

LECTURE 11

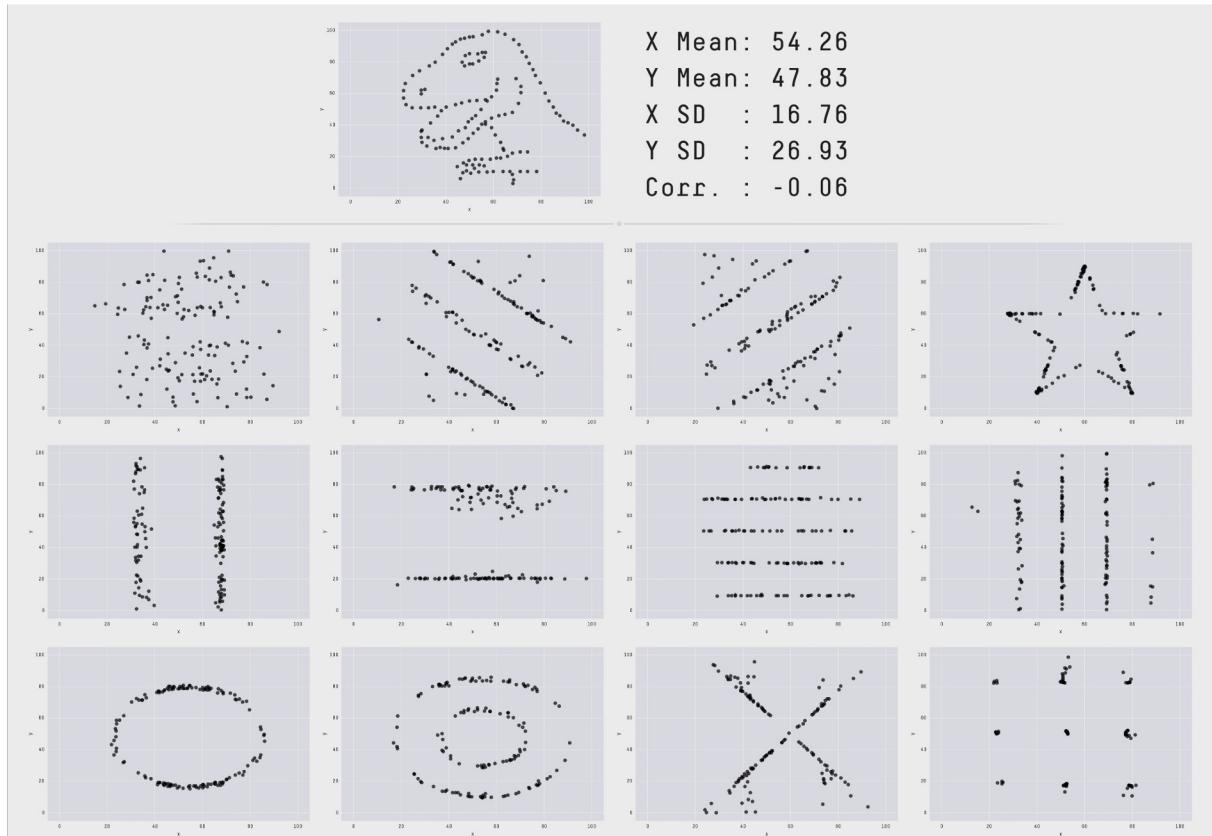
# Visualization, Part II

KDEs, Visualization Theory, and Transformations

# Visualizations Are More Expressive than Summary Statistics

Each of these 13 datasets has the same mean, standard deviation, and correlation coefficient.

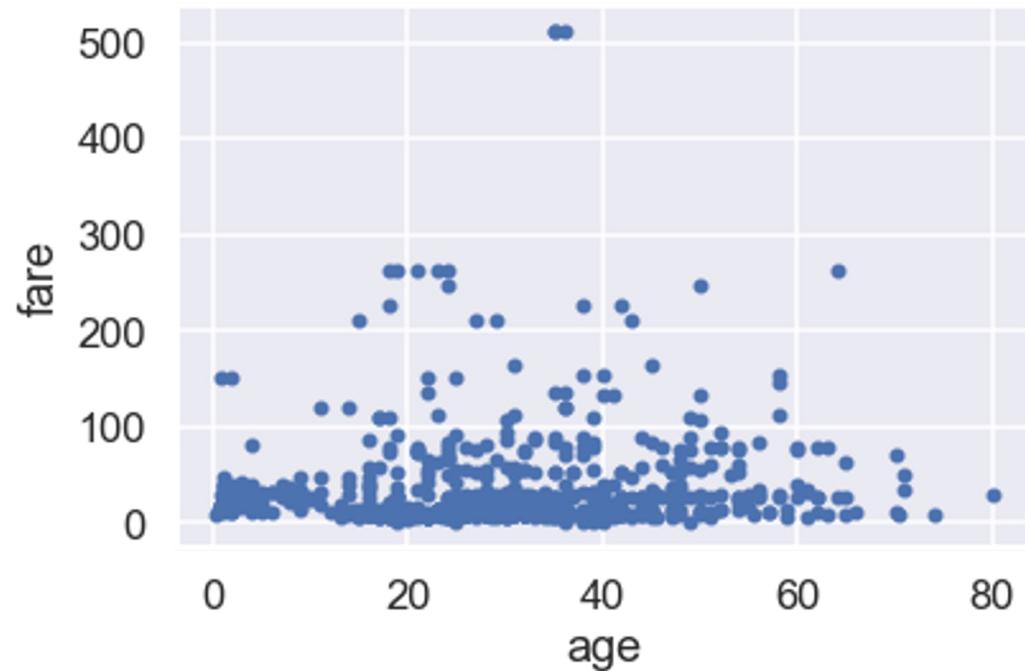
**Visualization complements statistics.**



<https://www.autodesk.com/research/publications/same-stats-different-graphs>



**"Looks like older people didn't spend more money on tickets for the Titanic than younger people."**



Note: A histogram or KDE would give stronger evidence than a scatter plot. 3

# Information Channels

---

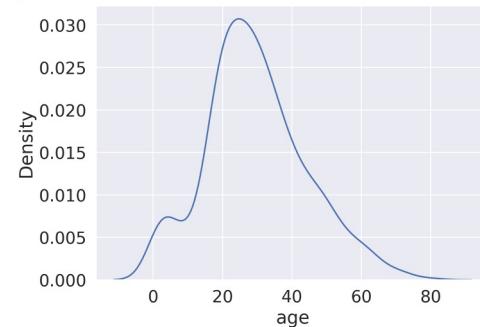
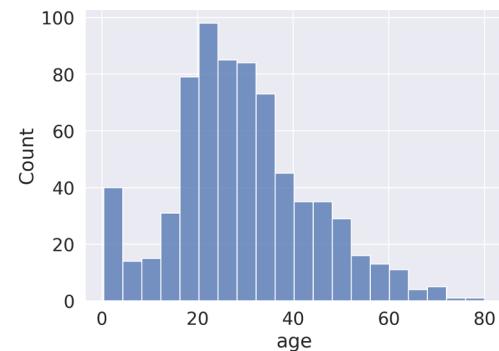
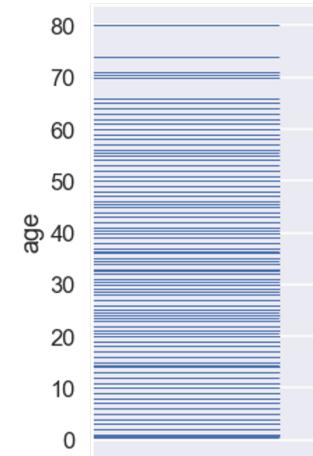
- Visualization Theory
  - **Information Channels**
  - Harnessing X/Y
  - Harnessing Color
  - Harnessing Markings
  - Harnessing Conditioning
  - Harnessing Context
- Transformations

# Take advantage of the human visual perception system

Data can be visualized in many ways!

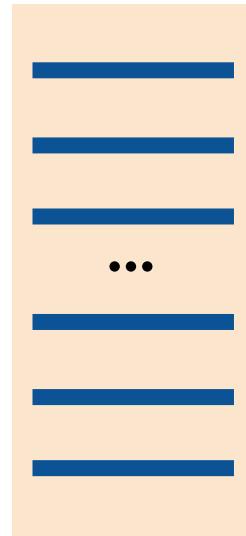
- Let's deconstruct the most basic plot types.

age	
0	22.0
1	38.0
2	26.0
...	
888	NaN
889	26.0
890	32.0



## Rug Plot: Encoding 1 Variable

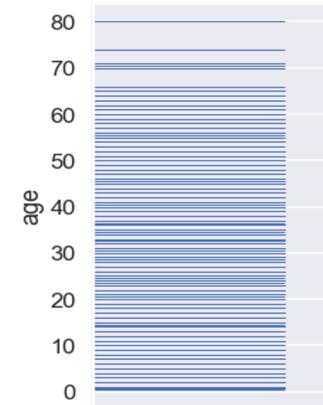
age	
0	22.0
1	38.0
2	26.0
...	...
888	NaN
889	26.0
890	32.0



Mark  
(Represents a datum)

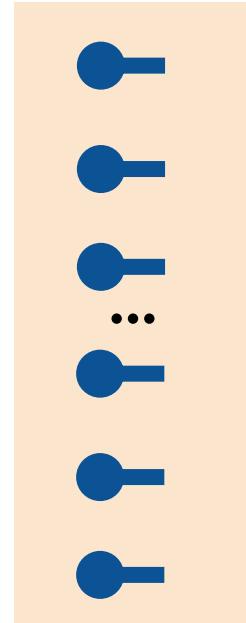
10px  
16px  
11px  
...  
**NONE**  
11px  
15px

Encoding  
(Maps datum to visual position)



## Rug Plot: Different Marks

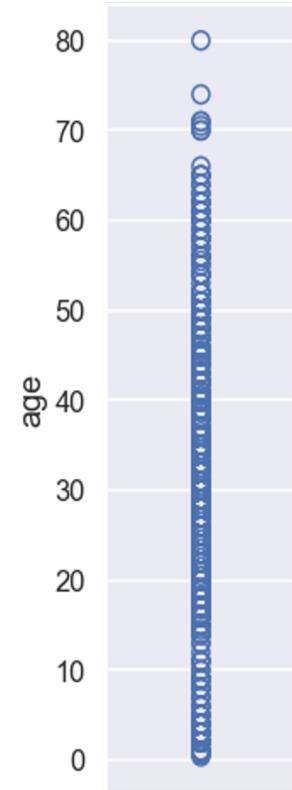
age	
0	22.0
1	38.0
2	26.0
...	...
888	NaN
889	26.0
890	32.0



**Mark**  
(Represents a datum)

**10px**  
**16px**  
**11px**  
...  
**NONE**  
**11px**  
**15px**

**Encoding**  
(Maps datum to visual position)



# Scatter Plot: Encoding 2 Variables

	age	fare
0	22.0	7.25
1	38.0	71.28
2	26.0	7.92
...	...	...
888	NaN	23.45
889	26.0	30.00
890	32.0	7.75



(10px, 7px)

(70px, 60px)

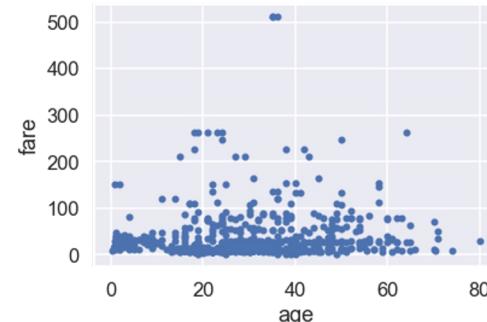
(45px, 9px)

...

(5px, 24px)

(45px, 37px)

(66px, 8px)



Mark  
(Represents a datum)

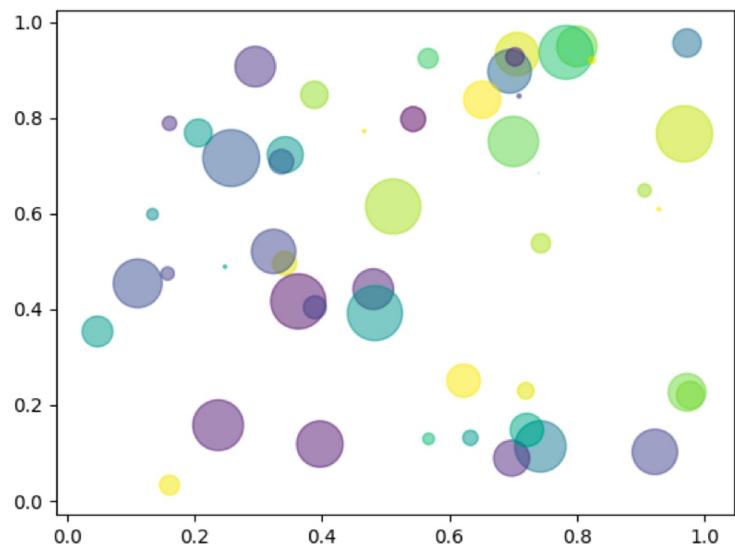
Encoding  
(Maps datum to visual position)

## Going Beyond 3D

---

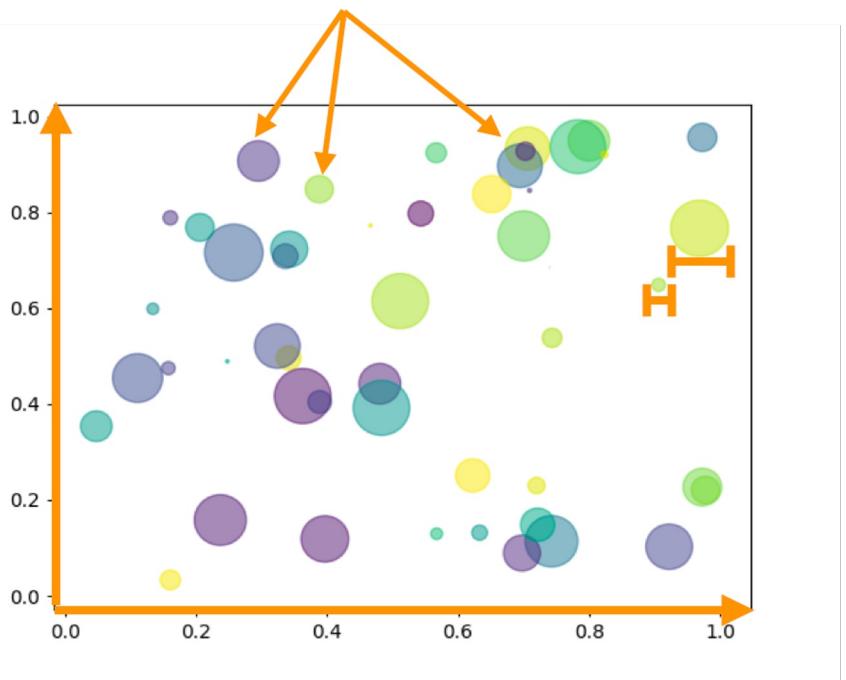
How many variables are we encoding here?

- In other words, how many “channels” of information are there?



How many variables are we encoding here?

- In other words, how many “channels” of information are there?



**Answer: 4.**

- x
- y
- area
- color

We could add even more: Shapes, outline colors of shapes, shading, etc. There are infinite possibilities.

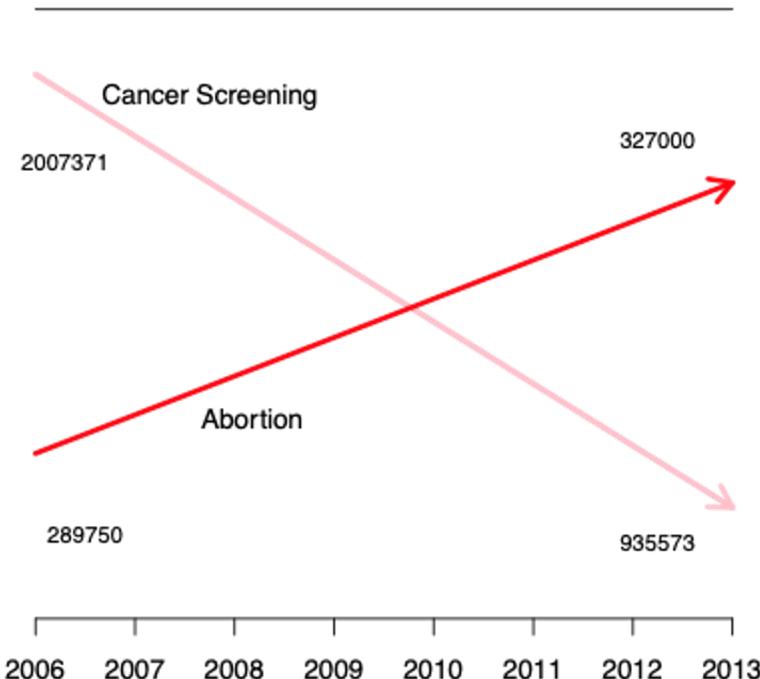
# Harnessing X/Y

---

- Visualization Theory
  - Information Channels
  - **Harnessing X/Y**
  - Harnessing Color
  - Harnessing Markings
  - Harnessing Conditioning
  - Harnessing Context
- Transformations

## Case Study: Planned Parenthood Hearing

---



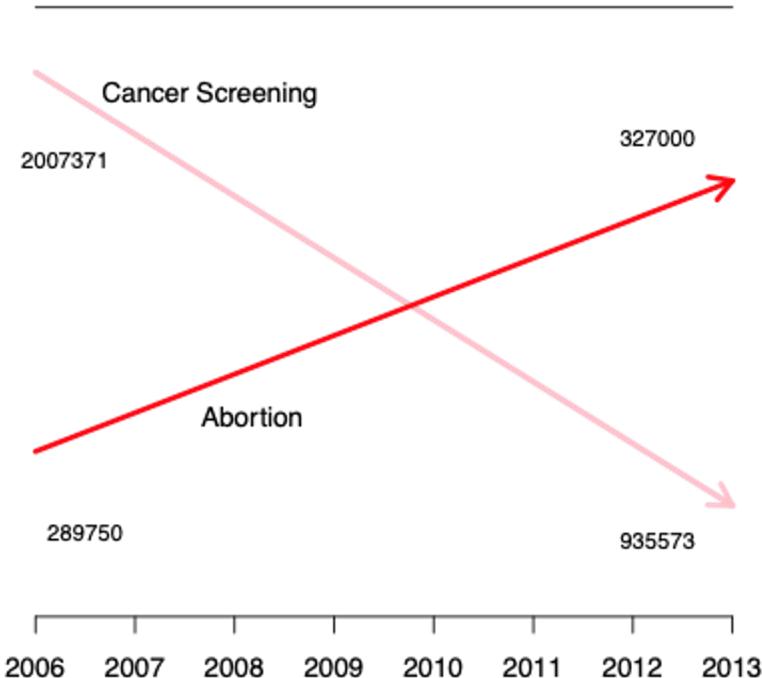
In 2015, Planned Parenthood was accused of selling aborted fetal tissue for profit.

Congressman Chaffetz (R-UT) showed this plot which originally appeared in a report by [Americans United for Life](#).

- What is this graph plotting?
- What message is this plot trying to convey?
- Is anything suspicious?

## Keep axis scales consistent

---



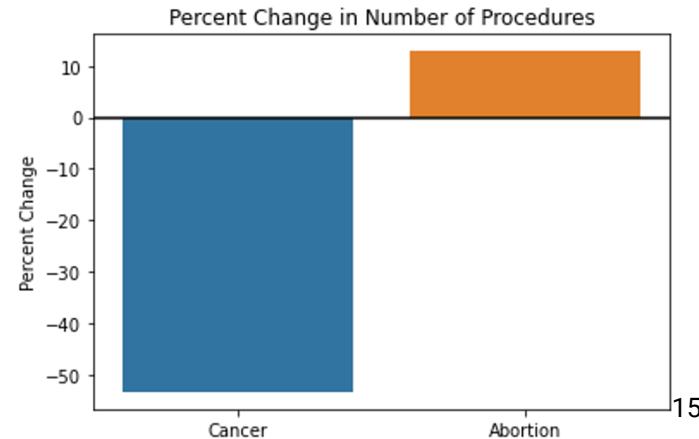
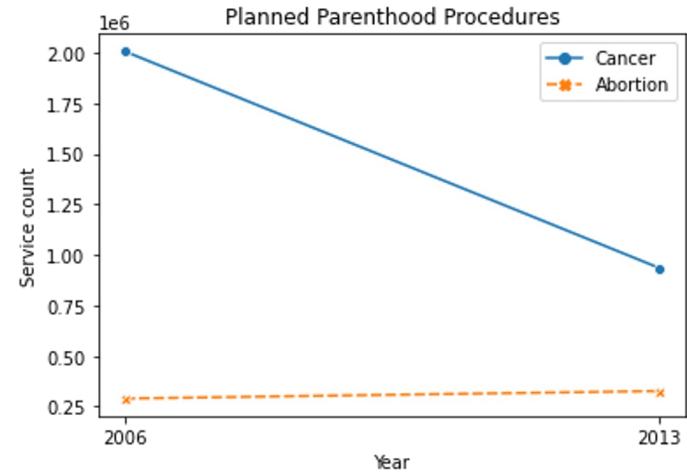
**The scales for the two lines are completely different!**

- 327000 is smaller than 935573, but appears to be way bigger.
- **Do not use two different scales for the same axis!**

## Consider scale of the data

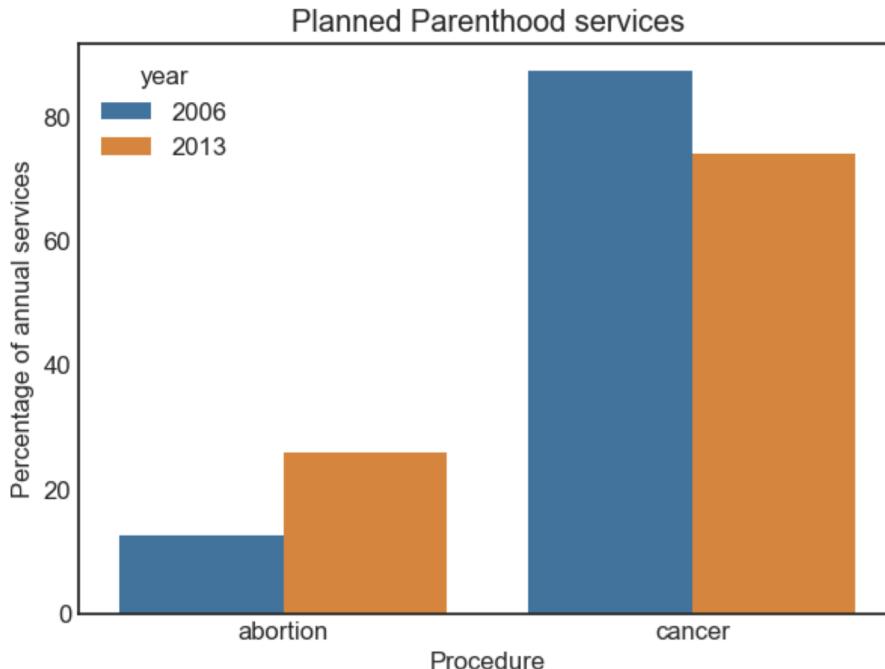
The top plot draws all of the data on the same scale.

- It clearly shows there was a dramatic drop in cancer screenings by Planned Parenthood.
- But there are still far more cancer screenings than abortions.
- Can plot percentage change instead of raw counts (bottom). This shows that cancer screenings have decreased and abortions have increased, without being misleading.



## Consider scale of the data

---



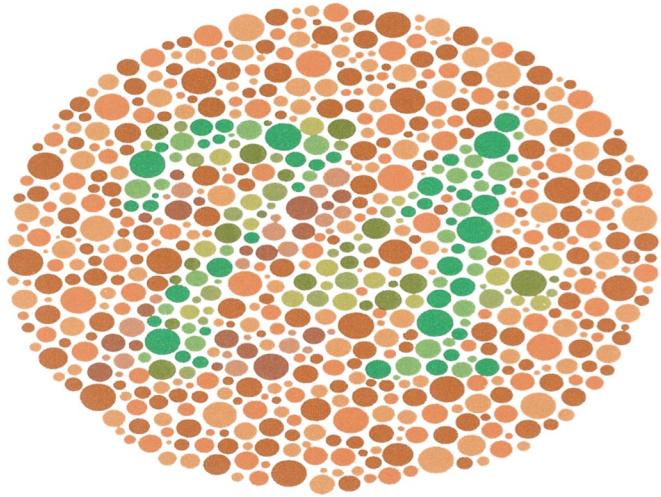
We could also visualize abortions and cancer screenings as a percentage of total procedures.

- Abortions increased from 13% to 26% of total procedures.

# Harnessing Color

---

- Visualization Theory
  - Information Channels
  - Harnessing X/Y
  - **Harnessing Color**
  - Harnessing Markings
  - Harnessing Conditioning
  - Harnessing Context
- Transformations

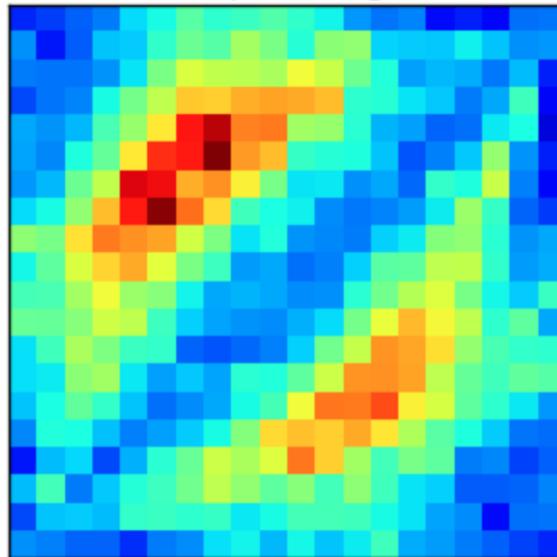


Choosing a set of colors which work together is a challenging task!

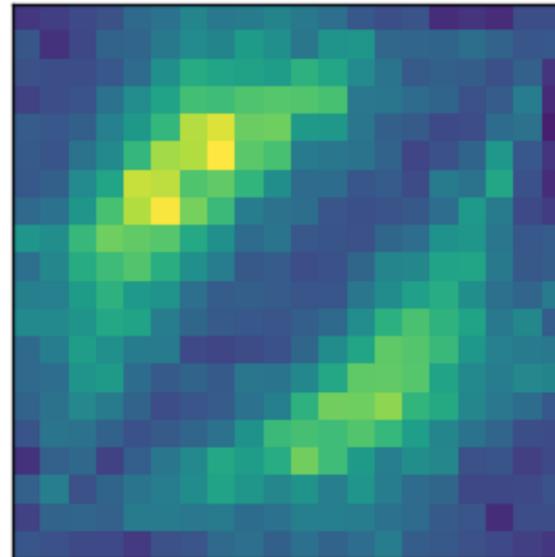
## Perception of Color

# Colormaps

---

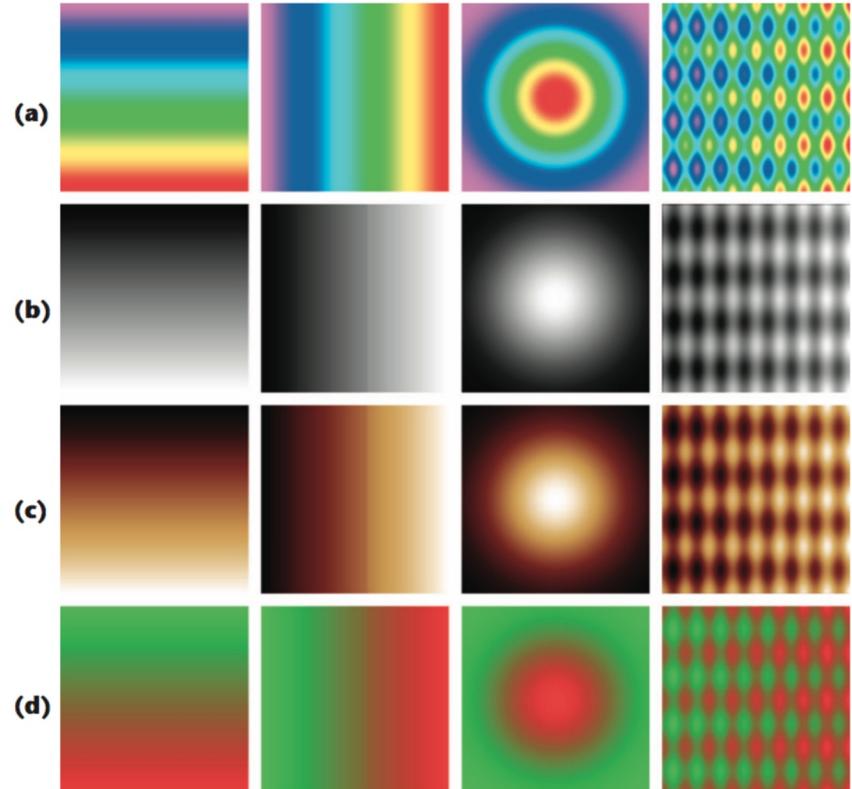
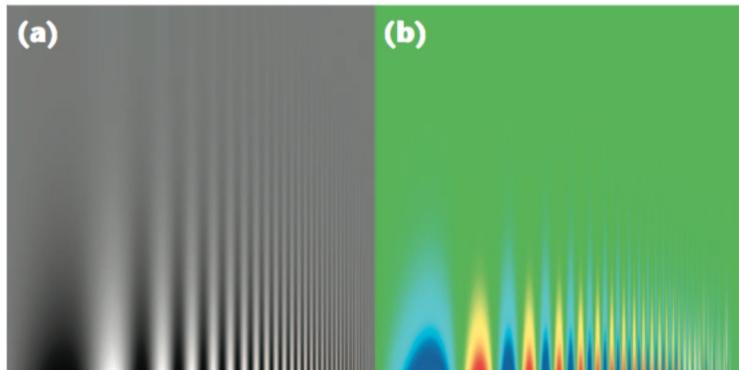
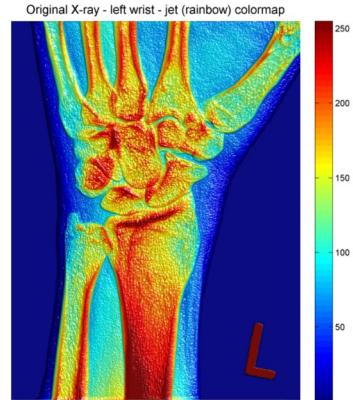


**Jet**



**Viridis**

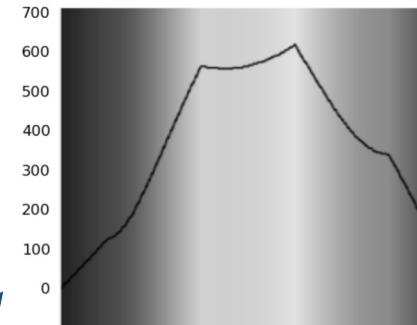
# The jet/rainbow colormap actively misleads



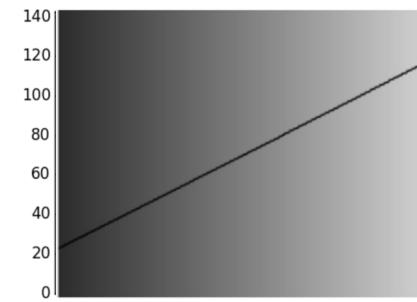
# Use a perceptually uniform colormap!

- **Perceptually uniform colormaps** have the property that if the data goes from 0.1 to 0.2, the **perceptual change** is the same as when the data goes from 0.8 to 0.9.
- Jet, the old matplotlib default, was far from uniform.
- Viridis, the new default colormap, is.
  - <https://bids.github.io/colormap/>
- Avoid combinations of red and green, due to red-green color blindness.

x-axis is color, y-axis is “lightness”



Bounces all over



Slope is constant

# Except when not :) The Google Turbo Colormap



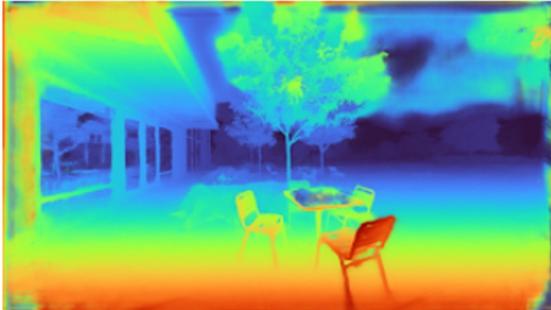
Turbo



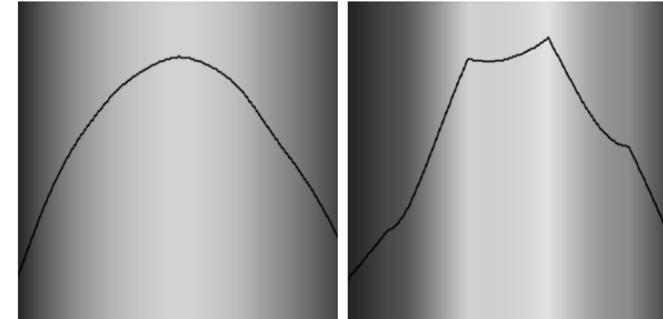
Jet



Inferno

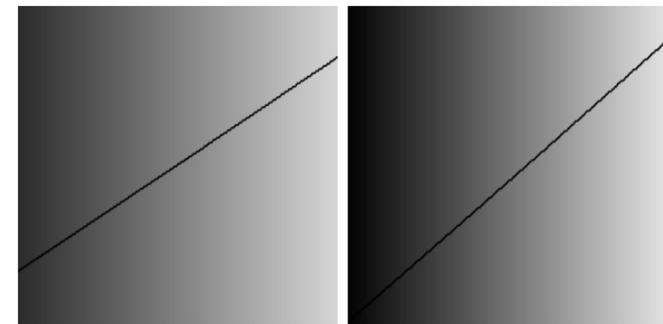


Turbo



Turbo

Jet



Viridis

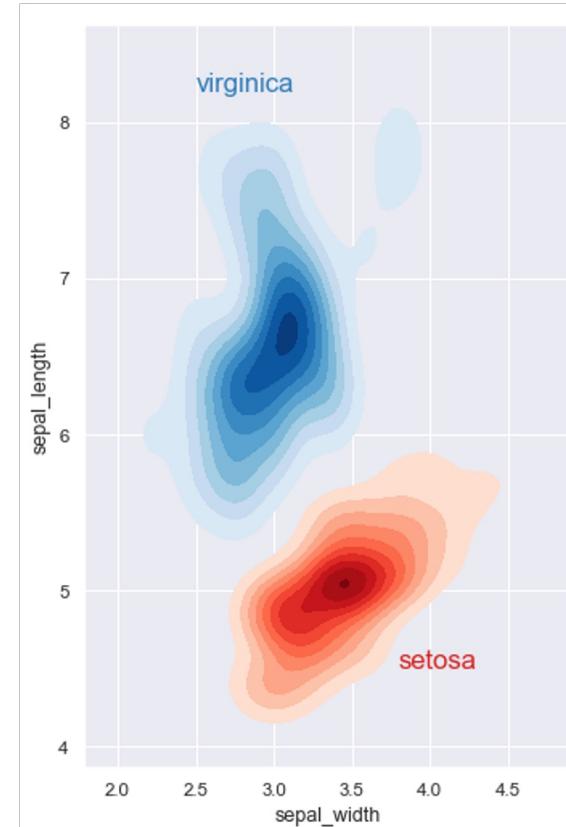
Inferno

X-axis is color, y-axis is "lightness"<sup>24</sup>



## Use color to highlight data type

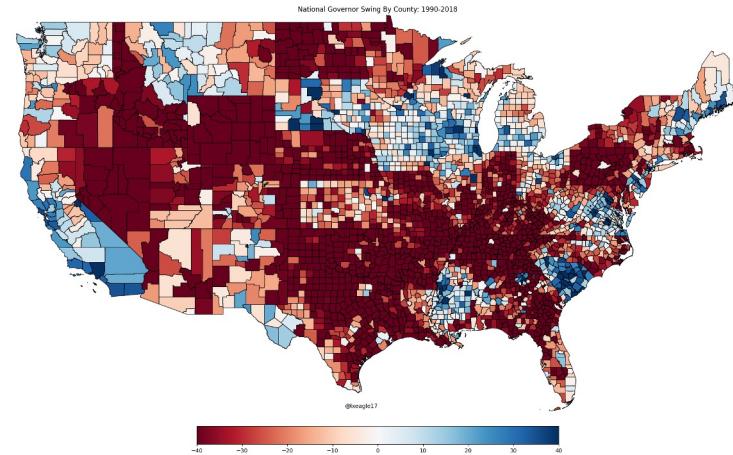
- **Qualitative:** Choose a qualitative scheme that makes it easy to distinguish between categories.
  - One category isn't "higher" or "lower" than another.
- **Quantitative:** Choose a color scheme that implies magnitude.
  - More on this in the next slide.
- The plot on the right has both!



# Sequential vs. diverging colormaps for quantitative data

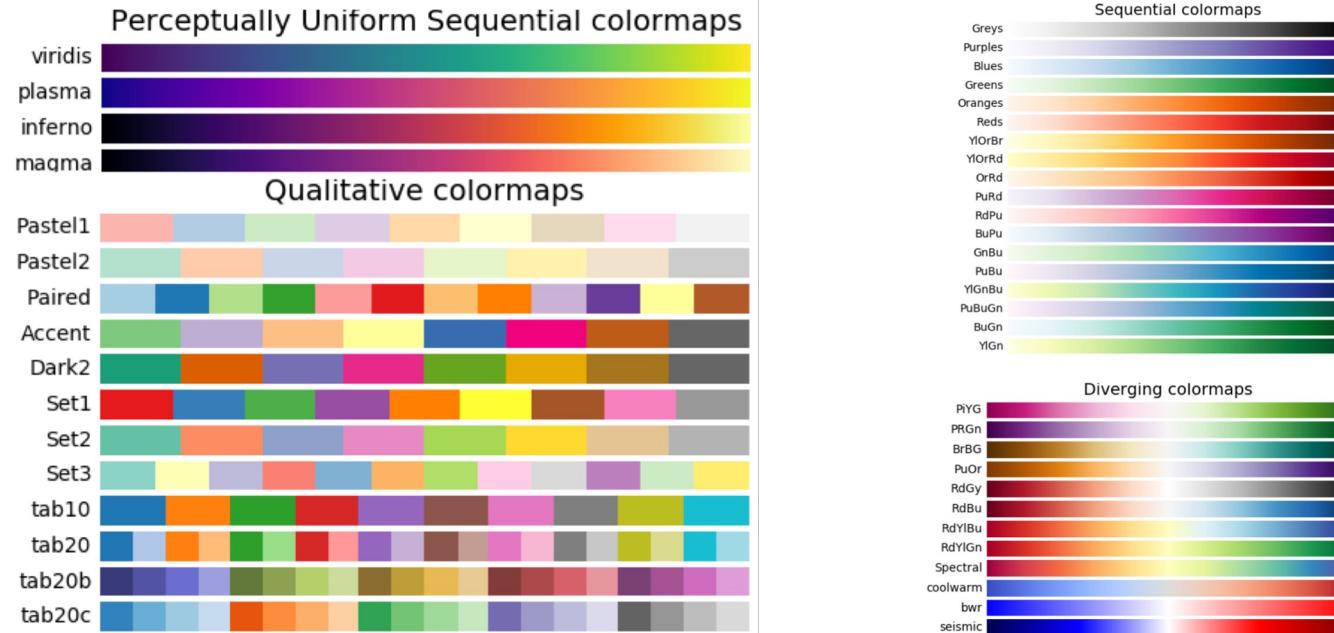


If the data progresses from low to high, use a **sequential** scheme where lighter colors are for more extreme values.



If low and high values deserve equal emphasis, use a **diverging** scheme where lighter colors represent middle values.

# Default matplotlib colormaps



Taken from [matplotlib documentation](#).

## Extra reading

---

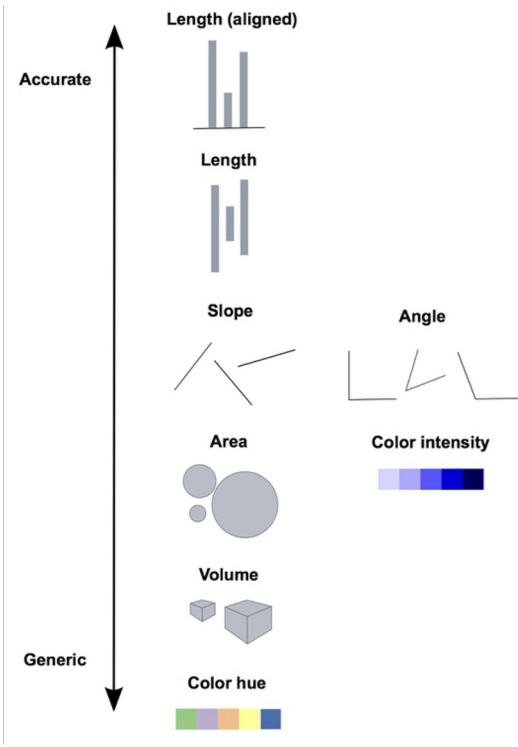
You may want to refer to these articles, which also discuss colormaps.

- Rainbow Colormap (Still) Considered Harmful - [paper](#) and [presentation slides](#).
- <https://eagereyes.org/basics/rainbow-color-map>
- <https://everydayanalytics.ca/2017/03/when-to-use-sequential-and-diverging-palettes.html>
- [https://web.natur.cuni.cz/~langhamr/lectures/vtfg1/mapinfo\\_2/bavy/colors.html](https://web.natur.cuni.cz/~langhamr/lectures/vtfg1/mapinfo_2/bavy/colors.html)

# Harnessing Markings

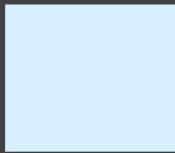
---

- Visualization Theory
  - Information Channels
  - Harnessing X/Y
  - Harnessing Color
  - **Harnessing Markings**
  - Harnessing Conditioning
  - Harnessing Context
- Transformations

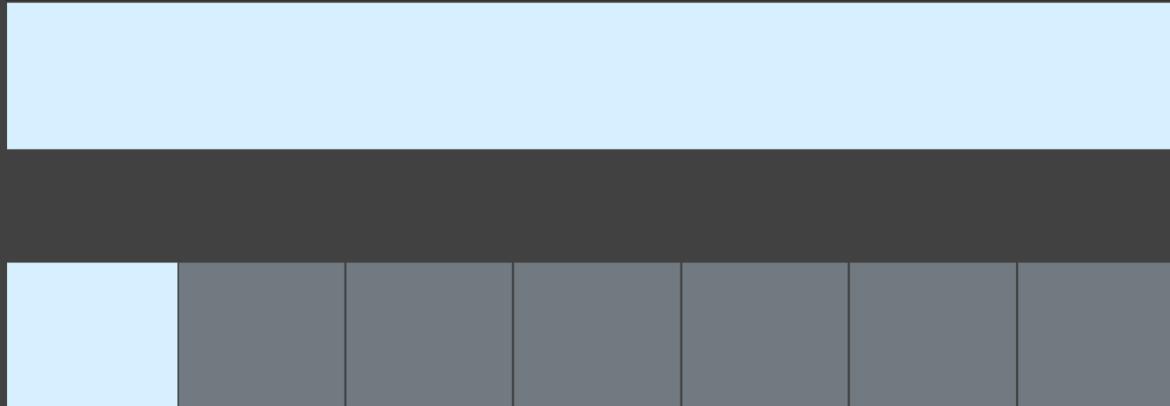


## Perception of Markings

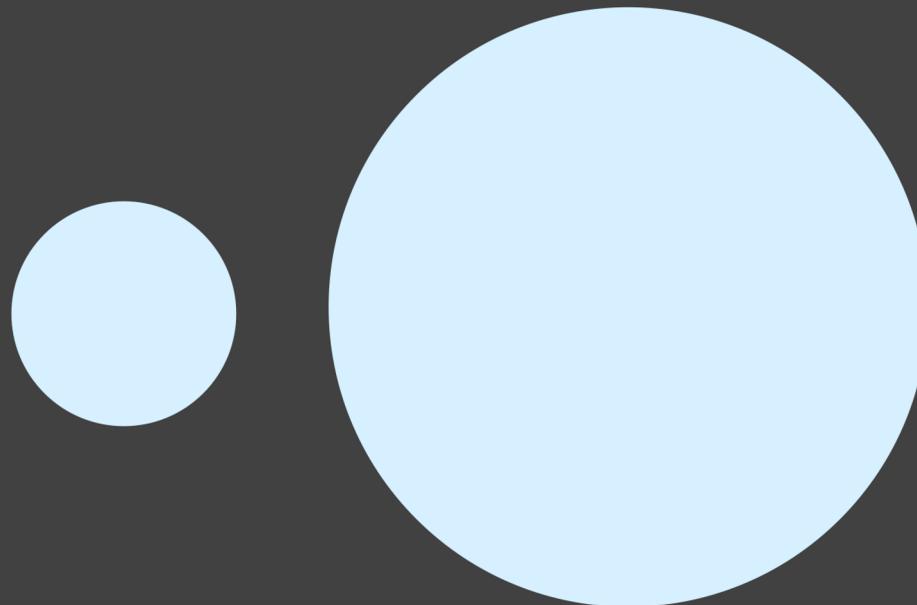
The accuracy of our judgements depend on the type of marking.



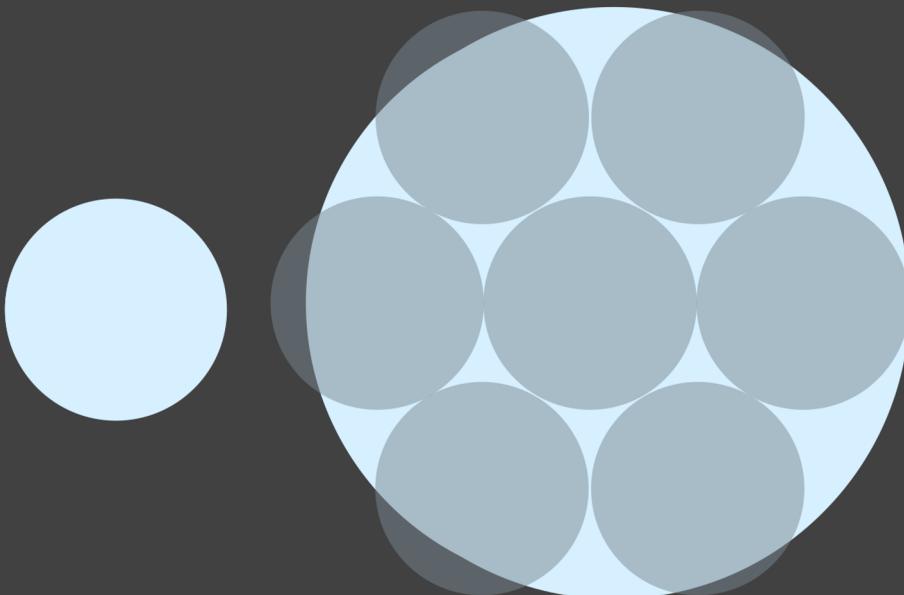
**How much longer is the top bar?**



The top bar is 7 times longer than the bottom bar.

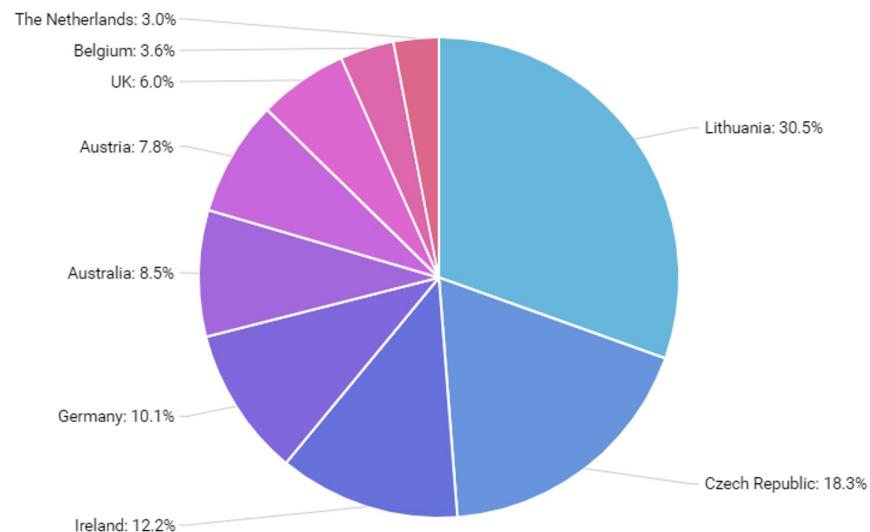
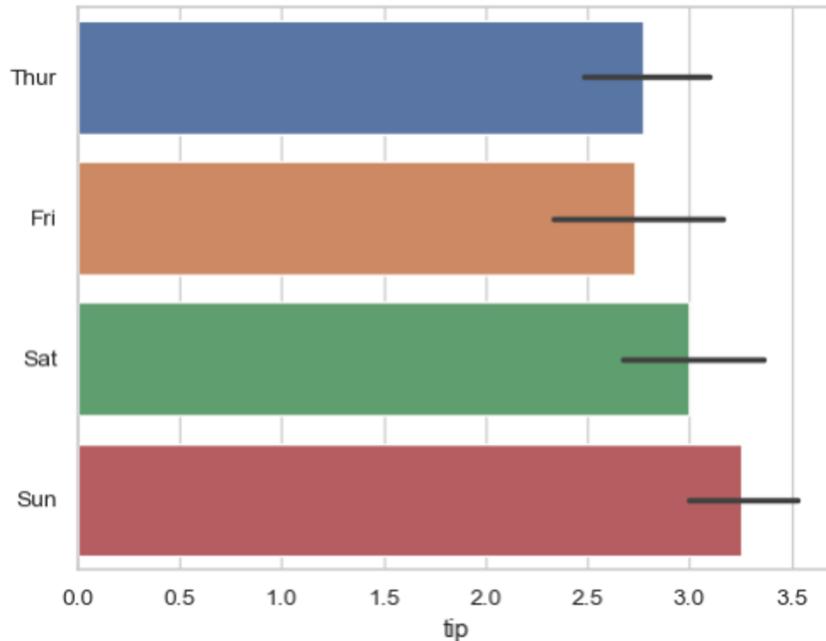


**How much bigger is the big circle?**



The area of the big circle is 7 times larger than the area of the small circle.

## Lengths are easy to distinguish; angles are hard



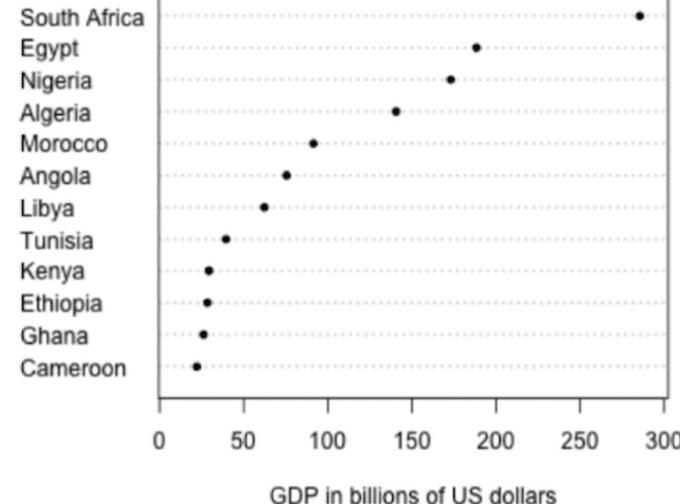
**Don't use pie charts!** Angle judgements are inaccurate.

# Areas are hard to distinguish

## African Countries by GDP

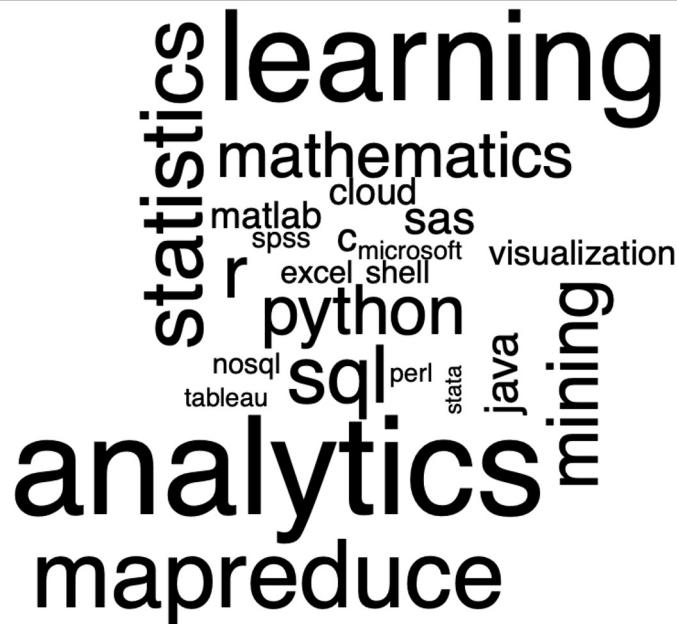


## African Countries by GDP

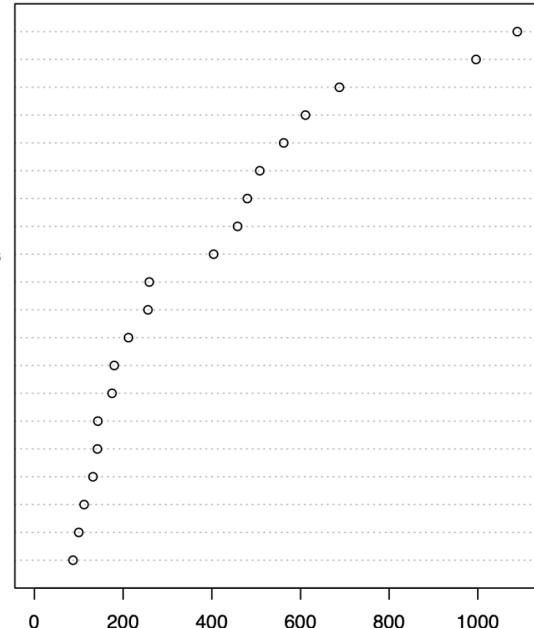


**Avoid area charts!** Area judgements are inaccurate. (For instance, South Africa has twice the GDP of Algeria, but that isn't clear from the areas.)

## Areas are hard to distinguish



analytics  
learning  
mapreduce  
statistics  
sql  
r  
mining  
python  
mathematics  
java  
sas  
c  
cloud  
matlab  
visualization  
shell  
excel  
nosql  
spss  
perl

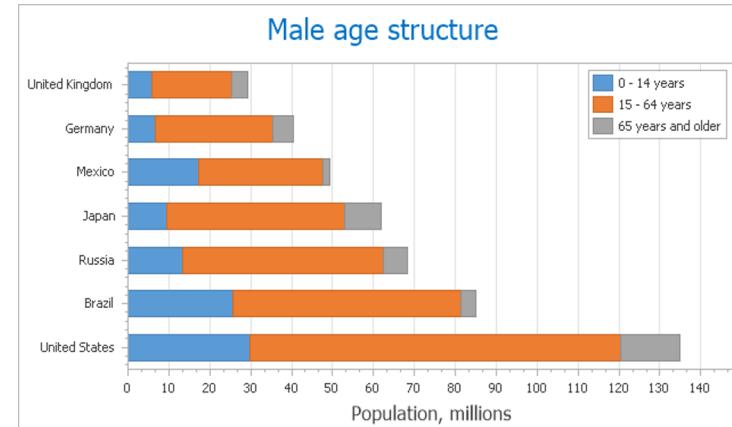
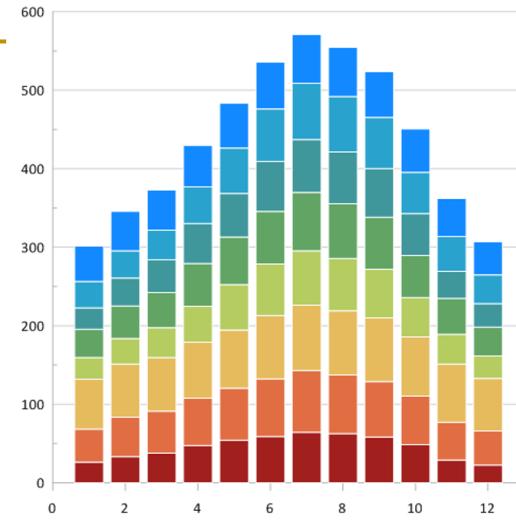


**Avoid word clouds too!** It's hard to tell the area taken up by a word.

## Avoid jiggling the baseline

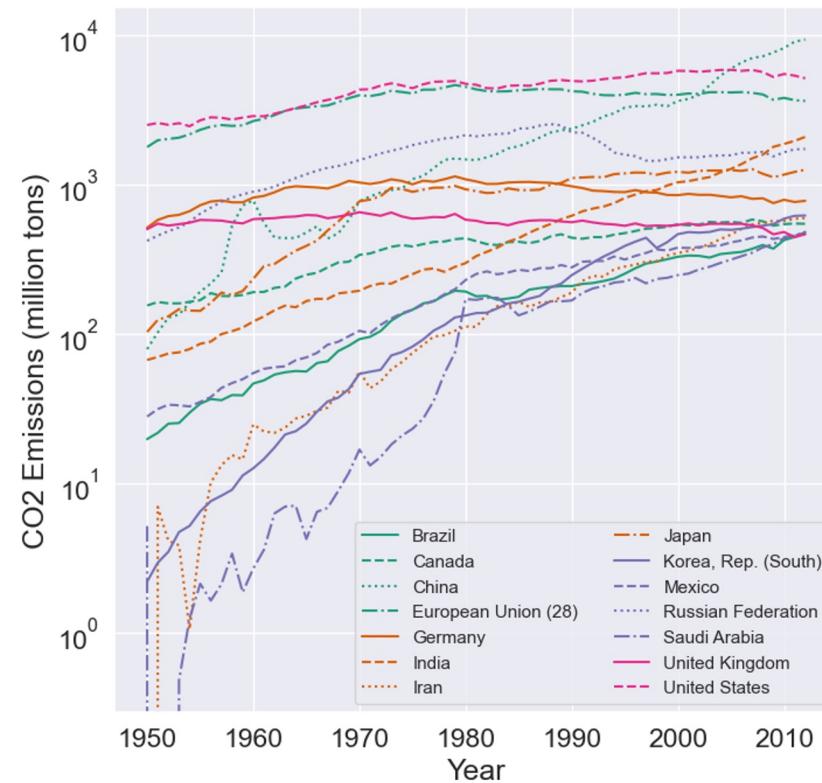
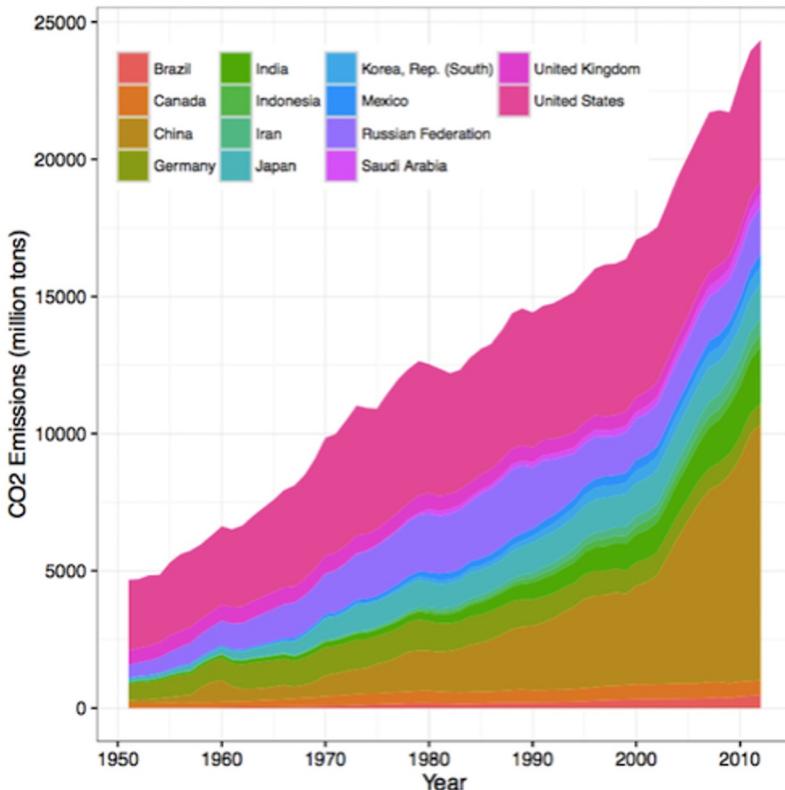
Stacked bar charts, histograms, and area charts are hard to read because the baseline moves.

- In the first plot, the top blue bars are all roughly of the same length. But that's not immediately obvious!
- In the second plot, comparing the number of 15-64 year old males in Germany and Mexico is difficult.



## Avoid jiggling the baseline

Here, by switching to a line plot, comparisons are made much easier.

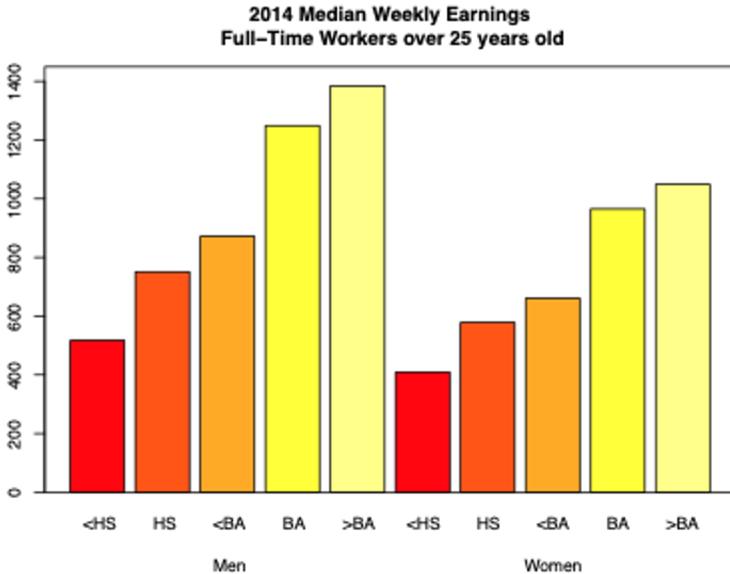


# Harnessing Conditioning

---

- Visualization Theory
  - Information Channels
  - Harnessing X/Y
  - Harnessing Color
  - Harnessing Markings
  - **Harnessing Conditioning**
  - Harnessing Context
- Transformations

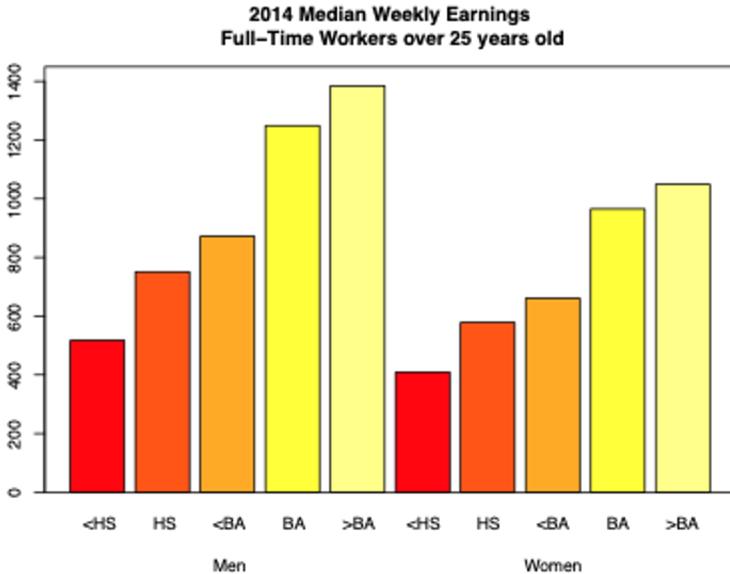
## Use Conditioning to Aid Comparison



This data comes from the [Bureau of Labor Statistics](#), who oversees surveys regarding the economic health of the US. They have plotted median weekly earnings for men and women by education level.

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

# Use Conditioning to Aid Comparison



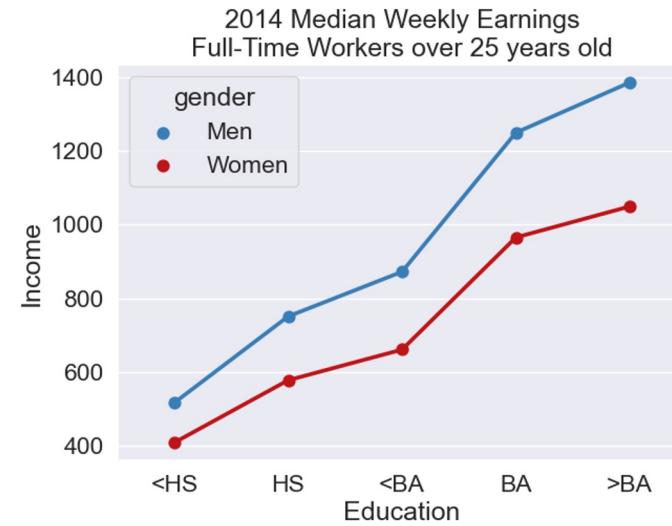
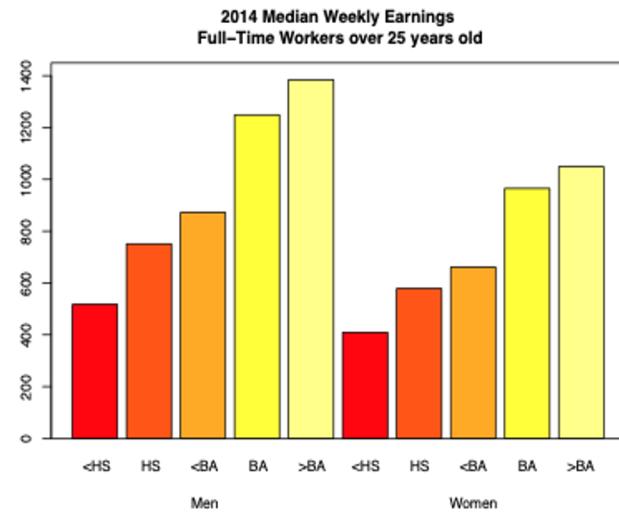
This data comes from the [Bureau of Labor Statistics](#), who oversees surveys regarding the economic health of the US. They have plotted median weekly earnings for men and women by education level.

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

- Easy to see the effect of education on earnings.
- Hard to compare between the two genders in the dataset.

How could we more easily make this difficult comparison?

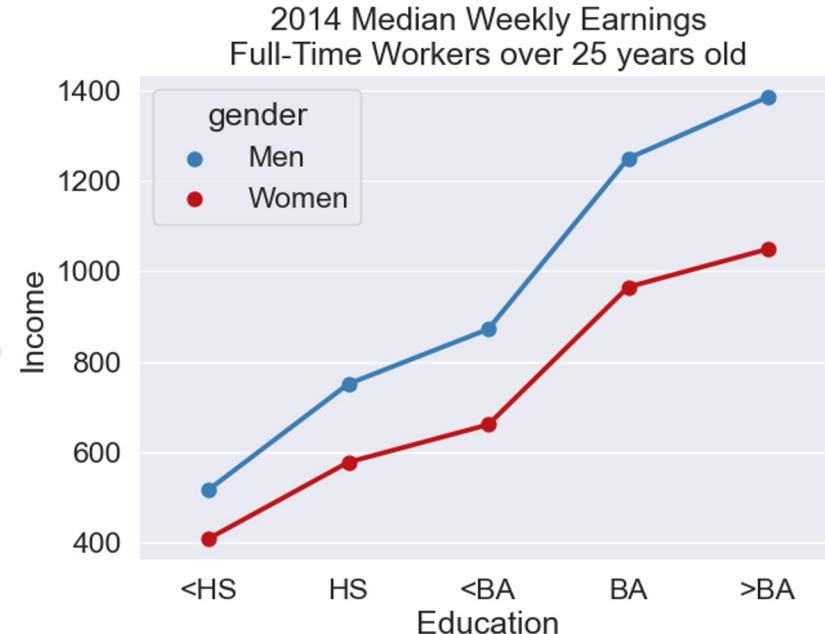
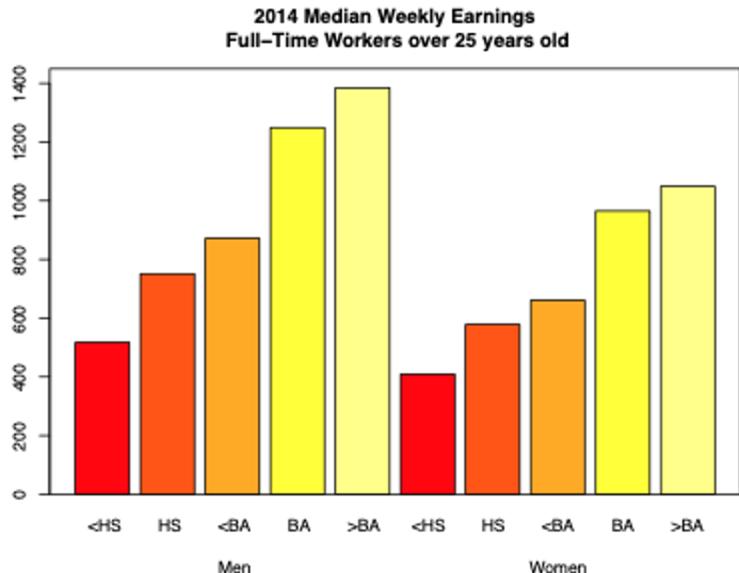
## Use conditioning to aid comparison



Left figure: Easy to see the effect of education on earnings. But hard to compare between the two genders in the dataset.

Right figure: Having two separate lines makes clear the wage difference between men and women.

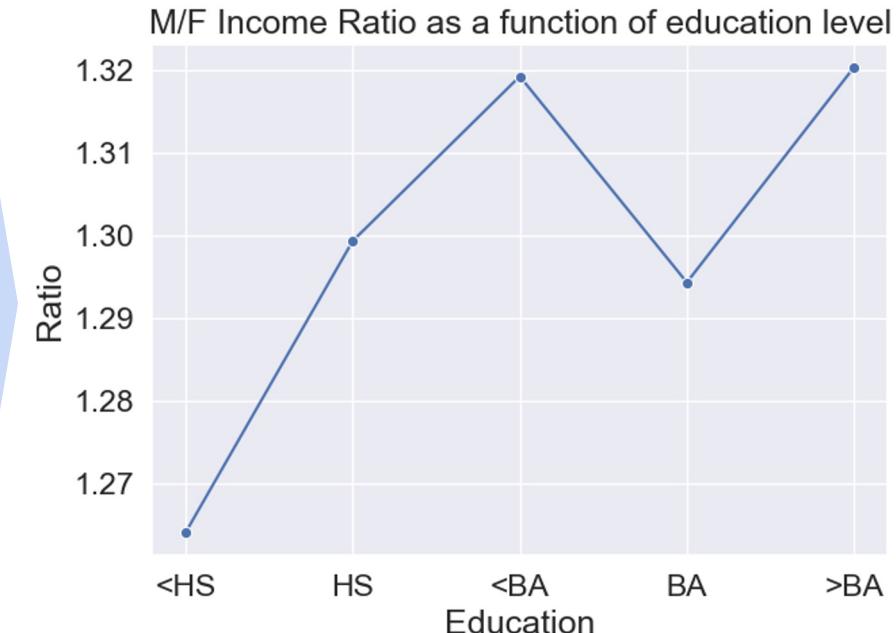
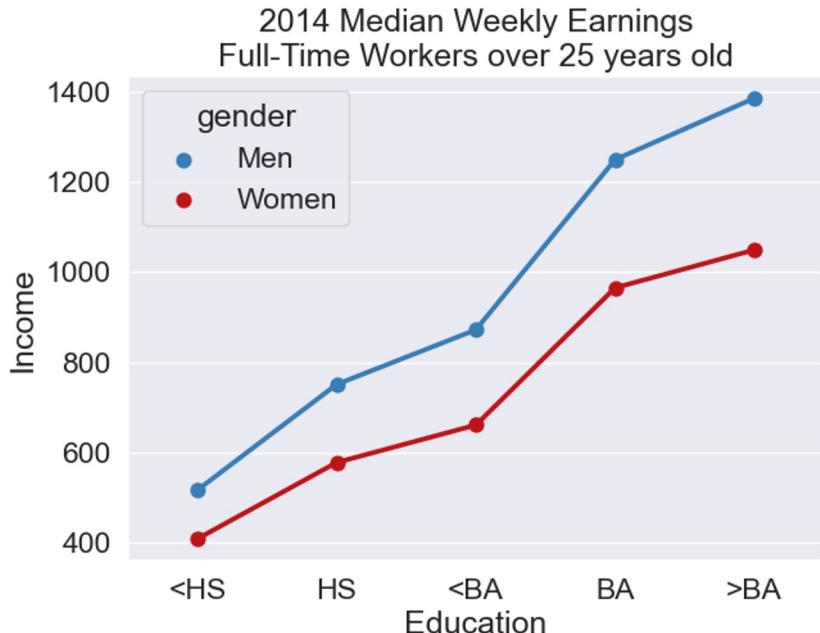
# Use Conditioning to Aid Comparison



- Easy to see the effect of education on earnings.
- Hard to compare between the two genders in the dataset.

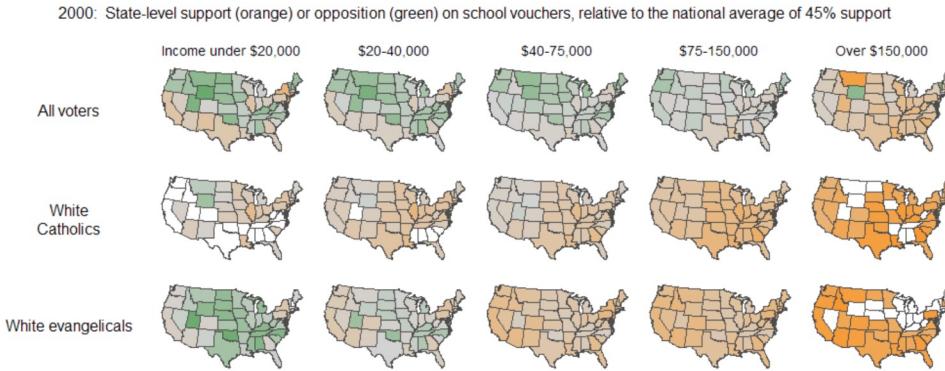
Having two separate lines makes clear the wage difference between men and women.

# How Does the Income Gap Increase with Education?



See notebook for how to get this figure with groupby!

# Superposition vs. Juxtaposition



An example of **small multiples**.

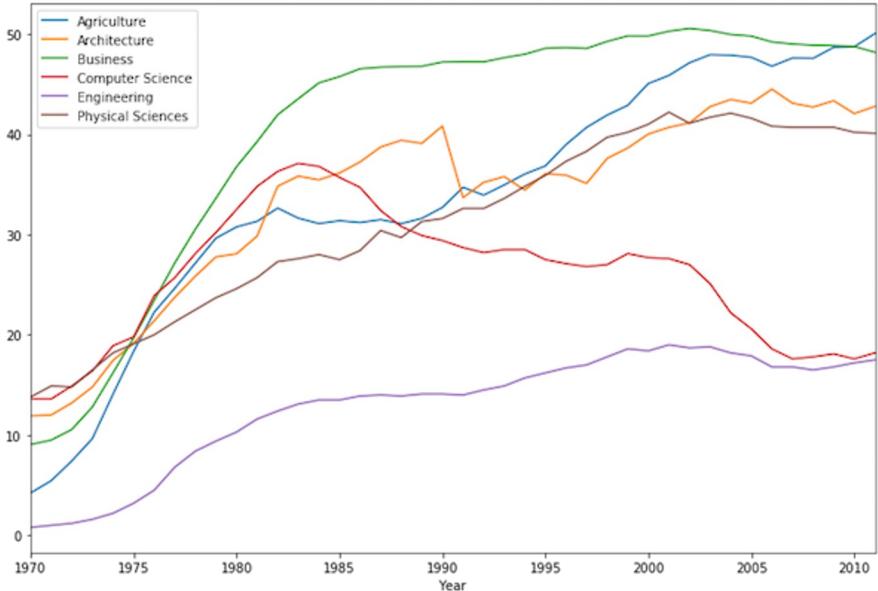
**Superposition:** placing multiple density curves, scatter plots on top of each other (what we've usually been doing)

**Juxtaposition:** placing multiple plots side by side, with the same scale (called "small multiples") (see left).

# Harnessing Context

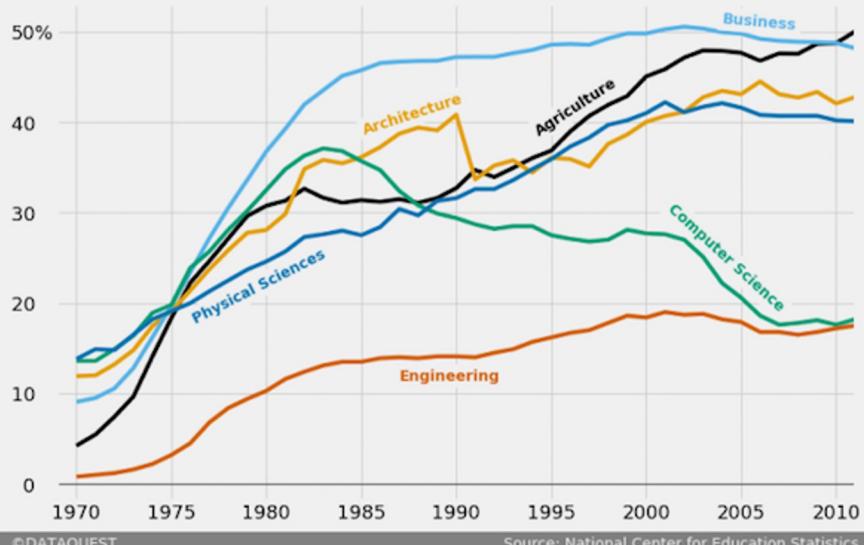
---

- Visualization Theory
  - Information Channels
  - Harnessing X/Y
  - Harnessing Color
  - Harnessing Markings
  - Harnessing Conditioning
  - **Harnessing Context**
- Transformations



## 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



©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, markers, and labels for important values.
- Legends, if appropriate.
- Captions that describe the data.

The plots you create in this class always need titles and axes labels, too.

## Captions

---

A picture is worth a thousand words, but not all thousand words you want to tell may be in the picture. In many cases, we need captions to help tell the story.

Captions should be:

- Comprehensive and self-contained.
- Describe what has been graphed.
- Draw attention to important features.
- Describe conclusions drawn from graph.

# Transformations

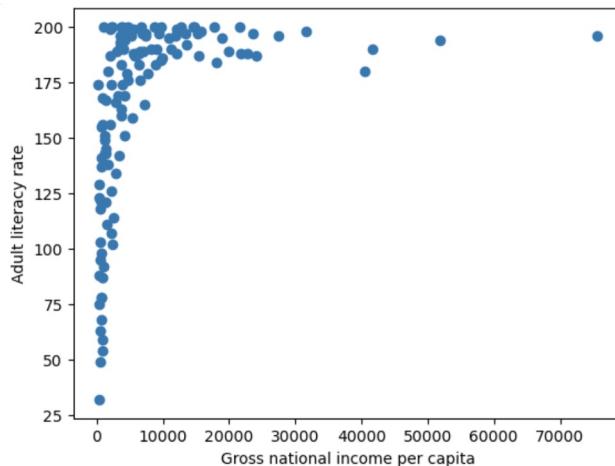
---

- Visualization Theory
  - Information Channels
  - Harnessing X/Y
  - Harnessing Color
  - Harnessing Markings
  - Harnessing Conditioning
  - Harnessing Context
- **Transformations**

Remember our goals of visualization:

1. To help your own understanding of your data/results.
2. To communicate results/conclusions to others.

These are influenced by our choice of visualization and our choices in *how to prepare data for visualization*.



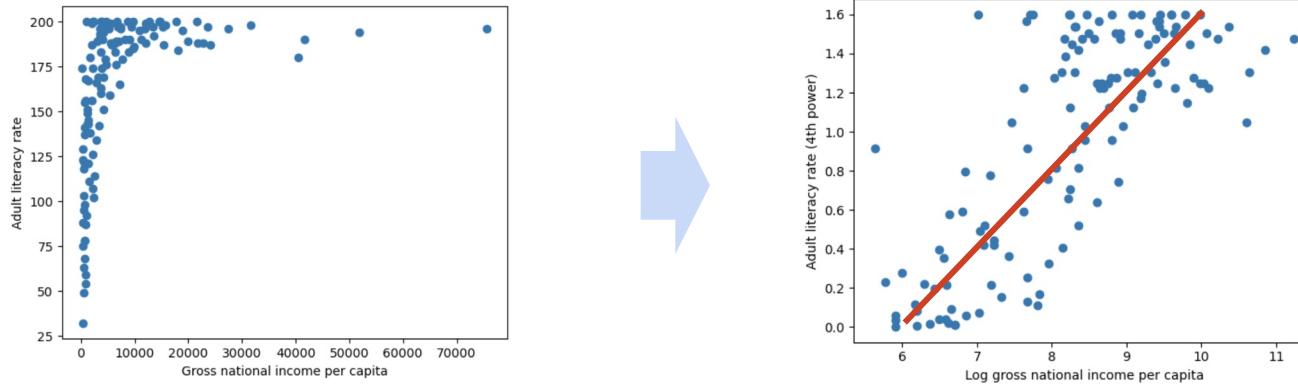
What problems are there here?

- Data is "smushed" – hard to interpret, even if we jittered.
- Difficult to generalize a clear relationship between the variables.

We often **transform** a dataset to help prepare it for being visualized.

# Linearization

When applying transformations, we often want to **linearize** the data – rescale the data so the x and y variables share a linear relationship.

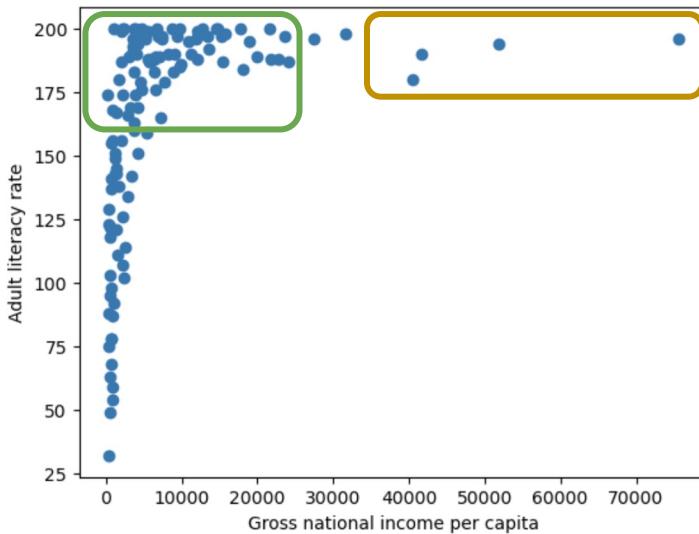


Why?

- Linear relationships are simple to interpret – we know how to work with slopes and intercepts to understand how two variables are related.
- Starting next week, we will start building linear models – these are more effective with linearized data.

# Applying Transformations

What makes this plot non-linear?



1. A few **large outlying x values** are distorting the horizontal axis.
2. Many **large y values** are all clumped together, compressing the vertical axis.

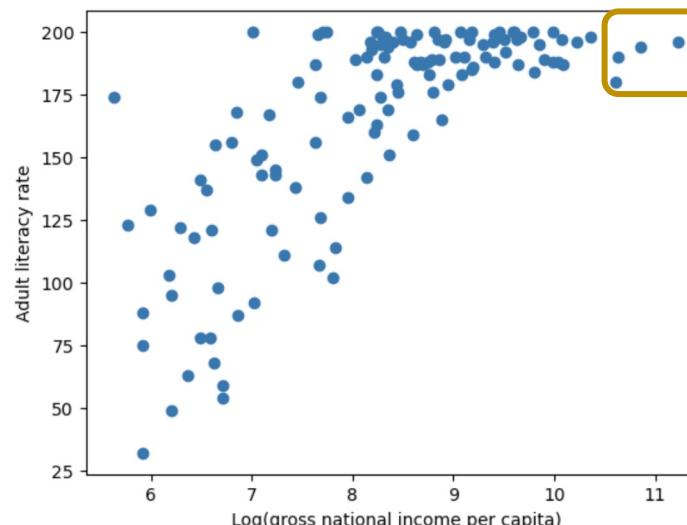
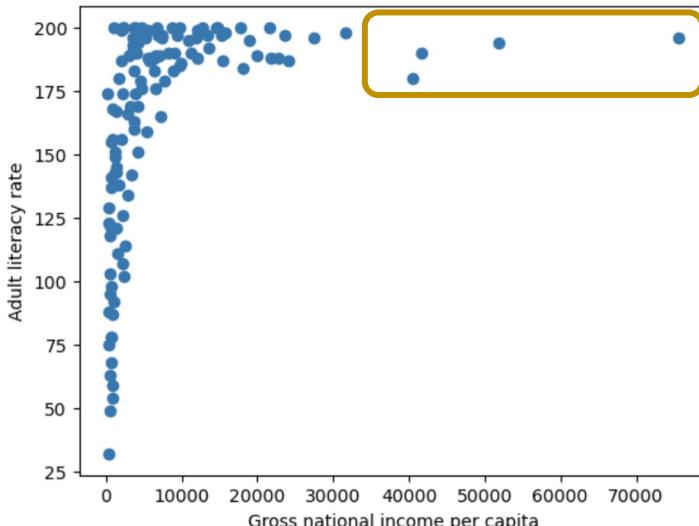
# Applying Transformations

What makes this plot non-linear?

1. A few **large outlying x values** are distorting the horizontal axis.

Resolve by log-transforming the x data:

- Taking the log of a large number decreases its value significantly.
- Taking the log of a small number does not change its value as significantly.



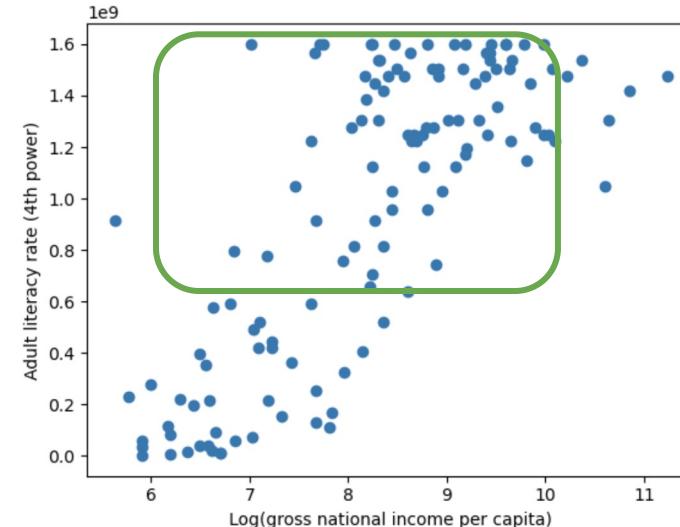
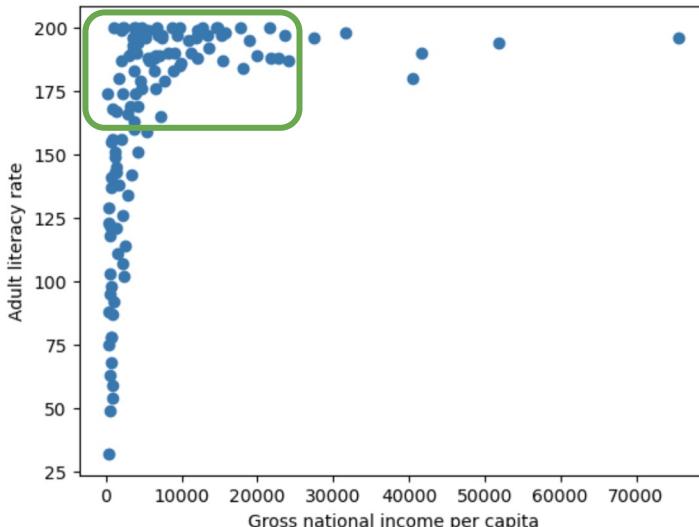
# Applying Transformations

What makes this plot non-linear?

2. Many **large y values** are all clumped together, compressing the vertical axis.

Resolve by power-transforming the x data:

- Raising a large number to a power increases its value significantly.
- Raising a small number to a power does not change its value as significantly.



## Interpreting Transformed Data

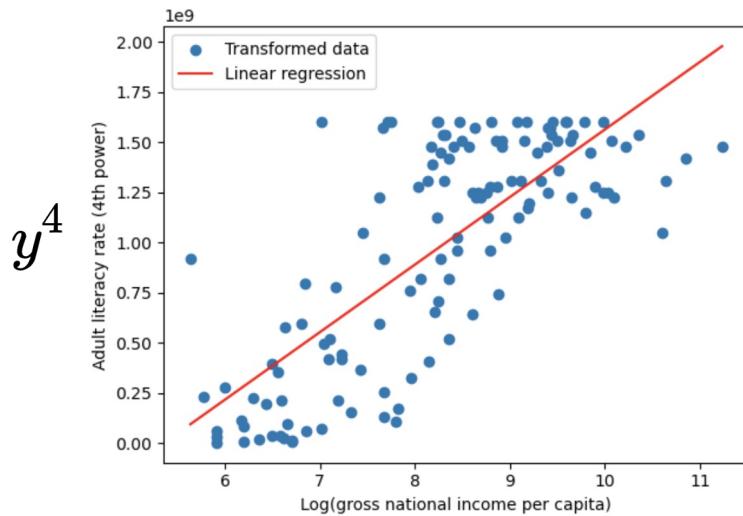
Now, we see a linear relationship between the transformed variables.

$$y^4 = m(\log x) + b$$

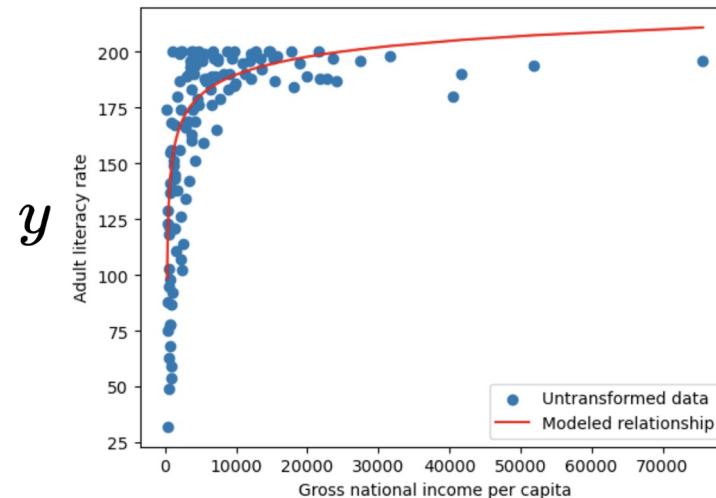


This tells us about the underlying relationship between the *original*  $x$  and  $y$ !

$$y = [m(\log x) + b]^{1/4}$$



$\log x$



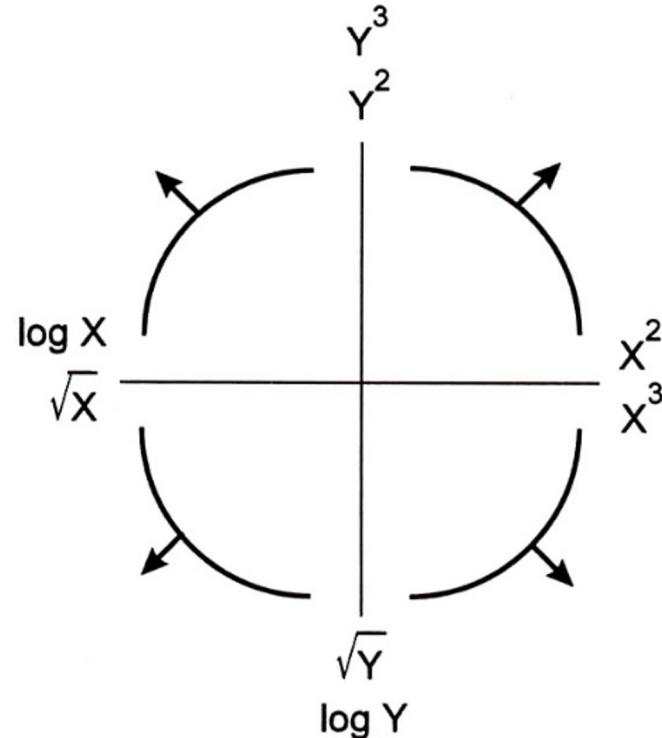
$x$

## Tukey-Mosteller Bulge Diagram

The **Tukey-Mosteller Bulge Diagram** is a guide to possible transforms to try to get linearity.

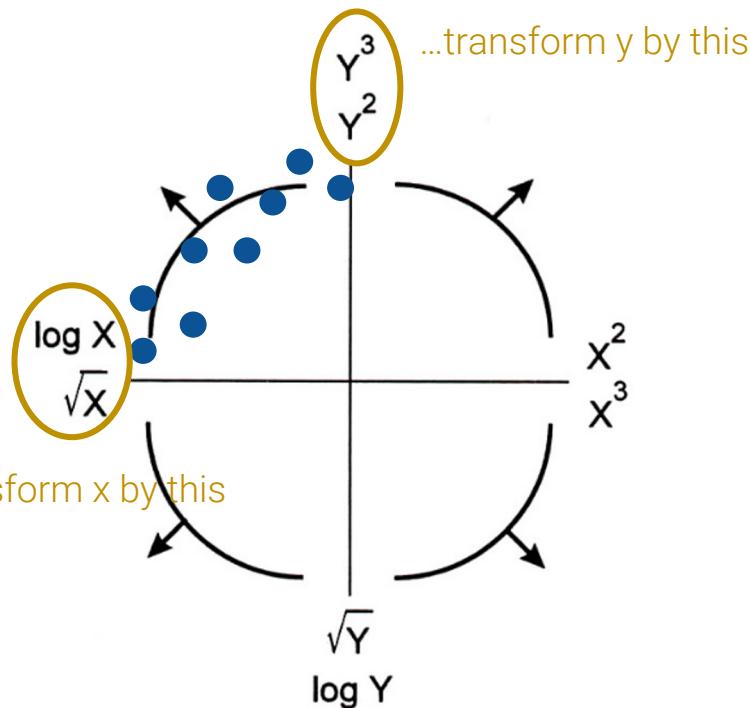
- A visual summary of the reasoning we just worked through.
- $\text{sqrt}$  and  $\log$  make a value "smaller".
- Raising to a value to a power makes it "bigger".
- There are multiple solutions. Some will fit better than others.

You should still understand the *logic* we just worked through to decide how to transform the data. The bulge diagram is just a summary.



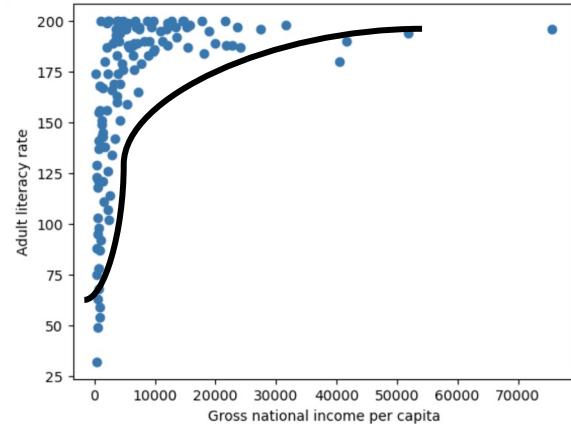
## Tukey-Mosteller Bulge Diagram

If the data bulges like this...



Applying to the data from before:

Could have transformed y by  $y^2, y^3$



Could have transformed x by  $\log(x)$ ,  $\sqrt{x}$

Some key ideas from today:

- Choose appropriate scales.
- Choose colors and markings that are easy to interpret correctly.
- Condition in order to make comparisons more natural.
- Add context and captions that help tell the story.
- Transforming our data can linearize relationships.
  - Helpful when we start linear modeling next lecture.
- **More generally – reveal the data!**
  - Eliminate anything unrelated to the data itself – “chart junk.”
  - It’s fine to plot the same thing multiple ways, if it helps fit the narrative better.