Coursera Capstone

IBM Applied Data Science Capstone

Recommender System to build an Indian restaurant in New York

By: Kavita Patil
August 2020

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 2018, 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.

In this project we will go through the entire neighbourhood of New York by analysing the chain of restaurants serving Indian food nearby. By doing this we will be able to decide whether it is wise to open an Indian restaurant. For analysis we will find the preferred cuisine of the population, most profitable area as to figure out the area where most of the population visits. If the area is profitable, where exactly the restaurant should be so that the business yields more profit to the owner.

Business Problem

The objective of this Capstone project is to analyse and select the best locations in the city of New York to open a new Indian restaurant. Using Data Science methodology and instruments such as Data Analysis and Visualization, this project aims to provide solutions to answer the business question: Where in the city of New York, should the investor open an Indian Restaurant?

Data

To solve the problem, we will need the following data:

- New York City data containing the neighbourhoods and boroughs.
- Latitude and longitude coordinates of those neighbourhoods. This is required to plot the map and get the venue data.
- Venue data, particularly data related to restaurants. We are going to use this data to perform further analysis of the neighbourhoods.

Data Sources:

a) https://cocl.us/new_york_dataset" csv file is used in the project to get all the geographical coordinates of the neighbourhoods.

b) To get location and other information about various venues in New York I'm using Four square's explore API. Using the Four square's explore API (which gives venues recommendations), I'm fetching details about the venues up present in Toronto and collected their names, categories and locations (latitude and longitude).
From Foursquare API (https://developer.foursquare.com/docs), following venues are retrieved:

1. **Name**: The name of the venue.
2. **Category**: The category type as defined by the API.
3. **Latitude**: The latitude value of the venue.
4. **Longitude**: The longitude value of the venue.

All data related to locations and quality of Indian restaurants will be obtained via the Foursquare API utilized via the Request library in Python. In the next Methodology section, we will discuss and describe any exploratory data analysis that we did, any inferential statistical testing that we performed, and what machine learning techniques were used.

Methodology

1. Data will be collected from https://cocl.us/new_york_dataset and cleaned and processed into a dataframe.
2. FourSquare be used to locate all venues and then filtered by Japanese restaurants. Ratings, tips, and likes by users will be counted and added to the dataframe.
3. Data will be sorted based on rankings.
4. Finally, the data be will be visually assessed using graphing from Python libraries.
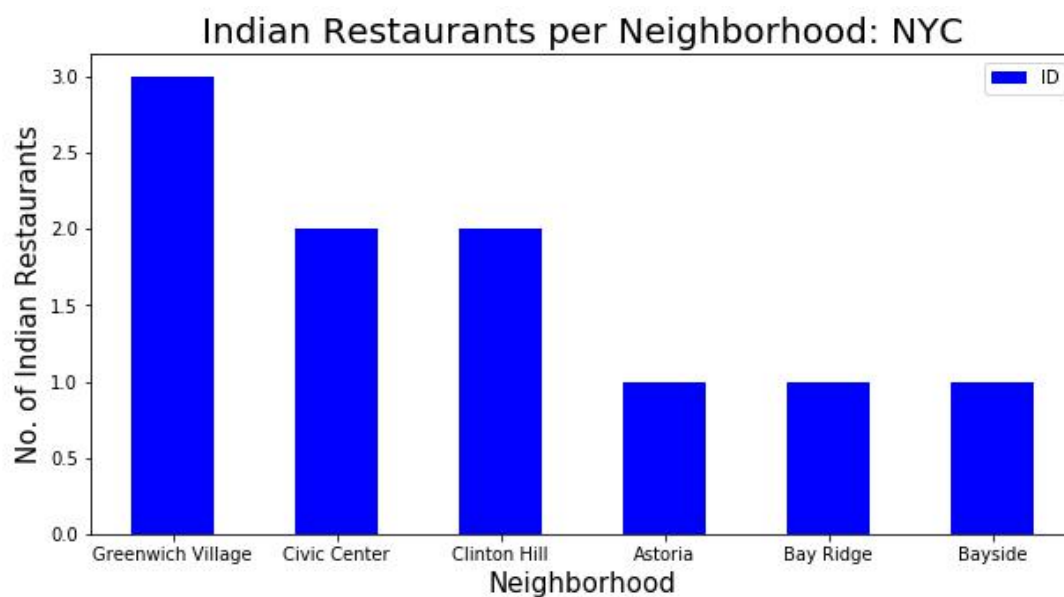
Results

Below chart shows the count of neighbourhoods per borough
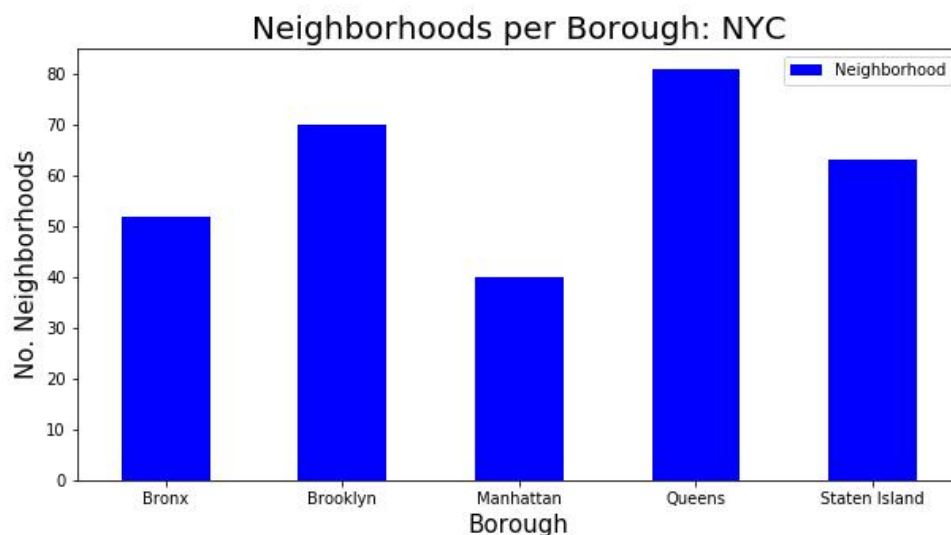
Manhattan has the least number of neighbourhoods while Queens has the highest.

Indian Restaurants per Borough: NYC
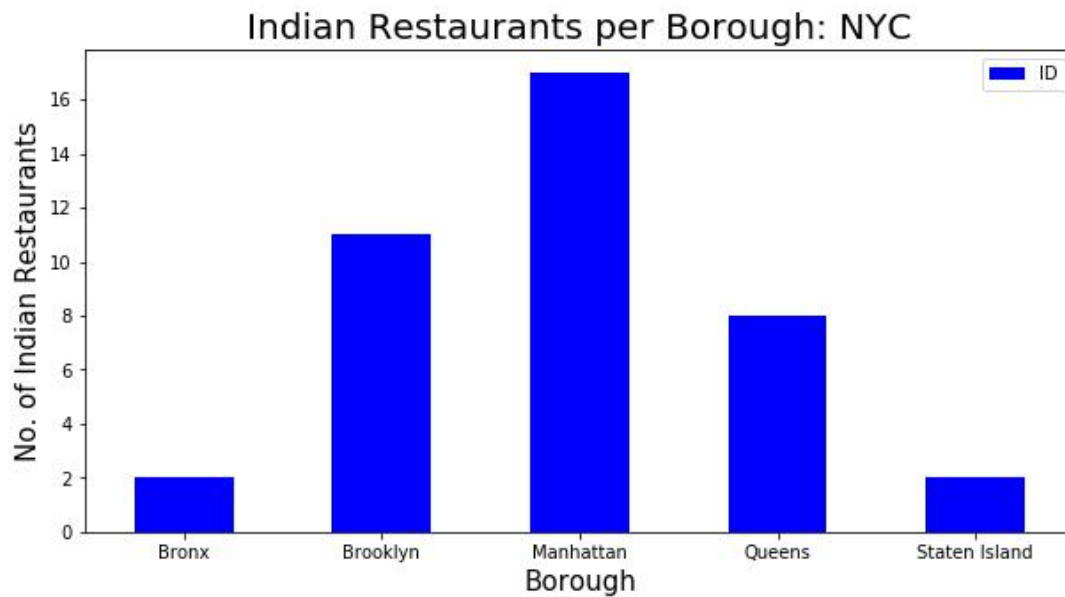
Despite of Manhattan having least number of neighbourhoods, it has highest number of Indian restaurants.

Indian Restaurants per Neighborhood: NYC

Greenwich in Manhattan has the highest number of Indian Restaurants with a total count of 3.

	Borough	Neighborhood	ID	Name
19	Manhattan	Greenwich Village	56c931b2cd10407d3ecbefda	Ananda
20	Manhattan	Greenwich Village	4593ed04f964a52050401fe3	The Kati Roll Company
21	Manhattan	Greenwich Village	4d992c1e942ba093ebb5968c	THELEwala

	Neighborhood	Average Rating
30	Tribeca	9.1
23	Noho	8.8
11	Fort Greene	8.8
26	Prospect Lefferts Gardens	8.7
8	Chelsea	8.7
14	Greenwich Village	8.6
25	North Side	8.5
34	West Village	8.5
21	Midtown	8.4
0	Astoria	8.4

Above are the top neighborhoods with the highest average ratings of Indian restaurants.

Discussion Section

Manhattan and Brooklyn have the best rated Indian restaurants on average. Staten Island and Queens have the least number of Indian restaurants per borough. However, of note, Greenwich Village in Manhattan has the highest number of Indian Restaurants in all of NY. Despite Manhattan having the least number of neighbourhoods in all five boroughs, it has the greatest number of Indian restaurants.

Based on the above information, I would state that Manhattan and Brooklyn are the best locations for Indian cuisine in NYC. To have the best shot of success, I would open an Indian restaurant in Brooklyn. Brooklyn has multiple neighbourhoods with average ratings exceeding 8.0 on a scale of 1.0 to 10.0 and has a smaller number of Indian restaurants than Manhattan, making competition easier. Also, we should keep in mind, that real estate prices in Brooklyn are much cheaper than in Manhattan. Finally, I would go to Kati Roll Company in Manhattan for the best Indian food based on 832 likes. As a final note, all of the above analysis is depended on the accuracy of Four-Square data. A more comprehensive analysis and future work would need to incorporate data from other external databases.

Limitations and suggestions for future researches

All of the above analysis is depended on the accuracy of Four-Square data. Besides, during this project, we used a free Sandbox Tier Account of Foursquare API that goes with limitations as to the number of API calls and results returned. To get better results, future research work and more comprehensive analysis could consider using a paid account to bypass these limitations as well as incorporating data from other external databases.

Conclusions

In the project we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing data analysis, and lastly providing recommendations to the investors/developers. During the project, we applied different data science methods and instruments to get the answer to our main question: "Where in the City of New York, should the investor open an Indian Restaurant?" The findings of this project will help the relevant investor better understand the advantages and disadvantages of different New York neighbourhoods/boroughs in terms of opening an Indian restaurant.