Analyzing the best Neighborhoods to open an Indian Restaurant chain.

By - Manohar Rao

1.Introduction

1.1 Background:

With all its exotic ingredients, unfamiliar dishes, and tongue-tingling flavors, Indian cuisine can be both exciting and intimidating. “It’s such a complete world of taste. You combine all the techniques from other cuisines and add magical spices to get a titillating food experience.

Indian Cuisine has been very famous in United states, Especially in Manhattan. People in US love to explore different Cuisine and they love the punch of Spices which Indian cuisine offers.

A well-known Restaurant owner who has multiple well known Restaurant franchises in India wanted to open his chain of Restaurants in different Neighborhoods of Manhattan. However, he didn’t not want to open his chain where Indian restaurants was already available. He wanted to open the restaurant in Neighborhoods where the people visiting most common Venues had different types of Restaurants and where there was a Lack of Indian Restaurant.

1.2 Problem:

The problem is to figure out on which Neighborhoods can he Open his Restaurants and succeed. But his only concern was he didn’t want to open his Restaurant in a Neighborhood where people don’t often go out seeking different Cuisine food. Also he wanted to open at least two Restaurants in Manhattan.

2. Data Acquisition

2.1 Data Source.

Foursquare is a site which provide all the venue details and many more details regarding the venues which can be then analyzed to figure out a) Most visited Venues in each Cities and find top 10 according to Neighborhoods. b) Ratings of the Venues especially if they are restaurants.

We would also need the data which contains the Borough, Neighborhood names and their latitude and longitude locations for both the places which will be parsed from the web for Toronto and using readily available data for NY.

Below are the links to the Data Sources used in this Project

1. Luckily, the dataset which contains all the New York Borouge and Neighbourhood exists for free on the web. Link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572
2. Foursquare.com for all the Venue Sorucing.

2.2 Data Cleaning.

For the NewYork data the information was readily available in a JSON format hence it just needed some cleaning and transforming the data into Pandas dataframe for best analysis.

3. Data Analysis

Here is how I analyzed the data.

1.After downloading the data from the source it was in a JSON format and noticed all the relevant data is in the *"features key"*. Displaying just the first Neighbourhood

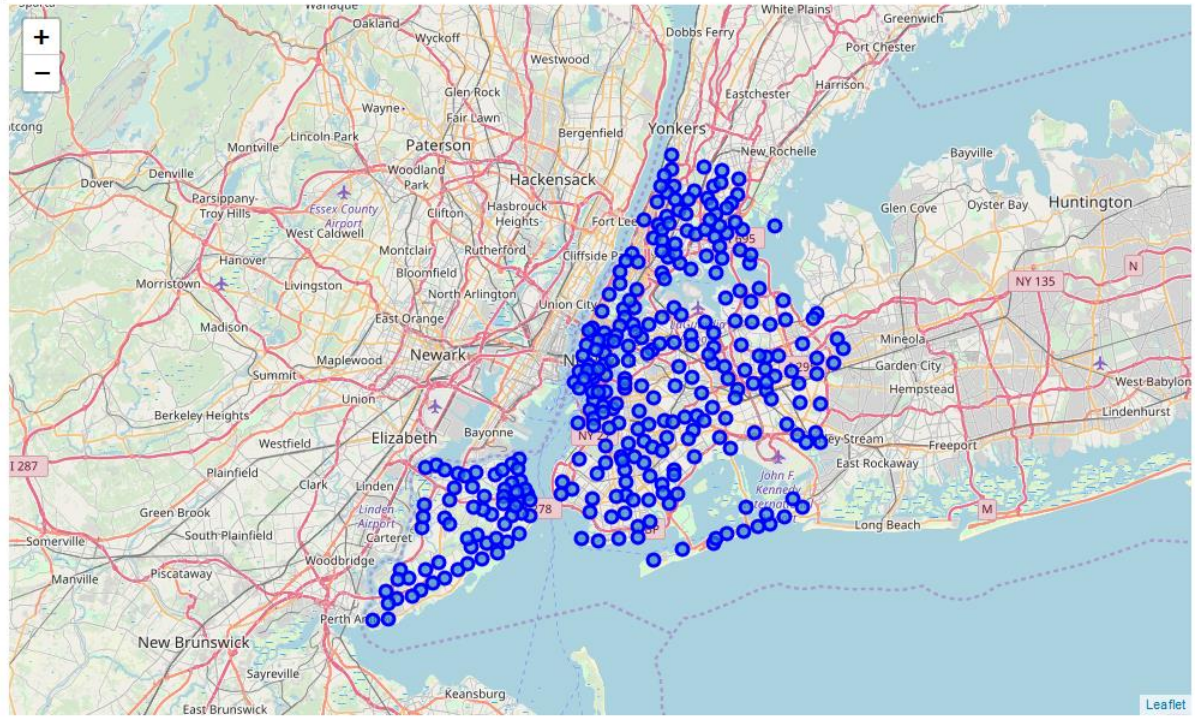
```
Out[6]: {'geometry': {'coordinates': [-73.84720052054902, 40.89470517661],
          'type': 'Point'},
         'geometry_name': 'geom',
         'id': 'nyu_2451_34572.1',
         'properties': {'annoangle': 0.0,
          'annoline1': 'Wakefield',
          'annoline2': None,
          'annoline3': None,
          'bbox': [-73.84720052054902,
           40.89470517661,
           -73.84720052054902,
           40.89470517661],
          'borough': 'Bronx',
          'name': 'Wakefield',
          'stacked': 1},
         'type': 'Feature'}
```

Now the challenge was to convert the above to a dataframe.

2. To convert the above to a dataframe, I created an empty DataFrame and started filling the DF by looping through data.

3. Using the Folium library briefly displaying the NewYork Map with the neighborhoods marked in Blue circles.

```
map_newyork
```



4. As we just need the Manhattan Neighborhoods I sliced the Df to contain only Manhattan Neighborhoods as below.

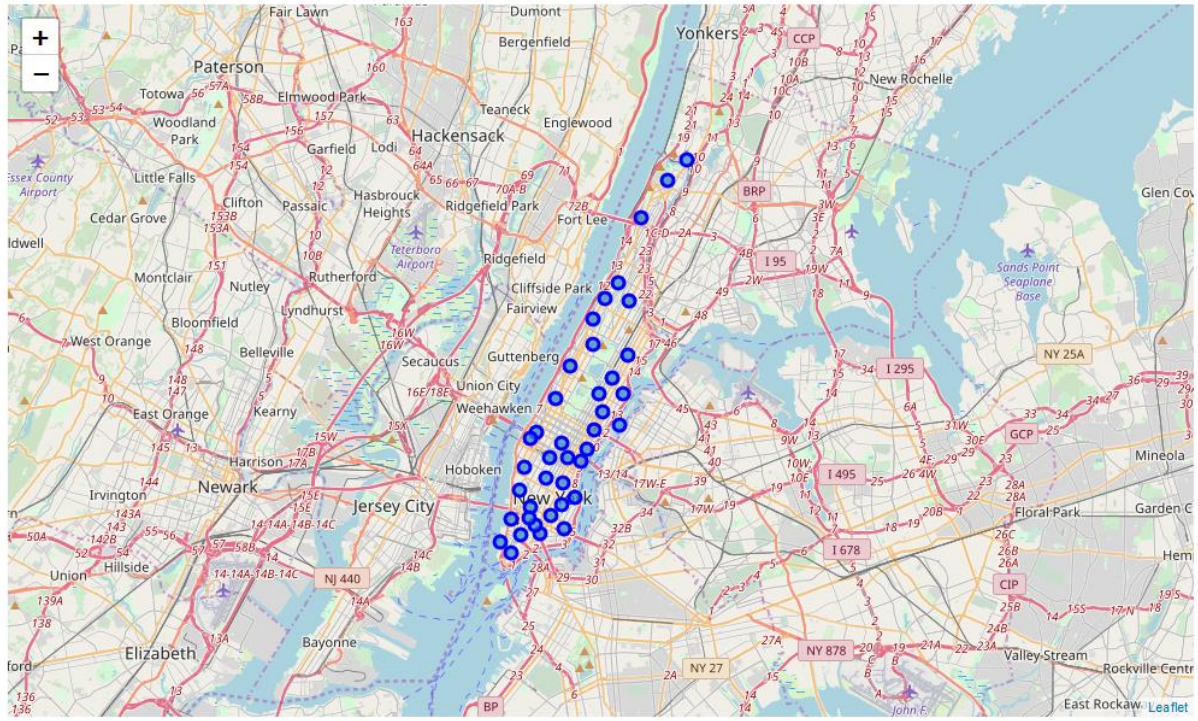
```
manhattan_data = neighborhoods[neighborhoods['Borough'] == 'Manhattan'].reset_index(drop=True)
display(manhattan_data.head()) #
display(manhattan_data.shape) #Gives us a count of Neighbourhoods.
```

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

```
(40, 4)
```

5. Now displaying the NY map which displays Neighborhoods of Manhattan.

```
map_manhattan
```



5. Now the Map data was ready with the help of Forusquare API I got the TOP 100 Venues for all the neighborhoods within 500Mts radius and the example looked like this.

```
results = requests.get(url).json()
results
```


```
{'meta': {'code': 200, 'requestId': '5c8aa99bdd57977684a56bba'},
 'response': {'groups': [{'items': [{'reasons': {'count': 0,
       'items': [{'reasonName': 'globalInteractionReason',
         'summary': 'This spot is popular',
         'type': 'general'}]},
     'referralId': 'e-0-4b4429abf964a52037f225e3-0',
     'venue': {'categories': [{'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/pizza_',
         'suffix': '.png'},
        'id': '4bf58dd8d48988d1ca941735',
        'name': 'Pizza Place',
        'pluralName': 'Pizza Places',
        'primary': True,
        'shortName': 'Pizza'}],
      'delivery': {'id': '72548',
       'provider': {'icon': {'name': '/delivery_provider_seamless_20180129.png',
         'prefix': 'https://fastly.4sqi.net/img/general/cap/',
```

6. As the information is in the items key. Before I proceed, I borrowed the get_category_type function from the Foursquare lab. Which basically returns the category of Venues. Callnng the function on only one Neighborhood I received 24 Venues. After cleaning the Json file by deleting all other columns which was not necessary for my analysis and saving it into a DF I got.

	name	categories	lat	lng
0	Arturo's	Pizza Place	40.874412	-73.910271
1	Bikram Yoga	Yoga Studio	40.876844	-73.906204
2	Tibbett Diner	Diner	40.880404	-73.908937
3	Starbucks	Coffee Shop	40.877531	-73.905582
4	Dunkin' Donuts	Donut Shop	40.876993	-73.906507

7. Now I looped through all the Neighborhoods and called the Foursquare API and created a DF which as below while displaying the DF head.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin' Donuts	40.876993	-73.906507	Donut Shop

And counting the number of Venues returned for each neighborhood shows us the below for few neighborhood

	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Neighborhood						
Battery Park City	99	99	99	99	99	99
Carnegie Hill	100	100	100	100	100	100
Central Harlem	44	44	44	44	44	44
Chelsea	100	100	100	100	100	100
Chinatown	100	100	100	100	100	100
Civic Center	100	100	100	100	100	100
Clinton	100	100	100	100	100	100
East Harlem	45	45	45	45	45	45
East Village	100	100	100	100	100	100
Financial District	100	100	100	100	100	100
Flatiron	100	100	100	100	100	100
Gramercy	100	100	100	100	100	100
Greenwich Village	100	100	100	100	100	100
Hamilton Heights	59	59	59	59	59	59
Hudson Yards	62	62	62	62	62	62
Inwood	54	54	54	54	54	54
Lenox Hill	100	100	100	100	100	100
Lincoln Square	100	100	100	100	100	100
Little Italy	100	100	100	100	100	100

8. As there were 330 Unique Venues returned the best approach was to figure out the frequency to find the top 100. Hence used the One hot encoding and created a frequency mean df from the data as below.

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	Art Gallery
0	Battery Park City	0.000000	0.00	0.00	0.000000	0.010101	0.00	0.00	0.00	0.000000	0.000000	0.000000
1	Carnegie Hill	0.000000	0.00	0.00	0.000000	0.010000	0.00	0.00	0.00	0.000000	0.010000	0.000000
2	Central Harlem	0.000000	0.00	0.00	0.068182	0.045455	0.00	0.00	0.00	0.000000	0.000000	0.022727
3	Chelsea	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.00	0.000000	0.000000	0.020000

9. So with the help of above when we print the top 5 for few Neighborhoods below is what it looks like.

```
----Battery Park City----
                venue  freq
0                Park  0.08
1         Coffee Shop  0.08
2               Hotel  0.05
3                 Gym  0.03
4  Italian Restaurant  0.03


----Carnegie Hill----
                venue  freq
0        Pizza Place  0.06
1        Coffee Shop  0.06
2               Café  0.04
3                Bar  0.04
4  French Restaurant  0.03


----Central Harlem----
                  venue  freq
0    African Restaurant  0.07
1     French Restaurant  0.05
2    Chinese Restaurant  0.05
3   American Restaurant  0.05
4  Gym / Fitness Center  0.05


----Chelsea----
                  venue  freq
0          Coffee Shop  0.07
1   Italian Restaurant  0.06
2       Ice Cream Shop  0.05
3               Bakery  0.04
4  American Restaurant  0.04
```

So the frequency have us a brief picture to figure out the top 100 venues in each Neighborhoods.

9. By creating a new DF and displaying the top 100.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Mc Comm Veni
0	Battery Park City	Coffee Shop	Park	Hotel	Gym	Wine Shop	Italian Restaurant	Shopping Mall	Fountain	BBQ Joint	Plaza
1	Carnegie Hill	Coffee Shop	Pizza Place	Café	Bar	Japanese Restaurant	French Restaurant	Gym	Yoga Studio	Bookstore	Cosmetic Shop
2	Central Harlem	African Restaurant	American Restaurant	French Restaurant	Gym / Fitness Center	Chinese Restaurant	Seafood Restaurant	Public Art	Pizza Place	Bookstore	Library
3	Chelsea	Coffee Shop	Italian Restaurant	Ice Cream Shop	Nightclub	American Restaurant	Bakery	Seafood Restaurant	Hotel	Theater	Asian Restaura
		Chinese	Vietnamese	Dim Sum	American	Cocktail	Ice Cream	Hotpot	Salon /		

Screenshot only contain few data as the data has 101 columns.

10. In the above dataframe I dropped the Rows(Neighborhoods) which already had Indian restaurants.

Let's check if the top 100 Venues has Indian Restaurant or not and drop the once which already has Indian.

```
#From 40 Neighbourhoods to just 20 Neighbourhoods which doesnt even have Indian Restaurants.
neighborhoods_venues_woindian_restaurant = neighborhoods_venues_sorted[neighborhoods_venues_sorted.iloc[:, 0:] != "Indian Res
taurant"]
neighborhoods_venues_woindian_restaurant.dropna(axis=0, inplace=True)
display(neighborhoods_venues_woindian_restaurant)
display(neighborhoods_venues_woindian_restaurant.shape)
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Battery Park City	Coffee Shop	Park	Hotel	Gym	Wine Shop	Italian Restaurant	Shopping Mall	Fountain	BBQ Joint	Plaza
2	Central Harlem	African Restaurant	American Restaurant	French Restaurant	Gym / Fitness Center	Chinese Restaurant	Seafood Restaurant	Public Art	Pizza Place	Bookstore	Library
4	Chinatown	Chinese Restaurant	Vietnamese Restaurant	Dim Sum Restaurant	American Restaurant	Cocktail Bar	Ice Cream Shop	Hotpot Restaurant	Salon / Barbershop	Noodle House	Bakery
6	Clinton	Theater	Gym / Fitness Center	Hotel	American Restaurant	Indie Theater	Gym	Italian Restaurant	Coffee Shop	Spa	Wine Shop
7	East Harlem	Mexican Restaurant	Bakery	Deli / Bodega	Latin American Restaurant	Thai Restaurant	Pharmacy	Street Art	Cocktail Bar	Clothing Store	Pet Store
8	East Village	Bar	Ice Cream Shop	Wine Bar	Mexican Restaurant	Vegetarian / Vegan Restaurant	Japanese Restaurant	Coffee Shop	Cocktail Bar	Ramen Restaurant	Pizza Place
11	Gramercy	Italian Restaurant	Bagel Shop	Thrift / Vintage Store	Pizza Place	Coffee Shop	Cocktail Bar	Restaurant	Mexican Restaurant	Bar	Grocery Store

Notice how the row index has changed displaying the missing deleted rows.

when we check the shape we just had 20 rows 20 Neighborhoods were dropped `(20, 101)`

11. Now that I have 20 Neighborhood which doesnt have Indian Restaurant I can just use the top 10 Venues for further analysis(which is to find out which Neighborhood is more centered to Restaurants in the top 10 venues)

12.Created an extra column to count the number of restaurants in each neighborhoods gave me the below

	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Total_Restaurants
	Coffee Shop	Park	Hotel	Gym	Wine Shop	Italian Restaurant	Shopping Mall	Fountain	BBQ Joint	Plaza	1
	African Restaurant	American Restaurant	French Restaurant	Gym / Fitness Center	Chinese Restaurant	Seafood Restaurant	Public Art	Pizza Place	Bookstore	Library	5
	Chinese Restaurant	Vietnamese Restaurant	Dim Sum Restaurant	American Restaurant	Cocktail Bar	Ice Cream Shop	Hotpot Restaurant	Salon / Barbershop	Noodle House	Bakery	5
	Theater	Gym / Fitness Center	Hotel	American Restaurant	Indie Theater	Gym	Italian Restaurant	Coffee Shop	Spa	Wine Shop	2
	Mexican Restaurant	Bakery	Deli / Bodega	Latin American Restaurant	Thai Restaurant	Pharmacy	Street Art	Cocktail Bar	Clothing Store	Pet Store	3
	Bar	Ice Cream Shop	Wine Bar	Mexican Restaurant	Vegetarian / Vegan Restaurant	Japanese Restaurant	Coffee Shop	Cocktail Bar	Ramen Restaurant	Pizza Place	4
	Italian Restaurant	Bagel Shop	Thrift / Vintage Store	Pizza Place	Coffee Shop	Cocktail Bar	Restaurant	Mexican Restaurant	Bar	Grocery Store	3
	Italian Restaurant	Coffee Shop	Restaurant	Gym / Fitness Center	American Restaurant	Café	Hotel	Theater	Dog Run	Art Gallery	3
	Mexican		Pizza			Frozen	Chinese			Deli /	

13. As the Owner was looking to open atleast 2 new Restaurant I had to select at least two Neighborhoods. Looking at a brief statistics of the final column we see that 75% of the Neighorhoods have More than 4 Restaurant Venues in their top 10.

```
#Filtering Neighborhoods with top 10 venues containing more than 4 Restaurants as the 75th percentile is 4 Restaurant I am ai
ming more than 4.
display(Neighborhood_top10_venues["Total_Restaurants"].describe())
f = Neighborhood_top10_venues[Neighborhood_top10_venues["Total_Restaurants"] > 4]
display(f)
```

```
count    20.000000
mean      2.900000
std       1.209611
min       1.000000
25%       2.000000
50%       3.000000
75%       4.000000
max       5.000000
Name: Total_Restaurants, dtype: float64
```

Hence Just considering Neighborhoods which only contain only more than 4 Restaurant.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	Total_Resta
1	Central Harlem	African Restaurant	American Restaurant	French Restaurant	Gym / Fitness Center	Chinese Restaurant	Seafood Restaurant	Public Art	Pizza Place	Bookstore	Library	5
2	Chinatown	Chinese Restaurant	Vietnamese Restaurant	Dim Sum Restaurant	American Restaurant	Cocktail Bar	Ice Cream Shop	Hotpot Restaurant	Salon / Barbershop	Noodle House	Bakery	5

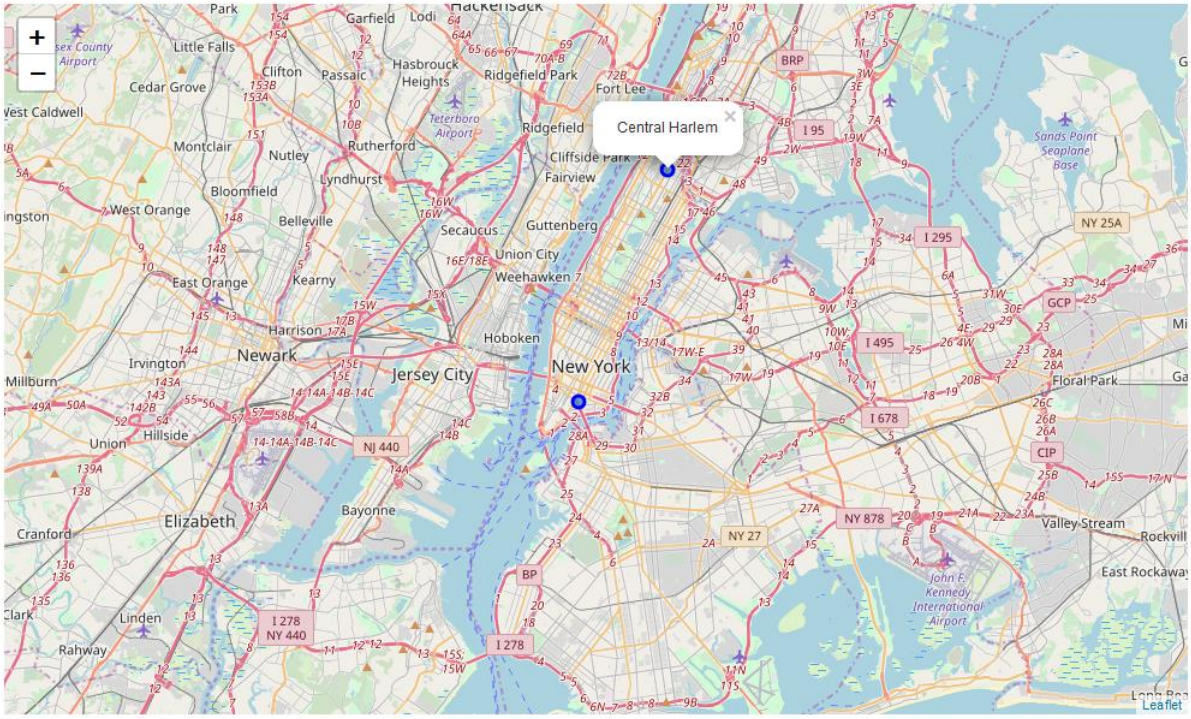
4. Result

The above dataframe shows that in Manhattan there are two Neighborhoods which has 5 out of 10 Venues as Restaurant and none of the Neighborhood has an Indian Restaurant.

5. Discussion

Displaying the Neighborhood on NY maps shows both are on unique location and not close to one another.

```
final_map_manhattan
```



6. Conclusion

Concluding that the Restaurant owner can open his two new Indian Restaurants in ***"Chainatown Manhattan"*** and ***"Central Harlem Manhattan"***.