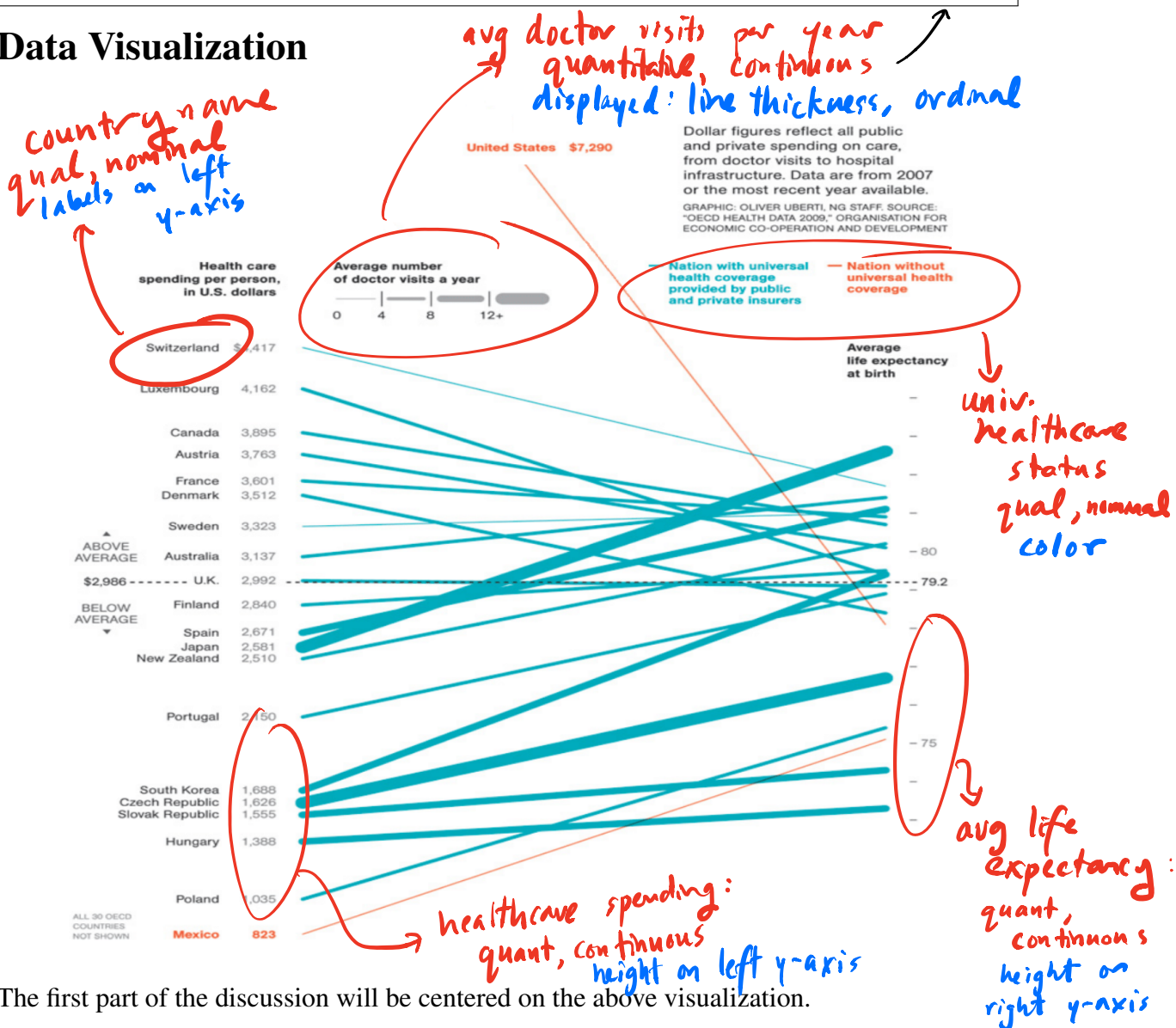
DS 100/200: Principles and Techniques of Data Science **Date: Feb 21, 2020**

Discussion #5

continuous because it's an average

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?

In red

(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.?

In blue

(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?

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

Context for timeframe

(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.

See solution!

(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?

See solution!

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

   (a) The distribution of population for various cities.

   (b) The distribution of income.

   See supplementary notebook

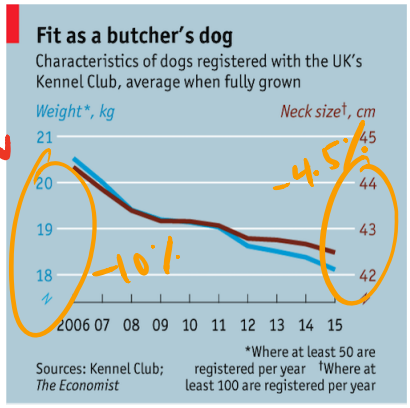
   (c) The relationship between income and life expectancy.

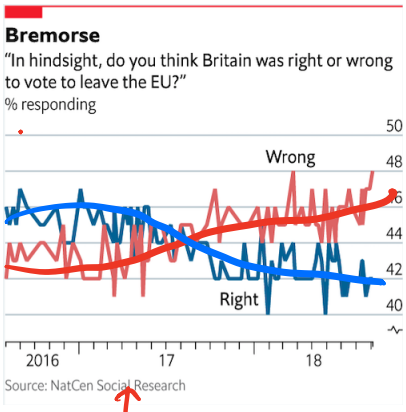
   (d) The relationship between income and city.

   (e) The relationship between income, life expectancy, smoking status(non-smoker, social smoker, smoker), and city

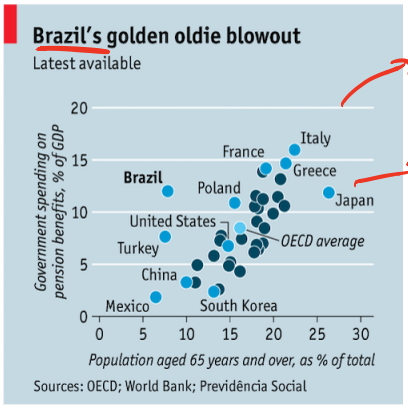
3. Creating visualizations that represent data accurately and that support the narrative we wish to create is no easy task. Even the journalists and editors at *The Economist*, a newspaper known for it's compelling, data-driven articles, have been known to make blunders. Three of their ill-thought-out plots are presented below. Consider what aspects of the visualizations are misleading, and think of ways in which you can remedy them.

should be rescaled so that %. differences are the same





should be smoother!



remove labels

remove colors (should be for categories)

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

Qualitative

Ordinal
categorical variables with an inherent ordering (e.g. class standing, movie ratings)

Nominal
categorical variables with no ordering (names, colors)

Quantitative
numerical data that we can perform operations on

Discrete
things that you count "how many" (family size, "number of ___")

Continuous
things you measure (weights, distances, etc.)
→ averages