Analyzing the Coffee Landscape:
New York City (NYC)

Author: Visaj Nirav Shah

September 2020

Problem Statement

Analyze neighborhood clusters of *New York City (NYC), USA* based on the number of coffeehouses and cafes present, the ratings, pricing and number of likes of the stores.

Purpose

NYC is the financial capital of *United States of America (USA)*, a popular tourist destination, has a huge population, and houses offices of important international institutions and major corporate headquarters. In such a landscape, coffeehouses are a crucial part of everyday life. People socialize in coffee shops, hold informal group meetings, relax after a long day, and so on. Hence, it is imperative to analyze the coffee scene in such a city where millions depend on a cup of fresh brew.

Target Audience

- Major Coffeehouse and Cafe Chains
- Prospective Franchisees and Coffeehouse Owners
- Offices and Public Places
- Tourists and Residents

Data

- *Places API - Foursquare Developers*
  - Features like Venue Ratings, Venue Likes, Venue Price Tier, and so on can be obtained from here.
- *2014 New York City Neighborhood Names*
  - The information regarding the neighborhoods of *NYC* and their location can be obtained from here.

Methodology

Getting the Neighborhoods of *New York City*

After downloading the dataset from the source mentioned before, the obtained JSON response is converted into a Pandas DataFrame named NYNghbor.

NYNghbor contains the Latitude, Longitude and Borough details of the neighborhoods.

NYNghbor

[81]:

	Borough	Latitude	Longitude
Neighborhood			
Wakefield	Bronx	40.894705	-73.847201
Co-op City	Bronx	40.874294	-73.829939
Eastchester	Bronx	40.887556	-73.827806
Fieldston	Bronx	40.895437	-73.905643
Riverdale	Bronx	40.890834	-73.912585
...	...	...	...
Hudson Yards	Manhattan	40.756658	-74.000111
Hammels	Queens	40.587338	-73.805530
Bayswater	Queens	40.611322	-73.765968
Queensbridge	Queens	40.756091	-73.945631
Fox Hills	Staten Island	40.617311	-74.081740

Neighborhoods Visualized on NYC Map
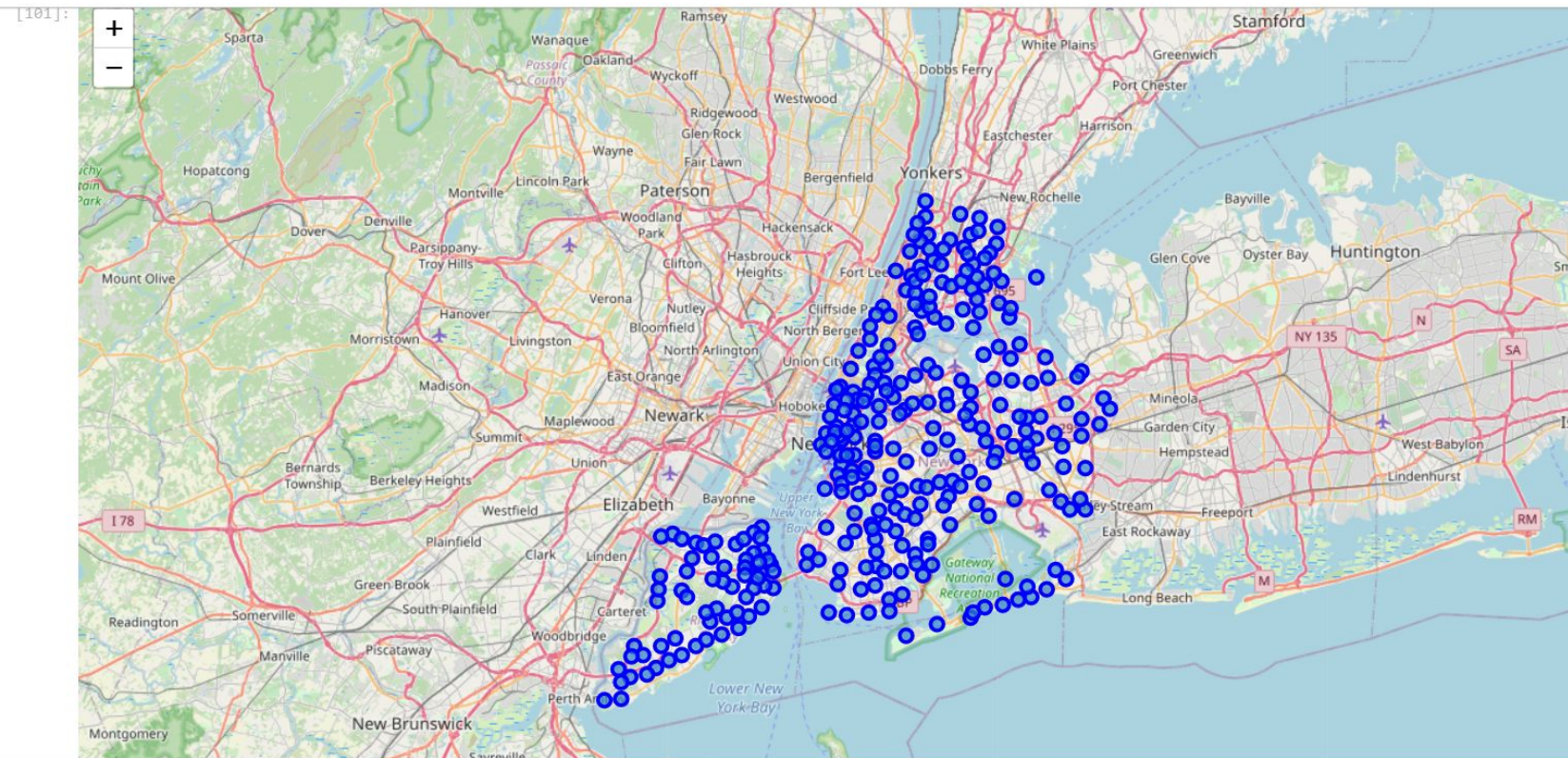
Getting Venues Information

We now use the Foursquare API to get details of various places in these neighborhoods. We will make a call to this API and then store the JSON response in a DataFrame NYVenues.

NYVenues.head()

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue ID	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	4c537892fd2ea593cb077a28	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Ripe Kitchen & Bar	4d375ce799fe8eec99fd2355	40.898152	-73.838875	Caribbean Restaurant
2	Wakefield	40.894705	-73.847201	Ali's Roti Shop	4c9e50e38afca09379b2ff15	40.894036	-73.856935	Caribbean Restaurant
3	Wakefield	40.894705	-73.847201	Jackie's West Indian Bakery	4c10f6aece57c92804a682d2	40.889283	-73.843310	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	Jimbo's	4c1bed4eb306c928140763b7	40.891740	-73.858226	Burger Joint

Filtering out Coffee shops and Cafes

NYVenues.loc[NYVenues['Venue Category'].str.contains('Coffee')]

NYVenues.loc[NYVenues['Venue Category'] == 'Cafe']

Now we make API calls to Foursquare API for getting individual details like ratings, price tier, likes, etc. of each coffee location. All these details will be stored in a DataFrame coffee.

coffee.head()

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue ID	Venue Latitude	Venue Longitude	Venue Category	Venue Price Tier	Venue Rating	Venue Count of Likes
0	Fieldston	40.895437	-73.905643	Mon Amour Coffee & Wine	5660c06b498e4003dba169a5	40.885009	-73.900332	Coffee Shop	1	8.4	22
1	Riverdale	40.890834	-73.912585	Mon Amour Coffee & Wine	5660c06b498e4003dba169a5	40.885009	-73.900332	Coffee Shop	1	8.4	22
2	Kingsbridge	40.881687	-73.902818	Mon Amour Coffee & Wine	5660c06b498e4003dba169a5	40.885009	-73.900332	Coffee Shop	1	8.4	22
3	Kingsbridge	40.881687	-73.902818	Starbucks	55f81cd2498ee903149fcc64	40.877531	-73.905582	Coffee Shop	1	8.2	24
4	Marble Hill	40.876551	-73.910660	Starbucks	55f81cd2498ee903149fcc64	40.877531	-73.905582	Coffee Shop	1	8.2	24

Grouped by Neighborhoods Data

We observe that there are 214 neighborhoods in coffee. We will be clustering these 214 neighborhoods. To get details for each neighborhood, we create a DataFrame NYNghborCoffee.

NYNghborCoffee.head()

[94]:

Neighborhood	Number of Coffee Locations	Price Tier	Average Rating	Number of Likes
Allerton	1	1.0	8.40	12.0
Annadale	2	1.0	7.75	19.0
Arden Heights	2	1.0	7.75	19.0
Arlington	1	1.0	8.40	12.0
Arverne	1	1.0	8.70	13.0

Clustering

For clustering, we will use the k-means algorithm. We will implement the algorithm using *scikit-learn* library available for *Python*.

To determine the perfect k (i.e. the number of clusters), we loop through a reasonable range of k values. Here, this range is [2, 10].

Elbow Method

For each k, we plot the SSE (i.e. Sum of Squared Distance) and then using the Elbow Method, we determine the required k.

According to the Elbow Method, the best k is the one after which the SSE vs. k plot is almost linear and has very little change in slope.

For better results, we will standardize the data before calculating the SSE. This means the mean of each column will be 0 and the standard deviation will be 1.

Plot for Elbow Method

[102]: [<matplotlib.lines.Line2D at 0x7f3875f6a240>]
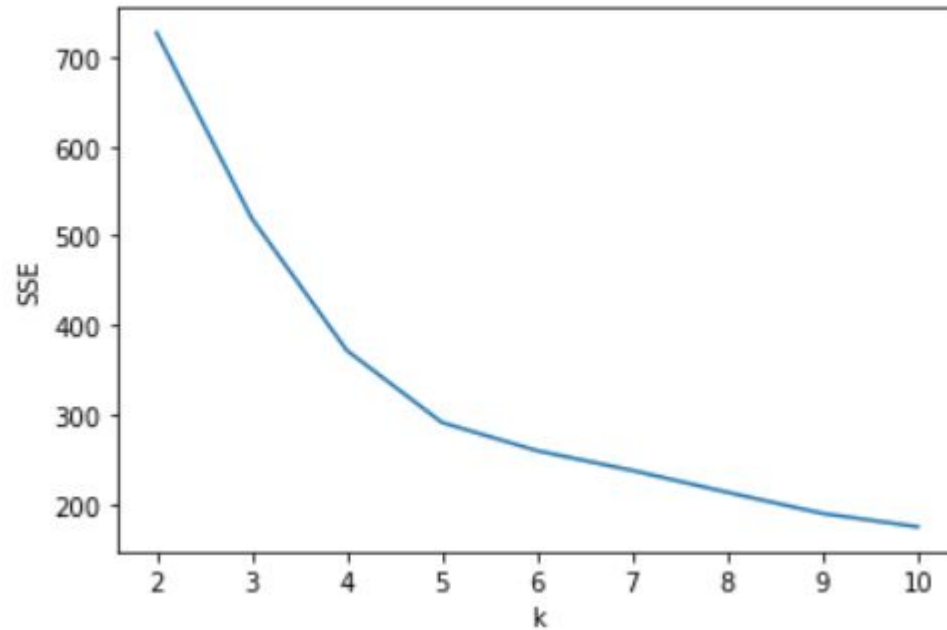
After looking at this plot, we can clearly say that k = 5 is a reasonable choice for number of clusters. We run the k-means algorithm for our data keeping **k = 5**.

Results

Output of k-means clustering

```
[97]: array([1, 2, 2, 1, 1, 4, 2, 2, 0, 1, 2, 4, 1, 0, 0, 2, 4, 2, 2, 1, 1, 1,
        2, 2, 1, 2, 1, 4, 2, 1, 1, 1, 2, 1, 0, 4, 1, 2, 4, 1, 0, 1, 1, 4,
        1, 2, 0, 4, 2, 0, 3, 2, 2, 2, 1, 0, 1, 1, 1, 2, 2, 1, 1, 0, 0, 1,
        2, 2, 1, 1, 2, 1, 1, 2, 1, 1, 1, 2, 2, 0, 2, 0, 1, 0, 1, 1, 1, 2,
        1, 2, 1, 4, 2, 4, 1, 2, 2, 0, 1, 1, 1, 0, 2, 1, 3, 2, 0, 2, 0, 2,
        0, 0, 0, 2, 4, 1, 2, 2, 2, 3, 1, 4, 2, 2, 0, 0, 2, 1, 0, 1, 2, 2,
        1, 1, 2, 4, 2, 2, 0, 2, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 4, 2,
        1, 0, 1, 2, 1, 2, 1, 2, 1, 4, 1, 2, 1, 1, 1, 1, 2, 1, 1, 0, 0, 1,
        2, 1, 1, 0, 2, 1, 1, 4, 4, 0, 0, 0, 1, 1, 1, 1, 4, 1, 0, 1, 1, 3,
        1, 3, 0, 0, 0, 1, 4, 2, 2, 2, 2, 2, 1, 1, 0, 0], dtype=int32)
```

NYNghborCoffee.head() after adding Cluster Num

[98]:

Neighborhood	Number of Coffee Locations	Price Tier	Average Rating	Number of Likes	Cluster Number
Allerton	1	1.0	8.40	12.0	1
Annadale	2	1.0	7.75	19.0	2
Arden Heights	2	1.0	7.75	19.0	2
Arlington	1	1.0	8.40	12.0	1
Arverne	1	1.0	8.70	13.0	1

Description of Each Cluster

[99]:

Cluster Number	Number of Coffee Locations	Price Tier	Average Rating	Number of Likes
0	3.731707	1.160163	8.299512	90.191870
1	1.418605	1.000000	8.514535	53.819767
2	1.587302	1.000000	7.823413	30.392857
3	2.600000	1.853333	8.633333	646.000000
4	1.894737	1.675439	8.442105	170.447368

Discussion

Cluster 3

Cluster 3 has the best coffee places in the city. With highest number of likes and highest average rating, these locations are extremely popular. The stores in this cluster are a bit pricey, but they serve products worth the price given the popularity and demand. Dicult for new stores to make a mark since there is little scope of improving customer experience.

Cluster 4

Cluster 4 is again a coffee-friendly cluster but with reduced number of locations. The density of stores is less, but customers have vouched for the quality of service (high ratings and number of likes). This cluster is a good choice for newcomers in the business who plan to start a non-franchise store. Given the demand but less number of locations, newcomers stand a good chance of attracting customers.

Cluster 0

Cluster 0 is not a good place for new businesses. The density of stores is very high as compared to others and popularity is low. The ratings and number of likes are lower than Cluster 3 and 4. Possible reasons for low popularity could include unsatisfactory service and products, or inherent nature of locations such that the demand for coffee is less. In either case, rather than new stores opening, some of the old ones may be pulling down the shutters forever. One possibility for high number of stores and low demand could mean that there are a lot of corporate coffee shops meaning the ones managed by companies exclusively for their employees and clients.

Cluster 1

Cluster 1 has a low density of stores but positive ratings. Given the low number of likes, one can deduce that not many are coffee-consumers, but those who do, have good options available. One could consider opening a store here since there is a demand (specific audience) but the number of options are really less. So, there is a high probability that people are willing to test new options and possibly make a shift. If you come up with some unique feature for your store, it could generate interest among both, regular customers, and new customers.

Cluster 2

Cluster 2 is clearly not a good place to open a coffee shop, or the best if you are a risk-taker and want to risk entering a completely new domain. These neighborhoods have very few stores, but since there are few people to drink, there are not many changes taking place in the landscape. Given the low ratings, people are probably not happy with their current options, so new stores could stand a chance of attracting the crowd, but they need to be really good to get people to drink coffee.

Conclusion

Possible Improvements and Future Scope

Clearly, this is not an exhaustive model for analysis. Some other factors which can be included are area of each neighborhood, population of individual neighborhood, real estate prices, on-field public response, and so on.

This study could be expanded to other major cities across the world as well. This should be of help, especially to multinational companies. Detailed analysis for entire countries can help them understand the soft spots of their business model and market.

All in all, after looking at the results and analyzing the process, one can decide on a location depending on their requirements. After understanding the Coffee landscape, people are now equipped to make intelligent business decisions and churn out maximum profits depending on the objectives. Also, as explained at the beginning, this study can empower tourists and citizens too, not only businesses.

Thank You