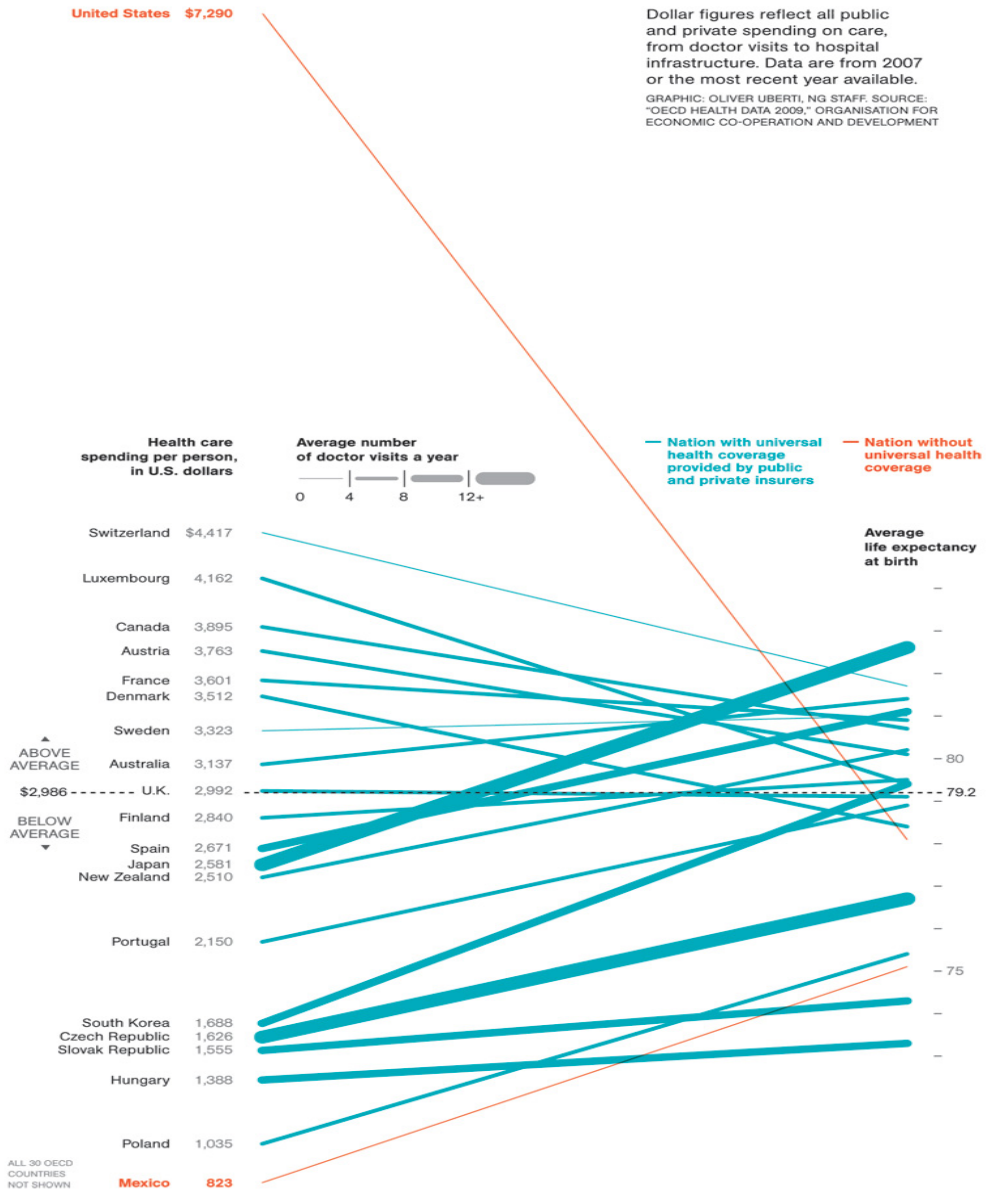
DS 100/200: Principles and Techniques of Data Science **Date: September 25, 2019**

Discussion #5 Solutions

Name:

Data Visualization



1. The first part of the discussion will be centered on the above visualization.

(a) Five variables are being represented visually in this graphic. What are they and what are their types?

Solution:

1. Country - qualitative/nominal (categorical)
2. Healthcare spending per person in USD - quantitative/continuous
3. Average number of doctor visits per year - quantitative/continuous
4. Universal health coverage status - qualitative/nominal
5. Average life expectancy at birth - quantitative/continuous

(b) How are the variables represented in the graphic, e.g., the variable `XXX` is mapped to the x-axis, the variable `WWW` is mapped to the y-axis, the variable `ZZZ` is conveyed through color, etc.?

Solution: This plot is called a parallel coordinate plot. The observations appear as connected line segments between variables. The variables are represented on parallel vertical axes.

1. "Country" appears as labels for the line segments (on the left y-axis)
2. "Healthcare spending per person in USD" is mapped to the y-axis on the left. It increases vertically.
3. "Average number of doctor visits per year" is presented as the thickness of the line connecting "healthcare spending per person in USD" to "average life expectancy at birth". The variable has been discretized.
4. "Universal health coverage status" is presented through the color of the line segment.
5. "Average life expectancy at birth" is presented on the second y-axis (on the right). It increases vertically.

(c) How can we figure out how to interpret the visual qualities of the plot, e.g., how do we know what a color represents?

Solution:

1. The left and right axes provide scales for "Healthcare spending per person in USD" and "Average life expectancy at birth"

2. Legends in the top right and top left regions of the graph provide information on ”Average number of doctor visits per year” and ”Universal health coverage status”

(d) What purpose does the comment at the top right of the plot serve?

Solution: It provides information on the source and temporality of the data. This data was collected between 2007-2009. How much would you expect this visualization to change if we considered similar data from 2017-2019?

(e) Make 3 observations about the figure. Describe the feature that you are basing your observation on.

For example, South Korea's expenditure on health care is comparable to Eastern European countries (and among the lowest of all countries plotted), but the life expectancy is much higher than the Eastern European countries. In the plot we see that the left endpoint of South Korea's line segment is near the Eastern European countries, but the slope of the line segment is much steeper.

Solution: There are many observations that can be made. Here are a few examples:

1. ”Healthcare expenditure per person” in the US is *much* higher than any other nation, but the average life expectancy at birth is below average. This is seen by the left endpoint being far above the others, with a huge gap between the US and the next highest country (Switzerland). And the steep decline of the line segment to a below average life expectancy.

2. The line segment representing the US is very thin, which represents very few (0 to 4) doctor's visits a year.

3. Japan has a very think line segment, which indicates a high average number of doctor's visits. It is interesting to note that Japan has the highest life expectancy, yet the healthcare expenditures are in the middle of the pack AND it has a high number of average doctor's visits.

4. Countries with below average ”healthcare expenditure per person” but with above average ”average number of doctor visits per year” tend to have higher ”average life expectancy at birth” than countries with similar levels of ”healthcare expenditure per person” but whose citizens visit their doctors less often.

5. All countries listed, with the exception of the US, have universal health coverage. NB: Mexico achieved universal health coverage in 2012. This information is presented by the ”universal health coverage status” variable.

(f) Consider the steep negative slope and narrowness of the line segment that represents the

data for the United States. What systemic, social, or societal issues might explain this?

> **Solution:** Access to health care. E.g. what if those who are uninsured were plotted separately, how might these two line segments be different?
>
> What if we created separate lines by race. How might they look? Why might it be misleading to draw conclusions about health based on race? e.g., are there any confounding factors that could be problematic?

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

   (a) The distribution of political party preference of voters.

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

   (b) The distribution of income.

   > **Solution:** If the sample size is manageable, a rug plot (or stripplot) that displays all of the data. If the sample size is large, a density plot or a boxplot that visualizes numeric summaries of the data. A log transformation is most likely appropriate for these data.

   (c) The relationship between income and height.

   > **Solution:** A scatterplot. with the income axis on a log scale.

   (d) The relationship between income and birth sex.

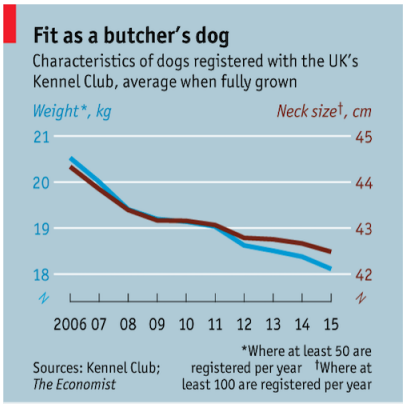
   > **Solution:** Side-by-side boxplots, overlayed densities, side-by-side violin plots. Overlayed histograms are not a good idea because it's hard to distinguish one group from the other.

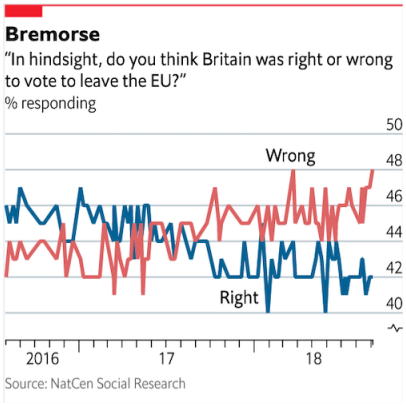
   (e) The relationship between income, height, race and birth sex.

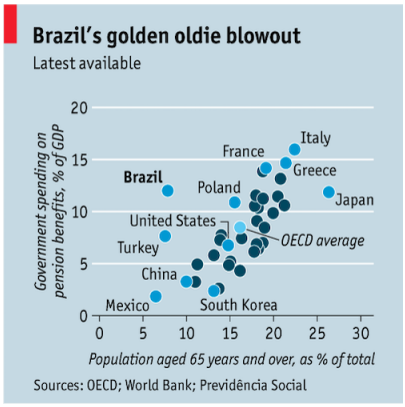
   > **Solution:** If you're investigating income vs. weight and controlling for race and birth sex, you can grid the race and birth sex combinations and make sub-scatterplots of income vs. weight. Or, you can also use different colors for birth sex and grid on race. Or, you can use different colors for race and different plotting symbols for birth sex.

3. Creating visualizations that represent data accurately and that support the narrative we wish to create is no easy task. Even the journalists and editors at *The Economist*, a newspaper known for it's compelling, data-driven articles, have been known to make blunders. Three of

their ill-thought-out plots are presented below. Consider what aspects of the visualizations are misleading, and think of ways in which you can remedy them.







Hint: The datapoints in the rightmost plot are shaded based on whether or not they are labeled.

Solution:

1. Fit as a butcher's dog: Though the scales on either side of the y-axis have the same increments (e.g. each tickmark corresponds to a change in one unit), they are not proportional. A change in weight from 20kg to 18kg corresponds to a 10% decrease in weight, but a decrease in neck size from 44cm to 42 cm roughly corresponds to a 4.5% decrease. The scales should be normalized to ensure a fair comparison.

2. Bremorse: The plot is attempting to display the trend of Brexit remorse over time by plotting the exact survey results at each time-point. Smoothed lines would reduce the amount of visual noise and better support the claim that Brexit remorse is increasing.

3. Brazil's golden oldie blowout: There a two main issues with this plot:

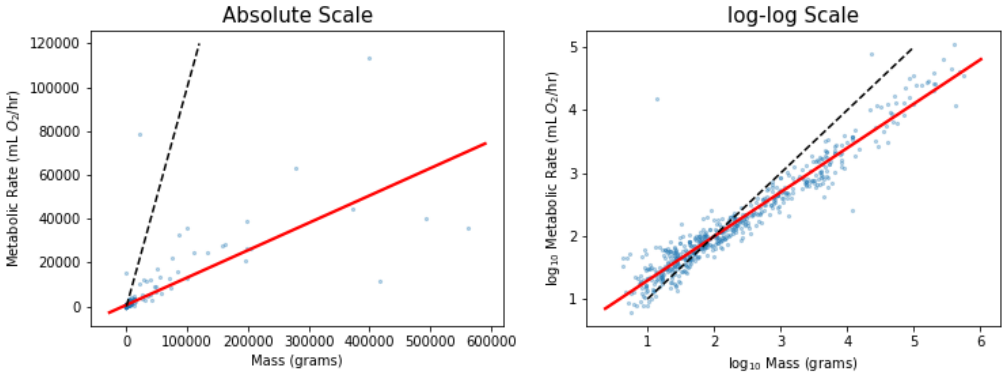
   (a) The shade of blue used to color datapoints serves only to indicate whether the datapoint is labeled. Colors should be used to illustrate categorical variables.

   (b) The inclusion a large number of labels contributes to the visual noise. Removing all but Brazil's label would make for a cleaner plot, all while supporting the narrative. Keep it simple!

Follow this link for more information on these plots, potential fixes, and other examples.

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?

◯ 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.