IBM – COURSERA
DATA SCIENCE SPECIALIZATION

CAPSTONE PROJECT – FINAL REPORT

The Battle of the Neighborhoods

SAJITH M P – SEPT 2019

CONTENT

- Introduction
- Business Problem
- 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 2018 population of 8,398,748 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
- 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.

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
- 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 – 7 Steps

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: Narrowing down to one of the Boroughs – Basis of Population/Density analysis- on the data available in Web. https://en.wikipedia.org/wiki/Demographics_of_New_York_City"

STEP 3: Deep Dive into the shortlisted Borough from Step 2 Using **FourSquare APIs**

STEP 4: 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 5: Deep Dive into the shortlisted neighborhoods using, **Word Cloud**, **Means of frequency** of each category of Restaurants & identifying the **Top5 Common** Restaurants/Hotels

STEP 6: Clustering the neighborhood using **K-means** & identifying the locations on the Map.

STEP 7: 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

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 [7]: NYneighborhoods.head()
        print('The dataframe has {} boroughs and {} neighborhoods.'.format(
                len(NYneighborhoods['Borough'].unique()),
                NYneighborhoods.shape[0]
            )
        )
        NYneighborhoods.head()
        # STEP 1 Completes

        The dataframe has 5 boroughs and 306 neighborhoods.
```

Out[7]:

	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

STEP 2: Narrowing down to one of the Boroughs – Basis of Population/Density analysis from - Web. https://en.wikipedia.org/wiki/Demographics_of_New_York_City"

Out[9]:

	Borough	County	Population Est(2017)	GDP-USD-Billions	Per-Capita-USD	LandArea-SqMile	LandArea-SqKM	Density-SqMiles	Density-SqMiles
0	The Bronx	Bronx	1,471,160	28.787	19,570	42.10	109.04	34,653	13,231
1	Brooklyn	Kings	2,648,771	63.303	23,900	70.82	183.42	37,137	14,649
2	Manhattan	New York	1,664,727	629.682	378,250	22.83	59.13	72,033	27,826
3	Queens	Queens	2,358,582	73.842	31,310	108.53	281.09	21,460	8,354
4	Staten Island	Richmond	479,458	11.249	23,460	58.37	151.18	8,112	3,132

Selected Borough for further deep dive is - **Brooklyn**

Methodology – Data Exploratory Analysis

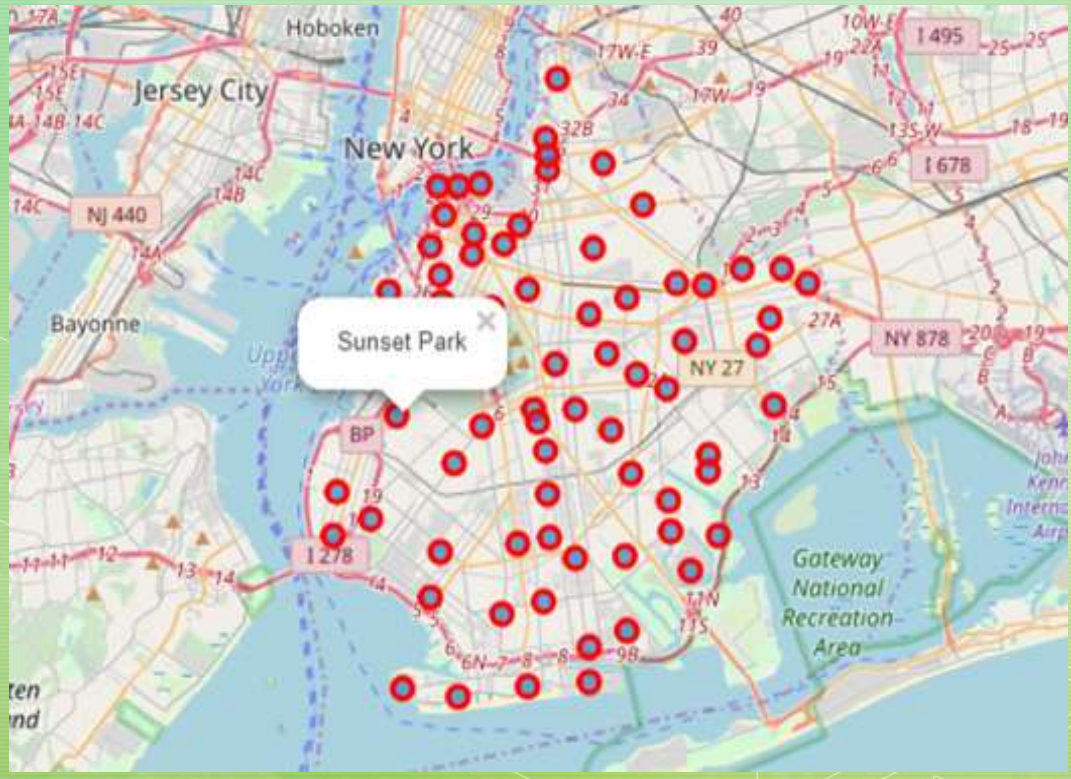
STEP 3: Deep Dive into the shortlisted Borough from Step 2 Using Four-square APIs

Brooklyn borough got 70 neighborhoods

Out[10]:

	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

Creating map of **Brooklyn** using latitude and longitude values



Methodology – Data Exploratory Analysis

STEP 4: 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 2838 Venues across 70 Neighborhoods

(2838, 7)

Out[24]:

	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	40.624740	-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	40.622967	-74.030663	Juice Bar
3	Bay Ridge	40.625801	-74.030621	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot
4	Bay Ridge	40.625801	-74.030621	Ho' Brah Taco Joint	40.622960	-74.031371	Taco Place

- There were 7 Neighborhood having 100+ Venues with 180 Unique Venue categories

	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	Cobble Hill	100	100	100	100	100	100
3	Downtown	100	100	100	100	100	100
4	Greenpoint	100	100	100	100	100	100
5	North Side	100	100	100	100	100	100
6	South Side	100	100	100	100	100	100

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

Out[34]:

		count
Neighborhood	Venue Type	
Brooklyn Heights	Restaurant	22
Carroll Gardens	Restaurant	24
Cobble Hill	Restaurant	25
Downtown	Hotel	2
	Restaurant	28
Greenpoint	Hotel	1
	Restaurant	23
North Side	Hotel	1
	Restaurant	24
South Side	Restaurant	31

Methodology – Data Exploratory Analysis

STEP 5: 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 Neighbourghood using means of Frequency of each category

	Neighborhood	American Restaurant	Arepa Restaurant	Argentinian Restaurant	Asian Restaurant	Caribbean Restaurant	Chinese Restaurant	Cuban Restaurant	Dumpling Restaurant	Eastern European Restaurant	Ethiopian Restaurant	Falafel Restaurant	Fast Food Restaurant
0	Brooklyn Heights	0.090909	0.000000	0.00	0.090909	0.000000	0.045455	0.000000	0.000000	0.045455	0.00	0.045455	0.045455
1	Carroll Gardens	0.000000	0.000000	0.00	0.000000	0.000000	0.000000	0.041667	0.041667	0.000000	0.00	0.000000	0.000000
2	Cobble Hill	0.040000	0.000000	0.04	0.000000	0.000000	0.000000	0.000000	0.040000	0.000000	0.04	0.040000	0.000000
3	Downtown	0.000000	0.000000	0.00	0.066667	0.033333	0.066667	0.033333	0.000000	0.000000	0.00	0.000000	0.000000
4	Greenpoint	0.041667	0.000000	0.00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.041667	0.000000
5	North Side	0.120000	0.040000	0.04	0.040000	0.000000	0.080000	0.000000	0.040000	0.000000	0.00	0.000000	0.000000
6	South Side	0.129032	0.032258	0.00	0.000000	0.000000	0.096774	0.000000	0.000000	0.000000	0.00	0.000000	0.000000

- Exploring each Neighbourhood along with the top 5 Common Restrnts /Hotels


```
----Brooklyn Heights----
                        venue  freq
0          Italian Restaurant  0.14
1        American Restaurant   0.09
2          Indian Restaurant   0.09
3            Thai Restaurant   0.09
4          Asian Restaurant    0.09
5          Sushi Restaurant    0.05
6          Ramen Restaurant    0.05
7   New American Restaurant    0.05
8  Middle Eastern Restaurant   0.05
9          Mexican Restaurant  0.05
```

```
----Carroll Gardens----
                        venue  freq
0          Italian Restaurant  0.46
1            Thai Restaurant   0.08
2         Seafood Restaurant   0.04
3         Spanish Restaurant   0.04
4                 Restaurant   0.04
5           Greek Restaurant   0.04
6          French Restaurant   0.04
7        Filipino Restaurant   0.04
8  Latin American Restaurant   0.04
9        Dumpling Restaurant   0.04
```

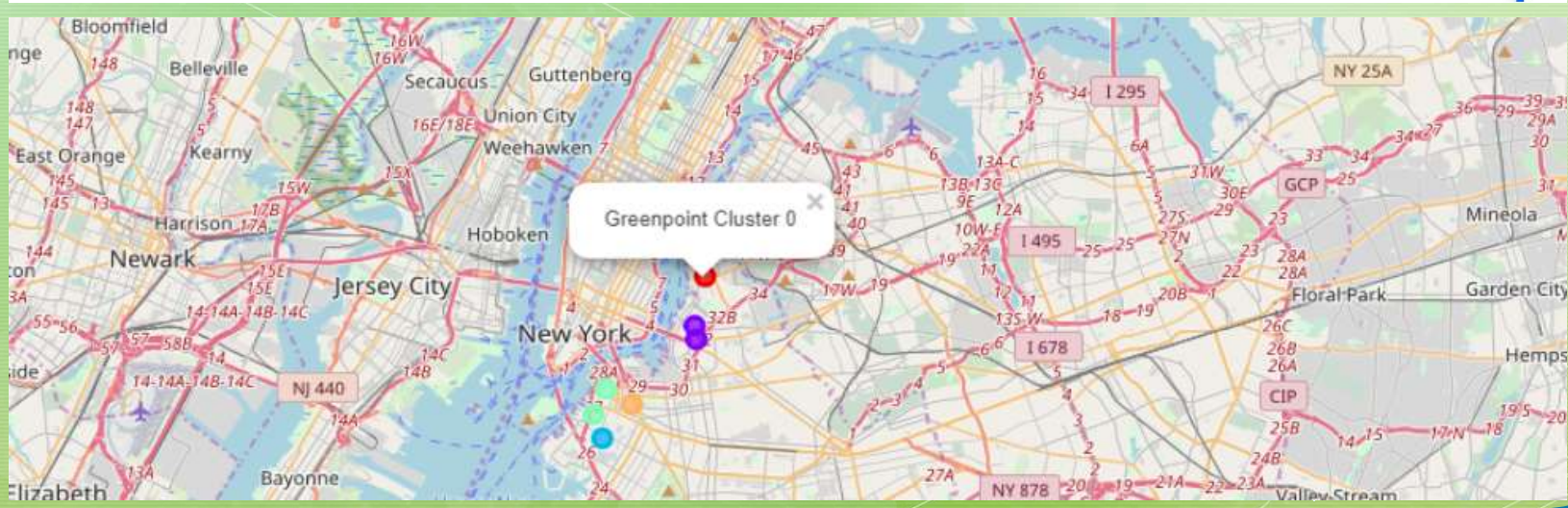
```
----Greenpoint----
                        venue  freq
0          French Restaurant   0.12
1         Mexican Restaurant   0.12
2           Sushi Restaurant   0.08
3                 Restaurant   0.08
4          Polish Restaurant   0.08
5   New American Restaurant    0.08
6          Italian Restaurant  0.08
7        American Restaurant   0.04
8  Vegetarian / Vegan Restaurant  0.04
9            Thai Restaurant   0.04
```

Methodology – Data Exploratory Analysis

STEP 6: Clustering the neighborhood using K-means , sorting the venues in the descending order & represent it in a cluster map .

	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
3	Brooklyn	Greenpoint	40.730201	-73.954241	0	French Restaurant	Mexican Restaurant	New American Restaurant	Sushi Restaurant	Italian Restaurant
18	Brooklyn	Brooklyn Heights	40.695864	-73.993782	3	Italian Restaurant	American Restaurant	Thai Restaurant	Asian Restaurant	Indian Restaurant
19	Brooklyn	Cobble Hill	40.687920	-73.998561	3	Italian Restaurant	Japanese Restaurant	Thai Restaurant	French Restaurant	Mediterranean Restaurant
20	Brooklyn	Carroll Gardens	40.680540	-73.994654	2	Italian Restaurant	Thai Restaurant	Cuban Restaurant	Restaurant	French Restaurant
40	Brooklyn	Downtown	40.690844	-73.983463	4	French Restaurant	Thai Restaurant	Asian Restaurant	Chinese Restaurant	Shanghai Restaurant
50	Brooklyn	North Side	40.714823	-73.958809	1	American Restaurant	Vegetarian / Vegan Restaurant	Chinese Restaurant	South American Restaurant	Seafood Restaurant
51	Brooklyn	South Side	40.710861	-73.958001	1	American Restaurant	Chinese Restaurant	Seafood Restaurant	Vegetarian / Vegan Restaurant	Korean Restaurant

Cluster Map



Methodology – Data Exploratory Analysis

STEP 7: Concluding the Choices of Restaurants & Locations basis of the data analysis in Step

Examining Cluster – 0 GREENPOINT

```
In [122]: # Examining the Clusters
          # Cluster =
          brooklyn_merged.loc[brooklyn_merged['Cluster Labels'] == 0, brooklyn_merged.columns[[1] + list(range(5, brooklyn_merged.shape[1]))]]
```

Out[122]:

	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	Mexican Restaurant	New American Restaurant	Sushi Restaurant	Italian Restaurant

Examining Cluster – 2 CARROL GARDENS

```
In [50]: # Examining the Clusters
         # Cluster = 2
         brooklyn_merged.loc[brooklyn_merged['Cluster Labels'] == 2, brooklyn_merged.columns[[1] + list(range(5, brooklyn_merged.shape[1]))]]
```

Out[50]:

	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	Cuban Restaurant	Restaurant	French Restaurant

Examining Cluster – 3 BROOKLYN HEIGHTS

```
In [51]: # Examining the Clusters
         # Cluster = 3
         brooklyn_merged.loc[brooklyn_merged['Cluster Labels'] == 3, brooklyn_merged.columns[[1] + list(range(5, brooklyn_merged.shape[1]))]]
```

Out[51]:

	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 0 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.

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