COURSERA CAPSTONE PROJECT

Exploring Japanese Restaurants in NYC

El Mehdi Selbi

INDEX

INTRODUCTION:

The idea behind developing this project was to understand the spread of different cuisines across NYC and evaluate it to solve a business problem. The business problem in this case is figuring out the right location in NYC for a client so that he’s able to open a Japanese restaurant at a place in the city that would yield him maximum profits by attracting more customers.

To solve this problem, it is important to understand the neighborhoods in the city and to figure out the distribution of variety of cuisines in such neighborhoods. Analyzing and understanding this will provide a lot of insight on the distribution of restaurants as per the cuisines and will thus prove to be very helpful for upcoming restaurants.

DATA:

The data used here contains information about the neighborhoods and boroughs present in NYC and is obtained from the following link, https://geo.nyu.edu/catalog/nyu_2451_34572, which is the NYU Spatial Data Repository. The data present here is in json format and it is converted into a data-frame with 5 boroughs and 306 neighborhoods.

Apart from this, the cuisine data for NYC was obtained from Wikipedia from Wikipedia, https://en.wikipedia.org/wiki/Cuisine_of_New_York_City. Here, too, the data was in json format and was converted into a data-frame. The data was further cleaned to obtain a data-frame with just names of the major towns like Manhattan, Bronx, Brooklyn and Staten Island and the cuisines in each one of them.

The cuisine data was used to visualize the distribution of cuisines with the help of a word-cloud and count of cuisines per borough. Later, this result is used to look at the distribution of restaurants and different public spots in one city.

METHODOLOGY:

There are two steps that are implemented as a part of the entire methodology for the project

Step1: Formation of word-clouds as per the cuisines

The initial data contained information about boroughs, street, zip codes, etc. So, a new data-frame was created with just the columns containing boroughs and the cuisine description. The data frame was then split into the cuisines in Manhattan, Brooklyn, Bronx and Staten Island.

Later, each of these locations is analyzed to figure out the distribution of cuisines and then a word cloud is visualized.

Manhattan Word cloud:

The Manhattan word cloud shows that widely consumed cuisine type across the city is American, as the font size of that cuisine is the biggest and we can see the Chinese and Japanese foods are also appear to be popular in this borough.

Brooklyn Word cloud:

The Brooklyn word cloud shows somewhat amount of equal distribution of American and Chinese cuisines. However, there seems to be more popularity of American cuisine here as well and no appearance for the Japanese food.

Bronx Word cloud:

The Bronx word cloud shows no options for Japanese food and we can see that Chinese food is also popular in this area and we can automatically say that it’s not a good idea to risk opening a Japanese restaurant here.

Staten Island Word cloud:

In the Staten Island word cloud again, the distribution of American cuisines seems to be the most and we can say the Japanese cuisine has its place also compared to the other borough we explored.

it's hard to decide from these word clouds the best borough to work on or identify as the best one we should focus to open the Japanese restaurant that's why were got the number of Japanese cuisine in each borough.

Number of Japanese cuisines per borough:

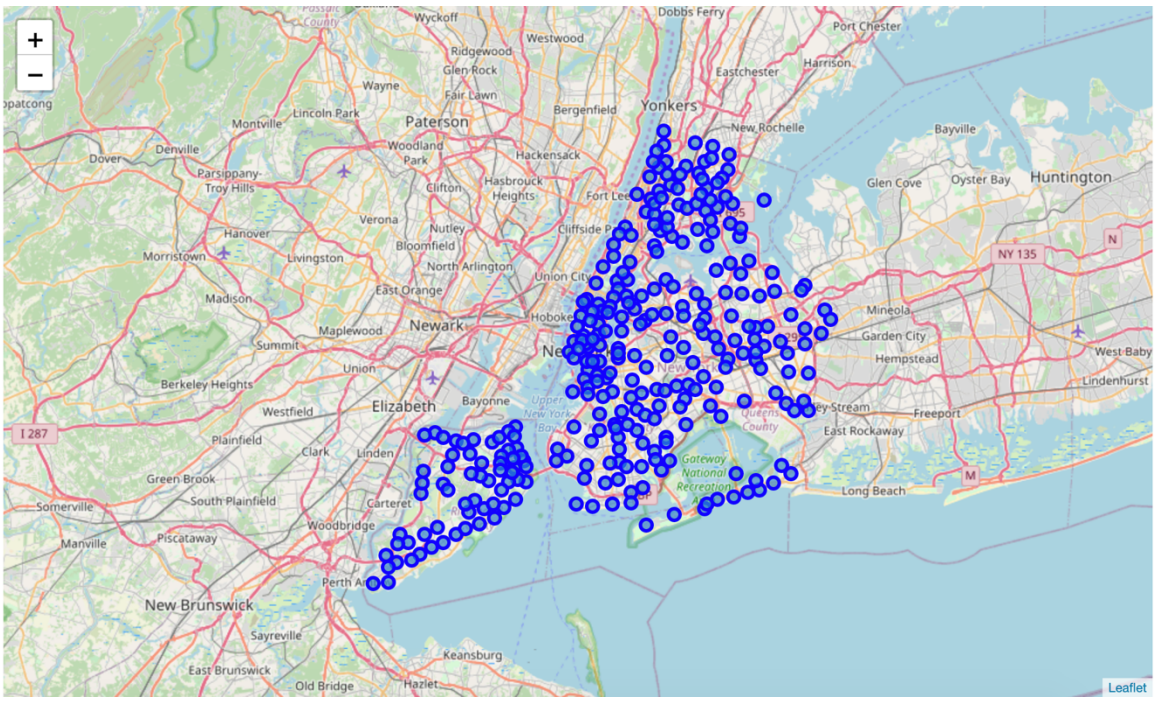
Manhattan	3603
Brooklyn	1405
Queens	1185
Staten Island	291
Bronx	153

It is clear now that we should work on Manhattan from world cloud and also in this ranking because it contains an important number of Japanese cuisines compared to the other boroughs.

Step2: Exploring New York City data for K-means clustering

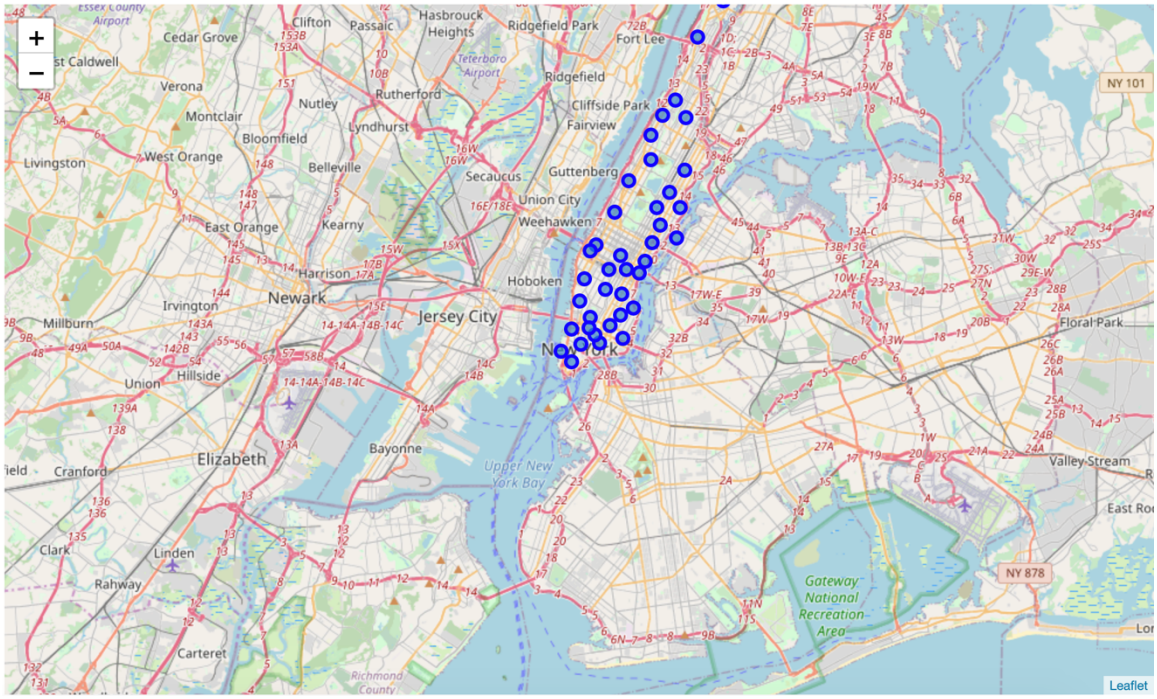
In this step, the json data of New York City was downloaded and was converted into a data-frame. For this, the json data is analyzed to find out the features of the neighborhoods and boroughs and an empty data-frame was created.

In the next step, the neighborhoods and boroughs in New York City were visualized on a map using Folium as shown below



From the exploration obtained in the first step, Manhattan has the greatest number of Japanese restaurants and thus doing the analysis further only on Manhattan makes sense.

Later, a map of Manhattan with the neighborhoods is plotted.



In the next step, Foursquare API was utilized to obtain the list of venues in these neighborhoods. This returns data in json format.

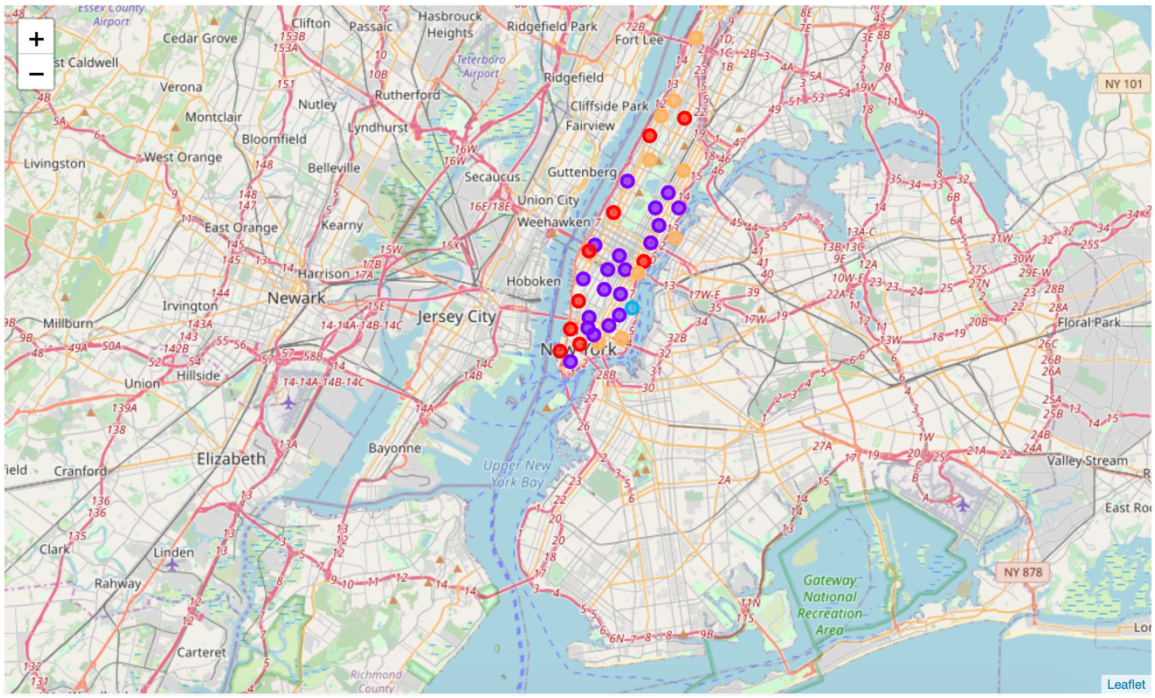
This data was then utilized to create a separate data-frame for Manhattan with latitude and longitude values of each neighborhood and with the name of the neighborhood and the category to which it belongs.

This data-frame was then combined with an output of a function which gave the list of nearby venues. This is converted into a data-frame and it is combined with data-frame of the category of the neighborhood.

Later, one-hot encoding is employed so that different types of venues are listed as columns and the rows represent the neighborhoods. This is then used to find the top 5 venues in each neighborhood.

Later, a data-frame is created with the columns representing the top 10 venues in each neighborhood, which are represented as rows.

In the last step, K-means clustering is employed with K=5 to find clusters with similar venues and return to a data-frame of those neighborhoods.



RESULTS:

As per the results obtained from the Jupyter notebook, different colors are given to each cluster. The majority of venues are common in cluster 1 (purple), and it mostly consists of food places and restaurants. Cluster 4 contains the second most (orange) number of places and it usually consists of public parks.

DISCUSSION:

Most of the restaurants could be found in cluster 1 and cluster 4. However, in cluster 4, the Japanese restaurants were the 5th most common venue. In cluster 1 however, in Midtown South and Flatiron neighborhoods, Japanese Restaurants were the 3rd most common venue, and this tells us that Japanese food is prominent in this area. Thus, opening a restaurant in these neighborhoods would be a profitable choice

CONCLUSION:

To solve this problem, it is important to understand the neighborhoods in the city and to figure out the distribution of variety of cuisines in such neighborhoods. Analyzing and understanding this will provide a lot of insight on the distribution of restaurants as per the cuisines and will thus prove to be very helpful for upcoming restaurants.