cs199 Outline/Take-aways/Project Guidance

1. Cultivate skepticism (car dealership mindset)
   1. Beware of confirmation bias
      1. The tendency to focus on data that supports your current belief.
   2. Entertain multiple hypothesis
      1. Think of many ways to explain the data/phenomenon, and try not to get attached to one in particular. Occam’s razor: of the set of hypotheses that best explain the data, bias toward the simplest one.
   3. Think about orders of magnitude
      1. Watch for huge/tiny numbers that may be blown out of proportion (too good/bad to be true). Remember that huge/tiny numbers are non-intuitive to reason about. Think, “compared to what?”
   4. Beware of unfair comparisons
      1. Remember that these can be subtle!
2. In data graphics, watch out for:
   1. Y-axis extension/truncation
      1. Generally, bar graphs should include 0 on y-axis; line graphs might not, but shrinking/extending can make changes look bigger/smaller.
   2. Manipulating bin sizes
      1. Default should be equal bin sizes (maybe with exceptions for larger/smaller than X) or you can tell any story you want!
   3. Ducks
      1. When style overwhelms substance – flashiness over information. Beware of ‘impressive’ looking graphs just for the sake of ‘impressiveness.’
   4. Glass slippers
      1. E.g. periodic tables (of anything other than elements), subway maps (not of subways), Venn diagrams that make no sense. Fitting the wrong type of information into a form to look ‘scientific.’
   5. Principle of Proportional Ink
      1. E.g. 3-d pie charts. The area of the shaded region should be directly proportional to what it represents numerically.
3. Statistical traps
   1. Right censoring (Musicians & Mortality case study)
      1. In data measuring lengths of something, watch for comparing samples where the data is being cut off before the entire length could be measured. (e.g. lifespan)
   2. Misuse of averages
      1. Pay attention to use of mean/median/mode – know the differences.
      2. Watch for outliers and/or skew affecting the mean value.
      3. Pay attention to the data type – does reporting a mean make sense? (e.g. no, for color, no for number of Y chromosomes)
   3. Inclusion of outliers (Senescence case study: world record holders)
      1. Not representative of average performance
   4. Trajectories of different individuals (Senescence case study)
      1. For claims about time series, all points in the time series should measure the same thing (which is consistent with the claim). E.g. not measurements of the thing in different people.
   5. Different sample sizes for trajectory points (Senescence case study)
      1. Beware of unequal sample sizes when comparing groups. The larger group may have more extreme values just by chance.
   6. Will Rodgers effect
      1. E.g. cancer screening and treatment. As better diagnostics move the hardest easy cases into the easiest hard cases, treatment efficacy can look like it changes even though it doesn’t.
   7. Simpson’s Paradox
      1. When aggregating averages, watch for unequal sample sizes in the points you’re aggregating AND watch for unmeasured variables that could affect the distribution of sample sizes and/or the outcome.
   8. Correlation and Causation
      1. Correlation is NOT the same as causation!
         1. Correlation: one variable gives information about another
         2. Causation: one variable influences another
      2. Potential pitfalls
         1. Is there a common cause? Could make two variables look correlated without a direct link between them.
         2. Spurious correlations: correlated just by chance. No relationship.
         3. Just because B happened after A doesn’t mean that A causes B
      3. Even if we DO have causation (e.g. from manipulative experiments)
         1. Do we have the most important cause?
         2. Could there be feedback loops?
   9. P-values and prosecutor’s fallacy
      1. Null hypothesis: a simple explanation of the data – how the data would be distributed if nothing is going on
      2. Significance level/threshold (‘alpha’): chance of incorrectly rejecting the null hypothesis
      3. P-value: the chance of seeing data at least this extreme under the null hypothesis
      4. Prosecutor’s fallacy: interpreting the p-value as telling us the truth value of the null hypothesis. This is incorrect!
   10. Big data
       1. Overfitting: taking the data ‘too seriously.’ Fitting not just the signal in the data, but also the noise, which will not translate well and generalize to new scenarios.
       2. Garbage in, garbage out: even if the modeling method is a black box, be skeptical of the data going in and conclusions coming out. If the input data is bad and/or biased, the conclusions may be too.
4. The scientific process IS CUMULATIVE
   1. Replication crisis: many scientific studies cannot be reproduced in other labs/settings
   2. Publication bias: authors and publishers preferentially report POSITIVE results (including false positives)
5. Checking your sources: don’t believe everything you read/hear
   1. Predatory journals: in some venues, you can basically pay to get your article published with little to no peer review. This does not produce good science!
   2. Fake news: it’s easy and cheap to produce fake articles as clickbait – beware.