Exploring New York City to Open an Indian Restaurant

IBM Data Science Capstone Project

Abhishek Mishra

June 2020

TABLE OF CONTENTS

1. INTRODUCTION

For this Capstone Project, I am using the hypothetical scenario for a concept Indian Entrepreneur who wants to open an Indian Restaurant in New York City (NYC). It might present a good opportunity for an Indian American already living in NYC and are well versed with the Places and the Neighborhoods. As Indian cuisine is quite popular with Americans and Indian Americans alike, there are already many restaurants most of which are a Franchise or a family owned business.

The New York City region is home to the largest Indian American population among metropolitan areas by a significant margin and represents the second-largest metropolitan Asian national diaspora both outside of Asia and within the New York City metropolitan area.

Ambience, menu, hygiene and of course taste are all important factors to be kept in mind before getting into the Hospitality Industry but these are all problems that can be tackled internally by the person(s) in charge. The location of a restaurant is also of utmost importance regardless of the history of a business or the taste of the food. If people don't come in to eat then none of the preparations matter. That is the problem I am tackling in this project.

2. PROBLEM STATEMENT

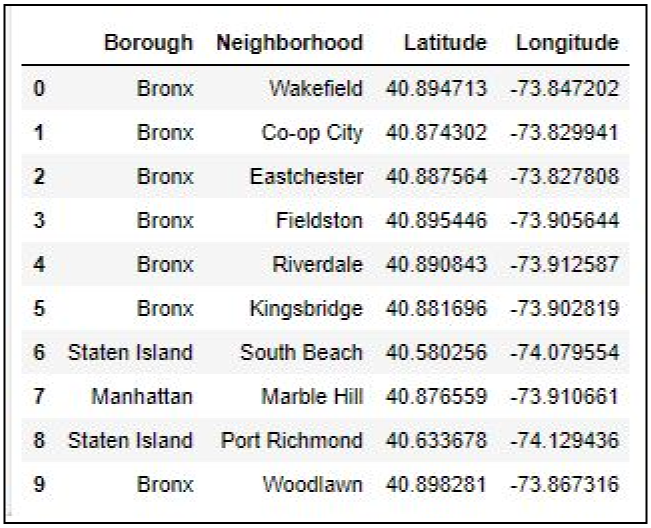
The objective is to find a suitable location(s) to open an Indian Restaurant in New York City, USA. This project makes use of various Data Science and Machine Learning methodologies (k-means Clustering) to provide a Solution to the client. The project aims to provide a Solution to the Question : 'Where should you consider opening an Indian Restaurant in New York City?'

3. DATA

3.1 Data

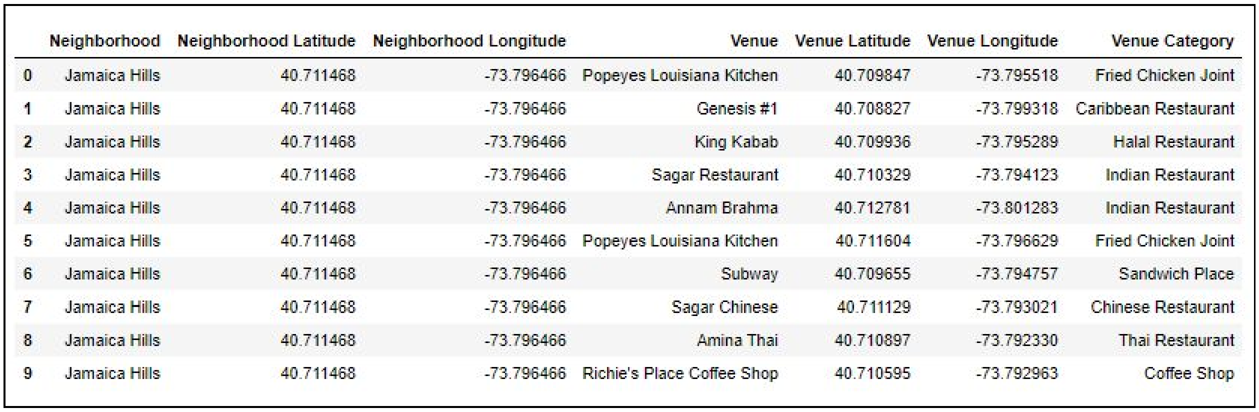
I have used the following Data for the completion of the project :

- List of Boroughs and Neighborhoods in NYC - This gives the coordinates of all the neighborhoods and is used to call the Foursquare API.
- List of Places and Venues in NYC - This contains data about all the nearby venues like Restaurants, Bars, Gym etc.
- Demographics of American Indians in New York City - Vital to understand the distribution of the target audience in NYC.
- Latitude and Longitude Data of the neighborhood(s) - To plot and visualize our data.



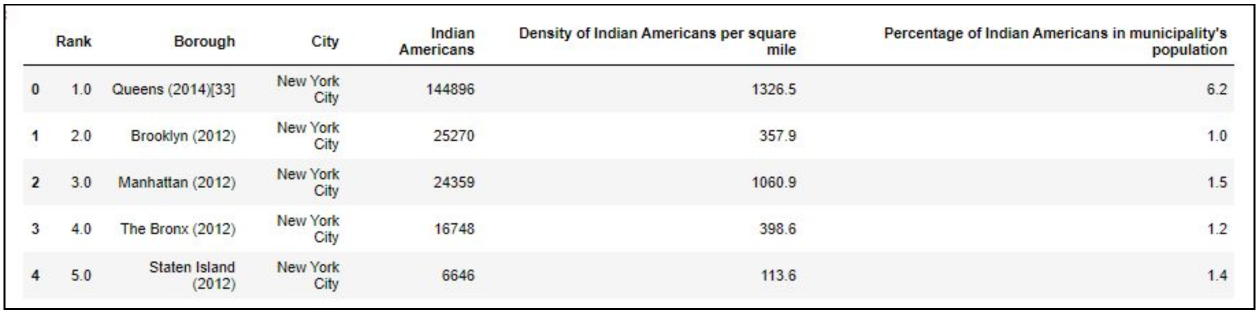
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894713	-73.847202
1	Bronx	Co-op City	40.874302	-73.829941
2	Bronx	Eastchester	40.887564	-73.827808
3	Bronx	Fieldston	40.895446	-73.905644
4	Bronx	Riverdale	40.890843	-73.912587
5	Bronx	Kingsbridge	40.881696	-73.902819
6	Staten Island	South Beach	40.580256	-74.079554
7	Manhattan	Marble Hill	40.876559	-73.910661
8	Staten Island	Port Richmond	40.633678	-74.129436
9	Bronx	Woodlawn	40.898281	-73.867316

Fig 3.1 Boroughs and Neighborhoods in NYC



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Jamaica Hills	40.711468	-73.796466	Popeyes Louisiana Kitchen	40.709847	-73.795518	Fried Chicken Joint
1	Jamaica Hills	40.711468	-73.796466	Genesis #1	40.708827	-73.799318	Caribbean Restaurant
2	Jamaica Hills	40.711468	-73.796466	King Kabab	40.709936	-73.795289	Halal Restaurant
3	Jamaica Hills	40.711468	-73.796466	Sagar Restaurant	40.710329	-73.794123	Indian Restaurant
4	Jamaica Hills	40.711468	-73.796466	Annam Brahma	40.712781	-73.801283	Indian Restaurant
5	Jamaica Hills	40.711468	-73.796466	Popeyes Louisiana Kitchen	40.711604	-73.796629	Fried Chicken Joint
6	Jamaica Hills	40.711468	-73.796466	Subway	40.709655	-73.794757	Sandwich Place
7	Jamaica Hills	40.711468	-73.796466	Sagar Chinese	40.711129	-73.793021	Chinese Restaurant
8	Jamaica Hills	40.711468	-73.796466	Amina Thai	40.710897	-73.792330	Thai Restaurant
9	Jamaica Hills	40.711468	-73.796466	Richie's Place Coffee Shop	40.710595	-73.792963	Coffee Shop

Fig 3.2 Venues returned by the Foursquare API



	Rank	Borough	City	Indian Americans	Density of Indian Americans per square mile	Percentage of Indian Americans in municipality's population
0	1.0	Queens (2014)[33]	New York City	144896	1326.5	6.2
1	2.0	Brooklyn (2012)	New York City	25270	357.9	1.0
2	3.0	Manhattan (2012)	New York City	24359	1060.9	1.5
3	4.0	The Bronx (2012)	New York City	16748	398.6	1.2
4	5.0	Staten Island (2012)	New York City	6646	113.6	1.4

Fig 3.3 Indian Demographic in NYC

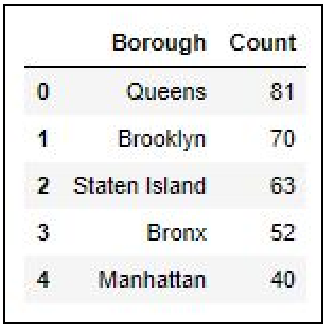
3.2 Data Sources

- New York City Neighborhoods Data from NYU website [1].
- Nearby Venues Data created using Foursquare API [2].
- The Demographics Data is scraped from Wikipedia [3].
- Latitude and Longitude values are obtained using the Geocoder package in python.

4. METHODOLOGY

4.1 Boroughs

The data section above clearly describes that our NYC data consists of Boroughs (a town or district) and Neighborhoods in these Boroughs. The data contains 5 Boroughs - Queens, Brooklyn, Bronx, Manhattan and Staten Island and over 300 neighborhoods in total. So before we begin our analysis of the Neighborhoods we select an appropriate Borough. This involves looking into all 5 of them. The data is filtered for each Borough and is used to make the call to the Foursquare API.



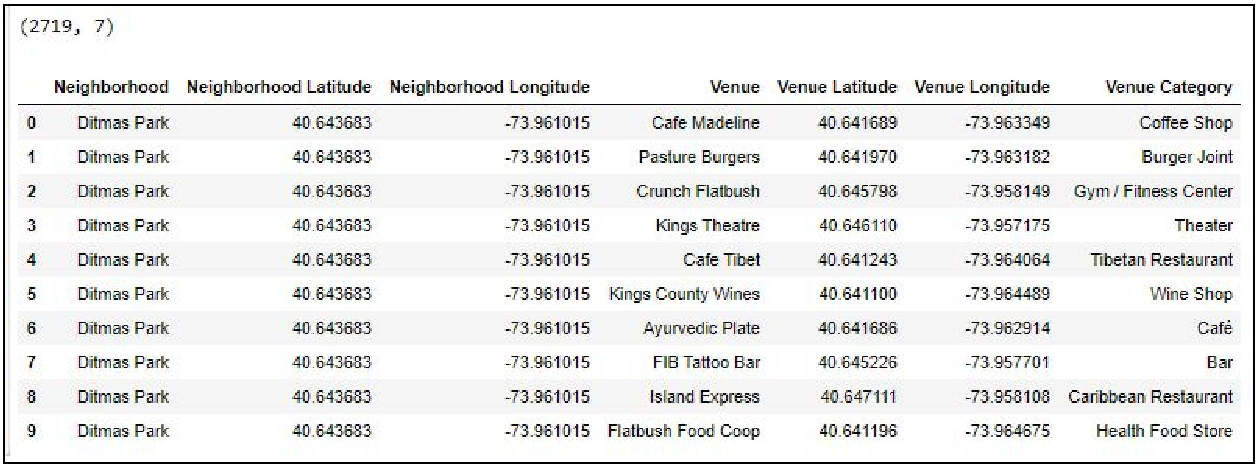
	Borough	Count
0	Queens	81
1	Brooklyn	70
2	Staten Island	63
3	Bronx	52
4	Manhattan	40

Fig 4.1 Count of Neighborhoods in the Boroughs

4.2 Foursquare API

The central part of this project involves making use of the Foursquare API to get various details of nearby venues, like - the Category (Pizza Place, Monument etc), The coordinates of the place (in Latitude and Longitude) and the Name of the Venue. We need to declare our Foursquare credentials like the Client ID and Client Secret. We assume a radius value of 500, which returns venues within a radius of half a kilometer. To prevent too many records being returned by the function call a limit of 100 is set.

The url is constructed with our declared credentials and a request call is made to the API. The data returned is in the form of a json payload. The pandas dataframe is then constructed by reading parts of this data. Therefore 5 dataframes are made - one for each Borough.



```
(2719, 7)
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Ditmas Park	40.643683	-73.961015	Cafe Madeline	40.641689	-73.963349	Coffee Shop
1	Ditmas Park	40.643683	-73.961015	Pasture Burgers	40.641970	-73.963182	Burger Joint
2	Ditmas Park	40.643683	-73.961015	Crunch Flatbush	40.645798	-73.958149	Gym / Fitness Center
3	Ditmas Park	40.643683	-73.961015	Kings Theatre	40.646110	-73.957175	Theater
4	Ditmas Park	40.643683	-73.961015	Cafe Tibet	40.641243	-73.964064	Tibetan Restaurant
5	Ditmas Park	40.643683	-73.961015	Kings County Wines	40.641100	-73.964489	Wine Shop
6	Ditmas Park	40.643683	-73.961015	Ayurvedic Plate	40.641686	-73.962914	Café
7	Ditmas Park	40.643683	-73.961015	FIB Tattoo Bar	40.645226	-73.957701	Bar
8	Ditmas Park	40.643683	-73.961015	Island Express	40.647111	-73.958108	Caribbean Restaurant
9	Ditmas Park	40.643683	-73.961015	Flatbush Food Coop	40.641196	-73.964675	Health Food Store

Fig 4.2 The data returned by the API for Brooklyn

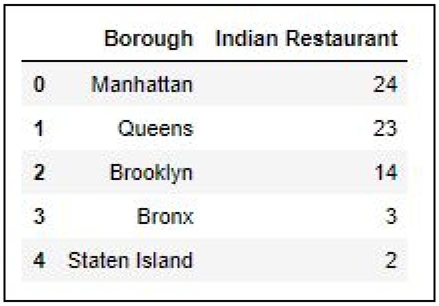
Now that the data has been structured for the preprocessing, we to decide on a Borough for the analysis and so we look into 2 aspects -

1. Pre-existing Indian Restaurants
2. Demographics of the Indian American Population

4.3 Pre-existing Indian Restaurants

Since we wish to open a new Indian Restaurant, it helps to look into ones that are already present. So we get the count of Indian Restaurants (from the Venue Category) in each Borough and merge them together to get an idea of the distribution or concentration of them. Logically, to avoid competition it would make sense to select a Borough with few Indian Restaurants.

It can be seen that Manhattan and Queens have the most number of Restaurants and Staten Island with the least.



	Borough	Indian Restaurant
0	Manhattan	24
1	Queens	23
2	Brooklyn	14
3	Bronx	3
4	Staten Island	2

Fig 4.3 Count of Indian Restaurants

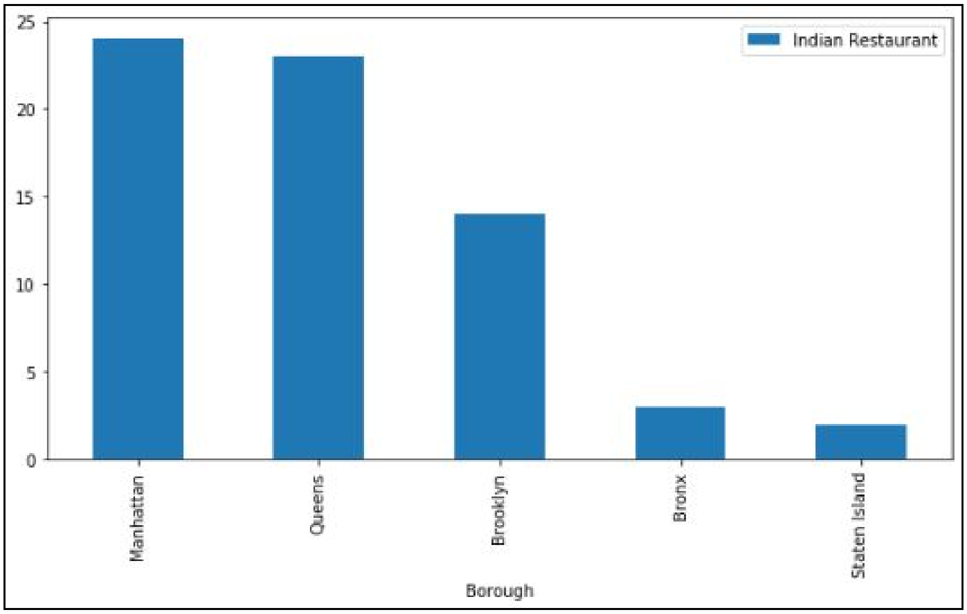
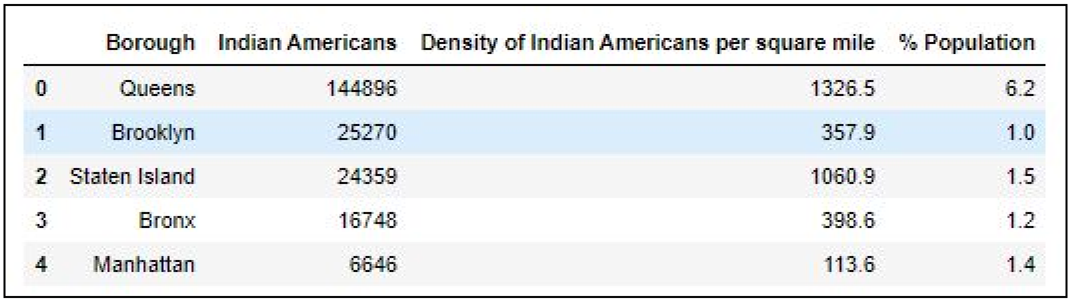


Fig 4.4 Plot of Indian Restaurants against Boroughs

4.4 Demographics of Indian Americans

An Indian Restaurant would primarily cater to the Indian American population and Indian tourists. So we look into the Indian American population in NYC. The data for the same was scraped from Wikipedia and is from a 2014 American Community Survey (that gathers census data including ethnicity). This helps us narrow down our location for the target population.

The raw data scraped contains some formatting and unnecessary columns that need to be cleaned before it can be used. Once completed, it looks like this -



	Borough	Indian Americans	Density of Indian Americans per square mile	% Population
0	Queens	144896	1326.5	6.2
1	Brooklyn	25270	357.9	1.0
2	Staten Island	24359	1060.9	1.5
3	Bronx	16748	398.6	1.2
4	Manhattan	6646	113.6	1.4

Fig 4.5 Indian Population in Boroughs

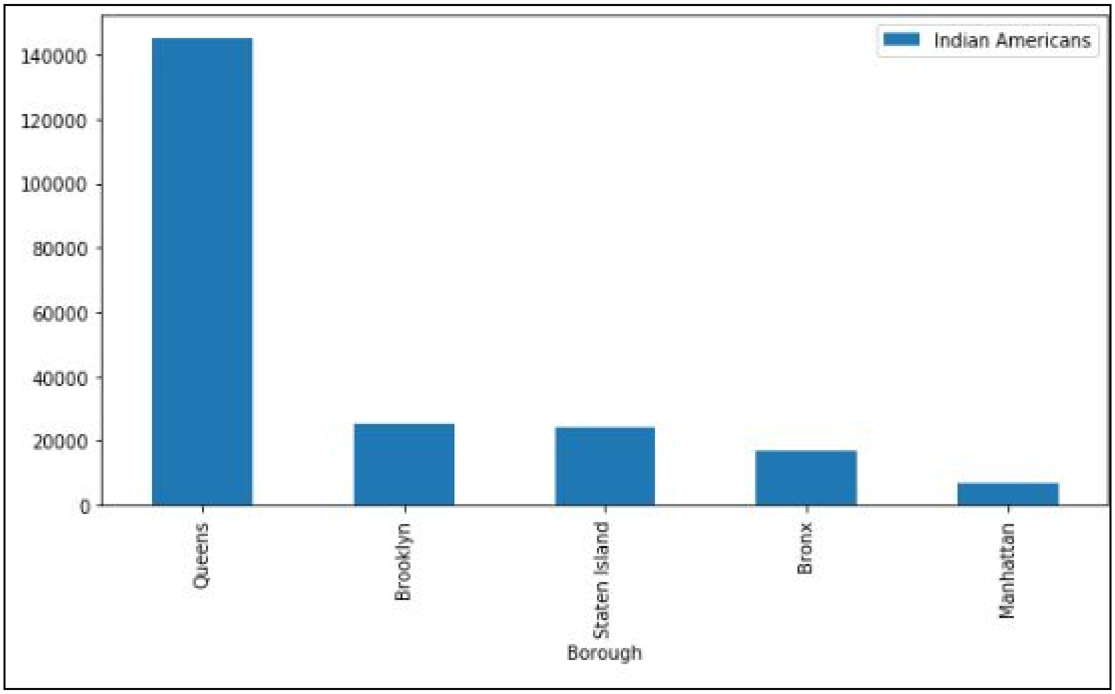
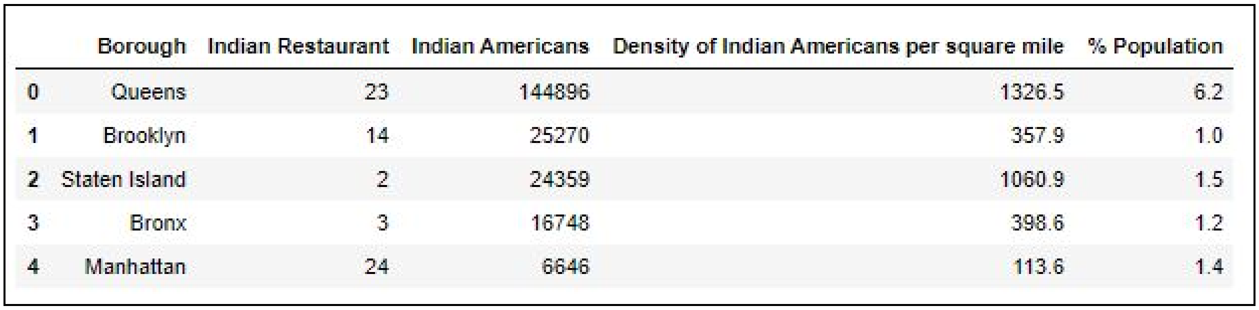


Fig 4.6 Plot of Indian Americans against the Boroughs

4.5 Initial Analysis



	Borough	Indian Restaurant	Indian Americans	Density of Indian Americans per square mile	% Population
0	Queens	23	144896	1326.5	6.2
1	Brooklyn	14	25270	357.9	1.0
2	Staten Island	2	24359	1060.9	1.5
3	Bronx	3	16748	398.6	1.2
4	Manhattan	24	6646	113.6	1.4

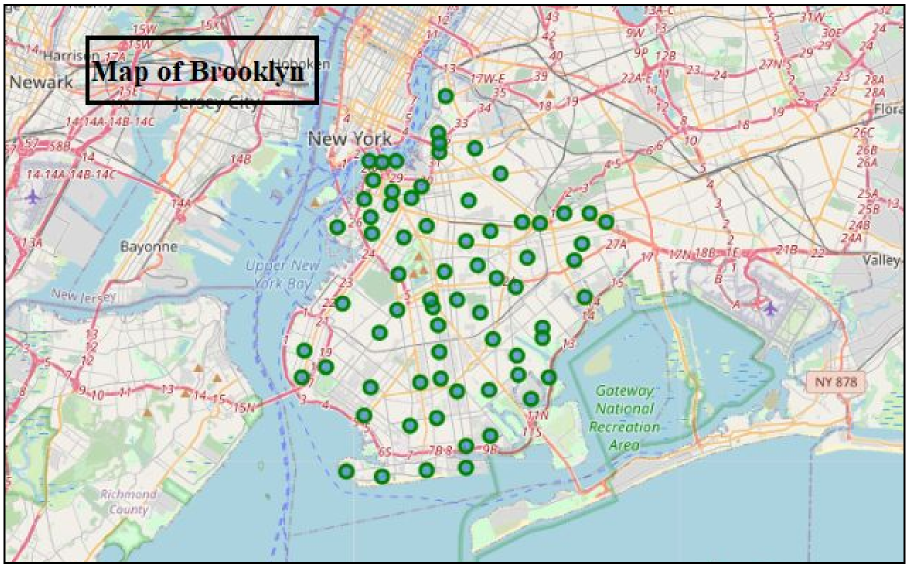
Fig 4.7 Merged data table showing Population and Indian Restaurants

Although Queens has the highest population of Indian Americans and the highest % population, we don't consider it as there are already numerous pre-existing restaurants. Manhattan has very few Indian Americans with a low % and also has the most no. of Indian Restaurants, so we eliminate it.

Brooklyn seems like a good first choice to begin our analysis as it does not have too many restaurants with a decent Indian Population.

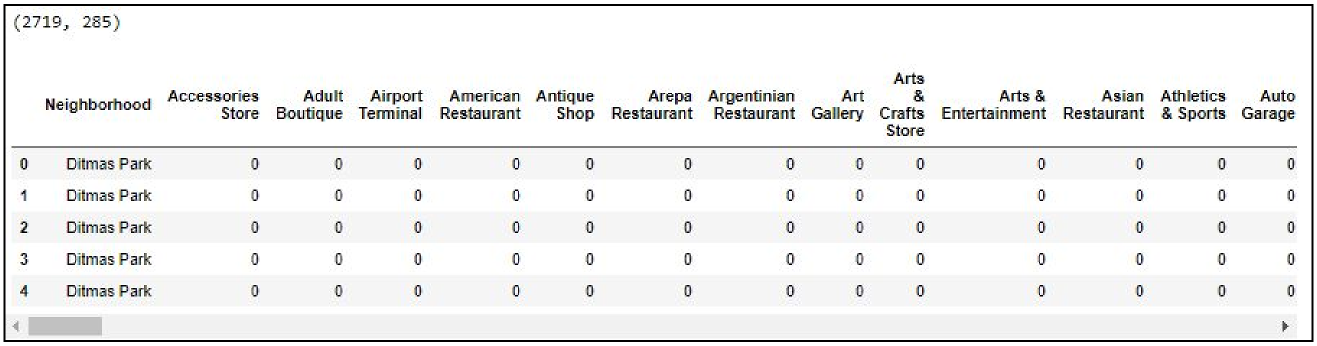
Staten Island can also be looked into next (High population density with very few places).

4.6 Preprocessing



4.6.1 One Hot Encoding

The data as mentioned above contains details of the nearby venues - Location, Category etc. This data needs to be transformed into a suitable format prior to Clustering. One Hot Encoding is first performed on the ‘Venue Category’ attribute. This is done using the pandas get_dummies() function. Encoding assigns a Nominal Value to our Categorical data so the model does not interpret any numbers as importance or weight.



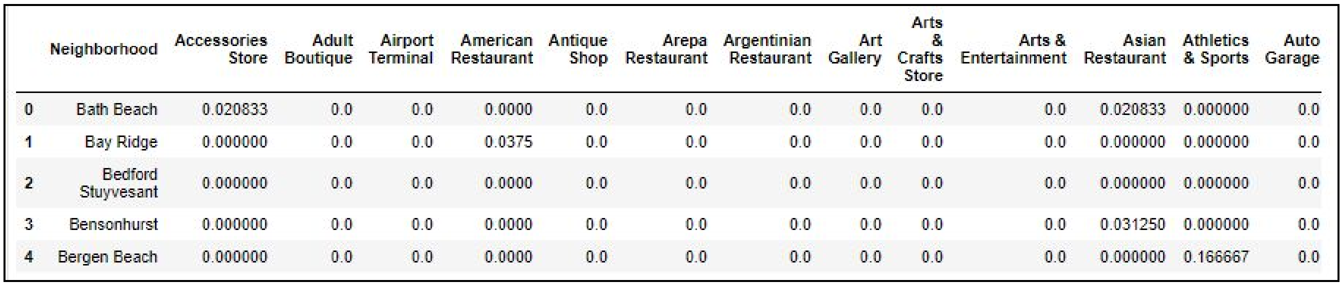
```
(2719, 285)
```

	Neighborhood	Accessories Store	Adult Boutique	Airport Terminal	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports	Auto Garage
0	Ditmas Park	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Ditmas Park	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Ditmas Park	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Ditmas Park	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Ditmas Park	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig 4.8 One-Hot Encoding

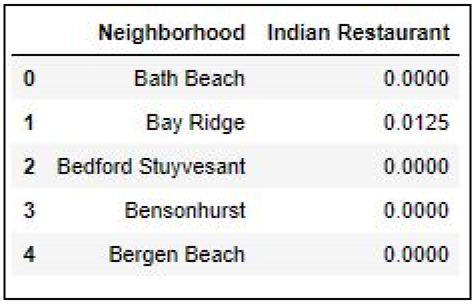
4.6.2 Grouping the Categories

The new dataframe is now grouped by Neighborhood and the mean for each Category is taken. This gives an average estimate for each Category in the neighborhood.



	Neighborhood	Accessories Store	Adult Boutique	Airport Terminal	American Restaurant	Antique Shop	Arepa Restaurant	Argentinian Restaurant	Art Gallery	Arts & Crafts Store	Arts & Entertainment	Asian Restaurant	Athletics & Sports	Auto Garage
0	Bath Beach	0.020833	0.0	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.020833	0.000000	0.0
1	Bay Ridge	0.000000	0.0	0.0	0.0375	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0
2	Bedford Stuyvesant	0.000000	0.0	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0
3	Bensonhurst	0.000000	0.0	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.031250	0.000000	0.0
4	Bergen Beach	0.000000	0.0	0.0	0.0000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.166667	0.0

Fig 4.9 Grouping the Categories by Neighborhood



	Neighborhood	Indian Restaurant
0	Bath Beach	0.0000
1	Bay Ridge	0.0125
2	Bedford Stuyvesant	0.0000
3	Bensonhurst	0.0000
4	Bergen Beach	0.0000

Fig 4.10 Filtering only for Indian Restaurants

Once this is done, we then select only the Indian Restaurants and Neighborhoods as the other Attributes are not of concern to us. This dataframe is used to cluster the data points.

4.7 Clustering

4.7.1 Selecting K Value

The ‘k’ stands for - number of clusters. It’s value in k-means Clustering is selected by the “Elbow Method”. The Elbow Method involves plotting the Cost vs k-value; where k is an integer > 1. The point where the curve makes a transition is generally chosen as the k-value. I have used a k value of 3 for the Analysis, although there was a transition at k=2, the cost decreased further at k=3 and this would give us more diverse Clusters to examine. The aim is to minimize the Within-Cluster-Sum-of-Squares - Cost by using the Inertia criteria in the sklearn library.

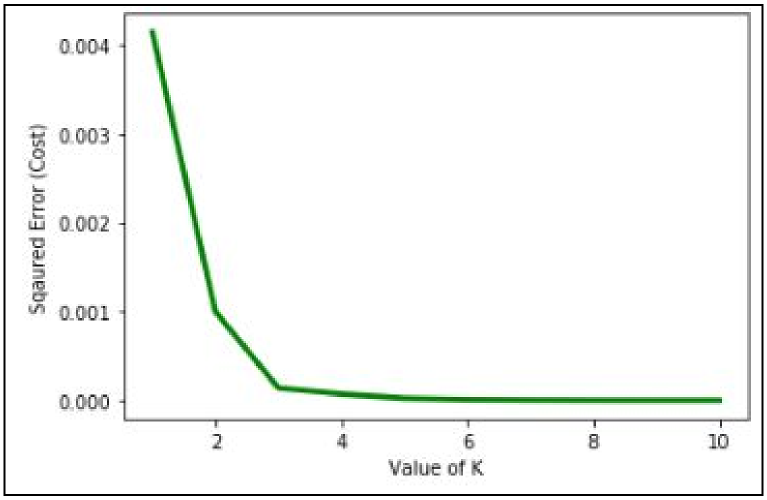
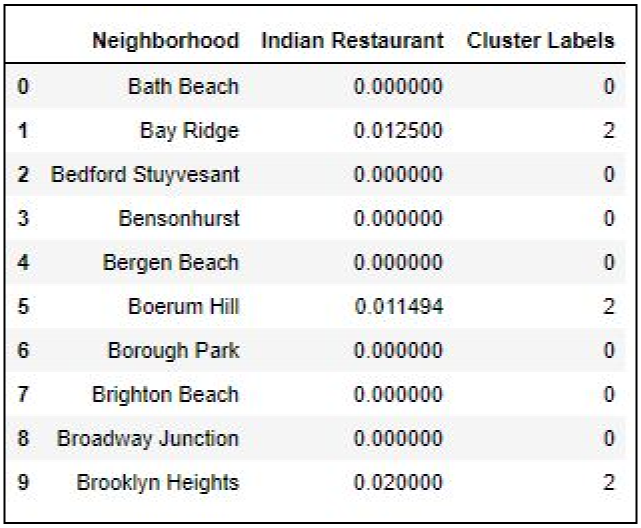


Fig 4.11 Elbow Plot of Cost Vs k

4.7.2 Cluster Labels

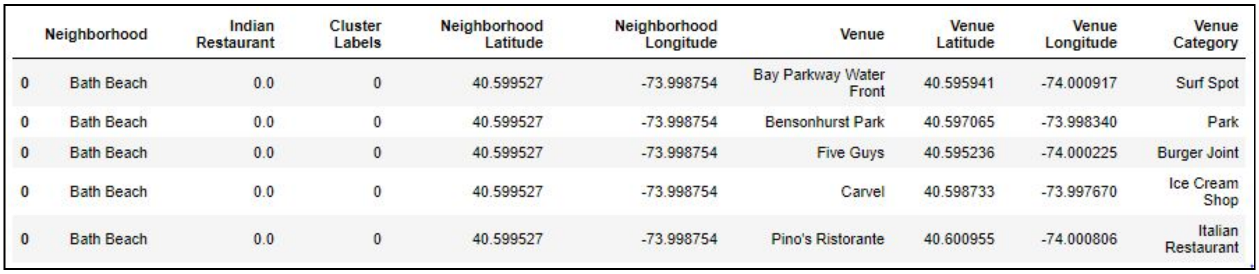
Next Clustering is performed and the Cluster Labels are saved. The Cluster Labels are merged with the previous dataframe containing only Indian Restaurants.



	Neighborhood	Indian Restaurant	Cluster Labels
0	Bath Beach	0.000000	0
1	Bay Ridge	0.012500	2
2	Bedford Stuyvesant	0.000000	0
3	Bensonhurst	0.000000	0
4	Bergen Beach	0.000000	0
5	Boerum Hill	0.011494	2
6	Borough Park	0.000000	0
7	Brighton Beach	0.000000	0
8	Broadway Junction	0.000000	0
9	Brooklyn Heights	0.020000	2

Fig 4.12 Cluster Labels for the Neighborhoods

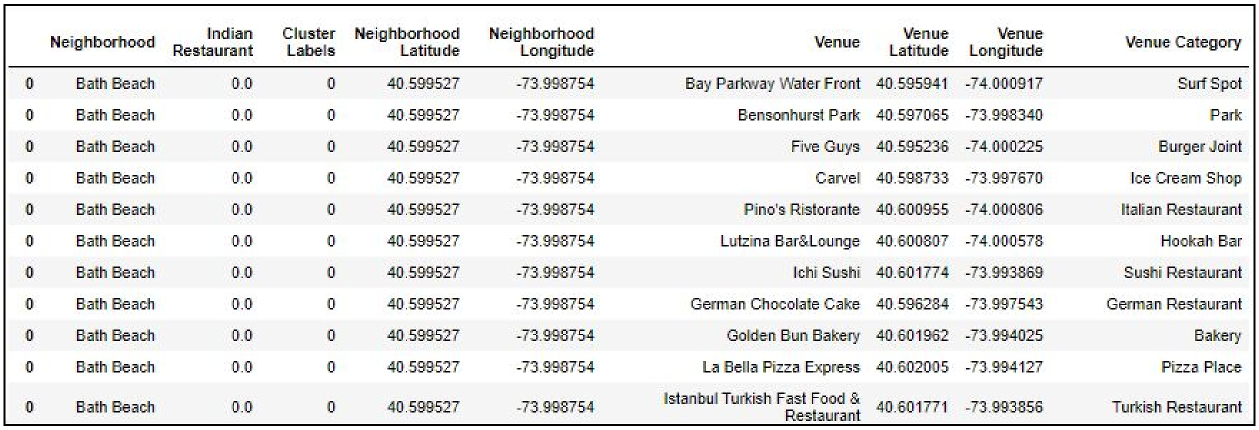
This dataframe is then joined with the Brooklyn Venues dataframe.



	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bath Beach	0.0	0	40.599527	-73.998754	Bay Parkway Water Front	40.595941	-74.000917	Surf Spot
0	Bath Beach	0.0	0	40.599527	-73.998754	Bensonhurst Park	40.597065	-73.998340	Park
0	Bath Beach	0.0	0	40.599527	-73.998754	Five Guys	40.595236	-74.000225	Burger Joint
0	Bath Beach	0.0	0	40.599527	-73.998754	Carvel	40.598733	-73.997670	Ice Cream Shop
0	Bath Beach	0.0	0	40.599527	-73.998754	Pino's Ristorante	40.600955	-74.000806	Italian Restaurant

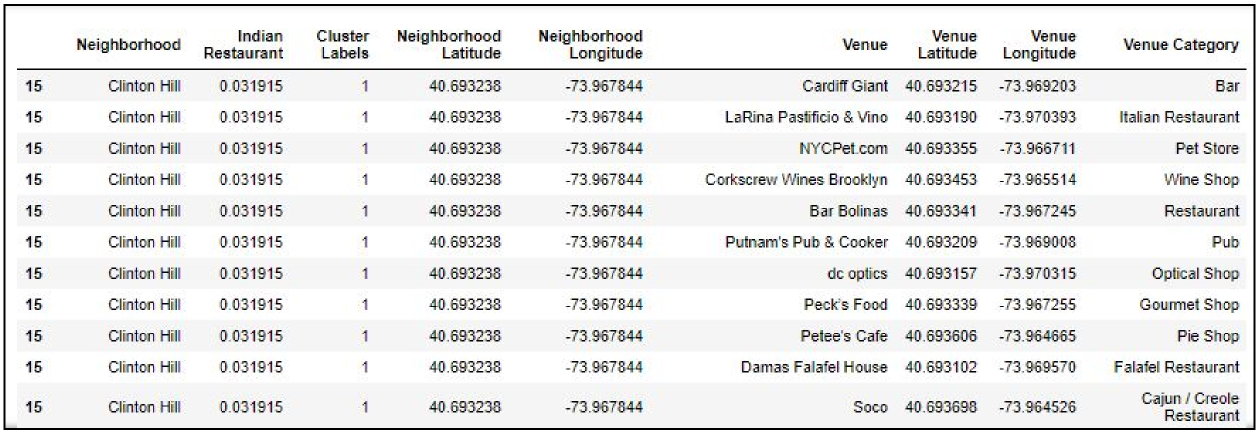
Fig 4.13 Merged Venues Dataframe with Labels

4.8 Clusters



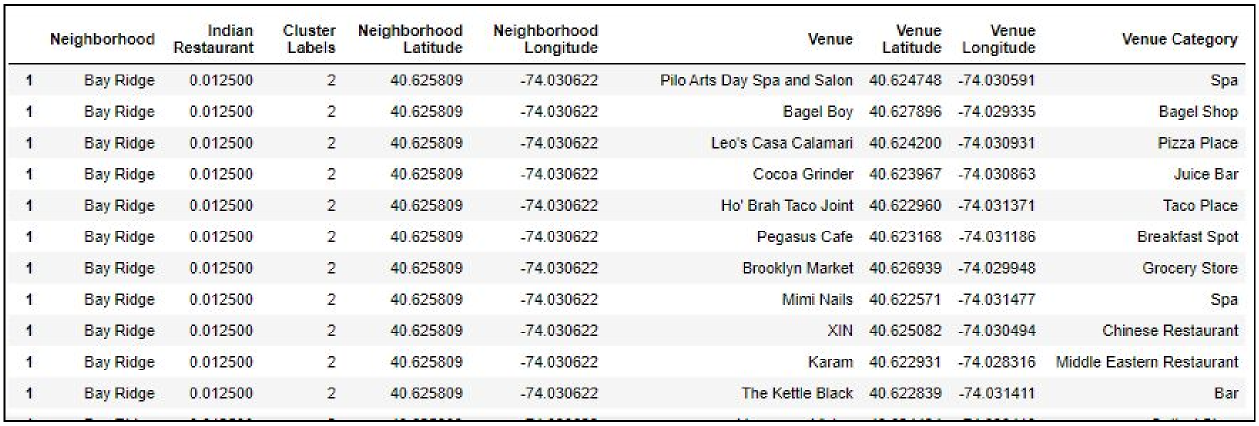
	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bath Beach	0.0	0	40.599527	-73.998754	Bay Parkway Water Front	40.595941	-74.000917	Surf Spot
0	Bath Beach	0.0	0	40.599527	-73.998754	Bensonhurst Park	40.597065	-73.998340	Park
0	Bath Beach	0.0	0	40.599527	-73.998754	Five Guys	40.595236	-74.000225	Burger Joint
0	Bath Beach	0.0	0	40.599527	-73.998754	Carvel	40.598733	-73.997670	Ice Cream Shop
0	Bath Beach	0.0	0	40.599527	-73.998754	Pino's Ristorante	40.600955	-74.000806	Italian Restaurant
0	Bath Beach	0.0	0	40.599527	-73.998754	Lutzina Bar&Lounge	40.600807	-74.000578	Hookah Bar
0	Bath Beach	0.0	0	40.599527	-73.998754	Ichi Sushi	40.601774	-73.993869	Sushi Restaurant
0	Bath Beach	0.0	0	40.599527	-73.998754	German Chocolate Cake	40.596284	-73.997543	German Restaurant
0	Bath Beach	0.0	0	40.599527	-73.998754	Golden Bun Bakery	40.601962	-73.994025	Bakery
0	Bath Beach	0.0	0	40.599527	-73.998754	La Bella Pizza Express	40.602005	-73.994127	Pizza Place
0	Bath Beach	0.0	0	40.599527	-73.998754	Istanbul Turkish Fast Food & Restaurant	40.601771	-73.993856	Turkish Restaurant

Cluster 0



	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
15	Clinton Hill	0.031915	1	40.693238	-73.967844	Cardiff Giant	40.693215	-73.969203	Bar
15	Clinton Hill	0.031915	1	40.693238	-73.967844	LaRina Pastificio & Vino	40.693190	-73.970393	Italian Restaurant
15	Clinton Hill	0.031915	1	40.693238	-73.967844	NYCPet.com	40.693355	-73.966711	Pet Store
15	Clinton Hill	0.031915	1	40.693238	-73.967844	Corkscrew Wines Brooklyn	40.693453	-73.965514	Wine Shop
15	Clinton Hill	0.031915	1	40.693238	-73.967844	Bar Bolinas	40.693341	-73.967245	Restaurant
15	Clinton Hill	0.031915	1	40.693238	-73.967844	Putnam's Pub & Cooker	40.693209	-73.969008	Pub
15	Clinton Hill	0.031915	1	40.693238	-73.967844	dc optics	40.693157	-73.970315	Optical Shop
15	Clinton Hill	0.031915	1	40.693238	-73.967844	Peck's Food	40.693339	-73.967255	Gourmet Shop
15	Clinton Hill	0.031915	1	40.693238	-73.967844	Petee's Cafe	40.693606	-73.964665	Pie Shop
15	Clinton Hill	0.031915	1	40.693238	-73.967844	Damas Falafel House	40.693102	-73.969570	Falafel Restaurant
15	Clinton Hill	0.031915	1	40.693238	-73.967844	Soco	40.693698	-73.964526	Cajun / Creole Restaurant

Cluster 1



	Neighborhood	Indian Restaurant	Cluster Labels	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Bay Ridge	0.012500	2	40.625809	-74.030622	Pilo Arts Day Spa and Salon	40.624748	-74.030591	Spa
1	Bay Ridge	0.012500	2	40.625809	-74.030622	Bagel Boy	40.627896	-74.029335	Bagel Shop
1	Bay Ridge	0.012500	2	40.625809	-74.030622	Leo's Casa Calamari	40.624200	-74.030931	Pizza Place
1	Bay Ridge	0.012500	2	40.625809	-74.030622	Cocoa Grinder	40.623967	-74.030863	Juice Bar
1	Bay Ridge	0.012500	2	40.625809	-74.030622	Ho' Brah Taco Joint	40.622960	-74.031371	Taco Place
1	Bay Ridge	0.012500	2	40.625809	-74.030622	Pegasus Cafe	40.623168	-74.031186	Breakfast Spot
1	Bay Ridge	0.012500	2	40.625809	-74.030622	Brooklyn Market	40.626939	-74.029948	Grocery Store
1	Bay Ridge	0.012500	2	40.625809	-74.030622	Mimi Nails	40.622571	-74.031477	Spa
1	Bay Ridge	0.012500	2	40.625809	-74.030622	XIN	40.625082	-74.030494	Chinese Restaurant
1	Bay Ridge	0.012500	2	40.625809	-74.030622	Karam	40.622931	-74.028316	Middle Eastern Restaurant
1	Bay Ridge	0.012500	2	40.625809	-74.030622	The Kettle Black	40.622839	-74.031411	Bar

Cluster 2

5. RESULTS

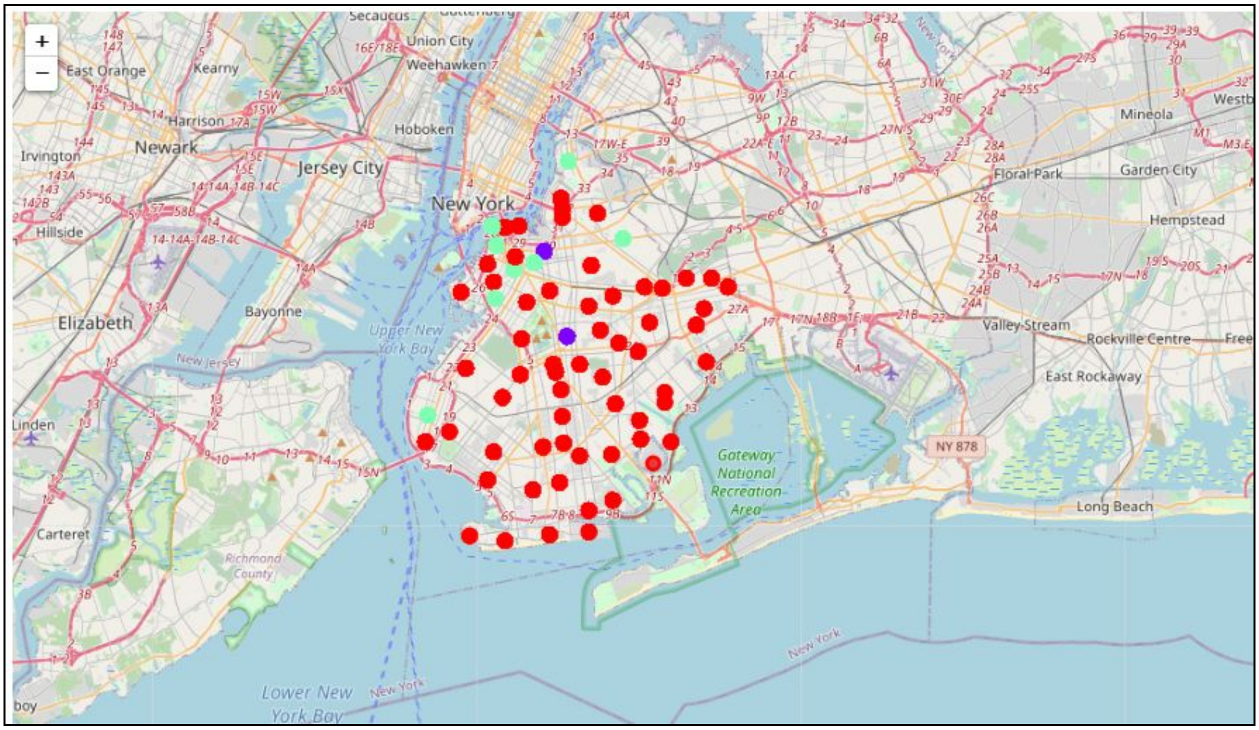


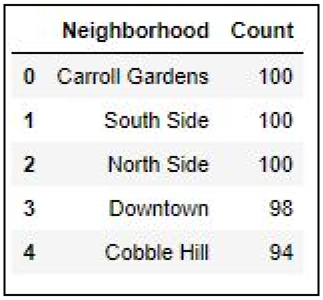
Fig 5.1 Plot of all the Clusters

Based on the Clustering,

Cluster 2 : has the most number of Indian Restaurants and is therefore not considered.

Cluster 1 : has a medium number of Restaurants.

Cluster 0 : is ideal as no restaurants are present. Therefore we can look into the places in this Cluster.



	Neighborhood	Count
0	Carroll Gardens	100
1	South Side	100
2	North Side	100
3	Downtown	98
4	Cobble Hill	94

Fig 5.2 Most common Neighborhoods in the Cluster

Looking at nearby venues, it seems Cluster 0 might be a good location as there are not a lot of Indian restaurants in these areas. There are 60 odd neighborhoods present in the Cluster and the most common ones being **Carroll Gardens, South Side, North Side, Downtown** and **Cobble Hill** in **Brooklyn**.

Therefore our Indian Restaurant can be opened in any of these neighborhoods with little to no competition. Nonetheless, if the food is affordable, authentic and has good taste, I am confident that it will have a great following everywhere.

6. DISCUSSION

Based on the analysis Carroll Gardens, South Side, North Side, Downtown and Cobble Hill are some of the neighborhoods to consider opening our restaurant. I also looked into Staten Island as it had a similar Indian Population as Brooklyn. Since there are only 2 restaurants, competition is very low. Staten Island also has a high density of Indian Americans per sq mile so foot traffic should not be a problem, but the lack of Indian Restaurants can also hint at various other problems like a Licensing, stringent community norms etc, something which should be looked into before making a decision. Manhattan has the most number of Indian Restaurants but the least number of Indian Americans, something that might be interesting to look into.

Some of the drawbacks of this analysis are — the clustering is completely based only on data obtained from Foursquare API and the data about the Indian population distribution in each neighborhood is also based on the 2014 census which is not up-to-date. Thus there is a huge gap in the population distribution data. Even Though there are lots of areas where it can be improved yet this analysis has certainly provided us with some good insights, preliminary information on possibilities & a head start into this business problem by setting the step stones properly.

7. CONCLUSION

We have worked on a business problem like how a real data scientist would. We used python libraries to fetch the data (json, requests etc), to manipulate the contents (pandas) & to analyze and visualize(matplotlib, Folium) those datasets. We have made use of the Foursquare API to explore the venues in neighborhoods of New York, then get data from Wikipedia which we scraped using the pandas library. We also applied machine learning techniques (Clustering) to predict the output given the data and used Folium to visualize it on a map.

Analysis can further be improved by using more recent data and making use of more complex Machine Learning Algorithms. This process however can be used as a baseline and be replicated for other cuisines or gyms, etc.

8. CITATIONS

[1] https://geo.nyu.edu/catalog/nyu_2451_34572

[2] https://foursquare.com/developers/apps

[3] https://en.wikipedia.org/wiki/Indians_in_the_New_York_City_metropolitan_region