Battle of the Neighborhoods: NYC Beach Neighborhoods

Introduction

Recently, there has been a massive influx of young people moving to and working in NYC, NY. We represent a large real estate firm for young professionals. A new client would like to move to a neighborhood in NYC that has access to beaches and yogo studios, as well as restaurants. Price range is not a limitation. The purpose of this project is to help this young professional determine which neighborhoods they should live in that has access to beaches and yogo studios and restaurants.

The Problem

The major purpose of this project is to provide a list of potential neighborhoods in NYC, NY for our client that matches her needs. We will use this data to suggest a neighborhood for her to move into and buy real estate.

Data

For this project we need the following data:

- New York City data that contains list Boroughs, Neighborhoods along with their latitude and longitude.
- Venues in each neighborhood of NYC, especially beaches, yoga studios, and restaurants
- GeoSpace data

The Location

New York City is a bustling metropolis that spans 5 boroughs in the State of New York. While it is known for it's nonstop nightlife, it also has many recreational studios, beaches, parks, and markets.

Foursquare API

This project would use Four-square API as its prime data gathering source as it has a database of millions of places, especially their places API which provides the ability to perform location search, location sharing and details about a business.

Methodology

We begin by collecting the New York city data from a json file from the IBM course at "https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json" We will find all venues for each neighborhood using FourSquare API. We will sort types of venues by neighborhood. We will use k-means clustering to visualize clusters of neighborhoods by venue category patterns. Based on the clustering patterns, we will make recommendations of neighborhoods to our client.

Clustering Approach

To make recommendations on where to live, we explored, segmented, and clustered the top types of venues within neighborhoods to make recommendations. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.

Libraries Used to Develop Project

Pandas: For creating and manipulating dataframes. Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map. Scikit Learn: For importing k-means clustering. JSON: Library to handle JSON files. XML: To separate data from presentation and XML stores data in plain text format. Geocoder: To retrieve Location Data. Beautiful Soup and Requests: To scrap and library to handle http requests. Matplotlib: Python Plotting Module.

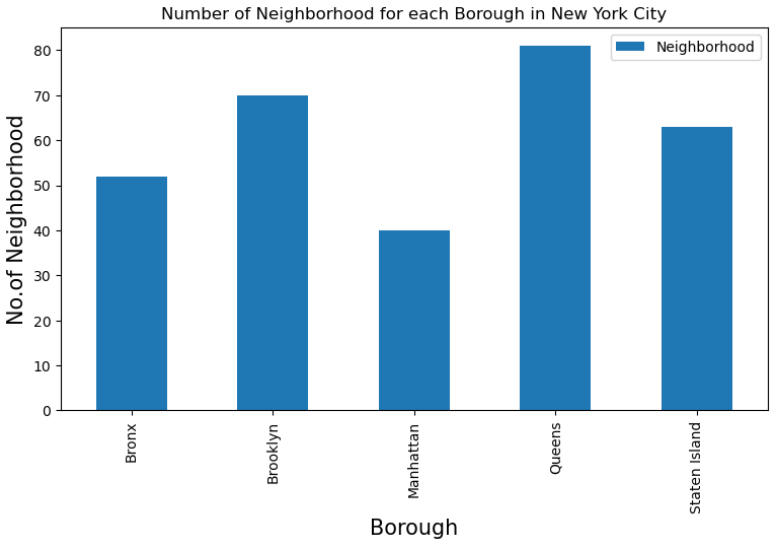
Results

1. Exploration of NYC data

From exploring the NYC data, it was calculated that NYC has 5 boroughs and 306 neighborhoods. We listed latitudes for each borough and neighborhood.

	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

The number of neighborhoods were then calculated per borough. Queens has the highest number of neighborhoods, followed by Brooklyn, Staten Island, Bronx, and Manhattan.

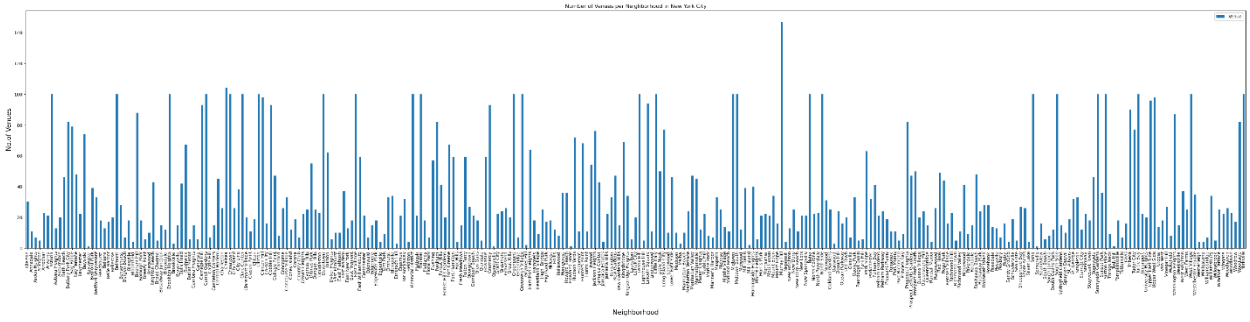


2. FourSquare scraping for venues by neighborhood and borough, data cleaning and processing

Next, FourSquare credentials were used to make a call to the FourSquare API. All venues in the geographic location of New York City were scraped and put into a Pandas dataframe. Latitudes and Longitudes were appended to the data frame as well.

	name	categories	lat	lng
0	The Bar Room at Temple Court	Hotel Bar	40.711448	-74.006802
1	The Beekman, A Thompson Hotel	Hotel	40.711173	-74.006702
2	Alba Dry Cleaner & Tailor	Laundry Service	40.711434	-74.006272
3	City Hall Park	Park	40.712359	-74.007493
4	Gibney Dance Center Downtown	Dance Studio	40.713923	-74.005661

Next, the venues were categorized by type. This included, for example, dessert shop, pharmacy, Japanese restaurants, beaches, yoga studios, and other restaurant and or recreation establishments. The number of venues per neighborhood was calculated and plotted.



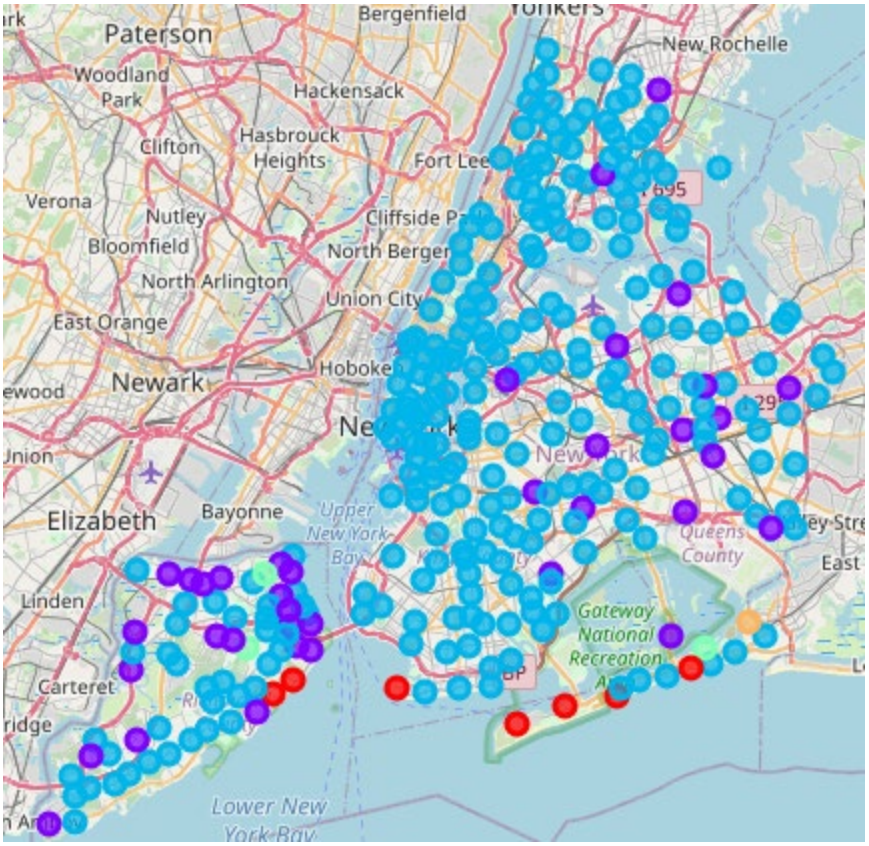
Each neighborhood was further analyzed for venue types. A dataframe was made that displayed the ratio of each venue type per neighborhood. The list of top five venues per neighborhood was printed. For example, Allerton neighborhood has a frequency of .13 for pizza places and .07 each for discount store, supermarket, pharmacy, and deli.

Per neighborhood, the top 10 most common venues were calculated and put into a dataframe.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Allerton	Pizza Place	Deli / Bodega	Supermarket	Discount Store	Pharmacy	Department Store	Chinese Restaurant	Martial Arts School	Bakery	Bus Station
1	Annadale	Pizza Place	Restaurant	Train Station	Bakery	Cosmetics Shop	Pub	Bar	American Restaurant	Diner	Pharmacy
2	Arden Heights	Rental Car Location	Deli / Bodega	Coffee Shop	Bus Stop	Pizza Place	Lawyer	Pharmacy	Fish & Chips Shop	Eye Doctor	Factory
3	Arlington	Grocery Store	Deli / Bodega	Coffee Shop	American Restaurant	Intersection	Yoga Studio	Factory	Falafel Restaurant	Farm	Farmers Market
4	Arrochar	Bus Stop	Italian Restaurant	Pizza Place	Deli / Bodega	Middle Eastern Restaurant	Supermarket	Sandwich Place	Liquor Store	Polish Restaurant	Cosmetics Shop

3. k-means clustering algorithm

To visualize patterns in neighborhoods by common venue types, k-mean clustering algorithm was calculated and a map was generated. 5 clusters were identified.



The algorithm calculated an output of array([2, 2, 1, 1, 1, 2, 2, 2, 2, 2], dtype=int32).

Cluster 1 included 7 neighborhoods. Cluster 1's top venues included beaches, yoga studios, bus stations, supermarket, deli, basketball and baseball areas, and pier. For 3rd and 4th most common venues included restaurants, trails, spa, diners, and dog run.

Cluster 2 included 37 neighborhoods. Top venues were bus stops, delis, pizza places, dog run, rental car location, liquor store.

Cluster 3 included 294 neighborhoods. This cluster had a higher frequency of bars, pubs, fast food restaurants, pizza places, international cuisine, café, gym, spa, coffee shops, cocktail bars and hotels.

Cluster 4 had 3 neighborhoods. They had mostly parks and yoga studios as the top venues. Exhibits, famers markets, and factors were also unique to these neighborhoods.

Cluster 5 only had one neighborhood. Top venue included playground and yoga studio.

Discussion

From the results, we recommend to our client to select one of the following neighborhoods to invest in real estate. These neighborhoods, all in Cluster #1, had a high concentration of beaches and yoga studios. This includes the neighborhoods of:

- Sea Gate
- Breezy Point
- Neponsit
- South Point
- Midland Beach
- Roxbury

Although Hammel has beach access, it is not a suitable neighborhood for our client as it has no yoga studios and a lesser number of restaurants.

Conclusion

From the results, we recommend to our client to buy real estate Sea Gate, Breezy Point, Neponsit, South Point, Midland Beach, Roxbury neighborhoods. Some additional analysis should be made to understand if properties are available in these neighborhoods.