Business opportunity in New York

Hoang Do

April 2020

I. Business Problem

1. Business interest

An international company want to invest in a service in New York. Two of the most important questions are:

1. What is the most interested service in New York?
2. Which areas (neighborhoods) in New York are in short of this service?

2. Resolution

To answer 02 important questions, data about neighborhoods and their venues in New York must be collected and analyzed.

1. Collect data about neighborhoods
2. Collect data about venues of these neighborhoods
3. Find the most interested venue category (the service is most interested in New York)
4. Analyze the data to find neighborhoods (areas) having no the most interested venue category

II. Data and Methodology

1. **Process data about neighborhoods in New York**

Data about neighborhoods in New York can be collected from official online sources such as https://cocl.us/new_york_dataset

The neighborhood data includes critical information such as names, latitude and longitude.

The neighborhood data must be cleaned to reach criteria of having unique identification, no duplicate data.

2. **Process data about venues of these neighborhoods**

Based on the neighborhood data with certain latitudes and longitude, data about venues are collected from source of Foursquare.

To avoid processing data too long, the venue data must be not excessive, so some limitations are:

i. Radius of a particular location (neighborhood with latitude and longitude) is 1000 kilometers

ii. Maximum number of venues of a particular neighborhood is 100

The venue data must show the interested rate of each venue category of a particular neighborhood

3. **Find the most interested venue category (the service is most interested in New York)**

The most interested service in New York is the venue category having maximum rate

4. **Analyze the data to find neighborhoods (areas) having no the most interested venue category**

Neighborhoods are clustered based on their similarity of venues

The answer is the cluster including no the most interested venue category

III. Processing data

- Load data from link https://cocl.us/new_york_dataset into file ***newyork_data.json***
- Create dataframe ***neighborhood_df*** from file ***newyork_data.json***

The dataframe ***neighborhood_df*** has 5 unique boroughs and 306 unique neighborhoods

	Neighborhood_ID	Neighborhood_name	Borough	Latitude	Longitude
0	Wakefield40.89470517661-73.84720052054902	Wakefield	Bronx	40,89470518	-73,84720052
1	Co-op City40.87429419303012-73.82993910812398	Co-op City	Bronx	40,87429419	-73,82993911
2	Eastchester40.887555677350775-73.82780644716412	Eastchester	Bronx	40,88755568	-73,82780645
3	Fieldston40.89543742690383-73.90564259591682	Fieldston	Bronx	40,89543743	-73,9056426
4	Riverdale40.890834493891305-73.9125854610857	Riverdale	Bronx	40,89083449	-73,91258546

- Create dataframe ***NY_venues*** including all venues in radius of 1,000 kilometers around each neighborhood in dataframe ***neighborhood_df***

Based on the neighborhood data with certain latitudes and longitude, data about venues are collected from source of Foursquare (maximum number of venues of a particular neighborhood is 100)

The dataframe ***NY_venues*** has 20,673 venues with 480 venue categories

	Neighborhood_ID	Neighborhood_name	Neighborhood_Lat	Neighborhood_Long	Venue	Venue_Lat	Venue_Long	Venue_Category
0	Wakefield40.894705176609996-73.84720052054901	Wakefield	40,89470518	-73,84720052	Lollipops Gelato	40,89412315	-73,84589162	Dessert Shop
1	Wakefield40.894705176609996-73.84720052054901	Wakefield	40,89470518	-73,84720052	Ripe Kitchen & Bar	40,89815169	-73,838875	Caribbean Restaurant
2	Wakefield40.894705176609996-73.84720052054901	Wakefield	40,89470518	-73,84720052	Ali's Roti Shop	40,8940357	-73,85693494	Caribbean Restaurant
3	Wakefield40.894705176609996-73.84720052054901	Wakefield	40,89470518	-73,84720052	Carvel Ice Cream	40,89048669	-73,84856773	Ice Cream Shop
4	Wakefield40.894705176609996-73.84720052054901	Wakefield	40,89470518	-73,84720052	Jimbo's	40,89174013	-73,85822585	Burger Joint

- Transform data of dataframe NY_venues to dataframe NY_onehot with rows are venues and columns are venue categories

Value of a column (category) is 1 if the venue belongs to that category, else 0.

The dataframe ***NY_onehot*** has 20,673 venues with 480 columns of venue categories and 1 column of ***neighborhood_id***

- Group venues by Neighborhood_ID from dataframe ***NY_onehot*** to dataframe ***NY_grouped*** with calculating the mean of the frequency of occurrence of each venue category

The dataframe ***NY_grouped*** has 306 neighborhoods with 480 venue categories

	Neighborhood_ID	ATM	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant
0	Allerton40.86578787802982-73.85931863221647	0	0	0	0	0
1	Annadale40.53811417474507-74.17854866165878	0	0	0	0	0
2	Arden Heights40.54928582278321-74.18588674583894	0	0	0	0	0
3	Arlington40.63532509911492-74.16510420241123	0	0	0	0	0
4	Arrochar40.596312571276734-74.06712363225573	0	0	0	0	0

- Find the most venue category (the service is most interested in New York)

The most interested service in New York is the venue category having maximum value in dataframe ***NY_grouped***

The most interested service in New York is ***Beach***

- Cluster dataframe ***NY_grouped*** by kMeans with k from 3 until finding at least a cluster having no the most interested venue category (***Beach***)

Create dataframe ***NY_merged*** by merging clustered dataframe ***NY_grouped*** with dataframe ***neighborhood_df*** and adding Cluster Label to each neighborhood

Dataframe ***NY_merged*** has full information including name and Cluster Label

	Neighborhood_ID	Neighborhood_name	Borough	Latitude	Longitude	Cluster Labels	ATM	Accessories Store	Adult Boutique
0	Wakefield40.894705176661-73.84720052054902	Wakefield	Bronx	40,89470518	-73,84720052	0	0	0	0
1	Co-op City40.87429419303012-73.82993910812398	Co-op City	Bronx	40,87429419	-73,82993911	0	0	0	0
2	Eastchester40.887555677350775-73.82780644716412	Eastchester	Bronx	40,88755568	-73,82780645	0	0	0	0
3	Fieldston40.89543742690383-73.90564259591682	Fieldston	Bronx	40,89543743	-73,9056426	3	0	0	0
4	Riverdale40.890834493891305-73.9125854610857	Riverdale	Bronx	40,89083449	-73,91258546	3	0	0	0

- Check number of neighborhoods having the most interested venue category in each cluster

Cluster 0 has 3 neighborhoods including Beach

Cluster 1 has 3 neighborhoods including Beach

Cluster 2 has 9 neighborhoods including Beach

Cluster 3 has 14 neighborhoods including Beach

Cluster 4 has 0 neighborhoods including Beach

Cluster 5 has 5 neighborhoods including Beach

- The ***target_clusters*** list all neighborhoods having no the most interested venue category (***Beach***). These neighborhoods are potential areas to invest services directly related to ***Beach***

	Neighborhood_ID	Neighborhood_name	Borough	Latitude	Longitude	Cluster Labels	Beach
207	Port Ivory40.63968297845542-74.17464532993542	Port Ivory	Staten Island	40,63968298	-74,17464533	4	0
227	Arlington40.63532509911492-74.16510420241124	Arlington	Staten Island	40,6353251	-74,1651042	4	0

IV. Conclusion

After processing and analyzing data, the ***target_clusters*** including neighborhoods of ***Port Ivory*** & ***Arlington*** has no ***Beach*** (the most interested venue in New York). In these areas, business opportunities should relate directly to ***Beach*** such as:

- Shuttle bus from neighborhood center or big hotels to interesting beaches
- Swimming pools or restaurants on top of buildings have view toward beautiful beaches