Coursera Capstone

Coursera IBM Data science Certificate

Inácio Kisu Sung

Dec 08th, 2019.

1. Introduction.

The New York City metropolitan region is one of the most important economic regions in the world, as a center of many industries, including finance, international trade, news and traditional media, real estate, education, fashion, entertainment, tourism, biotecnology, law, and manufacturing.

One unique caracteristic of New York city is the diversity of ethnics groups that have entered the United States. So almost all ethnic cuisines are well represented, both within and outside the various ethnics neighborhoods.

Americans and People from all over the world love the culinary diversity available in the city. And one in particular that is very appreciated is the Korean cuisine.

Business Problem.

Based on the introduction presented above, what if someone wanted to open a new Korean restaurant? Where would you recommend that they open it?

Target Audience.

This report is very interesting for investors, real estate owners, entrepreneurs and professional chefs that want to run their own business.

2. Data Section.

Description of the Data.

The following data is required to solve the problem:

A data set , which is a json file (newyork_data.json), that contains a list of boroughs and neighborhoods of New York with geographical coordinates such as latitude and longitude.

A list of Korean Restaurants and their respective geographical coordinates in New York using Foursquare API.

How the data will be used to solve the problem.

The data will be used as follows:

- Load, explore and transform the data into a dataframe to obtain a list of Boroughs and Neighborhoods and their respective geographical coordinates.
- Use geopy library to get the latitude and longitude values of New York City.
- Create a map of New York with neighborhoods superimposed on top.
- Define a query to search for Korean Restaurant that is within 5000 meters from New York city.
- Create a dataframe for Korean Restaurant data only.
- Use machine learning (K-means clustering) to cluster Neighborhoods with Korean Restaurants.
- Visualize all clusters in a map using Folium.

This data analysis will facilitate one to decide where to open a new Korean Restaurant based on the distribution of the actual ones in New York city.

3. Methodology.

First of all we downloaded the file "newyork_data.json" from this link https://geo.nyu.edu/catalog/nyu_2451_34572, which contains a list of neighborhoods with their respectives geographical coordinates.

We used geopy library to get the latitude and longitude values of New York City and then create a map of New York with neighborhoods superimposed on top.

We used Foursquare API to define a query to search for Korean Restaurants that is within 5000 meters from New York city and then create a dataframe with only Korean restaurants.

Then we created a function to get the top 100 venues that are in all neighborhoods within a radius of 5000 meters in New York city and check whether there are Korean Restaurants in the top 100 list or not.

After that we analyzed each neighborhood and we grouped rows by neighborhood and created a dataframe by taking the mean of the frequency of occurrence of Korean Restaurants.

So with this data, We used k-means clustering algorithm to cluster Neighborhoods with Korean Restaurants.

Finally We used the Folium library to visualize the Neighborhoods with Korean Restaurants in New York City.

4. Results.

The result from the k-means clustering algorithm showed us that it was reasonable to cluster Neighborhoods into 3 clusters.

	Neighborhood	Korean Restaurant	Cluster Labels	Borough	Latitude	Longitude
0	Allerton	0.000000	0	Bronx	40.865788	-73.859319
1	Annadale	0.000000	0	Staten Island	40.538114	-74.178549
2	Arden Heights	0.000000	0	Staten Island	40.549286	-74.185887
3	Arlington	0.000000	0	Staten Island	40.635325	-74.165104
4	Arrochar	0.000000	0	Staten Island	40.596313	-74.067124

Fig1. - Cluster 0 > Zero to low number of Korean Restaurants.

	Neighborhood	Korean Restaurant	Cluster Labels	Borough	Latitude	Longitude
174	Midtown South	0.150000	1	Manhattan	40.748510	-73.988713
185	Murray Hill	0.149660	1	Manhattan	40.748303	-73.978332
185	Murray Hill	0.149660	1	Queens	40.764126	-73.812763
197	Oakland Gardens	0.130435	1	Queens	40.745619	-73.754950

Fig2. - Cluster 1 > High number of Korean Restaurants.

	Neighborhood	Korean Restaurant	Cluster Labels	Borough	Latitude	Longitude
8	Auburndale	0.052632	2	Queens	40.761730	-73.791762
31	Brighton Beach	0.022222	2	Brooklyn	40.576825	-73.965094
81	East Village	0.020000	2	Manhattan	40.727847	-73.982226
100	Flushing	0.035088	2	Queens	40.764454	-73.831773
105	Fort Hamilton	0.015873	2	Brooklyn	40.614768	-74.031979

Fig3. - Cluster 2 > Medium number of Korean Restaurants.

We can see all 3 clusters in the map below.

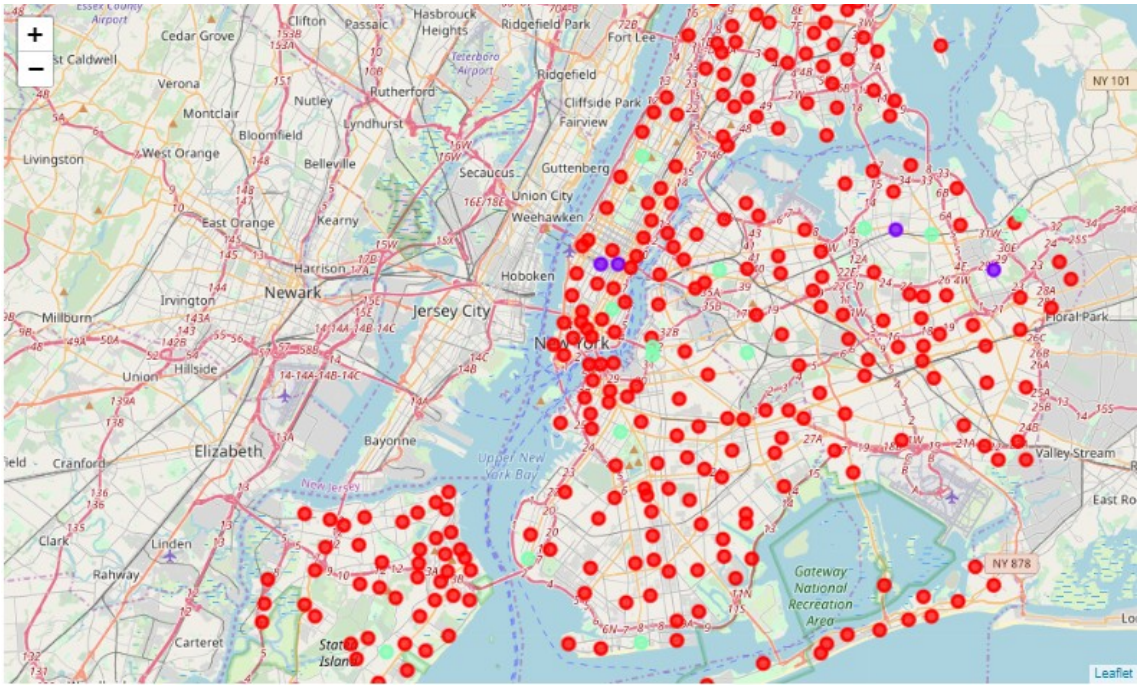

Fig.4 - Map with all 3 clusters > Cluster 0 - red / Cluster 1 - purple / Cluster 2 - green.

5. Discussion.

It is a very simple analysis based on a list of Neighborhoods of New York city and also on a list of Korean Restaurants and its locations extracted from Foursquare API.

We can definitely deduce from the map visualization that there are a lot of Neighborhoods in New York (Fig.4) that do not have a single Korean Restaurant. But that does not mean that it is a strong and good reason to open a new one in these areas.

The analysis is very helpful but it can not be the only source of information. Since this research was done from a limited source of data, for a complete analysis We could go further in this research. For example, it could be combined with others such as real estate prices, location of corporate buildings, location of Korean Community Neighborhoods, and also the location of top touristic Neighborhoods.

6. Conclusion.

Although We have some considerations as cited in the Discussion section, we could recommend a lot of Neighborhoods based on the result of our clustering which would be the list of Neighborhoods in Cluster 0.