Data Science
Brooklyn Neighborhoods

Hello and Welcome.

The project aims at defining a Business Problem, acquiring, wrangling, exploring and working with Data from external sources using Machine learning tools.

The business problem we have here is to identify a potential location for opening an Indian restaurant in the busy neighborhoods of Brooklyn. There are a various aspects of information around in the form of Data which helps us to complete an analysis on how successful the venture can be. In saying that, this is an analysis in terms of where we can ideally set up shop. There are many other factors which contribute to the overall success of any venture. Our goal here is to identify around which neighborhoods would be an ideal scope of opening a business which is centered on Indian cuisine.

Any individual when submitting a business case looks into certain aspects to ensure the success of the venture. In saying that, the project would give a more structured idea in terms of location based questions that are probed. Our goal here is to show how we can work with the available data and use Data science tools to derive the required results to have answers to the questions that we pose.

The below are the contents of the Brooklyn Capstone;

1. <u>Introduction</u>

New York is one of the major metropolis cities in the US. New York City has been described as the cultural, financial, and media capital of the world. The city consists of a diverse population of 20 Million people. **Brooklyn** is one of the 5 boroughs in New York. We are going to focus on Brooklyn as happens to be the second most densely populated county in the United States.
There are many factors to take into account when looking into any business opportunity. In our case, a restaurant poses many challenges when we dive deep into the process. We are going to focus on the location factor on our analysis. Brooklyn consists of 70 neighborhoods which would be our area of study. While venturing into the food industry can be exciting , it can be one of the toughest businesses to launch successfully. Brooklyn has a noted food scene, from artisanal eateries to ethnic eats; this stretch of NYC is a definite heaven for all noted foodies.

2. Data

The Data that has been used in this capstone has been acquired from various sources are combined into one table (Data frame). The Data process is the most integral part of the project. For any successful analysis, it is imperative that we collect the relevant data which would lead to the right validation. There can be many problems that are encountered in the sourcing and working with Data.

The data that has been used in the Brooklyn Capstone are as below;

- https://cocl.us/new_york_dataset (New York dataset with the list of Neighborhoods).
- Geocoder package (Function to provide the coordinates of a given location).
- Foursquare API (API which helps to acquire/explore a list of venues on a location).

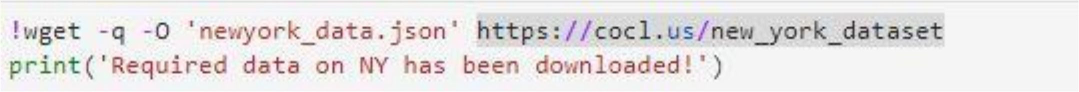
3. Methodology

Now we get into the Methodology of working on the data and arriving at the conclusion. In many cases we need to ensure that we are asking the right set of questions at the beginning before we dive into any project.

This section represents the main component of the report where the data is gathered, prepared for analysis. Please note the below methodology mainly consists of the Machine learning section of the capstone. The entire coding process is available on the Github link at the end.

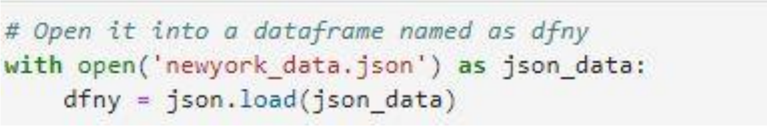
3.1 –The list of Neighborhoods in Brooklyn

Use the wget function to acquire the list of all neighborhoods in New York.



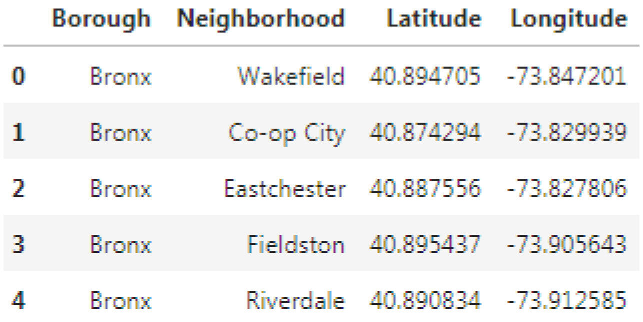
```
!wget -q -O 'newyork_data.json' https://cocl.us/new_york_dataset
print('Required data on NY has been downloaded!')
```

Then we open the json file into a data frame called dfny.



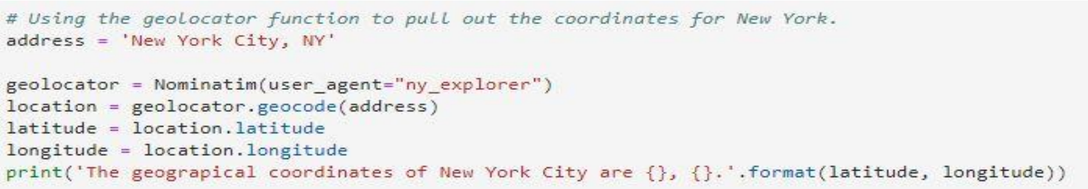
```
# Open it into a dataframe named as dfny
with open('newyork_data.json') as json_data:
    dfny = json.load(json_data)
```

After some data wrangling we arrive at the below dataset called ‘neighborhoods’.



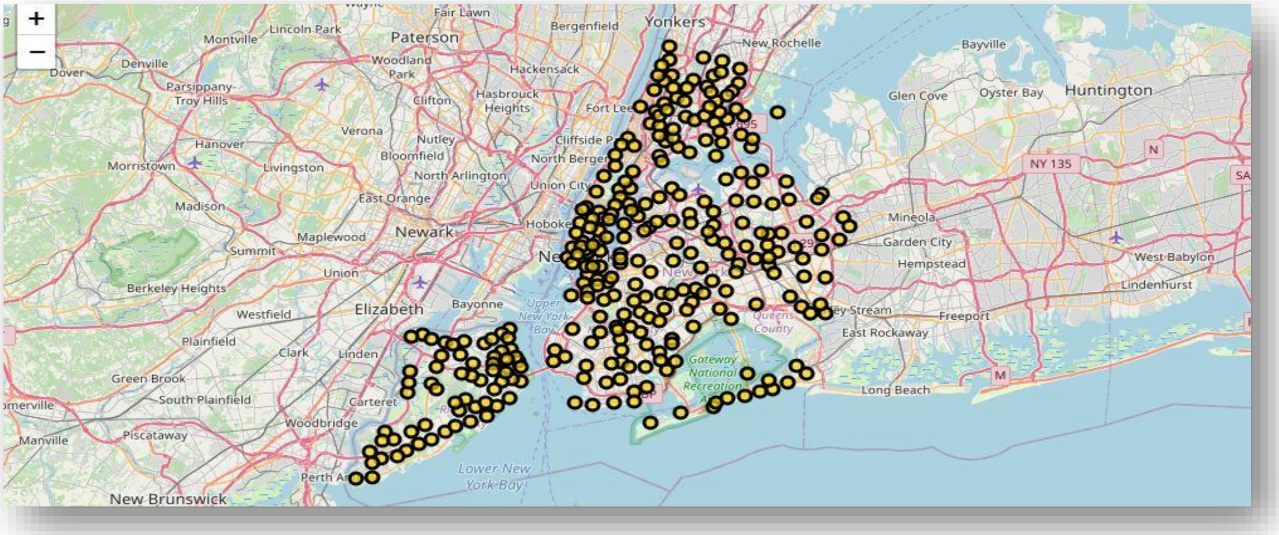
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

3.2 We will use to **Geocoder** function to acquire the co-ordinates for all the neighborhoods.

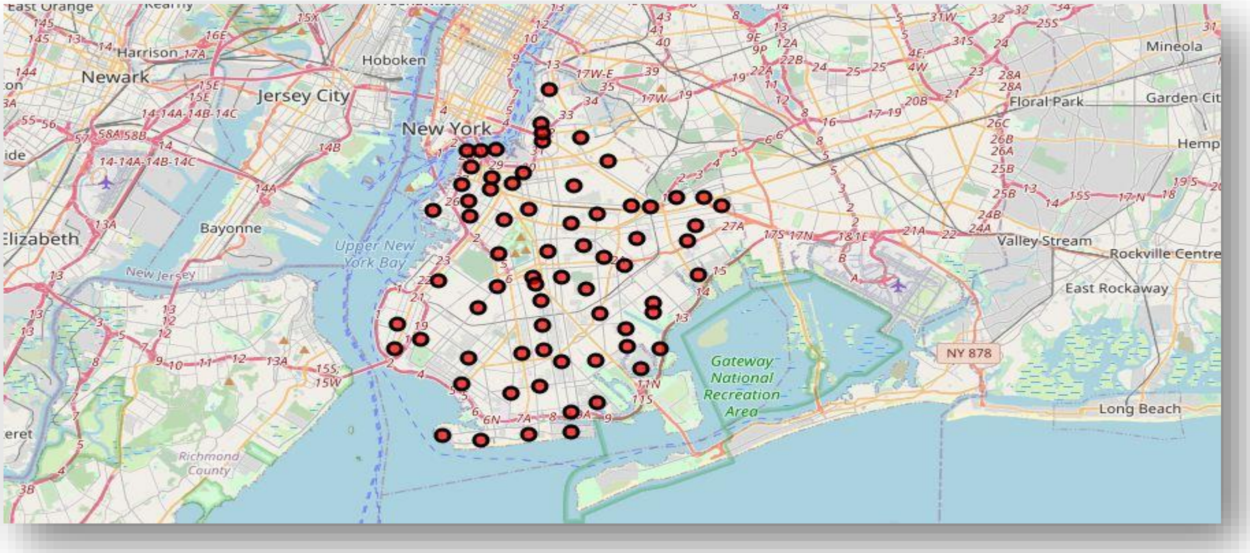


```
# Using the geolocator function to pull out the coordinates for New York.
address = 'New York City, NY'

geolocator = Nominatim(user_agent="ny_explorer")
location = geolocator.geocode(address)
latitude = location.latitude
longitude = location.longitude
print('The geograpical coordinates of New York City are {}, {}.'.format(latitude, longitude))
```

Now we have the list of all Neighborhoods in NY, let us visualize it using the folium maps. Folium is a mapping function which makes it easy to visualize data that's been manipulated in Python.

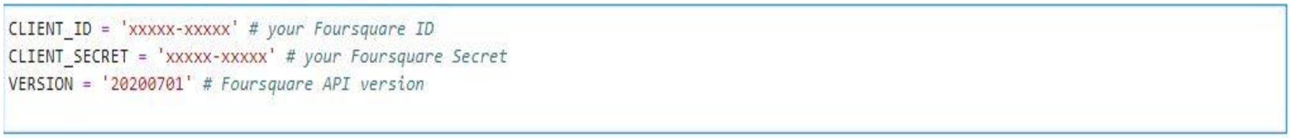


Next we need to extract the Neighborhoods in **Brooklyn** and visualize the same in the form of a map.

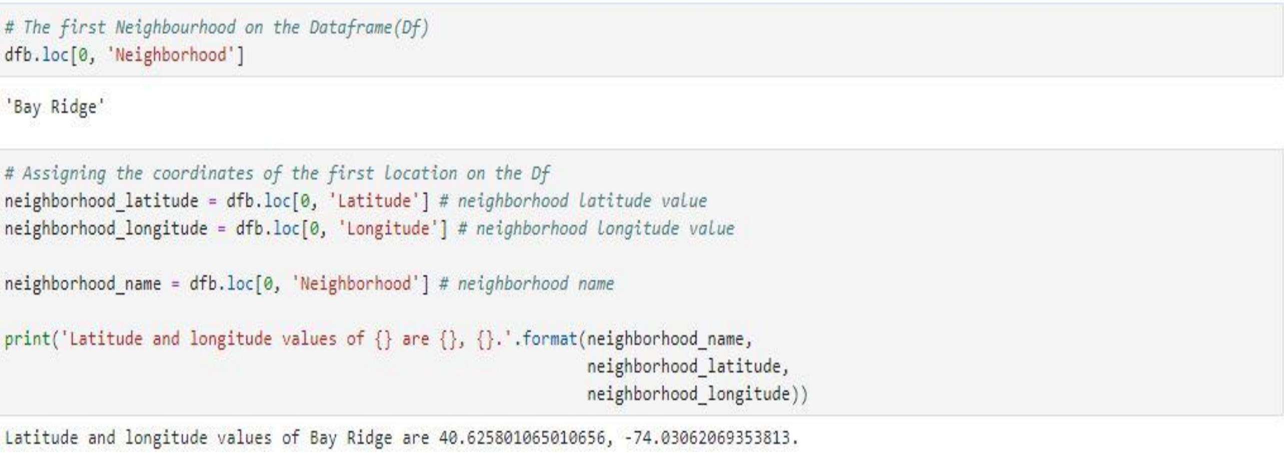


3.3 The Foursquare API.

The **Foursquare API** allows application developers to interact with the **Foursquare** platform. We are able to collect the data which are location based. In our case we need to explore all the venues around each neighborhood in Brooklyn. First we get the coordinated on the first borough of our Brooklyn data frame and acquire the venues. The first neighborhood on the list happens to be ‘Bay Ridge’.



```
CLIENT_ID = 'xxxxx-xxxxx' # your Foursquare ID
CLIENT_SECRET = 'xxxxx-xxxxx' # your Foursquare Secret
VERSION = '20200701' # Foursquare API version
```



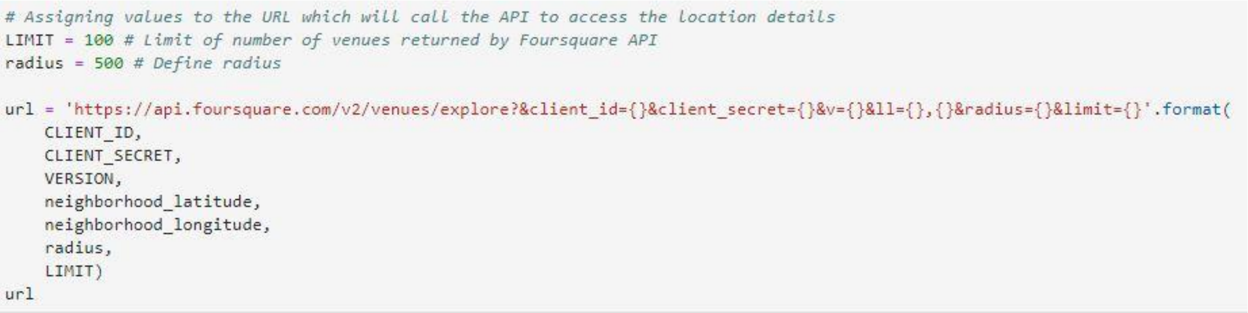
```
# The first Neighbourhood on the Dataframe(Df)
dfb.loc[0, 'Neighborhood']
```

```
'Bay Ridge'
```

```
# Assigning the coordinates of the first location on the Df
neighborhood_latitude = dfb.loc[0, 'Latitude'] # neighborhood latitude value
neighborhood_longitude = dfb.loc[0, 'Longitude'] # neighborhood longitude value

neighborhood_name = dfb.loc[0, 'Neighborhood'] # neighborhood name

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

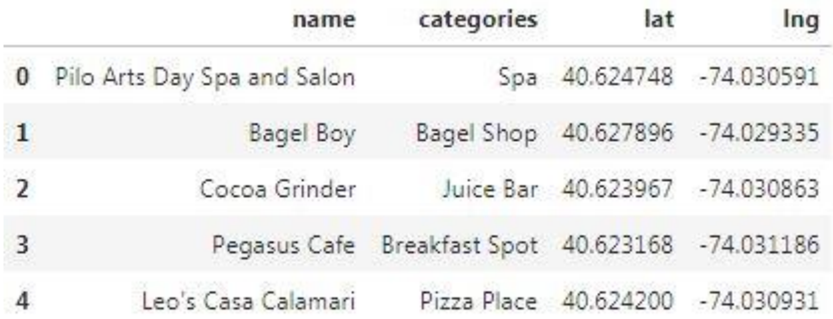
```
Latitude and longitude values of Bay Ridge are 40.625801065010656, -74.03062069353813.
```

Next we call the API to explore the list of venues in Bay Ridge.



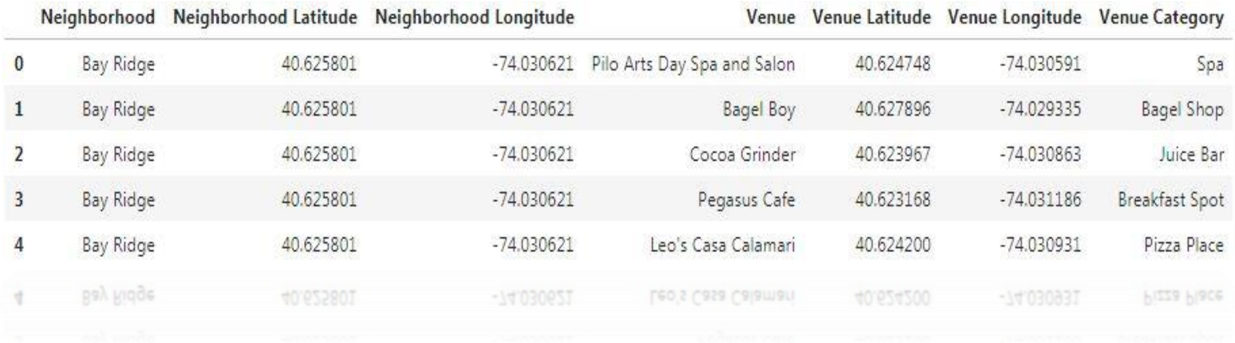
```
# Assigning values to the URL which will call the API to access the location details
LIMIT = 100 # Limit of number of venues returned by Foursquare API
radius = 500 # Define radius

url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(
    CLIENT_ID,
    CLIENT_SECRET,
    VERSION,
    neighborhood_latitude,
    neighborhood_longitude,
    radius,
    LIMIT)
url
```

Once we use the getresults function we arrive at the below list of venues;



	name	categories	lat	lng
0	Pilo Arts Day Spa and Salon	Spa	40.624748	-74.030591
1	Bagel Boy	Bagel Shop	40.627896	-74.029335
2	Cocoa Grinder	Juice Bar	40.623967	-74.030863
3	Pegasus Cafe	Breakfast Spot	40.623168	-74.031186
4	Leo's Casa Calamari	Pizza Place	40.624200	-74.030931

We define the getNearbyVenues function within a loop to explore all the locations per neighborhood in Brooklyn and come to the list which consists of all the venues.



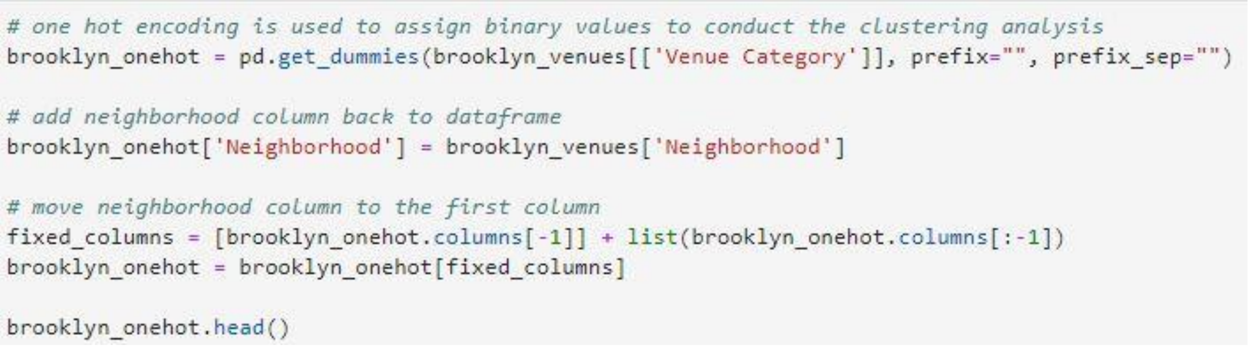
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bay Ridge	40.625801	-74.030621	Pilo Arts Day Spa and Salon	40.624748	-74.030591	Spa
1	Bay Ridge	40.625801	-74.030621	Bagel Boy	40.627896	-74.029335	Bagel Shop
2	Bay Ridge	40.625801	-74.030621	Cocoa Grinder	40.623967	-74.030863	Juice Bar
3	Bay Ridge	40.625801	-74.030621	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot
4	Bay Ridge	40.625801	-74.030621	Leo's Casa Calamari	40.624200	-74.030931	Pizza Place

3.4 Machine Learning Application

We venture into the exploratory part of the capstone. Now we have the list of the top 100 venues around each neighborhood. We would be using the clustering analysis on our data frame.

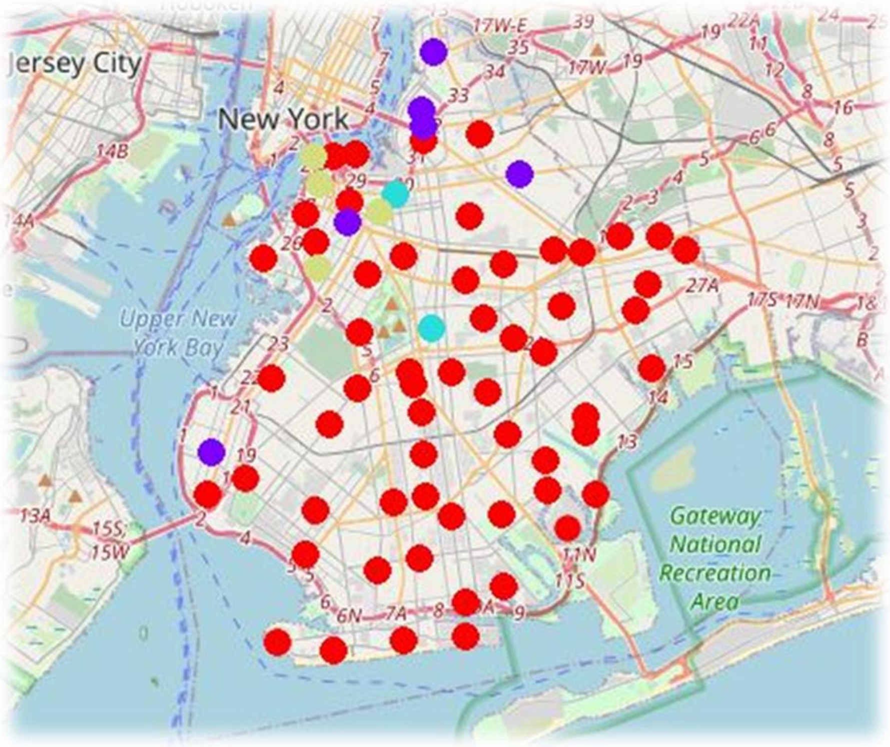
What is clustering?
Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. In Data Science, we can use clustering analysis to gain some valuable insights from our data by seeing what groups the data points fall into when we apply a clustering algorithm.

Since we are not able to cluster Geoospactial data we need to use the onehot encoding process to have our data frame ready to apply our Machine learning techniques. A **one hot encoding** allows the representation of categorical data to be more expressive. Many machine learning algorithms cannot work with categorical data directly. The categories must be converted into numbers.

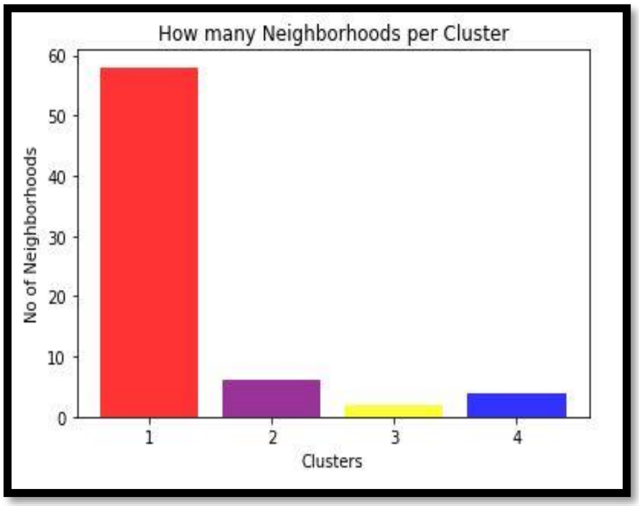


```
# one hot encoding is used to assign binary values to conduct the clustering analysis
brooklyn_onehot = pd.get_dummies(brooklyn_venues[['Venue Category']], prefix="", prefix_sep="")

# add neighborhood column back to dataframe
brooklyn_onehot['Neighborhood'] = brooklyn_venues['Neighborhood']

# move neighborhood column to the first column
fixed_columns = [brooklyn_onehot.columns[-1]] + list(brooklyn_onehot.columns[:-1])
brooklyn_onehot = brooklyn_onehot[fixed_columns]

brooklyn_onehot.head()
```

The venue category of ‘Indian Restaurants’ are extracted out and the preprocess of the data has been completed, we will use the clustering method to divide our data frame into 4 clusters.
Using Folium, the clusters are visualized on the map.

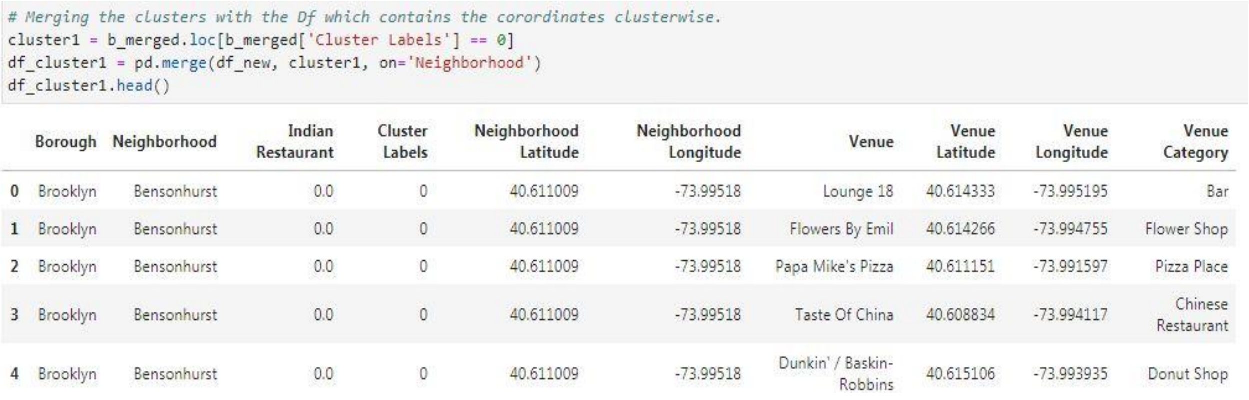


Once we have the data into the respective clusters, we plot the data in the form of a bar graph to visualize the count of neighborhoods per cluster.



4. Results

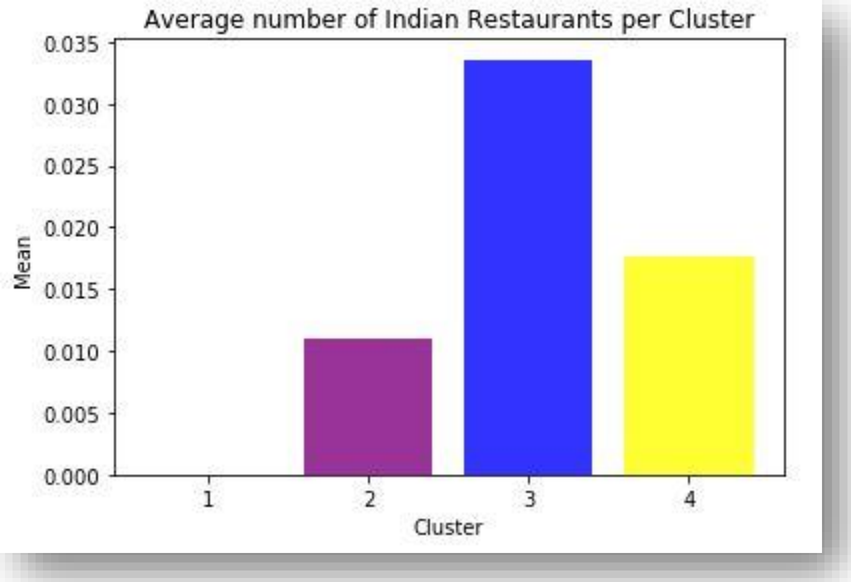
We have arrived to the penultimate section of our capstone.
Once we have applied the clustering analysis we have the clusters. We merge the location details cluster wise to get a clear indication of the venues that are present in each cluster.



```
# Merging the clusters with the Df which contains the corordinates clusterwise.
cluster1 = b_merged.loc[b_merged['Cluster Labels'] == 0]
df_cluster1 = pd.merge(df_new, cluster1, on='Neighborhood')
df_cluster1.head()
```

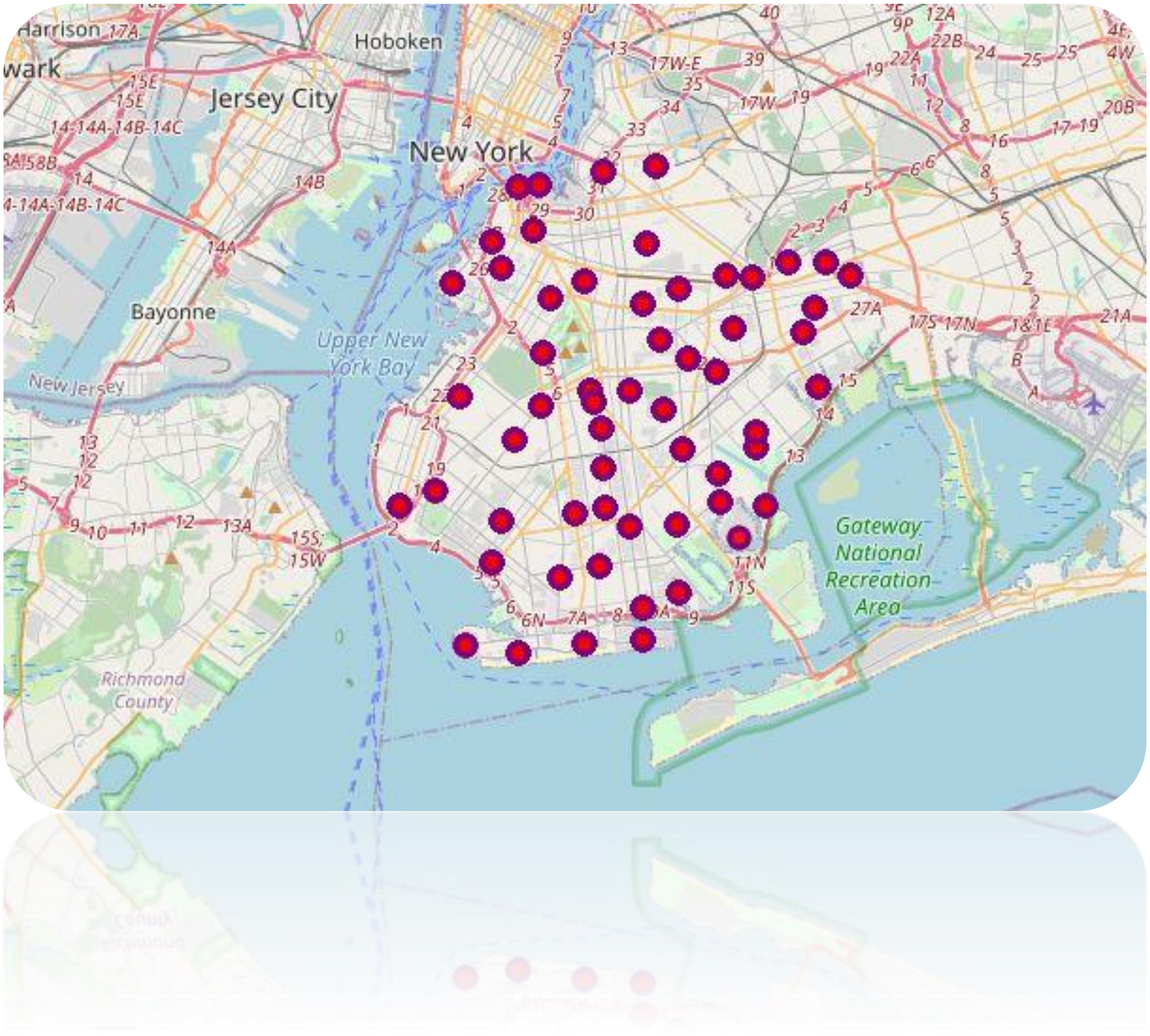
	Borough	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Brooklyn	Bensonhurst	0.0	0	40.611009	-73.99518	Lounge 18	40.614333	-73.995195	Bar
1	Brooklyn	Bensonhurst	0.0	0	40.611009	-73.99518	Flowers By Emil	40.614266	-73.994755	Flower Shop
2	Brooklyn	Bensonhurst	0.0	0	40.611009	-73.99518	Papa Mike's Pizza	40.611151	-73.991597	Pizza Place
3	Brooklyn	Bensonhurst	0.0	0	40.611009	-73.99518	Taste Of China	40.608834	-73.994117	Chinese Restaurant
4	Brooklyn	Bensonhurst	0.0	0	40.611009	-73.99518	Dunkin' / Baskin-Robbins	40.615106	-73.993935	Donut Shop

Once the similar process is completed for all the respective clusters, we look at the number of Indian restaurants in each cluster in the form of a bar graph. We take the mean of the 3rd column into consideration for each cluster. From here we have a clear indication that there are no restaurants which specialize in Indian cuisine from cluster 1 where as cluster 3 has the highest presence which is our case is considered competition.



5. Conclusion

In this Project, I have used the clustering analysis on Brooklyn neighborhoods with a vision to identify a potential location to open an Indian Restaurant. The cluster 1 was the most ideal on the fact that the following neighborhoods has the most venues around which ensure that there are potential customers to cater to. Another enticing factor is the absence of any competitors around these areas which gives the business an identity. I have visualized the neighborhoods which belong to cluster 1, in our case the most suitable locations to that can be considered when the focus is location.



The code is available on Github.

Thank you.
Nitin Kurian Chacko