DS 100: Principles and Techniques of Data Science **Date: July 1, 2019**

Discussion #3

Name:

Data Visualization and Scope

1. The first part of the discussion will be centered on this video

   `https://tinyurl.com/data100-rosling`

   Answer the following questions about the quality of the visualization in the video.

   The intended progression of this discussion is:

   1. What is represented?
   2. How is it represented?
   3. How do we make the representation interpretable?
   4. In making the representation interpretable, did we have to "change" the data?
   5. Did we miss anything by just looking at visualization?

   (a) How are the variables being represented visually?

   Solution: In order to prepare our mindsets for creating visualizations of the data, we should begin thinking of plots as a mapping from data onto a visual property. In this particular example, we have:

   1. Income $\rightarrow$ horizontal location (x)
   2. Life expectancy $\rightarrow$ vertical location (y)
   3. Time $\rightarrow$ text/plot frame
   4. Population $\rightarrow$ area of circular plotting symbol
   5. Country region $\rightarrow$ color of plotting symbol
   6. Country $\rightarrow$ text label

   Try organizing the students' thoughts on the board with a 2-part diagram Variable –¿ Visual Property.

(b) How do we interpret the visual qualities? In other words, how can we look at the image and know how to interpret the properties of the plot into data?

Solution:

1. Axes with scales are given for GDP per capita and life expectancy.
2. There is background text for each year as the video plays and labels for certain countries of interest.
3. We are told by Rosling how to interpret the color and size of the circles.

(c) Does it look like the raw values of the data were plotted or were they (numerically) transformed before plotting?

Solution:

1. Income has been hit by a log-transformation. The unlogged values are marked on the axes, though.
2. Life expectancy has been centered around the global average (axis does not start at 0).
3. The year progression is slowed/sped up to emphasize certain points in history.
4. Population was scaled so that the radius of the circle is the square root of the population count (hard to tell!).

(d) Is there any information present that is not represented visually?

Solution: Rosling's narration! He gives a selective account of the historical **context** of the data. Remember that each and every plot you present to others should tell a story of some sort.

(e) Write down your thoughts on the granularity, faithfulness, temporality, and scope of this dataset, including questions you would want to ask Rosling about the data.

Solution: It paints the general trend fairly well, but we should expect some issues in scope.

This video often generates quite a bit of discussion, and students are often surprised just how much we have to say about the visualization. This will take 15-20 minutes.

The data are compiled from a variety of sources. Data from high-income countries are mainly derived from registers, whereas surveys are a common source in low and middle-income countries. Such surveys are based on interviews with a representative sample of the population. Data for the 19th century are often based on various types of estimates.

Many countries had different borders or did not exist at all in the past. The data concern the area of the present day borders of the country.

Go to 1:30. Point out the some countries move a lot more in this period than others. This is because those countries' were actually collecting finer data at the time. For every other country, we have to make estimates!

2. Name and sketch some appropriate printed (on paper) 2D visualizations if your goal is to explore:

   (a) The distribution of a single categorical variable (e.g., political party preference of voters).

   > **Solution:** A dotplot or bar plot.

   (b) The distribution of a single continuous variable (e.g., income).

   > **Solution:** If the sample size is manageable, a stripplot that displays all of the data. If the sample size is large, a density plot or a boxplot that summarize the data.

   (c) The relationship between two continuous variables (e.g., income vs. weight).

   > **Solution:** A scatterplot.

   (d) The relationship between a discrete and a continuous variable.

   > **Solution:** Side-by-side boxplots, overlayed densities, side-by-side violin plots.

   (e) The relationship between two continuous variables and two discrete variables (e.g., income vs. weight by race and city).

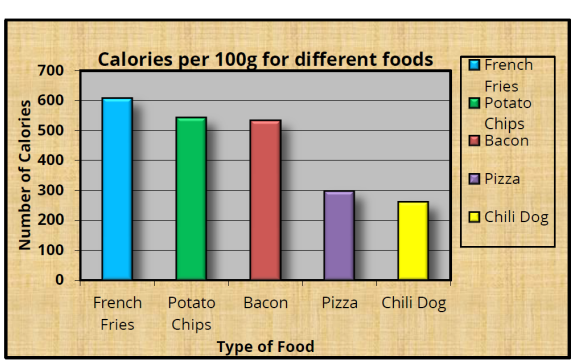
   > **Solution:** If you're investigating income vs. weight and controlling for race and city, grid the race and city and make sub-scatterplots of income vs. weight. You can also use different colors and/or plotting symbols for race and city.

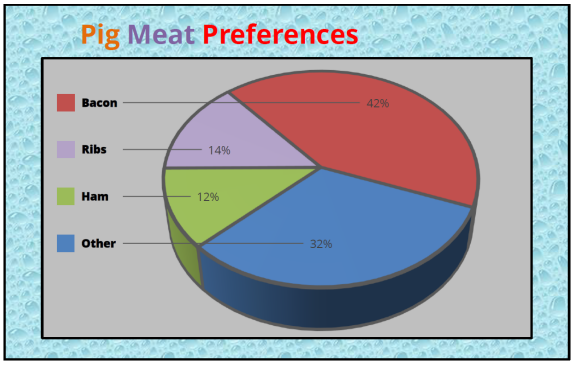
   Suggested: 5 minute group discussion, 5 minutes class discussion.

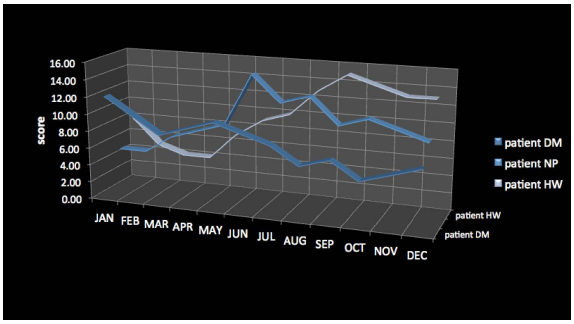
   Emphasize that these are starting guidelines. Sometimes the nature of the data will make some plot types uninformative.

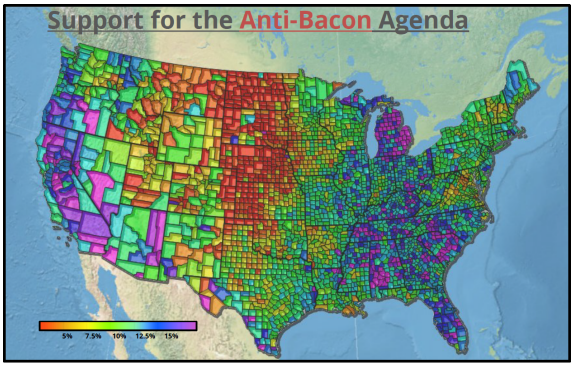
3. Discuss the problems with keeping the visualizations below as they are. Color versions are given in the document found on the course website. You may want to think about:

   - What could the plot be trying to communicate?
   - What visual qualities distract from the message?
   - If there is a comparison between different variables, how easy is it to compare relevant values?









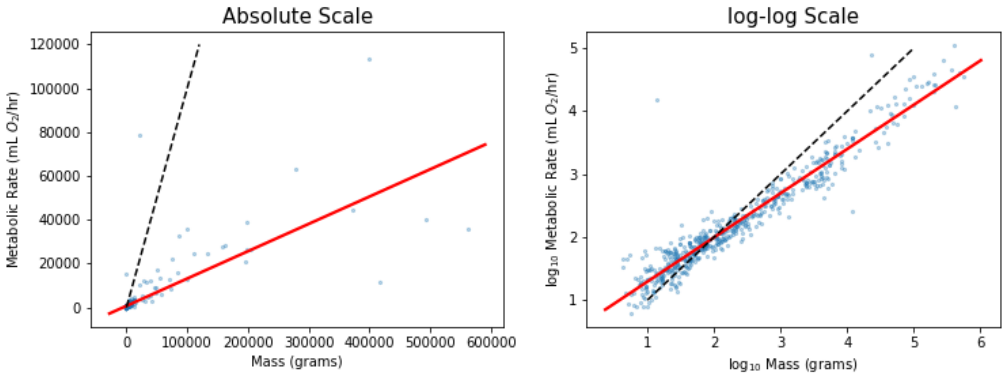
10-15 minutes. You'll have to triage this part a bit. It's more important that you get to the logarithmic transformations. Items to emphasize:

- The goal of every statistical visualization should be communication (sometimes of null results).
- If there are comparisons to be made, they should be immediate at a glance. See Cleveland paper on ordering angles
- Using color
  - Emphasizing important parts of a plot as in the bar and line plots
  - Sequential color scales for ordered values (chloropleth)
  - Divergent color scales for polarized, ordered values with a natural "neutral" value
  - Distinct colors for categorical variables (line plot)

Solution: Refer to the darkhorse pdfs for improvements to the plot.

Logarithmic Transformations

4. One of your friends at a biology lab asks you to help them analyze `panTHERIA`, a database of mammals. They are interested in the relationship between mass, measured in grams, and metabolic rate ("energy expenditure"), measured by oxygen use per hour. Originally, they show you the data on a linear (absolute) scale, shown on the left. You notice that the values on both axes vary over a large range with many data points clustered around the smaller values, so you suggest that they instead plot the data on a log-log scale, shown on the right. The solid red line is a "line of best fit" (we'll formalize this later in the course) while the black dashed line represents the identity line $y = x$.



The aim of this question is to have students think about the meaning of the log transformation rather than just "when to log-transform" the data. We give them the visual cue in the problem setup. For part (a), discuss with the students when we could use the other descriptions.

The data and notebook are available in the disc03 folder.

(a) Let C and k be some constants and x and y represent mass and metabolic rate, respectively. Based on the plots, which of the following best describe the pattern seen in the data?

A. $y = C + kx$ B. $y = C \times 10^{kx}$ C. $y = C + k\log_{10}(x)$ **D.** $y = Cx^k$

Solution: Starting with $y = Cx^k$, we can take the $\log_{10}$ of both sides to find the relationship between $\log_{10}(y)$ and $\log_{10}(x)$.

$$\begin{aligned}\log_{10}(y) &= \log_{10}(Cx^k) \\ &= \log_{10}(C) + \log_{10}(x^k) \\ &= \log_{10}(C) + k\log_{10}(x)\end{aligned}$$

Thus, $\log_{10}(y)$ and $\log_{10}(x)$ are linearly related, which matches what the log-log plot shows above.

(b) What parts of the plots could you use to make initial guesses on C and k?

Solution:

- C: 10^b, where b is the y-intercept of the solid red line in the log-log plot.
- k: slope of the solid red line log-log plot.

(c) Your friend points to the solid line on the log-log plot and says "since this line is going up and to the right, we can say that, in general, the bigger a mammal is, the greater its metabolic rate". Is this a reasonable interpretation of the plot?

Solution: Yes, the observation is equivalent to saying that the slope is positive, which means increases in x correspond to increases in y.

(d) They go on to say "since the slope of this line is less than 1, we see that, in general, mammals with greater mass tend to spend less energy per gram than their smaller counterparts". Is this a reasonable interpretation of the plot?

Solution: Yes, a slope between 0 and 1 means that k is likely between 0 and 1. Looking at $\frac{dy}{dx}$, we see that for these values of k, as x grows, its effect on y diminishes. In this case, it means that gram-for-gram larger mammals spend less energy than their smaller counterparts.

5. When making visualizations, what are some reasons for performing log transformations on the data?

Solution: Comparing orders of magnitude, when the underlying effects seems to be multiplicative and not additive. One heuristic is that "trimming outliers" doesn't seem to be helping the scale of the plot, i.e., new "outliers" appear when you truncate the data.

You have some domain knowledge about the variable, e.g., intensities measured on 16-bit scale.

Possible discussion:

The logarithm linearizes multiplication and scaling. Ink and paper exist in the physical realm, and unless you're traveling near the speed of light, distances act pretty linearly.

Let's say $x = \$1$ and $y = \$1000$. Now imagine that you're going to plot this as-is with a physical scaling of 1cm per dollar. You would need a piece of paper 10 meters wide to just barely fit the two data points on the same plot! On the other hand, if you were to plot $\log_{10} x = 0$ and $\log_{10} y = 3$ instead, we'd only need 3cm of space! Of course the trade-off is that we "distort" the scale.

Another issue is the interpretation of scaling. Income is a fairly canonical example. For a person making $20,000 per year, a $5,000 bonus would probably be a fairly big shock to their lifestyle. For a person making $30,000,000$ (how quickly can you read those 0's?), a $5,000 bonus would probably be insulting. Somehow the logarithm captures a better scale than just naively thinking about the raw values.

An aside: What happened to the units? Technically, when we use transcendental functions for our transformations, we actually want unitless arguments. In the logarithm case, we simply divide the data by a reference $\$1$ before we take the log. In practice, most everyone ignores the subtle point...