Applied Data Science Capstone Project

as part of the

IBM Data Science Professional Certificate

Exploring New York City's Diverse Food Cultures for Tourism Purposes

Shabbir Ahmad

Course: Applied Data Science Capstone

Offered By: Coursera

Contents

Introduction and Problem Description

New York City is the most populous city in the United States and the largest metropolitan area in the world by urban landmass (World Urban Areas, 2018). The city has remained a major entry point for immigrants over the years. According to recent statistics, about 36 percent of the city's population is foreign-born (NYC Mayor's Office, 2018). The ethnic diversity of New York City is reflected in its diverse cuisine. Almost all the ethnic cuisines are available in the city. John Corry, the New York Times reporter rightly pointed out that "One of the glories of New York is its ethnic food" (Food Reference, 2020).

Tourism industry in the city is in boom for over a decade. The city welcomed over 65 million tourists in 2019 which included over 13 million foreign tourists (NYC and Company, 2020). While visiting different sites, tourists often find it difficult to find the restaurants and eateries of their choice in those areas. For instance, in some neighbourhoods, there are numerous outlets of Italian and Chinese food but there might be only one or two Mexican or Caribbean restaurants. Moreover, sometimes the tourists are also overwhelmed by the food choices and the decision-making process may be time consuming.

In this project, I aim to explore all the boroughs of New York City and locate all the possible food options in specific areas and neighbourhoods of the city. Eateries for different food categories and cuisine belonging to different ethnic groups and countries will be located in the city.

Data Description

Requirements

For this project, data related to boroughs (including neighbourhoods) of New York City will be required. The Foursquare API will be used to extract data for different locations of the city. I will need address, latitude and longitude of the locations mainly. To create data frames and analyse data, I will use different libraries like pandas, numpy, matplotlib and folium (for mapping).

Data Sources

The data for this project will be collected from:

- New York City dataset: This dataset will provide the addresses of neighbourhoods of the city. A json file will be extracted from here. https://geo.nyu.edu/catalog/nyu_2451_34572

- Foursquare API: It is a location data provider and will be used to make API calls to extract data regarding different venues in New York City's neighbourhoods. To access the API, 'CLIENT_ID', 'CLIENT_SECRET' and 'VERSION' will be defined. https://developer.foursquare.com/docs/

Audience/Stakeholders

- The audience of this report is mainly local and foreign tourists in New York City who would like to find the eatery of their choice conveniently or explore the diverse food choices available at a particular locality.
- Findings of this report can also be used by the tourism department of the city as well as by tour operators to guide the tourists regarding the locations of the diverse food options.

Methodology

Methodology is comprised of the following steps.

- Downloading and exploring the dataset.
- API calls to Foursquare
- Importing Pickle library
- Data Cleaning
- Feature engineering (one hot encoding)
- Clustering data through k-means

These steps will be explained briefly here.

Downloading and exploring the dataset

The dataset is downloaded from the aforementioned URL. The URL returns the .json file containing the dataset in the form of a python dictionary. The requisite data i.e. the list of five boroughs and the neighbourhoods within these boroughs, is in the features key.

Dataframe 'neighbourhoods'

The dictionary is transformed into a dataframe by looping through the data. It is the first dataframe of this project. The dataframe rows are filled one at a time. It results into a dataframe with borough, Neighbourhood, Latitude and Longitude details of the city's neighbourhoods.

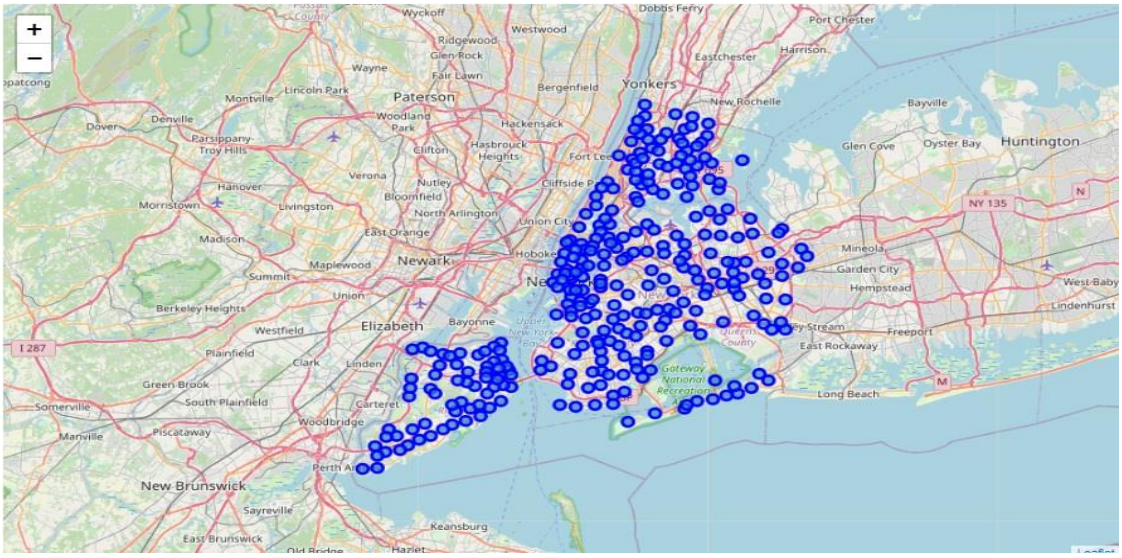
The dataframe has a total of five boroughs and 306 neighbourhoods. The first five rows of the table are shown as follows.

	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

First five rows of the dataframe (1)

The latitude and longitude of the New York City is found through 'Geopy' library. The latitude is 40.71 while the longitude is -74.01.

The dataframe is further used to generate a map of New York City with neighbourhoods superimposed on top. The map is generated by using the 'folium' library.



New York City map generated through folium

API calls to Foursquare

The neighbourhoods are explored and segmented through the Foursquare API. API can be accessed through defining 'CLIENT ID', 'CLIENT SECRET' and 'VERSION'. The first two parameters are unique for every developer while version is a 'v' parameter, which is actually a date. It is designed to give developers the freedom to adapt to Foursquare API changes on their own schedule. An API endpoint is also needed to send GET requests.

After extracting Foursquare venue category hierarchy, I got a list of ten categories of venues as shown below.

```
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
```

To explore New York City cuisines, the venues are extracted from 'Food' category. As shown below, a function is created to extract a dictionary with 'Category ID' and 'Category Name' of 'Food' and its sub-categories. 'Food' category ID is *'4d4b7105d754a06374d81259'*. It will be used in further analysis.

```
def flatten_Hierarchy(category_list, checkParentID, category_dict, parent_id = ''):
    for data in category_list:

        if checkParentID == True and data['id'] == parent_id:
            category_dict[data['id']] = data['name']
            flatten_Hierarchy(category_list = data['categories'], checkParentID = False, category_dict = category_dict)

        elif checkParentID == False:
            category_dict[data['id']] = data['name']
            if len(data['categories']) != 0:
                flatten_Hierarchy(category_list = data['categories'], checkParentID = False, category_dict = category_dict)

    return category_dict
```

The first neighbourhood (Wakefield) in the dataset is further explored to understand the outcome of GET request. The latitude of this neighbourhood is 40.89 and longitude is -73.85. A GET request URL is created to search for venue with the aforementioned 'Category ID' with a radius of 500 meters.


```
[{'id': '4c783cef3badb1f7e4244b54',
  'name': 'Carvel Ice Cream',
  'location': {'address': '1006 E 233rd St',
   'lat': 40.890486685759605,
   'lng': -73.84856772568665,
   'labeledLatLngs': [{'label': 'display',
     'lat': 40.890486685759605,
     'lng': -73.84856772568665},
    {'label': 'entrance', 'lat': 40.890438, 'lng': -73.848559}],
   'distance': 483,
   'postalCode': '10466',
   'cc': 'US',
   'city': 'Bronx',
   'state': 'NY',
   'country': 'United States',
   'formattedAddress': ['1006 E 233rd St',
    'Bronx, NY 10466',
    'United States']},
  'categories': [{'id': '4bf58dd8d48988d1c9941735',
    'name': 'Ice Cream Shop',
    'pluralName': 'Ice Cream Shops',
    'shortName': 'Ice Cream',
    'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/icecream_',
     'suffix': '.png'},
    'primary': True}],
  'referralId': 'v-1588011540',
  'hasPerk': False}]
```

As mentioned earlier, there are 306 neighbourhoods in New York City and to segment the neighbourhoods, fetching data from all the neighbourhoods is necessary. However, it is laborious to repeat the process for these neighbourhoods individually. To make this process easier, a new function 'getNearbyFood' is created. This function loops through all the neighbourhoods of New York City and creates an API request URL with radius = 500, LIMIT = 100. Limit is set to 100, so that a maximum of 100 venues are returned in the vicinity.

```
def getNearbyFood(names, latitudes, longitudes, radius=1000, LIMIT=500):
    not_found = 0
    print('***Start ', end='')
    venues_list=[]
    for name, lat, lng in zip(names, latitudes, longitudes):
        print(' .', end='')

        # create the API request URL
        url = 'https://api.foursquare.com/v2/venues/search?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&categoryId={}&
limit={}'.format(
            CLIENT_ID,
            CLIENT_SECRET,
            VERSION,
            lat,
            lng,
            radius,
            "4d4b7105d754a06374d81259", # "Food" category id
            LIMIT)

        try:
            # make the GET request
            results = requests.get(url).json()['response']['venues']

            # return only relevant information for each nearby venue
            venues_list.append([(
                name,
                lat,
                lng,
                v['name'],
                v['location']['lat'],
                v['location']['lng'],
                v['categories'][0]['name']) for v in results])
        except:
            not_found += 1

    nearby_venues = pd.DataFrame([item for venue_list in venues_list for item in venue_list])
    nearby_venues.columns = ['Neighborhood',
                'Neighborhood Latitude',
                'Neighborhood Longitude',
                'Venue',
                'Venue Latitude',
                'Venue Longitude',
                'Venue Category']
    print("\nDone*** with {} venues with incompelete information.".format(not_found))
    return(nearby_venues)
```

Dataframe 'nyc_venues'

In the next step, information for each nearby venue is extracted through a GET request to Foursquare API. This data is added to a python list which is then flattened to add it to the dataframe 'nyc_venues' being returned by the function. It is also worth mentioning that if a major category is specified in the GET request, the Foursquare API returns all the sub-categories.

It is evident that the data obtained for all the 306 neighbourhoods is a large amount of information and it may need some organization. To serialize this information, 'Pickle' library is

used. However, we may need to deserialize it again later because it will be necessary for retrieving an exact python object structure.

```
import pickle # to serialize and deserialize a Python object structure
try:
    with open('nyc_food_venues.pkl', 'rb') as f:
        nyc_venues = pickle.load(f)
    print("---Dataframe Existed and Deserialized---")
except:
    nyc_venues = getNearbyFood(names=neighborhoods['Neighborhood'],
                               latitudes=neighborhoods['Latitude'],
                               longitudes=neighborhoods['Longitude']
                              )
    with open('nyc_food_venues.pkl', 'wb') as f:
        pickle.dump(nyc_venues, f)
    print("---Dataframe Created and Serialized---")
```

```
***Start  . . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
Done*** with 0 venues with incompelete information.
---Dataframe Created and Serialized---
```

The size of this dataframe is 13612 venues and there are 187 unique sub-categories for 'Food' category. The first five rows of the dataframe look like this.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Central Deli	40.896728	-73.844387	Deli / Bodega
1	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop
3	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898083	-73.850259	Caribbean Restaurant
4	Wakefield	40.894705	-73.847201	SUBWAY	40.890468	-73.849152	Sandwich Place

Dataframe (2)

Data Cleaning

This project is aimed at locating diverse cuisine in New York City and make it easier for the tourists to find the eatery of their choice. So far in the data analysis, I have identified 187 categories of 'Food' but these also include general categories like coffee shops, ice cream shops, cafes and so on. These categories are common across different cultures and do not need to be identified in this analysis which is why these will be eliminated to refine the data. A list of these 'general' categories is compiled and then subtracted from the list of unique categories. It will give us the categories of our interest. After these steps, we are left with 101 unique categories in comparison to 187 categories before data cleaning.

The first five rows of the dataframe 'nyc_venues' now look like this.

	index	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	3	Wakefield	40.894705	-73.847201	Cooler Runnings Jamaican Restaurant Inc	40.898083	-73.850259	Caribbean Restaurant
1	5	Wakefield	40.894705	-73.847201	Popeyes Louisiana Kitchen	40.889322	-73.843323	Fried Chicken Joint
2	6	Wakefield	40.894705	-73.847201	Burger King	40.895540	-73.856460	Fast Food Restaurant
3	9	Wakefield	40.894705	-73.847201	Golden Krust Caribbean Restaurant	40.903773	-73.850051	Caribbean Restaurant
4	10	Wakefield	40.894705	-73.847201	McDonald's	40.892779	-73.857473	Fast Food Restaurant

Dataframe nyc_venues – First five rows

Previously we had ice cream and Dunkin donuts in the first five rows. Now we have only the restaurants and eateries in our list.

One hot encoding

One hot encoding is a function of the 'pandas' library and is usually used to convert categorical variables to a form that can be used as input for machine learning (ML) algorithms to efficiently predict the outcomes. In this project, it is used to analyse each neighbourhood individually to find out the most common cuisine within 500 meters of its radius.

Dataframe nyc_onehot

A new dataframe nyc_onehot is created. 'Neighbourhood' column. The size of this dataframe is 6635 data points including all venues for all the neighbourhoods. Now, we can also count the number of venues for each sub-category in every neighbourhood. For instance, the first five neighbourhoods of Allerton, Annadale and Arrochar have 6, 2 and 3 Chinese restaurants respectively in their 500 meters radius. Similarly, Annadale, Arden Heights and Arlington have 3, 3 and 2 American restaurants in their 500 meters radius respectively.

	Afghan Restaurant	African Restaurant	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Australian Restaurant	Austrian Restaurant	BBQ Joint	Beer Store	Brazilian Restaurant
Neighborhood											
Allerton	0	0	0	0	0	0	0	0	0	0	0
Annadale	0	0	3	0	0	0	0	0	0	0	0
Arden Heights	0	0	3	0	0	0	0	0	1	0	0
Arlington	0	0	2	0	0	1	0	0	0	0	0
Arrochar	0	0	0	0	0	0	0	0	0	0	0

The number of venues for each category in all neighbourhoods

To give us a clearer idea, the frequency of occurrence of each category is determined by taking mean of all values. The Foursquare API may return many venues since the limit is set to 100 but we can define a neighbourhood food habit by top five venues in its vicinity. A dataframe is created to show us the top five most common venues categories in each neighbourhood. It is shown as follows.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Allerton	Chinese Restaurant	Pizza Place	Mexican Restaurant	Fast Food Restaurant	Caribbean Restaurant
1	Annadale	Pizza Place	Sushi Restaurant	American Restaurant	Italian Restaurant	Japanese Restaurant
2	Arden Heights	Pizza Place	American Restaurant	Italian Restaurant	Sushi Restaurant	Mexican Restaurant
3	Arlington	Pizza Place	Peruvian Restaurant	American Restaurant	Fast Food Restaurant	Spanish Restaurant
4	Arrochar	Italian Restaurant	Pizza Place	Chinese Restaurant	Steakhouse	Middle Eastern Restaurant

Clustering data through k-means

Clustering through k-means is an unsupervised machine learning algorithm. A loose definition of clustering is that it is the process of organizing objects into groups whose members are similar in characteristics. K-means is one of the simplest unsupervised algorithms which classify a data set through a certain number of clusters. Determining the optimal number of clusters is of paramount importance in this process.

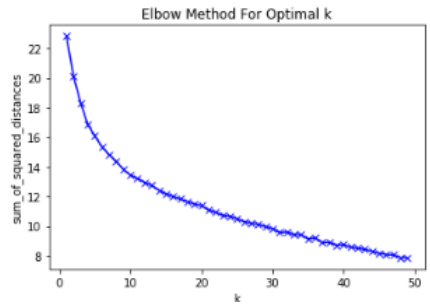
For this project, we will determine the optimal number of clusters through one of the most popular methods, i.e. The Elbow Method. It calculates the sum of squared distances of samples to their closest cluster center for different values of ‘k’. The optimal number of clusters is the value after which there is no significant decrease in the sum of squared distances. However, when this method was executed, we could not find a conclusive result as shown here.

```
In [62]: sum_of_squared_distances = []
K = range(1,50)
for k in K:
    print(k, end=' ')
    kmeans = KMeans(n_clusters=k).fit(nyc_grouped_clustering)
    sum_of_squared_distances.append(kmeans.inertia_)
```

```
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 4
6 47 48 49
```

Plotting the sum of squared distances

```
In [63]: plt.plot(K, sum_of_squared_distances, 'bx-')
plt.xlabel('k')
plt.ylabel('sum_of_squared_distances')
plt.title('Elbow Method For Optimal k');
```



Elbow Method

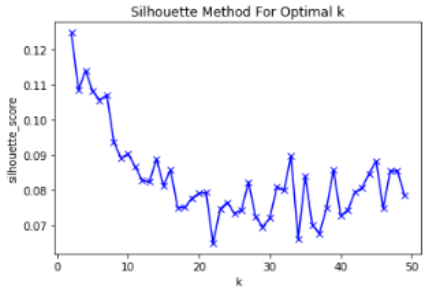
We can use another method which is known as Silhouette method. It “displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually. This measure has a range of [-1, 1]“ (Scikit-learn, 2019). Fortunately, this method worked for our data and we found the optimal number of clusters as shown below.

```
In [64]: from sklearn.metrics import silhouette_score

sil = []
K_sil = range(2,50)
# minimum 2 clusters required, to define dissimilarity
for k in K_sil:
    print(k, end=' ')
    kmeans = KMeans(n_clusters = k).fit(nyc_grouped_clustering)
    labels = kmeans.labels_
    sil.append(silhouette_score(nyc_grouped_clustering, labels, metric = 'euclidean'))
```

```
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
47 48 49
```

```
In [65]: plt.plot(K_sil, sil, 'bx-')
plt.xlabel('k')
plt.ylabel('silhouette_score')
plt.title('Silhouette Method For Optimal k')
plt.show()
```



Silhouette method

Three peaks can be seen in the given plot. These peaks are at k = 2, k = 4 and k = 8. Choosing the first two numbers may not be wise because it will result in excessively broader clusters, therefore we can choose k = 8.

We can assign the neighbourhoods to these clusters through k-means. Cluster 0 has 22 neighbourhoods, cluster 1 has 26, cluster 2 has 55, cluster 3 has 4, cluster 4 has 57, cluster 5 has 51, cluster 6 has 54 and cluster 7 has 33 neighbourhoods.

Setting the number of clusters to 8

```
kclusters = 8

# run k-means clustering
kmeans = KMeans(init="k-means++", n_clusters=kclusters, n_init=50).fit(nyc_grouped_clustering)

print(Counter(kmeans.labels_))
```

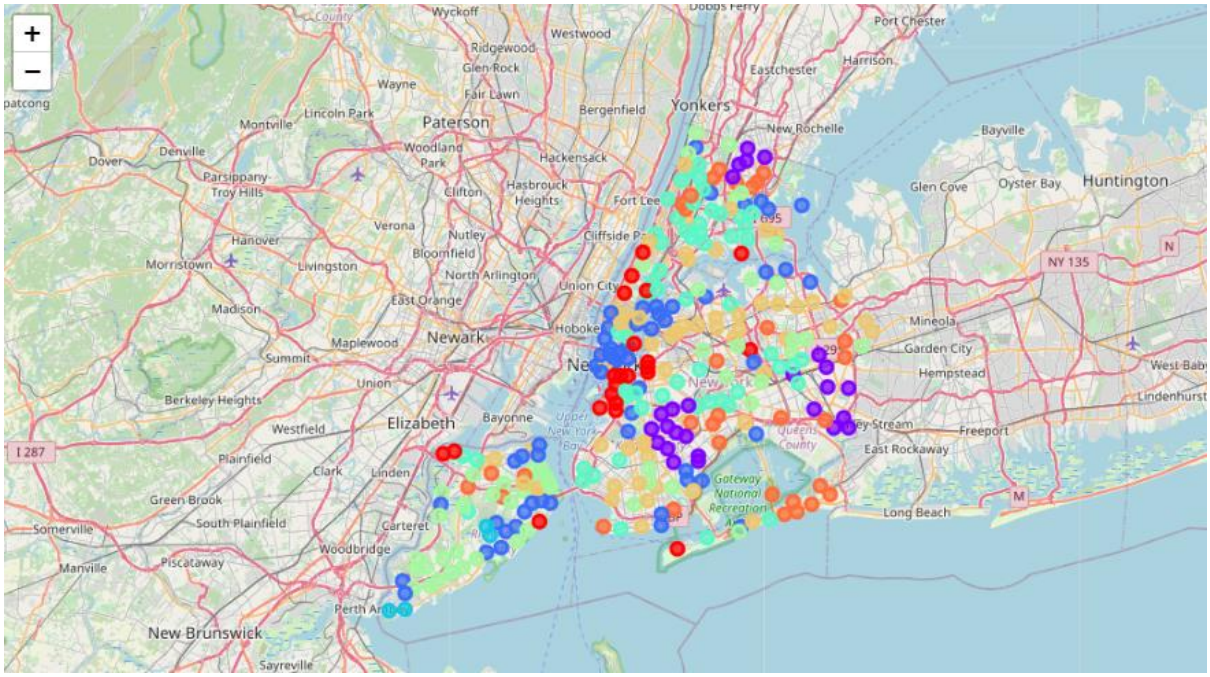
```
Counter({4: 57, 2: 55, 6: 54, 5: 51, 7: 33, 1: 26, 0: 22, 3: 4})
```

Now, we can add Borough, Latitude and Longitude for each neighbourhood to get a complete set of information regarding the venues available in different neighbourhoods. The cluster lables i.e. 0 to 7 are also added to the dataframe to get attain our objective of segmenting the neighbourhoods on the basis of most common venues in their vicinity.

```
# merging the neighborhoods_venues_sorted with nyc_data to add latitude/longitude for each neighborhood
nyc_merged = neighborhoods_venues_sorted.join(neighborhoods.set_index('Neighborhood'), on='Neighborhood')
nyc_merged.head()
```

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
0	7	Allerton	Chinese Restaurant	Pizza Place	Mexican Restaurant	Fast Food Restaurant	Caribbean Restaurant	Bronx	40.865788	-73.859319
1	5	Annadale	Pizza Place	Sushi Restaurant	American Restaurant	Italian Restaurant	Japanese Restaurant	Staten Island	40.538114	-74.178549
2	5	Arden Heights	Pizza Place	American Restaurant	Italian Restaurant	Sushi Restaurant	Mexican Restaurant	Staten Island	40.549286	-74.185887
3	4	Arlington	Pizza Place	Peruvian Restaurant	American Restaurant	Fast Food Restaurant	Spanish Restaurant	Staten Island	40.635325	-74.165104
4	2	Arrochar	Italian Restaurant	Pizza Place	Chinese Restaurant	Steakhouse	Middle Eastern Restaurant	Staten Island	40.596313	-74.067124

With the help of folium library, the neighbourhoods in different clusters can be visualized as follows.



Results

Result section will describe the outcomes for all the clusters.

Cluster 0

```
cluster_0 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 0, nyc_merged.columns[1:12]]
cluster_0.head(5)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
29	Breezy Point	American Restaurant	Pizza Place	Yemeni Restaurant	Himalayan Restaurant	Fast Food Restaurant	Queens	40.557401	-73.925512
35	Brooklyn Heights	Pizza Place	American Restaurant	Ramen Restaurant	French Restaurant	Korean Restaurant	Brooklyn	40.695864	-73.993782
44	Carroll Gardens	Pizza Place	Ramen Restaurant	French Restaurant	Seafood Restaurant	BBQ Joint	Brooklyn	40.680540	-73.994654
55	Clason Point	American Restaurant	Seafood Restaurant	South American Restaurant	Hawaiian Restaurant	English Restaurant	Bronx	40.806551	-73.854144
60	Cobble Hill	French Restaurant	Pizza Place	Korean Restaurant	American Restaurant	Japanese Restaurant	Brooklyn	40.687920	-73.998561

First five rows of the venue allocation in cluster 0

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

```
Pizza Place               12
American Restaurant        8
French Restaurant          1
Seafood Restaurant         1
Name: 1st Most Common Venue, dtype: int64
----------------------------------------------
American Restaurant           8
Pizza Place                   6
French Restaurant             1
Seafood Restaurant            1
BBQ Joint                     1
Mediterranean Restaurant      1
Mexican Restaurant            1
Italian Restaurant            1
Ramen Restaurant              1
Fast Food Restaurant          1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------------
Brooklyn           11
Manhattan           5
Staten Island       3
Queens              2
Bronx               1
Name: Borough, dtype: int64
----------------------------------------------
```

Overview of the different venues in cluster 0

Pizza Place is the most common in this cluster. It has 12 occurrences as 'First Most Common Venue' while 06 occurrences as 'Second Most Common Venue'. American Restaurant is the second most common venue in this cluster. It has 08 occurrences as 'First Most Common Venue' while 08 occurrences as 'Second Most Common Venue.' This cluster has also French Restaurant, Seafood, BBQ Joint, Mediterranean, Italian and Ramen Restaurants. It is also worth mentioning that majority of these neighbourhoods are in Brooklyn.

Cluster 1

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
36	Brookville	Fried Chicken Joint	Caribbean Restaurant	Pizza Place	Chinese Restaurant	Fast Food Restaurant	Queens	40.660003	-73.751753
41	Cambria Heights	Caribbean Restaurant	Fried Chicken Joint	Chinese Restaurant	Seafood Restaurant	Mexican Restaurant	Queens	40.692775	-73.735269
42	Canarsie	Caribbean Restaurant	Fast Food Restaurant	Chinese Restaurant	Pizza Place	Fried Chicken Joint	Brooklyn	40.635564	-73.902093
68	Crown Heights	Caribbean Restaurant	Fast Food Restaurant	Southern / Soul Food Restaurant	Fried Chicken Joint	Mexican Restaurant	Brooklyn	40.670829	-73.943291
77	East Flatbush	Caribbean Restaurant	Chinese Restaurant	Fried Chicken Joint	Pizza Place	Fast Food Restaurant	Brooklyn	40.641718	-73.936103

First five rows of the venue allocation in cluster 1

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

```
Caribbean Restaurant                  23
Fried Chicken Joint                    1
Pizza Place                            1
Southern / Soul Food Restaurant        1
Name: 1st Most Common Venue, dtype: int64
-----------------------------------------------
Fast Food Restaurant     8
Pizza Place              7
Chinese Restaurant       6
Caribbean Restaurant     3
Indian Restaurant        1
Fried Chicken Joint      1
Name: 2nd Most Common Venue, dtype: int64
-----------------------------------------------
Brooklyn     12
Queens        9
Bronx         5
Name: Borough, dtype: int64
-----------------------------------------------
```

Overview of the different venues in cluster 1

As evident, Caribbean Restaurant is by far the most common venue in this cluster. It has 23 occurrences as 'First Most Common Venue' while 03 occurrences as 'Second Most Common Venue'. Besides Caribbean Restaurant, this cluster also offer Chinese, fast food, Indian and Southern Food. Majority of these neighbourhoods are also in Brooklyn.

Cluster 2

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
4	Arrochar	Italian Restaurant	Pizza Place	Chinese Restaurant	Steakhouse	Middle Eastern Restaurant	Staten Island	40.596313	-74.067124
7	Astoria Heights	Italian Restaurant	Pizza Place	Chinese Restaurant	Greek Restaurant	Sushi Restaurant	Queens	40.770317	-73.894680
10	Battery Park City	Fast Food Restaurant	Pizza Place	Mexican Restaurant	Seafood Restaurant	Italian Restaurant	Manhattan	40.711932	-74.016869
12	Bay Terrace	Italian Restaurant	Pizza Place	American Restaurant	Chinese Restaurant	Asian Restaurant	Queens	40.782843	-73.776802
12	Bay Terrace	Italian Restaurant	Pizza Place	American Restaurant	Chinese Restaurant	Asian Restaurant	Staten Island	40.553988	-74.139166

First five rows of the venue allocation in cluster 2

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

```
Italian Restaurant       24
Pizza Place              12
Mexican Restaurant        4
Fast Food Restaurant      4
American Restaurant       3
Chinese Restaurant        3
Seafood Restaurant        2
Asian Restaurant          1
Steakhouse                1
Sushi Restaurant          1
Taco Place                1
Fried Chicken Joint       1
Name: 1st Most Common Venue, dtype: int64
----------------------------------------------------
Pizza Place                    16
Italian Restaurant             15
Mexican Restaurant              5
Fast Food Restaurant            5
Chinese Restaurant              4
American Restaurant             3
Vietnamese Restaurant           1
Seafood Restaurant              1
Korean Restaurant               1
Middle Eastern Restaurant       1
New American Restaurant         1
Sushi Restaurant                1
Thai Restaurant                 1
BBQ Joint                       1
Japanese Restaurant             1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------------------
Staten Island     16
Manhattan         16
Queens            11
Brooklyn           8
Bronx              6
Name: Borough, dtype: int64
```

Overview of the different venues in cluster 2

Italian Restaurant is the most common venue in this cluster. It has 24 occurrences as ‘First Most Common Venue’ while 15 occurrences as ‘Second Most Common Venue’. Pizza place is not far from Italian Restaurant and is the second most common venue in this cluster. It has 12 occurrences as ‘First Most Common Venue’ while 16 occurrences as ‘Second Most Common Venue.’ This cluster has also other varieties of food like Mexican, Vietnamese, Korean, Middle Eastern, Japanese and Thai food. However, this cluster is mainly suitable for those tourists who have craving for Italian food. Majority of these neighbourhoods are in Staten Island and Manhattan.

Cluster 3

```
cluster_3 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 3, nyc_merged.columns[1:12]]
cluster_3.head(4)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
40	Butler Manor	Italian Restaurant	Asian Restaurant	Fried Chicken Joint	BBQ Joint	Chinese Restaurant	Staten Island	40.506082	-74.229504
152	Lighthouse Hill	Italian Restaurant	Yemeni Restaurant	Himalayan Restaurant	Falafel Restaurant	Fast Food Restaurant	Staten Island	40.576506	-74.137927
230	Richmond Town	Italian Restaurant	Asian Restaurant	Mexican Restaurant	Pizza Place	Fast Food Restaurant	Staten Island	40.569606	-74.134057
271	Tottenville	Italian Restaurant	Mexican Restaurant	Asian Restaurant	Pizza Place	Chinese Restaurant	Staten Island	40.505334	-74.246569

First five rows of the venue allocation in cluster 3

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

```
Italian Restaurant    4
Name: 1st Most Common Venue, dtype: int64
-----------------------------------------------
Asian Restaurant      2
Yemeni Restaurant     1
Mexican Restaurant    1
Name: 2nd Most Common Venue, dtype: int64
-----------------------------------------------
Staten Island    4
Name: Borough, dtype: int64
-----------------------------------------------
```

Overview of the different venues in cluster 3

Cluster 3 has only 04 neighbourhoods. Italian Restaurant is the 'First Most Common Venue' in all the neighbourhoods while Asian Restaurant has 02 occurrences as 'Second Most Common Venue'. All the neighbourhoods in this cluster are located in Staten Island.

Cluster 4

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
3	Arlington	Pizza Place	Peruvian Restaurant	American Restaurant	Fast Food Restaurant	Spanish Restaurant	Staten Island	40.635325	-74.165104
6	Astoria	Fast Food Restaurant	Pizza Place	Ramen Restaurant	Italian Restaurant	Middle Eastern Restaurant	Queens	40.768509	-73.915654
11	Bay Ridge	Pizza Place	Fast Food Restaurant	Middle Eastern Restaurant	Mexican Restaurant	Fried Chicken Joint	Brooklyn	40.625801	-74.030621
17	Bedford Stuyvesant	Fried Chicken Joint	Fast Food Restaurant	Pizza Place	New American Restaurant	Southern / Soul Food Restaurant	Brooklyn	40.687232	-73.941785
27	Boerum Hill	Pizza Place	Fast Food Restaurant	Mediterranean Restaurant	French Restaurant	BBQ Joint	Brooklyn	40.685683	-73.983748

First five rows of the venue allocation in cluster 4

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

```
Pizza Place                31
Fast Food Restaurant       21
Fried Chicken Joint         4
Taco Place                  1
Name: 1st Most Common Venue, dtype: int64
-----------------------------------------------
Fast Food Restaurant         20
Pizza Place                  17
Fried Chicken Joint           8
Mexican Restaurant            2
American Restaurant           2
Spanish Restaurant            2
Italian Restaurant            1
Steakhouse                    1
New American Restaurant       1
Peruvian Restaurant           1
Korean Restaurant             1
Chinese Restaurant            1
Name: 2nd Most Common Venue, dtype: int64
-----------------------------------------------
Bronx            17
Brooklyn         15
Queens           11
Manhattan        11
Staten Island     3
Name: Borough, dtype: int64
```

Overview of the different venues in cluster 4

Pizza Place is the most common venue in this cluster. It has 31 occurrences as 'First Most Common Venue' while 17 occurrences as 'Second Most Common Venue'. Fast Food Restaurant has also a significant number of occurrences in this cluster and is the second most common venue in this cluster. It has 21 occurrences as 'First Most Common Venue' while 20 occurrences as 'Second Most Common Venue.' So, this cluster is suitable for those tourists who want to eat pizza or other types of fast food. Cluster 4 is the largest cluster with 57 neighbourhoods. Majority of these neighbourhoods are in Bronx.

Cluster 5

```
cluster_5 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 5, nyc_merged.columns[1:12]]
cluster_5.head(5)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
1	Annadale	Pizza Place	Sushi Restaurant	American Restaurant	Italian Restaurant	Japanese Restaurant	Staten Island	40.538114	-74.178549
2	Arden Heights	Pizza Place	American Restaurant	Italian Restaurant	Sushi Restaurant	Mexican Restaurant	Staten Island	40.549286	-74.185887
21	Bellerose	Pizza Place	Indian Restaurant	Chinese Restaurant	American Restaurant	Fast Food Restaurant	Queens	40.728573	-73.720128
26	Bloomfield	Pizza Place	Mexican Restaurant	Italian Restaurant	BBQ Joint	Yemeni Restaurant	Staten Island	40.605779	-74.187256
34	Bronxdale	Pizza Place	Italian Restaurant	Mexican Restaurant	Yemeni Restaurant	Seafood Restaurant	Bronx	40.852723	-73.861726

First five rows of the venue allocation in cluster 5

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

```
Pizza Place             49
Mexican Restaurant       1
Italian Restaurant       1
Name: 1st Most Common Venue, dtype: int64
-----------------------------------------------
Italian Restaurant       11
American Restaurant       9
Chinese Restaurant        8
Mexican Restaurant        4
Caribbean Restaurant      3
Japanese Restaurant       3
Fast Food Restaurant      3
Indian Restaurant         2
Sushi Restaurant          2
Spanish Restaurant        1
Pizza Place               1
Kosher Restaurant         1
BBQ Joint                 1
Taco Place                1
Asian Restaurant          1
Name: 2nd Most Common Venue, dtype: int64
-----------------------------------------------
Staten Island     27
Queens            13
Bronx              6
Brooklyn           4
Manhattan          1
Name: Borough, dtype: int64
-----------------------------------------------
```

Overview of the different venues in cluster 5

Pizza Place is by far the most common venue in this cluster. It has 49 occurrences as 'First Most Common Venue' while 01 occurrence as 'Second Most Common Venue'. Apart from Pizza Place, this cluster also offers a variety of other cuisines like American, Chinese, Mexican, Caribbean, Indian, to name a few. Over half of these neighbourhoods are located in Staten Island.

Cluster 6

```
cluster_6 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 6, nyc_merged.columns[1:12]]
cluster_6.head(5)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
8	Auburndale	Korean Restaurant	Pizza Place	Greek Restaurant	Italian Restaurant	Chinese Restaurant	Queens	40.761730	-73.791762
9	Bath Beach	Cantonese Restaurant	Chinese Restaurant	Vietnamese Restaurant	Fast Food Restaurant	Pizza Place	Brooklyn	40.599519	-73.998752
14	Bayside	Korean Restaurant	Pizza Place	Fast Food Restaurant	Asian Restaurant	Thai Restaurant	Queens	40.766041	-73.774274
23	Bensonhurst	Cantonese Restaurant	Chinese Restaurant	Asian Restaurant	Italian Restaurant	Sushi Restaurant	Brooklyn	40.611009	-73.995180
25	Blissville	Pizza Place	Chinese Restaurant	Mexican Restaurant	Italian Restaurant	Fast Food Restaurant	Queens	40.737251	-73.932442

First five rows of the venue allocation in cluster 6

```
Pizza Place                          22
Fast Food Restaurant                  6
Mexican Restaurant                    5
Korean Restaurant                     5
Indian Restaurant                     4
Chinese Restaurant                    4
Cantonese Restaurant                  2
Latin American Restaurant             2
Spanish Restaurant                    1
Seafood Restaurant                    1
New American Restaurant               1
Fried Chicken Joint                   1
Sushi Restaurant                      1
Eastern European Restaurant           1
Name: 1st Most Common Venue, dtype: int64
----------------------------------------------------
Chinese Restaurant                   16
Pizza Place                           9
Fast Food Restaurant                  7
Mexican Restaurant                    3
Caribbean Restaurant                  2
Fried Chicken Joint                   2
Thai Restaurant                       2
Italian Restaurant                    2
Japanese Restaurant                   2
Seafood Restaurant                    1
Indian Restaurant                     1
Middle Eastern Restaurant             1
Spanish Restaurant                    1
American Restaurant                   1
Sushi Restaurant                      1
Filipino Restaurant                   1
Empanada Restaurant                   1
Russian Restaurant                    1
Korean Restaurant                     1
Vegetarian / Vegan Restaurant         1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------------------
Queens            21
Brooklyn          14
Bronx              8
Manhattan          7
Staten Island      6
Name: Borough, dtype: int64
```

Overview of the different venues in cluster 6

Pizza Place is the most common venue in this cluster as well. It has 22 occurrences as 'First Most Common Venue' while 09 occurrence as 'Second Most Common Venue'. Chinese Restaurant is the second most common venue with 16 occurrences as the 'Second Most Common Venue.' Apart from these, this cluster also offers a variety of other cuisines like Thai, Middle Eastern, Japanese, Caribbean, Korean and many others. Majority of the neighbourhoods are located in Queens.

Cluster 7

```
cluster_7 = nyc_merged.loc[nyc_merged['Cluster Labels'] == 7, nyc_merged.columns[1:12]]
cluster_7.head(5)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Borough	Latitude	Longitude
0	Allerton	Chinese Restaurant	Pizza Place	Mexican Restaurant	Fast Food Restaurant	Caribbean Restaurant	Bronx	40.865788	-73.859319
5	Arverne	Chinese Restaurant	Pizza Place	Thai Restaurant	American Restaurant	Asian Restaurant	Queens	40.589144	-73.791992
13	Baychester	Pizza Place	Chinese Restaurant	Seafood Restaurant	BBQ Joint	Caribbean Restaurant	Bronx	40.866858	-73.835798
15	Bayswater	Chinese Restaurant	Fast Food Restaurant	Fried Chicken Joint	Pizza Place	Caribbean Restaurant	Queens	40.611322	-73.765968
16	Bedford Park	Pizza Place	Chinese Restaurant	Fast Food Restaurant	Mexican Restaurant	Fried Chicken Joint	Bronx	40.870185	-73.885512

First five rows of the venue allocation in cluster 7

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

```
Chinese Restaurant      16
Pizza Place             15
Fast Food Restaurant     2
Name: 1st Most Common Venue, dtype: int64
----------------------------------------------
Chinese Restaurant      15
Pizza Place             10
Fast Food Restaurant     3
Fried Chicken Joint      2
Caribbean Restaurant     2
American Restaurant      1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------------
Queens            14
Bronx              9
Brooklyn           6
Staten Island      4
Name: Borough, dtype: int64
----------------------------------------------
```

Overview of the different venues in cluster 7

Chinese Restaurant is the most common venue in this cluster. It has 16 occurrences as 'First Most Common Venue' while 15 occurrences as 'Second Most Common Venue'. Pizza Place is the second most common venue with 15 occurrences as the 'First Most Common Venue' and 10 most common occurrences as 'Second Most Common Venue.' Majority of the neighbourhoods are located in Queens.

Discussion

K-mean clustering algorithm is used to segment New York City's neighbourhoods. The optimal number of clusters is eight for this project as found out through 'The Silhouette Method.' Each of these clusters is analysed through determining the borough count, count of 'First Most Common Venue' and 'Second Most Common Venue.' Results for the eights clusters are summarized in the following table.

	Occurrences' Count within the Cluster		
Cluster	1st Most Common Venue	2nd Most Common Venue	Borough
0	Pizza Place	American Restaurant	Brooklyn
1	Caribbean Restaurant	Fast Food Restaurant	Brooklyn, Queens
2	Italian Restaurant	Pizza Place	Staten Island, Manhattan
3	Italian Restaurant	Asian Restaurant	Staten Island
4	Pizza Place	Fast Food Restaurant	Bronx, Brooklyn
5	Pizza Place	Italian Restaurant	Staten Island, Queens
6	Pizza Place	Chinese Restaurant	Queens, Brooklyn
7	Chinese Restaurant	Pizza Place	Queens, Bronx

As evident from the table that pizza is the most common food in New York City. Since it is considered as go-to food, we can set it aside and regard Italian food as the most common cuisine. Caribbean, Chinese, Asian and American foods are also common in different areas.

Conclusion

This project is aimed at exploring the boroughs of New York City to locate the most common food options in different neighbourhoods. Data is collected from 'New York City Dataset' and it is processed and cleaned through different python libraries. K-mean clustering algorithm is used to segment the data into clusters and after analysis, we find out the most common cuisines in different parts of the city. Pizza is the most common food but keeping in view the ubiquitous nature of this and other fast foods, it is set aside, and conclusion is drawn on the basis of most common foods next to pizza. It is evident from the analysis that American, Italian, Caribbean, Chinese and Asian cuisines are most common in the city.

Since this project used a dataset from 2018, the results can be improved by compiling a dataset based on the recent data. Detailed information can also be gained through exploring individual cuisines and boroughs by compiling a dataset through the Foursquare API. In this way, highly rated restaurants for a particular cuisine in a borough can be highlighted which will help the locals and tourists to make informed decision.

References

Food Reference (2020). Food Quotes. John Corry. NY Times Reporter.

NYC and Company (2020). Travel and Tourism trend report. https://indd.adobe.com/view/e91e777a-c68b-4db1-a609-58664a52cffd

NYC Mayor Office (2018). State of our immigrant city. Annual Report-2018. https://www1.nyc.gov/assets/immigrants/downloads/pdf/moia_annual_report_2018_final.pdf

Scikit-learn (2019). Selecting the number of clusters with Silhoutte analysis on KMeans clustering. Retrieved from https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

World Urban Areas (2018). Demographia World Urban Areas. http://demographia.com/db-worldua.pdf