Coursera Capstone Project:

The Battle of Neighborhoods
Final Report (Week 1 and 2)

By Murt Sayeed

Introduction Section

Background:

In United States, the City of New York is the most populous city with its diverse and financial district reputation. It is also very multicultural with many business opportunities and attracts many different players into the market. It is a global capital of everything, such as banking & finance, retailing, world trade, transportation, tourism, real estate, new media, traditional media, advertising, legal services, accountancy, insurance, theater, fashion, and the arts in the United States. The city is seen as most diverse city on the planet, having close to 9 million people and over 800 languages.

The city can also be very competitive. As it is highly developed city so cost of doing business is also one of the highest. Thus, any new business venture or expansion needs to be analyzed carefully. The insights derived from analysis will not only give good understanding of the business environment but also give general idea to a person on what kind of leisure activity can be done in the city.

Problem:

A restaurant is a business that prepares and/or serves food & drink to customers. In any major city, the food is an important part of an ethnically diverse urban complex. The City of New York is famous for its excellent cuisine. It's food culture includes an many different types of international cuisines influenced by the city's immigrant history. This city is home to over one thousand of the finest and most diverse restaurants. So it is evident that we can call New York a mini heaven for many tourists or any foodie who wants try meals that are different and in such search one needs to be strategic about what and where to try these meals.

The idea of this capstone project is to show various segments the neighborhoods within New York City with their most popular cuisines. Through this analysis, we can figure out if food has any relationship with the diversity of a neighborhood. Our analysis will help us understand the diversity of a neighborhood by leveraging venue data from Foursquare's 'Places API' and 'k-means clustering' unsupervised machine learning algorithm.

Target Audience:

We are here to understand how different cultures and cuisines are placed all over New York City' neighborhood. This information or process of results can be used by many organization and people. A simple tourist visiting New York City, a New Yorker or traveling organizations such as expedia or tripadvisor can help people find meals or cuisine. In addition, this information can be used as a business opportunities and can be utilized by a new business owner on where or where not to open a restaurant. A government entity can also utilized such

information to study overall impact and structure of immigrant, restaurants and/or diverse cuisine.

Data Section

We will analyzed New York City in this capstone project. The main dataset will be from link (https://geo.nyu.edu/catalog/nyu_2451_34572), with a total of 5 boroughs and 306 neighborhoods. We will segment the neighborhoods in the city and explore various restaurants with the latitude and longitude coordinates of each neighborhood. This dataset will provide the addresses of neighborhood of NYC in json format. We will also use Foursquare API, a location data provider, to make RESTful API calls to retrieve data about venues in different neighborhoods.

Methodology Section

Load and explore the data:

A dataset was required to segment the neighborhoods of New York City and this data should contain appropriate boroughs and the neighborhoods. It should also must contains the latitude and longitude coordinates. The data link is provided above. Once the .json file is downloaded, we will structure it to fully understand the relevant fields. The data will be transferred to pandas dataframe, by looping through the data and filling the dataframe rows one at a time using the following depicted loop.

New York City Data:
We have to format the city data to be readable and contains proper boroughs and its neighborhoods, all including their exact location via latitude and longitude coordinates. The data is online at https://geo.nyu.edu/catalog/nyu_2451_34572 and the json format https://cocl.us/new_york_dataset/newyork_data.json

Transform the data into a dataframe pandas:

The next step is to transform the data of nested Python dictionaries into a pandas dataframe. We will start by creating an empty dataframe and take a look at the empty dataframe to confirm that the columns/rows. Then loop through the data and fill the dataframe one row at a time. Finally, examine top rows of resulting dataframe.

	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

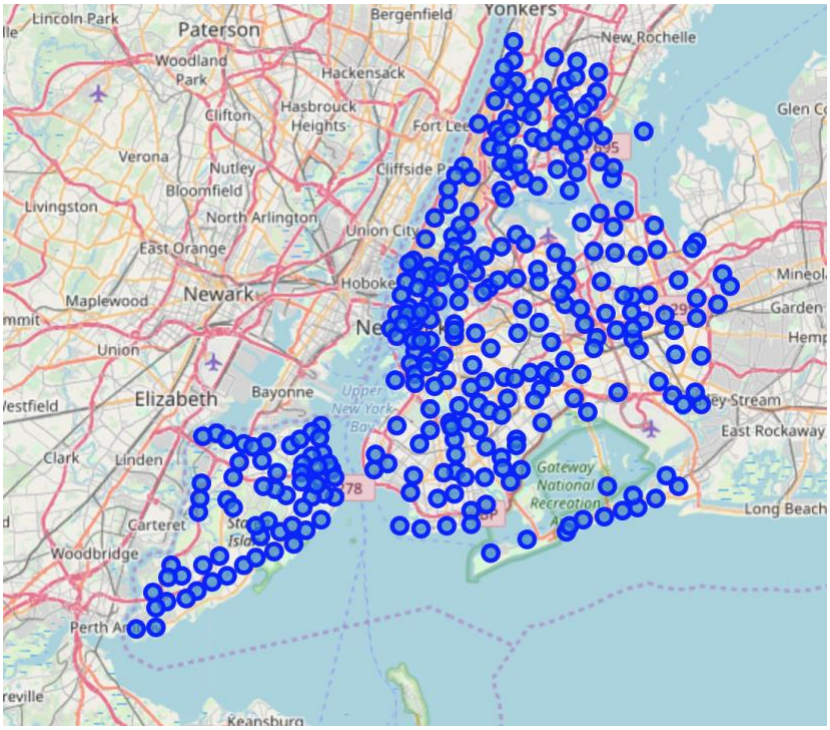
Add latitude and longitude:

Import latitude and longitude values of New York City via geopy library. Also, we will be creating a map through Folium tool. User can zoom in/out and click on the map to know the neighborhood and its borough.

```
address = 'New York City, NY'
location = None

#define an instance of the geocoder -> ny_explorer
while location == None:
    try:
        geolocator = Nominatim(user_agent="ny_explorer")
        location = geolocator.geocode(address)
        latitude = location.latitude
        longitude = location.longitude
    except:
        pass
print('The geograpical coordinate of New York City are {}, {}.'.format(latitude, longitude))
```

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



Foursquare:

Now, we will use the Foursquare API to explore the neighborhoods and segment them. There are many endpoints available on Foursquare for various GET requests. But, to explore the cuisines, it is required that all the venues extracted are from 'Food' category. Foursquare Venue Category Hierarchy is retrieved and the returned request is further analyzed. We found that there are 10 major or parent categories of venues.

```
for data in category_list:
    print(data['id'], data['name'])
```

```
4d4b7104d754a06370d81259 Arts & Entertainment
4d4b7105d754a06372d81259 College & University
4d4b7105d754a06373d81259 Event
4d4b7105d754a06374d81259 Food
4d4b7105d754a06376d81259 Nightlife Spot
4d4b7105d754a06377d81259 Outdoors & Recreation
4d4b7105d754a06375d81259 Professional & Other Places
4e67e38e036454776db1fb3a Residence
4d4b7105d754a06378d81259 Shop & Service
4d4b7105d754a06379d81259 Travel & Transport
```

Defining Food:

The ‘FOOD’ category in the above depiction is the matter of interest. We will create a function to return a dictionary with ‘Category ID’ & ‘Category Name’ of ‘Food’ & it's sub-categories. This function takes the parent ‘Category ID’ and returns the ‘Category Name’ and ‘Category ID’ of all the sub-categories.

Test the neighborhood results of GET Request:

We will try to get the neighborhood's name and its latitude and longitude values to confirm data. Then, we will request the 'Food' that is in neighborhood within a radius of 1000 meters. The first neighborhood returned is ‘Co-op City’ with Latitude 40.87 and Longitude -73.82. Afterwards, a GET request URL will be created to search for Venue with ‘Category ID’ ='4d4b7105d754a06374d81259', which is the ‘Category ID’ for ‘Food’, and radius = 1000 meters.

```
neighborhoods.loc[1, 'Neighborhood']
```

```
]: 'Co-op City'
```

```
neighborhood_latitude = neighborhoods.loc[1, 'Latitude'] #neighborhood latitude value
neighborhood_longitude = neighborhoods.loc[1, 'Longitude'] #neighborhood longitude value

neighborhood_name = neighborhoods.loc[1, 'Neighborhood'] #neighborhood name

print('Latitude and longitude values of {} are {}, {}.'.format(neighborhood_name,
                                                               neighborhood_latitude,
                                                               neighborhood_longitude))
```

```
Latitude and longitude values of Co-op City are 40.87429419303012, -73.82993910812398.
```

Creating a function:

In above example, the category name of the venue 'Dunkin'. We are targeting to segment NYC neighborhoods through the lens of various food. We will have to do similar process of this food data for all the 306 neighborhoods.

We need to create a function ‘getNearbyFood’ to repeat this process below for all neighborhoods:
-Loop through neighborhoods, make the API request URL with radius=1000, LIMIT=100; make the GET request
-Individual neighborhood, return only relevant information for each nearby venue
-Append all nearby venues to a list and unfold the list & append it to dataframe being returned

The categoryId parameter in the API request URL can be a comma seperated string. So, we create a comma separated string from category_dict. But this process is redundant because if a top-level category is specified in the GET Request, all sub-categories will also match the query.

But it is an interesting way to retrieve all the sub-categories ID with name. We will also use pickle library to serialize the information retrieved from GET requests. This step will counter any redundant requests to the Foursquare API.

Machine Learning: Exploratory Data

The combined dataframe 'nyc_venues' will have the required information with the size 13,935 venues. There are also 197 unique categories with Deli and Pizza on the top.

Sorting Data:

We should show the size of our dataframe and see its unique categories by removing general categories since we are planned to explore diverse food. First, we inputted all the unique categories into a python 'list' so after we can create our general list. We reduced the data noise by almost half, resulting in 101 unique food categories from 197.

```
In [34]: nyc_venues = nyc_venues[nyc_venues['Venue Category'].isin(food_categories)].reset_index()
         nyc_venues.head(5)
```

Out[34]:

	index	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	2	Wakefield	40.894705	-73.847201	McDonald's	40.902645	-73.849485	Fast Food Restaurant
1	4	Wakefield	40.894705	-73.847201	Burger King	40.895540	-73.856460	Fast Food Restaurant
2	5	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898083	-73.850259	Caribbean Restaurant
3	7	Wakefield	40.894705	-73.847201	Golden Krust Caribbean Restaurant	40.903773	-73.850051	Caribbean Restaurant
4	9	Wakefield	40.894705	-73.847201	McDonald's	40.892779	-73.857473	Fast Food Restaurant

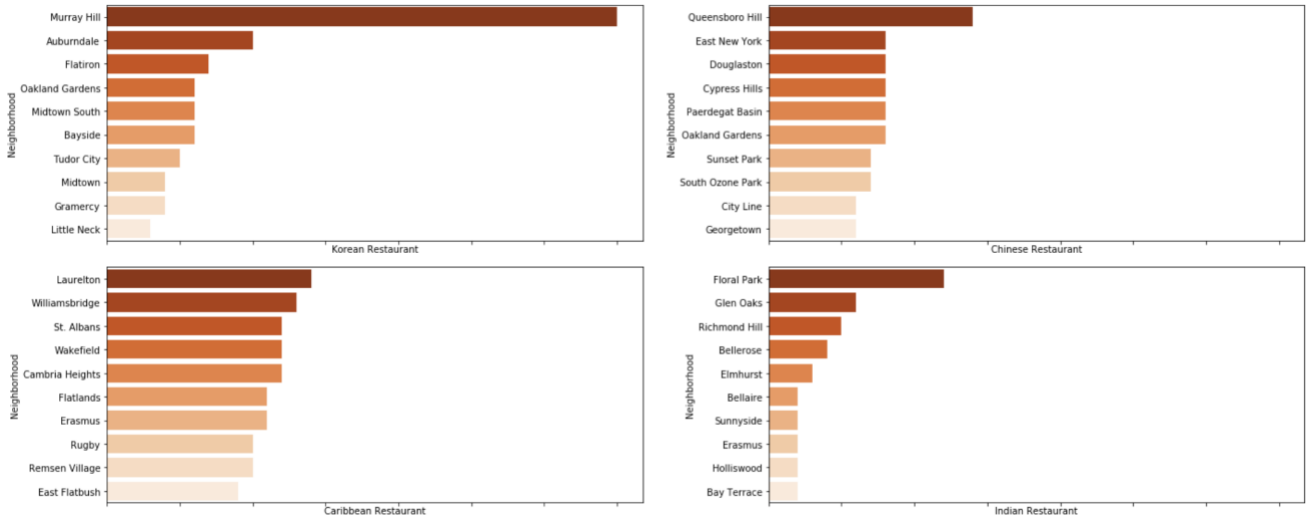
Looking further into each Neighborhood:

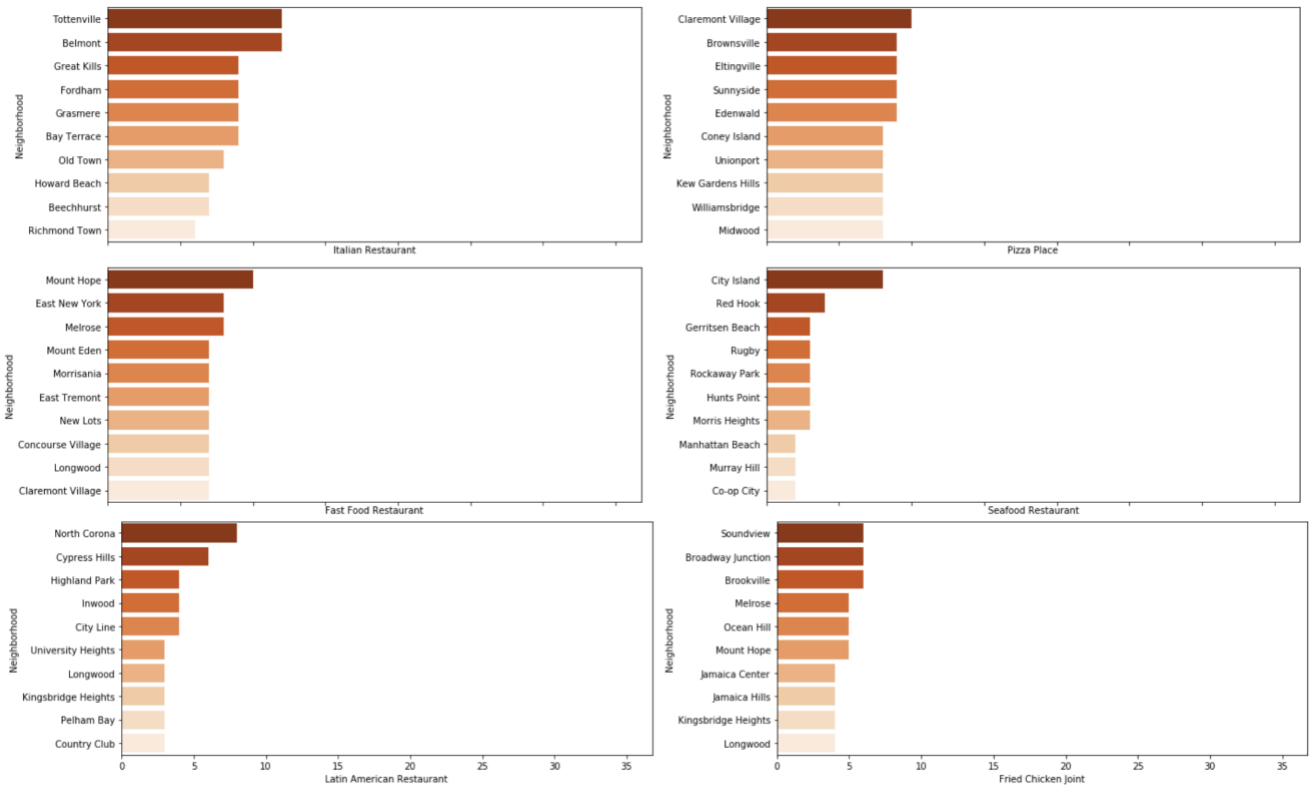
We will try to show top 10 food categories in NYC by counting venues of each category in each neighborhood. We will also show some statistical data of each neighborhood and the frequency of venues. The below process is taken forth by using 'one hot encoding' function of python pandas library, where it converts the categorical variables into a form that could be provided to ML algorithms to do a better job in prediction. Afterwards, we will add column (Neighborhood) to see the results. For example, we can see that in the first five neighborhoods of the dataframe, Annadale, Arden Heights and Arlington have 3, 3, 2 'American Restaurant' in its vicinity. With the help of some statistics below, we can find the top venue categories; such as Korean Restaurant, Chinese Restaurant and Caribbean Restaurant as top three.

	count	mean	std	min	25%	50%	75%	max
Korean Restaurant	302.0	0.364238	2.263339	0.0	0.0	0.0	0.00	35.0
Chinese Restaurant	302.0	2.069536	1.879722	0.0	1.0	2.0	3.00	14.0
Caribbean Restaurant	302.0	1.165563	2.551608	0.0	0.0	0.0	1.00	14.0
Indian Restaurant	302.0	0.291391	0.965101	0.0	0.0	0.0	0.00	12.0
Italian Restaurant	302.0	1.837748	2.093376	0.0	0.0	1.0	3.00	12.0
Pizza Place	302.0	3.427152	2.031232	0.0	2.0	3.0	5.00	10.0
Fast Food Restaurant	302.0	2.052980	1.950511	0.0	0.0	2.0	3.00	10.0
Seafood Restaurant	302.0	0.533113	0.825605	0.0	0.0	0.0	1.00	8.0
Latin American Restaurant	302.0	0.476821	0.987190	0.0	0.0	0.0	1.00	8.0
Fried Chicken Joint	302.0	0.956954	1.228726	0.0	0.0	1.0	1.75	6.0

Data Visualization:

We can visualize the top 10 categories via bar graph using python seaborn library. There is also a table of neighborhood grouped together and the frequency of each category is calculated by taking the mean.

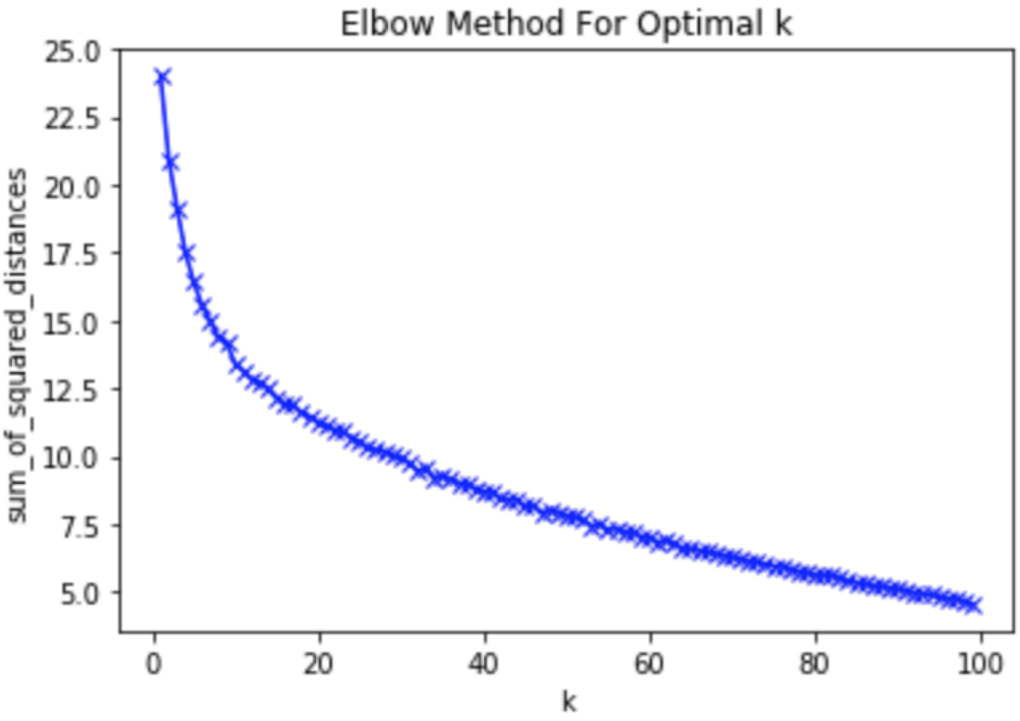




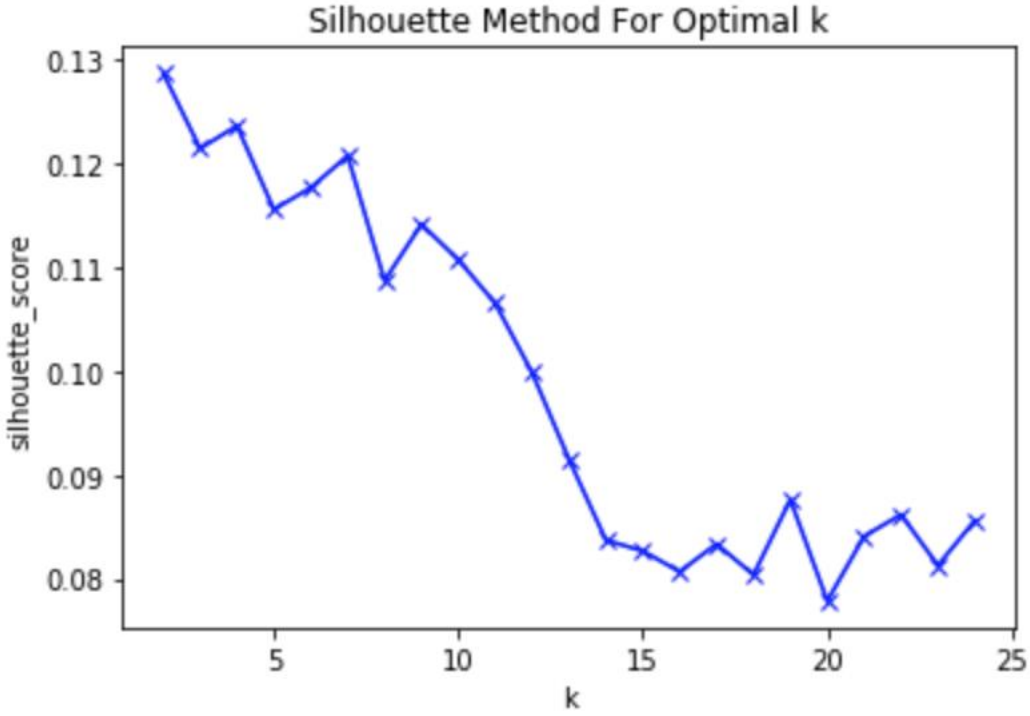
Cluster Neighborhoods:

Now, we will use k-means to count each neighborhoods cluster. 'k-means' is an unsupervised machine learning algorithm which creates clusters of data points aggregated those points together because of certain similarities. We need to find optimal number of clusters for k-means. We can use two methods and graph them to see which one works; the elbow method and silhouette method.

Elbow: calculates the sum of squared distances of samples to their closest cluster center for different values of k. The value of k after which there is no significant decrease in sum of squared distances is chosen. Below we will see that elbow method isn't useful to find the optimal number of clusters.

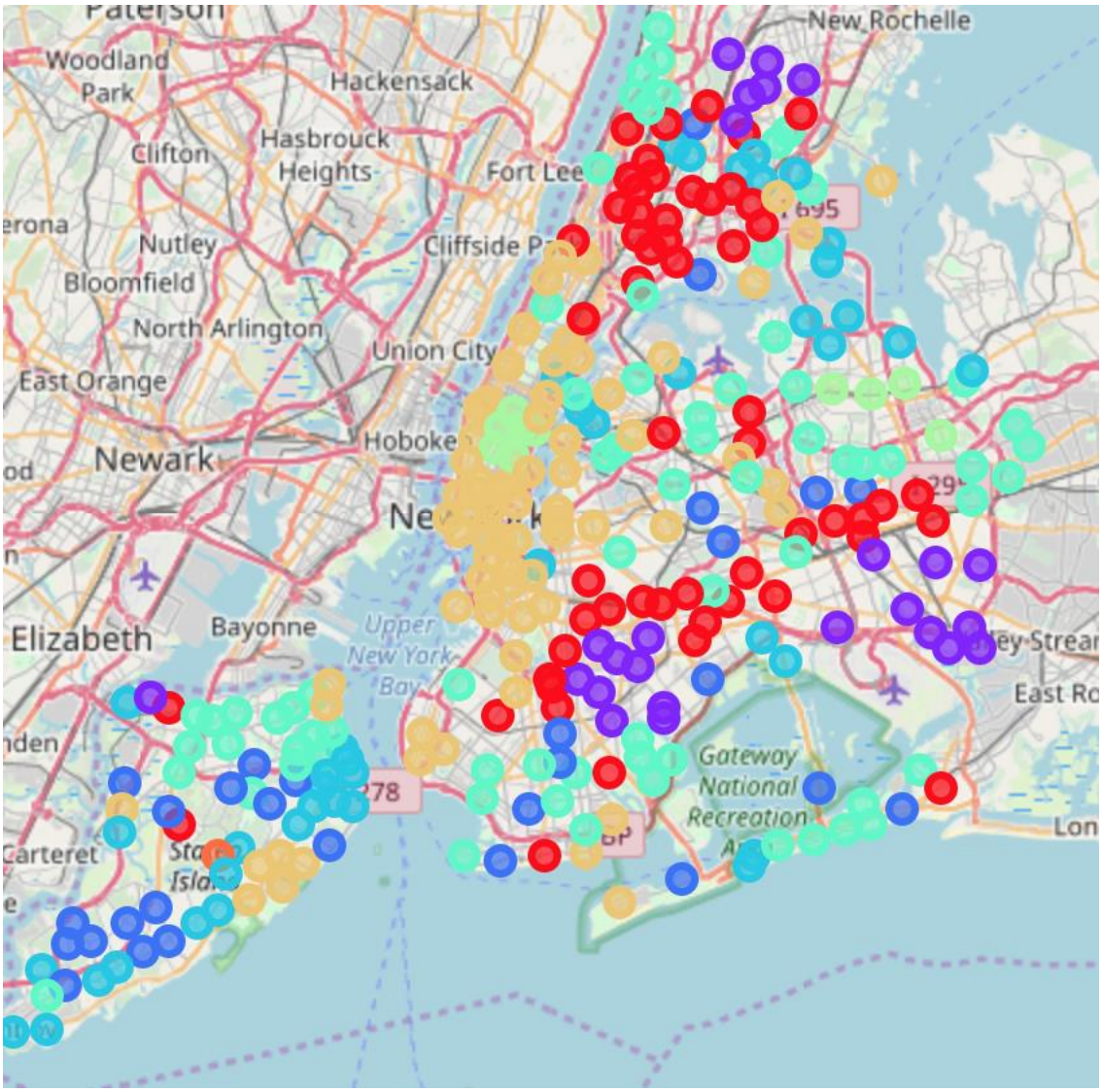


Silhouette: measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).



k-Means:

The highest values are at k = 2, k = 4, and k = 8. We will use set number of clusters 8, since 2 and 4 clusters will give us very broad classification of the venues. Then, we will create a new dataframe that includes the cluster as well as the top 10 venues for each neighborhood. We will also use this data to visualize the resulting clusters.



Results & Findings:

Cluster 0, 1 & 2:

- Cluster 0) Fast Food Restaurant is a dominant occurrence, with Pizza Place number two on the top venues.
- Cluster 1) Caribbean Restaurant is a massive venue with 18 occurrences, followed by Pizza Place with 7 occurrences.
- Cluster 2) Pizza Place holds a number spot with 27 occurrences for a long shot and no one comes close to it.

Cluster 3, 4 & 5:

- Cluster 3) Italian Restaurant is a dominant occurrence and Pizza Place as number two.
- Cluster 4) Pizza Place is a massive venue with 33 occurrences, followed by Chinese Restaurant with 22 occurrences.
- Cluster 5) Korean Restaurant holds a number spot with 8 occurrences, followed by American Restaurant with 4 occurrences.

To full incorporate the clusters, we have completed 03 analysis; count of Borough, count of 1st Most Common Venue and count of 2nd Most Common Venue.

There is scope to increase this analysis:

- We could look at economical PnL of each venue of restaurant to figure out what cuisine a potential inventor/owner should open.
- We could have also utilized word could library tool to visualize the cuisine type in each borough, the higher number of times specific cuisine word appears in a dataset the bigger and bolder it appears in the world cloud.

Conclusions

Through the method of machine learning, specially k-Means, we can understand, visualize the data and how the conclusion can be varied based off ever changing data. Segmenting the clusters in New York City helped us see various sides of boroughs and its neighborhoods. Now, we are able to see what are the popular cuisines within each area of New York City. In the future, we can improve our model by looking at daily live dataset. Overall, our data was limited data so our analysis could be biased. But in the future, we may need to get more data and expand the scope of this analysis to get better results. Brooklyn and Manhattan has high concentration of restaurant business.