

**Boston University**  
**CS506**  
**Bay State Banner**  
**Final Deliverable**

Zilin Zhang (U87038789)  
Mingyan Yang (U30215865)  
Yujing Chen (U70567267)  
Yicheng Li (U29503597)

## **1. Introduction**

The Bay State Banner is an independent newspaper primarily geared toward the readership interests of the African-American community in Boston, Massachusetts. It owns news weekly that reports on the political, economic, social and cultural issues that are of interest to African American and English speaking Latinos in Boston and throughout New England. Started in 1965 by Melvin B. Miller, a graduate of Harvard College and Columbia Law School, the Banner provides its estimated 120,000 weekly print readers and 300,000 online readers with reports on a mix of local, national, international news and cultural events. The Banner is the newspaper of record for this community and has chronicled the struggles and successes of the last 50 years.

To carry out the political plans, politicians need a large amount of money to support the activities. And they need donors. Actually, the real estate industry is the largest donor to Massachusetts politicians. It covers brokers, developers, architects, and anyone with a vested investment in building. They acquire a large number of profits from the real estate industry and become the power of the politicians to support the politicians' activities.

The real estate industry covers many people who have a vested investment in building, including brokers, developers, architects and ect.. Our client Yawu Miller, a journalist writing for the Bay State banner would like to examine the lobbying power of the real estate industry in Massachusetts.

## 2. Data Collection & Processing

### 2.1 Data Source

Our data source are The Secretary of Commonwealth's website:

<http://corp.sec.state.ma.us/corpweb/CorpSearch/CorpSearch.aspx>

Our task is to collect all business filings for the past 10 years that specify "real estate" as the industry. We need to type "a" or "b" or "c", etc...then view filings, annual report, extract PDFs from all links; and extract names of bulist "real estate" as industry. Actually we type "aa" or "ab",etc...to get results because of the large amount of data.

### 2.2 Scraping & Data Downloading

We use Python code to scrape the Secretary of Commonwealth's website.

Since we were not familiar with the scraping, we use Selenium IDE to create the basic skeleton of the code. Then we modified the code to the one that works for our scraping. And we choose the Chrome driver as the driver scraping the data, so we download the Chrome driver to help the scraping.

You can find our Python code in the "code" folder. Just run the *Scraping.py* and we can download all the PDFs automatically.

When we scrape data, the biggest problem is that we are rejected by the website due to the high frequency of visiting. To solve the problem, we added the *sleep()* to change the frequency of visiting. Also, we restarted the Chrome driver every 100 downloading so that we will not be rejected anymore.

### 2.3 Changing File Format

Just run the *Modify.py* in the "code" folder and we can change all the files format from ".aspx" into ".pdf".

### 2.4 Extracting

Run the *extracter.py* in the "code" folder and we can extract some information from all the PDFs. We save the information as a .csv file. Then we clean the information using the Excel and then get the final dataset.

We have two kinds of dataset extracted from the downloaded PDFs. One only contains name and description and the other contains many other detailed information like address and member information. Our project manager also provided us OCPF data to do the filtering. All of them are saved in excel files.

## 2.5 Filtering & Data Preparation

We use the following keywords related to real estate to filter the two kinds of data from the secretary website using the column “description”:

Realtor	Real Estate
Contractor	Construction
Builder	Building
Carpenter	Property
Painter	Property Management
Architect (ambiguous)	Surveyor
Developer (ambiguous)	Civil Engineer/ing
Planner	Realtor
Environmental	Realty
Reviewer	
Surveyor	

Also, we use the following name of PACS to filter the OCPF data using the name of PACS (This is only a part of it. We don't list all the names of PACS here. ). We also use the keywords shown above to filter the OCPF data using columns “Occupation” and “Employer”. Also, we use APPENDIX in the keywords document to filter the data using column “Last\_Name”:

Greater Boston Real  
Estate Board PAC  
NAIOP MA Commercial  
Real Estate  
Development Political  
Action Committee  
  
Real Estate Bar  
Association for MA PAC

The keywords of real estate can be find here:

<https://docs.google.com/document/d/1ITRwlff2TkTwBaKpSVPL5vO5QYOzsp2ib35SkLZGCdQ/edit>

We will show the filtered data in the next part.


### 3. Data Description & Visualization

You can view our filtered data here:

<https://drive.google.com/drive/folders/0AONGMu03aGUKUk9PVA>

#### 3.1 PDFs from the Secretary Website

According to the PDFs we have downloaded, most PDFs have the same format like this image:

		<b>The Commonwealth of Massachusetts</b> <b>William Francis Galvin</b>		Minimum Fee: \$100.00	
Secretary of the Commonwealth, Corporations Division One Ashburton Place, 17th floor Boston, MA 02108-1512 Telephone: (617) 727-9640					
<b>Annual Report</b> (General Laws, Chapter 156D, Section 16.22; 950 CMR 113.57)					
Identification Number: <u>463158640</u>					
1. Exact name of the corporation: <u>A&amp;B AUTO INC</u>					
2. Jurisdiction of Incorporation: State: <u>MA</u> Country:					
3.4. Street address of the corporation registered office in the commonwealth and the name of the registered agent at that office: Name: <u>WILLIAM J. PLASKO JR.</u> No. and Street: <u>10 GUILD ST</u> City or Town: <u>NORWOOD</u> State: <u>MA</u> Zip: <u>02062</u> Country: <u>USA</u>					
5. Street address of the corporation's principal office: No. and Street: <u>10 GUILD ST</u> City or Town: <u>NORWOOD</u> State: <u>MA</u> Zip: <u>02062</u> Country: <u>USA</u>					
6. Provide the name and addresses of the corporation's board of directors and its president, treasurer, secretary, and if different, its chief executive officer and chief financial officer.					
Title		Individual Name First, Middle, Last, Suffix		Address (no PO Box) Address, City or Town, State, Zip Code	
PRESIDENT		WILLIAM J PLASKO JR		12 ROSEDALE RD NORWOOD, MA 02062 USA	
PRESIDENT		WILLIAM J PLASKO JR		12 ROSEDALE RD NORWOOD, MA 02062 USA	
TREASURER		WILLIAM J PLASKO JR		12 ROSEDALE RD NORWOOD, MA 02062 USA	
SECRETARY		WILLIAM J PLASKO JR		12 ROSEDALE RD NORWOOD, MA 02062 USA	
VICE PRESIDENT		WILLIAM J PLASKO JR		12 ROSEDALE RD NORWOOD, MA 02062 USA	
DIRECTOR		WILLIAM J PLASKO JR		12 ROSEDALE RD NORWOOD, MA 02062 USA	
7. Briefly describe the business of the corporation: <u>OWN AND OPERATE A DRIVING BUSINESS AND FACILITIES</u>					
8. Capital stock of each class and series:					
Class of Stock	Par Value Per Share Enter 0 if no Par	Total Authorized by Articles of Organization or Amendments		Total Issued and Outstanding	

CNP	\$0.00000	Num of Shares 10,000	Total Par Value \$0.00	Num of Shares 10,000
-----	-----------	-------------------------	---------------------------	-------------------------

9. Check here if the stock of the corporation is publicly traded: ☐

10. Report is filed for fiscal year ending: 12/31/ 2016

Signed by WILLIAM J PLASKO JR, its PRESIDENT  
on this 27 Day of April, 2017

© 2001 - 2017 Commonwealth of Massachusetts  
All Rights Reserved

## 3.2 Extracted Data from PDFs

The two kinds of dataset collected after filtering are shown below.

### 3.2.1 The file containing name and description

Name	Description
JML HOLDING COMPANY I	REAL ESTATE
THE WALNUT HILL GROUP, LLC	PROVIDE REAL ESTATE BROKERAGE SERVICES AS WELL AS NOTARY SERVICES.
THE WARBY GROUP LLC	SCIENCE EDUCATION, AND ENVIRONMENTAL CONSULTING; AND ANYTHING ELSE LEGAL
THE WARREN GROUP, INC.	PUBLICATION OF REAL ESTATE & BANKING NEWS & DATA
THE WATER CONNECTION, LLC	REAL ESTATE HOLDING COMPANY
THE WATER PLANET COMPANY	ENVIRONMENTAL CONSULTING AND OPERATIONS
THE WATERMAN REALTY MANAGEMENT	COMMERCIAL REAL ESTATE MANAGEMENT
TWC LIQUIDATION, INC.	SPECIALTY SUBCONTRACTOR
THE WATERPROOFING COMPANY, LLC	SPECIALTY SUBCONTRACTOR WATERPROOFING, DAMPROOFING, CAULKING AND MASO
THE WAY CONSTRUCTION, INC	CONSTRUCTION
THE WEATHERIZE GUYZ, LLC	CONSTRUCTION
THE WECO GROUP, LTD.	REAL ESTATE
THE WEEPING BEECH, INC.	LANDSCAPE DESIGN & CONSTRUCTION
THE WELL FOR INTEGRATIVE MEDICINE, LLC	THE PURPOSES OF THE LIMITED LIABILITY COMPANY ARE TO RENT REAL PROPERTY AND
UL INFORMATION AND INSIGHTS INC.	OCCUPATIONAL AND ENVIRONMENTAL SAFETY & COMPLIANCE
THE WESTFORD LLC	SINGLE FAMILY HOME PURCHASE, RENOVATION, AND SALE. INVESTMENT PROPERTY.
THE WESTWOOD GROUP PROPERTY	PROPERTY OWNERSHIP, DESIGN AND CONSTRUCTION AND OTHER PURPOSES PERMITTED
THE WHALE LLC	REAL ESTATE
THE WHARF PROPERTIES LLC	REAL ESTATE
THE WHEELRIGHT, LLC	PURCHASING, MANAGING, AND SELLING REAL ESTATE, AND ANY OR ALL OTHER LAWFUL
THE WHITING-TURNER CONTRACTING COMPANY	CONSTRUCTION MANAGER/GENERAL CONTRACTOR
THE WILBUR GROUP, INC.	CONSTRUCTION/REMODELING
THE WILDER COMPANIES, LTD.	REAL ESTATE MANAGEMENT & LEASING
WILDER & CHAVEZ LLC	REAL ESTATE DEVELOPMENT
THE WILEMPAT, L.L.C.	PURCHASE, REPAIR, RENOVATE, RENT AND SELL ALL KINDS OF REAL ESTATE FOR INVEST
THE WILLI MANAGEMENT CORPORATION	MANAGEMENT OF REAL ESTATE
THE WILLIAM FALL GROUP, INC.	REAL ESTATE VALUATION
THE WILLIAM GERALD GROUP LLC	TO OWN, OPERATE, SELL AND DEVELOP REAL ESTATE
THE WILLIAMSVILLE CENTER, LLC	OWNERSHIP AND INVESTMENT IN REAL ESTATE AND OWNERSHIP AND OPERATION OF AD
THE WILMAT COMPANY, INC.	RESIDENTIAL/COMMERCIAL CONSTRUCTION AND REMODELING

### 3.2.2 The file containing many other detailed information



Name of Corporation	Street Address	Street Address of Corporation Principal Office	Registered Agent	Description	Member Title	Member Individual Name	Member Address
NETCOV MANAGED IT SERVICES LLC	3 ELECTRONICS	3 ELECTRONICS AVE UNIT 102, DANVERS, MA, 0192	TERRANCE L LEC	MANAGED SERVICE P	MANAGER	TERRANCE L LEDGER	19A BAYNS HILL ROAD BOXFORD, MA 01921 USA
NETCOV MANAGED IT SERVICES LLC	3 ELECTRONICS	3 ELECTRONICS AVE UNIT 102, DANVERS, MA, 0192	TERRANCE L LEC	MANAGED SERVICE PROVIDER FOR INFORMATI	TECHNOLOGY SERVICES		
NETCOV MANAGED IT SERVICES LLC	3 ELECTRONICS	3 ELECTRONICS AVE UNIT 102, DANVERS, MA, 0192	TERRANCE L LEC	MANAGED SERVICE PROVIDER FOR INFORMATION TECHNOLOGY SERVICES			
NAIL STYLUS CORPORATION		55 OAKWOOD AVE 2, REVERE, MA, 02151, USA		NAIL CARE SERVICE	PRESIDENT	BILLY K HUYNH	55 OAKWOOD AVE, 2 REVERE, MA 02151 USA
NAIL STYLUS CORPORATION		55 OAKWOOD AVE 2, REVERE, MA, 02151, USA		NAIL CARE SERVICE	PRESIDENT	BILLY KIM HUYNH	55 OAKWOOD AVE #2 REVERE, MA 02151 USA
NAIL STYLUS CORPORATION		55 OAKWOOD AVE 2, REVERE, MA, 02151, USA		NAIL CARE SERVICE	PRESIDENT	BILLY KIM HUYNH	55 OAKWOOD AVE #2 REVERE, MA 02151 USA
NAIL STYLUS CORPORATION		55 OAKWOOD AVE 2, REVERE, MA, 02151, USA		NAIL CARE SERVICE	PRESIDENT	BILLY KIM HUYNH	55 OAKWOOD AVE #2 REVERE, MA 02151 USA
NAIL STYLUS CORPORATION		55 OAKWOOD AVE 2, REVERE, MA, 02151, USA		NAIL CARE SERVICE	TREASURER	BILLY KIM HUYNH	55 OAKWOOD AVE #2 REVERE, MA 02151 USA
NAIL STYLUS CORPORATION		55 OAKWOOD AVE 2, REVERE, MA, 02151, USA		NAIL CARE SERVICE	SECRETARY	BILLY KIM HUYNH	55 OAKWOOD AVE #2 REVERE, MA 02151 USA
NAIL STYLUS CORPORATION		55 OAKWOOD AVE 2, REVERE, MA, 02151, USA		NAIL CARE SERVICE	DIRECTOR	BILLY KIM HUYNH	55 OAKWOOD AVE #2 REVERE, MA 02151 USA
NATWELLS TRADING, INC.		8 MIKAILA WAY, WOBURN, MA, 01801, USA		ENERGY TRADING	PRESIDENT	HONGYANG LIU	8 MIKAILA WAYWOBURN, MA 01801 USA
NATWELLS TRADING, INC.		8 MIKAILA WAY, WOBURN, MA, 01801, USA		ENERGY TRADING	TREASURER	BECK CHEN	8 MIKAILA WAYWOBURN, MA 01801 USA
NATWELLS TRADING, INC.		8 MIKAILA WAY, WOBURN, MA, 01801, USA		ENERGY TRADING	SECRETARY	BECK CHEN	8 MIKAILA WAYWOBURN, MA 01801 USA
NATWELLS TRADING, INC.		8 MIKAILA WAY, WOBURN, MA, 01801, USA		ENERGY TRADING	DIRECTOR	HONGYANG LIU	8 MIKAILA WAYWOBURN, MA 01801 USA
NATWELLS TRADING, INC.		8 MIKAILA WAY, WOBURN, MA, 01801, USA		ENERGY TRADING	DIRECTOR	BECK CHEN	8 MIKAILA WAYWOBURN, MA 01801 USA
NORTHEAST FLEX HEATERS, INC.		251 FORGE RD., WESTPORT, MA, 02790, USA		MANUFACTURER	PRESIDENT	CHARLES MICHAUD	251 FORGE RD. WESTPORT, MA 02790 USA
NORTHEAST FLEX HEATERS, INC.		251 FORGE RD., WESTPORT, MA, 02790, USA		MANUFACTURER	PRESIDENT	CHARLES MICHAUD	251 FORGE RD. WESTPORT, MA 02790 USA

### 3.3 OCPF Data

After filtering OCPF data in the “Last\_Name” column using the name of PACS, we can get the information under the “Committee” type as below.

Amount	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	
	First Name	Last Name	Exp/1003	Address	City	State	Zip	Occupation	Employer	Contributor	Date	myDate	Report ID	Full Name	Full Name, I District	District	Category	Report_Can	Report_Can	Report_Off	Report_Comm	Name			
250.00		Sheet Metal Workers	L 2015-01-22	1157 Adams	Dorchester	MA	02124			Committee	2015-01-22	2015/1/22	215677	Timothy J. Toomey Jr., 26th Middle		283 D		Timothy J.	Toomey Jr.	Cambridge	Toomey	Committee			
500.00		New England Regions	L 2015-03-30	750 Dorches	Boston	MA	02125-1132			Committee	2015-03-30	2015/3/30	217668	Jason Palitsch	Palitsch, Jas 11th Worces		354 N		Jason	Palitsch	House, 11th	Palitsch	Committee		
500.00		Assoc Builders & Contr	L 2015-03-19	100 Union	Woburn	MA	01801			Committee	2015-03-19	2015/3/19	308915	Hannah Kar Kane	Hann, 11th Worces		354 N		Hannah	Kane	House 11th	Kane	Committee		
250.00		Ma & Northern Ne La	L 2015-04-06	7 Labores	Hopkinton	MA	01748			Committee	2015-04-06	2015/4/6	537861	Susanah N Lee	Susan, 2nd Franklir		241 N		Susanah N lee		House 2nd	L Lee	Committee		
500.00		MA & Northern NE La	L 2015-05-18	7 Labores	Hopkinton	MA	01748			Committee	2015-05-18	2015/5/18	538302	Denise Prov	Provost, De 27th Middle		284 N		Denise	Provost	House 27th	Provost	Committee		
500.00		MA Labores' District	L 2015-04-30	7 Labores	Hopkinton	MA	01748			Committee	2015-04-30	2015/4/30	538777	Dennis Rossi	Rosa, Denni 4th Worces		347 N		Dennis	Rosa	House 4th	V Rosa	Committee		
500.00		MA & Northern Ne La	L 2015-05-22	7 Labores	Hopkinton	MA	01748			Committee	2015-05-22	2015/5/22	538882	Elen Story	Story, Ellen 3rd Hampsh		257 N		Elen	Story	House 3rd	H-Story	Committee		
250.00		MA Labores' District	L 2015-04-21	7 Labores	Hopkinton	MA	01748			Committee	2015-04-21	2015/4/21	538945	Fred J. Barr	Barnow, Fred 1st Bristol		228 N		Fred J.	Barnow	House 1st	B Barnow	Committee		
500.00		MA & Northern NE La	L 2015-04-06	7 Labores	Hopkinton	MA	01748			Committee	2015-04-06	2015/4/6	538961	Josh Cutler	Cutler, Josh 6th Plymoul		317 N		Josh	Cutler	House 6th	P Cutler	Committee		
125.00		Carpenters Local Unio	L 2015-06-03	21 Mazso	Randolph	MA	02368			Committee	2015-06-03	2015/6/3	538961	Josh Cutler	Cutler, Josh 6th Plymoul		317 N		Josh	Cutler	House 6th	P Cutler	Committee		
250.00		Sheet Metal Workers	L 2015-05-15	1157 Adams	Dorchester	MA	02124			Committee	2015-05-15	2015/5/15	538961	Josh Cutler	Cutler, Josh 6th Plymoul		317 N		Josh	Cutler	House 6th	P Cutler	Committee		
400.00		Ma Assoc. of Realtors	L 2015-03-21	1400 Main	S Waltham	MA	02451			Committee	2015-03-21	2015/3/21	538961	Josh Cutler	Cutler, Josh 6th Plymoul		317 N		Josh	Cutler	House 6th	P Cutler	Committee		
100.00		Int'l Union of Operati	L 2015-06-15	16 Trotter	D Medway	MA	02053			Committee	2015-06-15	2015/6/15	538961	Josh Cutler	Cutler, Josh 6th Plymoul		317 N		Josh	Cutler	House 6th	P Cutler	Committee		
250.00		Plumbers Union Local	L 2015-06-17	1240 Massac	Boston	MA	02125			Committee	2015-06-17	2015/6/17	538961	Josh Cutler	Cutler, Josh 6th Plymoul		317 N		Josh	Cutler	House 6th	P Cutler	Committee		
500.00		MA & Northern NE La	L 2015-05-01	7 Labores	Hopkinton	MA	01748			Committee	2015-05-01	2015/5/1	538972	Ruth B. Bal	Baler, Ruth 12th Middle		269 N		Ruth B.	Baler	House 12th	Baler	Committee		
300.00		MA Assoc. of Realtors	L 2015-04-06	1400 Main	S Waltham	MA	02451			Committee	2015-04-06	2015/4/6	538978	Stephen Kul	Kulik, Steph 1st Franklin		240 N		Stephen	Kulik	House 1st	F Kulik	Committee		
500.00		MA & Northern NE La	L 2015-04-18	7 Labores	Hopkinton	MA	01748			Committee	2015-04-18	2015/4/18	538978	Stephen Kul	Kulik, Steph 1st Franklin		240 N		Stephen	Kulik	House 1st	F Kulik	Committee		
200.00		Real Estate Bar Associ	L 2015-05-19	50 Congress	Boston	MA	02101			Committee	2015-05-19	2015/5/19	538980	Louis L. Kafil	Kafka, Louis 8th Norfolk		304 N		Louis L.	Kafka	House 8th	N Kafka	Committee		
200.00		Greater Boston Real Es	L 2015-05-19	One Center	Boston	MA	02108			Committee	2015-05-19	2015/5/19	538980	Louis L. Kafil	Kafka, Louis 8th Norfolk		304 N		Louis L.	Kafka	House 8th	N Kafka	Committee		
500.00		MA & Northern NE La	L 2015-05-19	7 Labores	Hopkinton	MA	01748			Committee	2015-05-19	2015/5/19	538980	Louis L. Kafil	Kafka, Louis 8th Norfolk		304 N		Louis L.	Kafka	House 8th	N Kafka	Committee		
500.00		MA Labores' District	L 2015-03-22	7 Labores	Hopkinton	MA	01748			Committee	2015-03-22	2015/3/22	538986	Angelo J. P.	Puppulo Jr., 12th Hampc		253 N		Angelo J.	Puppulo Jr.	House 12th	Puppulo Jr.	Committee		
100.00		Carpenters Local #108	L 2015-06-19	29 Oakland	Springfield	MA	01108			Committee	2015-06-19	2015/6/19	538986	Angelo J. P.	Puppulo Jr., 12th Hampc		253 N		Angelo J.	Puppulo Jr.	House 12th	Puppulo Jr.	Committee		
200.00		Pipefitters Local #537	L 2015-06-20	35 Travis	Str Allston	MA	02134			Committee	2015-06-20	2015/6/20	538986	Angelo J. P.	Puppulo Jr., 12th Hampc		253 N		Angelo J.	Puppulo Jr.	House 12th	Puppulo Jr.	Committee		
500.00		Ma & Northern Ne La	L 2015-05-26	7 Labores	Hopkinton	MA	01748			Committee	2015-05-26	2015/5/26	538987	Todd M. Smi	Smola, Todd 1st Hampde		242 N		Todd M.	Smola	House 1st	H Smola	Committee		
500.00		Massachusetts & Nort	L 2015-04-01	7 Labores	Hopkinton	MA	01748-2684			Committee	2015-04-01	2015/4/1	539023	Paul J. Don	Donato, Paul 35th Middle		292 N				Friends of Paul J. Donato				
250.00		Sheet Metal Workers	L 2015-04-03	1157 Adams	Dorchester	MA	02124-5710			Committee	2015-04-03	2015/4/3	539023	Paul J. Don	Donato, Paul 35th Middle		292 N				Friends of Paul J. Donato				
500.00		Ma & Northern Ne La	L 2015-03-18	7 Labores	Hopkinton	MA	01748			Committee	2015-03-18	2015/3/18	539040	Jennifer Ber	Benson, Jen 37th Middle		294 N		Jennifer	Benson	House 37th	Benson	Committee		
125.00		NIET Peabody Lip Pc	L 2015-06-01	1300 Clinor	Roxchester	NY	14604			Committee	2015-06-01	2015/6/1	539040	Jennifer Ber	Benson, Jen 37th Middle		294 N		Jennifer	Benson	House 37th	Benson	Committee		
500.00		International Brotherh	L 2015-06-02	256 Freepor	Boston	MA	02122			Committee	2015-06-02	2015/6/2	539040	Jennifer Ber	Benson, Jen 37th Middle		294 N		Jennifer	Benson	House 37th	Benson	Committee		
400.00		Ma Assoc. of Realtors	L 2015-06-09	1400 Main	S Waltham	MA	02451			Committee	2015-06-09	2015/6/9	539040	Jennifer Ber	Benson, Jen 37th Middle		294 N		Jennifer	Benson	House 37th	Benson	Committee		
500.00		Pipefitters Local #537	L 2015-06-09	35 Travis	Str Allston	MA	02134			Committee	2015-06-09	2015/6/9	539040	Jennifer Ber	Benson, Jen 37th Middle		294 N		Jennifer	Benson	House 37th	Benson	Committee		
500.00		Assoc Builders & Contr	L 2015-06-09	100 Union	Woburn	MA	01801			Committee	2015-06-09	2015/6/9	539040	Jennifer Ber	Benson, Jen 37th Middle		294 N		Jennifer	Benson	House 37th	Benson	Committee		
200.00		Greater Boston Real Es	L 2015-06-04	One Center	Boston	MA	02108			Committee	2015-06-04	2015/6/4	539040	Jennifer Ber	Benson, Jen 37th Middle		294 N		Jennifer	Benson	House 37th	Benson	Committee		
200.00		MA Systems Contracs	L 2015-06-06	1 Merchant	Sharon	MA	02067			Committee	2015-06-06	2015/6/6	539040	Jennifer Ber	Benson, Jen 37th Middle		294 N		Jennifer	Benson	House 37th	Benson	Committee		
200.00		MA Building Trade	L 2015-06-04	256 Freepor	Dorchester	MA	02122			Committee	2015-06-04	2015/6/4	539147	Denise Garl	Garlick, Den 13th Norfol		309 N		Denise	Garlick	House 13th	Garlick	Committee		
500.00		MA Labores' District	L 2015-03-25	7 Labores	Hopkinton	MA	01748			Committee	2015-03-25	2015/3/25	539142	Denise Garl	Garlick, Den 13th Norfol		309 N		Denise	Garlick	House 13th	Garlick	Committee		
500.00		MA & Northern NE La	L 2015-06-11	7 Labores	Hopkinton	MA	01748			Committee	2015-06-11	2015/6/11	539142	Cory Atkins	Atkins, Cory 14th Middle		271 N		Cory	Atkins	House 14th	Atkins	Committee		
50.00		MA & Northern NE La	L 2015-04-03	7 Labores	Hopkinton	MA	01748			Committee	2015-04-03	2015/4/3	539183	Tom Sannio	Sannicciandro, 7th Middle		264 N		Tom	Sannicciandro	House 7th	N Sannicciandro	Committee		
200.00		Associated Subcontr	L 2015-05-27	31 State	Str Boston	MA	02109			Committee	2015-05-27	2015/5/27	539183	Tom Sannio	Sannicciandro, 7th Middle		264 N				Sannicciandro	House 7th	N Sannicciandro	Committee	
100.00		Pipefitters Local #537	L 2015-02-27	35 Travis	Str Allston	MA	02134			Committee	2015-02-27	2015/2/27	539249	James Murp	Murphy, Jan 4th Norfolk		300 N		James	Murphy	House 4th	N Murphy	Committee		
100.00		Ironworkers Union Lo	L 2015-03-05	195 Old	Col South Bostc	MA	02127			Committee	2015-03-05	2015/3/5	539249	James Murp	Murphy, Jan 4th Norfolk		300 N		James	Murphy	House 4th	N Murphy	Committee		
250.00		International Brotherh	L 2015-03-09	256 Freepor	Boston	MA	02122			Committee	2015-03-09	2015/3/9	539249	James Murp	Murphy, Jan 4th Norfolk		300 N		James	Murphy	House 4th	N Murphy	Committee		
250.00		U.A. Plumbers and Ga	L 2015-03-16	1240 Massac	Boston	MA	02125			Committee	2015-03-16	2015/3/16	539249	James Murp	Murphy, Jan 4th Norfolk		300 N		James	Murphy	House 4th	N Murphy	Committee		
100.00		Labores Local 133 Po	L 2015-04-24	P.O. Box 696	Quincy	MA	02269			Committee	2015-04-24	2015/4/24	539249	James Murp	Murphy, Jan 4th Norfolk		300 N		James	Murphy	House 4th	N Murphy	Committee		
500.00		MA & Northern NE La	L 2015-04-24	7 Labores	Hopkinton	MA	01748			Committee	2015-04-24	2015/4/24	539249	James Murp	Murphy, Jan 4th Norfolk		300 N		James	Murphy	House 4th	N Murphy	Committee		

## 4. Data Presentation

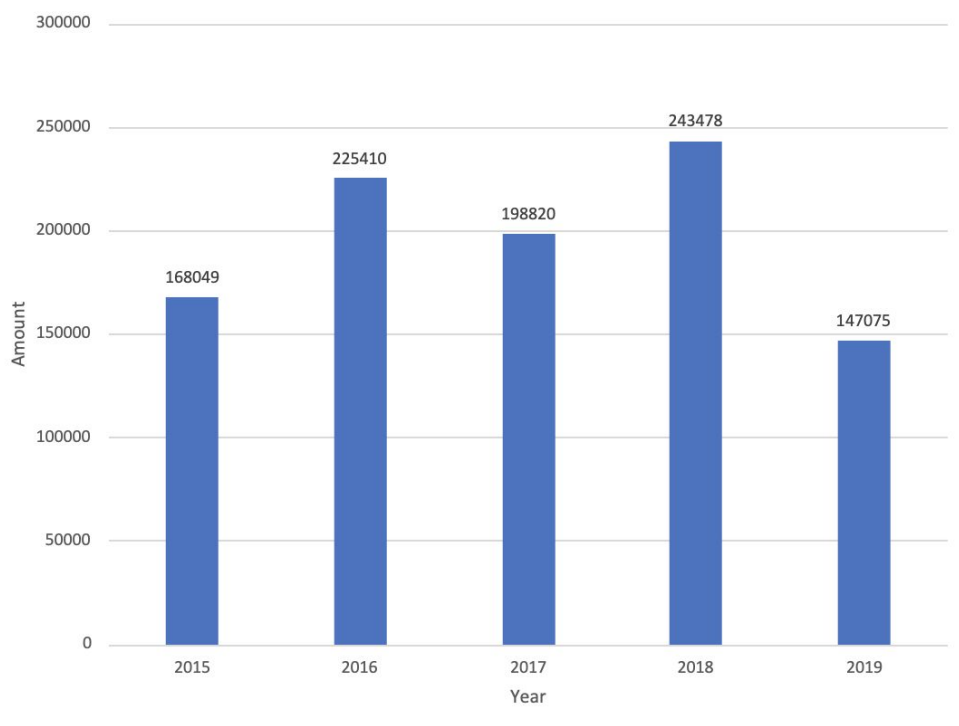


Figure 1: Total amount of money from 2015 to 2019

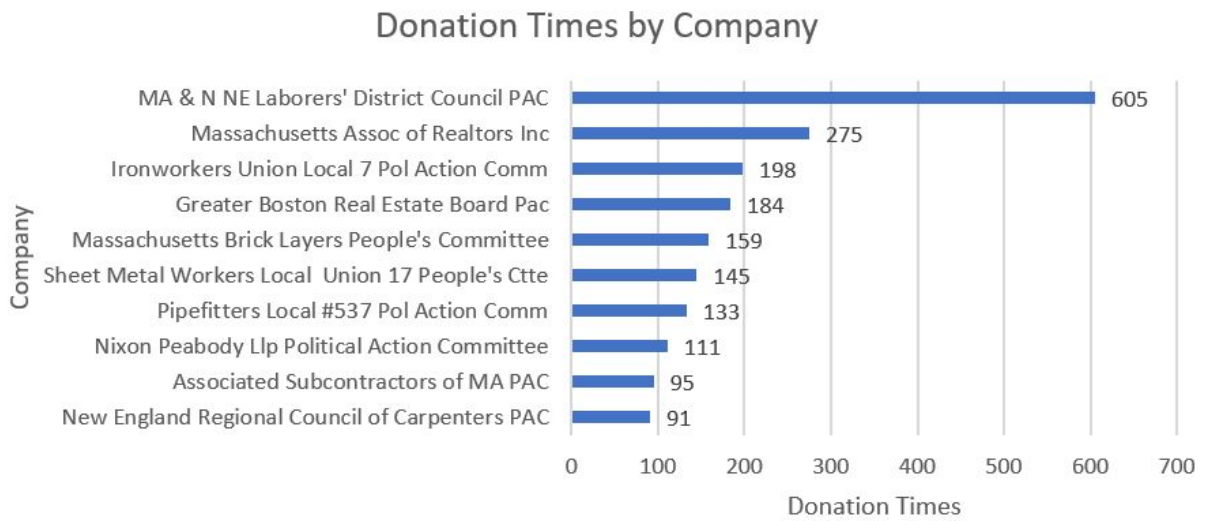


Figure 2: Top 10 companies which donated most times from 2015 to 2019



### Donation Amount by Company

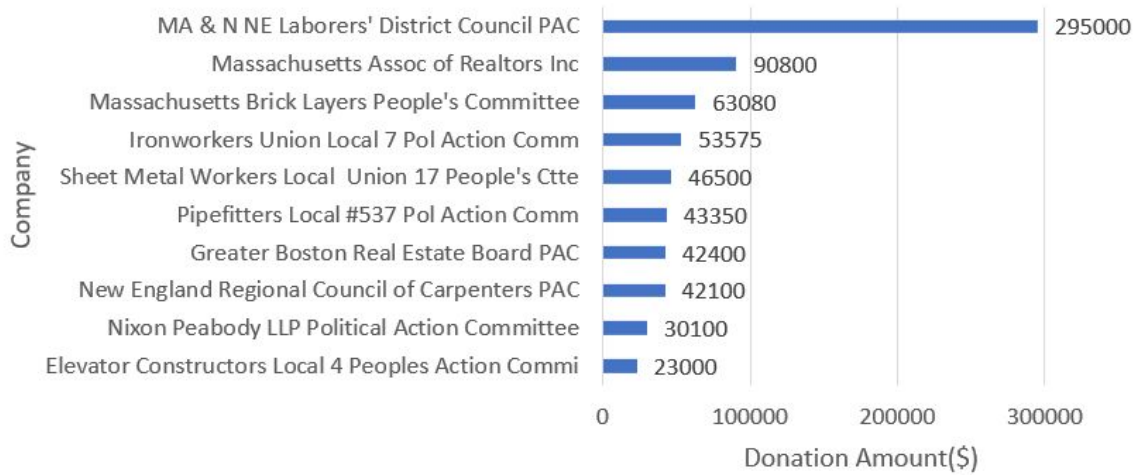


Figure 3: Top 10 companies which donated most money from 2015 to 2019

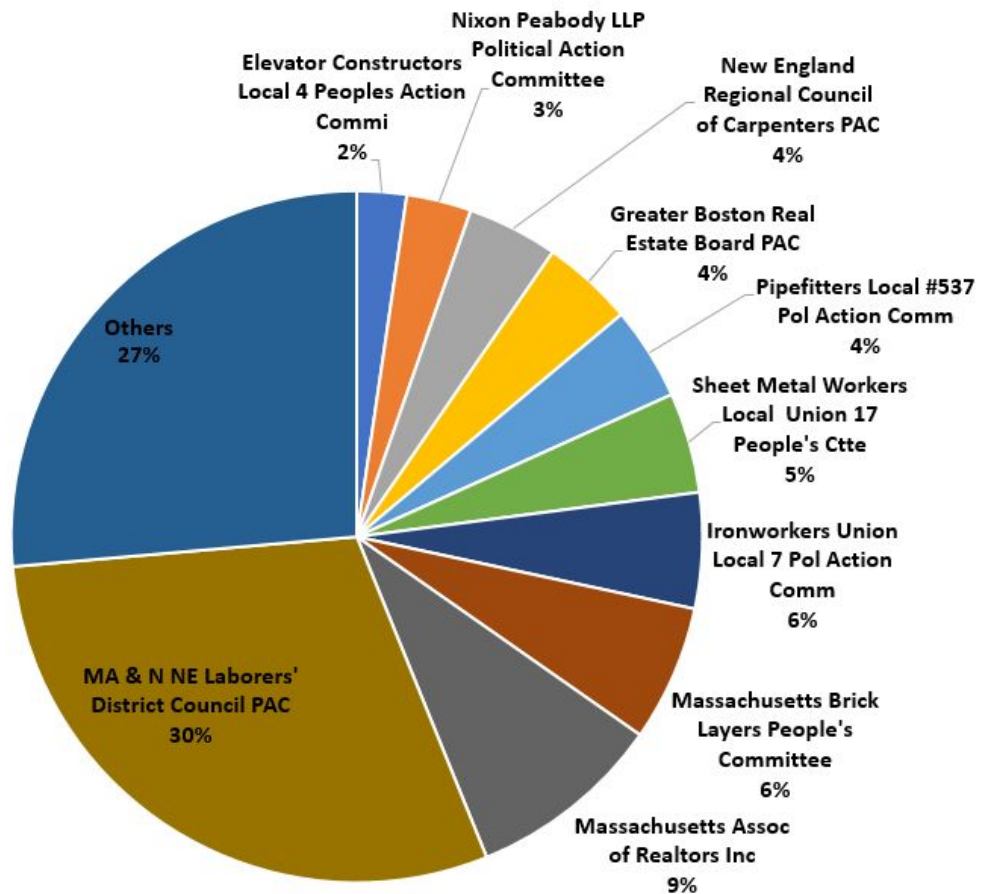


Figure 4: Distribution of donations from 2015 to 2019

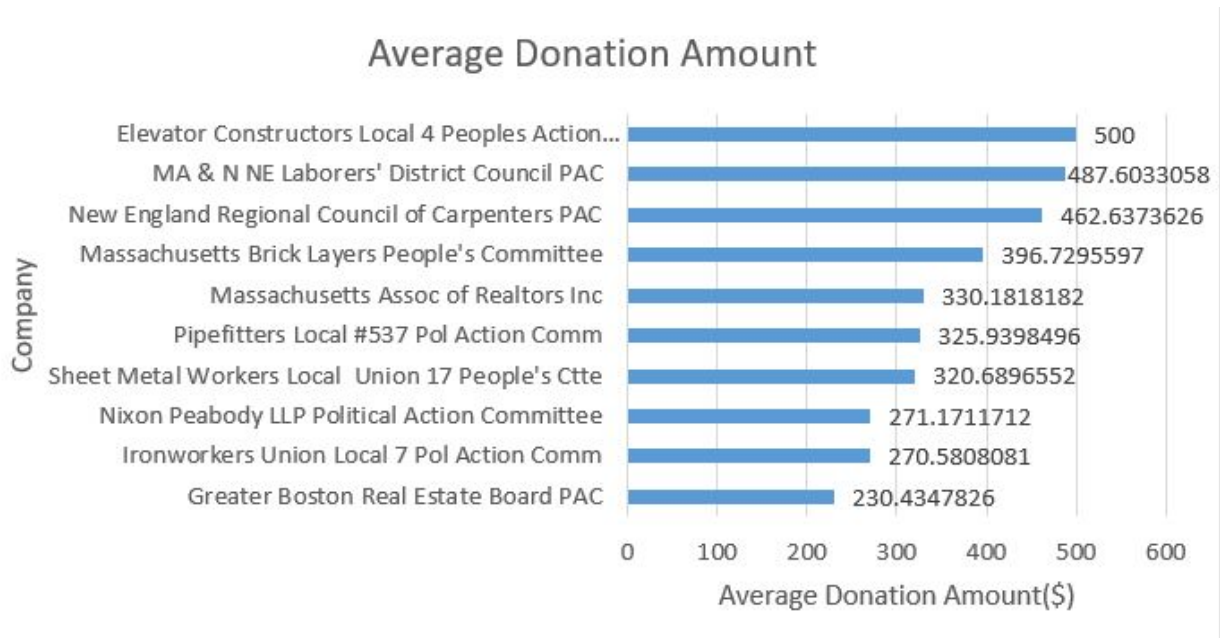


Figure 5: Top 10 average donation amount from 2015 to 2019

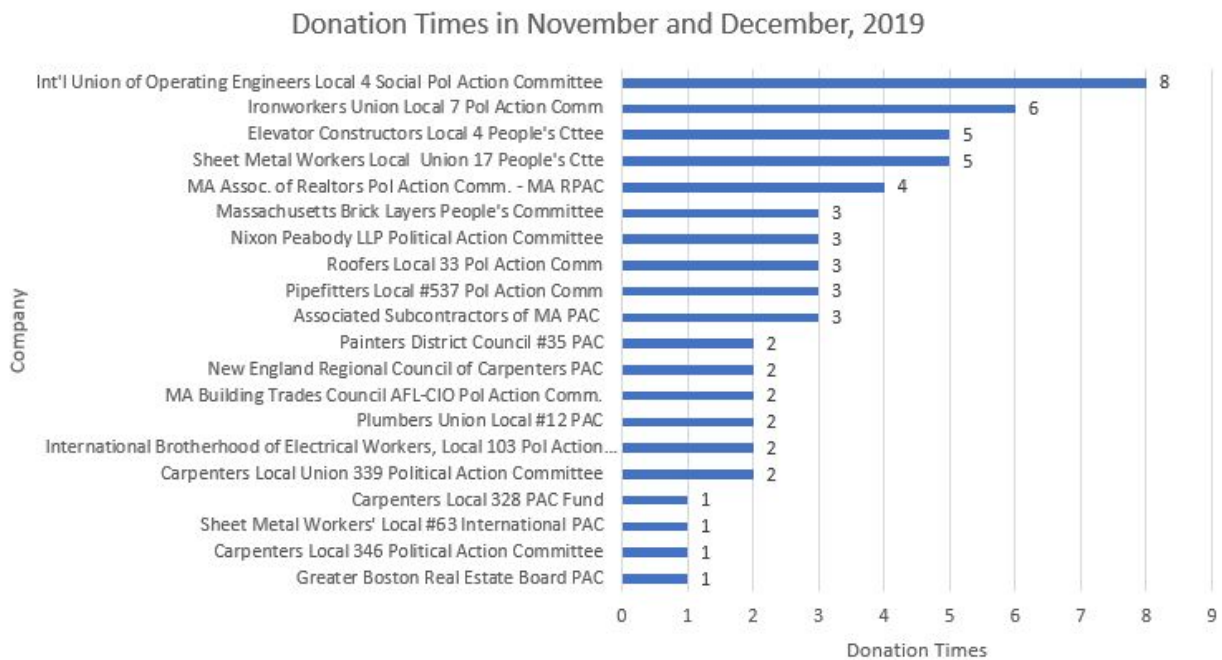


Figure 6: Companies which donated most times in November and December, 2019

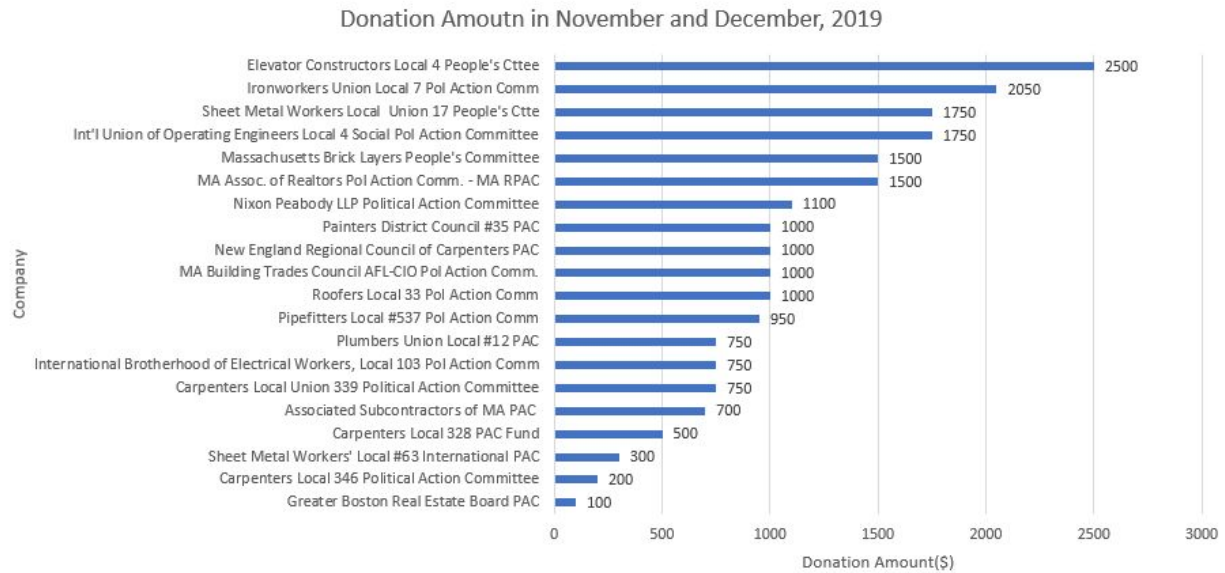


Figure 7: Companies which donated most money in November and December, 2019

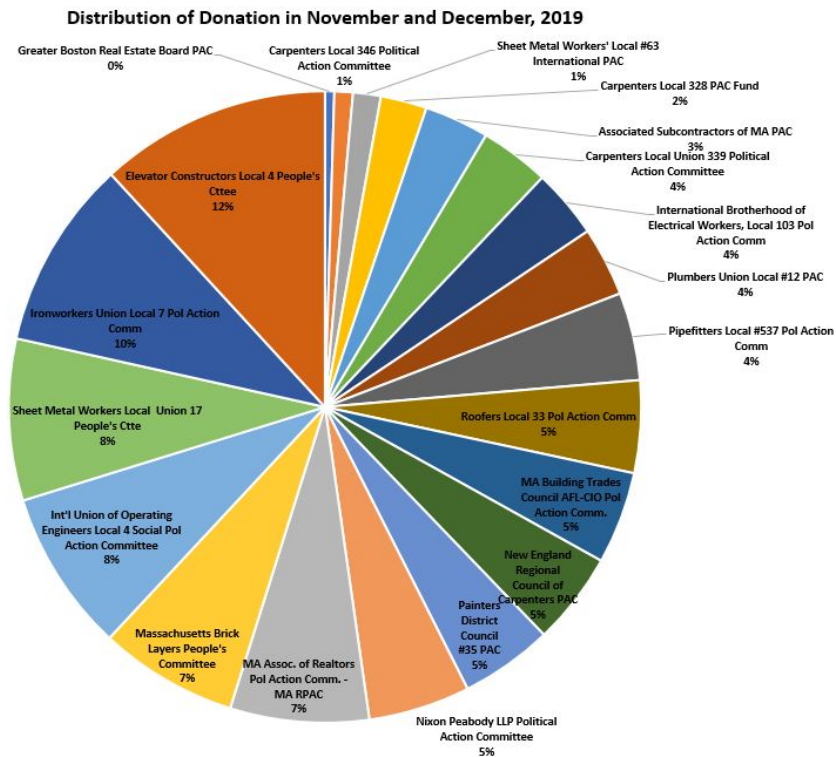


Figure 8: Distribution of donations in November and December, 2019

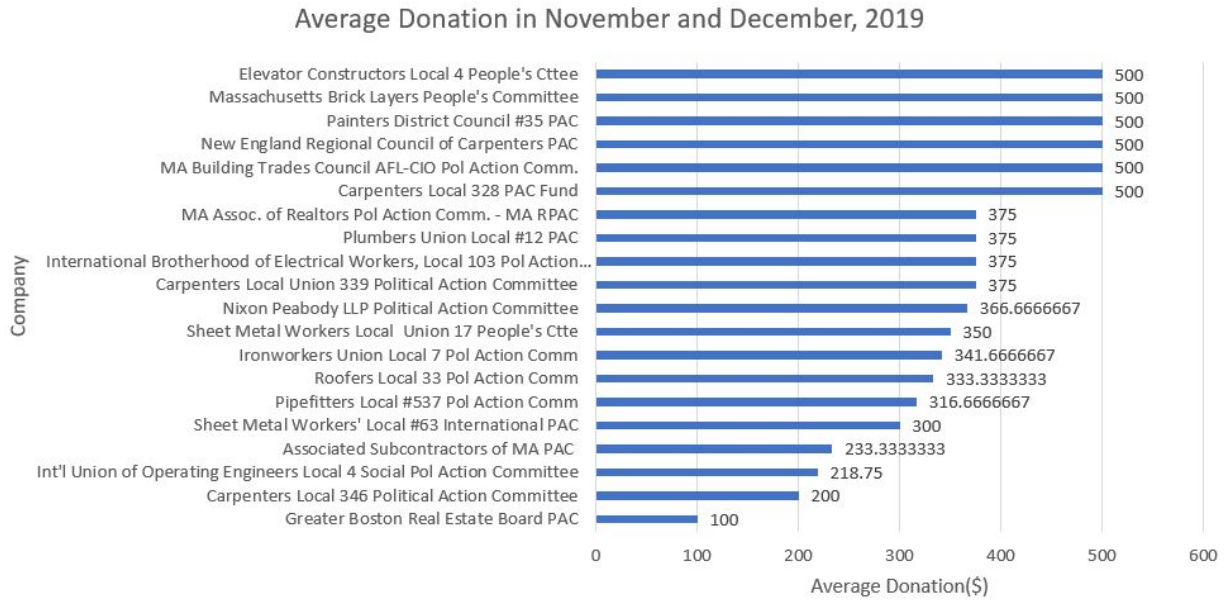


Figure 9: Average donation amount in November and December, 2019

## 5. Analysis

The question to be answered is “Which companies and unions and PACS are engaged in the political process through donations, i.e. donating most often to elected officials?”.

Actually, we have answered 3 related questions:

1) Who donated most often to elected officials?

From figure 2, we can see that *Ma & Northern NE Laborers' District Council PAC* donated most times from 2015 to 2019. It donated more than 600 times while others donated up to 275 times.

2) Who donated the most money?

From figure 3, we can see that *Ma & Northern NE Laborers' District Council PAC* donated most money from 2015 to 2019. It donated 295,000 dollars while others donated less than 100,000 dollars.

Also, from figure 4, there are more than 60 companies in total. “Others” include more than 50 companies, which took up only 27% of total donated money. However, the donation of *Ma & Northern NE Laborers' District Council PAC* took up 30%, which is more than the sum of 50 companies. No companies donated more than 10%.

3) Who made the most contributions?

From figure 5, *Ma & Northern NE Laborers' District Council PAC* also ranks top 2 in the average donation amount list. Combining figures 1, 2, 3 and 4, we can conclude that *Ma & Northern NE Laborers' District Council PAC* has made the most contributions from 2015 to 2019.

Also, we were asked to analyze the data in November and December, 2019.

1) Who donated most often to elected officials?

From figure 6, we can see that *Int'l Union of Operating Engineers Local 4 Social Pol Action Committee* donated most times in November and December, 2019. It donated 8 times while most companies donated less than 5 times.

2) Who donated the most money?

From figure 7, we can see that *Elevator Constructors Local 4 People's Cttee* donated most money in November and December, 2019. It donated 2,500 dollars while most companies donated less than 2,000 dollars.

Also, from figure 8, there are 20 companies in total. *Elevator Constructors Local 4 People's Cttee* took up 12% while others took up less than 10%.

3) Who made the most contributions?

From figure 9, there are 6 companies that have the same rank in the average donation amount list. *Elevator Constructors Local 4 People's Cttee* ranks 1st in

the average donation amount list, ranks 4th in donation times list, and ranks 1st in donation amount list. Combining figures 6, 7, 8 and 9, we can conclude that *Elevator Constructors Local 4 People's Cttee* has made the most contributions in November and December, 2019.



## 6. Project Complexity and Team Creativity

### 6.1 Challenges and obstacles

- 1) Because of the huge amount of data, we cannot go to the next page when searching for some letters which contain more than hundreds millions of links. We solved this problem by choosing to show all links when searching the letters. Also, we search “aa” or “ab” etc. to reduce the number of links each time.
- 2) It is our first time doing scraping and we are not familiar with it. We spent a lot of time studying it.
- 3) When we were scraping the website, at first we were using Octoparse to scrape the website since it is very convenient. However, we were rejected by the website due to the high frequency of visiting. Then we chose to write our own Python code to do the scraping since it can modify the code so that we can solve the problem.

We added a random *sleep()* function in the code to reduce the frequency of visiting. Also, we tried to restart the chrome driver every 100 times downloading so that we wouldn't be rejected. Then we could download PDFs without rejecting.

- 4) Another problem is that there are millions of PDFs to download. We have been downloading for a whole week and we got only about 100,000 PDFs. It may take more than a month to download all the PDFs. Then our project manager provided us the data from Campaign Finance Team since we are using the same website and data.

### 6.2 Limitation

Although we have solved the problems and collected data successfully, there are still some limitations in our project. All the PDFs from the Secretary of Commonwealth's website do not have exactly the same format. When we were extracting information from PDFs, we may skip some PDFs with different formats automatically. So we may lose some data, which may cause a slight bias to our results.

### 6.3 Team Creativity

When solving the rejecting problem, we have tried many ways to solve the problem. Finally we find our own way to avoid being rejected. Adding *sleep()* function and restarting chrome driver every 100 downloads can solve the problem well.

## **6.4 Conclusion**

In this project, we primarily learned two skills, scraping and extracting information from PDFs. These are completely new experiences to us. We have tried all kinds of ways to solve the problems in the process and then improved our ability of solving the problems. We really learned a lot in the project.