Finding the Optimal Neighborhood in Manhattan for Café Business

Ruby Wu

December 22, 2020

1 Introduction

1.1 Background

Manhattan, the most densely populated borough of the New York City, is a well-known area which serves as the economic and administrative center of the United States. It covers 33.58 square miles' land and has approximately 1,628,706 residents (by 2019) in total. People all around the world are attracted to Manhattan because of its leading economy, diverse culture, historical tourism and, more importantly, huge potential for success. Therefore, it is advantageous for us to explore deeper into Manhattan in terms of business opportunities.

1.2 Business Problem

There are lots of coffee shops locating in different neighborhoods of Manhattan. Their business performances highly depend on various factors including location, residence and other business categories. Therefore, based on these considerations, this project aims to find the optimal neighborhood for opening a coffee shop/café. Specifically, we will explore 40 neighborhoods in Manhattan and generate a best solution for stakeholders interested in coffee business.

2 Data

We will be using three types of data in this project: Manhattan neighborhood data, Manhattan population data and Foursquare venue data. In this section, we will explain our data sets in terms of their sources and content.

2.1 Manhattan Neighborhood Data

Manhattan neighborhood data can be found in *IBM Skills Network*. The data is in JSON format containing features including boroughs, neighborhoods, latitude and longitude across all 5 boroughs in the New York City. The dataset has 306 neighborhoods in total along with their coordinates. We will only use borough Manhattan data from this data set.

2.2 Manhattan Population Data

Manhattan population data can be scraped from the Neighborhoods in New York City page of Wikipedia. It contains area, population and population density for each community board which is the appointed advisory group of the community districts of Manhattan. Manhattan has 12 community boards (CB) in total. Based on this dataset, we can see which neighborhoods are assigned to the same community board and the population distribution across all community areas.

2.3 Foursquare Location Data

We will also use the Foursquare API to pull the venue information based on each Manhattan neighborhood's coordinates and to explore what kinds of venue categories are in the

neighborhood and their frequencies. We will also pay closer attention to coffee related categories specifically. According to the Foursquare documentation, venue categories related with coffee are named to Coffee Shop, Cafeteria, or Café. Therefore, when we explore coffee business in Manhattan, we will extract those venue data based on the three category names.

3 Methodology

3.1 Exploratory Data Analysis

To find a neighborhood that is the best to open a café, we will need to consider the following factors:

- The number of coffee shops already existing in the neighborhood
- The number of different venue categories in the neighborhood
- Population of the neighborhood

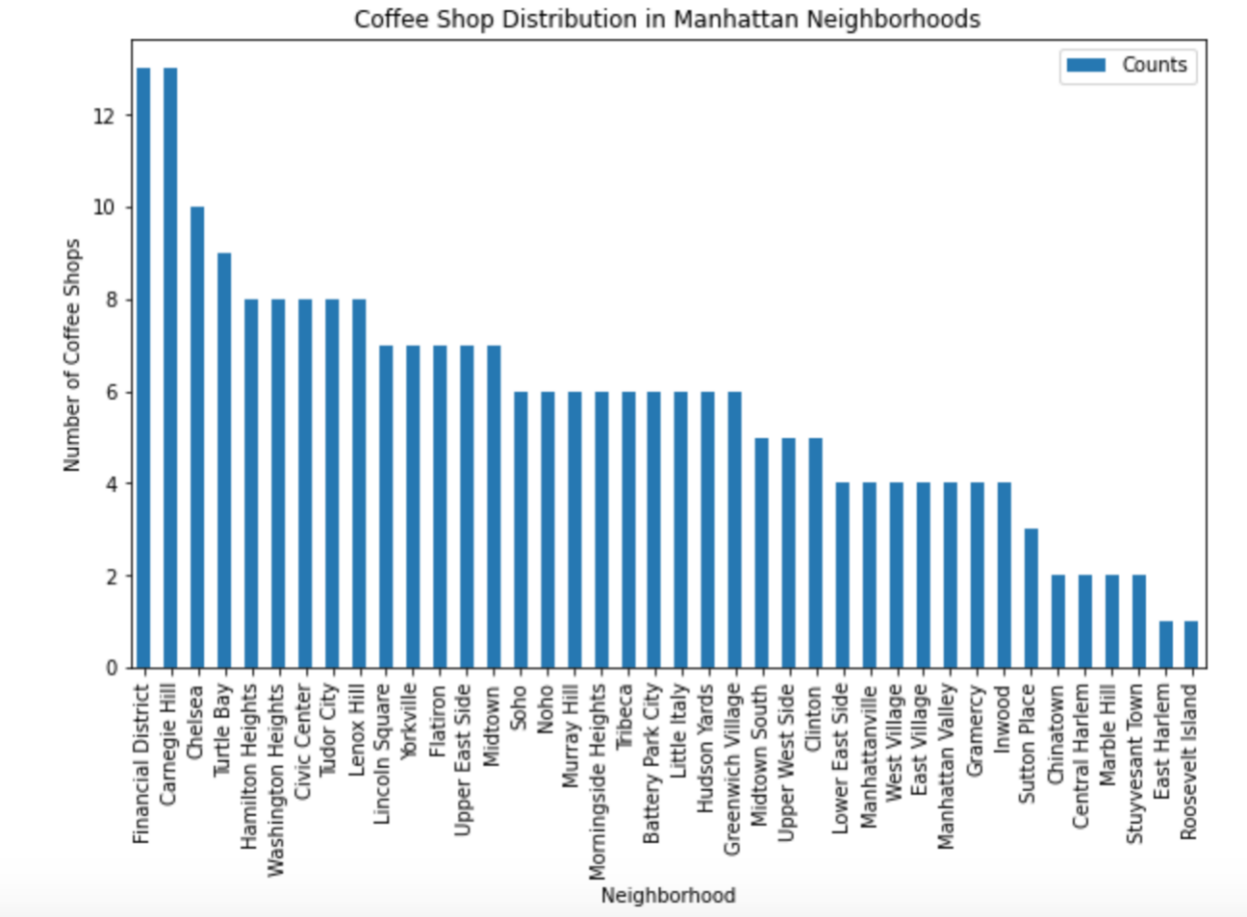
Specifically, we do not want to open our café in the place where there already are many coffee shops around as this may significantly increase our business competition/burden. We also want to look at the number of venue categories because this information can represent business variety in the neighborhood. The more venue categories the neighborhood has, the more diverse the environment can be, and the more potential the business can explore. Besides, we also consider population of the neighborhood as we want to find a neighborhood with more residence which may lead to more attention and popularity.

We start with 40 neighborhoods in Manhattan including Marble Hill, Chinatown, Washington Heights, Inwood, East Harlem, Midtown, etc. Our Manhattan neighborhood data gives us these neighborhoods' names and their coordinates. We then pass this data to the Foursquare API to return the venue data for each neighborhood. We limit the venue amount to be 100 and the radius to be 500 meters for each neighborhood from their given latitude and longitude. Here is a head of the data frame listing venue name and category.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop
4	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop

Factor 1: The number of coffee shops already existing in the neighborhood

There are 333 unique venue categories in Manhattan, and we specifically want to see how are the coffee shops distributed across all 40 neighborhoods. Therefore, we sum up the number of coffee shops in each neighborhood and the below is the bar plot.



According to the bar plot, we can see that neighborhoods such as Financial District, Carnegie Hill and Chelsea have the most coffee shops, while East Harlem and Roosevelt Island have the least amount.

As we stated above, we want to pick the neighborhoods that do not have heavy business competition. Therefore, we avoid those top few neighborhoods with the most coffee shops and restrict our candidates to the middle 9 neighborhoods.

Candidates:

Soho	Noho	Murray Hill	Morningside Heights	Tribeca	Battery Park City	Little Italy	Hudson Yards	Greenwich Village

Factor 2: The number of different venue categories in the neighborhood

As we limit our candidates to the above 9 neighborhoods, we will then look at their venue distributions. We group our Foursquare venue data by the 9 neighborhoods and get the below table.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Category
Soho	100	100	100	100	100	100
Noho	100	100	100	100	100	100
Murray Hill	100	100	100	100	100	100
Morningside Heights	43	43	43	43	43	43
Tribeca	86	86	86	86	86	86
Battery Park City	77	77	77	77	77	77
Little Italy	100	100	100	100	100	100
Hudson Yards	70	70	70	70	70	70
Greenwich Village	100	100	100	100	100	100

The table shows that Morningside Heights has the least number of venue categories and is much less than other neighborhoods. Therefore, we should remove it from consideration because of its limited venue variety.

We also look at the top 10 venue categories in our candidate neighborhoods for further details.

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Battery Park City	Cafe	Park	Hotel	Memorial Site	Clothing Store	Gym	Food Court	Gourmet Shop	Plaza	Playground
Greenwich Village	Italian Restaurant	Cafe	Clothing Store	Sushi Restaurant	American Restaurant	Dessert Shop	Indian Restaurant	Seafood Restaurant	Chinese Restaurant	Gym
Hudson Yards	Cafe	Gym / Fitness Center	American Restaurant	Hotel	Italian Restaurant	Burger Joint	Gym	Dog Run	Nightclub	Park
Little Italy	Cafe	Bakery	Bubble Tea Shop	Italian Restaurant	Chinese Restaurant	Ice Cream Shop	Sandwich Place	Salon / Barbershop	Cocktail Bar	Pizza Place
Murray Hill	Hotel	Cafe	Sandwich Place	Bar	American Restaurant	Italian Restaurant	Burger Joint	Japanese Restaurant	Gym / Fitness Center	Restaurant
Noho	Italian Restaurant	Cafe	Cocktail Bar	Hotel	Yoga Studio	Bookstore	Sandwich Place	Pizza Place	Mexican Restaurant	Grocery Store
Soho	Clothing Store	Cafe	Italian Restaurant	Mediterranean Restaurant	Asian Restaurant	Sporting Goods Shop	Boutique	Salon / Barbershop	Bakery	Pizza Place
Tribeca	Cafe	Park	Italian Restaurant	Spa	American Restaurant	Wine Bar	Men's Store	Greek Restaurant	Bakery	Gym / Fitness Center

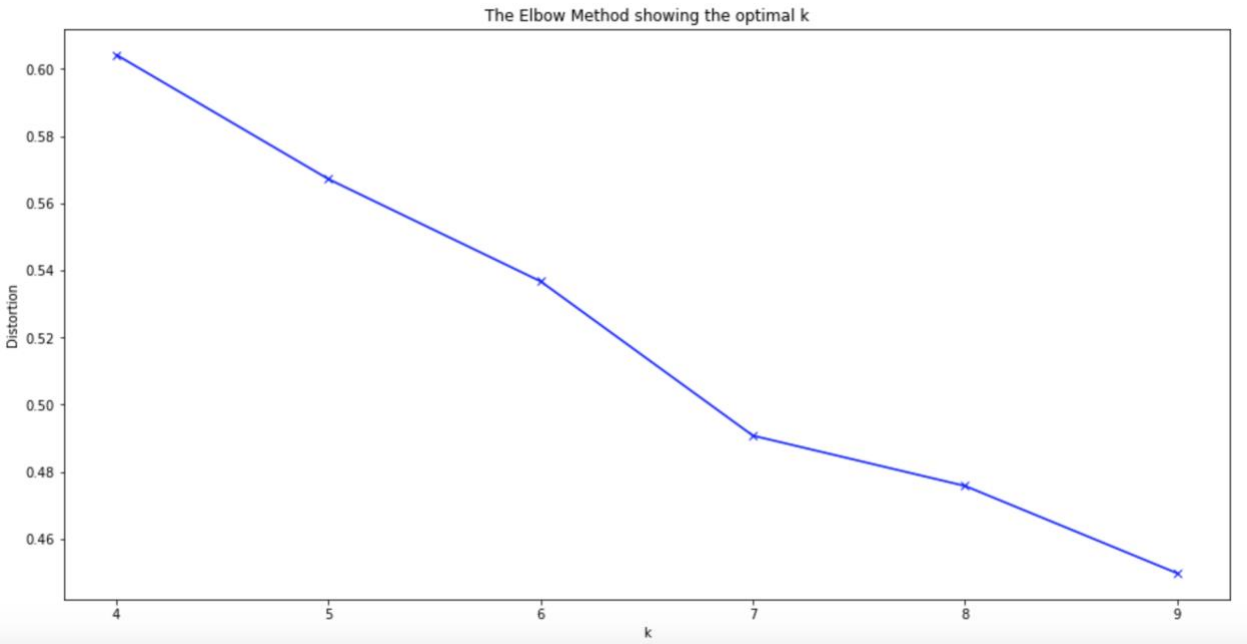
Based on the above table, we can see that most of the venue categories are restaurants. Some neighborhoods have parks, and some have spa, hotel and gym. Now our candidates are left with 8 neighborhoods.

Candidates:

Soho	Noho	Murray Hill	Tribeca	Battery Park City	Little Italy	Hudson Yards	Greenwich Village

3.2 K-means Clustering

At this step, we perform a clustering analysis on the most common venue data of all neighborhoods in Manhattan. We use k-means algorithm in order to categorize similar neighborhoods into clusters based on the similarities of venue categories. We first need to investigate the optimal k value for clusters.

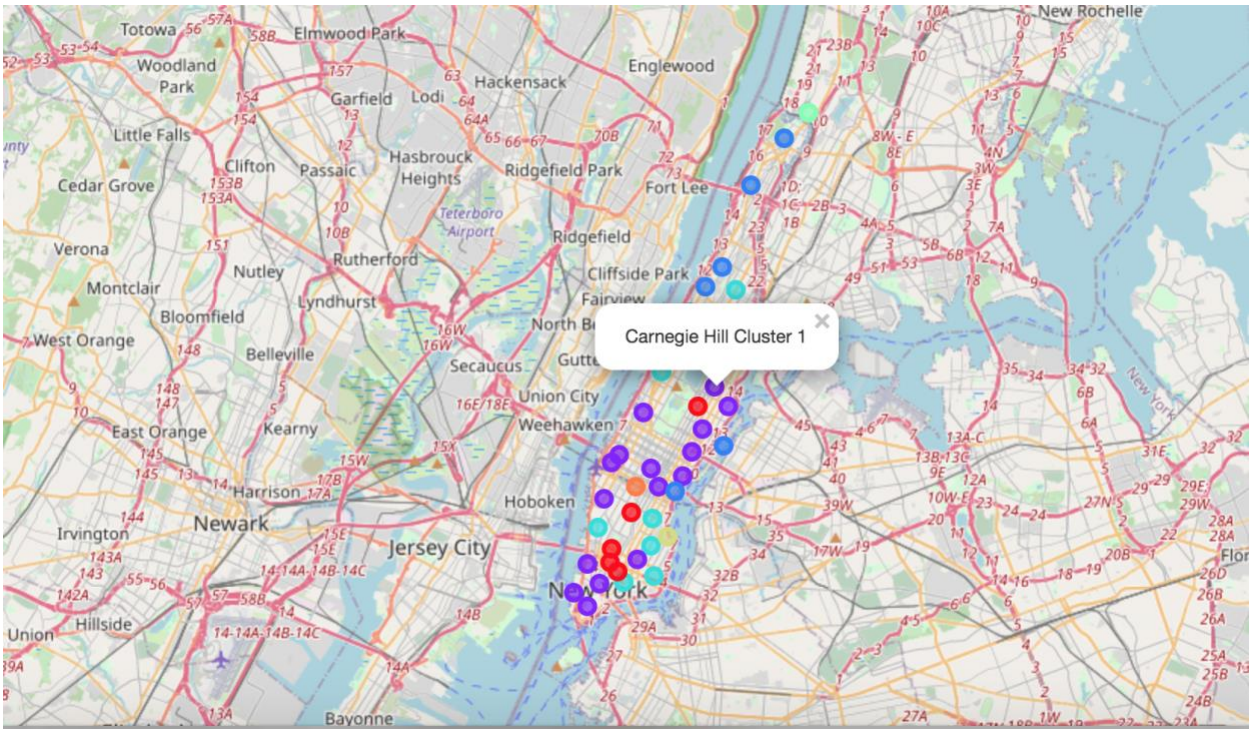


The elbow method plot overall does not show significant kink for k from 4 to 9, but it has a very small kink around k = 7. Therefore, to further investigate, we look into the Silhouette score for clustering for each k value as another metric to determine the optimal k. Silhouette score gives a perspective into the density and separation of the formed clusters as it measures how similar a point is to its own cluster compared to other clusters. Here is the Silhouette scores for each k.

```
For n_clusters = 4 The average silhouette_score is : 0.030734128667779876
For n_clusters = 5 The average silhouette_score is : 0.002508987600141415
For n_clusters = 6 The average silhouette_score is : 0.016984022724632723
For n_clusters = 7 The average silhouette_score is : 0.0131426919175024
For n_clusters = 8 The average silhouette_score is : 0.0026767148998263816
For n_clusters = 9 The average silhouette_score is : 0.006567279598424958
```

According to the output, k=6 and k=7 give the maximum scores. Therefore, considering both the elbow method and Silhouette score, we choose to run k-means with 7 clusters, which categorize Manhattan neighborhoods into 7 clusters.

The clustering plot is shown as below. Neighborhoods in Manhattan are clustered into 7 groups with different marker colors. Generally speaking, neighborhoods that are close to each other tend to belong to the same cluster.



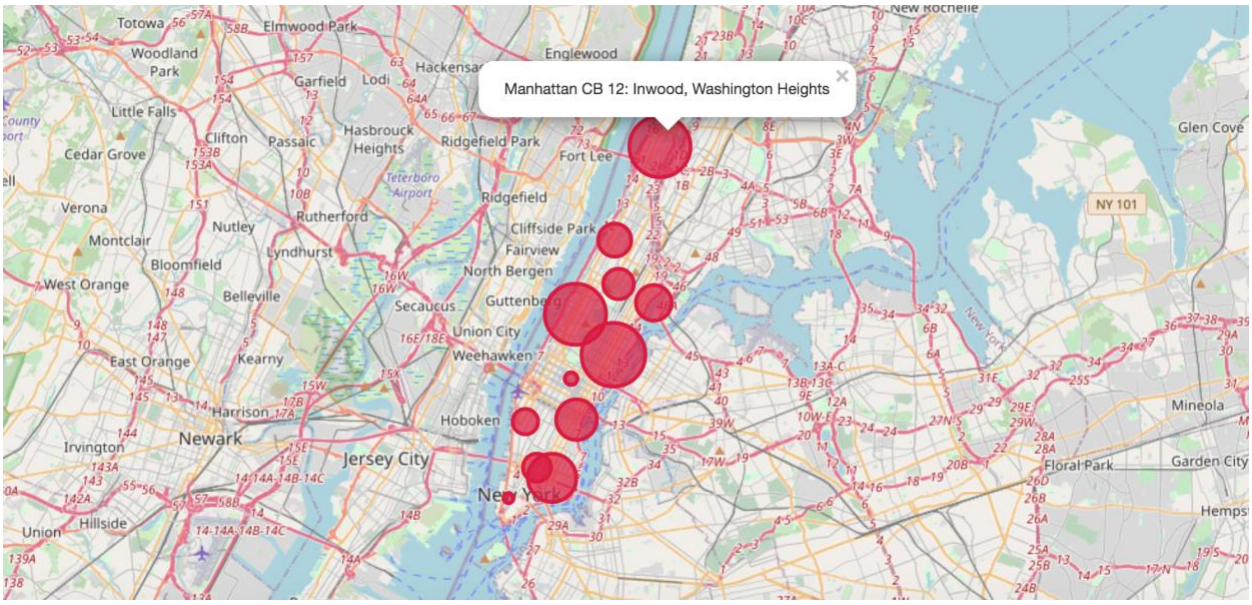
Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Manhattan	Murray Hill	40.748303	-73.978332	1	Hotel	Cafe	Sandwich Place	Bar	American Restaurant	Italian Restaurant	Burger Joint	Japanese Restaurant	Gym / Fitness Center	Restaurant
Manhattan	Greenwich Village	40.726933	-73.999914	0	Italian Restaurant	Cafe	Clothing Store	Sushi Restaurant	American Restaurant	Dessert Shop	Indian Restaurant	Seafood Restaurant	Chinese Restaurant	Gym
Manhattan	Tribeca	40.721522	-74.010683	1	Cafe	Park	Italian Restaurant	Spa	American Restaurant	Wine Bar	Men's Store	Greek Restaurant	Bakery	Gym / Fitness Center
Manhattan	Little Italy	40.719324	-73.997305	0	Cafe	Bakery	Bubble Tea Shop	Italian Restaurant	Chinese Restaurant	Ice Cream Shop	Sandwich Place	Salon / Barbershop	Cocktail Bar	Pizza Place
Manhattan	Soho	40.722184	-74.000657	0	Clothing Store	Cafe	Italian Restaurant	Mediterranean Restaurant	Asian Restaurant	Sporting Goods Shop	Boutique	Salon / Barbershop	Bakery	Pizza Place
Manhattan	Battery Park City	40.711932	-74.016869	1	Cafe	Park	Hotel	Memorial Site	Clothing Store	Gym	Food Court	Gourmet Shop	Plaza	Playground
Manhattan	Noho	40.723259	-73.988434	1	Italian Restaurant	Cafe	Cocktail Bar	Hotel	Yoga Studio	Bookstore	Sandwich Place	Pizza Place	Mexican Restaurant	Grocery Store
Manhattan	Hudson Yards	40.756658	-74.000111	1	Cafe	Gym / Fitness Center	American Restaurant	Hotel	Italian Restaurant	Burger Joint	Gym	Dog Run	Nightclub	Park

This table displays the clustering labels for our 8 candidate neighborhoods. They are assigned to two clusters: cluster 0 and 1. Three are in cluster 0, and five are in cluster 1. Based on the given information, we cannot explicitly see any difference among all neighborhoods. Therefore, we need to keep analyzing our third factor which is population data to help us narrow down solutions.

Factor 3: Population of the neighborhood

Wikipedia population data provide us with information about the coordinates, neighborhoods and area of each community board in Manhattan. The below table summarizes data for us. We find that CB 8 has the most population, which includes neighborhoods Lenox Hill, Roosevelt Island, Upper East Side and Yorkville. We also visualize the population density based on a folium map.

CB	Area	Population	Population/Area	Neighborhoods	Latitude	Longitude
Manhattan CB 8	5.13	217063	42312	Lenox Hill, Roosevelt Island, Upper East Side,...	40.769891	-73.955392
Manhattan CB 12	7.64	208414	27279	Inwood, Washington Heights	40.854728	-73.930358
Manhattan CB 7	5.46	207699	38040	Lincoln Square, Manhattan Valley, Upper West Side	40.786380	-73.975863
Manhattan CB 3	4.56	164407	36054	Alphabet City, Chinatown, East Village, Lower ...	40.719617	-73.988447
Manhattan CB 6	3.55	136152	38353	Gramercy Park, Kips Bay, Rose Hill, Murray Hil...	40.743383	-73.975280



The folium map helps us visualize the population density across all community boards in Manhattan. The higher the density, the bigger the radius of the mark. Therefore, we can clearly see that there are 5 community boards that have relatively larger population. As we want to figure out what neighborhoods are located in those community boards, we extract their names out from the data set. We then compare these 24 neighborhoods with our candidates, and realize that Murray Hill is the only one has relatively larger population. Therefore, we finalize our result to Murray Hill. It is the optimal neighborhood to open a café based on all our analysis.

Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Manhattan	Murray Hill	40.748303	-73.978332	1	Hotel	Cafe	Sandwich Place	Bar	American Restaurant	Italian Restaurant	Burger Joint	Japanese Restaurant	Gym / Fitness Center	Restaurant

4 Results

We've followed three factors as our standard of choosing a neighborhood that is optimal to open a café/coffee shop. We started with 40 neighborhoods in Manhattan and narrowed down to 9 as candidates based on the existing coffee business in the neighborhoods. We then looked at the diversity of venue categories in the neighborhoods to determine business potential and space for development. In addition, we took population density into consideration as one of the factors to measure residence condition and future exposure. By investigating all three factors, we got a solution that out of all neighborhoods in Manhattan, Murray Hill is the optimal one to open our café/coffee shop.

5 Discussion

To solve the business problem, we considered three factors: the number of existing coffee shops in the neighborhood, the number of venue categories in the neighborhood and the population density of the neighborhood. With these three factors, we were able to narrow our candidates from over 40 neighborhoods to 9 and to 8 and finally to just 1. Therefore, we can recommend our stakeholders that Murray Hill can be the ideal neighborhood to open a café/coffee shop.

However, there are some improvements that can be done in the future. We only analyzed the data for the Manhattan borough and performed a clustering algorithm to find those similar neighborhoods. One thing to note that the elbow method we implemented did not give us an explicit instruction of choosing the optimal k value, even though we saw a small kink around 7. Silhouette scores were not high enough to ensure us the best k as well. Therefore, we may need more information to run clustering, either more feature analysis or more metrics to compare. Besides, the result is based on the three factors we analyzed in this project. We should take additional factors into consideration such as housing price and the neighborhood center location, as they may influence the café business to some extent as well.

6 Conclusion

The goal of our project is to help our stakeholders find the optimal neighborhood in Manhattan to open a café/coffee shop. According to the analysis of the three factors, we can conclude that Murray Hill is the one that fulfills all the requirements. It does not have many coffee shops in the neighborhood for now, so we do not need to care about the business competition with other owners. Murray Hill also has many unique venue categories in it, which indicates business diversity and huge potential for growth. Relatively larger population density is also another advantage of Murray Hill as it represents popularity and exposure. Therefore, Murray Hill is recommended for our stakeholders. This can be used as a reference, and final decision will be made by stakeholders who are interested in opening a café in Manhattan. They will need to consider additional factors and improvements that we discussed above.