Exploring Chinese Restaurants in NYC

IBM Data Science Capstone Project

- Simaant Patil

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 Chinese 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 neighbourhoods in the city and to figure out the distribution of variety of cuisines in such neighbourhoods. Analysing 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 neighbourhoods 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 neighbourhoods.

Apart from this, the cuisine data for NYC was obtained from Wikipedia from the following website, 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. Later, this result is used to look at the distribution of restaurants and different public spots in one city.

Methodology:

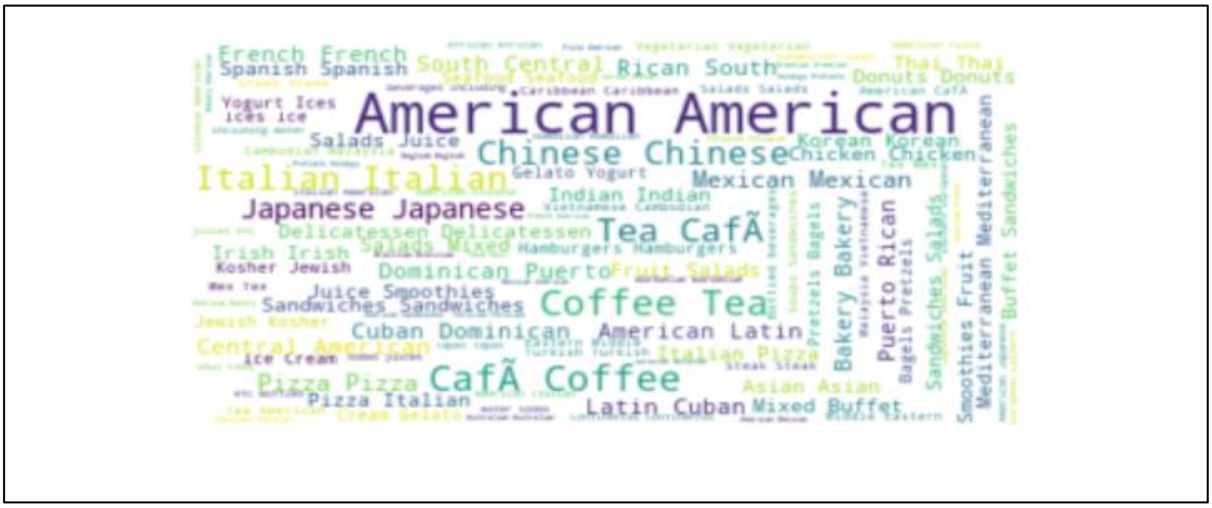
There are two steps that are implemented as a part of the entire methodology for the project

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.

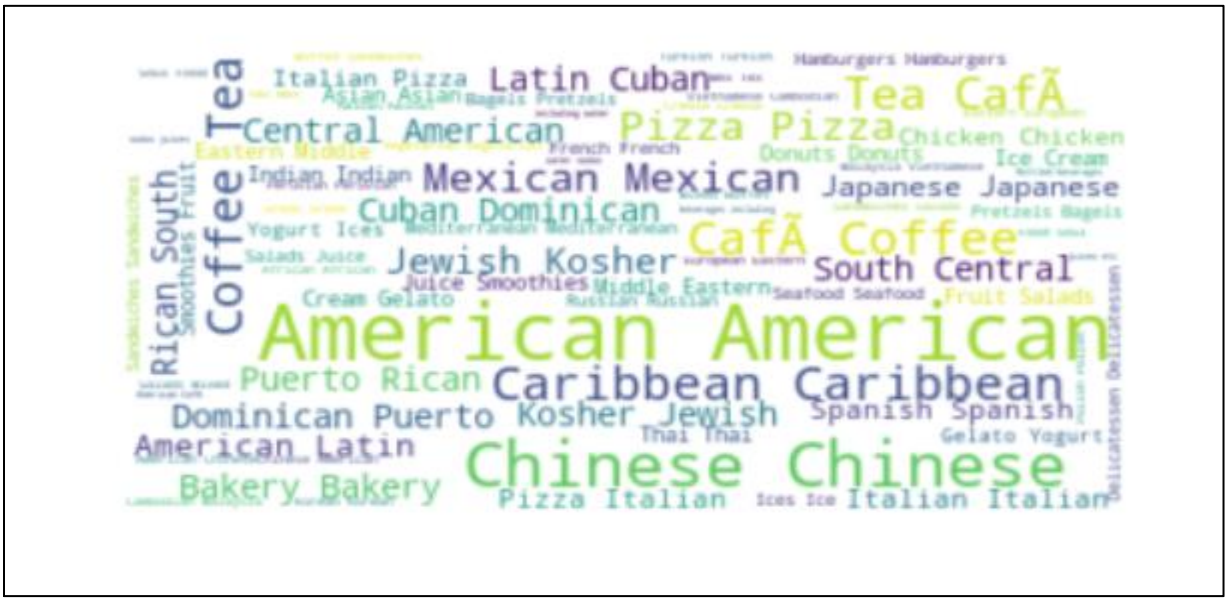
Later, each of these locations is analysed to figure out the distribution of cuisines and then a word cloud is visualized.

Manhattan Word cloud:



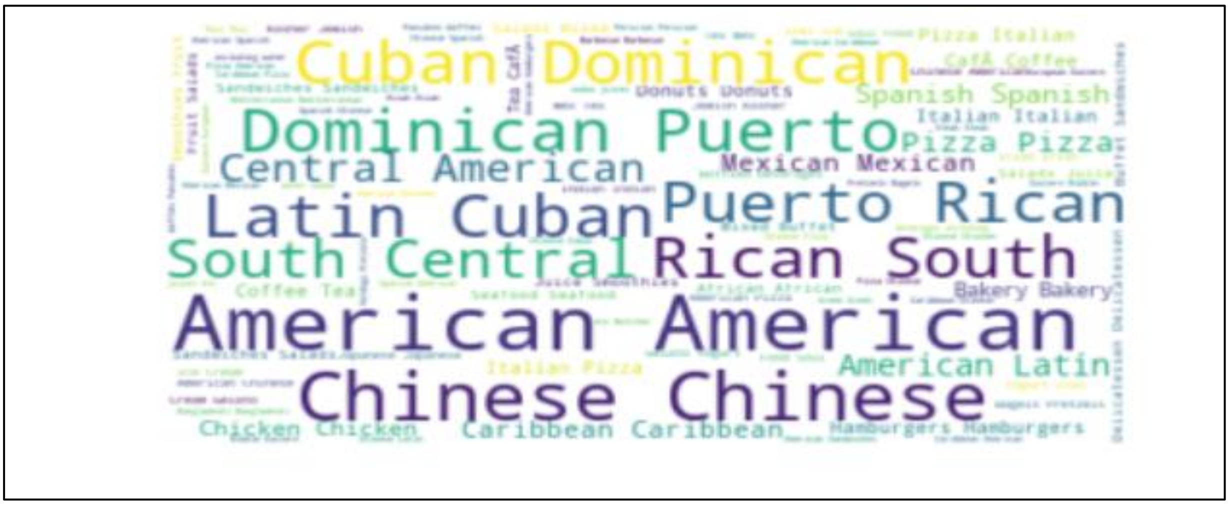
The Manhattan word cloud shows that widely consumed cuisine type across the city is American., as the font size of that cuisine is the biggest.

Brooklyn Word cloud:



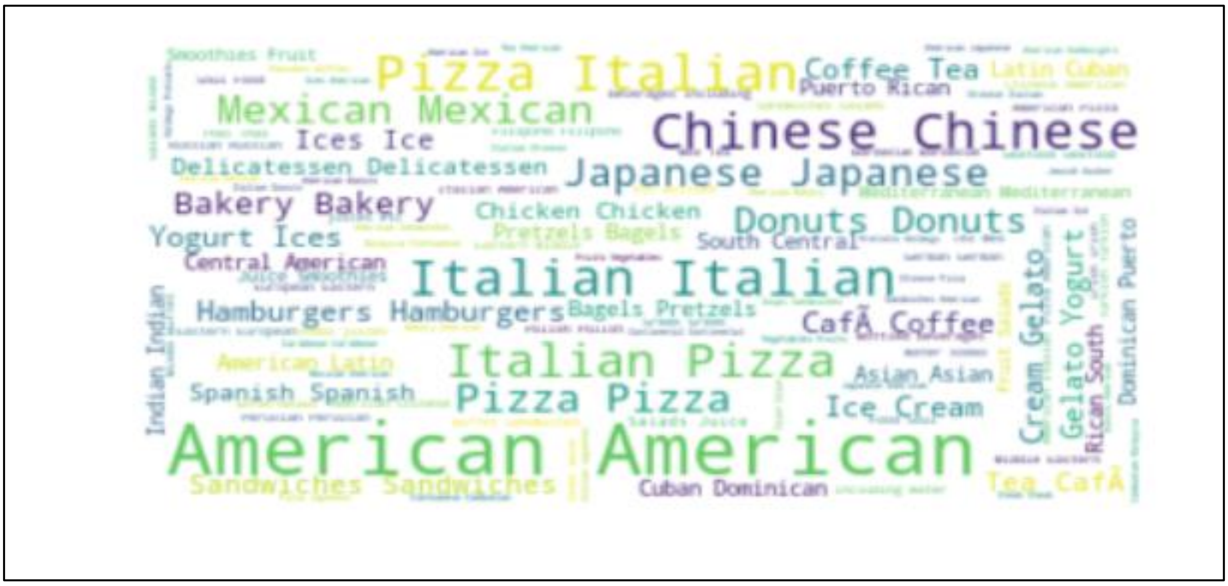
The Brooklyn word cloud shows somewhat amount of equal distribution of American and Chinese cuisines. However, there seems to be more popularity of American cuisine here as well.

Bronx Word cloud:



The Bronx word cloud shows a greater distribution of Chinese cuisine as compared to the Brooklyn cuisine and thus, it will be a good option to consider Bronx to open a Chinese restaurant.

Staten Island Word cloud:

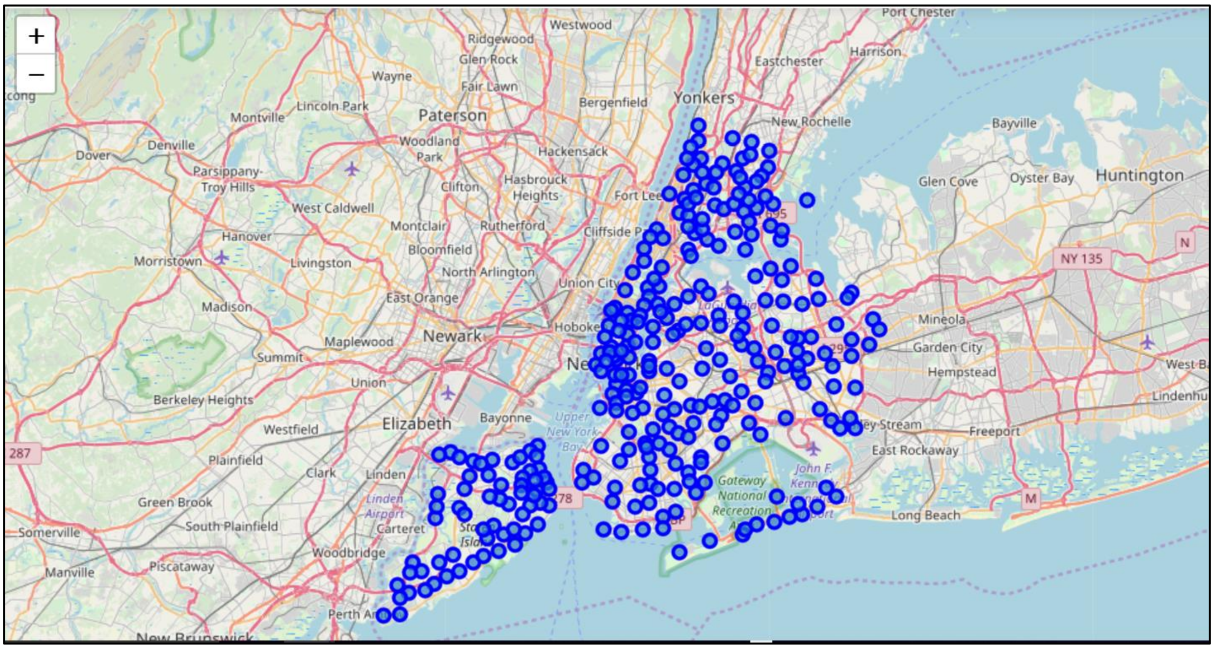


In the Staten Island word cloud again, the distribution of American cuisines seems to be the most.

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

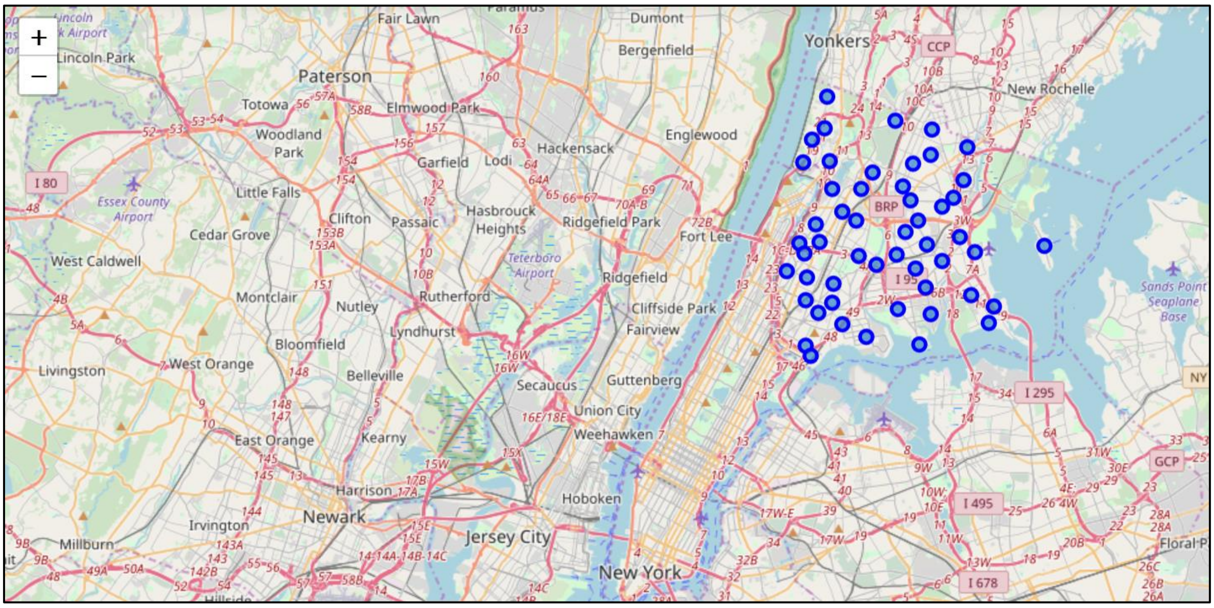
In this step, the json data of New York City was downloaded and was converted into a data-frame. For this, the json data is analysed to find out the features of the neighbourhoods and boroughs and an empty data-frame was created.

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



From the word clouds obtained above, Bronx has the greatest number of Chinese restaurants and thus doing the analysis further only on Bronx makes sense.

Later, a map of Bronx with the neighbourhoods is plotted.



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

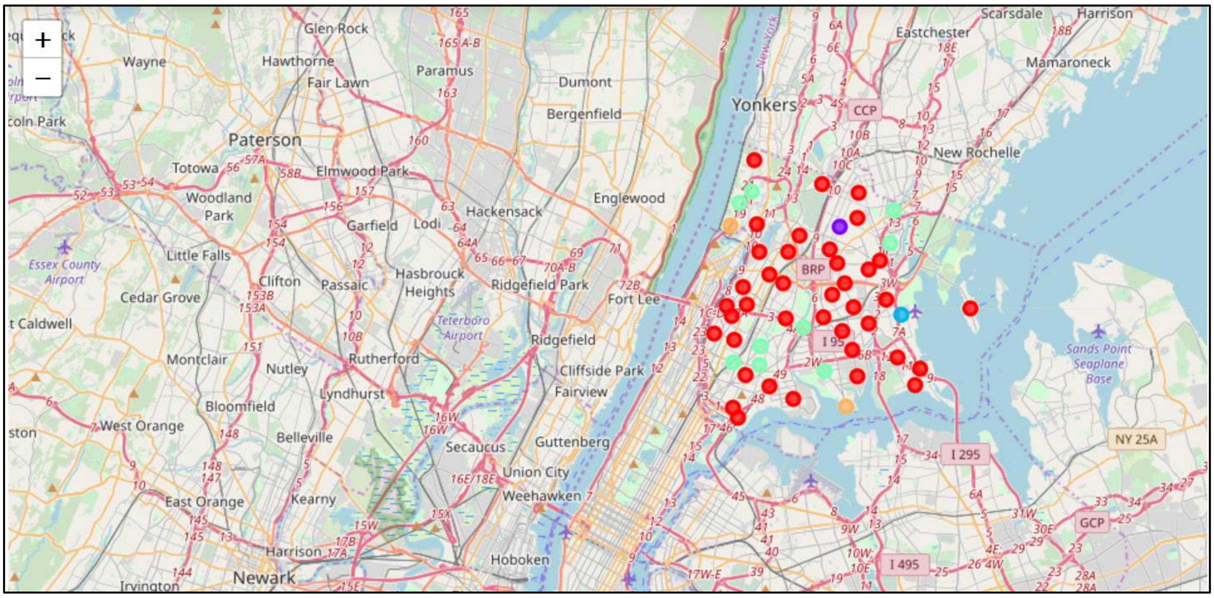
This data was then utilized to create a separate data-frame for Bronx with latitude and longitude values of each neighbourhood and with the name of the neighbourhood and the category to which it belongs.

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 neighbourhood.

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

Later, a data-frame is created with the columns representing the top 10 venues in each neighbourhood, 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 neighbourhoods.



Results:

As per the results obtained from the Jupyter notebook, different colours are given to each cluster. Majority of venues are common in cluster 0 (red), and it mostly consists of food places and restaurants. Cluster 3 contains the second most (green) number of places and it usually consists of public parks.

Discussion:

Most of the restaurants could be found in cluster 0 and cluster 3. However, in cluster 0, there weren't any Chinese restaurants which the first most common venue was. In cluster 3 however, in Soundview neighbourhood, Chinese Restaurant is the most common venue, and this tells us that Chinese food is prominent in this area. Thus, opening a restaurant in this neighbourhood would be a profitable choice

Conclusion:

As per the analysis, clustering the neighbourhoods in Bronx made it easy to determine type of venues in each neighbourhood, which led to accurately determining the venue for opening a Chinese restaurant.