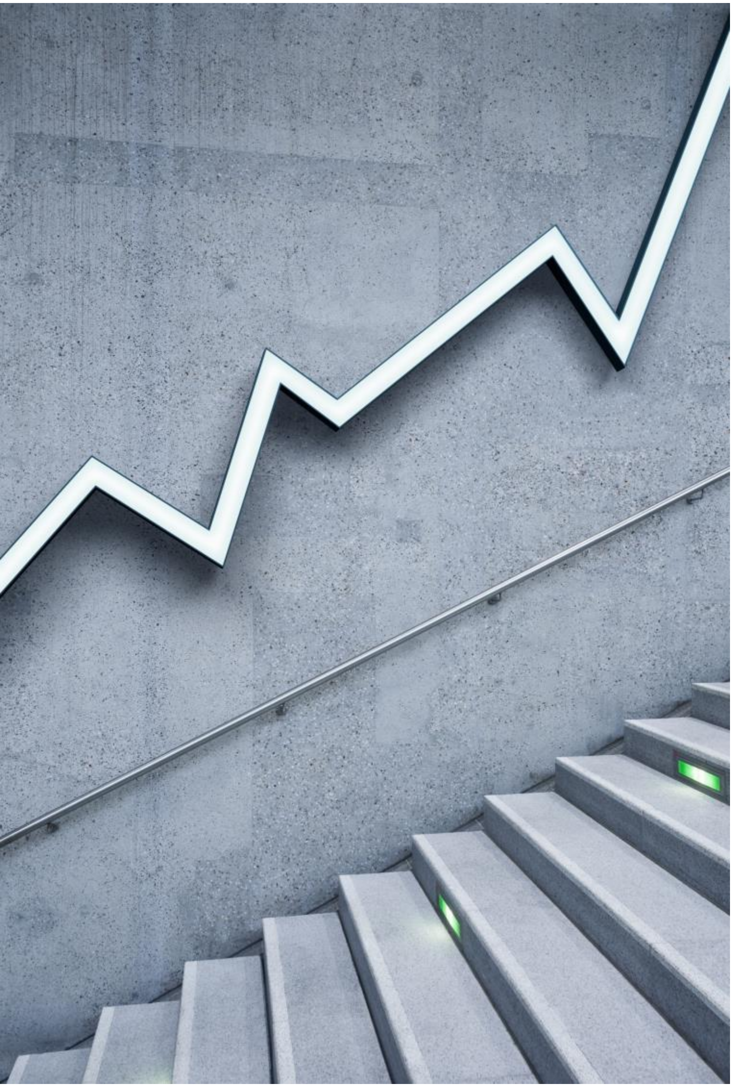


The Battle of Neighborhood

SYED SADAQAT ALI

Introduction

New York City is one of the most ethnically diverse cities in the world. It started accepting immigrants in 19th century and have since became a melting pot of the diverse languages, people and cultures. In 2019, it was estimated to have a population of 8.3 million which live in 5 main boroughs namely Brooklyn, Bronx, Manhattan, Queens and Staten Island.

New York is a very busy city, both in terms of its population and tourists. As per the data of 2019, the populations of Asians American is more than a million , which is about 12 % of the population of New York city. Add to it the people from subcontinent i.e. India, Bangladesh and Pakistan, Srilanka and Nepal and you have a very large Asian and south east Asian population that has come to New York for studying, jobs and businesses.

Business Problem Description

An entrepreneur has recently moved to New York's and is looking to open a restaurant in Manhattan. His initial market research reveals that there is good opportunity for a Thai food restaurant as it equally popular among the European and north American tourists who crave exotic food and will pay hefty prices for a upbeat expensive restaurant that provides good ambiance and Asian population who loves aromatic and spicy food at affordable prices. Manhattan is also very diverse in terms of earning and the entrepreneur is thus planning to open two Thai restaurants, an expensive version providing a fine dinning experience to wealthy residents and tourists and an express version for middle class clients.

The business man thus hires a data science firm to suggest to him the location for two restaurants, the upbeat expensive restaurant in an locality which offers the opportunity for less competition for Thai food at affordable prices in a neighbourhood which has low number of Thai food restaurants.

Target Audience

A entrepreneur who wants to open a Thai Restaurant in Manhattan,New York.

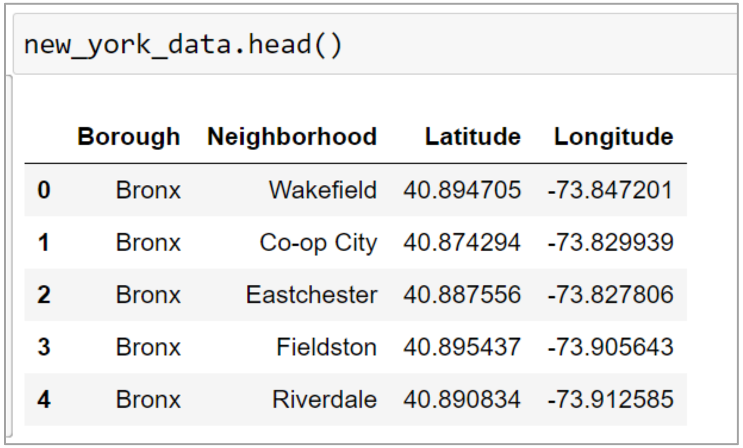
Data Sources

Geospatial data of the boroughs

New York population is distributed into 5 boroughs and 306 neighbourhoods. To explore the data, we need to get the access to the data containing the boroughs and their geospatial coordinates. I downloaded the data freely available from the website

https://cocl.us/new_york_dataset

in geojason format. This data will be transformed into Pandas data frame for easy data analysis and visualization.



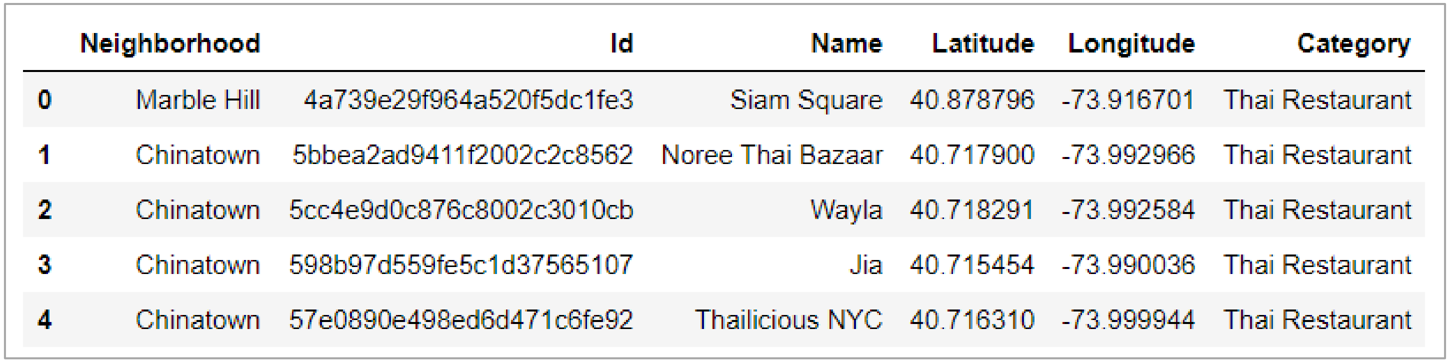
```
new_york_data.head()
```

	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

Data Sources (Continued)

Additionally from FourSquare API , we get Neighborhood names, ID, restaurant names, lat/long and restaurant categories.

The data was retrieved from Foursquare API that required a user account and the secret key and Client ID which allows to extract required data from Foursquare API



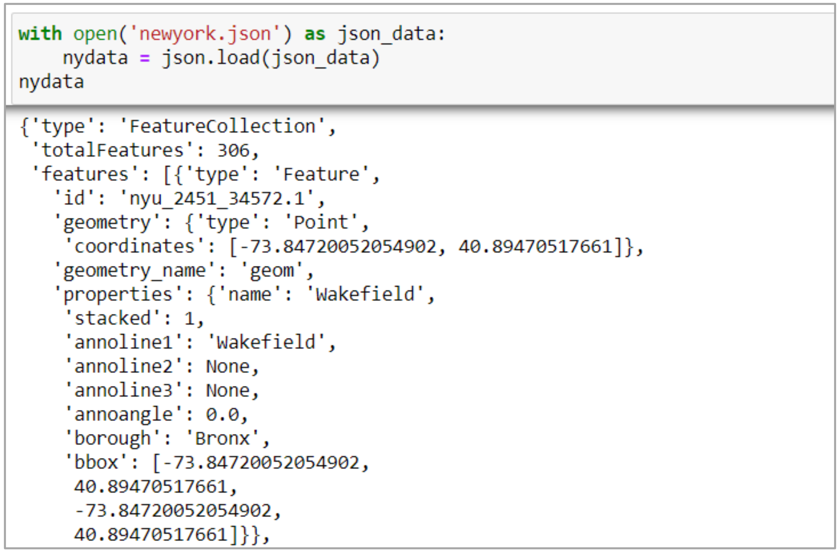
	Neighborhood	Id	Name	Latitude	Longitude	Category
0	Marble Hill	4a739e29f964a520f5dc1fe3	Siam Square	40.878796	-73.916701	Thai Restaurant
1	Chinatown	5bbea2ad9411f2002c2c8562	Noree Thai Bazaar	40.717900	-73.992966	Thai Restaurant
2	Chinatown	5cc4e9d0c876c8002c3010cb	Wayla	40.718291	-73.992584	Thai Restaurant
3	Chinatown	598b97d559fe5c1d37565107	Jia	40.715454	-73.990036	Thai Restaurant
4	Chinatown	57e0890e498ed6d471c6fe92	Thailicious NYC	40.716310	-73.999944	Thai Restaurant

Methodology – Data Retrieval

Data was retrieved from

https://cocl.us/new_york_dataset

The data is automatically downloaded to a file called “nyu_2451_34572-geojason.json”. The file was renamed to newyork.Jason for easy referral.

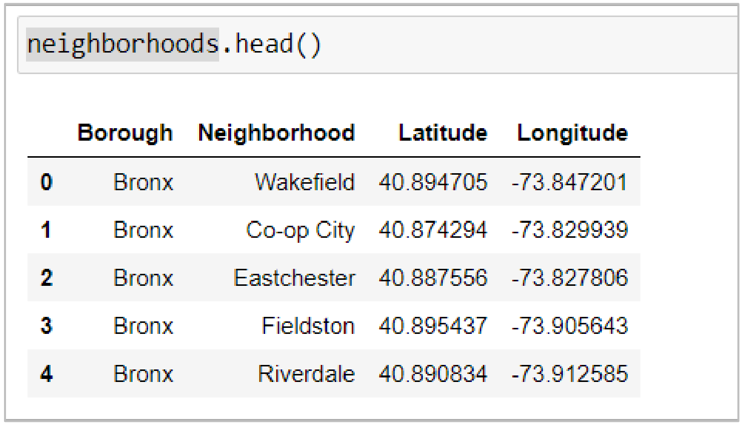


```
with open('newyork.json') as json_data:
    nydata = json.load(json_data)
nydata
```

```
{'type': 'FeatureCollection',
 'totalFeatures': 306,
 'features': [{'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]}},
```

Methodology – Data Transformation

Data was transformed from .json format to pandas dataframe called “neighborhoods”

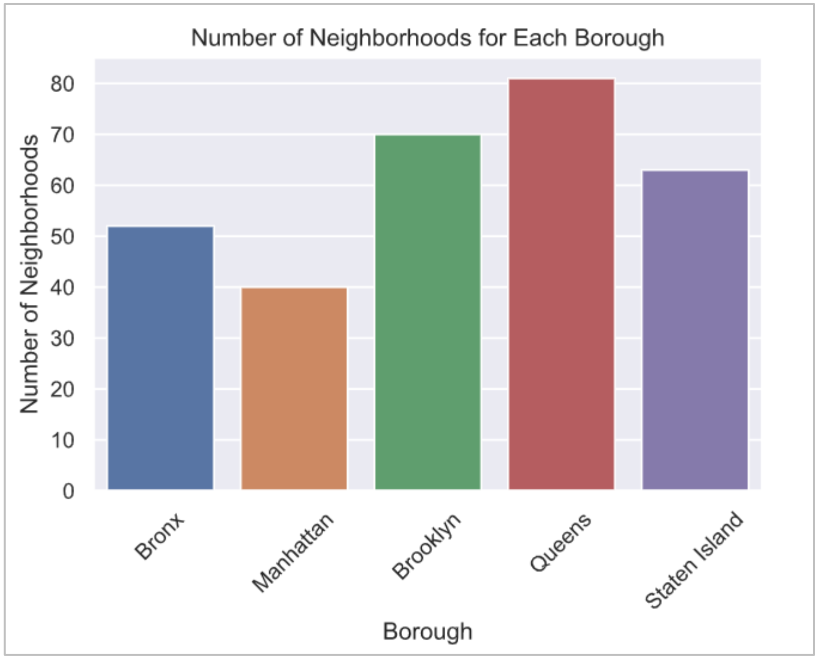


```
neighborhoods.head()
```

	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

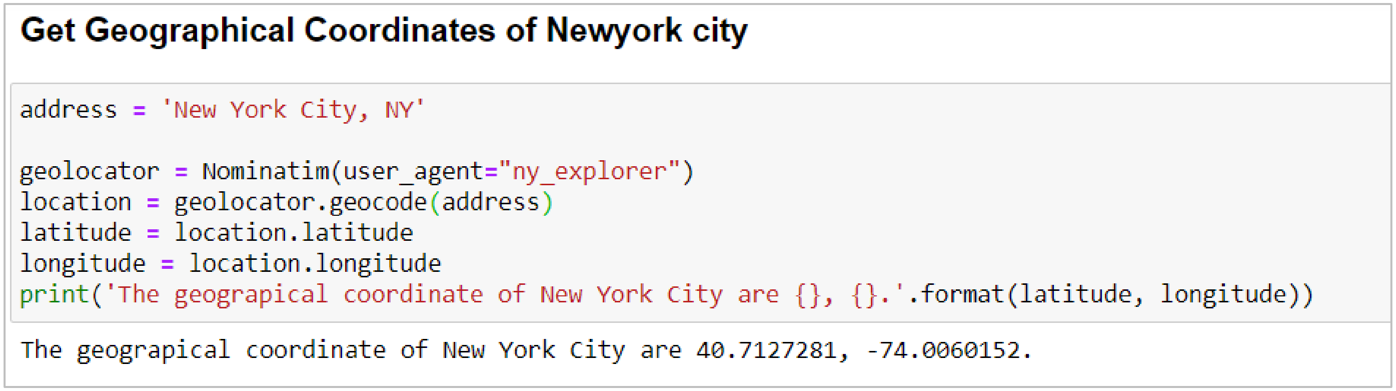
Methodology – Data Visualization

The transformed data was then plotted as bar plot using Seaborn plotting library. Boroughs were plotted against the neighborhoods to get a feel neighborhood distribution per burough



Methodology – Geocoder Locations

Geopy package helps to retrieve the location data from any selected area such as neighborhood or borrows



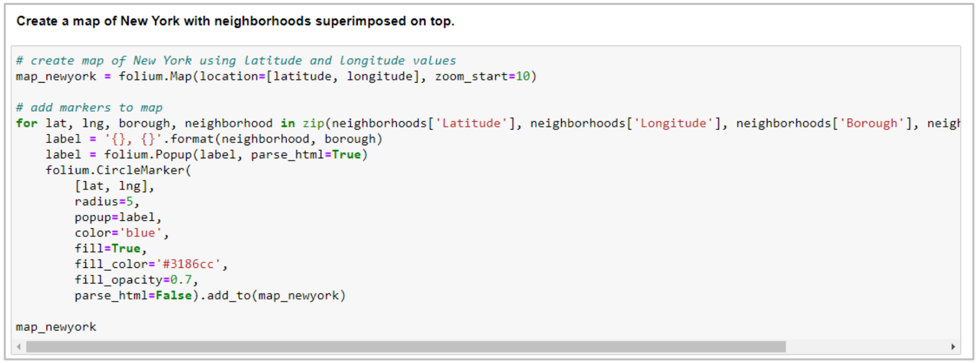
Get Geographical Coordinates of Newyork city

```
address = 'New York City, NY'

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinate of New York City are {}, {}.'.format(latitude, longitude))
```

```
The geograpical coordinate of New York City are 40.7127281, -74.0060152.
```

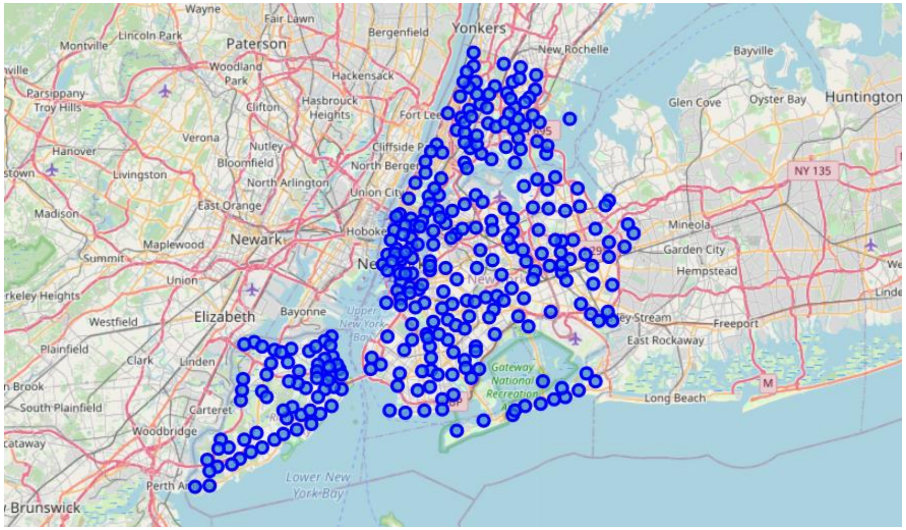
Methodology – Visualize Regional Geospatial Data (Newyork)

Folium library was used to create the map of Newyork and its neighborhoods

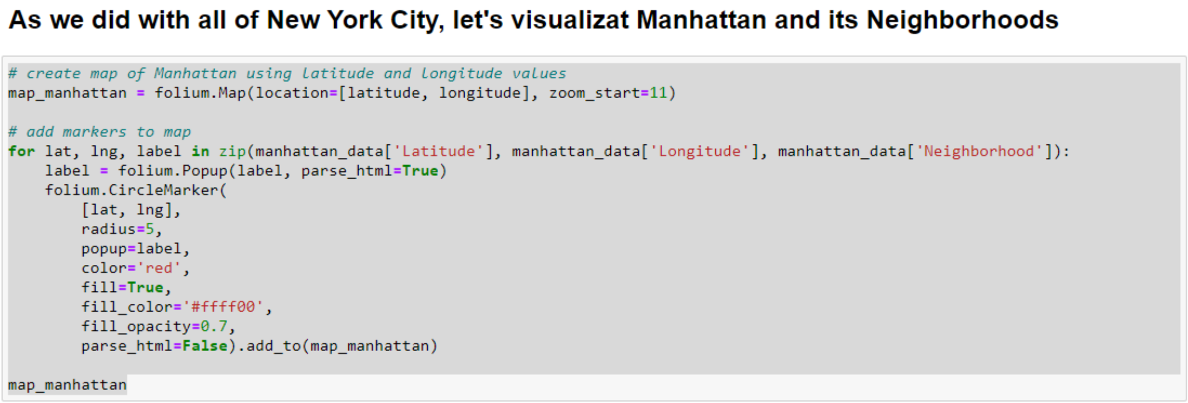


Create a map of New York with neighborhoods superimposed on top.

```
# create map of New York using latitude and longitude values
map_newyork = folium.Map(location=[latitude, longitude], zoom_start=10)

# add markers to map
for lat, lng, borough, neighborhood in zip(neighborhoods['Latitude'], neighborhoods['Longitude'], neighborhoods['Borough'], neigh
    label = '{}, {}'.format(neighborhood, borough)
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='blue',
        fill=True,
        fill_color='#3186cc',
        fill_opacity=0.7,
        parse_html=False).add_to(map_newyork)

map_newyork
```

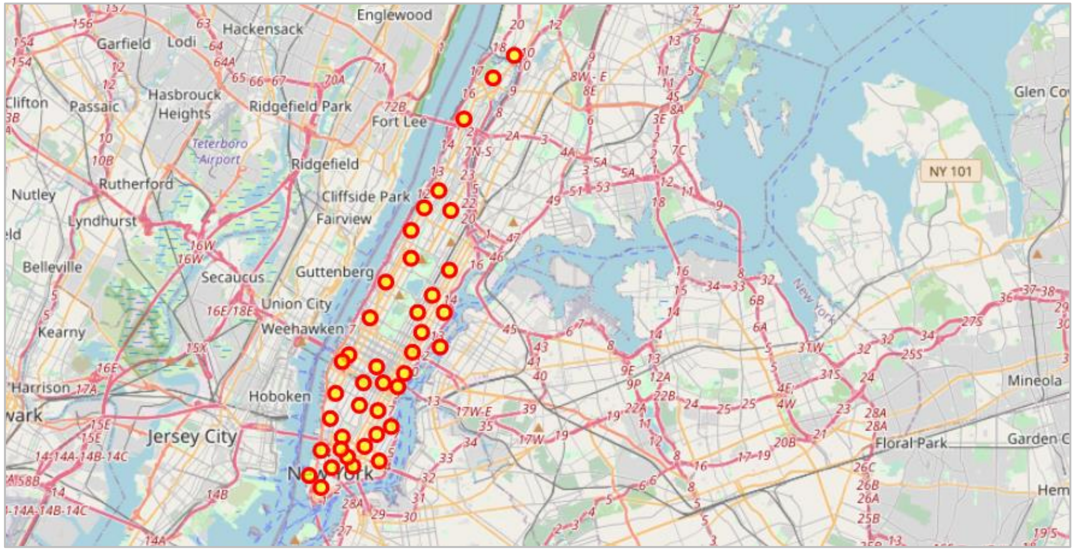


Methodology – Visualize Local Geospatial Data (Manhattan)



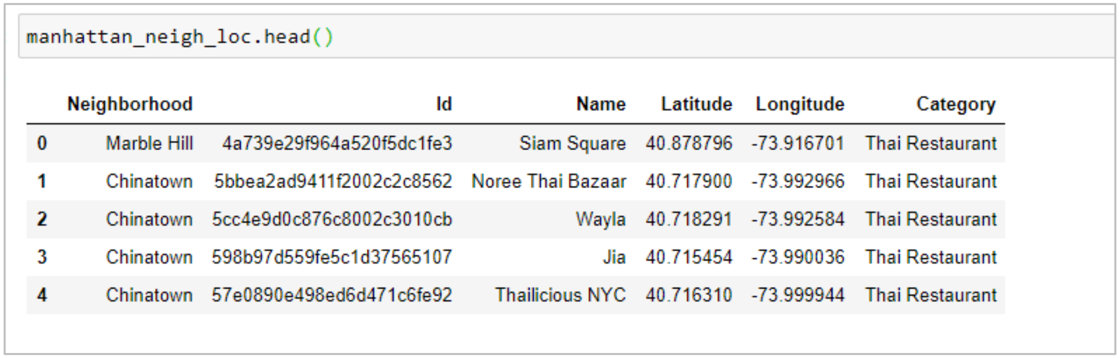
As we did with all of New York City, let's visualizat Manhattan and its Neighborhoods

```
# create map of Manhattan using latitude and longitude values
map_manhattan = folium.Map(location=[latitude, longitude], zoom_start=11)

# add markers to map
for lat, lng, label in zip(manhattan_data['Latitude'], manhattan_data['Longitude'], manhattan_data['Neighborhood']):
    label = folium.Popup(label, parse_html=True)
    folium.CircleMarker(
        [lat, lng],
        radius=5,
        popup=label,
        color='red',
        fill=True,
        fill_color='#ffff00',
        fill_opacity=0.7,
        parse_html=False).add_to(map_manhattan)

map_manhattan
```



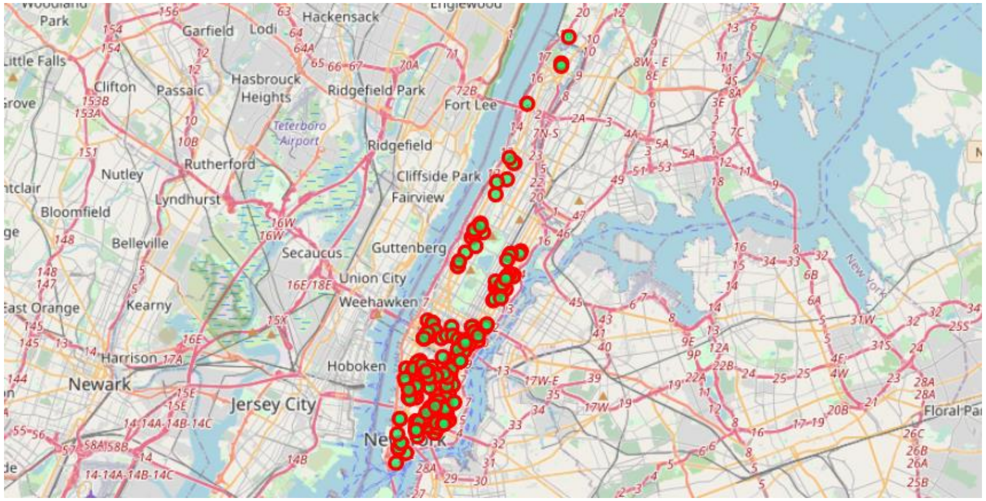
Methodology – Explore the Neighborhood

Data exploration of the neighborhood of the nearby venues of Manhattan was done and data was filtered based on Thai restaurants and data was plotted on the map for the neighborhoods with thai restaurants



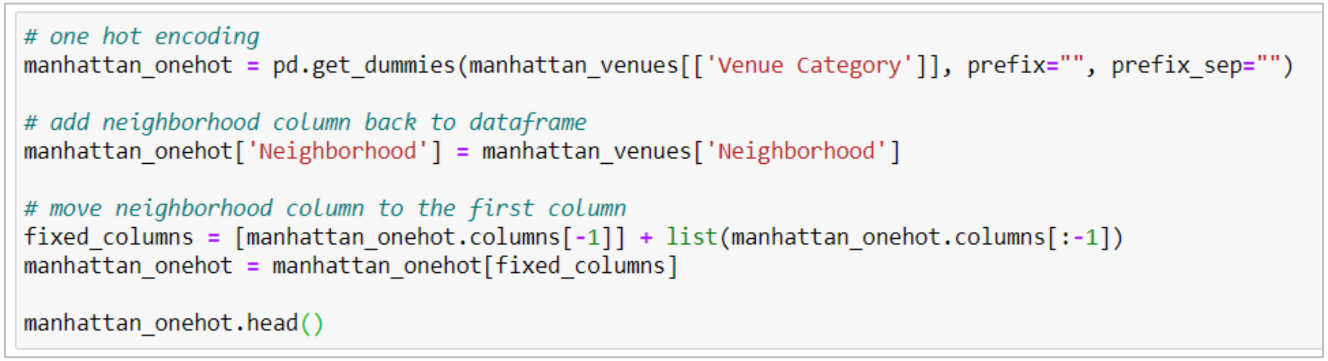
```
manhattan_neigh_loc.head()
```

	Neighborhood	Id	Name	Latitude	Longitude	Category
0	Marble Hill	4a739e29f964a520f5dc1fe3	Siam Square	40.878796	-73.916701	Thai Restaurant
1	Chinatown	5bbea2ad9411f2002c2c8562	Noree Thai Bazaar	40.717900	-73.992966	Thai Restaurant
2	Chinatown	5cc4e9d0c876c8002c3010cb	Wayla	40.718291	-73.992584	Thai Restaurant
3	Chinatown	598b97d559fe5c1d37565107	Jia	40.715454	-73.990036	Thai Restaurant
4	Chinatown	57e0890e498ed6d471c6fe92	Thailicious NYC	40.716310	-73.999944	Thai Restaurant

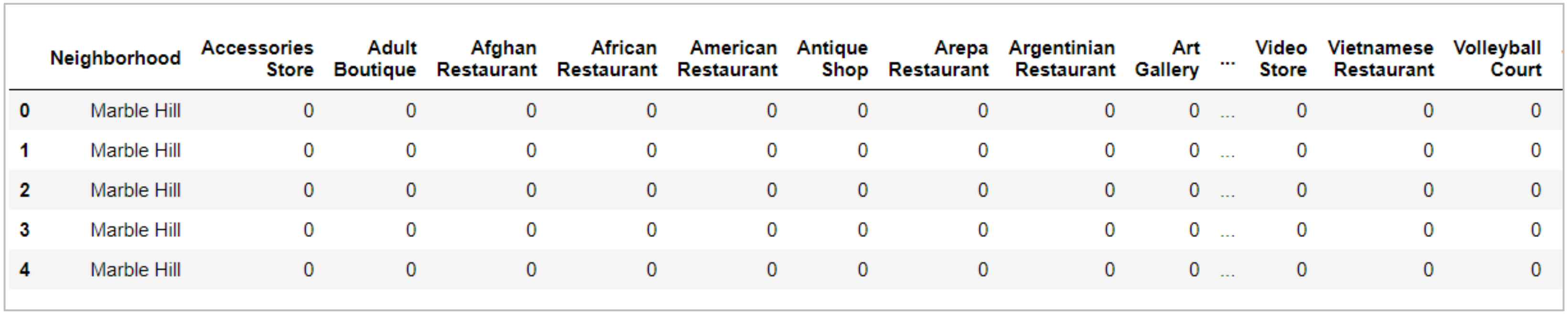


Methodology – Data Preparation for Machine Learning

Machine learning cannot be applied to the categorical data; thus the data needs to be transformed into numerical data using different technique. One such technique is called One Hot Encoding which helped convert the venues occurrence to frequency and how many venues were in each neighborhood



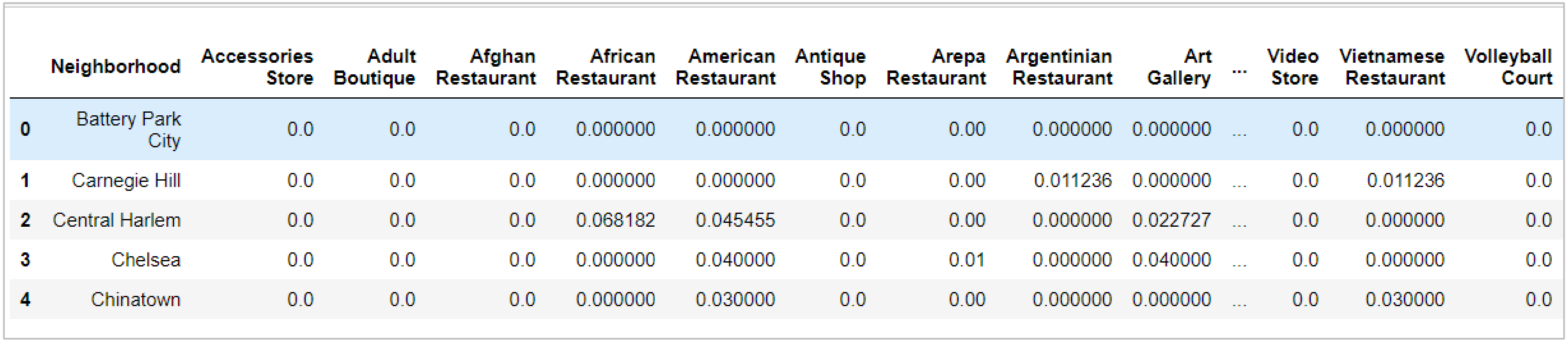
```
# one hot encoding
manhattan_onehot = pd.get_dummies(manhattan_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
manhattan_onehot['Neighborhood'] = manhattan_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [manhattan_onehot.columns[-1]] + list(manhattan_onehot.columns[:-1])
manhattan_onehot = manhattan_onehot[fixed_columns]

manhattan_onehot.head()
```



	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Video Store	Vietnamese Restaurant	Volleyball Court
0	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0
1	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0
2	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0
3	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0
4	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0	0

Methodology – Kmeans Clustering

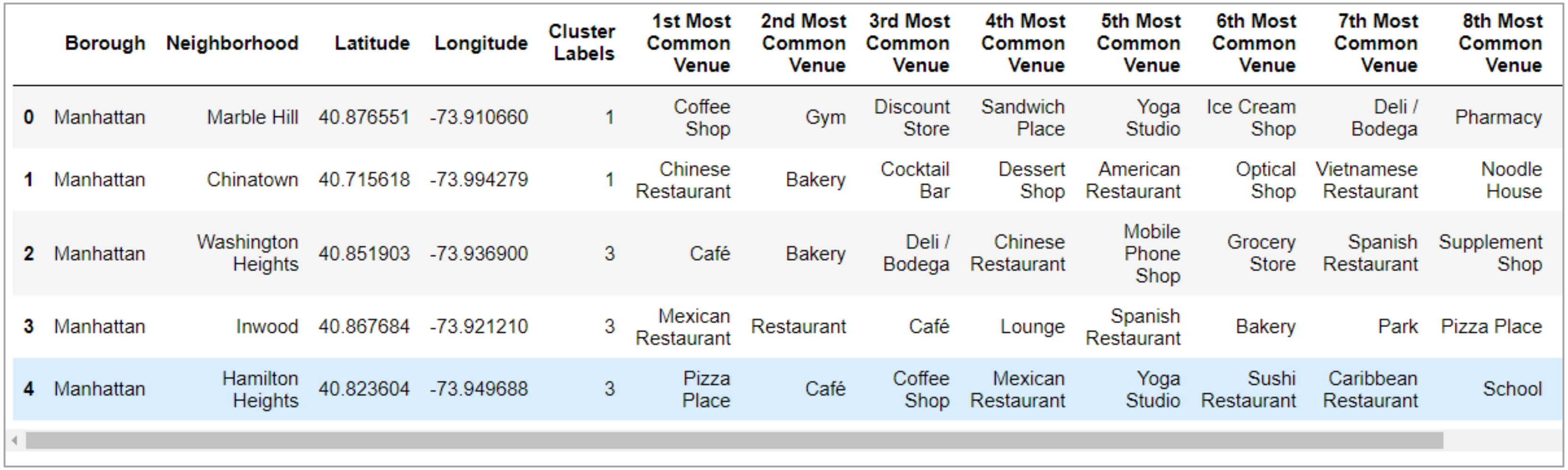
Kmeans clustering is an unsupervised machine learning algorithms that groups together a set of objects in a way that objects in the same cluster are more similar to each other than to objects in other clusters. Kmeans clustering was used to cluster the neighborhoods that has similar averages for Thai restaurants in that particular neighborhood



	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	...	Video Store	Vietnamese Restaurant	Volleyball Court
0	Battery Park City	0.0	0.0	0.0	0.000000	0.000000	0.0	0.00	0.000000	0.000000	...	0.0	0.000000	0.0
1	Carnegie Hill	0.0	0.0	0.0	0.000000	0.000000	0.0	0.00	0.011236	0.000000	...	0.0	0.011236	0.0
2	Central Harlem	0.0	0.0	0.0	0.068182	0.045455	0.0	0.00	0.000000	0.022727	...	0.0	0.000000	0.0
3	Chelsea	0.0	0.0	0.0	0.000000	0.040000	0.0	0.01	0.000000	0.040000	...	0.0	0.000000	0.0
4	Chinatown	0.0	0.0	0.0	0.000000	0.030000	0.0	0.00	0.000000	0.000000	...	0.0	0.030000	0.0

Methodology – Merged Dataframe with Venues and Clusters

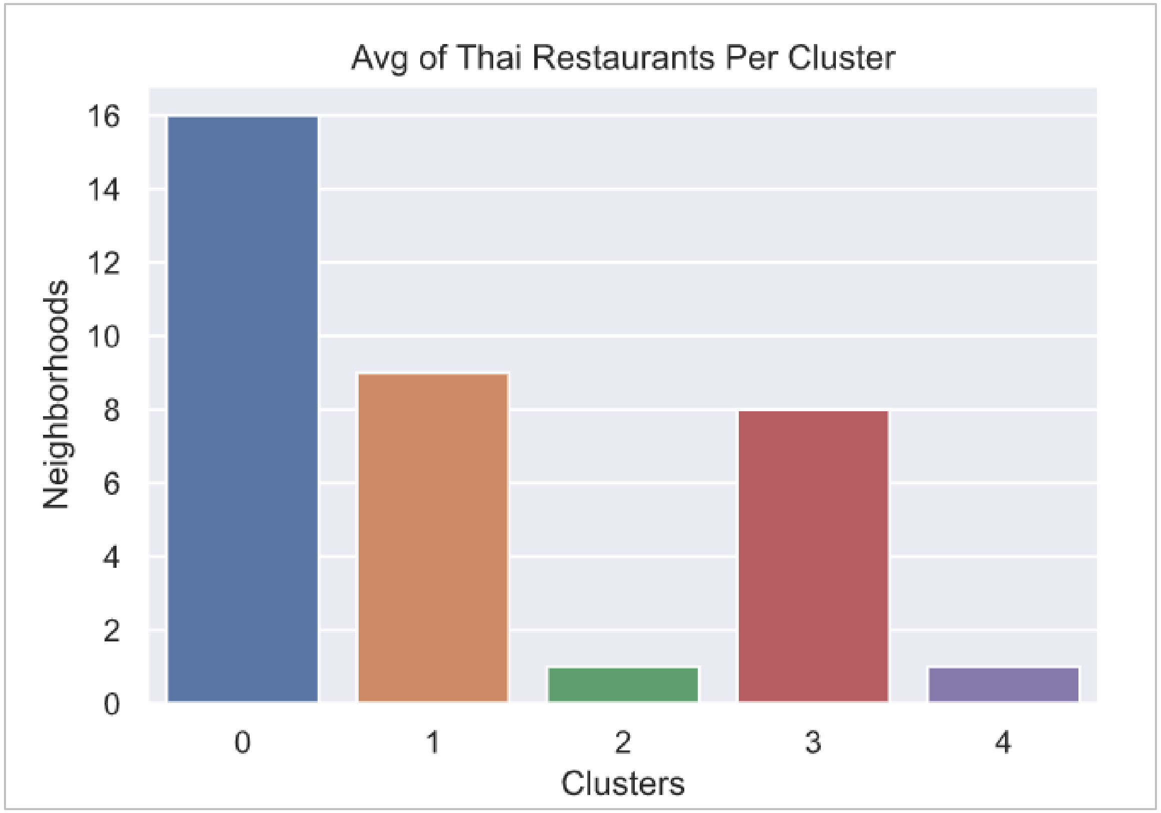
We then created a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood.



	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
0	Manhattan	Marble Hill	40.876551	-73.910660	1	Coffee Shop	Gym	Discount Store	Sandwich Place	Yoga Studio	Ice Cream Shop	Deli / Bodega	Pharmacy
1	Manhattan	Chinatown	40.715618	-73.994279	1	Chinese Restaurant	Bakery	Cocktail Bar	Dessert Shop	American Restaurant	Optical Shop	Vietnamese Restaurant	Noodle House
2	Manhattan	Washington Heights	40.851903	-73.936900	3	Café	Bakery	Deli / Bodega	Chinese Restaurant	Mobile Phone Shop	Grocery Store	Spanish Restaurant	Supplement Shop
3	Manhattan	Inwood	40.867684	-73.921210	3	Mexican Restaurant	Restaurant	Café	Lounge	Spanish Restaurant	Bakery	Park	Pizza Place
4	Manhattan	Hamilton Heights	40.823604	-73.949688	3	Pizza Place	Café	Coffee Shop	Mexican Restaurant	Yoga Studio	Sushi Restaurant	Caribbean Restaurant	School

Methodology – Cluster Visualization

The resulting clusters were then visualized as bar plot using seaborn library to show average number of Thai resturants in each cluster within Manhattan. We can see that most of the Thai restaurants are in cluster 0. Interpreting the plot, we can see that the most optimum cluster to open the thai restaurant are cluster number 2 or number 4. As per initial requirement, the entrepreneur needed to invest safely in an area with less competition.



Results

- Datascience methodology proved effective in recommending areas for opening a Thai Resturant with high probability of good return on investment.
- The exercise showed how data can be scraped from a website and used in python environment for data analysis, visualization and applying machine learning
- Data visualization provided excellent methods of graphically representing the data and using seaborn library and geospatial data was effectively visualized using the folium library.
- Kmeans provided good clustering algorithm for helping to recommend a location for thai Resturant that makes business sense.

Discussion

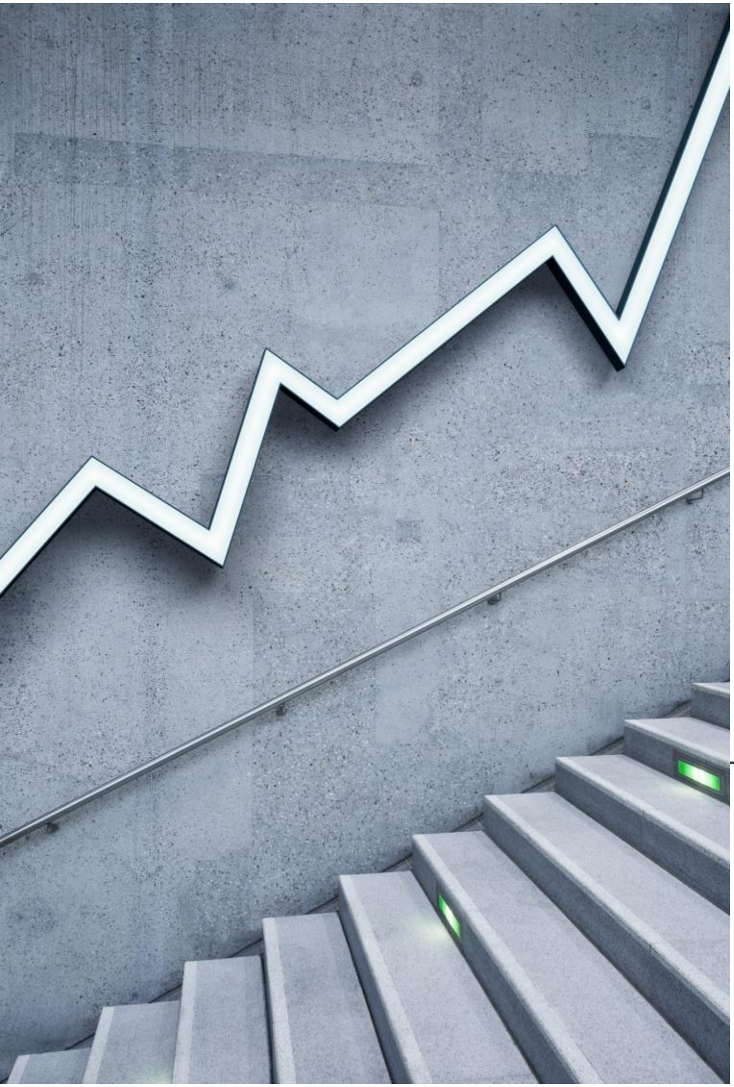
- There is room for the improvement as other features such as restaurant ratings, areas with best tips and user likes from foursquare would provide more data and better clustering based on multi-attribute analysis and clustering.

- The exercise showcased the power of datascience methodology and practice as recommender system and in visualization and data wrangling domains

Conclusion

- The exercise provided good opportunity to help recommend a best place / places for opening a restaurant in Manhattan, Newyork. This methodology can be applied to variety of similar problems requiring clustering and recommendations using unsupervised machine learning.

- We were able to predict the best location to start Thai resturant while ensuring the high rate of return and safe investment.



END

SYED SADAQAT ALI