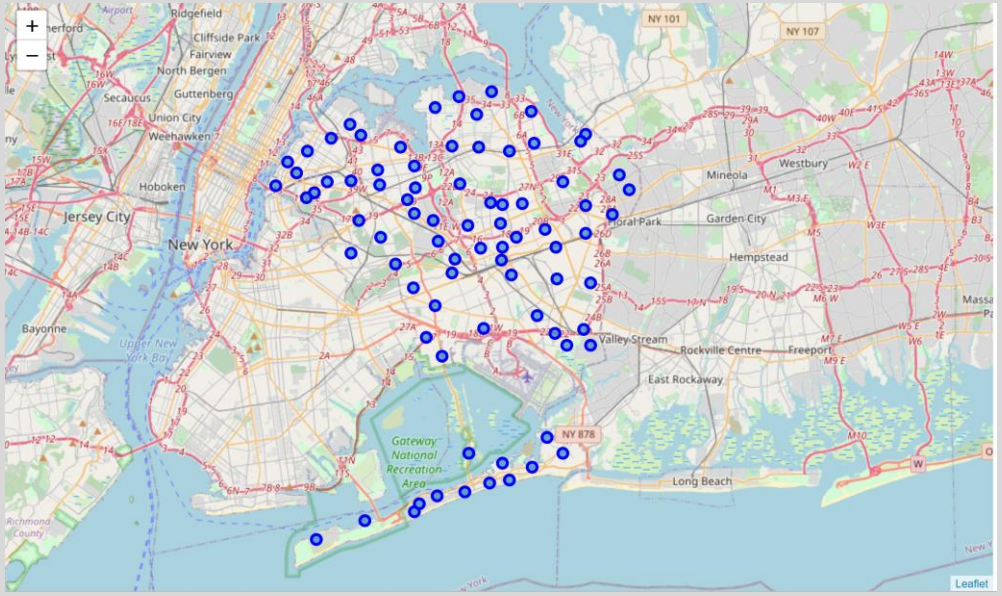

IDENTIFYING A LOCATION FOR A NEW RESTAURANT IN NEW YORK (QUEENS)

Identifying the best neighborhood(s) to open a new restaurant reduces risk

- The restaurant industry is a high-risk business.
- It helps to place the restaurant in a neighborhood with high visibility and traffic but low eateries saturation
- The data required includes New York neighborhood data (https://cocl.us/new_york_dataset) and Foursquare data on eateries in the neighborhoods.
- With the cluster analysis, the neighborhood(s) with a saturated eatery market are expected to appear together in a cluster while the others will offer possible options for the right neighborhood(s) for the new restaurant.



Data Sources

- The NYU Spatial Data Repository's 2014 New York City Neighborhood Names data has information on the 5 borough and 306 neighborhood that make up the city.
  - It includes the longitudes and latitudes of each of the neighborhoods.
- With the Foursquare APIs, the neighborhoods coordinates will be used to identify surrounding venues in these neighborhoods.
  - The venues include details on the categories which will inform whether a neighborhood is highly trafficked but not saturated by restaurants.

	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

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Astoria	40.768509	-73.915654	Favela Grill	40.767348	-73.917897	Brazilian Restaurant
1	Astoria	40.768509	-73.915654	Orange Blossom	40.769856	-73.917012	Gourmet Shop
2	Astoria	40.768509	-73.915654	Simply Fit Astoria	40.769114	-73.912403	Gym
3	Astoria	40.768509	-73.915654	Titan Foods Inc.	40.769198	-73.919253	Gourmet Shop
4	Astoria	40.768509	-73.915654	CrossFit Queens	40.769404	-73.918977	Gym

Data Cleaning

- New York City neighborhoods data as of 2014.
  - 'Features' category
  - Borough, Neighborhood, Longitude, and Latitude
  - **Five** boroughs and **306** neighborhoods.
- Filtered dataframe to only include neighborhoods in **Queens**.
- The data was clean without missing values or concerns about outlier and inconsistencies.
- The Foursquare data on the surrounding neighborhoods was extracted using an API call on the coordinates
  - Avoided overloading the call request (radius=500, limit=100)
  - Grouped by the respective neighborhoods
  - Used one hot encoding. The categories were indexed by the neighborhood so that each venue category was displayed as a column and each neighborhood as row
  - Transformed data to show the mean of the frequency of the occurrence of each venue category per neighborhood
- Both the spatial data on Queens and the Foursquare data on the nearest and most common venues were joined for the clustering process.

	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	Arverne	Surf Spot	Sandwich Place	Metro Station	Donut Shop	Beach	Thai Restaurant	Coffee Shop	Restaurant	Café	Board Shop
1	Astoria	Bar	Middle Eastern Restaurant	Greek Restaurant	Seafood Restaurant	Hookah Bar	Pizza Place	Indian Restaurant	Mediterranean Restaurant	Bakery	Bagel Shop
2	Astoria Heights	Italian Restaurant	Plaza	Playground	Hostel	Pizza Place	Bus Station	Motel	Laundromat	Burger Joint	Bowling Alley
3	Auburndale	Italian Restaurant	Mattress Store	Train	Fast Food Restaurant	Furniture / Home Store	Supermarket	Noodle House	Sushi Restaurant	Korean Restaurant	Miscellaneous Shop
4	Bay Terrace	Clothing Store	Women's Store	Cosmetics Shop	Bus Stop	Men's Store	Mobile Phone Shop	Kids Store	Bank	Donut Shop	Shoe Store

Methodology

- Following the data cleaning there were **40** neighborhoods and **330** venues ready for clustering.
  - k-means from the clustering stage (from sklearn.cluster import KMeans)
  - Initial run, the k-means was set to five to get five different clusters. Each neighborhood is assigned a cluster of zero to four to determine its cluster group.
- The cluster results with the venues is then joined to the neighborhood data which includes the information on the borough and the coordinates.

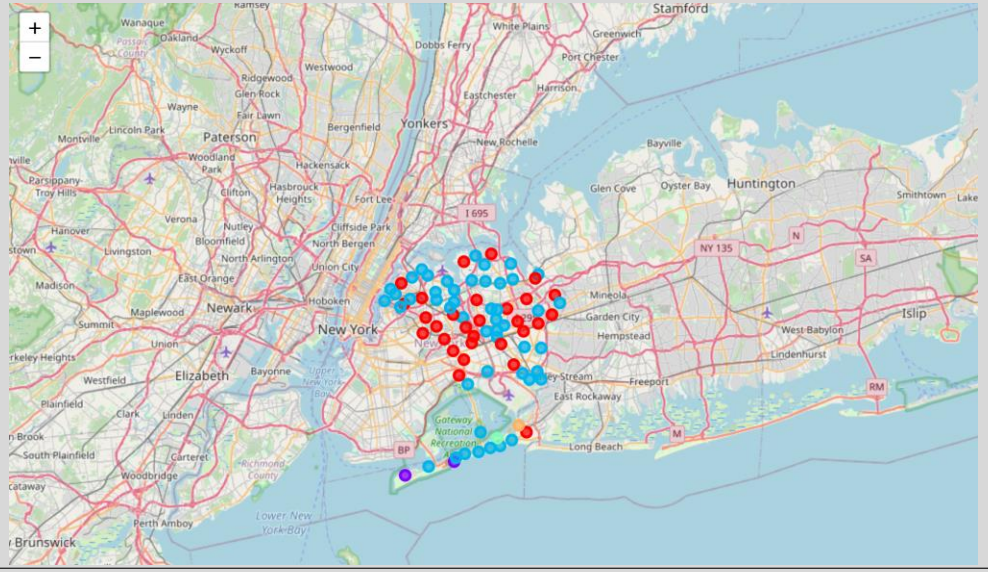
```
# set number of clusters
kclusters = 5

queens_grouped_clustering = queens_grouped.drop('Neighborhood', 1)

# run k-means clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(queens_grouped_clustering)

# check cluster labels generated for each row in the dataframe
kmeans.labels_[0:10]
```

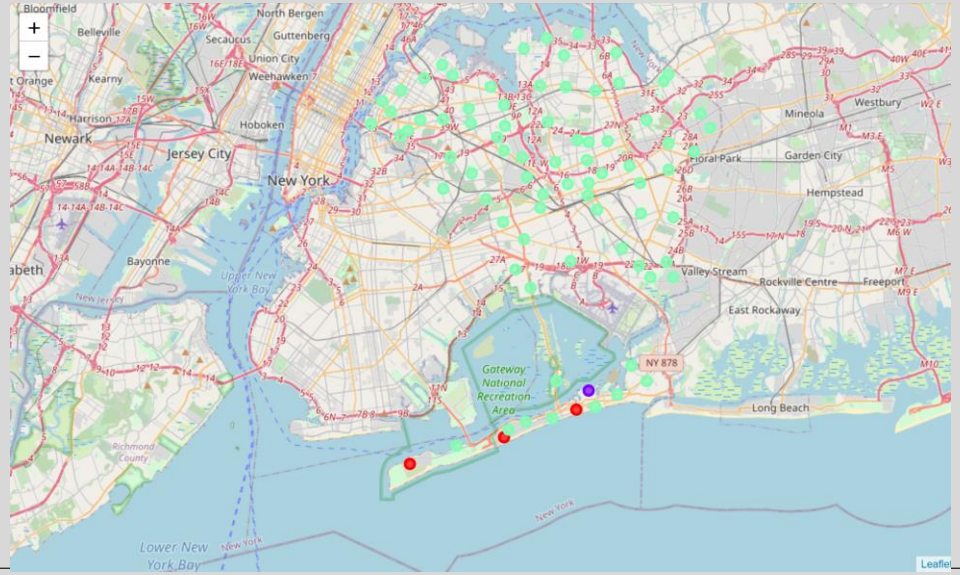
```
array([2, 2, 2, 2, 2, 2, 2, 2, 2, 2], dtype=int32)
```

```
# add clustering labels
neighborhoods_venues_sorted.insert(0, 'Cluster Labels', kmeans.labels_)

queens_merged = queens_data

# merge to add latitude/longitude for each neighborhood
queens_merged = queens_merged.join(neighborhoods_venues_sorted.set_index('Neighborhood'), on='Neighborhood')

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

Result

- The examination revealed inconclusive clusters for determining the right neighborhood(s)
  - Too many clusters for the neighborhoods and therefore the k-means should be reduced
  - Objective was to have a stand out cluster that was not like the rest
  - k-means was updated to three for the development of three clusters
- The three clusters included the following:
  - Cluster One: Breezy Point, Neponsit, and Hammels neighborhoods. Most common venue was the beach.
  - Cluster Two: Only includes the Somerville neighborhood with the park as the most common venue.
  - Cluster Three: High concentration of restaurants which include Mexican, Italian, Thai and Latin American.





Discussion

- Each of the three clusters have eateries represented in the top five most common venues which is a positive representation of the Queens area as a welcoming market for eateries. However, the distinguishing factor across the clusters is the type of eateries and surrounding non-eatery venues.
  - Cluster One: Low saturation of eateries but highly trafficked
  - Cluster Two: Medium concentration of restaurants
  - Cluster Three: High concentration of eateries
- Low proportion of eateries to non-eateries shows a great opportunity to distinguish the new restaurant from the other eateries in the area
- Considering that the client will be opening a Ghanaian-inspired restaurant having a neighborhood with few non-American options will make it a standout in the neighborhood.

```
queens_merged.loc[queens_merged['Cluster Labels'] == 0, queens_merged.columns[[1] + list(range(5, queens_merged.shape[1]))]]
```

	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
43	Breezy Point	Beach	Monument / Landmark	Trail	Empanada Restaurant	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Filipino Restaurant	Fish & Chips Shop
50	Neponsit	Beach	Lounge	Women's Store	Fish & Chips Shop	Farm	Farmers Market	Fast Food Restaurant	Filipino Restaurant	Fish Market	Event Space
78	Hammels	Beach	Food Truck	Deli / Bodega	Fried Chicken Joint	Gym / Fitness Center	Diner	Bus Station	Bus Stop	Dog Run	Shoe Store

Conclusion

- In this study, we analyzed New York city neighborhoods, specifically Queens, to determine the best neighborhoods to open a brand-new Ghanaian-inspired restaurant.
- We took a look at the neighborhood data to determine which neighborhoods made up the Queens borough and mapped out their respective coordinate to visualize their locations.
- We then transformed the Foursquare data into a dataframe which included the top ten venues across each of these neighborhoods.
- We then run a k-means cluster analysis to group the neighborhood into what turned out to be three informative clusters.
- The cluster chosen, cluster one, will be presented to the client to help her determine where in Queens to place the first of many restaurants.
- The clustering model can be replicated for any other restauranteurs interested in opening a restaurant in New York city.