Campaign Finance City Council and State Representative Scorecard

Yuang Liu (U99473611), Jun Xiao(U85900288), Qi Yin (U31787103), Yiming Miao(U76158067)

Department of Computer Science

Boston University

yaliu@bu.edu

Contents

1 Introduction

Our partener 'Progressive Massachusetts' is a statewide, member-driven grassroots organization built from the ground up by organizers and activists from across Massachusetts to advocate for progressive policy. We believe in a Massachusetts where social, racial, and economic justice; environmental sustainability; health care as a right; equal access to quality public services; respect for all residents; and accountable and transparent government are given top priority. And we work to make that vision a reality. Progressive Mass welcomes all progressives regardless of political affiliation.

Last year the students examined publicly available campaign contributions data to city council members and categorized them by industry using Bureau of Labor Statistics' Occupational Outlook Handbook and the NAICS' industry keywords to classify donor industries. They then created a scorecard for each city council member and classified donations by industry, donation size, and employment type. This year they want to focus specifically on donations from specific industries.

Our team is assigned to analyze the donations from different industries but will mainly focus on the real estate industry donations to the city committees. Besides, we'd like to answer some other strategy questions, including:

1. What is the average donation ($) to all state representatives by industry?
2. Which state representatives are the outliers i.e. larger volume of donations received? How much larger?
3. Is there a pattern or insights of these individuals? (E.g. time of year, donating to multiple candidates from the same companies, etc.)
4. Do members of the committees most relevant to a given industry receive a significantly larger volume of donations?
5. How much are PACs and Unions giving (look at this alone as well as by industry)?
6. Who are the largest industry donors? (include PACs, Unions, Companies, Individuals)
7. What percent of Reps overall receive donations from the same firm by industry. (E.g. 60% of reps receive from RE firms who make donations, while only 25% receive from Law Enf donating firms)

2 Data Collection

For this project, most analysis are based on four source of data: 'keywords for industry and pac', the OCPF (Massachusetts Office of Campaign and Political Finance) data, the Annual Report from Sec of State website (http://corp.sec.state.ma.us/) and city committees names from the Boston Government website (https://www.boston.gov/departments/city-council)

Both of the first two data (keywords and OCPF) are provided by partners via google drive.

The Annual Reports are scraped from the Sec of State website for all companies listed on that website. We finished the scraping job by using Selenium in python to emulate web-browser actions and Tor to rotate IP preventing be blocked by the website.

The names for city committees are simply copied from the Boston Government website.

What's more, as a backup plan, we also scrape the City of Boston website (https://www.cityofboston.gov/cityclerk/dbasearch/Default.aspx) to collect companies names related to real estate in case of we cannot scrape the Sec of State website.

The Sec of State website has a very strict and very robust anti-scraper mechanism: 1, each chrome controlled by selenium can only request up to 150 times on that website and then will meet anti-scraper robot bans. 2, the website will track the IP address, if one IP has a very frequent request on the website, it will simply denied and further request from that IP. (Actually, the IP from my apartment has still been banned even after nearly 1 month). To overcome the first anti-scraper strategy, we first download links for companies on the website and then use those links to download Annual Report to make sure each time we only request no more than 120 times. To handle the second strategy, we not only use Tor web browser to rotate my IP but also use VPN to generate virtual IP address. Besides the two final solutions, we actually tried many other methods to scrape the Sec of State such as changing

web browser, using virtual environment to generate virtual IP but none of them work. With the two final solutions, another obstacle of downloading the pdf is the limited Internet speed comparing with nearly 0.5 million links. We can only process up to 20 links per minute, and the whole downloading work finished on three computers for the whole 8 days and nights.

3 Data Preparation and Cleaning

3.1 OCPF data

1. Preparation

   The OCPF data is provided in Microsoft Access Database format, so the first thing we did was install a Windows Operating System on Virtual Box to manipulate the database.

   To extract relative entries from the database, we use pypyodbc package to establish connection between python and Access database. There are 10 tables in the OCPF database but we only use two of them: vUPLOAD_tCURRENT_RECEIPTS and vUPLOAD_MASTER.

   For this project we only focus on the donations from 2016-01-01 to 2019-12-31, so we only extract data within this period "where date$\geq$'2016-01-01' and date$\leq$'2019-12-31'". And not all columns in these two tables are useful, the same, we only extract the info we need, the needed info from the OCPF dataset are shown in Table 1 and 2.

Table 1: vUPLOAD_MASTER

Table	vupload_master	
Info	information of the receiver	
Attributes	report_id	the identifier for distinct report
	cpf_id	the identifier for distinct donation receiver
	full_name	the name of the receiver

Table 2: vUPLOAD_tCURRENT_RECEIPTS

Table	vupload_tcurrent_receipt	
Info	information of the contributor	
Attributes	report_id	which report this entry is belonging to
	date	when the contribution happened
	first_name	the first name of the contributor
	last_name	the last name of the contributor
	city	the city of the contributor
	state	the state of the contributor
	occupation	the occupation of the contributor
	employer	the company that the contributor works in
	amount	how many money does the contributor contribute

2. Clean
   Because the occupation, employer, full_name, first_name, last_name attributes of the OCPF data contain letters both in uppercase and lowercase, we general change all the letters in the uppercase. And there are some employer in different format but with the same meaning, such as 'AT HOME' and 'HOMEMAKER'; 'SELF', 'SELF EMPLOYED' and 'SELF-EMPLOYED' etc. We simply changed those words into the same form, however we can only clean a small portion of employer with different format because we can only implement this after noticing the different by ourselves.

3.2 Keywords for employer and pac

1. Preparation
   Simply copied the keywords from the google doc to local text file and csv file.
2. Clean
   In the keywords for employer, we change all the letters to uppercase. And some of the keywords for pac are the same but in different format such as "Carpenters Local #108 Pol Action Comm" and "Carpenters Local #40 Pol Action Comm", we simply change them into the same format in analyzing section and keep the different during data extraction.

3.3 Annual Report from Sec of State

1. Preparation
   Use pdfplumber and pdfminer3k packaged to extract information from the Annual Report.
2. Clean
   We simply discard the Annual Report in unreadable format such as image, scanned document etc.

4 Analysis

For this project, we should count the amount of contribution from different industries to different city committees. The analysis goes in a very straightforward way, we simply extract the data relating with one specific industry by filtering the occupation, employer's name that contains the keywords from the keywords list or the last name of the contributor is in the list of the pca. Then, we accumulate the data by the cpf_id, which is the identifier for each city committees. Also, we count the amount of donations by different year, season, city, state and the size of the donations.

5 Presentation

5.1 Results Analysis

Among all the OCPF donations data, the 'VERIZON NEW ENG-LAND INC.' company donates 157478 times with 313080 dollars to 'Int'l Brotherhood of Electrical Workers Local Union 2222 Pol Action Comm', which is the most times one donors donates to one specific receiver.

Besides SELF-EMPLOYED people and RETIRED people, the 'COMMONWEALTH OF MASSACHUSETTS' has donated to 473 different receivers, which is the most number of different receivers one donor has donated.

Besides SELF-EMPLOYED people and RETIRED people, the 'THE BAUPOST GROUP LLC' has donated 3979550.0 dollars, which is the largest total amount of the donations from one company. But when we look at average donations, the 'ARVEST BANK GROUP' has the largest average donation which is 1125000.0 dollars on average.

We first look into if the position (chairman or member) of a city committees is relating with the donation amount he/she receive or not. We found that the higher the committee's position in the corresponding field, the more inclined it is for him/her to receive more donations in this field.

Then we analyze the distribution of donations from Real Estate field over different months. We found that the change of donation amount from Real Estate industry has a very similar pattern with the change of the firsthand and secondhand house price in the Great Boston Area.

5.2 Reproduce Results

5.2.1 Scraper for the Sec of State part

i. Run **get_all_links.py** script to extract all the companies links for a specific letter, which you should specify in the python script. The script will save all the links to a txt file, where you should specify the name of the txt file such as **links_for_z.txt**

ii. Run **get_all_pdf.py** script to download all the annual reports for the companies. You have to specify the txt file storing the links for companies. The script runs very slow, it will take you one hour to download only 1,200 to 1,500 links. Sometimes, you have to use VPN to change your IP, for that the Sec of State website has a very strict anti-scraper management. (Actually, it takes us 8 whole days to download all pdfs for companies from N-Z on three computers).

iii. Run **analyze_pdf.py** script to get the companies name and corresponding business type. And run **analyze_pdf_extra_info.py** to get some extra information such as Street Address, Registered Agent, Street Address of Corporation Principal Office, Names and Addresses of Board Members or Executives.

Up to here, we finish downloading and analyzing all companies for one specific letter from the Sec of State website. If you want to download pdfs for other letters. Just go through all the above steps again.

5.2.2 Download city of boston data

Just run the **scrapy_for_city_of_boston.py** script, it will collect all the companies relating to the real estate field based on the keywords. Actually, this scraper is only a backup solution for fear that we are not able to handle the Sec of State website.

5.2.3 Analyzing OCPF data

i. Use **filter_ocpf.py** script to filter the ocpf data to get all the raw data from 01-01-2016 to 12-31-2019. The intermediate data will be stored in the file you specify in the script.

ii. Use **get_ocpf_by_pcas.py** to get ocpf data only relating to PCA. The intermediate data will be stored in the file you specify in the script.

5.2.4 Instructions of APIs

After the two steps above, we've already got the raw data to analyze, then run the **analyze_ocpf_data_with_keywords.py** script. There are many APIs in this script, here are some instructions over them:

(1) **get_filtered_data_xxx**: extract data relating with a specific field 'xxx' based on keywords

(2) **compute_amount_by_cpfid**: count the amount of money a receiver received and sort by the amount from lowest to highest

(3) **compute_amount_by_season**: count the amount of donations based on season

(4) **compute_amount_by_year**: count the amount of donations based on the year

(5) **computer_amount_by_size**: count the amount of donations based on the size of the donations (under $25, from $25-$50 etc)

(6) **computer_amount_by_city**: count the amount of donations based on city

(7) **computer_amount_by_state**: count the amount of donations based on state

(8) **computer_amount_by_contributor_type**: count the amount of donations based on contributor's type (individual or committees)

(9) **the_most**: get the top 10 companies donator regarding with total amount, average amount per donation and total numbers of donation

(10) **companies_insight**: see which companies donate the most number of different receivers and which companies donate the most time to one receiver

All steps above are the way to generate the result data. And the visualization jobs are mostly done with the help of Microsoft Excel.

5.3 Database

All the datasets are shown as follows:

5.3.1 OCPF data

https://drive.google.com/drive/folders/19Ne8561l4XdHYnpyJpWHziV0G74dKp2n?usp=sharing

5.3.2 Annual report for companies from N-Z

https://drive.google.com/drive/folders/1kMET_SQKGrKh81KeOt1xhzUta8-mpucM?usp=sharing

5.3.3 Business type for companies from N-Z

https://drive.google.com/drive/folders/12DA-eT82_1iDsWZdHoLKoA6Y2l1ty_Ub?usp=sharing

5.3.4 All info for companies from N-Z

https://drive.google.com/drive/folders/1SYcDUuOsn0aOQ-lFodipP0vDwmGNwcIF?usp=sharing

5.3.5 Info from City of Boston

https://drive.google.com/drive/folders/1CjNdmvsTaEkKzN_BylggM-eJoXcp1D0F?usp=sharing

5.4 Results and Visualizations

5.4.1 City Committees

ANDREA CAMPBELL(15931)

Real Estate

WAYS AND MEANS (MEMBER)

Law Enforcement

PUBLIC SAFETY AND CRIMINAL JUSTICE (COMMITTEE CHAIR)

Higher Education

EDUCATION (VICE CHAIR)

ANNISSA ESSAIBI GEORGE(15618)

Real Estate

WAYS AND MEANS (VICE CHAIR)

Law Enforcement

PUBLIC SAFETY AND CRIMINAL JUSTICE (MEMBER)

CIVIL RIGHTS (MEMBER)

Healthcare

STRONG WOMEN, FAMILIES, AND COMMUNITIES (MEMBER)

Higher Education

EDUCATION (COMMITTEE CHAIR)

ED FLYNN(14391)

Real Estate

WAYS AND MEANS (MEMBER)

HOUSING AND COMMUNITY DEVELOPMENT (MEMBER)

Law Enforcement

PUBLIC SAFETY AND CRIMINAL JUSTICE (MEMBER)

CIVIL RIGHTS (VICE CHAIR)

FRANK BAKER(15333)

Real Estate

PLANNING, DEVELOPMENT, AND TRANSPORTATION (VICE CHAIR)

WAYS AND MEANS (MEMBER)

HOUSING AND COMMUNITY DEVELOPMENT (MEMBER)

JULIA MEJIA(17092)

Real Estate

WAYS AND MEANS (MEMBER)

Law Enforcement

CIVIL RIGHTS (COMMITTEE CHAIR)

Healthcare

STRONG WOMEN, FAMILIES, AND COMMUNITIES (MEMBER)

KENZIE BOK(17164)

Real Estate

WAYS AND MEANS (COMMITTEE CHAIR)

PLANNING, DEVELOPMENT, AND TRANSPORTATION (MEMBER)

HOUSING AND COMMUNITY DEVELOPMENT (VICE CHAIR)

Energy

ENVIRONMENT, RESILIENCY, AND PARKS (MEMBER)

KIM JANEY(16537)

Higher Education

EDUCATION (MEMBER)

LIZ BREADON(17111)

Real Estate

PLANNING, DEVELOPMENT, AND TRANSPORTATION (MEMBER)

HOUSING AND COMMUNITY DEVELOPMENT (MEMBER)

Energy

ENVIRONMENT, RESILIENCY, AND PARKS (MEMBER)

Healthcare

STRONG WOMEN, FAMILIES, AND COMMUNITIES (COMMITTEE CHAIR)

LYDIA EDWARDS(16314)

Real Estate

PLANNING, DEVELOPMENT, AND TRANSPORTATION (MEMBER)

HOUSING AND COMMUNITY DEVELOPMENT (COMMITTEE CHAIR)

Energy

ENVIRONMENT, RESILIENCY, AND PARKS (MEMBER)

Higher Education

EDUCATION (MEMBER)

MATT O'MALLEY(14092)

Law Enforcement

PUBLIC SAFETY AND CRIMINAL JUSTICE (MEMBER)

CIVIL RIGHTS (MEMBER)

Energy

ENVIRONMENT, RESILIENCY, AND PARKS (COMMITTEE CHAIR)

Healthcare

STRONG WOMEN, FAMILIES, AND COMMUNITIES (VICE CHAIR)

Higher Education

EDUCATION (MEMBER)

MICHAEL FLAHERTY(12892)

Real Estate

WAYS AND MEANS (MEMBER)

Law Enforcement

PUBLIC SAFETY AND CRIMINAL JUSTICE (VICE CHAIR)

Higher Education

EDUCATION (MEMBER)

MICHELLE WU(15563)

Real Estate

PLANNING, DEVELOPMENT, AND TRANSPORTATION (COMMITTEE CHAIR)

Law Enforcement

CIVIL RIGHTS (MEMBER)

Energy

ENVIRONMENT, RESILIENCY, AND PARKS (VICE CHAIR)

Healthcare

STRONG WOMEN, FAMILIES, AND COMMUNITIES (MEMBER)

RICARDO ARROYO(17105)

Real Estate

PLANNING, DEVELOPMENT, AND TRANSPORTATION (MEMBER)

Law Enforcement

PUBLIC SAFETY AND CRIMINAL JUSTICE (MEMBER)

CIVIL RIGHTS (MEMBER)

Higher Education

EDUCATION (MEMBER)

5.4.2 Results

1. Distribution of Donations to City Committees These pie charts show the distribution of donations from different industry for each city committees listed above, the total amount is the sum of individual donations and PAC donations for that some industries are lack of PAC data.

   The pie charts are shown in Figure 1.

   According to the charts, the higher the committee's position in the corresponding field, the more inclined it is for him/her to receive more donations in this field. Generally speaking, the committees are more likely to receive donations from real estate, higher education, and healthcare. Besides, the amount of donation is not closely related to the committees' positions as we imagined before. It may be only related to the committee himself/herself, since the committee who received more donation from one category may also receive more from other categories.

2. A Deeper Look into the Real Estate Field
   (a) The distribution of donations from different PACs for each city committees are shown in Figure 2.
   (b) Compute by Time
   The result is shown in Figure 3.
   Based on the actual situation, continued shortage in the supply of homes for sale in Great Boston appreciated home values. According to the chart, the donation is extremely high on June 2017, and relatively low at January and February of 2019. In June 2017, housing prices in the Great Boston area increased significantly over the same period in previous years*, and there was also a shortage of supply. In this environment, more people tended to donate to real estate-related industries, hoping that the housing problem can be effectively alleviated.
   As for January 2019, despite increased inventory levels and a decline in mortgage rates since fall 2018, sales of detached single-family homes and condominium fell by double-digit percentages during January in Great Boston*. Although there is still a shortage of supply, due to factors such as stock market adjustments, housing prices in the Boston fell in January 2019. The increase in housing listings has made the industry environment more favorable to buyers, so relatively few donations were received in the real estate sector during the month. Since then, in February, despite the rebound in the real estate market, the overall situation has not changed, and there has been no greater fluctuation in the number of donations in related fields.
   Generally speaking, the donation in the real estate industry is more sensitive to the buyer's attitude. When the buyer hopes that the real estate market will be adjusted, the corresponding donation amount will also increase.
   **Data from the Greater Boston Association of REALTORS® (GBAR).*
   (c) Compute by Season
   The result is shown in Table 3 and Figure 4.
   (d) Compute by Year
   The result is shown in Table 4 and Figure 5.

Figure 1: Distribution of Donations to City Committees

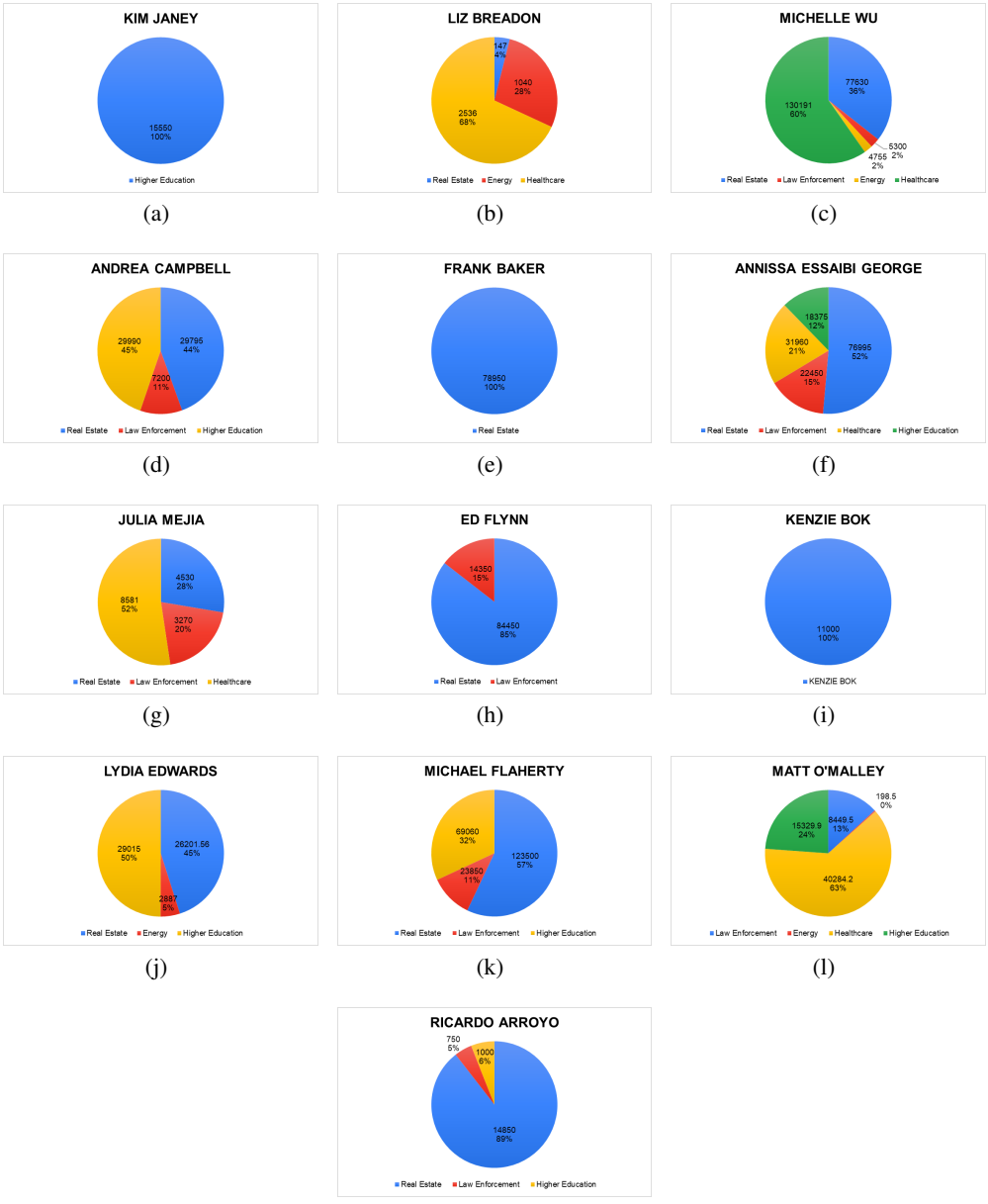
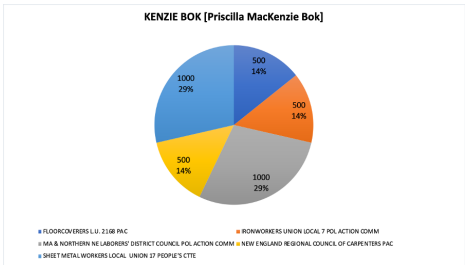
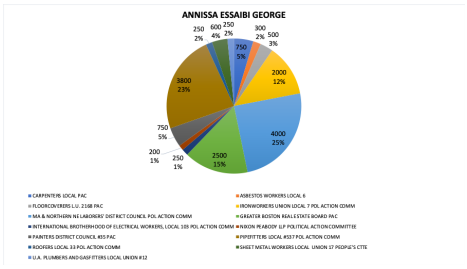


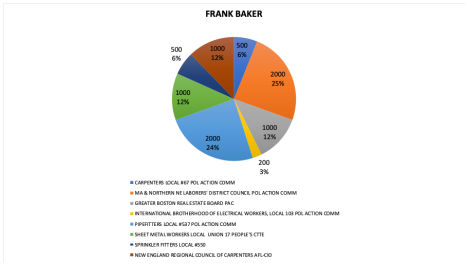
Figure 2: Distribution of donations from different PACs for each city committees



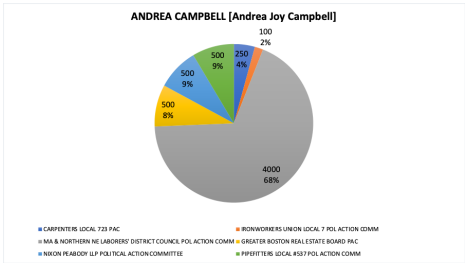
(a)



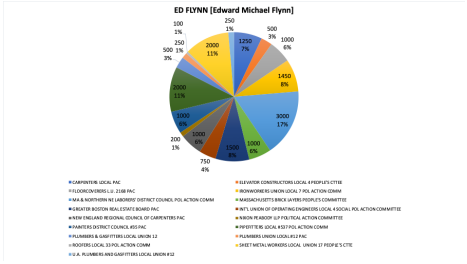
(b)



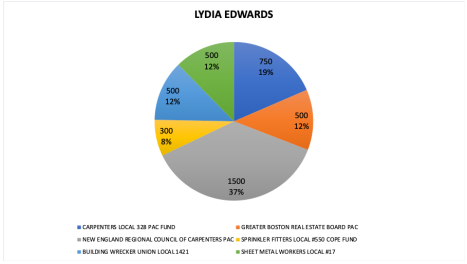
(c)



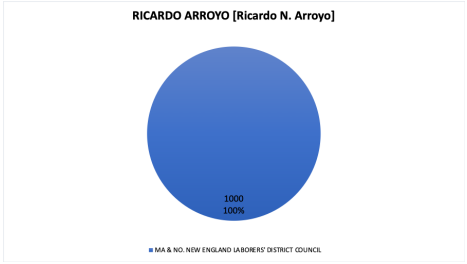
(d)



(e)



(f)



(g)

Figure 3: Compute by Time

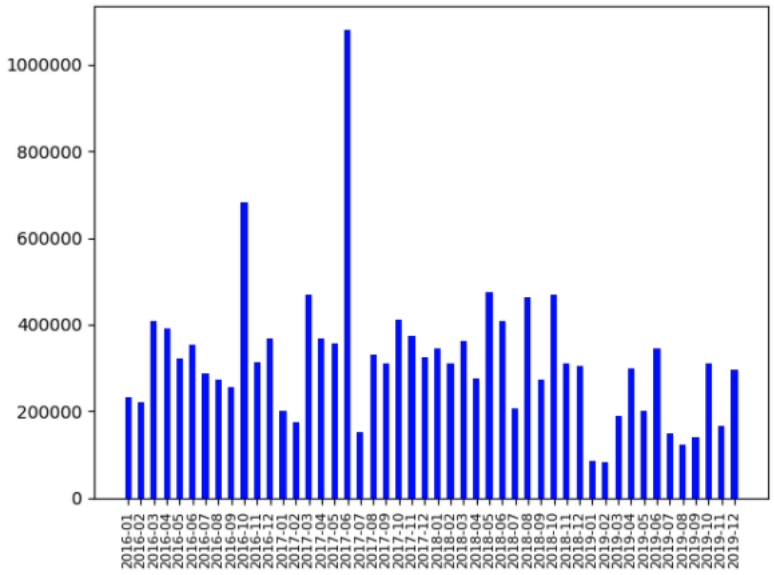


Table 3: Compute by Season

Season	Amount
1	$1134122.81
2	$1901463.20
3	$1347196.58
4	$2507028.23

Figure 4: Compute by Season

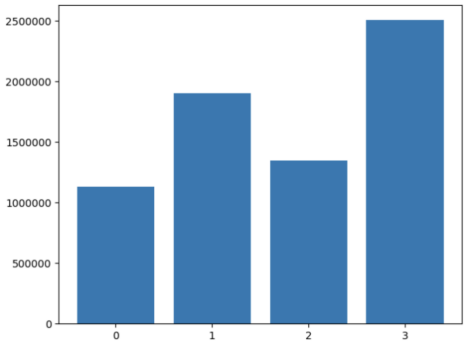
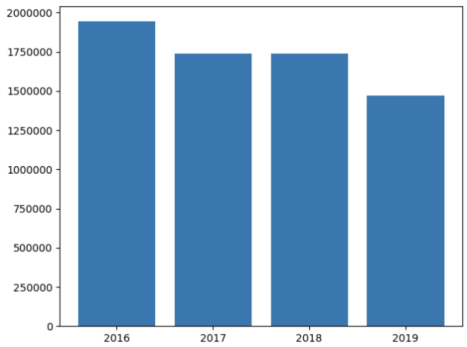


Table 4: Compute by Year

Year	Amount
2016	$1942829.81
2017	$1736838.23
2018	$1739569.30
2019	$1470573.48

Figure 5: Compute by Year

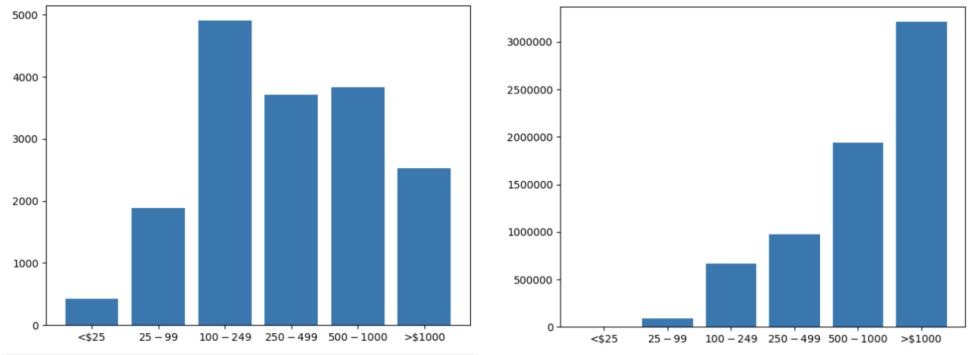


(e) Compute by Size
The result is shown in Table 5 and Figure 6.

Table 5: Compute by Size

Size	Number	Amount
$< \$25$	422	$4304.76
$\$25 - \$99$	1884	$86518.04
$\$100 - \$249$	4903	$669819.92
$\$250 - \$499$	3714	$976471.49
$\$500 - \$1000$	3833	$1942196.61
$> \$1000$	2530	$3210500.00

Figure 6: Compute by Size



(a) Number by Size

(b) Amount by Size

5.5 Answers to the Questions

1. What is the average donation ($) to all city committees by industry?
   The results are shown in Table 6.

   Table 6: The average donation ($) to all city committees by industry.

	Total	Average
Real estate	528048.6	356.1
Law Enforcement	85609.5	268.1
Higher Education	178319.9	303.5
Energy	8880.5	231.7
Healthcare	213552.2	195.6

2. Which city committees are the outliers i.e. larger volume of donations received? How much larger?
   The results are shown in Table 7.

Table 7: The outliers of city committees.

	People	Amount
Real estate	MICHAEL FLAHERTY	123500.0
Law Enforcement	ANNISSA ESSAIBI GEORGE	22450.0
Higher Education	MICHAEL FLAHERTY	69060.0
Energy	MICHELLE WU	4755.0
Healthcare	MICHELLE WU	130191.0

3. Is there a pattern or insights of these individuals? E.g. time of year, donating to multiple candidates from the same companies, etc.)
   (a) Here list (companies, donations receivers, how many times, the amount of donations), and shows the amount the companies donated and how many times they donated to one specific receiver. We only list donation times more than 2,000, and the results are shown in Table 8.

Table 8: The amount the companies donated and how many times they donated to one specific receiver.

Company	Donation Receiver	Time	Amount
VERIZON NEW ENGLAND INC.	Int'l Brotherhood of Electrical Workers Local Union 2222 Pol Action Comm	157478	313080
MASS PCA - CP MASS	1199 SEIU MA PAC	16540	138808.1
VERIZON SERVICES CORP.	Int'l Brotherhood of Electrical Workers Local Union 2222 Pol Action Comm	14149	20423.5
MASS PCA - STAVROS	1199 SEIU MA PAC	12673	97343.7
STATE OF MA	Local 509 Service Employees Int'l Union Comm on Pol Ed MA Workers' Pol Action Comm.	12284	33302.6
MASS PCA - NORTH SHORE	1199 SEIU MA PAC	7301	56961.8
STATE OF MASSACHUSETTS	Local 509 Service Employees Int'l Union Comm on Pol Ed MA Workers' Pol Action Comm.	6947	18864.8
RETIRED	Charles D. Baker	4085	1098682.2
VERIZON CORPORATE SVCS. CORP	Int'l Brotherhood of Electrical Workers Local Union 2222 Pol Action Comm	3858	5055
MCOFU	MA Correction Officers Federated Union PAC M.C.O.F.U PAC	3298	14611.8
STOP AND SHOP	Chapter 25 Associated the Nat'l DRIVE PAC of the Int'l Brotherhood of Teamsters	2620	21483.8
CITY OF BOSTON	Martin J. Walsh	2194	463308.2

   (b) Here list (companies name, number of different receivers), and shows each company donated to how many receivers. Here only list companies that donate to more than 200 different receivers and the results are shown in Table 9.

Table 9: Each company donated to how many receivers.

Company	Number of Different Receivers
SELF-EMPLOYED	1221
RETIRED	1029
COMMONWEALTH OF MASSACHUSETTS	473
NOT EMPLOYED	393
N/A	256
HARVARD UNIVERSITY	218
UNEMPLOYED	210
CITY OF BOSTON	210

4. Do members of the committees most relevant to a given industry receive a significantly larger volume of donations? The analysis and results are shown in Part 3.2.1.
5. How much are PACs and Unions giving (look at this alone as well as by industry)? The results are shown in Table 10.

Table 10: The amount PACs and Unions give.

	PAC	Individual
Real estate	103350.0	424698.6
Law Enforcement	12650.0	72969.5
Higher Education	N/A	178319.9
Energy	N/A	8880.5
Healthcare	200.0	213352.2

6. Who are the largest industry donors (include PACs, Unions, Companies, Individuals)
   (a) Top 10 companies in total amount are shown in Table 11

Table 11: Top 10 companies in total amount.

Name	Amount
SELF-EMPLOYED	9905121.9
RETIRED	8104532.1
THE BAUPOST GROUP LLC	3979550.0
HIGHFIELDS CAPITAL MANAGEMENT LP	2865350.0
NOT EMPLOYED	2537793.7
BAIN CAPITAL	1429300.0
ARVEST BANK GROUP	1125000.0
LAS VEGAS SANDS CORPORATION	1000000.0
CITY OF BOSTON	859949.6
BLOOOMBERG LP	740000.0

   (b) Top 10 companies in average amount are shown in Table 12.

Table 12: Top 10 companies in average amount.

Name	Amount
ARVEST BANK GROUP	1125000.0
LAS VEGAS SANDS CORPORATION	1000000.0
ACTION NOW INIATIVE LLC	510000.0
HIGHFIELDS CAPITAL MANAGEMENT LP	260486.3
DE SHAW & CO	250000.0
CENTAURUS ADVISORS	250000.0
PAR CAPITAL MANAGEMENT INC.	250000.0
BLOOOMBERG LP	246666.7
THE BAUPOST GROUP LLC	234091.2
REGENT ABLE ASSOCIATE	200000.0

(c) Top 10 companies in total times are shown in Table 13.

Table 13: Top 10 companies in total times.

Name	Times
VERIZON NEW ENGLAND INC.	157495
RETIRED	32545
SELF-EMPLOYED	25486
MASS PCA - CP MASS	216540
VERIZON SERVICES CORP.	14149
NOT EMPLOYED	13963
MASS PCA - STAVROS	12674
STATE OF MA	12366
MASS PCA - NORTH SHORE	7301
STATE OF MASSACHUSETTS	7076

7. What percent of Reps overall receive donations from the same firm by industry. (eg, 60% of reps receive from RE firms who make donations, while only 25% receive from Law Enf donating firms)

   The results are shown in Table 14.

Table 14: Percent of Reps overall receive donations from the same firm by industry.

	Total	Committees Received	Portion
Real estate	6889810.8	528048.6	7.66%
Law Enforcement	2077230.3	85609.5	4.12%
Higher Education	11340422.3	178319.9	1.57%
Energy	763041.1	8880.5	1.16%
Healthcare	16801269.5	213352.2	1.57%

5.6 Supplementation

1. Limitation

   Although the Sec of State website is very hard to scrape, the data from the website is not compatible with the OCPF data very well, we can only extract 968 data entries from the OCPF if we use the companies from Sec of State relating with real estate, in contrary, we can extract more than 16,000 data only with the keywords.

2. Assumptions

   We assume that the keywords for pcas are comprehensive enough. However, actually it only contains the name of the pca, but don't have the information relating with the companies name belonging to that pca.

3. Further Analyses

   We can go deep into the Annual Report from Sec of State website to find some potential contributor in the future. And moreover, we can analyze the location of the contributions to see is there some area has a denser contribution than other area and is there some area has abnormal contributes behaviors such as suddenly spur in the donation size.

4. Data Sources

   We can go to the website of each pca and to download all the companies name belonging to that pca. Then all the contributions from those companies will be counted as contributions from pca instead of from individual.

6 Project Complexity and Team Creativity

1. Challenges and obstacles

   The Sec of State website has a very robust and very strict anti-scraper mechanism.

   Scraping the Sec of State website is really time consuming, it takes nearly one and a half month to do this.

   The data from the Sec of State is not compatible with the OCPF data very well. With exact math, we can only get less than 1,000 data entries from the OCPF with the companies names from Sec of State. Even with fuzzy match, the number of data entries is still limited.

   The COVID-19 situation has forced us cancel 4 meetings continuously and introduced a lots of troubles for the group meeting. It's very inefficiently for working together online over zoom.

2. Team Creativity

   Our creativity is mostly relating with the scraper. The Sec of State website does not allow more than 150 consecutive requests and will ban an IP if the IP has abnormal behaviors. To make the scraper workers expectedly, our team made such modifications:

   (a) Instead of running the scraper for a specific character to download the Annual Report, we first analyze the web page containing all links for different companies to extract all the companies' link and save them into a text file.

   (b) Then we will use those links to download the Annual Report separately: for each time, we only use one chrome web browser controlled by Selenium to request 120 times with an 0.5 seconds interval and after those requests we will create another chrome web browser by Selenium to continue the download.

   (c) Actually, if we analyze the link to the Annual Report directly to get the PDF information, we will get nothing for it is in the ASPX format. To handle this, instead of directly analyzing, we change one attribute of chrome web browser to let it downloads PDF automatically when clicking on some links.

   (d) During the download, we will use Tor web browser and VPN to generate a virtual IP address.