Capstone project- The battle of neighborhoods

Optimal neighborhood recommender for food bank

Section: 1:

Problem Description:

Food wastage is a serious issue in United States, approximately 125 to 150 billion pounds of food or $220 billion is being wasted every year. Meanwhile, 12% of american households are food insecure according to https://foodprint.org/issues/the-problem-of-food-waste/. A major step to reduce food insecurity would be to divert or minimize this food wastage. Most of the food that is being wasted is edible and nutritious, this food majorly comes from restaurant chains, coffee houses, and bakeries.

Our aim here is to build a food bank that collects edible food from the above food hubs and keeps it distribution ready for people who visit the food bank. Some of the major challenges to this food bank would be;

* Proximity to restaurants

* How densely are the restaurants located?

* How famous or popular are these restaurants

* How easy is it to commute to this food bank for consumers using public transportation

We have chosen Boston city as high level location of the food bank, as Boston has lot of food chains and has a higher amount of food waste, so we could act upon this and divert the food to the needy. Now, we

need to identify the correct neighborhood or locality in which we need to start the food bank so that the consumers would be able to visit the food bank with less cost of transportation. Our end goal from this project is to leverage data obtained using foursquare API, and clustering methods to identify the best neighborhood to start the food bank.

Who would be interested in this project:

1. Food bank representatives who want to open a new office, this analysis will provide complete in-depth analysis of neighborhoods that are potential locations for the food bank
2. Social activists working to reduce hunger problems can understand the analysis and file petitions with respective associations to start a food bank
3. New Data Scientists who want to learn and implement exploratory data analysis techniques, scrape required data, perform analysis, and narrate a story on how the problem could be tackled

Section: 2:

Data Preparation:

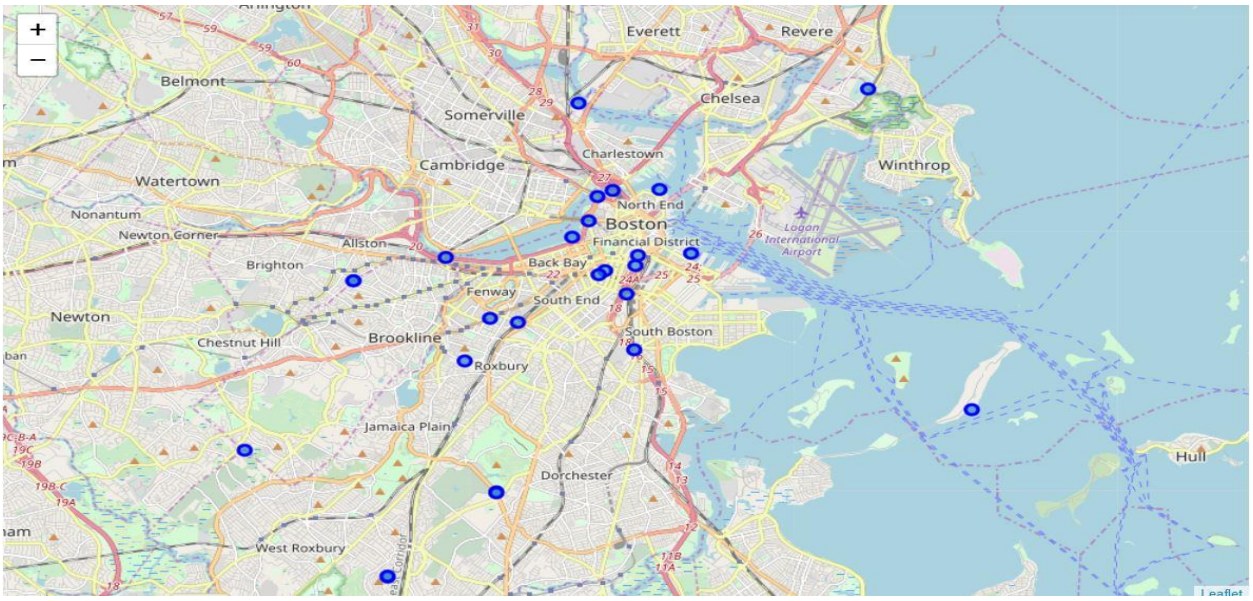
We extracted neighborhoods data from Boston Data web page. Using BeautifulSoup4 package we scraped the web page to build a data frame that consists of Neighborhood name, latitude, and longitude. Neighborhood names are in the form of 'Chinatown' and 'Downtown', we have a total of 25 neighborhoods, and we will analyze most visited venues for each of these neighborhoods to identify which is a potential place for opening the food bank. Initially after reading data from json file we obtain data that looks like in the below picture;


```
{'type': 'FeatureCollection',
 'crs': {'type': 'name',
  'properties': {'name': 'urn:ogc:def:crs:OGC:1.3:CRS84'}},
 'features': [{'type': 'Feature',
   'properties': {'Name': 'Roslindale', 'density': 5.58},
   'geometry': {'type': 'MultiPolygon',
    'coordinates': [[[[-71.12592656722312, 42.27200445346726],
       [-71.12574734036883, 42.2723385402892],
       [-71.12566364121143, 42.272474315069964],
       [-71.12555022588188, 42.272570167080346],
       [-71.12572827873186, 42.27248267618011],
       [-71.12637716166857, 42.272159141095266],
       [-71.12651460459686, 42.272090883284626],
       [-71.12659547750106, 42.27217280198417],
       [-71.12665777013348, 42.27223592040908],
       [-71.12690977451678, 42.27209079924884],
       [-71.12725672928872, 42.271994756984384],
       [-71.12739754630685, 42.271954634551136],
       [-71.12760296838418, 42.27190191373231],
       [-71.12786708093122, 42.271952962649934],
```

After further processing the json data we obtain a data frame that contains neighborhoods and their respective latitude and longitude values like in the below snapshot;

	Neighborhood	Latitude	Longitude
0	Roslindale	-71.125927	42.272004
1	Jamaica Plain	-71.104992	42.326093
2	Mission Hill	-71.090434	42.335761
3	Longwood Medical Area	-71.098108	42.336722
4	Bay Village	-71.066629	42.348774
5	Leather District	-71.058378	42.349822
6	Chinatown	-71.057905	42.352370
7	North End	-71.051995	42.368827
8	Roxbury	-71.096459	42.293224
9	South End	-71.068340	42.347742
10	Back Bay	-71.075688	42.356908
11	East Boston	-70.995462	42.393930
12	Charlestown	-71.074160	42.390503
13	West End	-71.068850	42.367187
14	Beacon Hill	-71.071409	42.361178
15	Downtown	-71.064776	42.368815
16	Fenway	-71.110108	42.352072
17	Brighton	-71.135174	42.346006
18	West Roxbury	-71.164702	42.303830
19	Hyde Park	-71.125881	42.272105

Using folium package, we plotted Boston city map using the above data frame, all the neighborhood locations are plotted on the map using respective coordinates, this can be seen in the below map:



Using foursquare API, I obtain the 100 most visited venues within 500 meters of each major neighborhoods in the above obtained data frame. Venues range from restaurants, coffee shops, gym, clothing stores, and movie theatre, below we can see how data from foursquare API looks like;

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Roslindale	42.272004	-71.125927	Weider Park	42.271045	-71.123697	Park
1	Roslindale	42.272004	-71.125927	Smith Field	42.271255	-71.129020	Baseball Field
2	Roslindale	42.272004	-71.125927	Conley School Playground	42.274607	-71.127877	Playground
3	Roslindale	42.272004	-71.125927	MBTA Commuter Rail- Province line	42.269557	-71.122011	Train Station
4	Roslindale	42.272004	-71.125927	Kelly's Liquors	42.272590	-71.119968	Liquor Store
5	Jamaica Plain	42.326093	-71.104992	The Frogmore	42.322489	-71.108238	Southern / Soul Food Restaurant

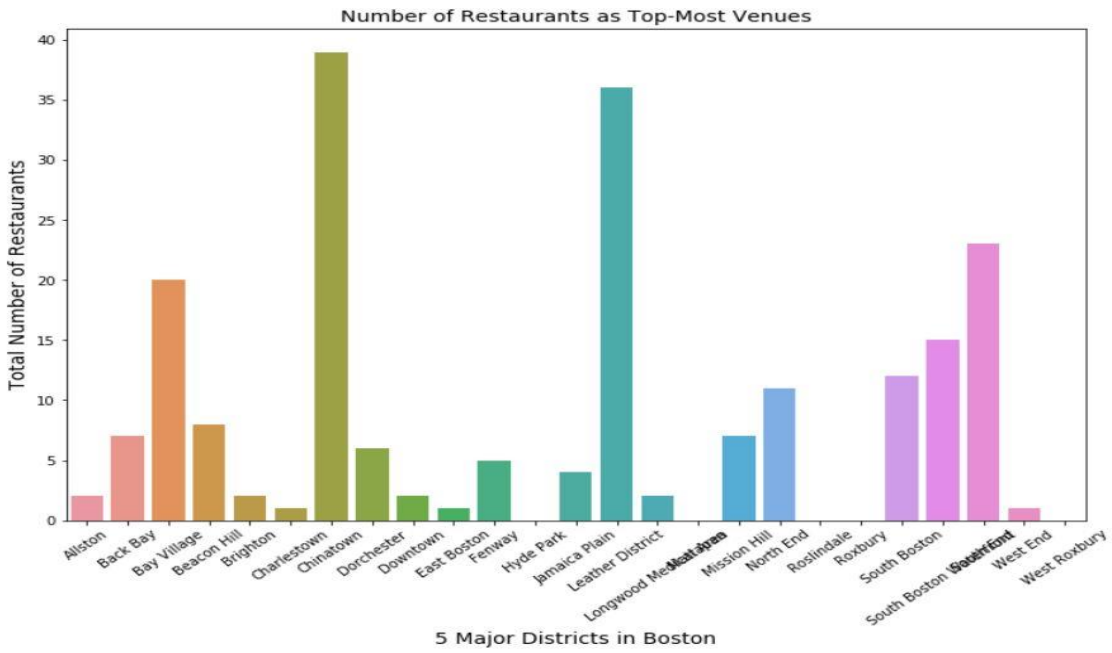
In the above data frame, we can see that we have neighborhood, latitude and longitude associated to that neighborhood, one of the 100 most visited venues, latitude and longitude associated with that venue, and the venue category.

Section: 3:

Methodology:

Exploratory Data Analysis:

For each neighborhood we have obtained top 10 most frequently visited venues and categories of each of the venues. For our problem, we are interested in venue categories that are associated with food. So, we calculate number of venues that have ‘restaurant’ keyword in their venue category and sum up total venues of this type for each neighborhood. In below picture we can find this visualization,



After obtaining the above graphic, we build a data frame that consists of top 5 venues for each neighborhood, we perform one hot encoding on venue categories to create a data frame, then use pandas groupby on neighborhood column, while grouping by we calculate mean value, after this we transpose the data frame and arrange the data frame in descending order, after performing the above operations we get the following data;

```
----Charlestown----
            venue  freq
0     Coffee Shop  0.12
1         Theater  0.06
2     Pastry Shop  0.06
3            Café  0.06
4  Clothing Store  0.06


----Chinatown----
                venue  freq
0         Coffee Shop  0.09
1  Chinese Restaurant  0.09
2    Asian Restaurant  0.07
3             Bakery  0.06
4      Sandwich Place  0.05


----Dorchester----
                 venue  freq
0          Coffee Shop  0.05
1  American Restaurant  0.05
2     Department Store  0.05
3        Metro Station  0.05
4                Hotel  0.05


----Downtown----
          venue  freq
0          Park  0.09
1           Bar  0.07
2   Donut Shop  0.07
3   Pizza Place  0.07
4          Café  0.04
```

From the above snapshot we can see that for each neighborhood we have the top 5 venues that have been most visited. Using further processing we obtain top 10 most common venues for each neighborhood, below is a snapshot of this dataset;

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allston	Pizza Place	Pharmacy	Bar	Convenience Store	Liquor Store	Music Venue	Plaza	Chinese Restaurant	Tea Room	Park
1	Back Bay	Pizza Place	Italian Restaurant	American Restaurant	Café	Gift Shop	Clothing Store	Scenic Lookout	Coffee Shop	Park	Breakfast Spot
2	Bay Village	Seafood Restaurant	Hotel	Theater	Bakery	Steakhouse	Sandwich Place	Coffee Shop	Mexican Restaurant	Performing Arts Venue	Pizza Place
3	Beacon Hill	Hotel Bar	Pizza Place	Food Truck	Gift Shop	Italian Restaurant	American Restaurant	Gourmet Shop	Plaza	Cycle Studio	Pub
4	Brighton	Pizza Place	Pharmacy	Bar	Convenience Store	Liquor Store	Music Venue	Plaza	Chinese Restaurant	Tea Room	Park

In the above picture we can see that there are different venues such as bar, park, restaurant, and theater.

Clustering:

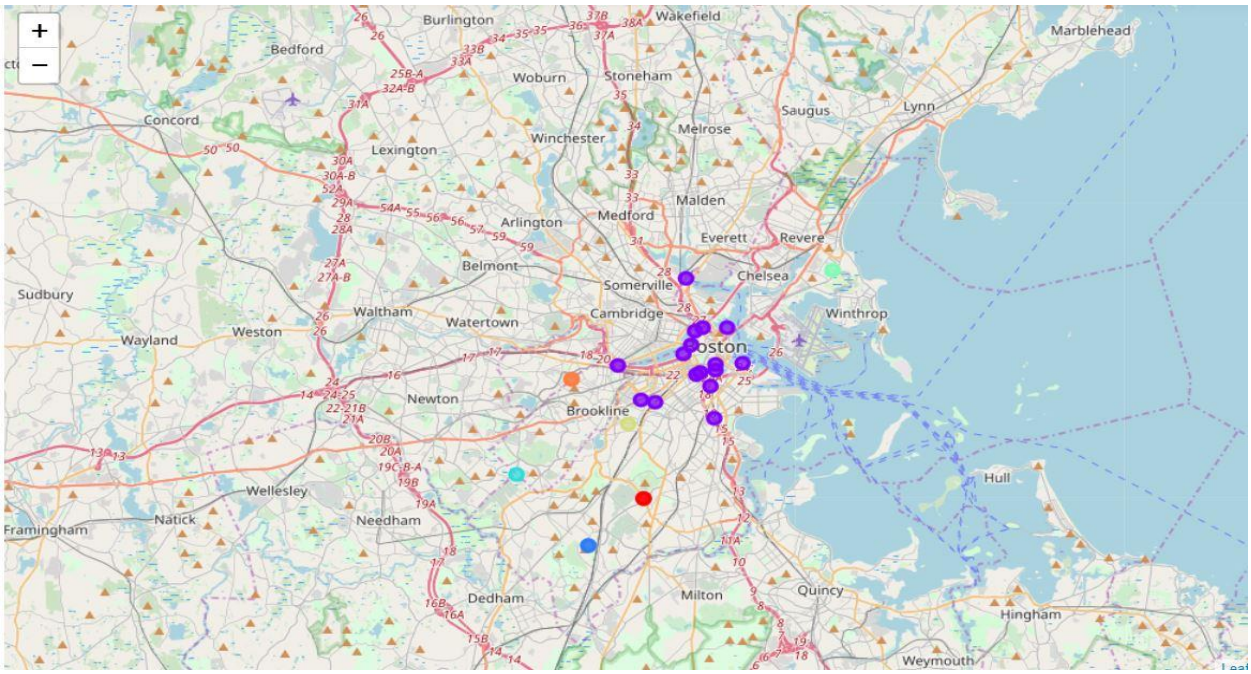
We have top 100 venues for each neighborhood, using this data we perform one hot encoding to obtain a data frame that contains numeric value for each venue that has been obtained after taking mean of this venue category across the neighborhood. Using this data frame, we perform K Means clustering. K means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters). After Iterating through different possible number of clusters, we have chosen number of clusters to be 7.

Here, we are trying to link each neighborhood to a cluster label, and identify which clusters have a greater number of food related venues in the top 10 venues for that neighborhood. If a neighborhood has a greater number of food related venues that means it is a food hub and possible food wastage is higher, and if this can be channeled properly to the food bank then we can reduce the wastage. If we can identify multiple neighborhood in the close vicinity that have majority of top 10 venues to be food related, then we suggest a location in this neighborhood locations for starting a food bank. As there are good number of food related places close by, we could collect good amount of food leftovers that can then be distributed for people in need, and this provides a good location for opening the new food bank like Greater Boston Food Bank.

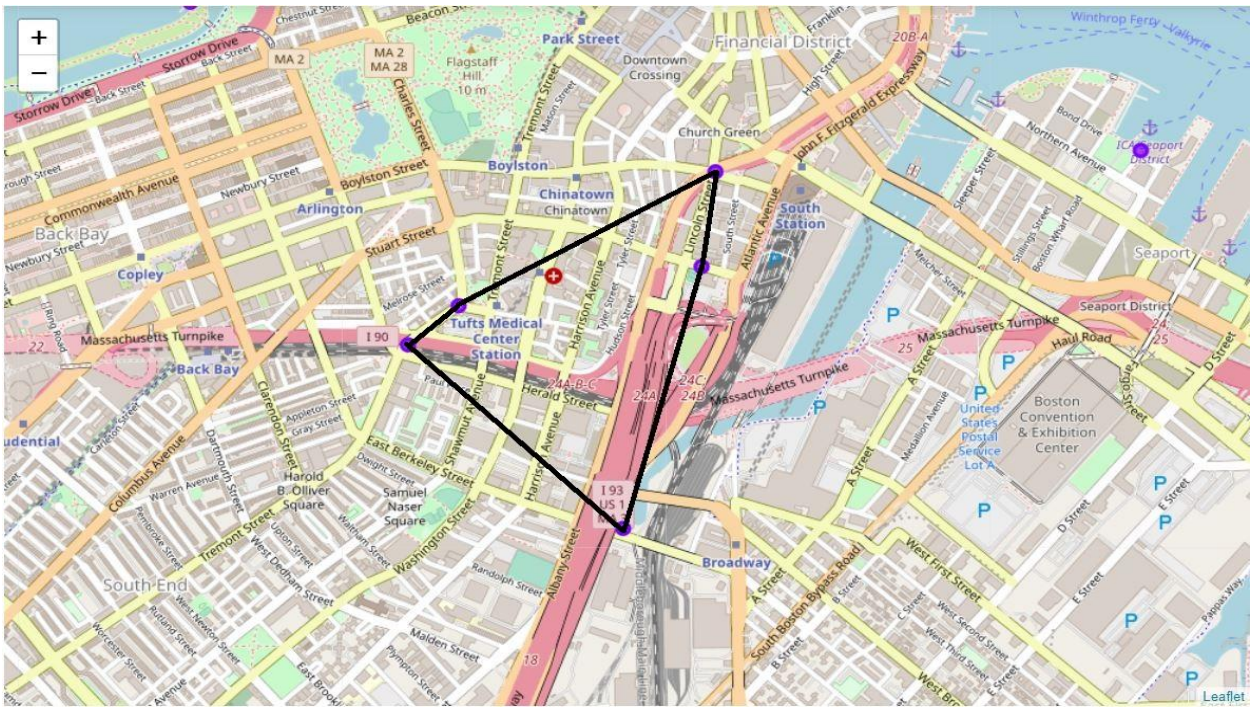
Section: 4:

Results:

After clustering we get mappings for neighborhoods and cluster labels, each cluster has its own trait. In the below graph we can see how different clusters are distributed across the map. After analyzing each of the clusters we can understand that cluster number 1 has majority of food related venues for its neighborhoods.



After careful analysis of neighborhoods in cluster 1, we can see that ‘Chinatown’, ‘Leather District’, ‘Bay Village’, ‘South End’, and ‘South Boston’ are located in close vicinity and all of these neighborhoods have majority of food places in their top 10 venues. Hence, opening a food bank in a region close to these neighborhoods would immensely reduce the amount of food wastage. In below graph we can see which places form the location in which food bank can be started. The boundary that has been built using these 5 neighborhoods serves as a good location for starting the food bank. Please find this region in the below graph;



The region marked in the above map with ‘Chinatown’, ‘Leather District’, ‘Bay Village’, ‘South End’, and ‘South Boston’ as its vertices represents the region in which starting the food bank would be the best idea. Commuting to this location is also easy as it is located at heart of the city.

There are lot of public transportation choices available in this area as well which would increase the ease of reaching the food bank. Hence, opening a food bank in the region depicted by the boundary in above graph would be a good choice to achieve our goals.

Section: 5:

Discussion:

According to the analysis, location marked between ‘Chinatown’, ‘Leather District’, ‘Bay Village’, ‘South End’, and ‘South Boston’ is the best choice for starting a food bank. As all these neighborhoods have a greater number of food related places, hence these areas could possibly have more food wastage compared to other areas. And if all this food wastage can be channeled in a right way, we will be able to feed lot of hungry people who would happily accept this delicious food.

There are also drawbacks for this analysis- land prices in these locations have not been considered in this analysis, as food bank is a nonprofit organization we can also apply for subsidy from the government. This analysis only considers most common venues obtained using foursquare API it does not take into account potential number of customers at each venue, as the number of customers will directly impacts the amount of food being prepared, this is an important variable that can be taken into consideration. We could also perform clustering using other techniques such as K Medoids, DBSCAN etc.

Section: 6:

Conclusion:

To conclude this project, we have got an idea about how real life data science projects are, I have used several python libraries to perform web scraping and also used Foursquare API to obtain necessary data, and then performed necessary data transformations, we have also used folium library to generate maps using geographical coordinates, in-depth analysis has been done on the data using K Means clustering technique. We have also discussed potentially best location to start a food bank and drawbacks in the analysis. After staying for more than 3 years in Boston and looking at the potentially possible locations for a food bank I am not surprised as all these locations are big food hubs in Boston city and sources of large quantities of food wastage which can be properly utilized through a food bank.