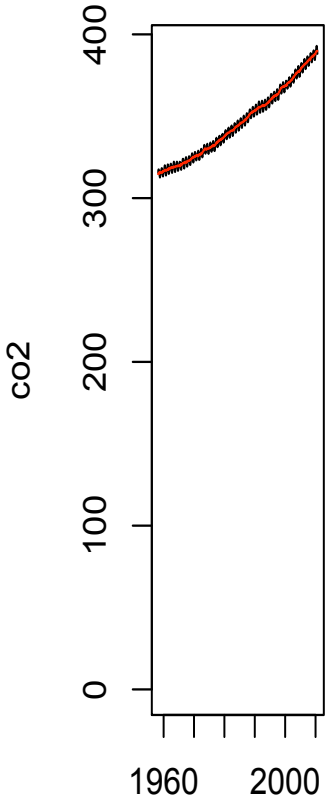
"It is a clever piece of chartology" because it can be read the wrong way.

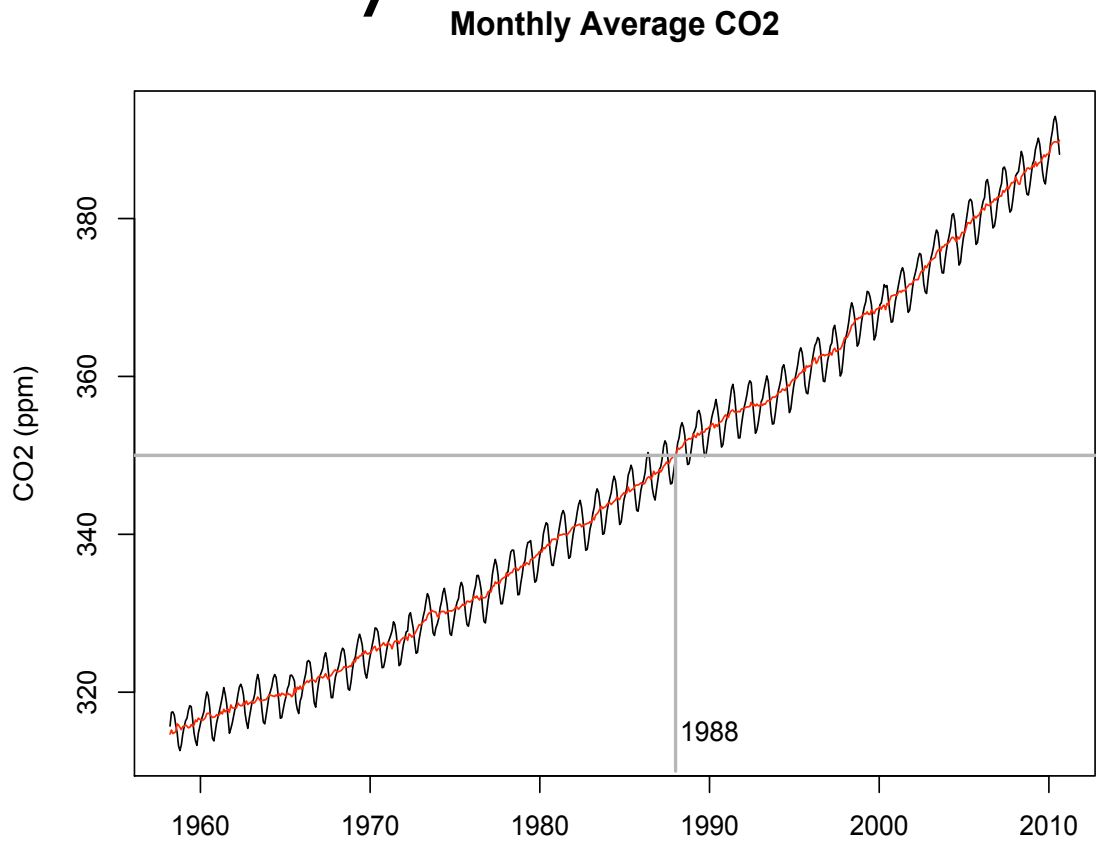
He continued (p96), "It is intellectually just exactly correct. It displays 315 going to 336, but it appears to be going from 0 to very large amounts."

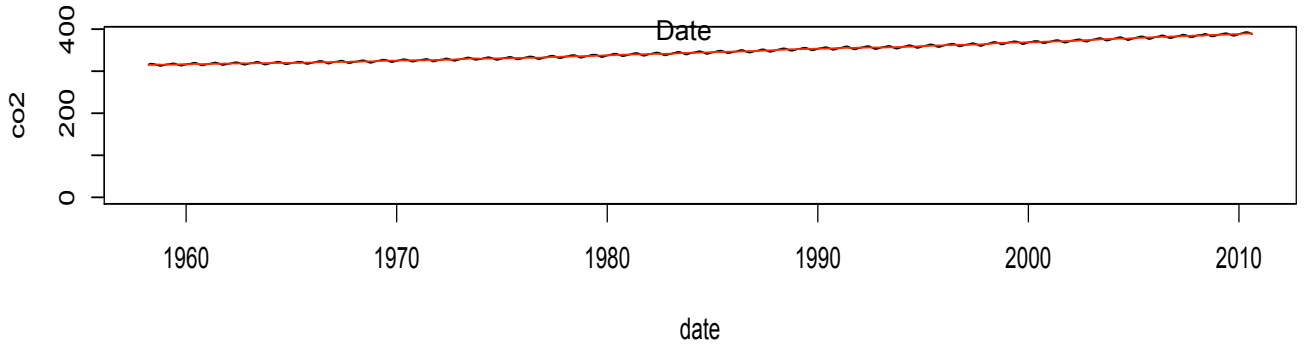
Steven Schneider (*Global Warming*) called Pewitt's objection "double talk"

Including 0 on the y-axis

Maintain the aspect ration







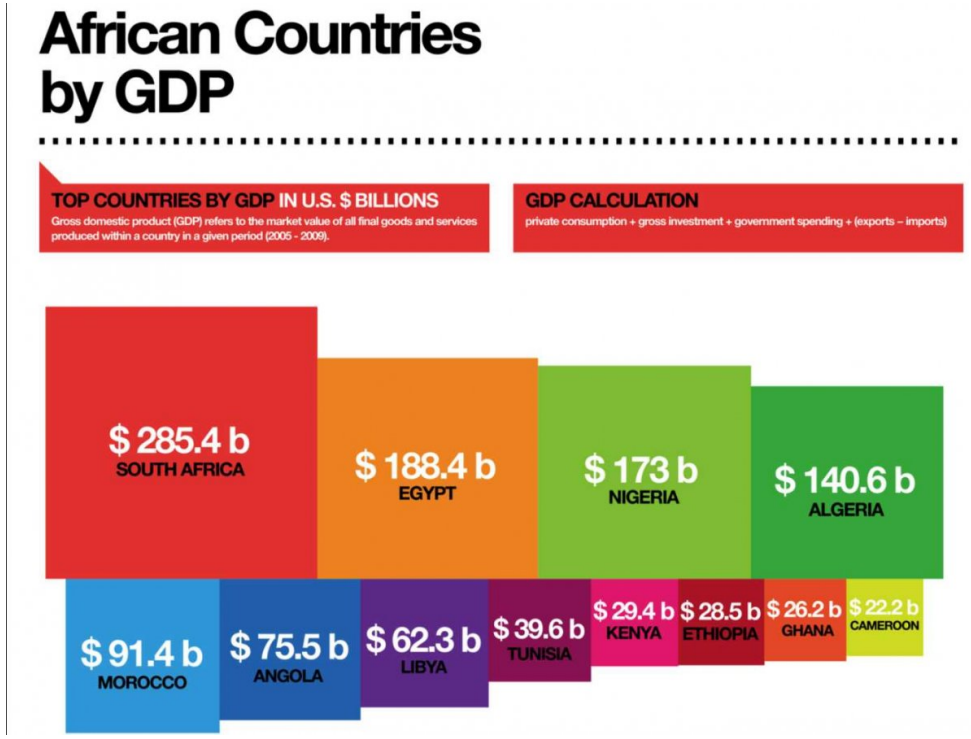
Fill the data region

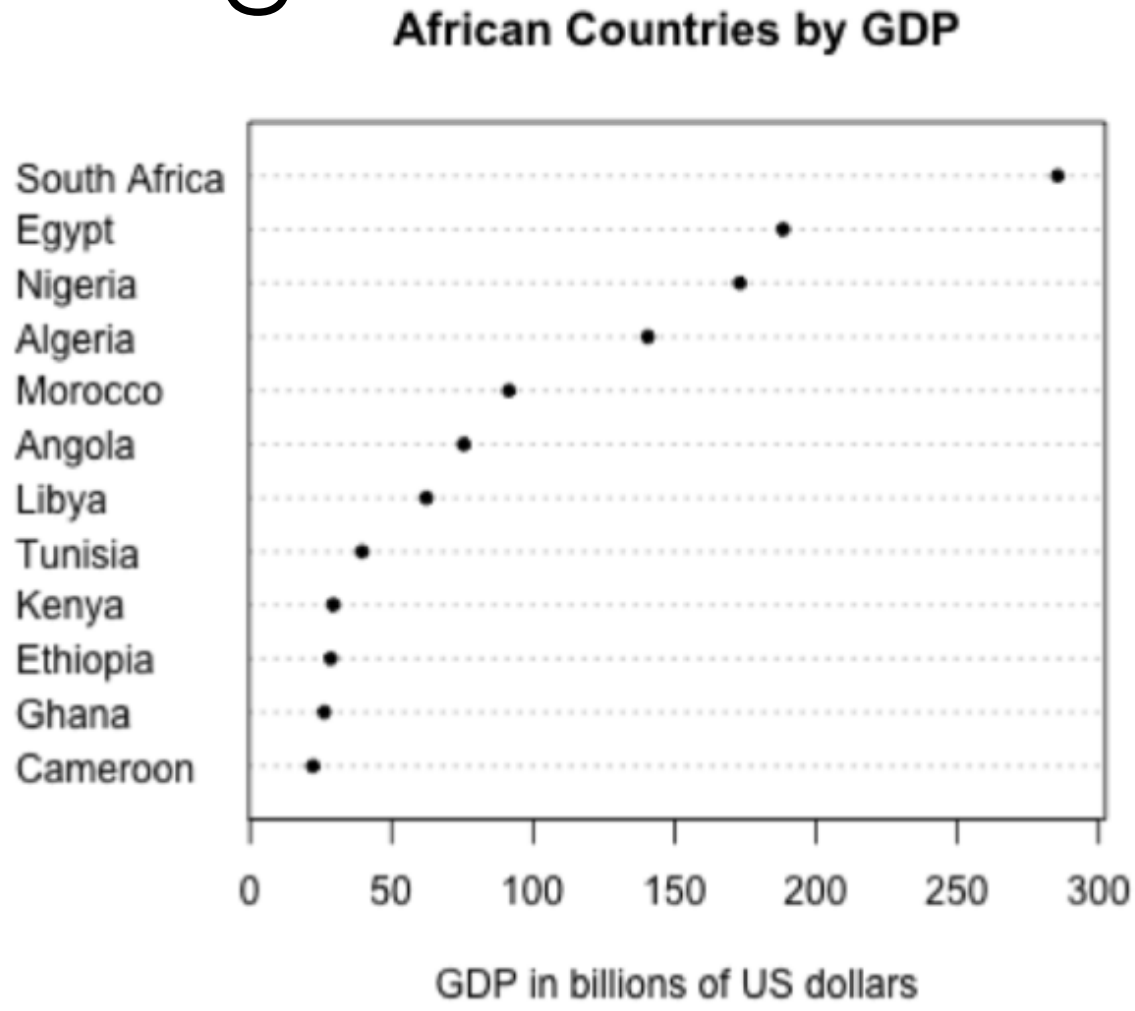
Including 0 on the y-axis

- When we include 0 and bank at 45 degrees, then the plot must be tall and narrow.
  - With this plot it's hard to see any of the features.
  - There is also a lot of empty space.

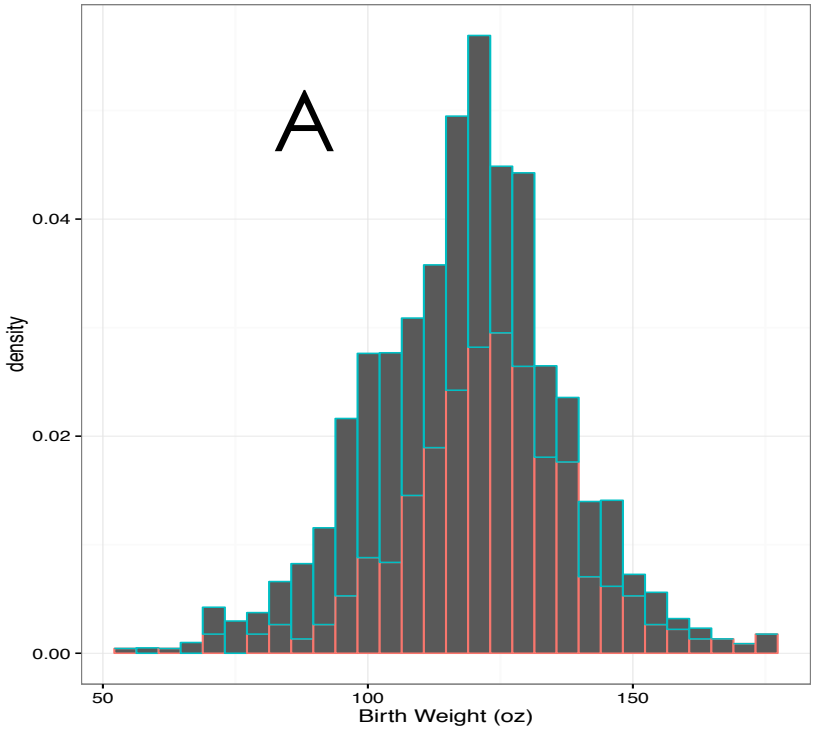
- To fill the space with data (and include 0), we need to stretch the data region to be wide and short.
  - Then, it's hard to see the most important feature, the curvature, because the banking is near 0.

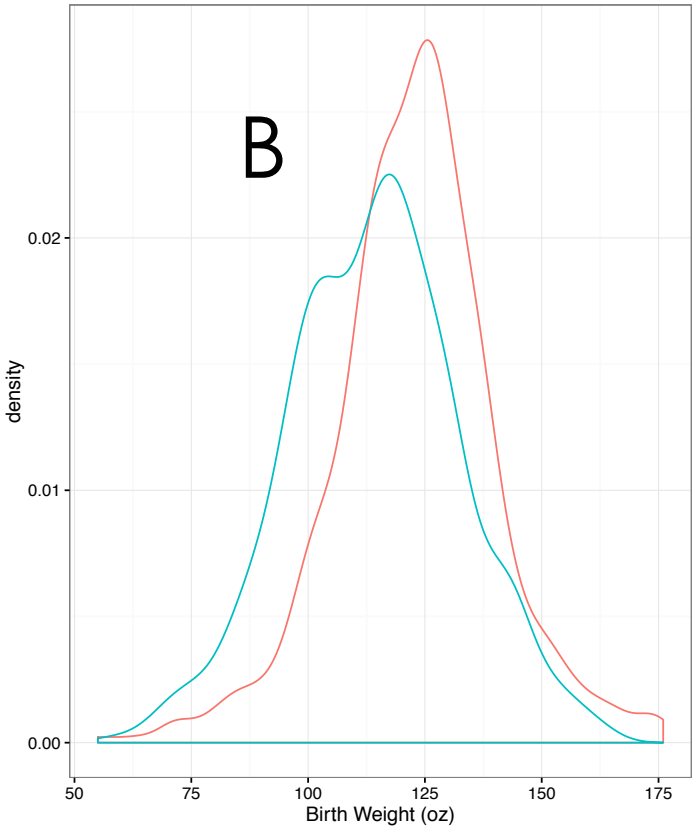
Area and Poor positioning

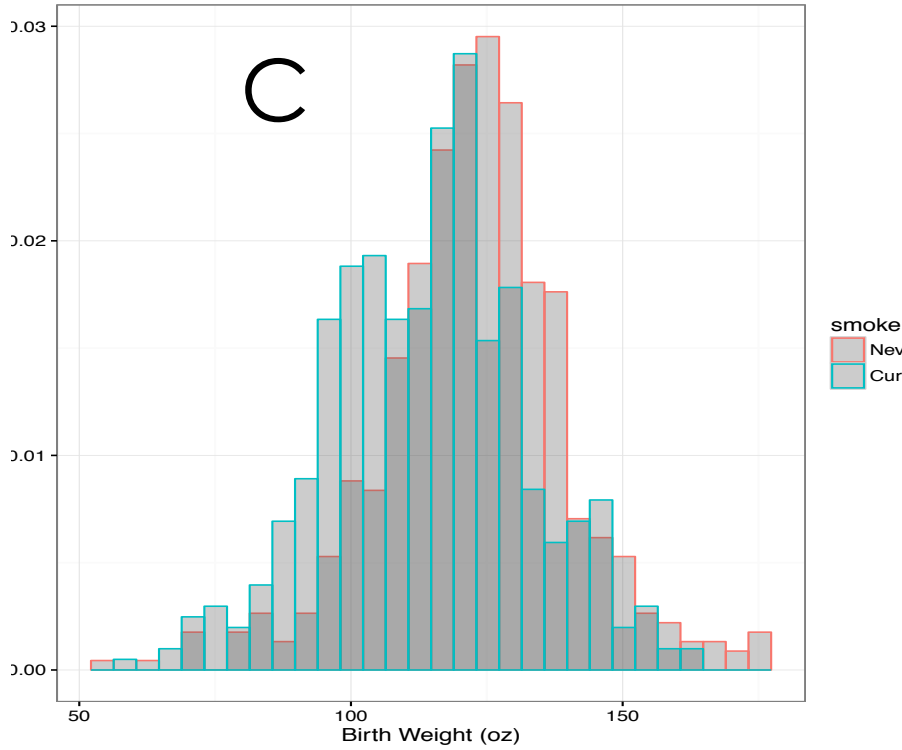




Which of these plots is easiest to read?



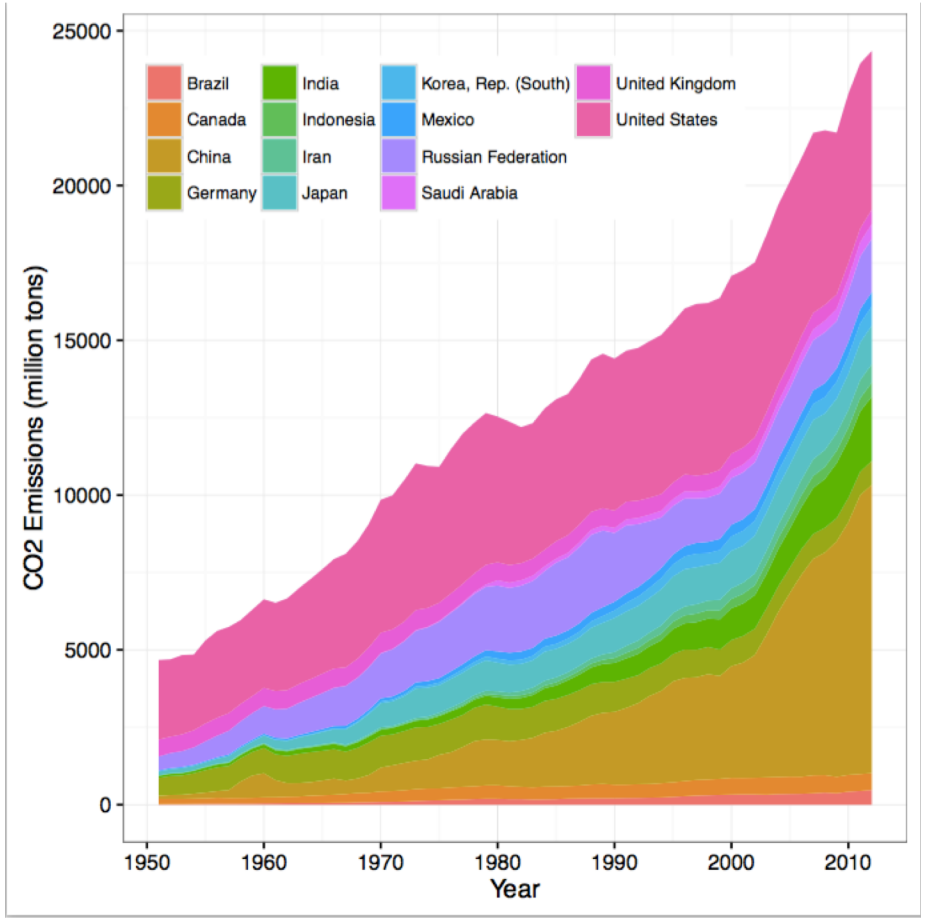


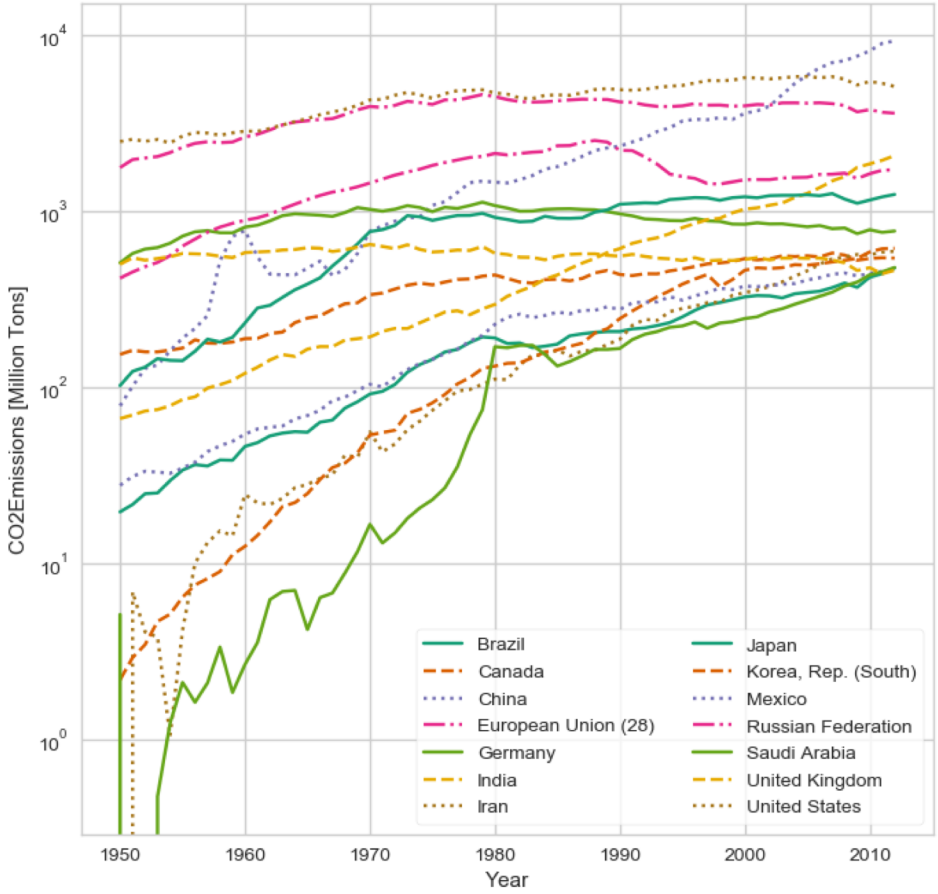


CO2 Emissions from Fuel Consumption

- 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/)

CO2 Emissions





Scale	Conditioning	Perception	Transformation	Context	Smoothing	Philosophy

Transformations

Why Transform Variables?

- Reveal distribution of most of the observations (otherwise much of the data is squashed in a small region)
- Reveal anomalies on the "other side" of the data
- Numerical summaries of transformed data are better summaries of a symmetric distribution

Log transformation: Swiss army knife

$$y = a^x \rightarrow \log(y) = x \log(a)$$

$$y = ax^k \rightarrow \log(y) = \log(a) + k \log(x)$$

$$e^y = bx \rightarrow y = \log(b) + \log(x)$$

Power Transformation

Effective when max/min > 5

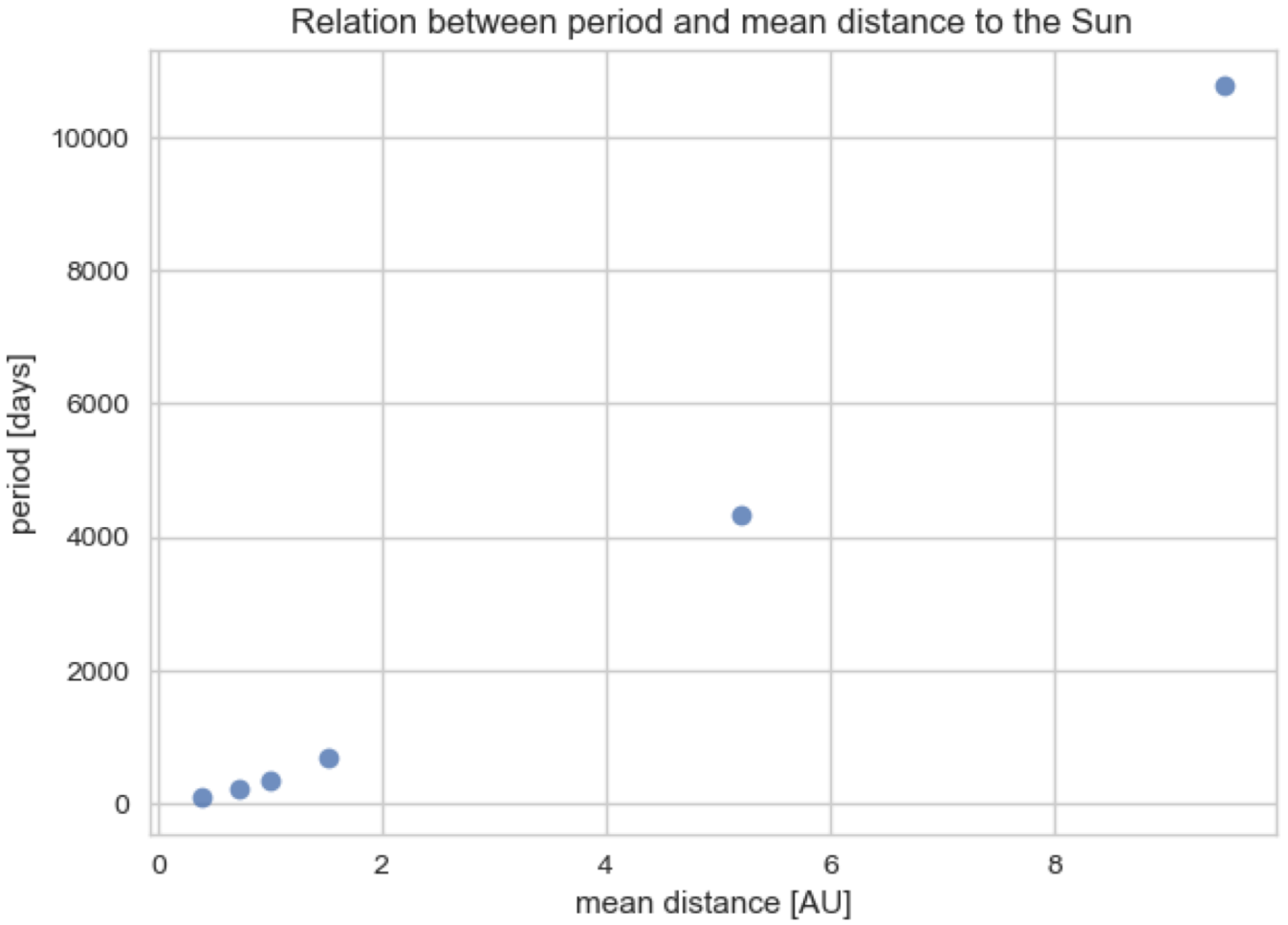
Sometimes add a shift before transform

Ratio of hinges can help select a transformation

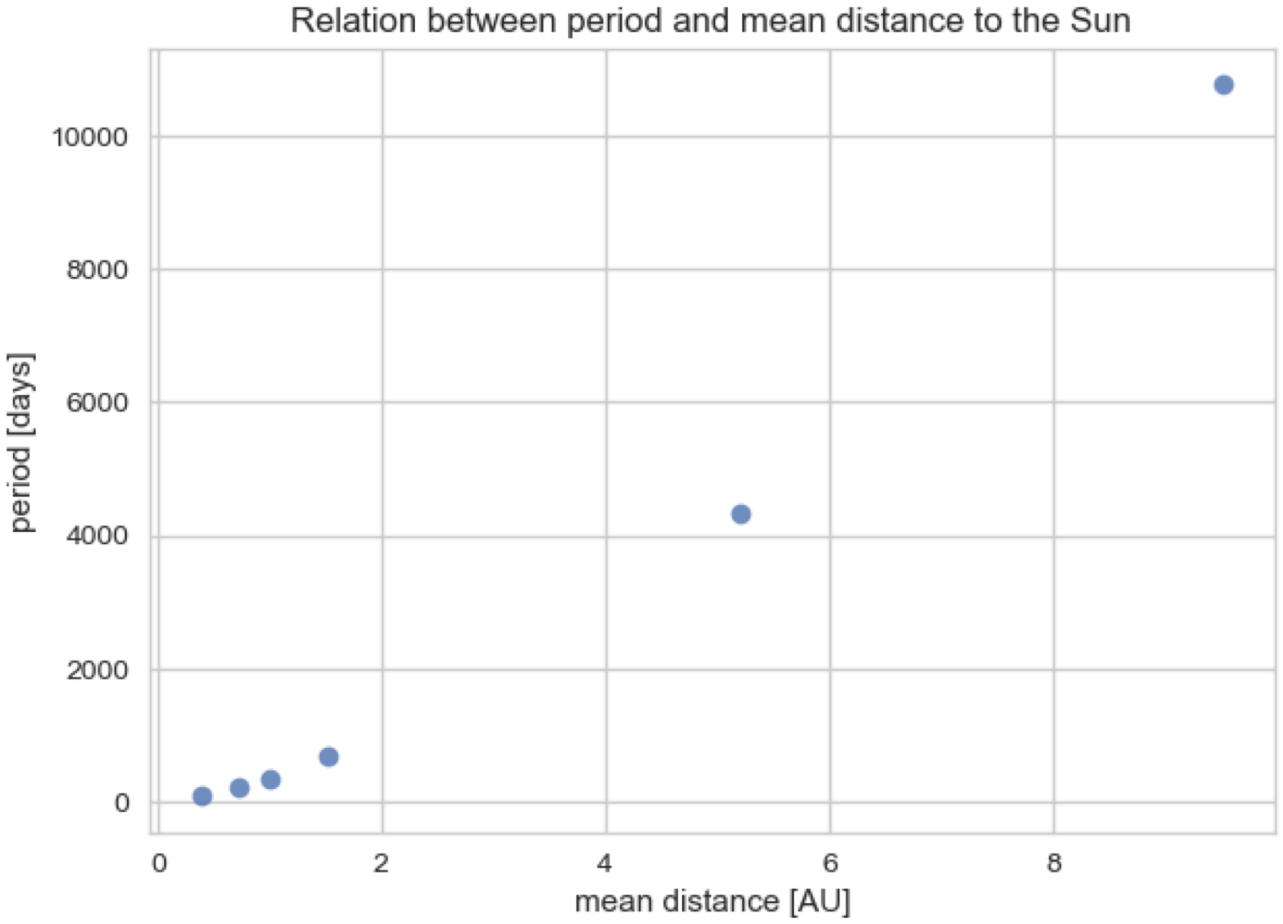
$$\frac{Upper\ Quartile - Median}{Median - Lower\ Quartile} \approx 1$$

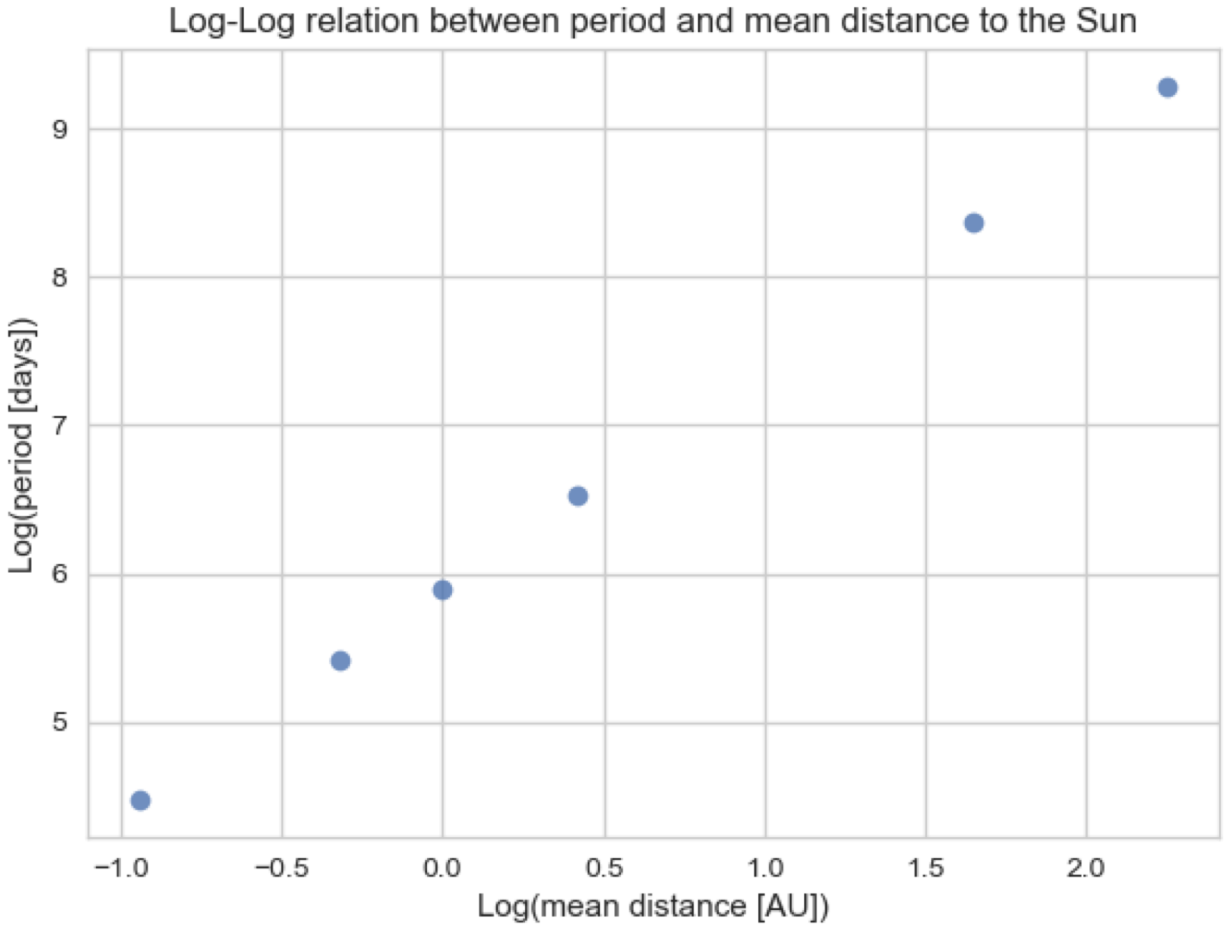
Logs and complex relations
Kepler's third law (1619)

Planet	Mean distance to sun [AU]	Period [days]
Mercury	0.389	87.77
Venus	0.724	224.70
Earth	1	365.25
Mars	1.524	686.95
Jupiter	5.2	4332.62
Saturn	9.510	10759.2



Logs and complex relations
Kepler's third law

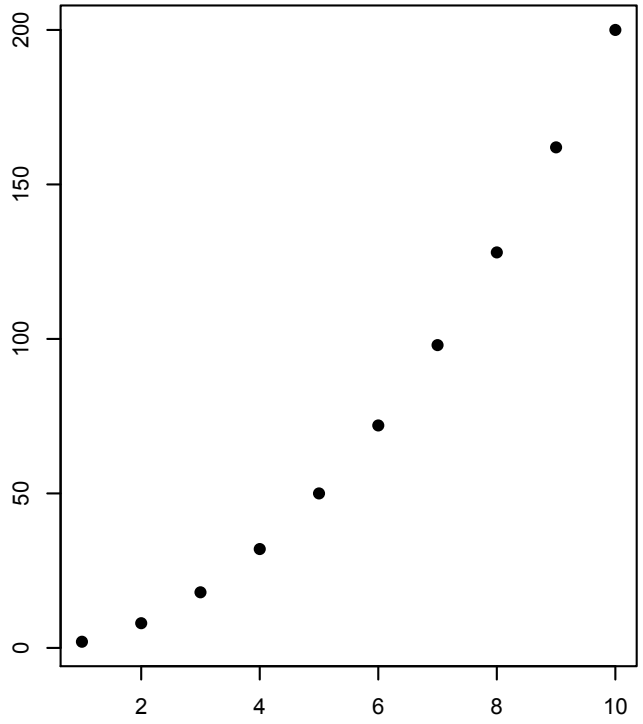


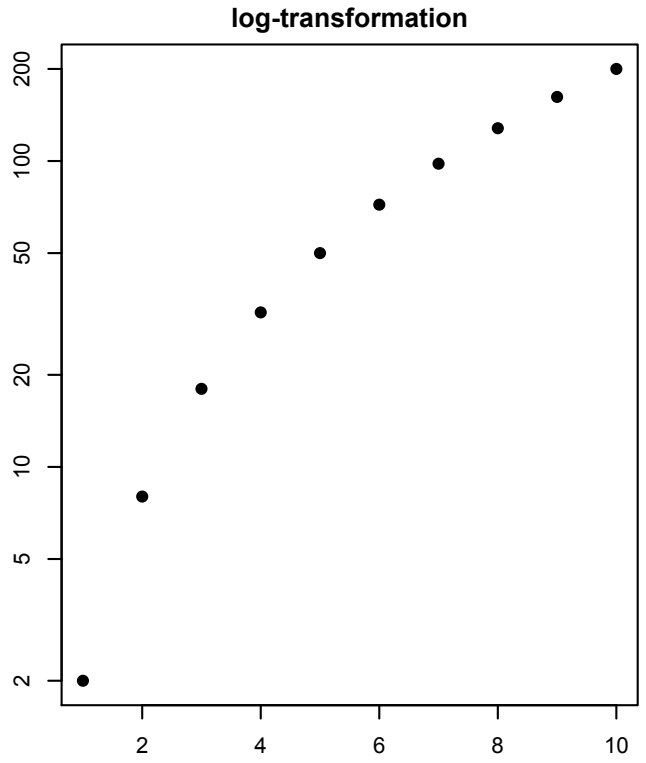


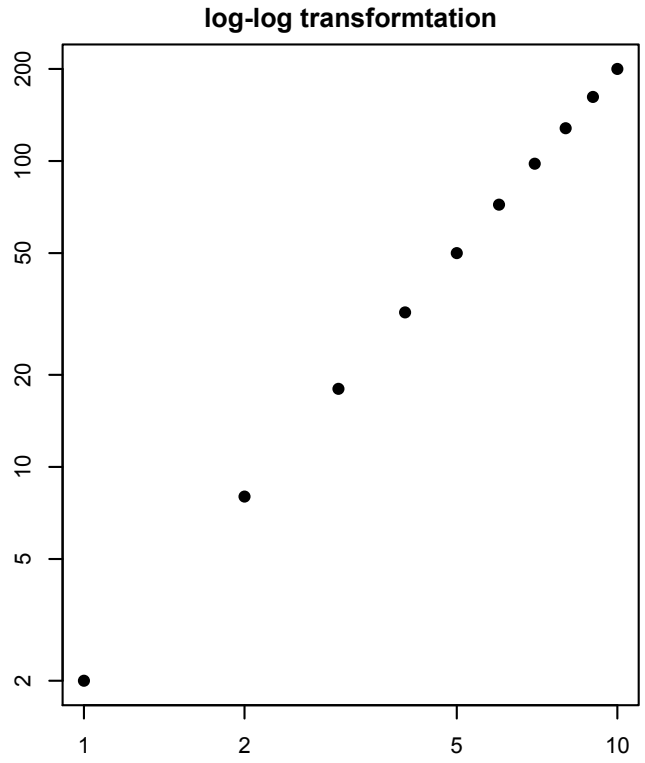
Why Straighten Relationships?

- 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, ½, 0 (log), -1

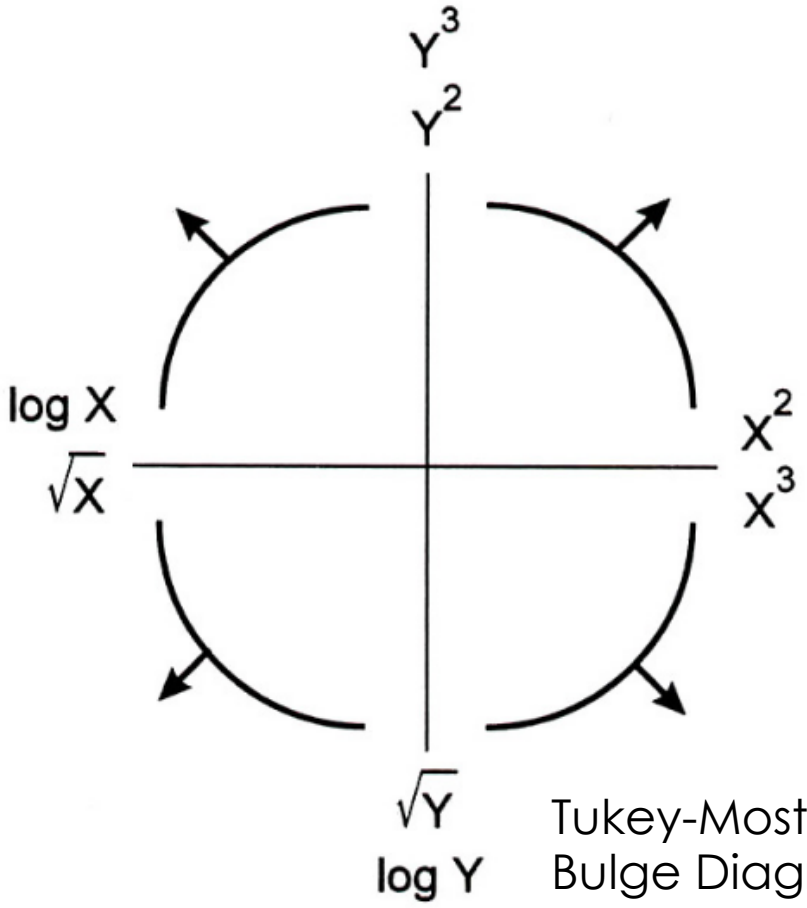
Straighten Relationships with Transformations



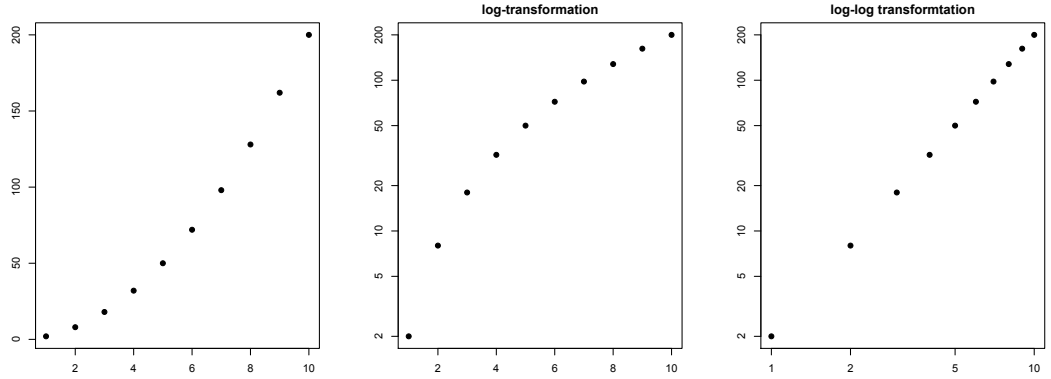




Straighten Relationships with Transformations

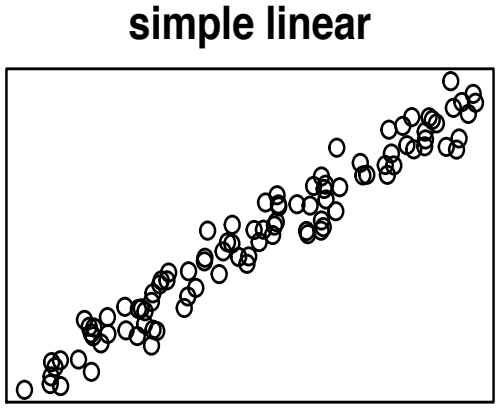


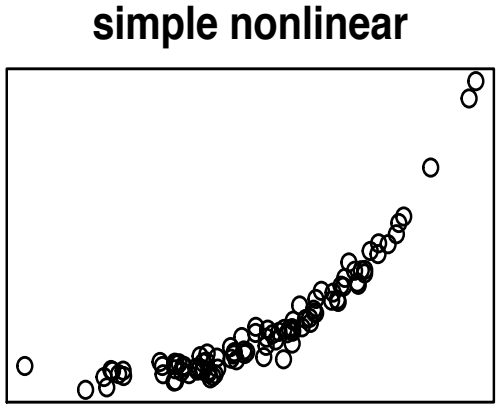
Tukey-Mosteller Bulge Diagram



Power transformations

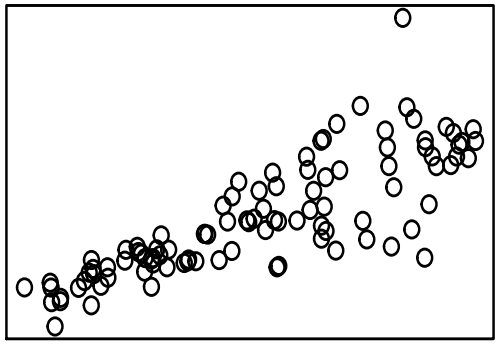
This is what we are looking for: monotone and linear

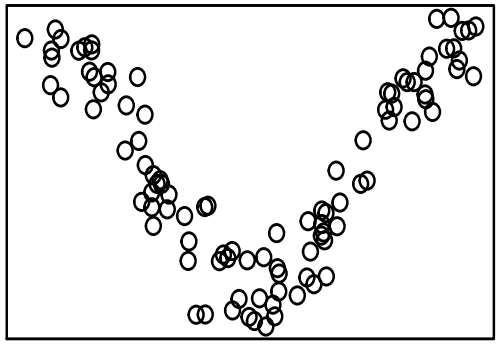




Bulge diagram says:
go up power ladder in x or down the power ladder in y or both

Some times a square root transformation can help





Shift positive and then transform can work

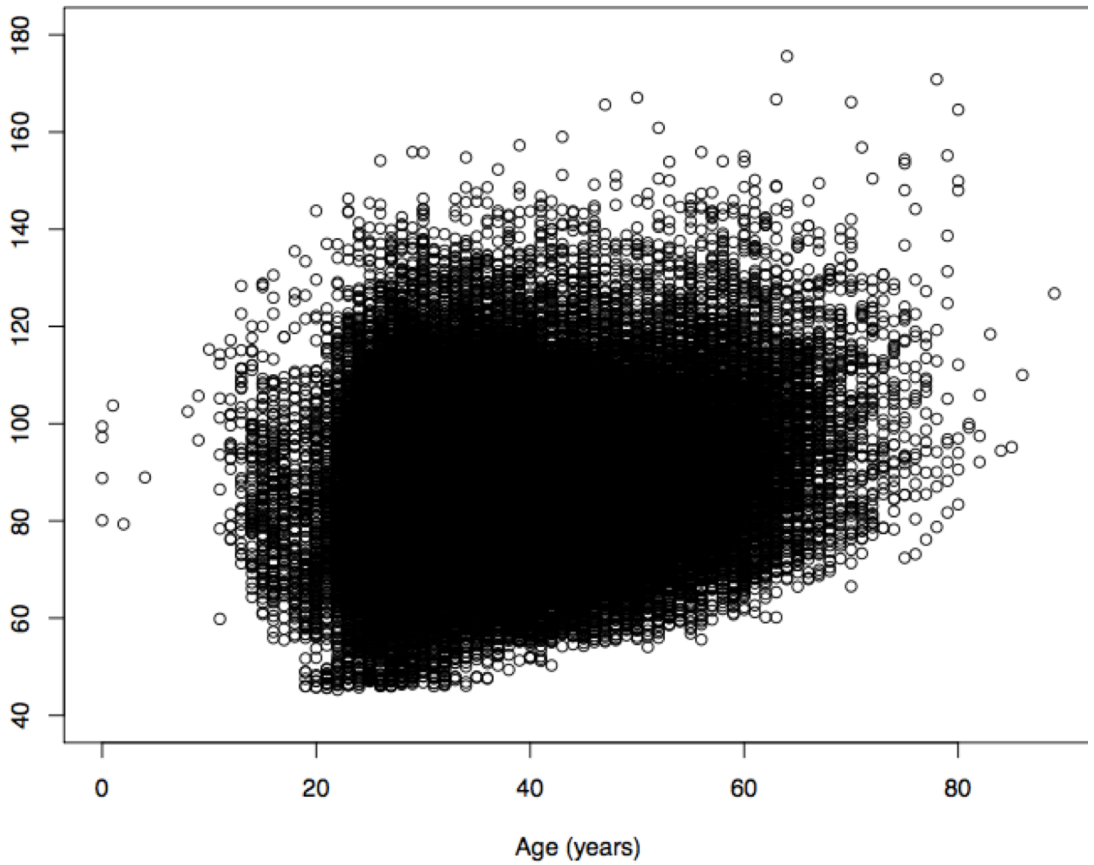
Add Context

Add Context

- Label axes, include 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

Large n (records)
Smoothing

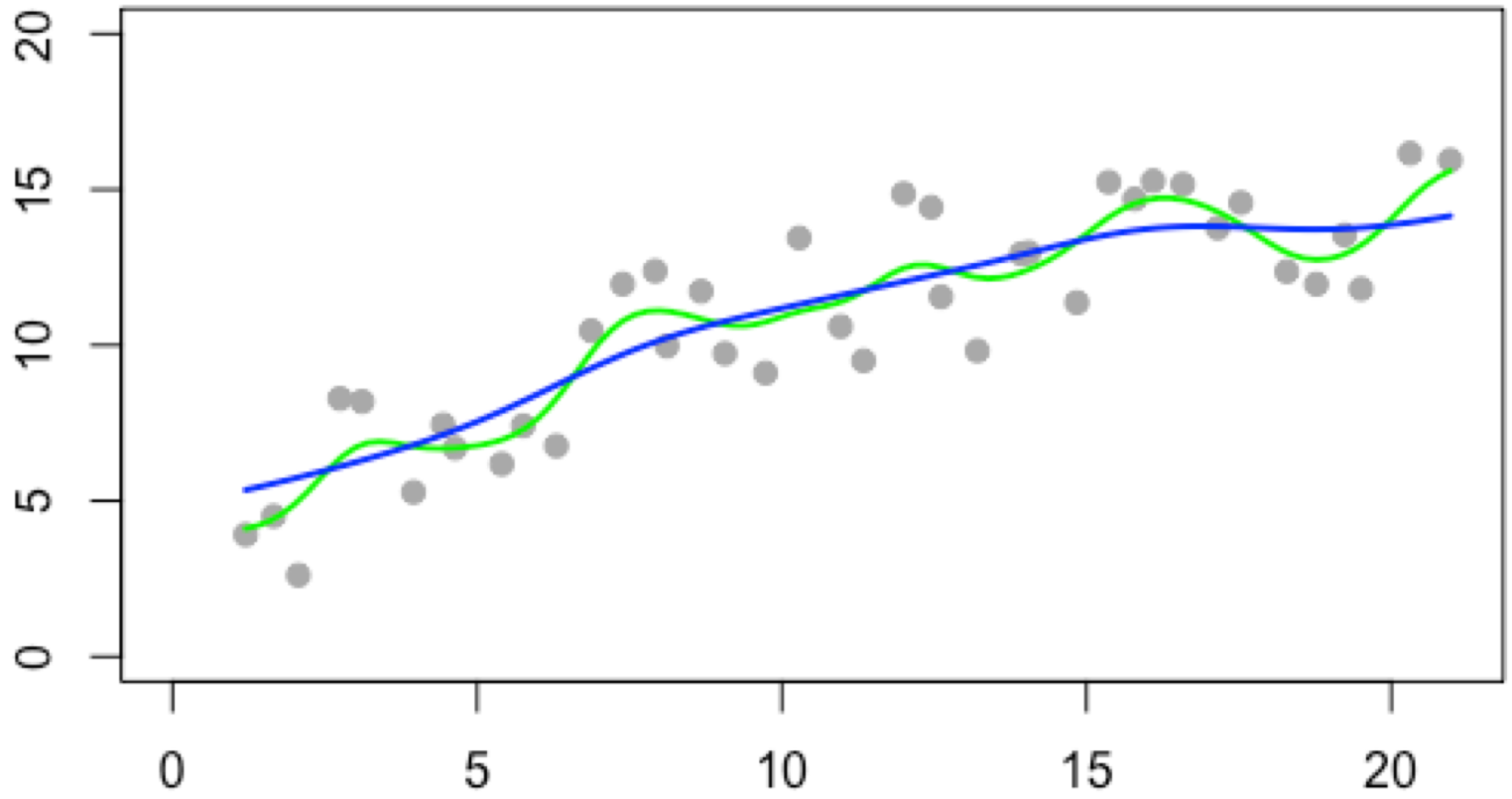
Cherry Blossom Run



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

Smoothing Scatter plots

$$g(x) = \sum_{i=1}^{n} \frac{K_h(x - x_i) y_i}{\sum K_h(x - x_i)}$$



For each x, we find g(x) by a weighted average of the y_i

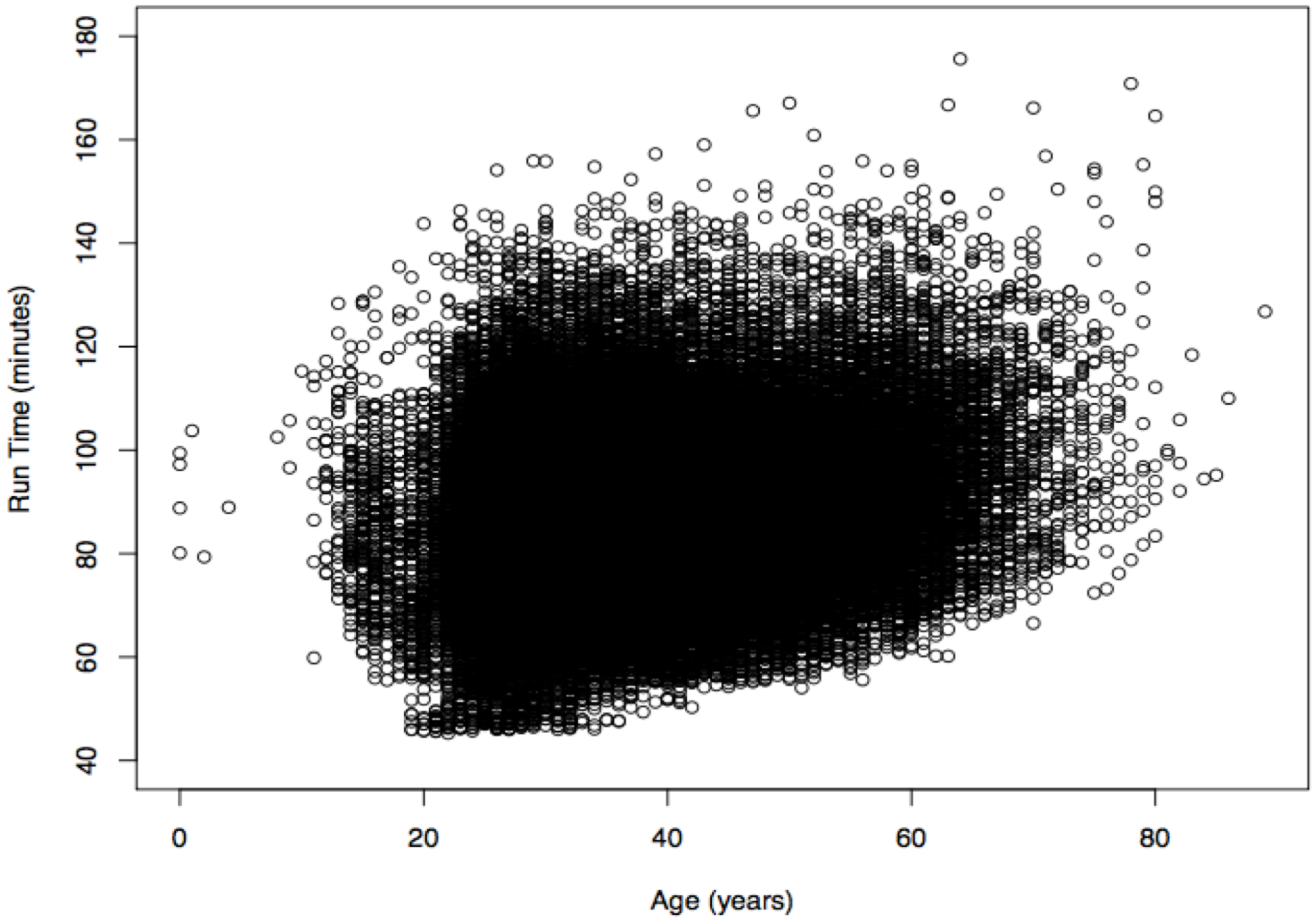
The y_i are weighted according to the kernel function.
So x_i far from x do not contribute much to g(x)

Local Smoothing

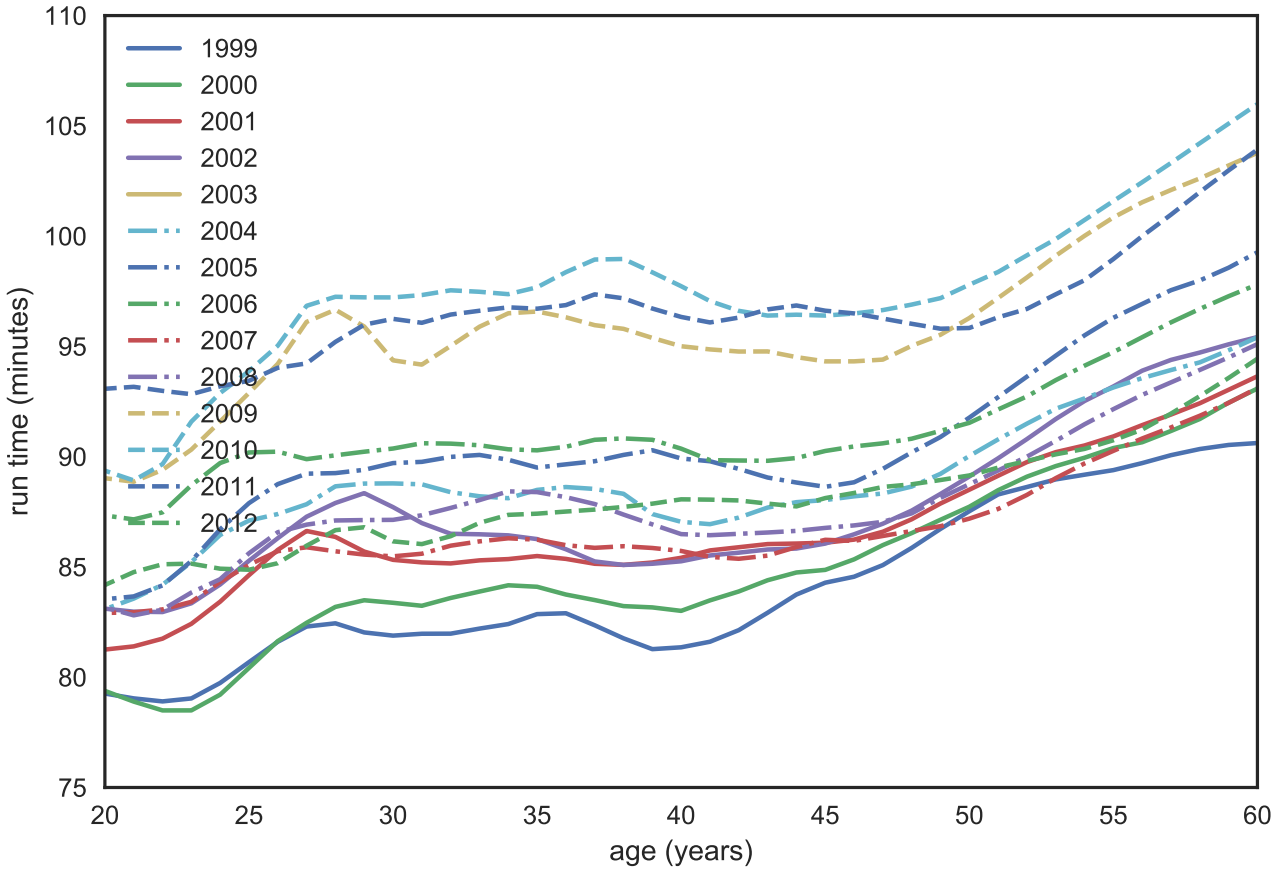
- Moving window
- Smooth/Average y values in the window
- Many different approaches for doing this:
  - kernel methods (what we just showed),
  - cubic splines, thin plate splines,
  - Locally weighted smooth scatterplot (lowess)

Allows us to see shape of the relationship between y and x

Cherry Blossom Run



Cherry Blossom Run

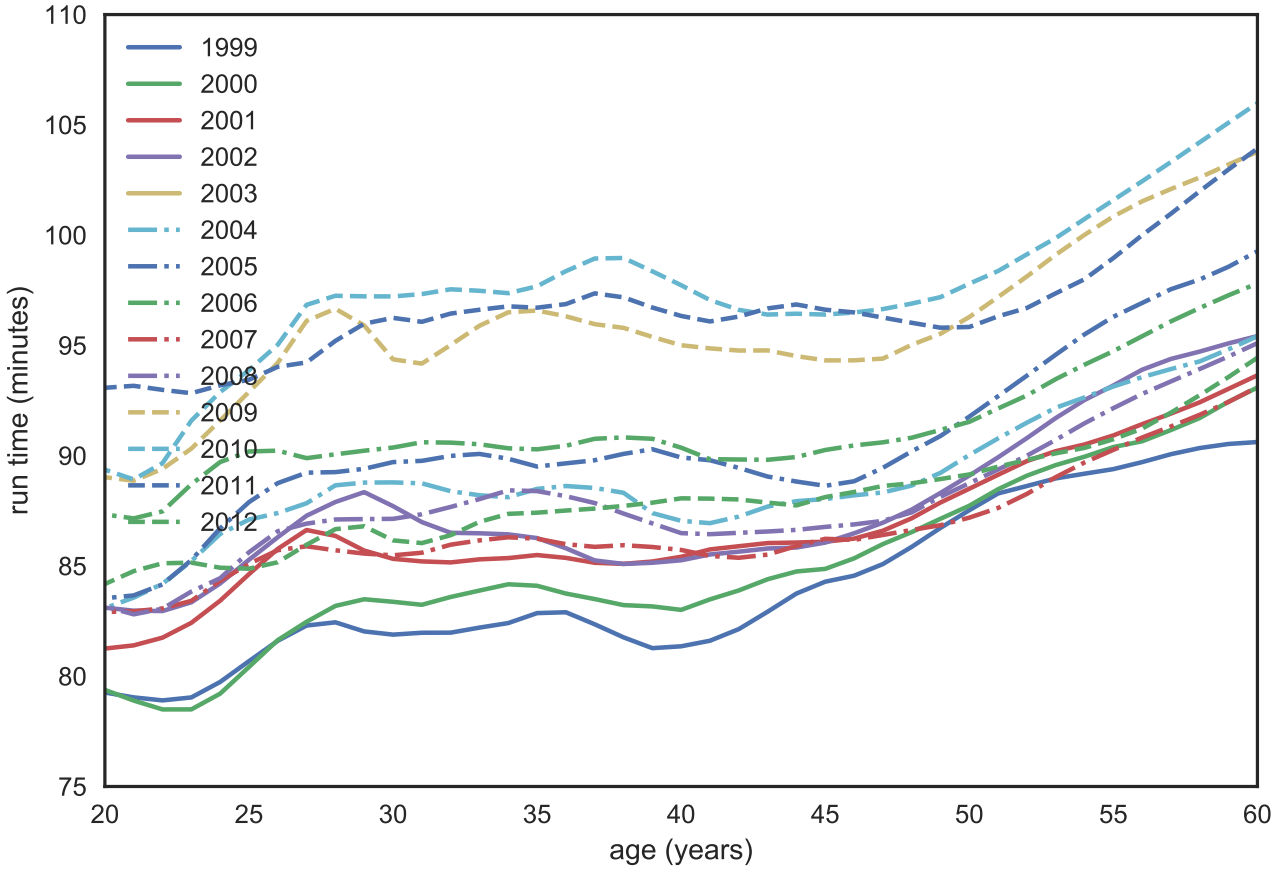


Local Smoothing helps us see the typical run time for each age

We have controlled for year by making separate smooth curves for runners in each yearly race

Notice anything unusual?

Cherry Blossom Run



How are the data generated?

All runners in each annual CB run from 1999 to 2012

Snapshots in time – one for each year – Runners in 1999 may be quite different than runners in 2012

If we control for year and only examine the relationship between time and age for 2012, we still have a problem. These are not longitudinal data where we follow the same people in time as they age.

Large p (variables) Dimension Reduction

PCA – in a couple of weeks

Philosophy

Reveal the Data

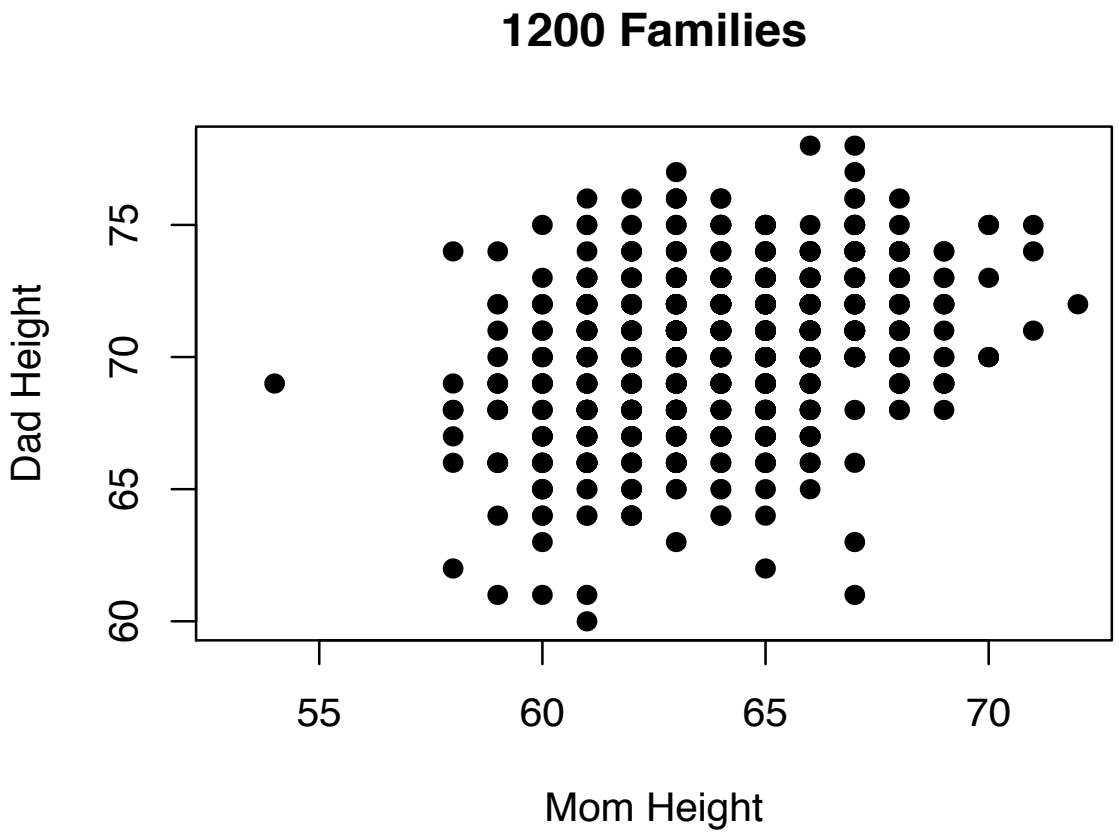
- Choose scale appropriately
- Avoid having other graph elements interfere with data
- Use visually prominent symbols
- Eliminate superfluous material, aka chart junk
- Avoid over-plotting

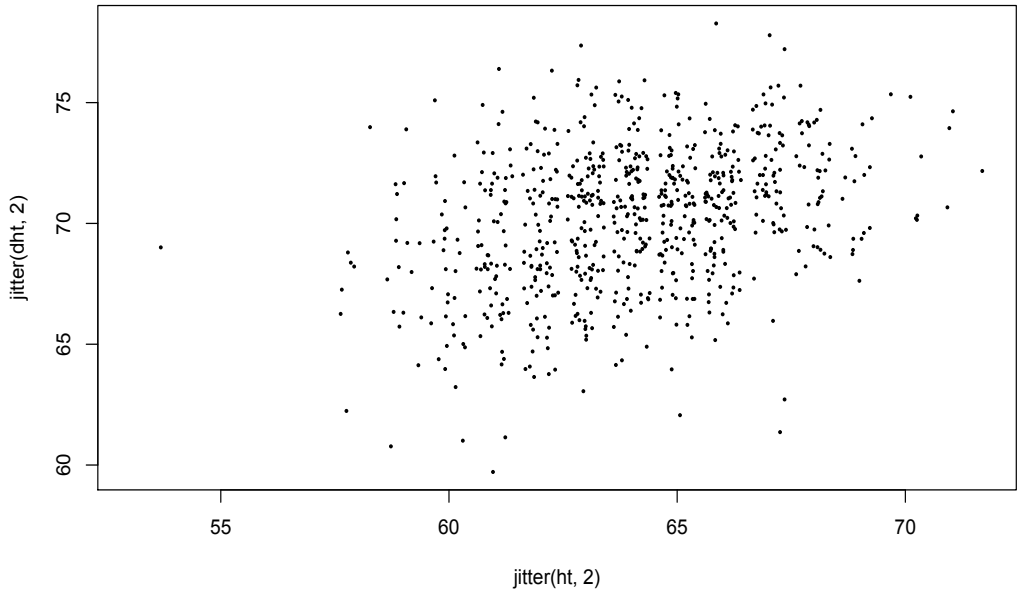
Avoid over-plotting

Why are there so few data points?

Jitter: Add random noise so the values aren't plotted on top of each other

Shrink the plotting symbol so they don't plot on top of each other





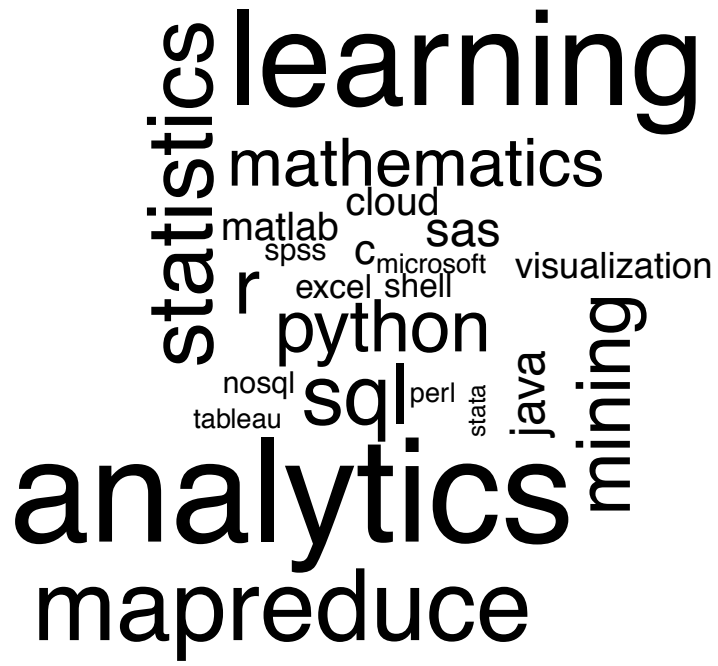
Facilitate Comparisons

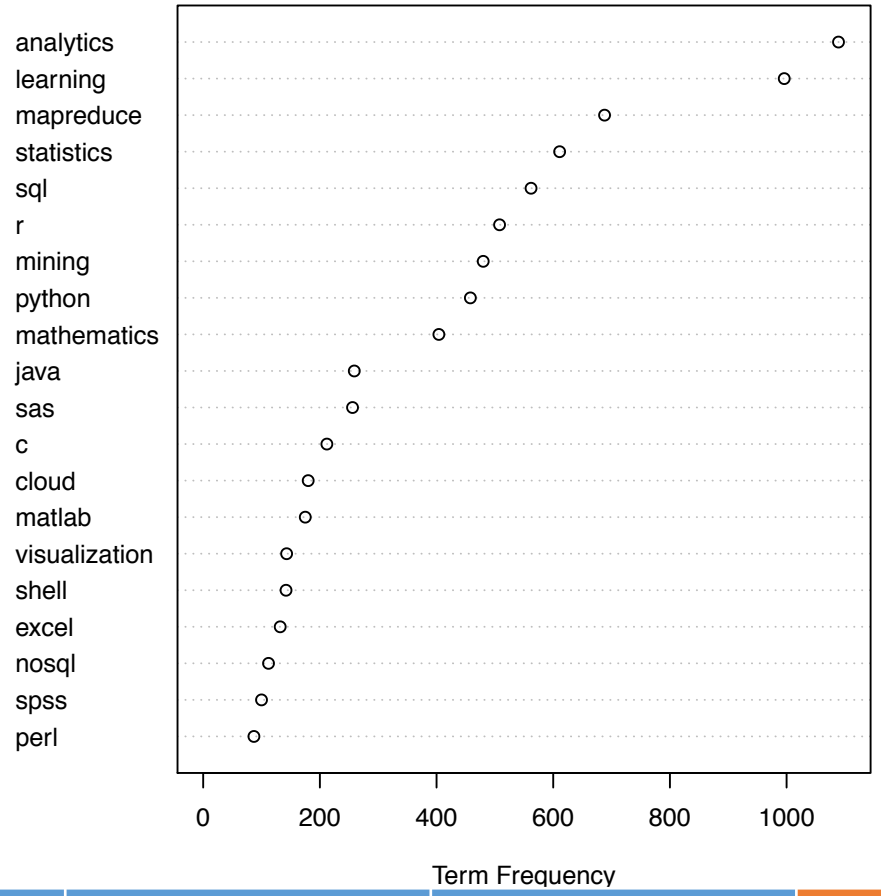
- Put Juxtaposed plots on same scale
- Make it easy to distinguish elements of *superposed* plots, e.g. with color, line type
- Avoid Stacking and Jiggling the baseline
- Avoid angles, extra dimensions (e.g., areas rather than lines)
- Don't break the visual metaphor, i.e., if use rectangles, then area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

Comparison: area vs length

Order of words/counts is random – makes it difficult to compare

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

Captions

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

Good Plot Making Practice

- Put major (**quantitative**) 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