Capstone Project – The Battle of Neighborhoods
Yange Bian

Introduction: Business Problem

As one of the 50 largest cities in USA, **Boston is the home for many intellectuals, scholars and students from all over the world.** According to general statistics, Boston's diversity index increased from 48.7 percent in 1980 to 70.2 percent in 2017, which indicates an above-average, 21.5 percentage point increase (Schuster and Ciurczak, 2018). More importantly, **the city of Boston remains home to the largest number of Asian Americans in the state (60,985) and neighborhoods in Greater Boston in 2016 had at least 4,000 Asian Americans.** Among Boston's neighborhoods, Dorchester has the most concentrated, flourishing Asian American communities, and their affluent economic businesses (Tanabe and Lo). In addition, because many colleges' off-campus housing are located around Back Bay, Fenway, Brookline, East Boston and Allston, these neighborhoods could have increasing number of Asian students and scholars. In this case, the local businesses would transform to accommodate those individuals' needs in order to get great profits. **One key indicator is the distribution of Asian restaurants in each neighborhood, since the restaurants types, quantities, locations, and popularity are always the true representation of the local neighborhoods 'culture, economic activity and ethnic diversity.**

Multiple data sources, data analysis tools and clustering algorithm will be used to explore the distribution of Asian restaurants in Boston neighborhoods and to analyze which neighborhoods are more embracing for Asian restaurants and Asian American communities. In other words, this report will help answer the following questions:

1. Where should I open an Asian restaurants to have the best chances of getting profits? Or what is the most popular neighborhood location for Asian cuisine or eatery?
2. What Asian restaurants have the high average rating and where are they located? Do they also locate in those popular neighborhoods?
3. To what extent does the distribution of Asian restaurants represent the local economic activities? Explore the other top venue/businesses categories in those popular neighborhoods where these Asian restaurants are located.

Target Audience

For any business personnel or local Asian immigrant family who wants to invest or open a restaurant in Boston, this analysis will be a comprehensive guide to start or expand restaurants of Asian cuisine in terms of local environment and economic/commerical activities. For freelancer who loves to open restaurant as part-time business. this analysis will give an idea, how beneficial it is to open a restaurant in nearby neighborhood. For Asian American individuals such as students, scholars or family who wants to find ideal option for dining, this analysis will certainly provide the most accurate information for geospatial distribution of restaurants and their ratings.

Data

In order to answer the above research problems, data on Boston neighborhoods, including boundaries, latitude, longitude, restaurants, and restaurant ratings and tips are required.

Boston neighborhoods, boundaries, latitudes and longitudes will be obtained from the data source via **geojson file**: https://data.boston.gov/dataset/boston-neighborhoods/resource/13ee2b65-6547-4168-b112-83995f138602

Foursquare API will be used to locate all venues and filtered by Asian restaurants(i.e. Chinese, Japanese, Korean, Indian, Thai, Vietnamese, Asian Fusion, etc) via using search, explore or trending queries. Locations/neighborhoods, ratings, tips and likes for each restaurant will be counted, merged and added to a dataframe.

Data will be organized, sorted, and analyzed based on ratings and neighborhoods into clean data frames. These data frames will be showcased and visualized in different types of plots. descriptive statistics will be used and analyzed for ratings or likes to identify popular neighborhood and top venue categories.

Most importantly, the cleaned data frames with venue information will be used to create geospatial visualizations to assess the local distribution of Asian American restaurants and neighborhoods, and to build K-means clustering model to better answer the research problems.

Data Preparation

In the preparation phase, we will load and extract from the original geographical data, and create, clean and view the full data frame. Initial geospatial visualization will also be created as a general overview of the local neighborhoods.

	Neighborhood	Latitude	Longitude
0	Roslindale	42.272	-71.1259
1	Jamaica Plain	42.3261	-71.105
2	Mission Hill	42.3358	-71.0904
3	Longwood	42.3367	-71.0981
4	Bay Village	42.3488	-71.0666
5	Leather District	42.3498	-71.0584
6	Chinatown	42.3524	-71.0579
7	North End	42.3688	-71.052
8	Roxbury	42.2932	-71.0965
9	South End	42.3478	-71.0683
10	Back Bay	42.3569	-71.0757
11	East Boston	42.3939	-70.9955
12	Charlestown	42.3905	-71.0742
13	West End	42.3688	-71.0648
14	Beacon Hill	42.3612	-71.0714
15	Downtown	42.3613	-71.0534
16	Fenway	42.3521	-71.1101
17	West Roxbury	42.3038	-71.1647
18	Hyde Park	42.2721	-71.1259
19	Mattapan	42.2932	-71.0963
20	Dorchester	42.3288	-71.059
21	South Boston Waterfront	42.3529	-71.0435
22	South Boston	42.3428	-71.061
23	Allston	42.346	-71.1352
24	Harbor Islands	42.3138	-70.9672
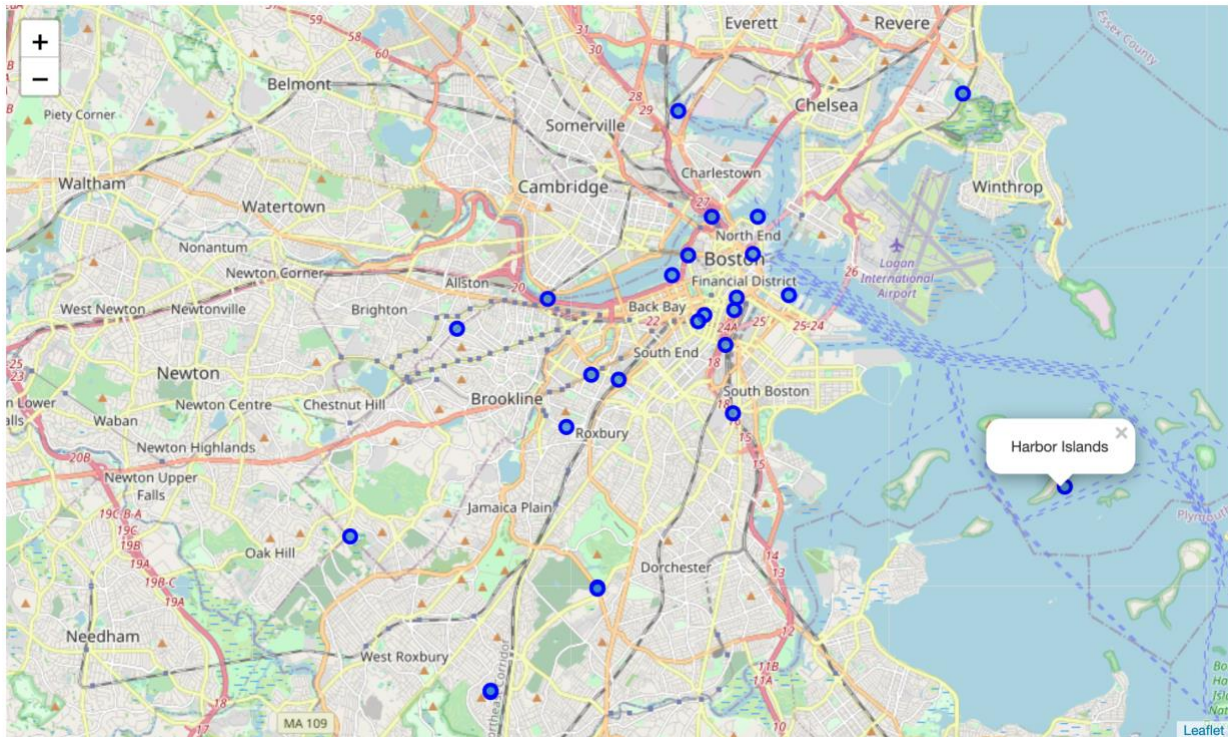
We used the latitude and longitude variables to locate folium map(the tool we use for geospatial visualization). Then we used for-loop for browsing through the cleaned data frame(neighborhoods) to locate and create labels for each neighborhood so that each location can be easily viewed as single blue dot on the map

Foursquare API and Venue Data Frames
The first 50 local restaurants (the middle column is the venue ID) are displayed below.

	Neighborhood	ID	Name
0	Roslindale	4b6caefaf964a520f04b2ce3	Misono
1	Roslindale	4a9dd1d2f964a520853920e3	Himalayan Bistro
2	Roslindale	50ef8d5c8acaed20c6c6c271	Shanti Taste of India Roslindale
3	Roslindale	4b107ba4f964a520687123e3	Phuket
4	Roslindale	4fc54e6ce4b0a6258efafba9	Banh Mi NGON
5	Jamaica Plain	4cb6134055af224b0a37977e	Lucky Boston Chinese
6	Jamaica Plain	4ae64951f964a52037a621e3	Food Wall
7	Jamaica Plain	4b6330bef964a52065692ae3	Yokohama Sushi
8	Mission Hill	4c045647310fc9b64534c761	Sunrise Chinese Food
9	Mission Hill	4a954ec2f964a520e52220e3	Thaitation
10	Mission Hill	4e83c45ebe7b87d6abc12316	Rod Dee 2
11	Mission Hill	56c4b6d2cd10583a69ad344c	Bangkok Pinto
12	Mission Hill	5152645cd86cb85b38e8a14e	Ginger Exchange at Symphony
13	Mission Hill	568f2ab6498e653e63bfd344	Shun's Kitchen
14	Mission Hill	4af9aa2af964a520131322e3	Symphony Sushi
15	Mission Hill	50b3e92b7043c62d4562074c	Mission Sushi & Wok
16	Longwood	55b2ba61498e454673a15379	Hojoko
17	Longwood	4bb4b379bc82a5935fb03e72	Basho Japanese Brasserie
18	Longwood	5a61341295d9863ac5bfcb9e	Sojuba
19	Longwood	4a954ec2f964a520e52220e3	Thaitation
20	Longwood	4e83c45ebe7b87d6abc12316	Rod Dee 2
21	Longwood	566a1870498e8921f5e9139d	Tiger Mama
22	Longwood	50b3e92b7043c62d4562074c	Mission Sushi & Wok
23	Longwood	5903816c06fb60796142faa1	Laughing Monk Cafe
24	Bay Village	49cf760ff964a520a95a1fe3	Myers + Chang
25	Bay Village	4a771079f964a52078e31fe3	Gourmet Dumpling House
26	Bay Village	3fd66200f964a52096ec1ee3	Peach Farm
27	Bay Village	575387bb498e80be5c09b17f	Gourmet China House
28	Bay Village	49f4d701f964a5205e6b1fe3	Oishii Boston
29	Bay Village	5a13605ee96d0c3ed5b426d8	Tora
30	Bay Village	53c1bcdb498e0567f9c29514	Crave - Mad for Chicken
31	Bay Village	4a932354f964a520e81e20e3	163 Vietnamese Sandwiches & Bubble Tea
32	Bay Village	4ac7f9fcf964a520fbba20e3	Shabu Zen
33	Bay Village	54340d65498efe796118e0a2	Double Chin
34	Bay Village	500f2993e4b0fe0c6fda6c61	Shojo
35	Bay Village	55995a86498e9e8cec3152d3	Banyan Bar & Refuge
36	Bay Village	4e4be732e4cde30b72898e6d	Irashai Sushi Teriyaki
37	Leather District	4a771079f964a52078e31fe3	Gourmet Dumpling House
38	Leather District	3fd66200f964a52096ec1ee3	Peach Farm
39	Leather District	4a74a6a2f964a5200bdf1fe3	Taiwan Cafe
40	Leather District	575387bb498e80be5c09b17f	Gourmet China House
41	Leather District	51eff56c498e1fe71b259d44	Gene's Chinese Flatbread Cafe
42	Leather District	3fd66200f964a52076ec1ee3	China Pearl Restaurant
43	Leather District	4c7c06403badb1f7ea9f5654	Dumpling Cafe
44	Leather District	4ab8fd1ff964a5208c7d20e3	Hong Kong Eatery
45	Leather District	5a13605ee96d0c3ed5b426d8	Tora
46	Leather District	53c1bcdb498e0567f9c29514	Crave - Mad for Chicken
47	Leather District	4a932354f964a520e81e20e3	163 Vietnamese Sandwiches & Bubble Tea
48	Leather District	4ac7f9fcf964a520fbba20e3	Shabu Zen
49	Leather District	500f2993e4b0fe0c6fda6c61	Shojo

Please note that among the total 25 neighborhoods, there are 22 neighborhoods have available Asian restaurants information.

In total, there are 168 Asian restaurants are returned with data information.

The unique neighborhoods are returned below:

```
array(['Roslindale', 'Jamaica Plain', 'Mission Hill', 'Longwood',
       'Bay Village', 'Leather District', 'Chinatown', 'Roxbury',
       'South End', 'Back Bay', 'East Boston', 'Charlestown',
       'Beacon Hill', 'Fenway', 'West Roxbury', 'Hyde Park', 'Mattapan',
       'Dorchester', 'South Boston Waterfront', 'South Boston', 'Allston',
       'Harbor Islands'], dtype=object)
```

Methodology

In this project, we will direct our effort on **identifying neighborhoods that are suitable for opening Asian restaurants**. To take a business approach, the solution could be polarized. **On one hand, the neighborhoods crowded with highly rated restaurants could be welcoming but competitive for opening a new restaurant**. On the other hand, the neighborhoods with less restaurants could be more welcoming for investing a new restaurant, which could also remain risky. Nonetheless, it is more likely that **the neighborhoods with low average ratings for Asian restaurants could reflect the local preferences for the other types of cuisine**. In the analysis, we will examine the neighborhoods based on the ratings for Asian restaurants to evaluate these assumptions before finalizing a business recommendation.

In first step, we have **created cleaned data frame for geographical information for neighborhoods and all available Asian restaurants.**

In second step, we will include **exploratory data analysis with descriptive statistics and basic visualizations** to show the **frequency distribution of Asian restaurants among neighborhoods** and to identify the potential popular neighborhoods. More importantly, **cleaned data frames for neighborhoods with restaurants ratings/likes/tips will be created and analyzed.** Note that the restaurants with the highest ratings/likes/tips will also be identified. The main geospatial visualization tool will be **Folium map and heatmap** to display ratings with each neighborhood, and **locate the populated regions for Asian restaurants**.

In third and final step, we will separate the neighborhoods based on ratings and analyze their top common venues to evaluate the business decision for opening a local Asian restaurant. Among the neighborhoods with **highly rated restaurants, the machine learning algorithm, k-means clustering**, will be used to further cluster the neighborhoods in order to **identify the most ideal group of neighborhoods** for opening an Asian restaurant. We will create a **folium map to display the clusters as well**.

If the neighborhoods have **low rated Asian restaurants but their top venues are other type of restaurants**, they shouldn't be the choices for opening Asian restaurants and **no further analysis would be done**. And if the neighborhoods have **low rated Asian restaurants and their top common venues are not restaurants-related**, opening Asian restaurant would be less competitive and possibly profitable. In this case, **further exploration could be done.**

Analysis

Let's first conduct the exploratory data analysis for the frequency of Asian restaurants in each neighborhood.
A bar plot is created to display the count of Asian restaurants for each neighborhood.
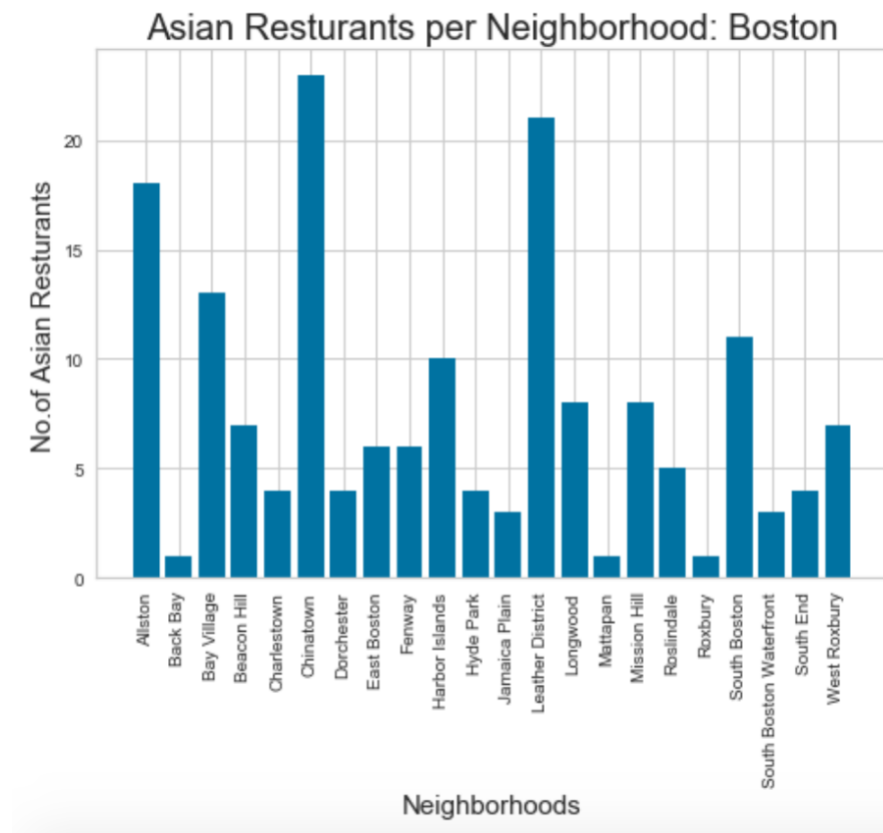
This following bar plot highlights the more populated neighborhoods with Asian restaurants. In other words, opening an Asian restaurants in these regions might seem satisfying but would be competitive in terms of available limited resources/customers.

Asian Resturants per Neighborhood: Boston
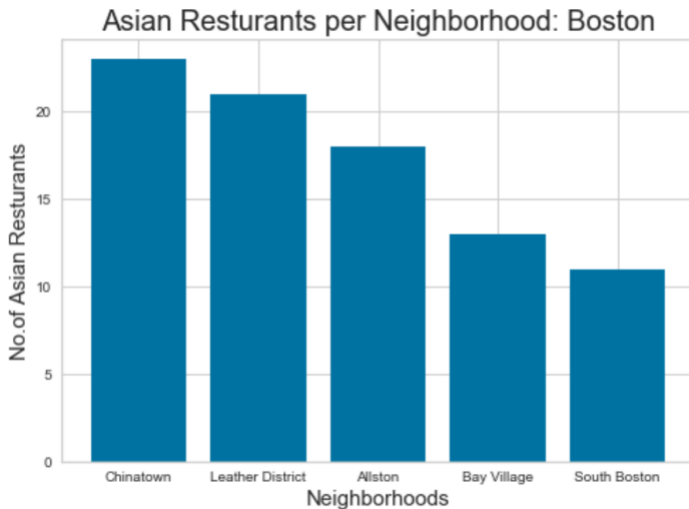
Now let's analyze the neighborhoods with restaurants ratings/likes/tips, the important metrics for customer feedback, to evaluate any possible business recommendation.

We first created an data frame to include all the neighborhood, and restaurant's venue ID, name, rating and tips. By calling get_venue_details function, we will call Foursquare API to return every venue details for each restaurant.

The dimension of the data frame has 168 rows and 6 columns. The first 10 rows are shown below:

	Neighborhood	ID	Name	Likes	Rating	Tips
0	Roslindale	4b6caefaf964a520f04b2ce3	Misono	20.0	7.2	10.0
1	Roslindale	4a9dd1d2f964a520853920e3	Himalayan Bistro	26.0	7.5	18.0
2	Roslindale	50ef8d5c8acaed20c6c6c271	Shanti Taste of India Roslindale	18.0	7.1	11.0
3	Roslindale	4b107ba4f964a520687123e3	Phuket	8.0	7.7	9.0
4	Roslindale	4fc54e6ce4b0a6258efafba9	Banh Mi NGON	5.0	7.3	7.0
5	Jamaica Plain	4cb6134055af224b0a37977e	Lucky Boston Chinese	6.0	7.1	4.0
6	Jamaica Plain	4ae64951f964a52037a621e3	Food Wall	8.0	6.3	9.0
7	Jamaica Plain	4b6330bef964a52065692ae3	Yokohama Sushi	5.0	7.6	7.0
8	Mission Hill	4c045647310fc9b64534c761	Sunrise Chinese Food	7.0	7.5	2.0
9	Mission Hill	4a954ec2f964a520e52220e3	Thaitation	67.0	8.6	27.0
10	Mission Hill	4e83c45ebe7b87d6abc12316	Rod Dee 2	43.0	8.0	24.0

```
array(['Roslindale', 'Jamaica Plain', 'Mission Hill', 'Longwood',
       'Bay Village', 'Leather District', 'Chinatown', 'Roxbury',
       'South End', 'Back Bay', 'East Boston', 'Charlestown',
       'Beacon Hill', 'Fenway', 'West Roxbury', 'Hyde Park', 'Mattapan',
       'Dorchester', 'South Boston Waterfront', 'South Boston', 'Allston',
       'Harbor Islands'], dtype=object)
```

The SAME number of neighborhoods are returned. This means that the neighborhoods with Asian restaurants ALL have available customer feedback.

The descriptive statistics for likes, rating and tips are provided. The average likes, rating and tips are 44.5, 7.9 and 20.5. The maximum are 449, 9.2 and 210.

	Unnamed: 0	Likes	Rating	Tips
count	168.000000	168.000000	168.000000	168.000000
mean	83.500000	78.202381	7.847619	34.363095
std	48.641546	94.345813	0.566421	41.791674
min	0.000000	4.000000	6.300000	0.000000
25%	41.750000	18.000000	7.500000	9.000000
50%	83.500000	44.500000	7.900000	20.500000
75%	125.250000	102.000000	8.200000	41.000000
max	167.000000	449.000000	9.200000	210.000000

Now we are interested in the restaurants that have the maximum value, and their neighborhoods might be useful for subsequent analysis.

```
Unnamed: 0                          24     Unnamed: 0                          107
Neighborhood               Bay Village     Neighborhood                     Fenway
ID         49cf760ff964a520a95a1fe3     ID          51703b83e4b06748d3b28952
Name                     Myers + Chang     Name            Gyu-Kaku Japanese BBQ
Likes                              449     Likes                             188
Rating                             8.6     Rating                            9.2
Tips                               210     Tips                               27
Name: 24, dtype: object                   Name: 107, dtype: object
```

Remember that Bay Village is one of the top populated for Asian restaurants. On the other hand, Fenway is not one of the populated neighborhoods. But its restaurant has the highest rating suggesting that the lack of competitors might improve customer feedback and economic profitability.

**Further analysis will be done in terms of the average rating.**

	Neighborhood	Average Rating
0	Allston	8.066667
1	Back Bay	8.100000
2	Bay Village	8.269231
3	Beacon Hill	7.914286
4	Charlestown	8.150000
5	Chinatown	7.756522
6	Dorchester	7.525000
7	East Boston	7.816667
8	Fenway	8.000000
9	Harbor Islands	7.660000
10	Hyde Park	7.400000
11	Jamaica Plain	7.000000
12	Leather District	7.871429
13	Longwood	7.562500
14	Mattapan	7.100000
15	Mission Hill	7.475000
16	Roslindale	7.360000
17	Roxbury	7.100000
18	South Boston	8.300000
19	South Boston Waterfront	7.666667
20	South End	8.600000
21	West Roxbury	7.671429

Due to the high rating for particular restaurants, Fenway seems to have fairly high average ratings(8.0).

Now let's create a bar graph to identify the neighborhoods with top ratings.

Average Rating for Asian Resturants per Neighborhood: Boston
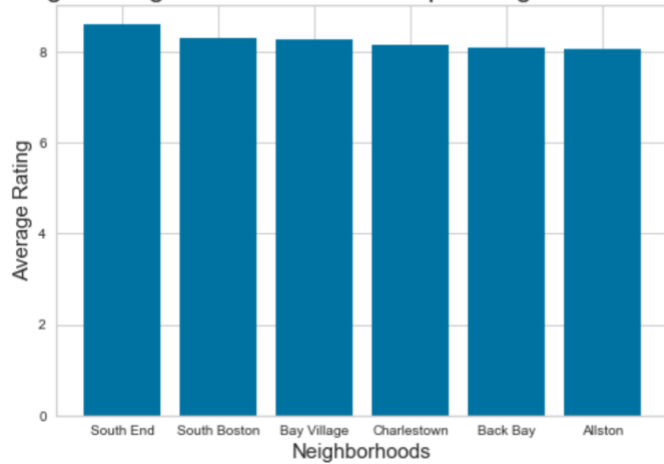
```
#Graph the top 5 neighborhoods
numm=asian_rest.groupby('Neighborhood').count()['ID'].nlargest()
numm
```

```
Neighborhood
Chinatown          23
Leather District   21
Allston            18
Bay Village        13
South Boston       11
Name: ID, dtype: int64
```

Compared to the bar graph of the top populated neighborhoods (most number of Asian restaurants), this bar graph shows that the top two neighborhoods are changed.

Chinatown and Leather District are no longer highly rated. This change again suggests that the number of restaurants or competitors influences customer ratings.

Another Folium map is created to show the location of each neighborhood and the average rating. Each neighborhood's label is formatted into a string started with the neighborhood's name and the average rating rounded to 2 decimal places. When we click each dot on the map, the rating will be displayed as well.
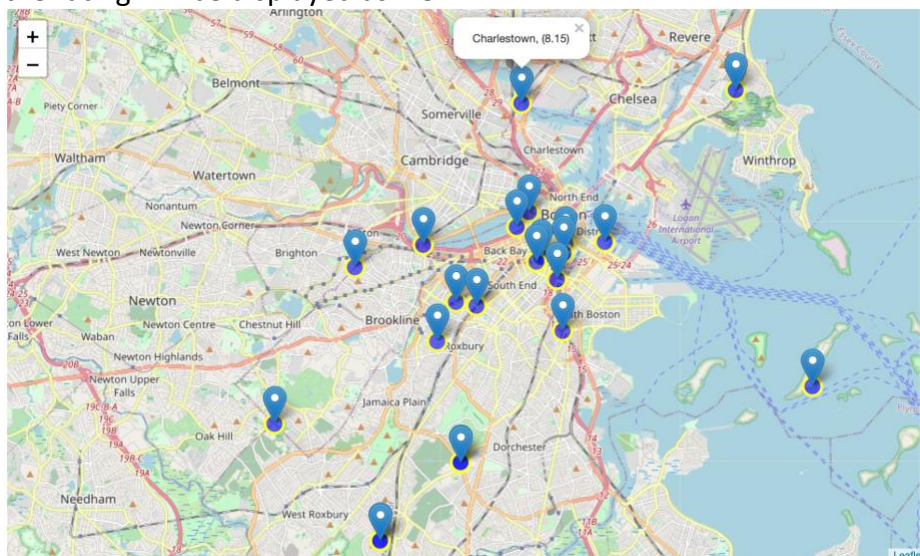
We also created a heat map to illustrate the average ratings terms of color density.
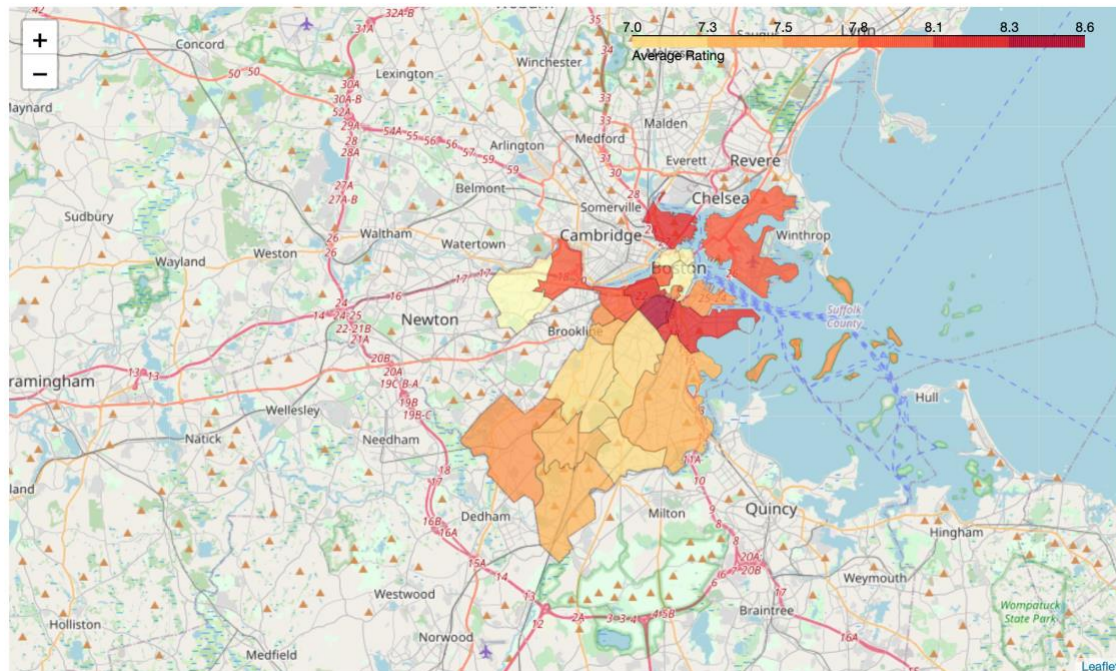
From this heatmap, we can approximately identify the regions with highly rated restaurants: Allston, Fenway, Back Bay, South End, South Boston, Charlestown and East Boston. These regions generally have the average rating of more than 7.8. They are also quite clustered around Boston's downtown area.

To take a more conservative approach, we will examine these neighborhoods with **average ratings greater than 7.5.** For these highly rated neighborhoods, we will analyze each of their top common venues, and create and analyze their clusters to evaluate our recommendation for opening new Asian restaurant.

We will create a separate data frame for only the highly rated neighborhoods with average rating greater than 7.5, and sort them in descending order.

Please note that there are **16 highly rated neighborhoods** after filtering.

	Neighborhood	Latitude	Longitude	Average Rating	Label
0	South End	42.347751	-71.068341	8.600000	South End, (8.6)
1	South Boston	42.342830	-71.061036	8.300000	South Boston, (8.3)
2	Bay Village	42.348783	-71.066629	8.269231	Bay Village, (8.27)
3	Charlestown	42.390512	-71.074160	8.150000	Charlestown, (8.15)
4	Back Bay	42.356917	-71.075689	8.100000	Back Bay, (8.1)
5	Allston	42.346015	-71.135174	8.066667	Allston, (8.07)
6	Fenway	42.352080	-71.110109	8.000000	Fenway, (8.0)
7	Beacon Hill	42.361187	-71.071409	7.914286	Beacon Hill, (7.91)
8	Leather District	42.349831	-71.058378	7.871429	Leather District, (7.87)
9	East Boston	42.393939	-70.995463	7.816667	East Boston, (7.82)
10	Chinatown	42.352379	-71.057906	7.756522	Chinatown, (7.76)
11	West Roxbury	42.303839	-71.164703	7.671429	West Roxbury, (7.67)
12	South Boston Waterfront	42.352935	-71.043477	7.666667	South Boston Waterfront, (7.67)
13	Harbor Islands	42.313772	-70.967178	7.660000	Harbor Islands, (7.66)
14	Longwood	42.336730	-71.098109	7.562500	Longwood, (7.56)
15	Dorchester	42.328758	-71.058955	7.525000	Dorchester, (7.52)

We will get nearby venues names, geographical information and categories by calling the function getNearbyVenues.
We will retrieve approximately 100 nearby venues per neighborhood, and python returned 226 unique venue categories and 1582 total venues.

```
print('There are {} uniques categories.'.format(len(bos_venues['Venue Category'].unique())))
```
There are 226 uniques categories.

```
print(bos_venues.shape)
bos_venues.head()
```
(1582, 7)

In order to sort top venue categories for each neighborhood and to create k-means clustering, the venue category column has to be spread out across the data frame. Each category should be a separate column, and there will be 226 additional columns. If a venue is belonged to a category, that column will be marked as 1; otherwise, the column will be marked as 0. The columns will be set as dummy variables using one hot encoding.
We group by each neighborhood, and take the mean for each column to show the average frequency of each venue category in the neighborhood.

	Neighborhood	Accessories Store	Afghan Restaurant	African Restaurant	Airport Lounge	Airport Service	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Aud
0	Allston	0.00	0.00	0.00	0.00	0.00	0.000000	0.00	0.00	0.00	0.00	0.01	0.010000	0.00	
1	Back Bay	0.00	0.00	0.00	0.00	0.00	0.010000	0.00	0.00	0.00	0.00	0.01	0.000000	0.01	
2	Bay Village	0.00	0.00	0.00	0.00	0.00	0.040000	0.01	0.00	0.00	0.00	0.01	0.040000	0.00	
3	Beacon Hill	0.00	0.01	0.00	0.00	0.00	0.040000	0.00	0.00	0.00	0.00	0.00	0.010000	0.00	
4	Charlestown	0.00	0.00	0.00	0.00	0.00	0.020000	0.00	0.00	0.00	0.00	0.00	0.010000	0.01	
5	Chinatown	0.00	0.00	0.00	0.00	0.00	0.010000	0.00	0.00	0.00	0.00	0.00	0.080000	0.01	
6	Dorchester	0.00	0.00	0.01	0.00	0.00	0.040000	0.01	0.00	0.02	0.00	0.00	0.010000	0.00	
7	East Boston	0.00	0.00	0.00	0.01	0.01	0.040000	0.00	0.00	0.00	0.00	0.00	0.010000	0.00	
8	Fenway	0.00	0.00	0.00	0.00	0.00	0.020619	0.00	0.00	0.00	0.00	0.00	0.000000	0.00	
9	Harbor Islands	0.00	0.00	0.00	0.00	0.00	0.035294	0.00	0.00	0.00	0.00	0.00	0.023529	0.00	
10	Leather District	0.00	0.00	0.00	0.00	0.00	0.030000	0.00	0.00	0.00	0.01	0.00	0.070000	0.00	
11	Longwood	0.00	0.00	0.01	0.00	0.00	0.040000	0.00	0.00	0.03	0.03	0.00	0.000000	0.00	
12	South Boston	0.01	0.00	0.00	0.00	0.00	0.050000	0.00	0.00	0.00	0.00	0.00	0.030000	0.00	
13	South Boston Waterfront	0.00	0.00	0.00	0.00	0.00	0.020000	0.00	0.02	0.01	0.02	0.00	0.020000	0.00	
14	South End	0.01	0.00	0.00	0.00	0.00	0.050000	0.01	0.00	0.01	0.00	0.01	0.010000	0.00	
15	West Roxbury	0.00	0.00	0.00	0.00	0.00	0.060000	0.00	0.00	0.00	0.00	0.00	0.000000	0.00	

```
bos_grouped.shape
```
(16, 226)

This data dimension correctly corresponds to **16 highly rated neighborhoods and 226 unique venue categories.**
Now, let's create a data frame displaying **top 10 venue categories for each neighborhood!**
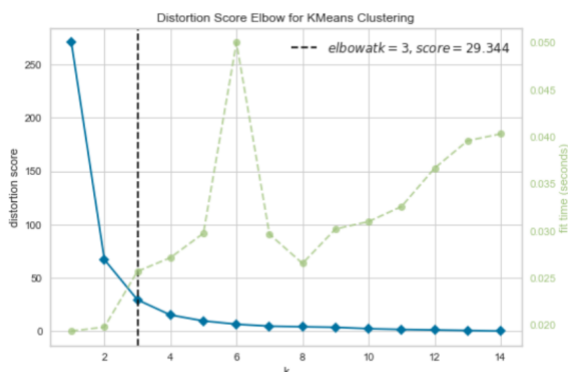
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allston	Bakery	Chinese Restaurant	Gastropub	Mexican Restaurant	Korean Restaurant	Sushi Restaurant	Pizza Place	Bar	Ice Cream Shop	Liquor Store
1	Back Bay	Spa	Italian Restaurant	Ice Cream Shop	Gourmet Shop	Hotel	Plaza	Bakery	Park	Jewelry Store	Coffee Shop
2	Bay Village	Coffee Shop	American Restaurant	Asian Restaurant	Theater	Spa	Chinese Restaurant	Italian Restaurant	Wine Shop	Hotel	Pizza Place
3	Beacon Hill	Science Museum	American Restaurant	Hotel	Hotel Bar	Italian Restaurant	Playground	Gym / Fitness Center	Pizza Place	Bar	Park
4	Charlestown	Brewery	Bakery	Pizza Place	Gastropub	Gym	Clothing Store	Sporting Goods Shop	Cosmetics Shop	Department Store	Pub
5	Chinatown	Chinese Restaurant	Coffee Shop	Asian Restaurant	Bakery	Sandwich Place	Sushi Restaurant	Gym / Fitness Center	Mexican Restaurant	Pizza Place	Vegetarian / Vegan Restaurant
6	Dorchester	Coffee Shop	Bar	Sandwich Place	Italian Restaurant	American Restaurant	Pizza Place	Liquor Store	Gym / Fitness Center	New American Restaurant	Gym
7	East Boston	Italian Restaurant	Pizza Place	Hotel	Sandwich Place	Beach	Ice Cream Shop	Chinese Restaurant	Donut Shop	American Restaurant	Coffee Shop
8	Fenway	Pizza Place	Gym / Fitness Center	Coffee Shop	Grocery Store	Chinese Restaurant	Bakery	Mexican Restaurant	Yoga Studio	Japanese Restaurant	Pub
9	Harbor Islands	Beach	Park	Seafood Restaurant	Chinese Restaurant	Breakfast Spot	Lighthouse	Bakery	Scenic Lookout	Bar	Diner
10	Leather District	Chinese Restaurant	Asian Restaurant	Bakery	Coffee Shop	Italian Restaurant	American Restaurant	Gym / Fitness Center	Sushi Restaurant	Sandwich Place	Theater
11	Longwood	Pizza Place	Coffee Shop	Café	Donut Shop	Sandwich Place	American Restaurant	Middle Eastern Restaurant	Art Museum	Art Gallery	Park
12	South Boston	Coffee Shop	American Restaurant	Bakery	Italian Restaurant	Theater	Wine Bar	Asian Restaurant	Pizza Place	Pet Store	Chinese Restaurant
13	South Boston Waterfront	Seafood Restaurant	Hotel	Park	Italian Restaurant	Steakhouse	Coffee Shop	Gym	Bakery	Bar	Brewery
14	South End	American Restaurant	Coffee Shop	Theater	Italian Restaurant	Gym	Wine Bar	Performing Arts Venue	Gourmet Shop	Hotel	Park
15	West Roxbury	Italian Restaurant	American Restaurant	Ice Cream Shop	Pizza Place	Gym / Fitness Center	Park	Burger Joint	Gym	Greek Restaurant	Chinese Restaurant

From the above data frame, most of these highly rated neighborhoods are crowded with restaurants/Pizza Place or other eatery, as the top 3 common venues are restaurants-related. A few neighborhoods such as Beacon Hill, Harbor Islands, Charlestown, and Dorchester have **less number of dining restaurants**, but **more common complementary facilities** such as bakery, coffee shops, beach, park or hotel.

We will need k-means algorithm to more precisely cluster the neighborhoods to separate the populated regions for restaurants from the other regions.

Now let's import matplotlib graphing tool and k-means algorithm to identify the best number of clusters and to fit the dataset.
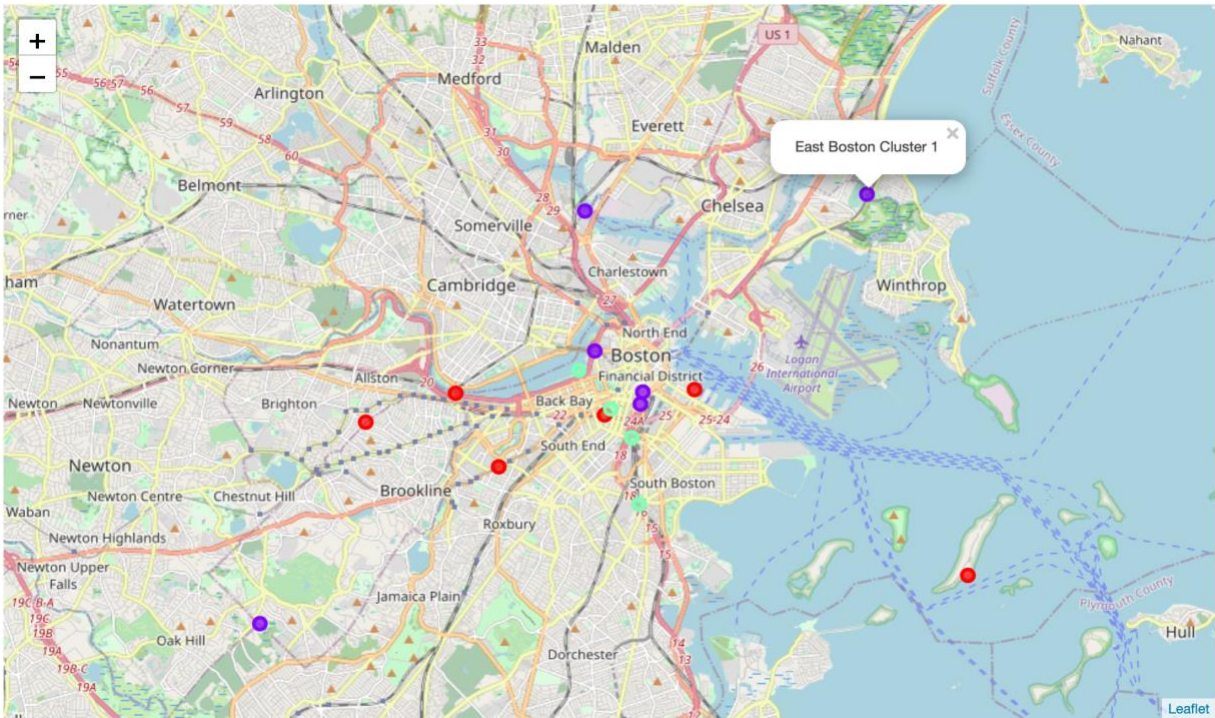
The elbow plot computes the distortion score, the sum of squared distances from each point to its assigned center.

The elbow plot shows that the best number of elbows should be 3. Let's fit the data using 3 clusters.

We created a Folium map to display the different 3 clusters labeled with different colors.

To identify the neighborhood for each cluster, we will display the corresponding data frames.

Cluster 1

```
boston_merged.loc[boston_merged['Cluster Labels'] == 0, boston_merged.columns[[0] + list(range(4, boston_merged.shape[
```

	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	South End	American Restaurant	Coffee Shop	Theater	Italian Restaurant	Gym	Wine Bar	Performing Arts Venue	Gourmet Shop	Hotel	Park
5	Allston	Bakery	Chinese Restaurant	Gastropub	Mexican Restaurant	Korean Restaurant	Sushi Restaurant	Pizza Place	Bar	Ice Cream Shop	Liquor Store
6	Fenway	Pizza Place	Gym / Fitness Center	Coffee Shop	Grocery Store	Chinese Restaurant	Bakery	Mexican Restaurant	Yoga Studio	Japanese Restaurant	Pub
12	South Boston Waterfront	Seafood Restaurant	Hotel	Park	Italian Restaurant	Steakhouse	Coffee Shop	Gym	Bakery	Bar	Brewery
13	Harbor Islands	Beach	Park	Seafood Restaurant	Chinese Restaurant	Breakfast Spot	Lighthouse	Bakery	Scenic Lookout	Bar	Diner
14	Longwood	Pizza Place	Coffee Shop	Café	Donut Shop	Sandwich Place	American Restaurant	Middle Eastern Restaurant	Art Museum	Art Gallery	Park

Cluster 2

```
n_merged.loc[boston_merged['Cluster Labels'] == 1, boston_merged.columns[[0] + list(range(4, boston_merged.shape[1]))]]
```

	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
3	Charlestown	Brewery	Bakery	Pizza Place	Gastropub	Gym	Clothing Store	Sporting Goods Shop	Cosmetics Shop	Department Store	Pub
7	Beacon Hill	Science Museum	American Restaurant	Hotel	Hotel Bar	Italian Restaurant	Playground	Gym / Fitness Center	Pizza Place	Bar	Park
8	Leather District	Chinese Restaurant	Asian Restaurant	Bakery	Coffee Shop	Italian Restaurant	American Restaurant	Gym / Fitness Center	Sushi Restaurant	Sandwich Place	Theater
9	East Boston	Italian Restaurant	Pizza Place	Hotel	Sandwich Place	Beach	Ice Cream Shop	Chinese Restaurant	Donut Shop	American Restaurant	Coffee Shop
10	Chinatown	Chinese Restaurant	Coffee Shop	Asian Restaurant	Bakery	Sandwich Place	Sushi Restaurant	Gym / Fitness Center	Mexican Restaurant	Pizza Place	Vegetarian / Vegan Restaurant
11	West Roxbury	Italian Restaurant	American Restaurant	Ice Cream Shop	Pizza Place	Gym / Fitness Center	Park	Burger Joint	Gym	Greek Restaurant	Chinese Restaurant

Cluster 3

```
n_merged.loc[boston_merged['Cluster Labels'] == 2, boston_merged.columns[[0] + list(range(4, boston_merged.shape[1]))]]
```

	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
1	South Boston	Coffee Shop	American Restaurant	Bakery	Italian Restaurant	Theater	Wine Bar	Asian Restaurant	Pizza Place	Pet Store	Chinese Restaurant
2	Bay Village	Coffee Shop	American Restaurant	Asian Restaurant	Theater	Spa	Chinese Restaurant	Italian Restaurant	Wine Shop	Hotel	Pizza Place
4	Back Bay	Spa	Italian Restaurant	Ice Cream Shop	Gourmet Shop	Hotel	Plaza	Bakery	Park	Jewelry Store	Coffee Shop
15	Dorchester	Coffee Shop	Bar	Sandwich Place	Italian Restaurant	American Restaurant	Pizza Place	Liquor Store	Gym / Fitness Center	New American Restaurant	Gym

After conducting k-means clustering, cluster 2 shows the **most populated neighborhoods** with Asian restaurants, Italian restaurants and American restaurants. In other words, opening new restaurants in these regions appears quite competitive and economic profits are not guaranteed. Though the top common venues suggest the local preferences for dining and the local ethnic diversity, customers, however, are more likely to be attracted to many competitors restaurants and the profit share will be even out.

Compared to cluster 2, **cluster 1 and 3** are more ideal regions for opening new restaurants. In particular, their top common venues could **help improve and complement** the local dining activities. These facilities are coffee shop, brewery, pizza place, beach, hotel, park or spa. After spending time at these places, people will find the surrounding restaurants as supportive or suitable to gain energy and take a rest. People can also enjoy the change of taste. For example, if people are tired of the common bakery for afternoon tea, walking at the park, or western restaurants for dining or seafood, Asian foreign restaurant will certainly be a more attractive, creative option for dining.

For those neighborhoods with lower ratings for Asian restaurants(average rating<=7.5), we will examine their top venues to see **if non-Asian restaurants are common and competitive(popular),** or if a opening a restaurant itself will be a potentially **less competitive business decision**. **(restaurant or eatery is not even the top local venue.)**

	Neighborhood	Latitude	Longitude	Average Rating
0	Mission Hill	42.335770	-71.090434	7.475
1	Hyde Park	42.272114	-71.125881	7.400
2	Roslindale	42.272013	-71.125927	7.360
3	Mattapan	42.293165	-71.096331	7.100
4	Roxbury	42.293233	-71.096460	7.100
5	Jamaica Plain	42.326102	-71.104992	7.000

The 6 lower-rated neighborhoods are shown above.

We will apply one hot encoding to create dummy variable columns for each venue category and compute each category's average frequency for each neighborhood as we did before.

	Neighborhood	Accessories Store	African Restaurant	American Restaurant	Antique Shop	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	BBQ Joint	Bagel Shop	Bakery	Bar	Baseb Fie
0	Hyde Park	0.00	0.000000	0.070000	0.00	0.000000	0.000000	0.00	0.00	0.01	0.00	0.000000	0.040000	0.020000	0.
1	Jamaica Plain	0.00	0.010101	0.030303	0.00	0.010101	0.010101	0.00	0.00	0.00	0.00	0.010101	0.010101	0.010101	0.
2	Mattapan	0.02	0.000000	0.060000	0.00	0.000000	0.000000	0.00	0.00	0.00	0.01	0.010000	0.050000	0.020000	0.
3	Mission Hill	0.00	0.020000	0.040000	0.01	0.020000	0.030000	0.02	0.02	0.00	0.00	0.000000	0.010000	0.020000	0.
4	Roslindale	0.00	0.000000	0.070000	0.00	0.000000	0.000000	0.00	0.00	0.01	0.00	0.000000	0.040000	0.020000	0.
5	Roxbury	0.02	0.000000	0.060000	0.00	0.000000	0.000000	0.00	0.00	0.00	0.01	0.010000	0.050000	0.020000	0.

```
print('There are {} unique venue categories.'.format(len(bos_venuess['Venue Category'].unique())))
```

There are 120 unique venue categories.

```
bos_groupedd.shape
```

(6, 121)

Including neighborhood as a column, the data frame has 121 columns and 6 unique neighborhoods.

We also look at the top 10 common venues for each neighborhood to confirm our finding that the top common venues are all related to American or western restaurants.

	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	Hyde Park	American Restaurant	Italian Restaurant	Pharmacy	Pizza Place	Ice Cream Shop	Bakery	Grocery Store	Discount Store	Liquor Store	Mexican Restaurant
1	Jamaica Plain	Pizza Place	Park	Gym	Italian Restaurant	Donut Shop	American Restaurant	Sandwich Place	Liquor Store	Coffee Shop	Mexican Restaurant
2	Mattapan	American Restaurant	Ice Cream Shop	Bakery	Coffee Shop	Park	Brewery	Caribbean Restaurant	Pizza Place	Gym / Fitness Center	Italian Restaurant
3	Mission Hill	Café	Sandwich Place	Pizza Place	American Restaurant	Middle Eastern Restaurant	Concert Hall	Italian Restaurant	Art Museum	Thai Restaurant	Coffee Shop
4	Roslindale	American Restaurant	Italian Restaurant	Pharmacy	Pizza Place	Ice Cream Shop	Bakery	Grocery Store	Discount Store	Liquor Store	Mexican Restaurant
5	Roxbury	American Restaurant	Ice Cream Shop	Bakery	Coffee Shop	Park	Brewery	Caribbean Restaurant	Pizza Place	Gym / Fitness Center	Italian Restaurant

The sorted dataframe shows that the top common venues are all western restaurants. **Because these neighborhoods have low average ratings for Asian restaurants(<7.5),** the local

preferences for dining are more domestic instead of diverse. **Opening a new Asian restaurant can be a risky business decision, since the local customers don't seem to prefer non-Western cuisine as their dining option.**

Results and Decision

Results

In this section, we will document all the findings from the data analysis and visualization. The essential business question we started off is to identify a suitable neighborhood to open a new Asian restaurant. We examined the distribution of Asian restaurants in each neighborhood in terms of frequency and location. We analyzed the average popularity for Asian restaurant cuisine for each neighborhood using ratings/likes/tips as metrics. We also performed data wrangling by creating, cleaning and sorting data frames to build geospatial visualization(folium map and heat map). We used the average rating as a key separator for highly rated and low rated neighborhoods to conduct further analysis using k-means clustering. We identified the regions/cluster of neighborhoods suitable for opening new Asian restaurants and even examined the top common venue categories within each cluster before making business recommendations.

- Among the total of 25 neighborhoods, 22 of them have local Asian restaurants.
- The top 5 neighborhoods with the most number of Asian restaurants are: **Chinatown, Leather Strict, Allston, Bay Village and South Boston. They are the highly populated regions for Asian restaurants.**
- The top neighborhoods with highest average ratings are (>8.0): **South End, South Boston, Bay Village, Charlestown, and Allston.**
- Using the **heat map** visualization tool, we can approximately identify the regions with highly rated restaurants: Allston, Fenway, Back Bay, South End, South Boston, Charlestown and East Boston.
- After sorting the data frame, **Bay Village and Fenway** have individual restaurants with highest rating, tips and likes.
- 16 highly rated neighborhoods have average rating of 7.5 and not all of their top common venues are restaurants-related.
- The top common venues are restaurants, coffee stop, bakery, brewery, park, beach, museum, spa, American or Italian restaurants.
- Using k-means clustering, 3 clusters are identified.
- **Clusters 1 and 3** are generally recommended regions for opening restaurants, since the top common venues for those neighborhoods are **complementary and supportive for restaurants dining**. The venues include coffee stop, bakery, brewery, park, beach, hotel, and spa. In other words, restaurants in these neighborhoods are **not as competitive** as the populated neighborhoods. Because these neighborhoods are also highly rated for Asian restaurants, the local customers still prefer Asian cuisine and these restaurants can still easily attract customers.

- **Within the clusters, the recommended neighborhoods are: South End, Allston, Fenway, South Boston Waterfront, Harbor Islands, Longwood, South Boston, Bay Village, Back Bay and Dorchester.**
- For those neighborhoods with average rating less than 7.5, their top common venues are non-Asian restaurants. Their popular economic activities are still dining/restaurant-related. This means that the local preferences are western cuisine, and opening Asian restaurants to attract local customers isn't as ideal.
- The following neighborhoods are therefore **not recommended: Hyde Park, Jamaica Plain, Mattapan, Mission Hill, Roslindale and Roxbury.**

In summary, the business recommendations are based on the number of Asian restaurants, high ratings of Asian restaurants(available online customer feedback) and less possibility of competitions.

Discussion

According to the analysis, South End has the highest average rating for Asian restaurant, and is one of the least populated competitive neighborhood for opening new Asian restaurant based on K-means clustering. From the geospatial visualization, South End is located around the populated neighborhoods for Asian restaurants such as Chinatown, Leather District Bay Village and South Boston, and this location can benefit from high customer visit possibility. Because Asian restaurants is not the top common venue within the neighborhood, South End could potentially be an ideal choice for starting a new Asian restaurants to attract nearby customers and to experience less competitions. As similarly ideal choices, most neighborhoods from Cluster 1 and Cluster 3 could also be less competitive but are highly rated for opening quality Asian restaurants.

There are some drawbacks of this analysis. As the main data source, Foursquare API data is not guaranteed to be up-to-date. Though 22 of 25 neighborhoods have available data, it doesn't mean that the 3 neighborhoods actually do not have any available Asian restaurants located. The analysis procedure is step-by-step and appropriate, but the results might be inaccurate and less comprehensive. If more similar online API interfaces are introduced for gathering local venue data, we could certainly compare the quantity and quality of multiple data sources and select the most comprehensive data source to improve the results. Yet this analysis so far has provided us with a lot of insightful, detailed findings as summarized in the Results section, and could be a important cornerstone for answering the business problem of identifying neighborhoods for opening new Asian restaurants.

Conclusion

The purpose of this project was to identify the best neighborhoods for opening Asian restaurants in Boston. The business recommendation is based on the local average ratings and low number of surrounding Asian restaurants(an indicator of local competitiveness). First, we gathered necessary geographical and venue data from online geojson file and foursquare API.

We then performed data wrangling to clean and extract relevant information to build pandas data frames. Using data sorting, we were able to get the top neighborhoods with the most number of Asian restaurants and those neighborhoods with the high average ratings. We also performed exploratory data analysis to get descriptive statistics and basic visualization to obtain an overview of the data. We created the geospatial visualizations such as Folium map and Heat map to display the geographical center of each neighborhood on the city map and display the popularity density using the sorted average ratings. We organized the data and ranked each neighborhood's common venue category. More importantly, we applied K-means clustering to group the neighborhoods based on their common venue categories, and identify multiple choices and the best choice of the neighborhood for opening Asian restaurant.

The analysis could certainly be improved by using multiple data sources to increase the venue information coverage. Different machine learning algorithm could be used and compared with the performance of K-means clustering. Nevertheless, this analysis procedure is detail-oriented and appropriate to apply on the similar business problem such as opening a new gym, new coffee shop or new school in a given city area. As a comprehensive initial guidance, this capstone project has successfully taken a data-driven approach to solve complex, real-life problem.