IBM – COURSERA
DATA SCIENCE SPECIALIZATION

CAPSTONE PROJECT – FINAL REPORT
The Battle of the Neighborhoods

~ Nandini Gupta

Table of content:

Contents

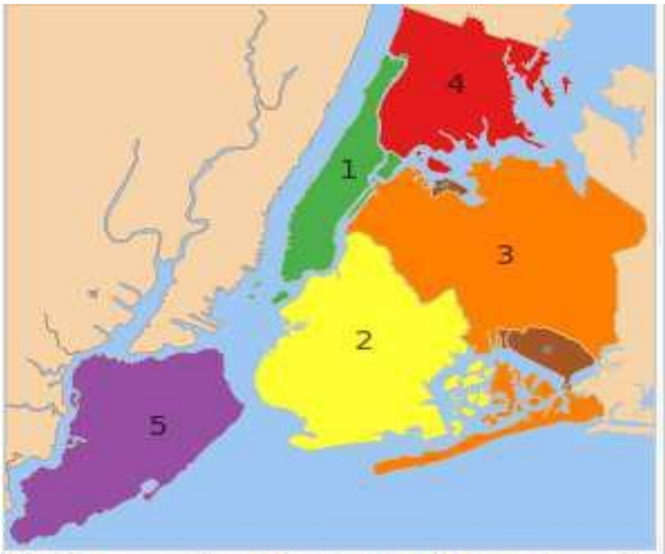
INTRODUCTION

The City of New York, usually called either New York City (NYC) or simply New York (NY), is the most populous city in the United States. With an estimated 2019 population of 8,336,817 distributed over a land area of about 302.6 square miles (784 km^2),

It is diverse and is the financial capital of USA. It is multicultural. It provides lot of business opportunities and business friendly environment. It has attracted many different players into the market. It is a global hub of business and commerce. The city is a major center for banking and 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. This also means that the market is highly 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 give good understanding of the business environment, which help in strategically targeting the market. This will help in reduction of risk and better control on the Return on Investment

New York is also the most densely populated major city in the United States. Located at the southern tip of the state of New York. A global power city, New York City has been described as the cultural, financial, and media capital of the world, and exerts a significant impact upon commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports.

NY is split up into five boroughs: the Bronx, Brooklyn, Manhattan, Queens, and Staten Island. Each borough has the same boundaries as a county of the state.



The five boroughs: 1: Manhattan, 2: Brooklyn, 3: Queens, 4: The Bronx, 5: Staten Island

BUSINESS PROBLEM

The City of New York is famous for it’s excellent cuisine. It's food culture includes an array of international cuisines influenced by the city's immigrant history. Italian & Indian restaurants have become so popular in the United States now it seems that there is one on every corner, not only in major cities but also in smaller cities. One of my friends who is thinking of starting a restaurant in the NY neighborhood, consulted with me to get some analysis done with the all-possible data available. Manhattan being the costliest place, it was decided to compare rest of the boroughs and pick one of the most suitable neighborhoods with in the shortlisted boroughs. Based on the data analysis, it is expected to logically conclude which restaurant type (Italian Or Indian) and its recommended location. All the choices to be rationalized with the data analysis
& it helps to distinguish the selections, securing long-term success.

Overall Problem Statement can be broken into the following

Exploring the Boroughs in NY and narrow down to one.

Explore the Venues in the neighborhoods across that specific Borough

Narrow down to handful of neighborhoods and then deep dive into the current Restaurants & Hotels landscape across those.

Venue clustering by filtered neighborhoods and analyze the best choice of the restaurant and the best fit location.

TARGET AUDIENCE

Any Business Entrepreneurs or Companies who would like to start a Restaurant business in NewYork. The objective is to narrow down to best possible, affordable neighborhood to start a restaurant. The model also look at picking a type of restaurants from multiple choices like Italian Vs Indian. The Solution is expected to rationalize the choices backed up with data and its analysis. For this project, all boroughs except Manhattan being considered due to high cost.

SOLUTION DESIGN APPROACH

Solution is approached in seven steps as listed below

STEP 1: Pull all the boroughs & the respective neighborhood details of the New York data using newyork_data.json.['newyork_data.json' - https://cocl.us/new_york_dataset]

STEP 2: Deep Dive into the shortlisted Borough from Step 1 Using FourSquare APIs

STEP 3: Explore Venues across the neighborhoods in that Borough & Narrow down to handful of it based on larger number of Venues Vs less number of Restaurants +Hotels

STEP 4: Deep Dive into the shortlisted neighborhoods using, Word Cloud, Means of frequency of each category of Restaurants & identifying the Top5 Common Restaurants/Hotels

STEP 5: Clustering the neighborhood using K-means & identifying the locations on the Map.

STEP 6: Concluding the Choices of Restaurants & Locations basis of the data analysis in Step

SUCCESS CRITERIA

The success criteria of this project will be a good recommendation of borough/neighborhood for the choice of a restaurant, to the Stakeholder from the Target Audience. All choices and recommendations should be rationalized with the data analysis and inferences made.

DATA

One City will be analyzed in this project : NewYork USA .

Data sources that's been analyzed in the projects are

Data1 : NewYork has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.

Data Source : newyork_data.json' https://cocl.us/new_york_dataset

Data 2 : To Narrow down to one of the boroughs , basis of population /density analysis of the data available in Wikipedia

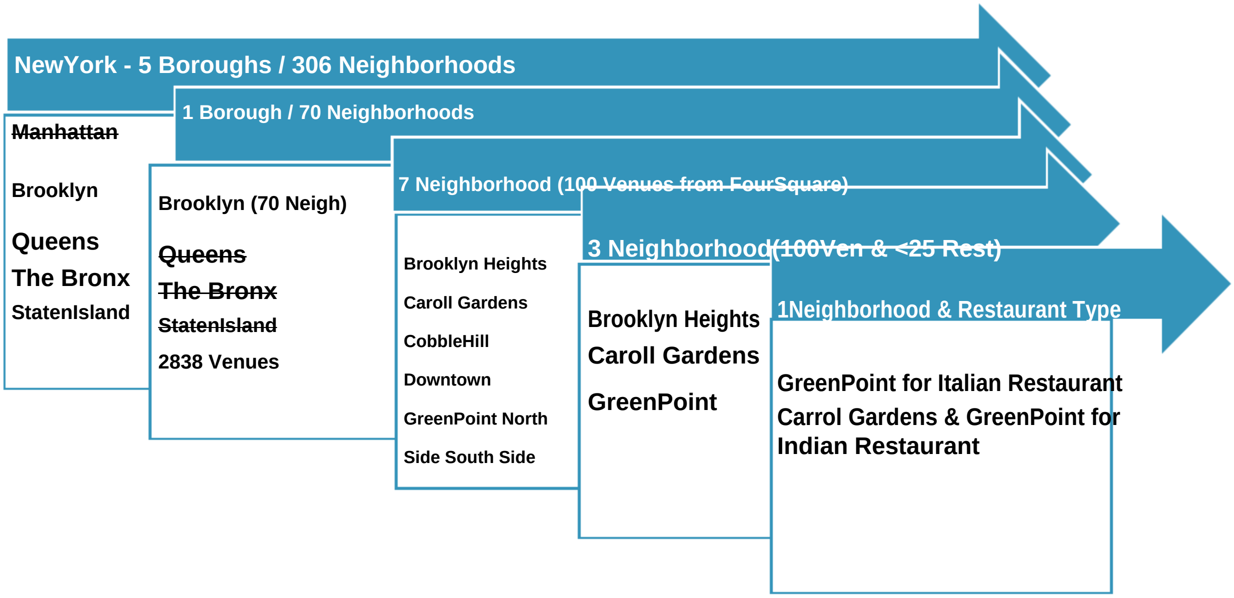
Data Source : https://en.wikipedia.org/wiki/Demographics_of_New_York_City

Data3 : Exploring the neighborhoods in one of the shortlisted boroughs using FourSquare APIS

METHODOLOGY

ANALYTIC APPROACH

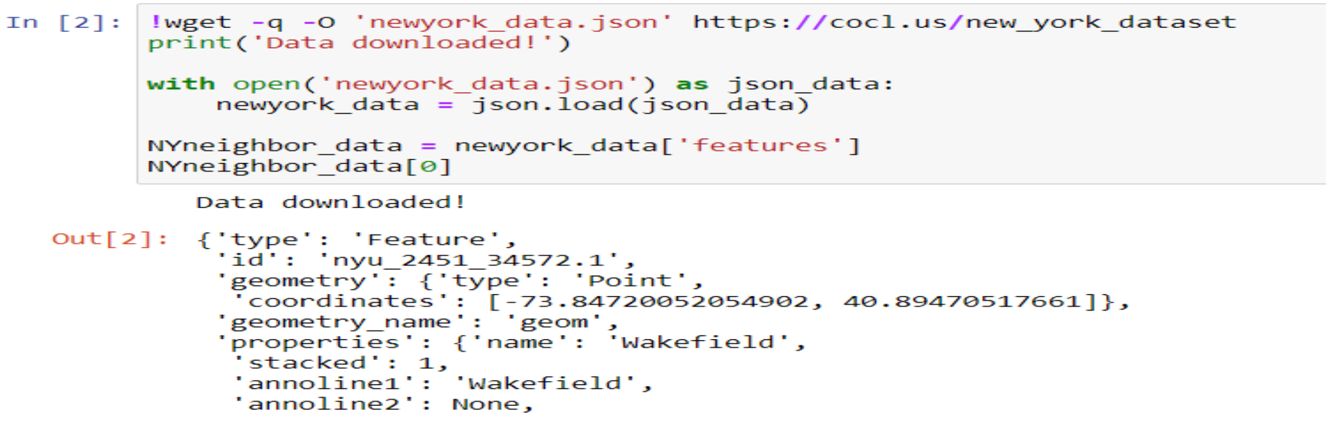
New York city neighborhood has a total of 5 boroughs and 306 neighborhoods. In this project we excluded Manhattan due to high cost and focus only on the rest of the 4 boroughs. From 300 + Neighborhoods across all the boroughs, we have applied the following analytic approach to narrow down to 3 Neighborhood in Brooklyn through multiple data exploratory analysis as explained below.


NewYork - 5 Boroughs / 306 Neighborhoods
~~Manhattan~~
Brooklyn
Queens
The Bronx
StatenIsland
1 Borough / 70 Neighborhoods
Brooklyn (70 Neigh)
~~Queens~~
~~The Bronx~~
~~StatenIsland~~
2838 Venues
7 Neighborhood (100 Venues from FourSquare)
Brooklyn Heights
Caroll Gardens
CobbleHill
Downtown
GreenPoint North
Side South Side
3 Neighborhood(100Ven & <25 Rest)
Brooklyn Heights
Caroll Gardens
GreenPoint
1Neighborhood & Restaurant Type
GreenPoint for Italian Restaurant
Carrol Gardens & GreenPoint for Indian Restaurant

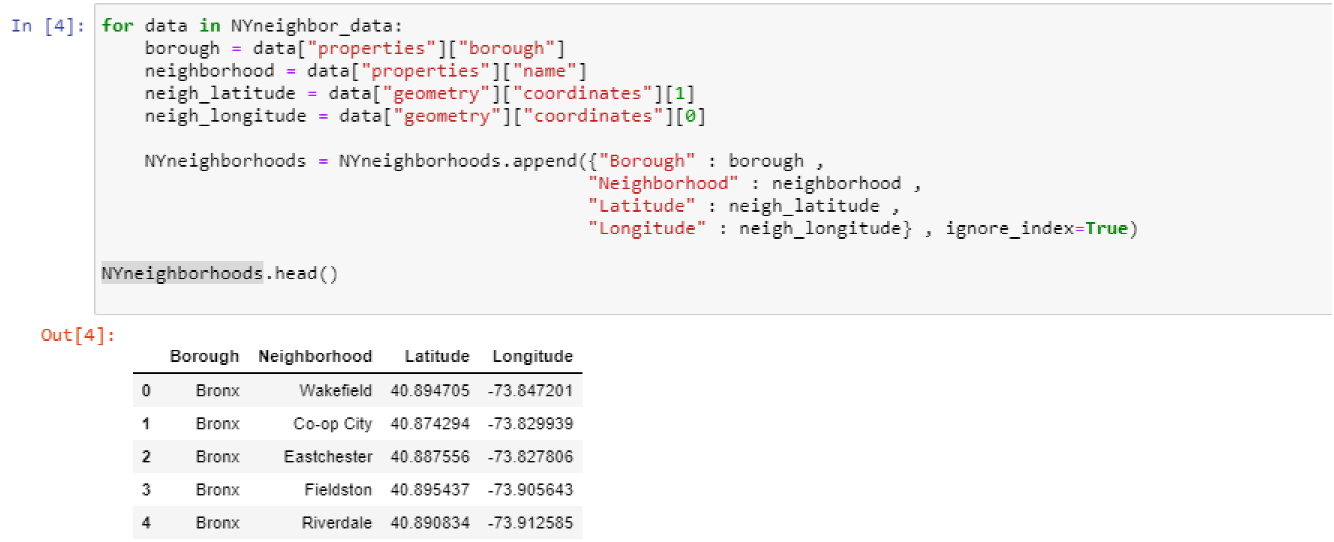
DATA EXPLORATORY ANALYSIS

Solution is approached in seven-step data exploratory analysis as explained below

STEP 1: Pull all the boroughs & the respective neighborhood details of the New York data using newyork_data.json.['newyork_data.json' - https://cocl.us/new_york_dataset]

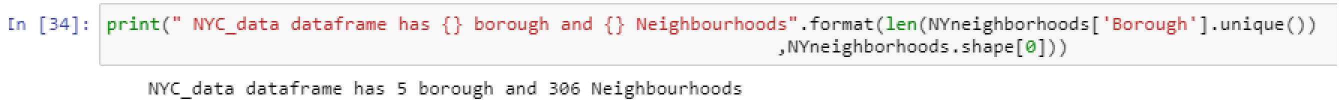


```
In [2]: !wget -q -O 'newyork_data.json' https://cocl.us/new_york_dataset
        print('Data downloaded!')

        with open('newyork_data.json') as json_data:
            newyork_data = json.load(json_data)

        NYneighbor_data = newyork_data['features']
        NYneighbor_data[0]

        Data downloaded!

Out[2]: {'type': 'Feature',
         'id': 'nyu_2451_34572.1',
         'geometry': {'type': 'Point',
          'coordinates': [-73.84720052054902, 40.89470517661]},
         'geometry_name': 'geom',
         'properties': {'name': 'Wakefield',
          'stacked': 1,
          'annoline1': 'Wakefield',
          'annoline2': None,
```



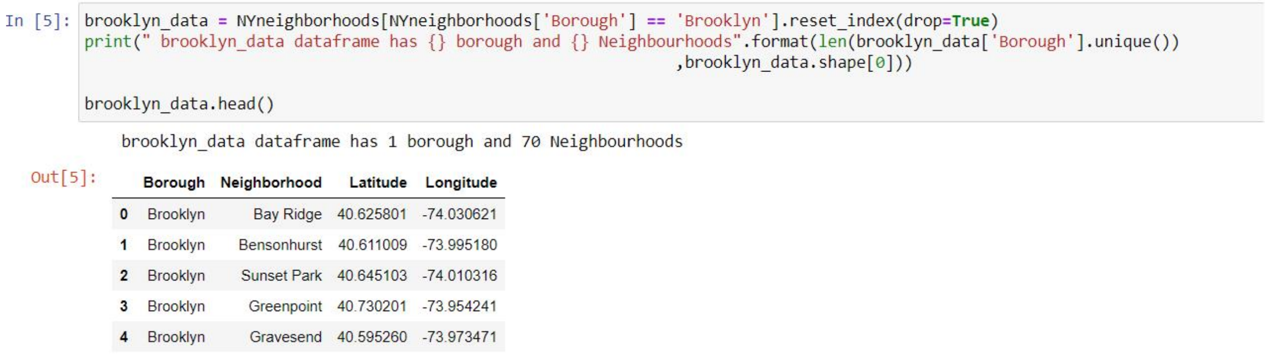
```
In [4]: for data in NYneighbor_data:
            borough = data["properties"]["borough"]
            neighborhood = data["properties"]["name"]
            neigh_latitude = data["geometry"]["coordinates"][1]
            neigh_longitude = data["geometry"]["coordinates"][0]

            NYneighborhoods = NYneighborhoods.append({"Borough" : borough ,
                                                   "Neighborhood" : neighborhood ,
                                                   "Latitude" : neigh_latitude ,
                                                   "Longitude" : neigh_longitude} , ignore_index=True)

        NYneighborhoods.head()

Out[4]:
```

	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



```
In [34]: print(" NYC_data dataframe has {} borough and {} Neighbourhoods".format(len(NYneighborhoods['Borough'].unique())
                                                                               ,NYneighborhoods.shape[0]))

          NYC_data dataframe has 5 borough and 306 Neighbourhoods
```

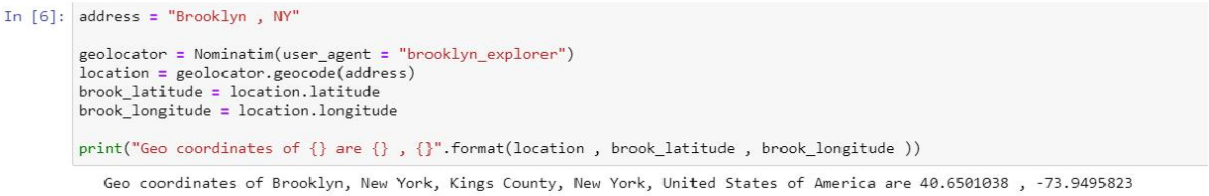
STEP 2: Deep Dive into the shortlisted Borough from Step 1 Using FourSquare APIs



```
In [5]: brooklyn_data = NYneighborhoods[NYneighborhoods['Borough'] == 'Brooklyn'].reset_index(drop=True)
        print(" brooklyn_data dataframe has {} borough and {} Neighbourhoods".format(len(brooklyn_data['Borough'].unique())
                                                                      ,brooklyn_data.shape[0]))

        brooklyn_data.head()

         brooklyn_data dataframe has 1 borough and 70 Neighbourhoods
```

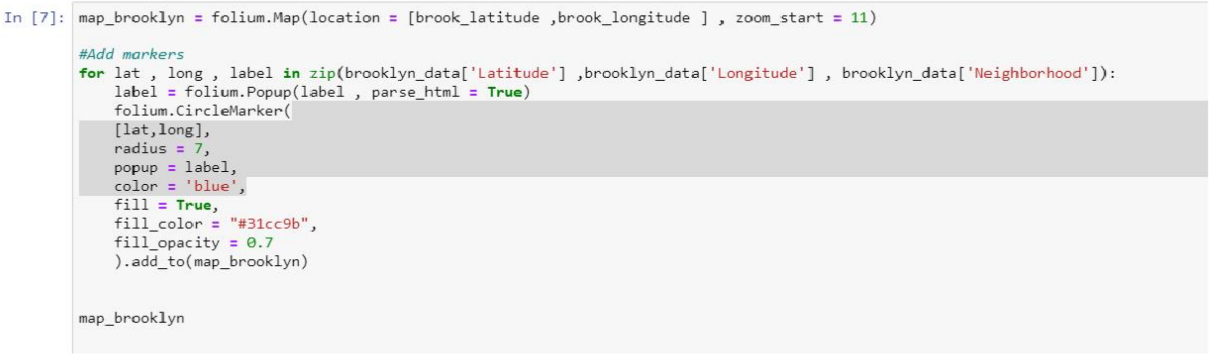
Out[5]:

	Borough	Neighborhood	Latitude	Longitude
0	Brooklyn	Bay Ridge	40.625801	-74.030621
1	Brooklyn	Bensonhurst	40.611009	-73.995180
2	Brooklyn	Sunset Park	40.645103	-74.010316
3	Brooklyn	Greenpoint	40.730201	-73.954241
4	Brooklyn	Gravesend	40.595260	-73.973471

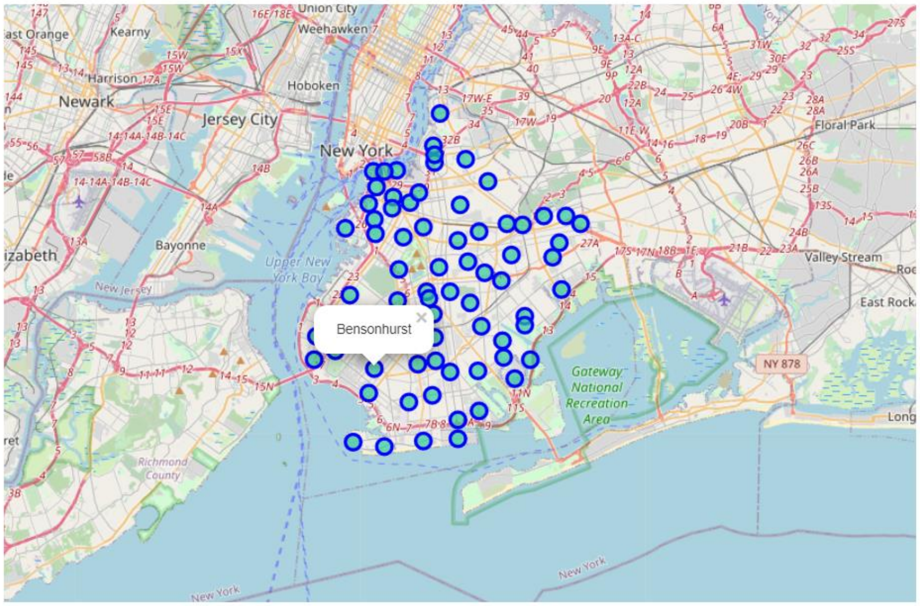


```
In [6]: address = "Brooklyn , NY"

        geolocator = Nominatim(user_agent = "brooklyn_explorer")
        location = geolocator.geocode(address)
        brook_latitude = location.latitude
        brook_longitude = location.longitude

        print("Geo coordinates of {} are {} , {}".format(location , brook_latitude , brook_longitude ))

           Geo coordinates of Brooklyn, New York, Kings County, New York, United States of America are 40.6501038 , -73.9495823
```

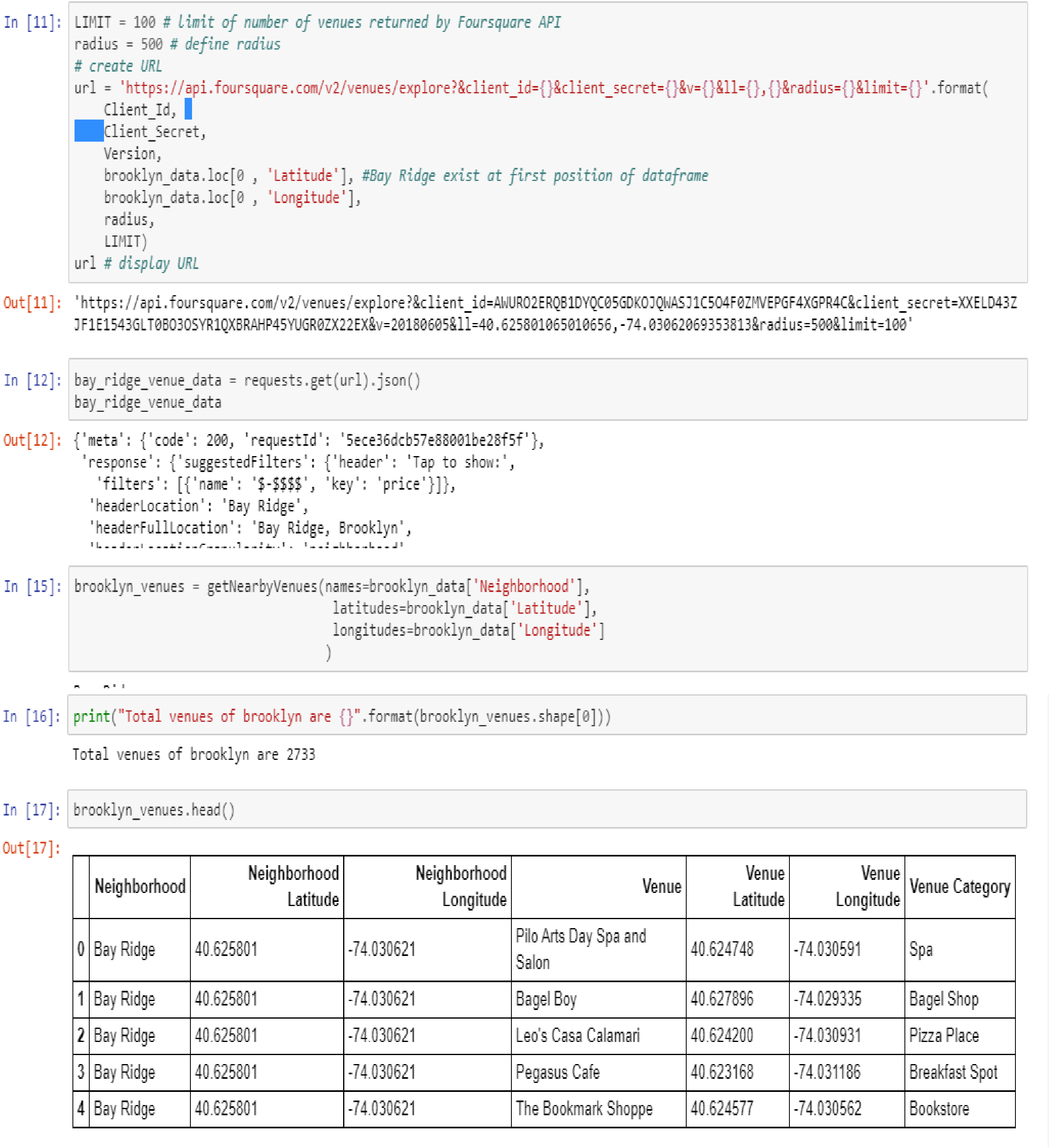
plotting map of brooklyn along with all neighbors



```
In [7]: map_brooklyn = folium.Map(location = [brook_latitude ,brook_longitude ] , zoom_start = 11)

        #Add markers
        for lat , long , label in zip(brooklyn_data['Latitude'] ,brooklyn_data['Longitude'] , brooklyn_data['Neighborhood']):
            label = folium.Popup(label , parse_html = True)
            folium.CircleMarker(
            [lat,long],
            radius = 7,
            popup = label,
            color = 'blue',
            fill = True,
            fill_color = "#31cc9b",
            fill_opacity = 0.7
            ).add_to(map_brooklyn)


        map_brooklyn
```



STEP 3: Explore Venues across the neighborhoods in that Borough & Narrow down to handful of it based on larger number of Venues Vs less number of Restaurants +Hotels



In [11]:
```
LIMIT = 100 # limit of number of venues returned by Foursquare API
radius = 500 # define radius
# create URL
url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    Client_Id, 
    Client_Secret,
    Version,
    brooklyn_data.loc[0 , 'Latitude'], #Bay Ridge exist at first position of dataframe
    brooklyn_data.loc[0 , 'Longitude'],
    radius,
    LIMIT)
url # display URL
```

Out[11]:
```
'https://api.foursquare.com/v2/venues/explore?&client_id=AWURO2ERQB1DYQC05GDKOJQWASJ1C5O4F0ZMVEPGF4XGPR4C&client_secret=XXELD43Z
JF1E1543GLT0BO3OSYR1QXBRAHP45YUGR0ZX22EX&v=20180605&ll=40.625801065010656,-74.03062069353813&radius=500&limit=100'
```

In [12]:
```
bay_ridge_venue_data = requests.get(url).json()
bay_ridge_venue_data
```

Out[12]:
```
{'meta': {'code': 200, 'requestId': '5ece36dcb57e88001be28f5f'},
 'response': {'suggestedFilters': {'header': 'Tap to show:',
   'filters': [{'name': '$-$$$$', 'key': 'price'}]},
  'headerLocation': 'Bay Ridge',
  'headerFullLocation': 'Bay Ridge, Brooklyn',
  'headerLocationGranularity': 'neighborhood',
```

In [15]:
```
brooklyn_venues = getNearbyVenues(names=brooklyn_data['Neighborhood'],
                                   latitudes=brooklyn_data['Latitude'],
                                   longitudes=brooklyn_data['Longitude']
                                  )
```

In [16]:
```
print("Total venues of brooklyn are {}".format(brooklyn_venues.shape[0]))
```

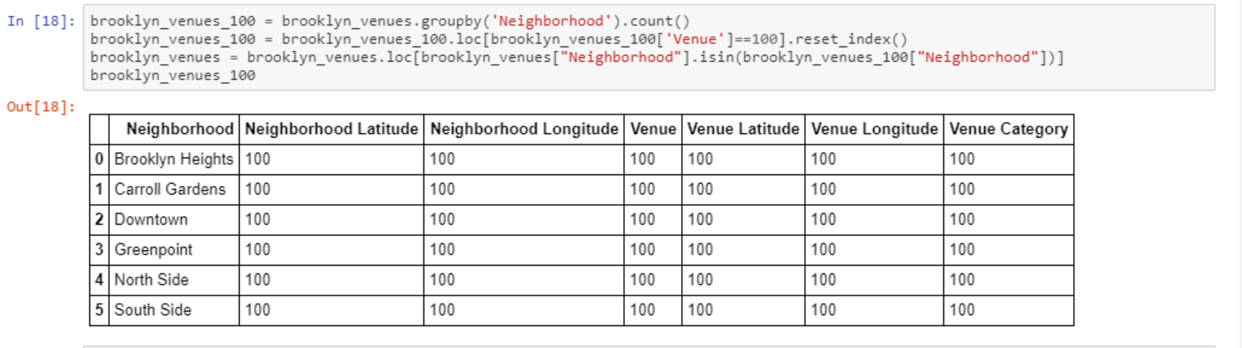
```
Total venues of brooklyn are 2733
```

In [17]:
```
brooklyn_venues.head()
```

Out[17]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bay Ridge	40.625801	-74.030621	Pilo Arts Day Spa and Salon	40.624748	-74.030591	Spa
1	Bay Ridge	40.625801	-74.030621	Bagel Boy	40.627896	-74.029335	Bagel Shop
2	Bay Ridge	40.625801	-74.030621	Leo's Casa Calamari	40.624200	-74.030931	Pizza Place
3	Bay Ridge	40.625801	-74.030621	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot
4	Bay Ridge	40.625801	-74.030621	The Bookmark Shoppe	40.624577	-74.030562	Bookstore

FILTERING NEIGHBORHOODS HAVING 100 VENUES

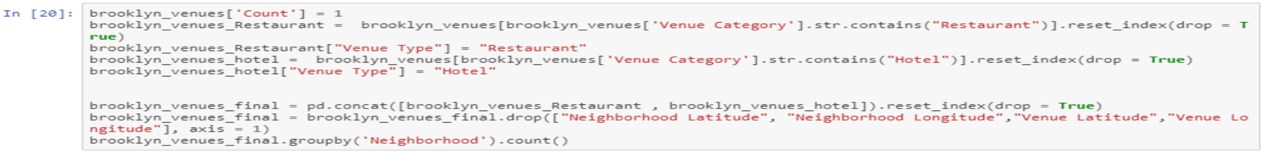


In [18]:
```
brooklyn_venues_100 = brooklyn_venues.groupby('Neighborhood').count()
brooklyn_venues_100 = brooklyn_venues_100.loc[brooklyn_venues_100['Venue']==100].reset_index()
brooklyn_venues = brooklyn_venues.loc[brooklyn_venues["Neighborhood"].isin(brooklyn_venues_100["Neighborhood"])]
brooklyn_venues_100
```

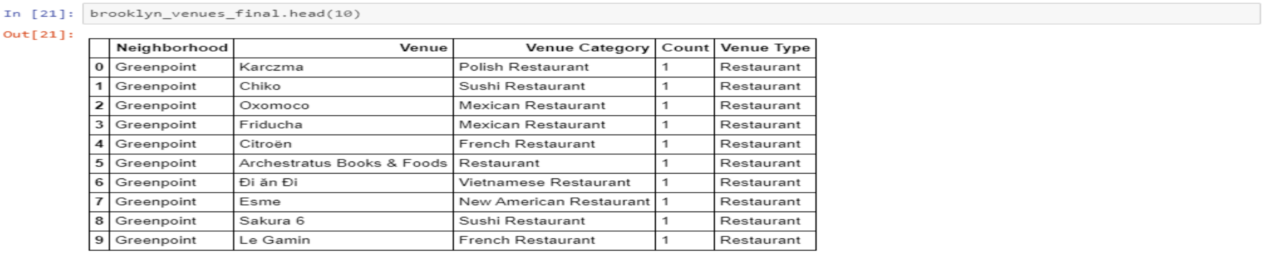
Out[18]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Brooklyn Heights	100	100	100	100	100	100
1	Carroll Gardens	100	100	100	100	100	100
2	Downtown	100	100	100	100	100	100
3	Greenpoint	100	100	100	100	100	100
4	North Side	100	100	100	100	100	100
5	South Side	100	100	100	100	100	100

FOCUSSING ON THE “RESTAURANTS & HOTELS” IN THE VENUE CATEGORY



In [20]:
```
brooklyn_venues['Count'] = 1
brooklyn_venues_Restaurant =  brooklyn_venues[brooklyn_venues['Venue Category'].str.contains("Restaurant")].reset_index(drop = T
rue)
brooklyn_venues_Restaurant["Venue Type"] = "Restaurant"
brooklyn_venues_hotel =  brooklyn_venues[brooklyn_venues['Venue Category'].str.contains("Hotel")].reset_index(drop = True)
brooklyn_venues_hotel["Venue Type"] = "Hotel"


brooklyn_venues_final = pd.concat([brooklyn_venues_Restaurant , brooklyn_venues_hotel]).reset_index(drop = True)
brooklyn_venues_final = brooklyn_venues_final.drop(["Neighborhood Latitude", "Neighborhood Longitude","Venue Latitude","Venue Lo
ngitude"], axis = 1)
brooklyn_venues_final.groupby('Neighborhood').count()
```



In [21]:
```
brooklyn_venues_final.head(10)
```

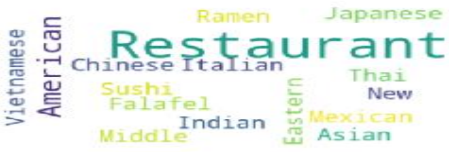
Out[21]:

	Neighborhood	Venue	Venue Category	Count	Venue Type
0	Greenpoint	Karczma	Polish Restaurant	1	Restaurant
1	Greenpoint	Chiko	Sushi Restaurant	1	Restaurant
2	Greenpoint	Oxomoco	Mexican Restaurant	1	Restaurant
3	Greenpoint	Friducha	Mexican Restaurant	1	Restaurant
4	Greenpoint	Citroën	French Restaurant	1	Restaurant
5	Greenpoint	Archestratus Books & Foods	Restaurant	1	Restaurant
6	Greenpoint	Đi ăn Đi	Vietnamese Restaurant	1	Restaurant
7	Greenpoint	Esme	New American Restaurant	1	Restaurant
8	Greenpoint	Sakura 6	Sushi Restaurant	1	Restaurant
9	Greenpoint	Le Gamin	French Restaurant	1	Restaurant

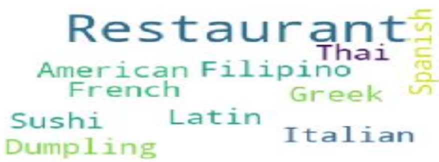
STEP 4: Deep Dive into the shortlisted neighborhoods using, Word Cloud, Means of frequency of each category of Restaurants & identifying the Top5 Common Restaurants/Hotels

a) WORD CLOUD to look at the Restaurant Types among the Seven Neighborhoods

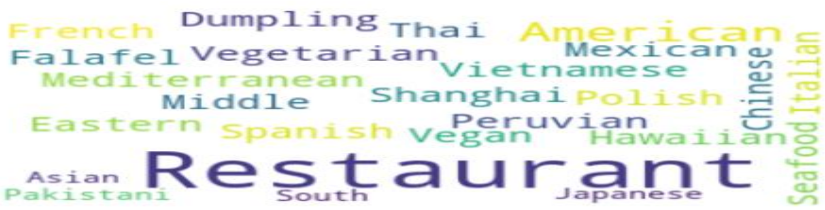
Analyziing Brooklyn Heights Neighborhood



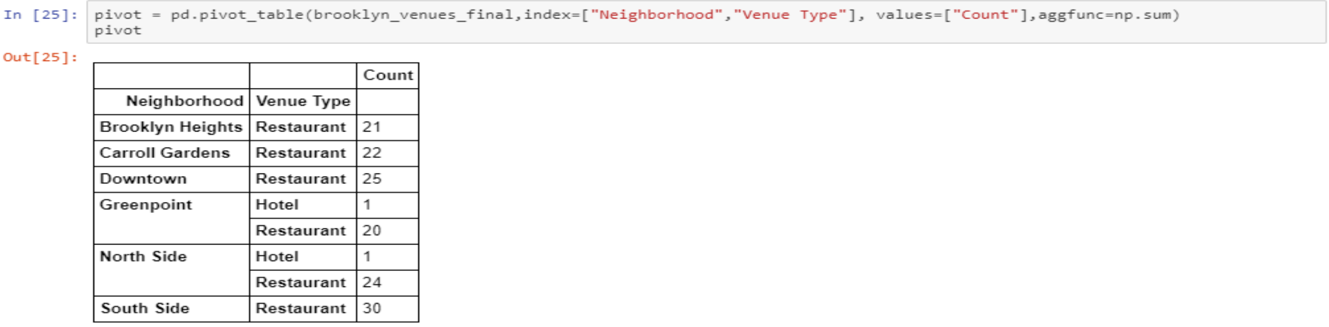
--------- Analyziing Carroll Gardens Neighborhood ----------



------------- Analyziing Downtown Neighborhood -------------



b) PIVOT to Look at the Less Restaurants/Hotels Venues with in the shortlisted 7 Neighborhoods

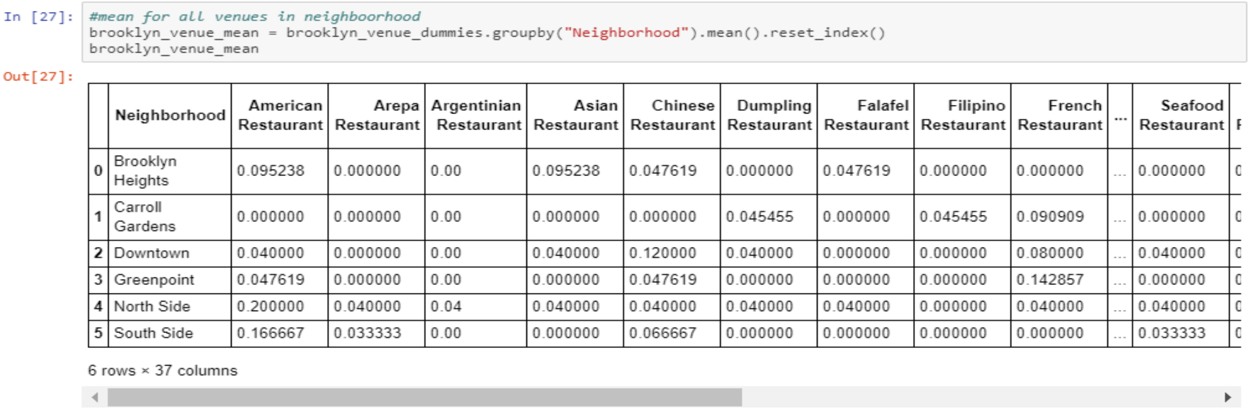


In [25]:
```
pivot = pd.pivot_table(brooklyn_venues_final,index=["Neighborhood","Venue Type"], values=["Count"],aggfunc=np.sum)
pivot
```

Out[25]:

		Count
Neighborhood	Venue Type	
Brooklyn Heights	Restaurant	21
Carroll Gardens	Restaurant	22
Downtown	Restaurant	25
Greenpoint	Hotel	1
	Restaurant	20
North Side	Hotel	1
	Restaurant	24
South Side	Restaurant	30

c) Grouping the Neighborhood Using Means of Frequency of each Category



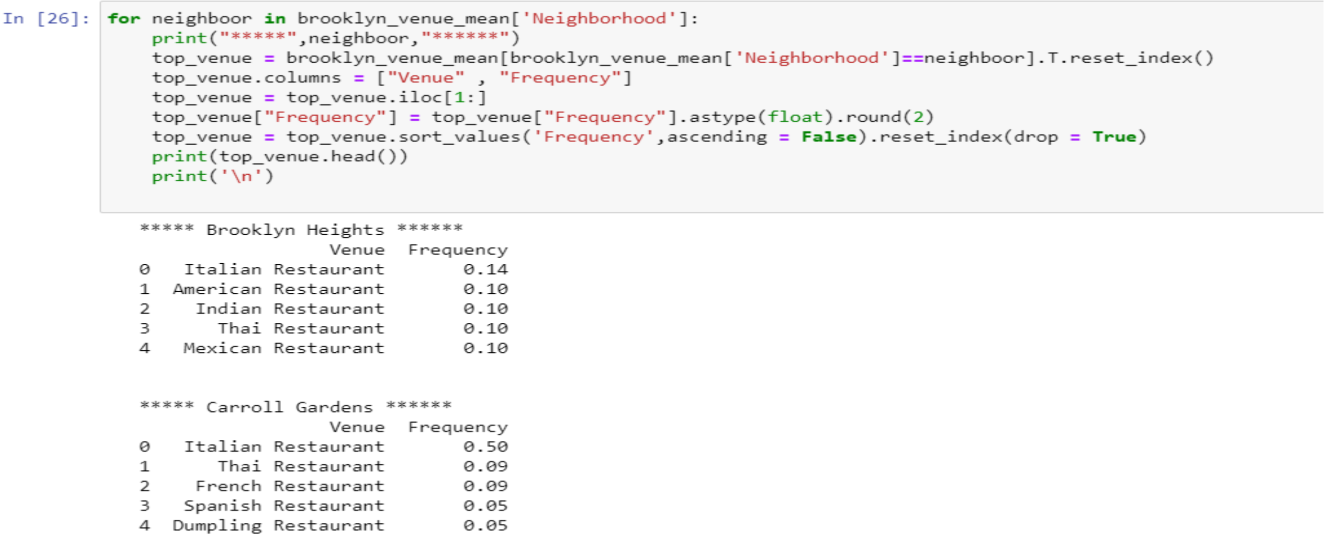
In [27]:
```
#mean for all venues in neighboorhood
brooklyn_venue_mean = brooklyn_venue_dummies.groupby("Neighborhood").mean().reset_index()
brooklyn_venue_mean
```

Out[27]:

	Neighborhood	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Chinese Restaurant	Dumpling Restaurant	Falafel Restaurant	Filipino Restaurant	French Restaurant	...	Seafood Restaurant	F
0	Brooklyn Heights	0.095238	0.000000	0.00	0.095238	0.047619	0.000000	0.047619	0.000000	0.000000	...	0.000000	0
1	Carroll Gardens	0.000000	0.000000	0.00	0.000000	0.000000	0.045455	0.000000	0.045455	0.090909	...	0.000000	0
2	Downtown	0.040000	0.000000	0.00	0.040000	0.120000	0.040000	0.000000	0.000000	0.080000	...	0.040000	0
3	Greenpoint	0.047619	0.000000	0.00	0.000000	0.047619	0.000000	0.000000	0.000000	0.142857	...	0.000000	0
4	North Side	0.200000	0.040000	0.04	0.040000	0.040000	0.040000	0.040000	0.000000	0.040000	...	0.040000	0
5	South Side	0.166667	0.033333	0.00	0.000000	0.066667	0.000000	0.000000	0.000000	0.000000	...	0.033333	0

6 rows × 37 columns

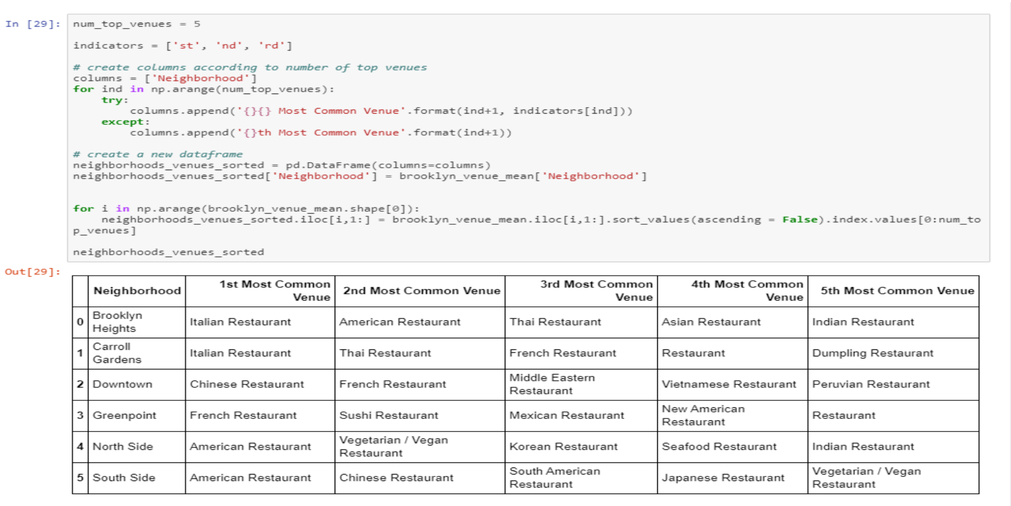
d) Exploring Each Neighborhood along with top 5 Common Restaurants/Hotels



In [26]:
```
for neighboor in brooklyn_venue_mean['Neighborhood']:
    print("*****",neighboor,"******")
    top_venue = brooklyn_venue_mean[brooklyn_venue_mean['Neighborhood']==neighboor].T.reset_index()
    top_venue.columns = ["Venue" , "Frequency"]
    top_venue = top_venue.iloc[1:]
    top_venue["Frequency"] = top_venue["Frequency"].astype(float).round(2)
    top_venue = top_venue.sort_values('Frequency',ascending = False).reset_index(drop = True)
    print(top_venue.head())
    print('\n')
```

```
***** Brooklyn Heights ******
                Venue  Frequency
0   Italian Restaurant       0.14
1  American Restaurant       0.10
2    Indian Restaurant       0.10
3      Thai Restaurant       0.10
4   Mexican Restaurant       0.10


***** Carroll Gardens ******
                Venue  Frequency
0   Italian Restaurant       0.50
1      Thai Restaurant       0.09
2    French Restaurant       0.09
3   Spanish Restaurant       0.05
4  Dumpling Restaurant       0.05
```

e) Sorting the Venues in the Descending Order

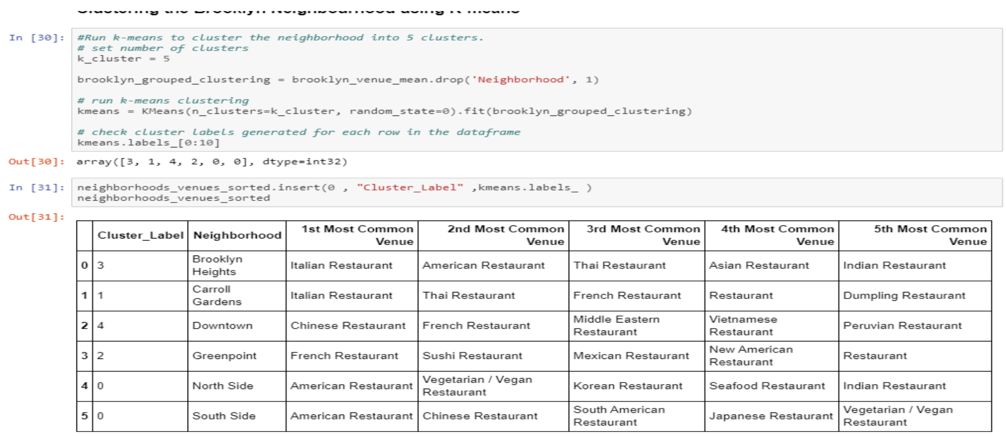


In [29]:
```
num_top_venues = 5

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

# create columns according to number of top venues
columns = ['Neighborhood']
for ind in np.arange(num_top_venues):
    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'] = brooklyn_venue_mean['Neighborhood']


for i in np.arange(brooklyn_venue_mean.shape[0]):
    neighborhoods_venues_sorted.iloc[i,1:] = brooklyn_venue_mean.iloc[i,1:].sort_values(ascending = False).index.values[0:num_to
p_venues]

neighborhoods_venues_sorted
```

Out[29]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Brooklyn Heights	Italian Restaurant	American Restaurant	Thai Restaurant	Asian Restaurant	Indian Restaurant
1	Carroll Gardens	Italian Restaurant	Thai Restaurant	French Restaurant	Restaurant	Dumpling Restaurant
2	Downtown	Chinese Restaurant	French Restaurant	Middle Eastern Restaurant	Vietnamese Restaurant	Peruvian Restaurant
3	Greenpoint	French Restaurant	Sushi Restaurant	Mexican Restaurant	New American Restaurant	Restaurant
4	North Side	American Restaurant	Vegetarian / Vegan Restaurant	Korean Restaurant	Seafood Restaurant	Indian Restaurant
5	South Side	American Restaurant	Chinese Restaurant	South American Restaurant	Japanese Restaurant	Vegetarian / Vegan Restaurant

STEP 5: Clustering the neighborhood using K-means & identifying the locations on the Map.



Clustering the Brooklyn Neighbourhood using K-means

In [30]:
```
#Run k-means to cluster the neighborhood into 5 clusters.
# set number of clusters
k_cluster = 5

brooklyn_grouped_clustering = brooklyn_venue_mean.drop('Neighborhood', 1)

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

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

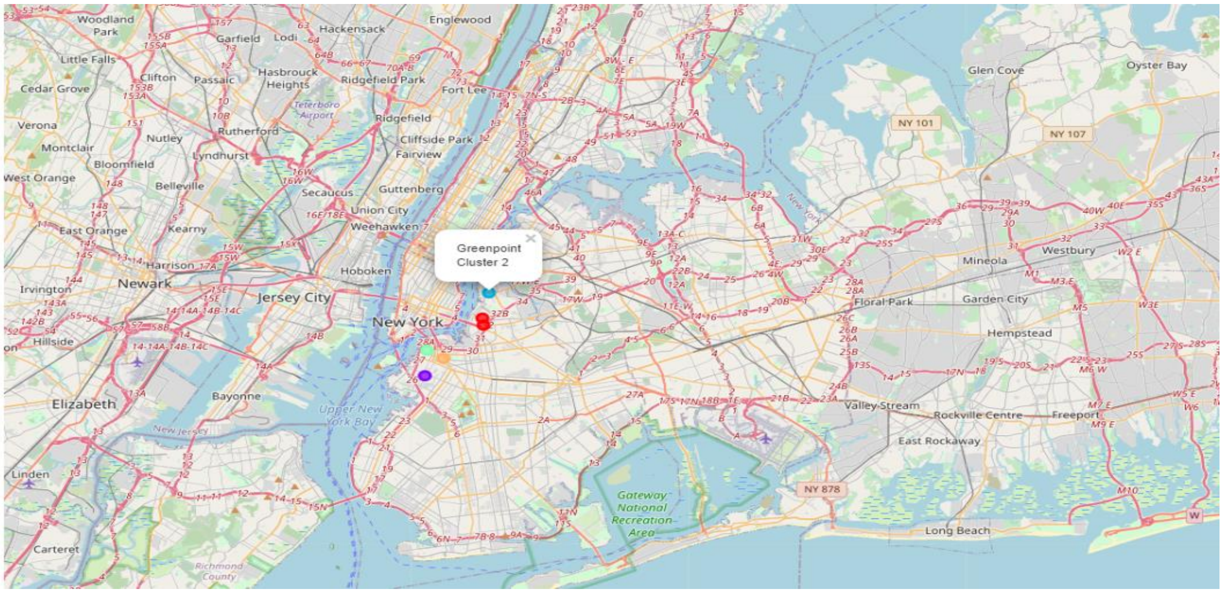
Out[30]:
```
array([3, 1, 4, 2, 0, 0], dtype=int32)
```

In [31]:
```
neighborhoods_venues_sorted.insert(0 , "Cluster_Label" ,kmeans.labels_ )
neighborhoods_venues_sorted
```

Out[31]:

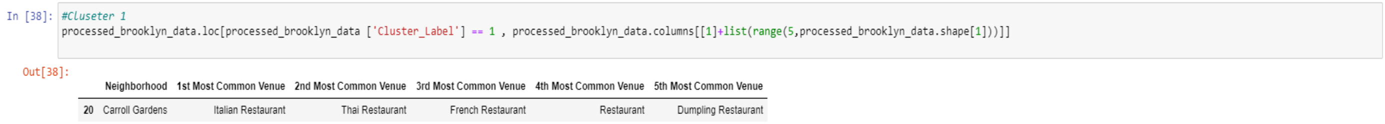
	Cluster_Label	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	3	Brooklyn Heights	Italian Restaurant	American Restaurant	Thai Restaurant	Asian Restaurant	Indian Restaurant
1	1	Carroll Gardens	Italian Restaurant	Thai Restaurant	French Restaurant	Restaurant	Dumpling Restaurant
2	4	Downtown	Chinese Restaurant	French Restaurant	Middle Eastern Restaurant	Vietnamese Restaurant	Peruvian Restaurant
3	2	Greenpoint	French Restaurant	Sushi Restaurant	Mexican Restaurant	New American Restaurant	Restaurant
4	0	North Side	American Restaurant	Vegetarian / Vegan Restaurant	Korean Restaurant	Seafood Restaurant	Indian Restaurant
5	0	South Side	American Restaurant	Chinese Restaurant	South American Restaurant	Japanese Restaurant	Vegetarian / Vegan Restaurant

CLUSTER MAP



STEP 6: Concluding the Choices of Restaurants & Locations basis of the data analysis in Step

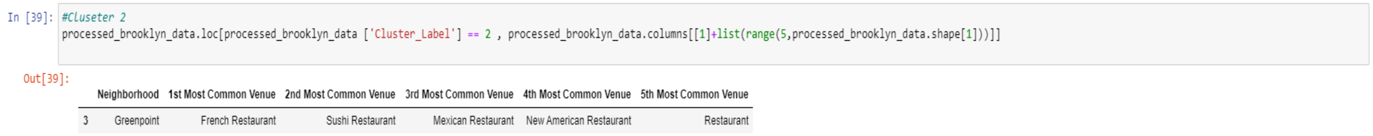
1) Examining the Cluster -1 – Carrol Gardens



```
In [38]: #Cluseter 1
         processed_brooklyn_data.loc[processed_brooklyn_data ['Cluster_Label'] == 1 , processed_brooklyn_data.columns[[1]+list(range(5,processed_brooklyn_data.shape[1]))]]

Out[38]:
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
20	Carroll Gardens	Italian Restaurant	Thai Restaurant	French Restaurant	Restaurant	Dumpling Restaurant

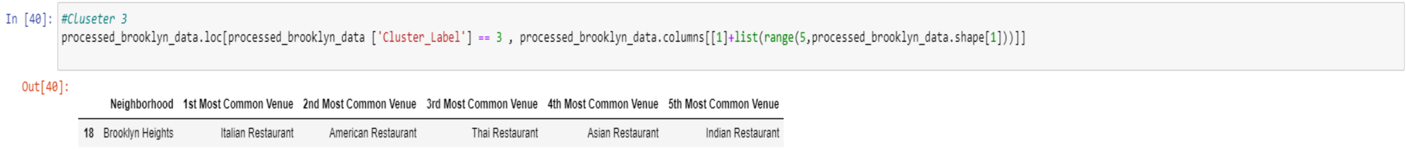
2) Examining the Cluster -2 - Green Point



```
In [39]: #Cluseter 2
         processed_brooklyn_data.loc[processed_brooklyn_data ['Cluster_Label'] == 2 , processed_brooklyn_data.columns[[1]+list(range(5,processed_brooklyn_data.shape[1]))]]

Out[39]:
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
3	Greenpoint	French Restaurant	Sushi Restaurant	Mexican Restaurant	New American Restaurant	Restaurant

3) Examining the Cluster -3 – Brooklyn Heights



```
In [40]: #Cluseter 3
         processed_brooklyn_data.loc[processed_brooklyn_data ['Cluster_Label'] == 3 , processed_brooklyn_data.columns[[1]+list(range(5,processed_brooklyn_data.shape[1]))]]

Out[40]:
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
18	Brooklyn Heights	Italian Restaurant	American Restaurant	Thai Restaurant	Asian Restaurant	Indian Restaurant

RESULTS

Out of those shortlisted three Neighborhoods, Asian & Indian Restaurants are not that common in Cluster 1 or in Cluster 2, whereas it's quite common in Brooklyn Heights. So Indian Restaurant would be preferred in Carrol Gardens or GreenPoint. If It's Italian Restaurant, best bet would be @ GreenPoint.

DISCUSSION

- **When combining data from multiple sources, inconsistent can happen. And lots of efforts are required to check, research and change the data before merge.**
- **For data obtained through API calls, different results are returned with different set of parameters and different point of time. Multiple trial and error runs are required to get the optimal result.**

- Even after the dataset has been constructed, lots of research and analysis are required to decide if the data should be kept as is or be transform by normalization or standardization.

It can be considered the most important process in the whole data science pipeline. Which can affect the most on the result.

On the other hand, choosing the suitable technique to construct the model is also a worthwhile process. As this report shows that, by applying a different method, the result can be improved.

CONCLUSION

It’s an attempt to explore the different possible analysis we could do in the available data and rationalize the decision. Although all of the goals of this project were met there is definitely room for further improvement by analyzing few more supplementary data points like demographic information, Average Spent of the population, Proximity of other crowd pulling venues like Malls, shopping complex, Cinema halls etc. However, this project could definitely be handy to narrow down a Neighborhood and a