IBM – COURSERA
DATA SCIENCE SPECIALIZATION

CAPSTONE PROJECT – FINAL REPORT The Battle of the Neighborhoods

~Nandini Gupta

Content

- INTRODUCTION
- BUSINESS
- SOLUTION DESIGN APPROACH
- METHODOLOGY
- RESULTS & CONCLUSION

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 km2).
- 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
- New York is also the most densely populated major city, located at the southern tip of the state of New York. 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.

Buisness 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
- 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

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 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

Solution Design Approach – 6 Steps

Solution is approached in six steps as listed below:

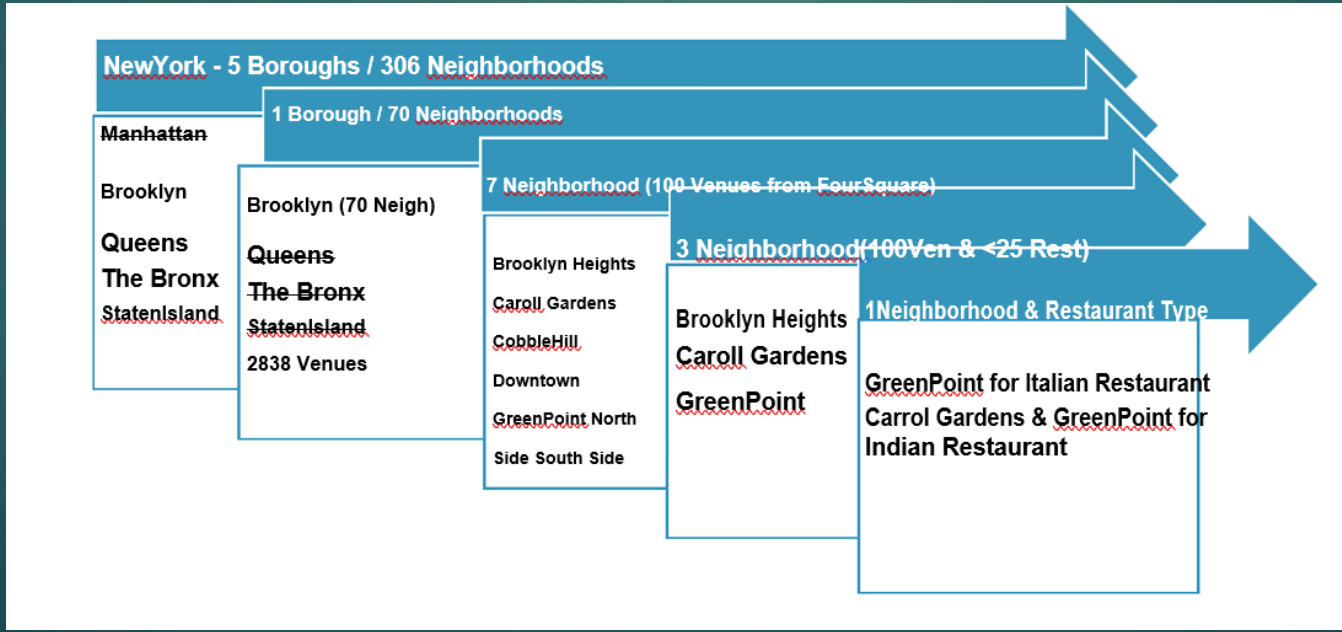
- STEP 1: Pull all the boroughs & the respective neighborhood details of the New York data using newyork_data.json.['newyork_data.json' - https://cocl.us/new_york_dataset]
- STEP 2: Deep Dive into the shortlisted Borough from Step 1 Using FourSquare APIs
- STEP 3: Explore Venues across the neighborhoods 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

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



Methodology – Data Exploratory Analysis

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,
```

Methodology – Data Exploratory Analysis

- STEP 2: Deep Dive into the shortlisted Borough from Step 1 Using Four-square APIs Brooklyn borough got 70 neighborhoods

Creating map of Brooklyn using latitude and longitude values

```
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]:

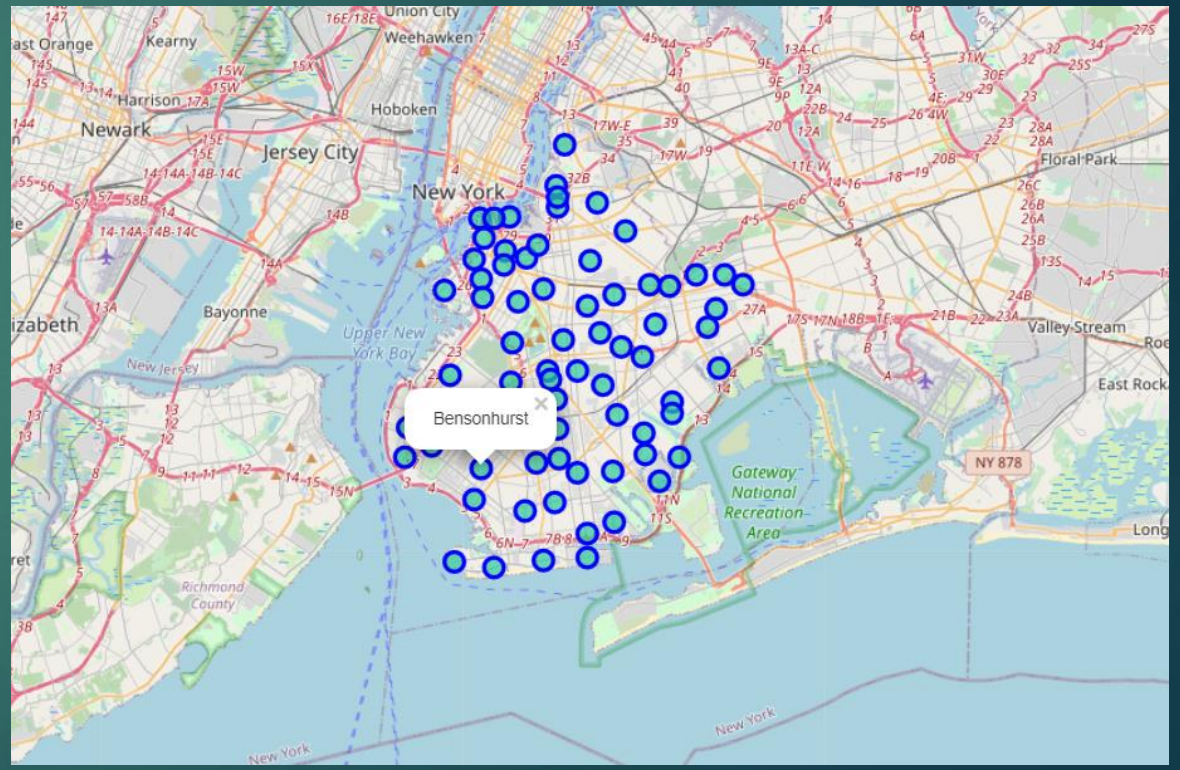
	Borough	Neighborhood	Latitude	Longitude
0	Brooklyn	Bay Ridge	40.625801	-74.030621
1	Brooklyn	Bensonhurst	40.611009	-73.995180
2	Brooklyn	Sunset Park	40.645103	-74.010316
3	Brooklyn	Greenpoint	40.730201	-73.954241
4	Brooklyn	Gravesend	40.595260	-73.973471

```
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
```



Methodology – Data Exploratory Analysis

- STEP 3: Explore Venues across the neighborhoods in Brooklyn & Narrow down to handful of it based on larger number of Venues Vs less number of Restaurants +Hotels. There were 2733 Venues across 70 Neighborhoods

```
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

- There were 6 Neighborhood having 100+ Venues with 166 Unique Venue categories

[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

```
[19]: print('There are {} uniques categories.'.format(len(brooklyn_venues['Venue Category'].unique())))
      brooklyn_venues.reset_index(drop = True)
```

There are 166 uniques categories.

- Filtering out only Restaurants & Hotels from the Venue Category
- Selecting 3 Neighborhood having Large Venues & but Less Restaurants/Hotels

		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

Methodology – Data Exploratory Analysis

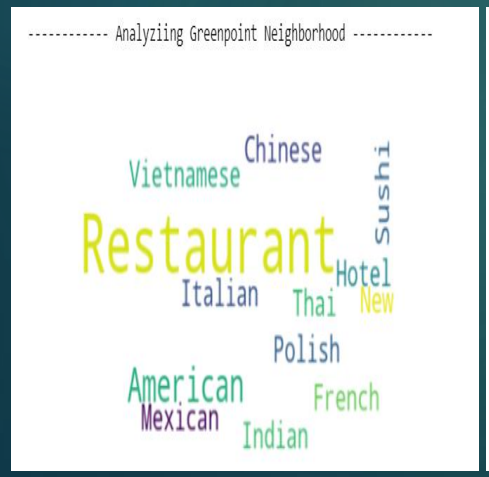
- STEP 4: Deep Dive into the shortlisted 3 neighborhoods using, Word Cloud, Means of frequency of each category of Restaurants & identifying the Top5 Common Restaurants/Hotels

Grouping the Neighborghood using means of Frequency of each category

	Neighborhood	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Chinese Restaurant	Dumpling Restaurant	Falafel Restaurant	Filipino Restaurant	French Restaurant	...	Seafood Restaurant	
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

Exploring each Neighborhood along with the top 5 Common Restaurants /Hotels



***** 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

***** Greenpoint ******

	Venue	Frequency
0	Mexican Restaurant	0.14
1	Sushi Restaurant	0.14
2	French Restaurant	0.14
3	Restaurant	0.10
4	New American Restaurant	0.10

***** Carroll Gardens ******

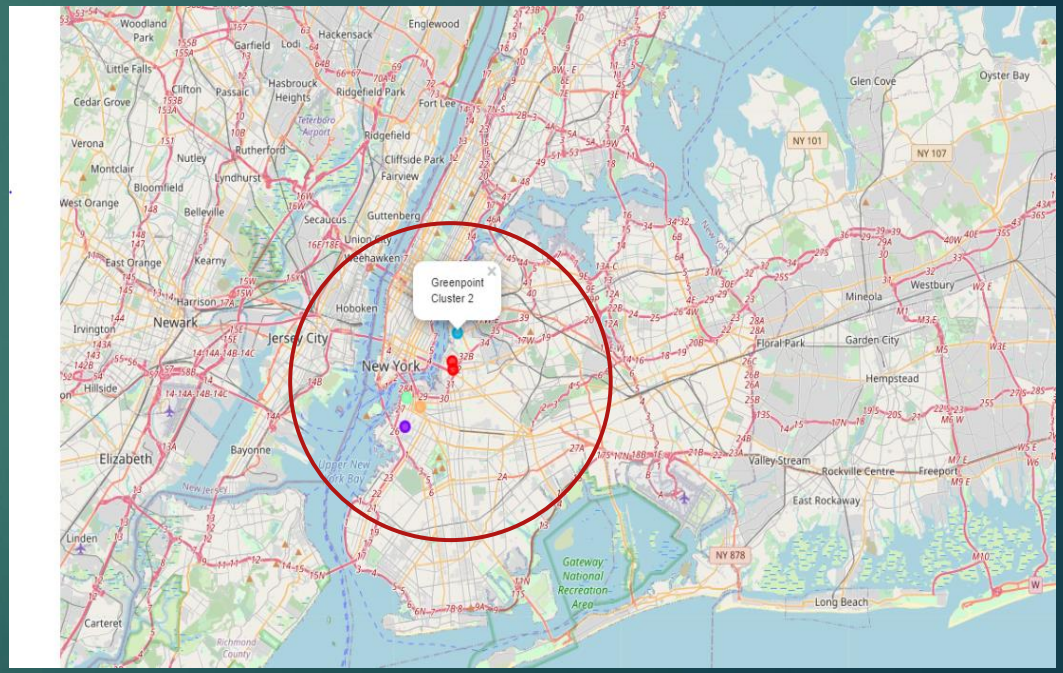
	Venue	Frequency
0	Italian Restaurant	0.50
1	Thai Restaurant	0.09
2	French Restaurant	0.09
3	Spanish Restaurant	0.05
4	Dumpling Restaurant	0.05

Methodology – Data Exploratory Analysis

- STEP 5: Clustering the neighborhood using K-means , sorting the venues in the descending order & represent it in a cluster map

1]:

	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



Methodology – Data Exploratory Analysis

- STEP 6: Concluding the Choices of Restaurants & Locations basis of the data analysis in Step
- Examining Cluster – 1 CARROLL 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

- Examining Cluster – 2 GREENPOINT

```
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

- Examining 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 & Conclusion

- **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 at GreenPoint.

- **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 type of Restaurant as a first step.

THANK YOU