Applied Data Science Capstone:
Exploration and Mapping of the Coffee Shops in Manhattan

Ashley Figatner
May 30th, 2021

1. Introduction/Business Problem

For many aspiring business owners, the decision of where to set down roots is crucial to the business strategy. This location impacts revenue, initial costs, operational costs, customer traffic and advertising. Depending on the city and business, there are also likely to be competitors nearby. Take New York City coffee shops, for example. Throughout New York City, there are hundreds of coffee shops, many of them within close proximity to each other. For a potential coffee shop owner, this competition may pose a threat to the business. Why not build a coffee shop in a location where this venue type is scarce? The location of a new business is likely the biggest factor in its success or failure, so it is important to spend time to explore and analyze the options. Since it is not always simple to determine the optimal location, we can turn to data science to guide these decisions.

Using Foursquare venue data, locations for a specific venue type can be mapped, and neighborhoods can be ranked by how many venues exist in a specific category, or even what percentage of venues are of a specific type. In this project, Foursquare data will be used to map out and guide business owners towards optimal neighborhoods to open coffee shops in New York City.

2. Data

In order to map the coffee shops in New York City, I combined Wikipedia data of New York City boroughs and neighborhoods with Foursquare location and venue data. I mapped neighborhoods in NYC using latitude and longitude coordinates. I utilized venues available in the Foursquare dataset in order to calculate a percentage of venues that are coffee shops in each neighborhood, and I ranked neighborhoods based on coffee shop frequency as well as the number of people per coffee shop.

	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

Venue	Venue Latitude	Venue Longitude	Venue Category
Bikram Yoga	40.876844	-73.906204	Yoga Studio
Arturo's	40.874412	-73.910271	Pizza Place
Tibbett Diner	40.880404	-73.908937	Diner
Dunkin'	40.877136	-73.906666	Donut Shop
Starbucks	40.877531	-73.905582	Coffee Shop

Figures 1&2: Manhattan neighborhood data from Wikipedia alongside venue data from Foursquare

In order to incorporate other parameters into this analysis, I computed population per coffee shop by utilizing Manhattan neighborhood population data found on WorldAtlas.com. This dataset uses the 2012 census estimate to show population per neighborhood.

WorldAtlas CONTINENTS COUNTRIES WORLD EDUCATION SOCIAL SCIENCE

Manhattan Neighborhoods By Population

Rank	Neighborhood	Population
1	Midtown	391,371
2	Lower Manhattan	382,654
3	Harlem	335,109
4	Upper East Side	229,688
5	Upper West Side	209,084
6	Washington Heights	158,318
7	East Harlem	115,921
8	Chinatown	100,000
9	Lower East Village	72,957
10	Alphabet City	63,347

Figure 3: Manhattan neighborhood population data from World Atlas, showing 2012 census information

https://www.worldatlas.com/articles/manhattan-neighborhoods-by-population.html

In addition, it was clear that a business owner might be interested in recent neighborhood growth. To help answer this point, I incorporated Real Estate sales activity data from PropertyShark.com, which shows percent change in number of real estate transactions from 2019 to 2020. This data could show potentially interesting neighborhoods based on recent growth in those areas.

PropertyShark

Rank	Borough	Neighborhood	Median Sale Price 2020	Y-o-Y Change in Median Sale Price	No of Transactions 2020	Y-o-Y Change in No of Transactions
1	Manhattan	Hudson Yards	$4,504,000	15%	23	-85%
2	Manhattan	TriBeCa	$3,157,000	-3%	198	-51%
3	Manhattan	Little Italy	$2,750,000	4%	35	25%
4	Manhattan	SoHo	$2,463,000	-10%	64	-45%
5	Manhattan	Hudson Square	$2,100,000	-15%	78	-26%
6	Manhattan	Theatre District - Times Square	$1,807,000	8%	127	-39%

Figure 4: Manhattan neighborhood real estate data from Property Shark, showing 2020 transaction data

https://www.propertyshark.com/Real-Estate-Reports/nyc-neighborhood-sales/#:~:text=Only%209%20NYC%20neighborhoods%20recorded,and%20Neponsit's%20medians%20exceeded%20%241%2C000%2C000

3. Methodology

The first aspect of location data that a potential business owner may be interested in is the frequency of venue type in each area. To compute this information, Foursquare venue data was collected for New York City, and these venues were filtered to focus on neighborhoods in Manhattan specifically. At this point, one hot encoding was used to group venues into categories. Venues were then grouped by neighborhood, where the output was the mean; this value is equivalent to the frequency of each venue type per neighborhood. From here, top venue categories were shown for each neighborhood, and a data-frame was created to show the top 10 most common venue types per neighborhood. With this dataset, it is possible to look qualitatively at the top most common venues to identify neighborhoods with high frequency of each venue.

```
----Battery Park City----
            venue  freq
0            Park  0.10
1     Coffee Shop  0.06
2           Hotel  0.06
3  Clothing Store  0.05
4             Gym  0.04


----Carnegie Hill----
         venue  freq
0  Coffee Shop  0.07
1         Café  0.06
2    Wine Shop  0.04
3          Gym  0.03
4    Bookstore  0.03
```

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue
0	Battery Park City	Park	Coffee Shop	Hotel	Clothing Store
1	Carnegie Hill	Coffee Shop	Café	Wine Shop	Gym
2	Central Harlem	African Restaurant	Seafood Restaurant	Gym / Fitness Center	American Restaurant
3	Chelsea	Coffee Shop	Art Gallery	Bakery	American Restaurant
4	Chinatown	Chinese Restaurant	Bakery	Cocktail Bar	American Restaurant

Figures 5&6: Output from venue category frequency analysis along with the first few rows from the common venues data-frame.

In this project, the objective was to discuss Coffee Shops and the frequency of this venue type in Manhattan. If an entrepreneur was looking to start a new coffee shop, they may care about which neighborhoods already have a lot of coffee shops. The dataset from Part 1 was sorted in order of frequency of coffee shop venues in Manhattan. With this dataset, the top 3 most frequent and bottom 3 least frequent neighborhoods were highlighted. In order to generate maps to visualize the data, Wikipedia data was merged with the Foursquare dataset. The folium package was utilized along with branca color mapping to generate the graph and to color-code the neighborhood markers on a color scale from red to yellow to green, where green is low frequency of coffee shops, and red is high frequency. This could help the potential business owner to determine areas with high numbers of coffee shops that they may wish to avoid.

	Neighborhood	Coffee Shop			
9	Financial District	0.100000	27	Noho	0.030000
30	Stuyvesant Town	0.095238	38	West Village	0.030000
25	Morningside Heights	0.093023	10	Flatiron	0.030000
22	Marble Hill	0.086957	37	Washington Heights	0.023529
21	Manhattanville	0.083333	32	Tribeca	0.022222
28	Roosevelt Island	0.074074	2	Central Harlem	0.021277
1	Carnegie Hill	0.072917	4	Chinatown	0.020000
5	Civic Center	0.070000	15	Inwood	0.018519
			7	East Harlem	0.000000

Figure 7: Top and bottom rows of the data-frame, which show highest and lowest frequency of coffee shops by neighborhood. The Financial District has the highest at 10%, and East Harlem has the lowest, at 0%.

While this information may be helpful as a first step, not all neighborhoods are the same size. It could also be important to look at which neighborhoods have the highest number of coffee shops per person. In this case, similar analysis to Part 1 and 2 was constructed, however instead of calculating a mean value when grouping by neighborhood, the new dataset calculated a sum, which is the total of all coffee shops found by Foursquare in each neighborhood. This data of number of coffee shops per neighborhood was merged with data of population per neighborhood in order to calculate a third column, Population per coffee shop. This is essentially providing a measurement of the number of people per coffee shop in each neighborhood. As the two datasets referred to the neighborhoods in Manhattan differently, it was important to combine similar neighborhoods, based on the name itself ("Soho" versus "SoHo" or "Gramercy" versus "Gramercy Park") as well as based on location ("Clinton" is the same geographic location as "Hell's Kitchen"). In these cases, the neighborhood in one dataset was replaced to match the second dataset, so that they would merge successfully.

After cleaning the data, it was possible to collect the three neighborhoods with the most people per coffee shop, and the three neighborhoods with the least people per coffee shop. In this case, the folium package was once again used to map the neighborhoods and color code based this time on number of people per coffee shop. Red signified low numbers of people per coffee shop, and green signaled high numbers of people per coffee shop.

	Neighborhood	Coffee Shop	Population	Population per shop
1	Central Harlem	1.0	335109.0	335109.000000
21	Midtown	4.0	391371.0	97842.750000
31	Washington Heights	2.0	158318.0	79159.000000
4	Civic Center	7.0	382654.0	54664.857143
3	Chinatown	2.0	100000.0	50000.000000
15	Inwood	1.0	46746.0	46746.000000
30	Upper West Side	6.0	209084.0	34847.333333
7	East Village	3.0	62832.0	20944.000000

Figure 8: Top rows of the data-frame to incorporate population data, with number of coffee shops, population, and ratio of population per shop. Central Harlem has the highest ratio, followed by Midtown.

From Part 2, a list of neighborhoods where coffee shops are the most/least frequent venues was collected. And from Part 3, a list of neighborhoods where there is a high/low ratio of people per coffee shop was found. In addition to the current state of a neighborhood, it may be useful to also take into account neighborhoods with highest growth from year to year. Neighborhoods with highest growth could generate larger revenues in the future. In this case, real estate transaction data was collected from PropertyShark to show which neighborhoods saw the most growth in 2020. Since this dataset had both positive and negative values as percentages, it was necessary to remove the "%" symbols and convert to float before sorting the data. After cleaning the data, the top 3 neighborhoods with highest percent growth and bottom 3 neighborhoods with lowest growth were collected.

	Rank	Borough	Neighborhood	Median Sale Price 2020	Y-o-Y Change in Median Sale Price	No of Transactions 2020	Y-o-Y Change in No of Transactions
2	3	Manhattan	Little Italy	$2,750,000.0	4%	35	25
6	7	Manhattan	Central Midtown	$1,790,000.0	44%	225	3
23	24	Manhattan	Financial District	$1,100,000.0	10%	212	-18
36	34	Manhattan	Sutton Place	$900,000.0	-16%	160	-20
20	21	Manhattan	Greenwich Village	$1,136,000.0	-19%	368	-24
12	13	Manhattan	Upper West Side	$1,303,000.0	18%	1429	-25
4	5	Manhattan	Hudson Square	$2,100,000.0	-15%	78	-26

Figure 9: Top rows of the data-frame to incorporate real estate transaction data. The final column, "Y-o-Y Change in No of Transactions" was used in this analysis. The only two neighborhoods with positive growth are Little Italy and Central Midtown.

4. Results

From this project, two interactive maps of the neighborhoods in Manhattan have been created. One shows frequency of coffee shops by neighborhood and the other shows a ranking of people per coffee shop. Both can be useful in the location strategy for an upcoming coffee shop. In the first map, the neighborhood markers are filled based on a color scale from red to yellow to green, where green is low frequency of coffee shops, and red is high frequency.

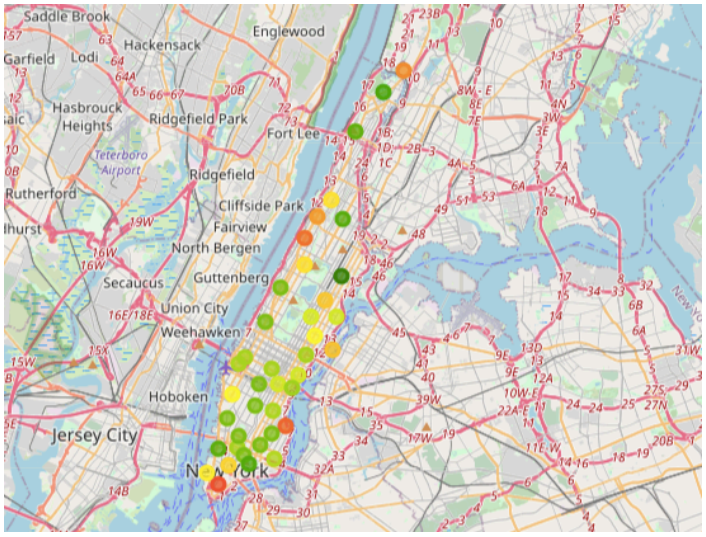

Figure 10: Image of the map of Manhattan, the markers showing each neighborhood color-coded based on frequency of coffee shops as venues in Foursquare data.

In the second map, the neighborhood markers are filled based on a color scale from red to yellow to green, where red signified low numbers of people per coffee shop, and green signaled high numbers of people per coffee shop.

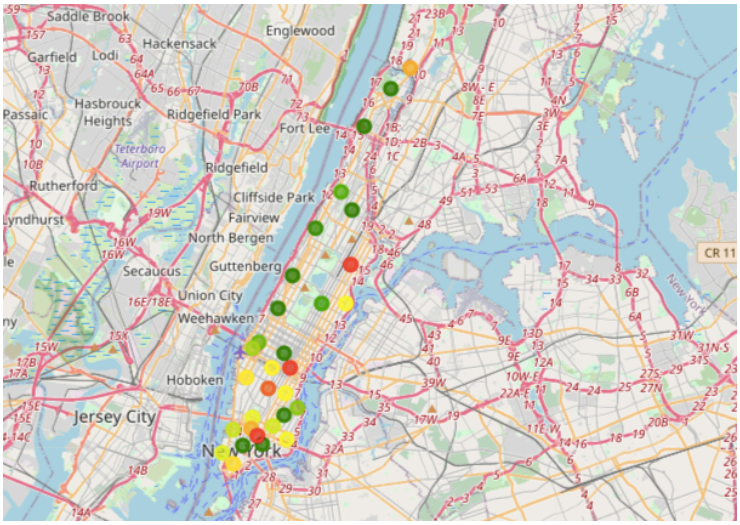

Figure 11: Image of the map of Manhattan, the markers showing each neighborhood color-coded based on number of people living in the neighborhood per coffee shop as found in 2012 census data.

Based on the analyses in this project, it is possible to ascertain neighborhood ranking based on a few key factors. The first is based on most common venue type. The neighborhoods where coffee shops are the most common venue type are, in alphabetic order: Carnegie Hill, Chelsea, Civic Center, Financial District, Manhattan Valley, Manhattanville, Marble Hill, and Morningside Heights. Out of these, the neighborhoods with the highest frequency of coffee shops are the Financial District, Stuyvesant Town, and Morningside Heights. Conversely, the neighborhoods with the lowest frequency of coffee shops are Chinatown, Inwood, and East Harlem.

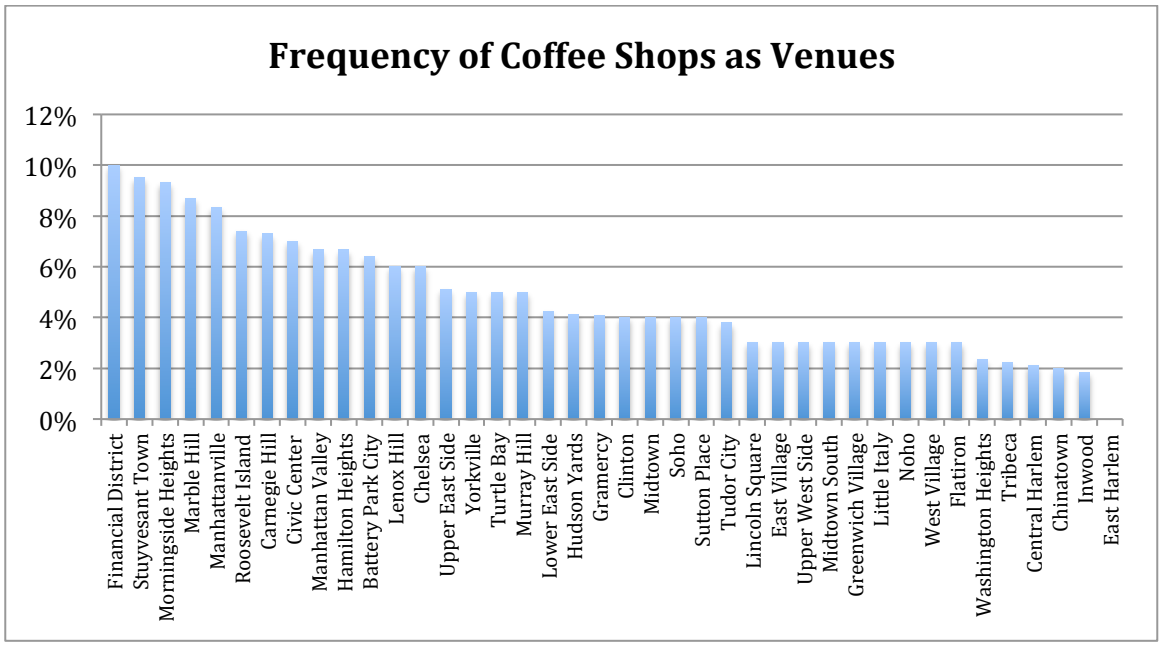

Figure 12: Graph showing the frequency of coffee shops as venues in Foursquare data per neighborhood in Manhattan. The Financial District showed highest frequency, and East Harlem showed the lowest.

The second ranking strategy is by population per coffee shop. This is a ratio of the population estimate of a neighborhood divided by the number of coffee shops returned by the Foursquare exploration. Based on the 2012 census data along with the Foursquare analysis, the neighborhoods with the most people per coffee shop are Central Harlem, Midtown, and Washington Heights. The neighborhoods with the least people per coffee shop are Murray Hill, Little Italy, and East Harlem.

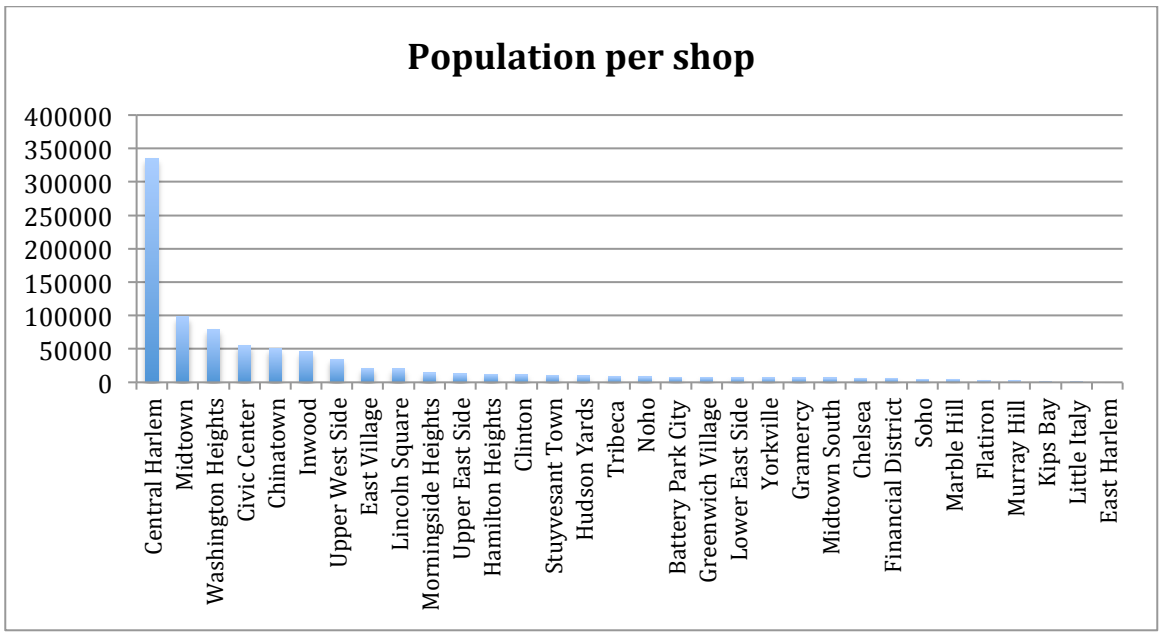

Figure 13: Graph showing the ratio of population to coffee shops in each neighborhood in Manhattan, using 2012 census data. Central Harlem and Midtown had the highest population per coffee shop.

The final ranking strategy incorporated into this project is neighborhood growth. Based on year-on-year growth in number of transactions from 2019 to 2020, the neighborhoods with the highest growth in 2020 are Little Italy, Central Midtown, and the Financial District. The neighborhoods with the largest drop in transactions are Chinatown, Lower East Side, and Hudson Yards, the lowest of which comes in at 85% drop in transactions.

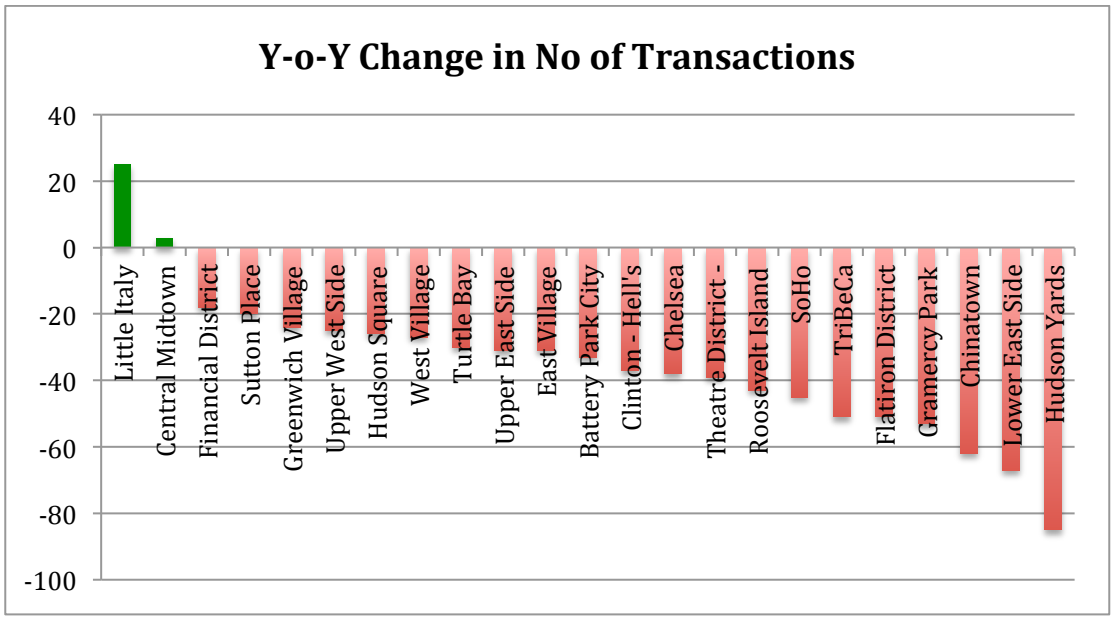

Figure 14: Graph showing real estate transaction growth per neighborhood in Manhattan, using Property Shark data from 2019 to 2020. Little Italy and Central Midtown are the only neighborhoods experiencing growth in 2020, compared to 2019

5. Discussion

When considering neighborhoods to open a new coffee shop, the ideal scenario would be to find a neighborhood with fewer coffee shops, a lot of people, and high potential (high growth). Looking at each parameter separately, it is clear that there is no “perfect” choice that satisfies each parameter equally. The neighborhoods with the highest growth in transactions in 2020 are Little Italy, Central Midtown, and Financial District. The neighborhoods with the lowest frequency of coffee shops are Chinatown, Inwood, and East Harlem. The neighborhoods with the most people per coffee shop are Central Harlem, Midtown, and Washington Heights.

Any of these analyses could be used to assist in decision-making. While there are no clear winners, since Midtown shows up in the top three neighborhoods in two out of three assessments, this could be an appropriate choice. It is also possible to run an analysis to optimize all three categories, which highlights Midtown as well as the Financial District as interesting options.

One aspect that is not addressed in this project is foot traffic in Manhattan. As New York City is a main tourist and commuter destination, some neighborhoods in the city experience an extremely high influx of people for business or vacation that do not live in the city and would not be counted in the census information. For future work, finding a dataset to quantify foot traffic in each neighborhood could add significantly to this project.

6. Conclusion

This project explored how Foursquare venue data can be harnessed along with other forms of location data to make informed business decisions. In this case, Manhattan was highlighted in an analysis of coffee shop locations throughout each neighborhood. Two interactive maps of the neighborhoods in Manhattan were created to show the frequency of coffee shops by neighborhood and a color map of the number of people per coffee shop living in each neighborhood. The top neighborhoods for opening a new coffee shop were identified using three key parameters: frequency of coffee shops, population, and growth rate. While there was no clearly superior neighborhood to optimize all three measurements, the most indicated neighborhood was Midtown, followed by the Financial District. This analysis of venue and location data could be used to determine the best neighborhood for opening a coffee shop.