# Battle of the Neighborhoods

## Segmentation of New York City Based on Cuisines

# Business Problem

The idea of this project is to categorically segment the neighborhoods of New York City into major clusters and examine their cuisines. A desirable intention is to examine the neighborhood cluster's food habits and taste.

**Stakeholders** would be the one who are interested to use this quantifiable analysis to understand the distribution of different cultures and cuisines over "the most diverse city on the planet - NYC". Also, this project can be utilized by a new food vendor who is willing to open his or her restaurant. Or by a government authority to examine and study their city's culture diversity better.

# Data Source

**New York City Dataset**

Link: https://geo.nyu.edu/catalog/nyu_2451_34572

Description: This New York City Neighborhood Names point file was created as a guide to New York City's neighborhoods that appear on the web resource, "New York: A City of Neighborhoods." Best estimates of label centroids were established at a 1:1,000 scale, but are ideally viewed at a 1:50,000 scale. This dataset will provide the addresses of neighborhood of NYC in json format.

**Foursquare API**

Link: https://developer.foursquare.com/docs

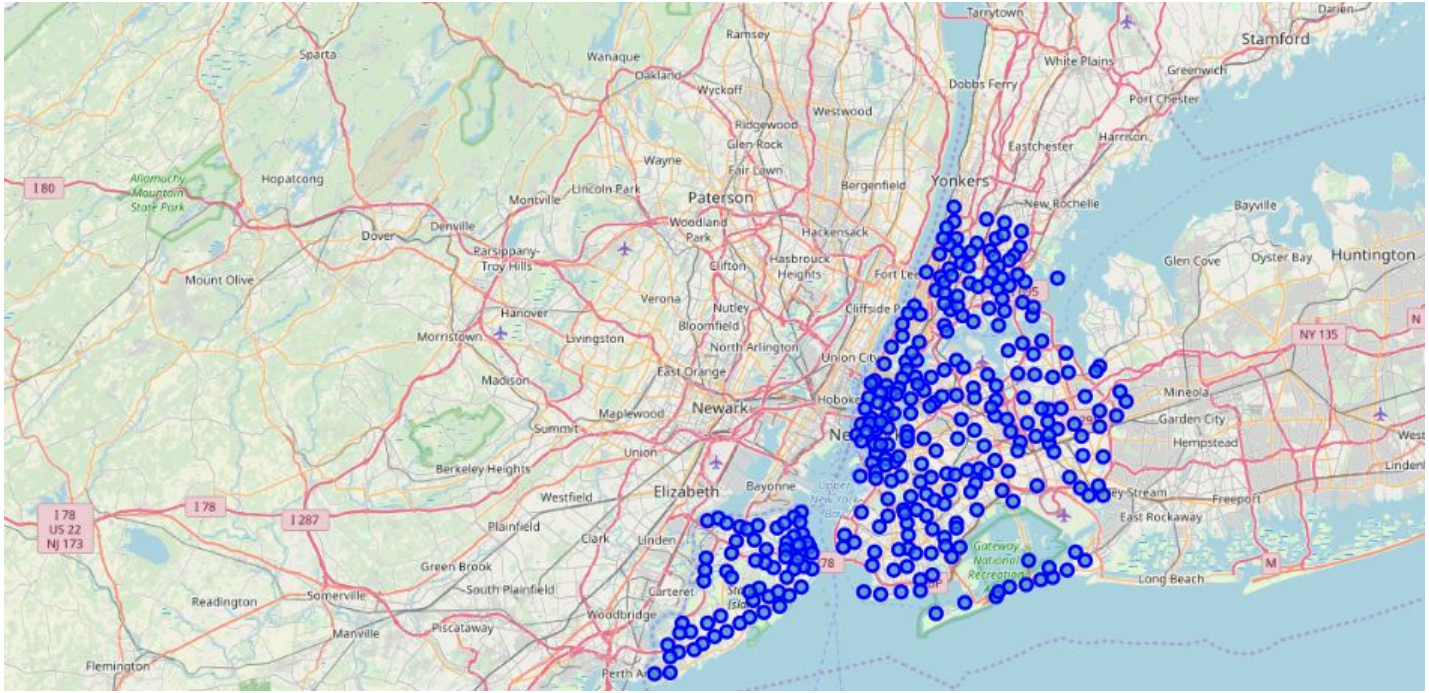
Description: Foursquare API, a location data provider, will be used to make RESTful API calls to retrieve data about venues in different neighborhoods. This is the link to Foursquare Venue Category Hierarchy. Venues retrieved from all the neighborhoods are categorized broadly into "Arts & Entertainment", "College & University", "Event", "Food", "Nightlife Spot", "Outdoors & Recreation", etc.

# Data Processing

Create a data frame with Borough, Neighborhood, Latitude and Longitude details of the New York City's neighborhood.

| | 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 visualization of New York City with neighborhoods superimposed on top



# Major 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
```

# 202 unique sub-categories under Food category

```
print('There are {} uniques categories.'.format(len(nyc_venues['Venue Category'].unique())))
nyc_venues.groupby('Venue Category')['Venue Category'].count().sort_values(ascending=False)
```

```
There are 202 uniques categories.
```

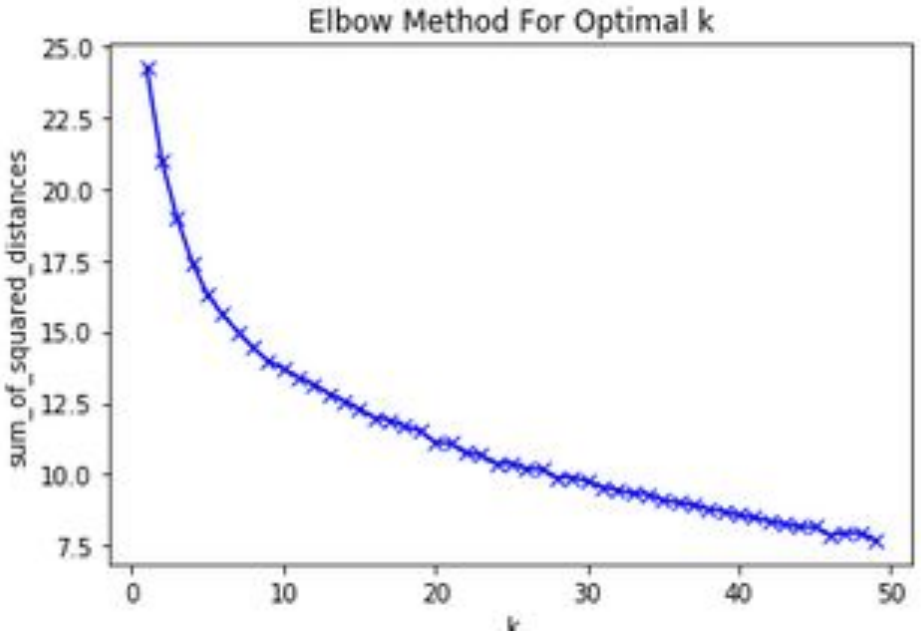
```
Venue Category
Deli / Bodega                    1046
Pizza Place                      1036
Coffee Shop                       905
Donut Shop                        638
Fast Food Restaurant              620
Chinese Restaurant                607
Bakery                            548
Italian Restaurant                544
American Restaurant               443
Café                              426
Caribbean Restaurant              361
Mexican Restaurant                337
Bagel Shop                        335
Sandwich Place                    325
Fried Chicken Joint               324
Diner                             289
```

# Machine Learning

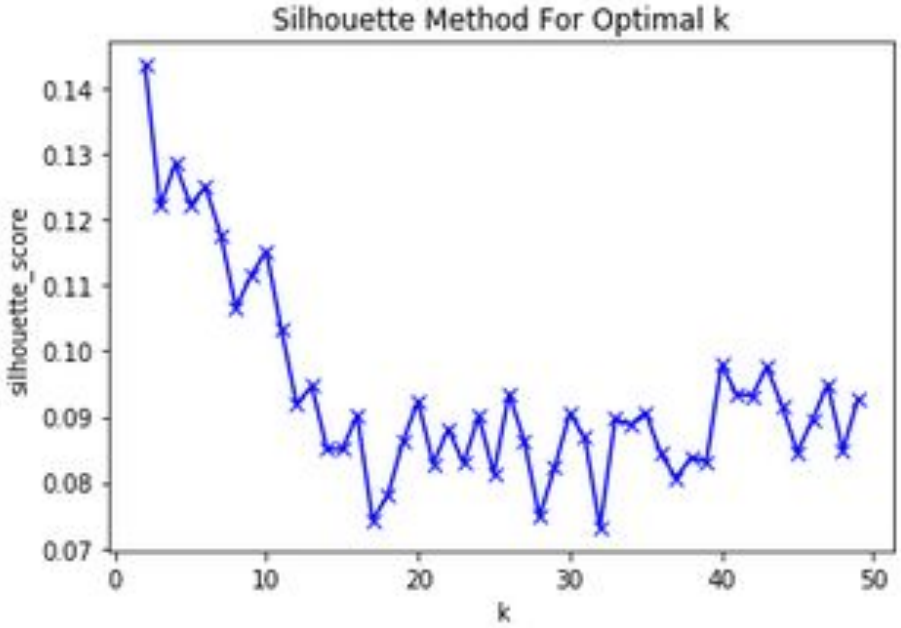
‘k-means’ is an unsupervised machine learning algorithm which creates clusters of data points aggregated together because of certain similarities. This algorithm will be used to count neighborhoods for each cluster label for variable cluster size.

To implement this algorithm, it is very important to determine the optimal number of clusters (i.e. k). There are 2 most popular methods for the same, namely ‘The Elbow Method’ and ‘The Silhouette Method’.

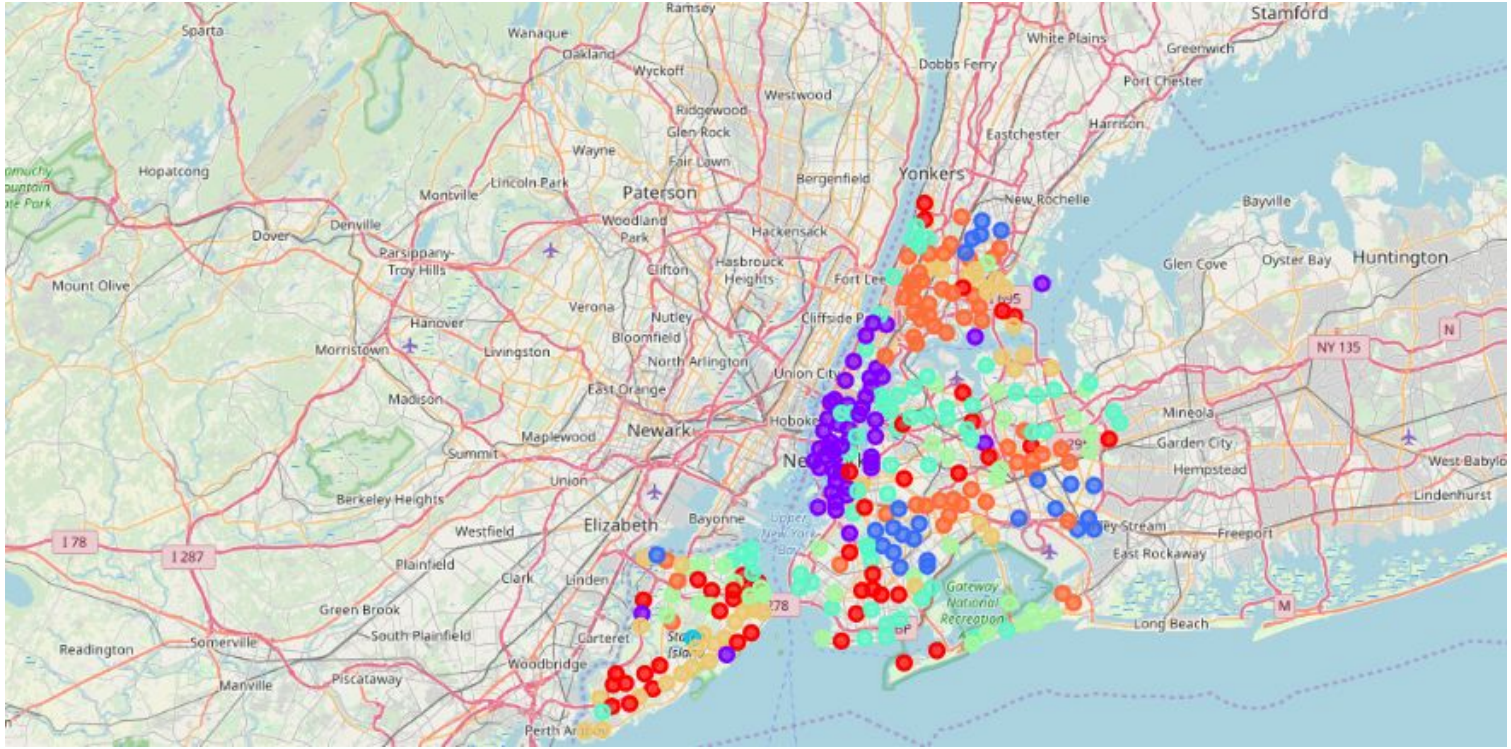
# Elbow Method



# The Silhouette Method



# Visualization after clustering



# Cluster 0

```
for col in required_column:
    print(cluster_0[col].value_counts(ascending = False))
    print("-----------------------------------------------")
```

```
Pizza Place             42
Taco Place               1
American Restaurant      1
Name: 1st Most Common Venue, dtype: int64
-----------------------------------------------
Italian Restaurant      11
American Restaurant      8
Fast Food Restaurant     6
Mexican Restaurant       4
Sushi Restaurant         3
Japanese Restaurant      2
Pizza Place              2
BBQ Joint                2
Chinese Restaurant       2
Taco Place               1
Asian Restaurant         1
Thai Restaurant          1
Spanish Restaurant       1
Name: 2nd Most Common Venue, dtype: int64
-----------------------------------------------
Staten Island    19
Brooklyn         11
Queens            9
Bronx             5
Name: Borough, dtype: int64
-----------------------------------------------
```

# Cluster 1

```
for col in required_column:
    print(cluster_1[col].value_counts(ascending = False))
    print("-----------------------------------------------")
```

```
American Restaurant                25
Pizza Place                         6
Italian Restaurant                  6
French Restaurant                   3
New American Restaurant             3
Dim Sum Restaurant                  2
Korean Restaurant                   1
Southern / Soul Food Restaurant     1
Seafood Restaurant                  1
Fast Food Restaurant                1
Name: 1st Most Common Venue, dtype: int64
-----------------------------------------------
American Restaurant         14
Italian Restaurant           9
Pizza Place                  8
Japanese Restaurant          2
Seafood Restaurant           2
Ramen Restaurant             2
Thai Restaurant              2
Mexican Restaurant           2
Chinese Restaurant           2
German Restaurant            1
Korean Restaurant            1
Mediterranean Restaurant     1
Dim Sum Restaurant           1
French Restaurant            1
BBQ Joint                    1
Name: 2nd Most Common Venue, dtype: int64
-----------------------------------------------
Manhattan        29
Brooklyn         14
Queens            2
Staten Island     2
Bronx             2
Name: Borough, dtype: int64
```

# Cluster 2

```
for col in required_column:
    print(cluster_2[col].value_counts(ascending = False))
    print("-----------------------------------------")
```

```
Caribbean Restaurant    21
Chinese Restaurant       2
Fried Chicken Joint      1
American Restaurant      1
Name: 1st Most Common Venue, dtype: int64
-----------------------------------------
Fast Food Restaurant    7
Fried Chicken Joint     5
Pizza Place             5
Chinese Restaurant      4
Caribbean Restaurant    3
Seafood Restaurant      1
Name: 2nd Most Common Venue, dtype: int64
-----------------------------------------
Brooklyn         11
Queens            8
Bronx             5
Staten Island     1
Name: Borough, dtype: int64
-----------------------------------------
```

# Cluster 3

```
for col in required_column:
    print(cluster_3[col].value_counts(ascending = False))
    print("-----------------------------------------")
```

```
Italian Restaurant    1
Name: 1st Most Common Venue, dtype: int64
-----------------------------------------
Yemeni Restaurant    1
Name: 2nd Most Common Venue, dtype: int64
-----------------------------------------
Staten Island    1
Name: Borough, dtype: int64
-----------------------------------------
```

# Cluster 4

```
for col in required_column:
    print(cluster_4[col].value_counts(ascending = False))
    print("-----------------------------------------------")
```

```
Pizza Place                   16
Italian Restaurant             9
Fast Food Restaurant           8
Mexican Restaurant             5
Korean Restaurant              5
Sushi Restaurant               3
Indian Restaurant              2
Thai Restaurant                2
Caribbean Restaurant           2
Ramen Restaurant               1
Greek Restaurant               1
Seafood Restaurant             1
Sri Lankan Restaurant          1
Filipino Restaurant            1
American Restaurant            1
Russian Restaurant             1
Eastern European Restaurant    1
Chinese Restaurant             1
Name: 1st Most Common Venue, dtype: int64
-----------------------------------------------
Pizza Place                   10
Italian Restaurant            10
Fast Food Restaurant           7
Chinese Restaurant             6
Mexican Restaurant             6
Latin American Restaurant      2
Caribbean Restaurant           2
Thai Restaurant                2
Vietnamese Restaurant          2
Greek Restaurant               2
American Restaurant            2
Russian Restaurant             2
Sushi Restaurant               1
Mediterranean Restaurant       1
Spanish Restaurant             1
```

# Cluster 5

```
for col in required_column:
    print(cluster_5[col].value_counts(ascending = False))
    print("-------------------------------------------")
```

```
Chinese Restaurant    18
Pizza Place           13
Italian Restaurant     2
Indian Restaurant      1
Greek Restaurant       1
Name: 1st Most Common Venue, dtype: int64
-------------------------------------------
Chinese Restaurant      12
Pizza Place             10
Italian Restaurant       4
Mexican Restaurant       2
Fried Chicken Joint      1
Asian Restaurant         1
American Restaurant      1
Japanese Restaurant      1
Cantonese Restaurant     1
Caribbean Restaurant     1
Korean Restaurant        1
Name: 2nd Most Common Venue, dtype: int64
-------------------------------------------
Queens           18
Staten Island    11
Brooklyn          5
Bronx             1
Name: Borough, dtype: int64
-------------------------------------------
```

# Cluster 6

```
for col in required_column:
    print(cluster_6[col].value_counts(ascending = False))
    print("----------------------------------------------")
```

```
Italian Restaurant     26
Pizza Place             9
American Restaurant     1
Asian Restaurant        1
Name: 1st Most Common Venue, dtype: int64
----------------------------------------------
Pizza Place                14
Italian Restaurant         10
Fast Food Restaurant        4
American Restaurant         3
Asian Restaurant            2
Mexican Restaurant          2
Japanese Restaurant         1
New American Restaurant     1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------------
Staten Island    20
Bronx             9
Queens            6
Brooklyn          2
Name: Borough, dtype: int64
----------------------------------------------
```

# Cluster 7

```
for col in required_column:
    print(cluster_7[col].value_counts(ascending = False))
    print("----------------------------------------------")
```

```
Pizza Place                  20
Fast Food Restaurant         20
Chinese Restaurant            7
Fried Chicken Joint           3
Seafood Restaurant            1
Spanish Restaurant            1
Latin American Restaurant     1
Caribbean Restaurant          1
Name: 1st Most Common Venue, dtype: int64
----------------------------------------------
Pizza Place                          14
Fast Food Restaurant                 12
Chinese Restaurant                   10
Fried Chicken Joint                   6
Caribbean Restaurant                  5
Mexican Restaurant                    3
Spanish Restaurant                    2
Southern / Soul Food Restaurant       1
American Restaurant                   1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------------
Bronx            27
Queens           13
Brooklyn          9
Staten Island     3
Manhattan         2
Name: Borough, dtype: int64
----------------------------------------------
```

# Discussions

We analysed the different clusters by using the following metrics

- Cont of Borough
- Count of 1st Most Common Venue
- Count of 2nd Most Common Venue

From the analysis, it is obvious that Pizza Place is the most common venue across all the clusters. Following could be the name of the clusters segmented by K-Means unsupervised machine learning algorithm.

- Cluster 0 : Pizza
- Cluster 1 : American
- Cluster 2 : Caribbean
- Cluster 3 : Italian
- Cluster 4 : Mix of Cuisines
- Cluster 5 : Chinese
- Cluster 6 : Italian
- Cluster 7 : Fast Food

# Conclusion

The Neighborhoods of New York City were very briefly segmented into eight clusters and upon analysis it was possible to rename the clusters based on the venues in and arond the neighborhood. Along with American cuisine, Italian & Chinese are very dominant in NYC.

The results of this project can be improved by using a current dataset along with API which is more interested in Food venues. The scope of the project can be expanded further to understand the dynamics of each neighborhood and suggest a new vendor a profitable venue to start his food place.

# References

Notebook created by Alex Aklson and Polong Lin for the Applied Data Science Capstone project.