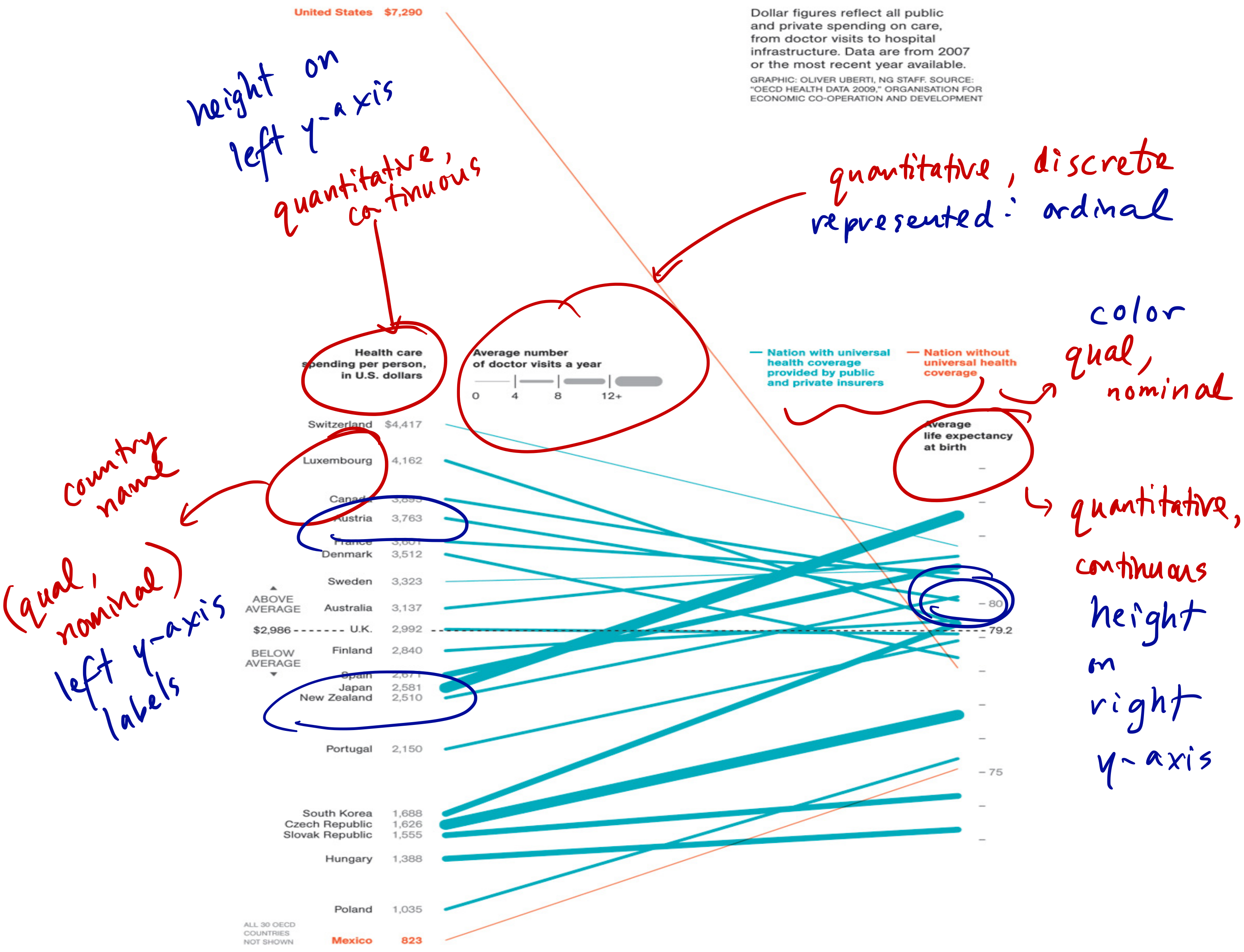
DS 100/200: Principles and Techniques of Data Science **Date: September 25, 2019**

Discussion #5

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 on previous page

(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 on previous page

(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 – e.g. Mexico got universal HC in 2012, but data is from 2007

(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 solutions!

(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?

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.

   (b) The distribution of income.

   at the end

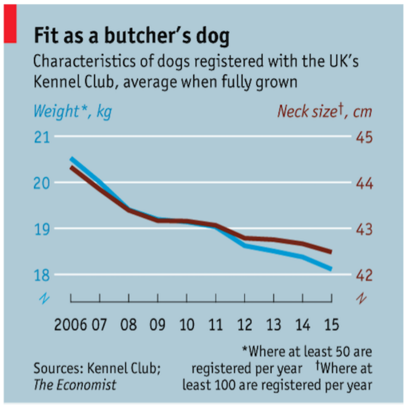
   (c) The relationship between income and height.

   (d) The relationship between income and birth sex.

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

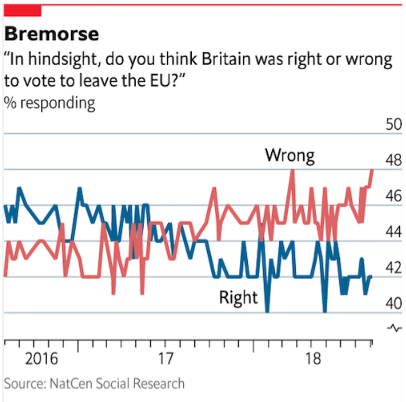


Fit as a butcher's dog

Characteristics of dogs registered with the UK's Kennel Club, average when fully grown

Sources: Kennel Club; *The Economist*

*Where at least 50 are registered per year †Where at least 100 are registered per year

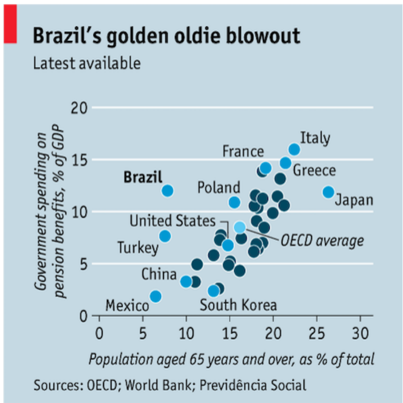


Bremorse

"In hindsight, do you think Britain was right or wrong to vote to leave the EU?"

% responding

Source: NatCen Social Research



Brazil's golden oldie blowout

Latest available

Sources: OECD; World Bank; Previdência Social

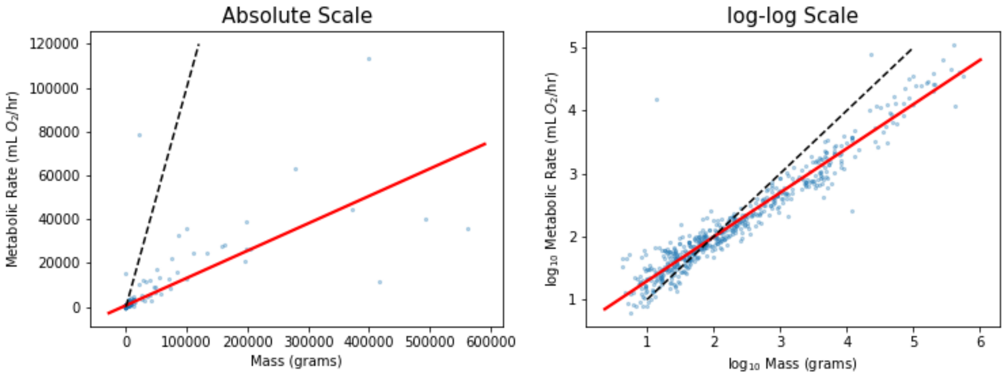
Hint: The datapoints in the rightmost plot are shaded based on whether or not they are labeled.

see solutions!

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 $y = x$.

$\log(y) = m \log(x) + b$



(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. $y = Cx^k$

→ only option that reduces to above relation

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

$C = 10^{\text{intercept of line}}$, k = slope of line

(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?

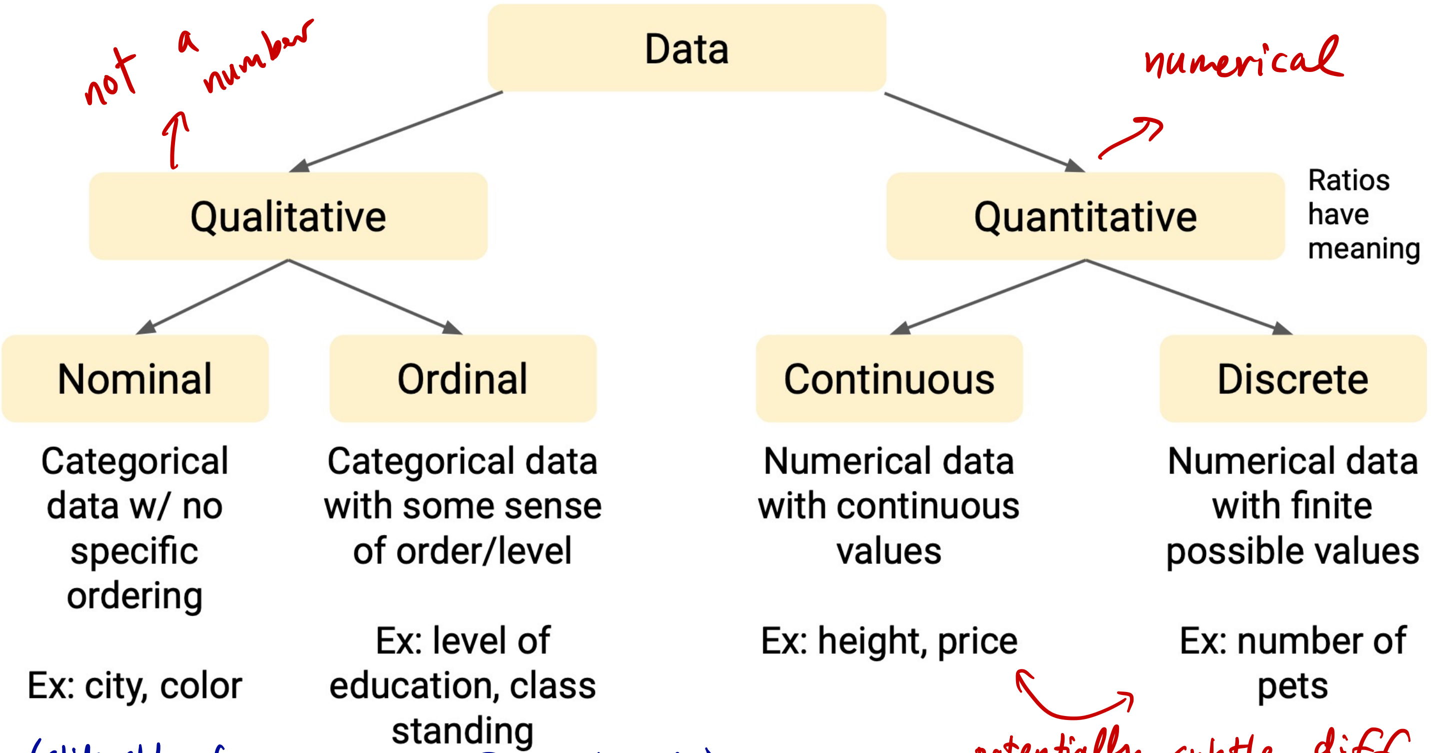
- it's easier to visually interpret linear relations
- future context: much easier to model things that are linear!

$$y = Cx^k$$

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

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

$$\log(y) = \underbrace{\log C}_{b} + \underbrace{k}_{m} \log(x)$$



Data

Qualitative

not a number

Quantitative

numerical

Ratios have meaning

Nominal

Categorical data w/ no specific ordering

Ex: city, color

Ordinal

Categorical data with some sense of order/level

Ex: level of education, class standing

Continuous

Numerical data with continuous values

Ex: height, price

Discrete

Numerical data with finite possible values

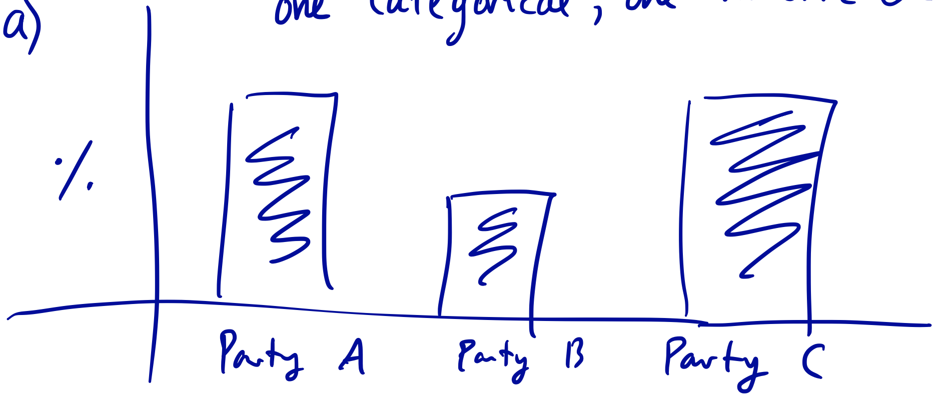
Ex: number of pets

potentially subtle diff.

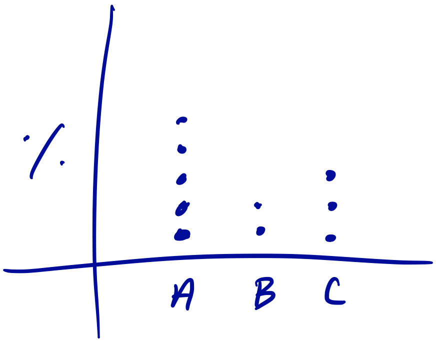
(slide stolen from emansishah@berkeley.edu)

Problem 2

a) one categorical, one numerical

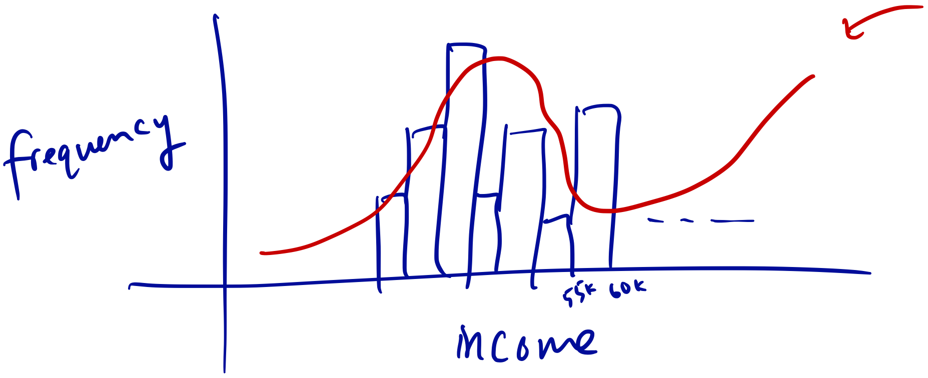


bar plot



dot plot

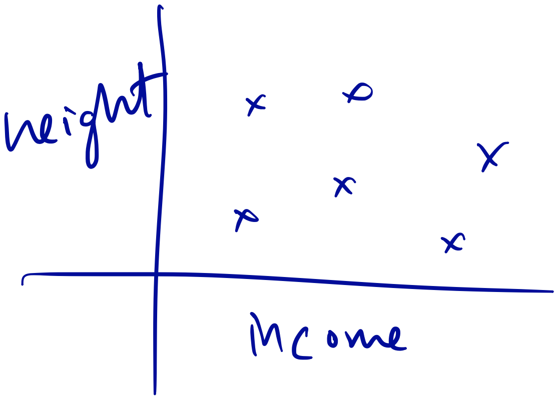
b) one continuous numerical value



KDE
density plot

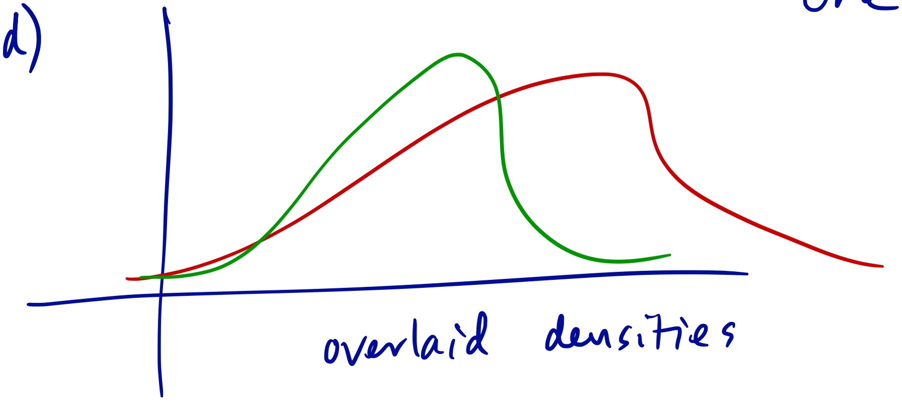
- density plot
- histogram
- rug plot

c) two continuous numerical → scatter plot!

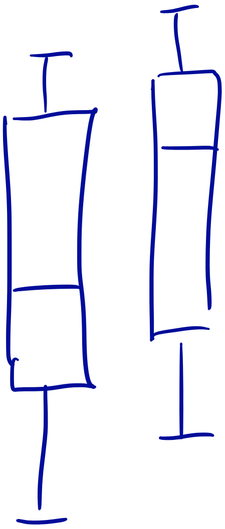


d)

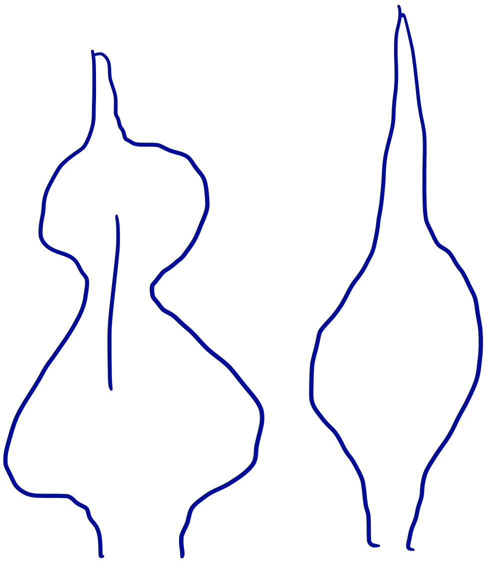
one categorical, one continuous



overlaid densities

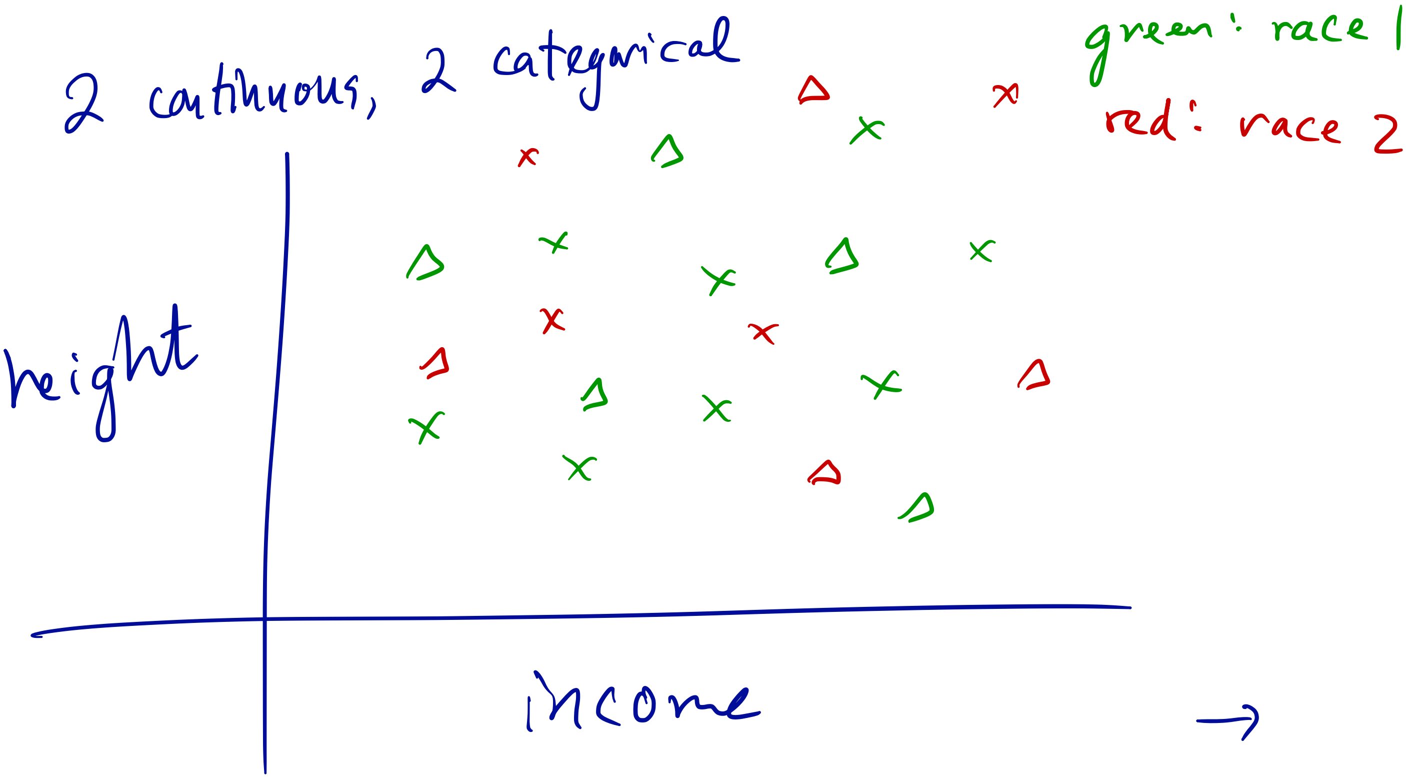


side-by-side boxplots

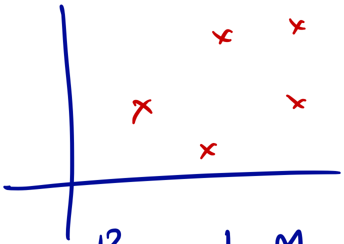


side-by-side violin plots

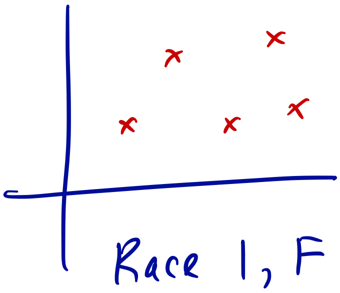
e) 2 continuous, 2 categorical



→

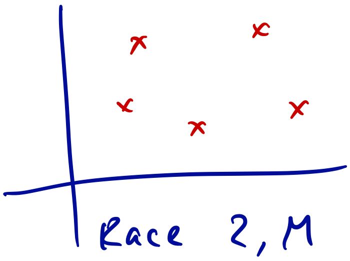


Race 1, M

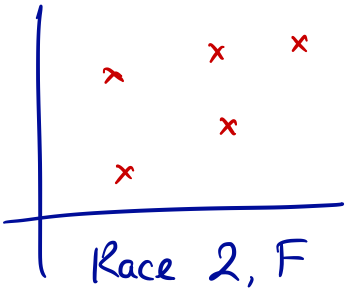


Race 1, F

another option for ⓔ



Race 2, M



Race 2, F

⋮