The Battle of the Neighborhoods

(New York City)

This report consists of four parts: description of problem, data, methodology, visualization and results. We will discuss about the problem and data in this week's submission.

Introduction

This project deals with discussing the neighborhoods of New York City, The Detroit of USA. This project would specifically help Business people planning to start Restaurants, Hotels, etc. in New York City, Manhattan, USA.

The Foursquare API is used to access the venues in the neighborhoods. Since, it returns less venues in the neighborhoods, we would be analysing areas for which countable number of venues are obtained. Then they are clustered based on their venues using Data Science Techniques. Here the k-means clustering algorithm is used to achieve the task. The optimal number of clusters can be obtained using silhouette score. Folium visualization library can be used to visualize the clusters superimposed on the map of New York city. These clusters can be analyzed to help small scale business owners select a suitable location for their need such as Hotels, shopping malls, Restaurants or even specifically restaurants or Coffee shops.

The major Target Audience would be small-scale business owners and stake holders planning to start their business at a location in Bangalore. This project would help them find the optimal location based on the category of their business such as

- What is the best location to start a new hotel in New York with restaurants around?
- Which area is best suitable for opening a Coffee Shop in New York City?

Data Requirements

- New York has multiple neighborhoods. The Opan Data City of New York website has a dataset which has the list of locations in New York City along with their Latitude and Longitude. In order to obtain the venue details in each neighborhood Foursquare API is used

- I used https://foursquare.com/ to get the most common venues of given Borough of New York City.

There is a total of 5,984 neighborhoods for Manhattan and Brooklyn. But the Latitude and Longitude data obtained are in Degrees Minute Seconds format which needs to be converted to Decimal Degrees Format. The following data are obtained from the Foursquare API.

- Venue
- Venue Latitude
- Venue Longitude
- Venue Category
- A total of 3252 venues for Manhattan and A total of 2732 venues for Brooklyn data have been obtained from Foursquare.

```
In [17]: manhattan_venues = getNearbyVenues(names=manhattan_data['Neighborhood'],
                                            latitudes=manhattan_data['Latitude'],
                                            longitudes=manhattan_data['Longitude']
                                           )

In [18]: print('There are {} venues in Manhattan.'.format(manhattan_venues.shape[0]))

         There are 3252 venues in Manhattan.
```

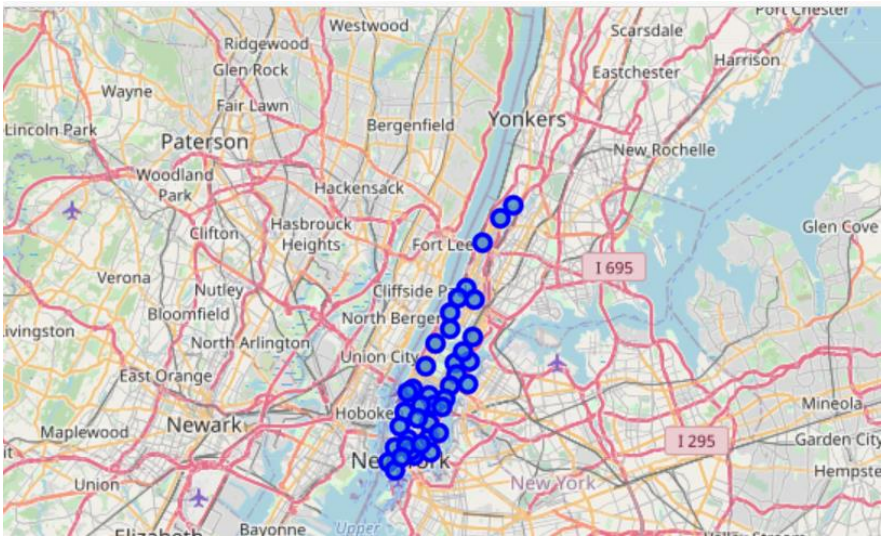
Explore the neighborhoods of Brooklyn using Foursquare API

```
brooklyn_venues = getNearbyVenues(names=brooklyn_data['Neighborhood'],
                                  latitudes=brooklyn_data['Latitude'],
                                  longitudes=brooklyn_data['Longitude']
                                 )

print('There are {} venues in Brooklyn.'.format(brooklyn_venues.shape[0]))

There are 2732 venues in Brooklyn.
```

Methodology

Now, we have the neighborhoods data of Manhattan (3252 neighborhoods). We also have the most popular venues in each neighborhood obtained using Foursquare API. A total of 5,984 venues have been obtained in the whole city and 1,547 unique categories. But as seen we have multiple neighborhoods 925 restaurants for Manhattan returned. In order to create a good analysis let's consider only the neighborhoods 622 for Brooklyn.

I used python folium library to visualize geographic details of New York City and its boroughs and I created a map of New York City with boroughs superimposed on top. I used latitude and longitude values to get the visual as below:



1.1 Map of the New York (Manhattan) neighborhoods.

1.2 Map of the New York (Brooklyn) neighborhoods.

We have some common venue categories in boroughs. In this reason I used unsupervised learning K-means algorithm to cluster the boroughs. K-Means algorithm is one of the most common cluster method of unsupervised learning.

Out[46]:

	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	Bath Beach	Chinese Restaurant	Pharmacy	Bubble Tea Shop	Italian Restaurant	Fast Food Restaurant	Donut Shop	Pizza Place	Gas Station	Mobile Phone Shop	Construction & Landscaping
1	Bay Ridge	Italian Restaurant	Spa	Pizza Place	Greek Restaurant	Bar	Bagel Shop	Chinese Restaurant	American Restaurant	Grocery Store	Sushi Restaurant
2	Bedford Stuyvesant	Bar	Coffee Shop	Pizza Place	Deli / Bodega	Café	Fruit & Vegetable Store	Fried Chicken Joint	Gourmet Shop	Gift Shop	Boutique
3	Bensonhurst	Chinese Restaurant	Sushi Restaurant	Ice Cream Shop	Donut Shop	Pharmacy	Italian Restaurant	Bank	Liquor Store	Coffee Shop	Noodle House
4	Bergen Beach	Harbor / Marina	Park	Athletics & Sports	Baseball Field	Trail	Playground	Farm	Farmers Market	Fast Food Restaurant	Field
5	Boerum Hill	Coffee Shop	Bar	Dance Studio	French Restaurant	Sandwich Place	Furniture / Home Store	Arts & Crafts Store	Bakery	Spa	Cocktail Bar
6	Borough Park	Bank	Pizza Place	Grocery Store	Deli / Bodega	Pharmacy	Café	Chinese	Fast Food	Bistro	Coffee Shop

In [48]:

```
brooklyn_merged = brooklyn_data

# add clustering labels
brooklyn_merged['Cluster Labels'] = brooklyn_kmeans.labels_

brooklyn_merged = brooklyn_merged.join(brooklyn_neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

brooklyn_merged.head() # check the last columns!
```

Out[48]:

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	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	Brooklyn	Bay Ridge	40.625801	-74.030621	6	Italian Restaurant	Spa	Pizza Place	Greek Restaurant	Bar	Bagel Shop	Chinese Restaurant	American Restaurant	Grocery Store	Sushi Restaurant
1	Brooklyn	Bensonhurst	40.611009	-73.995180	7	Chinese Restaurant	Sushi Restaurant	Ice Cream Shop	Donut Shop	Pharmacy	Italian Restaurant	Bank	Liquor Store	Coffee Shop	Noodle House
2	Brooklyn	Sunset Park	40.645103	-74.010316	7	Mexican Restaurant	Pizza Place	Latin American Restaurant	Bank	Bakery	Pharmacy	Gym	Fried Chicken Joint	Mobile Phone Shop	Grocery Store
3	Brooklyn	Greenpoint	40.730201	-73.954241	6	Coffee Shop	Bar	Pizza Place	Cocktail Bar	Yoga Studio	Furniture / Home Store	Deli / Bodega	French Restaurant	Nail Salon	Garden Center
4	Brooklyn	Gravesend	40.595260	-73.973471	0	Pizza Place	Lounge	Bakery	Chinese Restaurant	Bar	Sporting Goods Shop	Electronics Store	Spa	Bus Station	Diner

The clusters obtained can be analyzed to find the major type of venue categories in each cluster. This data can be used to suggest business people, suitable locations based on the category.

Conclusion

Purpose of this project was to analyze the neighborhoods of New York City and create a clustering model to suggest personals places to start a new business based on the category. The neighborhoods data was obtained from an online source and the Foursquare API was used to find the major venues in each neighborhood. But we found that many neighborhoods had less than 5,984 venues returned. In order to build a good Data Science model, we filtered out these locations. The remaining locations were used to create a clustering model. The best number of clusters i.e., 2 was obtained using the silhouette score. Each cluster was examined to find the most venue categories present, that defines the characteristics for that particular cluster.

A few examples for the applications that the clusters can be used for have also been discussed. A map showing the clusters have been provided. Both these can be used by stakeholders to decide the location for the particular type of business. A major drawback of this project was that the Foursquare API returned only few venues in each neighborhood. As a future improvement, better data sources can be used to obtain more venues in each neighborhood. This way the neighborhoods that were filtered out can be included in the clustering analysis to create a better decision model.

Libraries used to Develop the Project:

- Pandas: For creating and manipulating data frames.
- Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution using an 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.