1. Introduction

1.1 Background

New York city is a compact business center, as well as one of the most popular tourist destinations in the world. The Big Apple attracts on average 60 million visitors annually in the last few years and that number is still growing. New York city is currently ranked sixth, behind Bangkok, London, Paris, Dubai, and Singapore. Among American cities, New York ranks second only to Orlando in the number of visitors it draws. Orlando claims to attract more than 70 million visitors a year (1). In 2018, total visitor spending was estimated to be around $44 billion. Data published by VISA in 2016 suggested that 24% of all VISA spending was attributed to restaurants and food. It is clear that restaurants in New York city hold a big chunk of the pie. The success of a restaurant is greatly affected by its location and whether its location is in close proximity to other popular venues and destinations. It is highly advantageous to accurately predict which location would be best suited for a new restaurant to thrive. This information can be used to target desirable location to open a new restaurant.

1.2 Problem

Data that might contribute to determine the success of a restaurant might include its location, which other popular tourist destinations are in its proximity, which other popular restaurants are close by, and how many hotels are in its proximity. This project aims to predict the best location to open a new restaurant based on the data located to those criteria.

2. Data

2.1 Geographical Data

First, we will need to look at the geo data of New York city and explore the 5 boroughs of New York City to see where to open the new restaurant. We will be segmenting the neighborhoods and exploring them individually. To do that, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the the latitude and longitude coordinates of each neighborhood. The geo data used to analyze different boroughs and neighborhoods is available for free here: https://geo.nyu.edu/catalog/nyu_2451_34572. After segmenting and clustering the neighborhoods by borough, it will help to be able to visualize it using folium.

2.2 Foursquare Data

Utilizing the data available on Foursquare API, we will be exploring the venues in the proximity of the desired boroughs and neighborhoods. The data downloaded from Foursquare will most likely contain some irrelevant values. We will only look at the fields such as venues name, category, latitude, and longitude. There will be many types of businesses available on Foursquare. We will try to only look at the businesses with categories that most resemble restaurants. People usually dine in restaurants close to where they are staying. For this reason, we will also explore hotels in these neighborhoods. This could be a good factor to determine the prime location for a new restaurant.

2.3 Feature Selection

To predict where is the best location for our new restaurant, we need to examine the current restaurant scene in New York city. To do so, we need to get a list of restaurants in each borough and neighborhood. We can do that by using the Foursquare API. By using the latitude and longitude coordinates of each neighborhood, we will get the top 100 venues within the radius of 500 meters of each neighborhood. Below is what part of the result looks like:


```
{'reasons': {'count': 0,
   'items': [{'summary': 'This spot is popular',
     'type': 'general',
     'reasonName': 'globalInteractionReason'}]},
  'venue': {'id': '4b5357adf964a520319827e3',
   'name': "Dunkin'",
   'location': {'address': '5501 Broadway',
    'crossStreet': 'W 230th St',
    'lat': 40.87713584201589,
    'lng': -73.90666550701411,
    'labeledLatLngs': [{'label': 'display',
      'lat': 40.87713584201589,
      'lng': -73.90666550701411}],
    'distance': 342,
    'postalCode': '10463',
    'cc': 'US',
    'city': 'Bronx',
    'state': 'NY',
    'country': 'United States',
    'formattedAddress': ['5501 Broadway (W 230th St)',
     'Bronx, NY 10463',
     'United States']},
   'categories': [{'id': '4bf58dd8d48988d148941735',
     'name': 'Donut Shop',
     'pluralName': 'Donut Shops',
     'shortName': 'Donuts',
     'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/donuts_',
      'suffix': '.png'},
     'primary': True}],
   'photos': {'count': 0, 'groups': []}},
  'referralId': 'e-0-4b5357adf964a520319827e3-4'},
```


Figure 1 Sample of Foursquare data

A lot of the data available are not relevant to what we are trying to accomplish. We will only focus on the fields that would help us with the analysis. From looking at the Foursquare data, it looks like the information we need is in the *items* key. From the *items* key, we will grab the *name of the venue*, *latitude*, *longitude*, and *categories name*.

```
{'reasons': {'count': 0,
   'items': [{'summary': 'This spot is popular',
     'type': 'general',
     'reasonName': 'globalInteractionReason'}]},
  'venue': {'id': '4b5357adf964a520319827e3',
   'name': "Dunkin'",
   'location': {'address': '5501 Broadway',
    'crossStreet': 'W 230th St',
    'lat': 40.87713584201589,
    'lng': -73.90666550701411,
    'labeledLatLngs': [{'label': 'display',
      'lat': 40.87713584201589,
      'lng': -73.90666550701411}],
    'distance': 342,
    'postalCode': '10463',
    'cc': 'US',
    'city': 'Bronx',
    'state': 'NY',
    'country': 'United States',
    'formattedAddress': ['5501 Broadway (W 230th St)',
     'Bronx, NY 10463',
     'United States']},
   'categories': [{'id': '4bf58dd8d48988d148941735',
     'name': 'Donut Shop',
     'pluralName': 'Donut Shops',
     'shortName': 'Donuts',
     'icon': {'prefix': 'https://ss3.4sqi.net/img/categories_v2/food/donuts_',
      'suffix': '.png'},
     'primary': True}],
   'photos': {'count': 0, 'groups': []}},
  'referralId': 'e-0-4b5357adf964a520319827e3-4'},
```


Figure 2 Relevant data fields from Foursquare data

2.4 Data Cleaning

We will put the result into a pandas dataframe, joined with the name of the neighborhood, neighborhood latitude and longitude. We will end up with the dataframe like below:

	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

Figure 3 Putting Foursquare data into pandas dataframe

Foursquare data returned the top 100 venues for each neighborhood, regardless of the type of venue. There are venues like Yoga Studio, Park, Gym, Movie Theater among different types of restaurants in the data. To assess the restaurants activity in New York city, we need to isolate the restaurants from the rest of the venues. We can do that by removing the venue categories that do not resemble a restaurant, for

example: Pharmacy, Gas Station, Laundromat, Discount Store, Mattress Store, etc... After removing the venue categories that do not resemble a restaurant, we will be left with a dataframe of only the food venues. Below are some examples:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop
6	Wakefield	40.894705	-73.847201	Subway	40.890468	-73.849152	Sandwich Place
7	Wakefield	40.894705	-73.847201	Pitman Deli	40.894149	-73.845748	Food

Figure 4 Restaurants pandas dataframe example

3. Exploratory Data Analysis

3.1 Restaurants in New York City

To analyze the restaurants in each neighborhood, we need to use one hot encoding on the Venue Category column in our restaurant dataframe. After that, we group the rows by neighborhood and by taking the mean of the frequency of occurrence in each category.

From here, we will get the top 10 most common restaurant venues for each neighborhood and use k-means clustering for each neighborhood into 5 clusters. After adding the cluster labels, we merge this data with the geographical data of New York city.

With this, we can use Folium to visualize the data on the map.

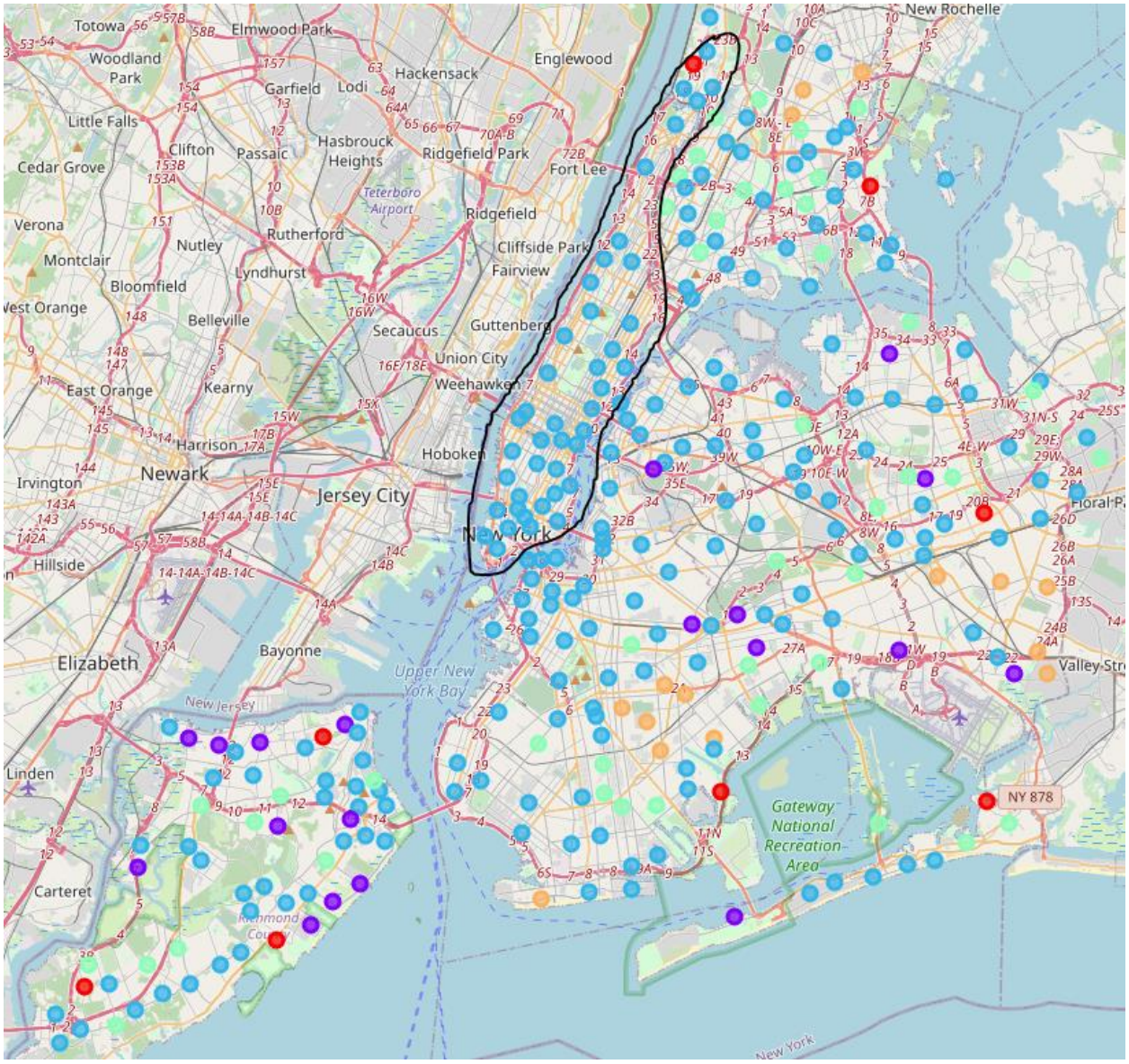


Figure 5 Map of k-means clusters for top 10 restaurants in each neighborhood

As we can see, Manhattan has the most condensed and populated number of restaurants out of all 5 boroughs (circled in black above).

3.2 Attractions in New York City

Remember we are not limited to just restaurants data. Even though we are looking to open a new restaurant, that does not mean the other venues are not useful in our analysis. Foursquare data contains other venue categories as well. If we can figure out which area attracts a lot of traffic, there could be good opportunities to open a restaurant in that area. For example, if people visit some attractions, chances are they would stop by the restaurants that are in close proximity afterwards. We will extract the attractions venues like Baseball Stadium, Park, Bowling Alley, Art Gallery, Museum,

Comedy Club, Concert Hall, etc... and put them in the *Attractions* dataframe to carefully analyze which of the 5 boroughs has more popular attraction venues. Below are a few examples of the *attractions*:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
17	Co-op City	40.874294	-73.829939	truman track n field	40.874963	-73.830847	Baseball Field
21	Co-op City	40.874294	-73.829939	The Park	40.877645	-73.830836	Park
31	Eastchester	40.887556	-73.827806	BowlerLand	40.886020	-73.823207	Bowling Alley
48	Riverdale	40.890834	-73.912585	Bell Tower Park	40.889178	-73.908331	Park
50	Riverdale	40.890834	-73.912585	Seton Park	40.887914	-73.916113	Park

Figure 6 Examples of attraction venues

To analyze the *Attractions*, we will perform the same analysis as we did with *Restaurants*. We will perform the one hot encoding for the attraction venues, and group rows by neighborhood and by taking mean of the frequency of concurrency for each category. We will get the top 10 most common Attractions in each neighborhood and run k-means clustering. After having the cluster label for each group, we merge the data with the neighborhood data to get the latitude and longitude coordinates and visualize the clusters on the map of New York city.

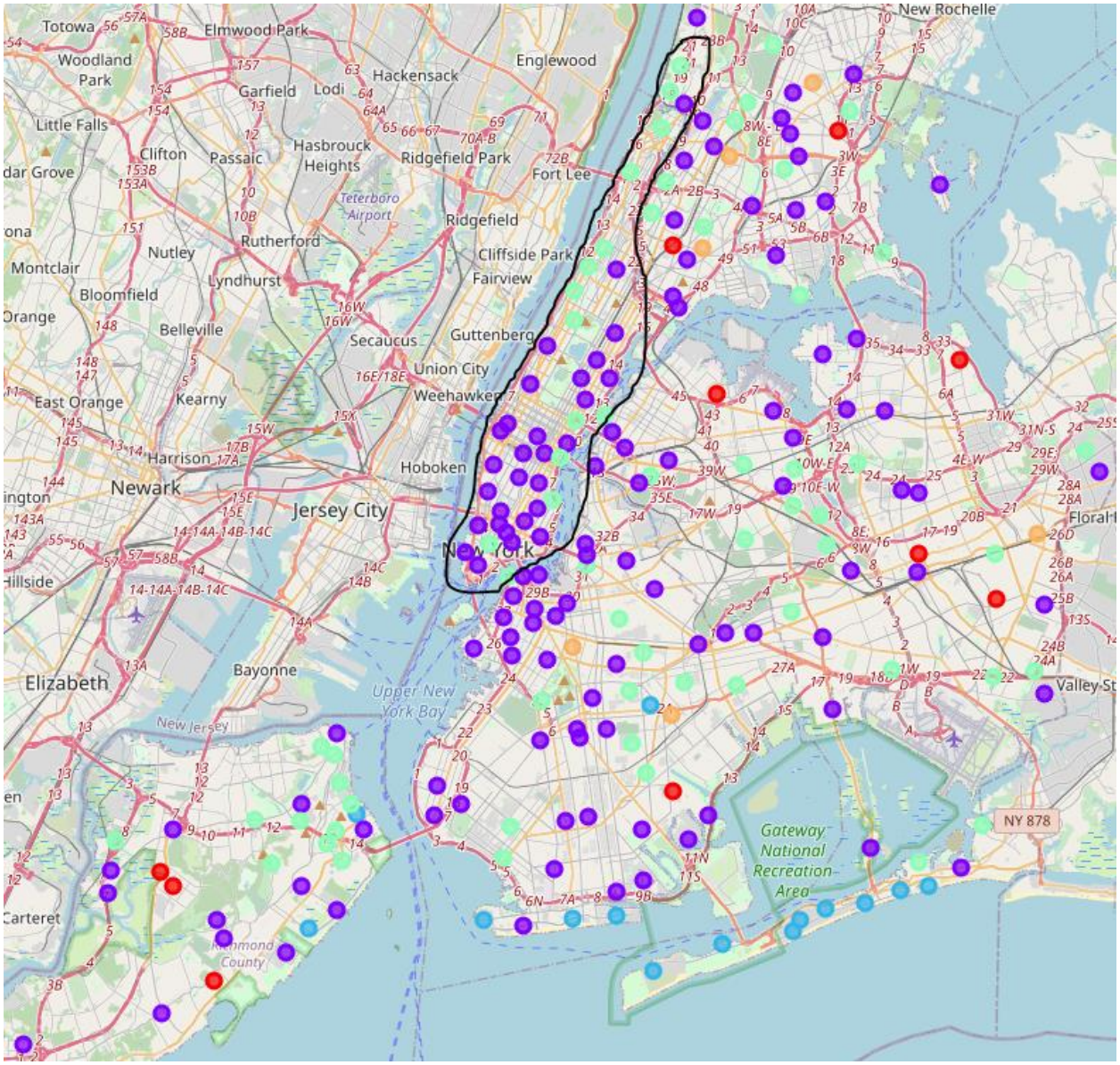


Figure 7 Map of k-means clusters for the top attractions in each neighborhood

From a quick glance, we can see the trend holds for attractions as well. There are more densely populated attractions within Manhattan, when compared to the other boroughs.

3.3 Hotels in New York city

If we follow the same logic, we should also look at the most popular hotels in each borough, as people would tend to dine in venues that are fairly close to where they are lodging. Similar to our approach with the attractions, we will extract the venue categories that resemble hotels or lodging venues and put them in the *Hotels* group. Unfortunately, the data we received from Foursquare did not contain many hotels. Below is the entire list of hotel venues that we have got from using Foursquare API.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Astoria Heights	1	1	1	1	1	1
Bellerose	1	1	1	1	1	1
East Williamsburg	1	1	1	1	1	1
Long Island City	1	1	1	1	1	1
Manhattan Valley	1	1	1	1	1	1
Queensbridge	1	1	1	1	1	1

Figure 8 All hotels from Foursquare dataset

Six hotel venues are not enough to perform our analysis. These would be the outliers in our data. Therefore, I decided to drop these venues from the analysis.

3.4 Other venues in New York city

Besides attractions and hotels, there are other venue categories that are worth looking at. Gym, Fitness Center, Business Center, Yoga Studio, Bus Station, Metro Station, etc... are locations that would attract lots of foot traffic, especially the customers who are hungry for a new, bold yet affordable culinary haven (hopefully). If we follow the same approach and perform the k-means clustering on the above-mentioned venue categories, we will end up with the following on the map of New York city.

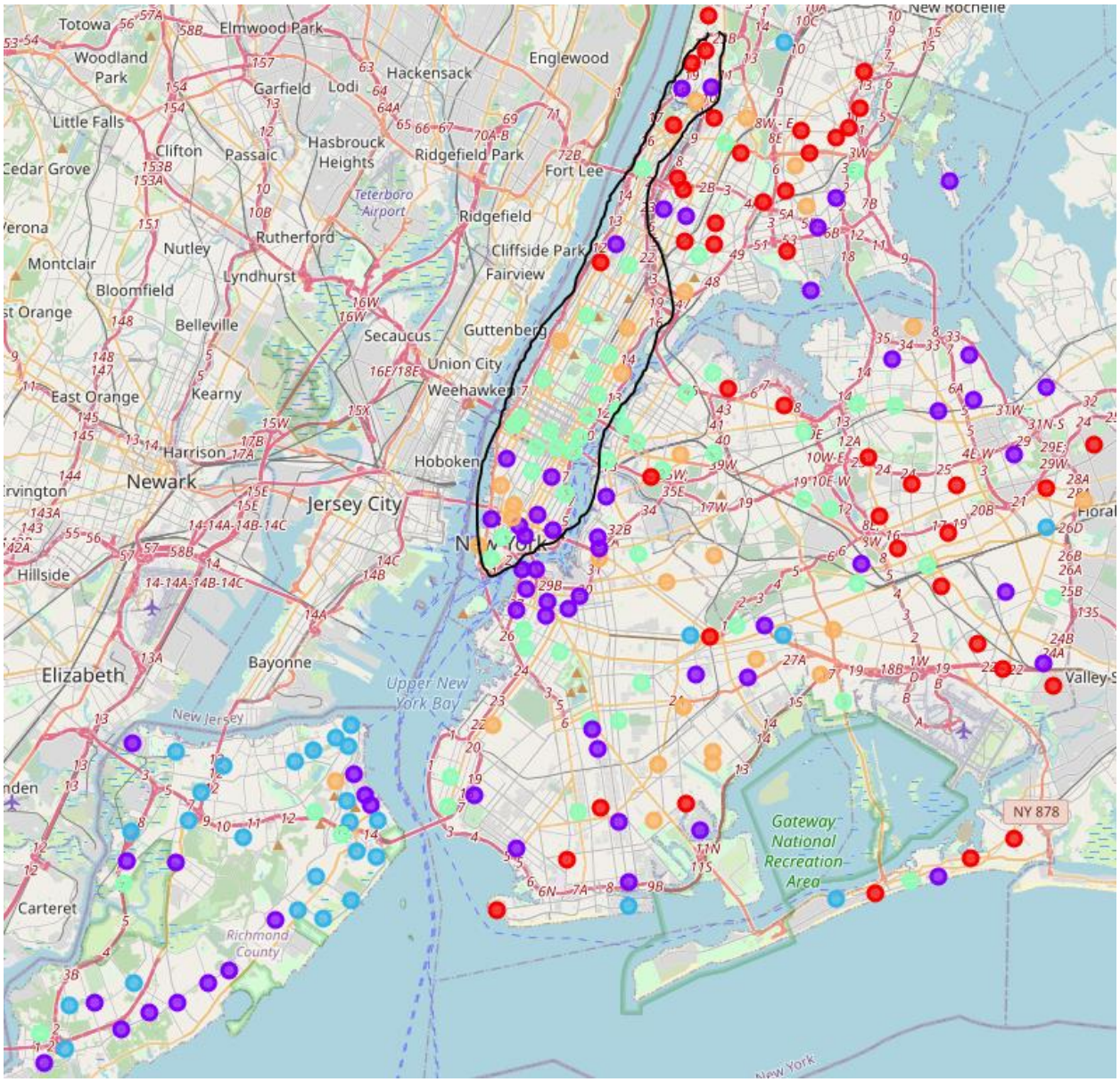


Figure 9 Map of k-means clusters for other venues that also attract lots of foot traffic

Again, there seems to be many venues in the Other categories within close proximity with one another in Manhattan. We can now confirm that Manhattan is the best borough to open our new restaurant. However, our work is not done yet. Our goal is to identify the best neighborhood within the best borough for our new restaurant. So, let us carefully look at Manhattan.

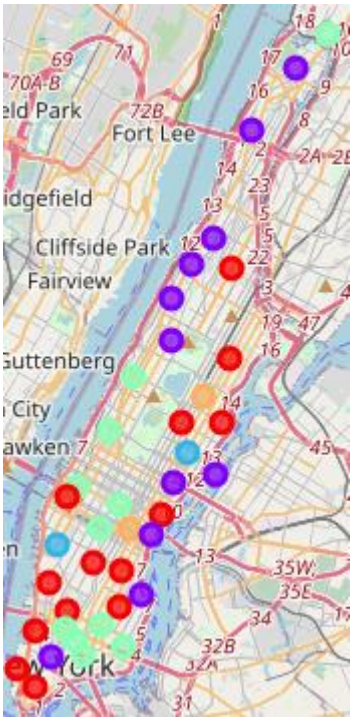
3.5 Manhattan

We do that by isolating Manhattan from our geographical dataset.

```
The geograpical coordinate of Manhattan are 40.7896239, -73.9598939.
```

Using the geographical coordinates, we can get the top venues for just Manhattan. Again, we will only analyze the venue categories that fit the following categories: Restaurants, Attracts, or Others (venues that attract lots of foot traffic, like gym, business center, metro station, etc...). Once we have that, we will grab the top 10 most common venues in each neighborhood and run k-means clustering on all 3 groups, and then visualize the results on the map of Manhattan.

Restaurants



Attractions

Others

4. Results

4.1 Closer look at the neighborhoods

It looks like the relevant venues are more densely populated in the south half of Manhattan. But it does not tell us which neighborhood is which. We need to look at it from a different angle. Let us try putting the number of venues for each neighborhood in Manhattan on a graph. To do this, we will simply count the number of venues in each neighborhood for 3 separate groups as above, Restaurants, Attractions, and Others.

Below is the graph of Restaurants in each neighborhood of Manhattan:

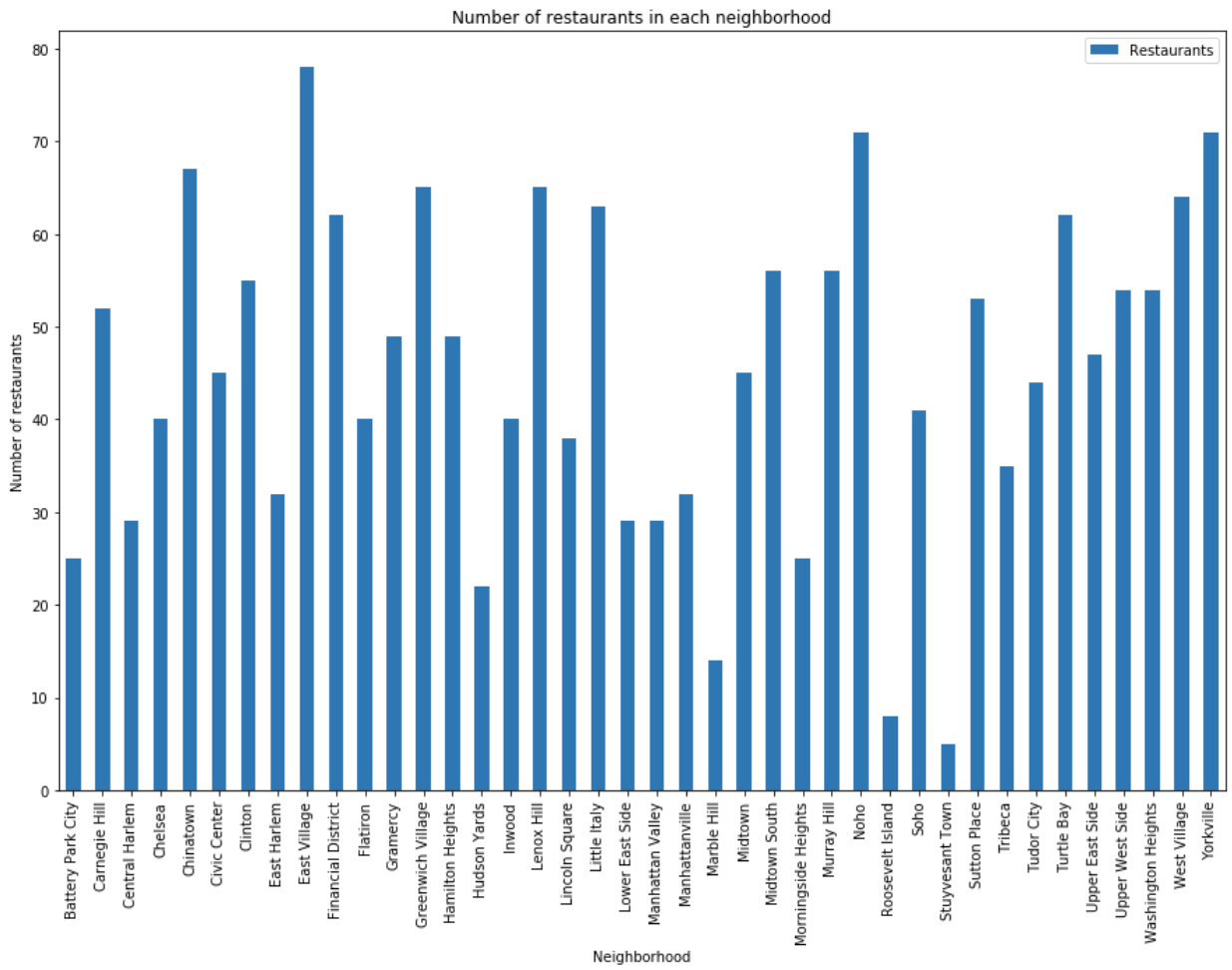


Figure 10 Number of restaurants in each neighborhood

The top 4 neighborhoods with the greatest number of restaurants are East Village, NoHo, Yorkville, and Chinatown, while Greenwich Village and Lenox Hill tie for 5th place. West Village, Turtle Bay, Little Italy, and Financial District are also good options as they are among the top 10 most populated neighborhood in term of dining venues. Let us examine the graphs of the other two groups.

4.2 Other things to consider

In the Attractions group, most of the venues are evenly spread out among the neighborhoods, except for a few areas such as Chelsea, Lincoln Square, Tribeca, West Village, Battery Park City, and Clinton.

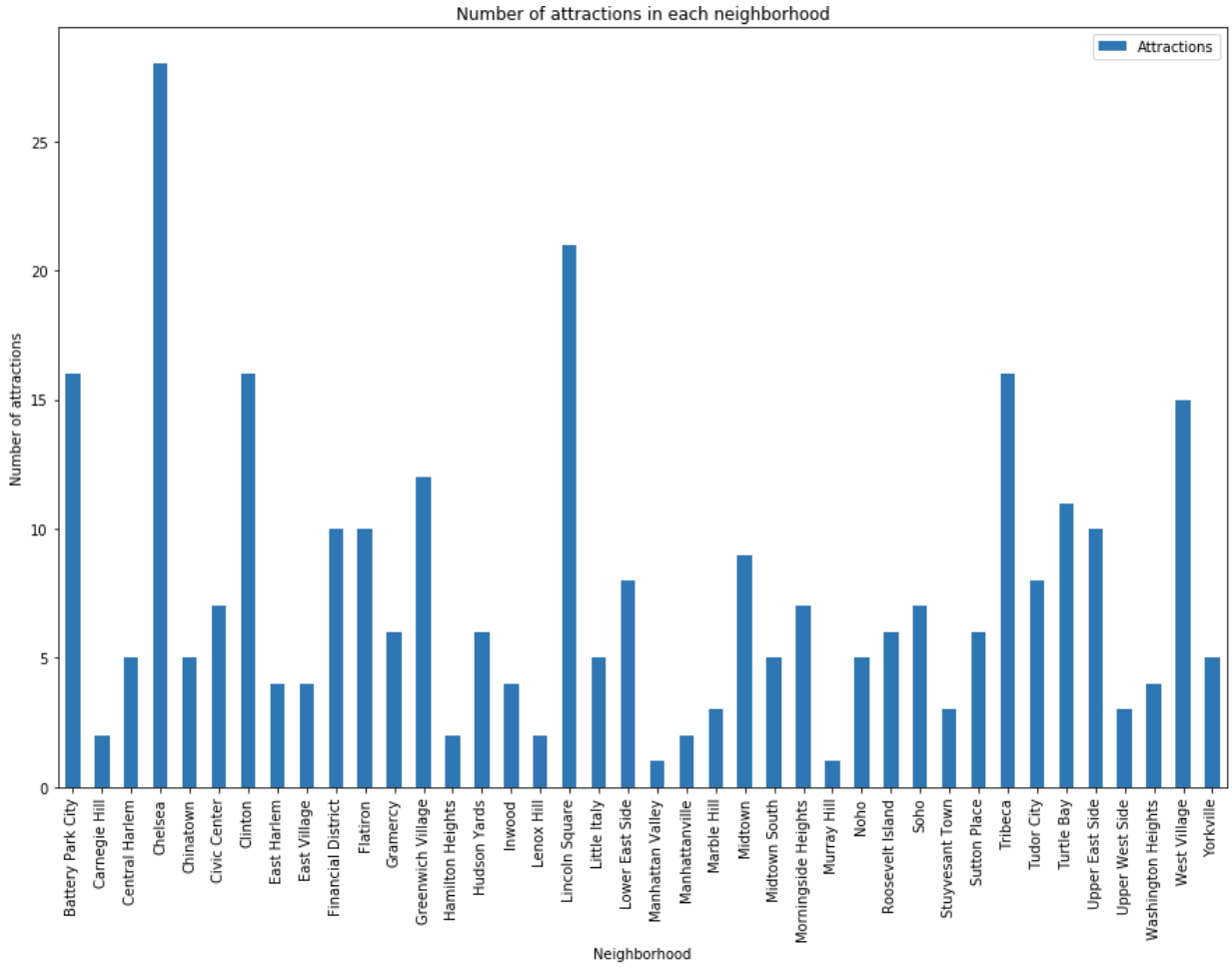


Figure 11 Number of attractions in each neighborhood

When we look at the Others group, it follows the same trend as the venues spread out evenly among the neighborhood, with an outlier in Flatiron.

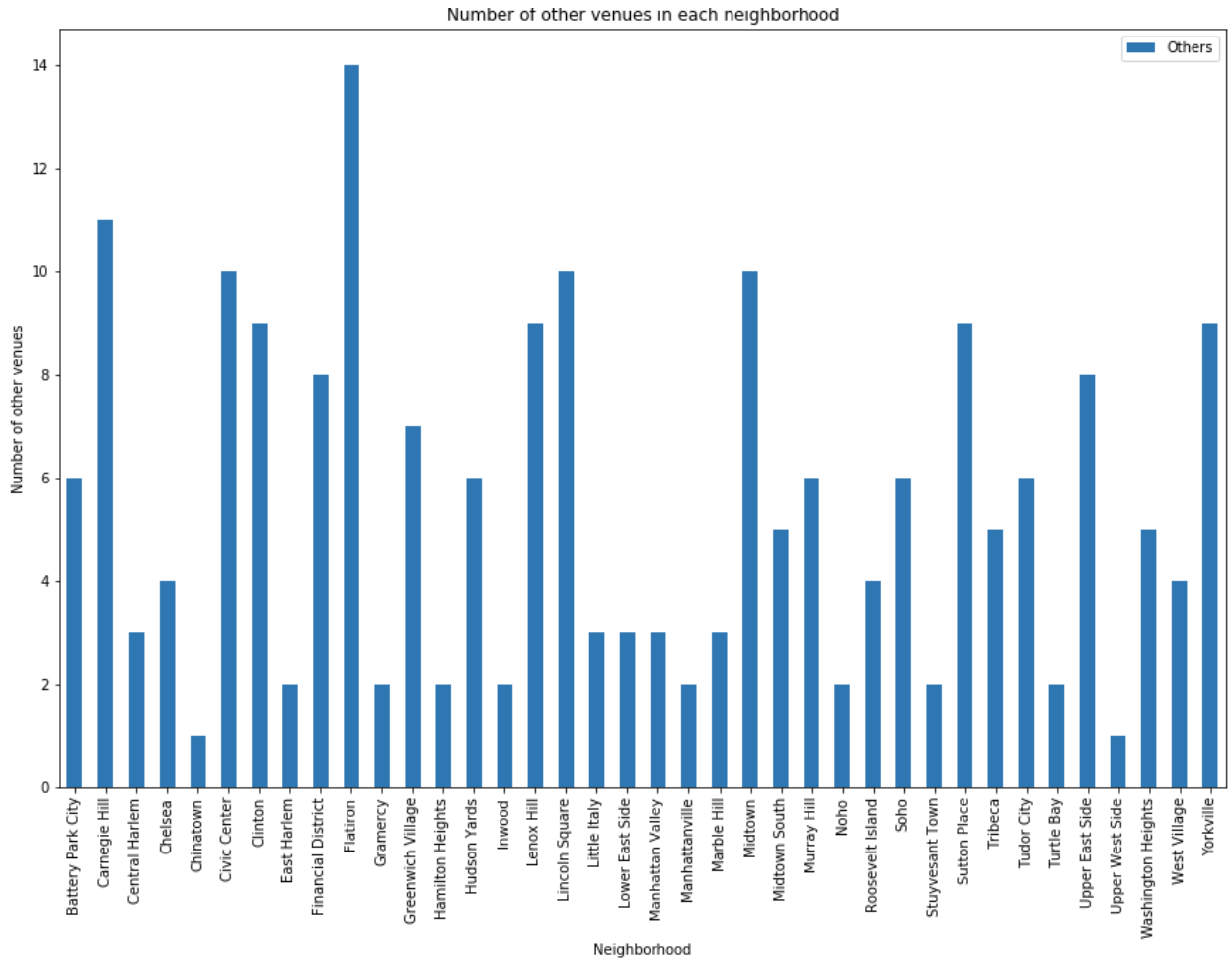


Figure 12 Number of other venues in each neighborhood

The above three graphs gave us different pieces of the puzzle. The Restaurants group lets us evaluate which neighborhoods are thriving in the culinary sector, while the Attractions and Other groups show us which neighborhoods are doing well in their own sectors. To complete the picture, we need to take it to the next level, by combining all three graphs into one, to see which neighborhood is the optimal option. We start by joining the 3 groups, using Neighborhood as key. Because some neighborhoods would be missing values, we could replace those missing values with *zero*. Here is a few examples of the joined dataframe:

	Restaurants	Attractions	Others
Battery Park City	25	16	6.0
Carnegie Hill	52	2	11.0
Central Harlem	29	5	3.0
Chelsea	40	28	4.0
Chinatown	67	5	1.0

Figure 13 Example of joined dataframe

This is the final graph, with all venues for each neighborhood of Manhattan.

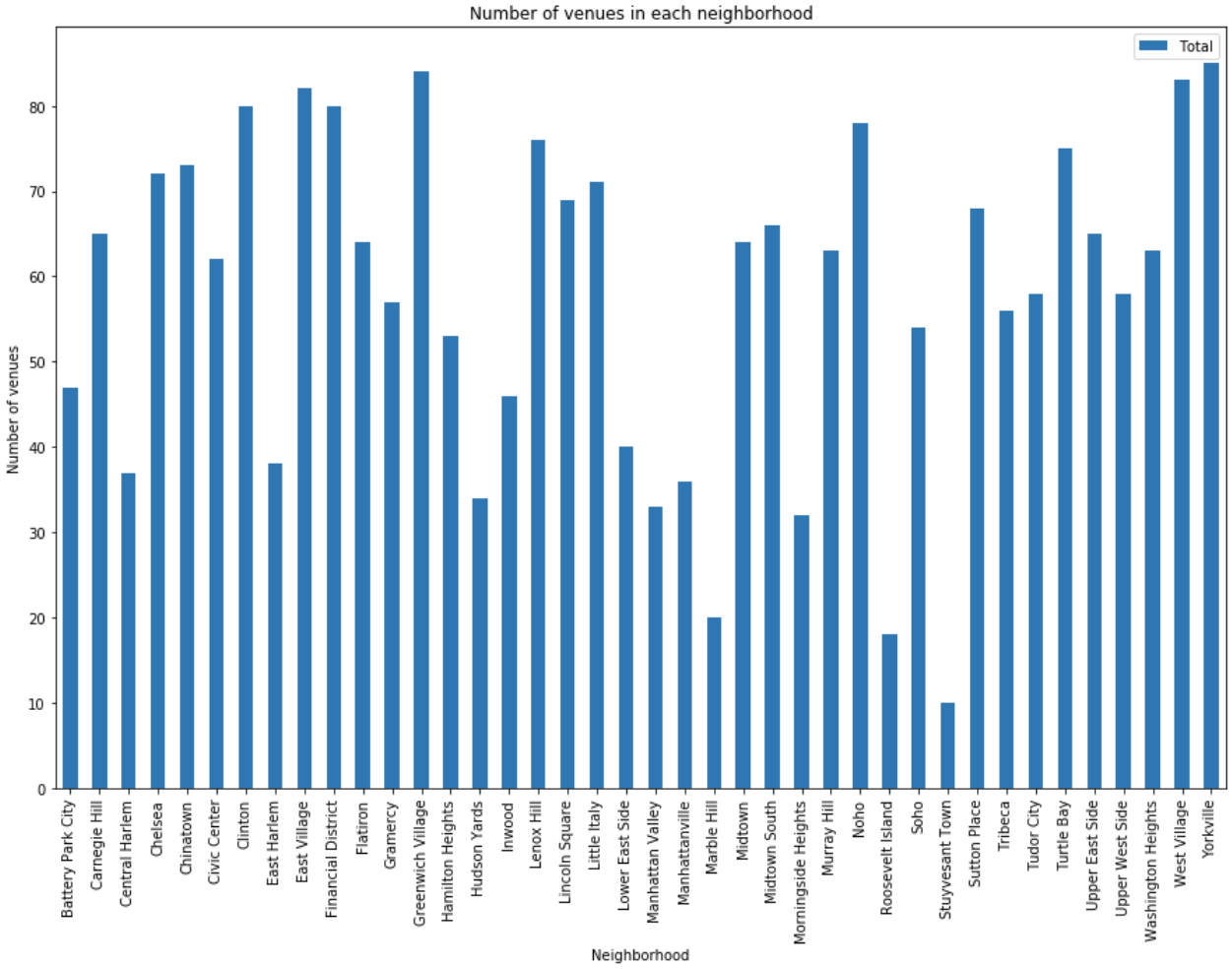


Figure 14 Total number of venues in each neighborhood

4.3 Recommendation

The top few neighborhoods are Greenwich Village, Yorkville, East Village, West Village, Financial District, Civic Center, and Noho. As we recall, East Village, NoHo, Yorkville, Greenwich Village and West Village are among the top neighborhoods with the greatest number of popular restaurant venues. Since these

neighborhoods also have the most foot traffic from other venue categories, they would be the perfect locations for a new restaurant to thrive.

5. Conclusion

In this analysis, we have been through the journey of finding out the optimal location for a new restaurant to thrive in as I have laid out the relationship between New York city geographical data and Foursquare data. We have examined and analyzed together the relevant data after carefully mining what we could from the data provided to us. We have witnessed together how many of the popular restaurants and other popular venue attractions were well located in proximity of the neighborhoods of Manhattan. Finally, we took a step further to close in on which neighborhoods were the best choices for a new restaurant.

References

(1) https://www.nytimes.com/2019/01/16/nyregion/nyc-tourism-record.html