Applied Data Science Capstone Project
The Battle of Neighborhoods

Contents

Finding an Ideal Location for a Vegan Restaurant in New York City

1 Introduction/Business Problem

The clients, a couple of vegan chefs, are looking for an ideal location to open a vegan restaurant in New York City, particularly in the boroughs of Manhattan and Brooklyn. There are various factors that need to be considered when choosing a location for a restaurant –

- Are there other restaurants in the area?
- Are there other similar restaurants in the area?
- Are there already several similar restaurants?
- Are there other businesses and facilities in the area that would complement this restaurant?

In this case, we will look for neighborhoods where there are a lot of restaurants, and maybe a few of them are vegan or vegetarian restaurants. In addition, we will look for neighborhoods that have other businesses and facilities that would indicate that a vegan restaurant would be a good fit for that area. Some of these are health food stores, gyms, yoga studios, farmer's markets and parks with jogging/hiking trails.

We will use New York City neighborhoods data and Foursquare location data to analyze the various neighborhoods and to identify an ideal location for a vegan restaurant.

2 Data Sources

2.1 Neighborhoods Data

In this section, I will discuss the data sources that will be used for this project. New York city has 306 neighborhoods spread out among 5 boroughs. This New York city neighborhoods data will be downloaded from https://geo.nyu.edu/catalog/nyu_2451_34572. This is a json file that contains neighborhood data in the features key that looks like this:

```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
```

```
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```


The neighborhood name, borough, latitude and longitude data will be extracted from this json file and loaded into a pandas dataframe that looks like this:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

2.2 Foursquare API

The Foursquare API provides location-based experiences with diverse information about venues, users, photos, and check-ins. We will use this API to get information about the venues in the various neighborhoods. The neighborhood coordinates from the neighborhoods dataframe will be used with the Foursquare API to analyze the neighborhoods. The response will be parsed and loaded in a dataframe that looks like this:

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
Marble Hill	40.876551	-73.91066	Rite Aid	40.875467	-73.908906	Pharmacy

This data will be analyzed to find neighborhoods in the boroughs of Manhattan and Brooklyn that will be suitable locations for the vegan restaurant.

3 Methodology

3.1 Data Acquisition

As discussed in the data sources section, the New York city neighborhoods data was downloaded from https://geo.nyu.edu/catalog/nyu_2451_34572 . This json file, newyork_data.json contains neighborhood data in the features key.

The other source of data is the location-based venues data obtained from the Foursquare API. This data was obtained using the explore endpoints calls in the Places API. The latitude and longitude data for the various neighborhoods in the boroughs of Manhattan and Brooklyn was used to get the venues information for those neighborhoods from Foursquare. Category IDs for health food store, vegetarian/vegan restaurant, farmers market, gym/fitness center, yoga studio, bike trail and trail to get data about the types of venues that we needed.

3.2 Data Wrangling

The neighborhood name, borough, latitude and longitude data was extracted from newyork_data.json file and loaded into the pandas data frame neighborhoods.

The results from the Get request to the Foursquare API is examined. The venues data is in the items key of the json format the results are in. The data in json format is cleaned up and loaded into the nearby_venues data frame. This is done for each neighborhood in the borough and the data is loaded into the data frame manhattan_venues for Manhattan and Brooklyn_venuues for Brooklyn.

3.3 Exploratory Data Analysis

The data frame was quickly examined to make sure that the data was loaded properly and data for all five boroughs and three hundred and six neighborhoods were present.

The venues data for all the neighborhoods in each borough is examined. The shape of the dataframes is examined.

3.4 Data Analysis/Machine Learning

The venues data for each neighborhood is analyzed by grouping the data for each neighborhood by taking the mean of the frequency of occurrence of each category of venue. The venues are then sorted in descending order of the mean. A new data frame is then created with the top 5 venues for each neighborhood. This data frame is called manhattan_venues_sorted for Manhattan and Brooklyn_venues_sorted for Brooklyn.

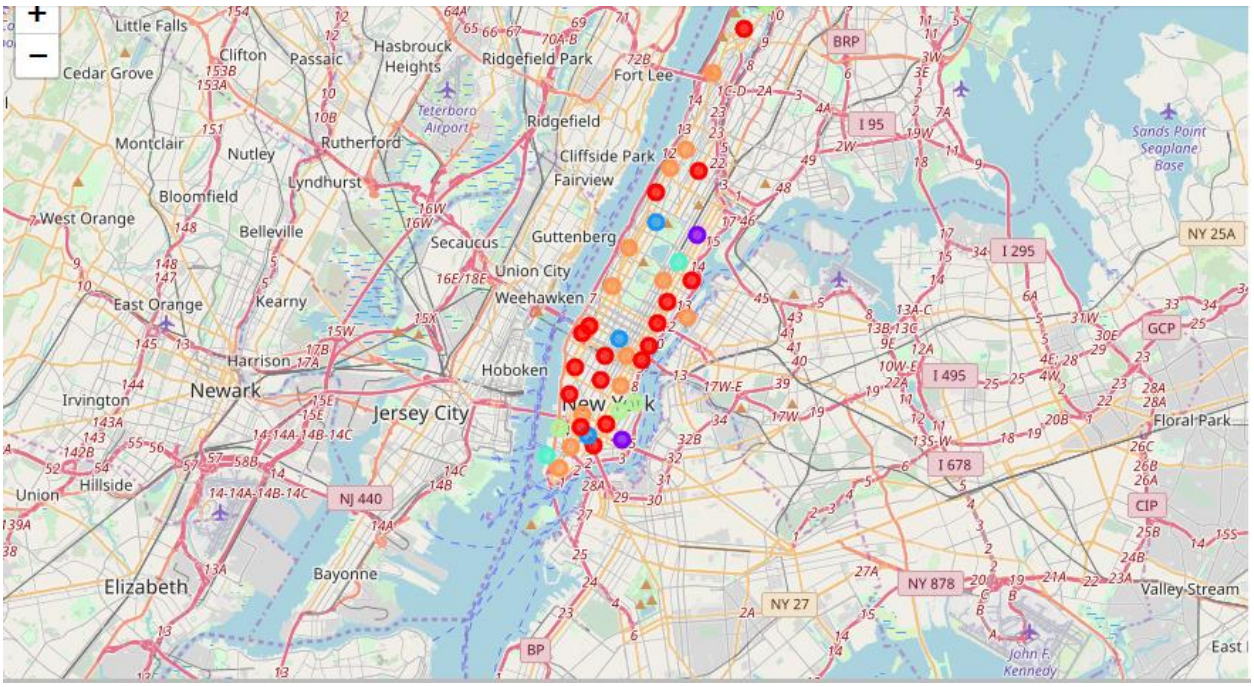
Then the neighborhoods in each borough are clustered using k-means clustering. The number of clusters, kclusters was set to 6. Method of initialization was set set to init = 'random'. The parameter n_init was set to 100.

Clustering was the machine learning technique that was used because the problem was to find locations that was suited for a vegan restaurant. By clustering the neighborhoods based on the top 5 venues in them, we can identify neighborhoods that have the types of venues that would also support a vegan restaurant.

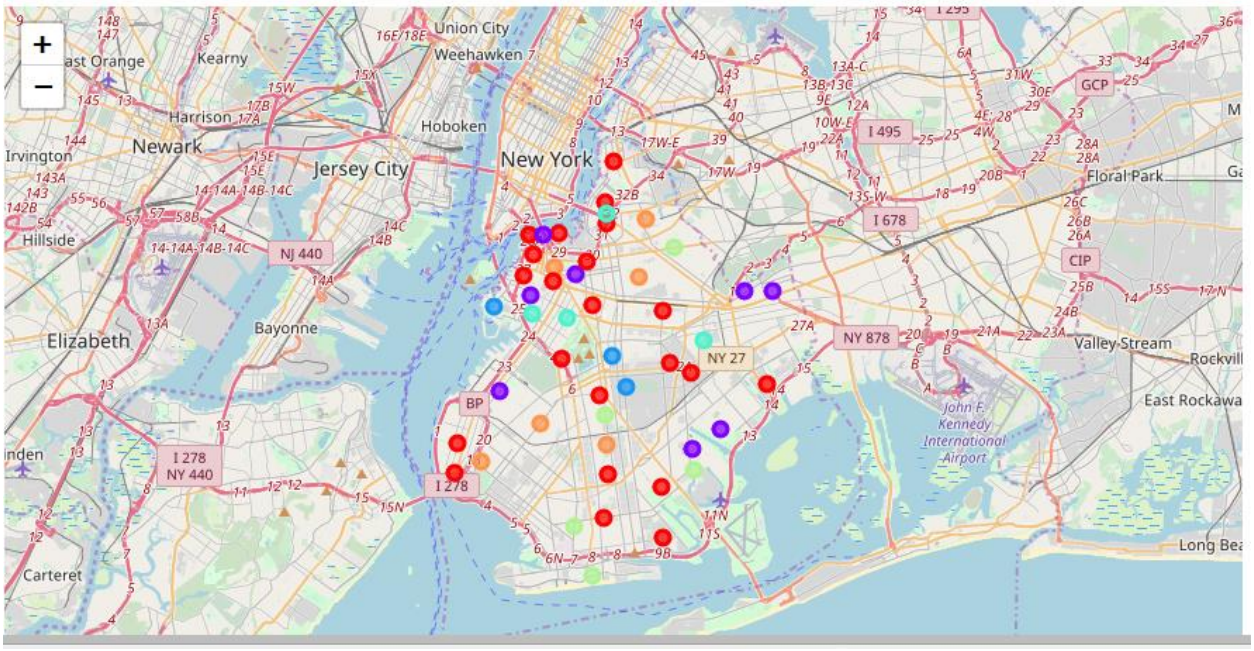
3.5 Data Visualization

The clusters for the boroughs are plotted on a map using folium map rendering library.

Clusters of Manhattan Neighborhoods.



Clusters of Brooklyn Neighborhoods



4 Results

The clusters for Manhattan and Brooklyn neighborhoods were examined. Cluster 0 and cluster 5 in the Manhattan clusters, have neighborhoods that have a good mix of the type of businesses and facilities that we were looking for. So, we can pick a neighborhood in these 2 clusters as location for the vegan restaurant in Manhattan. Similarly, cluster 0 and cluster 1 in the Brooklyn clusters, have neighborhoods that have a good mix of the type of businesses and facilities that we were looking for.

5 Discussion

At the start of the project, the assumption was made that by segmenting and clustering the neighborhoods based on the types of venues in those neighborhoods, we can select a neighborhood for the location of a vegan restaurant. After going through the analysis, we were able to identify a set of neighborhoods that would be suitable locations for a vegan restaurant. Additional demographic data for the neighborhoods would be useful in narrowing it down further.

6 Conclusion

The business problem that was addressed in this project was finding a location for a vegan restaurant in the boroughs of Manhattan and Brooklyn in New York city. The New York city neighborhoods data and foursquare location-based venues data were used to analyze the neighborhoods. K-means clustering was used to divide the neighborhoods into clusters. From these clusters, the clusters that had the neighborhoods that would be suitable for a vegan restaurant were selected.