Predicting the best location for opening a Restaurant in New York City

IBM Applied Data Science Project

1. Introduction

1.1 Background

Opening a new business in any field is a big challenge with heavy competition everywhere. With various trending customer preferences, one must explicitly adapt to the needs of these customers. This is where data science comes into picture where an informed decision on opening any business is advantageous. The need to leverage the data covering current customer preferences of restaurants and trending food categories aspects can be very crucial for opening a restaurant. One such way can be to analyse various locations that include several restaurants and provide details as to which locations can work for a particular food restaurant category (Italian, Japanese, Indian, etc.).

1.2 Problem

In this project, an attempt to find the best location for opening a restaurant business is carried out. The expected outcomes are:

1. Determine the best restaurant category (Chinese, Italian, Mexican, etc.) for a new business

2. Determine the best neighbourhood to open the restaurant

1.3 Interest

This project focuses on location analysis of opening a restaurant, hence, any stakeholders or entrepreneurs interested in doing a new business such as food restaurants which is a very demanding market. An overview of the venue feature analysis can help the stakeholders to open their restaurants in a particular location of high demand.

2. Data Acquisition and Cleaning

The city under consideration is NYC (New York City), which is divided into 5 boroughs and 306 neighbourhoods. In order to segment the neighbourhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighbourhoods that exist in each borough as well as the latitude and longitude coordinates of each neighbourhood. Luckily, this dataset exists for free on the web. Feel free to try to find this dataset on your own, but here is the link to the dataset: https://geo.nyu.edu/catalog/nyu_2451_34572

The relevant data of boroughs and neighbourhoods is extracted from the features key. The data thus obtained was then transformed to a Pandas dataset.

The Manhattan borough is selected for its most popularity with 3071 business venues. Amongst these businesses, we considered Restaurant business with 816 business venues. Next, all the existing restaurant businesses and their different categorical data is obtained from **Foursquare Location API Dataset**. The data consists of the location, restaurant names, neighbourhood names, and restaurant categories. There are also over 76 unique restaurants to analyse and setup

the most favourable restaurant category in NYC according to the neighbourhood location trends and advantages.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue_Category	Venue ID
0	Marble Hill	40.876551	-73.910660	Land & Sea Restaurant	40.877885	-73.905873	Seafood Restaurant	4b9c9c6af964a520b27236e3
1	Marble Hill	40.876551	-73.910660	Boston Market	40.877430	-73.905412	American Restaurant	585c205665e7c70a2f1055ea
2	Chinatown	40.715618	-73.994279	Kiki's	40.714476	-73.992036	Greek Restaurant	5521c2ff498ebe2368634187
3	Chinatown	40.715618	-73.994279	The Fat Radish	40.715323	-73.991950	English Restaurant	4c9d482e46978cfa8247967f
4	Chinatown	40.715618	-73.994279	Da Yu Hot Pot 大渝火锅	40.716735	-73.995752	Hotpot Restaurant	5d992946dbf3ca0008d05211

Foursquare Specific Venue Dataset is also used that contains more exploratory data containing the restaurant ratings, number of reviews, total likes, and price category. The data is further cleaned to the features of the top 5 restaurants that are most common in NYC.

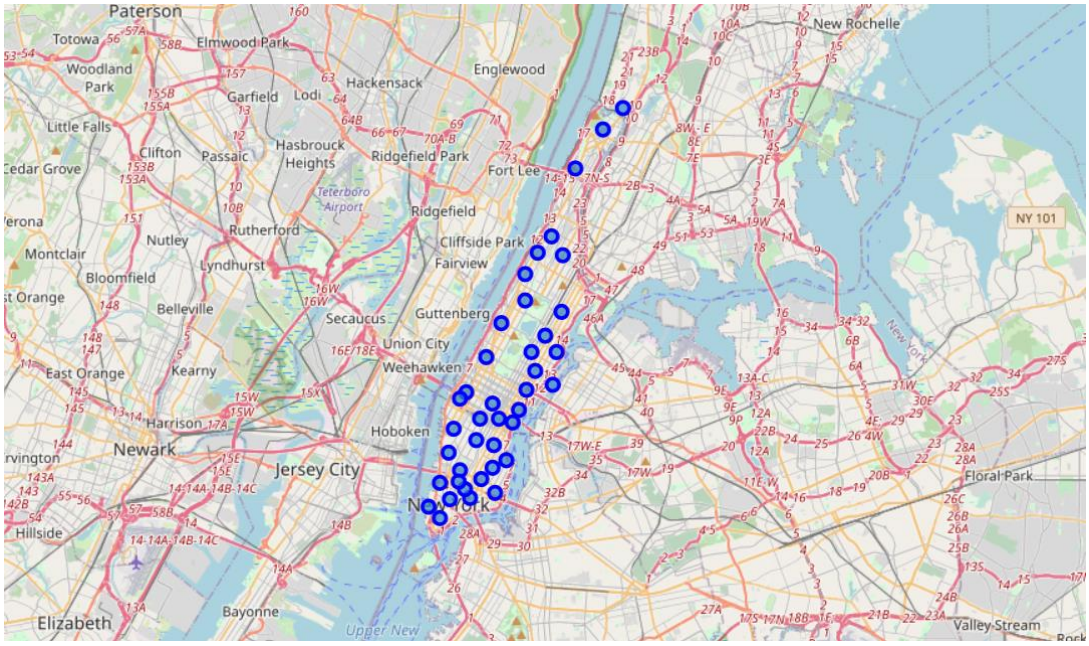
Feature Analysis: The features that were considered for the specific neighbourhood location analysis were: Average Ratings, Total Reviews, Total Likes, and the Price Category. These features are a good start for indicating which restaurant categories are most popular and unpopular amongst the customers. This data was gathered using the premium call from the Foursquare Venue API dataset for each venue.

	Neighborhood	Venue	Venue_Category	Venue ID	Price_Category	Ratings	Total_Reviews	Likes
0	Marble Hill	Land & Sea Restaurant	Seafood Restaurant	4b9c9c6af964a520b27236e3	Moderate	7.5	70	42
1	Marble Hill	Boston Market	American Restaurant	585c205665e7c70a2f1055ea	Moderate	7.1	5	4
2	Chinatown	Spicy Village	Chinese Restaurant	4db3374590a0843f295fb69b	Cheap	8.2	690	500
3	Chinatown	Wah Fung Number 1 Fast Food 華豐快飯店	Chinese Restaurant	4a96bf8ff964a520ce2620e3	Cheap	8.2	281	192
4	Chinatown	Xi'an Famous Foods	Chinese Restaurant	5894c9a15e56b417cf79e553	Cheap	8.5	138	104

3. Methodology

3.1 Visualizing Manhattan neighbourhoods

The Manhattan Neighbourhoods are visualized first



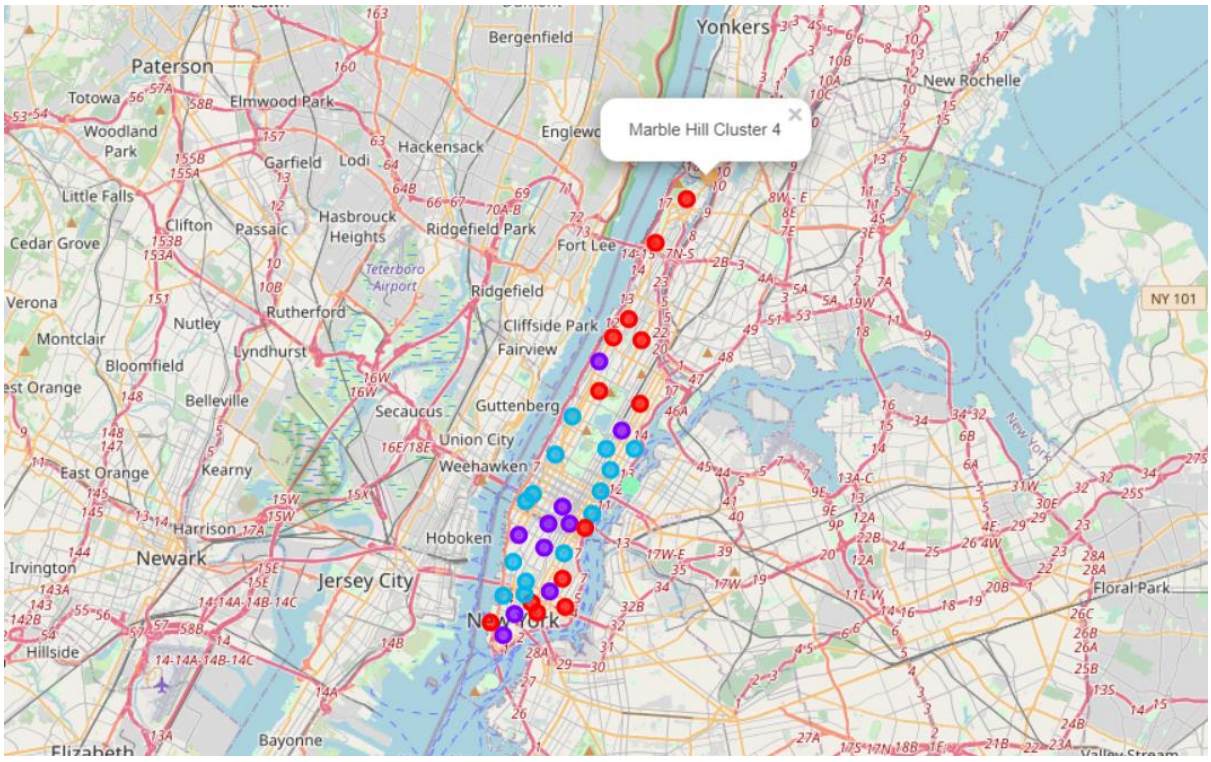
Next, the total business venues are 3071 roughly. From them, the restaurant business venues were selected that are 816 in total.

Further, the restaurant business venues are segmented according to which neighbourhood they belong being within a closed radius of around 500m in locality. The following table shown below represents the total restaurant venues belonging to a neighbourhood.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue_Category	Venue ID
Battery Park City	3	3	3	3	3	3	3
Carnegie Hill	19	19	19	19	19	19	19
Central Harlem	15	15	15	15	15	15	15
Chelsea	14	14	14	14	14	14	14
Chinatown	39	39	39	39	39	39	39
Civic Center	26	26	26	26	26	26	26
Clinton	17	17	17	17	17	17	17
East Harlem	15	15	15	15	15	15	15
East Village	36	36	36	36	36	36	36
Financial District	20	20	20	20	20	20	20

3.2 Grouping Neighbourhood using Machine Learning Algorithms

Now, to group each neighbourhood based on the type of restaurant categories, we use K-Means Clustering Machine Learning Algorithm to cluster the neighbourhoods with the most common restaurant venues between them. Hence, all the 38 neighbourhoods of NYC are clustered into 5 clusters.



The clusters are as follows:

Cluster 1:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
1	Chinatown	Chinese Restaurant	American Restaurant	Vietnamese Restaurant
2	Washington Heights	Chinese Restaurant	Spanish Restaurant	Latin American Restaurant
3	Inwood	Mexican Restaurant	Restaurant	Chinese Restaurant
4	Hamilton Heights	Mexican Restaurant	Sushi Restaurant	Indian Restaurant
5	Manhattanville	Seafood Restaurant	Italian Restaurant	Sushi Restaurant
6	Central Harlem	African Restaurant	Chinese Restaurant	American Restaurant
7	East Harlem	Mexican Restaurant	Latin American Restaurant	Thai Restaurant
19	East Village	Mexican Restaurant	Japanese Restaurant	Ramen Restaurant
20	Lower East Side	Chinese Restaurant	Vietnamese Restaurant	Mediterranean Restaurant
22	Little Italy	Chinese Restaurant	Mediterranean Restaurant	Thai Restaurant
25	Manhattan Valley	Mexican Restaurant	Vietnamese Restaurant	Caribbean Restaurant
28	Battery Park City	Chinese Restaurant	Mediterranean Restaurant	Mexican Restaurant
36	Tudor City	Mexican Restaurant	Sushi Restaurant	Greek Restaurant

Cluster 2:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
8	Upper East Side	Italian Restaurant	Sushi Restaurant	American Restaurant
9	Yorkville	Italian Restaurant	Sushi Restaurant	Japanese Restaurant
10	Lenox Hill	Italian Restaurant	Sushi Restaurant	Turkish Restaurant
12	Upper West Side	Italian Restaurant	Vegetarian / Vegan Restaurant	Indian Restaurant
13	Lincoln Square	Italian Restaurant	American Restaurant	French Restaurant
14	Clinton	Italian Restaurant	American Restaurant	Thai Restaurant
18	Greenwich Village	Italian Restaurant	Sushi Restaurant	Indian Restaurant
21	Tribeca	Italian Restaurant	American Restaurant	Greek Restaurant
23	Soho	Italian Restaurant	Mediterranean Restaurant	Sushi Restaurant
24	West Village	Italian Restaurant	American Restaurant	New American Restaurant
27	Gramercy	Italian Restaurant	American Restaurant	Thai Restaurant
34	Sutton Place	Italian Restaurant	Mexican Restaurant	Vegetarian / Vegan Restaurant
35	Turtle Bay	Italian Restaurant	Sushi Restaurant	French Restaurant
39	Hudson Yards	American Restaurant	Italian Restaurant	Restaurant

Cluster 3:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
11	Roosevelt Island	Greek Restaurant	Japanese Restaurant	German Restaurant

Cluster 4:

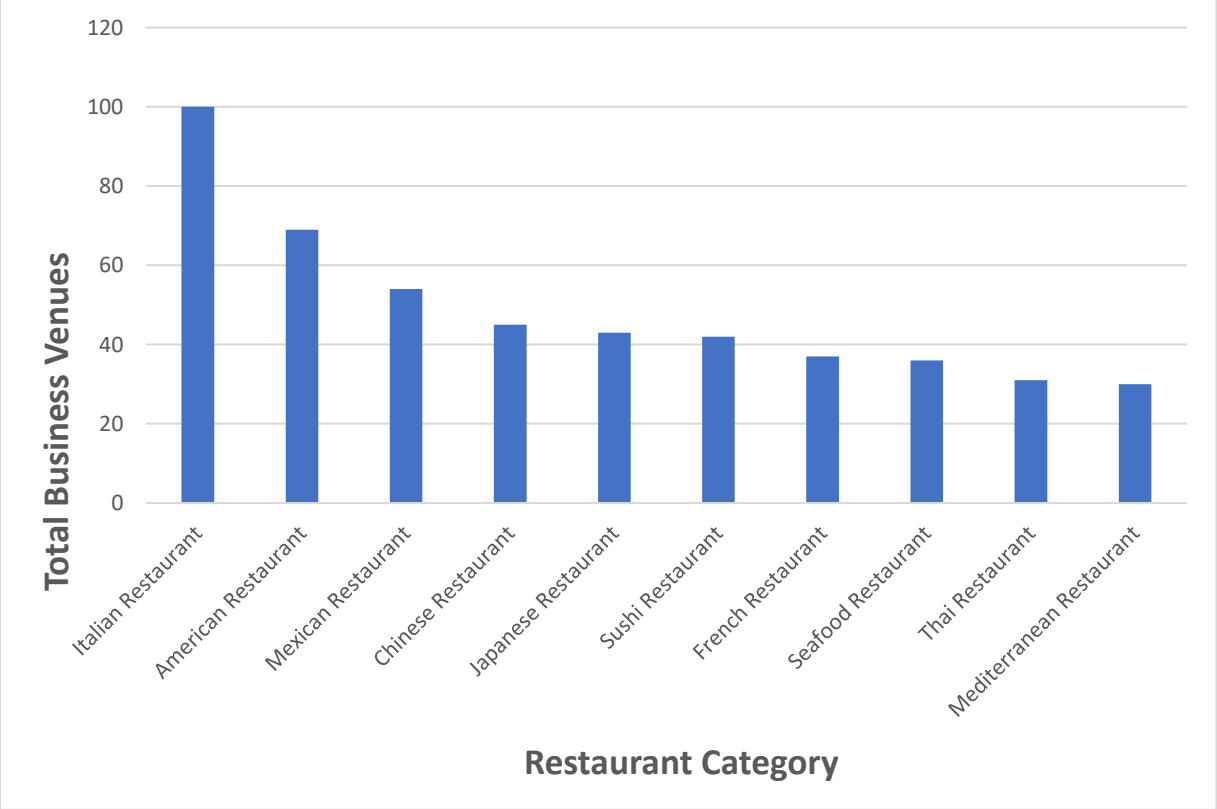
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
15	Midtown	Japanese Restaurant	Cuban Restaurant	Mediterranean Restaurant
16	Murray Hill	Japanese Restaurant	Mediterranean Restaurant	Indian Restaurant
17	Chelsea	American Restaurant	Italian Restaurant	Seafood Restaurant
26	Morningside Heights	American Restaurant	Mexican Restaurant	Ethiopian Restaurant
29	Financial District	American Restaurant	Falafel Restaurant	Japanese Restaurant
30	Carnegie Hill	Japanese Restaurant	Italian Restaurant	French Restaurant
31	Noho	Italian Restaurant	Japanese Restaurant	Mexican Restaurant
32	Civic Center	French Restaurant	American Restaurant	Sushi Restaurant
33	Midtown South	Korean Restaurant	Japanese Restaurant	American Restaurant
38	Flatiron	Italian Restaurant	Mediterranean Restaurant	Japanese Restaurant

Cluster 5:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	Marble Hill	American Restaurant	Seafood Restaurant	Vietnamese Restaurant

3.3 Analysing Venue Features Dataset

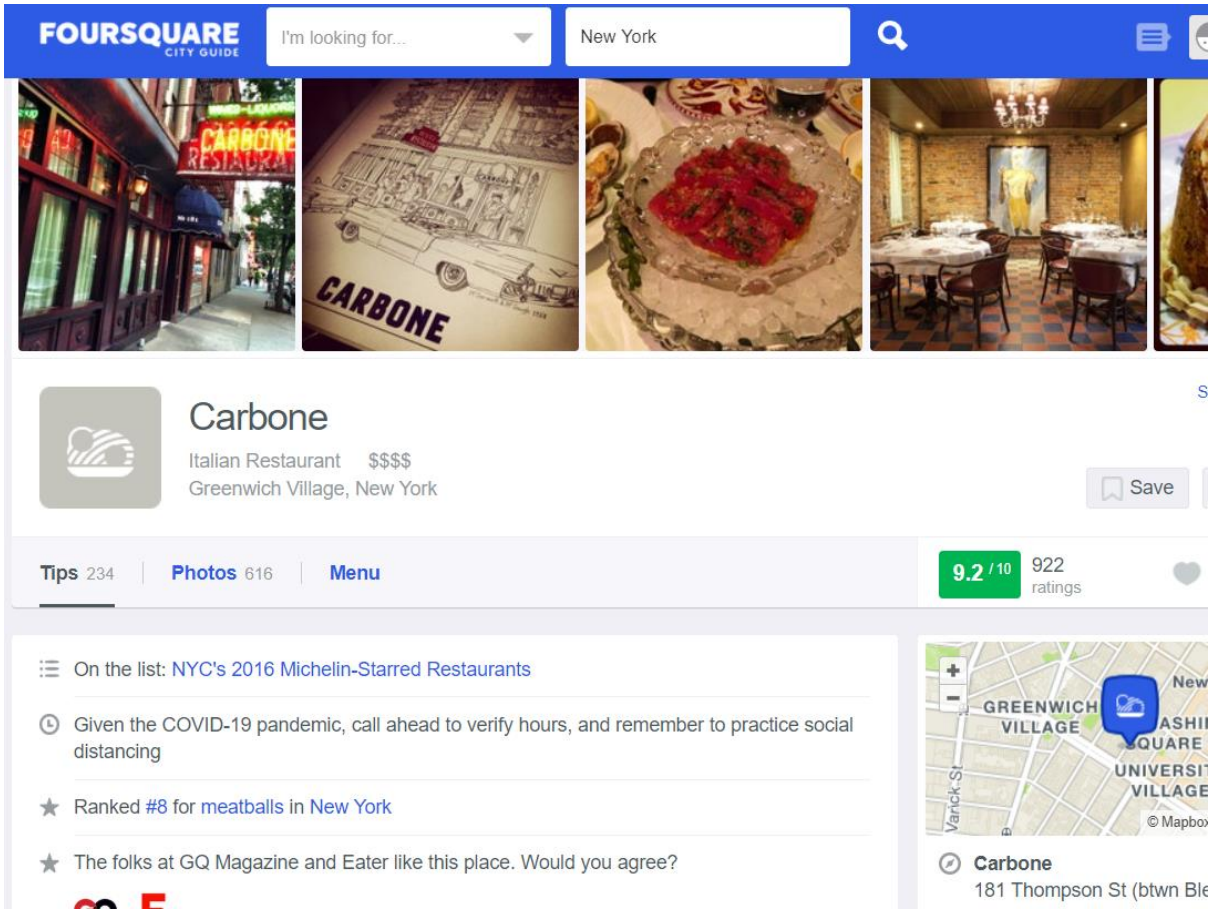
Next, we analyse the top 10 restaurants that are most common and famous in NYC.



We gather the top 5 restaurants venue data that include the following features:

a) Ratings
b) Total Reviews
c) Total Likes
d) Price Category (Cheap, Moderate, Expensive, Very Expensive)

This data is obtained from the Foursquare API Venue Dataset. The venue page on Foursquare website looks like this:



Using this dataset, we analyse the clusters and their most common restaurants. We assume that the restaurants with the most review counts are considered to be the most visited and famous among the customers. The list of neighbourhoods with the most number of reviews give and idea of where the customers are found the most.

	Neighborhood	Sum_Reviews	Total_Venues
0	Little Italy	18555	22
1	Soho	14153	17
2	West Village	11779	18
3	Chelsea	10680	11
4	Noho	10429	20
5	Greenwich Village	8585	16
6	Tribeca	7173	12
7	East Village	6862	13
8	Flatiron	6579	13
9	Lincoln Square	6272	16
10	Turtle Bay	5538	18
11	Murray Hill	5034	12

The clusters containing the top most common restaurant categories are the top five restaurants:

	Venue_Category	Total
0	Italian Restaurant	100
1	American Restaurant	69
2	Mexican Restaurant	54
3	Chinese Restaurant	45
4	Japanese Restaurant	43

Hence, we predict the best locations for these top five restaurants based on the venue features and neighbourhood clustering.

4. Discussions

Here, for each of the top five restaurants, we will predict two locations for opening the business:

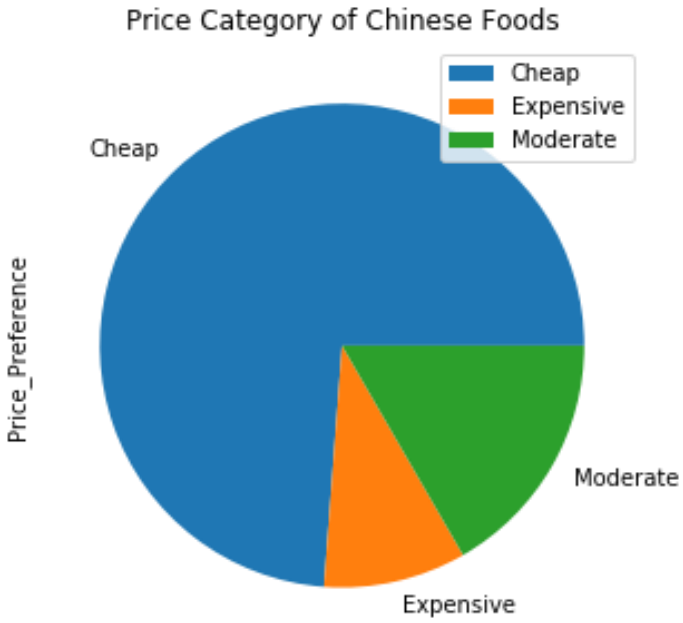
a) **Ideal Location**: A neighbourhood where the restaurant category is most reviewed, most rated that suggests the fact that this neighbourhood attracts the most customers.
b) **Potential Location**: A neighbourhood where the new stakeholder can establish his business that promises many aspects of success.

4.1 Chinese Restaurant Location

The five neighbourhoods in the cluster where the Chinese category shows on the most common venue is then analysed.

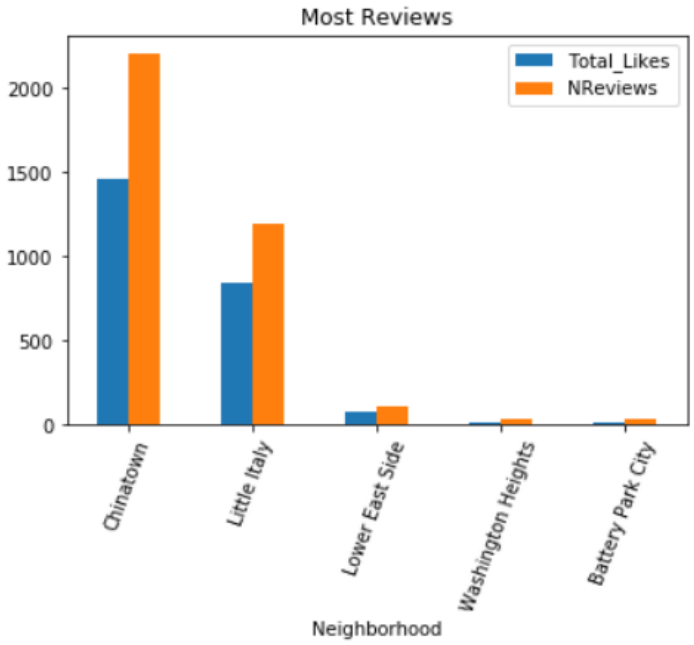
	Neighborhood	Price_Category	NRestaurants	Avg_Ratings	NReviews	Total_Likes
0	Chinatown	Cheap	7	8.100000	2202	1465
1	Little Italy	Cheap	5	8.240000	1195	850
2	Lower East Side	Cheap	3	7.633333	107	77
3	Washington Heights	Moderate	3	7.133333	35	16
4	Battery Park City	Cheap	1	7.300000	34	19

Most of the Chinese restaurants fall under the cheap price category and are also famous. Hence, a cheap price category is to be finalised for opening a Chinese Restaurant.



Now, here the most common is Chinatown neighbourhood where 100% of the venues are in Cheap Price Category. Hence, Chinatown is an ideal location to establish the business.

The potential location can be considered to be Little Italy for its most popularity amongst all based on the maximum likes and reviews. The fact that it is also second popular among Chinese restaurants makes it a potential place to open the business.

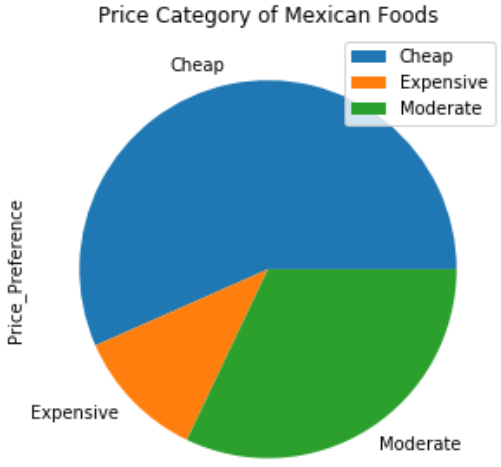


4.2 Mexican Restaurant Location

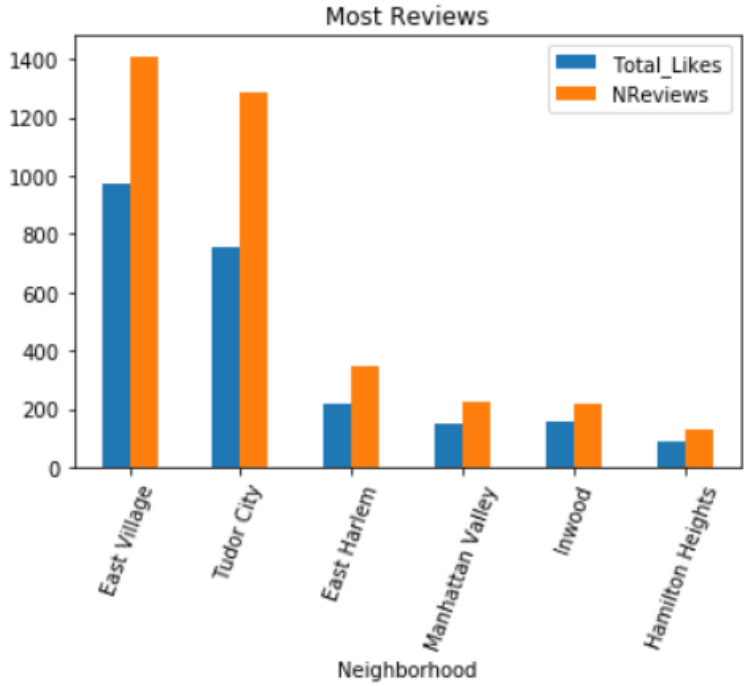
The six neighbourhoods in the cluster where the Mexican category shows on the most common venue is then analysed.

	Neighborhood	Price_Category	NRestaurants	Avg_Ratings	NReviews	Total_Likes	NLikesperReviewRatio
0	East Village	Cheap	4	8.275000	1410	970	0.687943
1	Tudor City	Moderate	4	7.550000	1286	757	0.588647
2	East Harlem	Moderate	5	7.920000	346	219	0.632948
3	Manhattan Valley	Cheap	2	8.000000	227	148	0.651982
4	Inwood	Moderate	4	7.750000	220	155	0.704545
5	Hamilton Heights	Cheap	3	7.966667	126	87	0.690476

The common price category among the Mexican restaurants is the Cheap price category



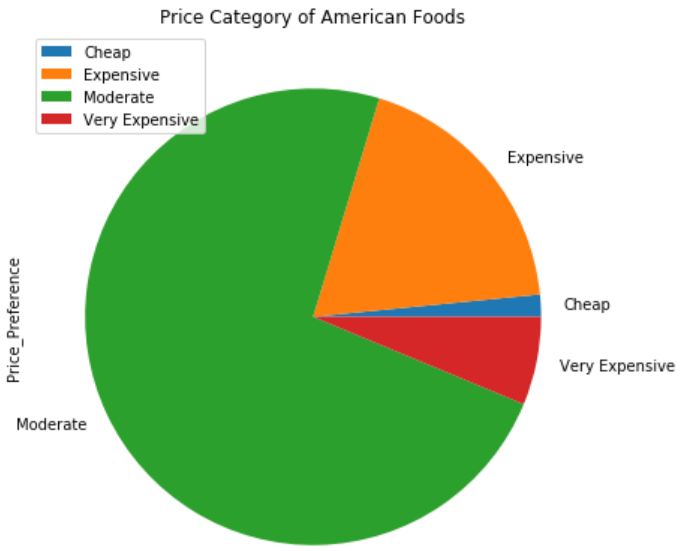
Here, East Village is the ideal location here with most reviews, ratings, likes and also cheap. Manhattan Valley is a potential location for its low competition, cheap price category and better in overall customer attraction than others that can prove beneficial.



4.3 American Restaurant Location

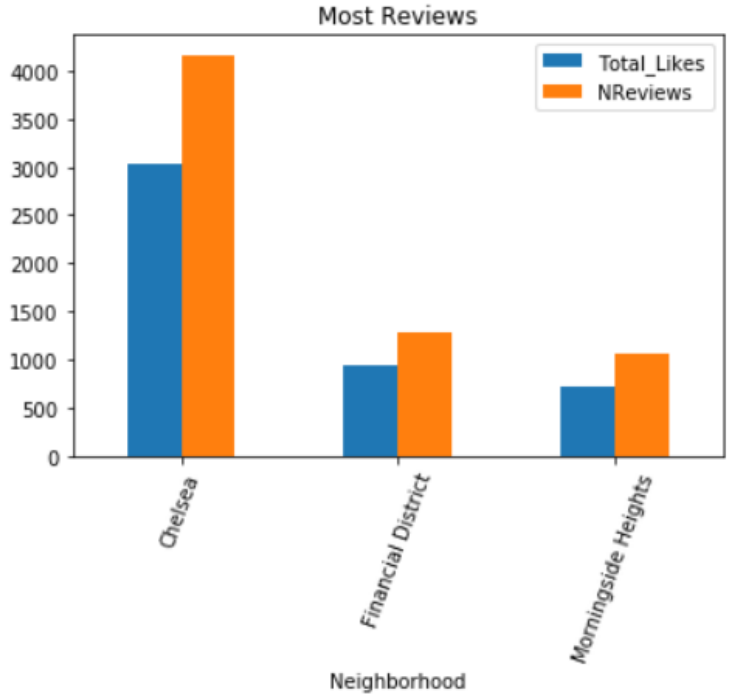
The three neighbourhoods in the cluster where the American category shows on the most common venue is then analysed.

	Neighborhood	Price_Category	NRestaurants	Avg_Ratings	NReviews	Total_Likes	NLikesperReviewRatio
0	Chelsea	Expensive	3	8.800000	4167	3042	0.730022
1	Financial District	Expensive	5	8.160000	1275	943	0.739608
2	Morningside Heights	Expensive	3	8.066667	1052	721	0.685361



Here, the price category is moderately expensive category.

Chelsea neighbourhood is dominant here with most reviews and likes and hence, the ideal location. For American, most of them are in moderate to expensive price category, and Morningside Heights is a potential location with 50% market in American restaurants category and less number of restaurants relatively.

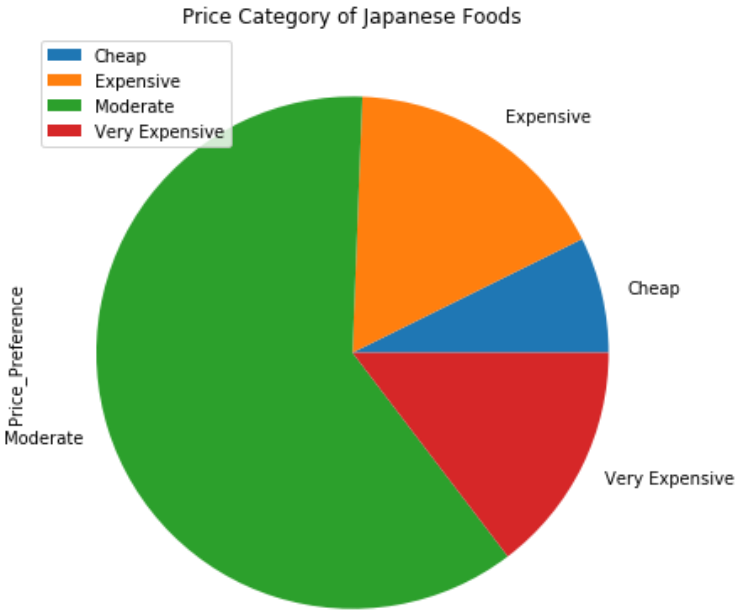


4.4 Japanese Restaurant Location

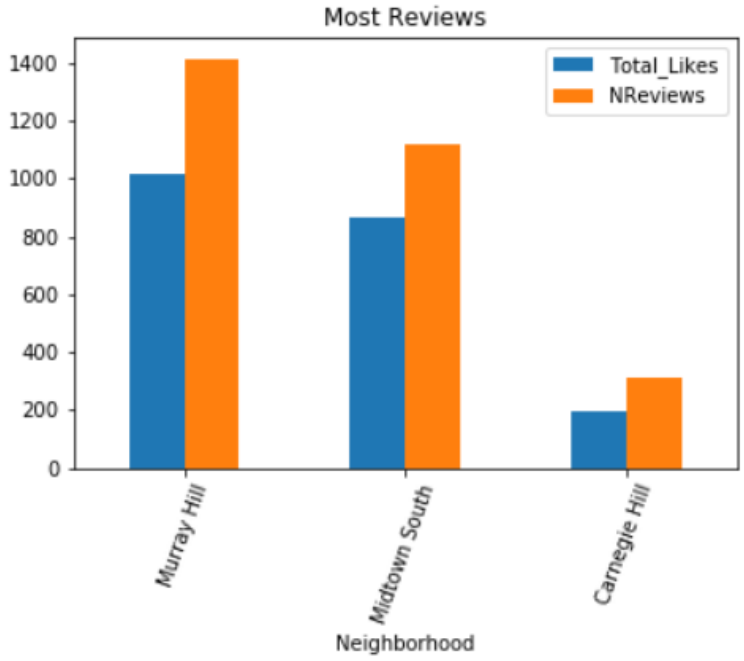
The three neighbourhoods in the cluster where the Japanese category shows on the most common venue is then analysed.

	Neighborhood	Price_Category	NRestaurants	Avg_Ratings	NReviews	Total_Likes	NLikesperReviewRatio
0	Murray Hill	Very Expensive	3	8.433333	1417	1020	0.719831
1	Midtown South	Moderate	4	8.550000	1122	865	0.770945
2	Carnegie Hill	Moderate	3	8.200000	313	197	0.629393

Here, the price category is moderately expensive category.



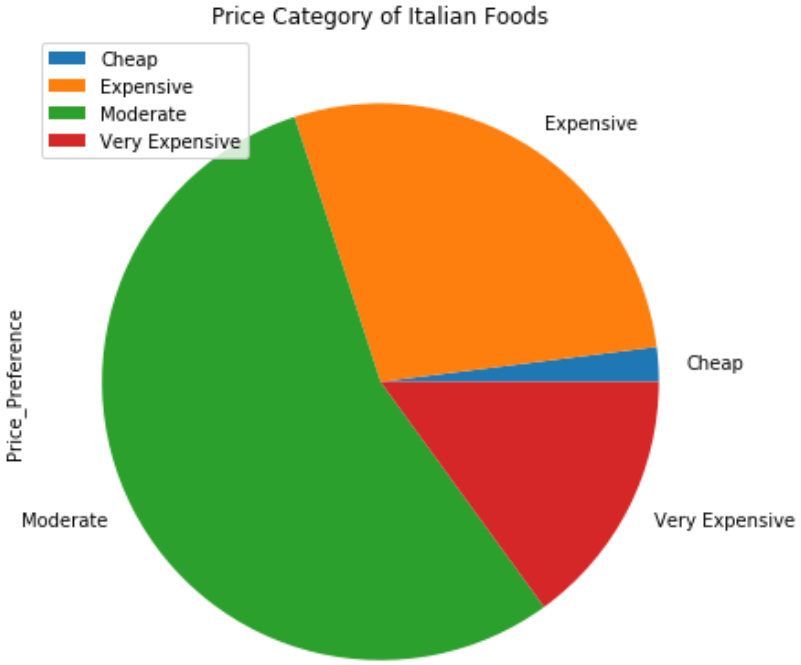
Murray Hill and Midtown South neighbourhoods have Moderate to expensive Price Categories and dominate the Japanese Market, hence the suitable locations.



4.5 Italian Restaurant Location

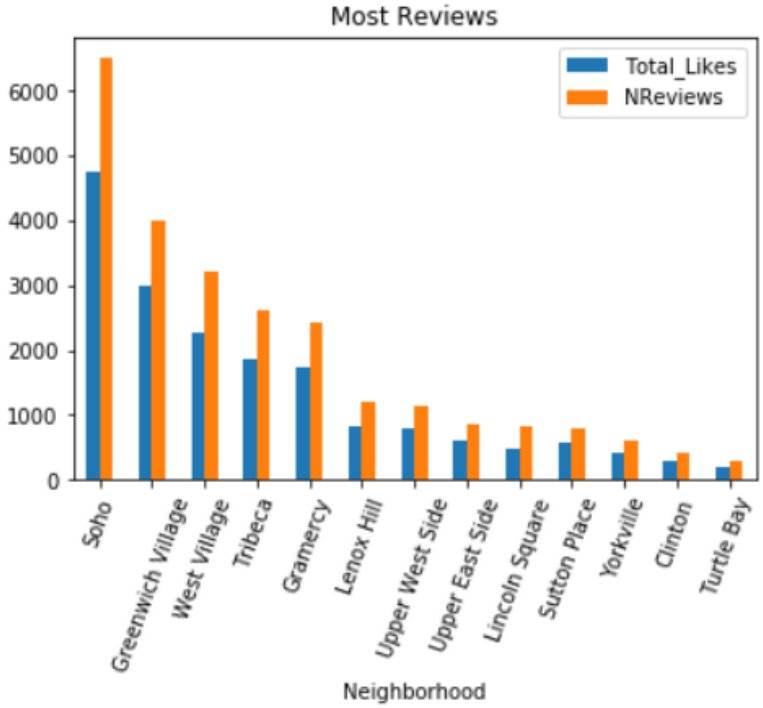
The thirteen neighbourhoods in the cluster where the Japanese category shows on the most common venue is then analysed.

Most of the Italian Restaurants are in Moderate to Expensive Price Category



Soho Neighbourhood is ideal for its most reviews and likes. (Almost double the second)

Tribeca has around 42% of its restaurant market in Italian and ranked 6th in most reviewed restaurants. Hence, it may be a potential.



Finally, the ideal and potential locations for the top five most common restaurants can be highlighted as:

	Restaurant Category	Ideal Neighborhood Location	Potential Neighborhood Location
0	Chinese	Chinatown	Little Italy
1	Mexican	East Village	Manhattan Valley
2	Japenese	Murray Hill	Midtown South
3	American	Chelsea	Morningside Heights
4	Italian	Soho	Tribeca

5. Conclusion

A data science methodology was used to predict the best location for opening a restaurant in New York City. The data input for the project was the location coordinates of the five boroughs and their neighbourhoods. The restaurant venue features were obtained from Foursquare API. Combining these two datasets, we were able to provide an insight into where one can ideally as well as potentially open his new business and also which specific restaurant business in the City of New York. The analysis can further be improved by more data analysis on the reviews and tips provided by the customers, find implications and work upon them in your new business to thrive at the neighbourhood. Also, the machine learning algorithms can also be used on the feature dataset to get better results.