DS2500
1/21 - Tues

## Admin

- Hw1 due Fri 1/24 9pm
- Hw2 out Fri 1/24
- Lab2 next Monday 1/27
- Mini Pressed: bit.ly/ds2500_mini_prez
  - in teams
  - submit 2 slides 2/23 9pm
  - present 2/24, 2/25
  - 5 min presentation

## Agenda

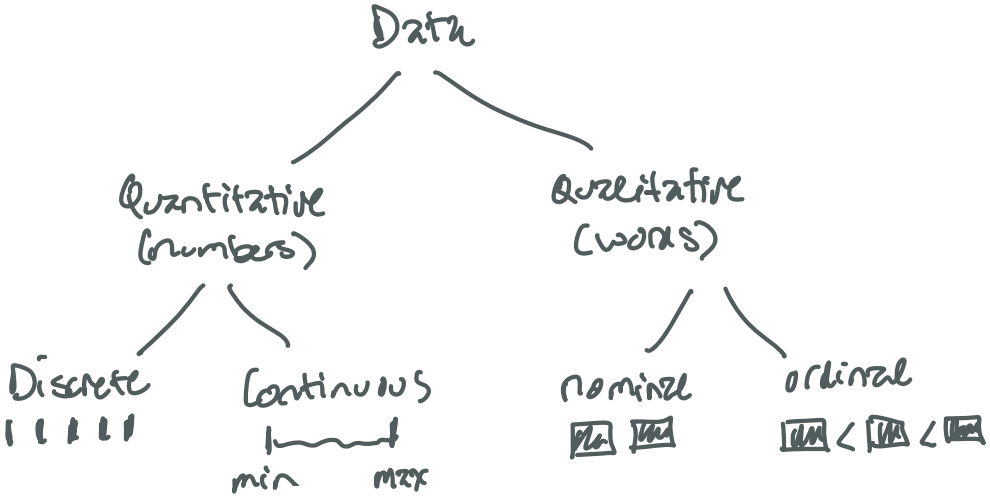
1. Types of Data
2. Asking & Answering Questions
3. Python

} exploring datasets
comparisons + similarity

## 1. Types of Data

↳ starting w/ a dataset, good **first step** is to categorize!

**Second step** - what questions can we ask/answer about the data?




(ex) weather data ~ Qualitative

- nominal: rainy, snowy, cloudy, clear, humid (no comparison)
- ordinal: chilly, cold, freezing, frigid (compare!)

nominal - no comparison  rainy < snowy? !! ✗

ordinal - comparison, no similarity → chilly < cold ✓, cold < freezing ✓

↳ [redacted] ? [redacted] ? [redacted]
chilly  cold  freezing

chilly - cold == cold - freezing?
✗ !!

(ex) weather data - quantitative ~ snowfall

discrete - comparison, similarity
- 1 inch of snow
- 2 inches of snow
- 3 inches of snow

1 < 2? ✓  1 == 1? ✓
1 → 2 == 2 → 3? ✓
1 → 2 < 1 → 3? ✓

continuous - comparison*, similarity
- 1.2 inches of snow
- 1.7831 inches of snow

1.2 < 1.7831 ✓
1.2 == 1.2 ✗ !!

min ——— max

1.2 → 1.3 < 1.3 → 1.5 ✓

similarity ⇝ meaningful distances
- quantitative
- qualitative* (sometimes?)

## 2. Asking + Answering Questions

↳ * if we have qualitative data, do we convert to quant. so we can measure similarity?

(ex) 1.4  1.8  1.35 → 140  180  135
== ✗ !!  == ✓ (lost precision)

This week's data/goal
↳ Find Lacey a place to eat lunch now that B. Good was shut down

- goal: find the place that is most similar to B. Good

What info can we have about restaurants? (attributes!)

- type of food – nominal
- ~~accessibility~~
- price – avg price entrée (cont), scale (disc)
- distance from Lacey's office – miles (cont), scale (disc)
- rating – # stars (disc)
- wait time – avg minutes (cont), scale (disc)
- hours – nominize, # hours (quant.)
- style of food – nominal
- vegan options, gf, etc. – nominal, or "vegan" 0/1
- name – nominal
- capacity – discrete
- menu – nominal
- dress code – ordinal

good for similarity

what about vegan – 0/1 → what type of data?

0/1 can be quantitative if they are numeric

0/1 are actually nominal if they rep categories yes/no

* can be used as 0/1 for similarity, but not with quantitative data

0/1

not quantitative!

raw in csv: 1.5, 4, 8, 15.5, 0

–1, –1, –1, –1, 1

–1, –1, –1, –1, 0

# 3. Python

## Data Science Stats

↳ statistics library (import, but no need to install)

Specific to Quantitative Data (mostly)

- range (min, max)
- mean (avg) ~ population with $N$ observations

$$\bar{x} = \frac{1}{N} \sum_i x_i$$

$\sum_i$ sum together all $x_i$ observations

- median – half population < median
  half population > median

  statistics.median

(ex) [3, 5, 1] median = 3 → [1, 3, 5]

(ex) [8, 6, 5, 10] median = 7 → [5, 6, 8, 10] → 7 (avg)

median when population has outliers

- mode – most common item in population

  statistics.mode

↳ good for: qualitative, discrete → [frigid, cold, frigid]

[not for continuous!] [1.3, 1.301, 3.8] XXX !!

[3, 5, 3, 3] 3

[1, 2, 3, 4, 5] no mode

bit.ly/food-recs

↳ What do we do in case of dupes?

How do we identify a dupe?