Introduction and Business Problem

New York

The idea about relocating to a different city is always terrifying. People need to research a lot before actually deciding a place to stay in city where you have never been to. The research basically means a variety of information needed to be gathered. This could be area, region, business, position size, among several other considerations like community study. This can be termed as a search algorithm request that usually returns the required features such as population rate, median house price, school ratings, crime rates, weather conditions, recreational facilities etc.

Getting an application which could render things simple by providing a comparative study with given variables within the locality would be helpful and good.

The user can use this project when renting apartment or buying house in a locality based on the distribution of various facilities available around the neighborhood. As an example, this project would compare 2 randomly selected neighborhoods and analyze the top 10 most common locations in each of those two neighborhoods based on the number of people visiting each of those locations. This project also uses the K-mean clustering of unsupervised machine learning algorithms to cluster locations based on location categories such as restaurants, parks, coffee shops, gyms, clubs etc.

DATA

Sources of Information:

FourSquare API:

This API has a database of more than 105 million places. This project would use Four-square API as its prime data gathering source. Many organizations are using to geo-tag their photos with detailed info about a destination, while also serving up contextually relevant locations for those who are searching for a place to eat, drink or explore. This API provides the ability to perform location search, location sharing and details about a business. Foursquare users can also use photos, tips and reviews in many productive ways to add value to the results.

HTTP requests would be made to this Foursquare API server using zip codes of the New York city neighborhoods to pull the location information (Latitude and Longitude).Foursquare API search feature would be enabled to collect the nearby places of the neighborhoods. Due to http request limitations the number of places per neighborhood parameter would reasonably be set to 100 and the radius parameter would be set to 700.

Folium- Python visualization library would be used to visualize the neighborhoods cluster distribution of New York city over an interactive leaflet map.Extensive comparative analysis of two randomly picked neighborhoods world be carried out to derive the desirable insights from the outcomes using python’s scientific libraries Pandas, NumPy and Scikit-learn.

- Unsupervised machine learning algorithm K-mean clustering would be applied to form the clusters of different categories of places residing in and around the neighborhoods. These clusters from each of those two chosen neighborhoods would be analyzed individually collectively and comparatively to derive the conclusions.

Packages:

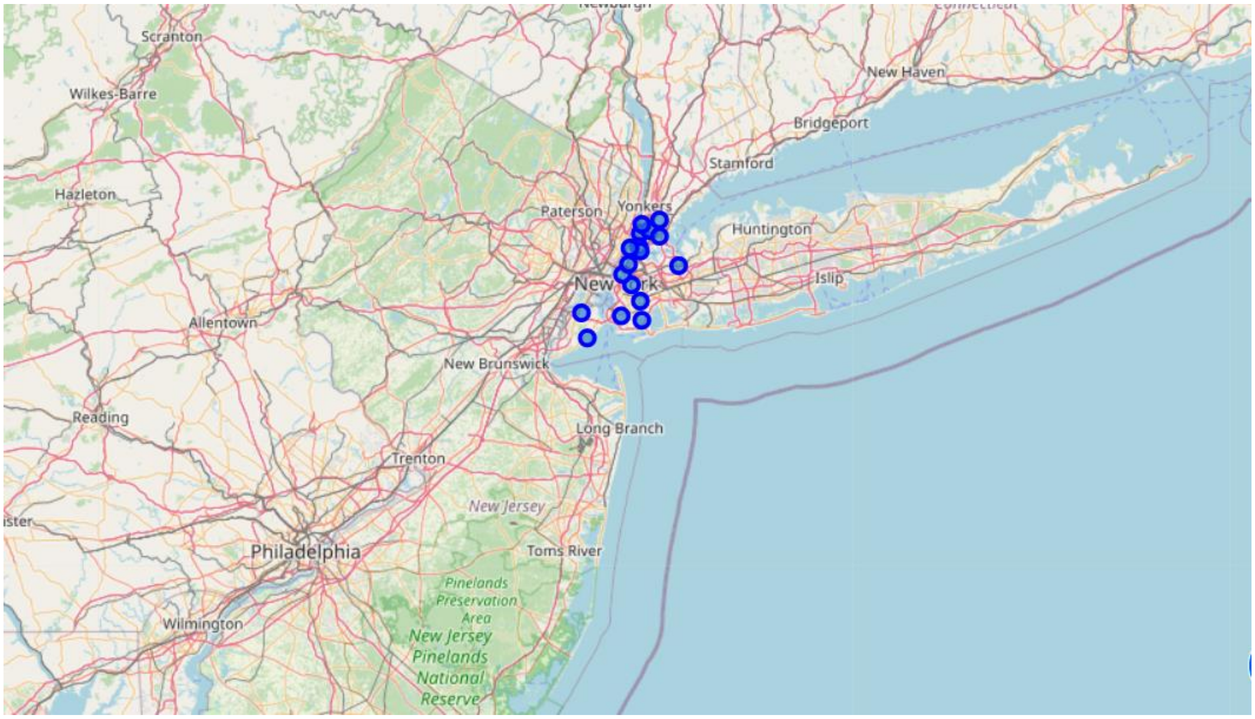
Pandas - Library for Data Analysis

- NumPy – Library to handle data in a vectorized manner
- JSON – Library to handle JSON files
- Geopy – To retrieve Location Data
- Requests – Library to handle http requests
- Matplotlib – Python Plotting Module
- Sklearn – Python machine learning Library
- Folium – Map rendering Library

Methodology:

The workflow of the project starts with the web scraping and data wrangling. Using the Beautiful Soup library, we extract the postal code and the neighborhood is processed to derive the latitude and the longitude of the New York City neighborhood given the CSV file.

With the folium Map, the latitude and longitude of the Chicago neighborhood provides the choropleth visualization.



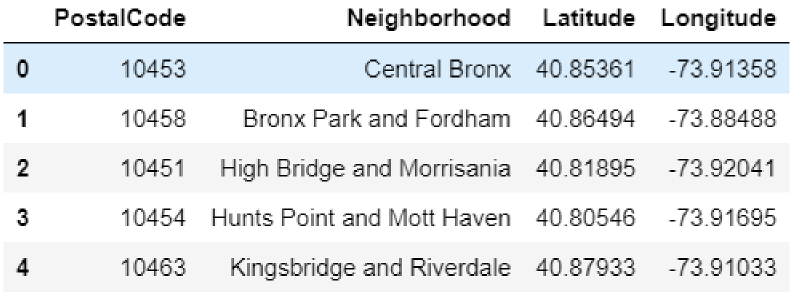
FourSquare API and K-means clustering methods are used to retrive the top trend venues of the New York neighborhood.

Elbow criterion method is used here to the optimum number of cluster present in the dataset. Silhoutte Cofficient analysis is also used to find the number of cluster.

In the Project, we find according to the neighborhoods, all the nearby top places for people to visit. So, according to a family's needs, a person can be on point while deciding the area he/she wants to move to.

Results:

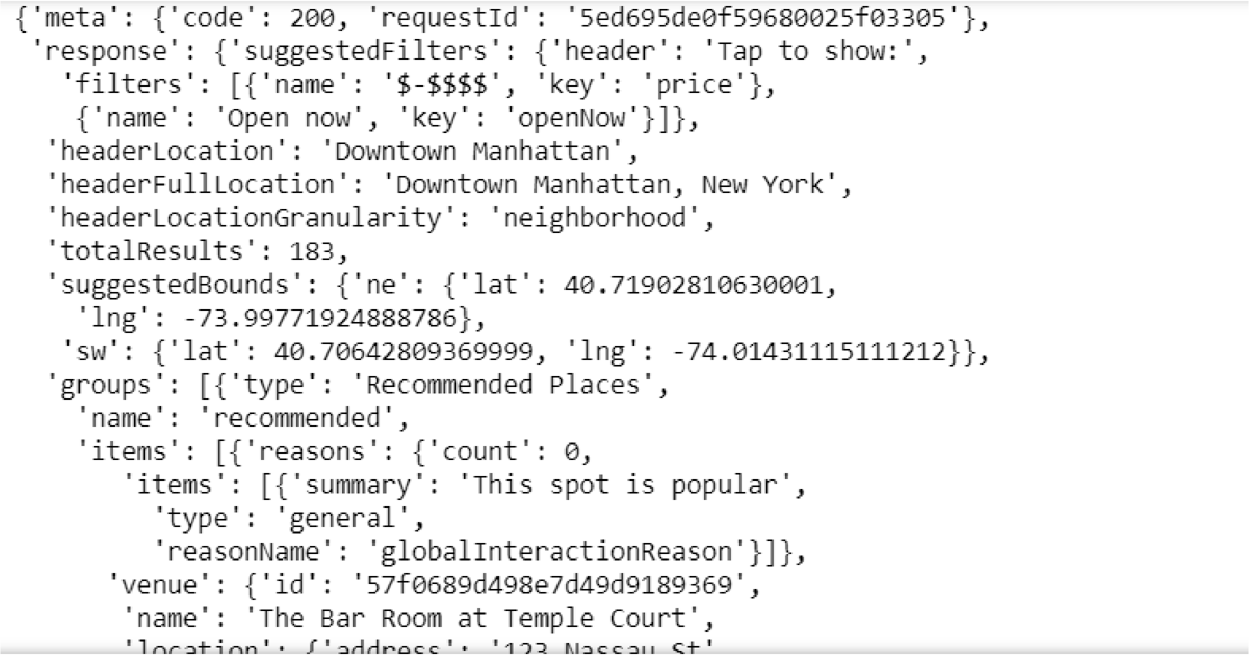
Initially, we find out all the latitudes and longitudes of a given neighborhood in the csv file.



	PostalCode	Neighborhood	Latitude	Longitude
0	10453	Central Bronx	40.85361	-73.91358
1	10458	Bronx Park and Fordham	40.86494	-73.88488
2	10451	High Bridge and Morrisania	40.81895	-73.92041
3	10454	Hunts Point and Mott Haven	40.80546	-73.91695
4	10463	Kingsbridge and Riverdale	40.87933	-73.91033

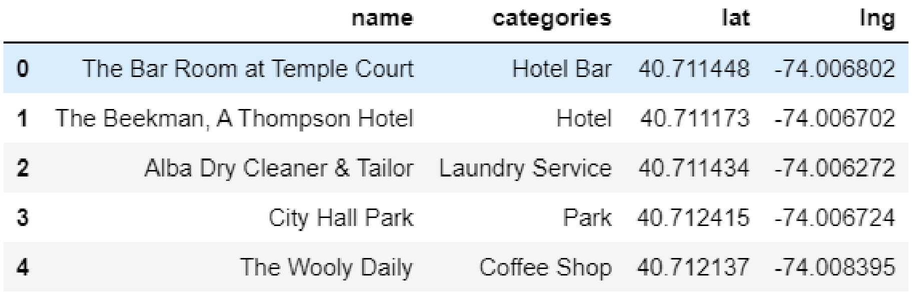
This will help generate the map using the folium library.

Now using the FourSquare API, we will pass on all the API credentials and using the latitude and longitude of the data frame, we will get the results in a JSON format.



```
{'meta': {'code': 200, 'requestId': '5ed695de0f59680025f03305'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': '$-$$$$', 'key': 'price'},
    {'name': 'Open now', 'key': 'openNow'}]},
  'headerLocation': 'Downtown Manhattan',
  'headerFullLocation': 'Downtown Manhattan, New York',
  'headerLocationGranularity': 'neighborhood',
  'totalResults': 183,
  'suggestedBounds': {'ne': {'lat': 40.71902810630001,
    'lng': -73.99771924888786},
   'sw': {'lat': 40.70642809369999, 'lng': -74.01431115111212}},
  'groups': [{'type': 'Recommended Places',
    'name': 'recommended',
    'items': [{'reasons': {'count': 0,
       'items': [{'summary': 'This spot is popular',
         'type': 'general',
         'reasonName': 'globalInteractionReason'}]},
      'venue': {'id': '57f0689d498e7d49d9189369',
       'name': 'The Bar Room at Temple Court',
       'location': {'address': '123 Nassau St',
```

Now, we need to parse the JSON properly and retrieve categories for hotspots in a neighborhood.



	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	40.712415	-74.006724
4	The Wooly Daily	Coffee Shop	40.712137	-74.008395

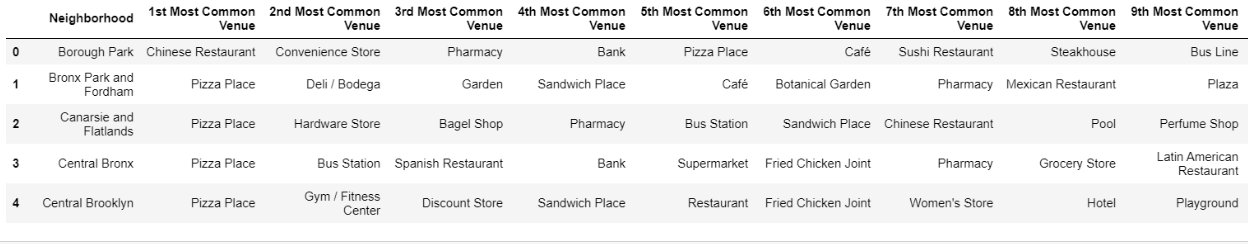
The next step would be to classify which venue fits in which neighborhood according to the given latitude and longitudes. So, we reshape the data frame and merge it.

```
----Borough Park----
                venue  freq
0  Chinese Restaurant  0.15
1                Bank  0.08
2            Pharmacy  0.08
3   Convenience Store  0.08
4         Pizza Place  0.05


----Bronx Park and Fordham----
            venue  freq
0     Pizza Place  0.08
1   Deli / Bodega  0.07
2          Garden  0.07
3  Sandwich Place  0.05
4            Café  0.05


----Canarsie and Flatlands----
            venue  freq
0     Pizza Place  0.12
1  Hardware Store  0.08
2      Bagel Shop  0.08
3            Pool  0.04
4          Bakery  0.04
```

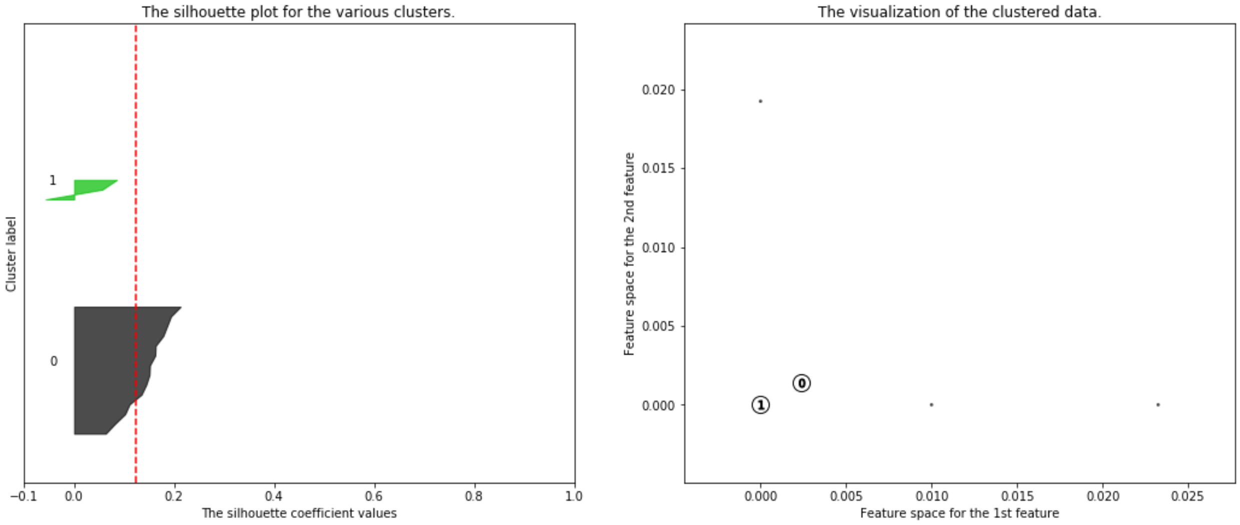
The given results will be transferred into a pandas data frame.



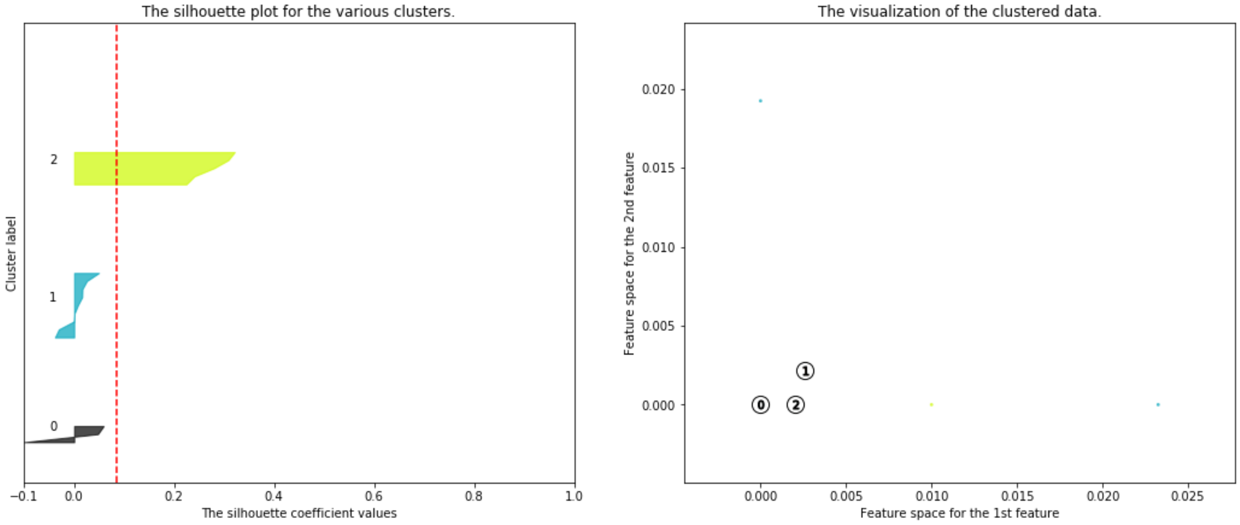
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue
0	Borough Park	Chinese Restaurant	Convenience Store	Pharmacy	Bank	Pizza Place	Café	Sushi Restaurant	Steakhouse	Bus Line
1	Bronx Park and Fordham	Pizza Place	Deli / Bodega	Garden	Sandwich Place	Café	Botanical Garden	Pharmacy	Mexican Restaurant	Plaza
2	Canarsie and Flatlands	Pizza Place	Hardware Store	Bagel Shop	Pharmacy	Bus Station	Sandwich Place	Chinese Restaurant	Pool	Perfume Shop
3	Central Bronx	Pizza Place	Bus Station	Spanish Restaurant	Bank	Supermarket	Fried Chicken Joint	Pharmacy	Grocery Store	Latin American Restaurant
4	Central Brooklyn	Pizza Place	Gym / Fitness Center	Discount Store	Sandwich Place	Restaurant	Fried Chicken Joint	Women's Store	Hotel	Playground

We used silhouette score to determine the number of clusters.

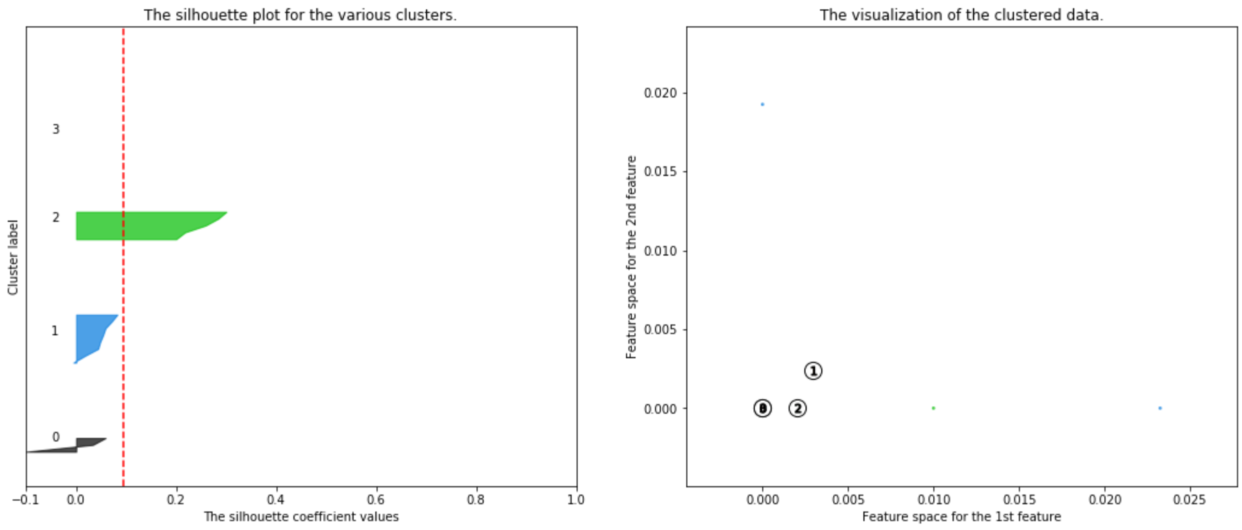
Silhouette analysis for KMeans clustering on sample data with n_clusters = 2



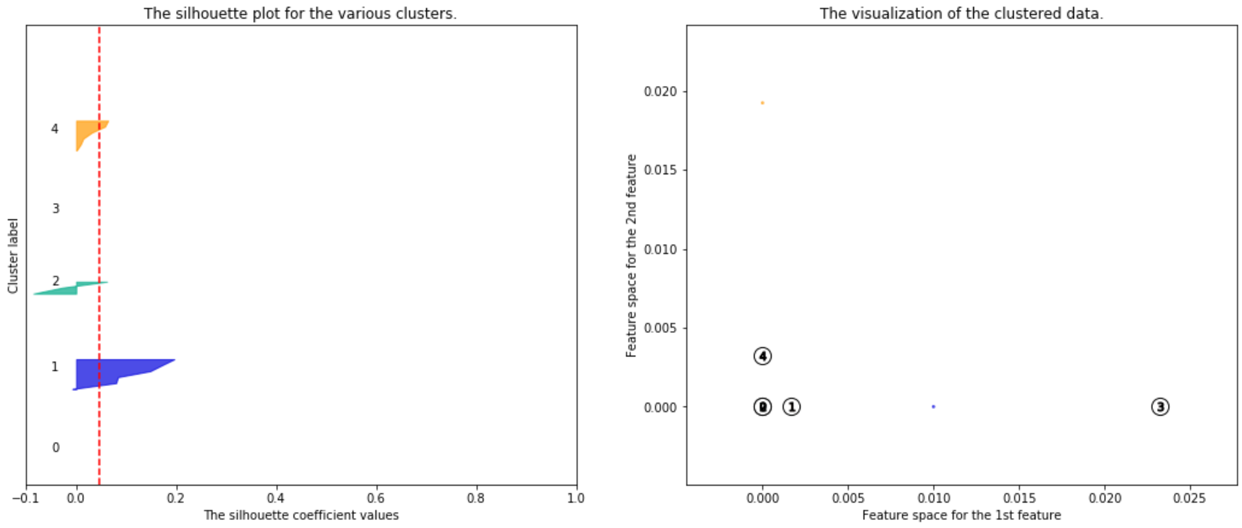
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3



Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



Discussion:

This project is beneficial in many terms, it will save the users time and money. Many a times, people when moved to the new place, they used to worry for settling down. This project recommend the better places in a very less time.

Conclusions

- I feel rewarded with the efforts, time and cash spent. I believe this course with all the topics lined is well worthy of appreciation.

- The mapping with geological formation could be a terribly powerful technique to consolidate data and create the analysis and call thoroughly and confidently. I'd suggest for use in similar things.

- One should keep up with recent tools for DS that continue to appear for application in many business fields.