Manhattan Coffee Shops Venues Data Analysis and Optimization

A. Introduction

Manhattan is a densely populated borough of New York City that's among the world's major commercial, financial and cultural centers. A client wants to open a coffee shop at Manhattan and would like to determine the location that is able to generate maximum revenue. As I have lived and worked at Manhattan for many years, I would like to give the client some advice. Clearly there are lots of things that could impact the decision, but this report will mainly focus on two factors, the number of coffee shops and total amount of residents within certain range of areas. More people could indicate more demands, while more coffee shops means more competition, which could have negative impact on revenue.

B. Data Description

1) Firstly get the latitude and longitude of Manhattan and each of the neighborhoods. The Json file has coordinates of the all boroughs of New York City. I cleaned the data and reduced it to borough of Manhattan and a few other useful fields, including latitude, longitude, etc.
2) Pull all the venues information using **Foursquare API**.
3) Filter by coffee shops and calculate the total amount of coffee shops in each neighborhood.

4) Scrape total population by neighborhood at Manhattan from website https://www.worldatlas.com/articles/manhattan-neighborhoods-by-population.html. Calculate the coffee shops per capita.

5) Other support evidence from https://www.6sqft.com/what-nycs-population-looks-like-day-vs-night/

C. Methodology

To start the analysis, we need to pull existing coffee shops in each neighborhood and

population data from some database. After data cleaning, we get the Manhattan neighborhoods location data.

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

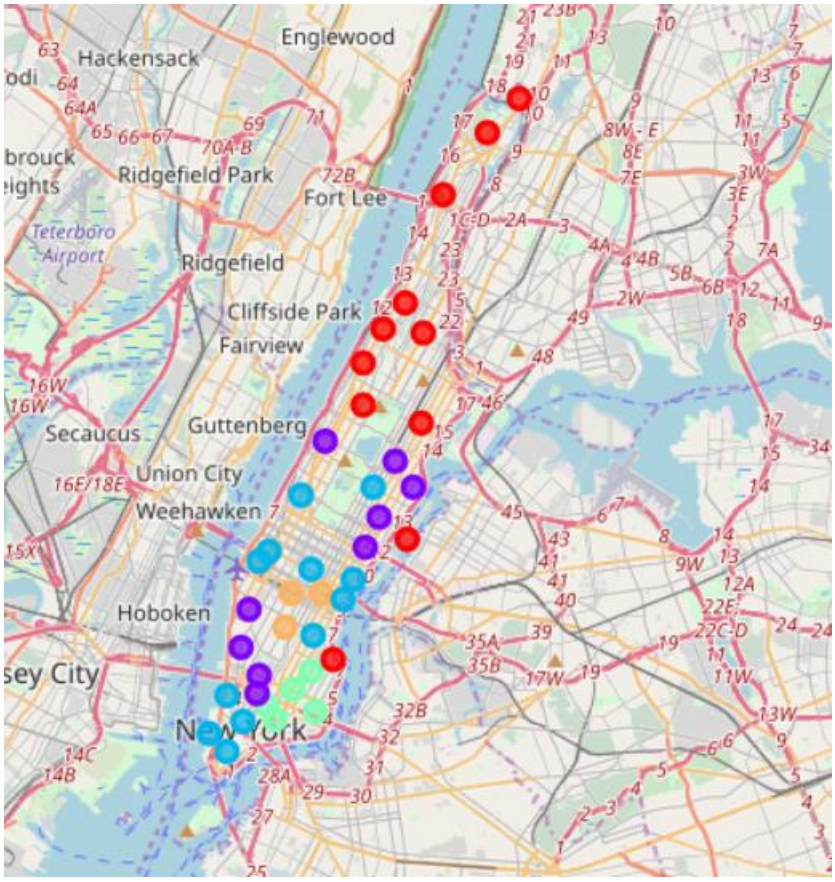
More data is pulled from Foursquare API with **all** the venues in each neighborhood at Manhattan with venue info and location. I designed the limit as 2000 venue and the radius 1000 meter for each borough from their given latitude and longitude information. Below is the snapshot of the data. There are around 4000 venues at Manhattan.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
1	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Sam's Pizza	40.879435	-73.905859	Pizza Place
4	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop

Since people move around across neighborhoods, it is important to figure out the similarities or differences among them. In this reason 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. Here is the merged table with cluster labels for each borough.

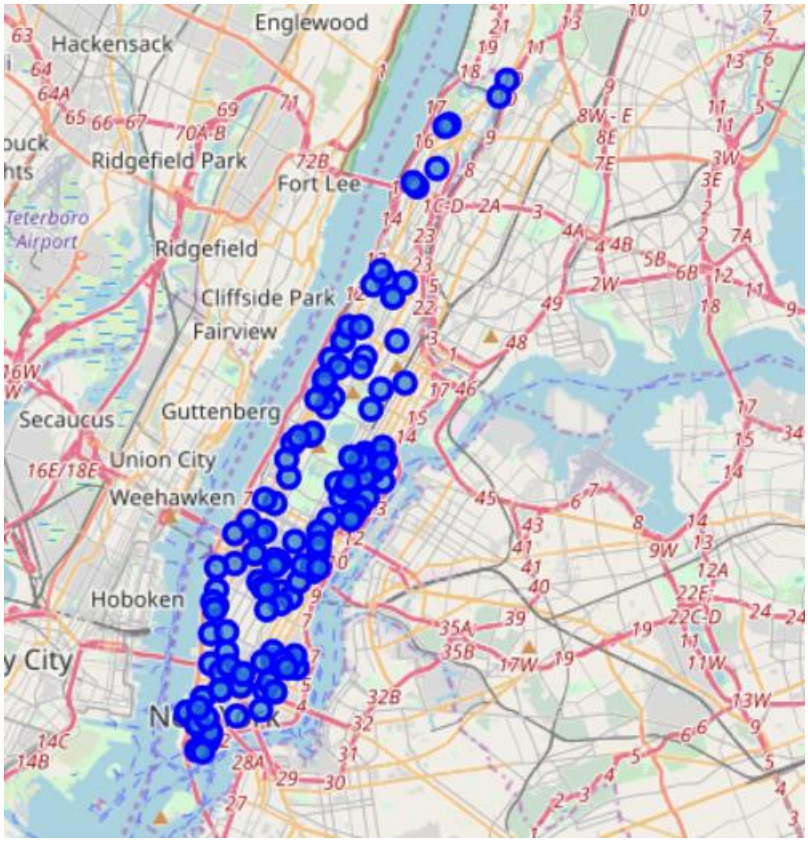
	Borough	Neighborhood	Latitude	Longitude	Cluster Labels
0	Manhattan	Marble Hill	40.876551	-73.910660	0
1	Manhattan	Chinatown	40.715618	-73.994279	3
2	Manhattan	Washington Heights	40.851903	-73.936900	0
3	Manhattan	Inwood	40.867684	-73.921210	0
4	Manhattan	Hamilton Heights	40.823604	-73.949688	0
5	Manhattan	Manhattanville	40.816934	-73.957385	0

Python folium library is used to visualize geographic details of Manhattan and its boroughs and a map of Manhattan with boroughs superimposed on top. Different clusters are showing in different colors.

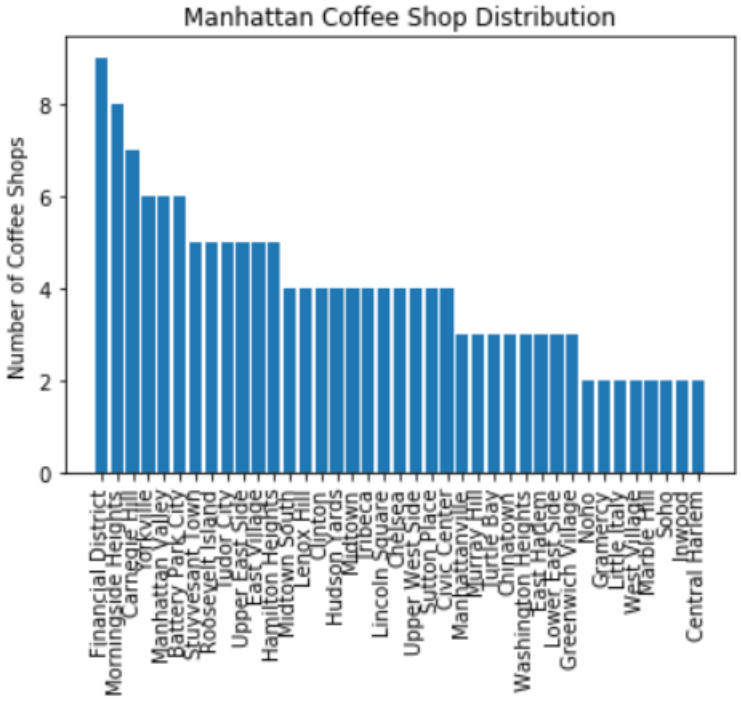


It is relatively clear that midtown and downtown regions are grouped as same cluster, as those areas are business areas with a lot of financial companies. Population around those areas tends to be higher.

As we need the data for coffee shops, I filtered by coffee shops on the dataset with all the venues information and lay out in the folium map as below:

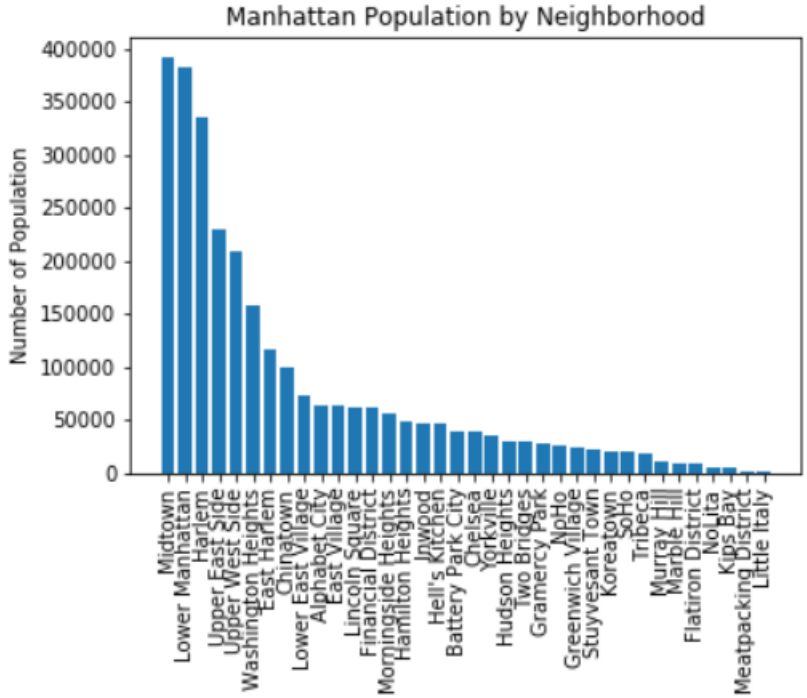


Coffee shops by boroughs bar chart is shown as below:

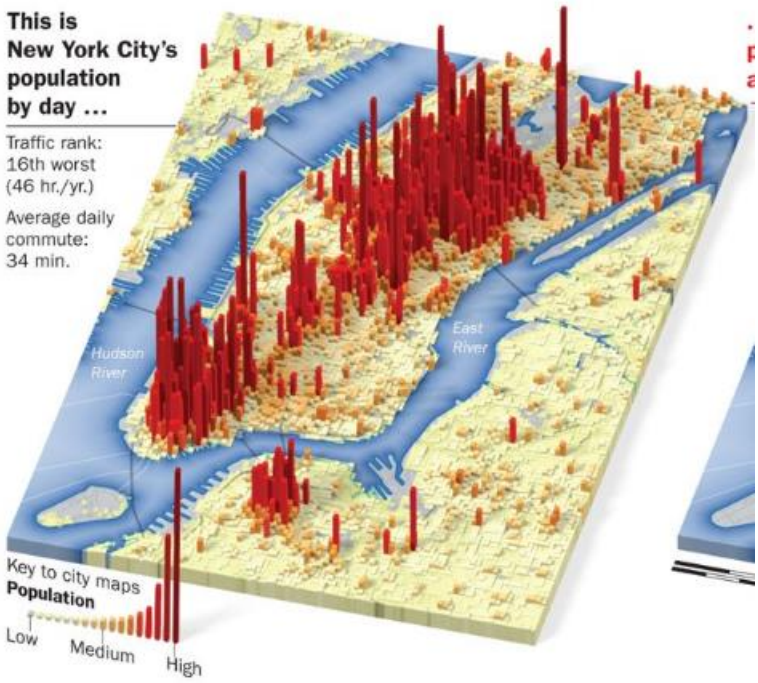


Total population by neighborhood at Manhattan is pulled from online website by leveraging python package ‘BeautifulSoup4’. After data cleaning, I get the table with neighborhood names and population. Below is the bar chart of population

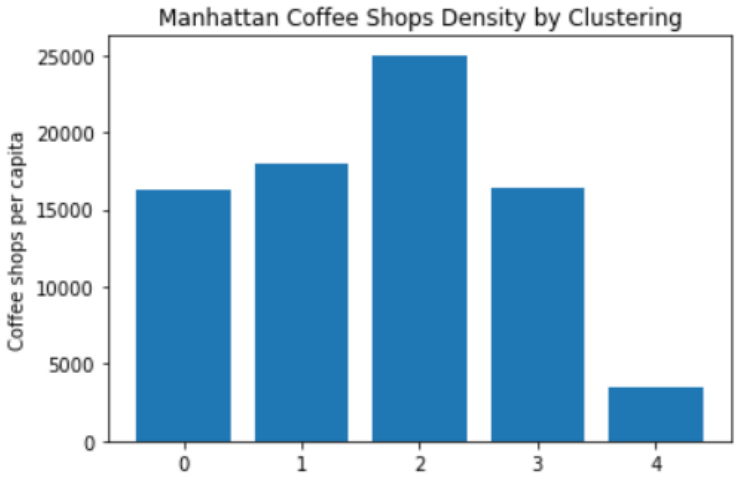
distribution.

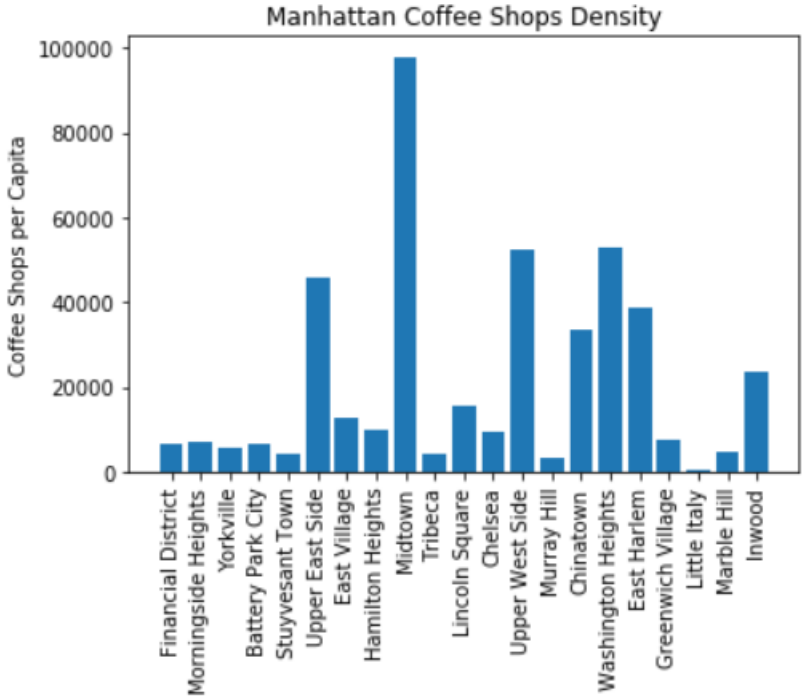


Other data of population distributions are as below. Midtown and downtown are having the most numbers of populations.



Average coffee shops per person is calculated by clustering and neighborhood respectively.





D. Results

The clustering results shows that some midtown areas and downtown areas are in the same group, while uptown are in a separated group. Midtown has the most population among all the Manhattan boroughs, while financial district has the most coffee shops. In terms of the results of average coffee shops per capita, cluster 2 is the highest. Below is the boroughs classified as cluster 2. Midtown has the highest coffee shops per capita if ranked by each borough.

	Neighborhood	Cluster Labels
8	Upper East Side	2
13	Lincoln Square	2
14	Clinton	2
15	Midtown	2
21	Tribeca	2
27	Gramercy	2
28	Battery Park City	2
29	Financial District	2
32	Civic Center	2
35	Turtle Bay	2
36	Tudor City	2
39	Hudson Yards	2

E. Discussion

As I have mentioned above, the numbers of coffee shops and the amounts of population are the main factors to determine the optimized location. The more people indicate more demands. Per capita data is a useful metric to measure the density of coffee shops. The higher coffee shot per capita means the more demand on coffee shops.

F. Conclusion

Midtown has highest coffee shops per capita both by neighborhood and by cluster. From the competitors/numbers of coffee shops and population density perspective, I recommend to open new coffee shop at Midtown as optimized location to potentially generate maximum profits.