Table of Contents

Return your materials within 48 hours
Please submit written responses as a PDF and code in a zipped directory.
You may access any reference materials you find helpful, but do not confer with others.
In your submission, please affirm that this is your own original work and note approximately how long you spent on the assignment in total.

- I affirm that this is my original work.
- I got a little carried away and took about 5 hours to do this. It was a fun exercise!

Section 1

a.

Going forward, we'll want to spruce this up a bit. In a GitHub gist or separate file, rewrite this query for clarity so it's easier to collaborate with teammates going forward (but don't worry about addressing the logic).

The author's goal was to identify the US Senate candidates with the most out-of-state donors. While the query probably returned the right answer for that specific question, its logic is over-simplified for our broader use cases.

BigQuery query, for ease, is here. GitHub script, for version control, is here.

b.

Identify a few assumptions this query makes about the underlying data. Are any of them problematic? You do not need to rewrite the query to account for the assumptions you identify. If you have an area of concern that would require more research into FEC data, just make a note of it.

Notes on Cleaning up the Query

Being Explicit

- It doesn't reference the aliases of the subqueries in the top SELECT statement or in the bottom WHERE statement. It also does a GROUP by the number of the columns in the top SELECT statement instead of explicitly identifying the columns. It is best to be explicit about the columns (table.column) when making code for others (or future you) to read so as to reduce time figuring out what the implications are.

Improving Efficiency of Queries

- It is best to avoid using SELECT * statements in BigQuery because it is columnar, which makes it very efficient as long as you don't use SELECT *!

Readability

- It is best to list one field/column name listed per line, which makes it very easy to see each column being used. *Controversially,* I am an advocate for leading columns so that you can more easily see that you're not missing any, but I appreciate that trailing columns are the industry standard and if that's already established I'm OK with using trailing columns.

Assumptions

This data only looks at contributions made during the 2020 election cycle. We likely want to have insights into other years because candidates running now may have run in years previous to 2020 and because insights into contentious elections in similar campaigns in previous years might provide insights into the requisite budget of future campaigns.

The query only looks at contributions where the filer is a candidate. It excludes donations made by committees, including instances where a candidate is associated with a committee but the committee filed instead of the candidate. Some committees are associated with multiple candidates so it is inaccurate to say that if a committee is associated with a candidate that the contribution necessarily went to the candidate but, for committees associated with only one candidate this is at least *more* likely to be the case. This query shows the general logic I would apply to pull in the filer information on contributions. In practice I would have an enriched version of the contributions records with the filer info.

The query assumes that all contributions are captured in the indivYY tables; I see that there are othYY tables that also capture contributions, which would also be

useful to know. I would want to dig into why there are these othYY tables and how they are different from the various types of contributions listed in the invdYY tables.

The query only includes contributions from individual people. There is an important difference between an “individual donation” meaning a distinct donation and a donation from an individual person. While there are instances where insights into contributions from individual people is highly valuable, that limitation does not seem consistent with the spirit of the original author’s intended universe of data: “-- *which Senate candidates had the most donors from outside their state in the past 16 years?*” It does not specify individuals and therefore is likely misleading to users. It would be helpful to have the information viewable by donor types (indv20.entity_tp).

It groups on the candidate’s state (cand_office_st), which may be different from year to year and also may be NULL. Also, it is unclear on *why* they group the results by the candidate’s office when that’s not indicated in the description of the query.

I see that the table indiv16_cm16 represents some sort of exception, although the nature of the exception is unclear. I notice that it has nearly double the columns as the other individual contribution tables and the transaction identified appear to be a different pattern, indicating that it may require significant adjustments to be cleaned and incorporated into the rest of the donor data.

Section 2

For a quick win, we want to start work on a simple dashboard of out-of-state individual donations to federal candidates.

After discussing the idea with our team, we have a few clear requirements: our dashboard must support filters by candidate name, candidate office (House, Senate, or both), and candidate state. It should display the number of donors, the number of donations, and the total amount donated to the candidates from out of state. These metrics should be visible in total and also by donor state.

In a gist or separate .sql file, draft a query for a table that can support this dashboard. We are only looking to assemble the underlying data; please do not create a dashboard!

- BigQuery query here, for ease.
- Same script, in GitHub, for version control and following instructions.

Even with carefully considered code, this dataset can’t produce a perfect reflection of these candidates’ donors. Draft a short paragraph describing the dashboard’s limitations, to be included alongside the dashboard itself.

- Regulatory Restrictions

    - The FEC has strict restrictions on the use of contribution data for fundraising. Although it is possible to review contributions, even on the donor level, the use of donor data for solicitation of fundraising is prohibited. So we would want to give some guidelines on do's and don'ts so we don't get in trouble and our clients don't get in trouble.
- Date Restrictions
    - For the purposes of this exercise, if I had more time on this dataset I would union the invdYY tables together for 2006-2020, to be able to look at historic data in order to see trends in campaign financing. However, this dataset still only goes to 2020. Due to the steep increase in funding going into campaigns these past 4 years of data are a big blind spot. This can be solved by using the openFEC API!
- Static Data
    - The FEC updates its contribution data on an ongoing and fairly current schedule. Rather than relying on this public dataset it would be better to use the openFEC API to pull in fresh data on an ongoing schedule into our own warehouse.

Section 3

a.

There are many different directions we could take this new product. What are your ideas? In a few sentences, summarize a few (2-3) ways we could get this data into users' hands in a valuable, time-saving form.

Aggregation for Adds

- Although direct solicitation to individual donors based on FEC data is prohibited, we *can* use aggregated information to inform our adds and localization of donor solicitation.
- Adds
    - We could pair this data with donor's demographics (from the Voter File and from Target Smart and other datasets) and aggregate on demographics such as age, sex, race and zipcode.
    - The actual FEC data has timestamps, so we could also use that to provide guidance on what time of day and days of the weeks adds should be active and emails should be sent.
    - As we have the city and zipcodes of the donors we could aggregate by location for billboard advertisement.

Out-of-District Partnerships

- Many campaigns are now boosted by out-of-district and out-of-state donors and orgs. Using this data to understand where outside funds are coming from can help campaigns allocate time to actually go to those outside locations to fundraise. In my experience doing data volunteerism for Tech for Campaigns and Sister District the most valuable fundraising actions are in-person fundraisers where the candidate is present. This data

can help candidates identify the orgs and the locations where they may need to go to get their biggest donors dollars.

Due to the restrictions on using this FEC donations data for direct solicitations of funds these were the ideas that are unfortunately not permissible.

- Creating donor-level activist codes (small, medium, large and very_large donor indicators) and applying them to the Voter File.
- Enriching the data with validated phone numbers and creating ready-to-go text and call lists.

b.

Compared to the raw source data, what do we need to do to make our end product accessible and approachable?

Automated ETL. We would use the openFEC API to ingest the data, run it through validation (whenever you have to union and join together a bunch of tables you want to check to see where you have duplicate and NULL values to understand what data is duplicated or missing, under which circumstances), transform the data into a few enriched tables with clear data dictionaries, and host the data with login access for our clients. Given that that most users do not need most of the data, it would be best for users to be able to filter a selection before downloading the data. Filters would be on:

- State of donor (drop-down)
- State of filer (drop-down)
- Donation range (a slider)
- Dates of donations (date range)
- Campaign cycle (drop-down)
- Candidate (search)
    - Candidates and committees associated with the chosen candidates (check box if the Candidate search box is used)
- Committee (search box)
    - Committees and candidates associated with chosen committees (check box if the Committee search box is used)
- Committee Designation (drop-down)
- Committee Type (drop-down)
- Committee Party (drop-down)
- Committee Interest Group Category (drop-down)

Most campaigns aren't going to have robust data resources, so one or more video tutorials on how to download and use the data would be helpful for most users.

What I Would Do with More Time

More validation!

- From what I have seen so far there are issues with duplicates, especially in the contribution records.
- I would want to do further validation on the linkage between candidate and committees, cross-referencing with the cmYY tables.
- I would also want to dig into the othYY tables to see what contribution data is there.
- I would want to dig more into the openFEC API and see what their limitations are on pulling from it.