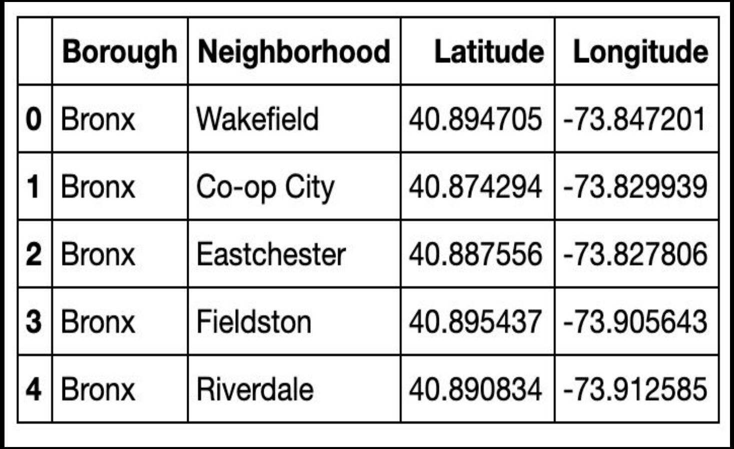
Clusters of New York City

Coursera Capstone Project
By Xuemei Xu

Introduction

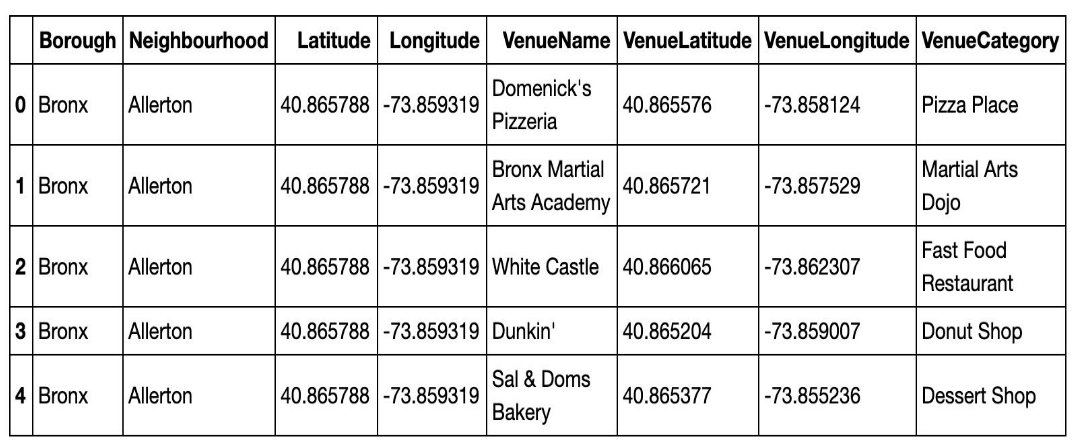
- New York City
  - The largest Population
  - The most diverse
  - The highest GDP
- Start a new business?
  - Understand business environment
- Stakeholders
  - Individuals want to start a small business
  - Organizations to conduct market research
  - Residents
  - Tourists

Data



	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

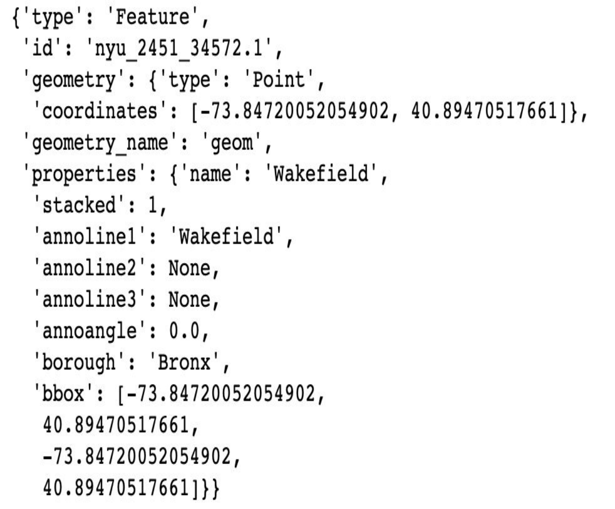
json file: 'newyork_data.json'
https://cocl.us/new_york_dataset



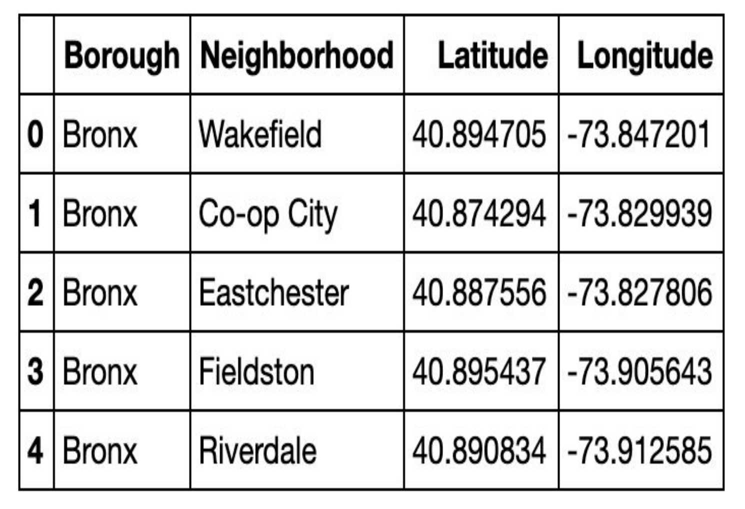
	Borough	Neighbourhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Bronx	Allerton	40.865788	-73.859319	Domenick's Pizzeria	40.865576	-73.858124	Pizza Place
1	Bronx	Allerton	40.865788	-73.859319	Bronx Martial Arts Academy	40.865721	-73.857529	Martial Arts Dojo
2	Bronx	Allerton	40.865788	-73.859319	White Castle	40.866065	-73.862307	Fast Food Restaurant
3	Bronx	Allerton	40.865788	-73.859319	Dunkin'	40.865204	-73.859007	Donut Shop
4	Bronx	Allerton	40.865788	-73.859319	Sal & Doms Bakery	40.865377	-73.855236	Dessert Shop

FourSquare API:
https://foursquare.com/developers/apps

Data Analysis



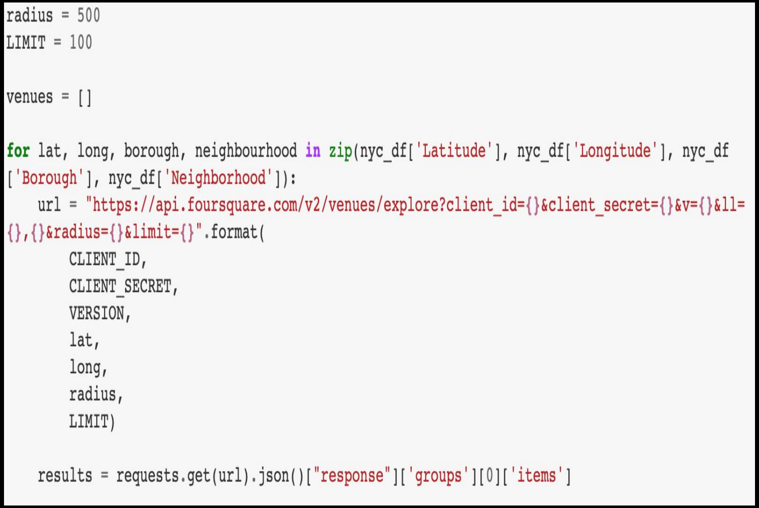
```
{'type': 'Feature',
 'id': 'nyu_2451_34572.1',
 'geometry': {'type': 'Point',
  'coordinates': [-73.84720052054902, 40.89470517661]},
 'geometry_name': 'geom',
 'properties': {'name': 'Wakefield',
  'stacked': 1,
  'annoline1': 'Wakefield',
  'annoline2': None,
  'annoline3': None,
  'annoangle': 0.0,
  'borough': 'Bronx',
  'bbox': [-73.84720052054902,
   40.89470517661,
   -73.84720052054902,
   40.89470517661]}}
```



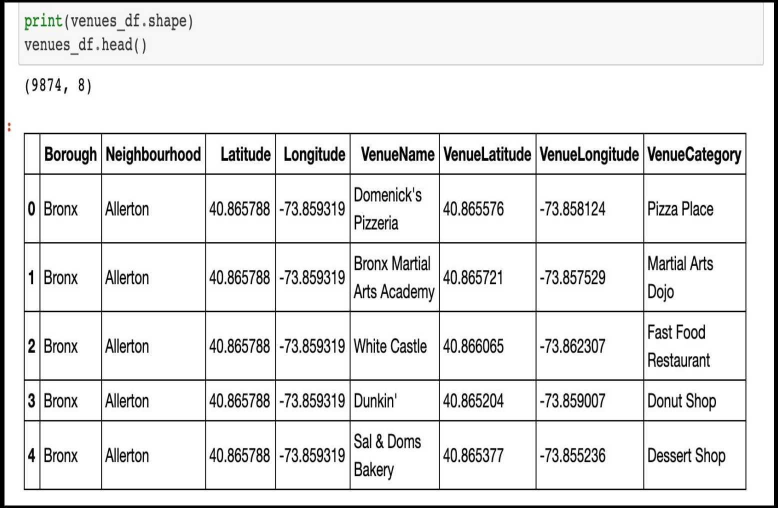
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Transforming dataset from json file to pandas dataframe for easier analysis.

Data Analysis (Cont.)



```
radius = 500
LIMIT = 100

venues = []

for lat, long, borough, neighbourhood in zip(nyc_df['Latitude'], nyc_df['Longitude'], nyc_df
['Borough'], nyc_df['Neighborhood']):
    url = "https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&v={}&ll=
{},{}&radius={}&limit={}".format(
        CLIENT_ID,
        CLIENT_SECRET,
        VERSION,
        lat,
        long,
        radius,
        LIMIT)

    results = requests.get(url).json()["response"]['groups'][0]['items']
```



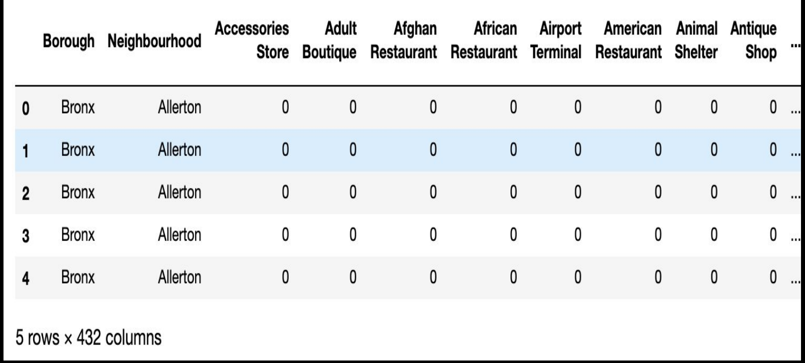
```
print(venues_df.shape)
venues_df.head()
```

```
(9874, 8)
```

	Borough	Neighbourhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Bronx	Allerton	40.865788	-73.859319	Domenick's Pizzeria	40.865576	-73.858124	Pizza Place
1	Bronx	Allerton	40.865788	-73.859319	Bronx Martial Arts Academy	40.865721	-73.857529	Martial Arts Dojo
2	Bronx	Allerton	40.865788	-73.859319	White Castle	40.866065	-73.862307	Fast Food Restaurant
3	Bronx	Allerton	40.865788	-73.859319	Dunkin'	40.865204	-73.859007	Donut Shop
4	Bronx	Allerton	40.865788	-73.859319	Sal & Doms Bakery	40.865377	-73.855236	Dessert Shop

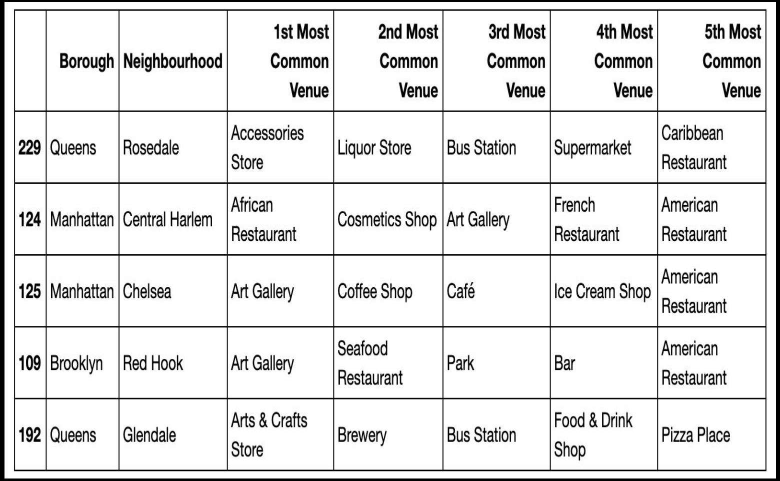
Calling explore results from FourSquare API and retrieve necessary information only to analyze.

Data Analysis (Cont.)



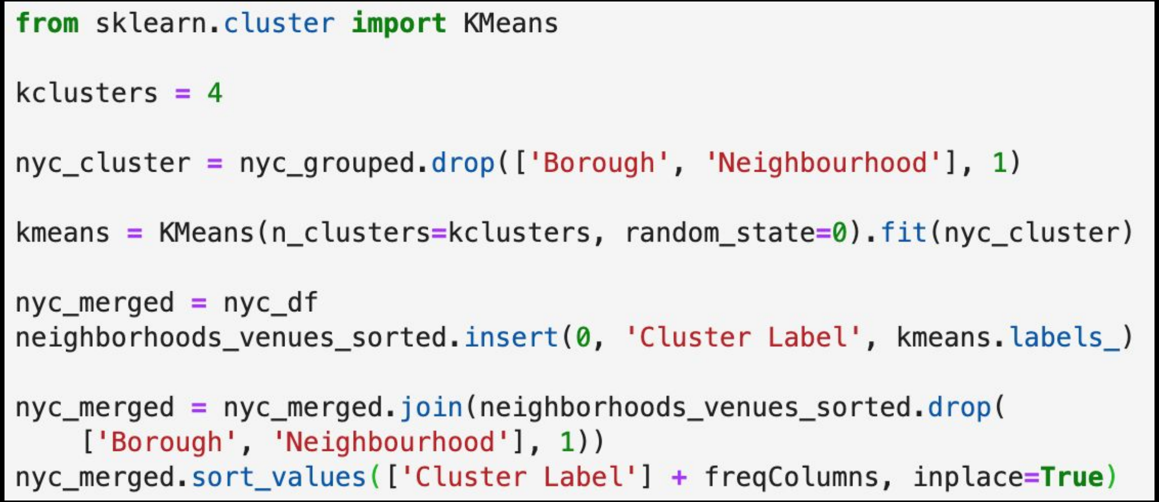
	Borough	Neighbourhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport Terminal	American Restaurant	Animal Shelter	Antique Shop	...
0	Bronx	Allerton	0	0	0	0	0	0	0	0	...
1	Bronx	Allerton	0	0	0	0	0	0	0	0	...
2	Bronx	Allerton	0	0	0	0	0	0	0	0	...
3	Bronx	Allerton	0	0	0	0	0	0	0	0	...
4	Bronx	Allerton	0	0	0	0	0	0	0	0	...

5 rows × 432 columns



	Borough	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
229	Queens	Rosedale	Accessories Store	Liquor Store	Bus Station	Supermarket	Caribbean Restaurant
124	Manhattan	Central Harlem	African Restaurant	Cosmetics Shop	Art Gallery	French Restaurant	American Restaurant
125	Manhattan	Chelsea	Art Gallery	Coffee Shop	Café	Ice Cream Shop	American Restaurant
109	Brooklyn	Red Hook	Art Gallery	Seafood Restaurant	Park	Bar	American Restaurant
192	Queens	Glendale	Arts & Crafts Store	Brewery	Bus Station	Food & Drink Shop	Pizza Place

Machine Learning (K-means)



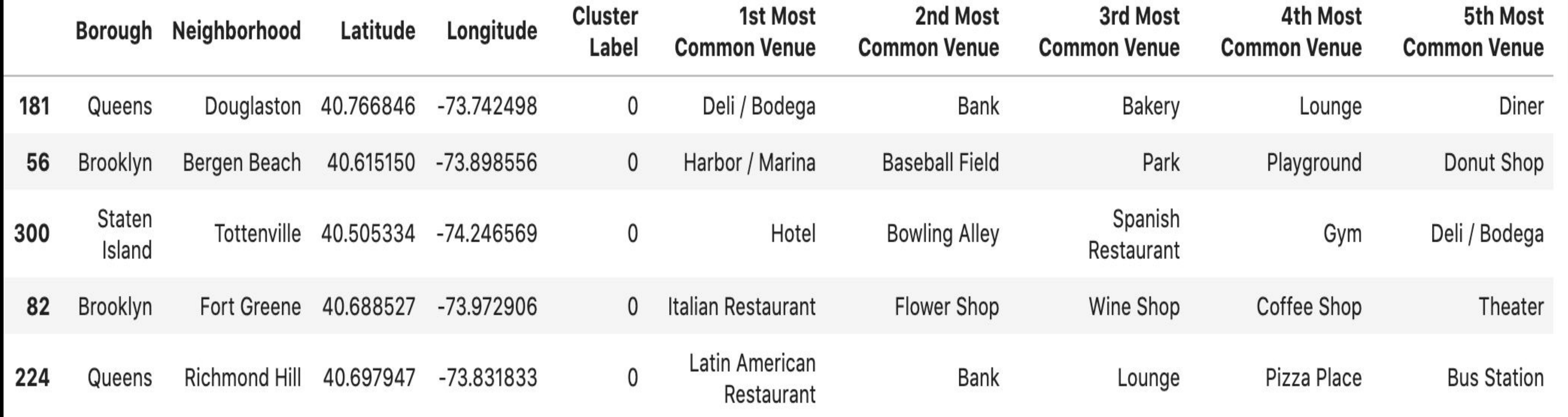
```
from sklearn.cluster import KMeans

kclusters = 4

nyc_cluster = nyc_grouped.drop(['Borough', 'Neighbourhood'], 1)

kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(nyc_cluster)

nyc_merged = nyc_df
neighborhoods_venues_sorted.insert(0, 'Cluster Label', kmeans.labels_)

nyc_merged = nyc_merged.join(neighborhoods_venues_sorted.drop(
    ['Borough', 'Neighbourhood'], 1))
nyc_merged.sort_values(['Cluster Label'] + freqColumns, inplace=True)
```

Why K-means?

❑ Unsupervised machine learning algorithm;

❑ Segment dataset into several clusters;

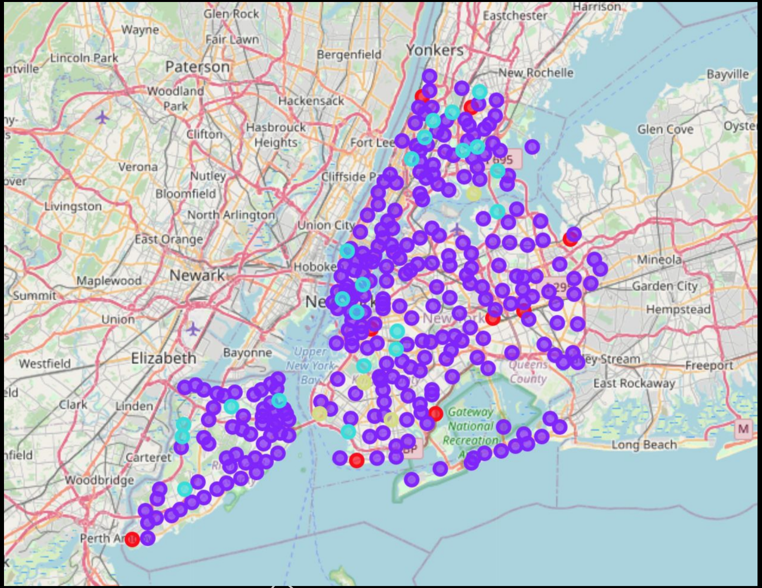
Machine Learning (K-means, Cont.)



	Borough	Neighborhood	Latitude	Longitude	Cluster Label	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
181	Queens	Douglaston	40.766846	-73.742498	0	Deli / Bodega	Bank	Bakery	Lounge	Diner
56	Brooklyn	Bergen Beach	40.615150	-73.898556	0	Harbor / Marina	Baseball Field	Park	Playground	Donut Shop
300	Staten Island	Tottenville	40.505334	-74.246569	0	Hotel	Bowling Alley	Spanish Restaurant	Gym	Deli / Bodega
82	Brooklyn	Fort Greene	40.688527	-73.972906	0	Italian Restaurant	Flower Shop	Wine Shop	Coffee Shop	Theater
224	Queens	Richmond Hill	40.697947	-73.831833	0	Latin American Restaurant	Bank	Lounge	Pizza Place	Bus Station

Due to high volume of the results, retrieve top 5 frequently occured venues.

Results



Four Clusters:

- ❑ Cluster 0 (Red) : entertainment areas
- ❑ Cluster 1 (Purple): living and business areas
- ❑ Cluster 2 (Light Blue): food places
- ❑ Cluster 3 (Yellow): fitness centers

Discussion

Observation:

- Cluster 0: located in the border of each borough;
- Cluster 1: evenly spreaded in five boroughs;
- Cluster 2: Queens borough has no specific neighborhood known for food places;
- Cluster 3: neighborhoods of fitness centers are only located in Brooklyn borough;

Recommendation:

- Cluster 0: great neighborhoods for residents, tourists and businesses relating to entertainment.
- Cluster 1: start businesses after understanding active people preferences in each neighborhood;
- Cluster 2: good for residents and tourists to look for food places;
- Cluster 3: good for competitive vendors to start fitness related businesses;

Conclusion

❏ The neighborhoods in cluster 1 are combined with food places, entertainment places, fitness centers, grocery stores, etc.

❏ In Cluster 1, each category in each category in each neighborhood is serving its specific customer segments.

❏ In the other three clusters, there are three corresponding businesses that can be practiced and get results of a high possibility of success.