Predicting the best place to open Restaurant in New York

Ritik Goel

June, 2019

1. Introduction

1.1 Background

In this project we analyze the dataset of New York City. We explore the dataset to find the Neighborhood. Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. After that using explore function we get the most common venue categories in each neighborhood. Using this data we predicted the best place to open a Restaurant in New York.

1.2 Problem

If in New York someone wants to open a restaurant then he/she must open it in such place where profit is more. And the place where people often visit Restaurant is best place to open it. This project aims to predict the best place to open a new Restaurant more over it will also predict which kind of restaurant it should be i.e. Indian, American, Mexican, Italian, etc.

1.3 Interest

Obviously, anyone who would like to open a new restaurant in New York will like to know where to open it so as to maximize the profit.

2. Data acquisition and cleaning

2.1 Data sources

Neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.

Luckily, this dataset exists for free on the web; here and here is the link of the dataset. Further we will use Foursquare API to fetch the most common places of every neighborhood to predict the best place to open restaurant.

2.2 Data cleaning and Analyzing

Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values from earlier seasons, because of lack of record keeping.

Since it is a json data we need all relevant data is in features key, which is basically a list of neighborhoods. So we defined a new variable that includes this data. After that we transform the data into Pandas dataframe as it is initially in Python Dictionaries. After cleaning dataframe will look like:

	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

Neighborhood has a total of 5 boroughs and 306 neighborhoods.

3. Data Processing

After cleaning of data I have used Foursquare API to explore the neighborhood in our Dataframe. I have taken top 100 venues that are in every neighborhood within radius of 500 meters, with their latitudes and longitudes. This looks like this:

	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'	40.877136	-73.906666	Donut Shop

Then I analyze each neighborhood using one hot encoding and group rows by neighborhood and the mean of frequency of occurrence of each category. Then I created a new dataframe that contains top 10 venues for each neighborhood. This dataframe looks like:

	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	Park	Coffee Shop	Hotel	Gym	Memorial Site	Wine Shop	Italian Restaurant	Clothing Store	Department Store	Men's Store
1	Carnegie Hill	Coffee Shop	Pizza Place	Café	Yoga Studio	Spa	Bar	Japanese Restaurant	Bookstore	Grocery Store	Gym
2	Central Harlem	African Restaurant	Gym / Fitness Center	French Restaurant	Seafood Restaurant	Public Art	Art Gallery	Chinese Restaurant	Cosmetics Shop	American Restaurant	Bookstore
3	Chelsea	Coffee Shop	Ice Cream Shop	Italian Restaurant	Bakery	Nightclub	Theater	Hotel	American Restaurant	Seafood Restaurant	Art Gallery
4	Chinatown	Chinese Restaurant	Cocktail Bar	American Restaurant	Dumpling Restaurant	Vietnamese Restaurant	Bubble Tea Shop	Ice Cream Shop	Salon / Barbershop	Optical Shop	Spa
5	Civic Center	Italian Restaurant	Gym / Fitness Center	French Restaurant	Coffee Shop	Sandwich Place	Yoga Studio	Sporting Goods Shop	Cocktail Bar	Bakery	Park
6	Clinton	Theater	Gym / Fitness Center	Italian Restaurant	Hotel	American Restaurant	Sandwich Place	Wine Shop	Spa	Coffee Shop	Thai Restaurant
7	East Harlem	Mexican Restaurant	Deli / Bodega	Bakery	Latin American Restaurant	Thai Restaurant	Convenience Store	Café	Gas Station	Taco Place	Steakhouse
8	East Village	Bar	Wine Bar	Ice Cream Shop	Mexican Restaurant	Chinese Restaurant	Cocktail Bar	Coffee Shop	Ramen Restaurant	Vegetarian / Vegan Restaurant	Pizza Place
9	Financial District	Coffee Shop	Wine Shop	Gym	Steakhouse	Cocktail Bar	Bar	Pizza Place	Italian Restaurant	Café	Juice Bar
10	Flatiron	Gym	Yoga Studio	Gym / Fitness Center	Japanese Restaurant	American Restaurant	Cycle Studio	Café	Salon / Barbershop	Clothing Store	Cosmetics Shop

4. Predictive Modeling

There are two types of models, regression and classification, that can be used to predict player improvement. Regression models can provide additional information on the amount of improvement, while classification models focus on the probabilities a player might improve. The underlying algorithms are similar between regression and classification models, but different audience might prefer one over the other.

Some graphs were made using folium library.

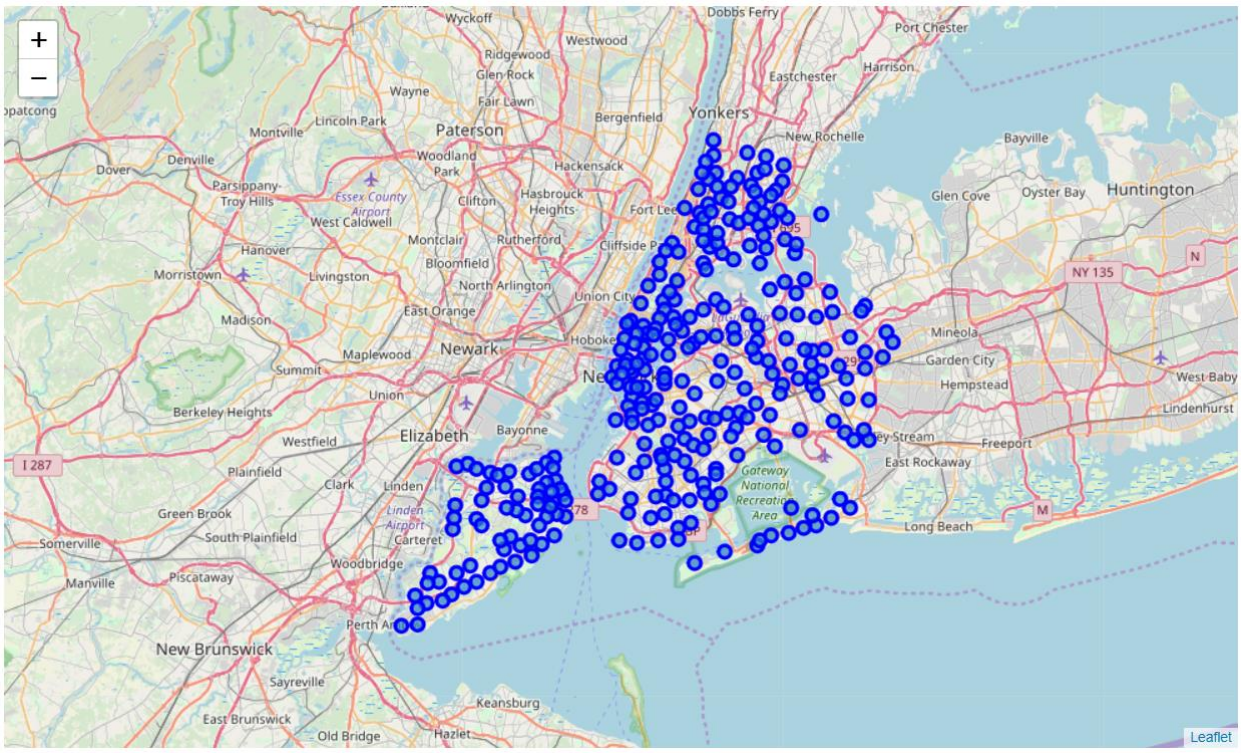
1. A map of New York with neighborhoods superimposed on top.

2. A map to visualize Manhattan and the neighborhoods in it.

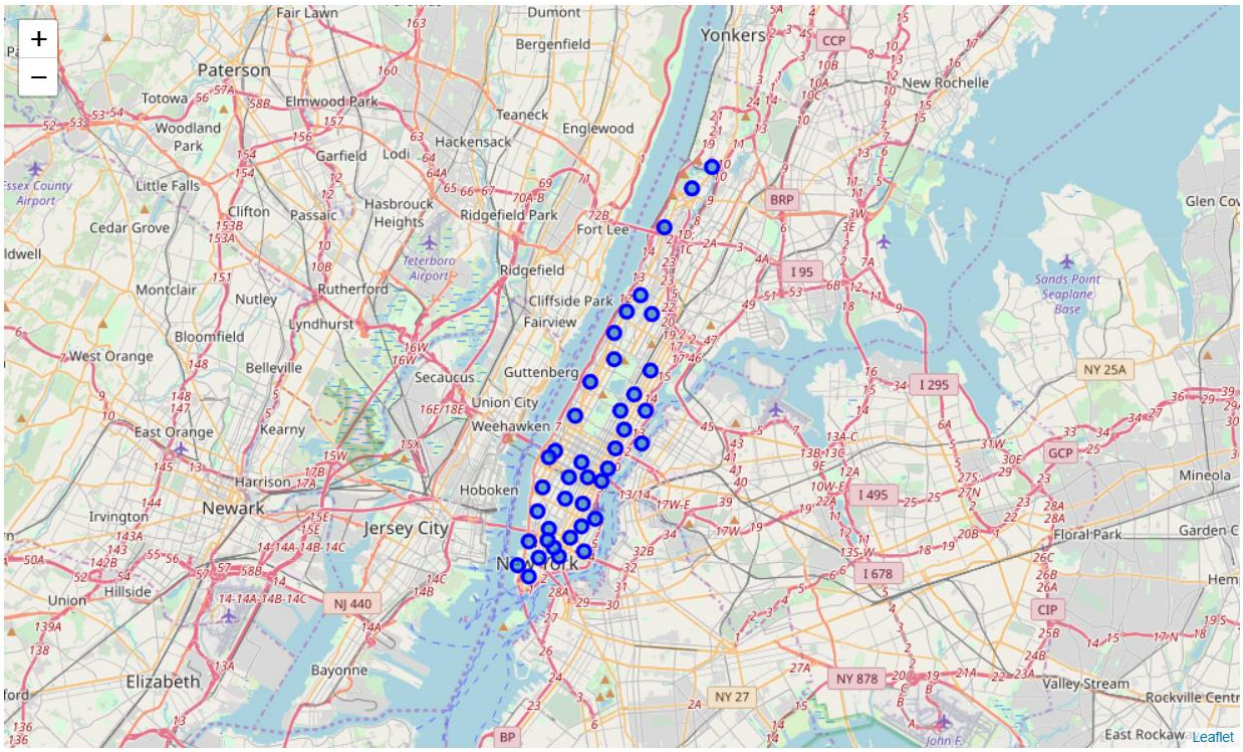
3. A map to visualize the resulting clusters.

4. A map to visualize the resulting restaurants on the preferred places.

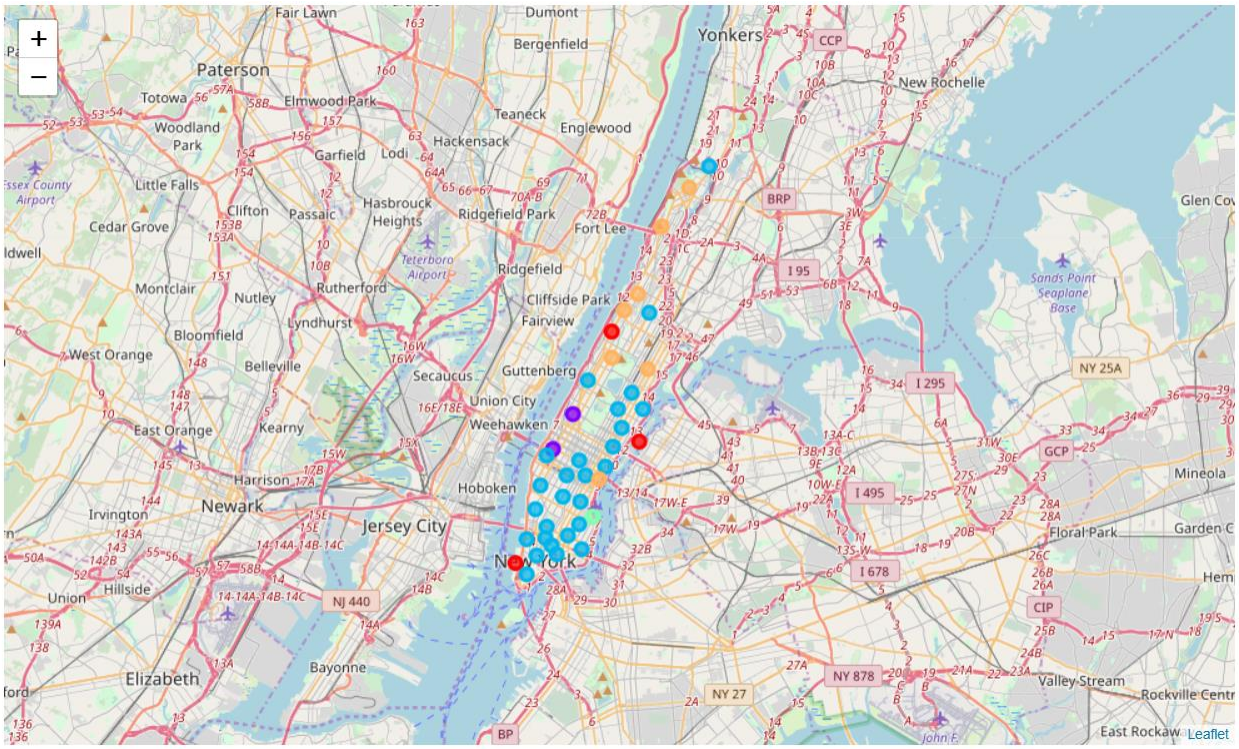
1.


Wyckoff
Dobbs Ferry
Port Chester
Ridgewood
Westwood
Glen Rock
Eastchester
Harrison
Wayne
Fair Lawn
Lincoln Park
Montville
Paterson
Bergenfield
Yonkers
New Rochelle
Bayville
Denville
Dover
Woodland Park
Hackensack
Parsippany-Troy Hills
Essex County Airport
Clifton
Hasbrouck Heights
Fort Lee
Glen Cove
Oyster Bay
Huntington
West Caldwell
Montclair
Rutherford
Cliffside P
Hanover
Bloomfield
Morristown
Livingston
North Arlington
Union City
NY 135
East Orange
Madison
Hoboke
Mineola
Newark
Maplewood
Garden City
Summit
Hempstead
West Baby
Union
Berkeley Heights
Lindenhurst
Elizabeth
Bayonne
Westfield
Freeport
I 287
Plainfield
East Rockaway
Clark
Linden
Gateway National Recreation Area
Green Brook
Linden Airport
Long Beach
Somerville
South Plainfield
Carteret
Woodbridge
Manville
Piscataway
Perth Am
New Brunswick
Sayreville
Keansburg
East Brunswick
Old Bridge
Hazlet
Leaflet

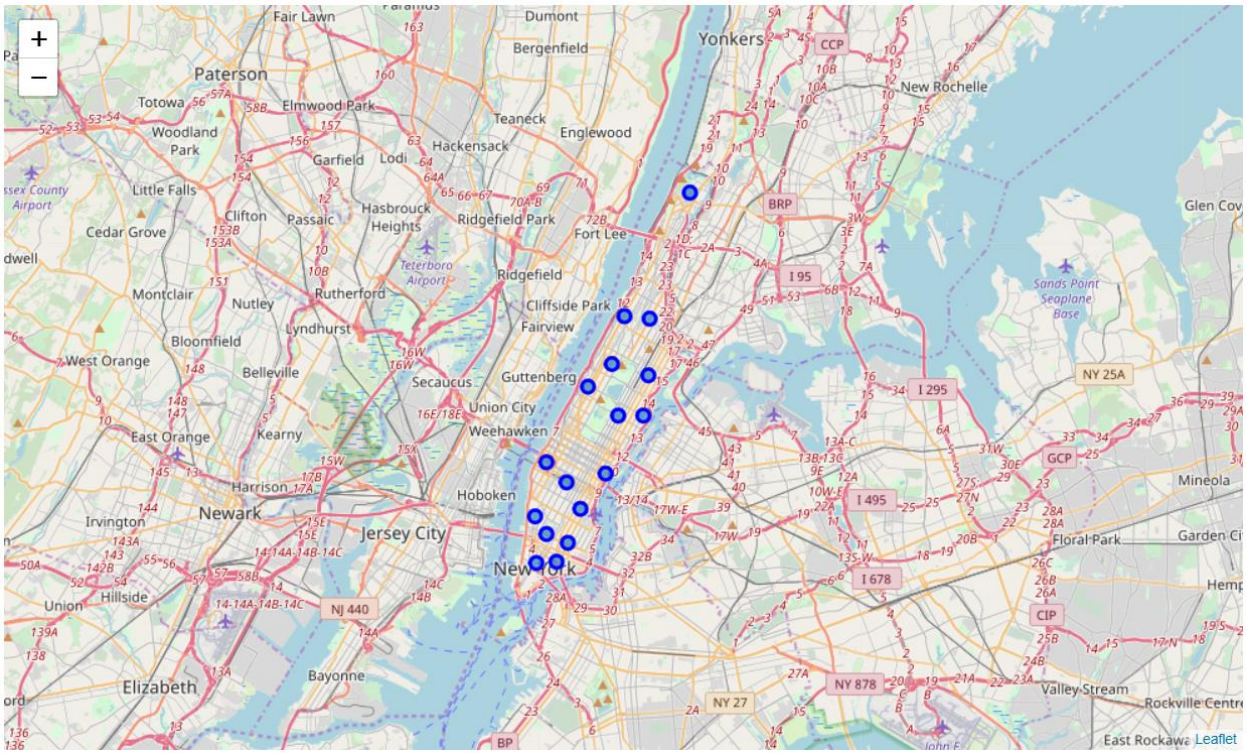
2.


Fair Lawn
Dumont
Bergenfield
Yonkers
CCP
Paterson
Totowa
Elmwood Park
New Rochelle
Woodland Park
Teaneck
Englewood
Hackensack
Garfield
Lodi
Essex County Airport
Little Falls
Clifton
Passaic
Hasbrouck Heights
Ridgefield Park
BRP
Cedar Grove
Fort Lee
Teterboro Airport
Ridgefield
I 95
Montclair
Nutley
Rutherford
Cliffside Park
Sands Point Seaplane Base
Lyndhurst
Fairview
Bloomfield
West Orange
Belleville
Secaucus
Guttenberg
NY 25A
I 295
Union City
Weehawken
East Orange
Kearny
GCP
Harrison
Hoboken
Mineola
I 495
Irvington
Newark
Jersey City
Floral Park
I 678
Union
Hillside
NJ 440
CIP
Elizabeth
Bayonne
NY 878
NY 27
Valley Stream
Rockville Centr
BP
East Rockaw
Leaflet

3.


Paterson
Totowa
Elmwood Park
Woodland Park
Garfield
Lodi
Hackensack
Teaneck
Englewood
Little Falls
Clifton
Passaic
Hasbrouck Heights
Ridgefield Park
Fort Lee
Cedar Grove
Teterboro Airport
Ridgefield
Montclair
Nutley
Rutherford
Cliffside Park
Fairview
Bloomfield
Lyndhurst
West Orange
Belleville
Secaucus
Guttenberg
Union City
Weehawken
East Orange
Kearny
Harrison
Hoboken
Newark
Irvington
Jersey City
New York
Union
Hillside
Elizabeth
Bayonne
Dumont
Bergenfield
Yonkers
New Rochelle
Glen Cove
Sands Point Seaplane Base
Mineola
Floral Park
Garden City
Valley Stream
Rockville Centre
East Rockaway
John F
I 95
I 295
I 495
I 678
NJ 440
NY 25A
NY 27
NY 878
BRP
CCP
GCP
CIP
BP
Leaflet

4.


Paterson
Totowa
Elmwood Park
Woodland Park
Garfield
Lodi
Hackensack
Teaneck
Englewood
Little Falls
Clifton
Passaic
Hasbrouck Heights
Ridgefield Park
Fort Lee
Cedar Grove
Teterboro Airport
Ridgefield
Montclair
Nutley
Rutherford
Cliffside Park
Fairview
Bloomfield
Lyndhurst
West Orange
Belleville
Secaucus
Guttenberg
Union City
Weehawken
East Orange
Kearny
Harrison
Hoboken
Newark
Irvington
Jersey City
New York
Union
Hillside
Elizabeth
Bayonne
Dumont
Bergenfield
Yonkers
New Rochelle
Glen Cove
Sands Point Seaplane Base
Mineola
Floral Park
Garden City
Valley Stream
Rockville Centre
East Rockaway
John F
I 95
I 295
I 495
I 678
NJ 440
NY 25A
NY 27
NY 878
BRP
CCP
GCP
CIP
BP
Leaflet

5. Result and Conclusion

In this project we predicted the best place to open a new Restaurant, along with their locations more over it will also predicted which kind of restaurant it should be i.e. Indian, American, Mexican, Italian, etc.

In this study, first we have cleaned the data then using Foursquare API we explore the neighborhood in our Dataframe. We have taken top 100 venues that are in every neighborhood within radius of 500 meters, with their latitudes and longitudes.

The result will look like this:

	Prefered Place	Prefered Restaurant	Latitude	Longitude
1	Chinatown	Chinese Restaurant	40.715618	-73.994279
3	Inwood	Mexican Restaurant	40.867684	-73.921210
5	Manhattanville	Italian Restaurant	40.816934	-73.957385
6	Central Harlem	African Restaurant	40.815976	-73.943211
7	East Harlem	Mexican Restaurant	40.792249	-73.944182
8	Upper East Side	Italian Restaurant	40.775639	-73.960508
9	Yorkville	Italian Restaurant	40.775930	-73.947118
12	Upper West Side	Italian Restaurant	40.787658	-73.977059
18	Greenwich Village	Italian Restaurant	40.726933	-73.999914
24	West Village	Italian Restaurant	40.734434	-74.006180
25	Manhattan Valley	Indian Restaurant	40.797307	-73.964286
27	Gramercy	Italian Restaurant	40.737210	-73.981376
31	Noho	Italian Restaurant	40.723259	-73.988434
32	Civic Center	Italian Restaurant	40.715229	-74.005415
33	Midtown South	Korean Restaurant	40.748510	-73.988713
35	Turtle Bay	Italian Restaurant	40.752042	-73.967708
39	Hudson Yards	American Restaurant	40.756658	-74.000111