Capstone Project - The Battle of the Neighborhoods

Applied Data Science Capstone by IBM/Coursera

Yu Liu

April 7, 2019

Table of Contents

Introduction

In this project I will try to find an optimal office location for a Hotel Tech Startup. Specifically, this report will be targeted to stakeholders interested in choosing a location in Manhattan, New York.

I will try to detect the location that surrounded with Subway Stations. I are also particularly interested in areas with Chinese restaurants since most of the employees are Asians. I would also prefer locations as close to hotels, assuming that first two conditions are met.

I will use our data science power to generate a few most promising office locations based on this criterion. Advantages of each area will then be clearly expressed so that best possible final location can be chosen by stakeholders.

Data

Based on definition of our problem, factors that will influence our decision are:

- number of potential office locations in Manhattan
- number of existing Subway Stations in the neighborhood (particularly ACE and NWQR)
- number of and distance to Chinese Restaurants in the neighborhood
- distance of neighborhood from Hotels groups

I decided to use regularly spaced grid of locations, centered around city center, to define our neighborhoods.

Following data sources will be needed to extract/generate the required information:

- Manhattan geo location will be generated by geopy library
- number of restaurants and their type and location in every neighborhood will be obtained using Foursquare API
- coordinate of Subway Station will be obtained using NYC OpenData

Methodology

In this project I will direct our efforts on detecting office spaces of Manhattan that have blue and yellow line and Chinese restaurants. I will limit our analysis to Chinese restaurants around each office location.

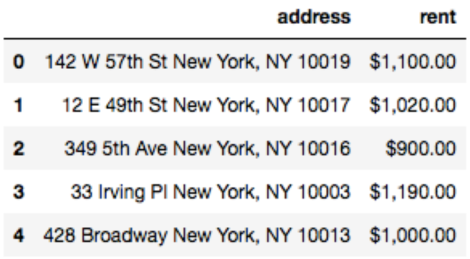
In first step I have collected the required data: location and Chinese restaurants of each office space. I have also identified Blue and Yellow subway line.

Second step in our analysis will be calculation and exploration of 'restaurant density' across different office locations of Manhattan - I will use heatmaps to identify a few promising areas close to center with low number of restaurants in general and focus our attention on those areas.

In third and final step I will focus on most promising areas and within those create clusters of locations that meet some basic requirements established in discussion with stakeholders: I will take into consideration locations that near the subway stations, and I want locations with Chinese restaurants nearby. I will present map of all such locations but also create clusters (using k-means clustering) of those locations to identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

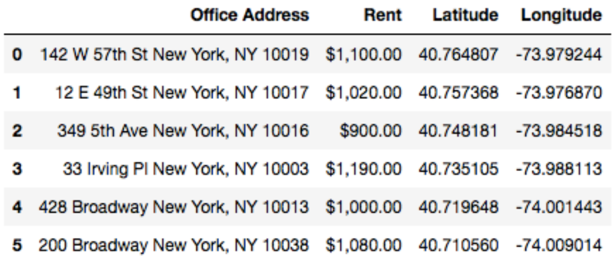
Analysis

For the office location, it only has two columns: address and rent.

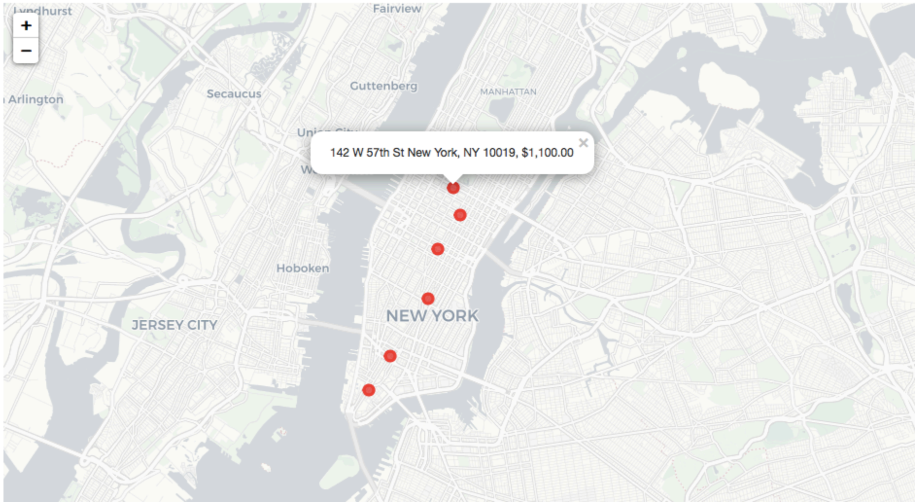


	address	rent
0	142 W 57th St New York, NY 10019	$1,100.00
1	12 E 49th St New York, NY 10017	$1,020.00
2	349 5th Ave New York, NY 10016	$900.00
3	33 Irving Pl New York, NY 10003	$1,190.00
4	428 Broadway New York, NY 10013	$1,000.00

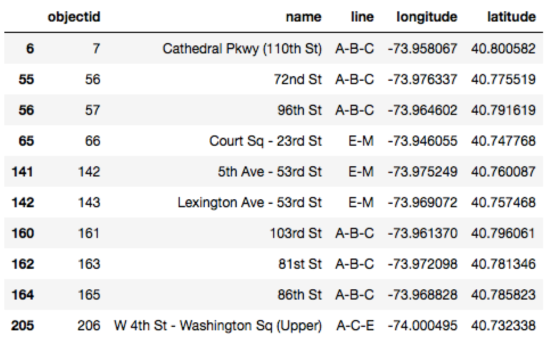
I first used geopy library to convert the address to Latitude and Longitude and the used python folium library to visualize the geographic location of offices with latitude and longitude on the map of Manhattan, New York.



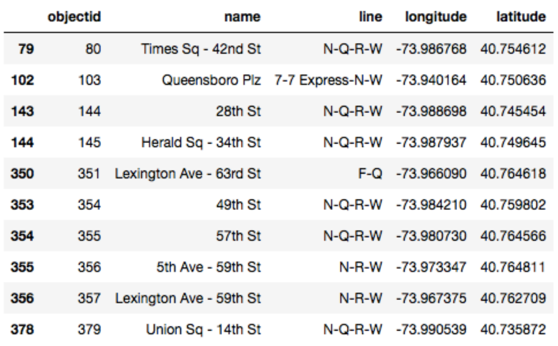
	Office Address	Rent	Latitude	Longitude
0	142 W 57th St New York, NY 10019	$1,100.00	40.764807	-73.979244
1	12 E 49th St New York, NY 10017	$1,020.00	40.757368	-73.976870
2	349 5th Ave New York, NY 10016	$900.00	40.748181	-73.984518
3	33 Irving Pl New York, NY 10003	$1,190.00	40.735105	-73.988113
4	428 Broadway New York, NY 10013	$1,000.00	40.719648	-74.001443
5	200 Broadway New York, NY 10038	$1,080.00	40.710560	-74.009014



I collected the all the New York subway station location data from NYC OpenData, cleaned up and removed irrelevant stations that are not ACE or NQRW

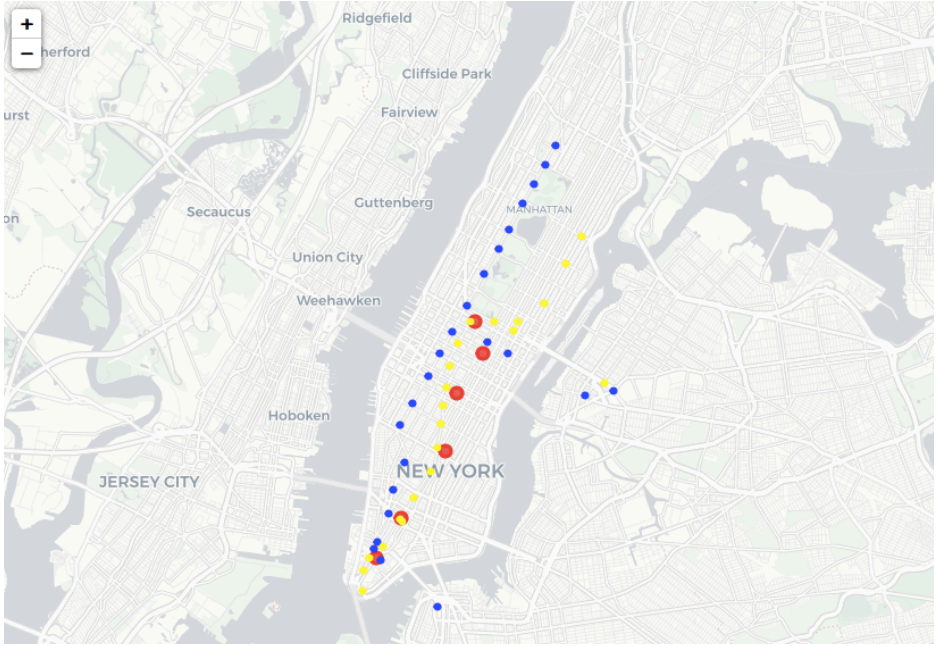


	objectid	name	line	longitude	latitude
6	7	Cathedral Pkwy (110th St)	A-B-C	-73.958067	40.800582
55	56	72nd St	A-B-C	-73.976337	40.775519
56	57	96th St	A-B-C	-73.964602	40.791619
65	66	Court Sq - 23rd St	E-M	-73.946055	40.747768
141	142	5th Ave - 53rd St	E-M	-73.975249	40.760087
142	143	Lexington Ave - 53rd St	E-M	-73.969072	40.757468
160	161	103rd St	A-B-C	-73.961370	40.796061
162	163	81st St	A-B-C	-73.972098	40.781346
164	165	86th St	A-B-C	-73.968828	40.785823
205	206	W 4th St - Washington Sq (Upper)	A-C-E	-74.000495	40.732338



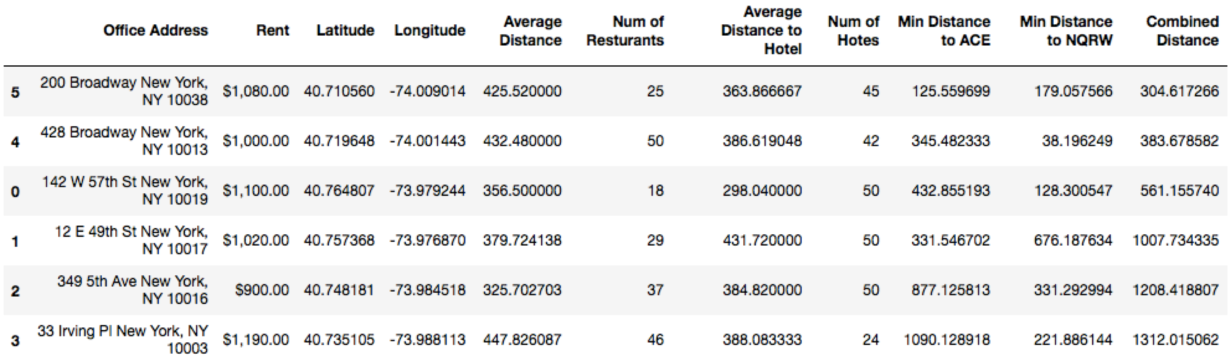
	objectid	name	line	longitude	latitude
79	80	Times Sq - 42nd St	N-Q-R-W	-73.986768	40.754612
102	103	Queensboro Plz	7-7 Express-N-W	-73.940164	40.750636
143	144	28th St	N-Q-R-W	-73.988698	40.745454
144	145	Herald Sq - 34th St	N-Q-R-W	-73.987937	40.749645
350	351	Lexington Ave - 63rd St	F-Q	-73.966090	40.764618
353	354	49th St	N-Q-R-W	-73.984210	40.759802
354	355	57th St	N-Q-R-W	-73.980730	40.764566
355	356	5th Ave - 59th St	N-R-W	-73.973347	40.764811
356	357	Lexington Ave - 59th St	N-R-W	-73.967375	40.762709
378	379	Union Sq - 14th St	N-Q-R-W	-73.990539	40.735872

I plot the stations in the map with the office location. Blue marker stands for ACE and yellow marker stands for NQRW.



I utilized the Foursquare API to explore the Chinese restaurants and hotel venues that near the office locations. I designed the limit as 100 venue and the radius 1000 meter for office from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API.

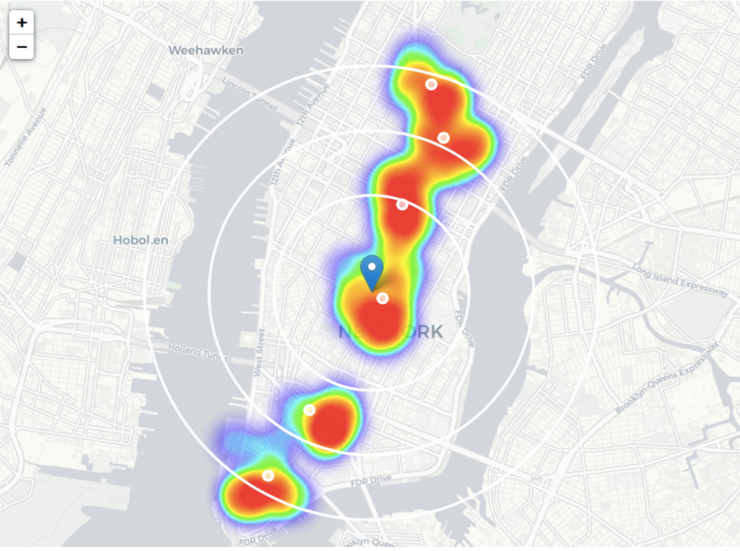
I then calculate the total number of restaurants and hotels, average distance, minimum distance to the nearest ACE and NQRW stations for each office locations.



	Office Address	Rent	Latitude	Longitude	Average Distance	Num of Resturants	Average Distance to Hotel	Num of Hotes	Min Distance to ACE	Min Distance to NQRW	Combined Distance
5	200 Broadway New York, NY 10038	$1,080.00	40.710560	-74.009014	425.520000	25	363.866667	45	125.559699	179.057566	304.617266
4	428 Broadway New York, NY 10013	$1,000.00	40.719648	-74.001443	432.480000	50	386.619048	42	345.482333	38.196249	383.678582
0	142 W 57th St New York, NY 10019	$1,100.00	40.764807	-73.979244	356.500000	18	298.040000	50	432.855193	128.300547	561.155740
1	12 E 49th St New York, NY 10017	$1,020.00	40.757368	-73.976870	379.724138	29	431.720000	50	331.546702	676.187634	1007.734335
2	349 5th Ave New York, NY 10016	$900.00	40.748181	-73.984518	325.702703	37	384.820000	50	877.125813	331.292994	1208.418807
3	33 Irving Pl New York, NY 10003	$1,190.00	40.735105	-73.988113	447.826087	46	388.083333	24	1090.128918	221.886144	1312.015062

I now can see the average distance and numbers of Chinese restaurants and hotels and also the minimum and combined distance to two subway lines. Offices on **200 Broadway,428 Broadway and 142 W 57th St** have a close distance to both subway lines.

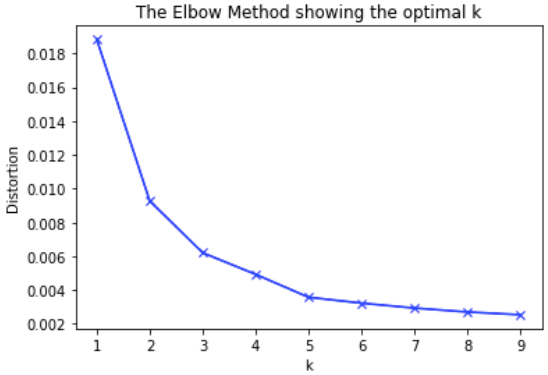
I then use the heatmap to plot the Chinese restaurant and add office locations as markers in the map.



I can see from the map that the almost all the offices are in the hot zone but only offices in **349 5th Ave New York, 33 Irving Pl New York** are in the middle of the zone.

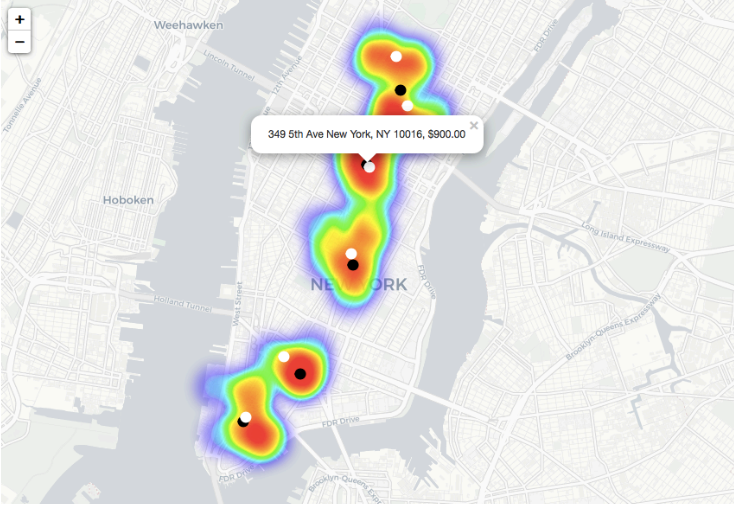
I want to show the center of those restaurant to evaluate the office location regards to the distance to Chinese restaurant. I used unsupervised learning **K-means algorithm** to cluster the boroughs. K-Means algorithm is one of the most common cluster methods of unsupervised learning.

First, I have to determine the best number of clusters using the **Elbow method**.



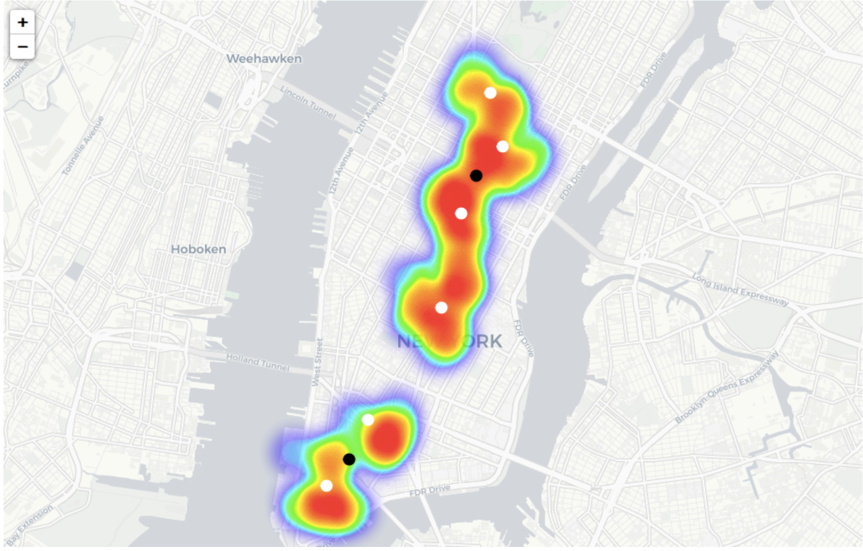
When K increases, the centroids are closer to the cluster's centroids. The improvements will decline, at some point rapidly, creating the elbow shape. That point is the optimal value for K. In the image above, K is range from 3 to 5.

I choose to use 5 clusters and plot the center of the cluster compared with office locations on a heat map.



I can see from the map that office **349 5th Ave, 33 Irving Pl and 200 Broadway** have a relatively perfect match.

I did same process with hotels for two clusters.



I can see from the heat map that there are two large hotel clusters and office in **349 5th Ave, 12 E 49th St, 428 Broadway and 200 Broadway** are relatively close to the location.

Results and Discussion

The analysis shows the result that there are number of Chinese restaurants and hotels near all six office locations and distance to both ACE and NQWR subway line varies.

From the heatmap of **Chinese Restaurant**, I can see that office on **349 5th Ave, 33 Irving Pl and 200 Broadway** have a relatively perfect location.

From the distance of **Subway Lines**, I can see that Offices on **200 Broadway,428 Broadway and 142 W 57th St** have a close distance to both subway lines.

From the heatmap of **Hotels**, I can see that office on **349 5th Ave, 12 E 49th St, 428 Broadway and 200 Broadway** are relatively close to the cluster center.

Conclusion

Based on the priority of the requirements from stakeholder, office on **200 Broadway New York, NY 10038** with a monthly rent of 1,080 dollars would be a great location for this startup company since it has the best location for Chinese restaurant, and relatively close distance to the required subway stations and in the middle of two hotel groups.

The second recommended location would be **349 5th Ave New York, NY 10016** with a monthly rent of 900 dollars and closer location to the midtown.

Final decision on optimal office location will be made by stakeholders based on specific characteristics of neighborhoods and locations in additional factors.