**Tuesday, January 11th, 2021:**

Kristopher:

Additional notes on datasets for Sebastian:

* As mentioned, all the data is based off the Form 990 ([see 2015 version here](https://www.irs.gov/pub/irs-prior/f990--2015.pdf))
  + Note: there are a couple different tax forms depending on the revenue size of the nonprofit. What we are focused on are nonprofits that submit the full Form 990. These are organizations that, on average, bring in $250,000 a year. So it excludes the smallest nonprofits (who only submit what is called a Form 990 EZ. The EZ does not require foreign expenditure disclosure so those filings are not helpful for the current project but are included in the data Jacob sent). There are also 990 PFs, these are for private foundations, so organizations whose sole purposes is to give away money. These are included in the data but are generally excluded from most nonprofit research because they are a distinct type of organization apart from nonprofits as we typically think of them.
* [If you look on OneDrive](https://princetonu-my.sharepoint.com/:f:/g/personal/kv7379_princeton_edu/Elk05Yg1gyhIk1UERMGSe8QBJVN2-MFW4oY4mjt6BpjW-Q?email=srojascabal%40princeton.edu&e=Z8ybVh). The datasets will follow different sections of the Form 990. I’m highlighting below what are most relevant for the current project:
  + **Return header**: this dataset includes all the administrative submission information such as a timestamp for when it was submitted, when the IRS put it on Amazon Web Services, the form type (990, EZ, PF). **It is important to note that each observation is the form submission.** That means the unique identifier within all these datasets will be a unique code noting the individual filing. So while most nonprofits only submit one per year, some nonprofits will submit multiple if they have to file a correction. So in 2020, for instance, Princeton University could have 3 submissions – one in February, one in May, and another in September – if we they had to re-file to correct previous mistakes. Therefore, always take the most recent filing within a year as being *the* observation we plan on using.
  + **Part 0**: This is the top portion of the 990 form:
    - Organization Name
    - Address
    - Employer Identification Number (EIN)
    - Etc.
    - Note: while when merging across datasets that Jacob sent the file ID is the unique identifier, my list of anti-LGBT organizations uses the EIN (which is the official code number the IRS gives each organization). So after we get to the most relevant organizations, EIN is what is going to be most important for identifying organizations.
  + **Part I:** Part 1 is all the summary financials. This is what we will need to determine overall revenues and overall expenditures if we want foreign expenditures to a percent of total budget rather than raw dollars.
  + **Part II:** Irrelevant
  + **Part III:** Part three includes the mission statement of the organization. This is extremely important and relevant text data. Especially as the project develops and we want to modify who “counts” as anti-LGBT, having these text data about the organization’s mission can be instrumental for identifying organizations we think should be included (e.g., any organization that has a religious ethos).
  + **Skip all other Parts:** So these “Parts” files represent the base form. This is part of the Form all nonprofits are expected to fill out. We will need some of these later for more detailed revenues and expenses, for example, but for the ASA submission and most immediate purpose, you can skip. Down the line, Part XIII (revenues) and Part IX (expenses) will become important. If you wanted to add them in now, you can go ahead but less of a priority.
  + **Schedule I:** So Schedule I lists other organizations the nonprofit gave money to ***within the United States.*** We may want this data if we want to eventually expand the project to ask: “After LGBT policy change occurs in some state (e.g., Texas), do nonprofits in that state start sending more money to other states within the U.S.?” So a variation of our research question be distributed across states and not countries. Similarly, you could add these data in now if you wanted.
  + **Schedule F:** So Schedule F is *the* part most relevant to us. This is the Schedule where nonprofits must detail foreign expenses ([example](https://projects.propublica.org/nonprofits/organizations/541660459/202101339349302205/IRS990ScheduleF)). What’s going to take the most effort as you can see from the data files is organizing these repeated data. So a nonprofit can have 8 expenses into Europe, each for different or for the same purposes, but they are all unique variables. I have yet to work with these data so I will defer to you on how best to make this workable because we may want to eventually break down how much money is going into each region. **For right now, if it is easiest to collapse all the information into two variables (# of individual grants; total $ amount) go ahead.** But if there is a way to both create summary variables while retaining individual allocation information, that will be best. Getting these data organized will be the hardest part.
  + **Everything else you can either ignore or incorporate based on your own curiosity.**

**Friday, December 24, 2021:**

Kristopher:

* Today, Jacob Fenton sent over the data for me to look at. Once I give him my approval, he will finish sending over the remainder.
* There are 41 individual data files. Not all are completely necessary for this project. But they are quite large as there are over 3.6 million nonprofit tax filing records total (but fewer within each dataset as not all nonprofits fill out every sevtion). So, while Github will serve as the central organizing repository, raw datafiles will need to be saved onto OneDrive.
* Once these larger dataset are limited down to just relevant observations, these smaller datasets which take up less space will be able to be saved onto Github. Once done, we will be able to make the GitHub folder our entire working directory for the entirety of the project.
* Notes to Sebastián for data:
  + General note: data may appear overwhelming, but we can slowly walk through it all. It’ll be less overwhelming than it first appears – I promise.
  + I was playing with some foreign expenditures data and there are some U.S. based organizations that simply never spend money abroad. My initial thinking is these should be excluded as if they are fundamentally U.S. focused, then nothing will influence foreign spending.
    - But bigger question about how best to precisely determine “anti-LGBT”
    - Quick reference: Right now, it’s a list of about 500 organizations I’ve collected that have participated in anti-LGBT/anti-gender campaigns, conferences, lobbying, etc.
  + Initial trendlines suggest that anti-LGBT organizations are actually bringing money in relative to overall nonprofits. May need to consider % rather than raw $ total.
    - Aside: If we actually demonstrate that anti-LGBT organizations move more money *back* into the U.S. that would be a powerful finding and contrary to so much public assumptions. So any significant effect is interesting!
    - But, if true, pro-LGBT organizations will be quite upset. So that makes the definition of who is counted within anti-LGBT very important.
  + We’ll need to convert all financials into real dollars so we don’t have to worry about inflation (not a big issue though since inflation was relatively low during this period)
  + I talked to Sam about doing some type of causal inference. He recommended [synthetic controls.](https://rss.onlinelibrary.wiley.com/doi/10.1111/rssb.12448) This is something for us to investigate down the line. [Corresponding R package.](https://github.com/ebenmichael/augsynth/blob/master/vignettes/singlesynth-vignette.md)
    - That said, if we want to do a causal approach, we’ll need to think through the treatment groups carefully to know the relevant comparisons. I can see one model comparing all anti-LGBT organizations across the U.S. and how state-level policy intervention influences foreign funding.
    - For federal policy, though, since all anti-LGBT organizations are exposed to the treatment. The more relevant comparison may be the nonprofit sector as a whole? Or similar conservative organizations?