BATTLE OF NEIGHBORHOODS

New York City

By Yadan Tang

Problem Statement

- The Catalysis Society of Metropolitan New York (CSMNY), a local chapter of North American Catalysis Society(NACS), need to find a location to host their annual meeting in Manhattan, New York City (NYC) in 2020. The number of attendees for this event is estimated to ~200 people and of various cultural background. The attendees are coming for professional exchange as well as sightseeing and leisure.
- The host needs a recommendation of a suitable neighborhood in Manhattan NYC with the capacity of hotels, and various options of restaurants and activities for such an event.

Outlines

- Data acquisition and process the Neighborhood data in Manhattan, New York City.
- Explore the venues of all the neighborhoods in Manhattan, New York City.
- Using K-mean Clustering to segment the neighborhoods
- Conclusions based on the K-mean Clustering results

Data acquisition and Processing

- Data sources: the json data that contains all the information of NYC neighborhoods is available online: https://cocl.us/new_york_dataset
- The neighborhood data is available under the list [‘features’], and it was converted into dataframe for further processing
- Create a dataframe contains the ‘Borough’, ‘Neighborhood’, ’Latitude’, ‘Longitude’ columns.

```
# instantiate the dataframe
neighborhoods = pd.DataFrame(columns=column_names)
```

```
[7]: for data in neighbor_data:
        borough = neighborhood_name = data['properties']['borough']
        neighborhood_name = data['properties']['name']

        neighborhood_latlon = data['geometry']['coordinates']
        neighborhood_lat = neighborhood_latlon[1]
        neighborhood_lon = neighborhood_latlon[0]

        neighborhoods = neighborhoods.append({'Borough': borough,
                                              'Neighborhood': neighborhood_name,
                                              'Latitude': neighborhood_lat,
                                              'Longitude': neighborhood_lon}, ignore_index=True)
```

- Double check that the dataframe contains 306 neighborhoods and 5 unique Boroughs, which is accurate.

Data acquisition and Processing con’t

- Use ‘groupby’ function and ‘Borough’ as the argument and count() to see how many neighborhoods are in each Borough. The result showed Manhattan has 40 neighborhoods.

```
[12]: neighborhoods.groupby('Borough').count()
```

[12]:

Borough	Neighborhood	Latitude	Longitude
Bronx	52	52	52
Brooklyn	70	70	70
Manhattan	40	40	40
Queens	81	81	81
Staten Island	63	63	63

- Create a dataframe that contains all the neighborhood information of Manhattan.

```
[13]: mht_data=neighborhoods[neighborhoods['Borough']=='Manhattan'].reset_index(drop=True)
mht_data.head()
```

[13]:

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

Connect to Foursquare API and explore the nearby venues

- Connect to Foursquare API using personal Foursquare ID and secrets
- Define a function to get all the nearby venues in Manhattan, NYC. Return only the relevant information of each nearby venue and save it into a dataframe called “nearby_venues” which includes the name and the coordinates of the neighborhoods and its nearby venues.
- Group the dataframe by venue category, the count show there are 331 unique venue categories.
- To create columns of venues for each neighborhood by one hot encoding and then group it by neighborhood and show the mean counts of each venue in each neighborhood.

Connect t0 Foursquare API and explore the nearby venues

- Show the top 10 venues in each Neighborhood

[51]:
```
# To list the top 10 venues for each Neighborhood in Manhattan
import numpy as np
top_venue_number = 10

indicators = ['st', 'nd', 'rd']

# create columns according to number of top venues
columns = ['Neighborhood']
for ind in np.arange(top_venue_number):
    try:
        columns.append('{}{} Most Common Venue'.format(ind+1, indicators[ind]))
    except:
        columns.append('{}th Most Common Venue'.format(ind+1))

# create a new dataframe
neighborhoods_venues_sorted = pd.DataFrame(columns=columns)
neighborhoods_venues_sorted['Neighborhood'] = mht_grouped['Neighborhood']

for ind in np.arange(mht_grouped.shape[0]):
    neighborhoods_venues_sorted.iloc[ind, 1:] = topVenues(mht_grouped.iloc[ind, :], top_venue_number)

print(neighborhoods_venues_sorted.shape[0])
neighborhoods_venues_sorted.head()
```

```
40
```

[51]:

	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	Battery Park City	Park	Coffee Shop	Hotel	Memorial Site	Gym	Boat or Ferry	Plaza	Gourmet Shop	Food Court	Shopping Mall
				Italian				Gym / Fitness			Vietnamese

K-means Clustering: segment the neighborhood of Manhattan

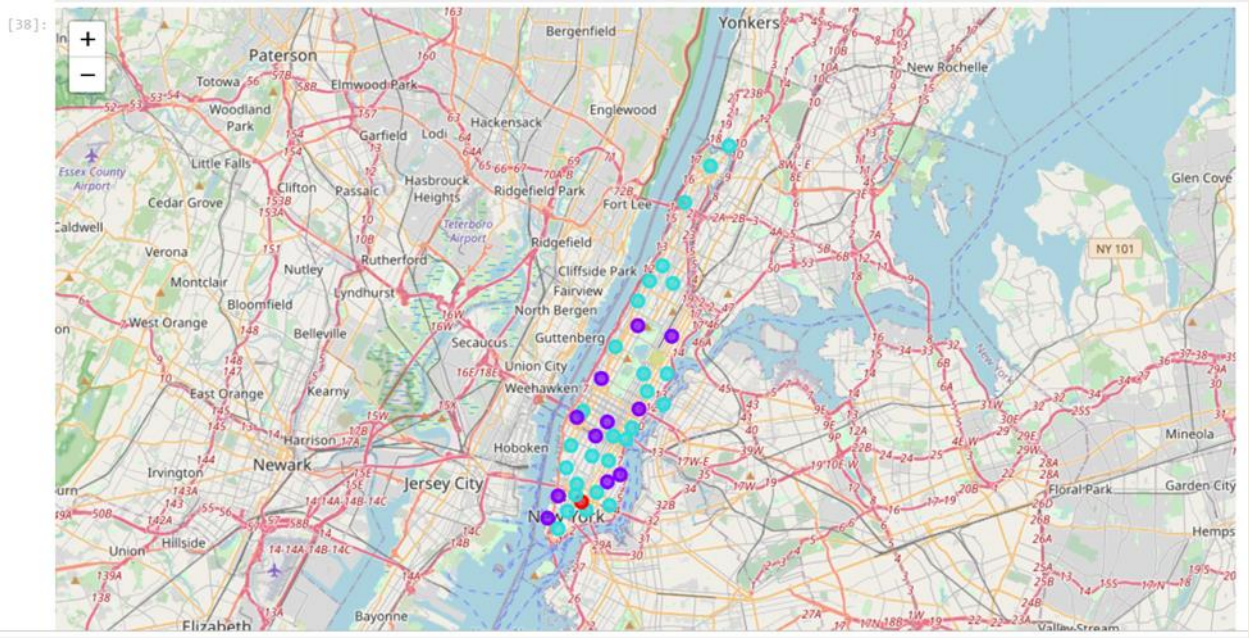
- The initial K-mean clustering is set to 4.

```
[23]: # import k-means from clustering stage
      from sklearn.cluster import KMeans
```

Use kcluster size of 4 to start with and create the labels and then add the labels as a new column back to the dataframe

```
[30]: kclusters=4

      mht_cluster=mht_grouped.drop('Neighborhood',1)

      kMeans=KMeans(n_clusters=kclusters, random_state=5).fit(mht_cluster)

      kMeans.labels_

[30]: array([2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 2, 1,
             0, 2, 2, 1, 2, 2, 1, 2, 3, 2, 2, 1, 1, 2, 2, 1, 2, 1], dtype=int32)
```

- Visualize the clustering result using folium and the clusters were classified and labelled in different colors and added to the map of Manhattan, NYC.



Explore the information in each cluster and find out the best options for the event

- Create dataframe for each cluster labels, showing one as an example:

```
[39]: cluster_1=mht_merged.loc[mht_merged['Labels']==1, mht_merged.columns[[1] + list(range(4, mht_merged.shape[1]))]]
cluster_1.head(10)
```

[39]:

	Neighborhood	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
7	East Harlem	1	Mexican Restaurant	Thai Restaurant	Deli / Bodega	Bakery	Latin American Restaurant	Sandwich Place	Beer Bar	Taco Place	Liquor Store	French Restaurant
13	Lincoln Square	1	Café	Plaza	Concert Hall	Performing Arts Venue	Italian Restaurant	Theater	Gym / Fitness Center	Gym	American Restaurant	Coffee Shop
15	Midtown	1	Coffee Shop	Hotel	Theater	Clothing Store	Bakery	Sandwich Place	Pizza Place	Tailor Shop	Sushi Restaurant	Steakhouse
19	East Village	1	Bar	Pizza Place	Mexican Restaurant	Cocktail Bar	Korean Restaurant	Wine Bar	Coffee Shop	Dessert Shop	Ice Cream Shop	Bagel Shop
21	Tribeca	1	Italian Restaurant	Park	American Restaurant	Wine Bar	Spa	Greek Restaurant	Coffee Shop	Café	Skate Park	Burger Joint
25	Manhattan Valley	1	Coffee Shop	Bar	Mexican Restaurant	Yoga Studio	Pizza Place	Bubble Tea Shop	Café	Peruvian Restaurant	Park	Arts & Crafts Store
28	Battery Park City	1	Park	Hotel	Coffee Shop	Gym	Memorial Site	Boat or Ferry	Gourmet Shop	Burger Joint	Food Court	Plaza
33	Midtown South	1	Korean Restaurant	Hotel	Japanese Restaurant	Dessert Shop	Gym / Fitness Center	Burger Joint	Café	Pizza Place	Flower Shop	Scenic Lookout

- Cluster_1 showed neighborhoods that has a good variety of restaurants/deli, gym/yoga studio, and hotels, plaza. Cluster_2 has good options of restaurants and activities but hotel is not listed as the top 10 venues.
- Cluster_1 are excellent options for the event and especially Battery Park City, Midtown, Midtown South and Murray Hill will have Hotels as the top 2 popular venue and still have a good combination of multicultural restaurants, deli/café, gyms and plaza.

Conclusions

- Based on the exercise, we have successfully explored the neighborhood of Manhattan, NYC and the nearby venues of each neighborhood. K-mean clustering helps to identify the neighborhoods in Cluster_1 are excellent options for the event and especially Battery Park City, Midtown, Midtown South and Murray Hill will have Hotels as the top 2 popular venue and still have a good combination of multicultural restaurants, deli/café, gyms and plaza. The host will be able to accommodate their attendees with good choice of hotels and enough entertainments for after-meeting activities.