Coursera Capstone

IBM Applied Data Science Capstone

Opening an Indian Restaurant in New York City, USA

By: Kunal Gupta

August 2020

Introduction

Delicious food is made all over the world and every region has its characteristics. From the hot and spicy Asian kitchen, through the exquisitely staged dishes of France to the barbeques of Northern America you can be ensured to never go hungry if you are open and willing to try new taste experiences.

Indian cuisine is very popular around the world. Hot curries with lots of chillies and a side of raita to cool down. Dishes are based on rice and often vegetarian or with seafood. Coriander, ginger, cumin, cardamom, saffron and nutmeg favoured flavour makers/spices.

As the popularity of different international cuisines is increasing worldwide and Indian diaspora being one of the largest among the world. Many international food chains do not want to miss this opportunity to set-up their restaurants. But setting up a restaurant requires many factors to be taken into consideration. Particularly, the location of the restaurant is one of the most important decisions that will determine whether the mall will be a success or failure.

Business Problem

The objective of this capstone project is to analyze and select the best locations in New York City, USA to open a Japanese cuisine restaurant. Using Data Science Methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the question: If a successful owner of multiple mid to high-end restaurants decided to open a new Indian restaurant in New York City, where would you recommend they open it?

Target Audience

This project is particularly useful for the owners of international food chains for Indian cuisine or other people who are looking to invest in setting up an Indian cuisine restaurant in New York City, USA.

Taking into account the price level at which the restaurant will operate, the intent is to find an optimal location in an area, where gastronomy is booming and which is easily accessible for Indians and local citizens as well.

Data

To perform this analysis, we will need the following data:

1. List of the Neighbourhoods of NYC
2. Geo-coordinates of the Neighbourhoods in NYC
3. Top venues in those Neighbourhoods

List of districts will be obtained from the following website

(https://yourneighborhood.co/neighborhood/list)

Geo-coordinates of districts will be obtained with the help of the geocoder tool in the notebook.

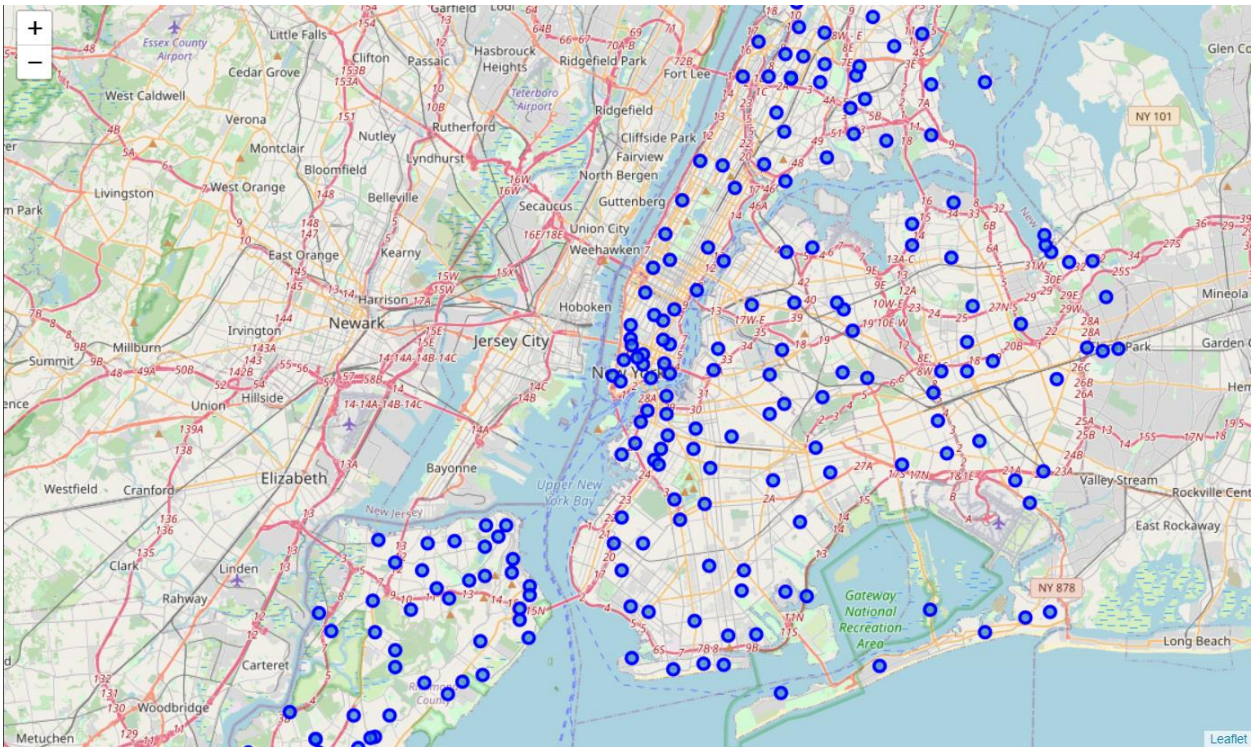
Top venues data will be obtained from Foursquare through an API.

Methodology

Firstly, we need to get the list of neighborhoods in the New York City. Fortunately, the list is available on this website yourneighborhood .

We will do web scrapping on this webpage using the Beautifulsoup package in python to extract the list of neighborhoods data. However, this is just a list of names. We will also need to get the geographical co-ordinates of these neighborhoods in the form of latitude and longitude in order to be able to use Foursquare API. For that we will use Geopy package in python that will allow us to convert address into geographical coordinates in the form of latitutde and longitude. After gathering the data, we will populate the data into pandas DataFrame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by geocoder are correctly plotted on the New York City.

Next, we will use the Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare developer account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighbourhood and examine how many unique categories can be curated from all the returned venues.



Now, In our first approach, Since we are analysing the “Indian Restaurant” data, we will filter the “Indian Restaurant” as venue category for the neighbourhoods.
Then we will perform clustering on the data by using the K-Means Clustering. K-Means clustering algorithm identifies the number of centroids, 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 most popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighbourhoods into 6 clusters based on their frequency of occurrence for “Indian Restaurant”. The results will allow us to identify which clusters have higher concentration of Indian Restaurants and which neighborhoods have fewer number of Indian Restaurants. Based of the occurrence of Indian Restaurants in

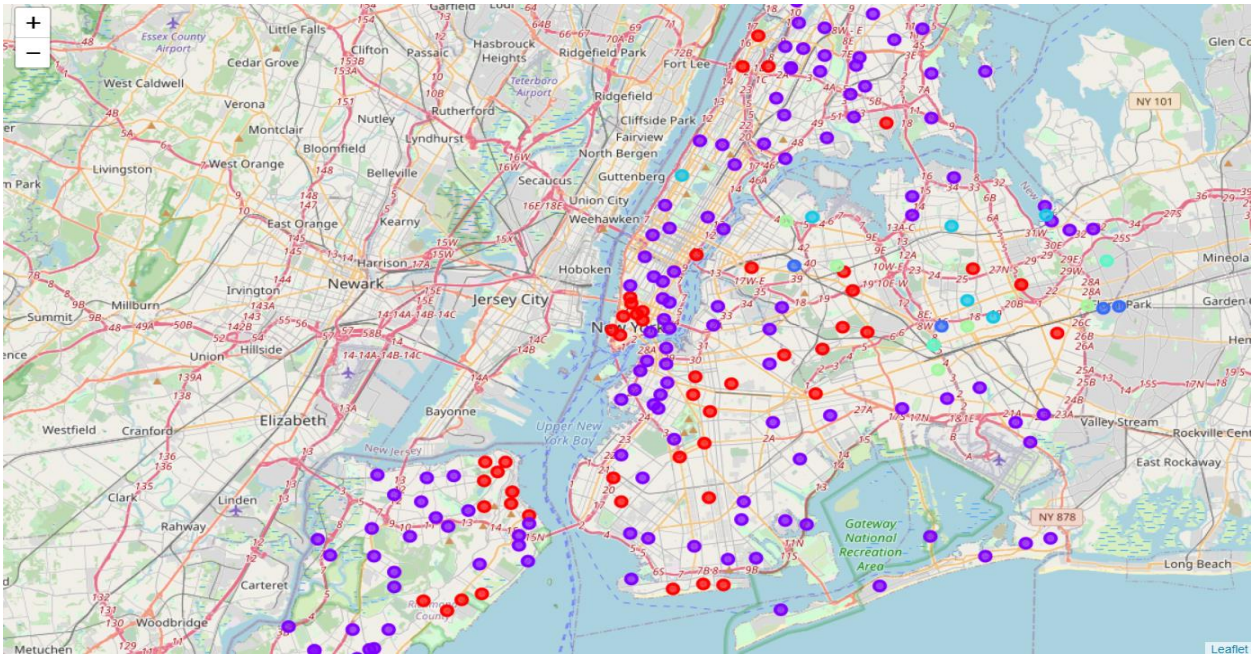
different neighborhoods, it will help us to answer the question as to which neighbourhoods are most suitable to open new Indian Restaurants.

In Second approach, we will take gastronomy of the people of a certain neighborhood in consideration. Firstly, we will find the top ten different venues in each neighborhood and then add them to a dataframe. Now, we will again use K-Means clustering algorithm to group these neighborhoods into 2 clusters based on the types of the top venues in those neighborhoods. This will result into two clusters and based on the types of top venues in these clusters we can tell about the gastronomy of the people of a certain neighborhoods and we can use these results to identify which neighborhoods are most suitable to open a new Indian Restaurant.

Results

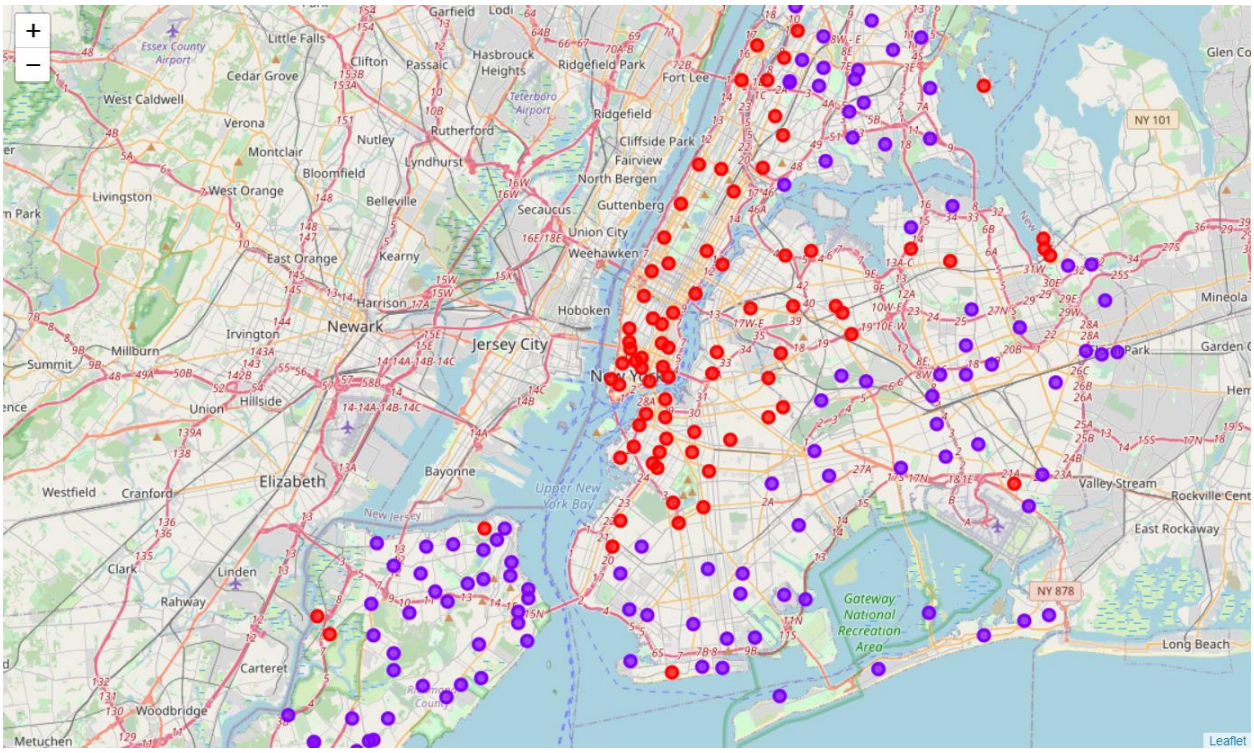
In first approach, the results from the k-means clustering show that we can categorize the neighborhoods into 6 clusters based on the frequency of occurrence of “Indian restaurant”:

- Cluster 0 and 3: These clusters have very less number of Indian restaurants.
- Cluster 1: This cluster have no Indian restuarants.
- Cluster 2 and 4: These clusters have moderate number of Indian restaurants.
- Cluster 5: This cluster have large number of Indian restaurants.



In second approach, the result of the k-means clustering algorithm groups the neighborhoods into 2 clusters based on the types top ten venues:

- Cluster 0: In this cluster, we can see that there are a lot of gastronomy related venues like coffee shop, pizza place, Thai restaurant, Mexican Restaurants, pub, etc.
- Cluster 1: In this cluster, gastronomy is not represented as pizza places and fast food are in top.



Discussion

In the first approach, as the obersvations noted from the map in result section,most of the Indian restaurants are located in the eastern and southern part of the New York City, with the highest number in cluster 5 and moderate number in cluster 2 and 4, very less number in cluster 0 and 3 with no Indian restaurant in cluster 0 which is the central area.

Cluster 5 might suffer from very high competition while there is moderate competition in cluster 2 and 4 and less competion in cluster 0 and 3. We are not so sure about the cluster 1 as there are no Indian restaurants which could either be a

good choice because of no competition or bad choice because of lack of gastronomy or popularity of Indian food among people.

In our second approach, as the obersvations noted from the map in result section, most of the gastronomy related venues are present in the northwestern area of the city which are part of cluster 0. While in cluster 1 there are venues which does not represent gastronomy which are present in all other parts of the city.

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 machine learning by clustering the data into a number of clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. Indian restaurant owners and International food chains that serve Indian cuisine and want to expand their business in New York city.

From first approach, we can conclude that Cluster 0 will be great to start an Indian restaurant because of less competition.

While from second approach we can say that, Cluster 0 of the second approach will be the best option to start a highend Indian resaurant.

Finally we can choose from the intersection of both clusters as they satisfy both approaches.