The Battle of the Neighborhoods

Annie Kang

June 20, 2019

1. Introduction/Business Problems

1.1 Introduction

New York City has population of 8.3MM estimated on 2019 and distributed on 5 boroughs - Brooklyn, Queens, Manhattan, the Bronx, and Staten Island. New York City has been described as the cultural, financial, and media capital of the world, significantly influencing commerce, entertainment, research, technology, education, politics, tourism, art, fashion, and sports. Above all, tourist attractions and landmarks in New York City are well known with record 62.8 MM tourists visited New York City in 2017.

Doing any business in New York City is very competitive, but rewarding experience with higher risk. So to mitigate the risk, thorough business preparation and understanding is very crucial.

1.2 Business Problem

Korean Tourism Organization under Ministry of Culture and Tourism wants to open a tourism office spaces to promote Korea Tourism and culture to domestic/international tourists visiting New York City. Thanks to K-drama and K-pop, Korean culture is well-known. Compare to China and Japan, Korea is still behind as tourist area. Because of COVID-19 pandemic, it seems tourism is in halt, however, human being are very adaptable and resilient, Korean government believe in that normal traveling will be eventually resumed. I strongly think it is good time to plan carefully and strategically in advance before tourism re-start.

The office space will have exhibition space, K-drama and K-pop view venues. It will be opened nearby places where most tourists are visited and have more numbers of attractions.

Which neighborhood will be the best place to open the office?

Where can be a culturally attuned area?

Questions/problems to review carefully by conducting the through data analytics;

- Trend of tourists visiting to NYC (up or down trending last 5 years)
- New York City population – can be interested in Korean Tourism and Culture
- Numbers of tourists visiting to each attraction; list of most visited attractions
- Neighborhoods of NYC attractions
- Building property price renting the building space near neighborhoods in attraction areas

1.3 Target Audience

Korean Tourism Organization appointed me as a team leader of data science team to conduct data analytics on expansion of promoting country's branding and for the first project I have to open the office space where can exhibit Korean Culture and Tourism.

However, the audience will be anyone or any country who wants to consider opening the office in a neighborhood of mostly visited attraction area in NYC.

2. Data

NYC has 5 boroughs and 306 neighborhoods. Since most visited attractions are concentrated in Manhattan borough. I will focus on Manhattan for finding the best office space to be opened.

2.1 Data sources

- Demographics of New York City: Population, Gross Domestic Product and Gross Domestic Product per capita of year 2019
  - Dataset: https://en.wikipedia.org/wiki/Demographics_of_New_York_City

- Tourism in New York City – Visitors data: Domestic/International visitors with visitor spending from year 1991 thru 2018 - page[0] on html below
  - Dataset: https://en.wikipedia.org/wiki/Tourism_in_New_York_City

- Tourism in New York City – Most visited attractions: name, numbers of visitors and locations of most visited attractions visited with 2MM visitors - page[1] on html below
  - Dataset: https://en.wikipedia.org/wiki/Tourism_in_New_York_City

- NYC neighborhoods: 5 boroughs and 306 neighborhoods with the latitude and longitude
  - Dataset: https://geo.nyu.edu/catalog/nyu_2451_34572

- All attractions in NYC: Foursquare API conjunction with neighborhoods' coordinates above
  - API : https://api.foursquare.com/v2/venues
  - Out of 331 venues categories from Foursquare API only 28 venue categories listed below will be used to extract

Art Museum	Monument / Landmark
Concert Hall	Museum
Cultural Center	Music Venue
Department Store	Opera House
Duty-free Shop	Outdoor Sculpture
Exhibit	Park
Garden	Pedestrian Plaza
Garden Center	Performing Arts Venue
Historic Site	Sculpture Garden
History Museum	Shopping Mall
Indie Movie Theater	Theater
Indie Theater	Theme Park Ride / Attraction
Library	Tourist Information Center
Memorial Site	Train Station

2.2 Data cleaning

For New York City population data the column names were changed accordingly and dropped the columns that are not used for this project in order to make the data as slim as possible. In the data for visitors the NaN

value was found in TotalVisitorSpendingBillions(US$) and imputed with the mean value of previous and forward rows. For the attraction list visited more than 2MM there were attraction names with (), so the attraction name cleaned by removing () in order to merge with venues retrieved from foursquare API later.

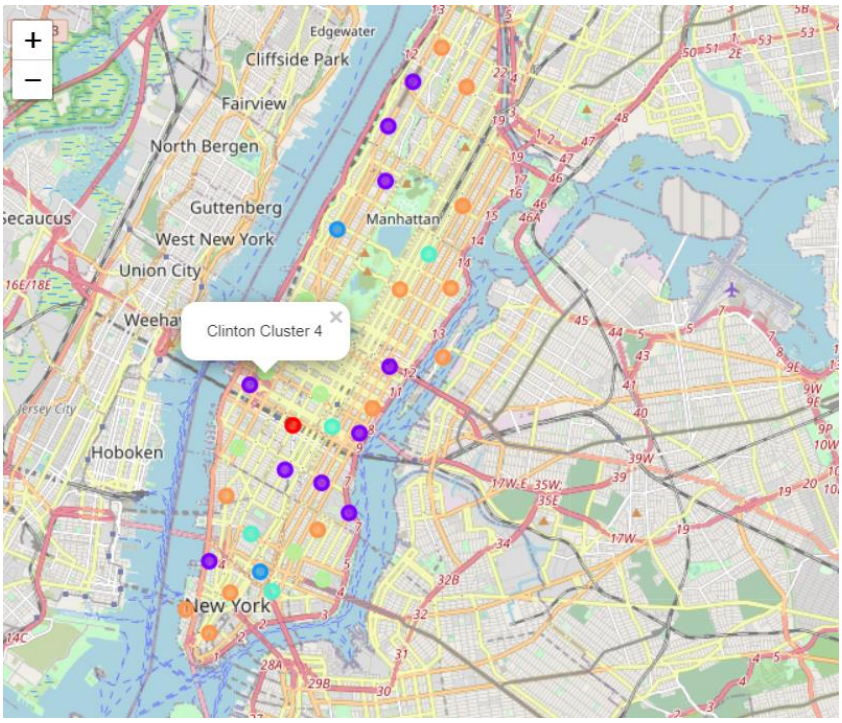
We used Manhattan neighborhood data with foursquare API to get the information of venues and venue categories in each neighborhood. Based on all available venue categories we selected venues categories applicable for this project. Further detail will be explained in the section on Methodology. We identified there were duplicate venues from foursquare and kept the first instance and dropped others.

3. Methodology

In this project we focused on neighborhood that had high venue categories tourists might be interested in, particularly those neighborhoods having similarity based on clustering with high number of attraction venues. We will limit our analysis to Manhattan.

First, we collected the data - New York City visitors, attraction, location, all available 3,142 venues and 331 venue categories using foursquare API. We have also identified venue categories that might be for attraction venues by domain expertise (according to foursquare categorization). It came as 27 venue categories among 331 venue categories listed on Data section. But actual venue categories came as 25 unique categories after the data cleaning with 38 neighborhood in Manhattan. We created the new data and compiled the top 10 venues for each neighborhood.

Secondly, we created clusters using k-means clustering with k=6 to identify general zones and see similarity of neighborhoods based on top 10 venues for each neighborhood within venue categories applicable to this analysis. We tried to find the optimal k-mean and came up k=6 as the best k. We identified cluster 4 as our main focus for further analysis.



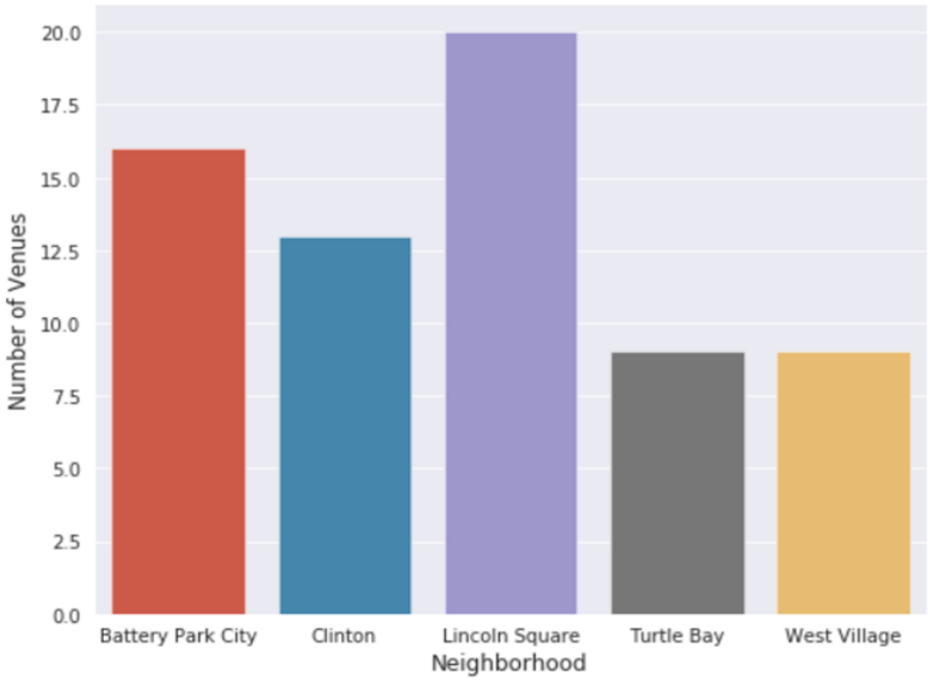
Neighborhood for cluster 4 are listed below:

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
Lincoln Square	Theater	Performing Arts Venue	Concert Hall	Indie Movie Theater	Park
Clinton	Theater	Park	Building	Performing Arts Venue	Indie Theater
Midtown	Theater	Train Station	Concert Hall	Park	Historic Site
Chelsea	Theater	Park	Indie Theater	Train Station	Indie Movie Theater
Lower East Side	Theater	Performing Arts Venue	Park	Train Station	Indie Movie Theater
Noho	Theater	Indie Movie Theater	Music Venue	Train Station	Building

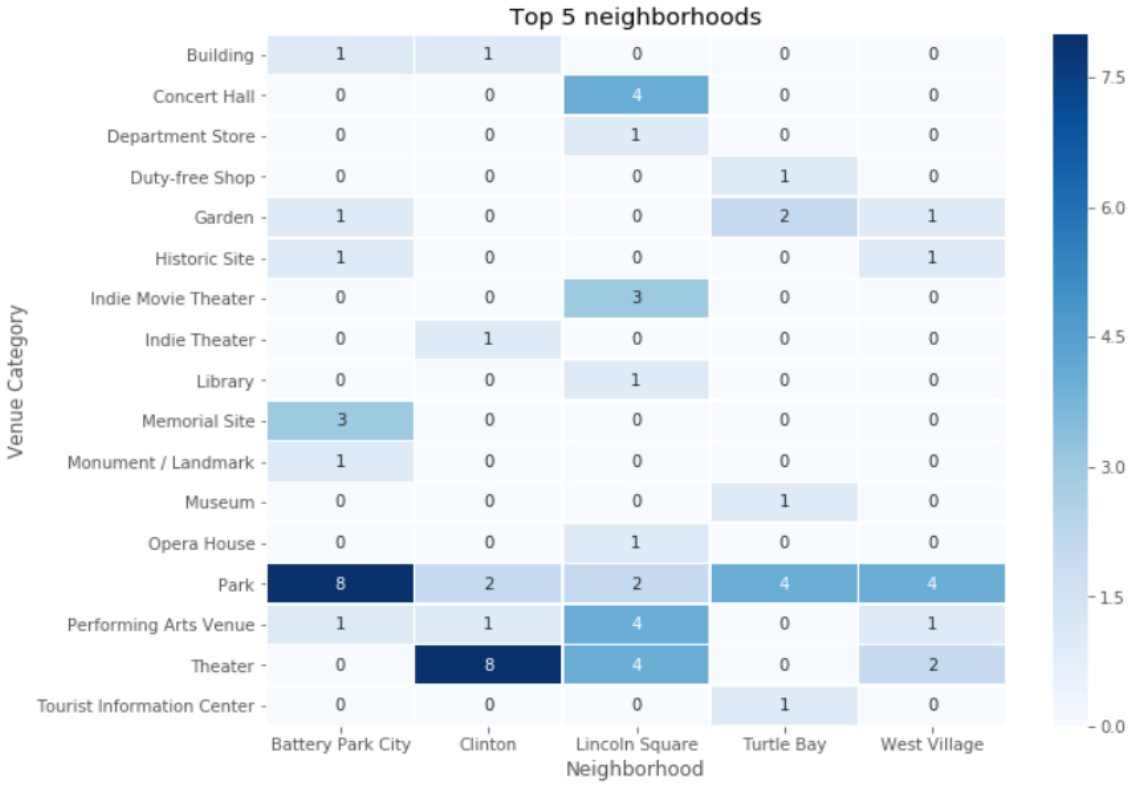
2 neighborhoods – Lincoln Square and Clinton - from cluster 4 were also in top 5 neighborhoods which those neighborhoods have the most venues mapped to tourist place.

Top 5 neighborhoods listed below:

	Neighborhood	Number of Venues
0	Lincoln Square	20
1	Battery Park City	16
2	Clinton	13
3	Turtle Bay	9
4	West Village	9



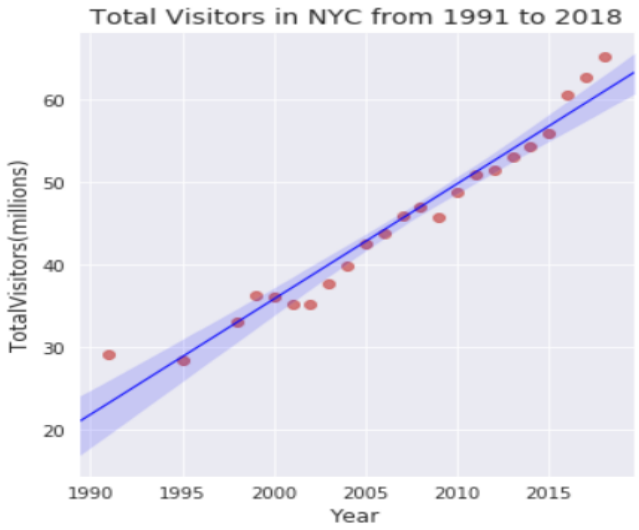
Top 5 neighborhoods with venue categories from foursquare API:



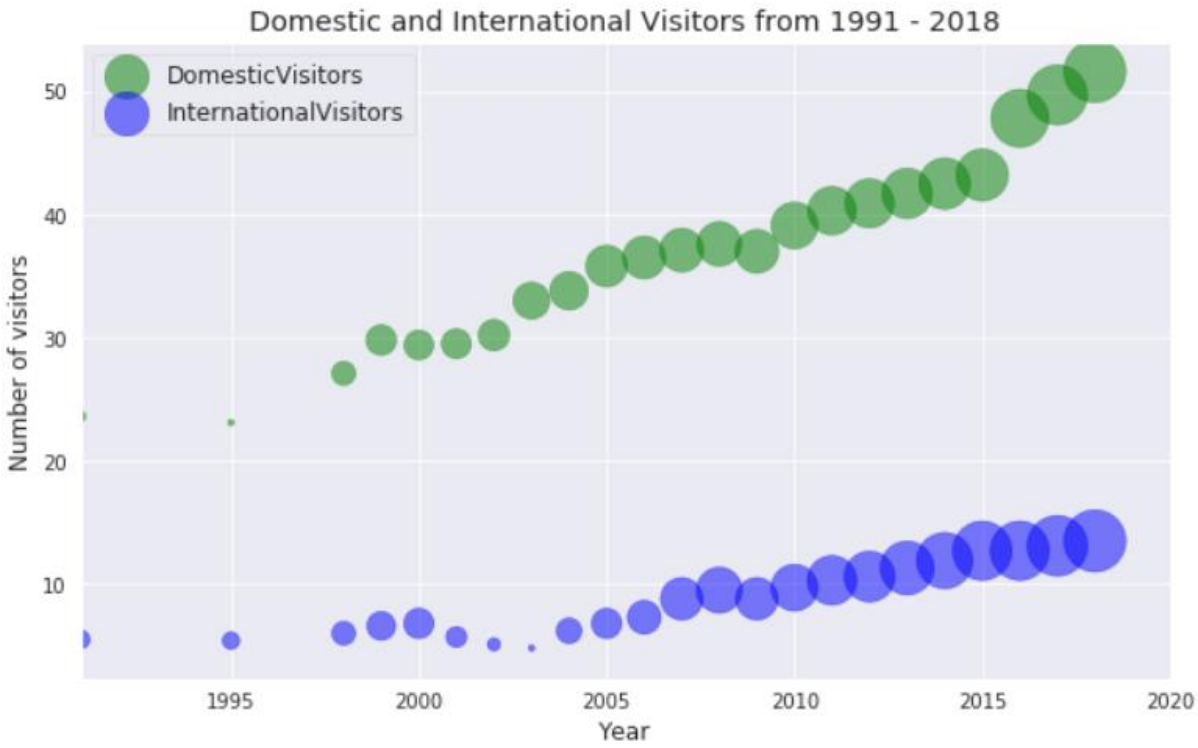
Lincoln Square and Clinton have more venues than other neighbors. Clinton has more theaters since it is located or nearby Theater Districts and Lincoln Square has more Performing Art Venues, Concert hall and theater since it has Lincoln Center complex.

Thirdly, for New York City visitors - domestic and international we conducted exploratory data analysis – distribution and profiling - with visualization using lmplot, regplot, jointplot and scatterplot using seaborn to see trends. This analysis is intended to confirm if it is worth to open the office in the event of pandemic.

Total visitors including domestic and international from 1991 to 2018:

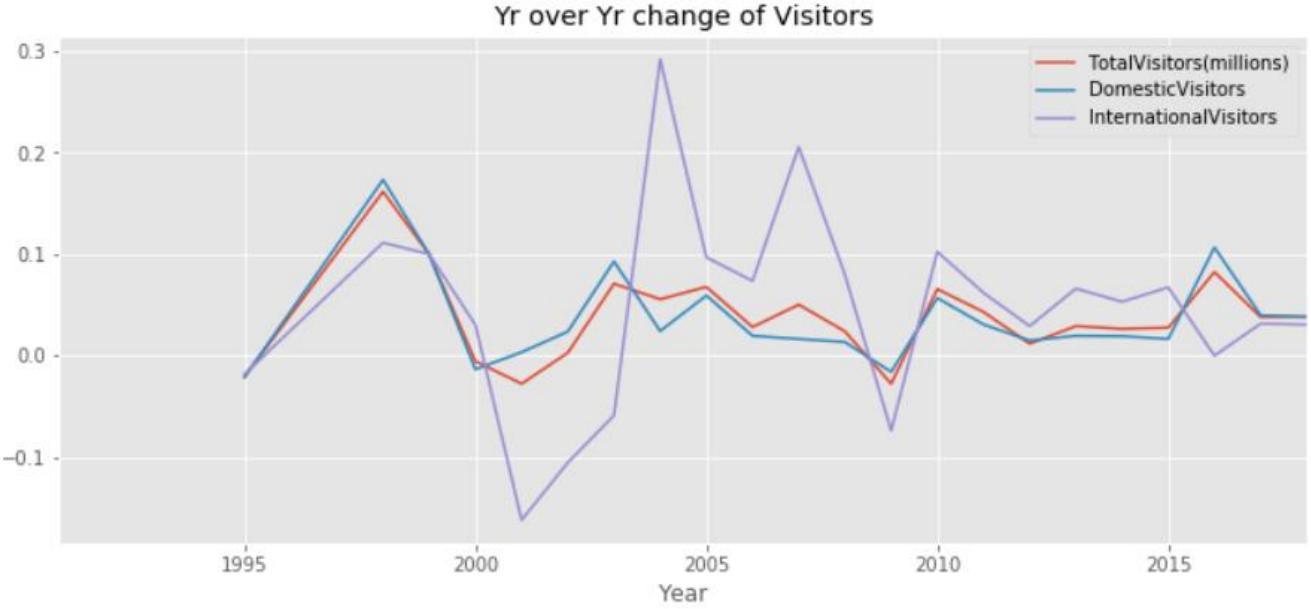


Total Visitors – Domestic and International from 1991 to 2018:



Big events like 9/11 and financial crisis have more impact on international visitors compared to domestic visitors. That's understandable that international travel costs more preparation and expenses.

Year over Year Change on Total Visitors – Domestic and International from 1991 to 2018:



After huge drop due to 9/11 international visitors jumped 29% on 2004 and another jump 20% on 2007 just before financial crisis. Pandemic can be like 9/11 since air travel has major impact due to fear of virus. So we can expect the trend can be recovered from big drop after 2 or 3 years.

Lastly, we applied linear regression to predict on the numbers of visitors from 2019 and forward since the data ends on 2018. We set 20% as test data and 80% as train data and trained the data. Then, we compared the actual values and predicted values to calculate the accuracy of a regression model since evaluation metrics provide a key role in the development of a model and it provides insight to areas that require improvement. R-squared is a popular metric for accuracy of the model. It represents how close the data are to the fitted regression line. The higher the R-squared, the better the model fits the data. Based on evaluation results below, it is pretty good R squared = 0.85.

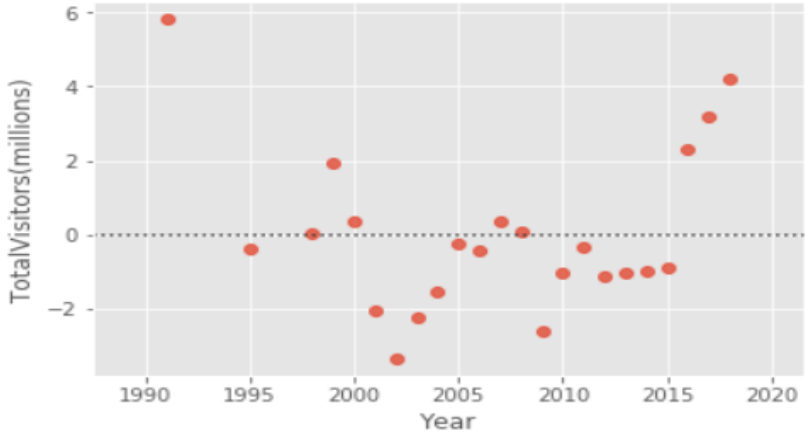
- Coefficients: [[1.33741868]]
- Intercept: [-2638.90219563]
- Mean absolute error: 1.73
- Residual sum of squares (MSE): 5.57
- **R2-score: 0.85**

Total visitors by year from 1991 to 2018 with Train dataset:

Our mean square error between our training and testing was pretty close as listed below.

- MSE with y_train: 4.73
- MSE with X_test and y_test: 5.57

We visualized with residual plot to see any outliers or abnormality.



Right after 9/11 and financial crisis the plots are further away to the line. More plots are randomly distributed around the line. We can conduct the further research the plots that are in further distance.

Based on the model we predicted ~70MM by 2025.

Predicted visitors:

	year	pred_visitors
0	2019	61.346120
1	2020	62.683539
2	2021	64.020957
3	2022	65.358376
4	2023	66.695795
5	2024	68.033213
6	2025	69.370632

4. Results

We examined the venues on foursquare to see how extensive foursquare data can be. However, we couldn't find Museum of Modern Art and Metropolitan Museum when we cross-checked wiki's attraction list into foursquare data. Only 6 attractions on foursquare data were hit from 19 attractions having 2MM visitors.

From top 5 neighborhoods having the most attractions based on foursquare data and 6 neighborhoods on cluster 4 tested similarity/dissimilarity from k-mean clustering method, there were 2 neighborhoods – **Clinton and Lincoln Square** - on both results. It is worth to pursue of opening the tourism office now based on the research explained on Methodology using total visitors trending.

5. Discussion

Our analysis shows that there is a great number of attractions tourists will visit in New York City, but foursquare data is not inclusive of all the attractions. As expected before the analysis, the clustering method provided the neighborhoods - **Lincoln Square and Clinton** neighborhoods as the location candidates.

Clinton, known as Hell's Kitchen, is a neighborhood on west of Midtown Manhattan, by 34th Street to the south, 59th Street to the north, Eighth Avenue to the Hudson River, adjacent to top5 attractions below.

Top 5 attractions:

Name	Visitors_millions
Central Park	42.0
Times Square	39.5
Grand Central Terminal	21.6
Theater District	13.0
Rockfella Center	12.8
Bryant Park	12.0

Also perfect place to find Airbnb with the right price and enjoy bars and clubs at night.

On the contrary, **Lincoln Square** is the home of Lincoln Center, a performing-arts venue with 30 indoor and outdoor performance facilities including:

- Metropolitan Opera House
- David Geffen Hall
- David H. Koch Theater
- Alice Tully Hall
- Vivian Beaumont Theater

Lincoln Center houses nationally and internationally renowned performing arts organizations including:

- The Chamber Music Society of Lincoln Center
- Film Society of Lincoln Center (sponsor of the New York Film Festival)
- Jazz at Lincoln Center
- Juilliard School
- Lincoln Center for the Performing Arts
- Lincoln Center Theater
- Metropolitan Opera
- New York City Ballet
- New York City Opera
- New York Philharmonic
- New York Public Library for the Performing Arts
- School of American Ballet
- Lincoln Center Education

This neighborhood is the surrounding neighborhood centered on the intersection of Broadway and Columbus Avenue, between West 65th and West 66th streets. It is bounded by Hell's Kitchen and Central Park.
Perfect place to experience music concert, opera, performing art and stroll Central Park with Natural history museum. Lincoln Square doesn't have popular venues among top 19 attractions having more than 2MM visitors, however, it has different venues in Lincoln Center complex and more art and cultural tourist destination.

6. Conclusion

This purpose of this project was to find the best location for the office building closer the tourist attractions having higher numbers of attraction venues where tourist might visit in order to help stakeholders make better decision in narrowing down the search for optimal location for the tourism office.

With k-mean clustering method and top 5 neighborhoods having the most attraction venues we have identified neighborhood candidates - Lincoln Square and Clinton. Based on other data analysis for visitors' trends we confirmed this is the right time to open the tourism office in New York City regardless pandemic.

We will use Warren Buffet's quotes, a legendary investor.
'We simply attempt to be fearful when others are greedy and to be greedy only when others are fearful.'

Final decision on optimal location will be made by stakeholders based on specific characteristics of neighborhoods and locations in the two recommended neighborhoods - Clinton and Lincoln Square, taking into consideration additional factors like attractiveness of each location-real estate availability, prices, social and economic dynamics of neighborhood etc.