Coursera IBM Data Science Capstone Project : Opening a new Korean Restaurant in New York

2020.05

Introduction

For this Capstone project, a Korean international student would like to spare his time to develop his own entrepreneurship by opening up a new Korean cuisine. Korean cuisine is popular among students as well as attracting those residences from Korea who are temporarily living in New York. Hence, to locate the populations who have the strongest desire for Korean cuisine, we may attempt to find out where could be a good place for his business.

Business Problem

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Korean restaurant in New York, China. By using data science methods and machine learning methods such as clustering, this project aims to provide solutions to answer the business question: In New York, if an entrepreneur wants to open a Korean restaurant, where should they consider opening it?

Target Audience

The entrepreneur who wants to find the location to open authentic Korean restaurant in New York.

Method:

To solve this problem, we deploy a very intuitive method that choosing the location where is sufficiently proximate to current Korean restaurants. The gathering of a certain type of cuisine in can convincingly indicate the taste of neighbourhood's residences, vice versa, this neighbourhood may be famous for such gathering of this certain type of cuisine.

Data :

To tackle this problem, I will deploy:

- List of neighbourhoods in New York, China
- Latitude and Longitude of these district.
- Venue data related to Asian restaurants, especially for Korean and Japanese cuisines.

Data Acquisition :

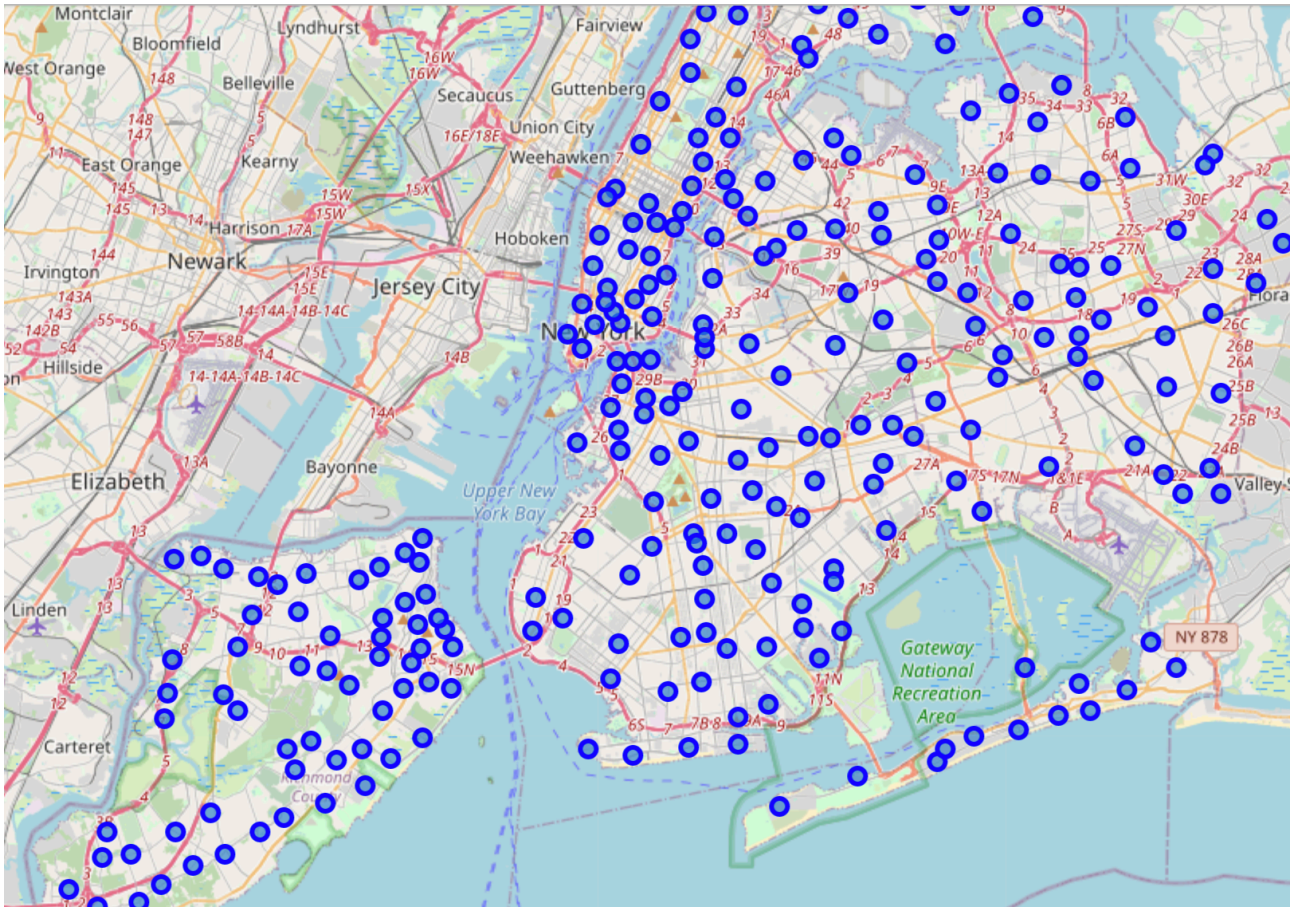
- Scrapping of New York neighbourhoods and postal codes via Wikipedia
- Getting Latitude and Longitude data of these neighbourhoods via Geocode package
- Using Foursquare API to get venue data related to these neighbourhoods

Methodology :

First, I need to get the list of neighbourhoods in New York, US. This is possible by extracting the list of neighbourhoods from wikipedia page
I did the web scraping by utilising pandas html table scraping method as it is easier and more convenient to pull tabular data directly from a web page into data-frame.

However, it is only a list of neighbourhood names and postal codes. I will need to get their coordinates to utilise Foursquare to pull the list of venues near these neighbourhoods. To get the coordinates, I tried using Geocodes package but it was not working so I used the csv file provided by IBM team to match the coordinates of New York neighbourhoods. After gathering all these coordinates, I visualised the map of New York using Folium package to verify whether these are correct coordinates.

Next, I use Foursquare API to pull the list of top 100 venues within 500 meters radius. I have created a Foursquare developer account in order to obtain account ID and API key to pull the data. From Foursquare, I am able to pull the names, categories, latitude and longitude off the venues. With this data, I can also check how many unique categories that I can get from these venues. Then,



I analyse each neighbourhood by grouping the rows by neighbourhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later.

Here, I made a justification to specifically look for “Korean restaurants”.

Lastly, I performed the clustering method by using k-means clustering. K-means clustering algorithm identifies k number of centeriods, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighbourhoods in New York into 3 clusters based on their frequency of occurrence for “Korean food”. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

Results :

Clusters

The results from k-means clustering show that we can categorise New York neighbourhoods into 5 clusters based on how many Korean restaurants are in each neighbourhood:

- Cluster 0: Neighbourhoods with almost no Korean restaurants

```
#Cluster 0
to_merged.loc[to_merged['Cluster Labels'] == 0].head()
```

	Neighborhood	Korean Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Allerton	0.0	0	40.865788	-73.859319	Domenick's Pizzeria	40.865576	-73.858124	Pizza Place
170	Middle Village	0.0	0	40.716415	-73.881143	Juniper Valley Park	40.720281	-73.881258	Park
170	Middle Village	0.0	0	40.716415	-73.881143	Juniper Valley Park Playground South	40.718504	-73.882652	Playground
169	Melrose	0.0	0	40.819754	-73.909422	3rd Avenue & East 163rd Street	40.824073	-73.908714	Intersection
169	Melrose	0.0	0	40.819754	-73.909422	Linda's Pizzeria	40.823760	-73.908632	Pizza Place

- Cluster 1: Neighbourhoods with the most Korean restaurants

```
#Cluster 1
to_merged.loc[to_merged['Cluster Labels'] == 1].head()
```

	Neighborhood	Korean Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
184	Murray Hill	0.190083	1	40.764126	-73.812763	Cafe de Cupping	40.765261	-73.814368	Coffee Shop
184	Murray Hill	0.190083	1	40.764126	-73.812763	H Mart	40.763239	-73.809126	Supermarket
173	Midtown South	0.160920	1	40.748510	-73.988713	Xi'an Famous Foods	40.748165	-73.984003	Chinese Restaurant
184	Murray Hill	0.190083	1	40.748303	-73.978332	Masala King	40.747184	-73.981853	North Indian Restaurant
184	Murray Hill	0.190083	1	40.748303	-73.978332	Wolfgang's Steakhouse	40.746531	-73.981953	Steakhouse

- Cluster 2: Neighbourhoods with few Korean restaurants

```
#Cluster 2
to_merged.loc[to_merged['Cluster Labels'] == 2].head()
```

	Neighborhood	Korean Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
262	Sunnyside Gardens	0.03	2	40.745652	-73.918193	Ariyoshi Japanese Restaurant	40.743914	-73.922969	Japanese Restaurant
262	Sunnyside Gardens	0.03	2	40.745652	-73.918193	Los Verdes	40.742389	-73.917970	Burger Joint
262	Sunnyside Gardens	0.03	2	40.745652	-73.918193	Doma	40.744171	-73.923192	Korean Restaurant
262	Sunnyside Gardens	0.03	2	40.745652	-73.918193	Zio Luigi	40.743591	-73.921973	Italian Restaurant
262	Sunnyside Gardens	0.03	2	40.745652	-73.918193	Taco Rey De Oro	40.744273	-73.912947	Food Truck

- Cluster 3: Neighbourhoods with some Korean restaurants

```
#Cluster 3
to_merged.loc[to_merged['Cluster Labels'] == 3].head()
```

	Neighborhood	Korean Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
205	Park Slope	0.017857	3	40.672321	-73.97705	Bareburger	40.671900	-73.977620	Burger Joint
205	Park Slope	0.017857	3	40.672321	-73.97705	Un Posto Italiano	40.672068	-73.976964	Gourmet Shop
205	Park Slope	0.017857	3	40.672321	-73.97705	Sounds	40.672450	-73.976784	Accessories Store
205	Park Slope	0.017857	3	40.672321	-73.97705	Blue Bottle Coffee	40.670600	-73.978458	Coffee Shop
205	Park Slope	0.017857	3	40.672321	-73.97705	Norman & Jules	40.672300	-73.977469	Toy / Game Store

- Cluster 4: Neighbourhoods with the second largest amount of Korean restaurants

```
#Cluster 4
to_merged.loc[to_merged['Cluster Labels'] == 4 ].head()
```

	Neighborhood	Korean Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
196	Oakland Gardens	0.125	4	40.745619	-73.75495	Imperial Taste Chinese Restaurant	40.749524	-73.755031	Chinese Restaurant
196	Oakland Gardens	0.125	4	40.745619	-73.75495	Red Arrow Card & Stationary	40.748499	-73.756339	Gift Shop
196	Oakland Gardens	0.125	4	40.745619	-73.75495	SUBWAY / Nathans / Bubble Tea	40.748683	-73.756058	Sandwich Place
196	Oakland Gardens	0.125	4	40.745619	-73.75495	Key Food	40.747926	-73.756323	Supermarket
196	Oakland Gardens	0.125	4	40.745619	-73.75495	Dunkin'	40.747459	-73.756093	Donut Shop

Recommendations :

Most of Korean restaurants are in Cluster 1 which is around Murray Hill, Midtown South. Also, there are good opportunities to open near Oakland Gardens. Looking at nearby venues, it seems Cluster 3 might be a good location as there are not a lot of Asian restaurants in these areas. Therefore, this project recommends the entrepreneur to open an authentic Burmese restaurant in these locations with little to no competition.

Limitations and Suggestions for Future Research :

In this project, I only take into consideration of one factor: the occurrence / existence of Korean Restaurants in each neighbourhood. There are many factors that can be taken into consideration such as population density, income of residents, rent that could influence the decision to open a new restaurant. However, to put all these data into this project is not possible to do within a short time frame for this capstone project. Future research can take into consideration of these factors.

Conclusion :

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing the machine learning by utilising k-means clustering and providing recommendation to the stakeholder.

References :

Foursquare Developer Documentation: https://developer.foursquare.com/docs

New York Venues: https://cocl.us/new_york_dataset