Analysis Challenge 1 | Due: Mon. Feb. 5th by Midnight

The Challenge

In this first analysis challenge, imagine that you are an analyst for either the RNC or DNC (it's up to you which one you want to work for).

The party has collected data on the amount of money they've spent on various U.S. House elections from 2018, along with some details about those districts and the candidates that ran for office. They want you to **look at districts where their candidates *won* to assess whether *party spending* did anything to help**.

Write up a short research note (no more than 700 words) where you: (1) explain to party leaders why the analysis they want you to perform is a bad idea and (2) show them using the data they've given you why only looking at districts where they won will be particularly misleading for looking at the link between party spending and margin of victory for their candidates.

For (1), discuss some of the ideas we talked about in class on Monday and in our reading on Wednesday dealing with selection bias. Since you're talking to party elites (who don't know or like technical jargon), try to talk about selection bias in plain terms so that they understand the intuition for why selection on the DV is conceptually a bad idea.

For (2), create two scatter plots of party spending. For the first, filter the data down to just the districts where the party won (where they got more than half the vote). For the second, use the full dataset. Each figure should show party spending on the x-axis and the party vote share on the y-axis, and you should show the trend using a simple linear model. Talk about the differences you see, speculate about why these differences exist, and reinforce for party elites the importance of looking at all the data.

Conclude with a paragraph summarizing your recommendations for party leaders.

Some ground rules for your data viz and submission:

- Before each figure provide a description of it (what kind of plot is it? and what are the variables?) and what the figure tells us (are the variables positively or negatively correlated? is there no correlation?). Your figures should have clear and informative x- and y-axis labels and a title.
- You'll render your report from a .qmd file to a word document. Make sure that in the final rendered version your code is hidden. We should only see the text of your report and your data viz. To make sure this happens, your very first code chunk in your report should look like this:

```
#| include: false
knitr::opts_chunk$set(
  warning = F,
  message = F,
  echo = F,
  dpi = 500,
  out.width = 80%
)
```

The next page has a link to the data and a summary of what it contains.

The Data

link: https://raw.githubusercontent.com/milesdwilliams15/Teaching/main/DPR%20201/Data/HouseElectionsSpending2018.csv

Details (Metadata):

This dataset shows electoral outcomes in U.S. House races for 2018. Observations in the data are at the district level. There are a total of 331 observations in the data. The data contains the following variables per district:

- state: Acronym for the state in which a district is located.
- dist: Numeric ID for a district within a state.
- incumbent: Indicator for whether the incumbent is a Democrat ("D") or Republican ("R").
- repvoteshare: The proportion (or share) of votes received by the Republican candidate (hint: the Democratic candidate's vote share is just 1 minus this).
- repspending: Total spending by the Republican campaign in the district.
- demspending: Total spending by the Democratic campaign in the district.
- trumpvoteshare: The share of the vote that Donald Trump received in 2016.
- lagrepvoteshare: The share of the vote that the Republican candidate received in the last election (2016).