Electoral Funding and Congressional Voting
Capstone Project Plan

Jonas Joehnk

PURPOSE

Summary of Topic

The United States spends more money on elections than any other nation, and even more so following the controversial Supreme Court decision in *Citizens United v. Federal Election Committee* in 2010 that prohibited the government from restricting independent expenditures on political campaigns by third-party groups.

In this project, the focus is to explore the relationship between funding that comes from special interest groups, in particular from the Super PACs made possible by the *Citizens United* decision, and the voting behavior of elected congress members who either receive their funding directly or benefit indirectly from their expenditures.

Hypothesis

If we test the variance between the perceived interests of special groups and the voting patterns of congress people they support, there will be little to no difference in ideology between the two groups. Furthermore, if we compare the voting patterns of the congress people who receive funding from the same parties, there will similarly be little to no difference in ideological support between the congress members.

Benefit

There is a wealth of data from the Federal Election Commission on contributions and expenditures by committees during U.S. elections, and this project would create a data extraction pipeline into a database and Tableau application for accessible description and analysis.

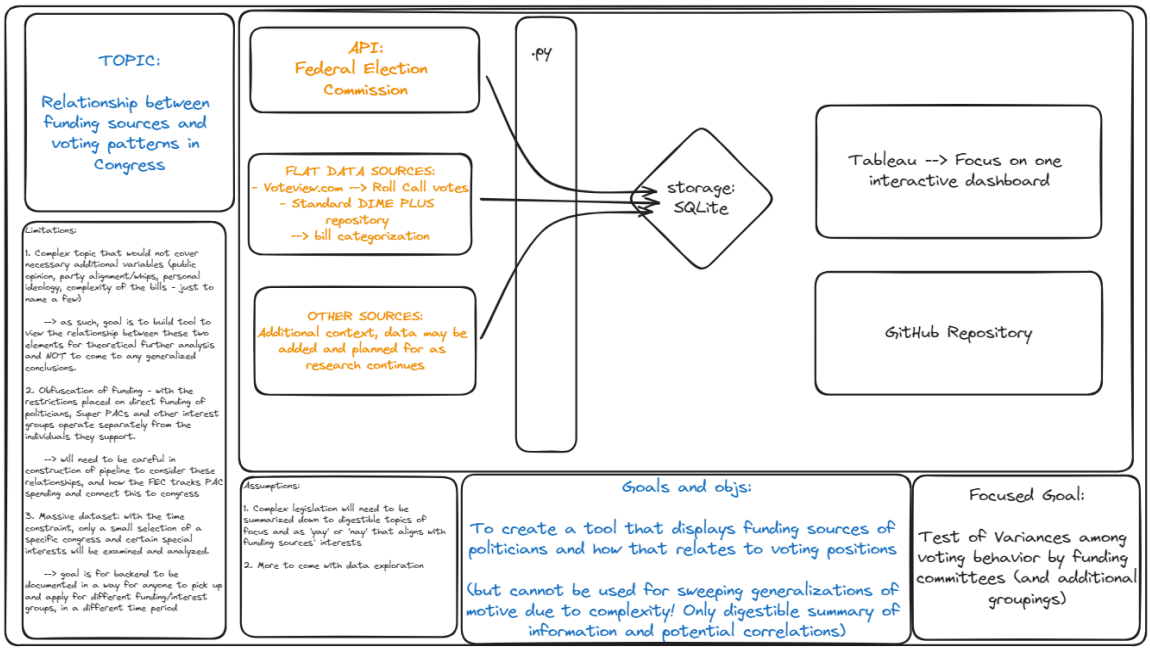
In creating a database of electoral funding alongside congressional voting information, this further allows for the analysis of patterns and correlations of these two elements in our political landscape.

Desired Outcomes

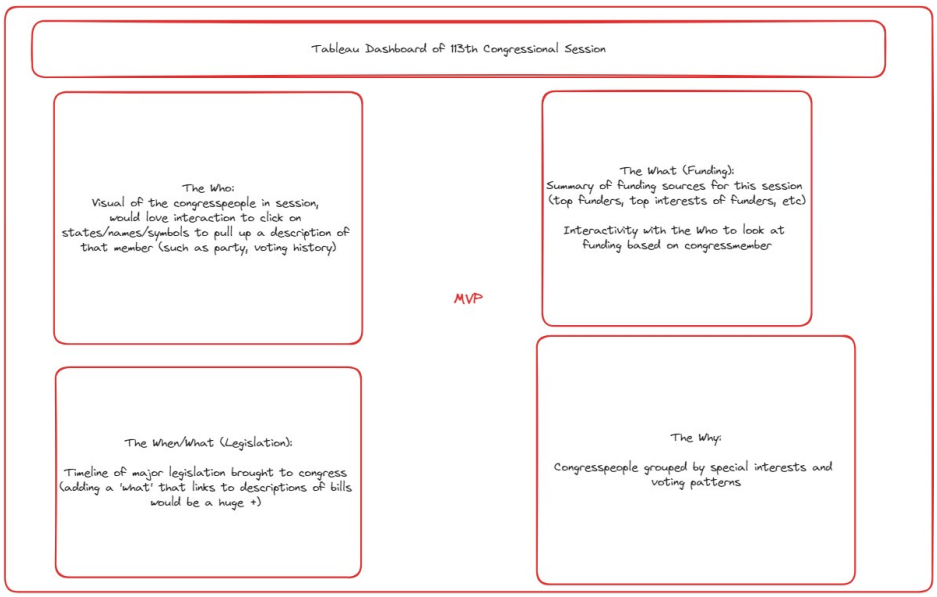
The first goal of this project is to build a strong back-end database of data. The FEC holds easily billions of rows of information that is not accessible for immediate analysis. While this project could never attempt to host all of that data, the idea is to practice developing strong documentation and road-mapping for extraction, loading, and transformation of data.

The second goal is to take the molded data and test for variances in voting behavior as would be expected by the special interests of their funding groups and display the results of the summarized data and the hypothesis testing in an interactive tableau, along with clear documentation for repeatability in a github repository.

This illustration shows a breakdown of anticipated flow of data from raw data through Python in Jupyter, stored in SQLite and then presented in Tableau and GitHub. Click on the image for an expanded view.



Below details an estimation of the information that would be presented in Tableau. Click on the image for an expanded view.



DATA

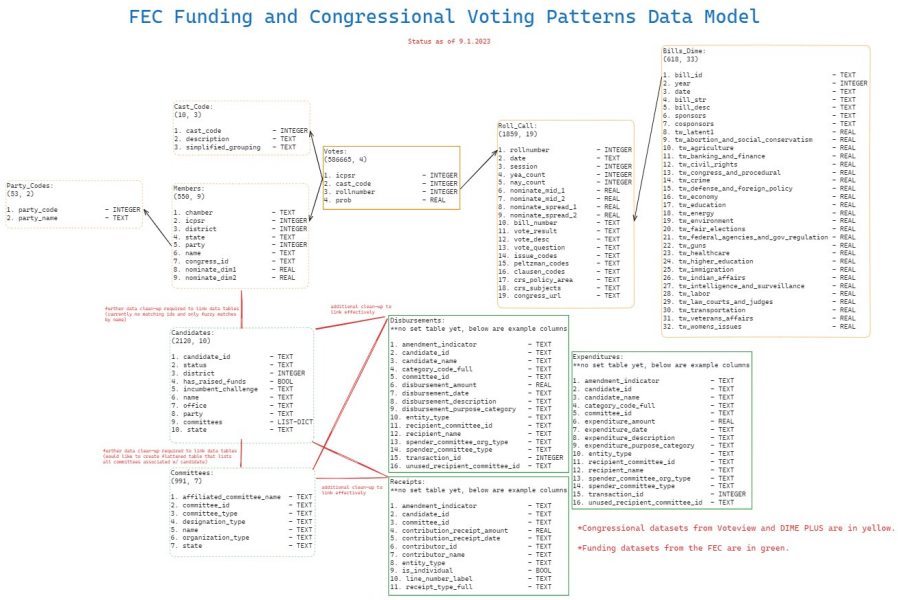
Datasets

- Voteview Roll Call Datasets:
  - https://voteview.com/data
  - Referential table on member descriptions
  - Referential table on legislation put forth for a vote
  - Transactional table itemizing votes put forth by each congress member for every bill
- DIME PLUS Datasets:
  - https://data.stanford.edu/dime-plus
  - Ideological stances of legislation put forth in congress, with details on sponsors and cosponsors
  - Transcripts of all discussions, proceedings, and debates on the congressional
- Federal Election Commission API:
  - https://api.open.fec.gov/developers/
  - All U.S. electoral candidates and committees with details of their receipts of contributions, disbursements, and expenditures

Data Model

An early focus of this project is to create an organized and efficient data model. Below shows what the model currently looks like, but this will evolve considerably.

Click on the image to view a larger version on Excalidraw.



“Story of One Row”

The transactional tables that describe key points are:

- Congressional Voting Tables:
    - *The below data provides ample information on the voting behavior of the 113th congressional session, and multiple issue categorization methods to describe the bills voted on.*
    - **Votes**
        - Each row describes a single congressperson’s vote for each roll call, along with the probability they would have made this vote (as determined by the NOMINATE scale from Keith Poole and Howard Rosenthal, original authors of Voteview)
    - **Roll Call**
        - Each row describes a vote brought forth to congress (bills, amendments, and other motions), along with multiple classification methods of each bill’s content (from Poole & Rosenthal, Congressional Research Service, Clausen, and Peltzman) and NOMINATE dimensional estimations.
    - **Bills Dime**
        - A secondary source that also details each vote brought forth to congress that categorizes their subject’s based on an alternative ‘topic weight calculation’, as well as details of a bill’s sponsor and cosponsors.
- Funding Tables:
    - *The below data details all electoral financial statements in the period of 2011-2012 (when the 113th session of congress was elected).*
    - Disbursements
        - Each row describes itemized disbursements of how a committee spends their money. Once filtered, this can show money directly contributed to a political candidate.
    - Expenditures
        - Each row describes line item expenditures of a committee. Once filtered, this can show money spent for a political candidate or against their opposition that did not directly go to an individual running for office.
    - Receipts
        - Each row describes itemized receipts, including all contributions from individuals and committees to a particular candidate.

All other tables are referential and offer additional information on member backgrounds, dictionaries of coded categories in transactional tables and so on.

Fields and Calculations of Insight

The first group of anticipated columns for major insight would be ‘cast_code’ from votes in conjunction with the issue code and topic weight columns from Roll Call and Bills Dime, as that would lend to a summarized picture of the political ideology and voting behavior of all congress members.

Next, looking at the disbursement/contribution/expenditure amounts detailed from the FEC that flow between each congress member and special interest committees and comparing that to the above detailed voting behavior allows for testing the hypothesis that the way congress people vote will not deviate from the interests of the committees that fund them.

Data Dictionaries

Voteview Datasets

Party Codes

1. Party_Code → INTEGER
    a. Unique id for all U.S political parties
2. Party_Name → TEXT
    a. Long-form text name of U.S. political parties

Cast Codes

1. Cast_Code → INTEGER
    a. Numerical representation of vote responses
2. Description → TEXT
    a. Long-form text description of vote responses
3. Simplified_Grouping → TEXT
    a. Self-defined grouping of all 10 possible responses into ‘Yea’, ‘Nay’, ‘Present’, or ‘Abstain’

Members

1. Chamber → TEXT
    a. House or Senate
2. ICPSR_ID → INTEGER
    a. Unique id for all U.S. elected officials used by Voteview
3. District → TEXT
    a. The U.S. district the congressperson represents
4. State → TEXT
    a. The U.S. state the congressperson represents
5. Party → INTEGER

    a. The party code of a congressperson
6. Name → TEXT
    a. The name of the elected official
7. Congress_ID → TEXT
    a. The unique id of all elected officials as represented on government websites
8. Nominate_Dim1 → REAL
    a. The estimated placement of a congressperson along the first dimension of the NOMINATE scale. This dimension (often represented along the x-axis) corresponds to the left-right or liberal-conservative spectrum on economic matters.
9. Nominate_Dim2 → REAL
    a. The estimated placement of a congressperson along the second dimension of the NOMINATE scale. This dimension (often represented along the y-axis) corresponds to social, salient issues of the time, such as civil rights, abortion, gun rights, and so on.

Votes

1. ICPSR_ID → INTEGER
    a. Unique id for all U.S. elected officials used by Voteview
2. Cast_Code → INTEGER
    a. Numerical representation of vote responses
3. Rollnumber → INTEGER
    a. Serial id of roll call vote that begins at 1 the start of every new congressional session
4. Prob → REAL
    a. Estimated probability, based on NOMINATE, of the member making the vote as recorded

Roll Call

1. Rollnumber → INTEGER
    a. Serial id of roll call vote that begins at 1 the start of every new congressional session
2. Date → TEXT
    a. Date a roll call vote took place
3. Session → INTEGER
    a. Either 1 or 2, corresponding to which session of congress the vote took place during
4. Yea_Count → INTEGER
    a. The number of votes in favor of a decision
5. Nay_Count → INTEGER
    a. The number of votes against a decision
6. Nominate_mid_1 → REAL
    a. First-dimension midpoint estimate based on the Nominate scale
7. Nominate_mid_2 → REAL

    a. Second-dimension midpoint estimate based on the Nominate scale
8. Nominate_spread_1 → REAL
    a. First-dimension spread estimate based on the Nominate scale
9. Nominate_spread_2 → REAL
    a. Second-dimension spread estimate based on the Nominate scale
10. Bill_number → TEXT
    a. Unique id of the bill
11. Vote_Result → TEXT
    a. Official result of the roll call
12. Vote_Desc → TEXT
    a. Description of the roll call as assigned by Congressional staff
13. Vote_Question → TEXT
    a. Question addressed by the roll call
14. Issue_Codes → TEXT
    a. Poole and Rosenthal specific issue codes.
15. Peltzman_Codes → TEXT
    a. Peltzman (1984) issue-area codes.
16. Clausen_Codes → TEXT
    a. Clausen (1973) issue-area codes.
17. CRS_Policy_Area → TEXT
    a. Congressional Research Service policy area.
18. CRS_Subjects → TEXT
    a. Congressional Research Service subject area.
19. Congress_url → TEXT
    a. Link to the bill description on congress.gov

Stanford DIME PLUS Datasets

Bills Dime

1. Bill_ID → TEXT
    a. Unique id of the bill
2. Year → INTEGER
    a. Year roll call vote took place
3. Date → TEXT
    a. Date roll call vote took place
4. Bill_Desc → TEXT
    a. Description of the bill
5. Sponsors → TEXT
    a. Congressperson that sponsored the bill
6. Cosponsors → TEXT
    a. Additional sponsors of the bill
7. TW_Latent1 → REAL

    a. Topic weight for the latent category calculated by the authors of the DIME PLUS repository. All below columns are the topic weight calculations for their corresponding header name.
8. TW_Abortion_and_Social_Conservatism → REAL
9. TW_Agriculture → REAL
10. TW_Banking_and_Finance → REAL
11. TW_Civil_Rights → REAL
12. TW_Congress_and_Procedural → REAL
13. TW_Crime → REAL
14. TW_Defense_and_Foreign_Policy → REAL
15. TW_Economy → REAL
16. TW_Education → REAL
17. TW_Energy → REAL
18. TW_Environment → REAL
19. TW_Fair_Elections → REAL
20. TW_Federal_Agencies_and_Gov_Regulation → REAL
21. TW_Guns → REAL
22. TW_Healthcare → REAL
23. TW_Higher_Education → REAL
24. TW_Immigration → REAL
25. TW_Indian_Affairs → REAL
26. TW_Intelligence_and_Surveillance → REAL
27. TW_Labor → REAL
28. TW_Law_Courts_and_Judges → REAL
29. TW_Transportation → REAL
30. TW_Veterans_Affairs → REAL
31. TW_Womens_Issues → REAL

Federal Election Commission Datasets

Candidates

1. Candidate_ID → TEXT
    a. Unique id from the FEC of all candidates who run for office in the U.S.
2. Status → TEXT
    a. Whether or not a candidate is statutory, or in the future/past
3. District → INTEGER
    a. Corresponding district a candidate is running for election in
4. Has_Raised_Funds → BOOLEAN
    a. Whether a candidate has received funds or not
5. Incumbent_Challenge → TEXT
    a. Whether a candidate is an incumbent or challenger
6. Name → TEXT
    a. Name of candidate
7. Office → TEXT

    a. House, Senate, or Presidency
8. Party → TEXT
    a. U.S. Political Party candidate is running under
9. Committees → LIST OF DICTIONARIES
    a. This contains all the committee ids a candidate is associated with. Needs considerable cleaning at this time.
10. State → TEXT
    a. Corresponding state a candidate is running for election in

Committees

1. Affiliated_Committee_Name → TEXT
    a. Affiliated organization name (such as the company or association)
2. Committee_ID → TEXT
    a. Unique ID from the FEC for committees that donate to federal election candidates and/or expend funds for election purposes
3. Committee_Type → TEXT
    a. Type of committee, such as PAC or Political Party
4. Designation_Type → TEXT
    a. Subcategory type of committee such as Lobbyist PAC
5. Name → TEXT
    a. Name of the committee as registered with the FEC
6. Organization_Type → TEXT
    a. Type of the affiliation organization
7. State →TEXT
    a. Location the committee is registered in

Disbursements, Receipts, Expenditures

** These datasets are not yet clearly defined at the time of the project plan.

ANALYSIS

Key Performance Indicators

- Amount of funds contributed to congress members
- Amount of funds disbursed to, or expended on behalf of, congress members
- Calculation of votes by congress members in alignment with or deviating from special interest groups

Main Questions of Analysis

- Do voting patterns of members correspond to the interests of those that fund them? If there is deviation or assimilation, is it more or less so for certain issue areas/specific members?
- Where do members of the 113th congressional session receive funding from?
- What committees, and more broadly what issue areas, spend the most on congressional candidates?
- Which congress members are funded the most? The least? Is there any pattern to derive from the result?
- What is the statistical significance of any variance found in voting patterns?
- How does the proportional funding by interest groups correspond to the subject matters of bills brought forth?

Process for Observations

The first key step is continued data modeling. As of now, the two groups of datasets - congressional votes and FEC funding - do not communicate to one another. Determining how to link all of the tables and modify existing ones to create an organized flow of data is crucial to the efficiency of any analysis.

Following that, the goal is to capture summary data of the roll call votes: what were the most frequent topics discussed, what was the make-up of voting decisions for each congressperson by topic. The same process is then needed for funding data: who are the largest funders, what is the average donation amount in total/by member, what are the descriptive statistics of disbursements to candidates or expenditures for them.

With a foundational understanding of the above, then the final step is to test for variance from expectations (all determined based on who funds the corresponding member), and thus testing if there is a correlation between funding and voting patterns.

Required Calculations/Statistical Measures

Calculations necessary for this will include: variance from expected values, variance from mean (when grouping members based on similar funding sources), and others to understand the congruence or deviation of behavior from expected.

Models or “What If” Scenarios

This project is not focused on building what-if scenarios to predict future scenarios, or map alternative ones. A potential question of interest would be to ask if changes in committee support and subsequent funding could impact how a congress member votes, but that is frankly outside of the scope of this project.

More importantly, as this project is a snapshot of one particular congressional session, and focuses on exploring the potential correlation between two variables, it wouldn’t be accurate or the right direction to build out models.

Analysis Methods

This project utilizes descriptive analytics to summarize funding and voting patterns in 2013-2014, and diagnostic analytics to speak to the relationship between these two variables in explaining members’ ideological actions.

There will not be a focus on predictive analytics to understand likelihoods in the future - especially as this case study looks just at one period of time a decade ago with limited historical context - nor does it aim to use prescriptive analytics and make any recommendations on a course of action based on the findings.

Assumptions

First, using outside sources as a means of understanding the content of roll call legislation. Significant background research is going into understanding the methodologies put forth by Poole & Rosenthal in Voteview and Bonica at DIME PLUS, but ultimately their scales will be assumed accurate.

Second, this analysis assumes that the 113th session is a statistically normal congress to observe and draw correlation from. While there is significant emphasis on the fact that no strong conclusions should be drawn from the set-up of this project, even any preliminary findings communicated suggest a normality of this case study.

Limitations

This is a complex topic, and the limited scope of this project means that the plethora of context and additional variables necessary to address are ignored here. Congressional voting behavior is determined by public opinion, party alignment, personal ideology, among many others.

With that in mind, the goal is to build an accessible database and subsequently a toll that looks into the relationship between funding and voting, but in no way comes to generalized conclusions

Other limitations include:

- Obfuscation of funding due to Super PACs acting as third-party entities to candidates
- Massive source of data (over a billion rows of data from the FEC) means that only a selection of details will be wrangled and analyzed in the case study.
- Availability of NOMINATE and DIME scaling for congressional sessions ends after the 113th, which determined the choice of session to use as a case study.