Deliverable 3

Xiaoyu An
Qing Han
Qihao Sun
Tingyi Zhang
Ruihong Zhu

Data Collection:

Approved Building Permits:

1. Raw data collection: filter data by mixed and Commercial and Mixed occupancy type, and download the csv file.
2. Extract features: use DataFrame to read csv file and extract: OccupancyType, Owner, Applicant, Address, City, State, Zip, Parcel_id.
3. Merge columns: merge “Address, City, State, Zip” to one column whose header is “Project Address”.

BPDA:

1. Raw data collection: Download all data as csv file.
2. Extract features: use DataFrame to read csv file and extract: ProjectID, ProjectName, Developer, Contact, Title, MailingAddress, Phone, eMail, ProjectStreetName, ProjectStreetNum, ProjectNeighborhood, ProjectZipCode.

City of Boston:

1. Raw data collection: Download data from City of Boston website as csv file.

ZBA Meeting:

1. Raw data collection: Download pdf documents from ZBA website for 2017, 2018 and 2019.
2. Extract features: Use the Tika library to parse the pdf files and use regular expressions to match and extract the value of BOA number, Applicant name and Applicant address. Write the three attributes to csv files.
3. Merge: Combine all csv files together to one file and clean the data by removing empty spaces and invalid values.

ZBA Decision:

1. Raw data collection: ZBA Decisions have less and cleaner data than ZBA Meetings, so we just copied all of it directly from the ZBA website to Excel and marked it as approved or denied.

2. Extract features: Used Text to Column function in Excel to split BOA number and address by delimiters.

OCPF:

1. Raw data collection: by searching keywords of Contributor/Occupation/Employer we can find out the raw data including attributes: Date, Contributor, Address, City, State, Zip, Occupation, Employer, Principal Office, Amount, CPF_ID, Recipient, Tender_Type_ID, Tender_Type_Description, Record_Type_ID, Record_Type_Description, Source_Description.
2. Duplication deleting and abnormal data replacement: remove duplication by judging if Date, Contributor, Address, Amount, CPF_ID are the same at the same time; replace empty attributes of entries with NaN.
3. Data integration: integrate cleansed data together, we can get OCPF_All_2019 (No Keywords filtering), OCPF_RealEstate_2019.
4. Data merging: extract the names of people who received donations from Master Data, extract the terms of OCPF from OCPF_All_2019 according to the names, then remove duplicates, we can get OCPF_temp_2019; Merge OCPF_temp_2019 with OCPF_RealEstate_2019, remove duplicates, we can get OCPF_RealEstateAll_2019.
5. Remove punctuations, remove stopwords, tokenize, lemmatize and get the word-stem for some particular attributes of terms of OCPF_RealEstateAll_2019: For example, a same company might have several similar but different representations among terms of OCPF data: TOM & John Law Firm, Tom and John Law Firm, Tom & John Law Firm, LLC. We know these are the same companies, but when doing data analysis, they would be recognized as three different companies which cause errors. So, to avoid these kinds of errors, we further processed data of OCPF_RealEstateAll_2019 as stated at the beginning of this graph.

Data Cleaning

Master data file:

The data from Approved Building Permits, BPDA, City of Boston, and ZBA is merged together to form a master data file, which can be viewed as a database which contains all the people who are affiliated to the real-estate industry. The attributes include Data Source, Project, Case Number, Applicant, Applicant Address, Decision, Developer and Project Address.

Final Result**:**

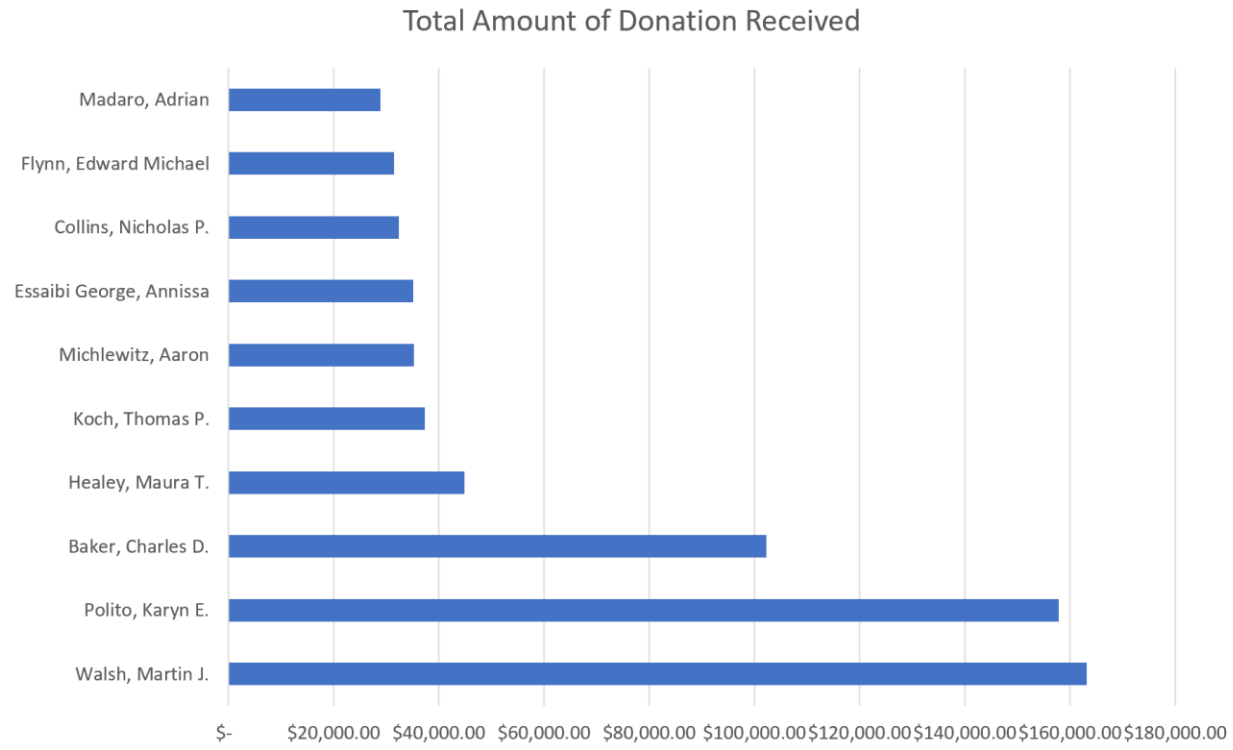
1. The comparison between OCPF and Master data file: The applicants who show up in the master data file should be considered as someone who connected to the real-estate industry. Therefore, the full 2019 OCPF file is filtered by the applicant names present in the master data file and only those individual type records which have the contributor name also show up in the master data file will be keeped.
2. The OCPF keyword data: the full 2019 OCPF file is filtered again by using the Occupation and Employer keywords provided by Spark team and the results also includes many people who are in the real-estate industry.
3. These two parts of OCPF records are merged together and all the duplicates are removed based on index. The final result contains 7146 OCPF records where all the contributors should be considered as someone connected to the real-estate industry.

Analysis:

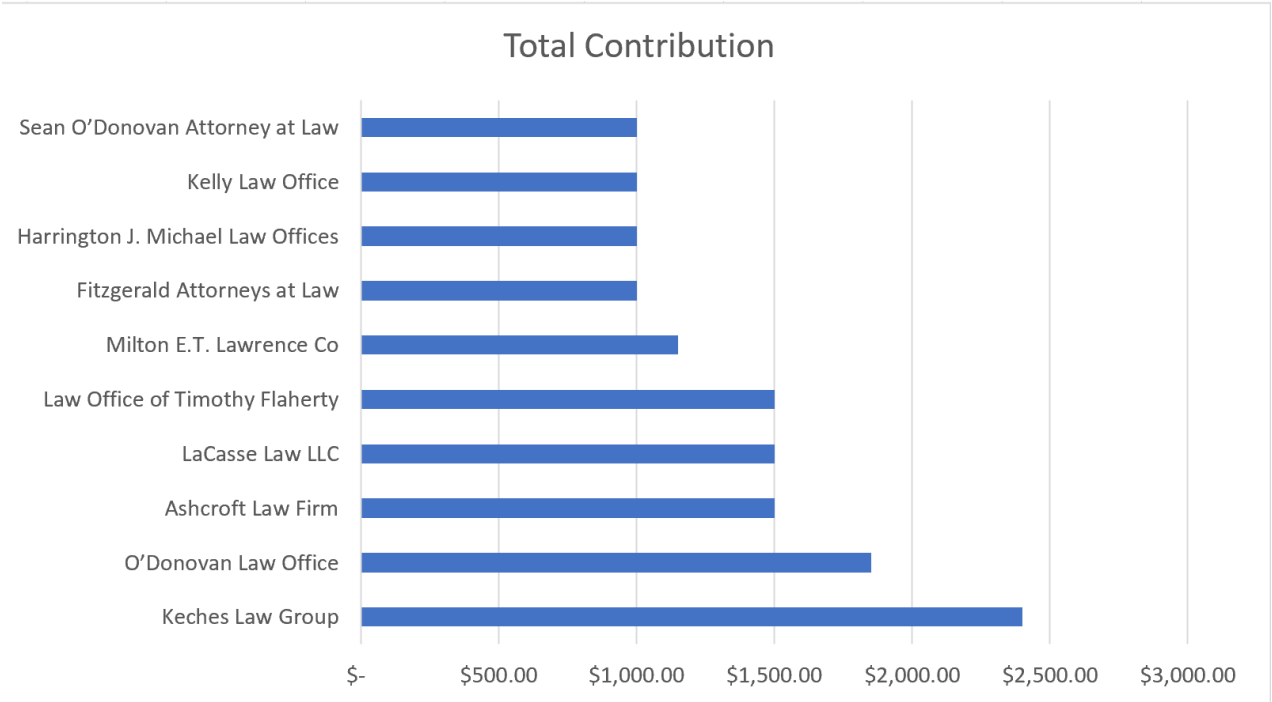
1. We have found out which group of people receive the most amount of contribution from the real-estate industry. In order to do that, we take the final result we have, group it by the Recipient column and calculate the sum of donations received by each of them. The ten people who receive the most amount of donations are shown below in the Presentation section. These ten people are all key members of the State of Massachusetts and City of Boston, including Mayor of Boston, Governor, lieutenant Governor and Attorney General of MA, and several City Councillors and State Representatives in the Boston area. It is clear that these key members have significant influence in the real-estate industry and based on the rankings, the members of the Executive Branch are the most important ones to the real-estate industry since the top three recipients are all from this Branch. Followed by the Judicial Branch and the legislative Branch. It is reasonable to conclude that the people who are in charge of the daily operation of the government have the most amount of influence in the real-estate industry.
2. Also, there are some works we are doing that can reveal more details and relationships between political powers and real-estate industry. As stated above, we have found that which part of Boston political powers may have the biggest influences on real-estate industry (based on the amount and times of contributions they have received), also we found what real-estate industry related companies are most active in the contribution activities, so we expect to find a positive correlation between these influential political powers and the activeness of these stated companies, the final results will be presented in the final report.

Presentation:

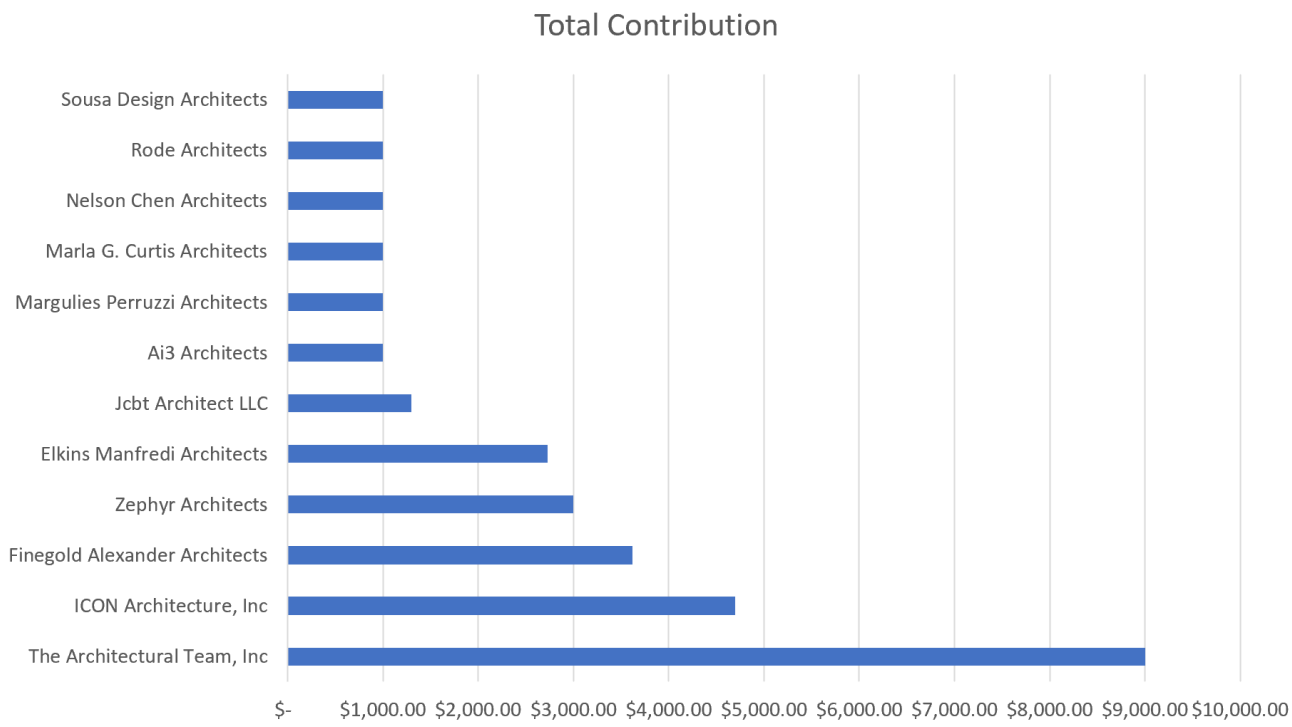
Top 10 recipients who receives the most amount of donations from the real-estate industry:



The Top 10 Law Firms who donate the most amount of money:



The Top 12 Architect Firms who donate the most amount of money:



Data Limit:

Our data only focus on the donation activities in 2019 and it will be better to extend it to recent three or five years in order to show a more accurate result. We can also analyze how the donation pattern changed over these years and try to find out what causes these changes.

Another limit is the amount of Master Data. If we have more time, we will be able to collect more real-estate related data and extract more people from this industry. We only extract about 7000 OCPF donation records by using the current Master Data file as the filter and it's difficult to launch deeper analysis with this amount of data. For example, we tried to analyze the relationship between ZBA approval rate and the amount of donation contributed by applicants, but then we realized that there's no overlap between ZBA applicants and OCPF contributors in 2019. If we use more OCPF data from previous years or find another source for approval rate, then we will be able to finish this analysis.

Project Complexity and Team Creativity

Challenges:

1. Since the data provided is huge, the process of extracting features we need and clear data is crucial. The representors for projects in different sources are

different, ex. applicants, owner, developer, so we have to figure them out and unify the term of it. Sometimes the address and even the email became the key to identify the projects.

2. Since our data sources are very complicated which include website content, PDF files, Excel files, txt files and so on, there is a great amount of preprocessed work to do in order to get useful data. Also, it requires a lot of time to merge, match and analyze all these data.
3. Since we cannot meet with each other in person during this period, it's difficult to communicate and collaborate efficiently to get a good result, sometimes there are duplicates and gaps between our works, which need extra time to be fixed.

Creativity:

We use different technologies based on the format of data. For instance, the data in Approved Building Permits and BPDA is well formatted and clear, so the data can be imported and processed using DataFrame. While the content in OCPF and ZBA Meeting is much more complicated, we extract useful information by using regular expressions, removing punctuations, removing stopwords, tokenize, lemmatize and get the correct value we want for certain attributes to clean the data.