# Discussion #3

## Data Visualization and Scope

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

$$\texttt{https://tinyurl.com/data100-rosling}$$

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

(a) What variables are being represented in the graphic?

> **Solution:** Income, Life Expectancy, Time, Population, Country region, Country

(b) How are the variables being represented visually?

> **Solution:** In order to prepare our mindsets for programmatically 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 $\rightarrow$ Circle size
>
> 5. Country region $\rightarrow$ circle color
>
> 6. Country $\rightarrow$ Text label

(c) 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 sizes of the circles

(d) 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. We see the absolute values 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!)

(e) 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.

(f) Write down your thoughts on the granularity, faithfulness, temporality, and scope of this data set, 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.

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

(a) The distribution of a single discrete variable (e.g. population of Cities)

> **Solution:** Bar plot

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

> **Solution:** Histogram, density, boxplot, rugplot

(c) The relationship between two continuous variables (e.g. Income vs Weight)

> **Solution:** Density plot, scatter plot

(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.

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?
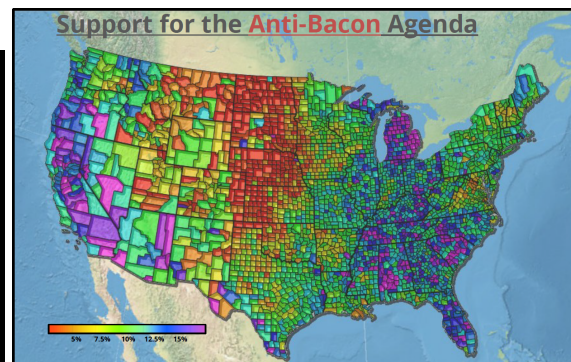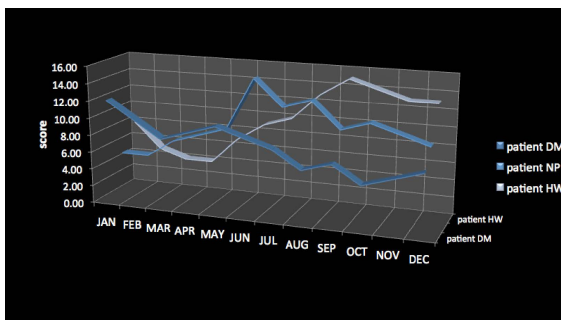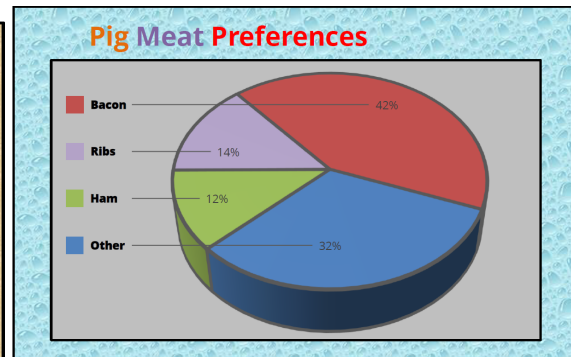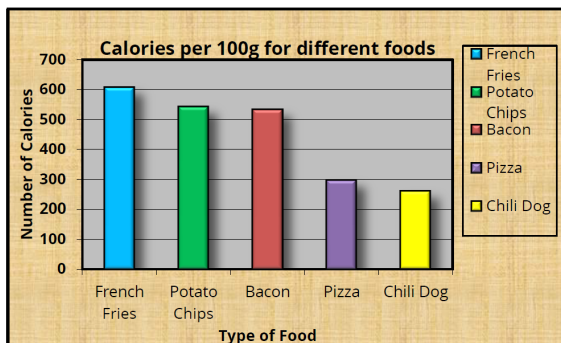




> **Solution:** Refer to the darkhorse pdfs for improvements to the plot.

# 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 segment represents the line $y = x$.
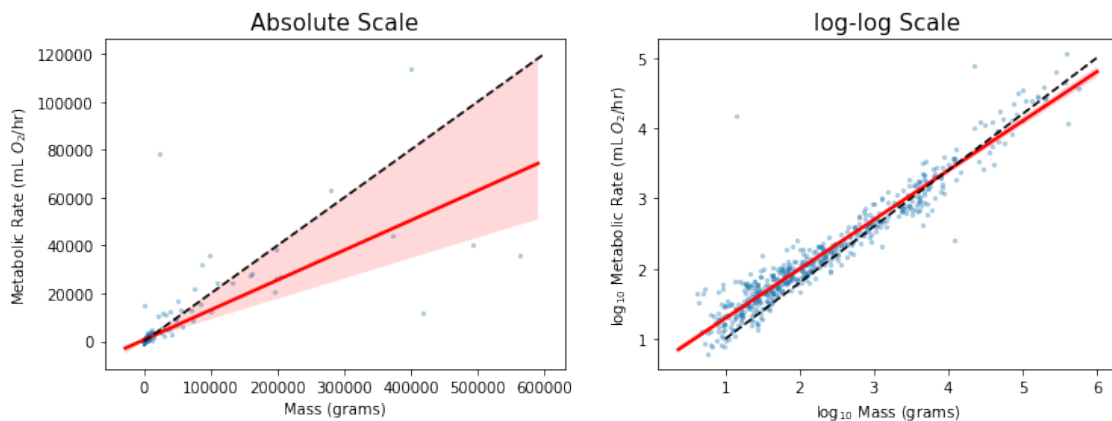


(a) Let $C$ and $k$ be some constants and $x$ and $y$ represent mass and metabolic rate, respectively. Based on the plots, which of the following best describe the pattern seen in the data?

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

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

> **Solution:**
>
> - C: 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.