Best Place to open an Indian Restaurant in New York

Introduction

A person wants to open an Indian restaurant in New York. The idea behind this project is that there may not be enough Indian restaurants in New York and it might present a great opportunity for this person who is based in New York. As Indian food is very similar to other Asian cuisines, this person is thinking of opening this restaurant in locations where Asian food is popular. With the purpose in mind, finding the location to open such a restaurant is one of the most important decisions for this person and I am creating a model to help the person in finding the best location to open the restaurant in New York.

Business Problem

In New York, if someone wants to open an Indian restaurant, where should they consider opening it?

Target Audience

The person who wants to find the location to open authentic Indian restaurant in New York

Data

For model designing following data will be required:

- List of neighborhoods in New York.

- Latitude and Longitude of these neighborhoods.
- Venue data related to Indian and Asian restaurants. This will help us find the best places that are most suitable to open a Indian restaurant.

Following data sources will be needed to extract/generate the required information:

- The information regarding the neighborhoods in New York and their corresponding latitude and longitude coordinates can be downloaded from the following link "**https://geo.nyu.edu/catalog/nyu_2451_34572**".
- Use **geopy** library to get the latitude and longitude values of New York City.
- number of restaurants and their type and location in every neighborhood will be obtained using **Foursquare API.**

After loading the data from "**https://geo.nyu.edu/catalog/nyu_2451_34572**" we came to know that all the relevant data is in the features key, which is basically a list of the neighborhoods.So, let's define a new variable that includes this data.The next task is essentially transforming this data of nested Python dictionaries into a *pandas* dataframe.

Foursquare

Now that we have our location candidates, let's use the Foursquare API to get info on restaurants in each neighborhood.

We're interested in venues in 'food' category, but only those that are proper restaurants - coffee shops, pizza places, bakeries etc. are not direct competitors so we don't care about those. So we will include in outlist only venues that have 'restaurant' in category name, and we'll make sure to detect and include all the subcategories of specific Indian restaurant' category, as we need info on Indian restaurants in the neighborhood.

Methodology

First, I need to get the list of neighborhoods in New York. Which can be downloaded from the following link "**https://geo.nyu.edu/catalog/nyu_2451_34572**".

I have created the dataframe of the data we downloaded earlier which contains Borough,Neighborhood and their respective latitude and longitude.Then I slice the original dataframe and create a new dataframe of the Manhattan data.

Next, I use use geopy library to get the latitude and longitude values of Manhattan.In order to define an instance of the geocoder, we need to define a user_agent. We will name our agent, as shown below.

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 of the venues. With this data, I can also check how many unique categories that I can get from these venues.Then I slice the dataframe and create a new dataframe consisting data related to Indian and Asian restaurant.

After creating and gathering all the coordinates, I visualized the map of New York using Folium package to verify correct coordinates of the restaurants.

Next I again use the Foursquare API to get the rating of all the Indain and Asian restaurants in the New York.This is to prepare clustering to be done later.

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 most popular unsupervised machine learning algorithms and it is highly suited for this project as well. I have clustered the neighborhoods in New York into 3 clusters based on their rating. Based on the results (the concentration of clusters), I will be able to recommend the ideal location to open the restaurant.

Results

Clusters
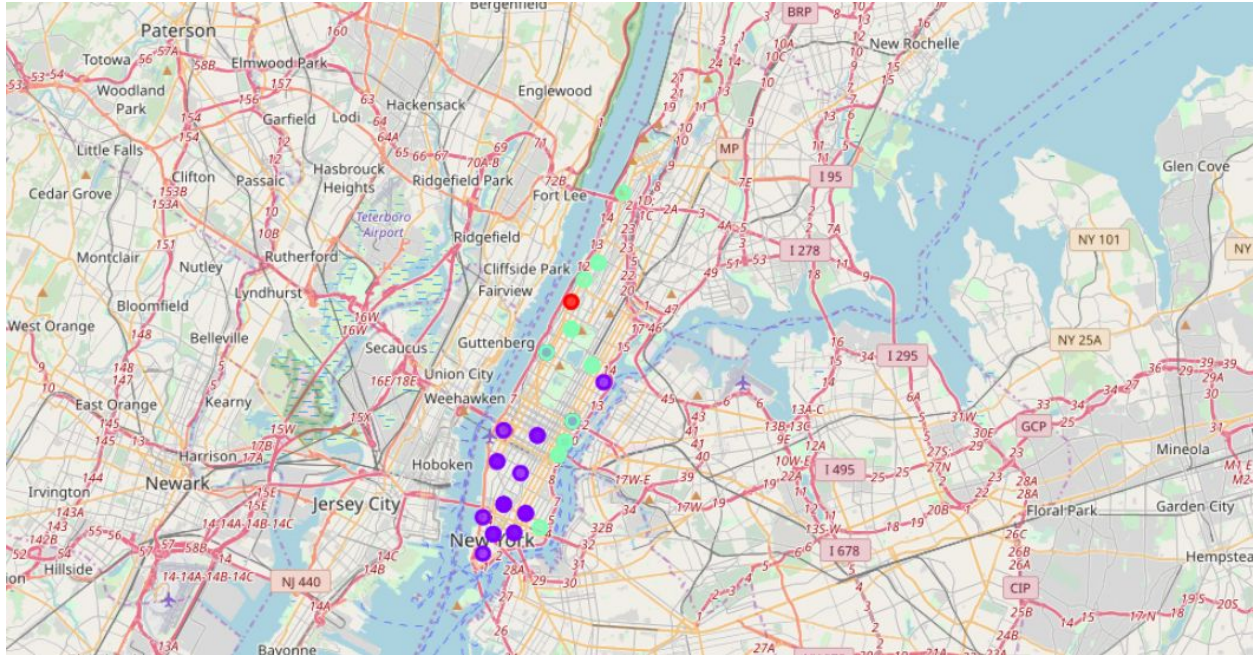
The results from k-means clustering show that we can categorize New York neighborhoods into 3 clusters based on how many Indian and Asain restaurants are in each neighborhood:

- Cluster 0: Restaurants having rating less than 7
- Cluster 1: Restaurants having rating between 7 and 8
- Cluster 2:Restaurants having rating greater than 8

The results are visualized in the above map with Cluster 0 in red color, Cluster 1 in light green color and Cluster 2 in blue color.

Recommendations

Most of Indian and Asian restaurants are in Cluster 0 and 1 which is around Manhattan Valley area have lowest rating on foursquare. So, there is a good opportunity to open near this area as the competition seems to be low.Therefore, this project recommends to open an authentic Indian restaurant in these locations with little to no competition.

Nonetheless, if the food is authentic, affordable and good taste, I am confident that it will have great following everywhere.

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 utilizing k-means clustering and providing a recommendation to the stakeholder.