The Battle of Neighborhoods

Capstone Project

July 2019

Introduction

In this project, I will try to find out the best place to start the restaurant business in New York City, New York. New York is the most populous and densely populated city with estimated population of 8,500,000 people. Also, millions of tourists visit the city each year, making it a good place to start a restaurant business. However, the problem is that starting a restaurant business in the city becomes more competitive every year due to rising costs and visitors' expectations. In this notebook, we will be going over some of the features of New York City and find out which type of cuisines would best fit each district. Current/future restaurant owners would benefit from the analysis.

Data

Data were downloaded and scraped from several sources including New York City and Toronto Neighborhood dataset. We also used many open-sourced API such as Geocode and Foursquare. Those data were combined into one table for the ease of use.

We used following data sources and libraries to solve the problem:

- New York City neighborhood dataset
    - https://geo.nyu.edu/catalog/nyu_2451_34572
    - https://cocl.us/new_york_dataset
- Toronto neighborhood dataset
    - https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- Wikipedia Zip Code data
- Geocode
    - Convert Zip Code data into longitude, latitude pair

- Foursquare API
    - Get relevant information based on the location
- List of Cuisine Types
    - https://en.wikipedia.org/wiki/List_of_cuisines
- folium
    - Visualize data

We spent most of the time cleaning and preparing the data. Raw data is not sufficiently clean and adequate to run the procedures described later. Thus, preprocessing is necessary to prepare it for another processing procedures. We have simply converted raw data into more structured and organized data as described below.

Cleaning

Data cleansing process was necessary to correct and remove any unnecessarily coarse data. We did the process in step by step to determine which features are valuable. We got geographic information of New York City using Wikipedia, Geocode, and other available dataset and use Foursquare API to obtain the restaurant information of the locations. Also, we visualized our cleaned data using folium, which is an open-source map library.

Methodology

After analyzing our data, we used two different data science methods to find out the meaningful location to start a new restaurant business. For this report, we chose Battery Park City as the location which restaurant owner wants to open a new business.

K-Nearest Neighbor

At first, we used K-nearest neighbor method to group each neighborhood with its closest (with similar restaurants/features) ones in New York City. K-means clustering is the most common type of method used for clustering. The goal of k-means clustering is to find K different group with data points clustered based on the similarity between points. We have used the implementation from scikit-learn libraries. We first decided to use K-means clustering since other clustering method are slow and memory-intensive. We set n cluster parameter to be 5 for the ease of analysis.

Cosine Similarity

Then, we found cosine similarity between one of the cities in New York (Battery Park City) and each neighborhood in Toronto. Cosine similarity measures the similarity between two vectors by calculating the cosine of the angle between inner product spaces. Cosine similarity ranges between 0 and 1, 0 being the farthest and 1 being the closest.

Results

K-Nearest Neighbors

Cluster 0:

Wakefield, Co-op City, Kingsbridge, Woodlawn, Norwood, Baychester, Bedford Park, University Heights, Morris Heights, Fordham, East Tremont, High Bridge, Melrose, Mott Haven, Longwood, Morrisania, Parkchester, Westchester Square, Morris Park, North Riverdale, Schuylerville, Castle Hill, Pelham Gardens, Unionport, Manhattan Terrace, Crown Heights,

Cypress Hills, Starrett City, Manhattan Beach, Borough Park, City Line, Bergen Beach, Midwood, Prospect Park South, Richmond Hill, East Elmhurst, Maspeth, Glendale, Ozone Park, Glen Oaks, Bellerose, Kew Gardens Hills, Fresh Meadows, Rochdale, Springfield Gardens, Far Rockaway, Beechhurst, Edgemere, Arverne, Floral Park, Holliswood, Lindenwood, Rockaway Park, St. George, Castleton Corners, New Springville, Great Kills, Eltingville, Annadale, Dongan Hills, Grant City, Pleasant Plains, Rossville, Greenridge, Heartland Village, Bulls Head, New Lots, Utopia, Pomonok, Claremont Village, Mount Eden, Mount Hope, Manor Heights, Sandy Ground, Prince's Bay, Allerton, Kingsbridge Heights

Cluster 1:

Van Nest, East New York, Marine Park, Ocean Hill, South Ozone Park, Whitestone, Douglaston, Briarwood, Broad Channel, Brookville, North Corona, New Brighton, South Beach, Port Richmond, Mariner's Harbor, Travis, Tottenville, Arrochar, Arden Heights, Shore Acres, Randall Manor, Elm Park, Blissville, Willowbrook, Roxbury, Highland Park, Madison

Cluster 2:

Country Club, Graniteville, Queensbridge, Fox Hills

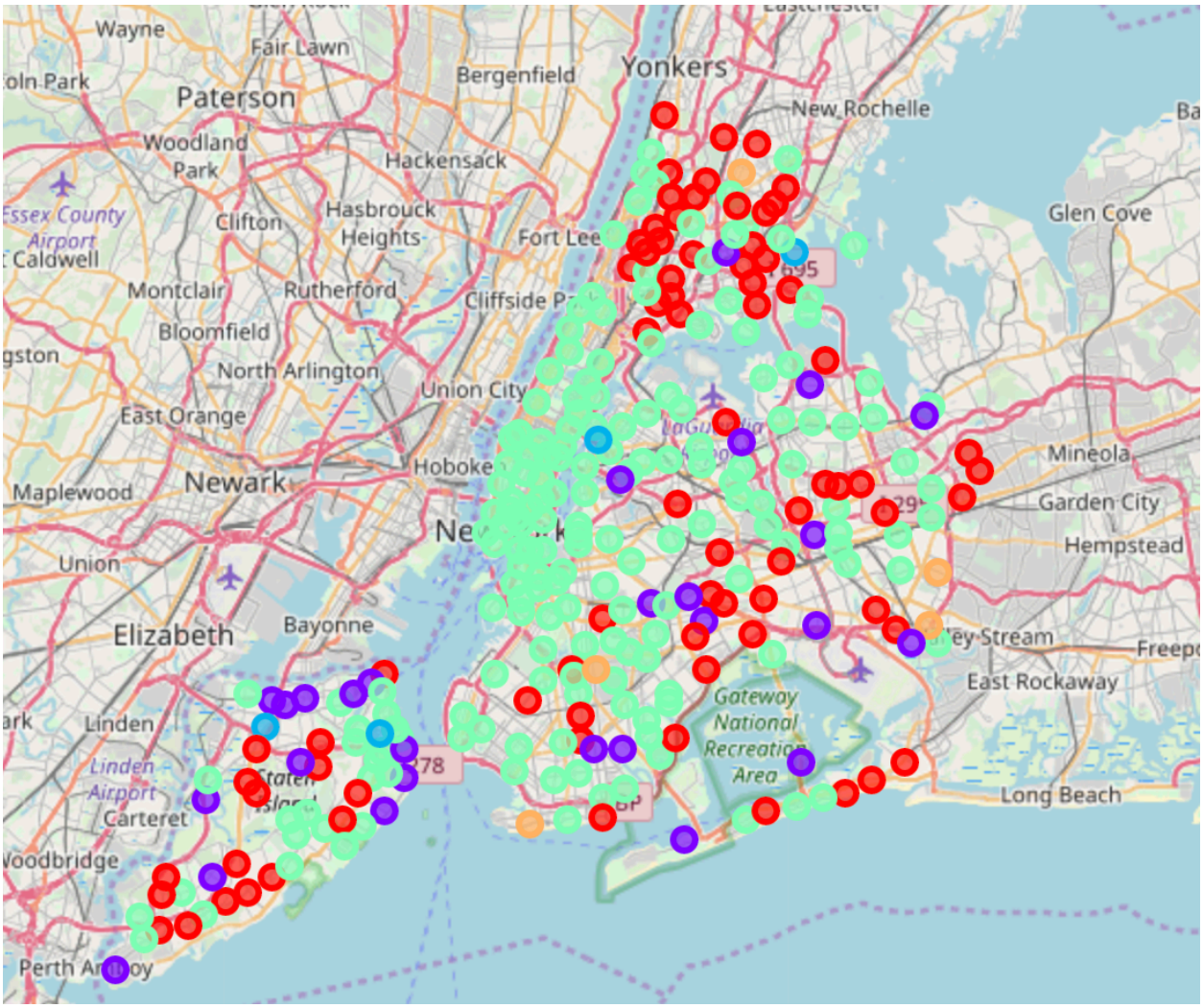
Cluster 3:

Eastchester, Riverdale, Marble Hill, Pelham Parkway, City Island, West Farms, Port Morris, Hunts Point, Soundview, Clason Point, Throgs Neck, Belmont, Spuyten Duyvil, Pelham Bay, Edgewater Park, Olinville, Concourse, Bay Ridge, Bensonhurst, Sunset Park, Greenpoint, Gravesend, Brighton Beach, Sheepshead Bay, Flatbush, East Flatbush, Kensington, Windsor Terrace, Prospect Heights, Brownsville, Williamsburg, Bushwick, Bedford Stuyvesant, Brooklyn Heights, Cobble Hill, Carroll Gardens, Red Hook, Gowanus, Fort Greene, Park Slope, Canarsie, Flatlands, Mill Island, Bath Beach, Dyker Heights, Gerritsen Beach, Clinton Hill, Downtown,

Boerum Hill, Prospect Lefferts Gardens, Georgetown, East Williamsburg, North Side, South Side, Ocean Parkway, Fort Hamilton, Chinatown, Washington Heights, Inwood, Hamilton Heights, Manhattanville, Central Harlem, East Harlem, Upper East Side, Yorkville, Lenox Hill, Roosevelt Island, Upper West Side, Lincoln Square, Clinton, Midtown, Murray Hill, Chelsea, Greenwich Village, East Village, Lower East Side, Tribeca, Little Italy, Soho, West Village, Manhattan Valley, Morningside Heights, Gramercy, Battery Park City, Financial District, Astoria, Woodside, Jackson Heights, Elmhurst, Howard Beach, Corona, Forest Hills, Kew Gardens, Flushing, Long Island City, Sunnyside, Ridgewood, Rego Park, Woodhaven, College Point, Bayside, Auburndale, Little Neck, Jamaica Center, Oakland Gardens, Queens Village, Hollis, South Jamaica, St. Albans, Rosedale, Steinway, Bay Terrace, Rockaway Beach, Murray Hill, Queensboro Hill, Hillcrest, Ravenswood, Lefrak City, Belle Harbor, Bellaire, Forest Hills Gardens, Stapleton, Rosebank, West Brighton, New Dorp, Woodrow, Tompkinsville, Silver Lake, Sunnyside, Ditmas Park, Wingate, Rugby, Park Hill, Arlington, Grasmere, Old Town, Midland Beach, New Dorp Beach, Bay Terrace, Huguenot, Charleston, Chelsea, Carnegie Hill, Noho, Civic Center, Midtown South, Richmond Town, Clifton, Concord, Remsen Village, Paerdegat Basin, Mill Basin, Jamaica Hills, Astoria Heights, Concourse Village, Sutton Place, Hunters Point, Turtle Bay, Tudor City, Stuyvesant Town, Flatiron, Sunnyside Gardens, Fulton Ferry, Vinegar Hill, Weeksville, Broadway Junction, Dumbo, Egbertville, Homecrest, Middle Village, Lighthouse Hill, Richmond Valley, Malba, Bronxdale, Hudson Yards, Hammels

Cluster 4:

Williamsbridge, Coney Island, Cambria Heights, Laurelton, Erasmus



(K-Nearest Neighbor Result in Map)

Cosine Similarity

	Neighborhood	cossim
70	Swansea, Runnymede	0.792118
10	Maryvale, Wexford	0.784633
49	Victoria Hotel, Commerce Court	0.772061
69	Roncesvalles, Parkdale	0.723370
52	Yorkville, North Midtown, The Annex	0.702311
39	St. James Town, Cabbagetown	0.693559

We found that Swansea and Runnymede in Toronto is closest to Battery Park City, NY with cosine similarity of 0.79. However, Maryvale and Wexford are also very similar to Battery Park City with cosine similarity of 0.78.

Discussion

Overall, we have discussed two different approaches and got some meaningful results. We could have used cosine similarity on other New York City, or use K-Nearest Neighbor method on Toronto neighborhoods. Searching for the correct value for *K* in k-nearest neighbor method is not easy as a small value of k means that noise will have a higher influence on the result and a large value make it computationally expensive. Using different *K* would result different outcomes and would be valuable to further investigate those outcomes. We also could have tried different methods of finding similarities. For instance, we could have used either Euclidean Distance or Manhattan Distance to calculate the similarity between two features

Conclusion

In this study, I analyzed the possible location for restaurant owners to go over before they start a new business. I simply focused on the number of restaurants in each neighborhood, and primarily used them as most important features. I used K-nearest neighbor and cosine similarity to find out the most similar neighborhoods. Though, I only have analyzed results using Battery Park City, same model can be used with different cities as well. These results can be very useful in helping business starters in many different ways. For example, new restaurant owners in Battery Park City can look over Swansea and Runnymede in Toronto to figure out what kind of restaurants would be popular in the area.