IBM Applied Data Science Capstone

Coursera Capstone

<u>Opening a New Restaurant in Manhattan, NYC</u>

By: Boyin ZHU, April 2020

Introduction

Manhattan has attracted lots of famous restaurants all over the world. People can enjoy everything that they can imagine whether it is domestic home cook or special foreign dishes. China Town, Korea street, Japanese street, Little Italy etc. become popular because most of the authentic restaurants gather there. However, some other top restaurants choose to be separated to distinguish themselves as noble. They might have their own parking areas, special decorations, might lie in the luxury shopping malls.

For the restaurant owner, they want to take advantage of the congesting effect which will bring a stable volume of customers, but they also want to avoid a fierce competition with other restaurants. Therefore, choosing the location of the restaurant is one of the most important decisions that will have great influence on whether the restaurant will be a success or a failure.

Business Problem

As for a restaurant owner, where should he open a new restaurant becomes a tough question. Choose the street that has all kinds of different foods or pick up a place where there are few competitions. Also, it is quite useful to other venues such as movie, theater, hotel and office that provide stable customers to have lunch or dinner.

The objective of this capstone project is to analyze and select the best locations in the Manhattan, NYC to open a new restaurant. Using data science analytical approach and machine learning techniques as clustering, this project aims to provide the solutions to answer the question : In the Manhattan, NYC, if a restaurant owner is looking for a place to open a new restaurant, where would you recommend to open it?

Data

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

1. List of neighborhoods in Manhattan, NYC. This defines the scope of this project, which is confined to the Manhattan, NYC.
2. Latitude and longitude coordinates of those neighborhoods. This is required in order to plot the map and to get the venue data.
3. Venue data, particularly data related to restaurant. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods to extract them

This Wikipedia page (https://en.wikipedia.org/wiki/List_of_Manhattan_neighborhoods) contains a list of neighborhoods in Manhattan, NYC, with a total of 25 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful-soup packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

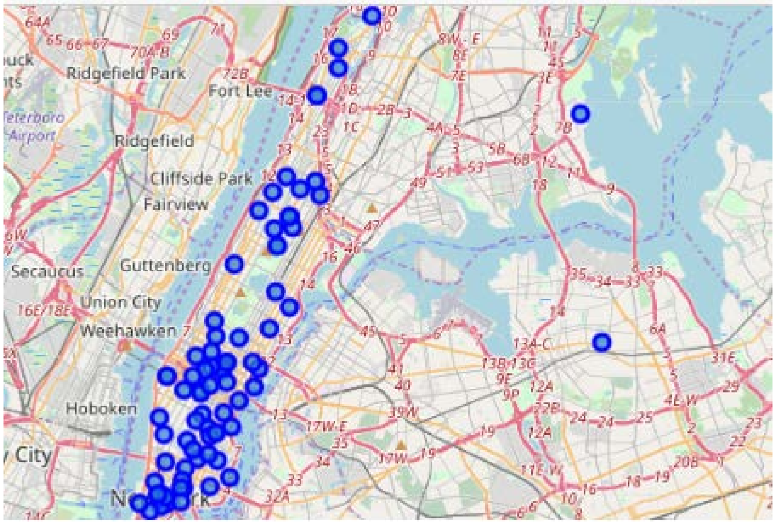
After that, we will use **Foursquare API** to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 125,000 developers. Foursquare API will provide many categories of the venue data, we are particularly interested in the Shopping Mall category in order to help us to solve the business problem put forward.

This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

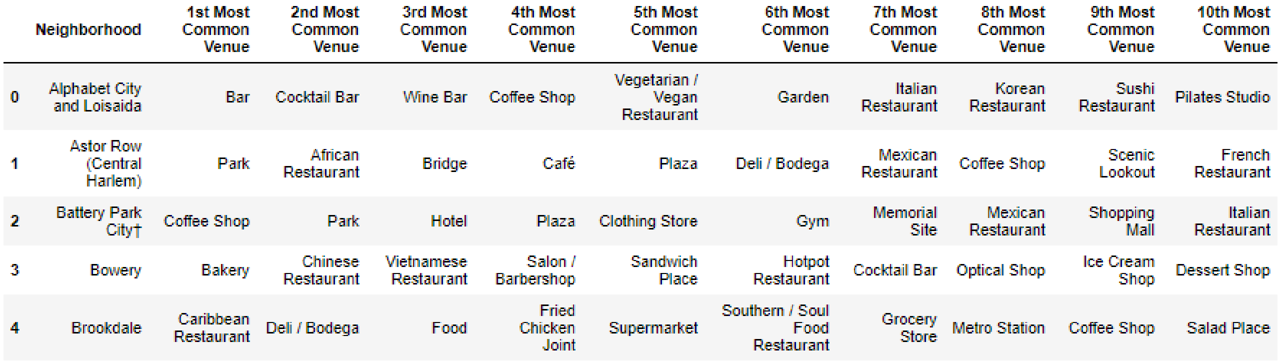
Methodology

After collecting data from Wikipedia , there are 85 Neighborhoods in mainly Manhattan New York City. With geocoder package , each neighborhood gets a latitude and longitude for further finding nearby venues.



Several points are noise in the dataset, it is partially because the name of the neighborhood cannot be found in the geocoder package. However, it will not do any harm to our further analysis since we are using unsupervised machine learning, clustering, to label the neighborhood.

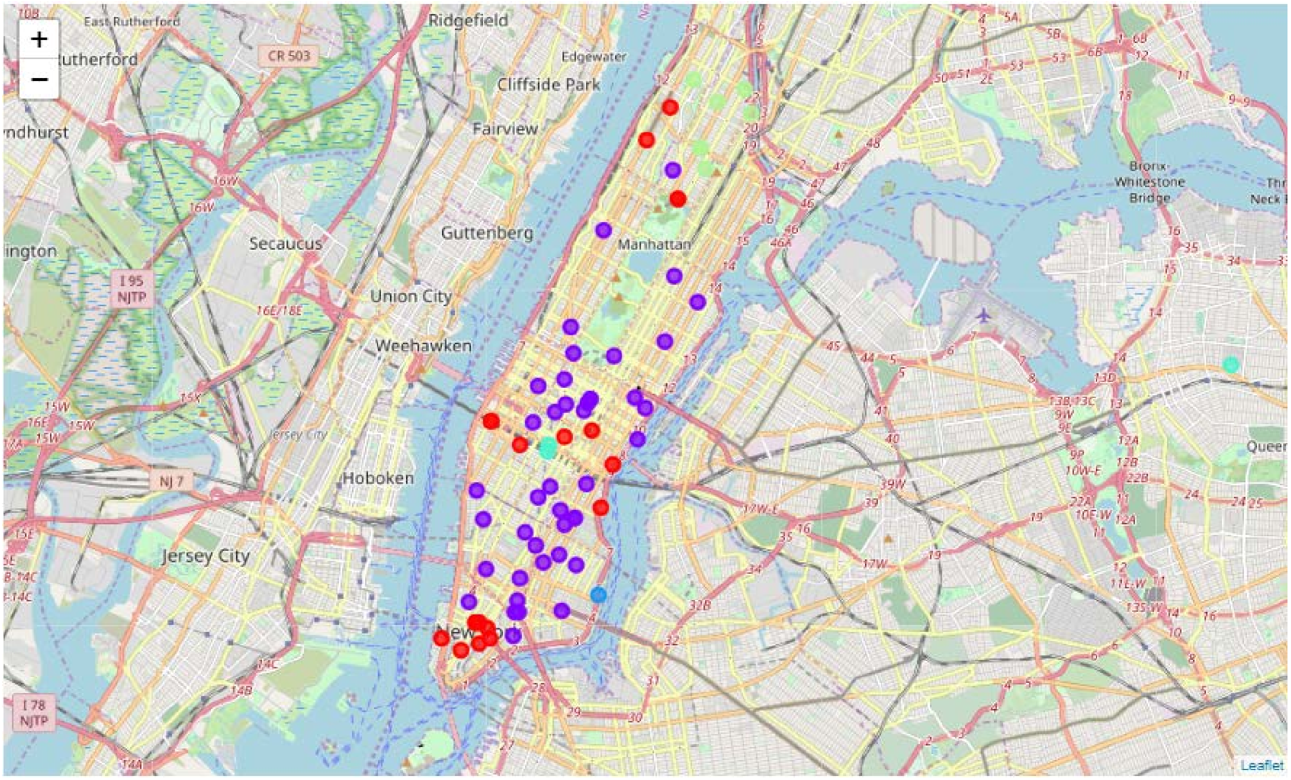
To solve the problem, how to define a cluster, we use the similarity of its nearby venues to calculate the distance for clustering. It will show the character of the neighborhood and imply whether it is good place to launch a new restaurant. The following dataset shows a sample of a top 10 venue category of each neighborhood.



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Alphabet City and Loisaida	Bar	Cocktail Bar	Wine Bar	Coffee Shop	Vegetarian / Vegan Restaurant	Garden	Italian Restaurant	Korean Restaurant	Sushi Restaurant	Pilates Studio
1	Astor Row (Central Harlem)	Park	African Restaurant	Bridge	Café	Plaza	Deli / Bodega	Mexican Restaurant	Coffee Shop	Scenic Lookout	French Restaurant
2	Battery Park City†	Coffee Shop	Park	Hotel	Plaza	Clothing Store	Gym	Memorial Site	Mexican Restaurant	Shopping Mall	Italian Restaurant
3	Bowery	Bakery	Chinese Restaurant	Vietnamese Restaurant	Salon / Barbershop	Sandwich Place	Hotpot Restaurant	Cocktail Bar	Optical Shop	Ice Cream Shop	Dessert Shop
4	Brookdale	Caribbean Restaurant	Deli / Bodega	Food	Fried Chicken Joint	Supermarket	Southern / Soul Food Restaurant	Grocery Store	Metro Station	Coffee Shop	Salad Place

Results

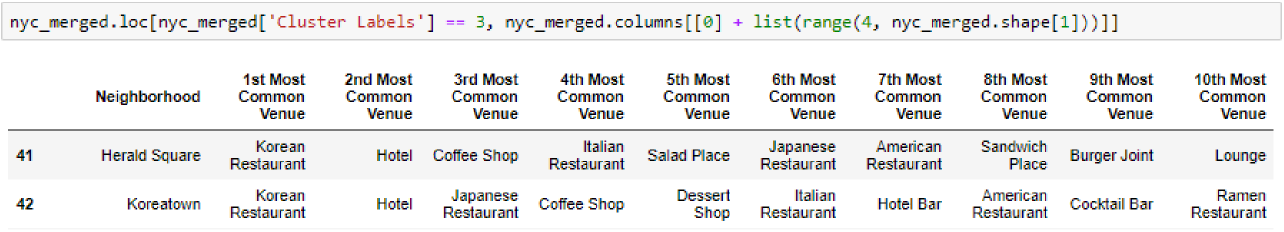
By applying K=5, with K-means clustering machine learning algorithm, we are able to cluster all the neighborhoods of Manhattan.



Discussion

It makes sense if two points are closer enough, they share a similar set of venues. This will lead them to be put in one cluster. Therefore, if two points are closer but in different cluster, we need make a deep dive to find the difference and figure it out the characteristics of different clusters.

There are 2 light blue points in the middle town and are very different from others, cluster 4. By taking a look the details as following, the 2 points are Korea town and there are plenty of Korean restaurants there, which make these 2 points unique.



```
nyc_merged.loc[nyc_merged['Cluster Labels'] == 3, nyc_merged.columns[[0] + list(range(4, nyc_merged.shape[1]))]]
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
41	Herald Square	Korean Restaurant	Hotel	Coffee Shop	Italian Restaurant	Salad Place	Japanese Restaurant	American Restaurant	Sandwich Place	Burger Joint	Lounge
42	Koreatown	Korean Restaurant	Hotel	Japanese Restaurant	Coffee Shop	Dessert Shop	Italian Restaurant	Hotel Bar	American Restaurant	Cocktail Bar	Ramen Restaurant

For cluster 1 (Red spots), there are mostly coffee shops and hotels, gathering in downtown mid-town and upper town west.
For cluster 2 (Purple spots), there are mostly different kinds of restaurants.
For cluster 5 (Light Green spots) there are mostly Grocery stores and a sparse of restaurants, gathering focusing on the upper areas.
For cluster 3,6, there is only one neighborhood in the cluster, it is unnecessary to take a deep dive.

Based on the characteristics of each cluster, there will be less competition to choose the cluster 1 (red spots) to launch a new restaurant in a busy neighborhood or to choose cluster 5 to launch one in a less commercial area, and will be more competition to launch restaurants in cluster 2 (Purple spots). However, it will be a good choice to open a new restaurant in cluster 4 (light blue spots) if the owner wants to open a Korean restaurant where will promise the customers' volume

Conclusion

This project collects the neighborhood dataset applying clustering machine learning algorithm to answer the questions: where will be a good place to launch a new restaurant?
By calculating the similarity of the nearby venues, we have 4 useful clusters telling the characteristics of each clusters. Based on these information cluster 1&5 are recommended for launching a new restaurant since there is less competition of restaurants nearby and there is a stable costumers' volume in that area.