THE BATTLE OF NEIGHBOURHOODS

IBM APPLIED DATA SCIENCE CAPSTONE

By: AAYUSH DUA

FEB 2019

INTRODUCTION

The City of New York, is the most populous city in the United States. It is diverse and is the financial capital of USA. It is multicultural. It provides lot of business opportunities and business friendly environment. It has attracted many different players into the market. It is a global hub of business and commerce. The city is a major center for banking and finance, retailing, world trade, transportation, tourism, real estate, new media, traditional media, advertising, legal services, accountancy, insurance, theatre, fashion, and the arts in the United States. This also means that the market is highly competitive. As it is highly developed city so cost of doing business is also one of the highest. Thus, any new business venture or expansion needs to be analysed carefully. The insights derived from analysis will give good understanding of the business environment which help in strategically targeting the market. This will help in reduction of risk. And the Return on Investment will be reasonable.

BUSINESS PROBLEM

The objective of this project is to recommend the most favourable neighbourhood for setting up an Indian Restaurant in New York City. The major issues faced while opening any restaurant are rentals and competition. Hence to open up a Indian Restaurant in NYC we will need to recommend a neighbourhood that has reasonable rent and least competition from other Indian Restaurants. Furthermore we will try to suggest restaurants that are in close proximity to high rentals neighbourhoods since these neighbourhoods will be the tourist attraction and place for many offices, university and hotels. Hence having a restaurant nearby to these neighbourhoods will attract residents and tourists from these locations to the restaurant.

TARGET AUDIENCE

Stakeholders: Any Businessman that wants to open a Indian Restaurant in New York City. This project will help them decide which neighbourhood will be the best option to open up the restaurant.

Consumers: The residents of New York City and tourists who come to visit New York City who want to taste Indian food. Hence opening a restaurant in a neighbourhood which has less proximity to high rent neighbourhoods where usually the hotels, offices and universities are located will give the restaurant an advantage to attract consumers who reside in these areas as these areas will be more populous when compared to other areas thus increasing the check-in count of the restaurant.

DATA

To solve the problem, we will need the following data:

1. List of neighbourhoods in New York City. Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data. https://geo.nyu.edu/catalog/nyu_2451_34572
2. The rent of neighbourhoods in New York City.(https://www.zillow.com/research/data/)
3. From Foursquare we will need to find the venues related to Indian Restaurants. By using data we will be able to cluster the neighbourhoods based on the number of Indian Restaurants in each neighbourhood. This will help us to determine which neighbourhood will have least competition for establishing an Indian Restaurant.
4. From Distance Matrix API of Google Cloud Platform we will extract distance matrix between different neighbourhoods.

We will then leverage the data in order to determine which locality is the most appropriate in order to locate the Indian Restaurant.

METHODOLOGY

1. The first step is to list all the neighborhoods along with their latitude and longitude. For this we will make use of the json file which was provided in the cognitive lab session. Using this json file we will extract the features required to list the neighborhoods and their respective coordinates. After extracting we will convert them to dataframe for convenience to access the values.
2. Next using the geopy library we will access the coordinates of New York City. Using Folium library we will use the circle markers to mark the neighborhoods of New York City on the map of New York for visualization of where they are located.
3. After we have visualized the map we will set up my foursquare credentials. Foursquare API will be used to access the top 100 venues in every category within a radius of 500metres of every neighborhood. The Foursquare API will return a json format which we will have to format into pandas dataframe.
4. Now that we have the venue dataset of each neighborhoods, we will analyze each neighborhood for its mean frequency occurrence of Indian Restaurants. We will only extract the Neighborhood and the Indian Restaurant Column for our further exploration.
5. Next step is to find the rent of every neighborhood and find the top 5 neighborhoods having the highest rental. For this we have used the Zillow Rental Index provided from Zillow Research website. The top 5 neighborhoods with highest rentals are Tribeca, North Hollywood, South Hollywood, West Village, and Carnegie Hill. We need to find neighborhoods that are in close proximity to these neighborhoods and at the same time have reasonable rents.
6. For finding the distance between the neighborhoods, we use Distance Matrix API provided by Google Cloud Platform. The response is a json format which we will convert into pandas dataframe. Next we will eliminate all those neighborhoods that do not lie within the 5km distance of the neighborhoods with the highest rentals.
7. Now we will club the rent data with the frequency occurrence of the Indian Restaurant in each neighborhood. Before performing the kmeans clustering we will perform normalization of rent and frequency so that both have equal effect in cluster formation.

8. Finally we will cluster the neighborhoods according to their rents and frequency and visualize the clusters on the map of New York City for better understanding of the result.

RESULTS

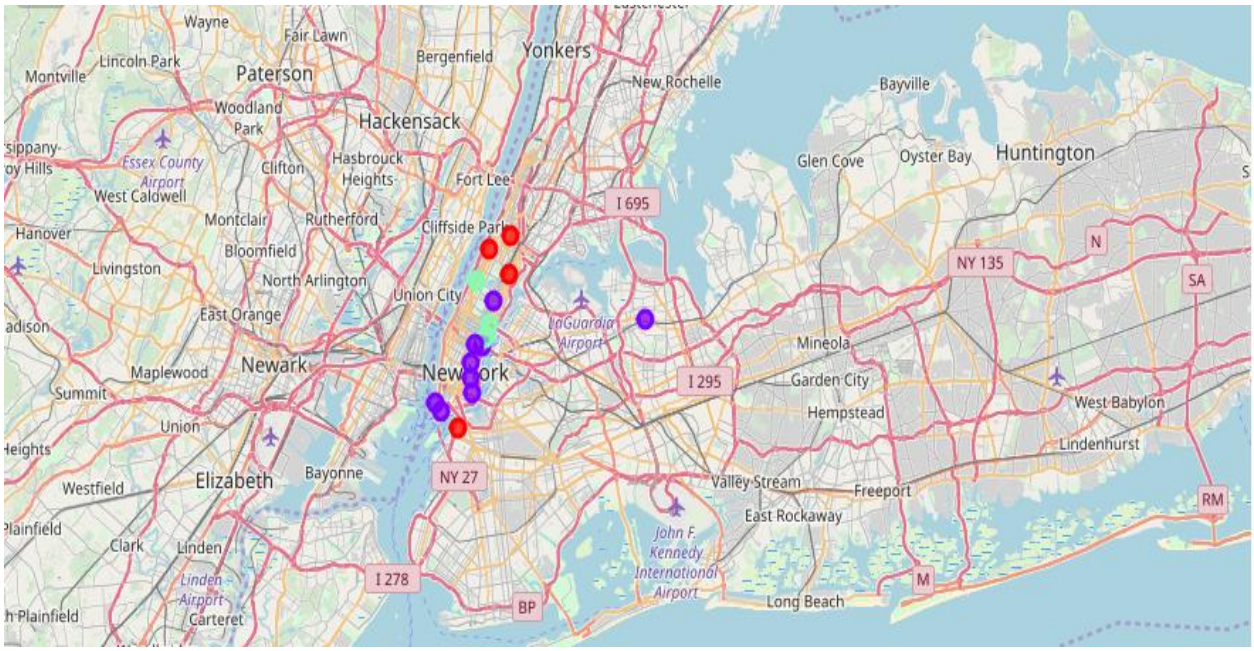
The results from the k-means clustering show that we can categorize the neighbourhoods into 3 clusters based on the frequency of occurrence for “Indian Restaurant” and the Rentals of Each Neighborhood:

• Cluster 0: Neighbourhoods with lowest rent and least competition.

• Cluster 1: Neighbourhoods with highest rent and moderate competition.

• Cluster 2: Neighbourhoods with moderate rent and very high competition.

The results of the clustering are visualized in the map below with cluster 0 in red color, cluster 1 in purple color, and cluster 2 in mint green color.



Cluster 0:

	Neighborhood	Re	Indian Restaurant	Re_zscore	Indian Restaurant_zscore	Cluster Labels	Latitude	Longitude
12	Morningside Heights	3171	0.023256	-1.016734	0.911422	0	40.808000	-73.963896
13	Brooklyn Heights	2949	0.020000	-1.621868	0.671574	0	40.695864	-73.993782
14	Central Harlem	2879	0.000000	-1.812676	-0.801777	0	40.815976	-73.943211
15	East Harlem	2869	0.000000	-1.839934	-0.801777	0	40.792249	-73.944182

Cluster 1:

	Neighborhood	Re	Indian Restaurant	Re_zscore	Indian Restaurant_zscore	Cluster Labels	Latitude	Longitude
0	East Village	3920	0.00	1.024912	-0.801777	1	40.727847	-73.982226
3	Battery Park City	3825	0.01	0.765958	-0.065102	1	40.711932	-74.016869
4	Lower East Side	3814	0.00	0.735974	-0.801777	1	40.717807	-73.980890
5	Financial District	3798	0.01	0.692360	-0.065102	1	40.707107	-74.010665
6	Tudor City	3757	0.00	0.580602	-0.801777	1	40.746917	-73.971219
7	Gramercy	3732	0.00	0.512456	-0.801777	1	40.737210	-73.981376
9	Murray Hill	3621	0.00	0.209889	-0.801777	1	40.748303	-73.978332
10	Murray Hill	3621	0.00	0.209889	-0.801777	1	40.764126	-73.812763
11	Upper East Side	3500	0.00	-0.119936	-0.801777	1	40.775639	-73.960508

Cluster 2:

	Neighborhood	Re	Indian Restaurant	Re_zscore	Indian Restaurant_zscore	Cluster Labels	Latitude	Longitude
1	Sutton Place	3843	0.04	0.815023	2.144926	2	40.760280	-73.963556
2	Upper West Side	3826	0.03	0.768684	1.408250	2	40.787658	-73.977059
8	Turtle Bay	3656	0.03	0.305293	1.408250	2	40.752042	-73.967708

FINAL RECOMMENDATION

From the above set of clusters we can see that cluster 0 is the most favorable set of neighborhoods to open up a Indian Restaurant with least competition and least rent.

Especially Harlem has the least rent and least competition. Hence it will be a very favorable neighborhood. However when we see the map we observe that Harlem is only close to Carnegie Hill. On the other hand Brooklyn Heights also resides in the cluster 0 with less rent and moderate competition. The advantage of Brooklyn Heights is that it comes within the 5km distance of Tribeca, Soho, Noho and West Village. Hence it has a very advantageous location attracting tourists from these areas. Hence I will recommend **Brooklyn Heights** as the most favorable neighborhood to open up the Indian Restaurant.

CONCLUSION

In this project we will able to successfully recommend the location that will be most favorable for opening up a Indian Restaurant. We used Foursquare API to generate venues and their frequency. By this we were able to extract the mean frequency of Indian Restaurant in each neighbourhood. Using Rent Dataset we were able to find neighborhoods that had reasonable rents and using Distance Matrix API we were able to find the neighborhoods having reasonable rent and located in the vicinity of high rental neighborhoods. Finally we performed Kmeans clustering on the rent and completion values of the final list of neighborhoods thus generating clusters and choosing the neighborhood that had the least rent and least completion.

For future work I would like to incorporate many other factors while recommending the location such as the proximity of Farmers Market from the neighborhood which will reduce a lot of transportation cost. By knowing the racial decomposition of population of each neighborhood we can suggest neighborhood having high racial decomposition in favor of

Asians since they will be the major customers. Finally we can classify the neighborhood based on the other venues that lie in the neighborhoods such as how many offices, universities and schools and residential places are in each neighborhood since we will be able to know our target audience.