Capstone Project - The Battle of Neighborhoods: Find an Optimal Chinese Restaurant Location in New York City

Coursera IBM Data Science Professional Certificate

Zhiyuan Yang

Table of Contents

Introduction

New York City is a major metropolitan area in America with more than 8 million people. It has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

New York City's food culture includes an array of international cuisines influenced by the city's immigrant history. As of 2019, there were 27,043 restaurants in the city. With its diverse culture, comes diverse food items. There are many restaurants in New York City, each belonging to different categories like Chinese, Indian, French etc.

For the final assignment of Coursera IBM Data Science Professional Certificate, the business problem is to find an optimal location to open a Chinese restaurant in New York City.

The targeted people who would be interested in this project could be stakeholders who want to open Chinese restaurants in New York City, and tourists who want to enjoy Chinese food in this city.

Data

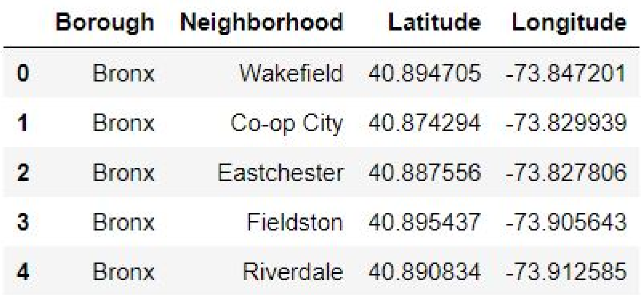
- New York City data containing neighborhoods, boroughs, latitudes and longitudes: https://cocl.us/new_york_dataset
- New York City neighborhood venues data: Foursquare API

Methodology

- Fetch New York City data from https://cocl.us/new_york_dataset.
- Visualize New York City data on map using Folium.
- Fetch New York City neighborhood venues data using Foursquare API.

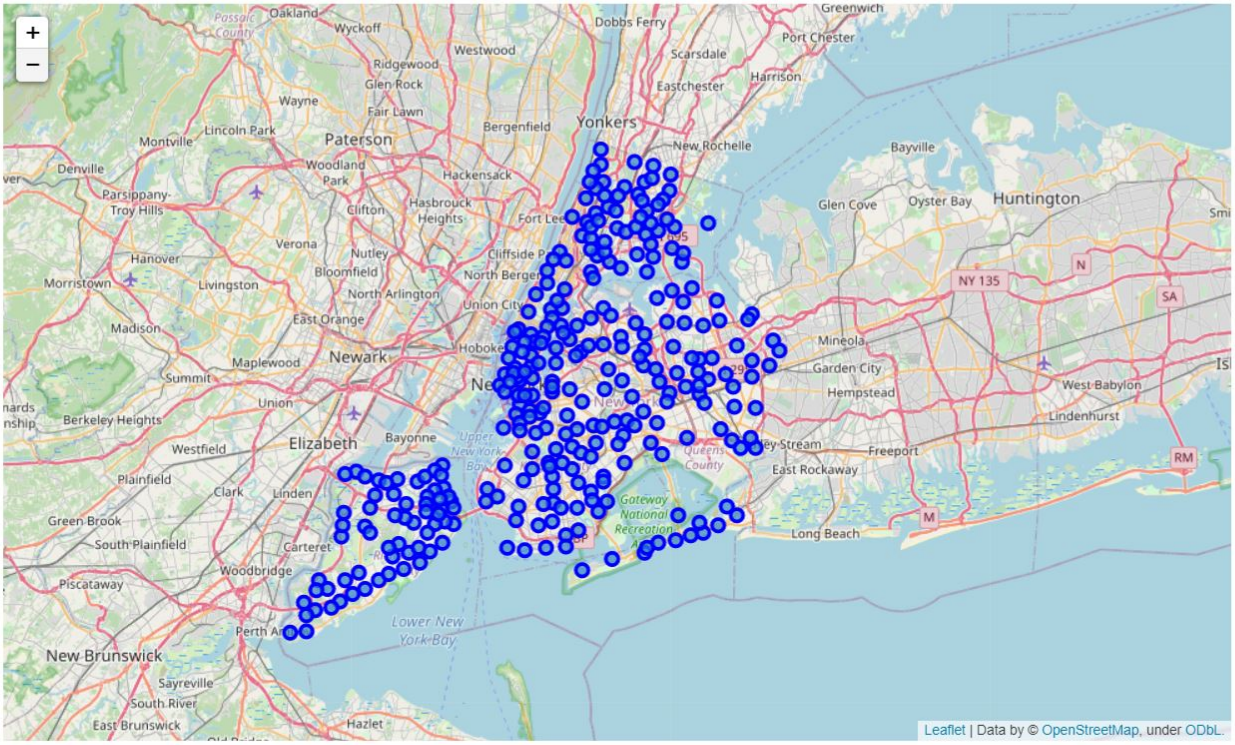
- Perform data preprocessing using one hot encoding and take the mean of the frequency of occurrence of each venue category.
- Filter out only Chinese restaurants and visualize them based on boroughs and neighborhoods using histograms and bar charts.
- Build a k-means clustering model and visualize results on map using Elbow Method.
- Compare neighborhoods to find an optimal location to open a Chinese restaurants.

Fetch New York City data from https://cocl.us/new_york_dataset and store the data into a pandas data frame.

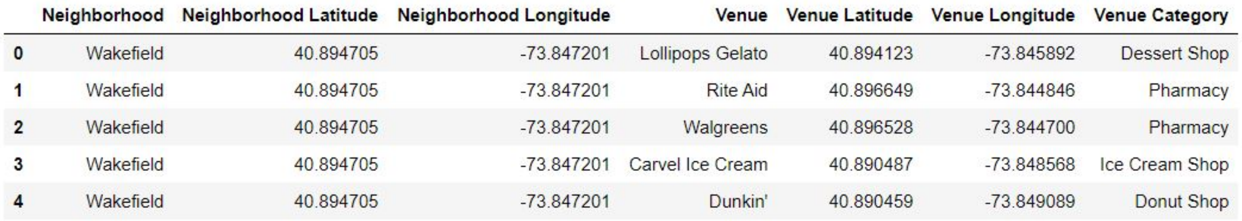


	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Visualize New York City data on map using Folium.

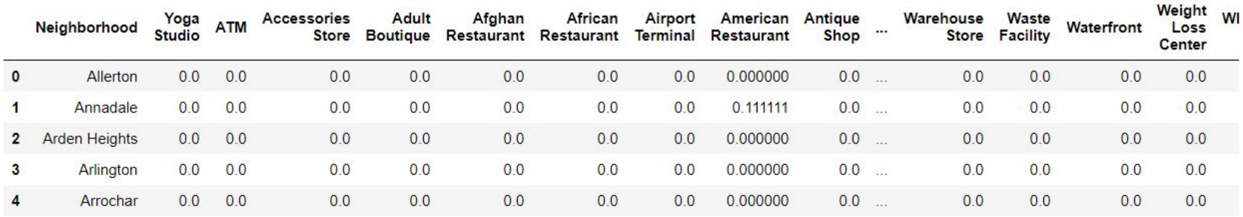


Fetch the top 100 venues in New York City neighborhoods within a radius of 500 meters using Foursquare API.



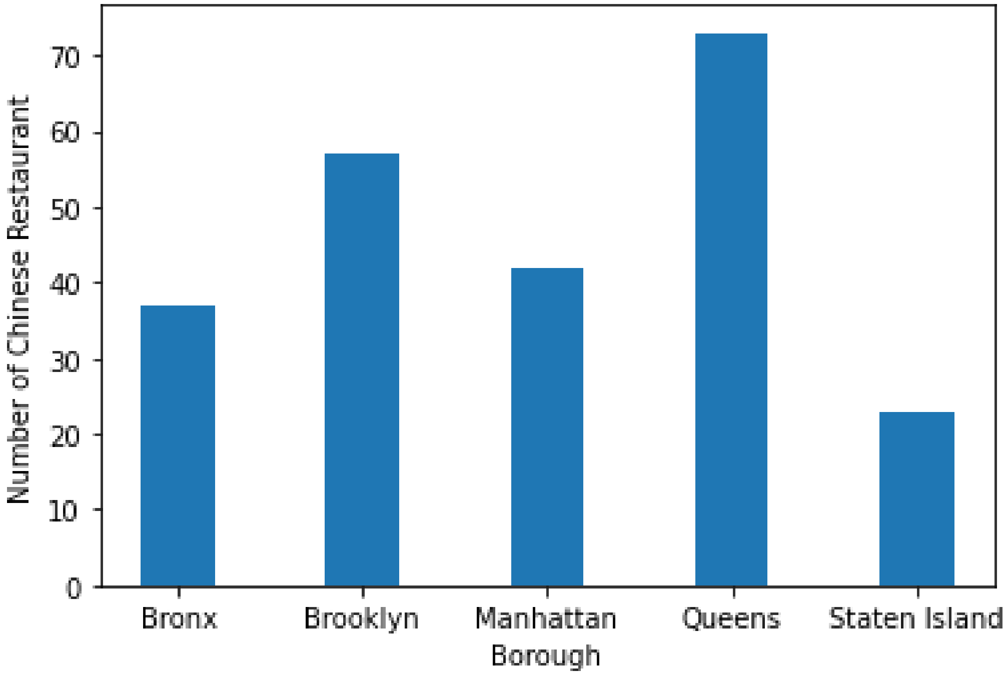
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

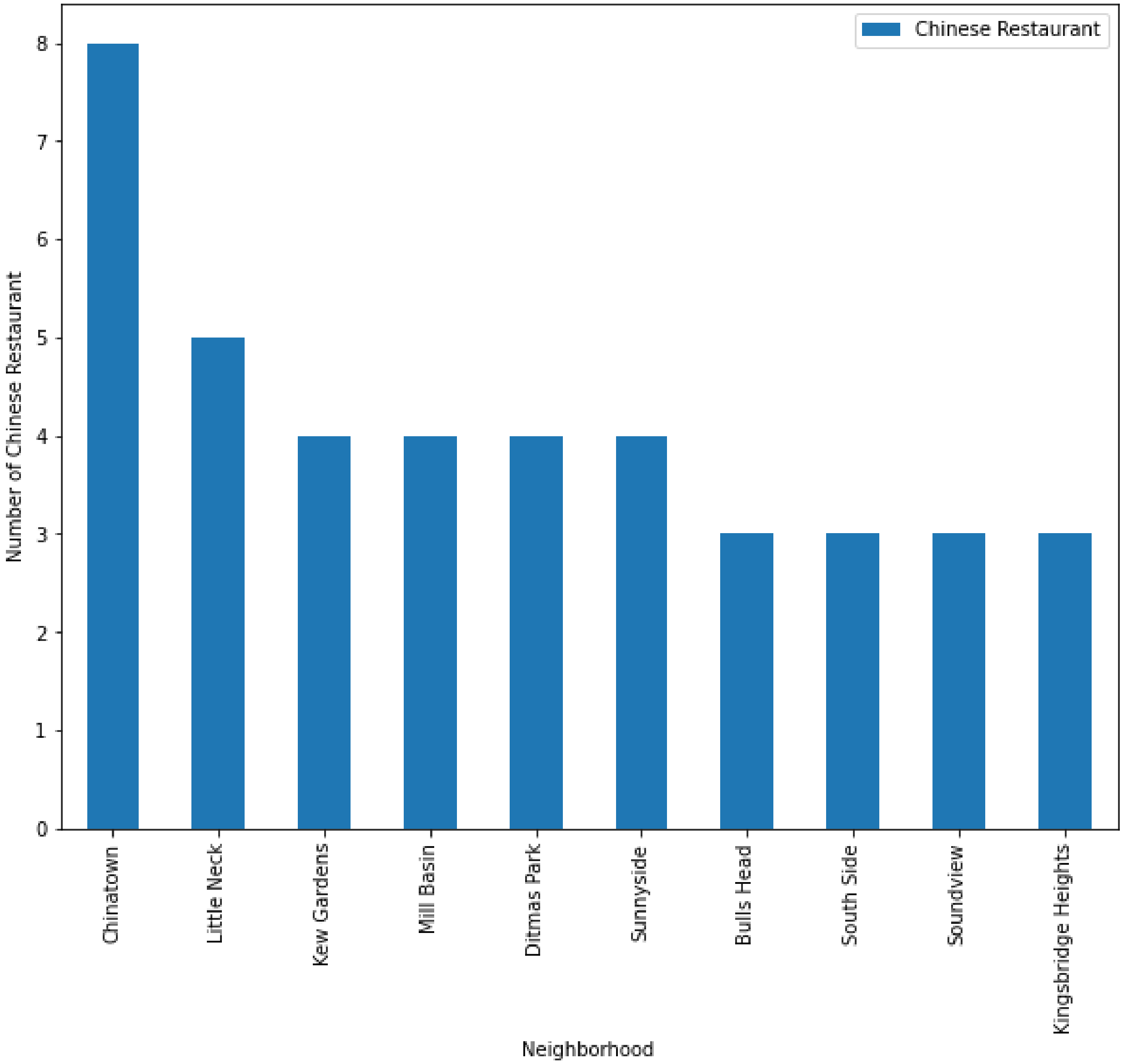
Perform data preprocessing using one hot encoding and take the mean of the frequency of occurrence of each venue category.



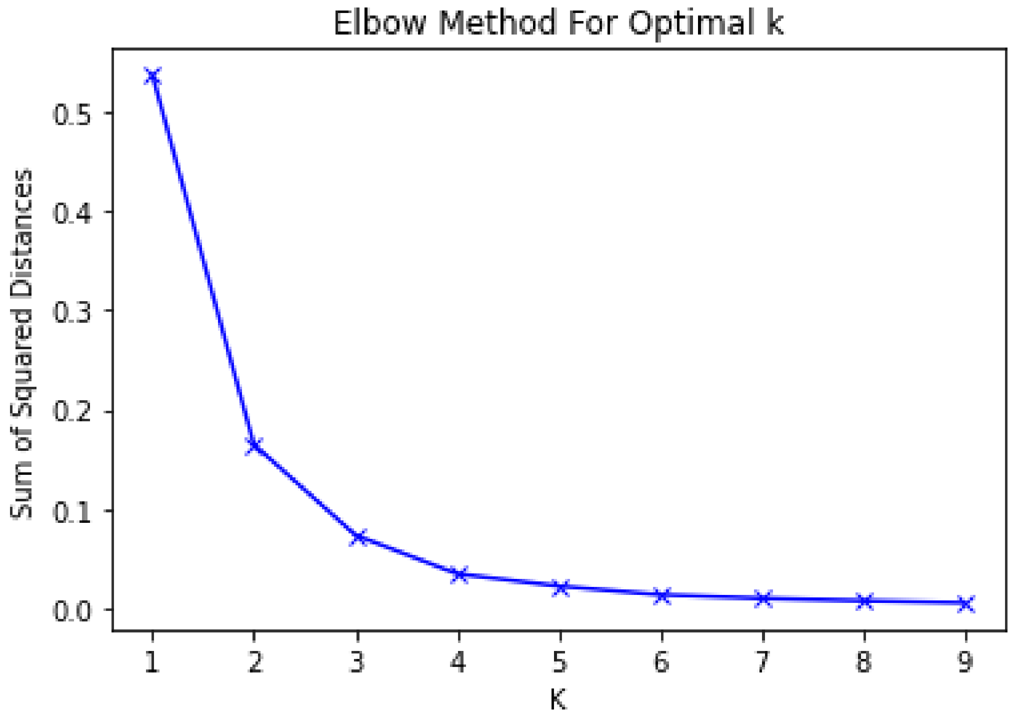
	Neighborhood	Yoga Studio	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Antique Shop	...	Warehouse Store	Waste Facility	Waterfront	Weight Loss Center	Wl
0	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	
1	Annadale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.111111	0.0	...	0.0	0.0	0.0	0.0	
2	Arden Heights	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	
3	Arlington	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	
4	Arrochar	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	

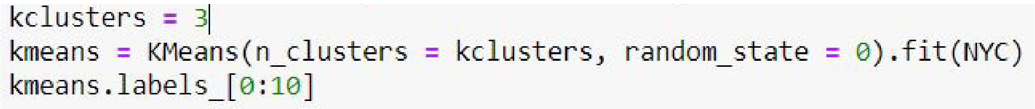
Filter out only Chinese restaurants and visualize them based on boroughs and neighborhoods using histograms and bar charts.



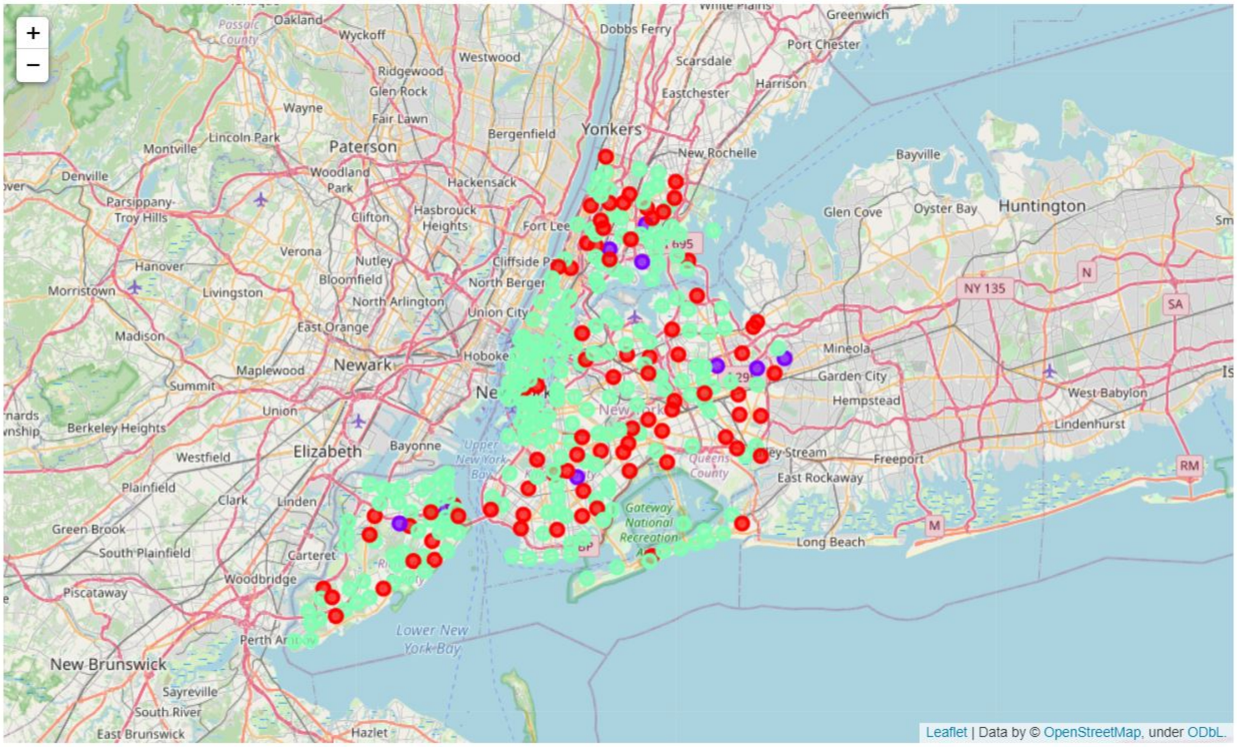


Build a k-means clustering model and visualize results on map using Elbow Method.



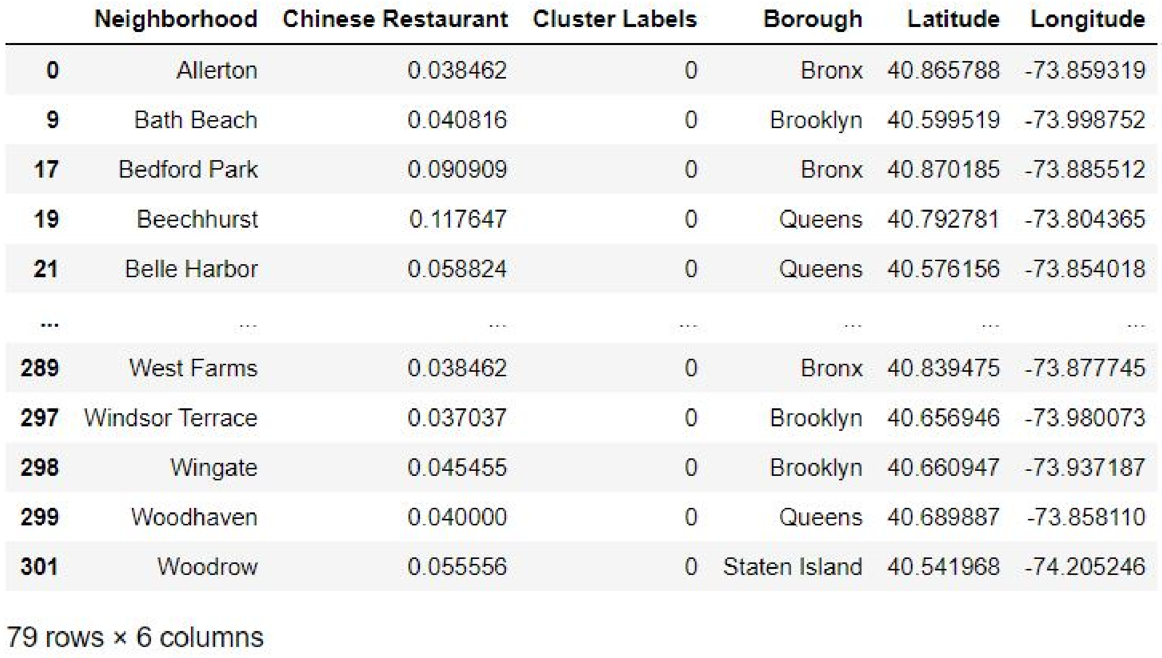


```
kclusters = 3
kmeans = KMeans(n_clusters = kclusters, random_state = 0).fit(NYC)
kmeans.labels_[0:10]
```



Compare neighborhoods to find an optimal location to open a Chinese restaurant.

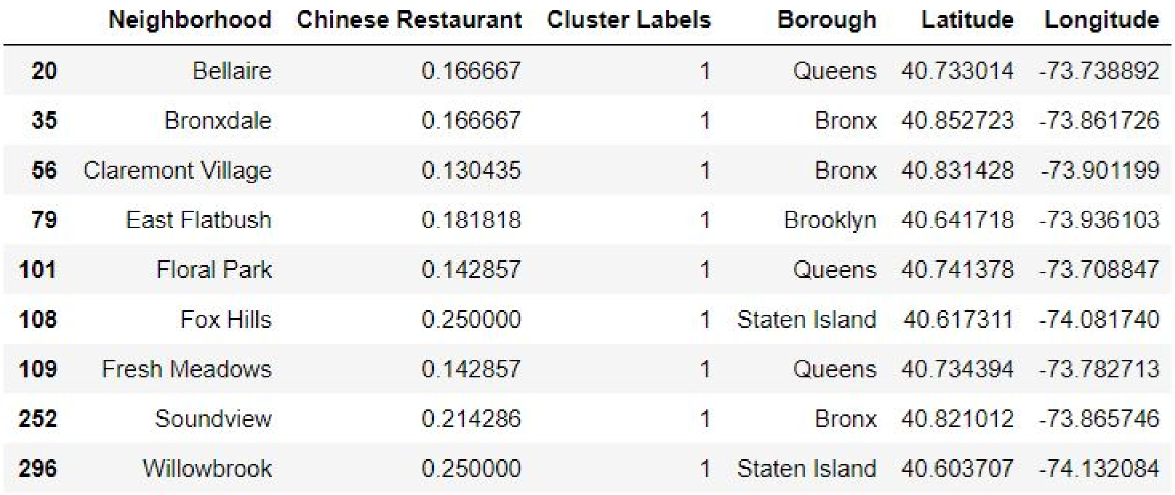
Cluster 0 (red).



	Neighborhood	Chinese Restaurant	Cluster Labels	Borough	Latitude	Longitude
0	Allerton	0.038462	0	Bronx	40.865788	-73.859319
9	Bath Beach	0.040816	0	Brooklyn	40.599519	-73.998752
17	Bedford Park	0.090909	0	Bronx	40.870185	-73.885512
19	Beechhurst	0.117647	0	Queens	40.792781	-73.804365
21	Belle Harbor	0.058824	0	Queens	40.576156	-73.854018
...	...	...	...	...	...	...
289	West Farms	0.038462	0	Bronx	40.839475	-73.877745
297	Windsor Terrace	0.037037	0	Brooklyn	40.656946	-73.980073
298	Wingate	0.045455	0	Brooklyn	40.660947	-73.937187
299	Woodhaven	0.040000	0	Queens	40.689887	-73.858110
301	Woodrow	0.055556	0	Staten Island	40.541968	-74.205246

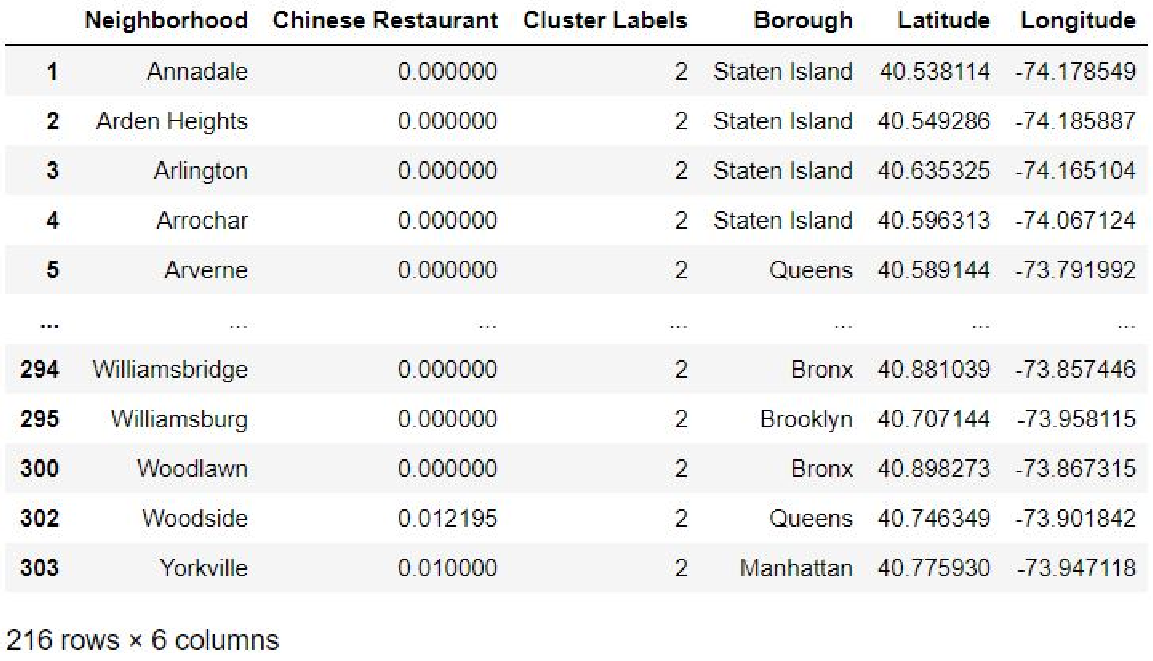
79 rows × 6 columns

Cluster 1 (purple).



	Neighborhood	Chinese Restaurant	Cluster Labels	Borough	Latitude	Longitude
20	Bellaire	0.166667	1	Queens	40.733014	-73.738892
35	Bronxdale	0.166667	1	Bronx	40.852723	-73.861726
56	Claremont Village	0.130435	1	Bronx	40.831428	-73.901199
79	East Flatbush	0.181818	1	Brooklyn	40.641718	-73.936103
101	Floral Park	0.142857	1	Queens	40.741378	-73.708847
108	Fox Hills	0.250000	1	Staten Island	40.617311	-74.081740
109	Fresh Meadows	0.142857	1	Queens	40.734394	-73.782713
252	Soundview	0.214286	1	Bronx	40.821012	-73.865746
296	Willowbrook	0.250000	1	Staten Island	40.603707	-74.132084

Cluster 2 (light green).



	Neighborhood	Chinese Restaurant	Cluster Labels	Borough	Latitude	Longitude
1	Annadale	0.000000	2	Staten Island	40.538114	-74.178549
2	Arden Heights	0.000000	2	Staten Island	40.549286	-74.185887
3	Arlington	0.000000	2	Staten Island	40.635325	-74.165104
4	Arrochar	0.000000	2	Staten Island	40.596313	-74.067124
5	Arverne	0.000000	2	Queens	40.589144	-73.791992
...	...	...	...	...	...	...
294	Williamsbridge	0.000000	2	Bronx	40.881039	-73.857446
295	Williamsburg	0.000000	2	Brooklyn	40.707144	-73.958115
300	Woodlawn	0.000000	2	Bronx	40.898273	-73.867315
302	Woodside	0.012195	2	Queens	40.746349	-73.901842
303	Yorkville	0.010000	2	Manhattan	40.775930	-73.947118

216 rows × 6 columns

Results

To discuss the results, based on the k-means clustering model and its visualization, Cluster 1 (purple) neighborhoods have the largest density of Chinese restaurants; Cluster 2 (light green) neighborhoods have the least density of Chinese restaurants; Cluster 0 (red) neighborhoods have the middle density of Chinese restaurants. So, the optimal location to open a Chinese restaurant in New York City would be these purple dots on map, as most Chinese restaurants are casual restaurants, and *casual restaurants do benefit by clustering near existing ones under the condition that demand is not severely hurt by competition. (To cluster or not to cluster: Understanding geographic clustering by restaurant segment. By Sangwon (Sean) Jung and SooCheong (Shawn) Jang. https://www.sciencedirect.com/science/article/abs/pii/S0278431918302123)*

Discussions

To discuss any observations noted, based on visualizations of Chinese restaurants based on boroughs and neighborhoods using histograms and bar charts, Queens has a high density of Chinese restaurants, while Chinatown (in Manhattan) has the highest number of Chinese restaurants.

Some drawbacks of this analysis are the clustering is completely based on data fetched from Foursquare API, also the analysis does not take into consideration of the Chinese population across neighborhoods as this can play a huge factor while finding an optimal location to open a Chinese restaurant.

So, the recommendation based on the results for local stakeholders and tourists is to open a casual Chinese restaurant in one of these neighborhoods: Bellaire, Floral Park, Fresh Meadows, as they both satisfy the k-means clustering model as well as observations.

Conclusions

To conclude the report, it’s a great opportunity on a business problem, and it’s tackled in a way that it’s similar to how a genuine data scientist would do: using numerous Python libraries to fetch the information, control the content and break down and visualize datasets, using Foursquare API to investigate the settings in neighborhoods of New York City, using different plots present in Matplotlib library, and using Folium to picture on map.

Places that have room for improvement or certain drawbacks meaning that this project can be additionally improved with the assistance of more information and distinctive machine learning strategies. Additionally, this project can be used to investigate any situation, for example, opening an alternate cuisine or opening a movie theater and so forth. Ideally, this project acts as an initial direction to tackle more complex real-life problems using data science.