POL 212
Winter 2024
Assignment 3

1. Load the ANES 2022 Pilot Study data in R (anes_pilot_2022_stata_20221214.dta). It will also be helpful to review the questionnaire and/or user guide to get a sense of which variables are included and how they are coded.

2. Select one of the feeling thermometers (all begin with "ft"; e.g., "ftjourn" asks about journalists; "ftbiden" asks about President Biden).

    a. For that feeling thermometer, produce three histographs in ggplot with different bin widths. NOTE: 0 (coldest) to 100 (warmest) are valid values for feeling thermometer ratings. MAKE SURE TO DEAL WITH MISSING/NA VALUES (i.e., they should be coded as NAs)!

3. Select another variable that you think might be related to the feeling thermometer you selected in (2). This could be a policy issue attitude, demographic characteristic, or whatever. It should, however, include at least five valid response categories (for instance, "libcon" [ideology] includes seven valid categories). MAKE SURE TO DEAL WITH MISSING/NA VALUES (i.e., they should be coded as NAs)!

    a. Using the feeling thermometer you selected in (2) and the second variable you just selected, create three scatterplots in ggplot. The three scatterplots should also include loess smoother lines (geom_smooth) with different values for span (which controls the bandwidth/amount of smoothing).

CHALLENGE QUESTION: ANSWER C1 or C2

C1.) Adapted from Regression and Other Stories, exercise 5.4. Demonstration of the Central Limit Theorem: Let $X = x_1 + ... x_{25}$, the sum of 25 independent uniform(0, 2) random variables. In R, create 100 simulations of X and plot their histogram.

(a) What is the normal approximation to this distribution provided by the Central Limit Theorem?
(b) Overlay a graph of the normal density on top of the histogram.

(c) Comment on any differences between the histogram and the curve.
(d) Repeat parts 2(a)–2(c) above, but using 5000 simulations. Compare the two sets of results (those obtained using 100 simulations and 5000 simulations).

C2.) In a 1–2 page essay, explain the bias-variance tradeoff. You may use any combination of theoretical discussion, math, illustrated diagrams, hypothetical examples, etc. How does the bias-variance tradeoff relate to “underfitting” and “overfitting” statistical models, and how does it relate to the results you obtained for question 3(b) above?