The Battle of Neighbourhoods

IBM Data Science Capstone Project

Fayaz Ahmad

1. Introduction

New York City's demographics show that it is a large and ethnically diverse metropolis. It is the largest city in the United States with a long history of international immigration. New York City was home to nearly 8.5 million people in 2014, accounting for over 40% of the population of New York State and a slightly lower percentage of the New York metropolitan area, home to approximately 23.6 million. Over the last decade the city has been growing faster than the region. The New York region continues to be by far the leading metropolitan gateway for legal immigrants admitted into the United States.

Throughout its history, New York City has been a major point of entry for immigrants; the term "melting pot" was coined to describe densely populated immigrant neighbourhoods on the Lower East Side. As many as 800 languages are spoken in New York, making it the most linguistically diverse city in the world. English remains the most widely spoken language, although there are areas in the outer boroughs in which up to 25% of people speak English as an alternate language, and/or have limited or no English language fluency. English is least spoken in neighbourhoods such as Flushing, Sunset Park, and Corona.

With it's diverse culture, comes diverse food items. There are many restaurants in New york City, each belonging to different categories like Chinese, Indian , French etc.

So as part of this project, we will list and visualize all major parts of New York City that has great Indian restaurants

2. Analytic Approach

Data Science Methodology: From Problem Approach to Analytic Approach! Selecting the right analytic approach depends on the question being asked. The approach involves seeking clarification from the person who is asking the question, so as to be able to pick the most appropriate path or approach. Once the problem to be addressed is defined, the appropriate analytic approach for the problem is selected in the context of the business requirements. This is the second stage of the data science methodology. Once a strong understanding of the question is established, the analytic approach can be selected. In Analytical Approach, there would be certain questions that will be discussed with the client. Phrase the problem as a question to be answered using data

Can we automatically determine the best neighbours in New York City for Indian Restaurants? Focus would be on below questions.

What is the best location in New York City for Indian Cuisine?

Which areas have a potential Indian Restaurant Market?

Which are some of the best neighbourhoods for Indian cuisine?

Which is the best place to stay if you prefer Indian Cuisine?

3. Data Requirement

In this stage, data scientist needs to work closely with the clients to tell them the all the data requirements. All the questions raised during analytical approach must be catered in the phase.

New York City data that contains list Boroughs, Neighbourhoods along with their latitude and longitude.

Data source: https://cocl.us/new_york_dataset

Description: This data set contains the required information. And we will use this data set to explore various neighbourhoods of New York City.

Indian restaurants in each neighbourhood of New York City.

Data source: Foursquare API

Description: By using this API we will get all the venues in each neighbourhood. We can filter these venues to get only Indian restaurants

GeoSpace data

Data source: https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm

Description: By using this geo space data we will get the New york Borough boundaries that will help us visualize Choropleth maps.

4. Data Collection and Methodology

Next stage is collection, this is very iterative and consume hefty amount of time of the project. In this data scientist need to discuss with client all his requirements and collect the data for his model. It's very critical stage and all data must be collected correctly and completely.

1. Collect the New York City data from https://cocl.us/new_york_dataset

2. Using FourSquare API we will find all venues for each neighbourhood.

3. Filter out all venues that are Indian Restaurants.

4. Find rating, tips and like count for each Indian Restaurants using FourSquare API.

5. Using rating for each restaurant, we will sort that data.

6. Visualize the Ranking of neighbourhoods using folium library(python)

Moreover, analysis was done with following main python libraries

i. pandas and numpy for handling data.
ii. request module for using FourSquare API.
iii. geopy to get co-ordinates of City of New York.
iv. folium to visualize the results on a map

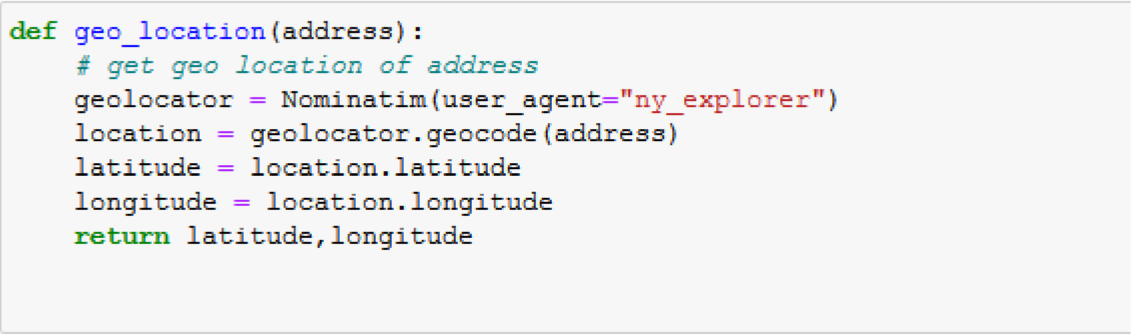
```
import pandas as pd
import numpy as np
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
import requests
import sys
from bs4 import BeautifulSoup
import os
!conda install -c conda-forge folium=0.5.0 --ye
import folium # map rendering library
from geopy.geocoders import Nominatim
import matplotlib.pyplot as plt
import matplotlib.cm as cm
import matplotlib.colors as colors
%matplotlib inline


print('Libraries imported.')
```

5. Data Understanding and Preparation

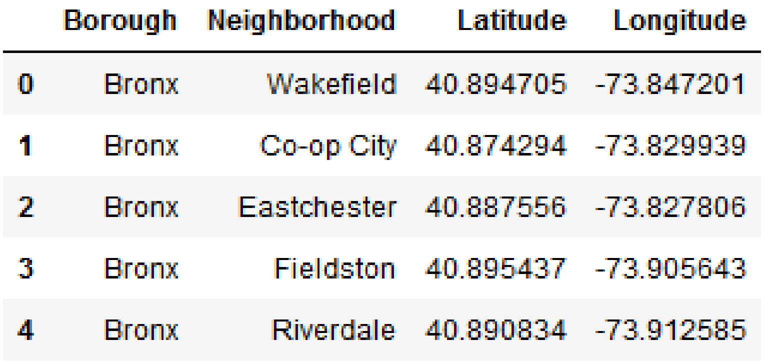
Data understanding and preparation is also very important stage of the model building. Data scientist should be very familiar with the data. If he has any confusion then he should clear it with client and understand it fully. Further, data scientists should transform the data in reasonable formats so that it can be used by machine learning algorithms. All the required data must be in good format and can be processed easily. If it not is required format then need to process it and change to desired format.

Different python functions and visualisation techniques were used to extract and prepare the data for easy understanding



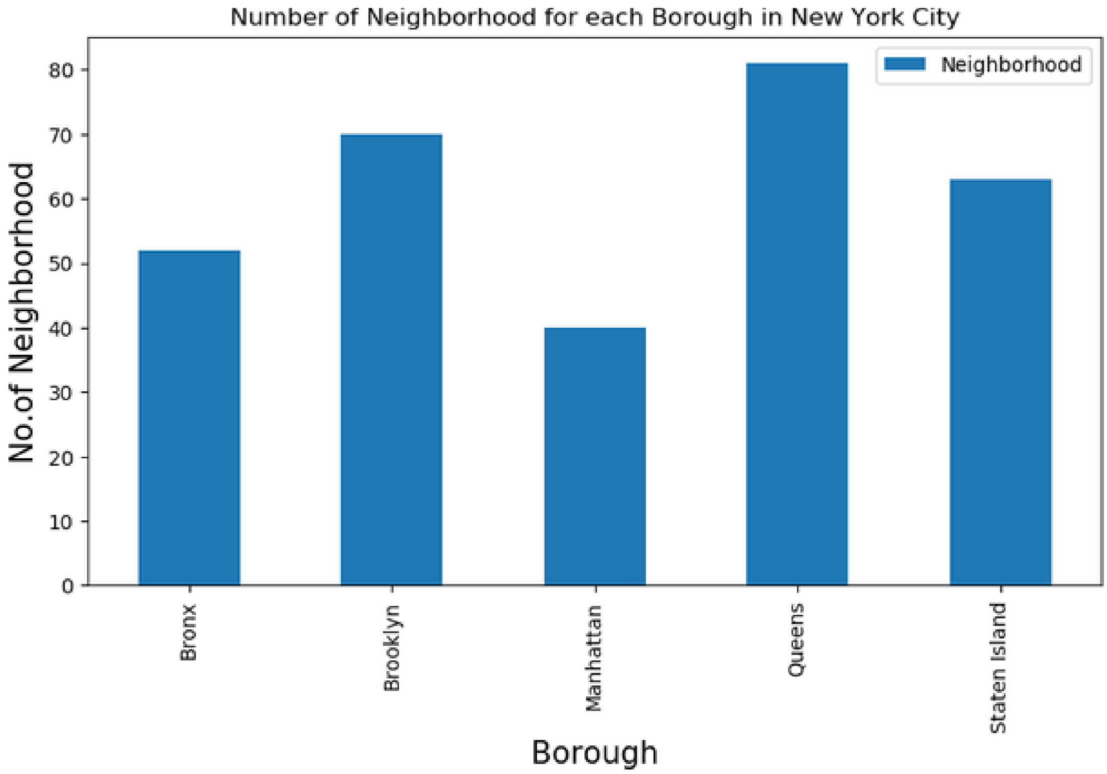
```
def geo_location(address):
    # get geo location of address
    geolocator = Nominatim(user_agent="ny_explorer")
    location = geolocator.geocode(address)
    latitude = location.latitude
    longitude = location.longitude
    return latitude,longitude
```

Data was collected and transformed into data frames



	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

Data was prepared for Borooghs and found and found the number of Neighbourhoods of each Bproguh as below

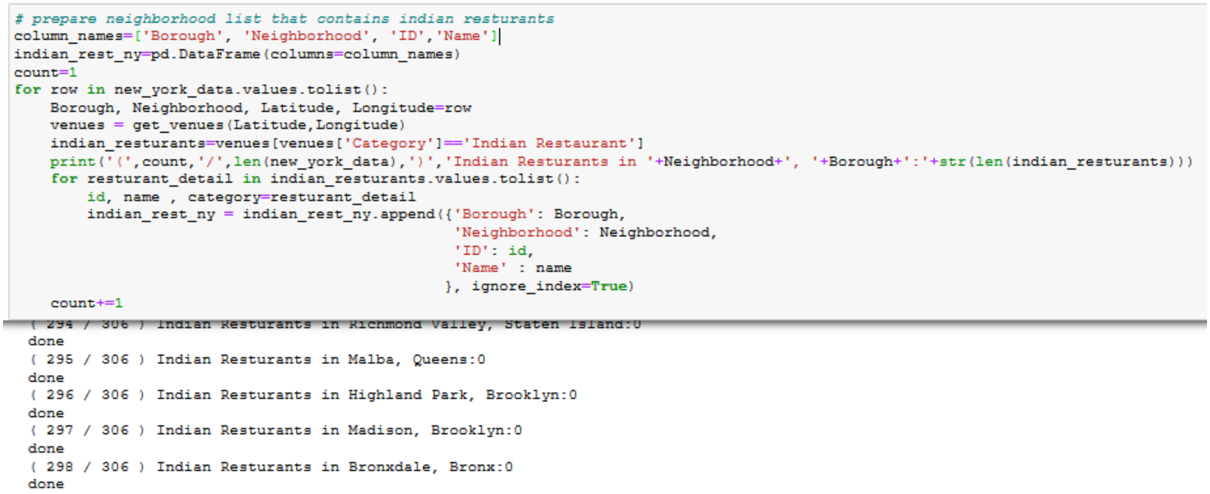


We can see that Queens Borough has the high number of Neighbourhoods

6. Modelling and Evaluation

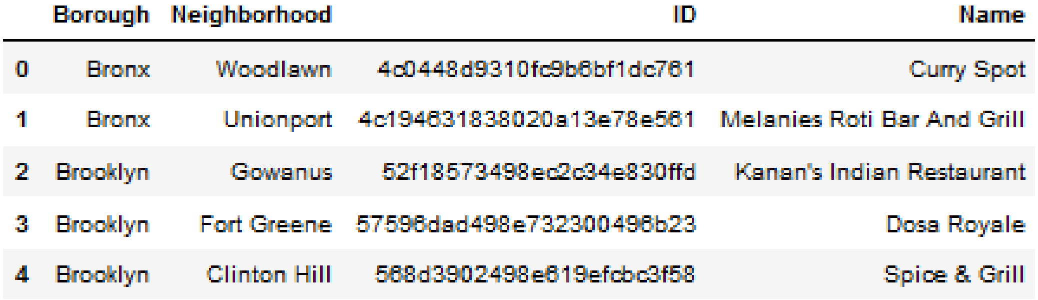
This is the final stage of data science where you finally build the model based on your data collection and preparations. In this stage, data scientists use some data science techniques to get the desired results. Depending on the data different models can be used. For the Battle of Neighbourhoods we will use clustering and segmentations techniques especially K-Means Clustering.

Collect Indian restaurants for each Neighbourhood



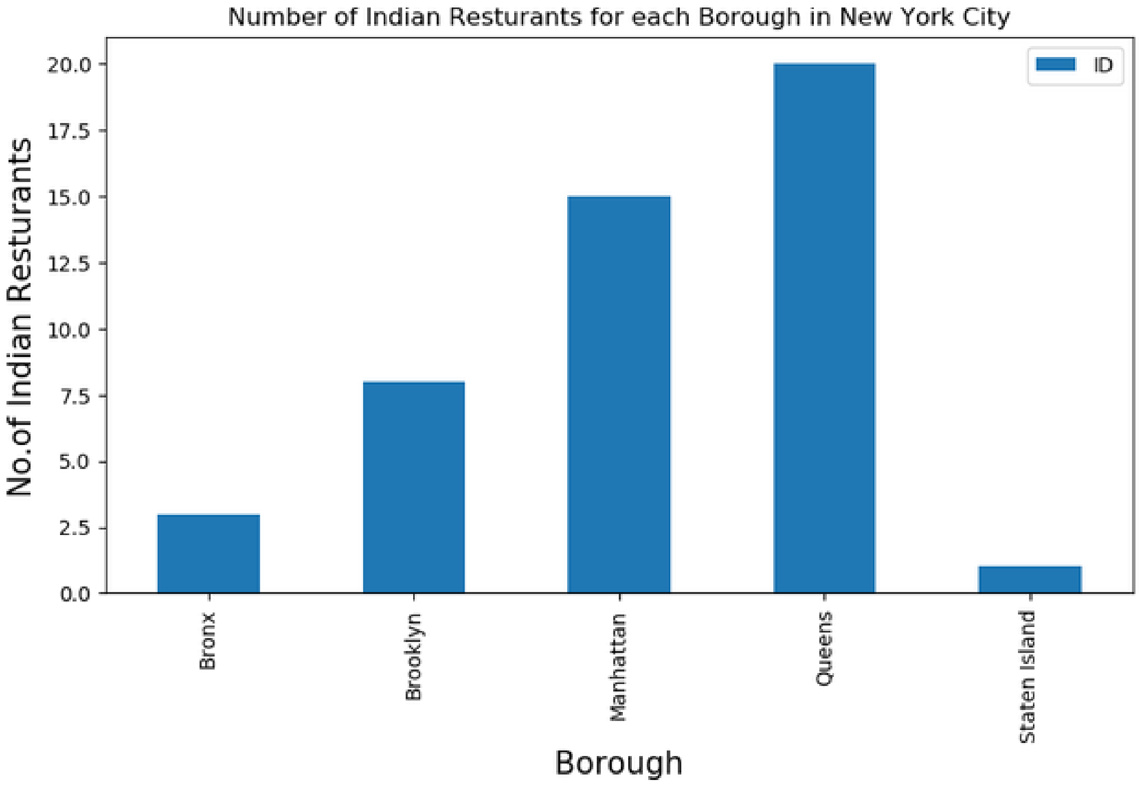
```
# prepare neighborhood list that contains indian resturants
column_names=['Borough', 'Neighborhood', 'ID','Name']
indian_rest_ny=pd.DataFrame(columns=column_names)
count=1
for row in new_york_data.values.tolist():
    Borough, Neighborhood, Latitude, Longitude=row
    venues = get_venues(Latitude,Longitude)
    indian_resturants=venues[venues['Category']=='Indian Restaurant']
    print('(',count,'/',len(new_york_data),')','Indian Resturants in '+Neighborhood+', '+Borough+':'+str(len(indian_resturants)))
    for resturant_detail in indian_resturants.values.tolist():
        id, name , category=resturant_detail
        indian_rest_ny = indian_rest_ny.append({'Borough': Borough,
                                                'Neighborhood': Neighborhood,
                                                'ID': id,
                                                'Name' : name
                                               }, ignore_index=True)
    count+=1
```

```
( 294 / 306 ) Indian Resturants in Richmond Valley, Staten Island:0
done
( 295 / 306 ) Indian Resturants in Malba, Queens:0
done
( 296 / 306 ) Indian Resturants in Highland Park, Brooklyn:0
done
( 297 / 306 ) Indian Resturants in Madison, Brooklyn:0
done
( 298 / 306 ) Indian Resturants in Bronxdale, Bronx:0
done
```

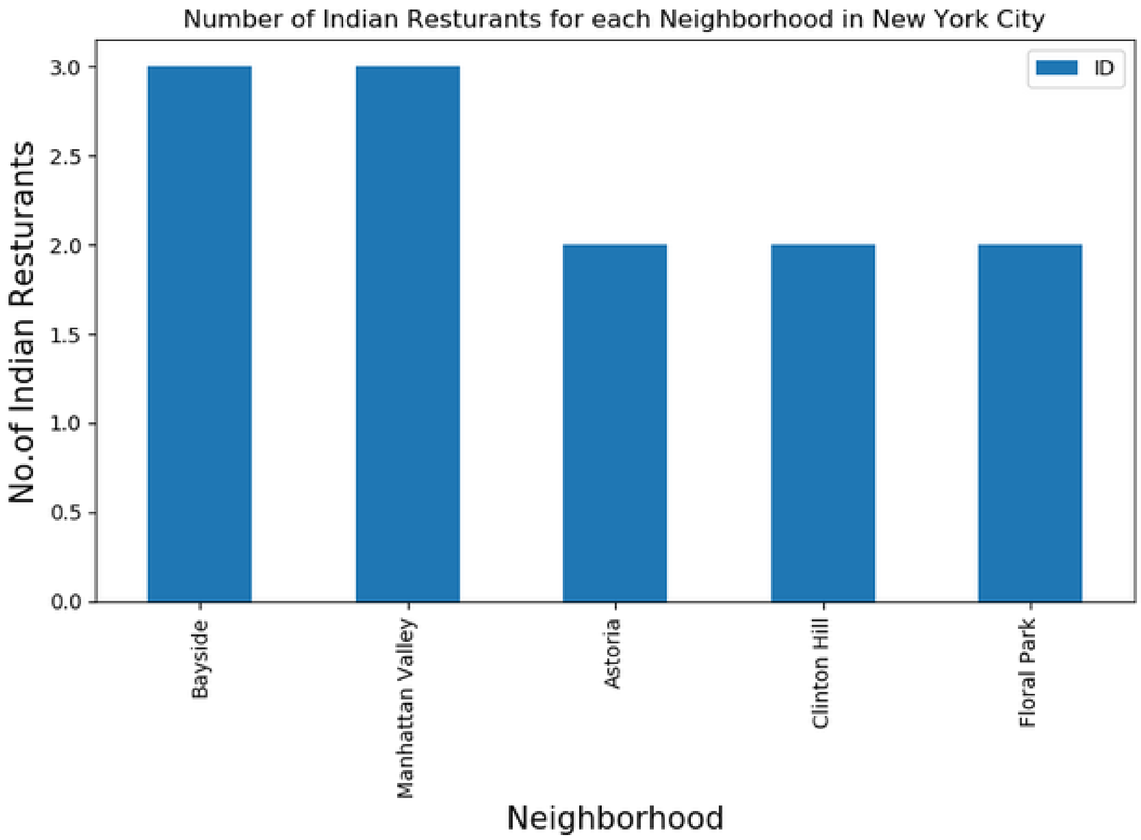


	Borough	Neighborhood	ID	Name
0	Bronx	Woodlawn	4c0448d9310fc9b6bf1dc761	Curry Spot
1	Bronx	Unionport	4c194631838020a13e78e561	Melanies Roti Bar And Grill
2	Brooklyn	Gowanus	52f18573498ec2c34e830ffd	Kanan's Indian Restaurant
3	Brooklyn	Fort Greene	57596dad498e732300496b23	Dosa Royale
4	Brooklyn	Clinton Hill	568d3902498e619efcbc3f58	Spice & Grill

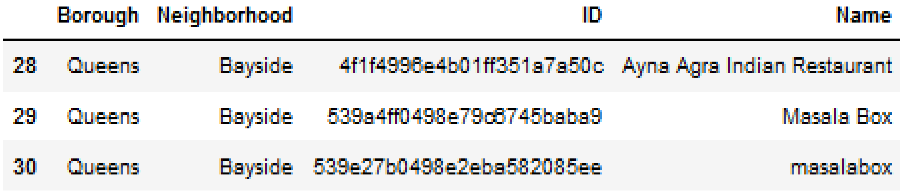
We can see that Queens in New York has highest number of restaurants i.e 20



Collect the Indian restaurants in each Neighbourhood

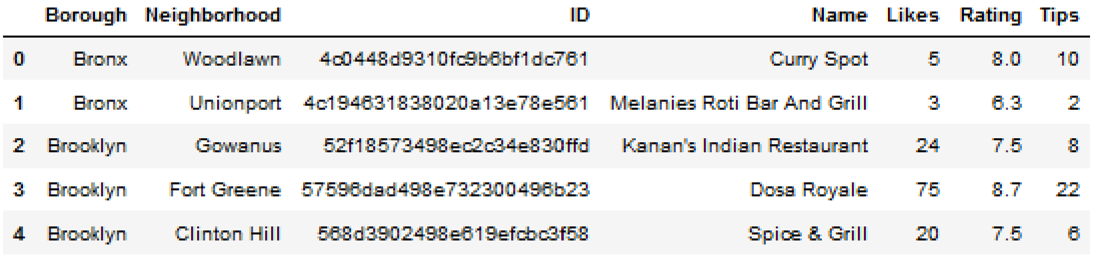


Thus Bayside in Queens has highest number of restaurants i.e 3. Below is the list of Indian Restaurants in Bayside, Queens Borough



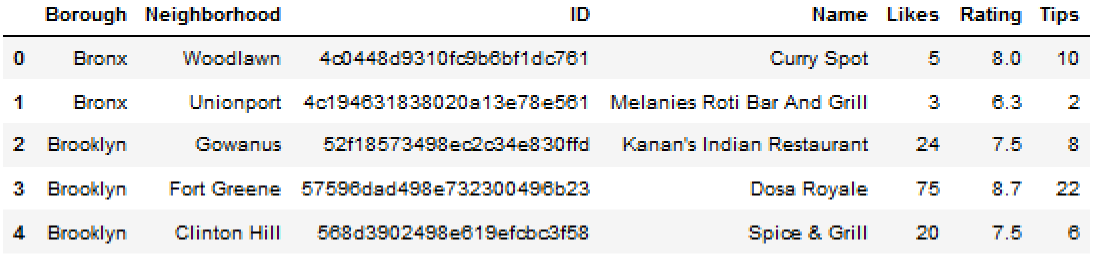
	Borough	Neighborhood	ID	Name
28	Queens	Bayside	4f1f4996e4b01ff351a7a50c	Ayna Agra Indian Restaurant
29	Queens	Bayside	539a4ff0498e79c6745baba9	Masala Box
30	Queens	Bayside	539e27b0498e2eba582085ee	masalabox

Now find the Name, Likes, Rating and Tips for each restaurant by using machine learning algorithms and Foursquare API

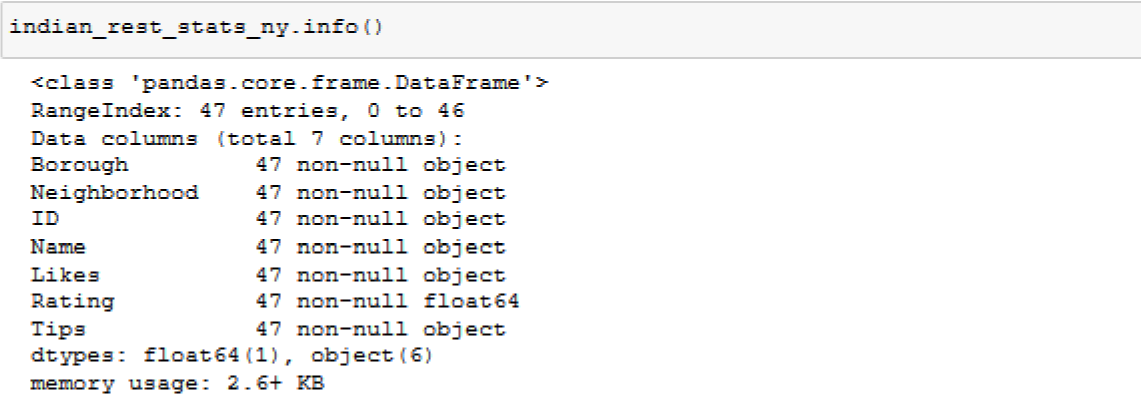


	Borough	Neighborhood	ID	Name	Likes	Rating	Tips
0	Bronx	Woodlawn	4c0448d9310fc9b6bf1dc761	Curry Spot	5	8.0	10
1	Bronx	Unionport	4c194631838020a13e78e561	Melanies Roti Bar And Grill	3	6.3	2
2	Brooklyn	Gowanus	52f18573498ec2c34e830ffd	Kanan's Indian Restaurant	24	7.5	8
3	Brooklyn	Fort Greene	57596dad498e732300496b23	Dosa Royale	75	8.7	22
4	Brooklyn	Clinton Hill	568d3902498e619efcbc3f58	Spice & Grill	20	7.5	6

Store the data in csv file and also transform the data in required formats

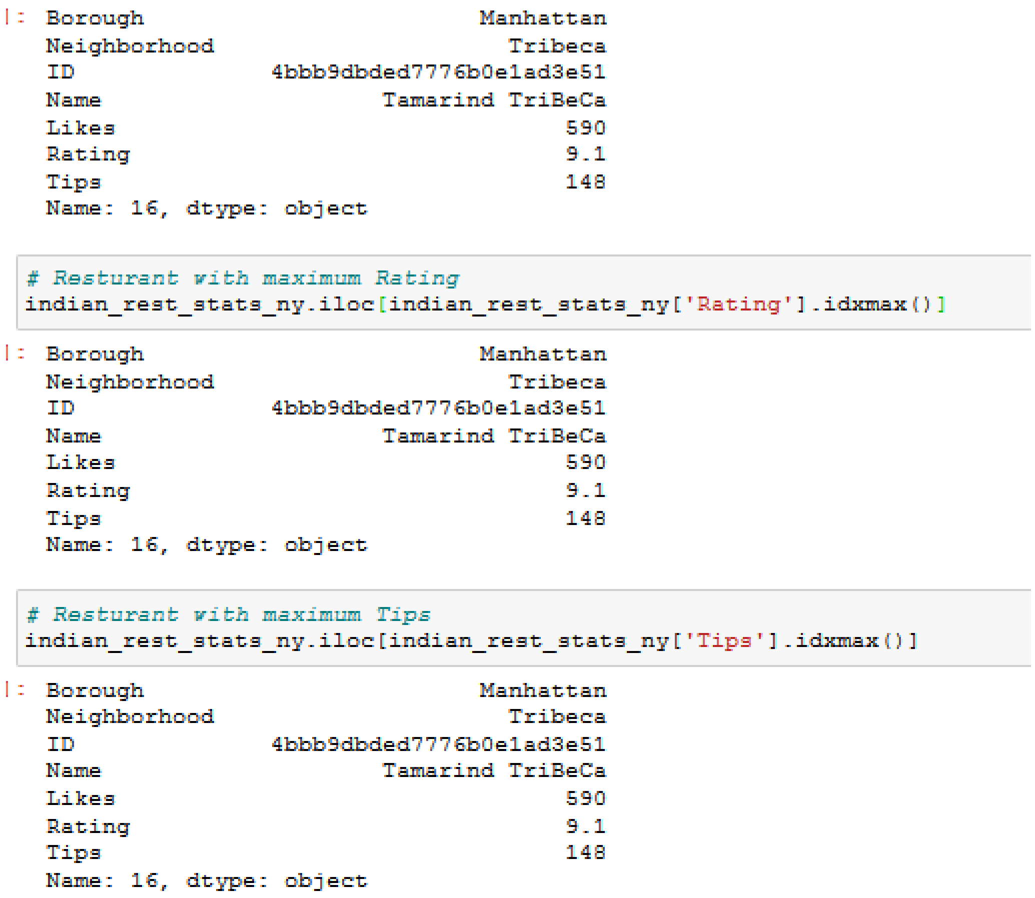


	Borough	Neighborhood	ID	Name	Likes	Rating	Tips
0	Bronx	Woodlawn	4c0448d9310fc9b6bf1dc761	Curry Spot	5	8.0	10
1	Bronx	Unionport	4c194631838020a13e78e561	Melanies Roti Bar And Grill	3	6.3	2
2	Brooklyn	Gowanus	52f18573498ec2c34e830ffd	Kanan's Indian Restaurant	24	7.5	8
3	Brooklyn	Fort Greene	57596dad498e732300496b23	Dosa Royale	75	8.7	22
4	Brooklyn	Clinton Hill	568d3902498e619efcbc3f58	Spice & Grill	20	7.5	6



```
indian_rest_stats_ny.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 47 entries, 0 to 46
Data columns (total 7 columns):
Borough         47 non-null object
Neighborhood    47 non-null object
ID              47 non-null object
Name            47 non-null object
Likes           47 non-null object
Rating          47 non-null float64
Tips            47 non-null object
dtypes: float64(1), object(6)
memory usage: 2.6+ KB
```



```
]:  Borough                        Manhattan
    Neighborhood                     Tribeca
    ID              4bbb9dbded7776b0e1ad3e51
    Name                    Tamarind TriBeCa
    Likes                                590
    Rating                               9.1
    Tips                                 148
    Name: 16, dtype: object
```

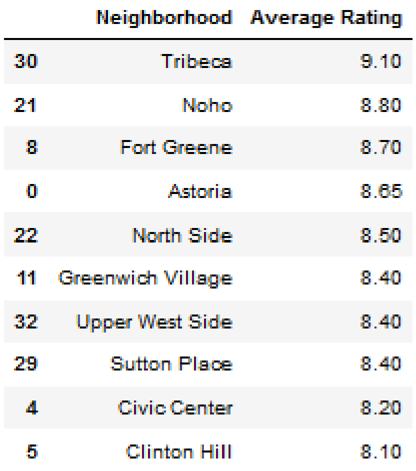
```
# Resturant with maximum Rating
indian_rest_stats_ny.iloc[indian_rest_stats_ny['Rating'].idxmax()]
```

```
]:  Borough                        Manhattan
    Neighborhood                     Tribeca
    ID              4bbb9dbded7776b0e1ad3e51
    Name                    Tamarind TriBeCa
    Likes                                590
    Rating                               9.1
    Tips                                 148
    Name: 16, dtype: object
```

```
# Resturant with maximum Tips
indian_rest_stats_ny.iloc[indian_rest_stats_ny['Tips'].idxmax()]
```

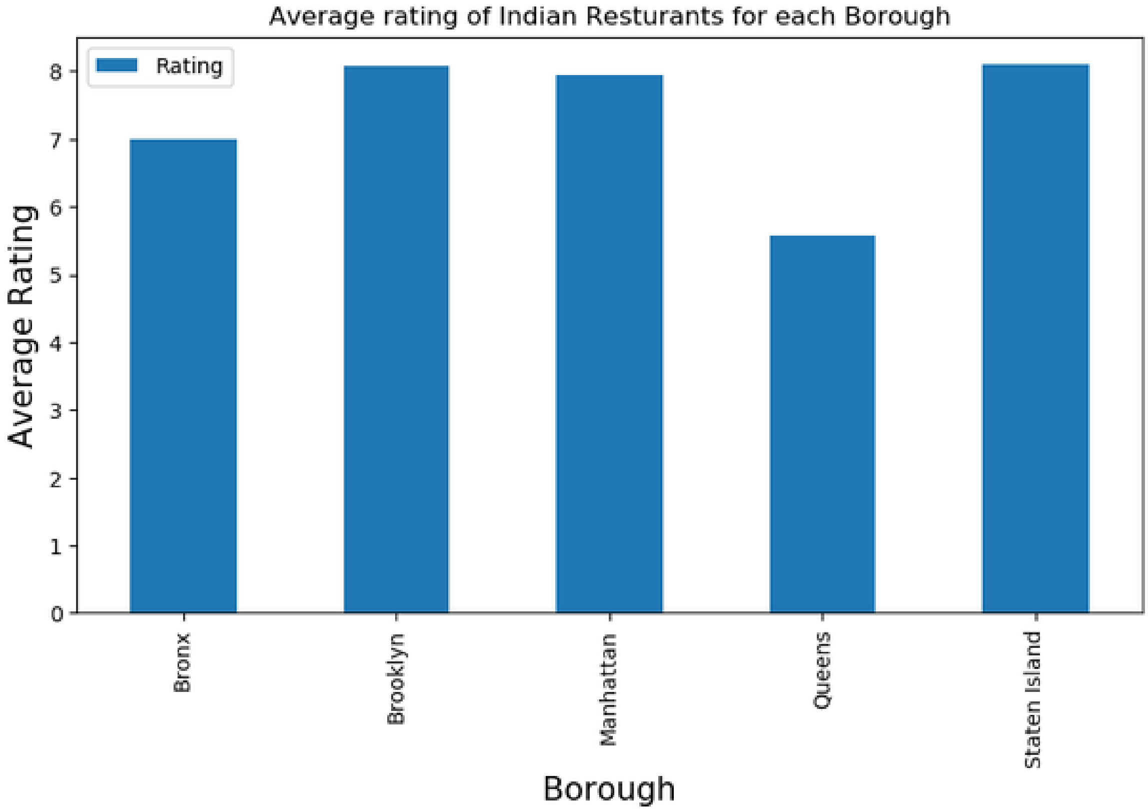
```
]:  Borough                        Manhattan
    Neighborhood                     Tribeca
    ID              4bbb9dbded7776b0e1ad3e51
    Name                    Tamarind TriBeCa
    Likes                                590
    Rating                               9.1
    Tips                                 148
    Name: 16, dtype: object
```

Now find the rating of top 10

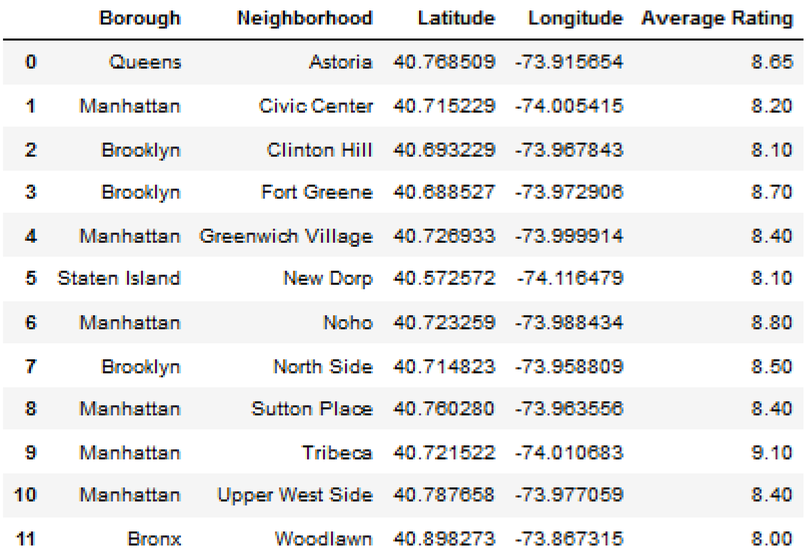


	Neighborhood	Average Rating
30	Tribeca	9.10
21	Noho	8.80
8	Fort Greene	8.70
0	Astoria	8.65
22	North Side	8.50
11	Greenwich Village	8.40
32	Upper West Side	8.40
29	Sutton Place	8.40
4	Civic Center	8.20
5	Clinton Hill	8.10

Next found the average rating of Indian Restaurant in each Borough

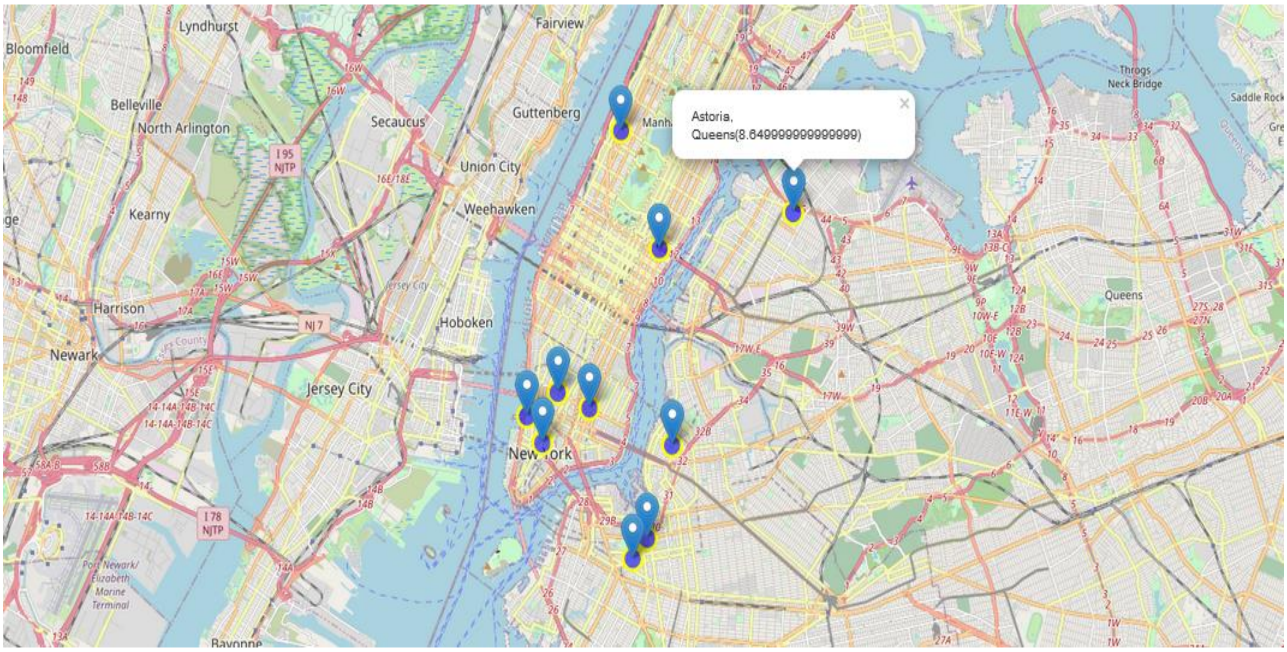


Display the average rating of each neighbourhood along with Lat and Long



	Borough	Neighborhood	Latitude	Longitude	Average Rating
0	Queens	Astoria	40.768509	-73.915654	8.65
1	Manhattan	Civic Center	40.715229	-74.005415	8.20
2	Brooklyn	Clinton Hill	40.693229	-73.967843	8.10
3	Brooklyn	Fort Greene	40.688527	-73.972906	8.70
4	Manhattan	Greenwich Village	40.726933	-73.999914	8.40
5	Staten Island	New Dorp	40.572572	-74.116479	8.10
6	Manhattan	Noho	40.723259	-73.988434	8.80
7	Brooklyn	North Side	40.714823	-73.958809	8.50
8	Manhattan	Sutton Place	40.760280	-73.963556	8.40
9	Manhattan	Tribeca	40.721522	-74.010683	9.10
10	Manhattan	Upper West Side	40.787658	-73.977059	8.40
11	Bronx	Woodlawn	40.898273	-73.867315	8.00

Show these highly rated neighbourhoods on map



7. Results¶

The results are carried out to give the answer of the following questions:

What is the best location in New York City for Indian Cuisine?

Which areas have a potential Indian Restaurant Market?

Which are some of the best neighbourhoods for Indian cuisine?

Which is the best place to stay if you prefer Indian Cuisine?

8. Discussion

The analysis has been carried out to list and visualize all major parts of New York City that has great Indian restaurants. The restaurants with varous attribues like maximum likes, rating and tips is also provided. The accuracy of data is completely depends on the data provided by FourSquare.

9. Conclusion

Below are the conclusions that can drawn from the analysis done using machine learning algorithms and Foursquare API.

Queens has the largest number of Indian Restaurants.

The Bayside in Queens has the highest number of Indian restaurants i.e 3.

Astoria (Queens), Civic Centre (Manhattan), Clinton Hill (Brooklyn) are some of the best neighbourhoods for Indian cuisine.

Queens and Manhattan are the best places to stay if you prefer Indian Cuisine.

Astoria (Queens) has the highest rating that is 8.65

Though Manhattan ranks 2nd and Brooklyn ranks 3rd in ranking ; the difference in their average rating is minimal i.e Manhattan is with 8.20 and Brooklyn is with 8.10