**DS 100: Principles and Techniques of Data Science** **Date: February 8, 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.

   (a) How are the variables being represented visually?

   population : size of dot
   income : x-axis
   continent/region : color of dot
   life expectancy : y-axis
   year : background text
   country : dot (label)

   (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?

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

   x-axis : log scale
   y-axis : starts at 25
   sizes of dots not exactly proportional

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

   historical context from narration

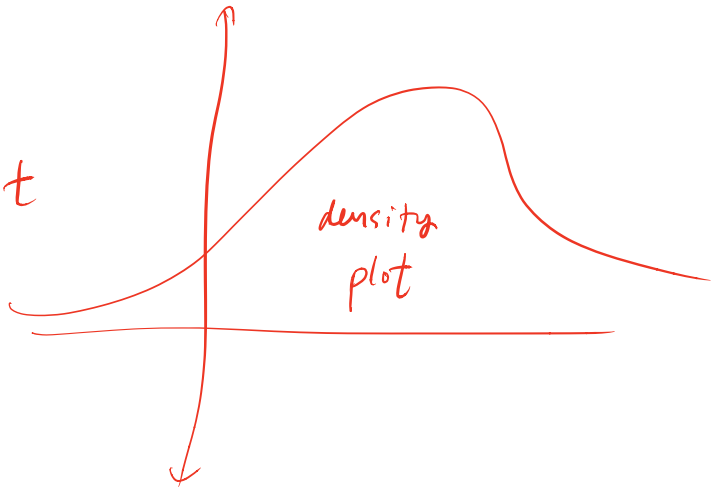
(e) Write down your thoughts on the granularity, faithfulness, temporality, and scope of this dataset, including questions you would want to ask Rosling about the data.

- unreliable data collection in earlier times / 3rd world
- borders change over time
- within each country is a lot of variance

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).
   (b) The distribution of a single continuous variable (e.g., income).
   (c) The relationship between two continuous variables (e.g., income vs. weight).
   (d) The relationship between a discrete and a continuous variable.
   (e) The relationship between two continuous variables and two discrete variables (e.g., income vs. weight by race and city).

a) bar plot

b) histogram, density plot

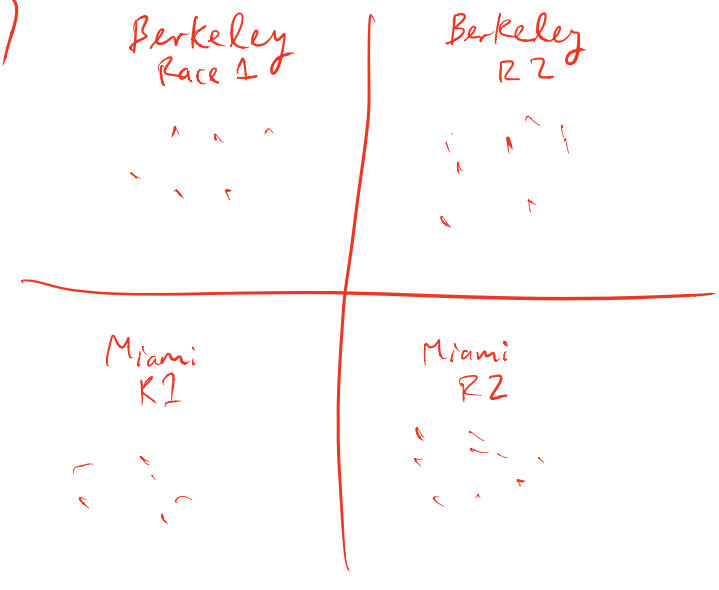


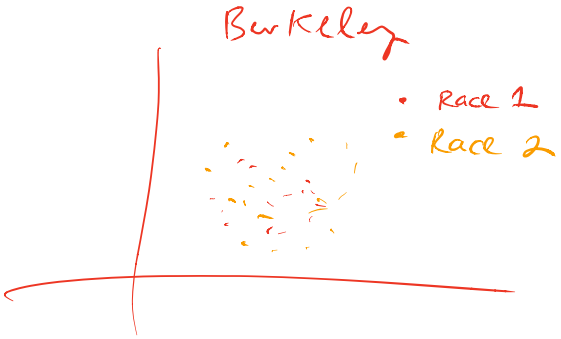

c) scatter plot

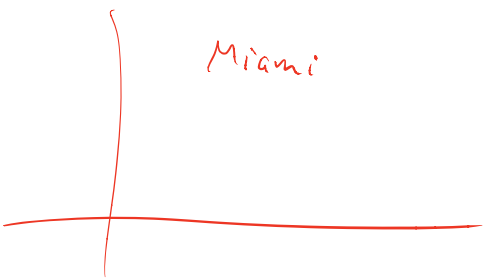
d) side-by-side
boxplots, violin plots

e)

grid of scatter plots

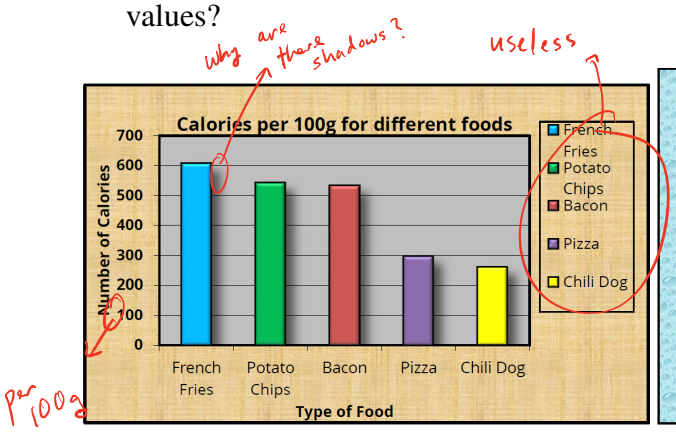


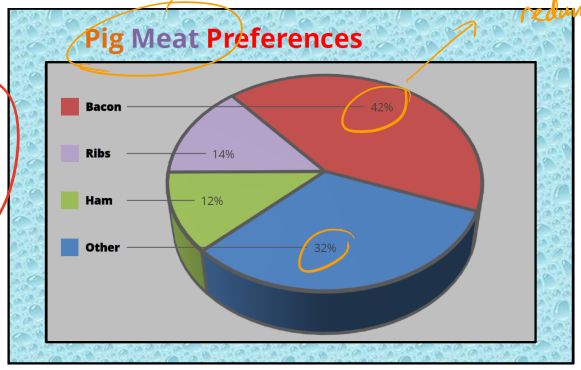



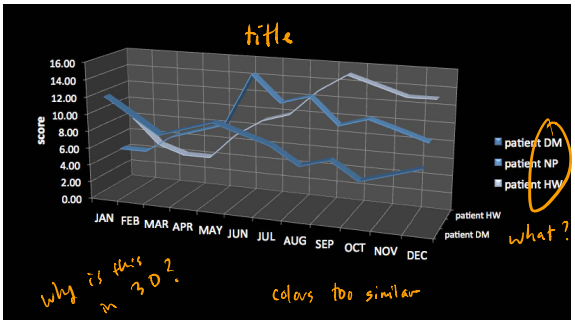



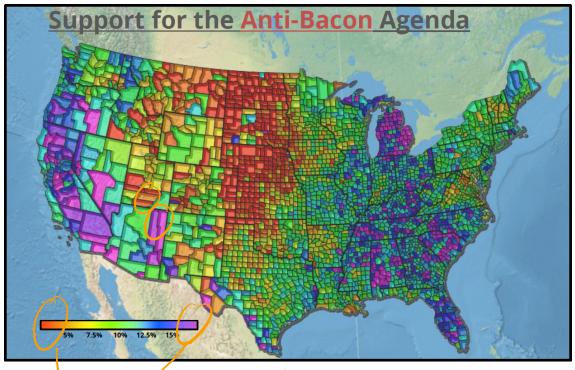
3. Discuss the problems with keeping the visualizations below as they are. Color versions are given in the document found on the course website. You may want to think about:

   - 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?



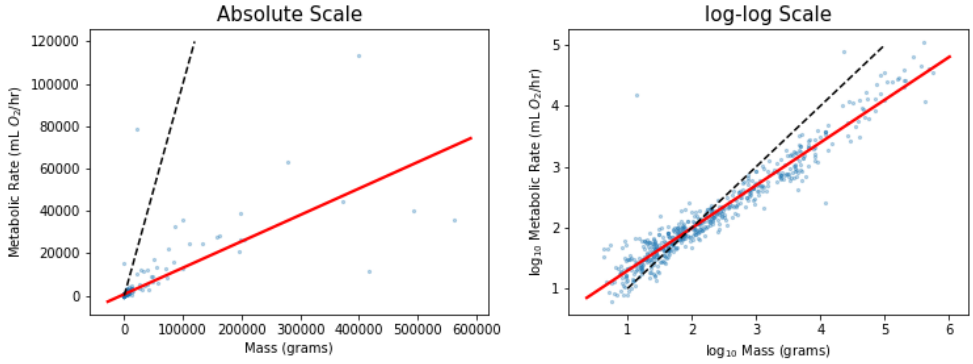



# Logarithmic Transformations

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 $y = x$.




   (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?

   ○ $y = C + kx$ ○ $y = C \times 10^{kx}$ ○ $y = C + k\log_{10}(x)$ ● $y = Cx^k$

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

   (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?

   (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?

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

4.

a) $y = Cx^k$

$\log y = \log(Cx^k)$

$= \log(C) + \log(x^k)$

$\boxed{\log y = \log C + k \log x}$

$\log C$: some constant

b) $y\text{-intercept} = b = \log_{10} C$

$\Rightarrow C = 10^b$

slope $= k$