CMPT 733 – Big Data Programming II

Visualization Principles for Data Science

Instructor Steven Bergner

Course website https://coursys.sfu.ca/2025sp-cmpt-733-g1/pages

Source: Ch. 6.4 - 6.6 of "Principles of Data Science" by Lau, Gonzales, Nolan

Defining Visualization (Vis)

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

["Visualization Analysis and Design" by T. Munzner, 2014]

Why have a human in the loop?

- Not needed when automatic solution is trusted
- Good for ill-specified anlaysis problems
  - Common setting: "What questions can we ask?"

Why have a human in the loop?

Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

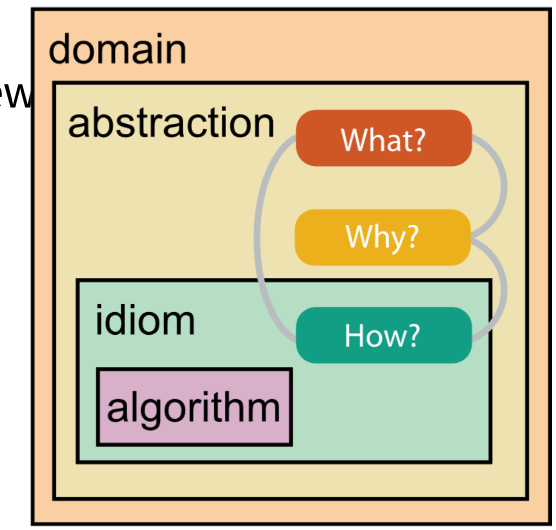
Munzner, T. (2014)

Long-term use
- Exploratory analysis of scientific data
- Presentation of known results

Short-term use
- For **developers** of automatic solutions:
  - Understand requirements for model development
  - Refine/debug and determine parameters
- For **end users** of automatic solutions: verify, build trust

Analysis framework: four levels

- ***Domain*** situation: Who are the target users?
- ***Abstraction***: Translate from specifics of domain to vocabulary of vis
- **What** is shown? *Data abstraction*
  - Don't just draw what you're given: transform to new
- **Why** is the user looking at it? *Task abstraction*
- **How** is it shown? ***Idiom (Vis technique)***
  - Visual encoding idiom: How to draw
  - Interaction idiom: How to manipulate
- ***Algorithm***: efficient computation

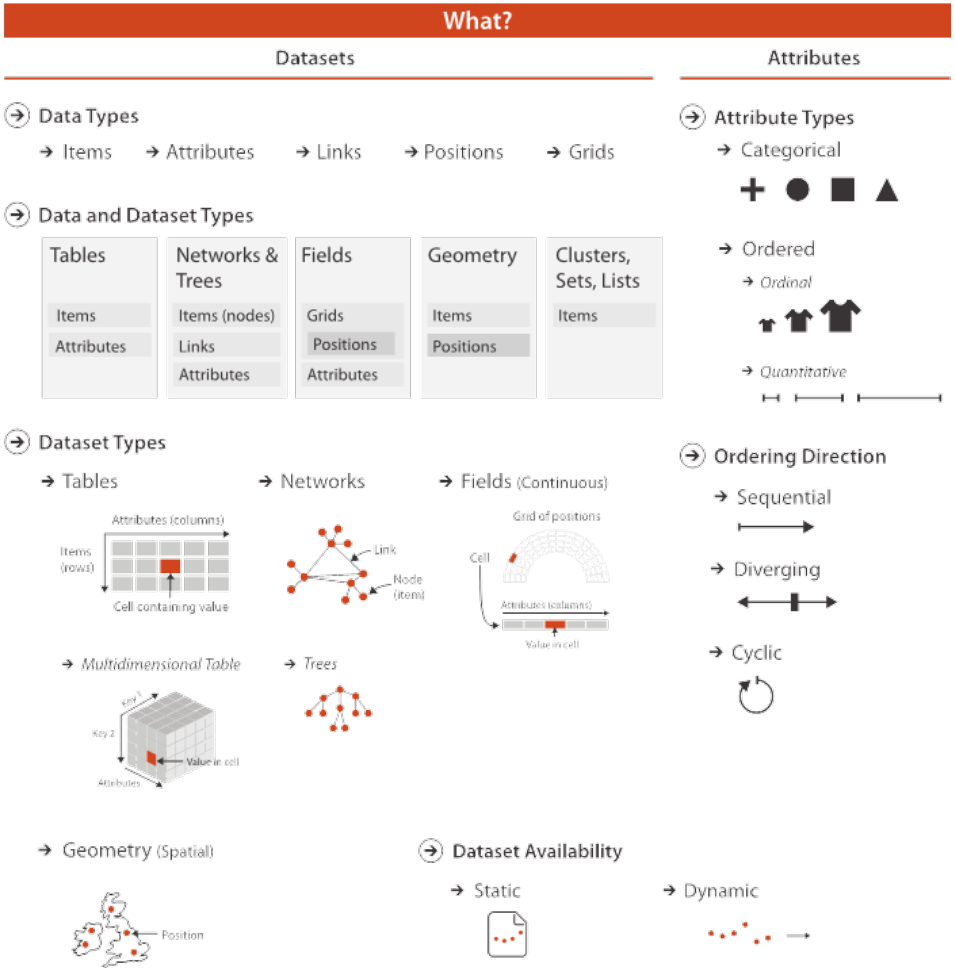


[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

Resource limitations

- **Computational** limits
  - Processing time and system memory
- **Human** limits
  - Human attention and memory
  - Understanding abstractions
- **Display** limits
  - Pixels are precious
  - Information density tradeoff: Info encoding vs unused whitespace

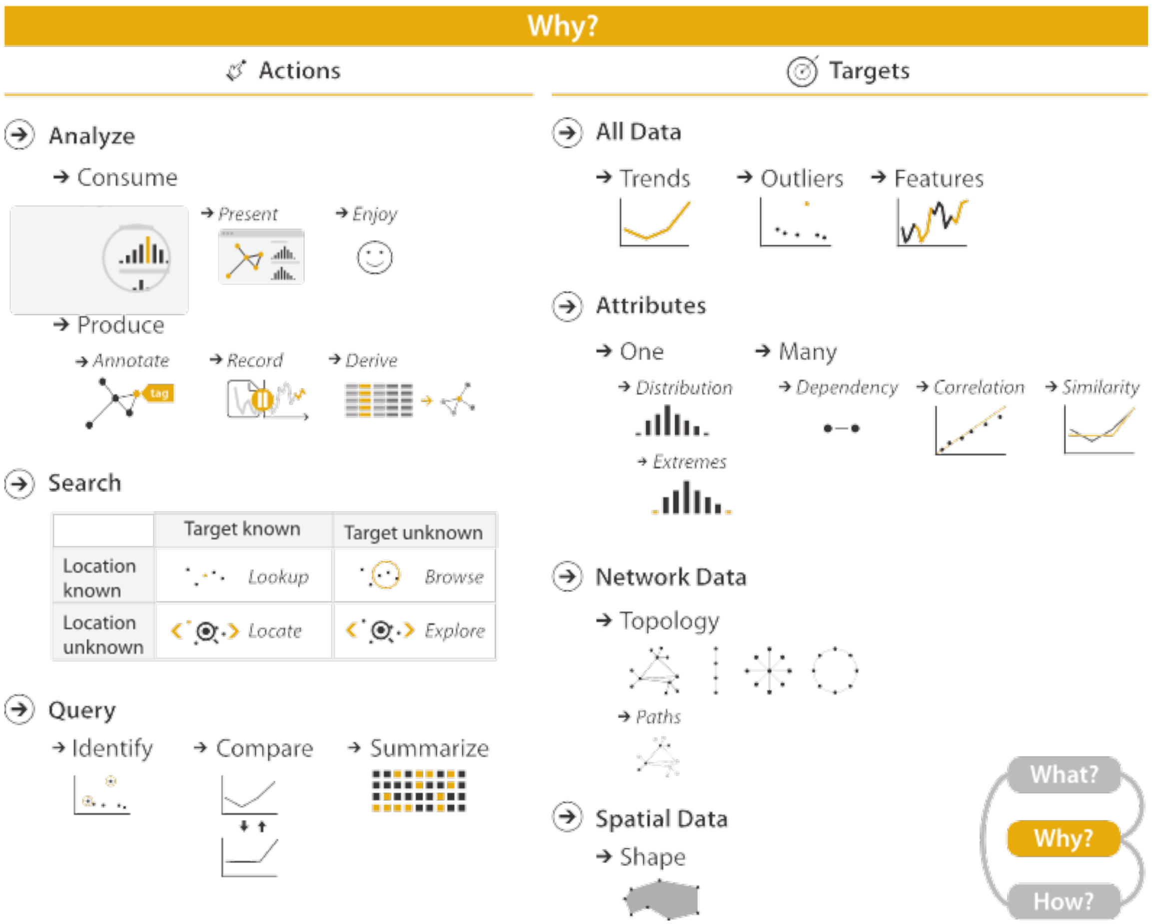
Understand
Data, Task, and Encoding



[T. Munzner, 2014]

Data Types

- Items and attributes as rows and columns of tables
- Position and time are special attributes
- Spatial data on grids makes computation easier



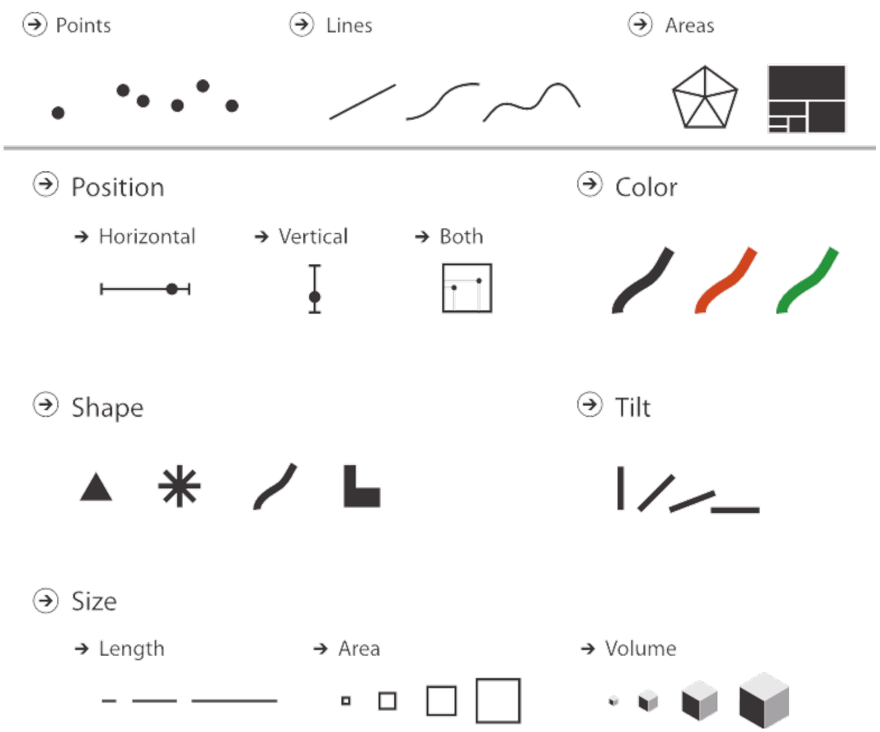
[T. Munzner, 2014]

Tasks

- Actions
  - Analyze
  - Search
  - Query
- Targets
  - Item & Attributes
  - Topology & Shape

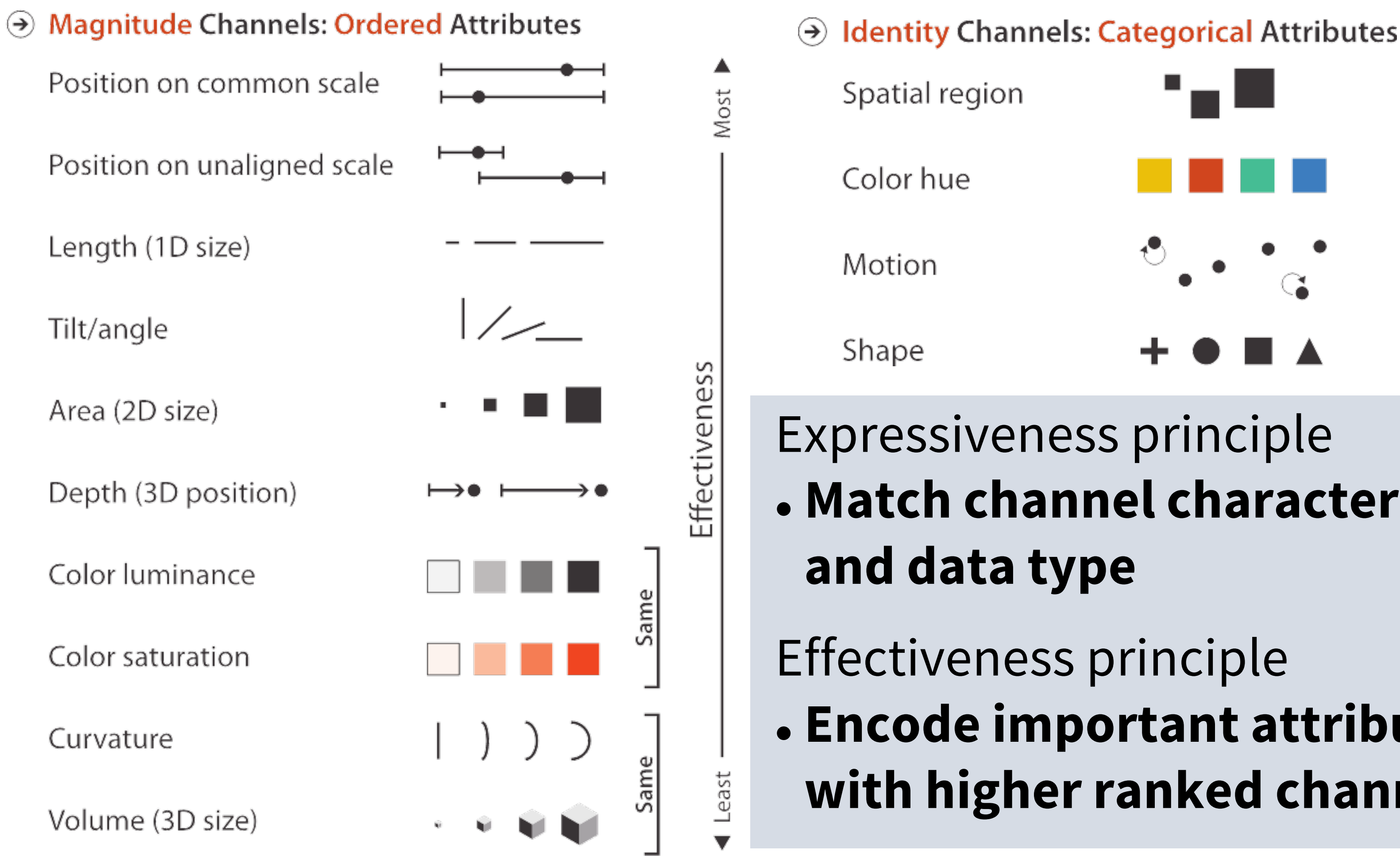
Visual Encoding – How?

- Marks
  - Geometric primitives
- Channels
  - Appearance of marks
  - Redundant coding of data with multiple channels is possible



[T. Munzner, 2014]

Design Principles for Task Effective Visualization



Expressiveness principle

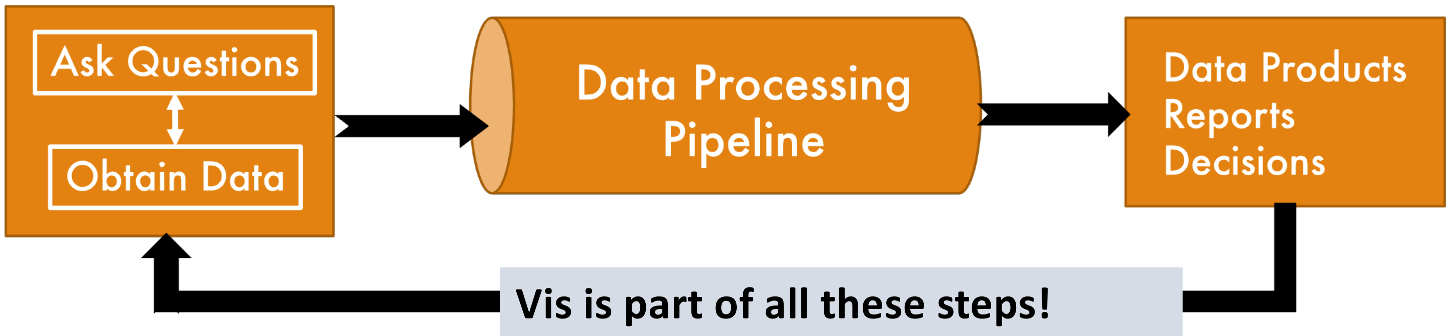
- **Match channel characteristics and data type**

Effectiveness principle

- **Encode important attributes with higher ranked channels**

[T. Munzner, 2014]

Recap: Data Science Lifecycle



Related Processes

Big Data Journey

- Business transformations as a company becomes more data-centric

Data Visualization *Process*

- Acquire, Parse, Filter, Mine, Represent, Refine, Interact [**Ben Fry '07**, Visualizing Data]

Data Visualization *Pipeline*

- Analyse (Wrangling), Filter, Map to visual properties, Render geometry

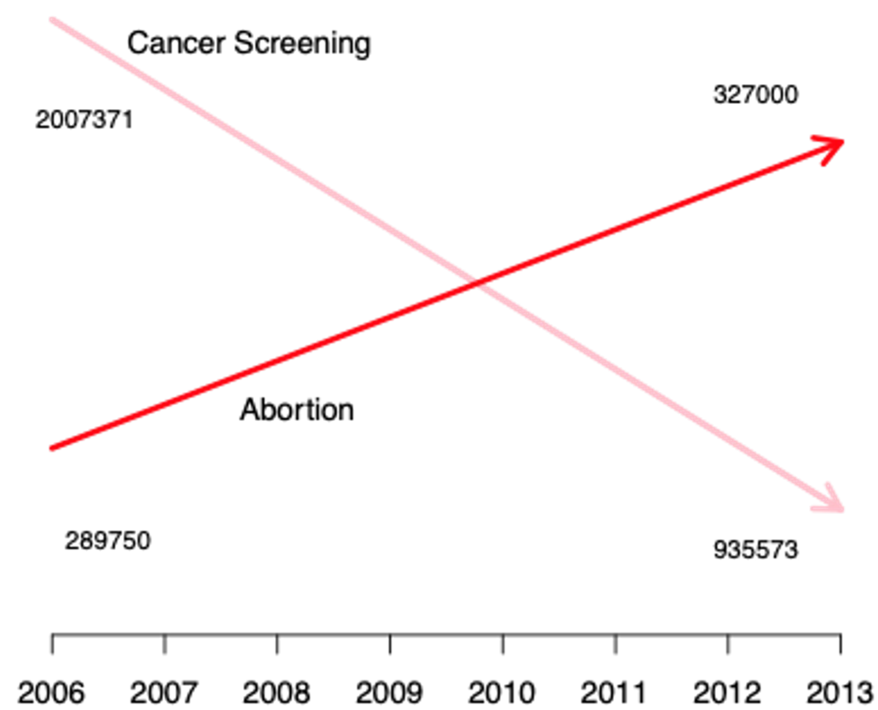
Six Principles Today

1. Scale
2. Conditioning
3. Perception
4. Transformations
5. Context
6. Smoothing

Explored via three case studies.

Case 1: Planned Parenthood 2015 Hearing

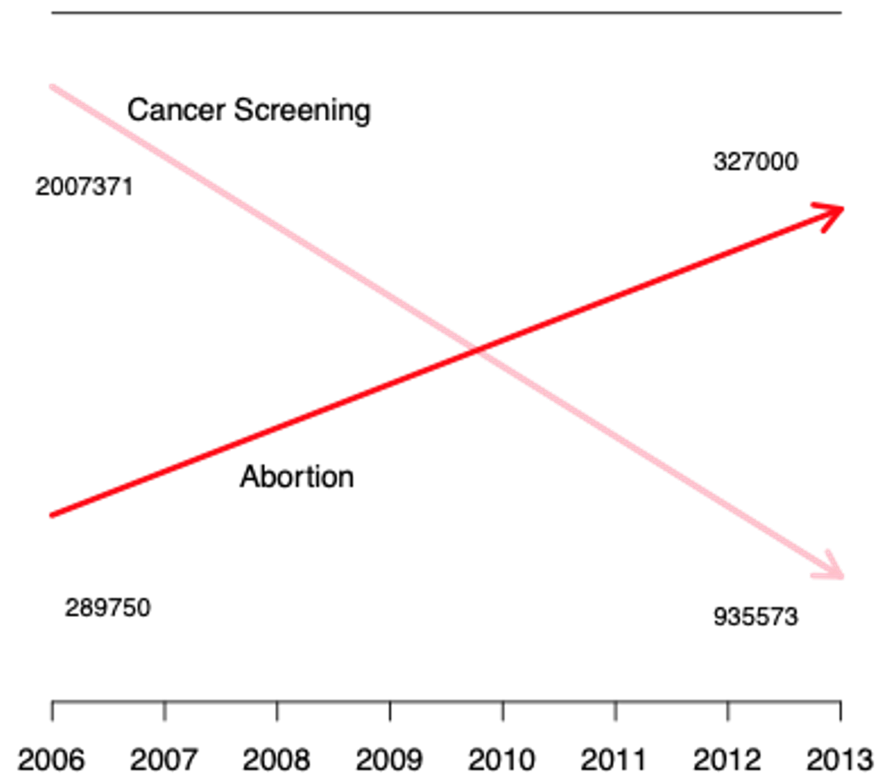
- Investigation of federal funding of Planned Parenthood in light of fetal tissue controversy
- Congressman Chaffetz (R-UT) showed plot which originally appeared in a report by Americans United for Life (http://www.aul.org/)



Full Report available at https://oversight.house.gov/interactivepage/plannedparenthood/ .

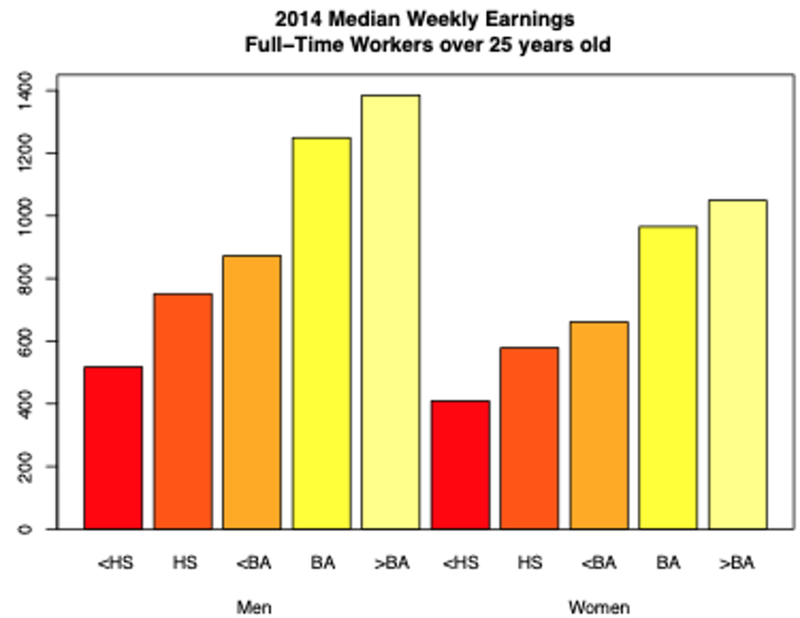
Case 1: Planned Parenthood 2015 Hearing

- Procedures: cancer screenings and abortions
- How many data points are plotted?
- What is suspicious?
- What message is this plot trying to convey?



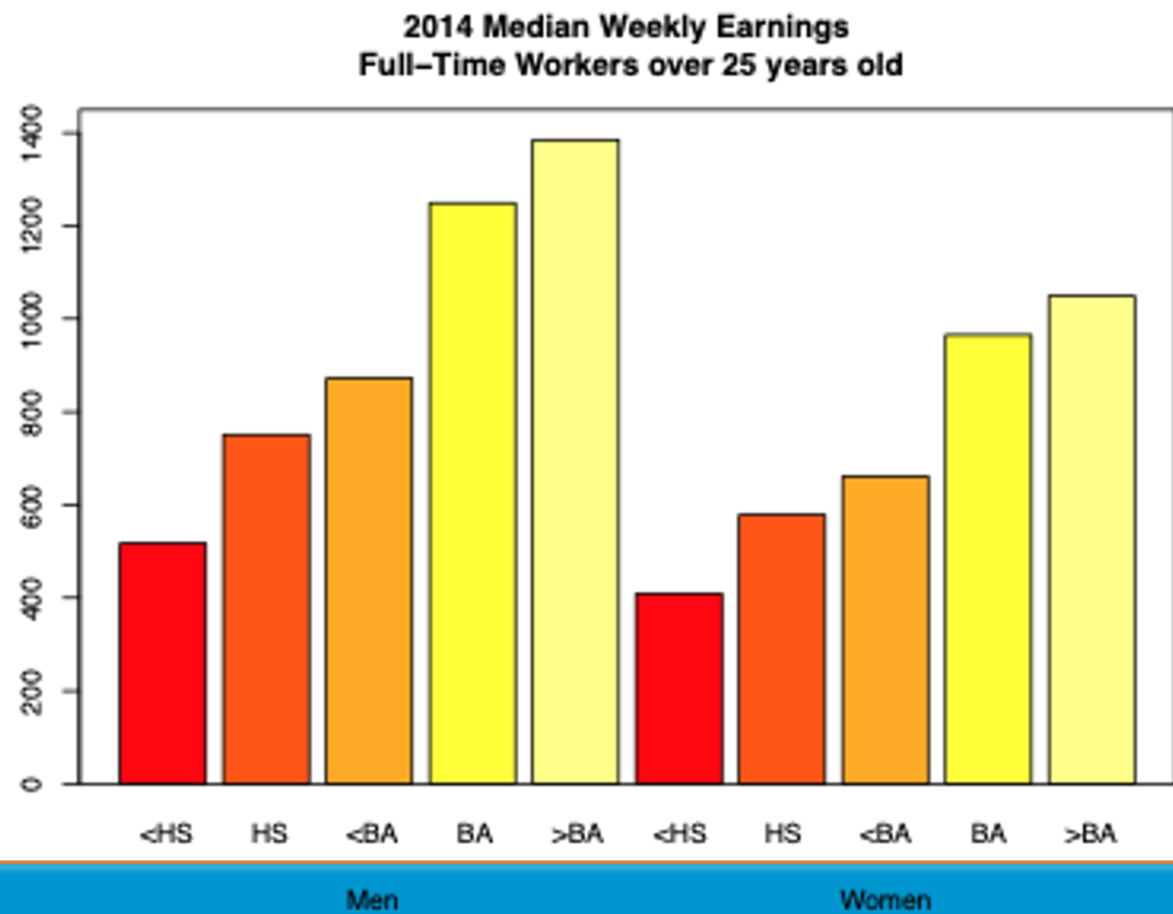
Case 2: Median Weekly Earnings

- Bureau of Labor Statistics surveys economics of labor
- www.bls.gov - Web interface to a report generating app
- Plot of median weekly earnings for males and females by education level



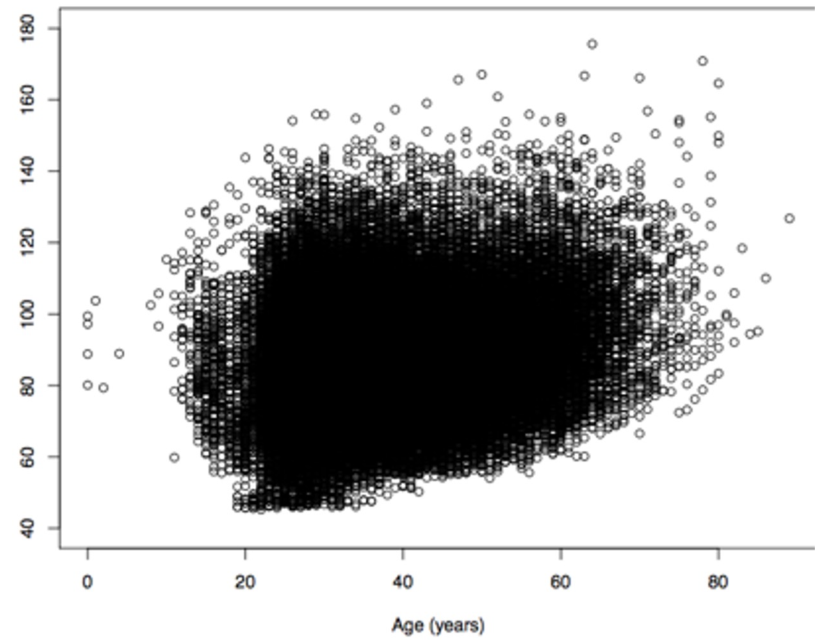
Case 2: Median Weekly Earnings

- What comparisons are easily made with this plot?
- What comparisons are most interesting and important?



Case 3: Cherry Blossom Runners

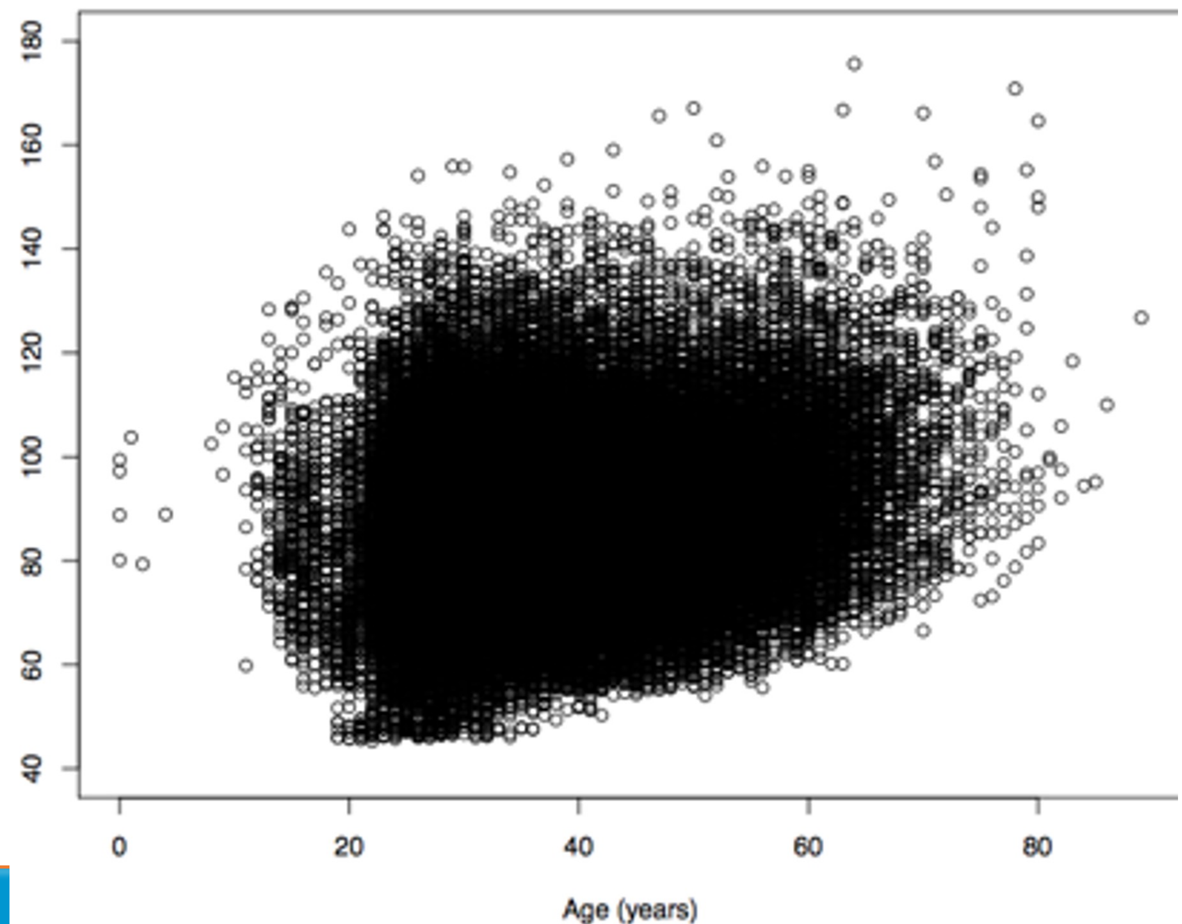
- 10 mi run in DC every April
- Results available from 1999-2019
- In 2019 over 17,000 runners
- Scatter plot of run time (min) against age (yrs)



http://www.cherryblossom.org/

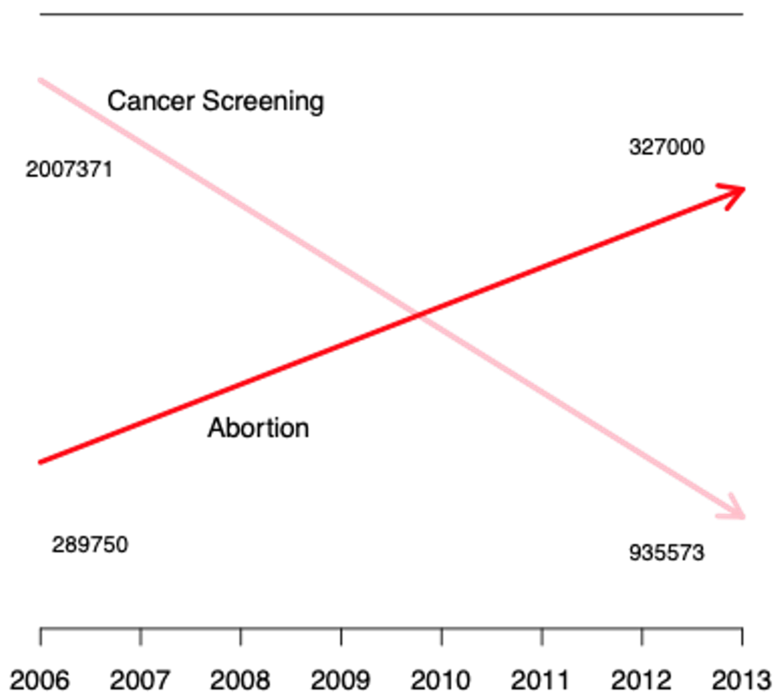
Case 3: Cherry Blossom Runners

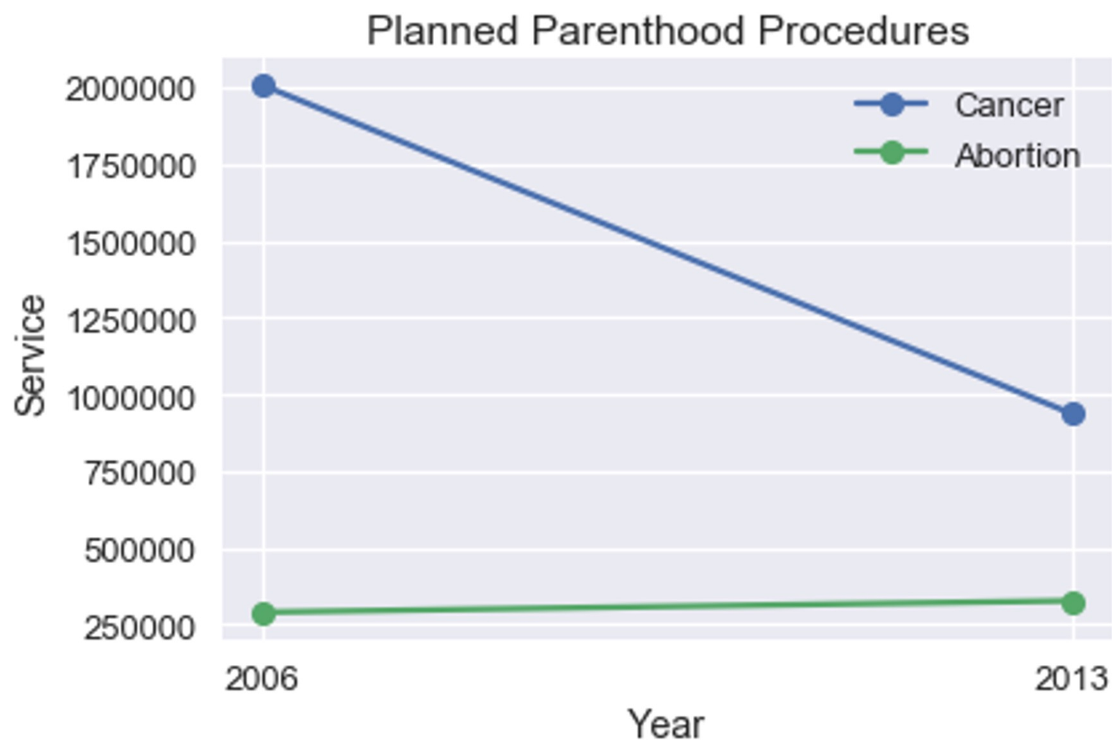
- 70,000+ points in the plot!
- What's the relationship between run time and age?



Principles of Scale

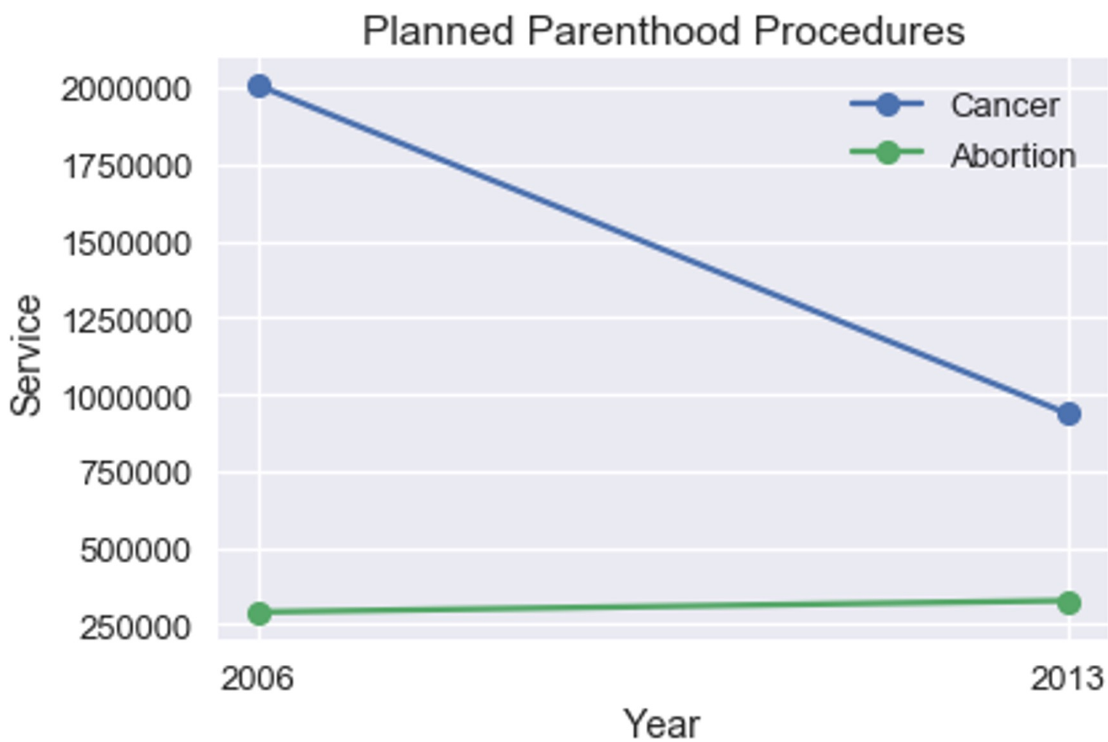
Scale


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


Planned Parenthood Procedures
2000000
1750000
1500000
1250000
1000000
750000
500000
250000
Service
Cancer
Abortion
2006
2013
Year

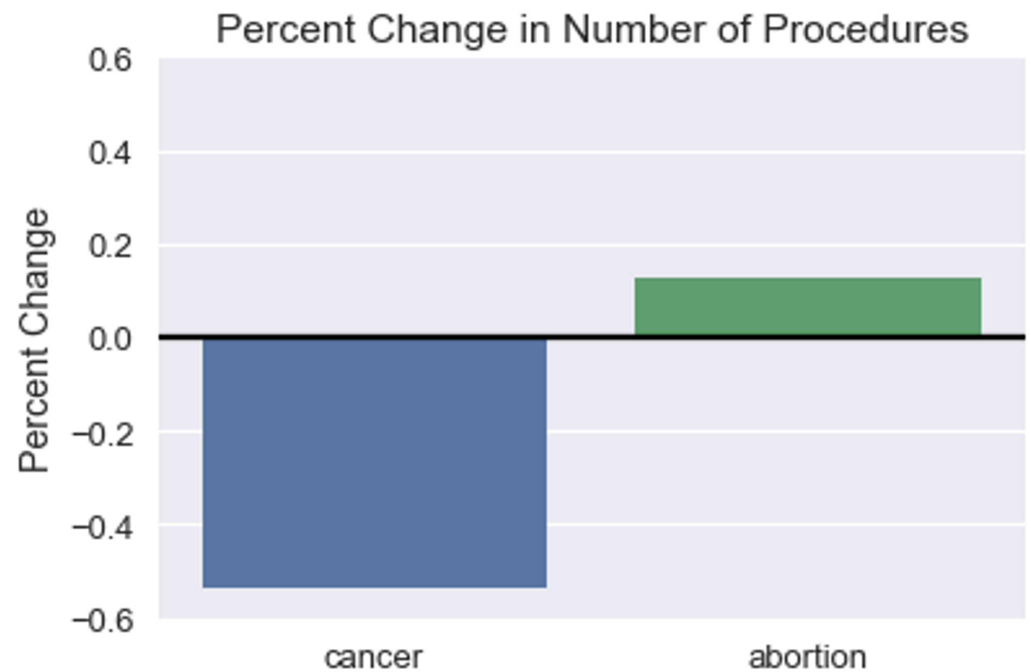
Keep consistent axis scales

- Don't change scale mid-axis
- Don't use two different scales for same axis
- How does this plot change perception of information?



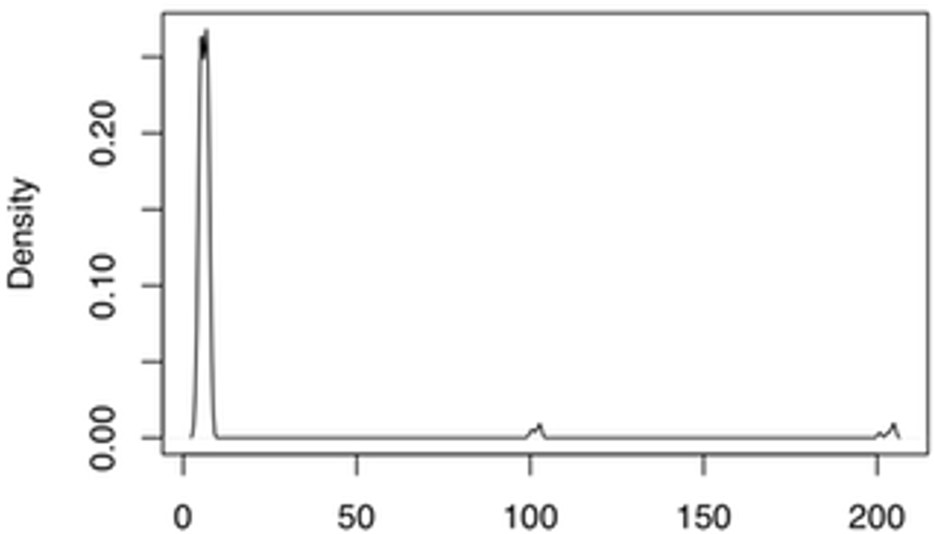
Consider Scale of Data

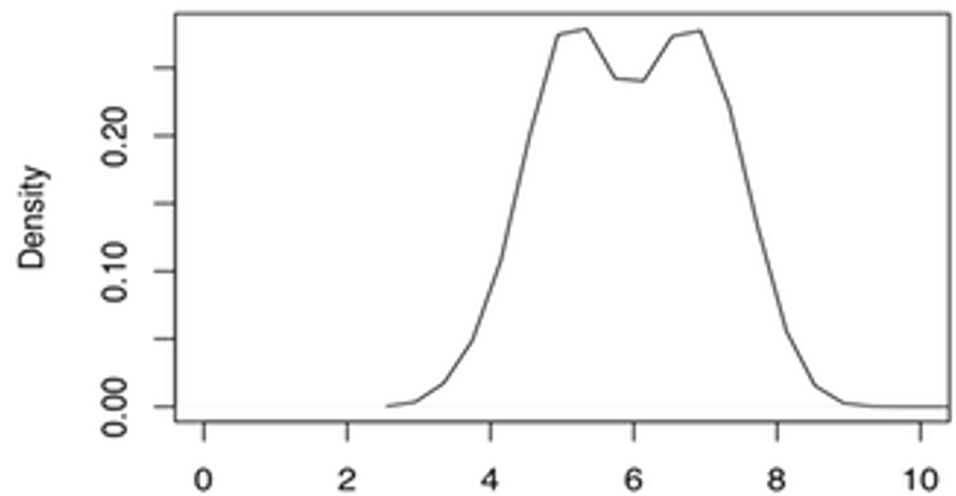
- Scales of cancer screenings vs. abortions quite different
- Can plot percent change instead of raw counts



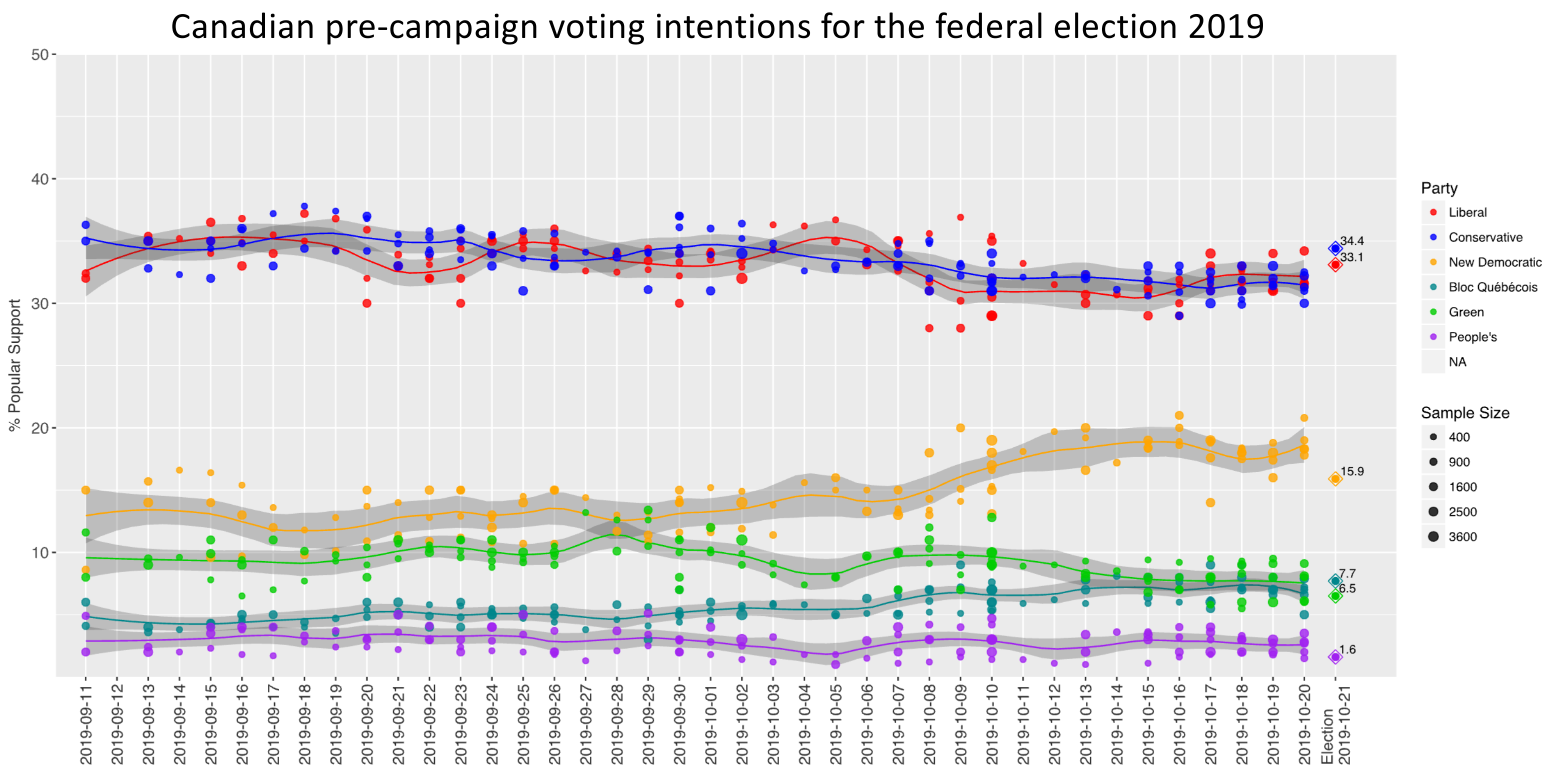
Reveal the Data

- Choose axis limits to fill plot
- If necessary, zoom into region with most of data
  - Can make separate plots for different regions





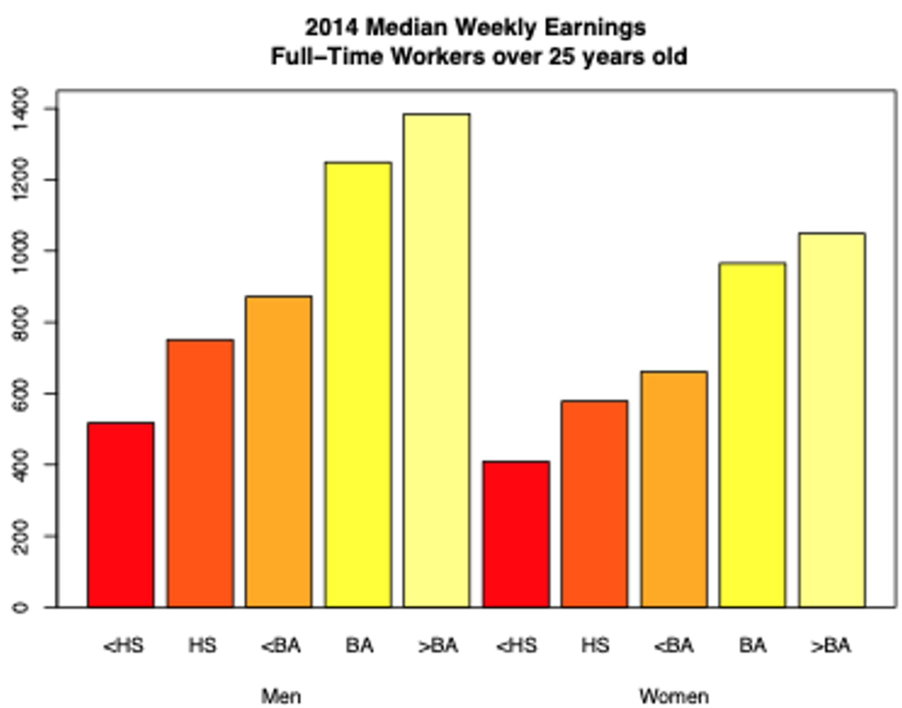
Time Scale

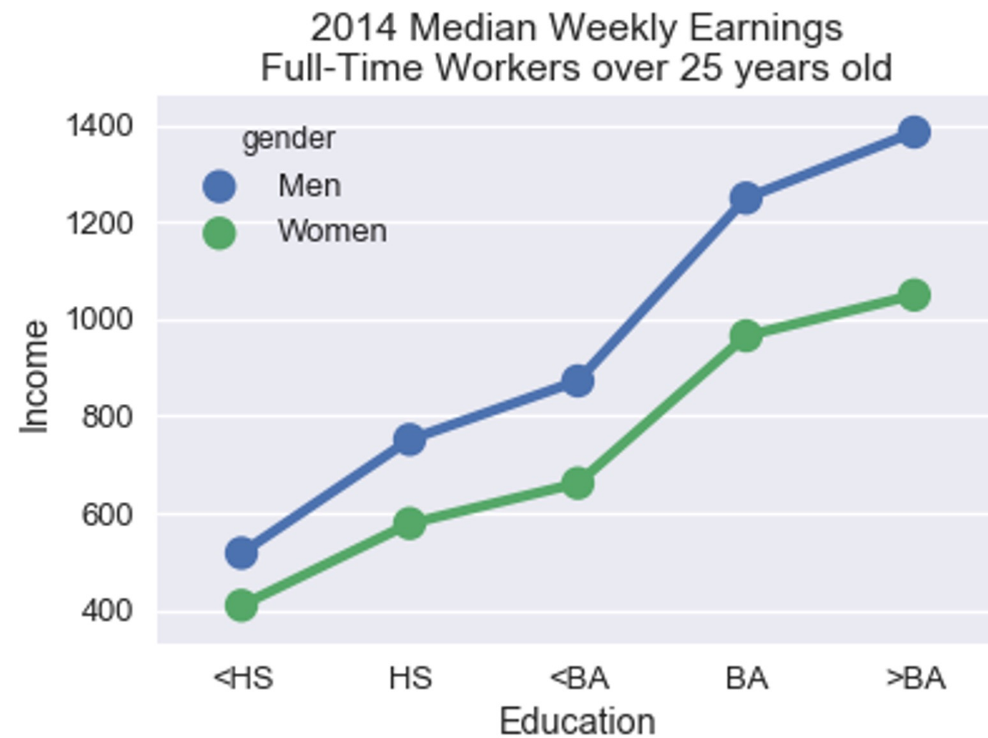


Code available at https://github.com/tylerecouture/wikiplot/blob/master/canadian-federal-polls-pre-43rd.R

Principles of Conditioning

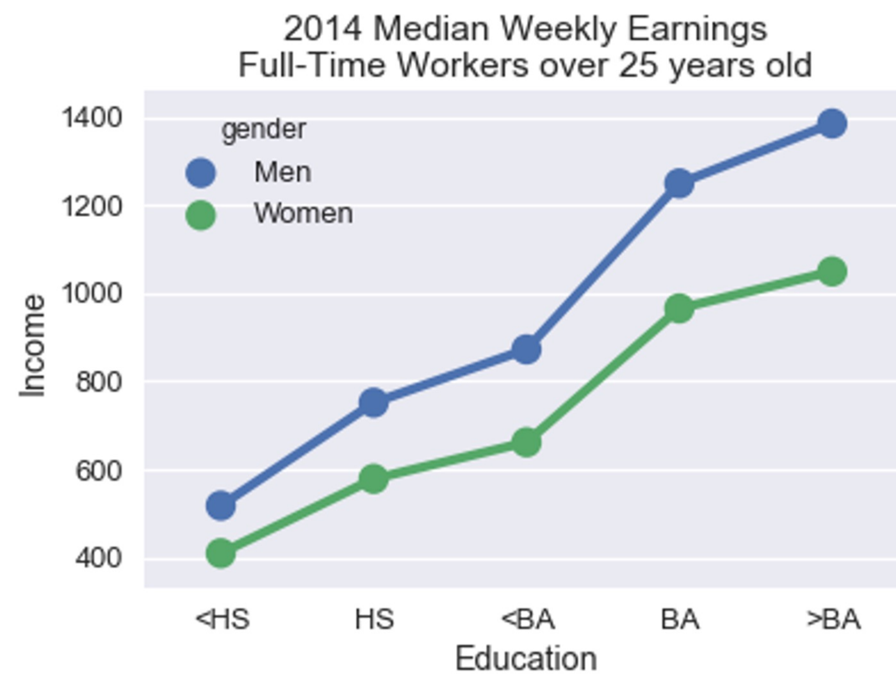
Conditioning





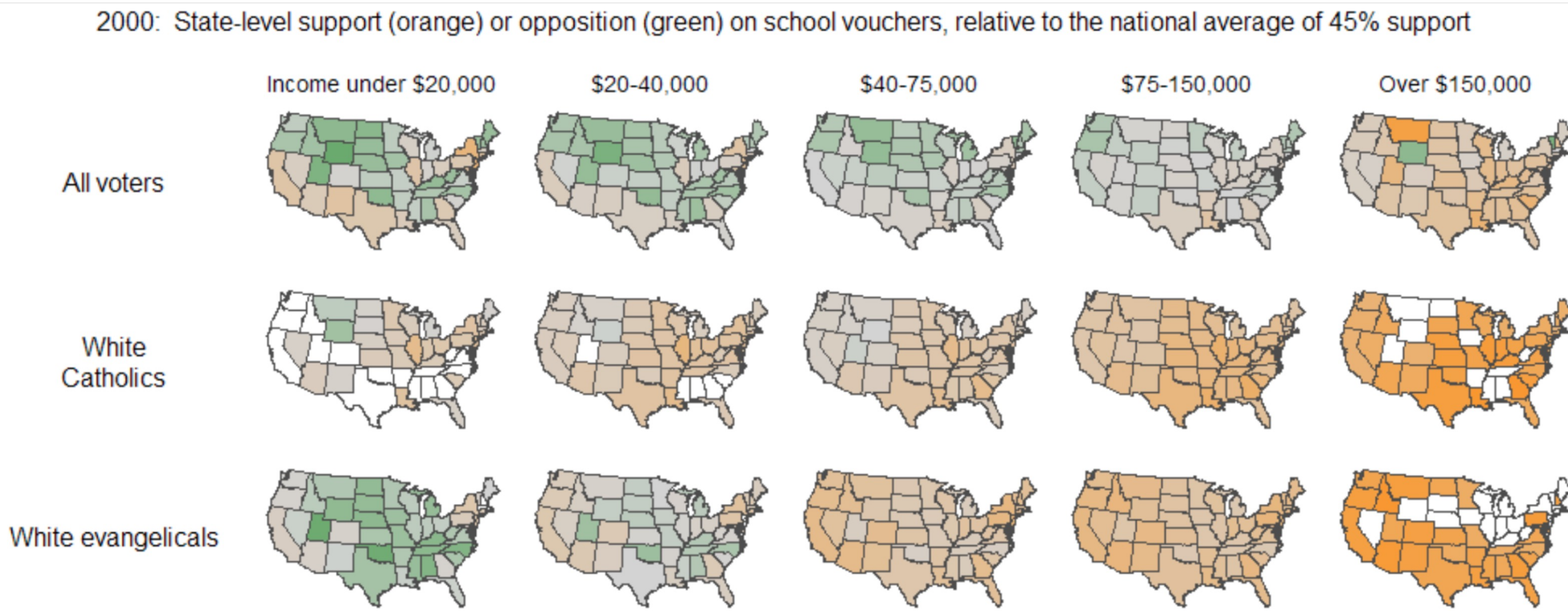
Use Conditioning To Aid Comparison

- Conditioning on male/female aligns points on x-axis
  - What does it reveal?
  - Why is this interesting?



Use Small Multiples To Aid Comparison

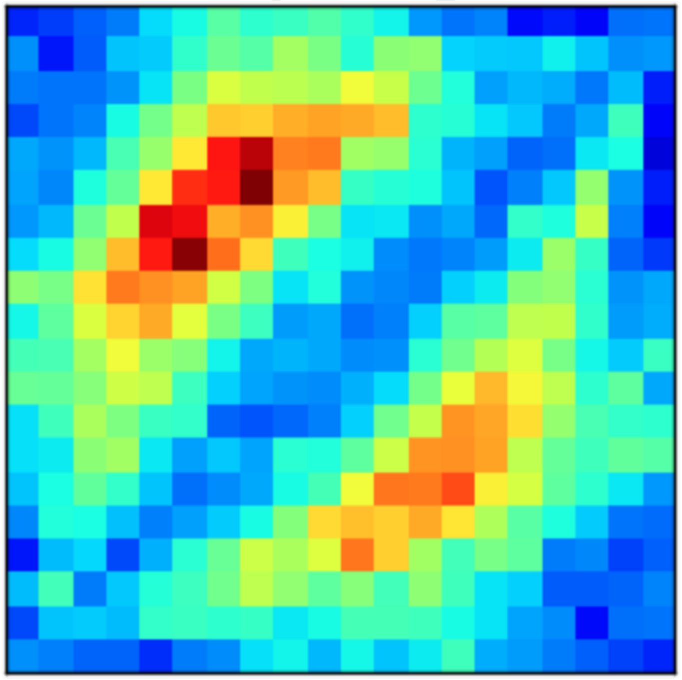
- Faceted plots that share scales are easy to compare
  - https://statmodeling.stat.columbia.edu/2009/07/15/hard_sell_for_b/



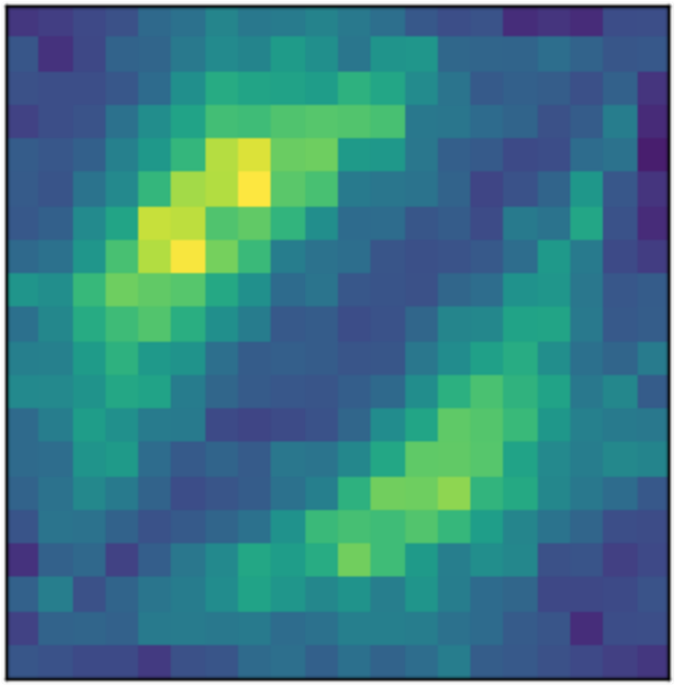
Principles of Perception

Color Choices Matter!

Jet Colormap



Viridis Colormap

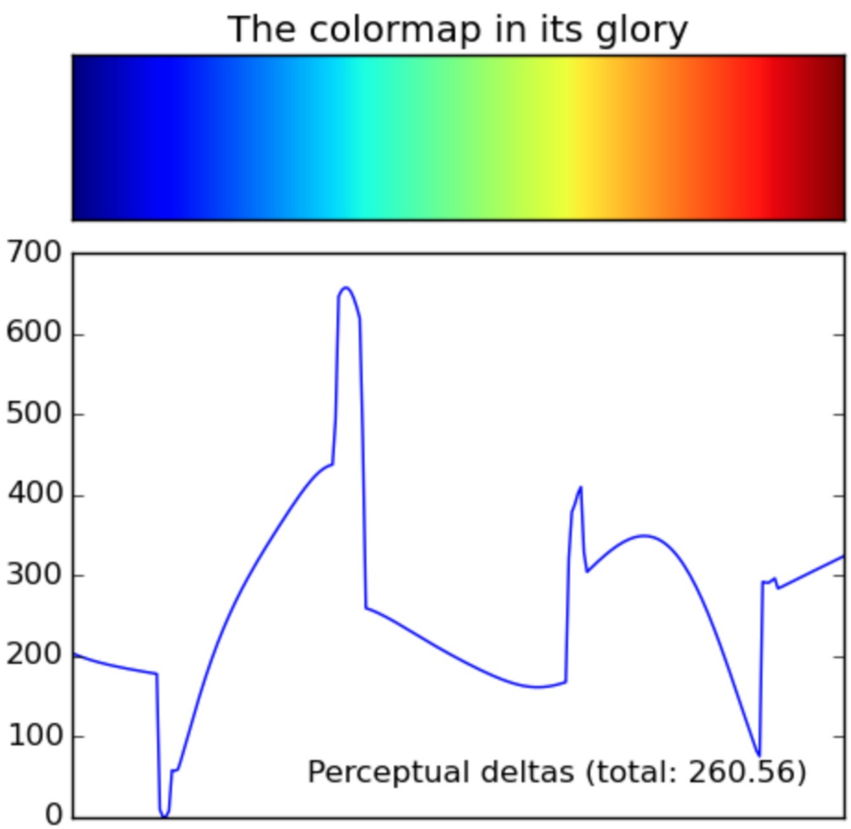


Use a Perceptually Uniform Color Map

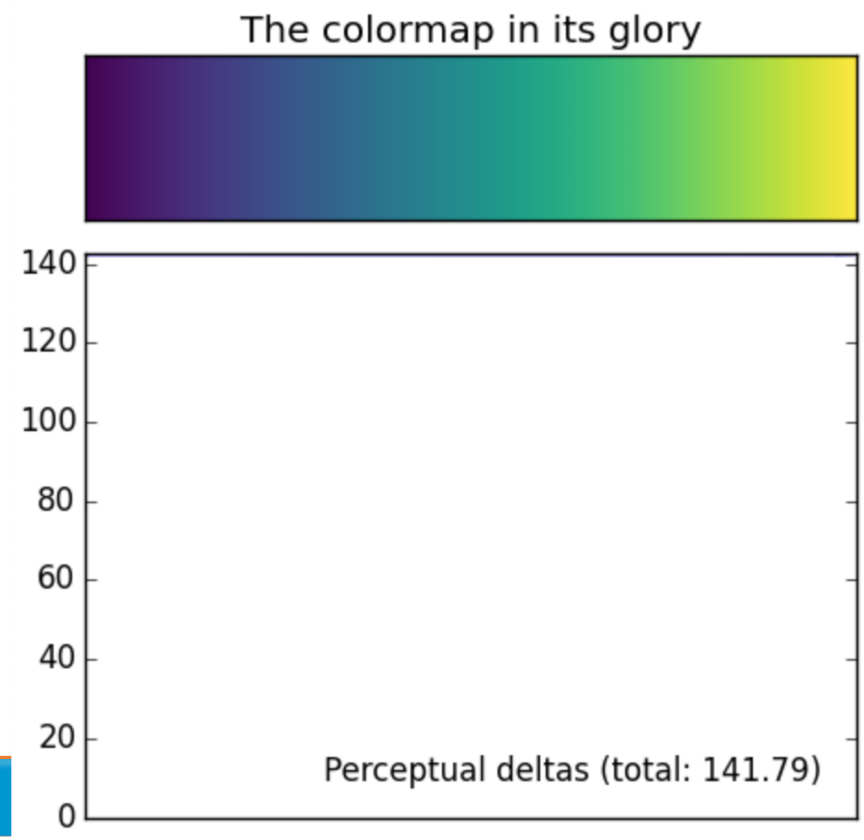
- Perceptually uniform:
  - Changing data from 0.1 to 0.2 appears similar to change from 0.8 to 0.9.
  - Measure by running experiments on people!
- Jet, the old matplotlib default, was far from uniform!
- Now fixed in MPL: https://bids.github.io/colormap/ (Eric Firing et al.)
- Also, avoid red + green since many people are colorblind

Use a Perceptually Uniform Color Map

Jet Colormap

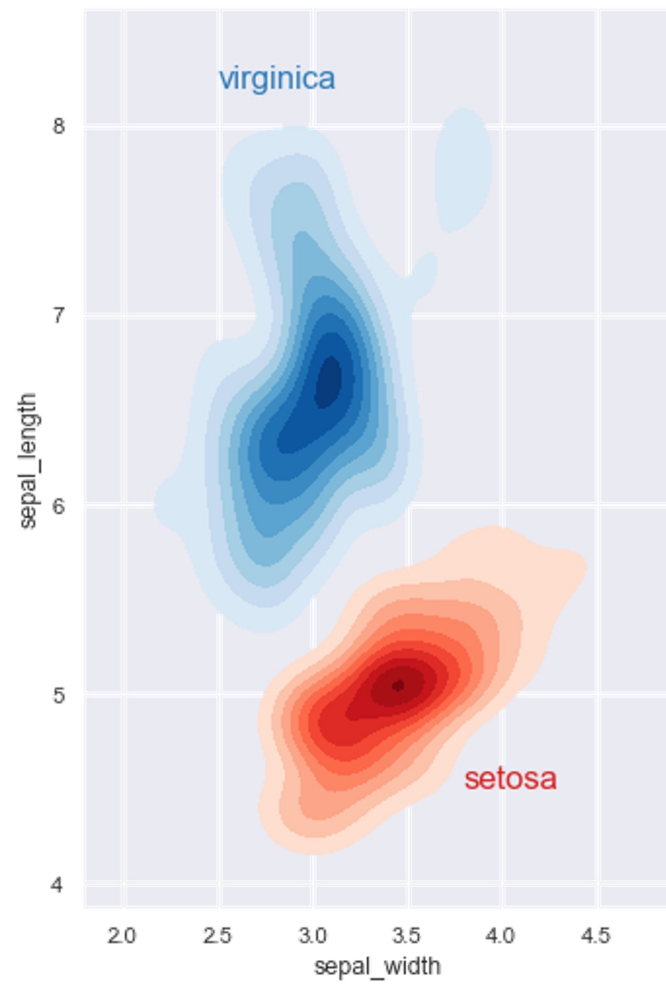


Viridis Colormap



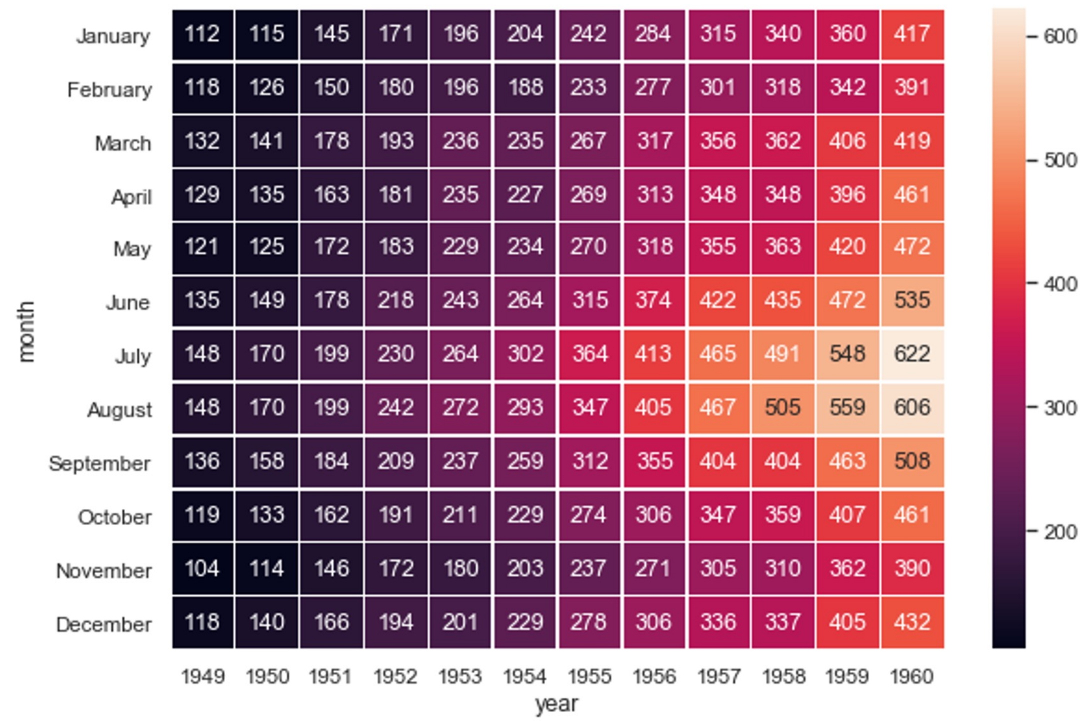
Use Color to Highlight Data Type

- Qualitative: Choose a qualitative scheme that makes it easy to distinguish between categories
- Quantitative: Choose a color scheme that implies magnitude.
- Plot on right has both!



Use Color to Highlight Data Type

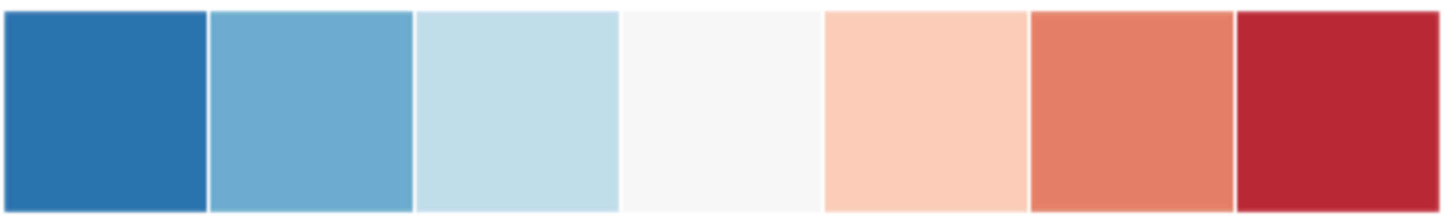
- Does the data progress from low to high?
- Use a sequential scheme where light colors are for more extreme values



Use Color to Highlight Data Type

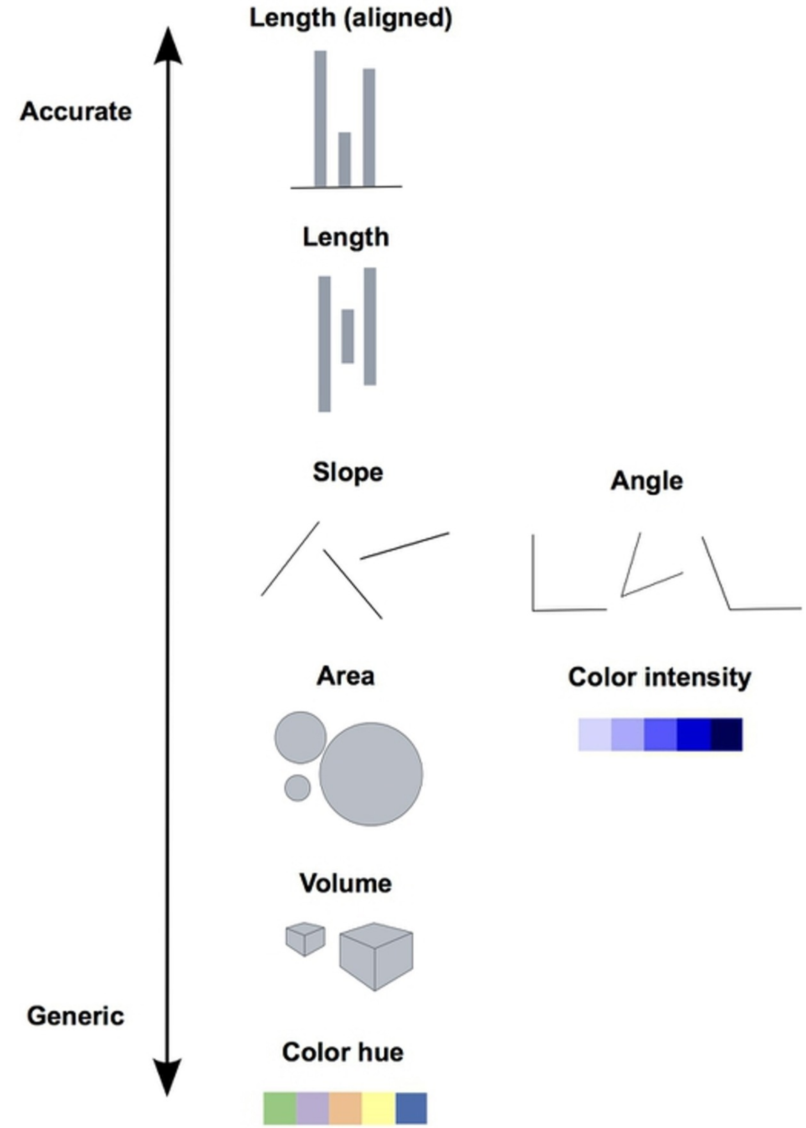
- Do both low and high value deserve equal emphasis? Use a diverging scheme where light colors represent middle values

```
sns.palplot(sns.color_palette("RdBu_r", 7))
```



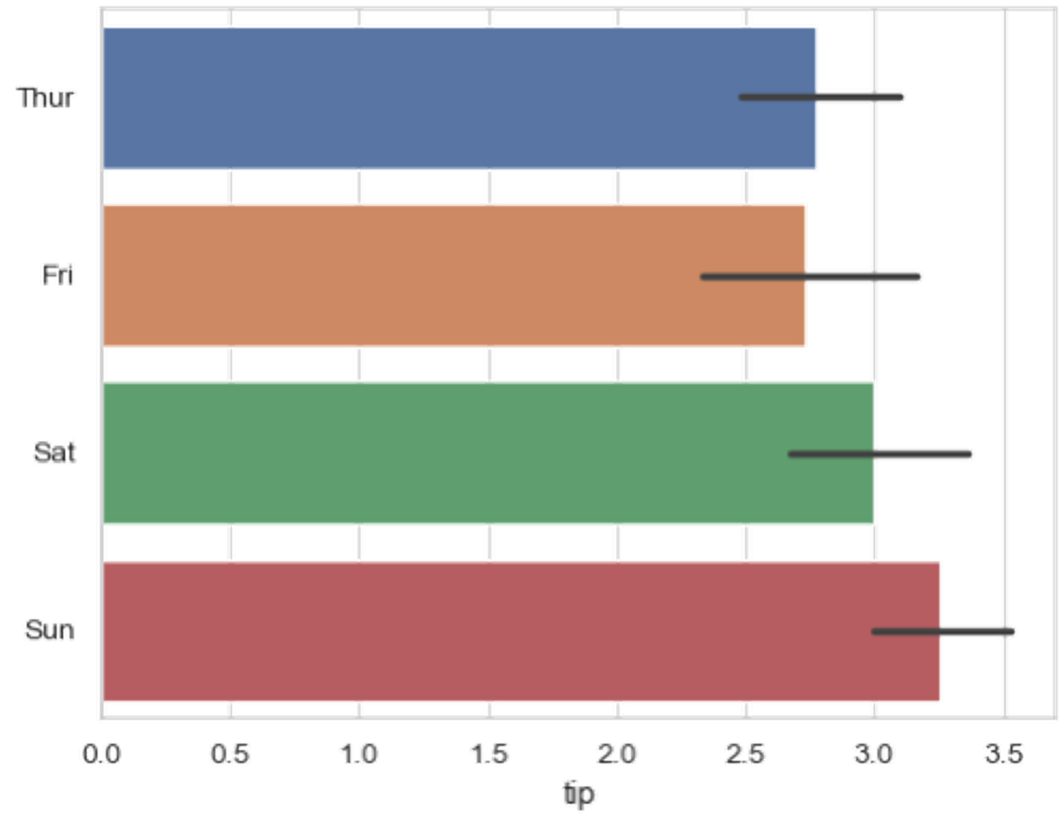
Not All Marks Are Good!

- Accuracy of judgements depend on the type of mark
- Aligned lengths most accurate
- Color least accurate



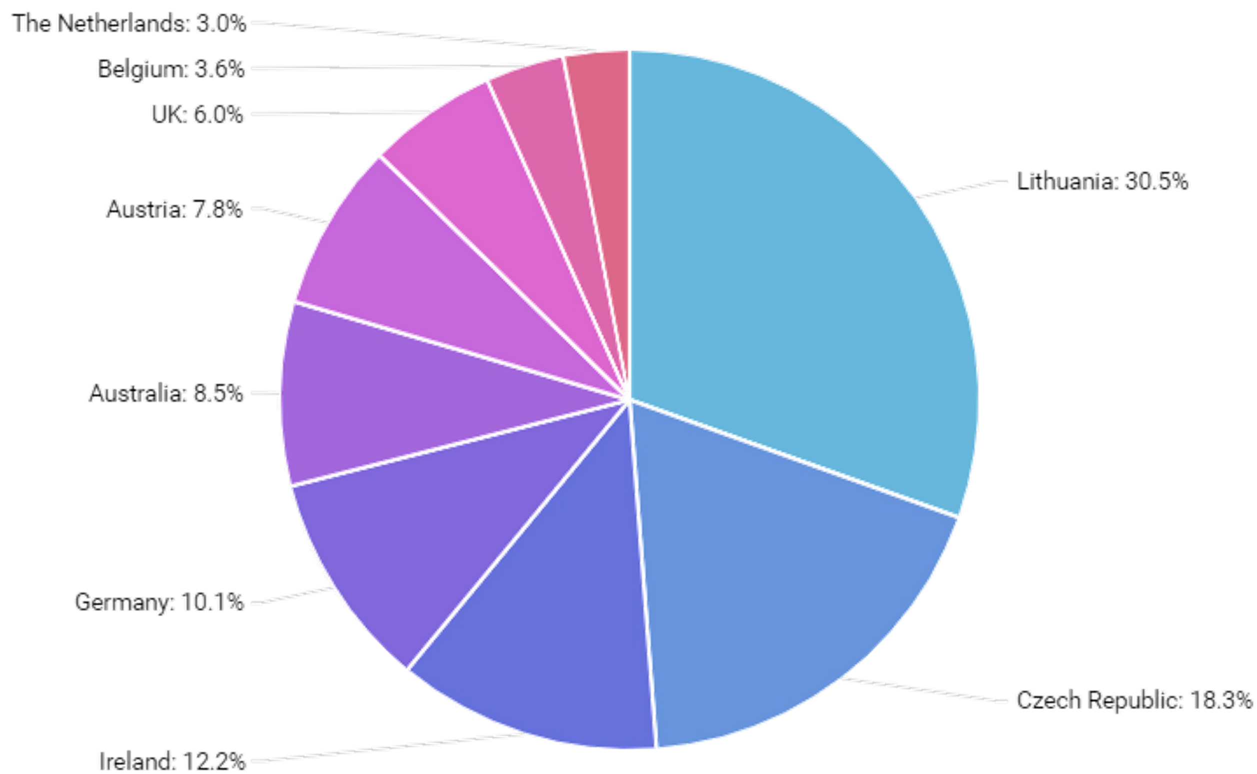
Lengths are Easy to Understand

- People can easily distinguish two different lengths
- E.g. Heights of bars in bar chart

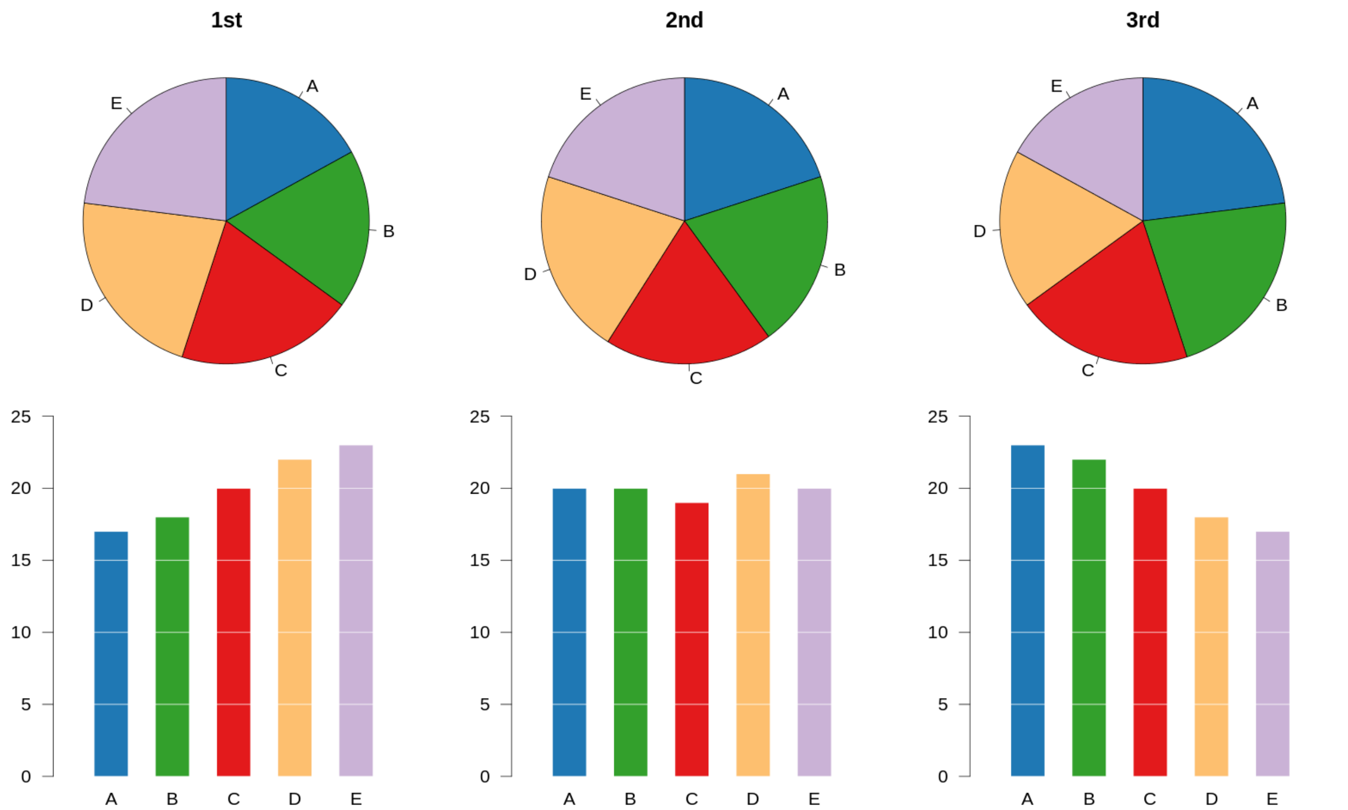


Angles are Hard to Understand

- Avoid pie charts!
- Angle judgements are inaccurate
- In general, underestimate size of larger angle

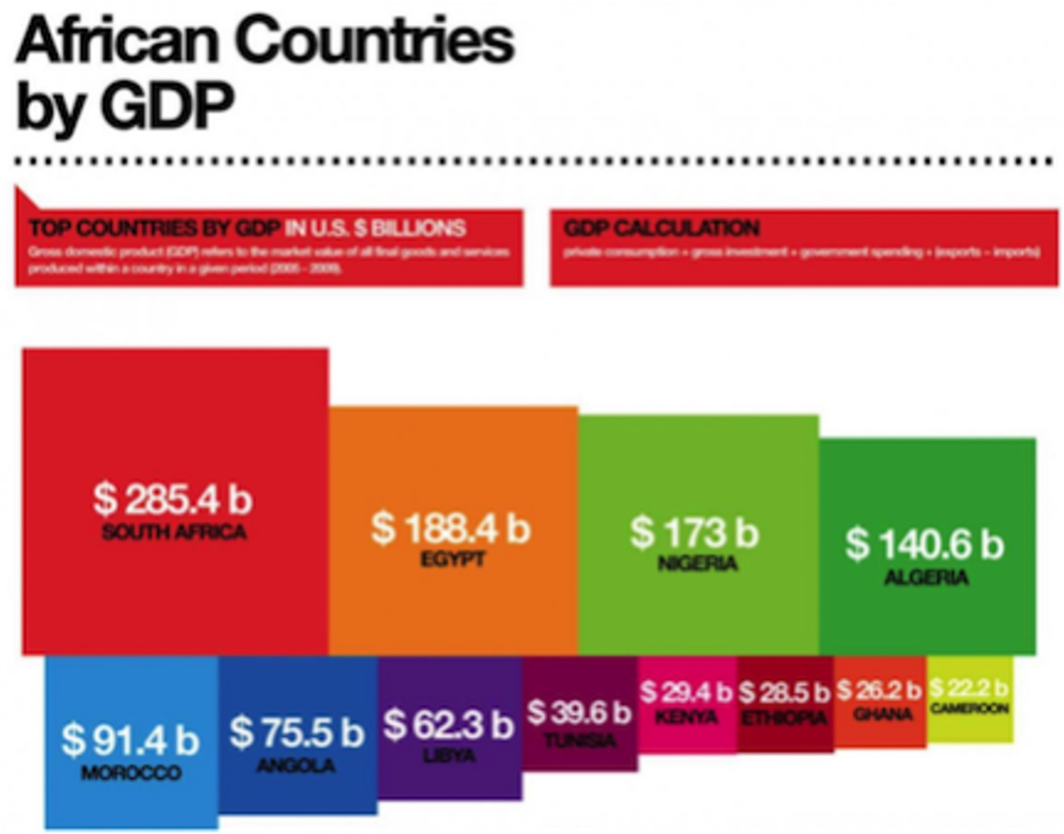


Bar vs Pie Chart



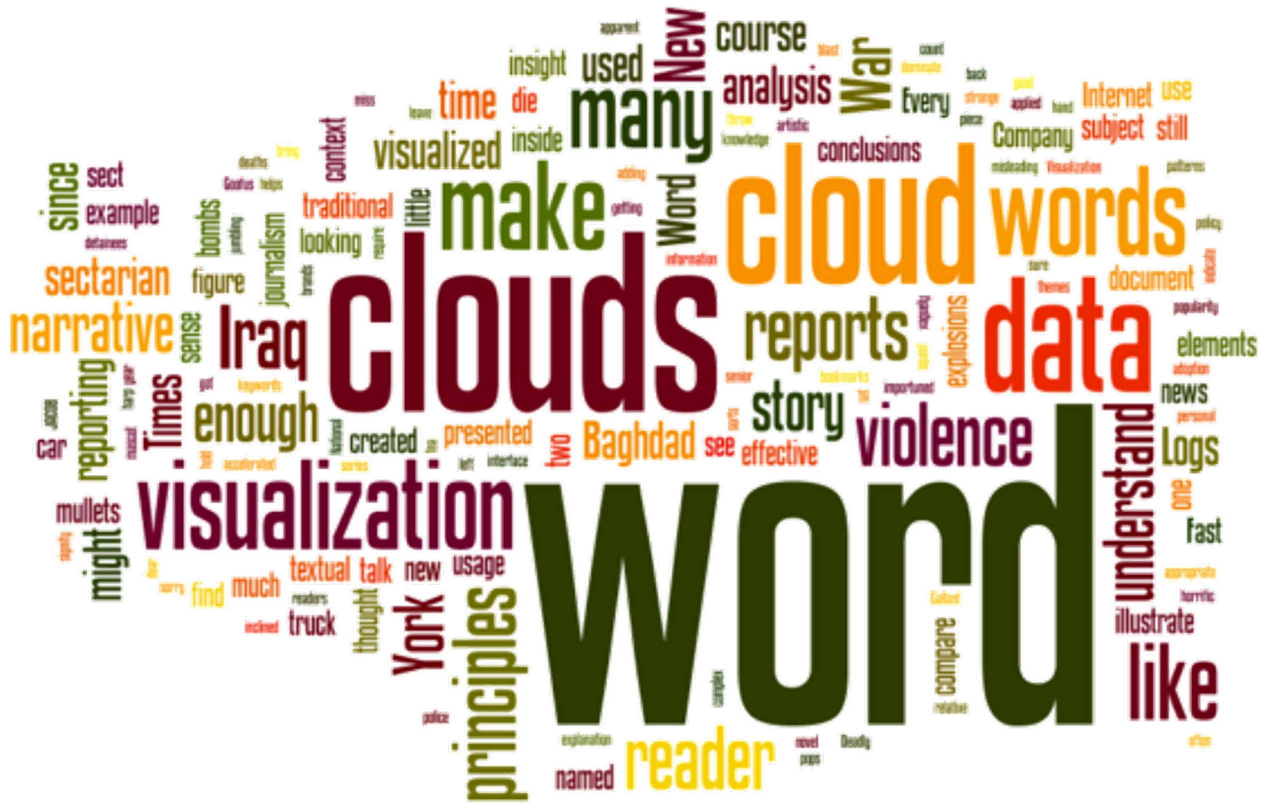
Areas are Hard to Understand

- Avoid area charts!
- Area judgements are inaccurate
- In general, underestimate size of larger area



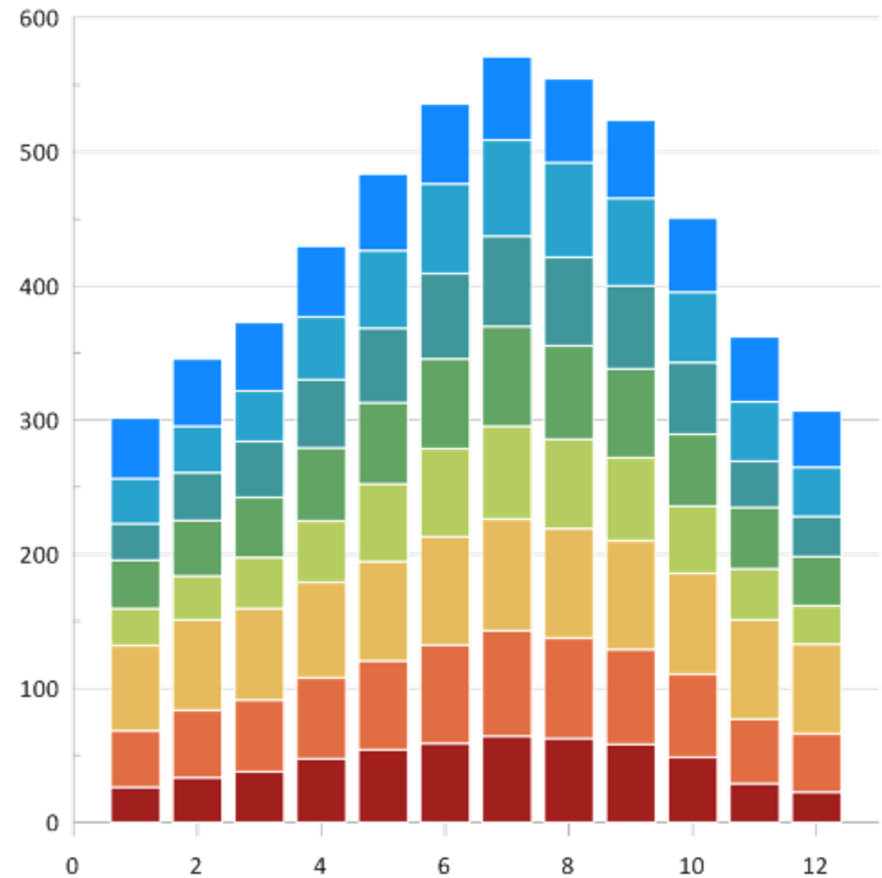
Areas are Hard to Understand

- Avoid word clouds!
- Hard to tell the “area” taken up by a word



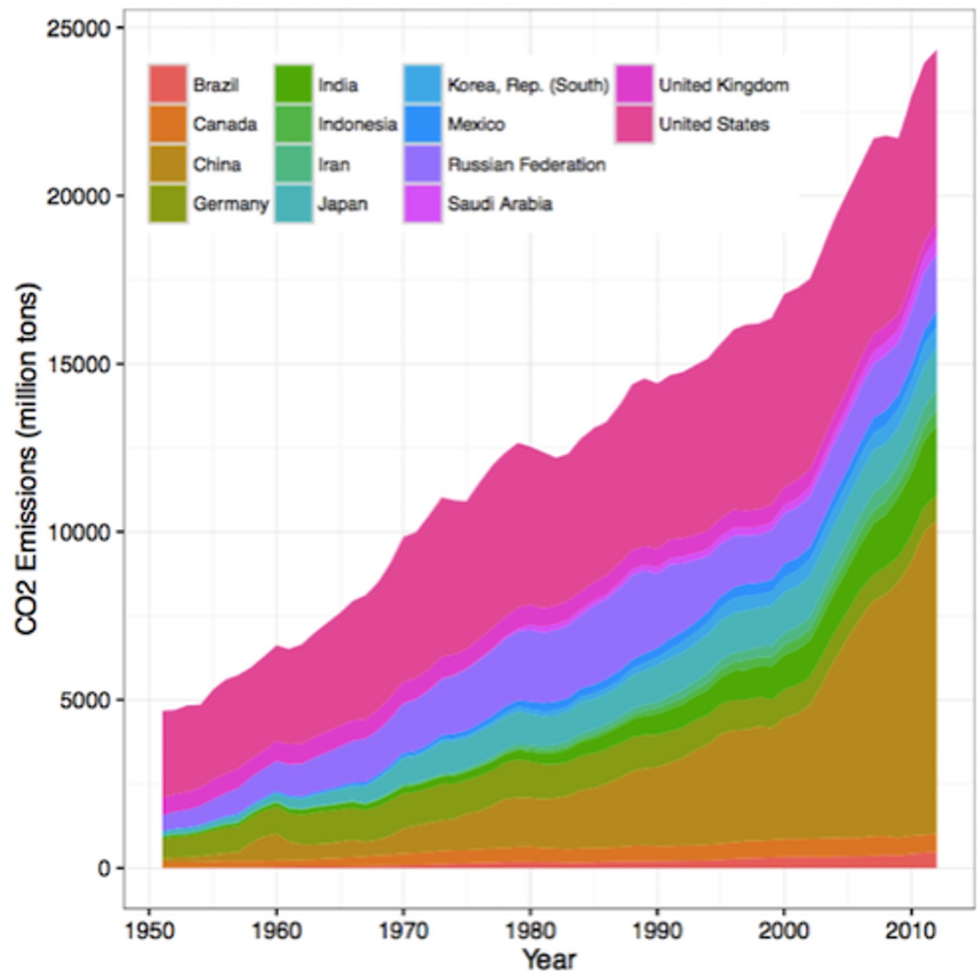
Avoid Jiggling Baseline

- Stacked bar charts / histograms hard to read because baseline moves
- Notice that top bars are all about the same height



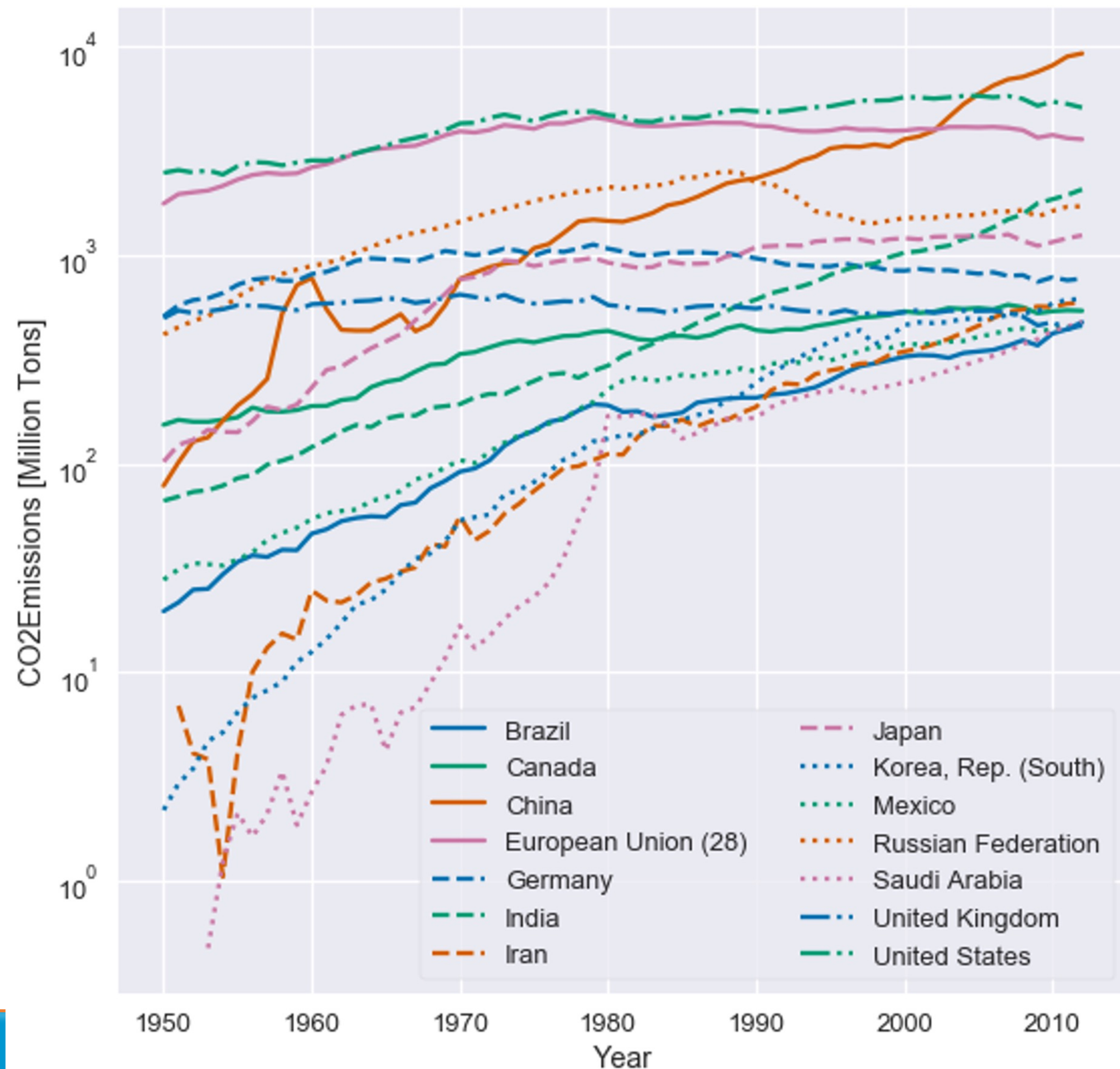
Avoid Jiggling Baseline

- Stacked area charts hard to read because baseline moves



Avoid Jiggling Baseline

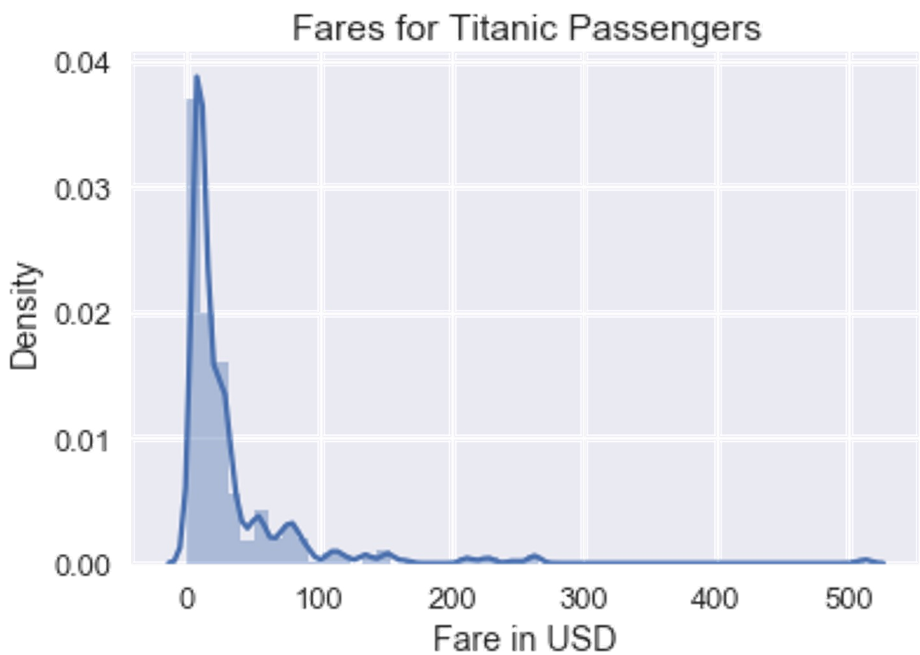
- Instead, plot lines themselves

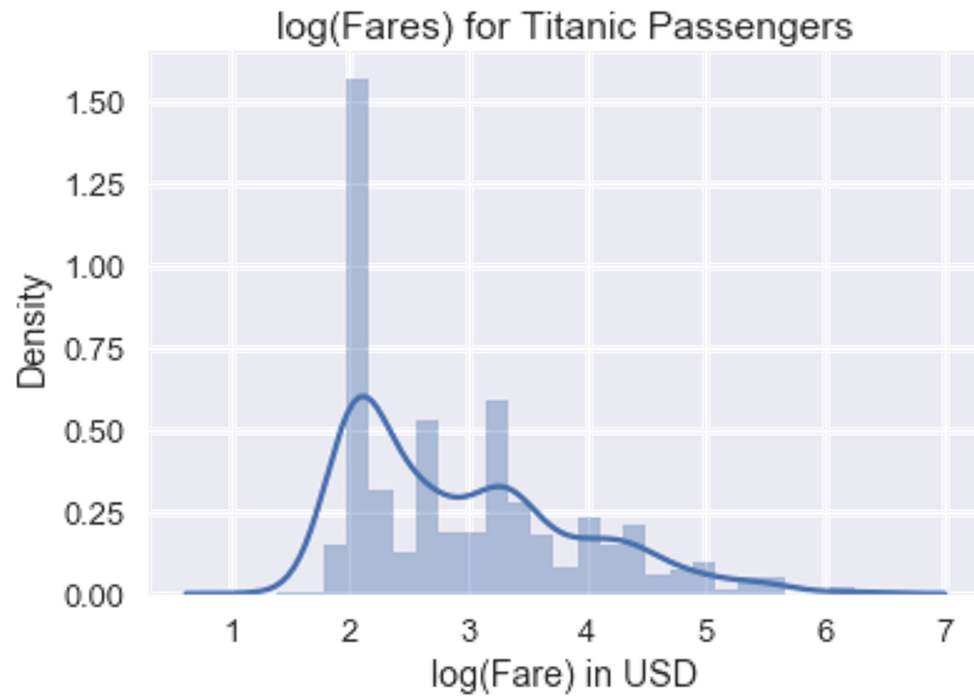


Principles of Transformation

Transforming Data Can Reveal Patterns

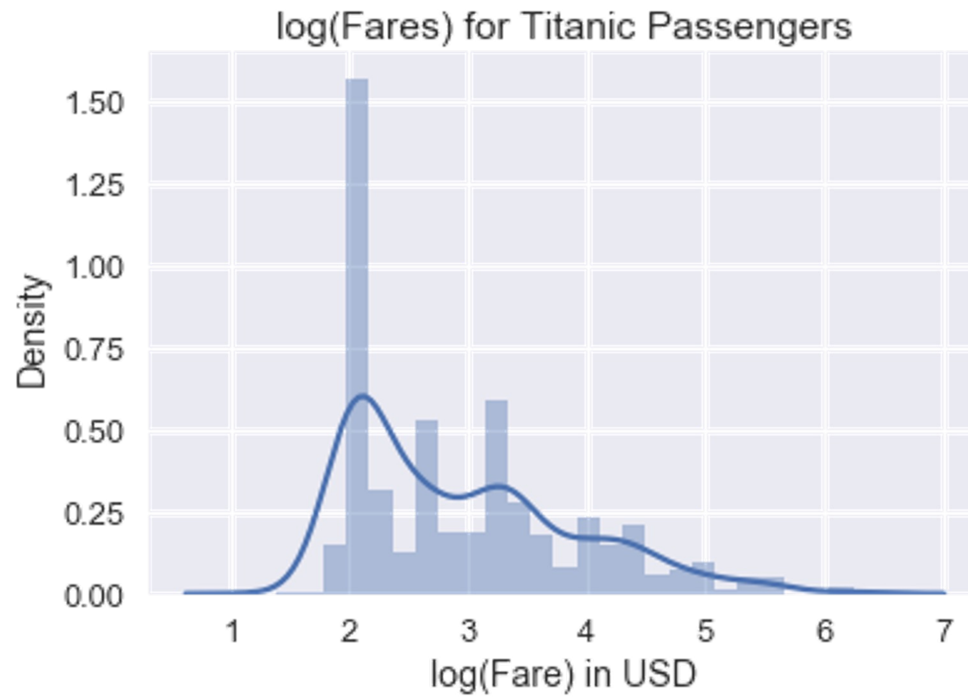
- When data are heavy tailed, useful to take the log and replot





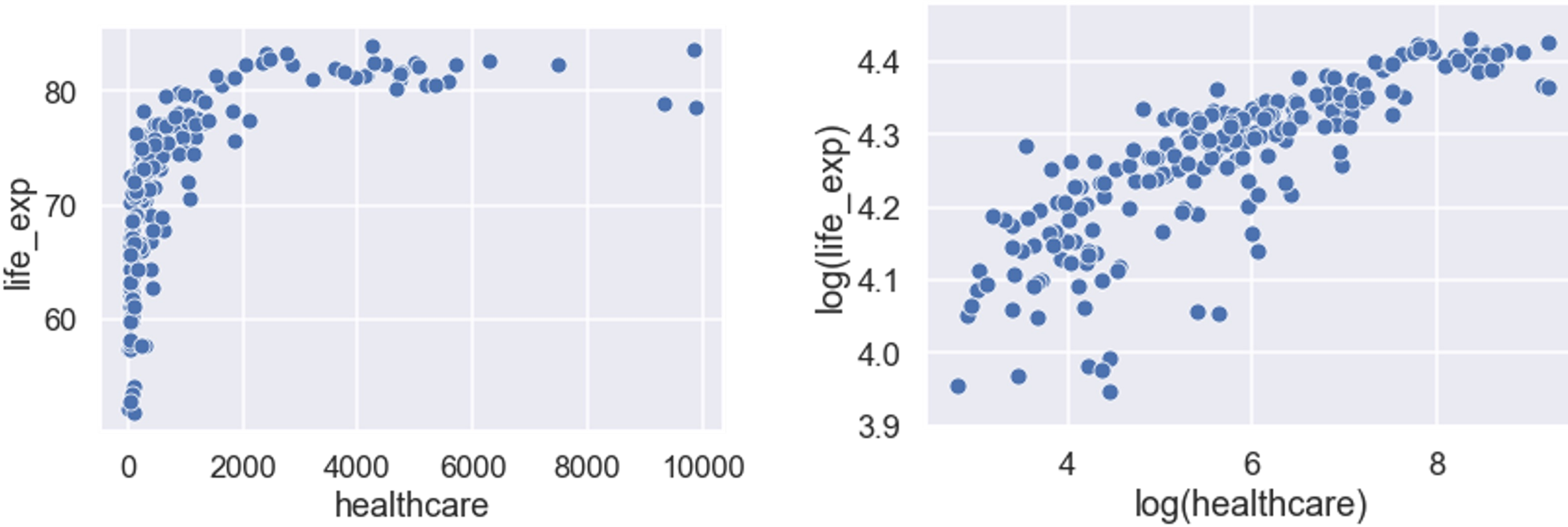
Transforming Data Can Reveal Patterns

- Shows a mode when log(fare) = 2 and a smaller mode at 3.4.
- What do these correspond to in actual dollars?
- exp(2) = $7.4
- exp(3.4) = $30



Transforming Data Can Reveal Patterns

- Log of nonlinear data can reveal pattern in scatter plot!

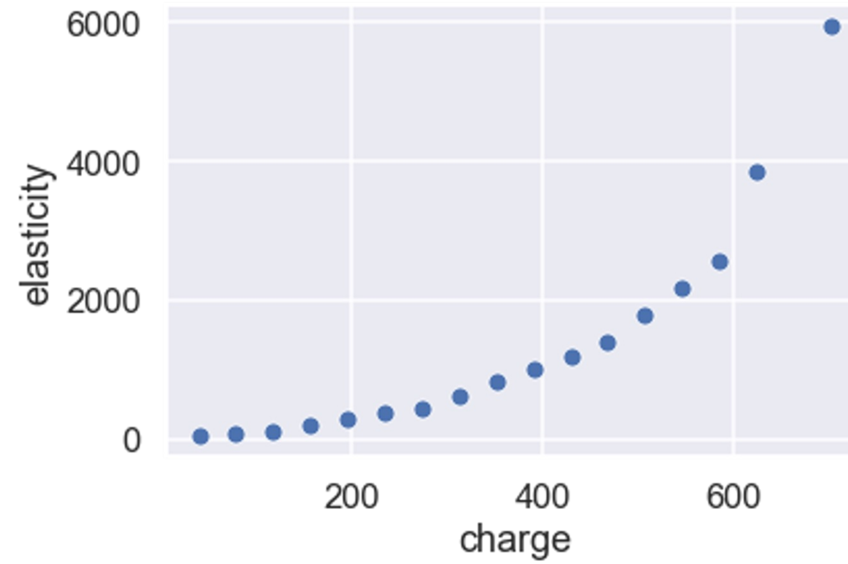


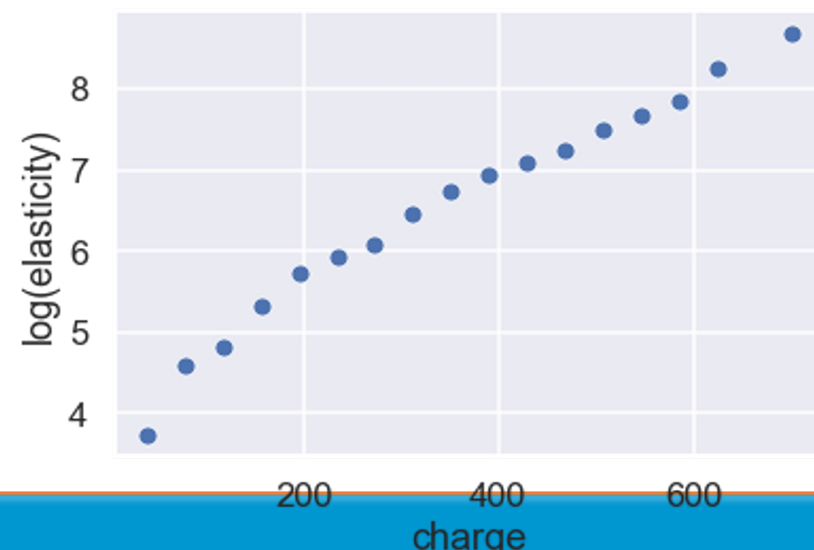
Log of y-values

- Fit line to log of y-values:

$$\log y = ax + b$$
$$y = e^{ax+b}$$
$$y = e^{ax}e^{b}$$
$$y = Ce^{ax}$$

- Linear relationship after log of y-values implies exponential model for original plot





Log of both x and y-values

- Fit line to log of x and y-values:

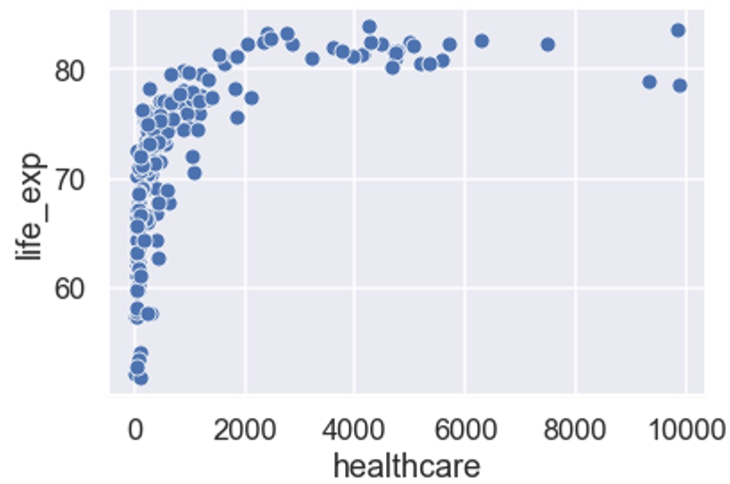
$$\log y = a \cdot \log x + b$$

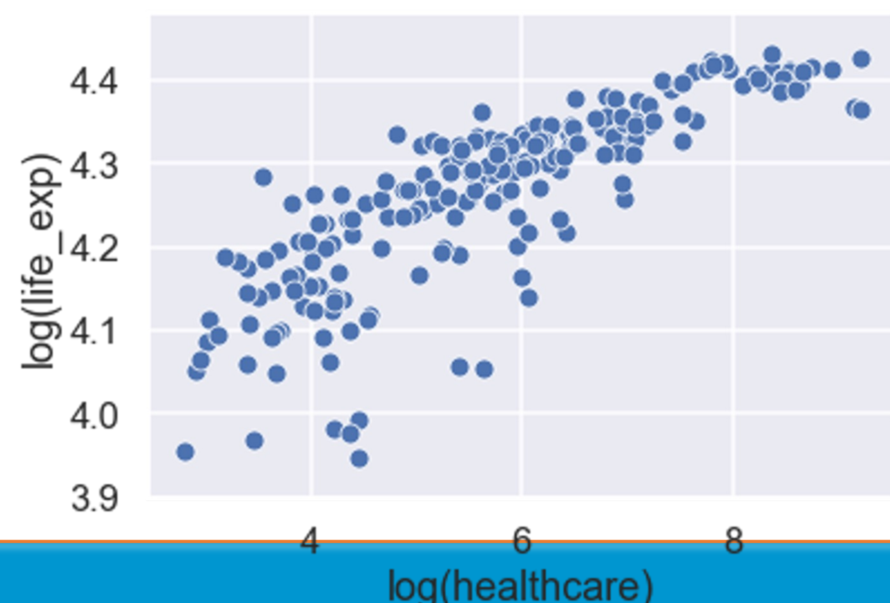
$$y = e^{a \cdot \log x + b}$$

$$y = Ce^{a \cdot \log x}$$

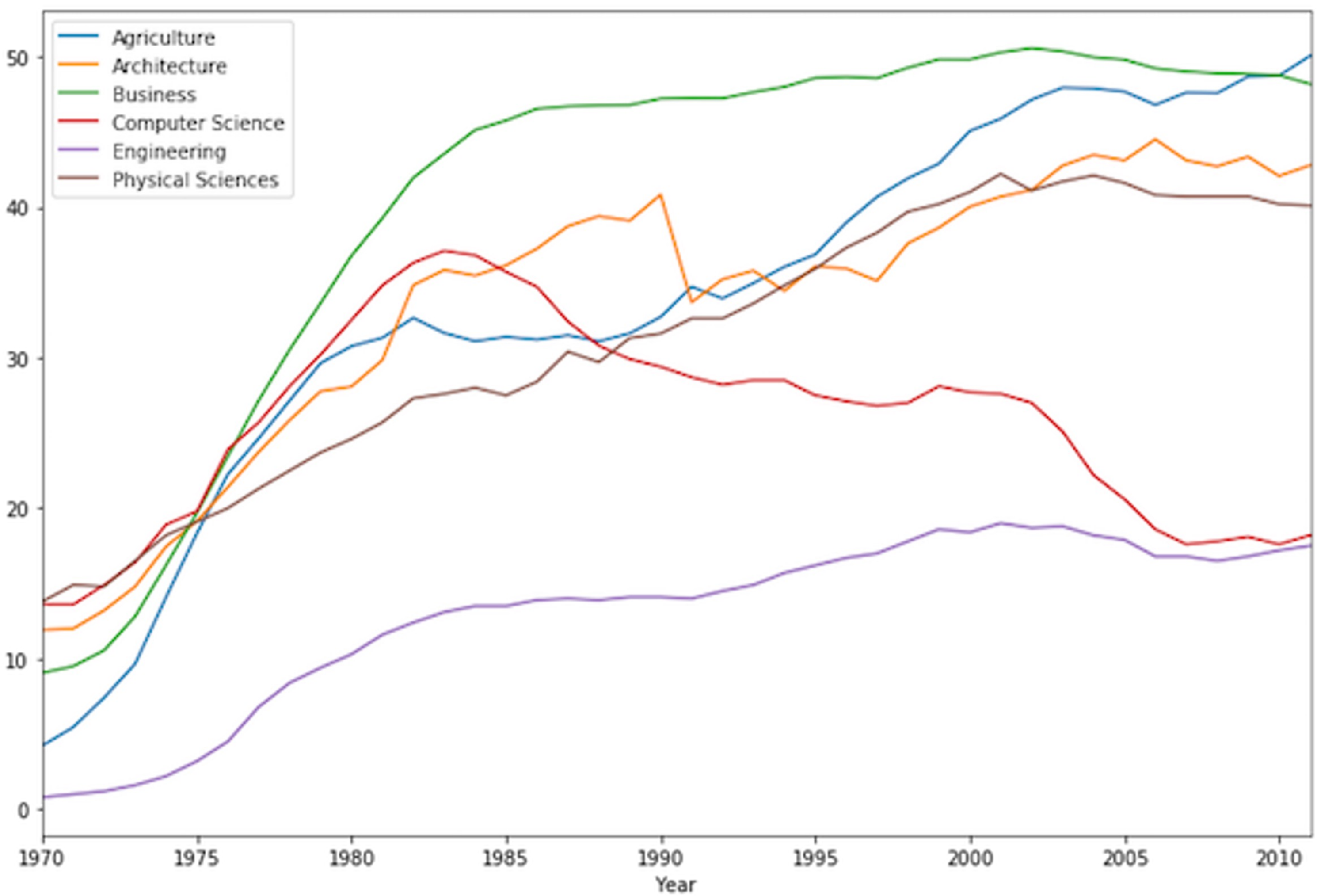
$$y = Cx^a$$

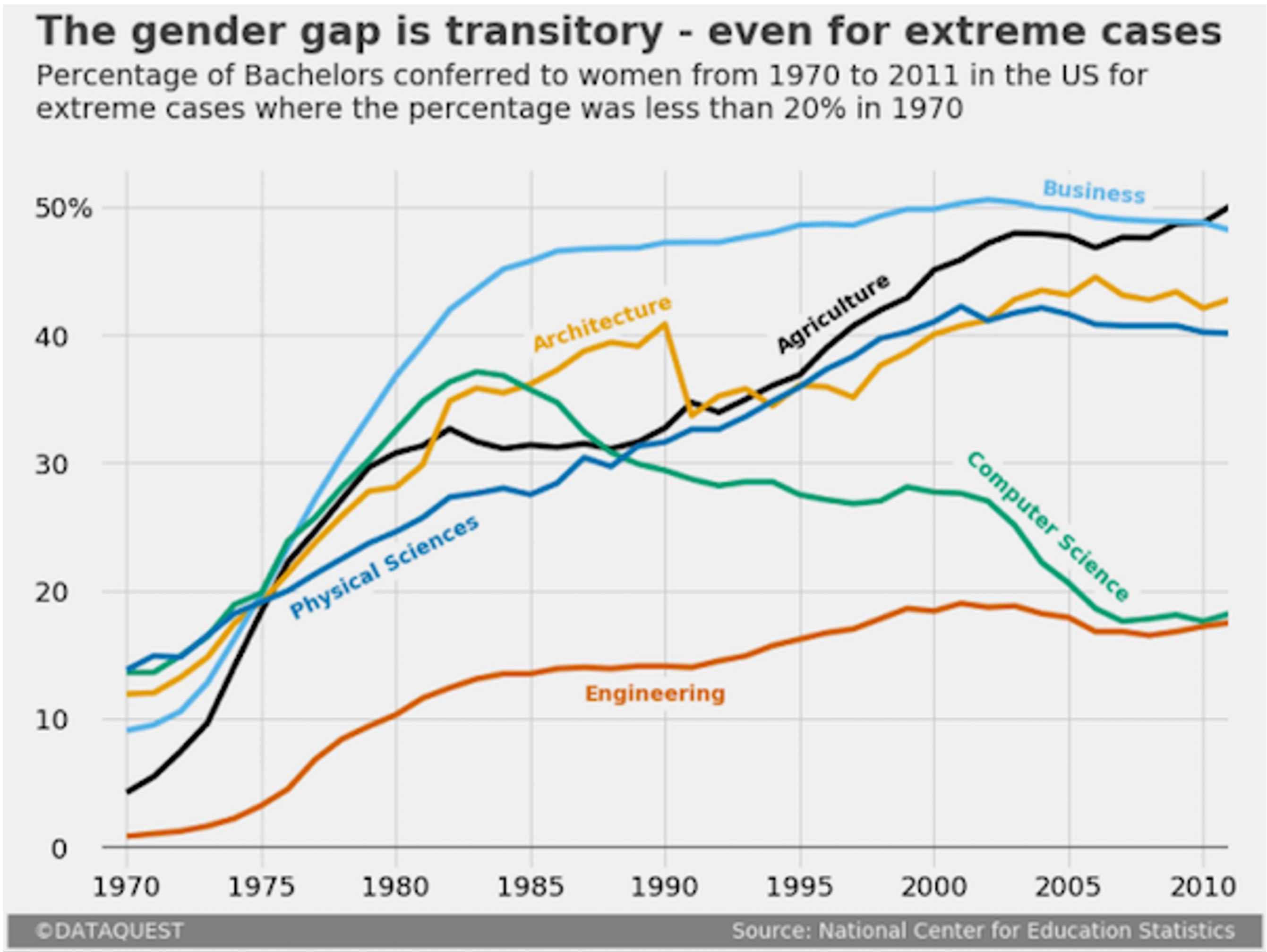
- Linear relationship after log of x and y-values implies polynomial model for original plot





Principles of Context


Agriculture
Architecture
Business
Computer Science
Engineering
Physical Sciences
50
40
30
20
10
0
1970
1975
1980
1985
1990
1995
2000
2005
2010
Year


The gender gap is transitory - even for extreme cases
Percentage of Bachelors conferred to women from 1970 to 2011 in the US for extreme cases where the percentage was less than 20% in 1970
50%
40
30
20
10
0
1970
1975
1980
1985
1990
1995
2000
2005
2010
Business
Agriculture
Architecture
Computer Science
Physical Sciences
Engineering
©DATAQUEST
Source: National Center for Education Statistics

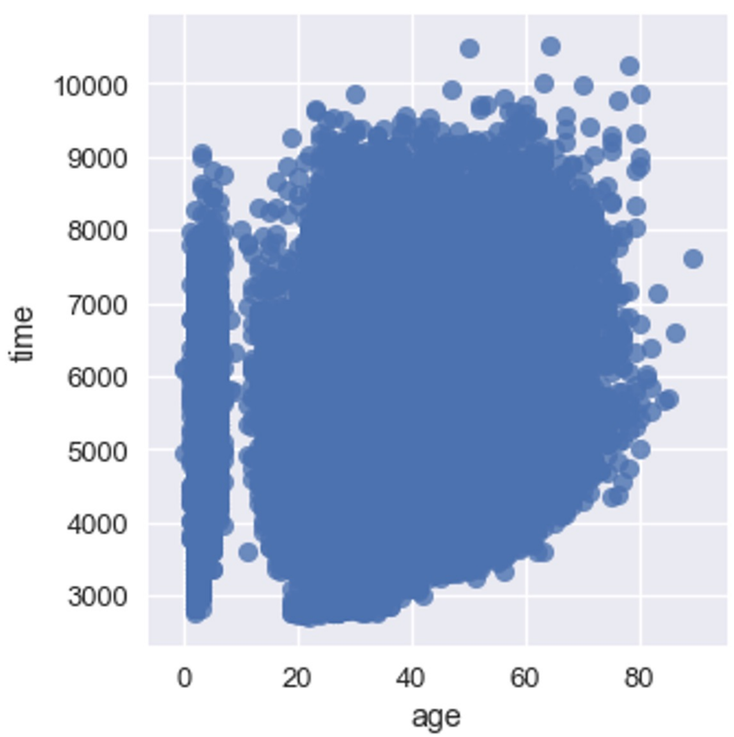
Add Context Directly to Plot

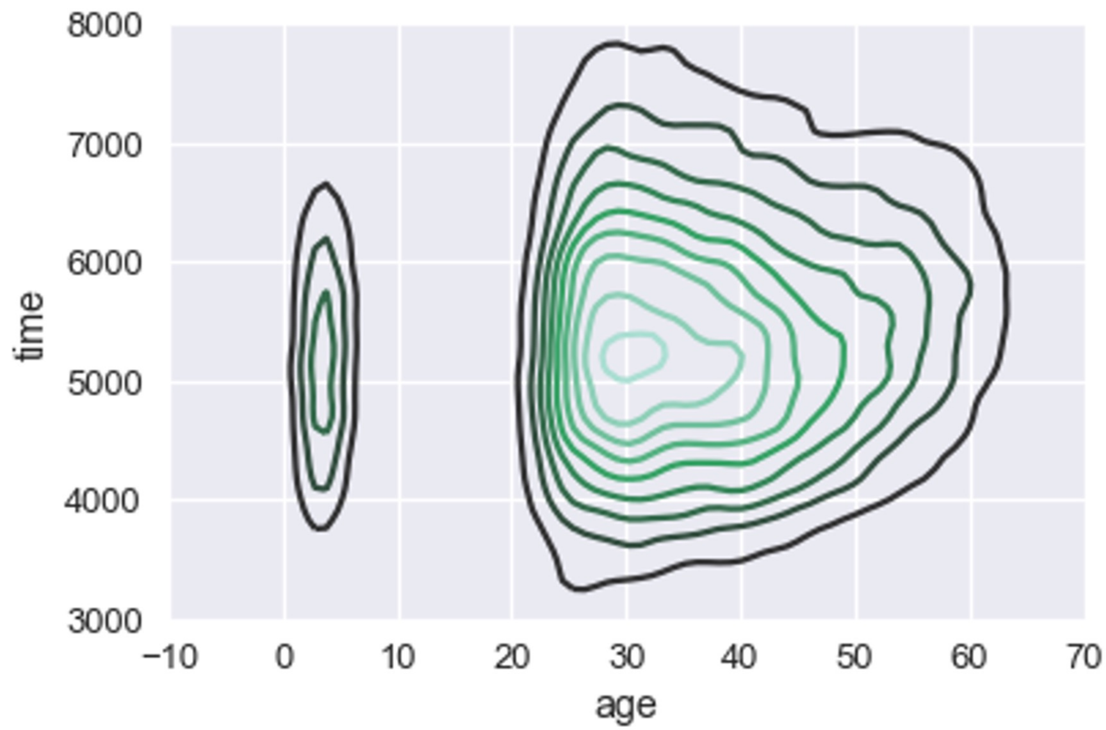
A publication-ready plot needs:

- Informative title (takeaway, not description)
  - “Older passengers spend more on plane tickets” instead of “Scatter plot of price vs. age”.
- Axis labels
- Reference lines and markers for important values
- Labels for unusual points
- Captions that describe data

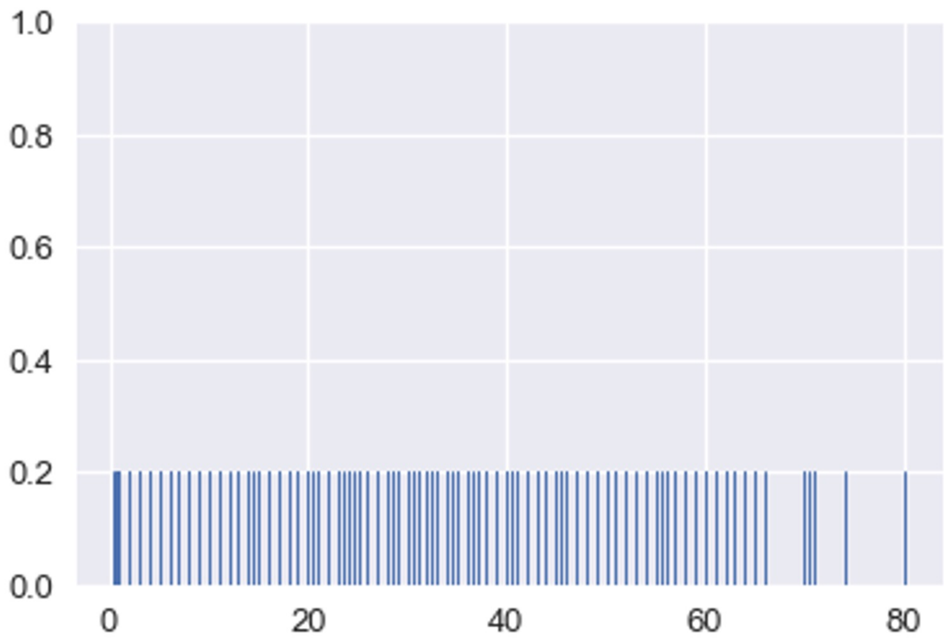
Principles of Smoothing

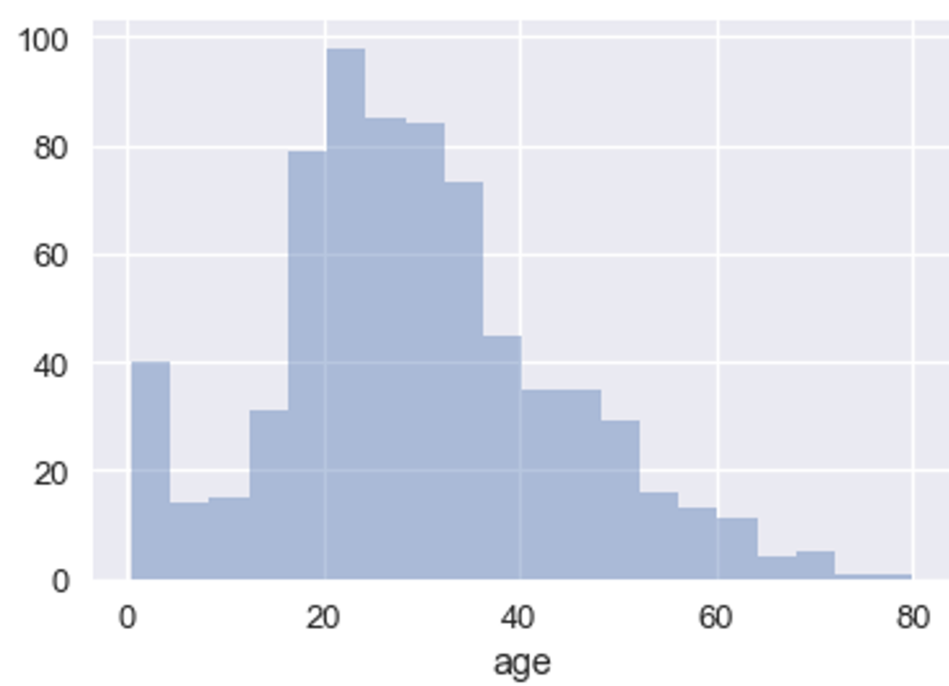
Apply Smoothing for Large Datasets



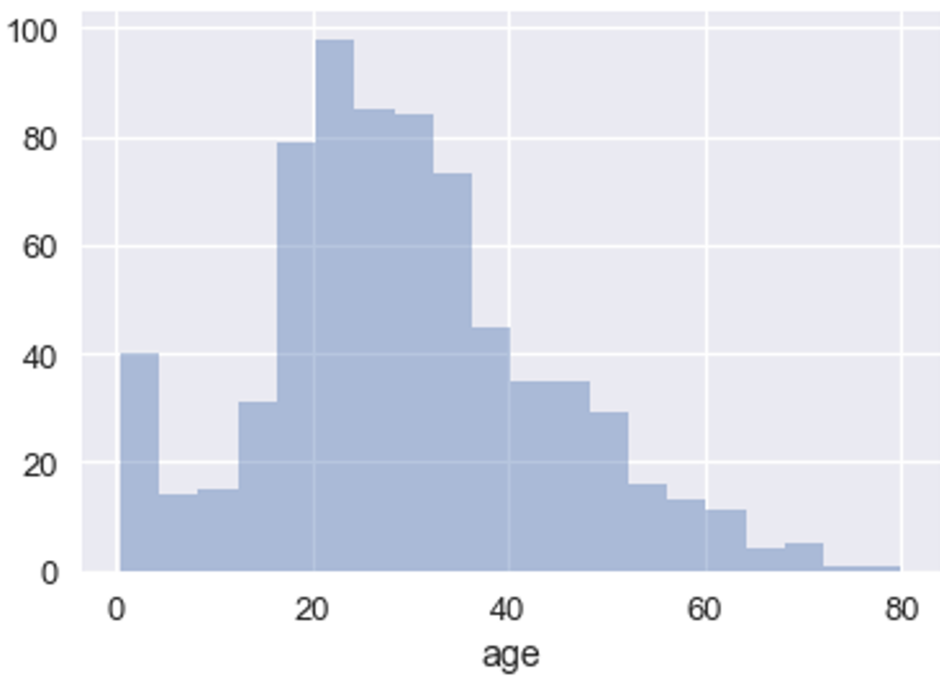


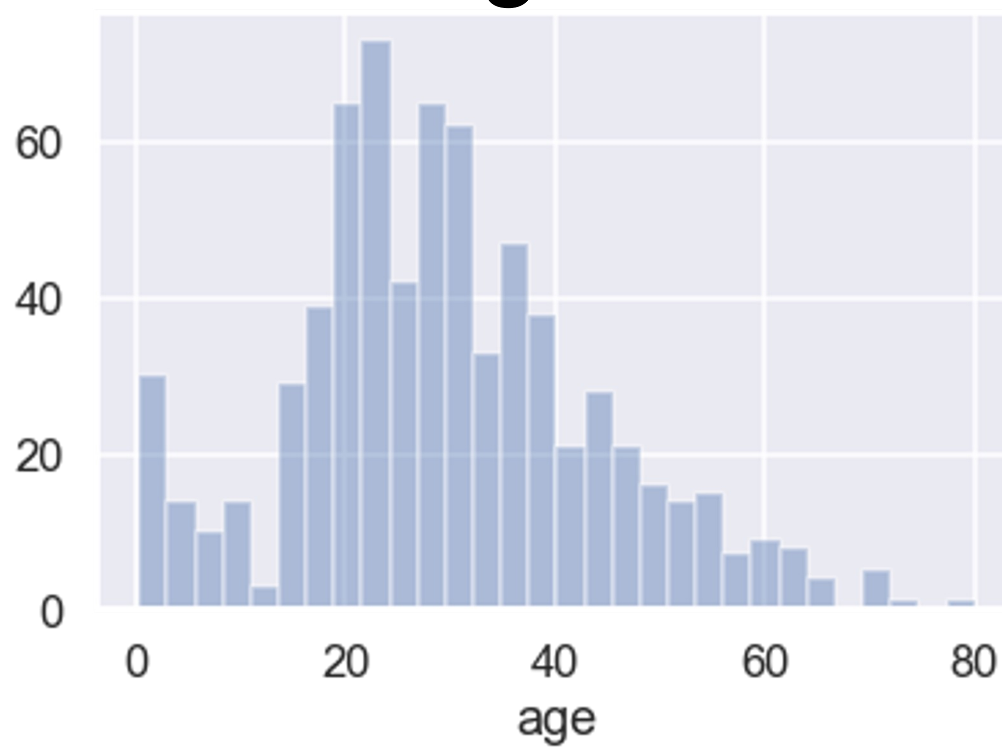
A Histogram is a Smoothed Rug Plot





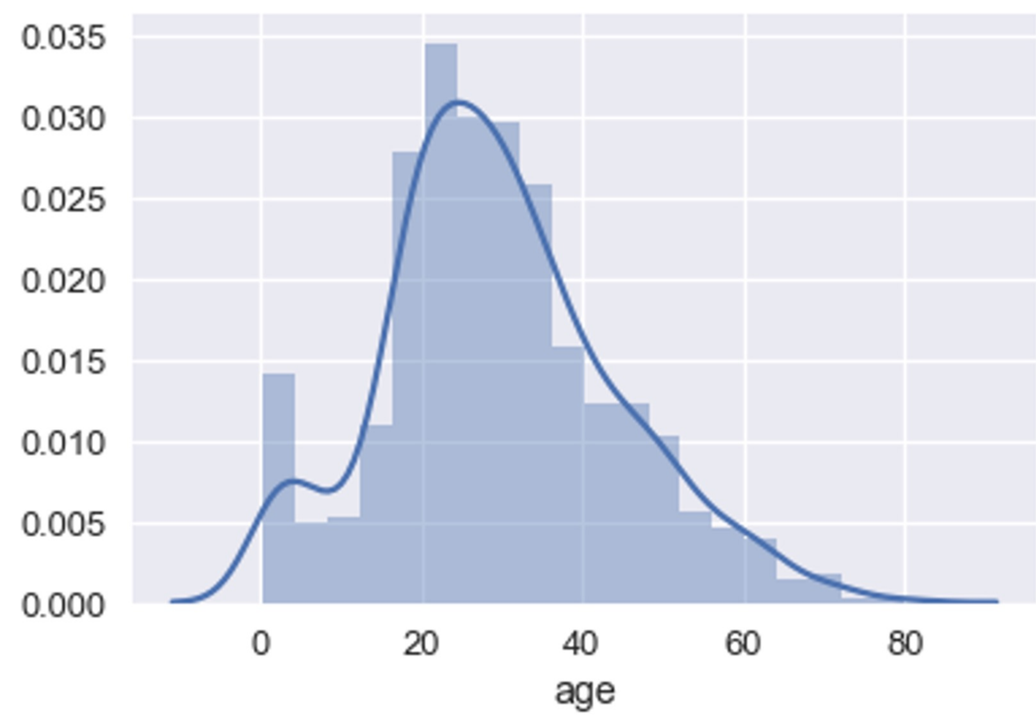
Smoothing Needs Tuning





Kernel Density Estimation (KDE)

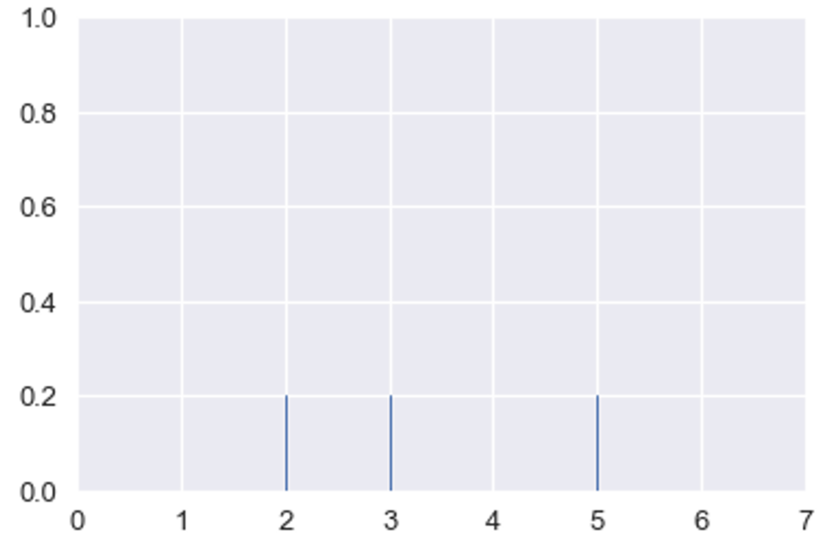
- Sophisticated smoothing technique
- Used to estimate a probability density function from a set of data

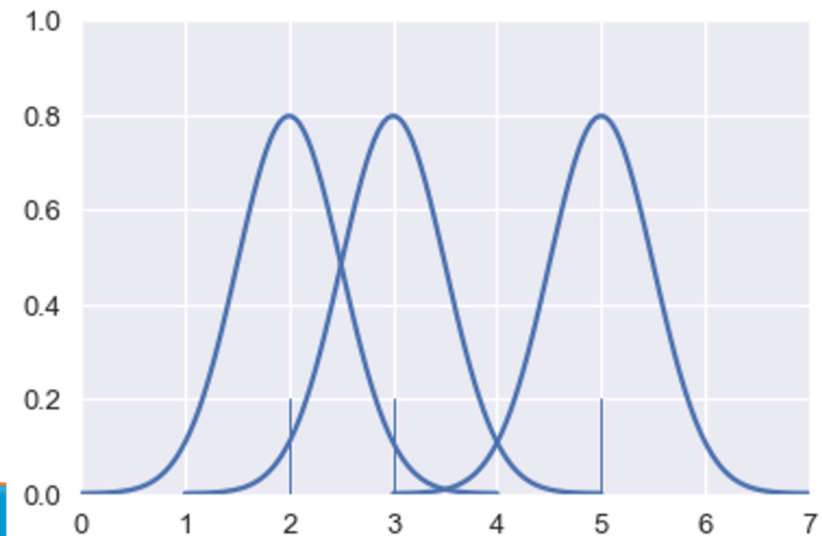


Kernel Density Estimation

Intuition:

1. Place a “kernel” at each data point

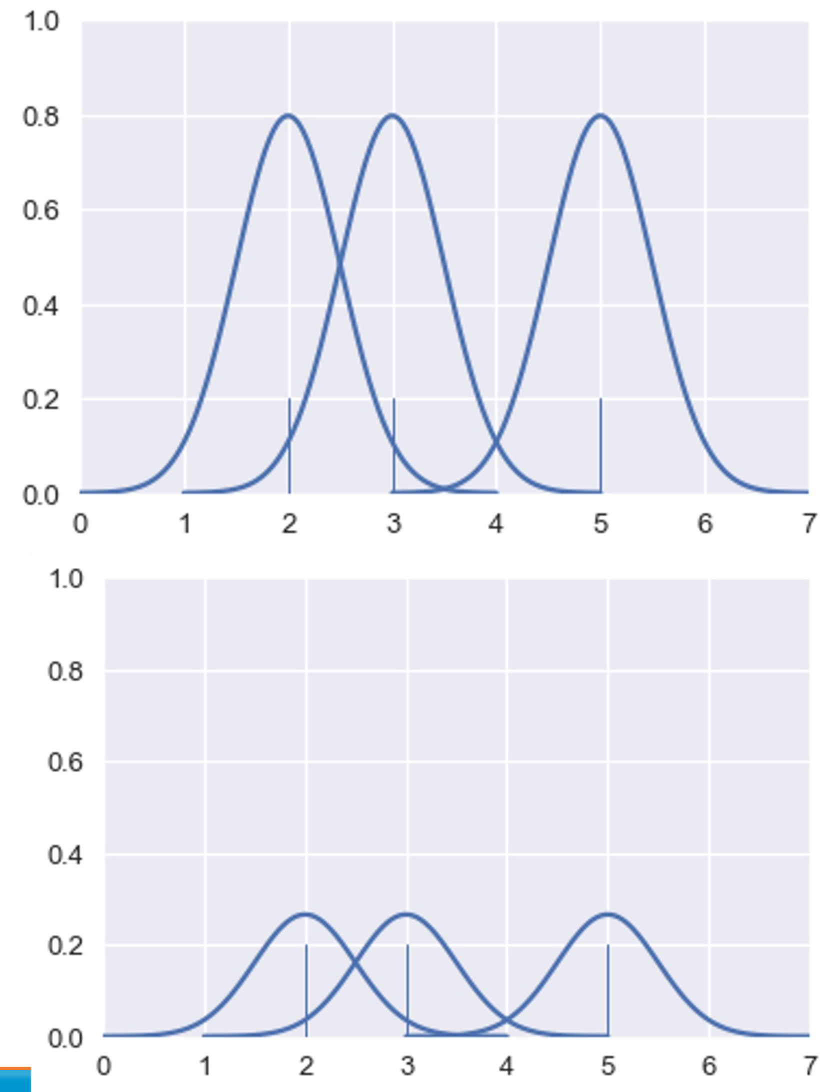




Kernel Density Estimation

Intuition:

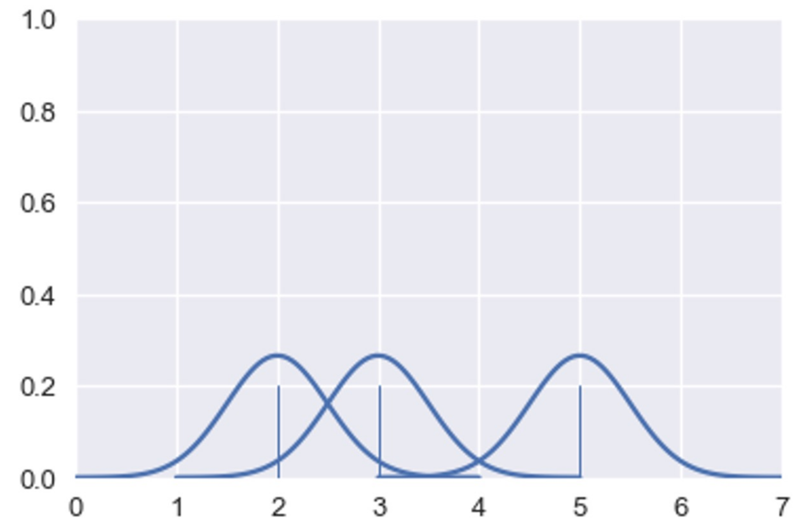
1. Place a “kernel” at each data point
2. Normalize kernels so that total area = 1

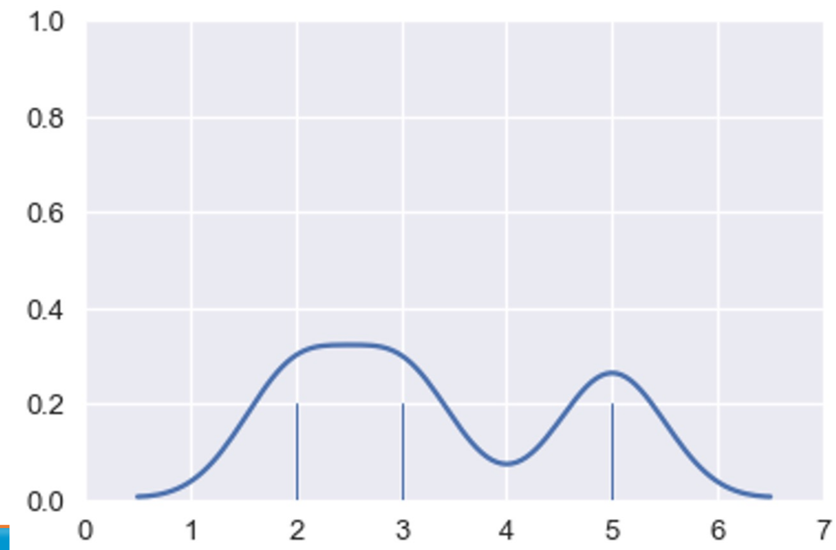


Kernel Density Estimation

Intuition:

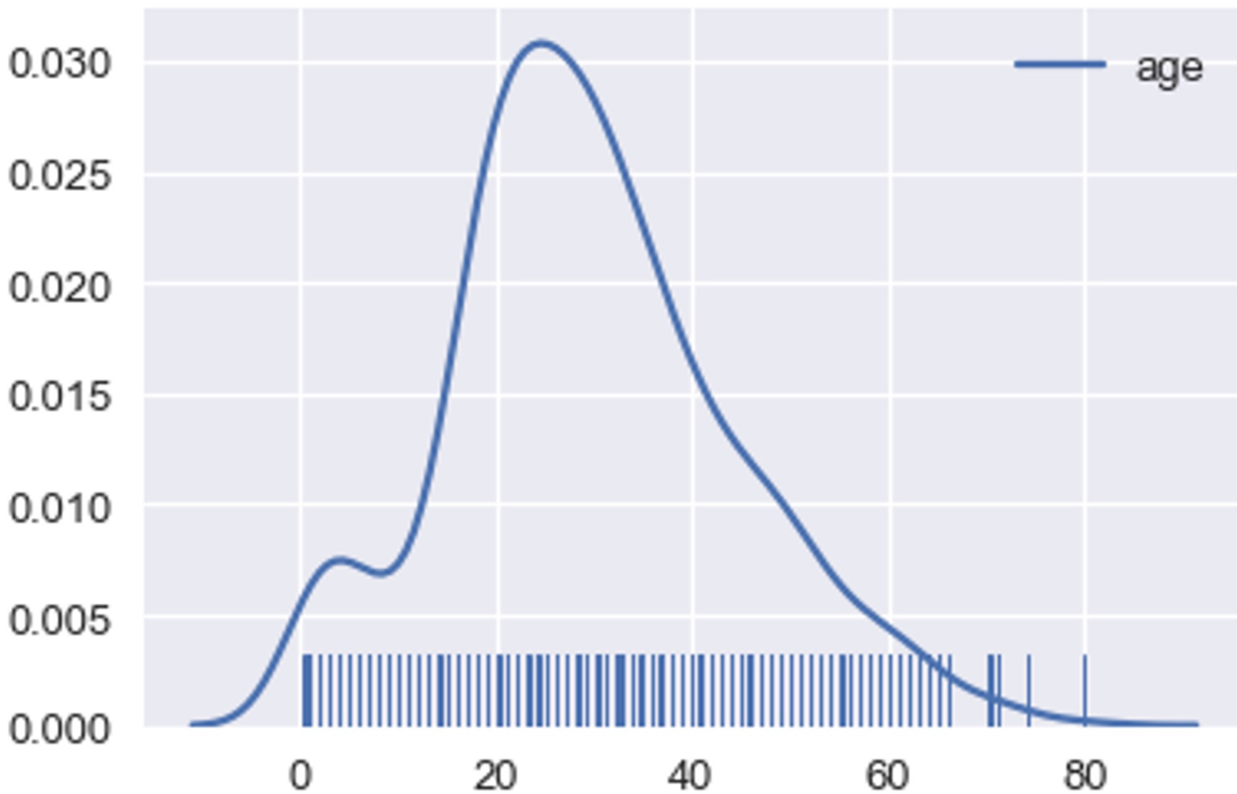
1. Place a “kernel” at each data point
2. Normalize kernels so that total area = 1
3. Sum all kernels together





Kernel Density Estimation

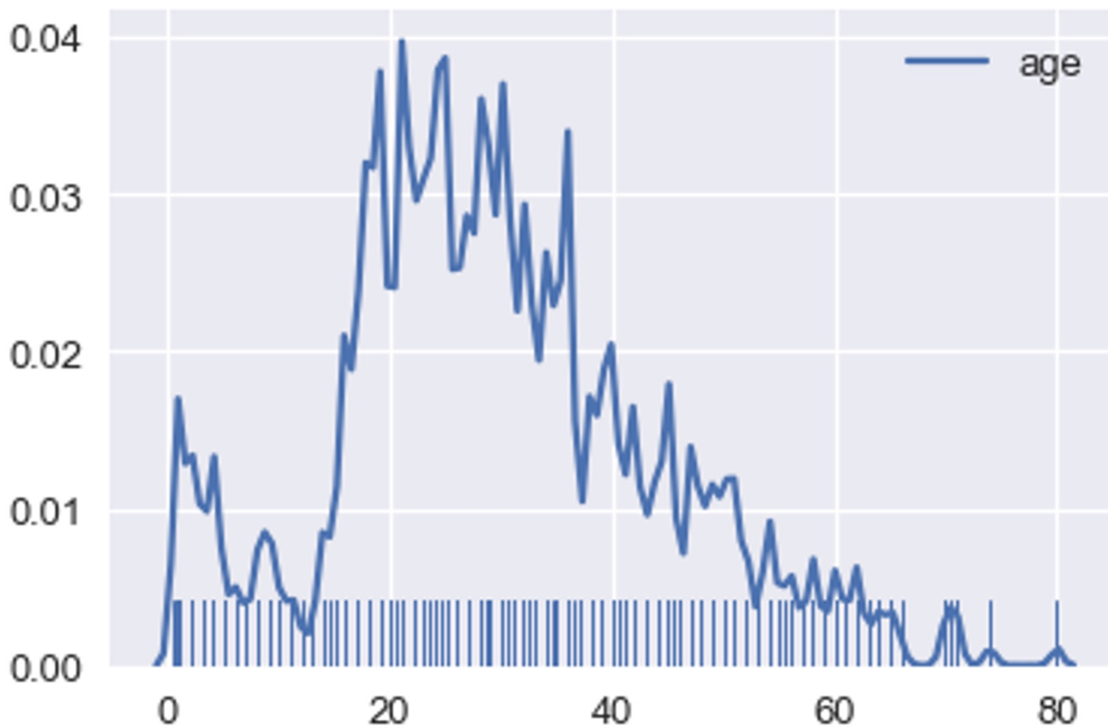
Gaussian kernel most common (default for seaborn).



Kernel Density Estimation

Changing width of each kernel = changing bandwidth

Narrow bandwidth is analogous to narrow bins for histogram

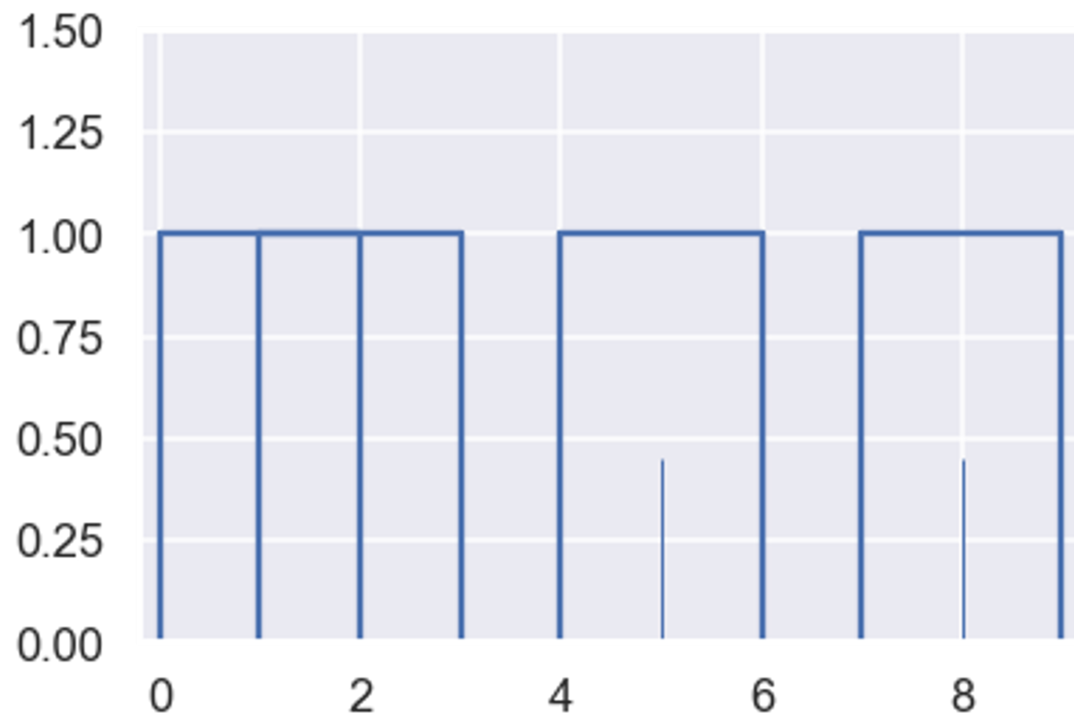


KDE Example – Uniform Kernel

Uniform kernel with bandwidth of 2.

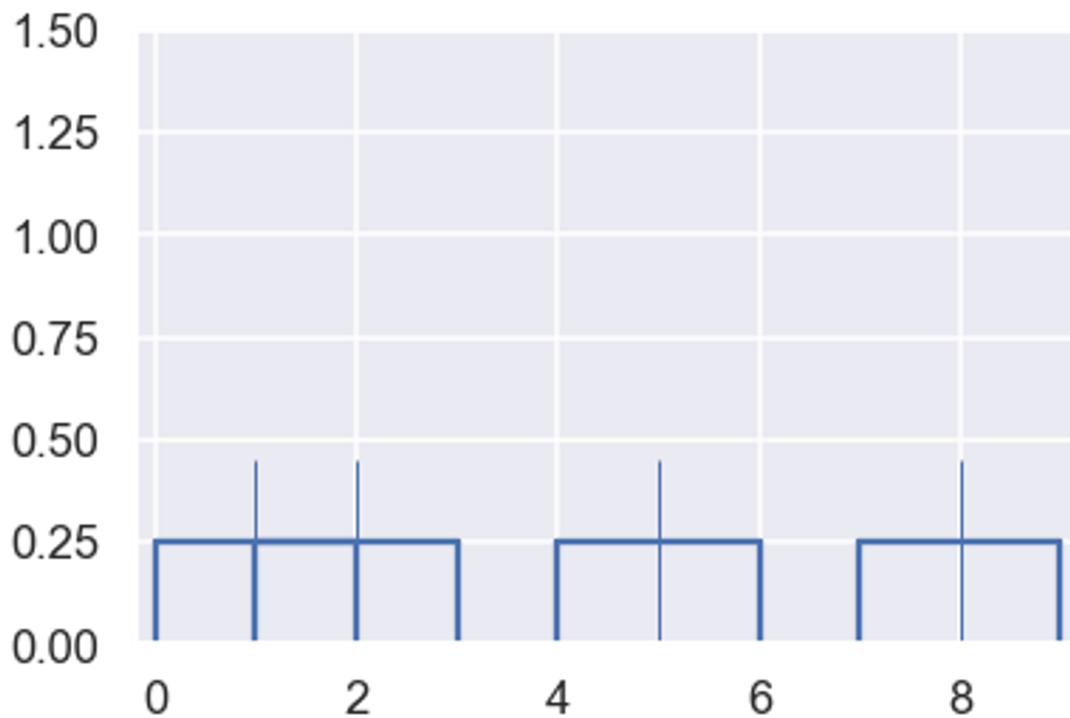
Data points at:

Kernel at each x: $x = [1, 2, 5, 8]$



KDE Example – Uniform Kernel

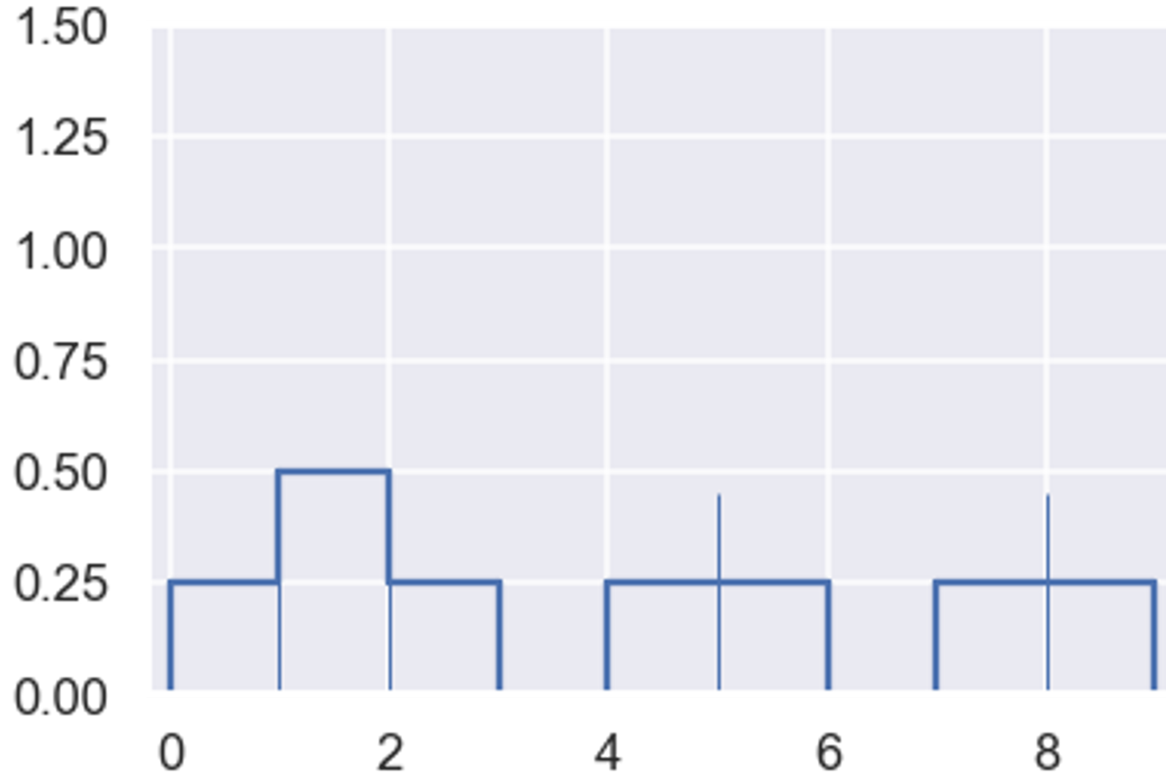
Scale each kernel by 1/4 since there are four points:



KDE Example – Uniform Kernel

Add kernels together:

Height at 1.5? 0.5



Summary

- When choosing a visualization, consider the principles of Scale, Conditioning, Perception, Transformation, Context, and Smoothing!
- In general: show the data!
  - Maximize data-ink ratio: cut out everything that isn't data-related

Sources

Books

- Tamara Munzner ”Visualization Analysis and Design”, 2014
- Lau, Gonzalez, Nolan “Principles and Techniques of Data Science”

Slides

- Torsten Möller’s Visualization course, Spring 2018
- UC Berkley Data 100 (Lau, Nolan, Dudoit, Perez)