Opening an Ice Cream Shop in NYC – Battle of the Neighborhoods

IBM Data Science Certificate Capstone Course – Coursera

Elizabeth P.
4/5/20201

Introduction:

According to a recent report by the Office of the New York State Comptroller on the restaurant industry in New York City (NYC), there were over 23,000 food establishments in 2019. The food and drink industry in is a major part of the social lifestyle in the city and nearly any type of food and cuisine from around the world can be found here. However, since the COVID-19 pandemic began in early 2020, the restaurant industry has taken a huge hit and many places have closed down.

As a New Yorker, I thought it would be interesting create a scenario for someone looking to open up new eatery in the city. This report will provide information to an ambitious person(s) looking to open up a new ice cream shop when the city starts back up. In this hypothetical scenario, the new shop owner(s) is/are looking for good areas in the borough of Manhattan to open up a shop.

The report will look at the areas with high and low number of competing shops, such as cafes, bakeries, and other types of dessert shops. It will also look at the proximity of potential areas to parks and plazas, as these are high pedestrian traffic, which can expose the shop to more customers. The report will provide a starting guide of potential locations to open a shop, before looking into rental and leasing prices, as well as other factors that influence the decision.

Data:

Manhattan is a large borough and contains many unique neighborhoods. To get a better understanding of the different regions in Manhattan, I used a json file that contains the geographical coordinates of the approximate centers of each neighborhood in the borough. The json file was provided in another assignment in the course. A dataframe containing the 'Borough', 'Neighborhood', 'Latitude', and 'Longitude was created for neighborhoods only found in Manhattan.

The next step was finding the parks and plazas in Manhattan. I used the Foursquare API to retreive a list of all the outdoor spaces near each neighborhood center. The resulting data was put into a dataframe containing information about the venue id, venue name, venue latitude,

venue longitude, venue category, neighborhood, neighborhood latitude, and neighborhood longitude. The dataframe was check for null values and none existed. The dataframe was also checked for duplicates using the venue id because the radius used to search for venues in each neighborhood overlaps in the area covered with nearby neighborhoods. Many duplicates were found and removed. The resulting dataframe contained over 1000 outdoor spaces and 91 unique categories. However, many of these categories didn't exactly fit the type of outdoor venues I was looking for. Since the shop owner(s) wanted the ice cream shop to be close to parks and plazas and other areas were people city and gather, I reduced the type of outdoor category to contain only parks, plazas, and pedestrian plazas.

The Foursquare API returned many results even just for parks, so I decided to focus only on major parks because these are the most popular areas (with both locals and tourists) and would they generate the most exposure for our ice cream shop. To get an idea of the major parks in Manhattan, I scraped the New York City Department of Parks and Recreation website to get featured parks in Manhattan. The results were placed into a list and used to check for matches with the Foursquare results. The problem I encountered was that some of the parks were named differently in both the datasets. I then compared them for the differences and made a list of the ones that appeared in the Foursquare dataset, but were named differently in the NYC Parks and Rec. featured parks list. In addition, after looking through the Foursquare dataset, I added several more locations. The larger parks, such as the Hudson River Park and Central Park, cover a lot of area. To make sure different regions along the park were covered, I added the venues that covered different regions along the parks. I also added 'Oculus Plaza', 'Waterfront Plaza, Brookfield Place' because these are also high foot traffic places.

The final dataset for all the parks in Manhattan used for this report contains 40 parks and plazas. It includes the parks that were included in both the Foursquare datasets and featured parks list, as well as the added park venues for the larger parks, pedestrian plazas from the Foursquare results, and a couple of popular plazas.

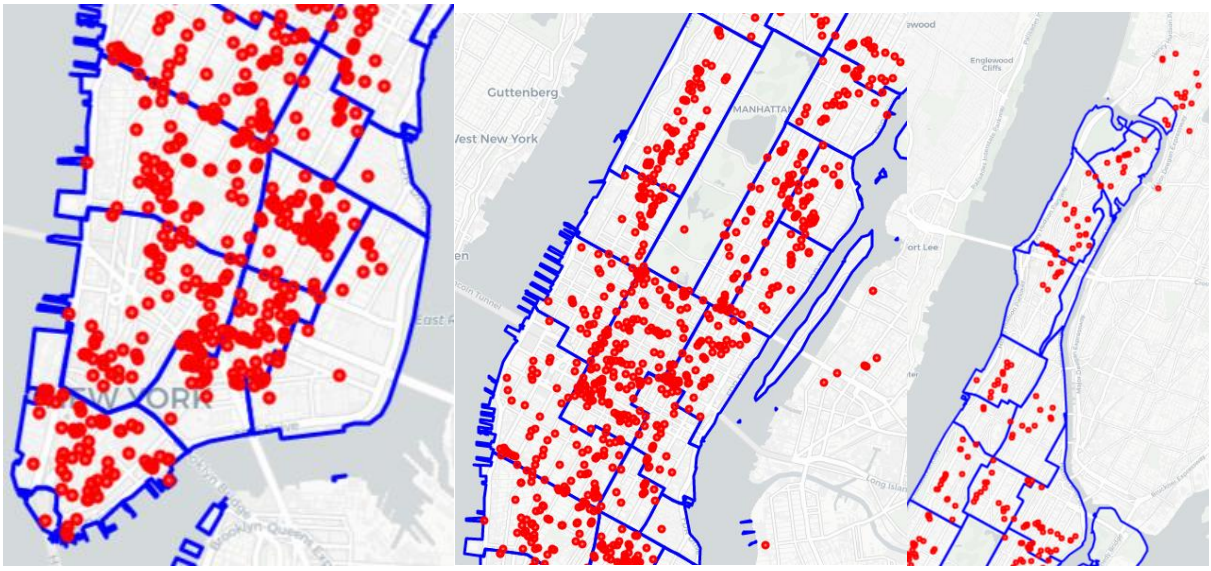
NYC Department of Transportation (DOT) created a program to turn underused streets into neighborhood plazas across the city for people to enjoy outdoor spaces. The location of these plazas can be found on the NYC Open Data Site. The geojson fie for the pedestrian plazas was imported and used. I created a geodataframe of the plazas and found the coordinates of the center of each plaza. There are a total of 31 pedestrian plazas created by the DOT.

To find competing venues for an ice cream shop, the Foursquare API was used to find nearby venues for each neighborhood. Similar to the Foursquare results for outdoor spaces, many results were returned. The dataset was checked for nulls and the duplicates were removed. Looking at the types of unique categories, many of them didn't fit the type of venue that would compete with an ice cream shop. I limited the competition to only include relevant competition, such as other ice cream shops, bakeries, dessert shops, etc. The final dataset of all the competition in Manhattan resulted in 1091 competing venues.

Information about the neighborhood boundaries was also found on NYC Open Data site. There is a geojson file provided by the NYC Department of City Planning for the Neighborhood Tabulation Areas, which gives the coordinates for the boundaries of each neighborhood.

Methodology:

To find ideal locations to open an ice cream shop, I began by plotting all the competition on a map using Folium. An initial look at the map shows there is a lot of competition.



(Left to Right: Downtown; Midtown, Upper East Side, Upper West Side; North of Central Park (Harlem, Washington Heights, etc.))

With individual markers it's a little difficult to see which areas have a lower number of competitors. A heatmap of the competition with the NTA boundaries made it easier to determine where the lower density areas are.

(TOP Left to Right: Downtown; Midtown| BOTTOM Left to Right: Upper East Side, Upper West Side; North of Central Park (Harlem, Washington Heights, etc.))

After zooming in closer on individual neighborhoods, pockets of low density can be found. Next, I narrowed down which neighborhoods seemed like the best choice based on both neighborhoods that are popular/less residential and those that have more areas with a low density of competition. This resulted in choosing the following NTAs:

- 'Battery Park City-Lower Manhattan'
- 'Hudson Yards-Chelsea-Flatiron-Union Square'
- 'West Village', 'SoHo-TriBeCa-Civic Center-Little Italy'
- 'Central Harlem South'
- 'Central Harlem North-Polo Grounds'

These neighborhood boundaries included the following:

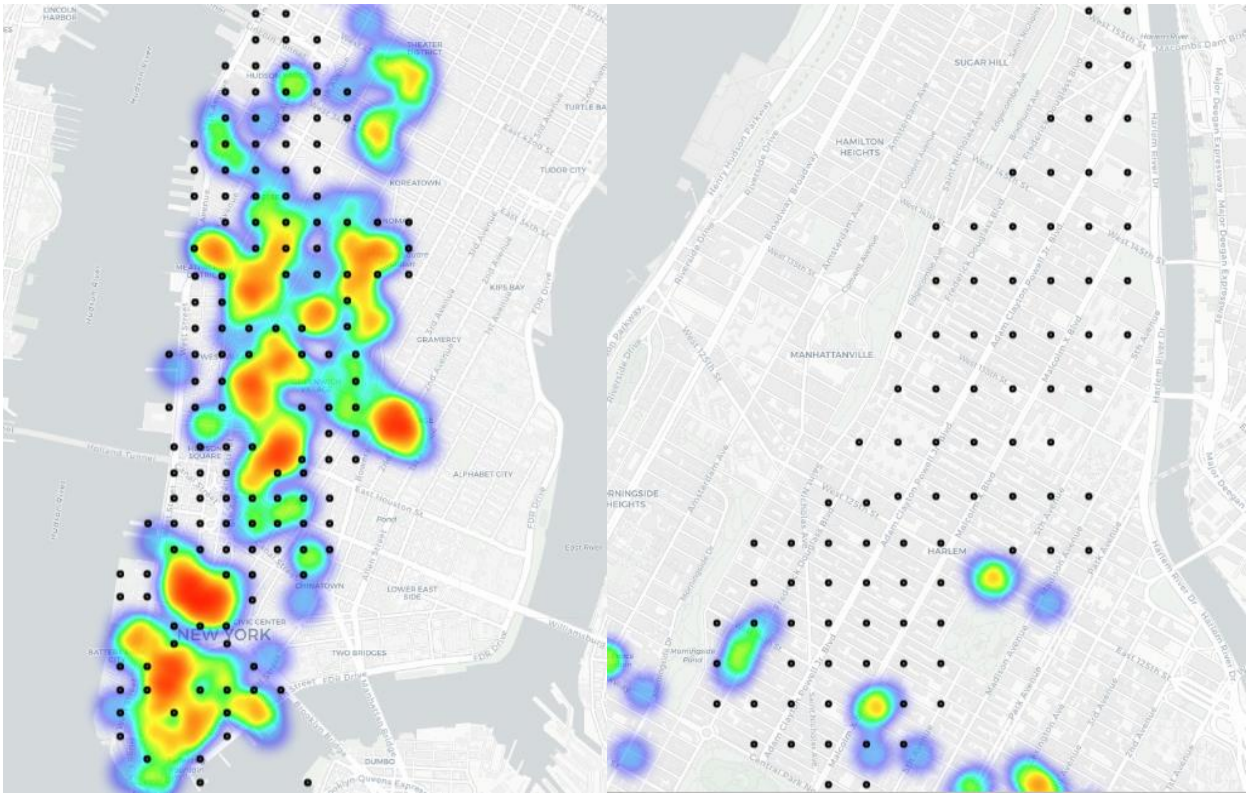
- 'Battery Park City'
- 'Financial District'
- 'Civic Center'
- 'Tribeca', 'Little Italy'
- 'Soho'
- 'Greenwich Village'
- 'West Village'
- Flatiron'
- 'Chelsea'
- 'Hudson Yards'
- 'Central Harlem'

'Manhattan Valley' and ‘East Harlem' were included so venues just north of Central Park could be counted.

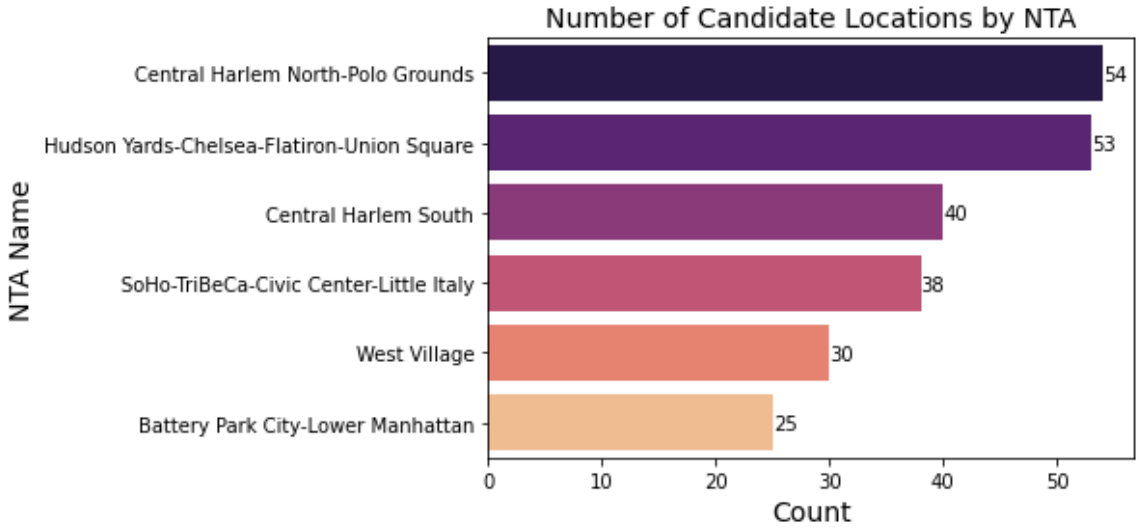
To find out which areas of each neighborhood had a low density, I made a grid of evenly spaced latitude-longitude dots for the chosen NTAs. These were mapped on the heatmap of the competition. I decided that a ‘low competition’ area would be classified as a latitude-longitude dot that had fewer than 3 competing venues within 150 meters. This distance of each dot to its closest park or plaza, in meters, was also calculated. After, I used K-means clustering to see if the locations could be grouped into further categories. The resulting cluster labels were assigned to each dot and each cluster was analyzed.

Results:

Location dots that did not meet the criteria were eliminated as potential candidate locations, this resulted in around 240 possible locations.

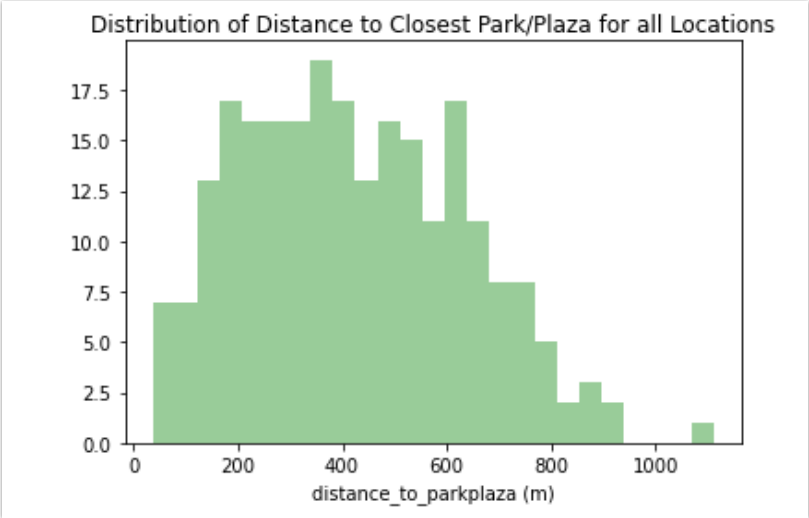


After breaking these dots down by NTA, we can see which neighborhoods contain the most candidate locations.

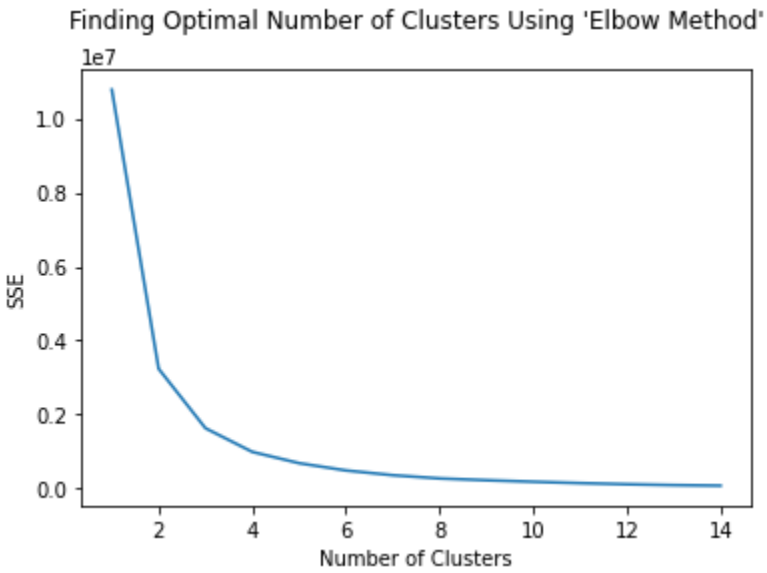


As the graph above shows, most of these locations are in 'Central Harlem North-Polo Grounds' and 'Hudson Yards-Chelsea-Flatiron-Union Square.' It is important to note that some of the locations may represent residential areas, so a closer look at the chosen neighborhoods must be taken to identify, which areas are more residential than others.

After calculating the distance of each dot to its closest park or plaza, the range of distance between location and park/plaza was between 37.0693 meters and 1112.0673 meters. The median distance is 403.6970 meters from a park or plaza.

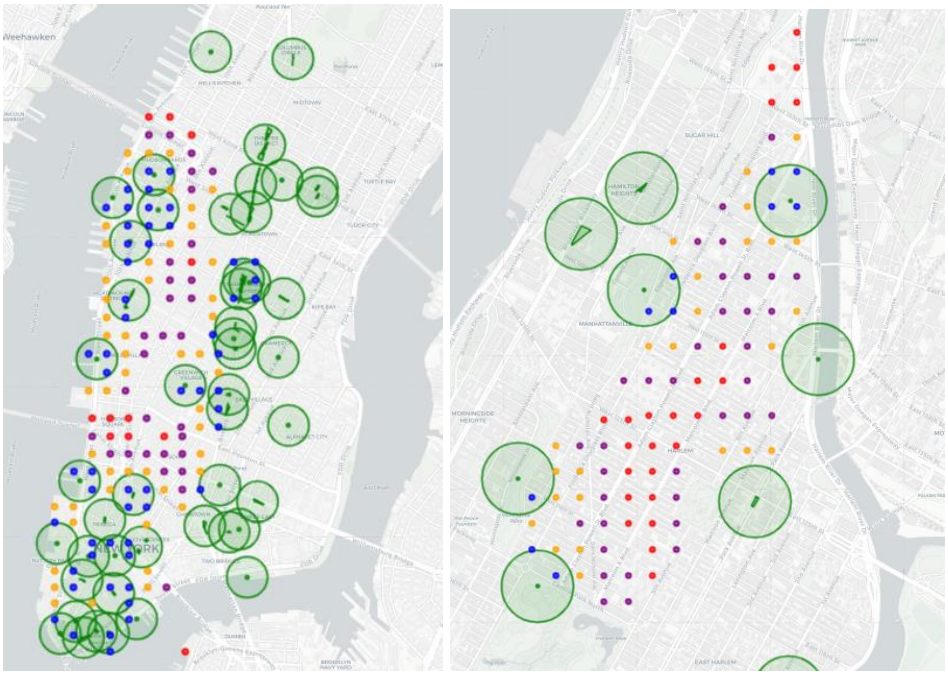


To see if the locations could be grouped further, I used the 'elbow method' to determine the ideal number of clusters.



From the plot, it seems like the inertia begins to decrease approximately linear at 4 clusters. Therefore, I chose 4 as the optimal number of clusters and fitted the model to the dataset using 4 clusters.

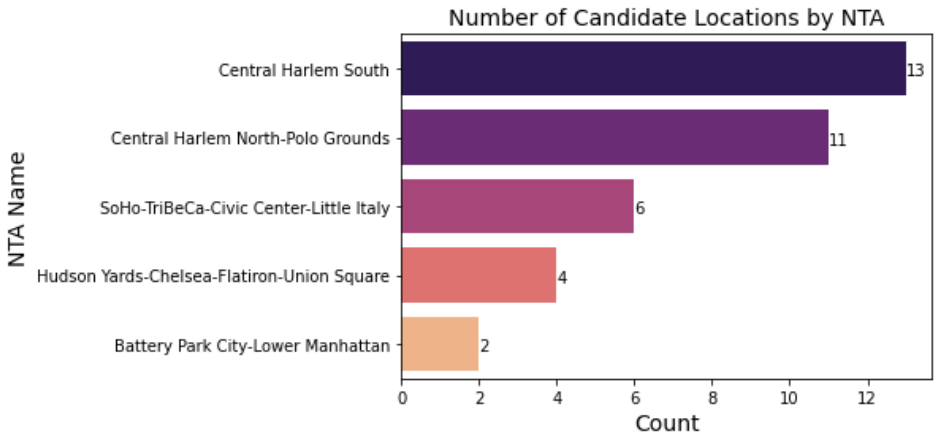
After assigning the cluster labs to each candidate location, I mapped the locations again on map and colored them according the cluster label. The map also shows a 250 meter radius around each park or plaza.



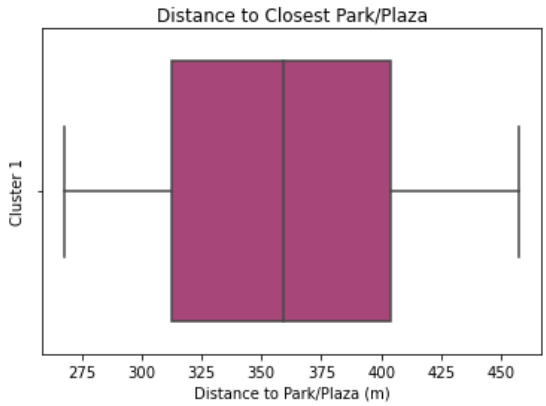
Looking at the maps above, the locations clustered based on distance to a park or plaza. The closest locations to a park are blue, then yellow, then purple, then red (farthest).

Cluster 0 contains candidate locations that are the furthest away from a park or plaza. It has 36 locations, the closest park being around 668 meters and the furthest being approximately 1112 meters. The median distance from a park/plaza for this cluster was around 739 meters.

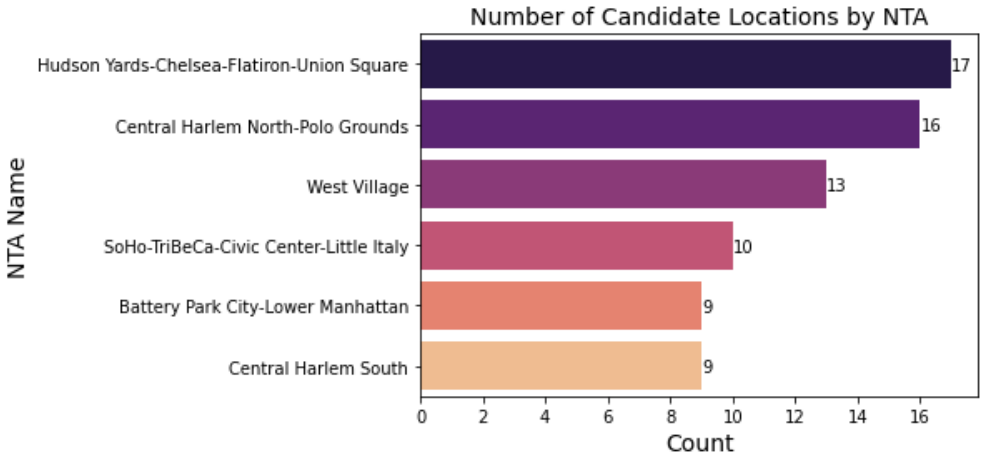
The most candidate locations in this cluster were located in both the Harlem neighborhoods.



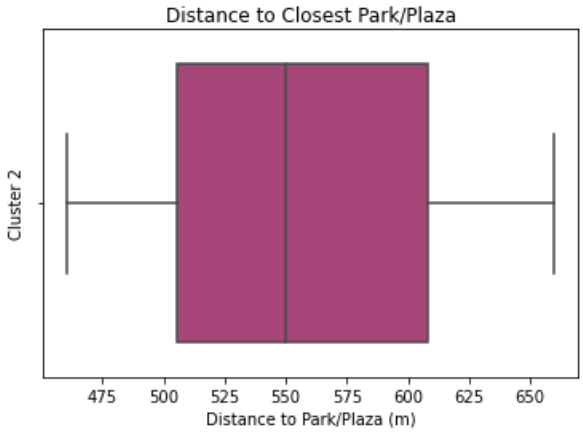
Cluster 1 contains candidate locations that are the second closest to a park or plaza. It has 74 locations, the closest park being around 268 meters and the furthest being approximately 457 meters. The median distance from a park/plaza for this cluster was around 359 meters.



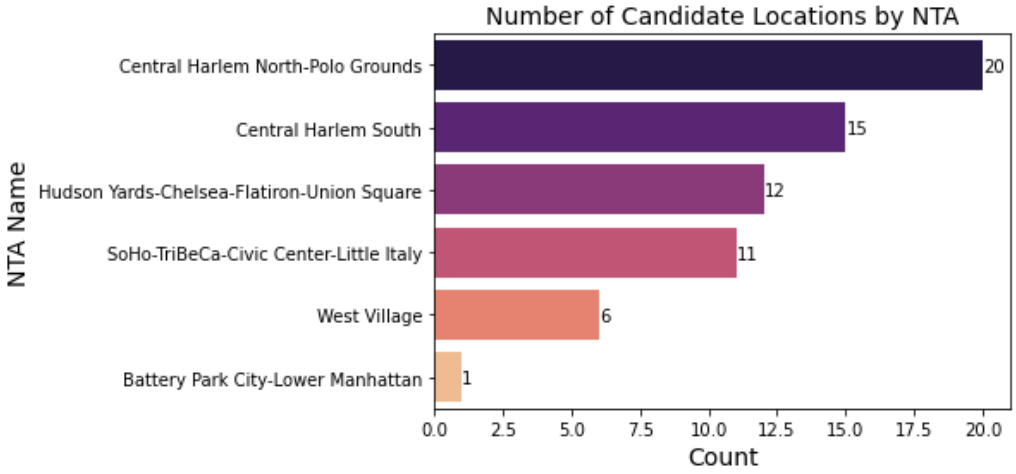
This clusters contained the most candidates from ‘Hudson Yards – Chelsea – Flatiron – Union Square’.



Cluster 2 contains candidate locations that are the second farthest to a park or plaza. It has 65 locations, the closest park being around 460 meters and the furthest being approximately 660 meters. The median distance from a park/plaza for this cluster was around 550 meters.

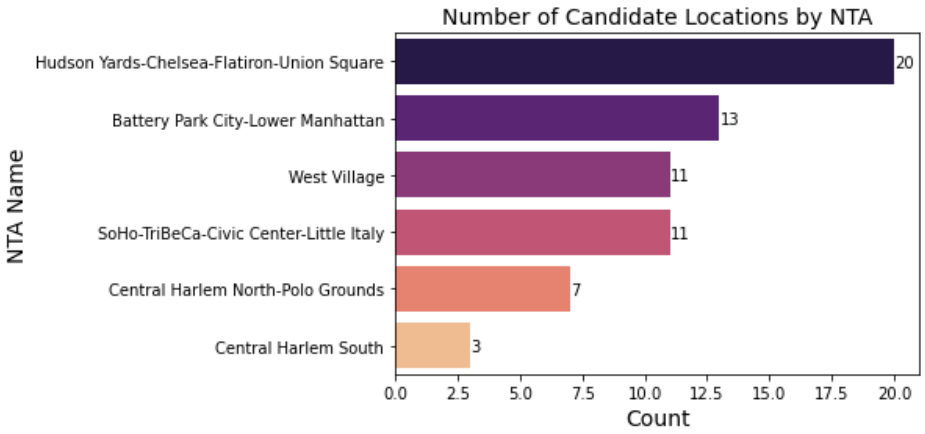


Again, the most candidate location are from both the Harlem neighborhoods.



Finally, cluster 3 contained locations that were the closest to a park or plaza. It also has 65 locations, the closest park being around 37 meters and the furthest being approximately 262 meters. The median distance from a park/plaza for this cluster was around 178 meters.

Again, the neighborhood with the most locations is ‘Hudson Yards – Chelsea – Flatiron – Union Square’.



Discussion:

From the results, it can be seen that the most locations are in the Harlem neighborhoods, however, these locations tend to be the further from a major park or plaza. From the heatmap and competition map, we can also see that the Harlem neighborhoods also have lower competition.

The downtown neighborhoods are closer to major parks and plazas, but have fewer candidate locations and more competition. Out of the neighborhoods that are downtown, ‘Hudson Yards – Chelsea – Flatiron – Union Square’ seems to have the most candidate locations in the two closest clusters to parks and plazas (Cluster 3 and Cluster 1).

Conclusion:

For an initial guide for picking a location to open a new ice cream shop, this report some candidate locations to begin exploration. There are candidate location in each of the neighborhoods that are both close to parks/plazas and have a low number of nearby competitors. Out of the potential neighborhoods that were chosen, it seems like the ‘Hudson Yards – Chelsea – Flatiron – Union Square’ NTA would be the best place to begin exploring. These neighborhoods have the most candidate locations and are in clusters that are the closer to major parks and plazas. Although the Harlem candidates tend to be further from major parks, they may be close to other venues of interest. In the more residential locations, being close to playgrounds, smaller public parks, or other high pedestrian traffic areas may help generate more exposure. Further exploration into characteristics of these locations and other types of venues should be included in the process. As always, other factors outside the scope of this report needs to be considered, such as rental/leasing prices and availability, proximity to public transport, etc.

References:

- The Restaurant Industry in New York City: Tracking the Recovery: https://www.osc.state.ny.us/files/reports/osdc/pdf/nyc-restaurant-industry-final.pdf

- List of New York City Parks: https://www.nycgovparks.org/park-features/parks-list?boro=M

- NYC DOT Pedestrian Plazas: https://data.cityofnewyork.us/Transportation/NYC-DOT-Pedestrian-Plazas/k5k6-6jex

- Neighborhood Tabulation Areas (NTA): https://data.cityofnewyork.us/City-Government/Neighborhood-Tabulation-Areas-NTA-/cpf4-rkhq

- Foursquare API: https://developer.foursquare.com/docs/api-reference/venues/search/