Battle Of The Neighborhoods
An analysis of Manhattan Restaurant Distribution
Mohamed Naseef, Data Science Aspirant

Introduction

Background

New York City is the most populous city in the United States and the center of the New York metropolitan area. Situated on one of the world's largest natural harbors, New York City is composed of five boroughs - Brooklyn, Queens, Manhattan, the Bronx, and Staten Island - each of which is a county of the State of New York.

New York City's food culture includes an array of international cuisines influenced by the city's immigrant history. Central and Eastern European immigrants, especially Jewish immigrants from those regions, brought bagels, cheesecake, hot dogs, knishes, and delicatessens (or delis) to the city. Italian immigrants brought New York-style pizza and Italian cuisine into the city, while Jewish immigrants and Irish immigrants brought pastrami and corned beef, respectively. Chinese and other Asian restaurants, sandwich joints, trattorias, diners, and coffeehouses are ubiquitous throughout the city. Some 4,000 mobile food vendors licensed by the city, many immigrant-owned, have made Middle Eastern foods such as falafel and kebabs examples of modern New York street food.

As of 2019, there were 27,043 restaurants in the city, up from 24,865 in 2017.

Problem

Due to the ever-increasing competition, it is inferred that to open up a new restaurant in New York is quite a challenge. With highly rated restaurants as neighbors, the business can hardly bud, let alone bloom. However, neglecting the cuisine, if we are able to open up a restaurant in a neighborhood with lesser no of high-rated restaurants, the chance to prosper would be fairly higher. This is exactly what we are trying to achieve using data - recommending the best neighborhood to open up a new restaurant. For simplicity, I am restricting my analysis to Manhattan.

Interest

Any current or future restaurant owners in Manhattan would find it interesting to have a segmentation of the neighborhoods based on the restaurant distribution and more importantly, a list of neighborhoods that are optimal to open up a new restaurant branch.

Data

Data Sources

1. New York City Neighborhood Dataset

The New York City neighborhood has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, I 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.

This dataset exists for free on the web, provided by NYU Spatial Data Repository.

https://geo.nyu.edu/catalog/nyu_2451_34572

For convenience, I will be downloading a copy of the same from the IBM server (https://cocl.us/new_york_dataset) using wget command.

2. Foursquare API

The Foursquare API provides different functions such as search for a specific type of venues, to explore a particular venue, to explore a Foursquare user, to explore a geographical location, and to get trending venues around a location. In this case, I will utilize the API to search for restaurants in a neighborhood and get the ratings corresponding to these restaurants.

This will help us in segmenting the neighborhoods and thus identifying the best options at hand to open up a new restaurant.

Data Cleansing

The New York City Neighborhood Dataset was a geojson file. This was converted to a dataframe taking only the relevant parameters - Borough, Neighborhood, Latitude and Longitude of the Neighborhood. The relevant features available in the 'features' key of the json file were converted to a dataframe using a user-defined function.

The resulting dataframe was checked for anomalies. All 306 neighborhoods were confirmed to be in the dataframe.

For illustration purposes and due to API constraints, I filtered the dataframe to get the neighborhoods belonging to Manhattan.

To get the nearby restaurants' (within a radius of 500m) analysis of each neighborhood, I define a function that uses Foursquare API. The function searches nearby restaurants and gets the rating for each restaurant. To simplify, the average ratings of the top 5 and bottom 5 restaurants are found within the function. Also the number of restaurants near the neighborhood is recorded. These parameters are added to the Manhattan Neighborhood dataframe.

Methodology

Prior to getting the nearby restaurant information, the dataframe is verified to contain the list of Manhattan neighborhoods along with their coordinates. This is visualized in a map using Folium library *(See Fig. 1).*

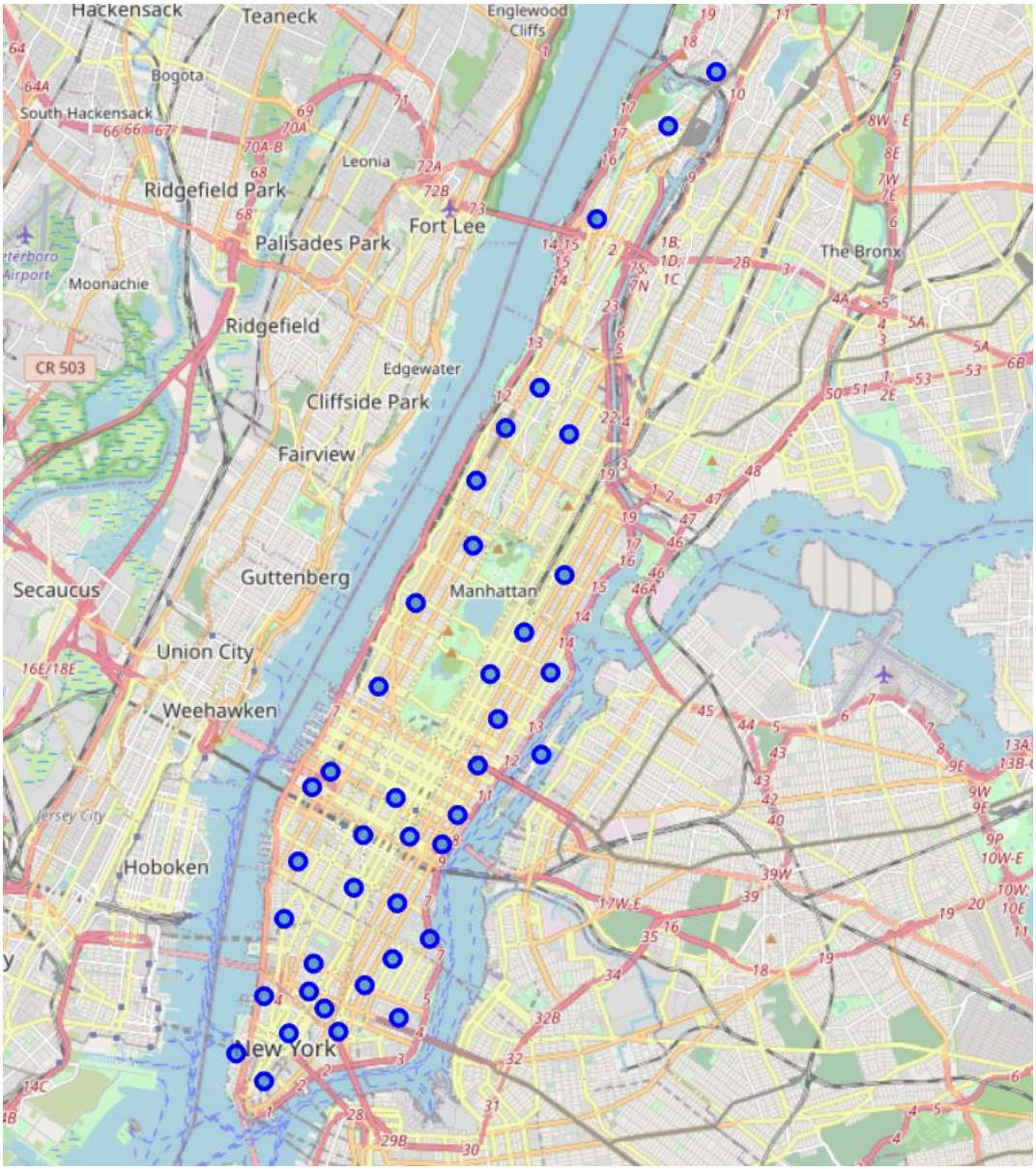

Fig 1: Manhattan Neighborhoods

The restaurant analysis function was designed to simplify my work. The function itself calculated the averages of the top 5 restaurants and the least rated 5 restaurants within a 500m radius of each neighborhood. This avoided any later steps in finding the same. The resultant dataframe contained all the relevant information we needed *(See Table 1).*

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	No of Restaurants	Average Top Rating	Average Low Rating
Marble Hill	40.88	-73.91	14	7.82	7.38
Chinatown	40.72	-73.99	30	7.84	6.24
Washington Heights	40.85	-73.94	30	8.04	6.92
Inwood	40.87	-73.92	26	7.41	6.75
Hamilton Heights	40.82	-73.95	30	7.06	5.94
Manhattanville	40.82	-73.96	21	7.47	6.67
Central Harlem	40.82	-73.94	24	7.96	6.5
East Harlem	40.79	-73.94	30	7.47	6.5
Upper East Side	40.78	-73.96	30	8.22	5.86
Yorkville	40.78	-73.95	30	8.36	5.96
Lenox Hill	40.77	-73.96	30	8.24	6.2
Roosevelt Island	40.76	-73.95	0	0	0
Upper West Side	40.79	-73.98	30	7.84	6.08
Lincoln Square	40.77	-73.99	25	8.38	6.12
Clinton	40.76	-74	30	7.42	5.5
Midtown	40.75	-73.98	30	7.32	5.62
Murray Hill	40.75	-73.98	30	6.67	6.48
Chelsea	40.74	-74	30	7.92	6.48
Greenwich Village	40.73	-74	30	8.52	6.22
East Village	40.73	-73.98	30	7.72	6.56
Lower East Side	40.72	-73.98	25	7.95	6.32
Tribeca	40.72	-74.01	30	8.28	5.96
Little Italy	40.72	-74	30	7.94	6.02
Soho	40.72	-74	30	8.16	6.02
West Village	40.73	-74.01	30	8.72	6.42
Manhattan Valley	40.8	-73.96	30	7.55	6.31
Morningside Heights	40.81	-73.96	12	7.23	6.3
Gramercy	40.74	-73.98	30	7.7	5.92
Battery Park City	40.71	-74.02	23	7.66	6.26
Financial District	40.71	-74.01	30	6.87	6.03
Carnegie Hill	40.78	-73.95	30	8.3	5.62
Noho	40.72	-73.99	30	7.98	6
Civic Center	40.72	-74.01	30	7.92	6.36
Midtown South	40.75	-73.99	30	7.73	6.3
Sutton Place	40.76	-73.96	30	7.73	6.98
Turtle Bay	40.75	-73.97	30	7.29	6.36
Tudor City	40.75	-73.97	30	7.58	6.82
Stuyvesant Town	40.73	-73.97	5	6.9	6.9
Flatiron	40.74	-73.99	30	8.04	7.01
Hudson Yards	40.76	-74	25	7.26	5.76

Table 1 : Manhattan Neighborhood Restaurant Analysis

K-Means Clustering

The next step is to group these neighborhoods based on their features. For this we use a clustering algorithm

There are many models for clustering available. In this project, I will use k-Means model that is considered the one of the simplest model among them. Despite its simplicity, k-means is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from unlabeled data.

I will pass the 3 features - number of restaurants, average high rating and average low rating - into the K-Means clustering algorithm. K-means will partition the neighborhoods into three groups since

I specified the algorithm to generate 3 clusters. Though I cannot be certain, based on the selected features, the 3 clusters should ideally predict high-risk, medium-risk and low-risk areas for starting up a new restaurant.

After running the algorithm, the neighborhoods are labeled Clusters 1,2 and 3. First, I visualized these clusters on a map, which gave a better idea of the distribution of the clusters. *(See Fig. 2)*

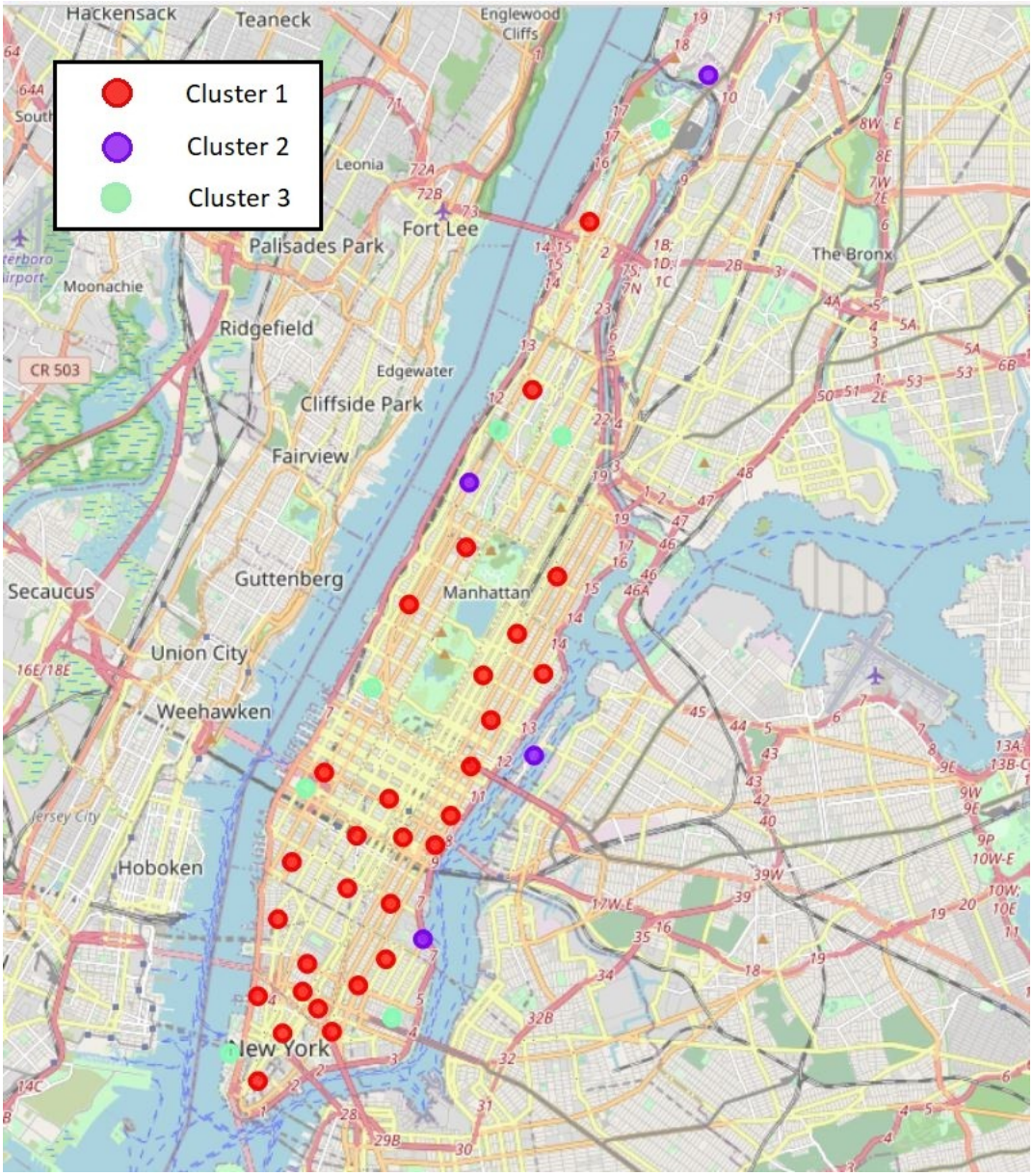

Fig 2: Manhattan Neighborhoods Clustered

Results

On examining the clusters, we can conclude that :

Cluster-1, which contains most of the neighborhoods, already have a large number of restaurants with a fairly high rating *(See Table 2)*. These neighborhood would prove highly risky for a new business.

Neighborhood	No of Restaurants	Average Top Rating	Average Low Rating	
Chinatown		30	7.84	6.24
Washington Heights		30	8.04	6.92
Hamilton Heights		30	7.06	5.94
East Harlem		30	7.47	6.5
Upper East Side		30	8.22	5.86
Yorkville		30	8.36	5.96
Lenox Hill		30	8.24	6.2
Upper West Side		30	7.84	6.08
Clinton		30	7.42	5.5
Midtown		30	7.32	5.62
Murray Hill		30	6.67	6.48
Chelsea		30	7.92	6.48
Greenwich Village		30	8.52	6.22
East Village		30	7.72	6.56
Tribeca		30	8.28	5.96
Little Italy		30	7.94	6.02
Soho		30	8.16	6.02
West Village		30	8.72	6.42
Manhattan Valley		30	7.55	6.31
Gramercy		30	7.7	5.92
Financial District		30	6.87	6.03
Carnegie Hill		30	8.3	5.62
Noho		30	7.98	6
Civic Center		30	7.92	6.36
Midtown South		30	7.73	6.3
Sutton Place		30	7.73	6.98
Turtle Bay		30	7.29	6.37
Tudor City		30	7.58	6.82
Flatiron		30	8.04	7

Table 2 : Cluster 1 – High Risk Neighborhoods

Cluster-3, which contain a few neighborhoods including Manhattenville and Hudson Yards, have an average number of restaurants and a fair rating of restaurants *(See Table 3)*.

Neighborhood	No of Restaurants	Average Top Rating	Average Low Rating	
Inwood		26	7.41	6.75
Manhattanville		21	7.47	6.67
Central Harlem		24	7.96	6.5
Lincoln Square		25	8.38	6.12
Lower East Side		25	7.95	6.32
Battery Park City		23	7.66	6.26
Hudson Yards		25	7.26	5.76

Table 3 : Cluster 3 – Medium Risk Neighborhoods

Cluster-2, which includes 4 neighborhoods -Marble Hill, Roosevelt Island, Morningside Heights, Stuyvesant Town - have very few restaurants and average rating *(See Table 4)*. These neighborhoods are ideal for starting a new restaurant.

Neighborhood	No of Restaurants	Average Top Rating	Average Low Rating	
Marble Hill		14	7.82	7.38
Roosevelt Island		0	0	0
Morningside Heights		12	7.23	6.3
Stuyvesant Town		5	6.9	6.9

Table 3 : Cluster 2 – Low Risk Neighborhoods

Discussions

The approach I used here considered the number of neighborhoods and the ratings of the restaurants - the top 5 average and the bottom 5 average. These features gave an idea of the scope for starting up a new restaurant, the competition it will face and if it can survive despite low ratings respectively. Also clustering was done into 3 groups, which in fact neglected a lot of possible neighborhoods and only brought out the least risky neighborhoods.

There is room for plenty of improvement on this approach. If we group into a higher number of clusters, we would be able to identify those neighborhoods, despite having an average number of restaurants, has very low rated restaurants. These can also be considered since the presence of low rated restaurants implies possibility for a new fair rated restaurants to prosper.

Incorporating the cuisine of the restaurants as a feature will help in identifying what restaurant cuisine has better chance of survival in any neighborhood. This will be helpful for cuisine specific restaurants to identify their best neighborhoods.

Also other major factors which can be included in the analysis is the population and demographic. For example, a neighborhood with more Indian migrants indicates the scope for a restaurant with Indian cuisine.

Conclusion

From the analysis, I was able to conclude the following:

A large number of neighborhoods, especially in lower Manhattan, already have a sufficient number of restaurants. Starting a new restaurant here would prove to be a high-risk.

There are a few neighborhoods with an average number of restaurants with some having low rated restaurants as well. These neighborhoods, mostly coastal, are of comparatively low risk.

However the best neighborhoods in Manhattan that have fewer restaurants and an average or low rating would be the most ideal neighborhoods to start a new restaurant. These were found to be:

- Marble Hill
- Roosevelt Island
- Morningside Heights
- Stuyvesant Town

These findings should be helpful to anybody planning to start their restaurant branch in Manhattan as it gives an overall idea of the restaurant distribution across the neighborhoods.

References

https://en.wikipedia.org/wiki/New_York_City
https://smallbusiness.chron.com/gain-competitive-advantage-restaurant-business-24162.html