Office Location Recommendation

Justin Lee

April 2020

1. Introduction

1.1 Background

Manhattan, New York City. It is the most densely populated of the five boroughs in New York City. Manhattan is also described as the cultural, financial, media and entertainment capital of the world. It is of no surprise that businesses would want to set up shop in Manhattan.

Derek, a good friend of mine is deciding to open up an office in Manhattan. However, he is clueless of where he should open it. After a brief discussion with his colleagues, he decided that the most important factor that should be taken into account is the number of food places around the area. Since his colleagues and himself love to eat and would like to have a wide variety of restaurants and joints to choose from during lunch time, he would want to have an office with the most food places around the neighborhood in Manhattan. As Derek is aware of my newly acquired data analytics skills, he has approached me to come up with a recommendation of which neighborhood that he should open up his office in.

1.2 Interest

Any individual that would want to set up an office or would like to know which area in Manhattan has the most food places around would find this analysis useful.

2. Data Acquisition

2.1 Data Sources

We would need various sets of data to ensure accurate analysis. First, we would need a dataset of New York neighborhoods which is already available to us from a previous exercise that can be obtained from this link: https://cocl.us/new_york_dataset.

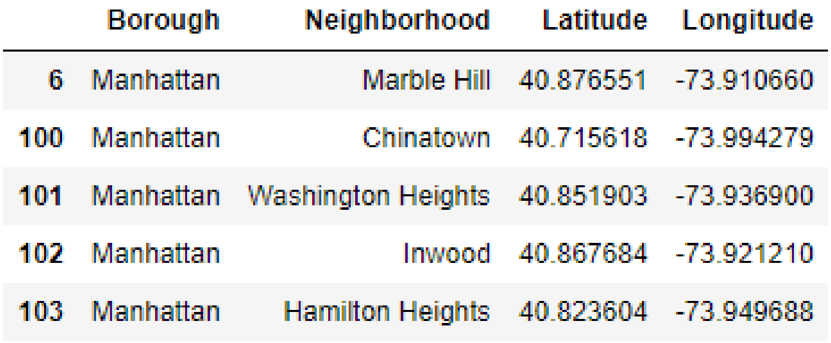
In the dataset, there are features that we would need like the boroughs and neighborhoods of New York so that we can find which neighborhoods reside in Manhattan. Next, we would need to make some foursquare API calls to obtain the data of the neighborhood venues in Manhattan. We would need to extract the venue category, longitude and latitudes of the venues.

A combination of the New York dataset from the link and foursquare API would allow us to obtain the longitude and latitudes of the different venues in the neighborhoods in Manhattan. Thus, it allows us to have enough information to identify clusters of food places further in our analysis.

2.2 Data Cleaning

2.2.1 Extracting Manhattan neighborhoods

After acquiring the data from the link, what we have is a set of features of the different locations around New York City. However, a lot of these data are redundant and the first thing I did was to extract only what I needed from the dataset like the borough, neighborhoods, latitude and longitude and put them into a data frame. The next step was to filter the data and extract the neighborhoods, latitude and longitude residing in the borough Manhattan. Now, we have a cleaned data frame of Manhattan neighborhoods and their longitude and latitudes (Table 1).

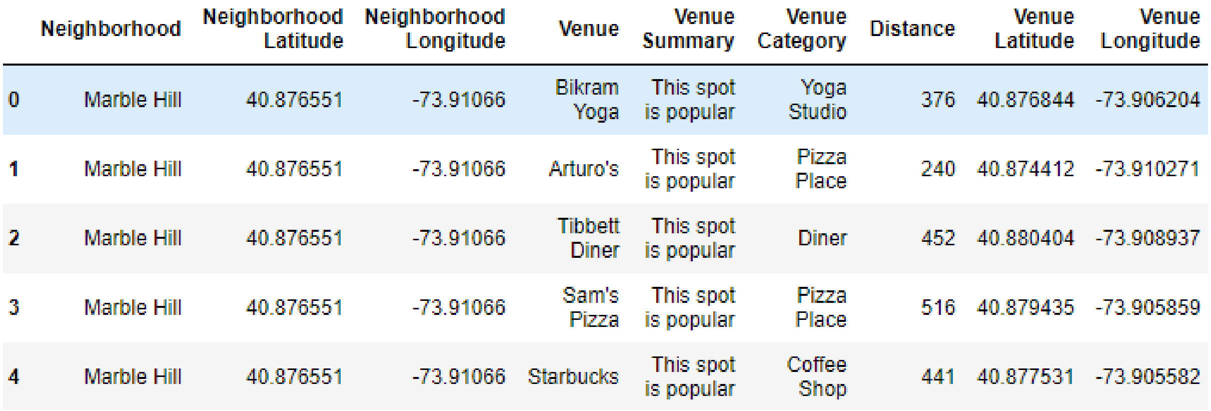


	Borough	Neighborhood	Latitude	Longitude
6	Manhattan	Marble Hill	40.876551	-73.910660
100	Manhattan	Chinatown	40.715618	-73.994279
101	Manhattan	Washington Heights	40.851903	-73.936900
102	Manhattan	Inwood	40.867684	-73.921210
103	Manhattan	Hamilton Heights	40.823604	-73.949688

Table 1: Cleaned data frame of Manhattan Neighborhoods with their Latitude and Longitudes.

2.2.2 Extracting information from Foursquare API

Next, we crawled the internet with foursquare API for the venues in Manhattan. The dataset consists of many pieces of information that we do not need. Thus, we extracted only the information we need once again and merged it with the data frame that we made previously (Table 2).



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Summary	Venue Category	Distance	Venue Latitude	Venue Longitude
0	Marble Hill	40.876551	-73.91066	Bikram Yoga	This spot is popular	Yoga Studio	376	40.876844	-73.906204
1	Marble Hill	40.876551	-73.91066	Arturo's	This spot is popular	Pizza Place	240	40.874412	-73.910271
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	This spot is popular	Diner	452	40.880404	-73.908937
3	Marble Hill	40.876551	-73.91066	Sam's Pizza	This spot is popular	Pizza Place	516	40.879435	-73.905859
4	Marble Hill	40.876551	-73.91066	Starbucks	This spot is popular	Coffee Shop	441	40.877531	-73.905582

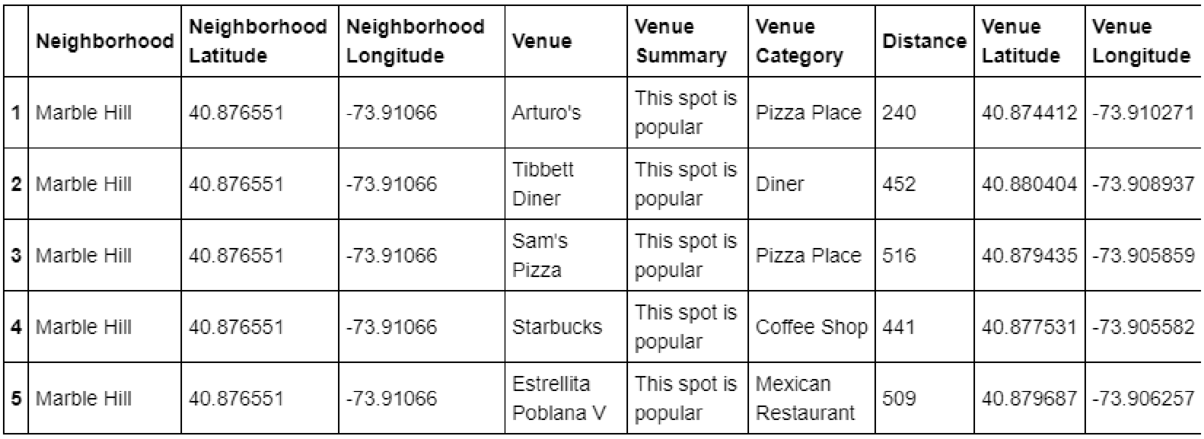
Table 2: Merged dataset of Manhattan neighborhoods and different venues.

2.2.3 Identifying unique venue categories in the dataset

The first thing I need to do is to identify the various unique venue categories that we have in the dataset. This will allow us to individually identify venue categories related to food places that we are trying to identify. The results showed that we have 307 unique venue categories in the dataset and a list of the venue categories identified.

2.2.4 Manually selecting features for food places in Manhattan.

From the 307 unique venue categories, we have to identify the venues that are food places and those that are not. The fastest way to do this is to manually identify them one by one as not all restaurants have ‘Restaurants’ in the venue category as some restaurants can be a ‘Steakhouse’ as well. Venues like ‘Yoga Studio’ and ‘Park’ are removed completely. It is a tedious task but it is the best way to do it. There are other features related to the venues as we would still need like their neighborhood, venue name, latitude and longitude so I’ve included them as well and updated the data frame (Table 3).



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Summary	Venue Category	Distance	Venue Latitude	Venue Longitude
1	Marble Hill	40.876551	-73.91066	Arturo's	This spot is popular	Pizza Place	240	40.874412	-73.910271
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	This spot is popular	Diner	452	40.880404	-73.908937
3	Marble Hill	40.876551	-73.91066	Sam's Pizza	This spot is popular	Pizza Place	516	40.879435	-73.905859
4	Marble Hill	40.876551	-73.91066	Starbucks	This spot is popular	Coffee Shop	441	40.877531	-73.905582
5	Marble Hill	40.876551	-73.91066	Estrellita Poblana V	This spot is popular	Mexican Restaurant	509	40.879687	-73.906257

Table 3: Updated data frame with only venue of food places in Manhattan

3. Exploratory Data Analysis

3.1 Visualizing the data

I've plotted the longitude and latitude of the food places into a scatter plot to visualize the locations of the venues (Figure 1). This gives me a rough understanding of the food places around Manhattan. However, the clusters in the plot are not clear as there are more than a thousand food places in Manhattan.

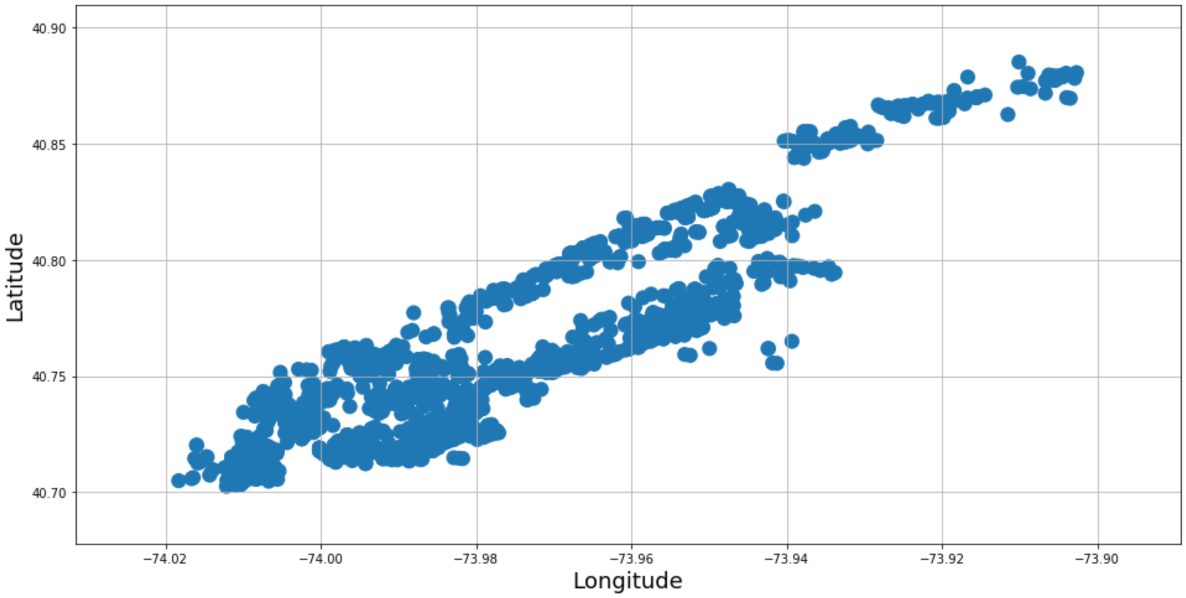


Figure 1: Longitude and latitude of food places around Manhattan.

3.2 Using K-means clustering to cluster the data

I’ve decided to use k-means clustering to cluster the data into 5 clusters (Figure 2). This will give me clusters of similar sizes which is ideal as I am trying to identify clusters and to find out the cluster with the most food places available. I’ve also included the center of each cluster which will eventually be the recommended location for the best cluster.

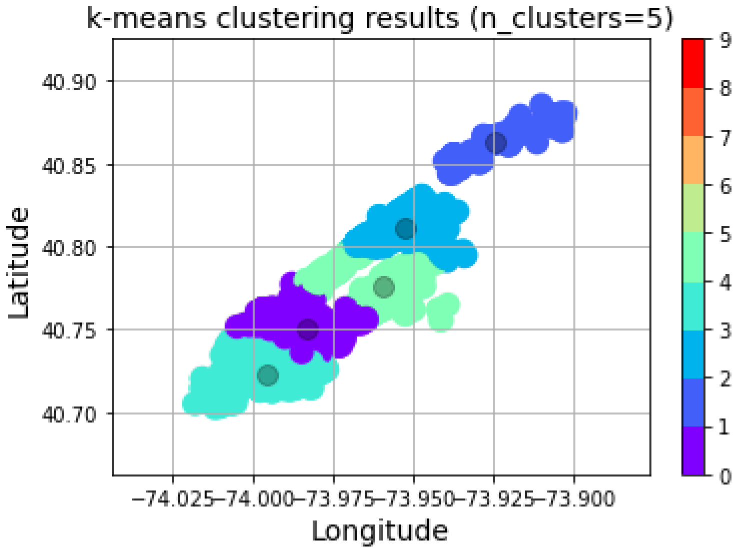
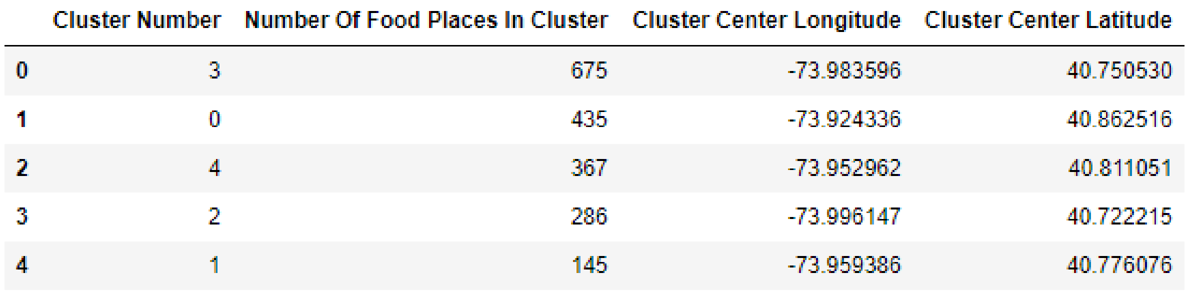


Figure 2: K-means clustering results

4. Results

I’ve generated a table to visualize the results of the analysis (Table 4). I’ve also included the cluster centers’ longitude and latitude and we can later visualize the exact location of the center of the cluster. The results show that the cluster with the greatest number of food places is cluster 3 with 675 food places in the cluster, the second-best cluster is cluster 0 with 435 food places and the third best is cluster 4 with 367 number of food places in the cluster.



	Cluster Number	Number Of Food Places In Cluster	Cluster Center Longitude	Cluster Center Latitude
0	3	675	-73.983596	40.750530
1	0	435	-73.924336	40.862516
2	4	367	-73.952962	40.811051
3	2	286	-73.996147	40.722215
4	1	145	-73.959386	40.776076

Table 4: Number of food places in each cluster with the longitude and latitude of their centers

5. Discussion

Finding an office near the center of the cluster 3 would be the best place for Derek to set up his office if he wants the best location possible for food places around the area (Figure 3). With 675 food places around the area of cluster 3, I am confident that Derek and his colleagues will never run out of food choices during lunch time. Derek could also consider cluster 0 as a viable choice for his office as it has a significant number of food places around the area with 435 food places.

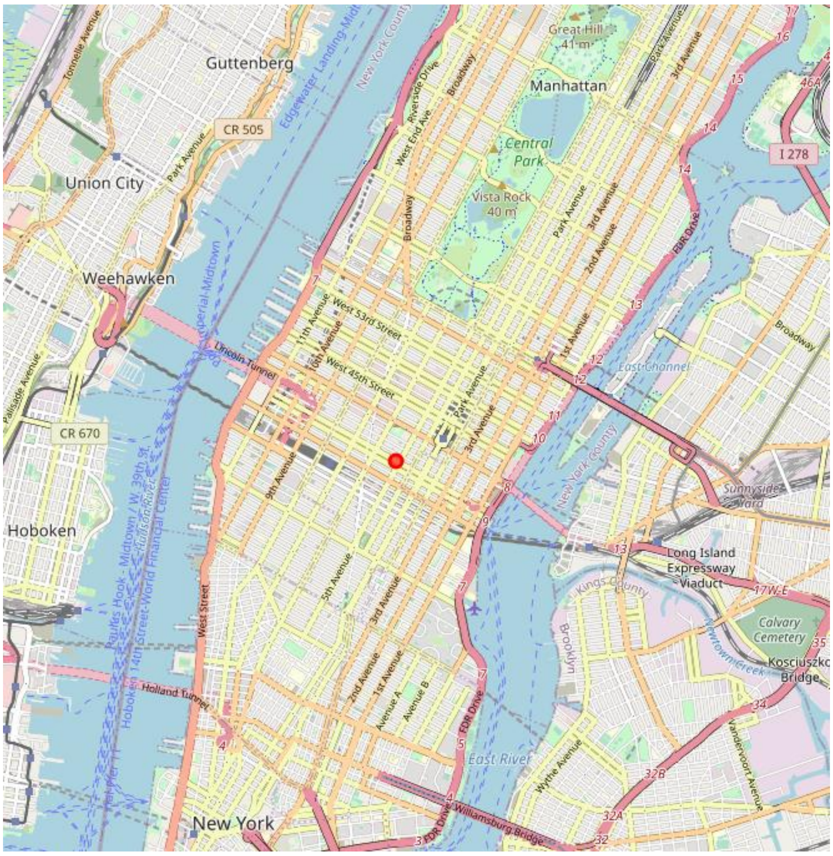


Figure 3: Location of the center of cluster 1

6. Conclusion

Any individual who wants to find the best location with the most food places around the area will find this analysis useful. Whether they want to set up a restaurant, an office or find an apartment with the most food places around. However, the model can be improved as the food places are widely scattered around the cluster with a big variety of food types. An individual might need to travel a distance just to get their favorite kebab. It would be better if there was a survey of preferences that Derek colleagues have to improve on the results of the analysis.