Analyzing Minority Business Enterprises in the Greater Boston Area

Justin Ingwersen | Umang Desai | Gahouray Dukuray | Ashwini Kulkarni

Introduction & Datasets

For our final Data Mechanics project, our group decided to work with the Pacesetters program to conduct research and analysis of small and large Minority Business Enterprises (MBE) in the Boston, Massachusetts area. The Pacesetters Program is a partnership of large and mid sized Greater Boston Chamber of Commerce member organizations who use their collective purchasing power to create opportunities at scale for local enterprises of color. One of the goals of the Pacesetters Program is to make sure minority-owned businesses, businesses owned and operated by members of minority groups such as African American, Native American, Asian, or Hispanic American, have MBE Certifications provided by the federal, state or local government.

We were specifically tasked to use datasets, provided to us by our Pacesetters Program partner, that gave us data on MBEs, some examples of data we were provided are the number of certified MBEs are in the Boston area and the types of industries these MBEs classify themselves with. In our project, we used the following datasets:

- MBEs that are certified provided by MassHousing
- MBE Businesses provided by Secretary of Commonwealth
- Zip Codes in the Boston area we are focusing our research on provided by our Pacesetters partner
- MBEs and non-MBEs along with the industries they are associated with provided by FactFinder
- Number of MBEs by state provided by Minority Business Development Agency(MBDA)

In order to better utilize the datasets we were provided with, there was a lot of transformations (of the datasets). Transformations including selections and projections were in heavy use to get two big masterlist of datasets we wanted to have in the end: a dataset of MBEs and a dataset of non-MBEs in our area of interest. We also wanted the MBE dataset to have a column that would note if the business was certified or not.

Algorithms & Analysis

In order to complete the specifications laid out by our partner, we had to perform several data manipulations and execute many algorithms.

Our first algorithm was used to get data from our web sources. We used MassHousing and the Secretary of Commonwealth as our two primary sources. Using a projection on the data fed in, we projected out the non-MBEs so that the result was a master list of all our certified MBEs. One issue we had to overcome was that the zip codes from the Secretary of Commonwealth were stored as integers instead of strings. To overcome this, we used a pandas method to convert the column into all strings, and then used a lambda function as shown to append 0s as needed:

```
validZipsDF['Zip'].apply(lambda zipCode: ((5 - len(zipCode))*'0' + zipCode \
                                         if len(zipCode) < 5 else zipCode)[:5])
```

By the end of the algorithm, the master list was set after several more projections to get the right columns we wanted as well as a selection to remove all the zip codes our partner was not interested in.

Building off of this master list, we aimed to also standardize the industry column and worked to build an algorithm to sort companies into various industries. Prior to this, we were limited by the datasets and the information they provided about where each company fit in industry-wise. One dataset was very vague and the other contained a verbose description. This algorithm, located in mergedList.py, aimed to merge together both MBEs and non-MBEs as well as categorize all the companies into industry “buckets”. We used a dictionary with 46 keys where the value of each key was a list of keywords to look for. Once a company fit into one of the keywords of a specific key, we set the industry of that company to that key (using a brand new column to be appended to our dataframe). This dataset would allow us to do much more advanced analyses and calculations.

The next part of our project involved an optimization problem. The code to accomplish this can be found in optimalLocation.py. Our goal was to find the zip codes where the most MBEs could move to or start their company in. To do this, we needed to add multiple constraints to our problem. The first constraint was that in order for an MBE of a particular industry to be hypothetically added to a zip code, there needed to be a non-MBE business of that industry already existing. This ensures that the MBE could survive in this particular area and that it will not starve as a business because its industry is not suited for the environment. The second and more obvious constraint we checked for is whether or not another MBE existed in that zip code already with the same industry. If this was true, we did not attempt to add the new business as the competition would not be healthy for both businesses.

To begin finding the number of MBEs that could be added to each zip code before it became saturated and the constraints could no longer be satisfied, we first split our master merged list of companies by zip code, breaking them off into separate dataframes. Next, we used a loop to build two sets - one for the banned industries, i.e. where MBEs already occupy, and the other for the non-MBE industries that do exist. Then, this allowed us to simply iterate through the master merged list and attempt to add each MBE business into the particular zip code assuming it meets our constraints. Finally, the algorithm spits out a dataframe with this count and the zip code as part of a dataframe.

Furthermore, we performed some statistical analysis on our data to better understand how the industries are interacting in the Greater Boston Area. Our goal was to find the correlation coefficient between each pair of industries to figure out how correlated these two particular industries are to each other. In other words, we wanted to see where two industries thrive together, or are completely absent at the same time. We needed two vectors per pairing of industry with each dimension as the count of the industry in a particular zip code. To be efficient, we used a dictionary to first parse through our data and calculate the count of each industry in the zip codes. This data structure would speed up our calculations as we looped through and compared industries directly to one another. After building up vectors for each pairing, we had to remove duplicates as our code was producing pairings such as “Food:Architecture” and “Architecture:Food”. These two would essentially produce the same correlation coefficient, so we only needed one of these. After calculating the coefficient between each industry, we appended the pairing key, i.e. “Architecture:Food” into a column as well as the coefficient value into its own column at the same index. Thus, at the end of our algorithm, we were able to output this statistical analysis as a dataset.

Summary

Problem statement:

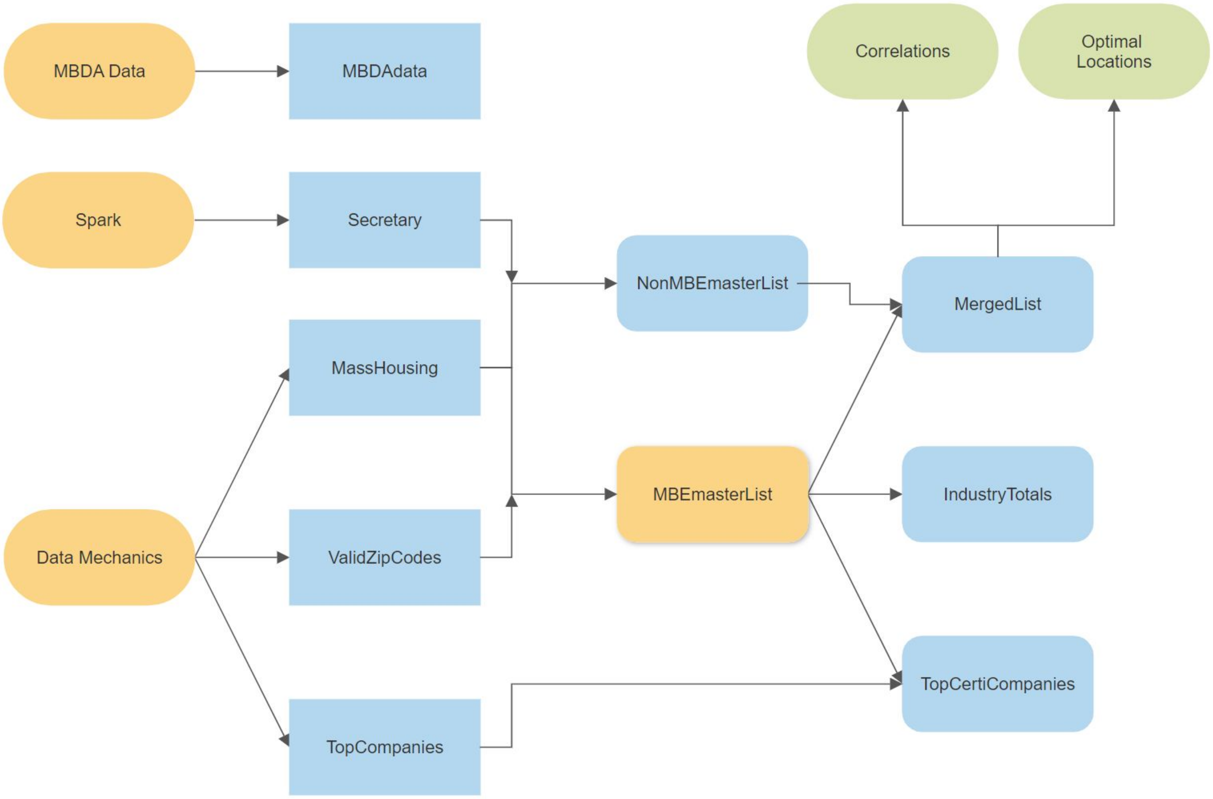
While working with the Pacesetters, we were tasked with finding a to locate and support upcoming MBEs based current data about established MBEs around the I-98 belt.

Our approach to a solution:

We decided that the best way to use this data was to optimally locate zipcodes where MBEs in particular industries can be successful. To do this, we assessed the locations of MBEs and non-MBEs by 46 industries. We then generated correlation statistics between every two pairs of industries in each zip code. Let us use an example to see how we can use this information: let's say that on average, if there are 10 restaurants and 18 businesses per zip code and we found that the correlation is 0.8, we can conclude that there is a strong correlation between location of businesses and restaurants. In a another zip code if we find that there are 10 businesses but only 1 restaurant, we can suggest that an MBE in the food industry could benefit but opening in this given zip code. We then generated some optimal zip codes where MBEs could be successful.

Data Sets and creation summary:

Below is a flow diagram showing how we first extracted data from our three data portals and generated different data frames to find optimal locations for the addition of new MBEs.

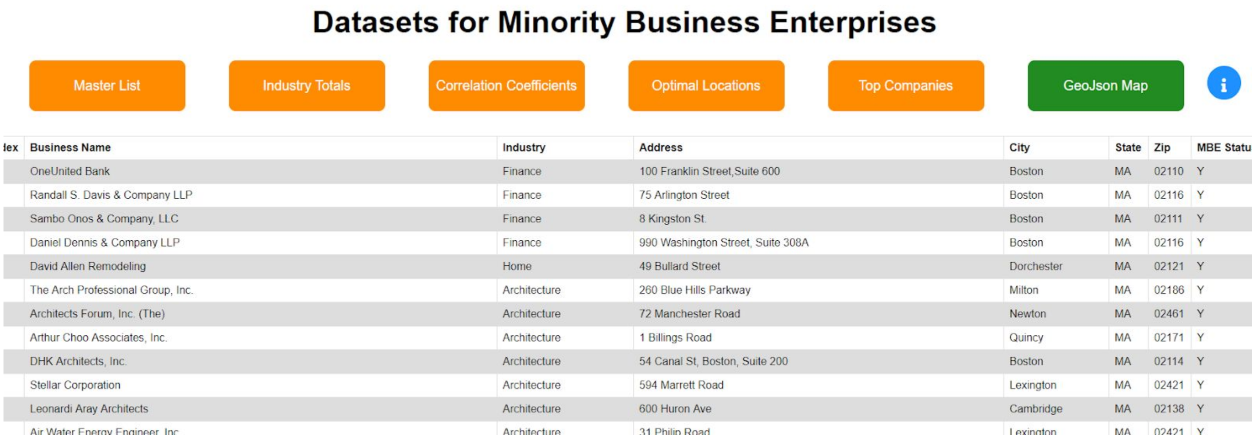


We used data provided by our Spark partners and from other governmental sources to generate specific data sets about the relevant companies and locations we aimed to target. Specifically, differentiating MBEs and non-MBEs. Once we had extracted these specific records, we generated some non-essential data points such as “Top Certified MBEs” and “Industry-wise total for MBEs” for general demographics for MBEs around the I-98 belt. Using the master lists for the MBEs and non-MBEs we computed two things: the correlation between each industry with every other industry (1080 total pairs) and the optimal zip codes where a new MBE can flourish with the number of additions per zip code.

Conclusion and Findings:

Out of 1080 pairs of industries, the most sensible pair having the highest positive correlation is Carpentry -- Janitorial Services with a correlation of 0.66. This means that, around the i-98 belt, blue collar jobs are usually located in a similar location! Another pair to that supports this claim is Masonry -- Landscaping which has a positive correlation of 0.62.

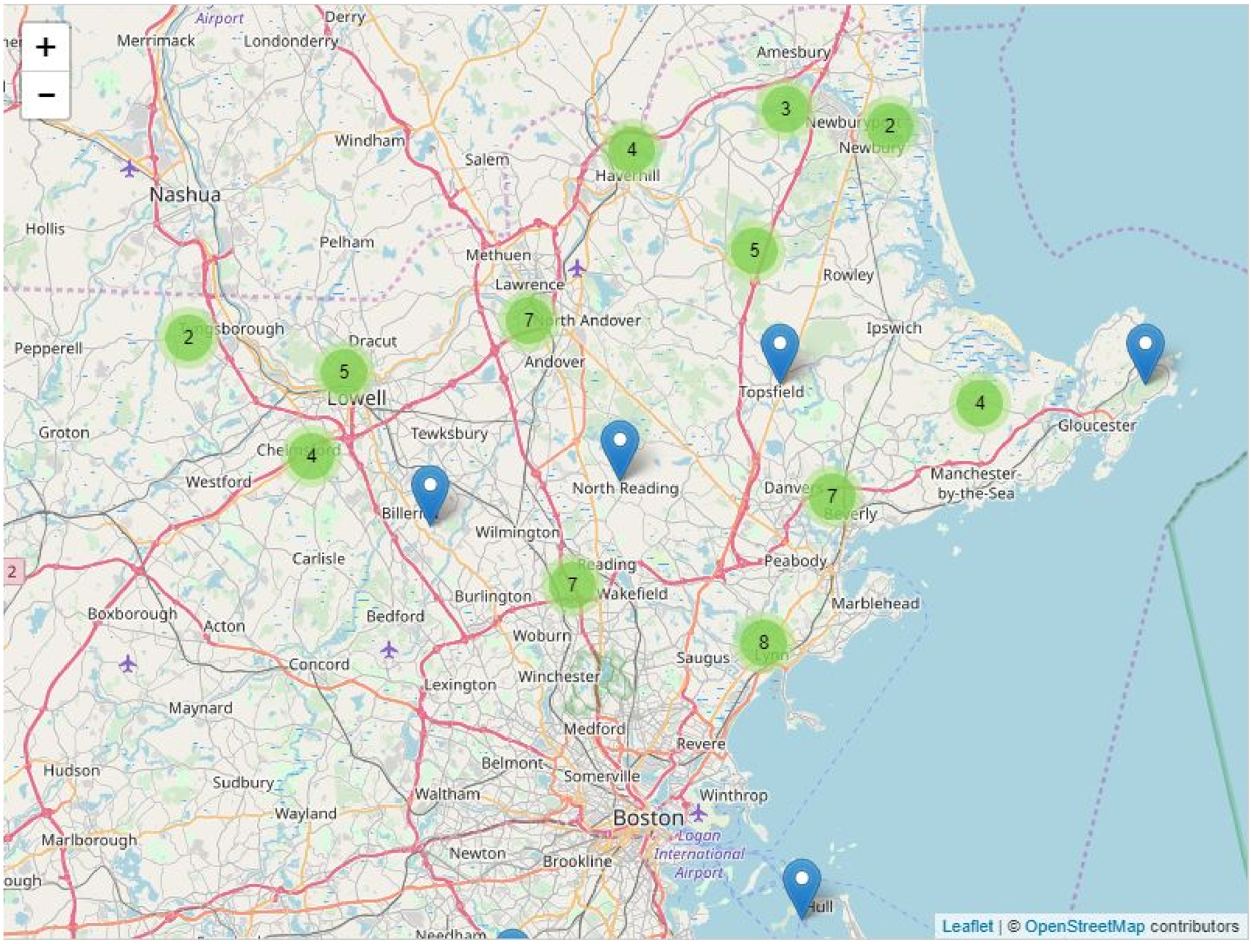
Out of the 80 zip codes surrounding the I-98 belt, the zip codes with the most number of MBE additions we suggest are: 01801, 02169, and 01880 (8 additions/zip code). This means that the cities of Woburn, Quincy, Braintree, and Wakefield and good locations for new MBEs to start and be successful.



Datasets for Minority Business Enterprises

dex	Business Name	Industry	Address	City	State	Zip	MBE Statu
	OneUnited Bank	Finance	100 Franklin Street,Suite 600	Boston	MA	02110	Y
	Randall S. Davis & Company LLP	Finance	75 Arlington Street	Boston	MA	02116	Y
	Sambo Onos & Company, LLC	Finance	8 Kingston St.	Boston	MA	02111	Y
	Daniel Dennis & Company LLP	Finance	990 Washington Street, Suite 308A	Boston	MA	02116	Y
	David Allen Remodeling	Home	49 Bullard Street	Dorchester	MA	02121	Y
	The Arch Professional Group, Inc.	Architecture	260 Blue Hills Parkway	Milton	MA	02186	Y
	Architects Forum, Inc. (The)	Architecture	72 Manchester Road	Newton	MA	02461	Y
	Arthur Choo Associates, Inc.	Architecture	1 Billings Road	Quincy	MA	02171	Y
	DHK Architects, Inc.	Architecture	54 Canal St, Boston, Suite 200	Boston	MA	02114	Y
	Stellar Corporation	Architecture	594 Marrett Road	Lexington	MA	02421	Y
	Leonardi Aray Architects	Architecture	600 Huron Ave	Cambridge	MA	02138	Y
	Air Water Energy Engineer, Inc	Architecture	31 Philip Road	Lexington	MA	02421	Y

Screenshot of our webservice API. Each of the orange buttons will hit our API and and the API will give back a dataset as it’s response.



Screenshot of our data visualization. This was created using Leaflet and allows the user to see the number of MBEs/nonMBEs per zip code.

Future Work:

The inequalities faced by the minority communities in terms of business opportunities, financial aid and the like, have served as obstacles in regards to limiting the height that those individuals can reach. As our country journeys towards a better future with flatter playing fields, a tool that identifies and suggests areas where MBEs are not as present could certainly go a long way. We started the project on a relatively small scale, with but a portion of zip codes in Boston, MA, and the room for development and growth is unimaginable, especially since we barely got the time to get our feet off the ground in the span of a semester. Under the Spark! Pacesetters project, we were able to learn of the types of resources & datasets that span this niche (in terms of readily-available information, a potential for non certified MBEs, etc.) and believe that with this knowledge and modification to our product, we can expand across the city, cities, states, and ultimately the country. By providing such assistance to offices, such as that which we worked with in Boston, they can reach and target a far greater audience and provide accurate data to show the reality of community situations and to aid those who need a helping hand in getting the proper government funding they are eligible for. In the end, the goal is to provide MBEs (and potentially other groups) a chance, and the role of our field is to build help build that bridge.