CS506 Project Final Report

Project Title: **Campaign Finance Scorecard - State Reps**
Team Members: Alikhan Nurlanuly, Yiquan Xing, Michael Harding
Code: https://github.com/alikhanlab/Campaign-Finance-Scorecard-Project-State-Reps

Project Problem Statement

To develop a Campaign Finance Scorecard for elected Massachusetts State Representatives based on publicly available campaign contribution data. The Scorecard will be an objective presentation of donors' contributions for each State Rep to answer the following political research questions:

Question 1: Number of unique donors for all the quarters for the past 2 election cycles from 2015-2018

Question 2: Distribution of contributions across 5 contribution bands (<$25, $25-$99, $100-$249, $250-$499, $500-$1000)

Question 3 Geographical distribution of donors (in-district, out of district, out of state)

Question 4: Monthly donations by PACs (Political Action Committees)

Question 5: PAC donations by industry sector, focusing on

1. Real Estate - including developers and construction.
2. Healthcare - including hospitals and health insurance.
3. Biotech/Pharmaceuticals

Question 6: Donations by Employer Types (excluding PACs):

- Examples - police, private education
- Industry sectors of interest:
    a. Real Estate, Construction
    b. Healthcare, Hospitals, Health Insurance Companies
    c. Pharmaceuticals, Biotech Companies

Question 7: Donation by Donor Type:

- Donations from individuals, PACs, Unions, Lobbyists and from Candidates to themselves

Insights which we hope to gain are:

1. Which employers and PACs are spending most money and how much are they giving? What industry are they in? How has their giving changed over time? What are the peaks?

2. Who are the wealthiest donors (i.e. giving maximum amount of $1,000), what are the patterns associated with their giving (e.g. average per year, geographic concentrations, favorite candidates)?

Methodology

Datasets Collected & Used

1. Political Contributions from **OCPF (Office of Campaign and Political Finance)**
    a. **Representative data** called ‘Master’ table detailing the Rep’s data such as name, address and office held
    b. **Contribution data** called ‘Receipts’ table detailing donors’ data such as name, address, occupation, employer, and contribution amount & date
2. MA State Reps list from from **MA Legislature website**
    a. Data on elected MA State Representatives such as name, district they represent, date assumed office and party affiliation
3. Industry Categories and Keywords for classifying Contributors into Industries
    a. Referenced the **Industry categories** from **NAICS (North American Industry Classification System)**
    b. **Industry keywords** collected from analyzing the **Bureau of Labor Statistics’s Occupational Outlook Handbook**, **OCPF** data and Internet research
4. Company Names and their Industry
    a. List of Massachusetts Real Estate related companies from **NAIOP (The Commercial Real Estate Development Association)**
    b. List of US companies based in Massachusetts and their industry classification from **LinkedIn** (which is very comprehensive and includes healthcare, pharmaceuticals, transportation, public administration, etc.)
5. Lobbyist Names
    a. List of Lobbyists names and entities maintained on the **Secretary of the Commonwealth of Massachusetts** website

Data Exploration & Classification

One of the key things we needed was to classify the source of donations:

1. What **Industry** are the donors from?
2. Which donors are **Lobbyists**, **PACs** **(Political Action Committees)** or **Unions**?

For **Lobbyists**, we can identify them by comparing the donor's name and employer to the list of Lobbyists maintained on the website of the Secretary of the Commonwealth of Massachusetts.

For **PACs** and **Unions**, we can identify them by keywords on donor's name and employer (e.g. PAC, union, association, etc).

Industry classification is less straightforward as there are multiple ways to determine a donor's industry due to the characteristics of our data:

- Firstly, we needed to establish a **standard set of Industry categories** for classification. We chose to reference the **NAICS (North American Industry Classification System)** as we see it as an authoritative source.
- Secondly, we needed to identify **attributes which give information about industry**. After analysing the **OCPF Contributions** data, we found that if the donor is a company, the donor name (i.e. company name) can be used to determine the industry; and if the donor is an individual, employer and occupation could be used.
- Thirdly, there are some **keywords** which are representative of specific industries which we can use to identify those industries (e.g. 'realtor' or 'realty' is characteristic of Real Estate industry, 'hospital' is characteristic of Healthcare). We compiled this list of keywords by referencing the **Bureau of Labor Statistics's Occupational Outlook Handbook** and also analysing and extracting from the **OCPF Contributions** data.
- Finally, specific companies belong to specific industries so if we could find or compile a list of **companies names and their industry,** we will be able to identify the industry of the donor through the donor's name or employer. We searched and found **LinkedIn companies** data, and from it extracted Massachusetts companies.

Next, in order to use LinkedIn's companies & industry data to help in classification, we needed to map LinkedIn's industry category to NAICS industry category which we have chosen to be the authoritative source. We did this manually and got the following mapping table:

Industry Category (referenced from NAICS)	**LinkedIn Industry Category**
Biotech/Pharma	'biotechnology', 'pharmaceuticals'
Education	'higher education', 'education management', 'primary/secondary education'

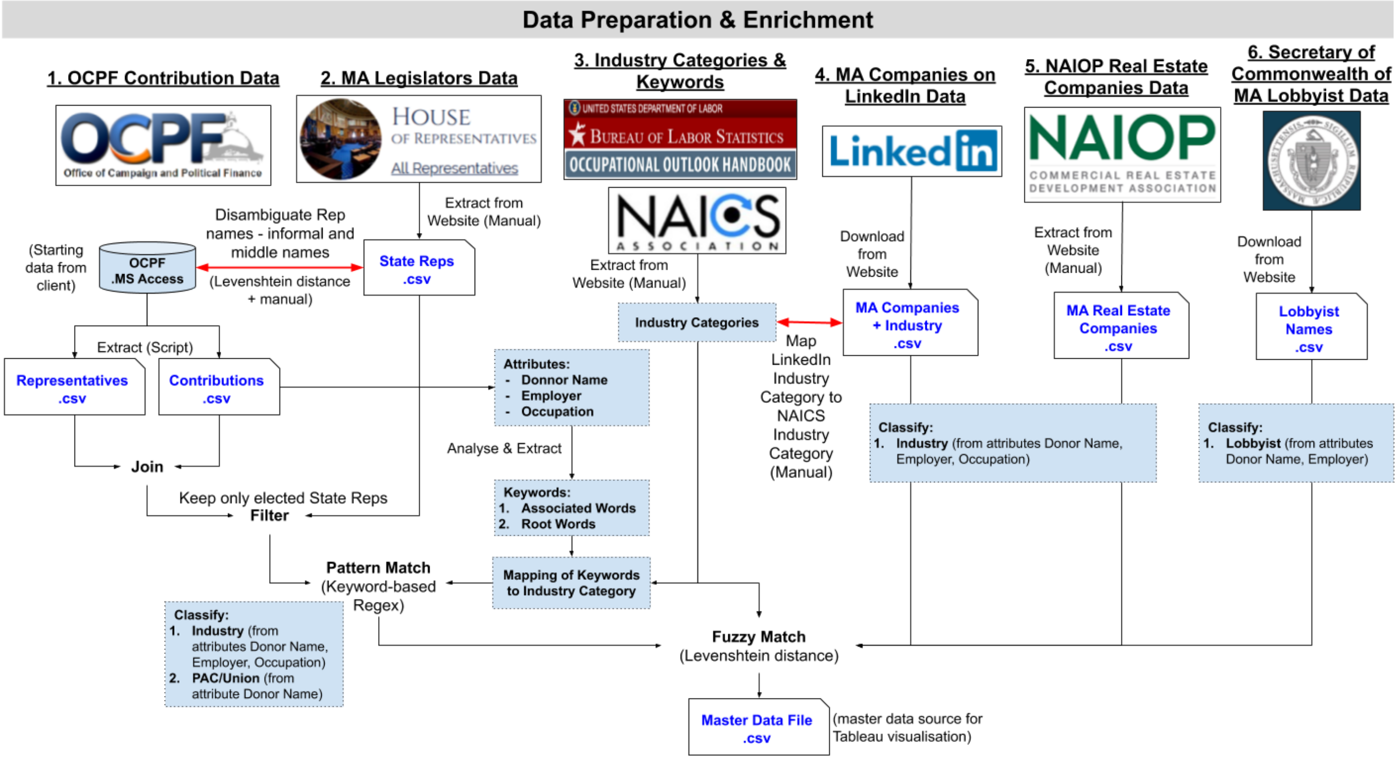
Healthcare	'hospital & health care', 'health, wellness and fitness', 'medical practice', 'mental health care', 'medical devices'
Real Estate	'real estate', 'construction', 'commercial real estate', 'architecture & planning', 'facilities services', 'building materials'
Law Practice	'law practice', 'legal services'
Government Relations	'government relations'
IT	'information technology and services', 'computer software', 'internet', 'semiconductors', 'telecommunications', 'computer & network security', 'graphic design'
Government	'government administration', 'defense & space'
Finance, Banking and Insurance	'financial services', 'banking', 'insurance', 'accounting', 'investment management', 'venture capital & private equity', 'investment banking'
Accomodation and Food Services	'restaurants', 'hospitality', 'entertainment'
Business Management	'management consulting', 'public relations and communications'
Retail and Wholesale Trade	'wholesale', 'food & beverages', 'wine and spirits', 'retail', 'food production'
Transportation	'transportation/trucking/railroad', 'automotive', 'warehousing'
Utilities	'oil & energy', 'utilities'

With all the data in place above, we proceeded to perform 2-tier classification to identify a donor’s Industry by doing the following in sequence:

1. **Pattern matching** keywords in donors’ name, employer and occupation. This is done using Python’s **regular expression** ‘re’ library.
2. **Fuzzy matching** company names in donors’ name and employer. This is done using the Python library ‘fuzzywuzzy’ which is based on **Levenshtein distance.**

Data Collection, Preparation & Enrichment Workflow

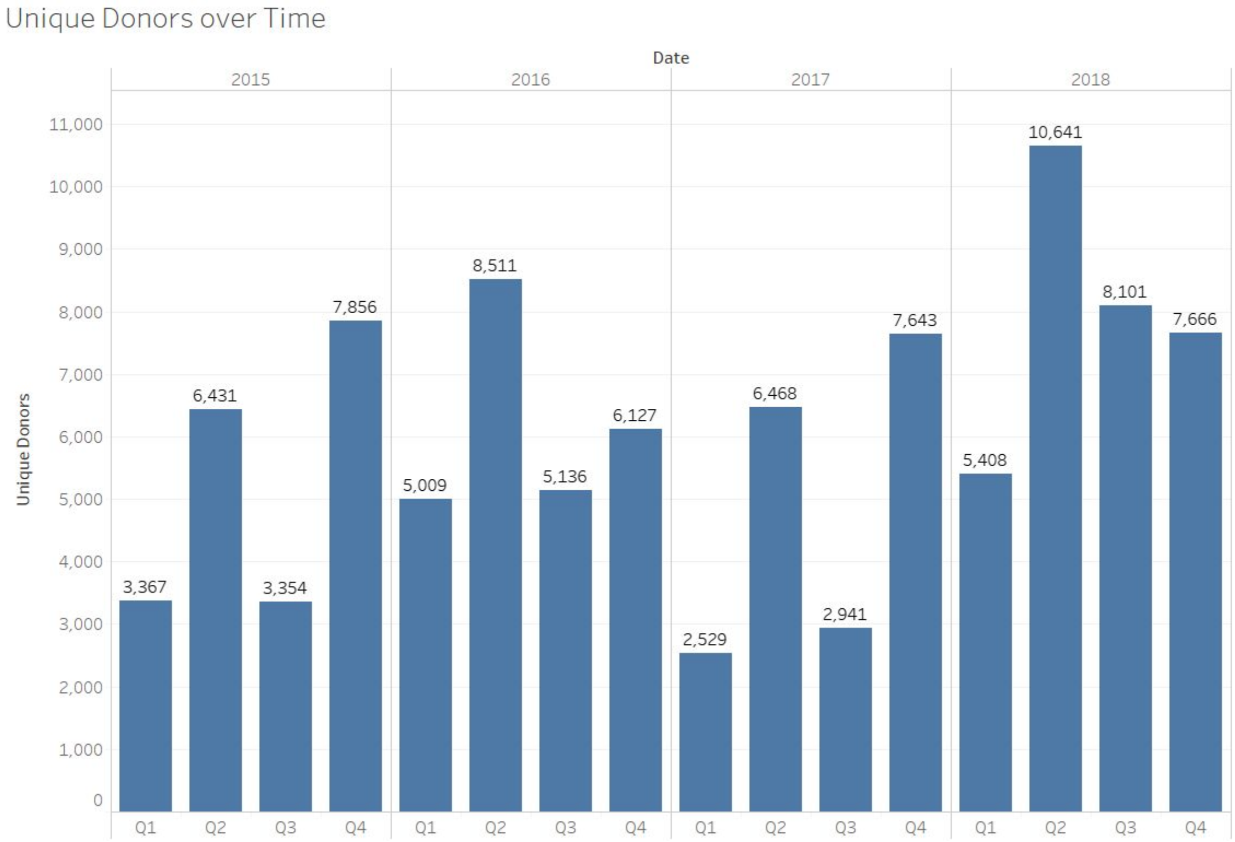
The following diagram gives an overview of the data collection, preparation and enrichment process. The output is the master data file (**contributions_processed.csv**) which is used for visualisation in **Tableau** public edition software to answer client's questions.


Data Preparation & Enrichment
1. OCPF Contribution Data
OCPF
Office of Campaign and Political Finance
2. MA Legislators Data
HOUSE OF REPRESENTATIVES
All Representatives
3. Industry Categories & Keywords
UNITED STATES DEPARTMENT OF LABOR
BUREAU OF LABOR STATISTICS
OCCUPATIONAL OUTLOOK HANDBOOK
NAICS ASSOCIATION
4. MA Companies on LinkedIn Data
LinkedIn
5. NAIOP Real Estate Companies Data
NAIOP
COMMERCIAL REAL ESTATE DEVELOPMENT ASSOCIATION
6. Secretary of Commonwealth of MA Lobbyist Data
(Starting data from client)
OCPF .MS Access
Disambiguate Rep names - informal and middle names
(Levenshtein distance + manual)
Extract from Website (Manual)
State Reps .csv
Extract (Script)
Representatives .csv
Contributions .csv
Join
Keep only elected State Reps
Filter
Extract from Website (Manual)
Industry Categories
Attributes:
- Donnor Name
- Employer
- Occupation
Analyse & Extract
Keywords:
1. Associated Words
2. Root Words
Mapping of Keywords to Industry Category
Pattern Match
(Keyword-based Regex)
Classify:
1. Industry (from attributes Donor Name, Employer, Occupation)
2. PAC/Union (from attribute Donor Name)
Download from Website
MA Companies + Industry .csv
Map LinkedIn Industry Category to NAICS Industry Category (Manual)
Extract from Website (Manual)
MA Real Estate Companies .csv
Classify:
1. Industry (from attributes Donor Name, Employer, Occupation)
Download from Website
Lobbyist Names .csv
Classify:
1. Lobbyist (from attributes Donor Name, Employer)
Fuzzy Match
(Levenshtein distance)
Master Data File .csv
(master data source for Tableau visualisation)

Findings

Visualisations

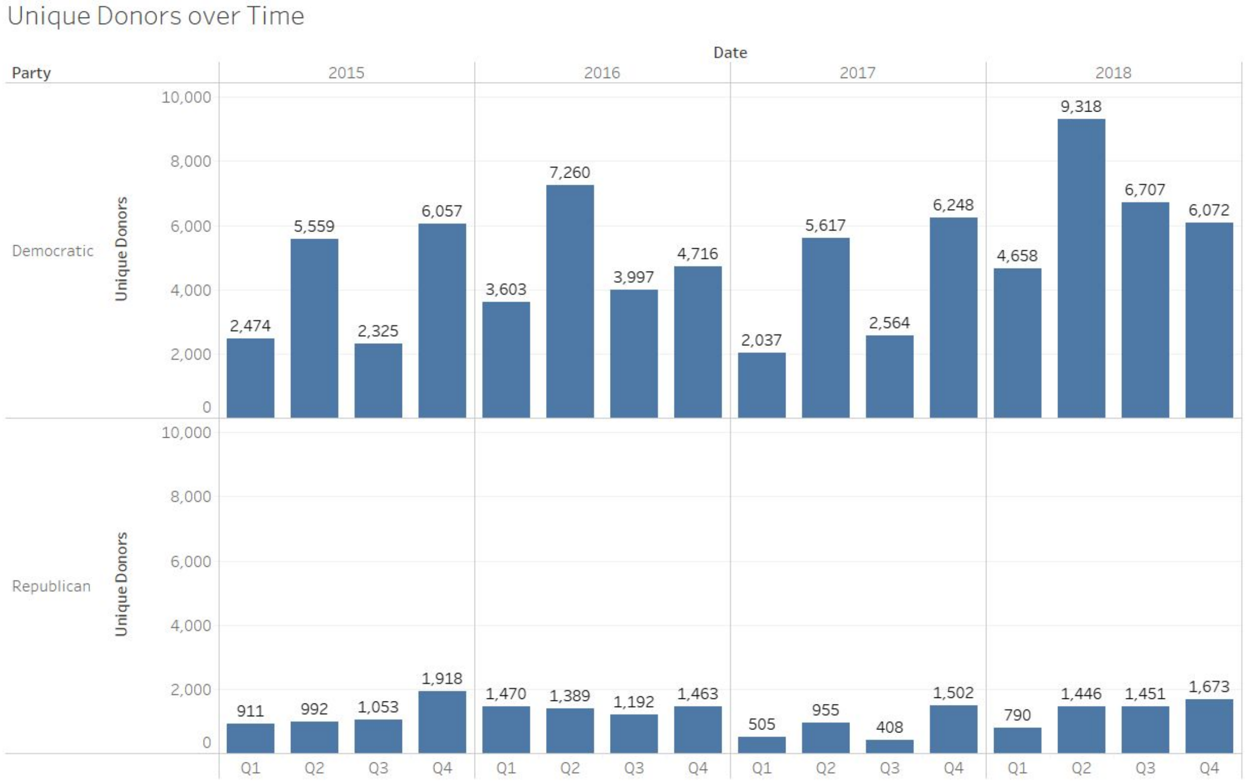
Question 1: Number of unique donors for all the quarters for the past 2 election cycles from 2015-2018



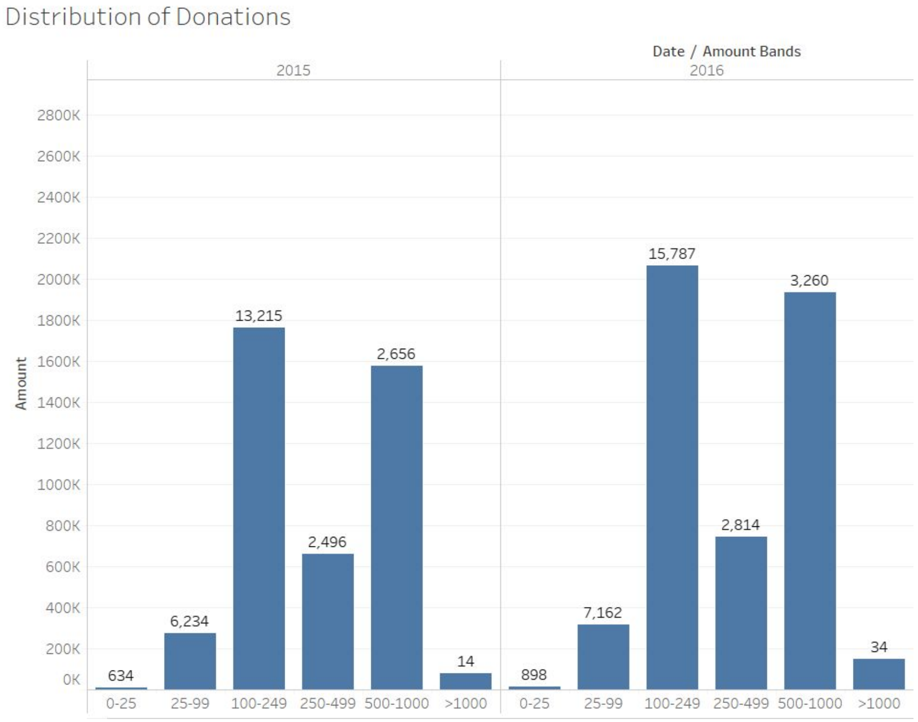
From this chart we can see that the number of unique donors tend to drop significantly in the year following an election, such as in 2015 and 2017. In those years, Q1 and Q3 have about half the number of unique donors as Q2 and Q4, and Q2 is typically 80% of Q4.

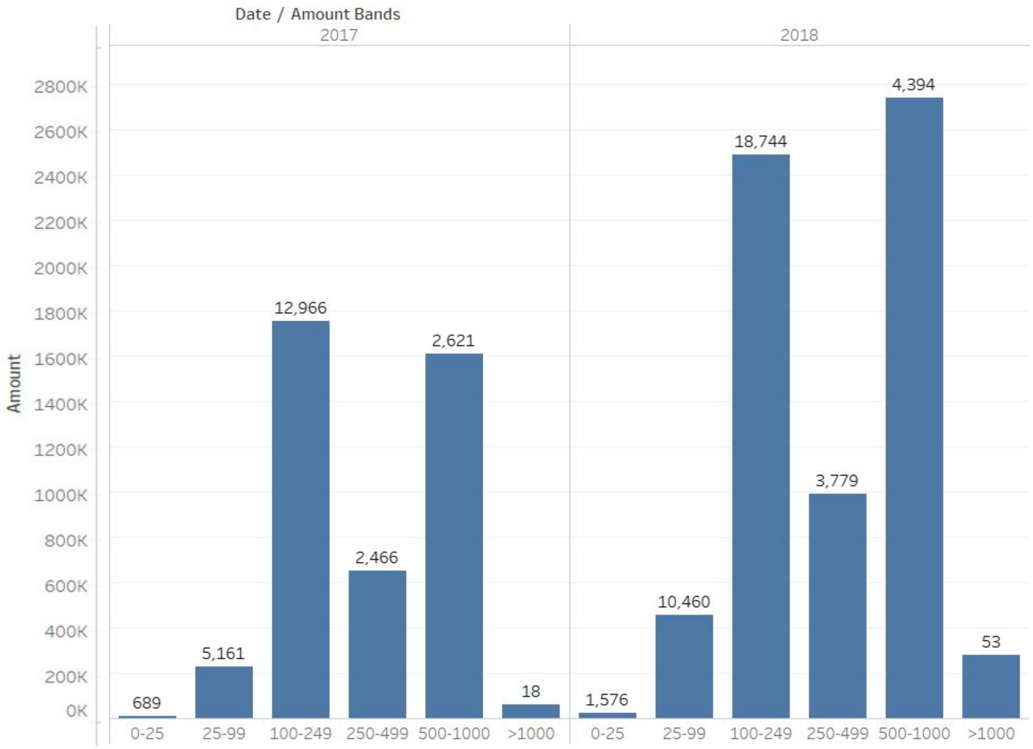
In the election years of 2016 and 2018, we also see that Q1 and Q3 tend to have about half the unique donors in Q2, and Q4 is about 70% of Q2.

The exception to these trends is Q3 2018 which had 8,101 unique donors, which is about 3,000 (or 50%) more than expected. If we further split the graph by party (refer next chart), we can see that this trend is mainly attributable to Democrat state reps although it is not linked to any single representative. Perhaps there were some events or bills passed around the quarter which caused more unique donors to contribute to Democrat state reps.



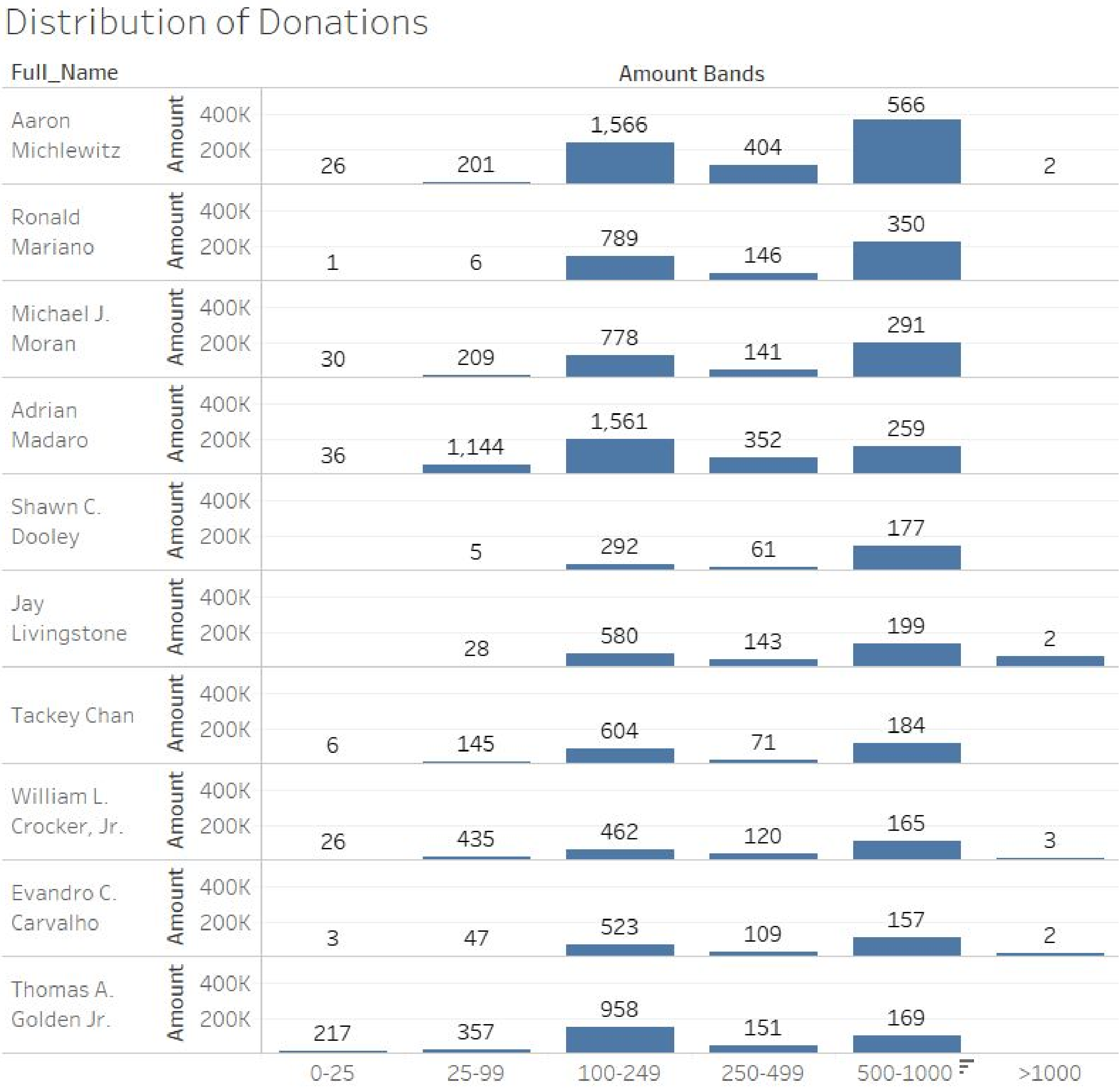
Question 2: Distribution of contributions across 5 contribution bands (<$25, $25-$99, $100-$249, $250-$499, $500-$1000)





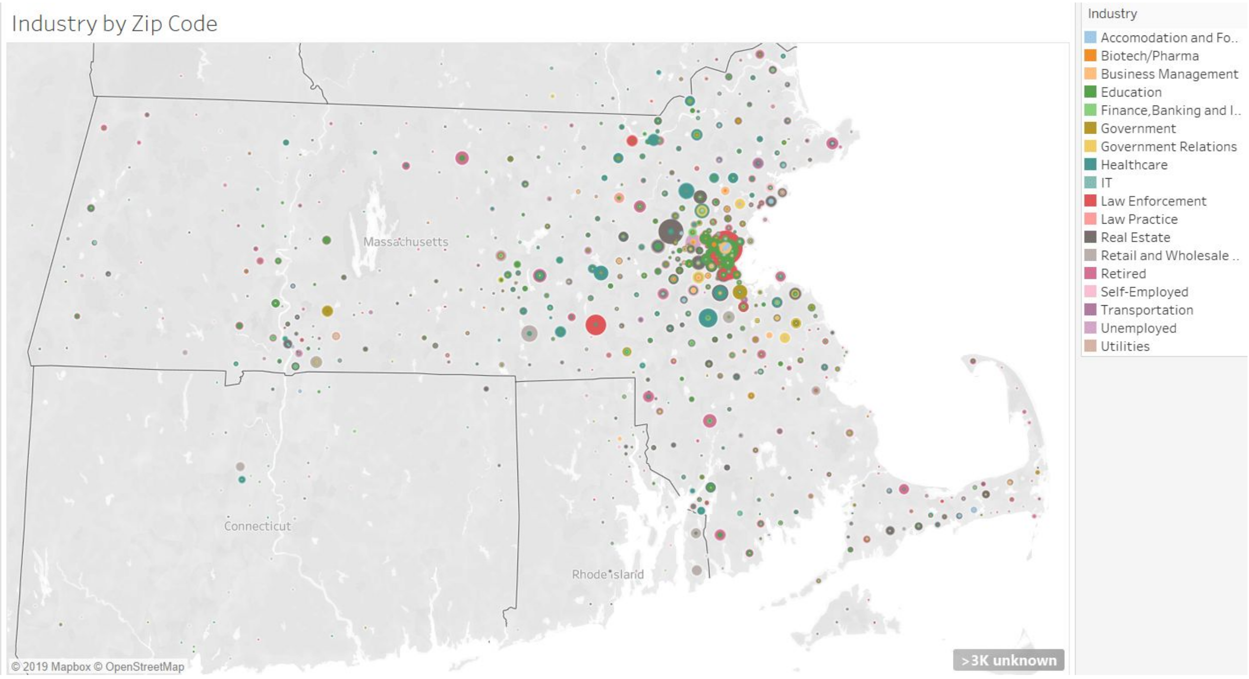
On the top of the bars it is number of donations.

And if we split by State Rep to see who has the top 10 most contributions in the highest donation band of $500-1000:

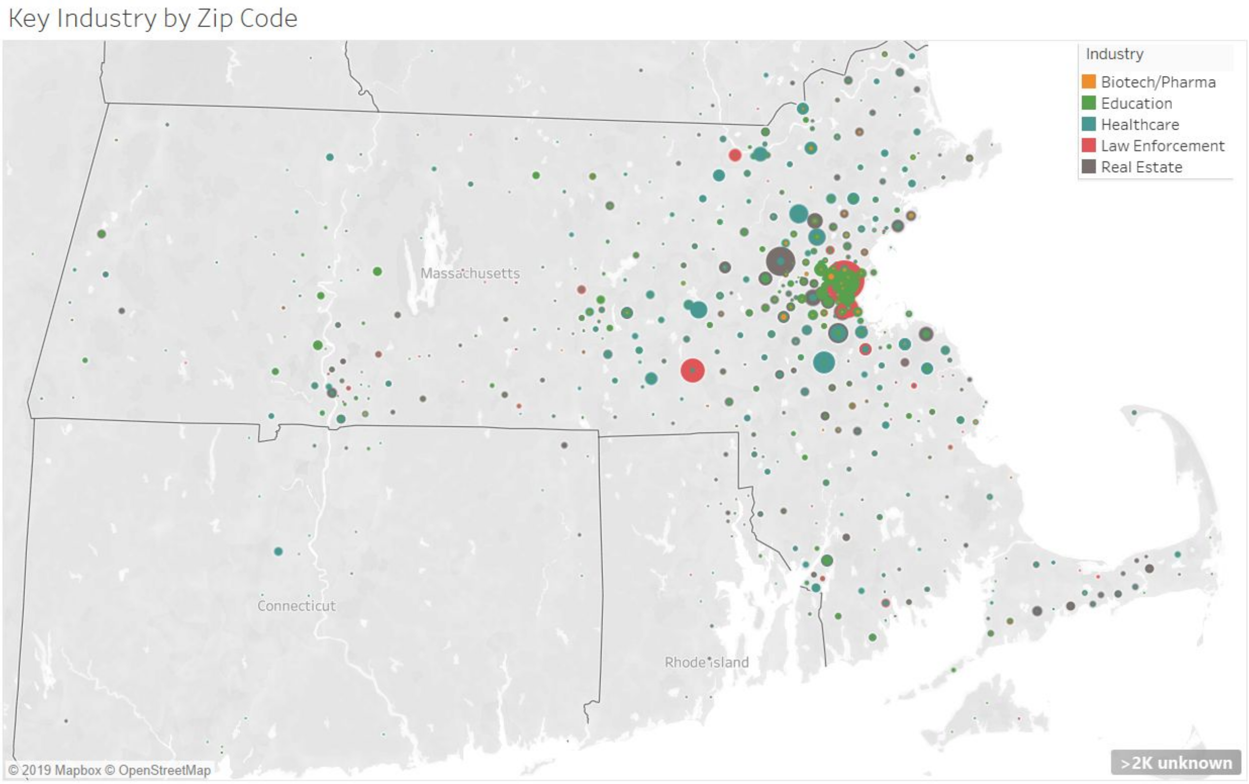


Question 3 Geographical distribution of donors (in-district, out of district, out of state)

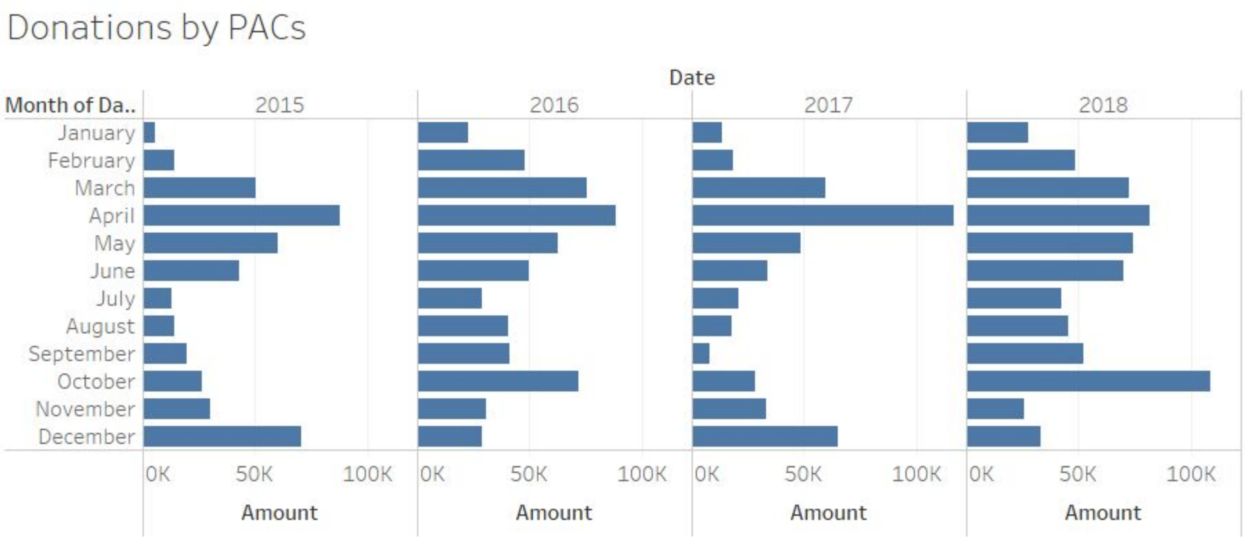
The following map shows the geographical distribution of contributions based on ZIP code. Each circle represents a contribution and the size is proportional to the contribution amount. Each circle is also color coded by the industry the donor is from.



And if we filter by the key industries of interest (Real Estate, Education, Healthcare and Law Enforcement, Biotech/Pharma):



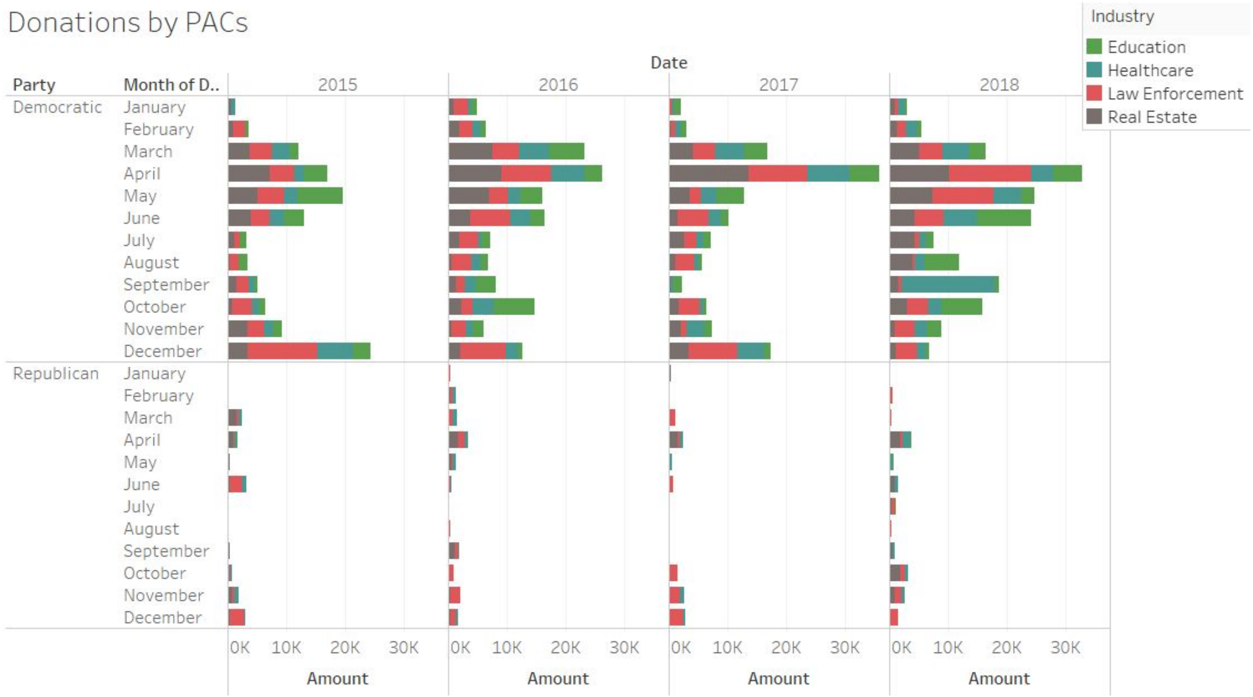
Question 4: Monthly donations by PACs (Political Action Committees)



From the chart above, we can see that from 2015 to 2018, PACs tend to follow a similar pattern of making increasing contributions from Jan till Apr, followed by a decreasing trend till Jul, followed by an increase until the end of the year. In the election years of 2016 and 2018, we can also observe that Nov and Dec contributions tend to decrease significantly after the elections.

An interesting anomaly happens in Apr 2017, where PAC contributions hit a high of $116K for Apr from a norm of $88K (30% over the norm). Another anomaly happens in Oct 2018 where PAC contributions hit a high of $109K for Oct from a norm of $71K (taking reference from previous election year Oct 2016, it is a 53% increase).

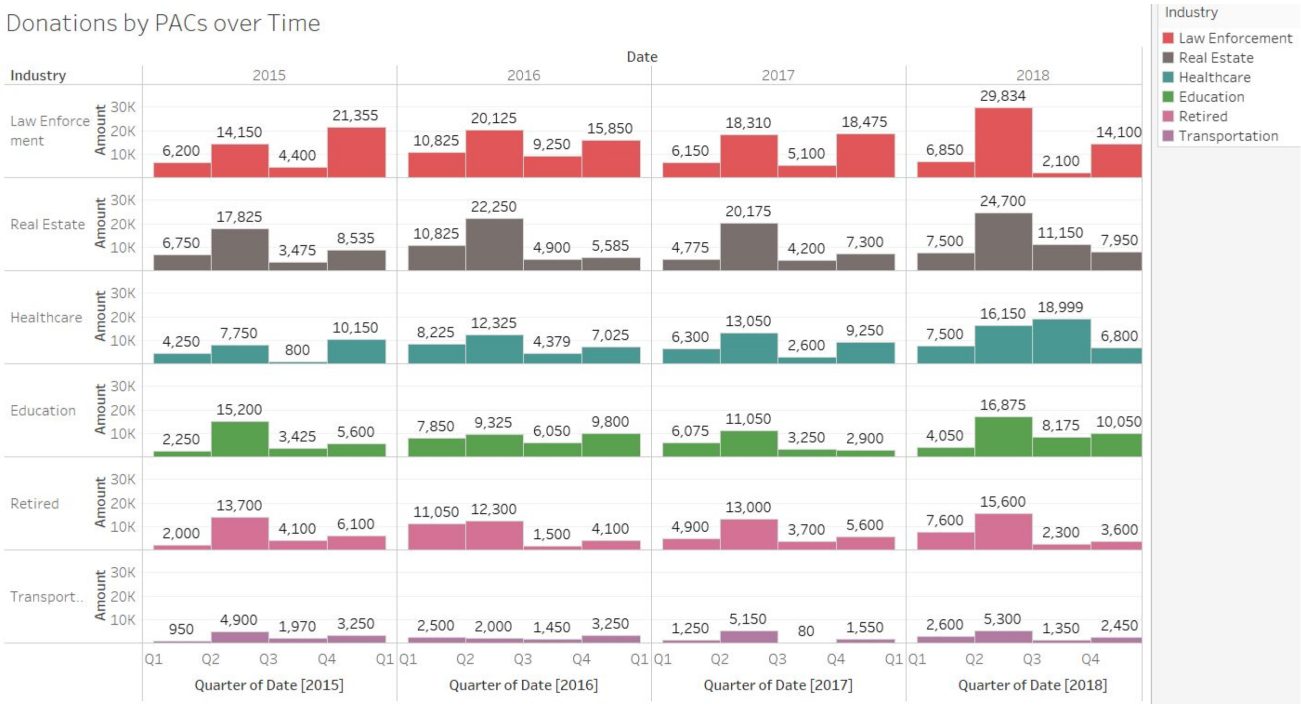
By further breaking down this chart by party and the key industries (healthcare, real estate, biotech/pharma, law enforcement, education) in the next chart, we can see that most PAC contributions go to Democrat state reps which is expected as majority of elected state reps are from the Democrat party. Also we can see that most of the PAC contributions come from outside of the key industries which we did not focus on. This could be an area to expand on for future work.



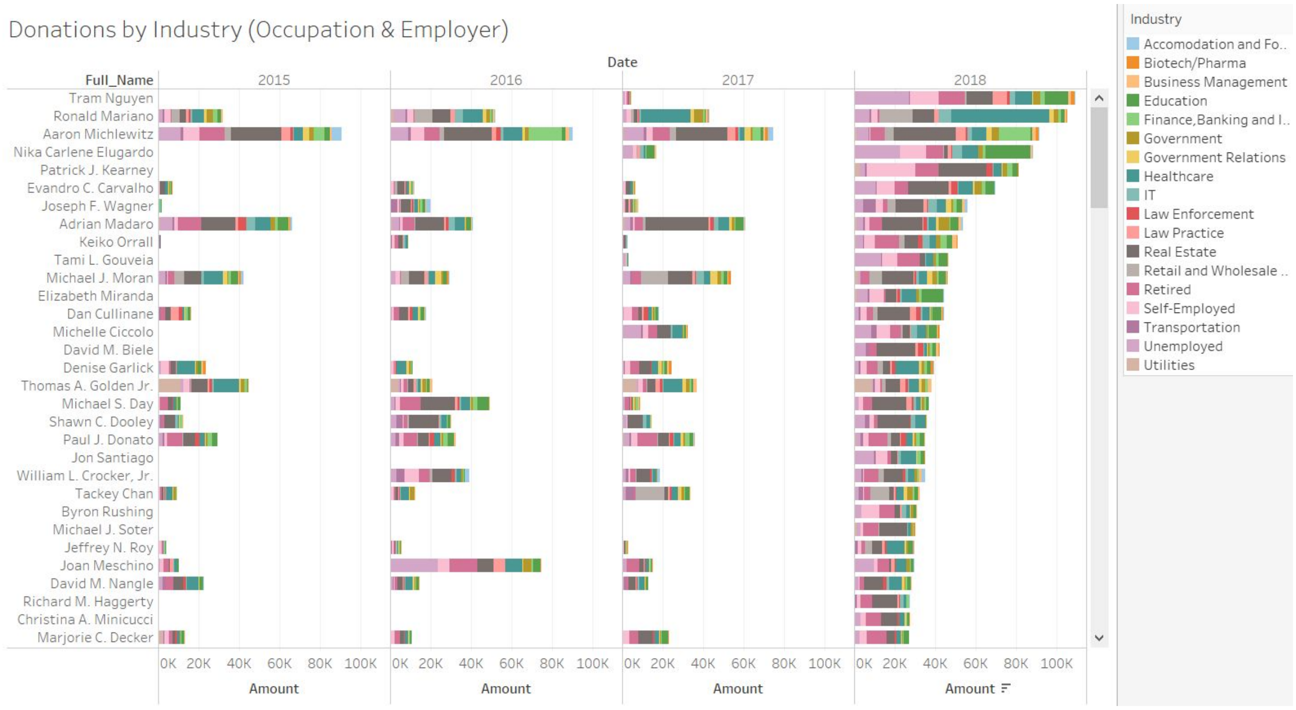
Question 5: PAC donations by industry sector

This graph shows PAC contributions over time for the key industries of:

1. Real Estate - including developers and construction.
2. Healthcare - including hospitals and health insurance.
3. Biotech/Pharmaceuticals (we did not find this category of PAC contribution)

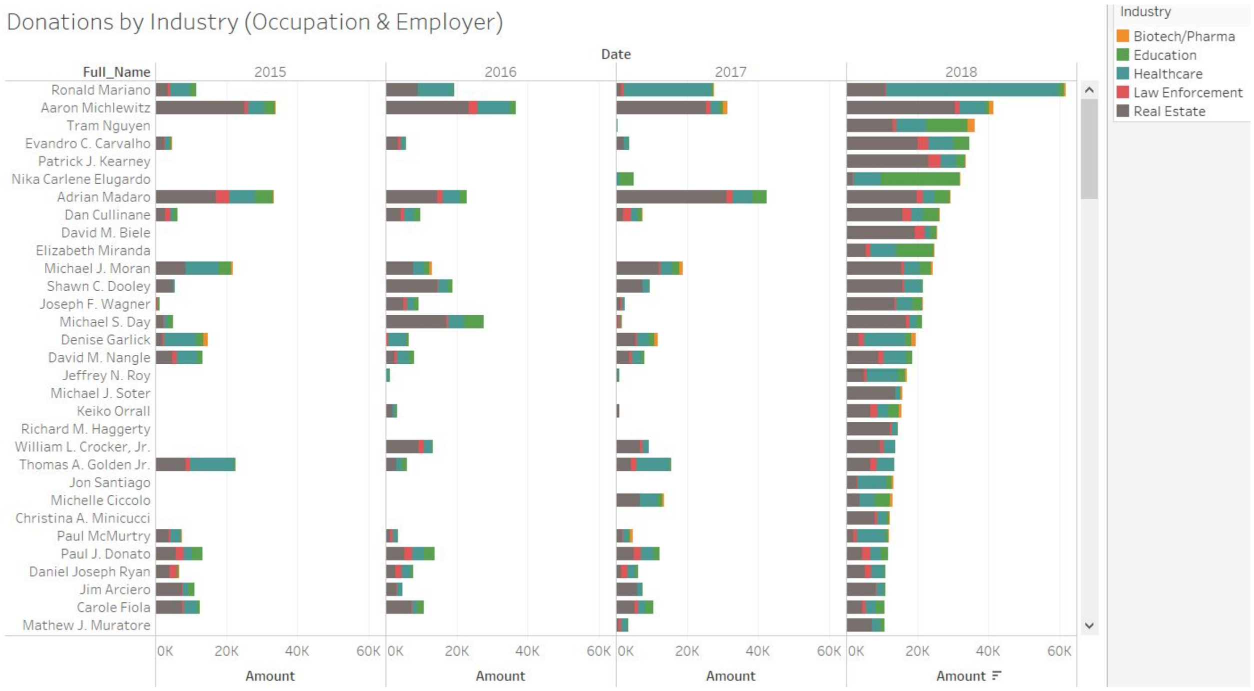


Question 6: Donations by Employer Types (excluding PACs)



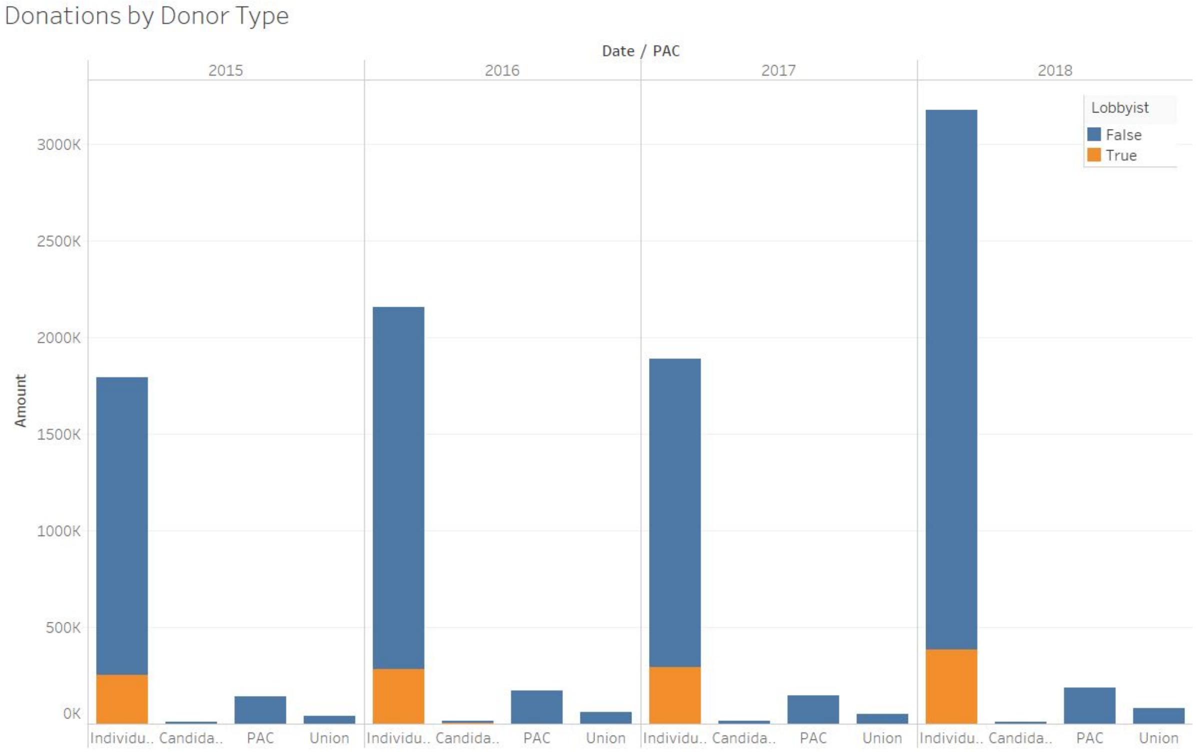
If we filter to get industries of interest:

a. Real Estate, Construction
b. Healthcare, Hospitals, Health Insurance Companies
c. Pharmaceuticals, Biotech Companies

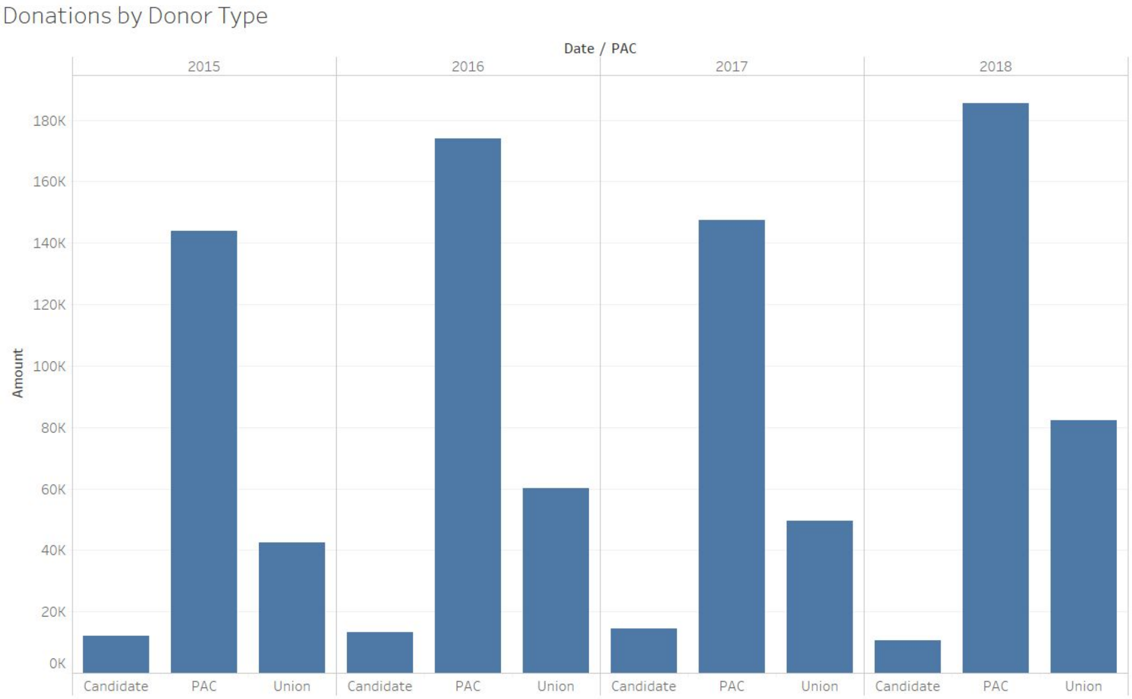


Question 7: Donation by Donor Type

- Donations from individuals, PACs, Unions, Lobbyists and from Candidates to themselves

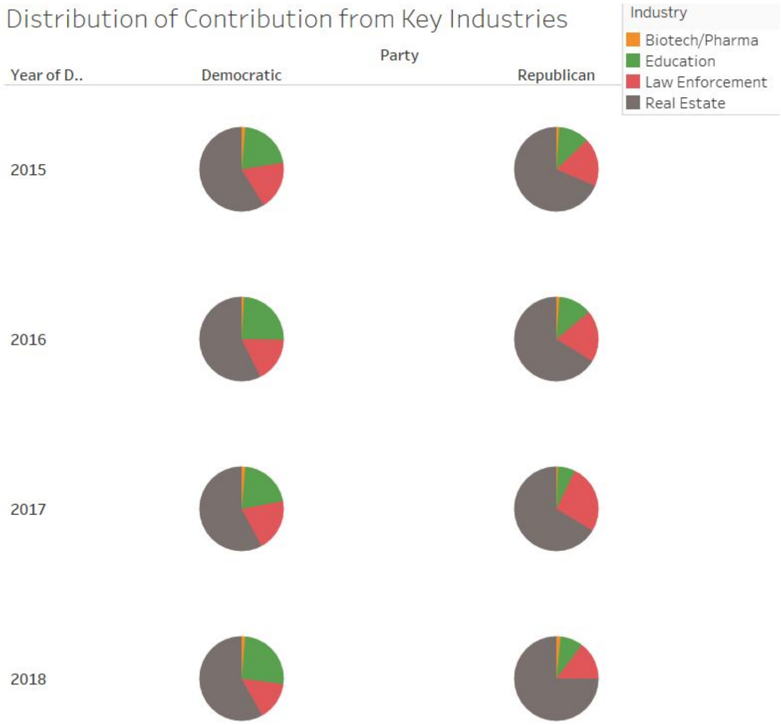


If we exclude Individuals to see the rest more clearly:

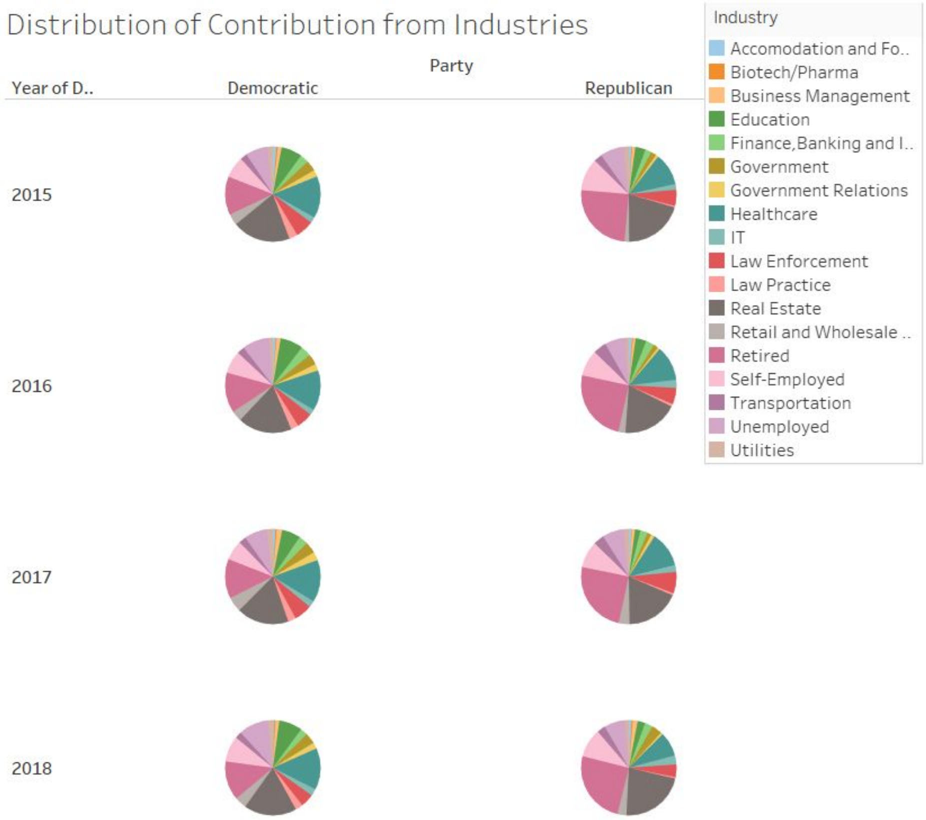


Proportion of Contribution from Different Industries

To see the impact of the key industries, we generated some additional visualisations. The first below shows the relative weight of the key industries based on contribution amount.



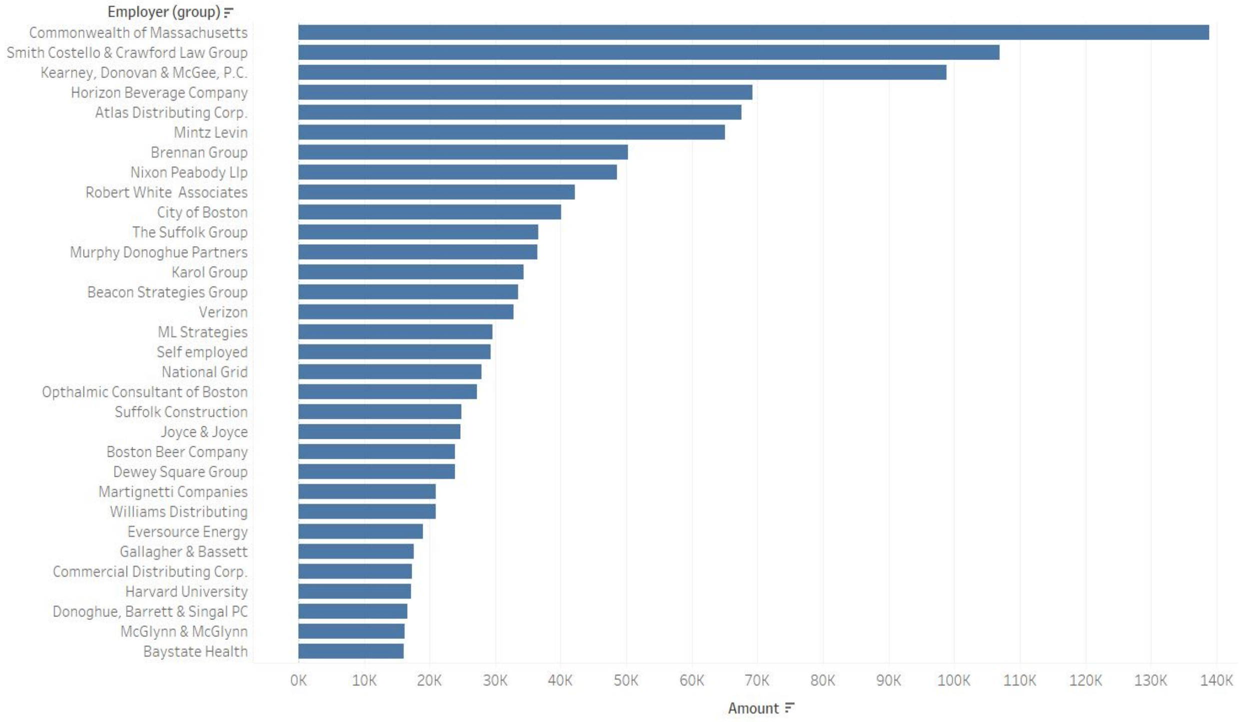
The next visualisation shows the relative weights of all industries we classified based on contribution amount. We can see that there are possible new areas of interest such as contributors who are Retired, Unemployed, Self-Employed or in the industries of Transport, and Retail & Wholesale. Some insights of interest can be garnered from the chart below, such as retired contributors seemed to contribute more to Republican state reps.



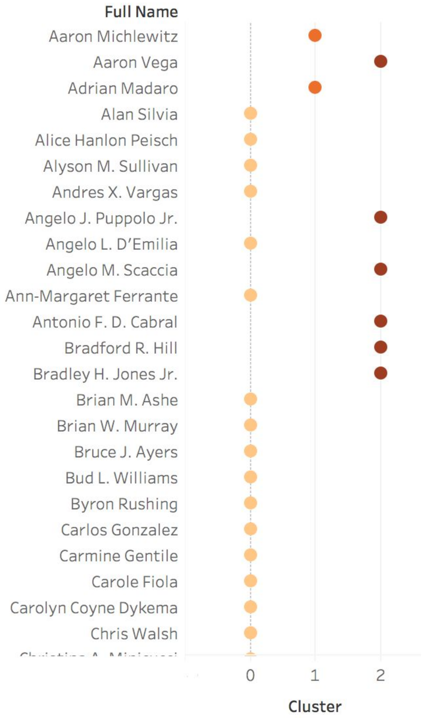
Top Companies in Contributions (as Employer)

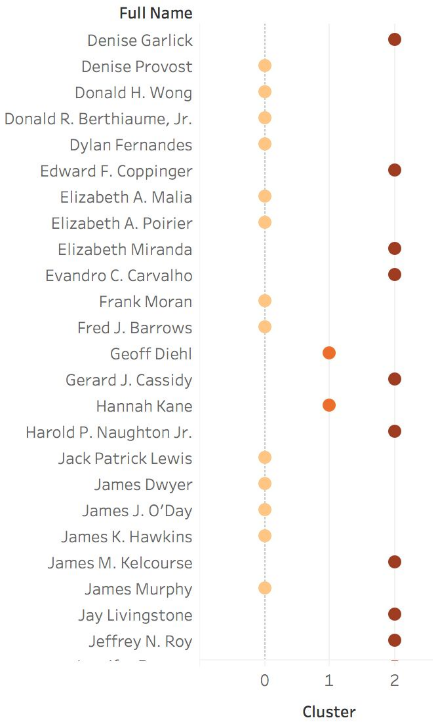
This graph shows the companies and the total amount of contributions their employees made to state reps. Of interest is that many of the companies who made the most amount of contributions are law firms or consultancies in the area of lobbying and/or government relations (e.g. Smith Costello & Crawford Law Group, Mintz Levin, etc.) . It may be of interest to discover what interests these companies are representing or if they are acting as intermediaries for another entity.

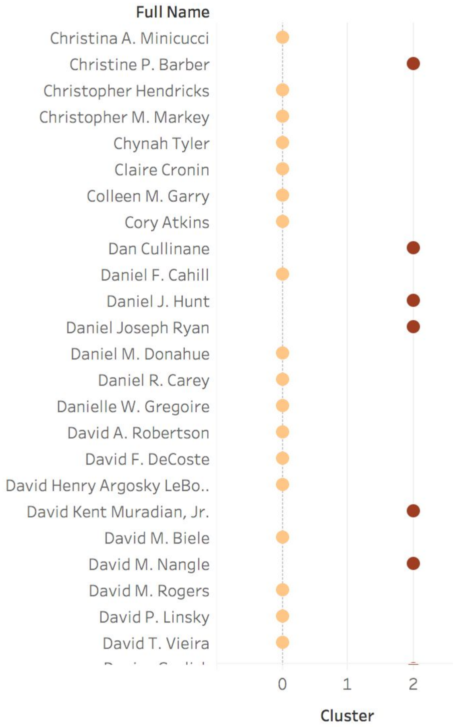
Top Company Contributors

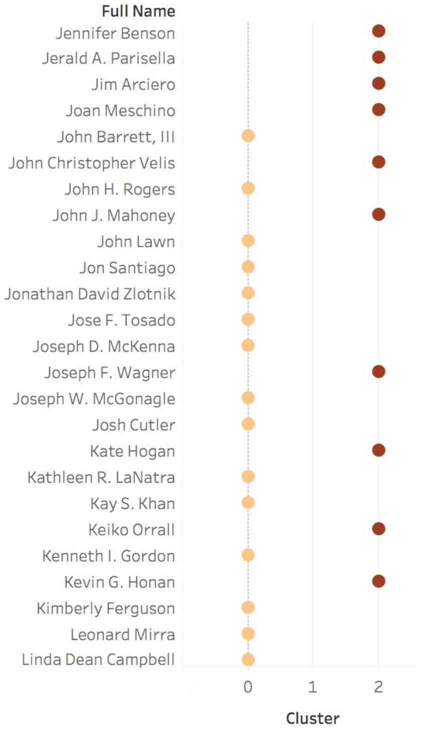


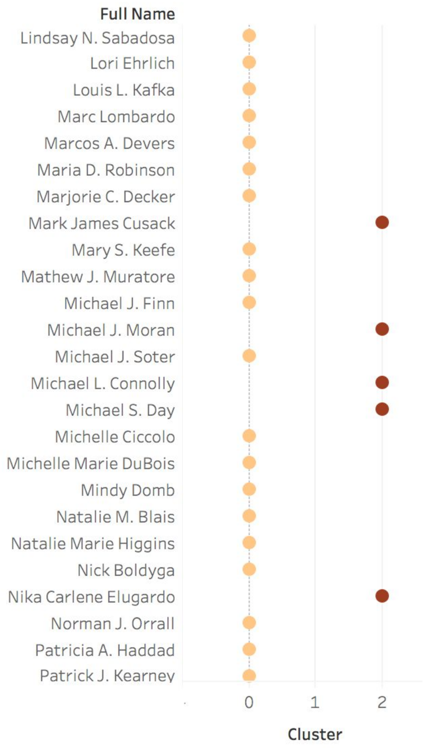
Clustering based on Number of Unique Donors over Time (by Quarters)

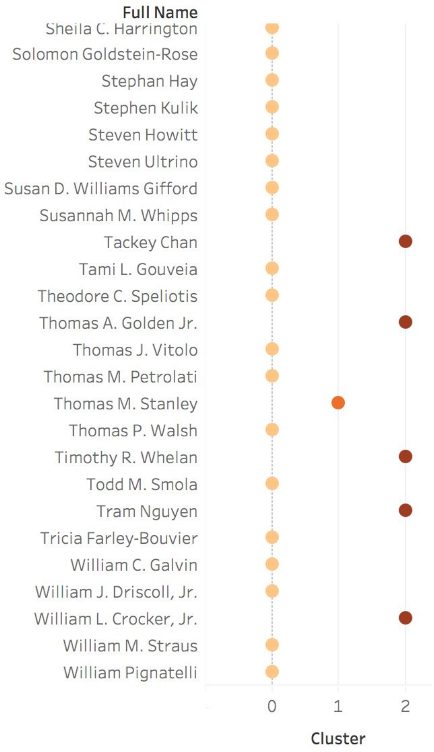


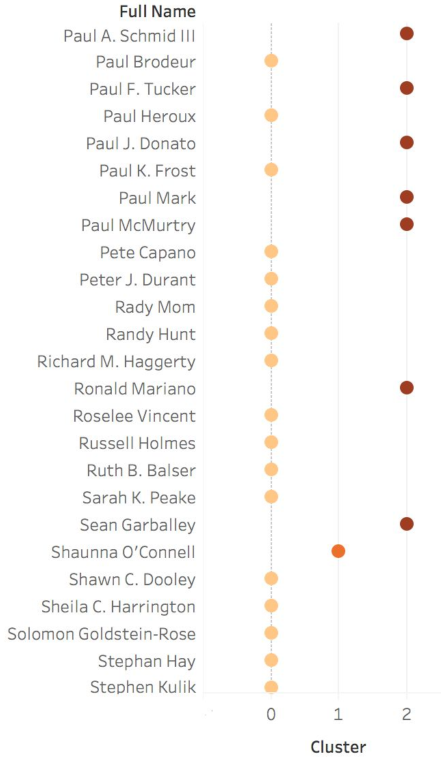







Full Name
Lindsay N. Sabadosa
Lori Ehrlich
Louis L. Kafka
Marc Lombardo
Marcos A. Devers
Maria D. Robinson
Marjorie C. Decker
Mark James Cusack
Mary S. Keefe
Mathew J. Muratore
Michael J. Finn
Michael J. Moran
Michael J. Soter
Michael L. Connolly
Michael S. Day
Michelle Ciccolo
Michelle Marie DuBois
Mindy Domb
Natalie M. Blais
Natalie Marie Higgins
Nick Boldyga
Nika Carlene Elugardo
Norman J. Orrall
Patricia A. Haddad
Patrick J. Kearney
0
1
2
Cluster


Full Name
Sheila C. Harrington
Solomon Goldstein-Rose
Stephan Hay
Stephen Kulik
Steven Howitt
Steven Ultrino
Susan D. Williams Gifford
Susannah M. Whipps
Tackey Chan
Tami L. Gouveia
Theodore C. Speliotis
Thomas A. Golden Jr.
Thomas J. Vitolo
Thomas M. Petrolati
Thomas M. Stanley
Thomas P. Walsh
Timothy R. Whelan
Todd M. Smola
Tram Nguyen
Tricia Farley-Bouvier
William C. Galvin
William J. Driscoll, Jr.
William L. Crocker, Jr.
William M. Straus
William Pignatelli
0
1
2
Cluster


Full Name
Paul A. Schmid III
Paul Brodeur
Paul F. Tucker
Paul Heroux
Paul J. Donato
Paul K. Frost
Paul Mark
Paul McMurtry
Pete Capano
Peter J. Durant
Rady Mom
Randy Hunt
Richard M. Haggerty
Ronald Mariano
Roselee Vincent
Russell Holmes
Ruth B. Balser
Sarah K. Peake
Sean Garballey
Shaunna O'Connell
Shawn C. Dooley
Sheila C. Harrington
Solomon Goldstein-Rose
Stephan Hay
Stephen Kulik
0
1
2
Cluster

We perform k-means clustering and used number of unique donors over quarters (2 election cycles) as a feature set per candidate. There are 3 clusters, because we assume candidate fundraising ability and popularity can be binned into three categories (high, medium, low). The candidates in the same cluster means they are similar to each other and different than candidates from other clusters.

Challenges Faced & Overcame

The main challenges we faced was in the preparation and enriching of data we have obtained so that we can obtain meaningful insights.

Firstly for the Representative data, only 50 out of 160 State Rep Candidate names in the OCPF data matched the names posted on MA Legislature webpage. There were abbreviations of middle names and informal spellings (e.g. Jim vs James, Gerry vs Gerard) of the remaining 109 names which had to be manually resolved. This was complicated as some politicians who received donations and have similar names (e.g. William F Galvin, MA Secretary of the Commonwealth, and William C Galvin, State Rep for 6 Norfolk District). We had to manually search and verify the data to resolve such cases.

Secondly was the classification of Industry of contributors. It was difficult to find a standardised set of Industry categories and a way to classify individual data records, because there were data that could fit in more than one category (e.g. should ambulance drivers be part of Healthcare or Transport? Law firms that deal with housing should be Law Practice or Real Estate?). We got around this by analysing the OCPF data and referencing the authoritative sources NAICS and Bureau of Labor Statistics. Furthermore, to get good matching results we had to iterate many times to adjust keywords and parameters which led to the development of our 2-tier classification method based on keyword pattern matching (words and root words characteristic of an industry) followed by fuzzy matching (against LinkedIn company data).

Conclusion - Lessons Learnt & Future Work

Based on our meetings with clients, we prioritised the data preparation and enrichment to answer the 7 questions. These questions focused on the key industries of real estate, healthcare, law enforcement, education, biotech/pharmaceuticals.

As we started doing the exploratory data visualisation and analysis, we realised that answering the 7 questions brought up more questions. For example, we discovered as shown in “proportion of contribution from different industries” that that there were other industries (e.g. transport) or types of contributors (e.g. retired) which made significant contributions to state reps. We suggest that if they are of interest, perhaps they could be included in the future work.

Thus far we have been tracking contributions made directly to a state rep. However in our exploration of the data as can be seen in the “top companies in contribution” chart, we discovered that many of the individuals making contributions are employees of either law firms and/or consultancies in the area of lobbying or government relations. It may be interesting to give more scrutiny to this area and, more generally, to trace the flow of money between entities (e.g. between law firms, consultancies, PACs, unions, candidates).

Through working on this project we have gained experience in dealing with real world data problems such as data preparation (resolving names of state reps from different sources) and classification (identifying the industry of the contributor). We have also realised the importance of having domain knowledge of the data, in this case it was very useful to know the cycle of campaign/political finance and how it works.

Lastly, it has been an enjoyable and rewarding experience working on this project and we would like to thank our clients Johnathan and Steve from ProgressiveMass, Professor Lance and teaching staff, and Director Ziba and BU Spark staff for their expertise and patience in guiding us through this project.