# Campaign Finance City Council And State Representative Scorecard

**Aaron Jacob Varghese - aaronjv@bu.edu, U60942775**
**Raj Vipani - rvipani@bu.edu, U11275347**
**Sun Diwen - jack123@bu.edu, U37462503**

**Link to data -**
**https://drive.google.com/drive/folders/1E0cY1iGYHMwfR30olp5J8CM-kXOnEg6F?usp=sharing**

**All code is uploaded to github project repo**

## Project Background

In this project we aim to analyze the donations made to the city and state representatives of Massachusetts for the years 2016 - 2019. Out of the two teams working on the same project, we divided the work and our team is analyzing the state representatives data, specifically, the donations they received from the healthcare industry. Owned by “Progressive Mass”, a state-wide member driven grassroots organization built by organizers and activists regardless of political affiliations from across Massachusetts to advocate for progressive policy, this project aims at answering questions and correlating specific industries for instance healthcare, Law enforcement and real-estate to the donations being contributed by people of these industries.

Our goal in this project was to continue the work done last semester, specifically focusing on the healthcare industry and the funds transferred between entities that operate in this industry. We were also focused on analyzing the funds received by state representatives.
Some of the questions we seeked to answer were -

- How much each representative receives.
- How the amount they receive corresponds to their power i.e. are they on special committees where they wield a lot of power?
- Is there a time of year of years where the donated amount goes up?
- How much do the Political Action Committees and Unions donate and to whom?

Being a continuation of the same project from last semester, we started by analyzing and verifying the results done by the previous semester team. We continued by scraping the Secretary of State website, collecting data from the OCPF website and the data from Boston Businesses. Furthermore, we cleaned the collected data and after the preliminary analysis we are working forward on a deeper analysis of the data acquired and producing visualizations of the consolidated results.

# Approach Adopted

- **Verification of Results**

  As a continued project, we started with looking at the results produced by the team from last semester. This included running some of their code scripts, inspecting the datasets collected by them, and analyzing the reports produced. This step gave us a detailed insight into our project and helped us develop a sense of understanding for the approach to adapt.

- **Data Collection**

  For our project we used the data collected from the following three sources:

  i. **OCPF Website Data** - This data was provided to us by the clients from ProgressiveMass in the form of Microsoft Access databases. This dataset contains details of donors and the candidate they are donating to along with other details such as date of donations, report id and address of the donor.
  ii. **Secretary of State Data** - This website has records of every business entity operating within Massachusetts, which includes annual reports of all active entities. Our goal was to download the annual report PDF's which contained details like the type of business, executives and addresses. This collected data was later used as keywords to filter data from the OCPF dataset.
  iii. **Boston Businesses** - This was a website that contained the listings of businesses in the city of Boston, filtered by neighborhood. Our goal was to scrape the site for the details of every listed company.

- **Data Cleaning and Merging**

  After clearing the datasets of missing values we merged them in the following fashion:

  i. **Merging of Master And Receipts Table in OCPF Database** - We worked on the analysis of the OCPF database by merging the Master and Receipts tables in the OCPF database on the basis of the "Report ID", the id assigned to the individual donors after the donations made.
  ii. **Use Secretary of State website PDF's to get business types and use that to augment the filtering process in OCPF data** - We went through each of the downloaded PDF files to obtain the details of each active business entity. Most importantly, we needed the type of business, which could be used to filter through the OCPF data, as our team was focused on the healthcare industry.

- **Data Analysis**

  i. **Preliminary Analysis of OCPF data** - For the preliminary analysis of the OCPF data, we created histogram charts of the overall contributions made to candidates per year and analysed the donation ranges.

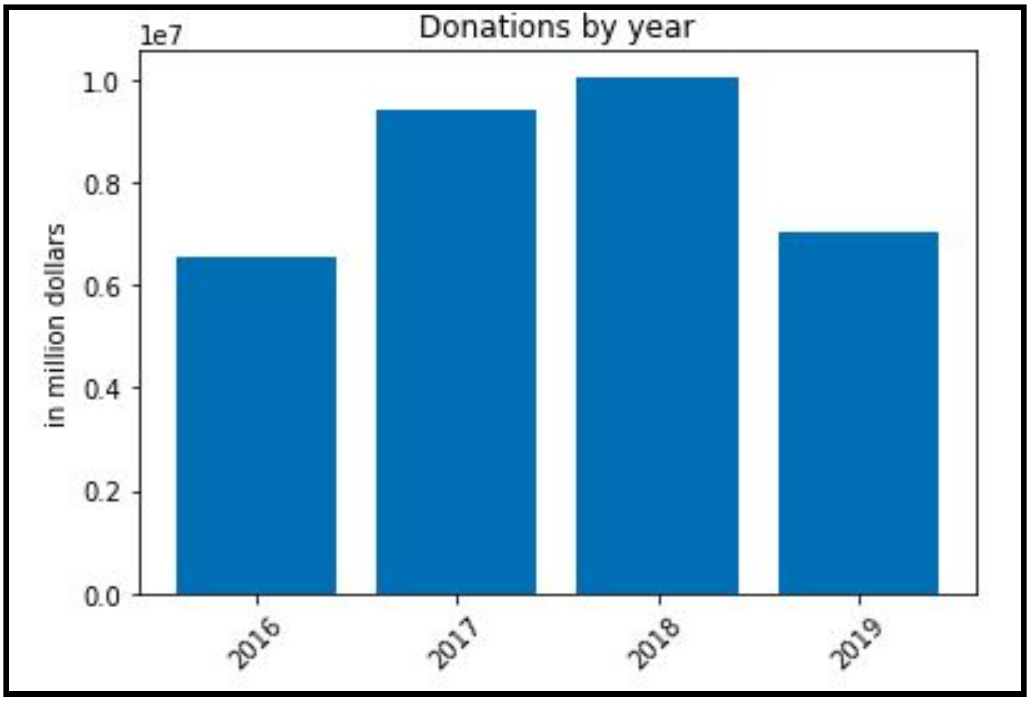



**Fig: Total amount of donations (in millions) made between the years 2016 - 2019**

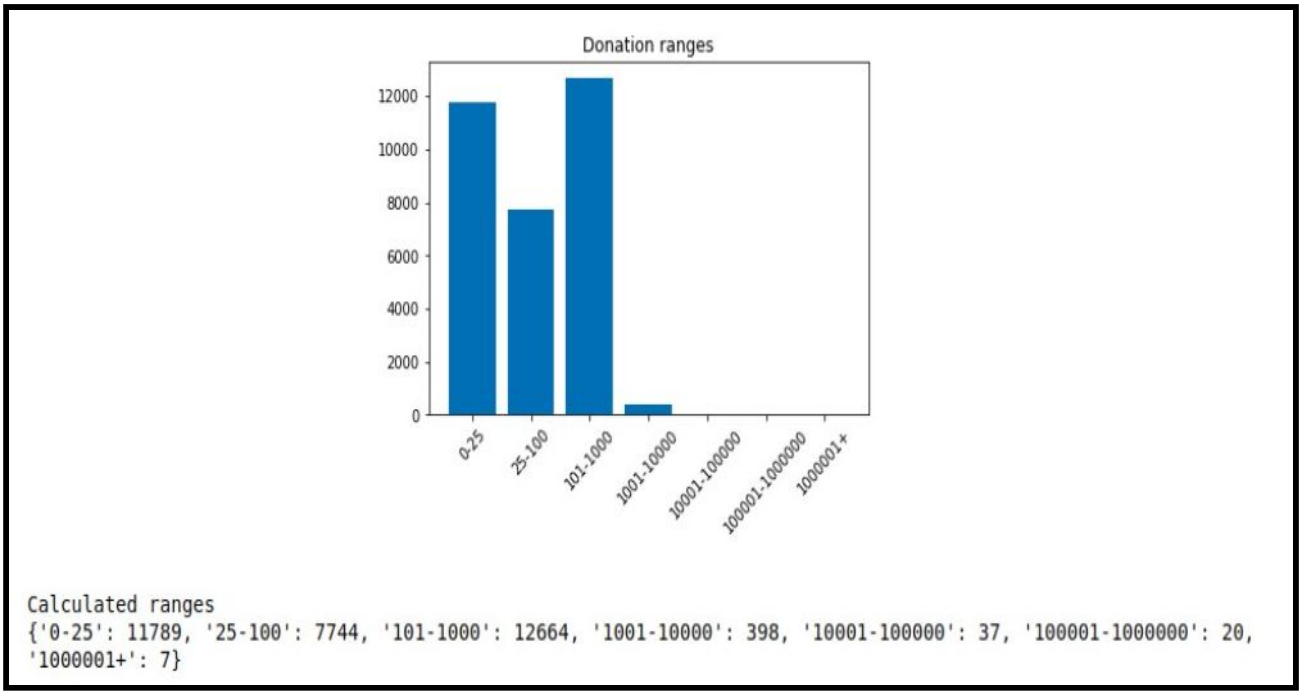



**Fig: No of Donations made during 2016 - 2019 grouped by the above mentioned ranges**

ii. **Analysis using Secretary of State website data to filter OCPF -** We coupled the data collected from the Secretary of State website using keywords extracted such as the industry the businesses in Massachusetts are registered in and the key personnel in these businesses to discover trends in the donations to the state representative candidates and see if there are any particular underlying trends.

iii. **Using keywords provided by the client to analyze data -** We were provided certain keywords such as the Political Action Committee names, Employer Names, Occupation keywords by the client to apply as filters to the OCPF data and analyze them.

## Tools and Packages used:

We used python scripting language throughout our project. Specifically, the following packages were largely used:

- **Selenium** - We used the Selenium python package to create scripts that were used to scrape through the Secretary of State website in order to collect the annual report filings in pdf format.
- **Pandas** - We used the Pandas Dataframe Package to load our datasets in the python environment.
- **PyPDF2** - We used this package to extract information from the pdf collected from the Secretary of State Website.
- **Jupyter Notebook -** We used the Jupyter Notebook as the python environment for our project.
- **Matplotlib -** We used this package to create graphs for visualization purposes.
- **FuzzyWuzzy -** We used this package to obtain a fuzzy match between keywords and data in the dataset. It was found that there were differences in the names collected and the names in the dataset. Examples of there are candidates having middle names as characters in one set and not in another. Fuzzy match gave us a match ratio between two strings.

## Results and Discussions

- **PAC Donation Analysis**

  We analysed the OCPF data on the basis of donations by the Political Action Committees

from the State of Massachusetts. These PACs were specifically associated with the healthcare industry. The names of the PACs were provided by the client.

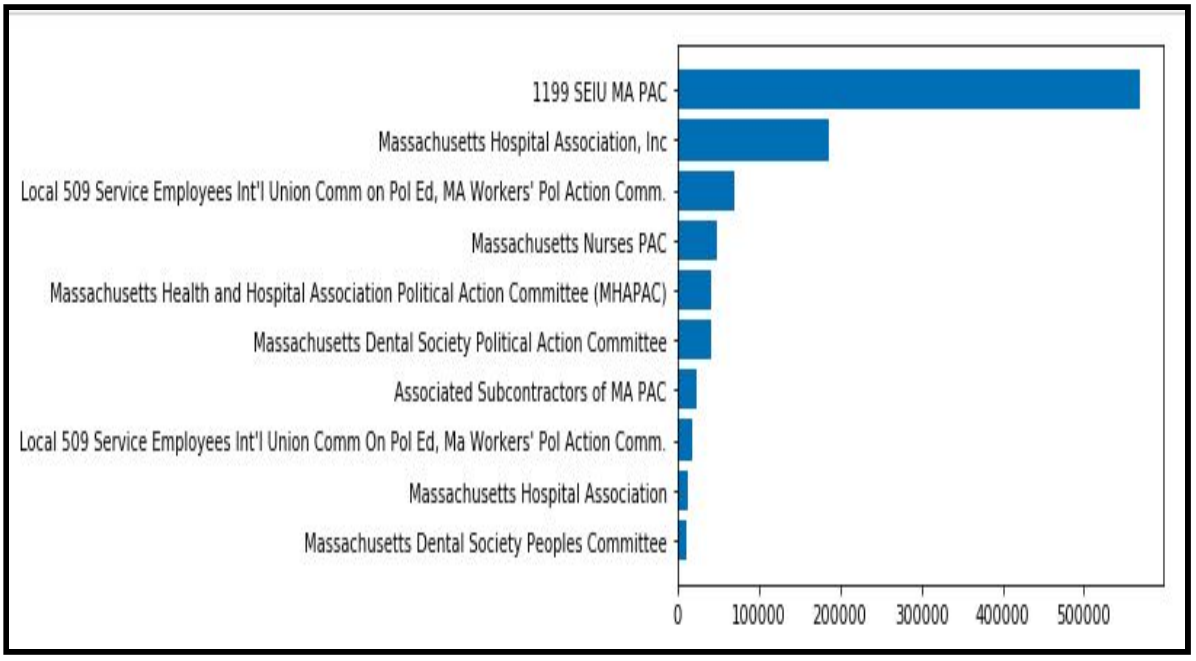



**Fig: Analysis of Donations from PACs associated with the healthcare industry sorted by amount donated for the top ten donors. The above figure shows donations made by these PACs for the years 2016 - 2019 and the total amount donated by them. We can specifically see that the 1199 SEIU MA PAC donated more than twice the amount donated by the next big donor.**

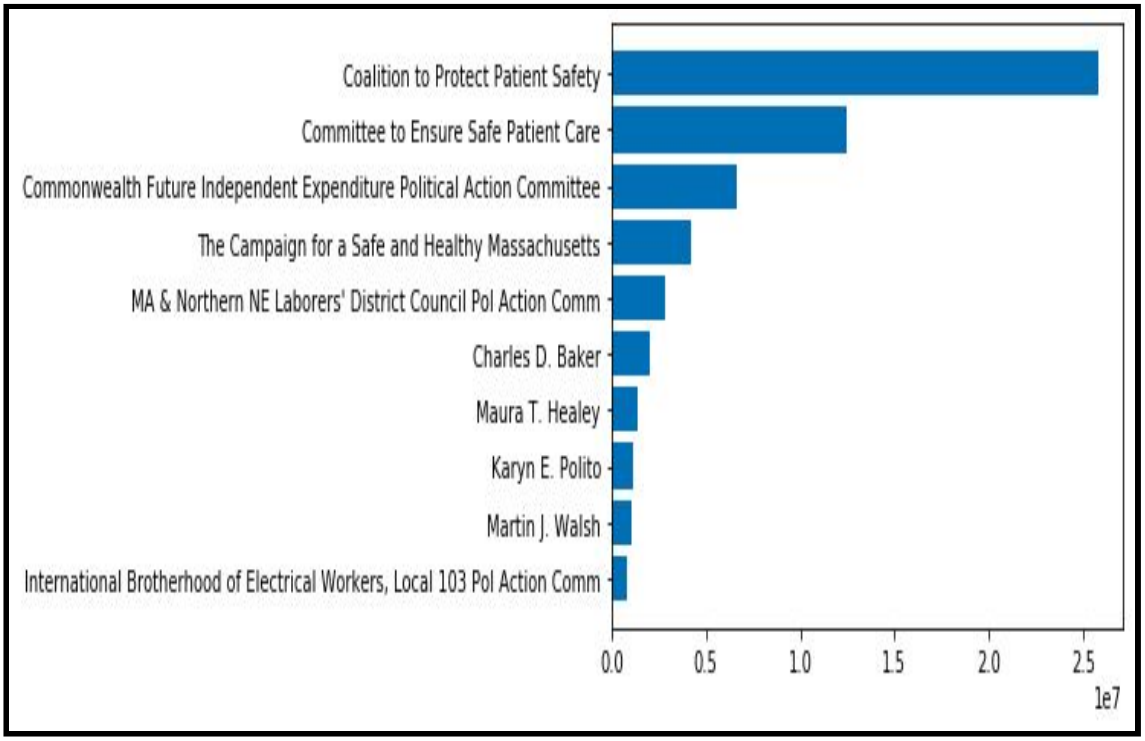



**Fig: Analysis of top ten candidates and committees receiving donations by PACs associated with the healthcare industry sorted in decreasing order of donations received. The above figure shows the recipients of donations made by these PACs for the years 2016 - 2019 and the total amount donated to them. We can specifically see that the Coalition to Protect Patient Safety received more than twice the amount donated by the next big donor.**

- **Analysis of Total Donations Made By Healthcare Industry Over The Last Four Years**

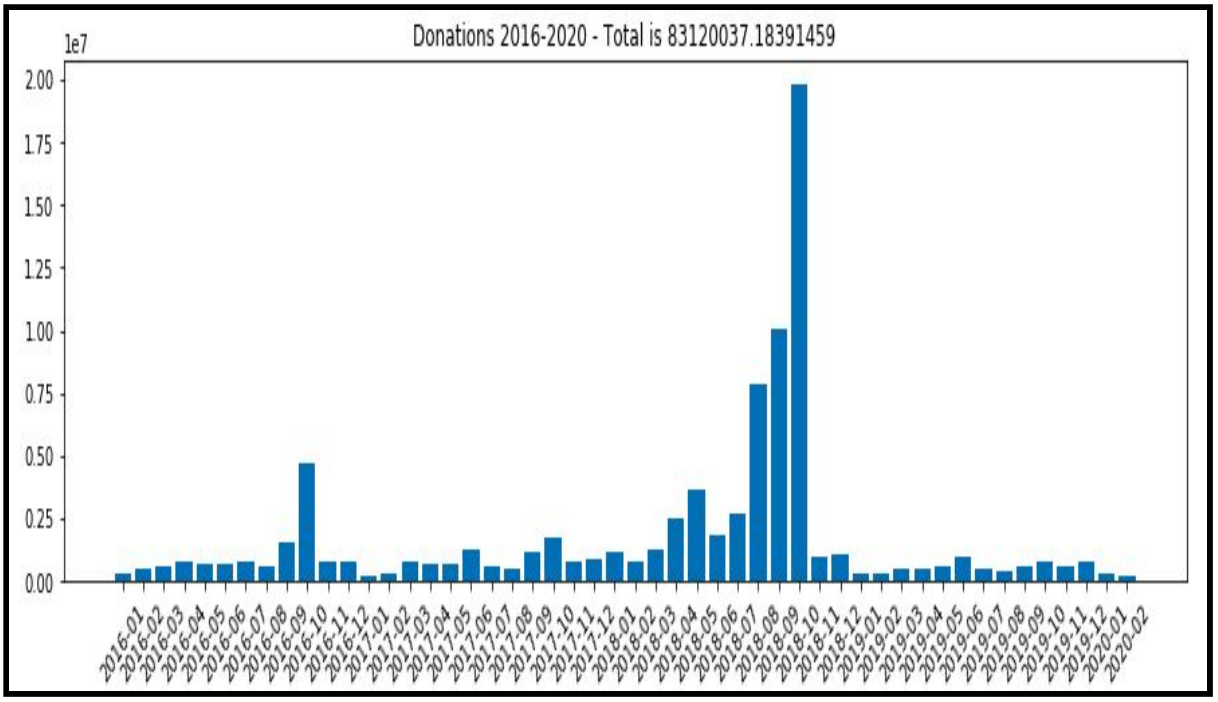



**Fig: Analysis of donations made between 2016 - 2019 grouped by the month.**

From the graph, we can see that there is a massive spike in the donations made in September of 2018. To uncover the reason, we peered into the data around those dates and found that the bulk of the donations were made by the Massachusetts Health & Hospital Association and the Massachusetts Nurses Association to the Coalition to Protect Patient Safety. The latter organization was created specifically to fight against a proposal to introduce government mandated rigid nurse staffing ratios in hospitals. This proposal was set to be voted on in a November ballot, which explains the spike in donations before that time.

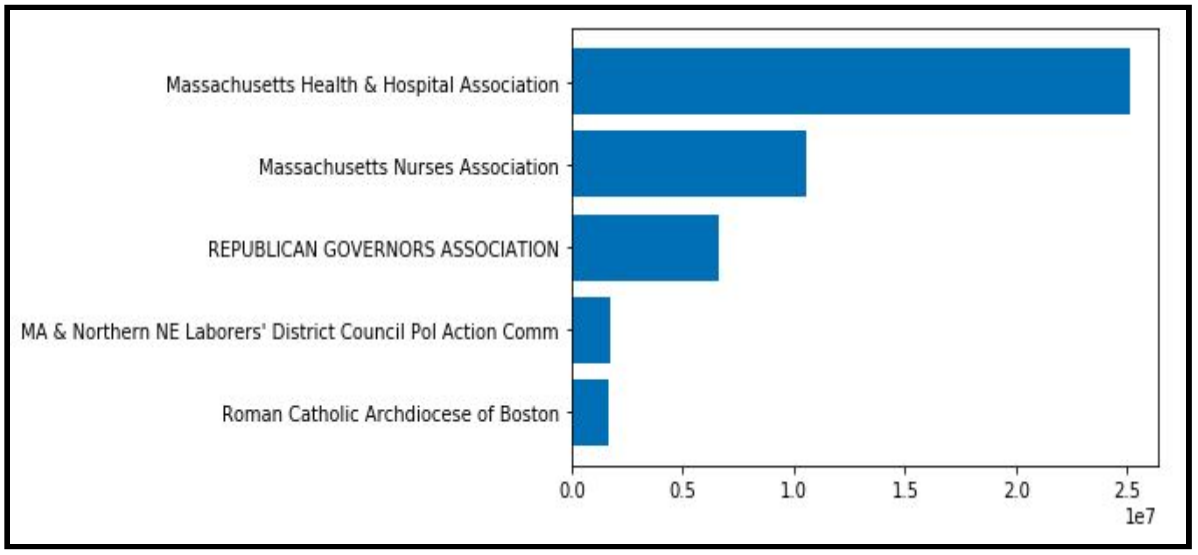
Another thing to note is the spikes around September 2016 and September 2018, both of which lead up to elections, decreasing sharply after the elections are over, as can be seen in the November and December donation amounts.

- **Analysis of Donors Contributing Largest Sums of Donations**

As can be seen in the figure below, Massachusetts Health & Hospital Association and the Massachusetts Nurses Association are the two highest donors, due to the Coalition to Protect Patient Safety. The total amount of donations made by these two associations alone surpass the donations by all other organizations combined by several times.

| Donor_Name | Amount_Donated |
|---|---|
| Massachusetts Health & Hospital Association | 25099250.00 |
| Massachusetts Nurses Association | 10543601.46 |
| Republican Governors Association | 6625000.00 |
| Roman Catholic Archdiocese Of Boston | 1700000.00 |
| Sheldon Adelson | 1000000.00 |
| Massachusetts Nurses Association Region 5 | 750000.00 |
| 1199 Seiu Ma Pac | 569050.00 |
| Laura And John Arnold | 510000.00 |
| Steward Health Care System Llc | 505100.00 |
| Massachusetts Nurses Association Region 2 | 350000.00 |
| Knights Of Columbus | 300000.00 |
| New England District Council Health & Welfare ... | 279120.90 |
| Strong Economy For Growth | 228000.00 |
| New England District Council Health & Welfare | 203409.11 |
| Partners Healthcare | 200000.00 |
| Wynn Resorts, Limited | 200000.00 |
| Massachusetts Nurses Association Region 3 | 200000.00 |
| Gmmb, Inc | 190737.42 |
| Partners Healthcare System, Inc | 190385.60 |
| Massachusetts Hospital Association, Inc | 185947.55 |

**Fig: Total amount contributed during the years 2016 - 2019 grouped by the year donating organization.**




**Below - Bar Graph of total amount contributed during the years 2016 - 2019 grouped by the year donating organization**

- **Analysis of Donations Received by State Representatives**

  If we analyze the 10 state representatives who received the highest amounts of donations for the years 2016 - 2019, we can see that the highest recipient, Robert A. DeLeo received the most donations by a huge margin. He has served as the Speaker of the Massachusetts House since 2009.

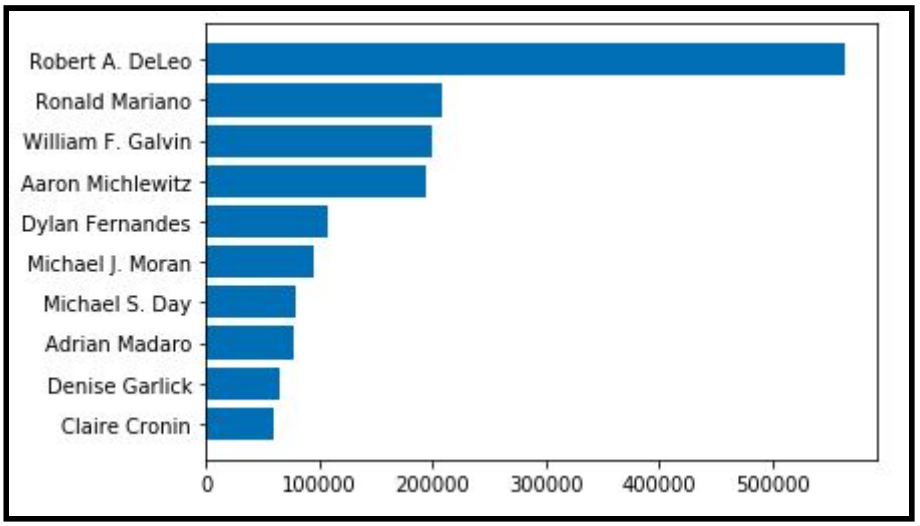



**Fig: Bar Chart of the total donations received by state representatives during the years 2016 - 2019**

- **Time Series Analysis of Donations Received By The 5 Highest State Representatives**

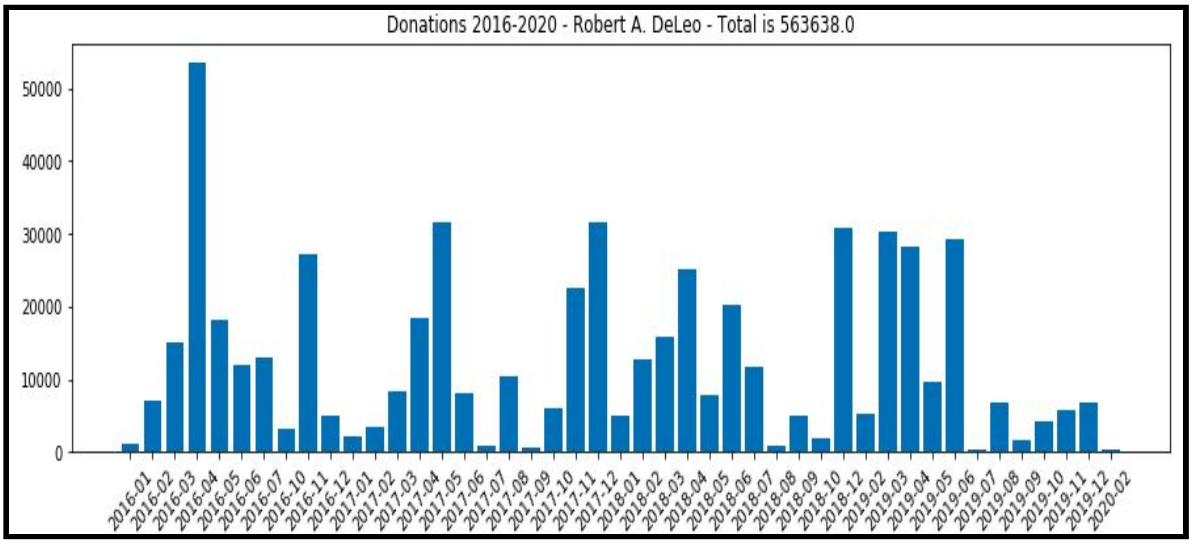



**Fig: Bar Chart of Donations received by Robert A. DeLeo for the years 2016 - 2020. Total Amount Received is $563638.00**

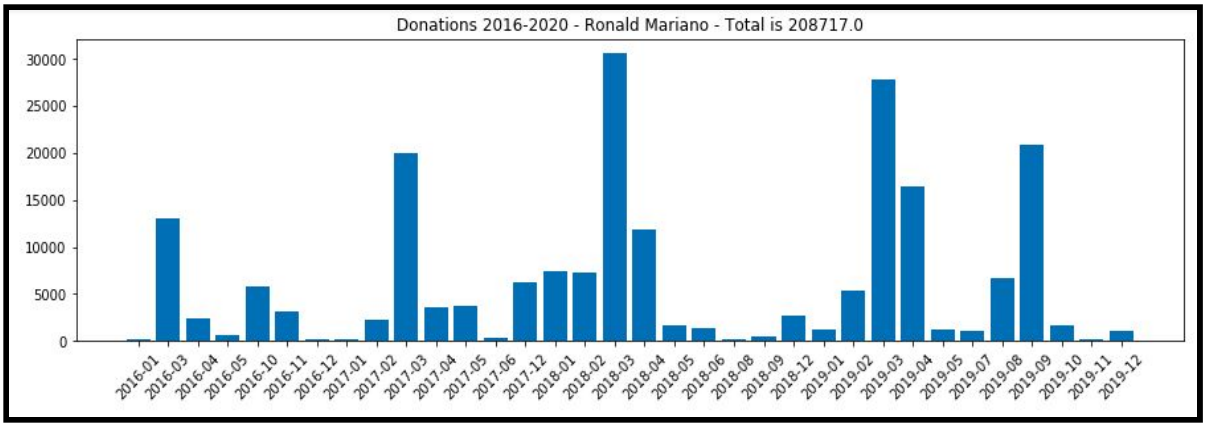



**Fig: Bar Chart of Donations received by Ronald Mariano for the years 2016 - 2020. Total Amount Received is $208717.00**

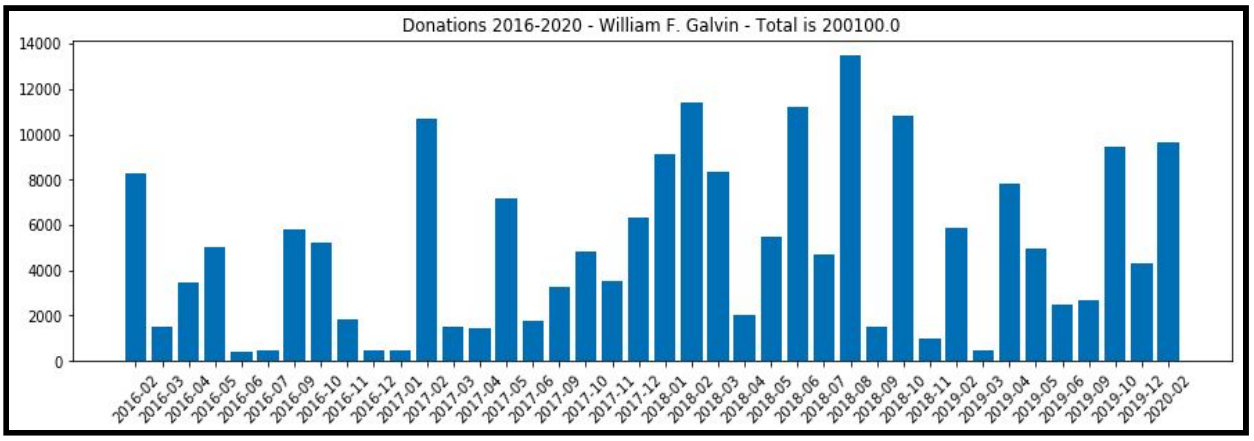



**Fig: Bar Chart of Donations received by William F. Galvin for the years 2016 - 2020. Total Amount Received is $200100.00**

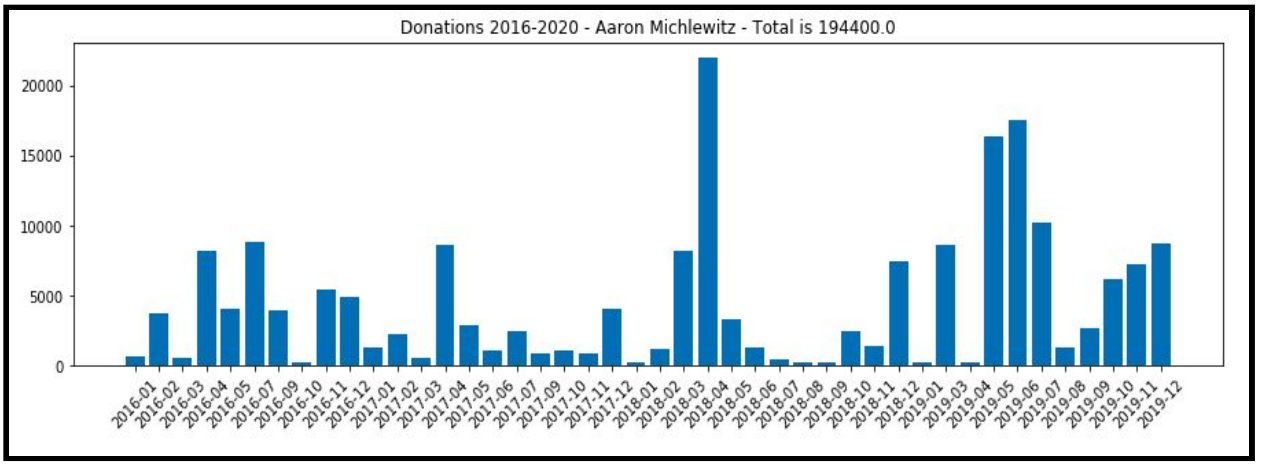



**Fig: Bar Chart of Donations received by Aaron Michlewitz for the years 2016 - 2020. Total Amount Received is $194400.00**

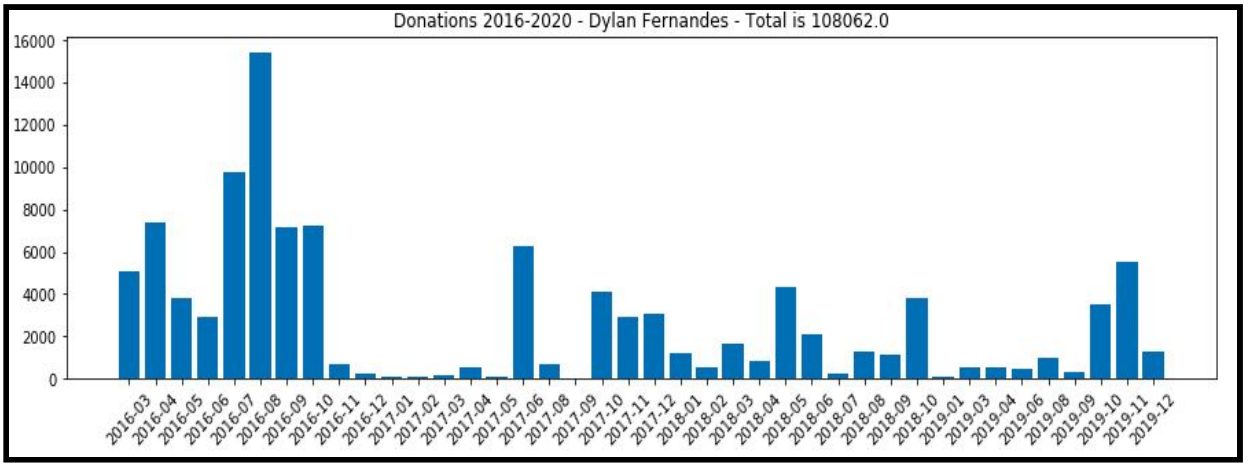



**Fig: Bar Chart of Donations received by Dylan Fernandes for the years 2016 - 2020. Total Amount Received is $108062.00**

While they have their differences, there does seem to be a pattern of spikes around

March of each year. At least some of these spikes correspond to fundraisers held by the individual, but further analysis could lend more insight.

- **Distribution of Donations From Healthcare Industry for State Representatives**

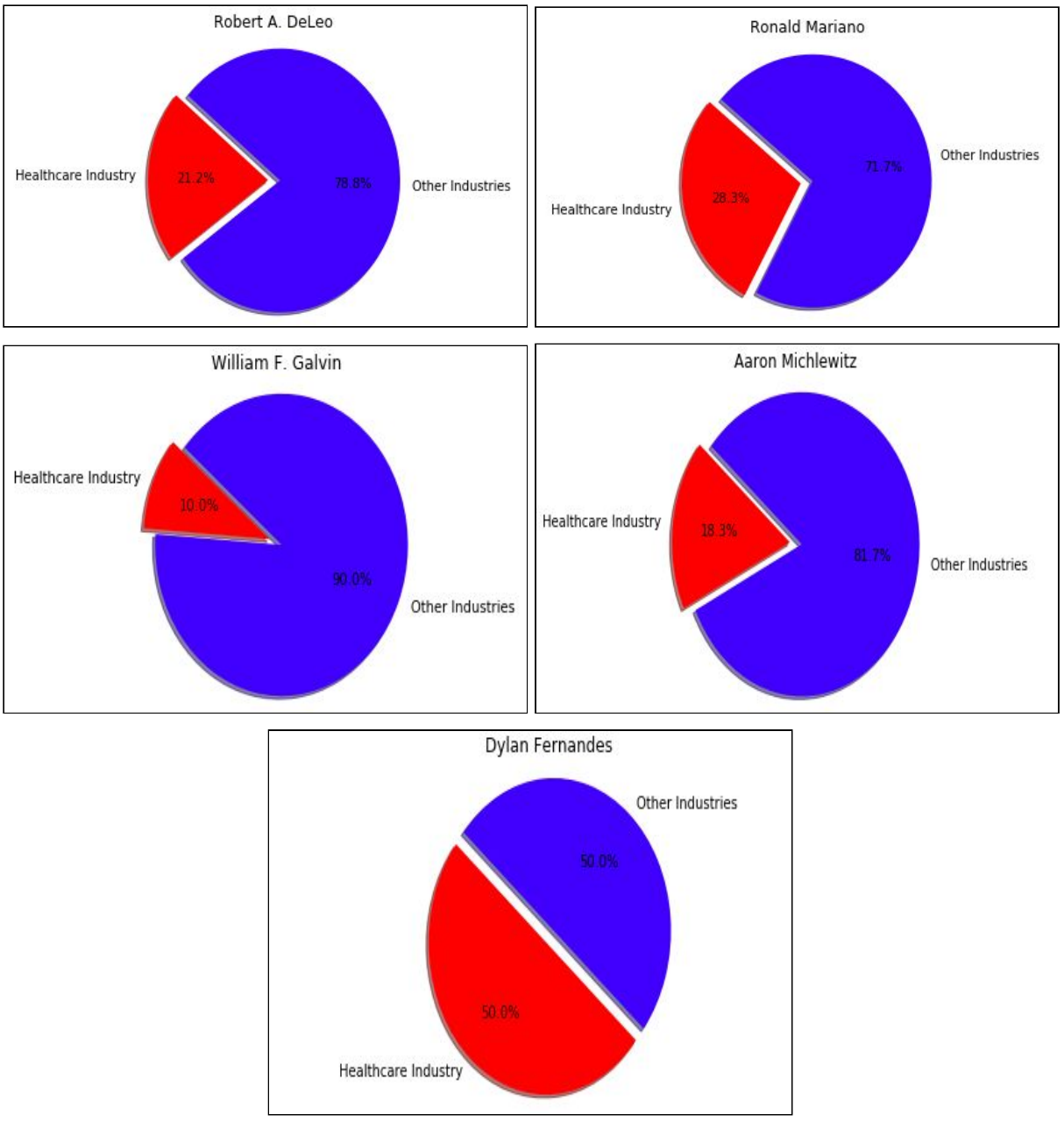


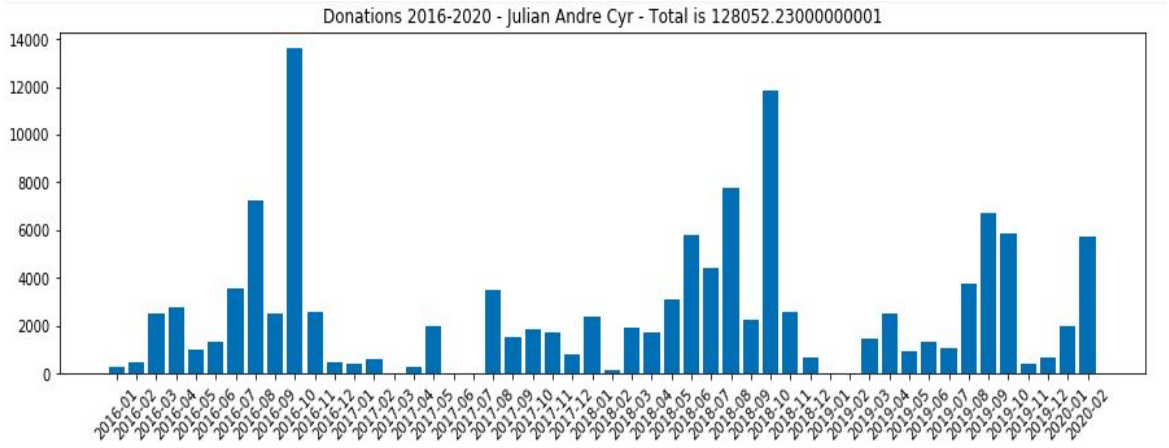

**Fig: Distribution of donations from healthcare industry for top 5 donation recipients**

Among the two highest state representative recipients, the healthcare industry accounted for around a quarter of all the donations they received. Overall, it varies a lot, going as high as 50% of the total donations.

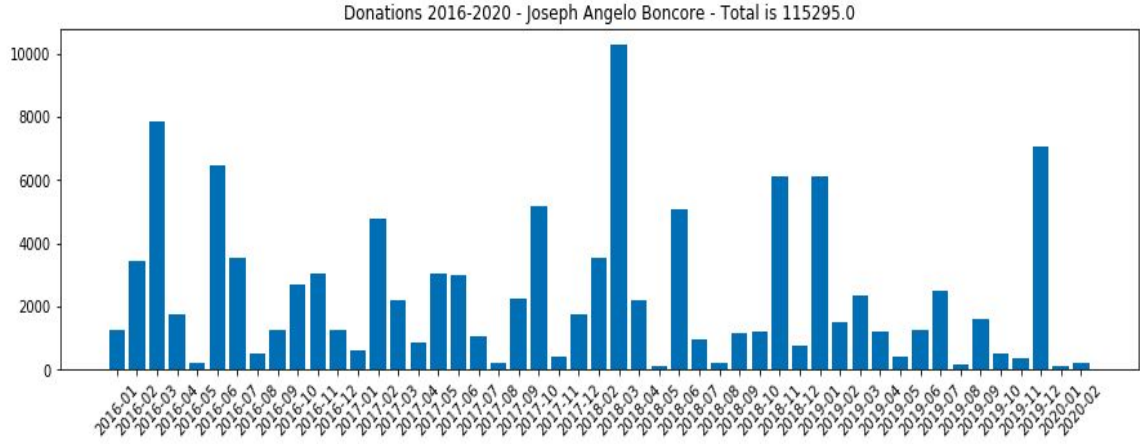
- **Donations received by state representatives on healthcare committees**

  Here we looked at three committees in particular

  i. Healthcare Finance
  ii. Mental Health and Recovery
  iii. Public Health




Julian Andre Cyr is on all three committees and is also the Chair of the Mental Health committee




Joseph Angelo Boncore is on the Healthcare Financing Committee

For most of the other committee members, the received amounts were smaller, likely due to the lack of filtered data.

- **Analysis of Non - State Representative Recipients**

  This is a plot of the highest recipients that were not state representatives and as explained previously, the coalition to protect patient safety was the highest recipient.

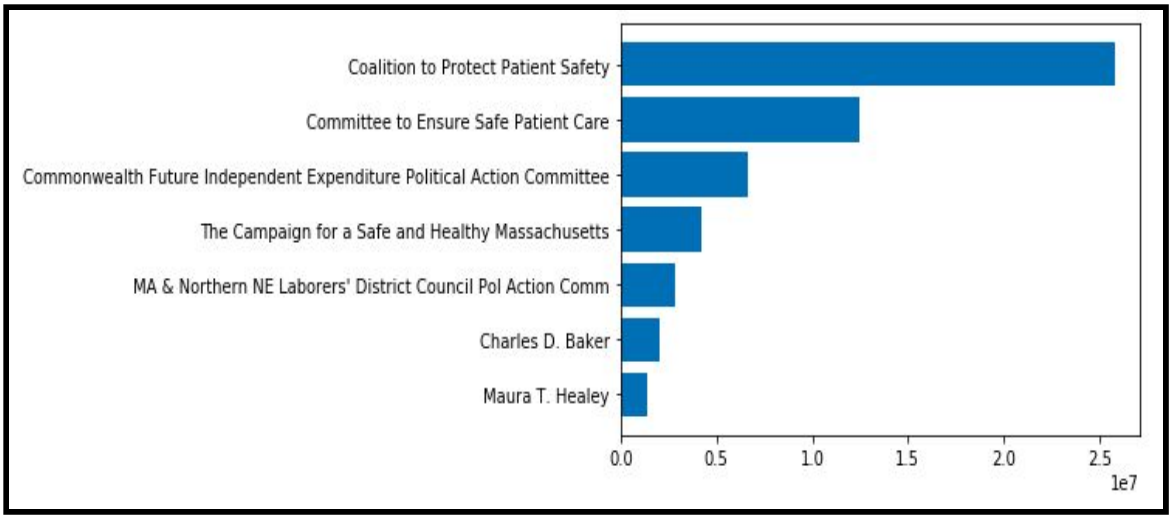



**Fig: Bar Graph of Non - State Representative Recipients**

## Issues and Challenges

- **Secretary of State website slow and buggy** - For us, the part of collecting annual reports was divided into two phases. First we collected the links to the annual reports. The code we developed for this used to often crash and we'd have to change the cookie and viewstate parameters in our code. Second was the phase of downloading the PDFs, which was slow and it took longer than estimated because of several bugs in the website and an anti-scraper that we had to bypass.
- **Data download process was delayed as the website was changed** - The process to download the data from the secretary of state website was further delayed because midway through the process the website was changed, which complicated our data collection code a lot, forcing us to change it and introduce some measures that made the overall process slightly slower.
  For both the issues discussed above, we immediately reached out to our project PM and the clients from Progressive mass and had kept them in the loop as we worked to get through it.
- **Lots of different formats used in the PDF's, hence hard to collect data** - Another difficulty that we faced and had to overcome was the different formats of the annual reports on the secretary of state website. It took quite a bit of trial and error analysis to model the format of the PDF's as accurately as possible.
- **Name matching isn't exact** - Given that matching names and keywords is not an exact procedure, it is highly likely we would have missed out on data that could have proved significant.

## Conclusion and Future Work

- We constructed a dataset of Annual Reports that can reveal a lot of data about companies operating in Massachusetts.
- Through weekly meetings with the clients, we cleaned the data and generated analysis that would be relevant to the healthcare industry.
- A good portion of the PDF documents were in unreadable format. A future team could potentially use an OCR library to further improve results and generate more relevant analysis.
- It would also be interesting to do geographical analysis of donors, which could highlight some outliers or interesting data points that we would have otherwise missed.
- It has been enjoyable working on this project. We have learnt a lot on working with real world data and performing analysis on it.