LECTURE 9

Visualization III

KDEs, Transformations, and Visualization Theory

Data Science, Spring 2024 @ Knowledge Stream

Sana Jabbar

Plotting Distributions - Revisited

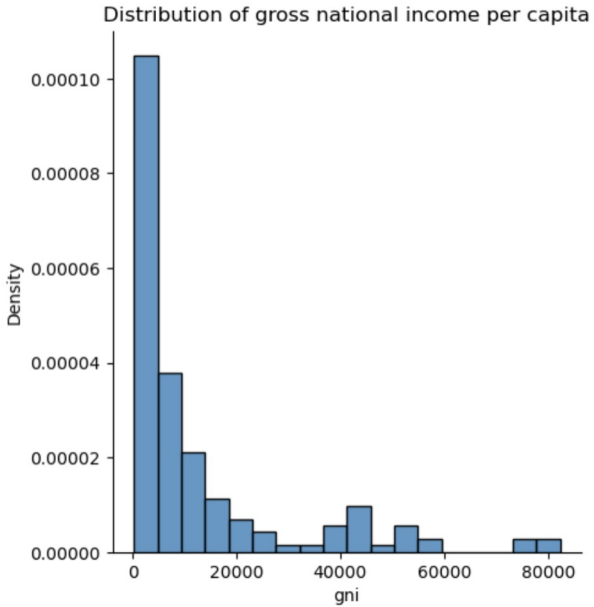
Lecture 09

- **Kernel Density Functions**
  - KDE Mechanics
  - Kernel Functions and Bandwidth
  - **Plotting Distributions**
- Relationships between Quantitative Variables
  - Transformations

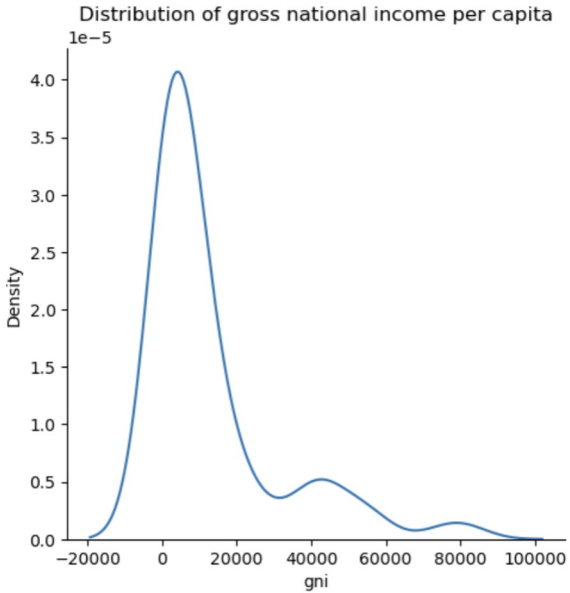
displot

displot is a wrapper for **histplot**, **kdeplot**, and **ecdfplot** to plot distributions.

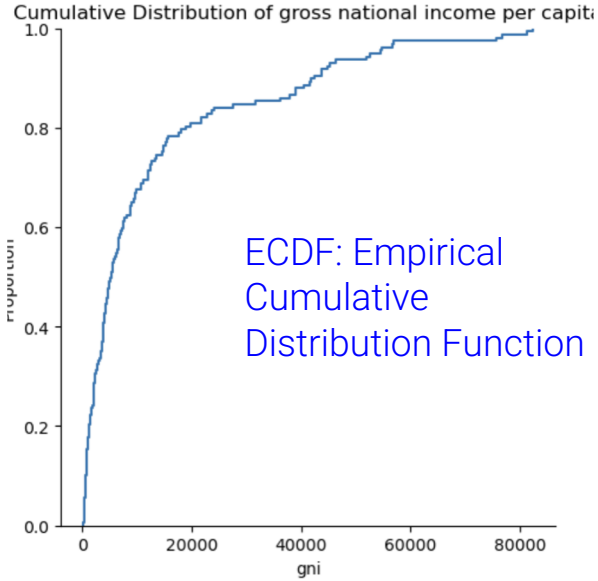
```
sns.displot(data=wb,
            x="gni",
            kind="hist",
            stat="density")
```




```
sns.displot(data=wb,
            x="gni",
            kind="kde")
```




```
sns.displot(data=wb,
            x="gni",
            kind="ecdf")
```



ECDF: Empirical Cumulative Distribution Function

Relationships between Quantitative Variables

Lecture 09

- Kernel Density Functions
  - KDE Mechanics
  - Kernel Functions and Bandwidth
  - Plotting Distributions - Revisited
- **Relationships between Quantitative Variables**
  - Transformations

From Distributions to Relationships

Up until now, we focused exclusively on visualizing variable distributions.

Now we will visualize **relationships** between variables. In other words, how do sets of two (or more) variables vary in relation to one another?

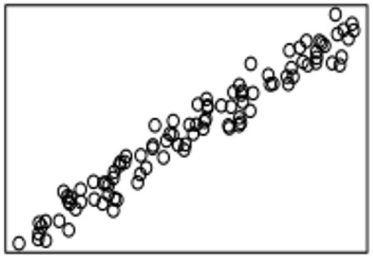
Scatter Plots

Scatter plots are used to reveal relationships between **pairs** of numerical variables.

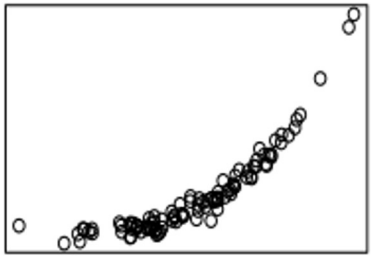
- Visual assessment may help us decide how to model these relationships.

- Example: Linear model
  - Linear Regression
  - Good for the left two, not so much for the right two.
- Reminder: "Correlation does not imply causation." A linear relationship is a mathematical one.

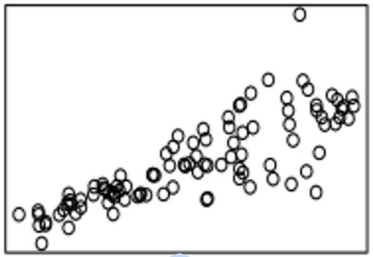
simple linear



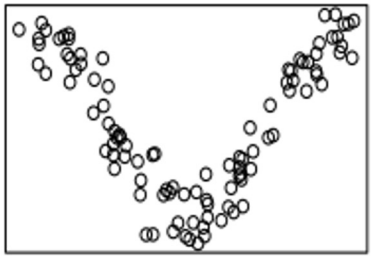
simple nonlinear



linear, spreading



v-shaped

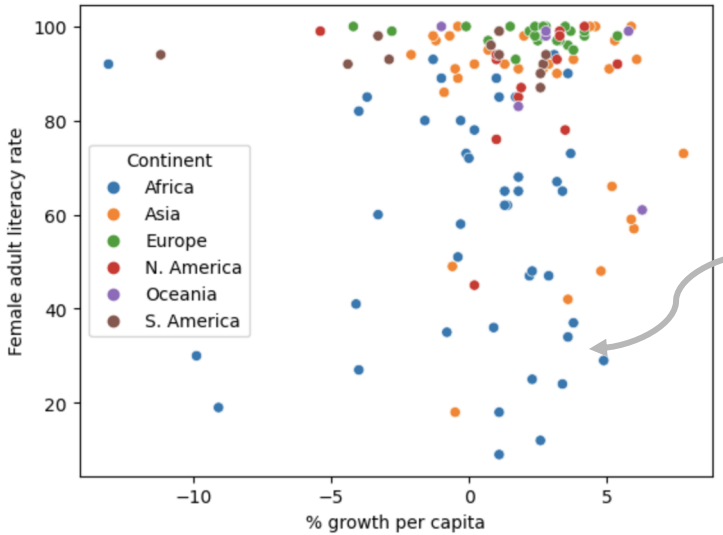


relationship appears linear, but with increasing spread as x gets larger

Scatter Plots

Scatter plots are used to reveal relationships between two quantitative variables.

- Plot one quantitative continuous variable on the x-axis, and second quantitative continuous variable on the y-axis.
- Each scatter point represents one datapoint in the dataset.



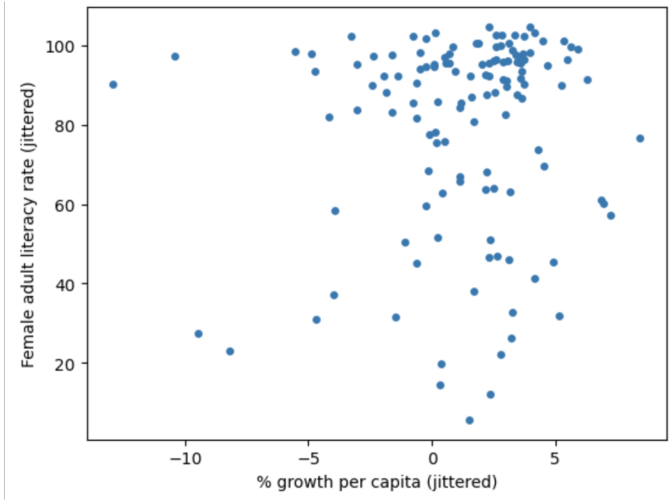

```
plt.scatter(x_values, y_values)
```

```
sns.scatterplot(data=df, x="x_column", \
                y="y_column", hue="hue_column")
```

Overplotting

The plot on the previous slide suffered from **overplotting** – scatter points all stacked on top of one another are difficult to see.

Jittering: adding a small amount of random noise to all x and y values to slightly move each scatter point. Main trends are still present, but individual datapoints are easier to distinguish.



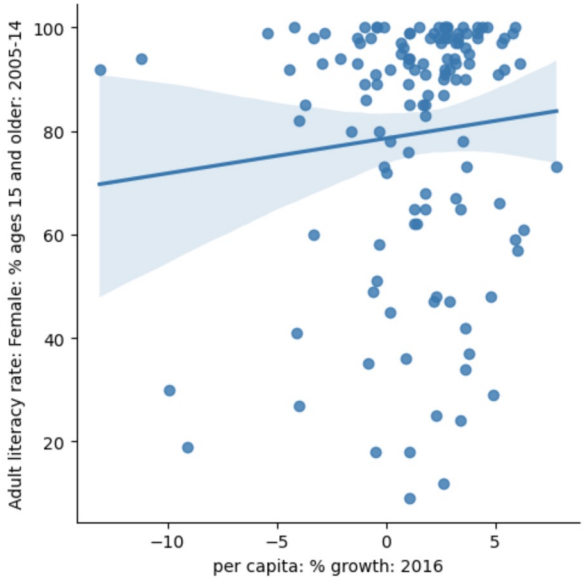

```
x_noise = np.random.uniform(-1, 1, len(wb))
y_noise = np.random.uniform(-5, 5, len(wb))

plt.scatter(wb['% growth'] + x_noise, \
            wb['Literacy rate: Female'] + y_noise, \
            s=15);
```

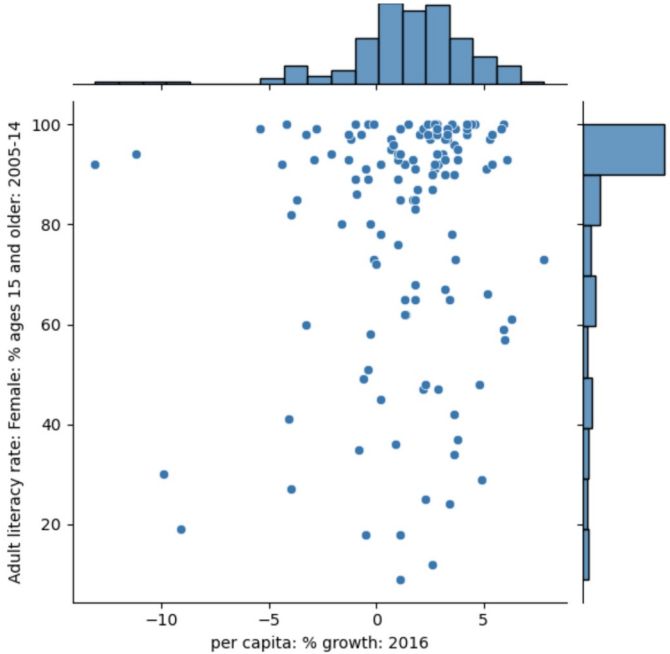
Decreasing point size also helps. `s` specifies the marker size in Matplotlib.

Scatter Plot Alternatives

Seaborn includes several built-in functions for making more complex scatter plots.




```
sns.lmplot(data=df, \
x="x_column", y="y_column")
```




```
sns.jointplot(data=df, \
x="x_column", y="y_column")
```

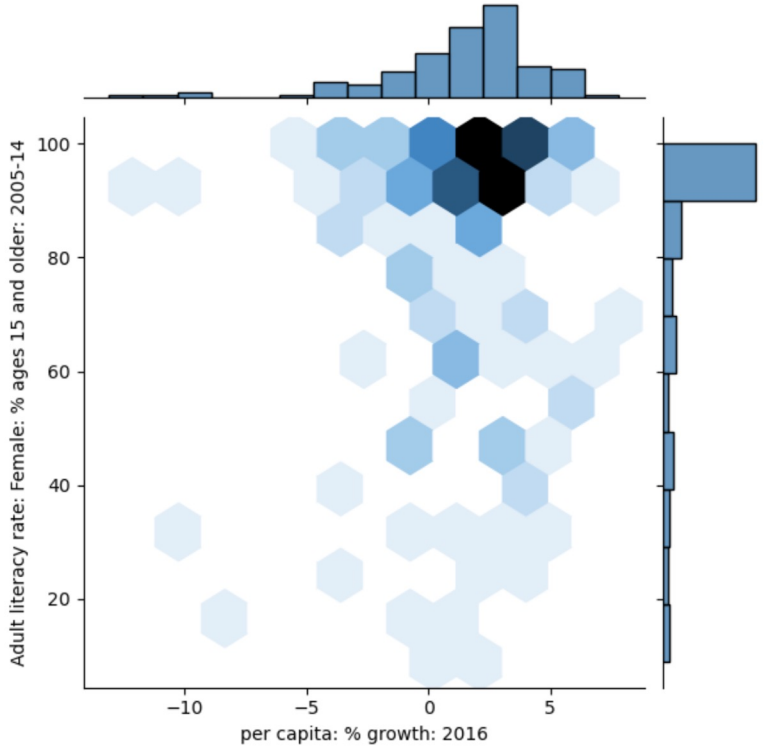
Hex Plots

Rather than plot individual datapoints, plot the *density* of their joint distribution.

Can be thought of as a two dimensional histogram.

- The xy plane is binned into hexagons.
- More shaded hexagons typically indicate a greater density/frequency = more datapoints lie in that spot

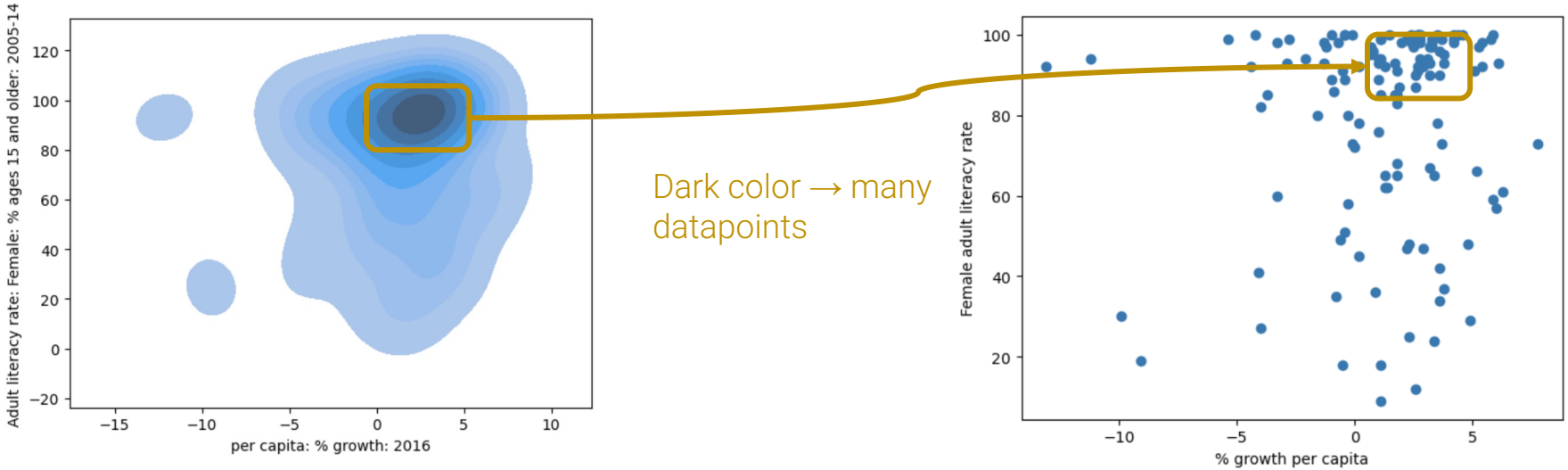
```
sns.jointplot(data=df, x="x_column", \
y="y_column", kind="hex")
```



Contour Plots

2-dimensional version of a KDE plot.

Similar to a topographic map – contour lines represent an area that has the *same density* of datapoints throughout. Darker colors indicate more datapoints in the region.




```
sns.kdeplot(data=df, x="x_column", y="y_column", fill=True)
```

Summary

- **Visualization requires a lot of thought!**
- Many tools for visualizing distributions.
  - Distribution of a single variable: rug plot, histogram, density plot, box, violin.
  - Joint distribution of two quantitative variables: scatter plot, hex plot, contour plot.
- This class primarily uses seaborn and matplotlib.
  - Pandas also has basic built-in plotting methods.
  - Many other visualization libraries exist. **`plotly`** is one of them.
    - plotly will occasionally appear in lecture code and assignments!

Transformations

Lecture 09

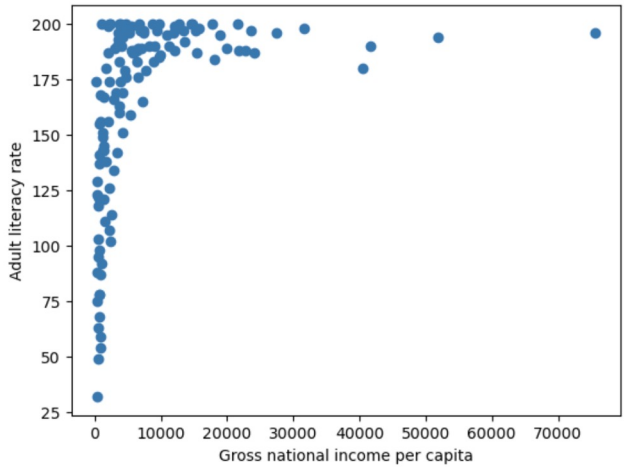
- Kernel Density Functions
  - KDE Mechanics
  - Kernel Functions and Bandwidth
  - Plotting Distributions - Revisited
- **Relationships between Quantitative Variables**
  - **Transformations**

Visualization Theory

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*.



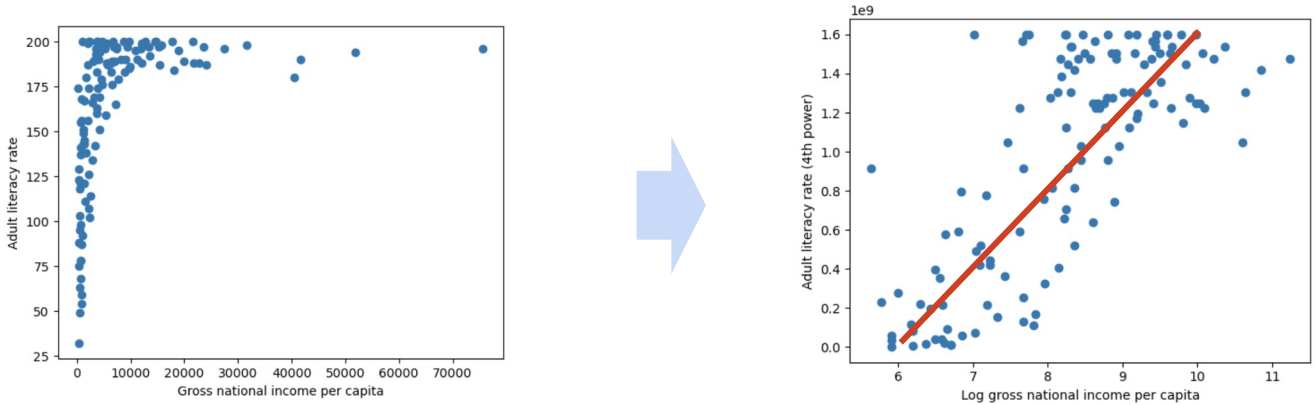
What problems are there here?

- 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.

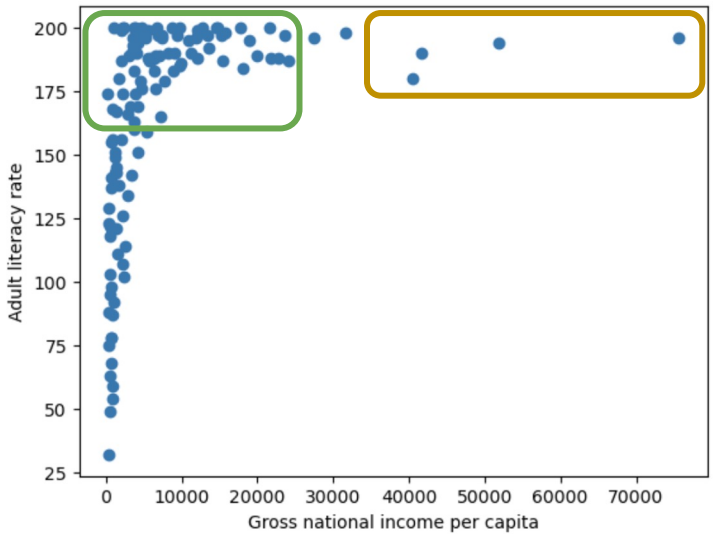


Why?

- Linear relationships are simple to interpret – we know how to work with slopes and intercepts to understand how two variables are related.

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.

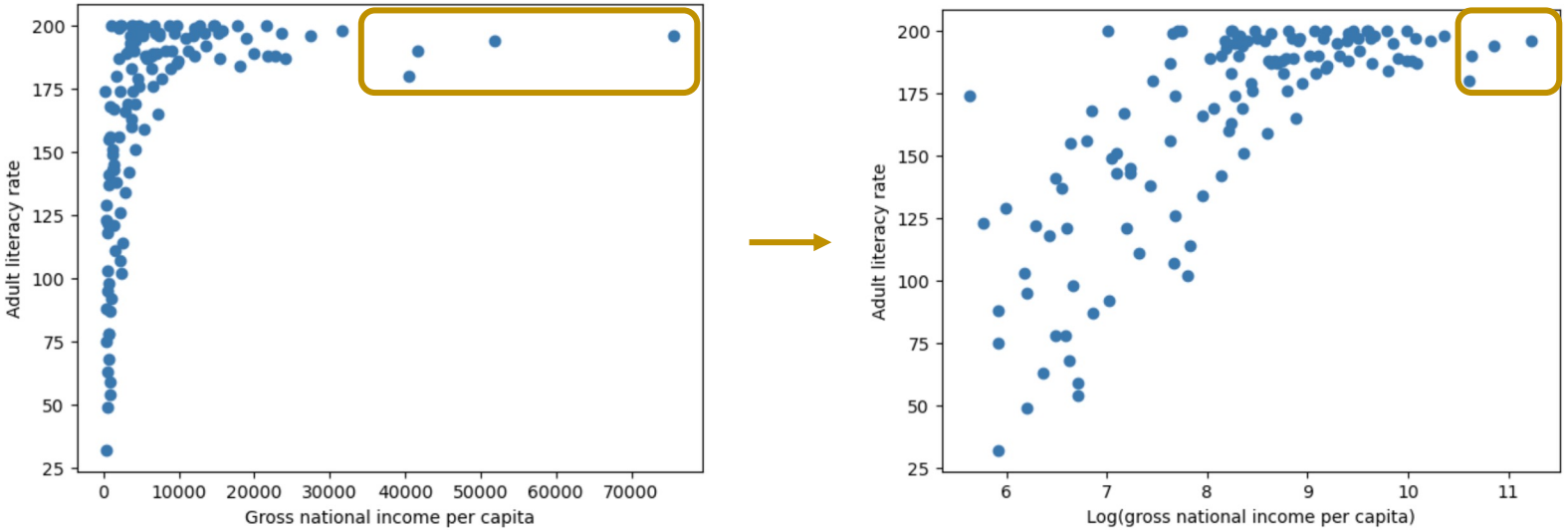
Applying Transformations

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.



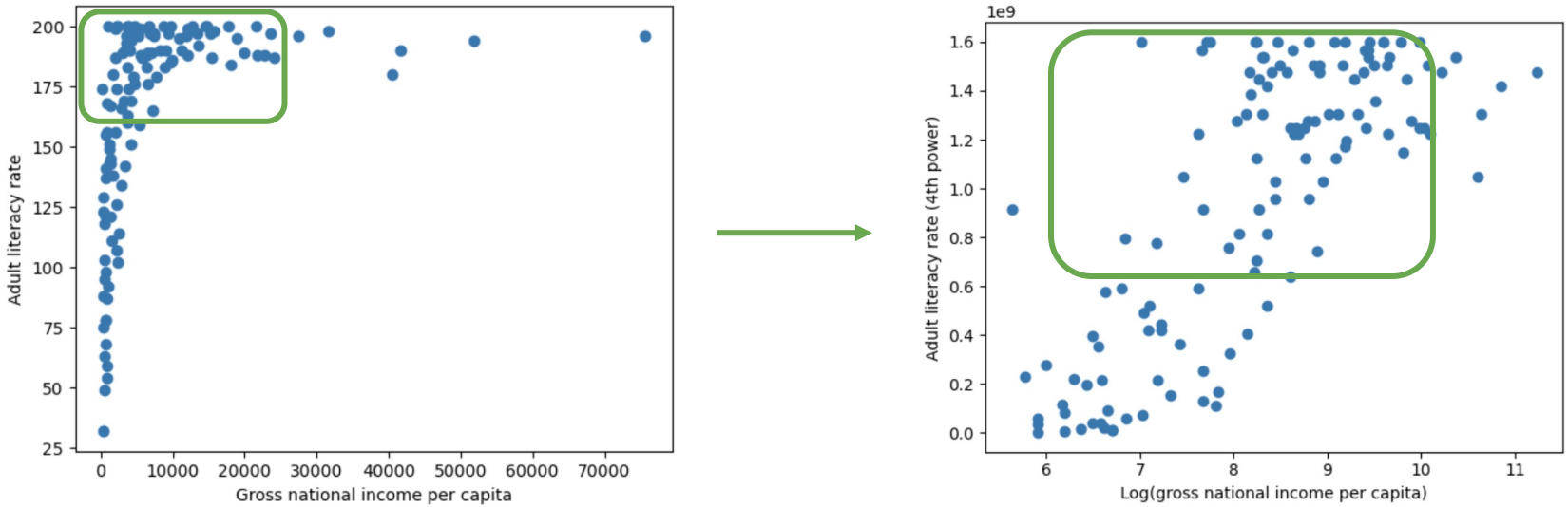
Applying Transformations

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.

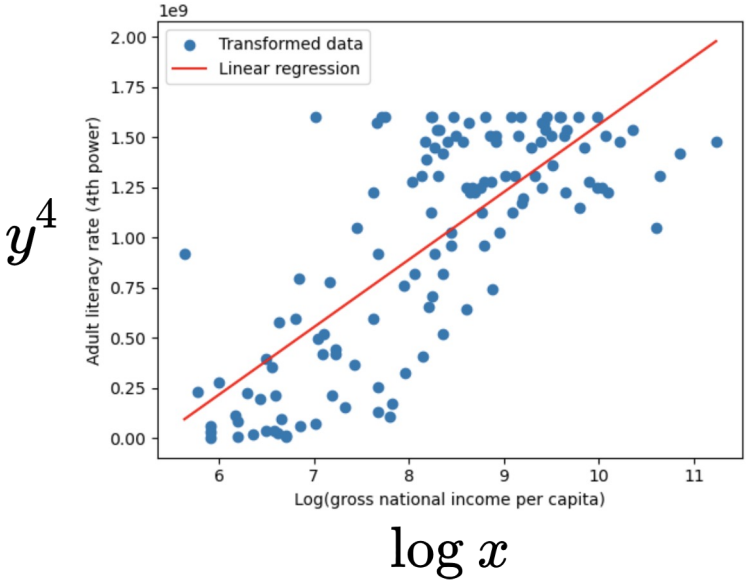


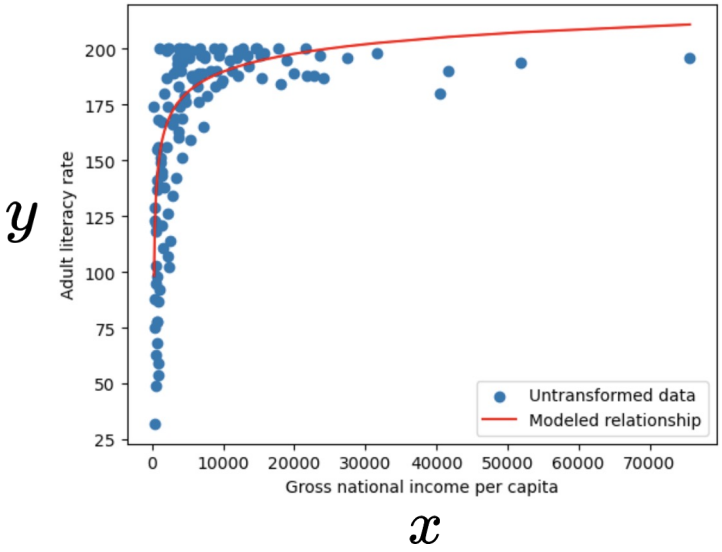
Interpreting Transformed Data

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

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

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



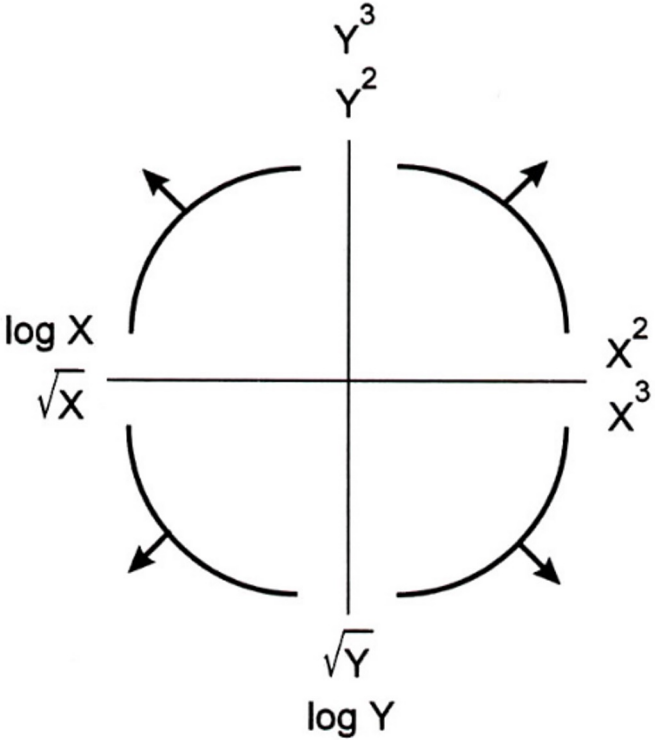


Tukey-Mosteller Bulge Diagram

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

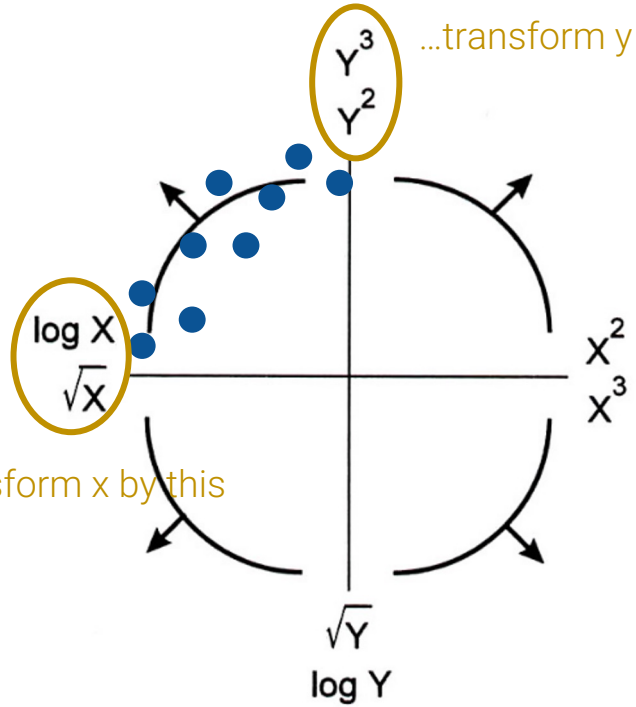
- A visual summary of the reasoning we just worked through.
- sqrt and log make a value "smaller".
- Raising to a value to a power makes it "bigger".
- There are multiple solutions. Some will fit better than others.

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



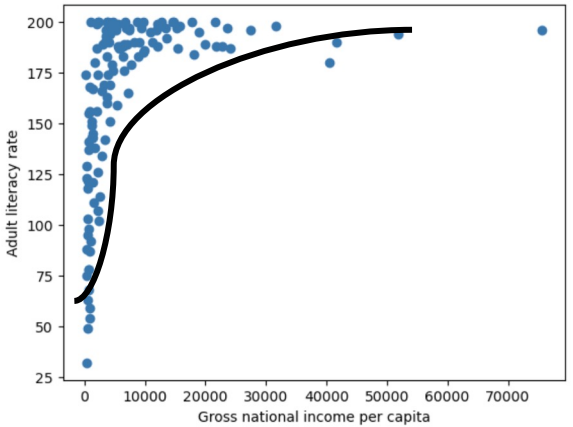
Tukey-Mosteller Bulge Diagram

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)

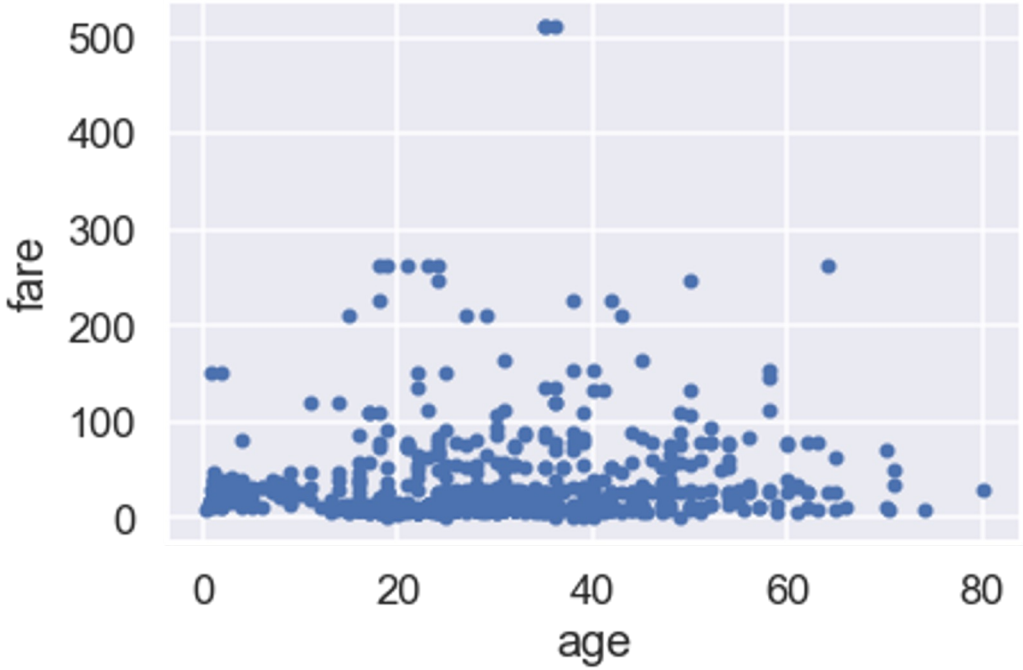
Visualization Theory

Lecture 09

- Kernel Density Functions
  - KDE Mechanics
  - Kernel Functions and Bandwidth
  - Plotting Distributions - Revisited
- Relationships between Quantitative Variables
  - Transformations
- **Visualization Theory**
  - Information Channels
  - Harnessing X/Y
  - Harnessing Color
  - Harnessing Markings
  - Harnessing Conditioning
  - Harnessing Context

Visualizations Are For Humans

“Looks like older people didn’t spend more money on tickets for the Titanic than younger people.”

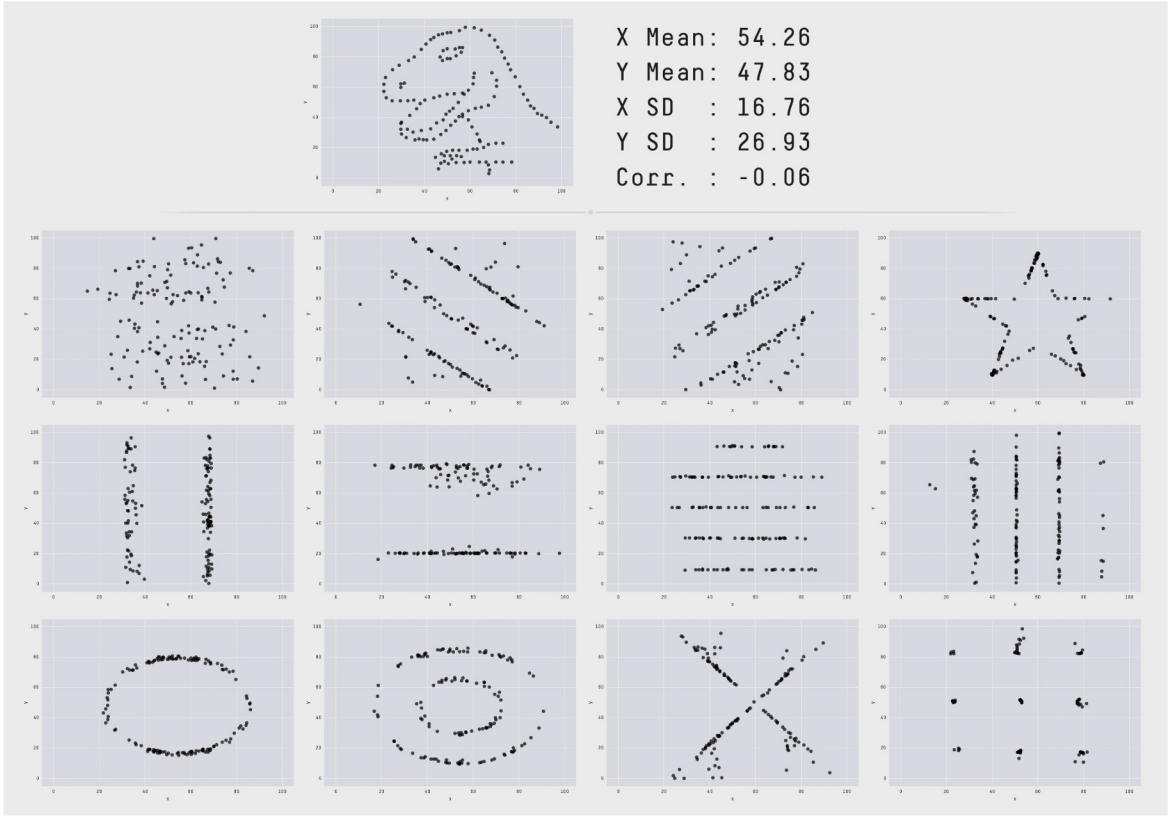


(Note: A histogram or KDE would give stronger evidence than a scatter plot.)

Visualizations Are More Expressive than Summary Statistics

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

Visualizations complement statistics.



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

Information Channels

Lecture 09

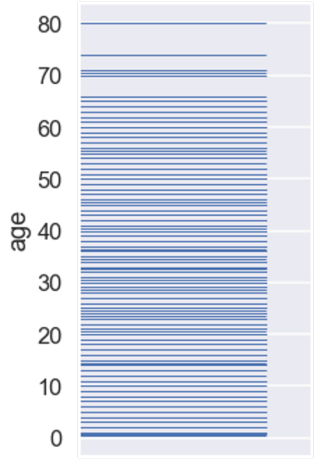
- Kernel Density Functions
  - KDE Mechanics
  - Kernel Functions and Bandwidth
  - Plotting Distributions - Revisited
- Relationships between Quantitative Variables
  - Transformations
- Visualization Theory
  - **Information Channels**
  - Harnessing X/Y
  - Harnessing Color
  - Harnessing Markings
  - Harnessing Conditioning
  - Harnessing Context

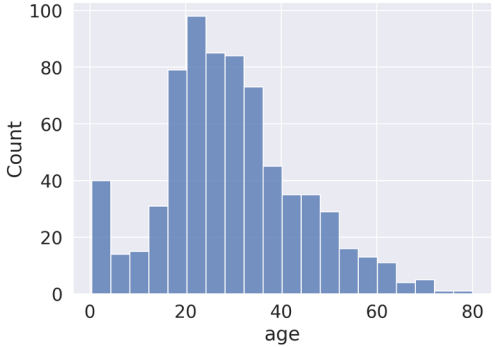
Take Advantage of the Human Visual Perception System

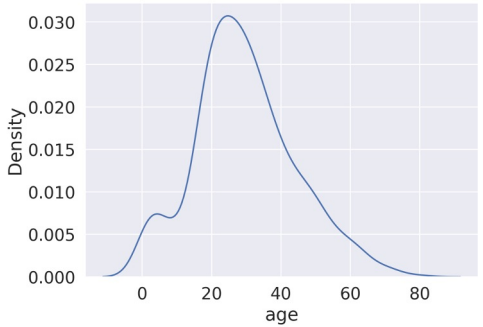
Data can be visualized in many ways!

- 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



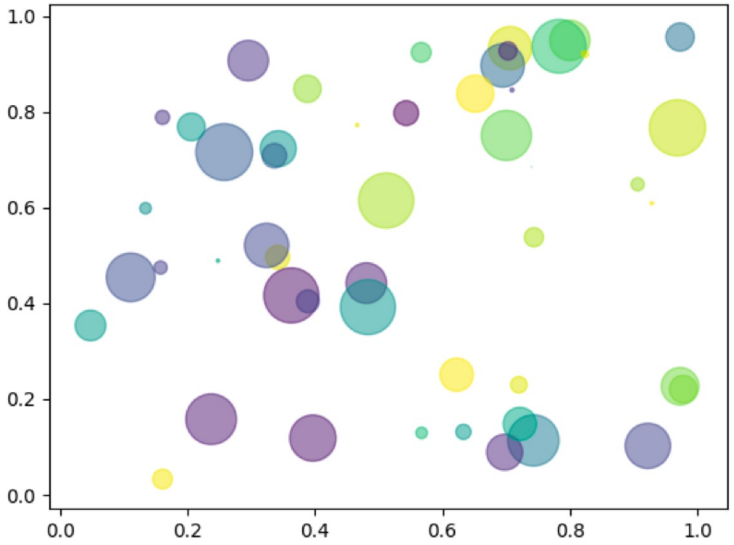




Going Beyond: Encoding 3+ Variables

How many variables are we encoding here?

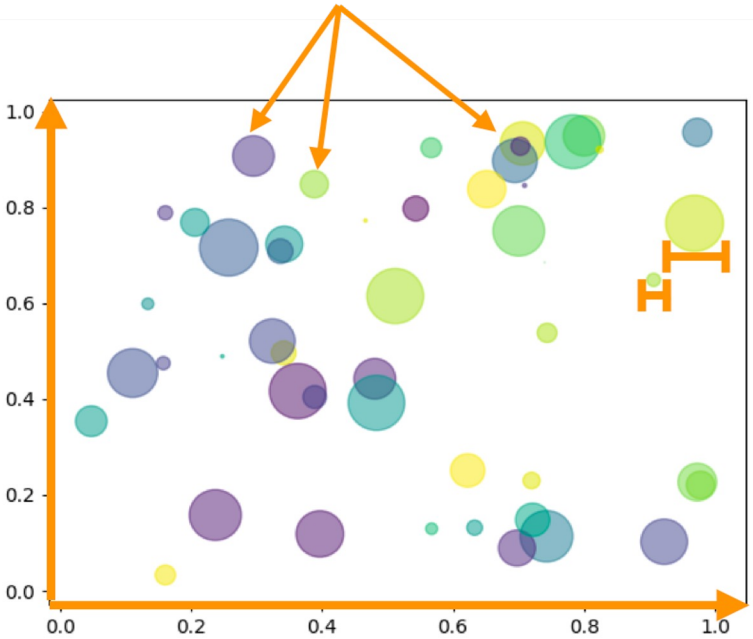
- In other words, how many "channels" of information are there?



Going Beyond: Encoding 3+ Variables

How many variables are we encoding here?

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



Answer: 4.

- x
- y
- area
- color

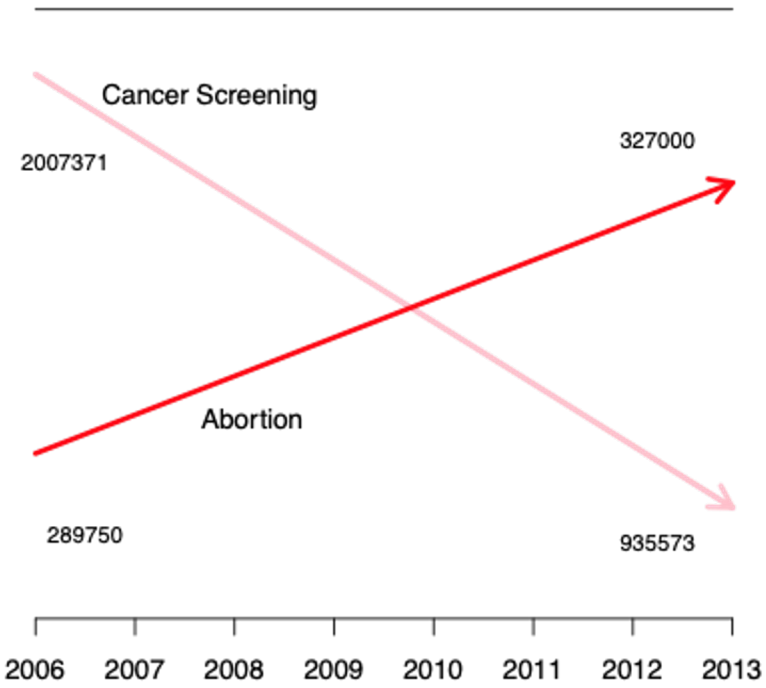
We could add even more: Shapes, outline colors of shapes, shading, etc.
There are infinite possibilities!

Harnessing X/Y

Lecture 09

- Kernel Density Functions
  - KDE Mechanics
  - Kernel Functions and Bandwidth
  - Plotting Distributions - Revisited
- Relationships between Quantitative Variables
  - Transformations
- Visualization Theory
  - Information Channels
  - **Harnessing X/Y**
  - Harnessing Color
  - Harnessing Markings
  - Harnessing Conditioning
  - Harnessing Context

Case Study: Planned Parenthood Hearing

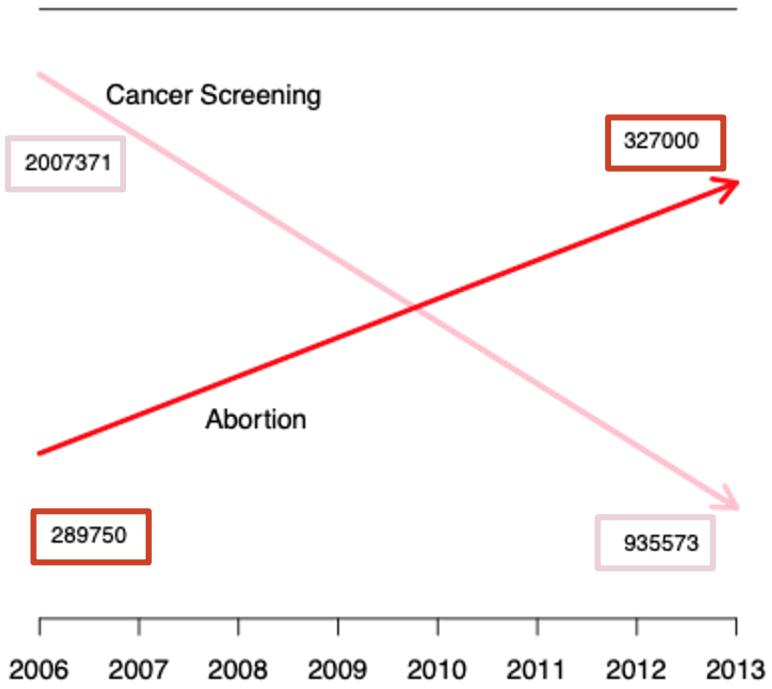


In 2015, Planned Parenthood was accused of selling aborted fetal tissue for profit.

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!

In 2013:

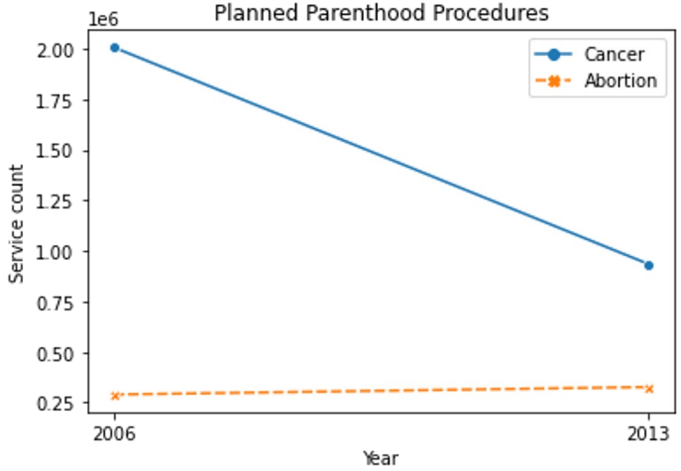
- 327000 is smaller than 935573…
- …but appears to be way bigger??

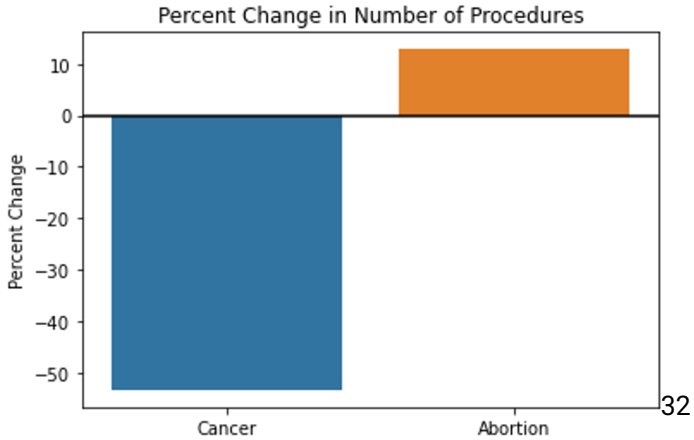
Do not use two different scales for the same axis!

Always Consider the Scale When Comparing "Similar" Data

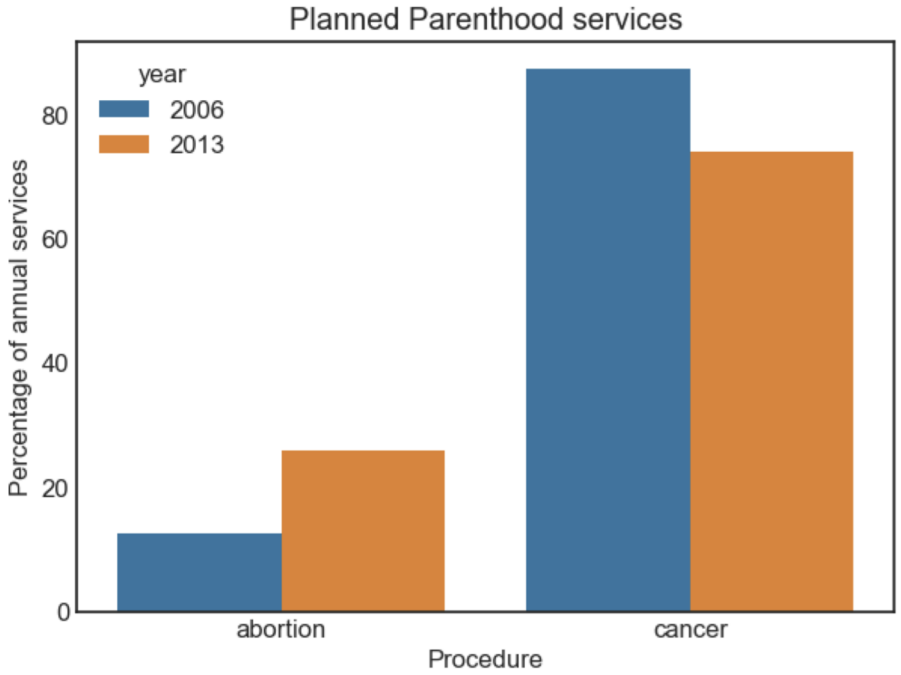
The top plot draws all of the data on the same scale.

- It clearly shows there was a dramatic drop in cancer screenings by PP.
- 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.





Always Consider the Scale When Comparing "Similar" Data



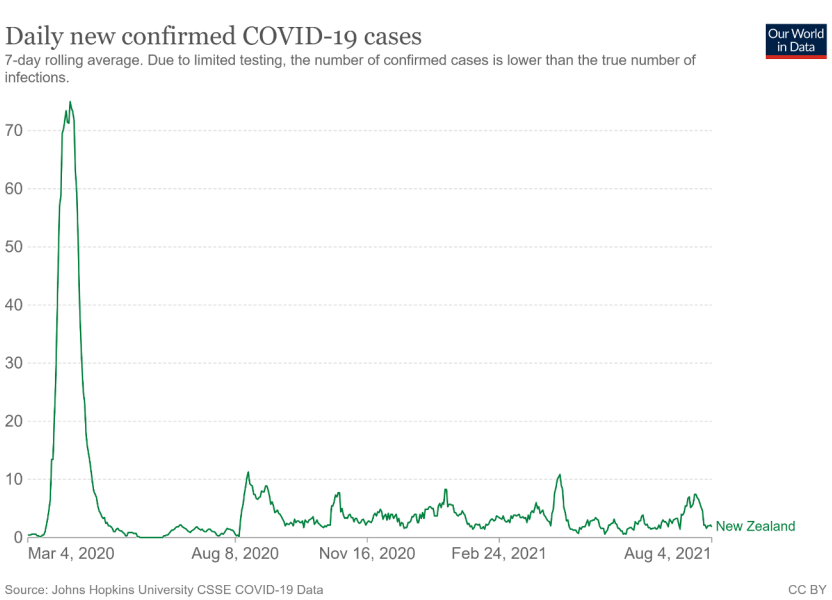
We could also visualize abortions and cancer screenings as a percentage of total procedures.

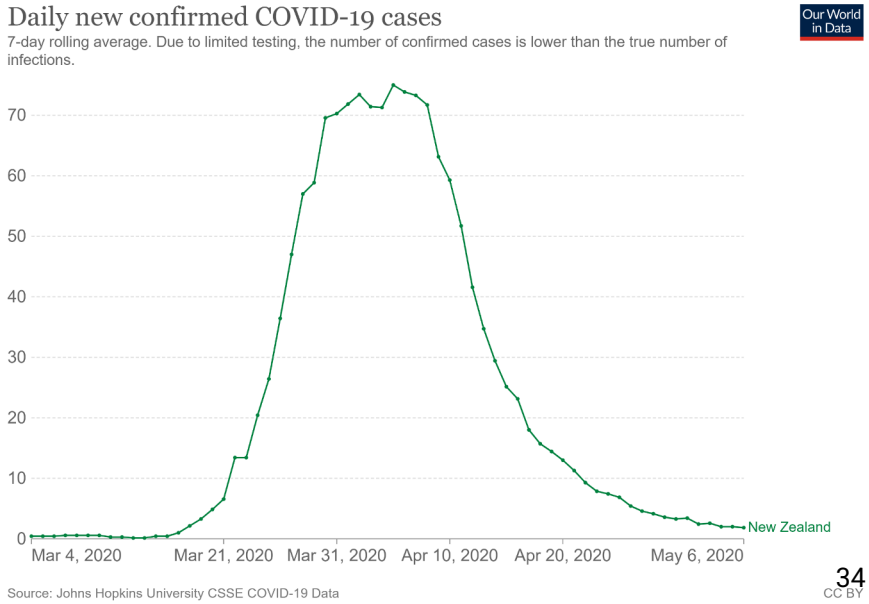
- Abortions increased from 13% to 26% of total procedures.

Reveal the Data

Recommendations:

- Choose axis limits to fill the visualization.
- **You don’t have to visualize all of the data at once:**
  - Zoom in on the bulk of the data (it's ok to not include 0!) if only one part matters.
  - Can also create multiple plots to show different regions of interest.





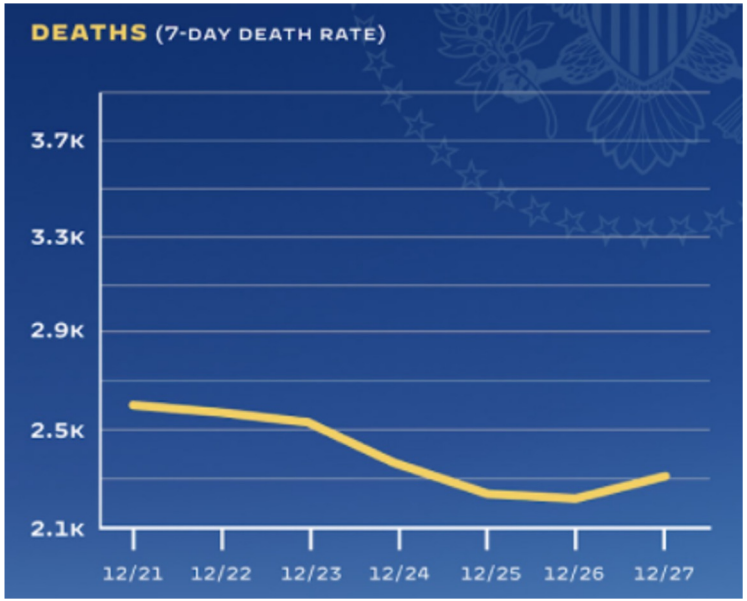
Reveal the Data

Recommendations:

- **Choose axis limits to fill the visualization.**
- You don't have to visualize all of the data at once:
  - Zoom in on the bulk of the data (it's ok to not include 0!) if only one part matters.
  - Can also create multiple plots to show different regions of interest.

Terrible White House COVID-19 visualization:

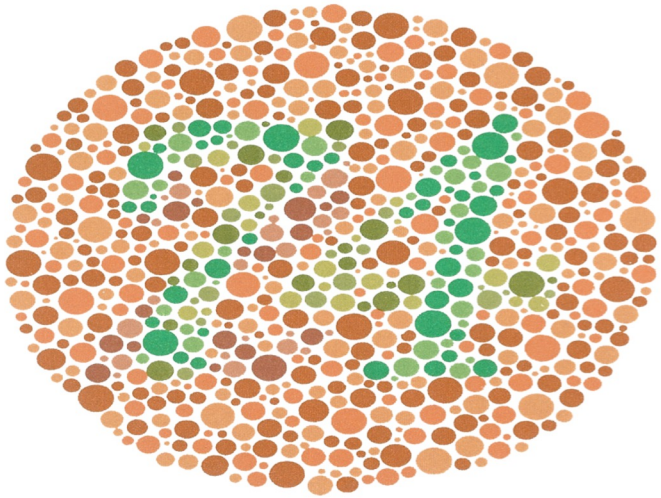
- Mysterious maximum value on y-axis.



Harnessing Color

Lecture 09

- Kernel Density Functions
  - KDE Mechanics
  - Kernel Functions and Bandwidth
  - Plotting Distributions - Revisited
- Relationships between Quantitative Variables
  - Transformations
- Visualization Theory
  - Information Channels
  - Harnessing X/Y
  - **Harnessing Color**
  - Harnessing Markings
  - Harnessing Conditioning
  - Harnessing Context

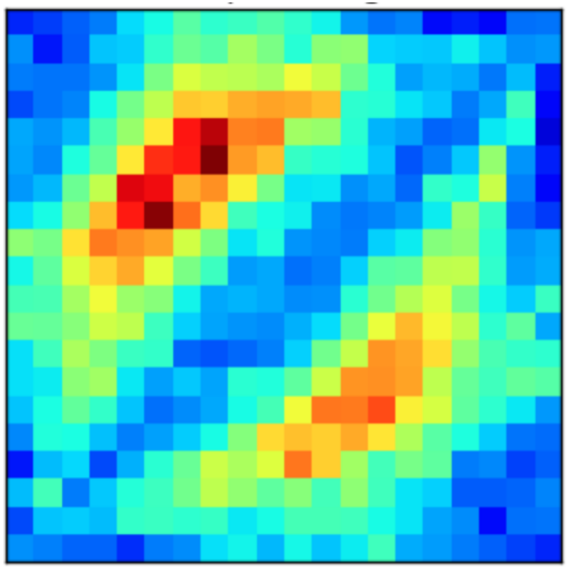


Choosing a set of colors which work together is a challenging task!

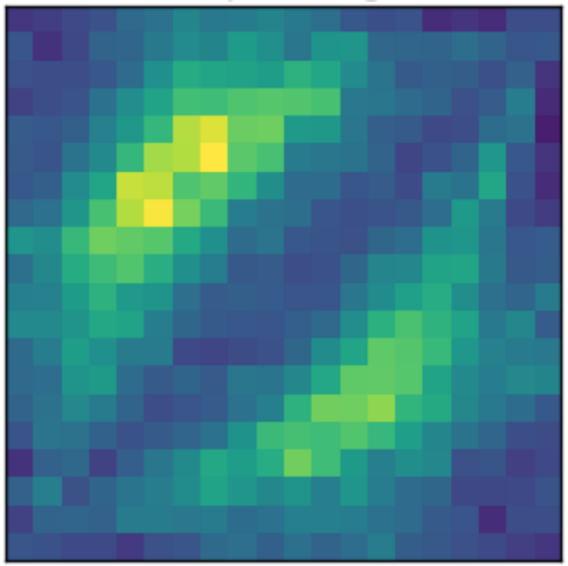
Perception of Color

Download the Color Oracle App to simulate common color vision impairments.

Colormaps

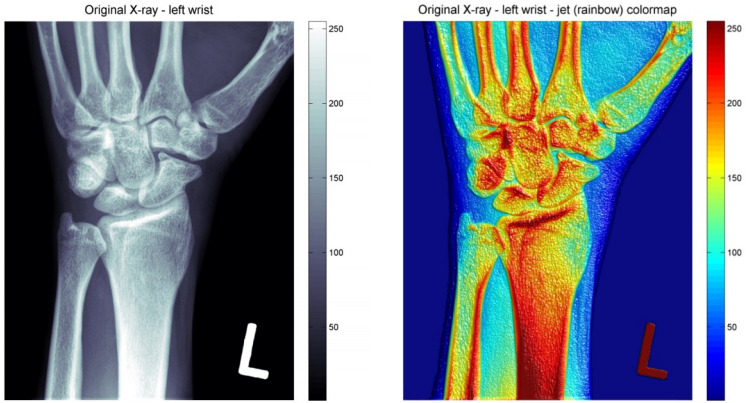


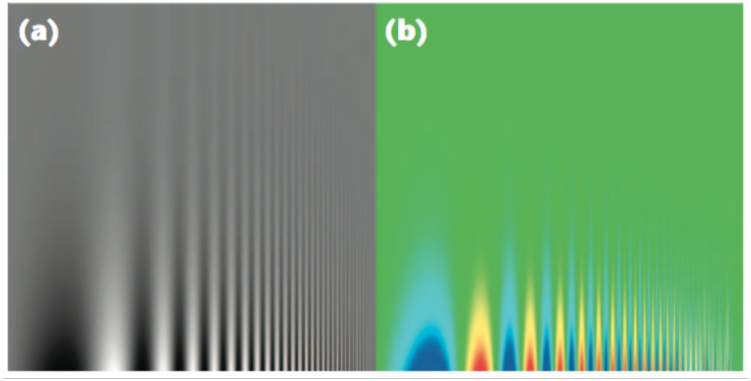
Jet

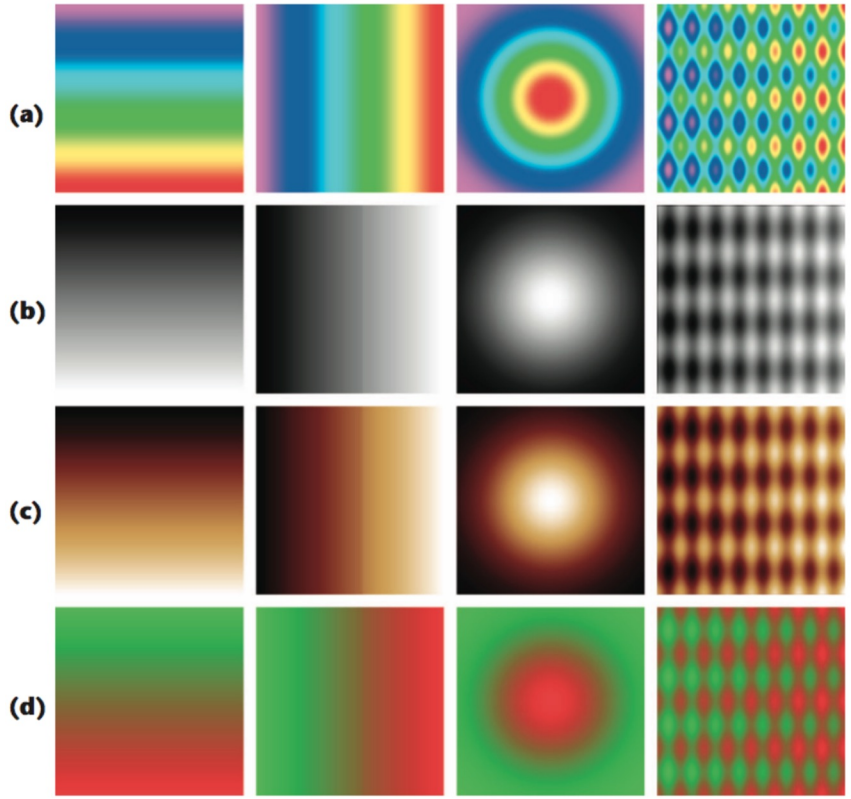


Viridis

The Jet/Rainbow Colormap Actively Misleads



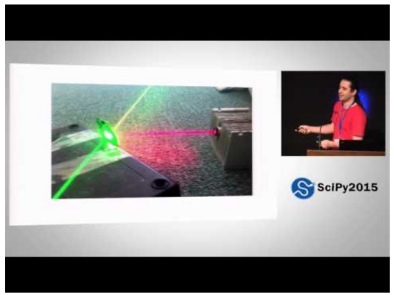


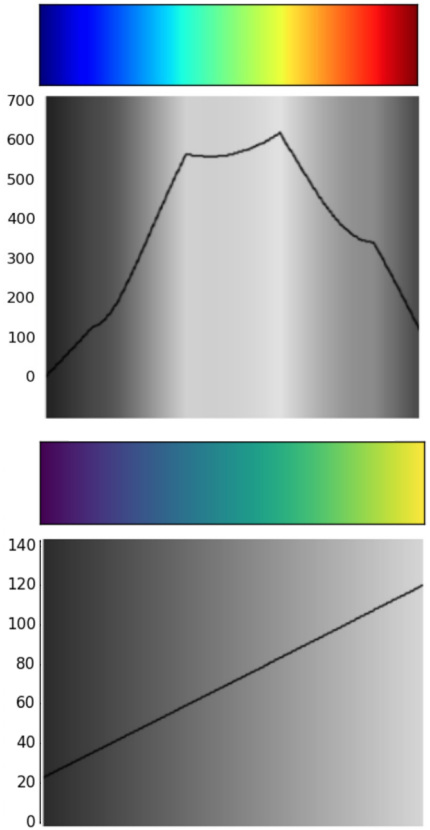


"Rainbow Colormap (Still) Considered Harmful", Borland and Taylor, 2007.

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
- Avoid combinations of red and green, due to red-green color blindness.





Bounces all over

Slope is constant

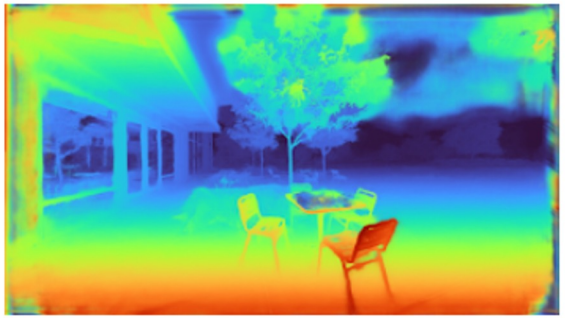
x-axis is color,
y-axis is "lightness"

Except when not :) The Google Turbo Colormap

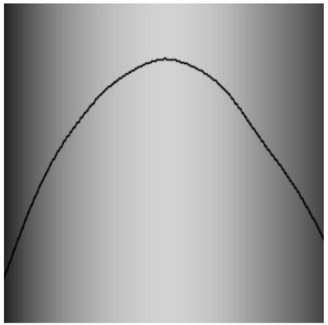
Turbo

Jet

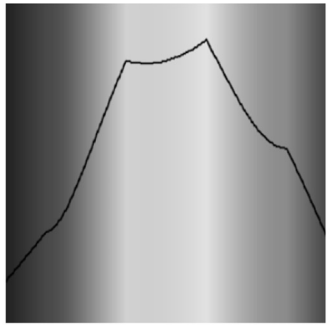
Inferno



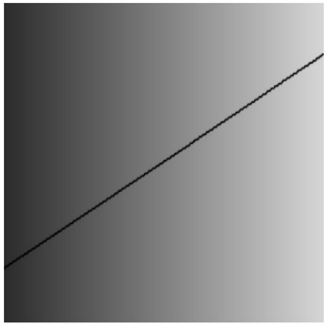
Turbo



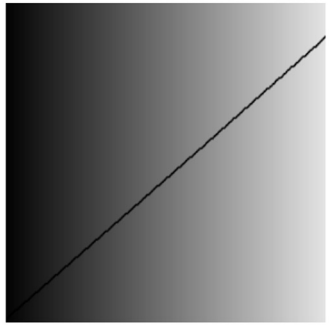
Turbo



Jet



Viridis



Inferno

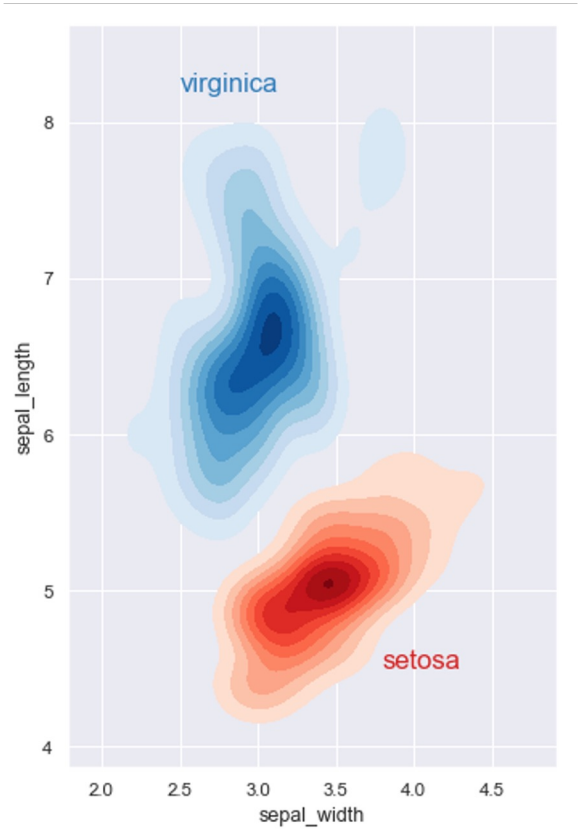
X-axis is color, y-axis is “lightness”

https://ai.googleblog.com/2019/08/turbo-improved-rainbow-colormap-for.html

Use Color to Highlight Data Type

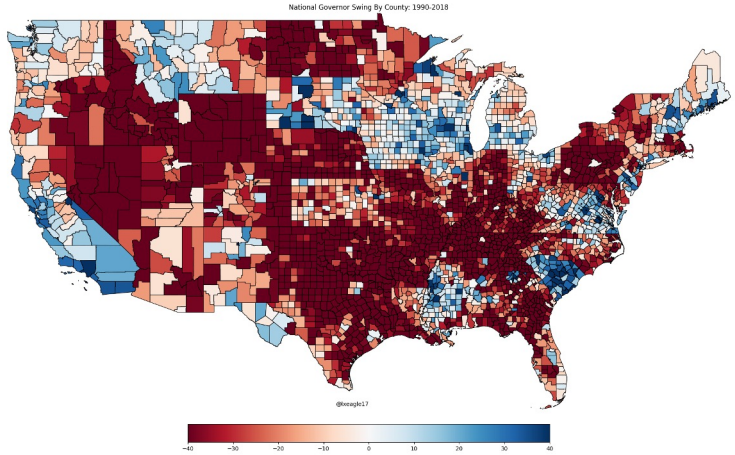
- **Qualitative**: Choose a qualitative scheme that makes it easy to distinguish between categories.
  - One category isn't "higher" or "lower" than another.
- **Quantitative**: Choose a color scheme that visualizes magnitude of change.

The plot on the right has both distinctions!



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.

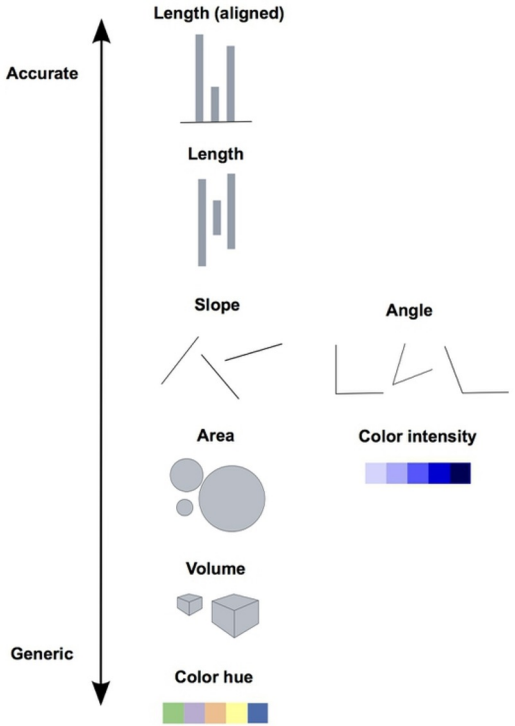


If low and high values deserve equal emphasis, use a **diverging** scheme where lighter colors represent middle values.

Harnessing Markings

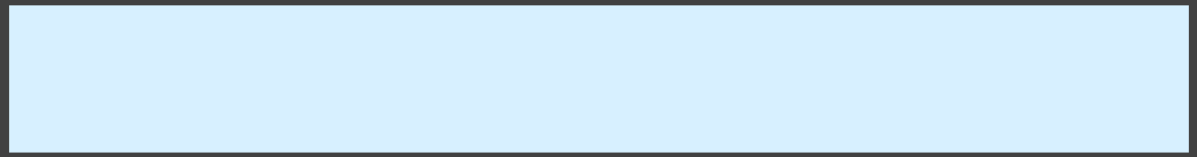
Lecture 09

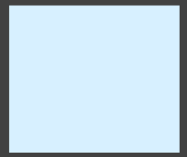
- Kernel Density Functions
  - KDE Mechanics
  - Kernel Functions and Bandwidth
  - Plotting Distributions - Revisited
- Relationships between Quantitative Variables
  - Transformations
- Visualization Theory
  - Information Channels
  - Harnessing X/Y
  - Harnessing Color
  - **Harnessing Markings**
  - Harnessing Conditioning
  - Harnessing Context



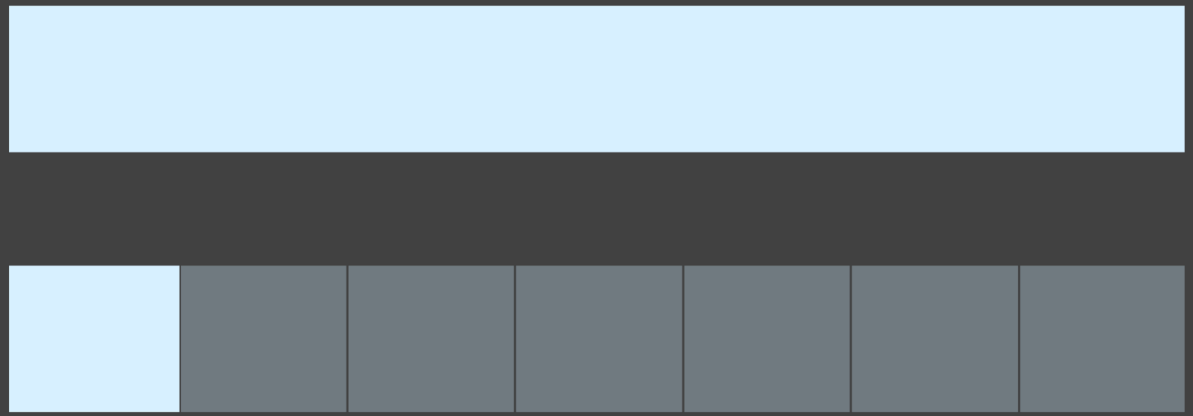
The accuracy of our judgements depend on the type of marking.

Perception of Markings

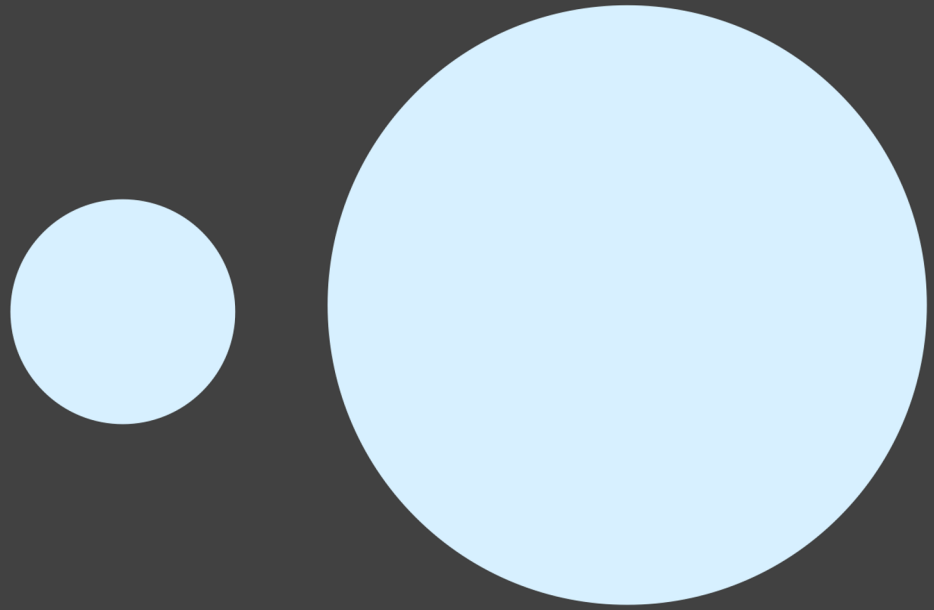




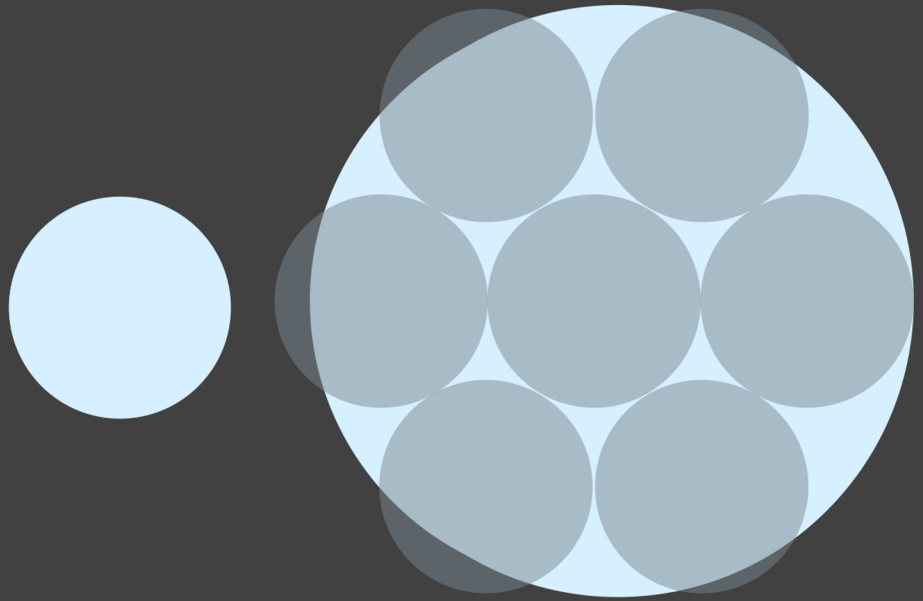
How much longer is the long bar?



The long bar is 7 times longer than the short bar.

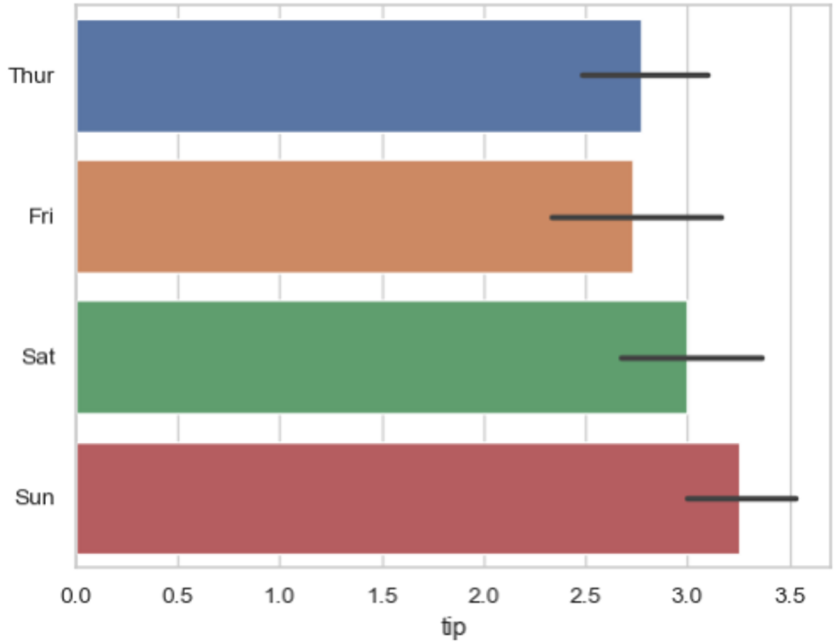


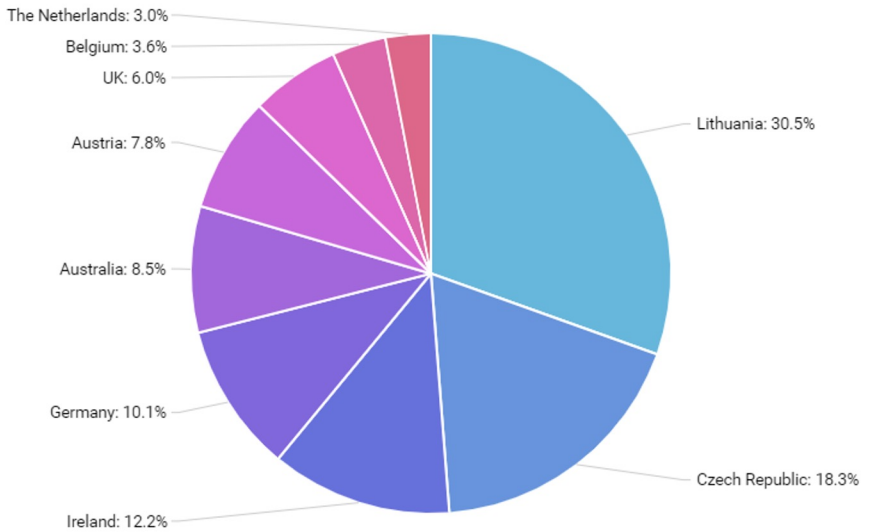
How much bigger is the big circle?



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

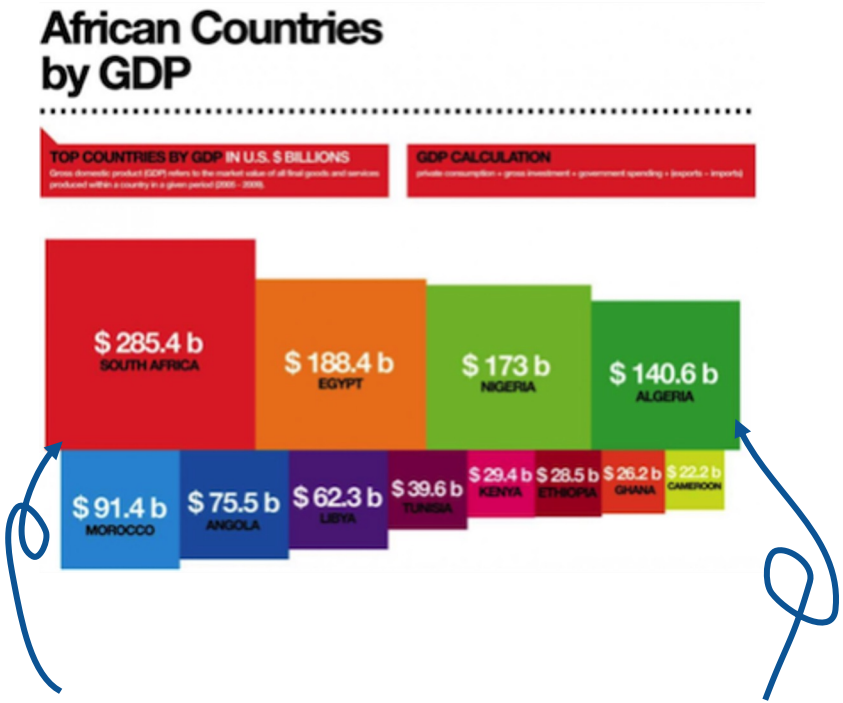
Lengths Are Easy to Distinguish. Others, Like Angles, Are Hard.



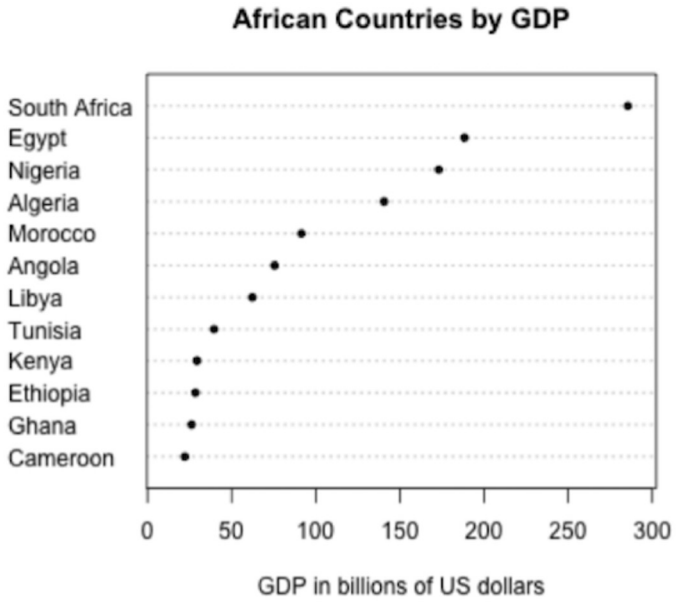


Don't use pie charts! Visual angle judgments are inaccurate.

Areas Are Hard to Distinguish

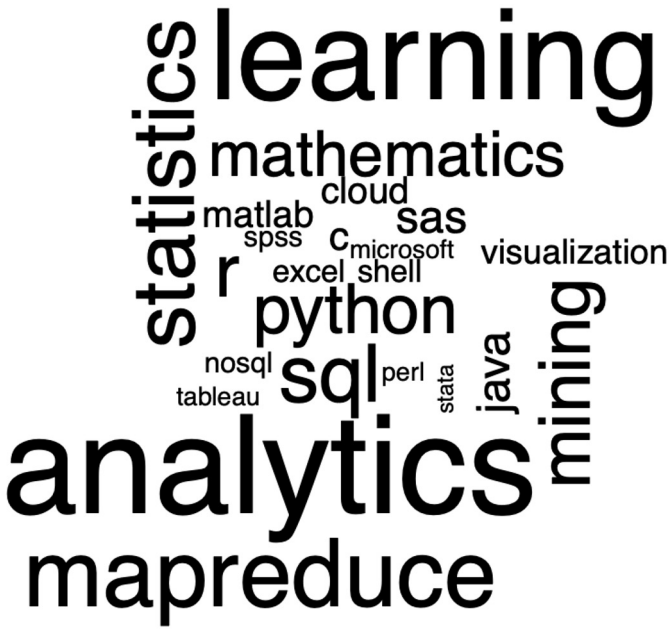


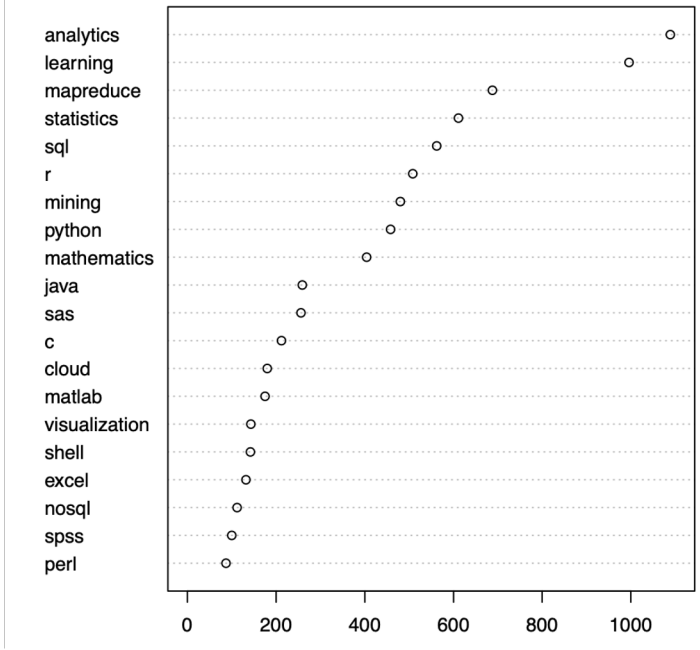
(South Africa has twice the GDP of Algeria, but that isn't clear from the areas.)



Avoid area charts!
Visual area judgments are inaccurate.

Areas Are Hard to Distinguish





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

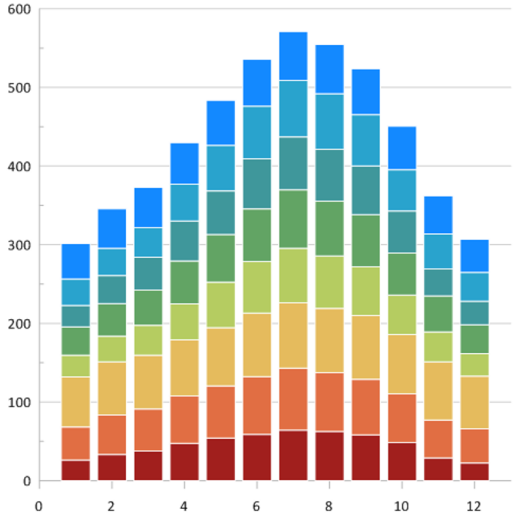
…that being said, if you are not trying to make quantifiable comparisons, then word clouds are useful for "the idea."

Avoid "Jiggling" the Baseline!

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

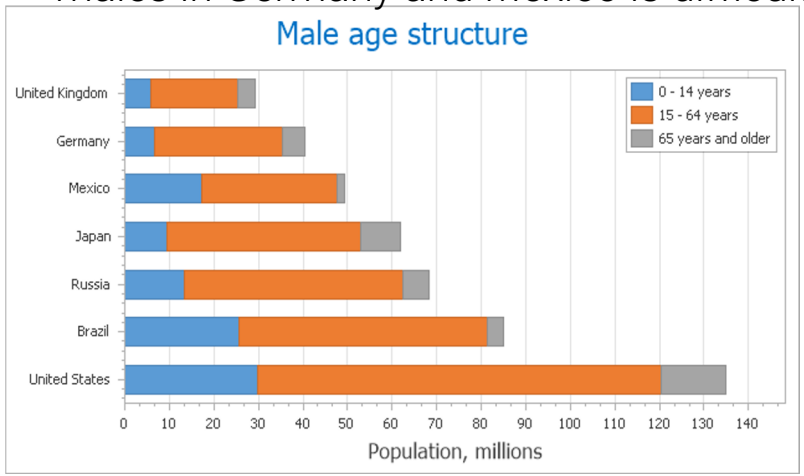
In the first plot:

- The top blue bars are all roughly of the same length.
- Not immediately obvious!



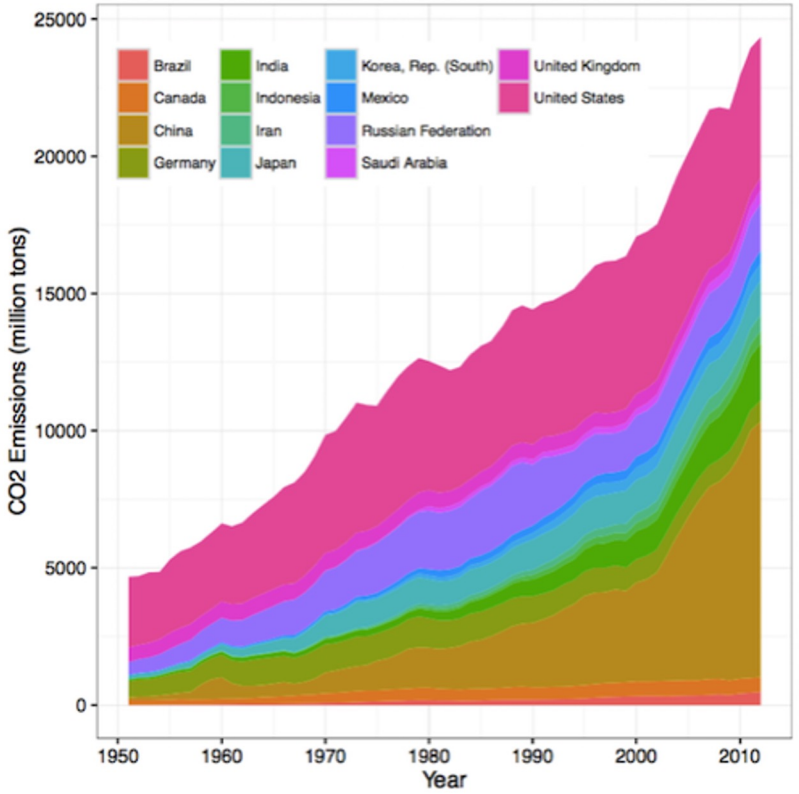
In the second plot:

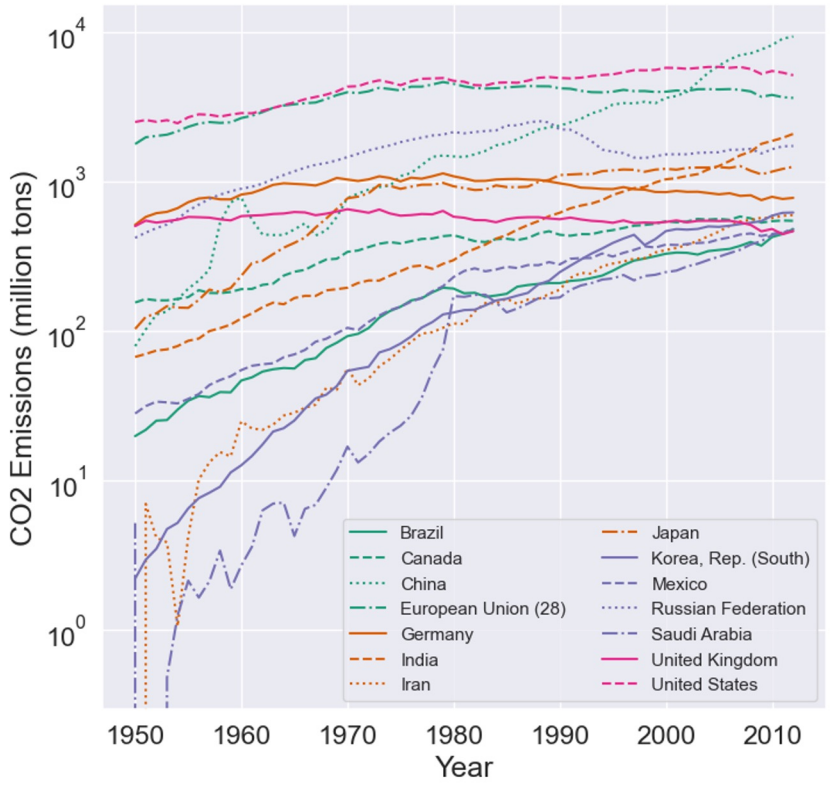
- Comparing the number of 15-64 year old males in Germany and Mexico is difficult.



Avoid Jiggling the Baseline

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



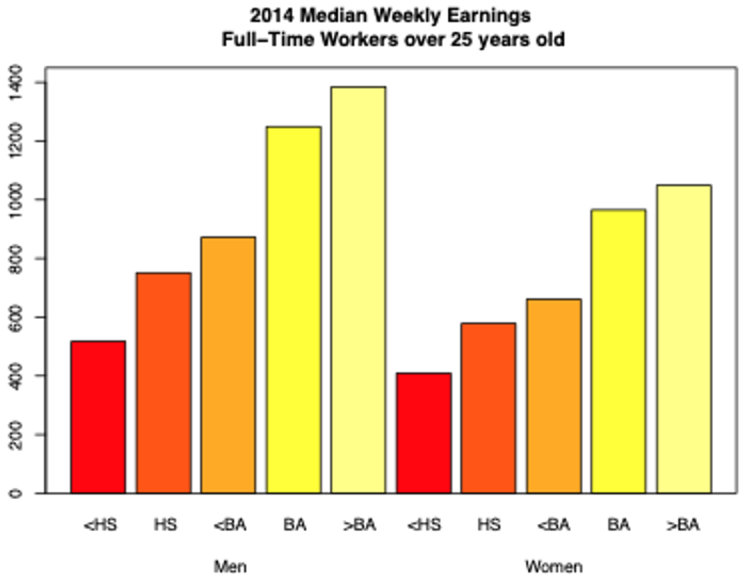


Harnessing Conditioning

Lecture 09

- Kernel Density Functions
  - KDE Mechanics
  - Kernel Functions and Bandwidth
  - Plotting Distributions - Revisited
- Relationships between Quantitative Variables
  - Transformations
- Visualization Theory
  - Information Channels
  - Harnessing X/Y
  - Harnessing Color
  - Harnessing Markings
  - **Harnessing Conditioning**
  - Harnessing Context

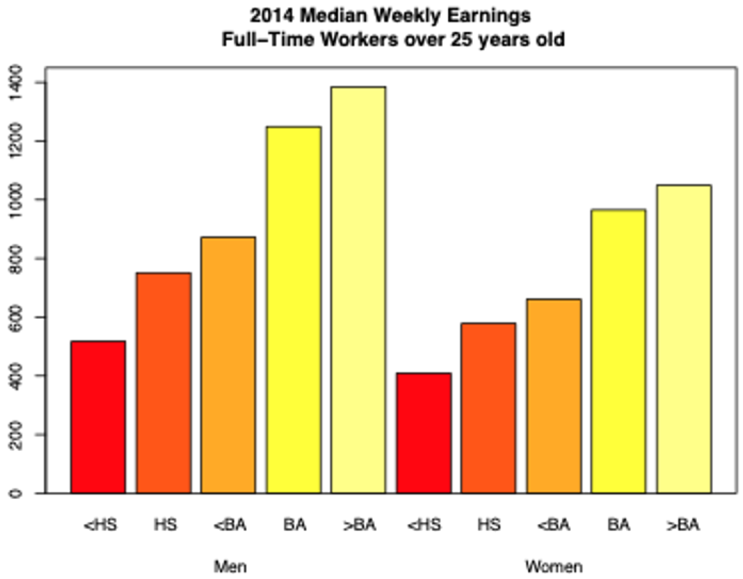
Use Conditioning to Aid Comparison



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



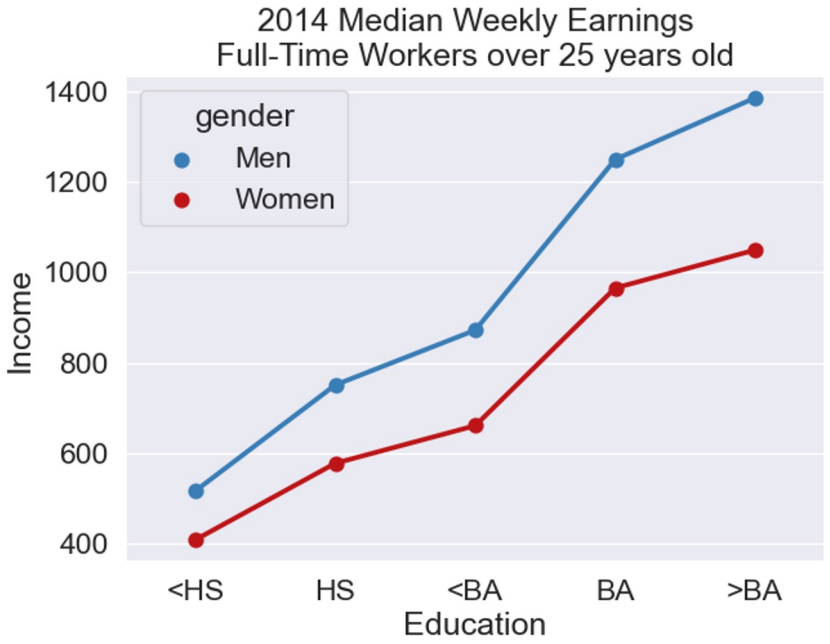
This data comes from the Bureau of Labor Statistics, who oversees surveys regarding the economic health of the US. They have plotted median weekly earnings for men and women by education level.

- 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?

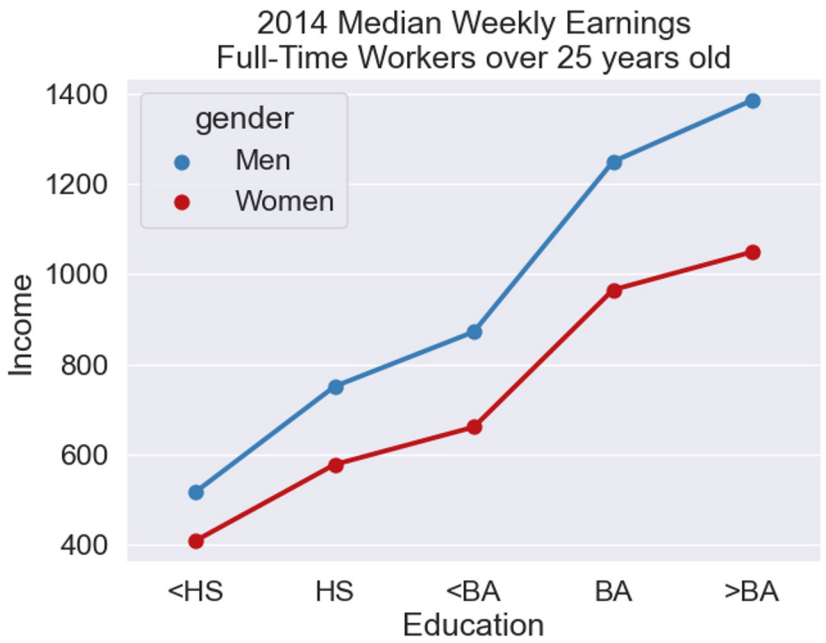
Use Conditioning to Aid Comparison

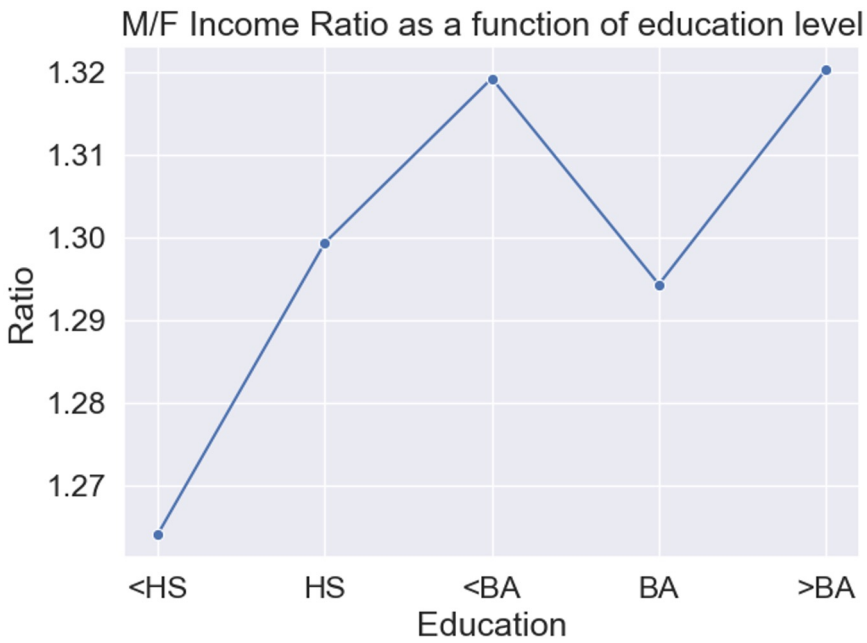


- Easy to see the effect of education on earnings.
- Hard to compare between the two genders in the dataset.

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

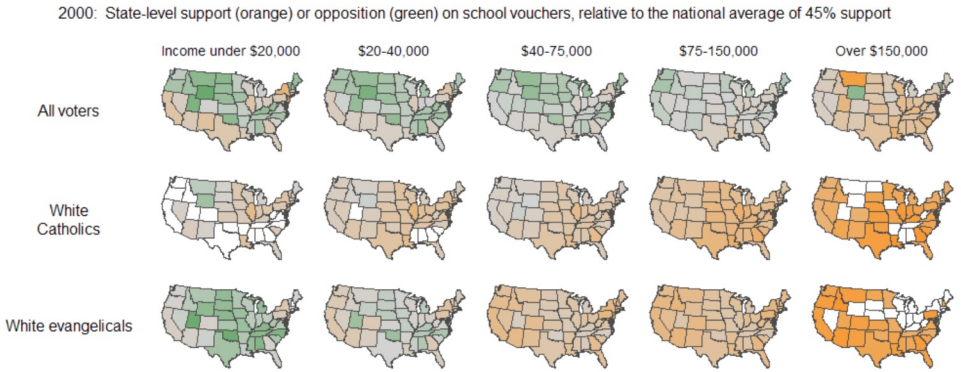
How Does the Income Gap Increase with Education?





See notebook for how to get this figure with groupby!

Other Notes: 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)

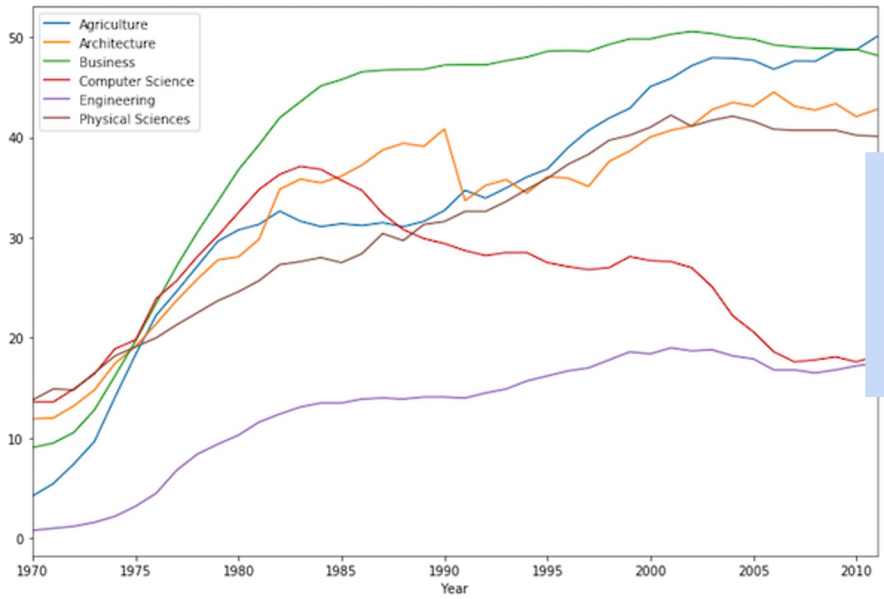
Juxtaposition: placing multiple plots side by side, with the same scale (called "small multiples") (see left).

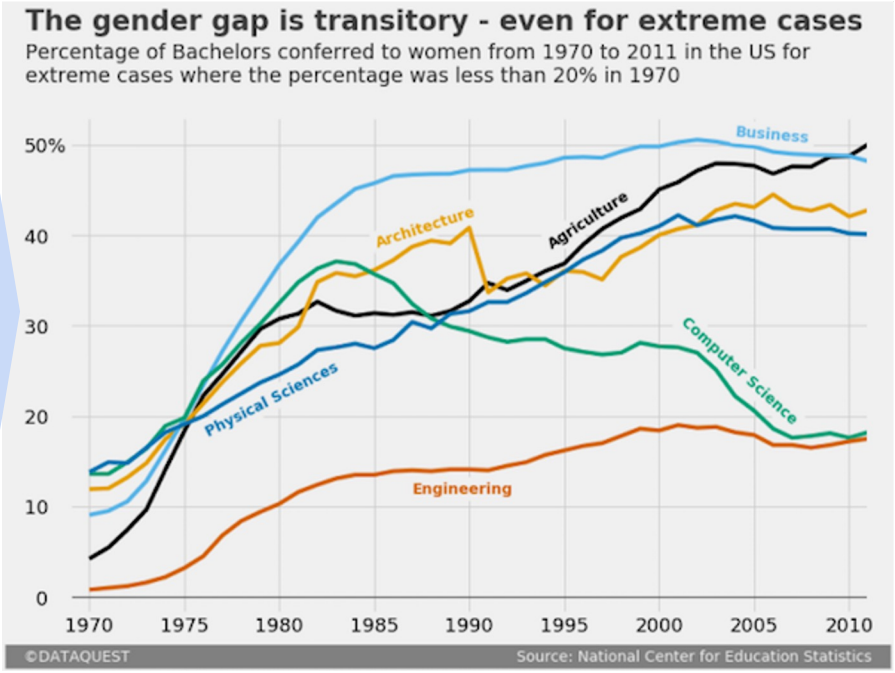
Harnessing Context (for Publication)

Lecture 09

- Kernel Density Functions
  - KDE Mechanics
  - Kernel Functions and Bandwidth
  - Plotting Distributions - Revisited
- Relationships between Quantitative Variables
  - Transformations
- Visualization Theory
  - Information Channels
  - Harnessing X/Y
  - Harnessing Color
  - Harnessing Markings
  - Harnessing Conditioning
  - **Harnessing Context**

Getting Ready for Publication





Publication-Ready: 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 **axis labels**.

Publication-Ready: Captions

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 **axis labels**.

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:

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

A Captioned, Publication-Ready (Famous) Figure

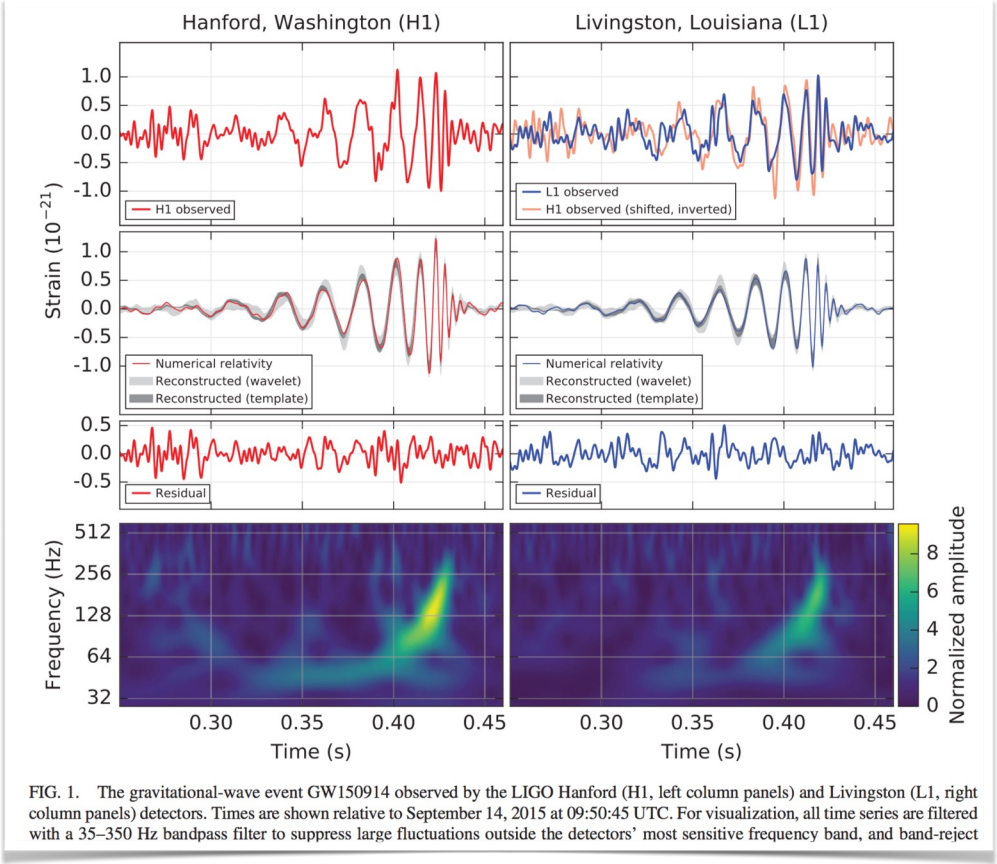


FIG. 1. The gravitational-wave event GW150914 observed by the LIGO Hanford (H1, left column panels) and Livingston (L1, right column panels) detectors. Times are shown relative to September 14, 2015 at 09:50:45 UTC. For visualization, all time series are filtered with a 35–350 Hz bandpass filter to suppress large fluctuations outside the detectors' most sensitive frequency band, and band-reject

Figure 1

The gravitational-wave event GW150914 observed by the LIGO Hanford (H1, left column panels) and Livingston (L1, right column panels) detectors. Times are shown relative to September 14, 2015 at 09:50:45 UTC. For visualization, all time series are filtered with a 35–350 Hz bandpass filter to suppress large fluctuations outside the detectors' most sensitive frequency band, and band-reject filters to remove the strong instrumental spectral lines seen in the Fig. 3 spectra. *Top row, left:* H1 strain. *Top row, right:* L1 strain. GW150914 arrived first at L1 and $6.9^{+0.5}_{-0.4}$ ms later at H1; for a visual comparison, the H1 data are also shown, shifted in time by this amount and inverted (to account for the detectors' relative orientations). *Second row:* Gravitational-wave strain projected

"Observation of Gravitational Waves from a Binary Black Hole Merger" - 2017 Nobel Prize in Physics.

https://www.gw-openscience.org/tutorials