Assignment 1: Data Visualization

Insert Name Here PSC 8101, Assignment 1
Due Friday, Sept. 12, 2025

Instructions

Answer the questions within this document. Make sure to answer the questions in sufficient detail. Keep the questions included below in your final document. Because the "format" in this document's YAML header is set to "pdf," the document you render and turn in will be a .pdf.

- **Workflow alert**: Before working with your .qmd file (Quarto), put all of your graphing commands in the "Assignment 1.R" script file. Use your script file to make sure all your graphs execute correctly first. Then you can copy and paste those commands over to your .qmd file and include your interpretations.
- **In your rendered document (pdf),** include your R code, output/graphs, and your explanations (in addition to the questions asked). Do not include any R warnings or errors in the output (this is already specified in the YAML under "execute").
- For all graphs, *make sure to label your axes with descriptive names instead of the raw variable names*. Provide descriptive titles for your graphs, too.
- Submit your rendered .pdf in Blackboard.

1. Histograms, Boxplots, Density Plots for Continuous Variables

Use the "**vdem**" data to examine the distribution of country wealth. This is the Varieties of Democracy data that is based on country experts' assessments of various political factors. For this problem, you'll analyze an empirical indicator of country GDP for the year 2019 (from an article by Farris et al. 2021, *Journal of Conflict Resolution*). The observations/units of analysis are countries around the world; N=179. You'll examine a measure of GDP per capita, "**e_gdppc**", which is measured in thousands of dollars.

A. Generate a histogram and a density plot of GDP. For the histogram, (1) experiment with different numbers of "bins" and explain why you chose the number you did and (2) use either "count" or "percent" on the vertical axis but justify your choice. For the density plot, use a "fill" color that is transparent (specify using "alpha"). Interpret what you visualize.

B. Generate a boxplot of GDP. Include the dots for the underlying data. Interpret what you visualize.

C. Generate a density plot of GDP across different regime types using a "facet wrap." Use the "**v2x_regime**" variable: 0=closed autocracy, 1=electoral autocracy, 2=electoral democracy, 3=liberal democracy. Interpret what you visualize. In addition, use a "ridge plot." Which do you think provides a better visualization of the results?

D. Now generate a boxplot by regime type. Include the dots for the underlying data. Interpret what you visualize.

E. Subsetting the data like we do in parts C and D implies an independent variable and a dependent variable. Identify each and briefly discuss the nature of the relationship.

2. Barplots for Discrete Variables

Use the "**nes**" data to analyze the distribution of party identification across American voters. The data come from the 2020 American National Election Study (ANES); N=8,280. Party identification (pid) is measured by asking voters two questions: (1) Are you a Democrat, Independent, or Republican? (2) For Dems and Reps, are you are "strongly" Dem/Rep or not strongly Dem/Rep? For Independents, do you lean toward being a Democrat or Republican (or neither)? You'll analyze two variables:

- **V201228**: Voters' answer to the first question in the pid measure. 1=Dem, 2=Rep, 3=Ind.
- **V201231x**: A 7-category variable based on voters' answers to both questions 1 and 2 in the pid measure. 1=strong Dem, 2=not strong Dem, 3=independent leaning Dem, 4=independent (no lean), 5=independent leaning Rep, 6=not strong Rep, 7=strong Rep.

A. Generate a barplot of the 3-category measure. Report percentages on your y-axis. Interpret what you visualize.

B. Generate a barplot of the 7-category measure. Report percentages on your y-axis. Interpret what you visualize. What in particular would you conclude about independents from comparing this graph to the graph from part A?

3. Bivariate Scatterplots

Open the "**states**" data. The observations/units of analysis are the 50 states. Examine the influence of **education** on **voter turnout** in the 2016 election. In the states data, use these variables:

- **ba_or_more_2015**: percentage of population in each state with a BA degree or higher (x-axis).
- **vep16_turnout**: percentage of the population in each state who voted (y-axis).

A. Which variable is the independent variable? Which is the dependent variable?

B. Generate a scatterplot showing the effect of education on voter turnout. Interpret what you visualize from this graph.

C. Now generate another graph that adds a "line of best fit" to your scatterplot from part B (remove the confidence interval). Interpret what you visualize from this graph.

D. Finally, generate the same graph from part C and add state labels using the "StateID" variable (two-letter abbreviation). Use ggrepel so they don't cluttered. Identify any outliers.

4. Bivariate Linegraphs

Use the "**scotus**" data to examine Supreme Court liberalism/conservatism over time, from 1947-2022. Here are the variable descriptions:

- **year**: year, from 1947-2022
- **sclib**: percentage of **all** Supreme Court decisions in each term that were decided in the liberal direction. Higher values (above 50%) = more liberal; lower values (below 50%) = more conservative.
- **sclib_sal**: percentage of **highly salient** Supreme Court decisions in each term that were decided in the liberal direction. Salient decisions are those that received prominent media coverage (front page of the *New York Times*). Higher values (above 50%) = more liberal; lower values (below 50%) = more conservative.

A. Generate a linegraph showing Supreme Court liberalism for all cases (**sclib**). Make sure your y-axis ranges from 0 to 100. Interpret what you visualize.

B. Generate a linegraph showing Supreme Court liberalism for salient cases (**sclib_sal**). Again, make sure your y-axis ranges from 0 to 100. Interpret what you visualize.

C. Now generate two additional graphs, and for each graph, add a "smoothed" line to your linegraph that allows you to visualize time trends more clearly. For each graph, interpret what you visualize.