“A cup of coffee, please!”

— A Data-Oriented Guide to Coffee Establishment in Manhattan

Capstone Project Report, Nov 2019

Data

Data Collection

To obtain the data needed for further analysis, I first use the ‘search’ endpoint in the API of Foursquare to retrieve a list of coffee establishments in Manhattan. To facilitate the search, I also borrow the geographical data of New York from the Week 3 module of this class so that the search can be conducted around each neighborhoods in Manhattan. This dataset also allow me to assign each coffee establishment to a nearby neighborhood. Afterward, the ‘detail’ endpoint will be used to collect information on the popularity and quality of each coffee establishments. The detailed procedure of the data collection is summarized in Figure 1.

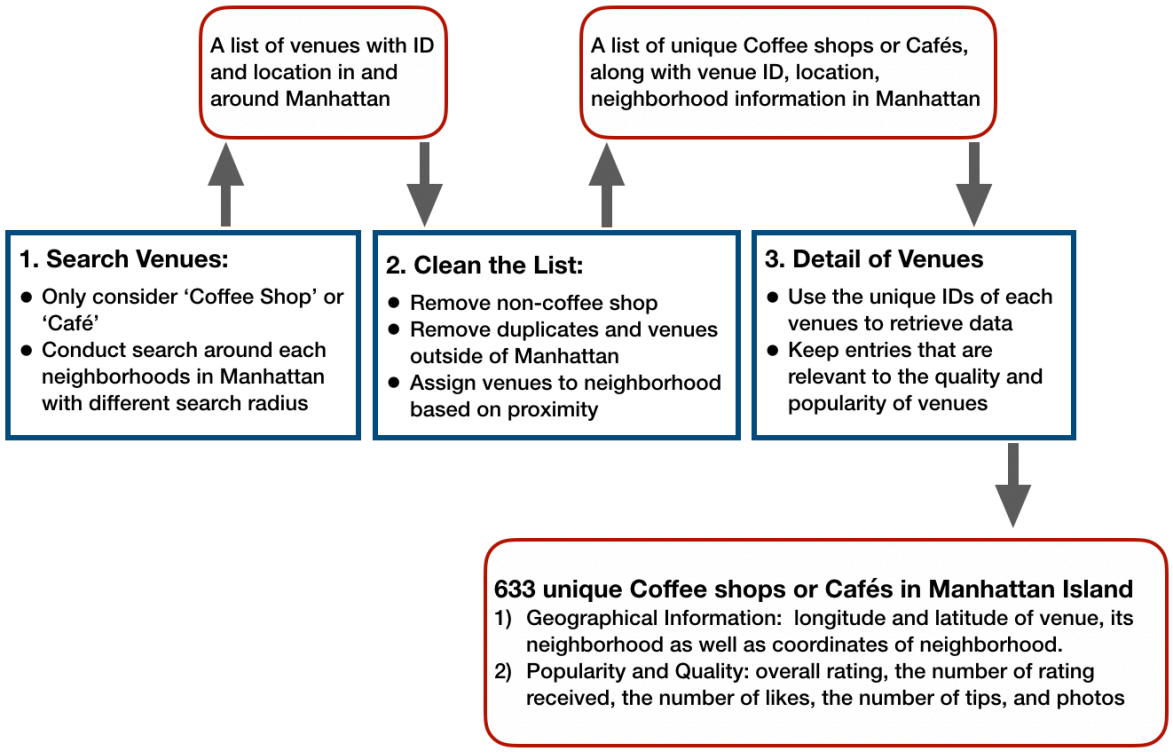

Figure 1: Procedure of Data Collection

Below I will focus on a few key considerations for collecting this data set.

1) The Definition of 'Coffee Establishment'

One fundamental question of this study is, how do we define 'Coffee Establishments'? In real life, people use different phrase to name a coffee establishment, including but not limited to 'Coffee Shop', 'Café', 'Coffee House' and 'Coffee Bar'. In the Foursquare database, each venue belongs to a particular category, which is chosen from a list of venue categories[1] maintained and updated by Foursquare.

Judging from this list, it is plausible that Foursquare would categorize a coffee establishment as either 'Coffee Shop' or 'Café'. Unfortunately, it is not clear how Foursquare defines the difference between these two categories. If we draw from the common sense, we might argue that 'Coffee Shop' is used to refer shop that focuses on coffee-related drinks along with a small selection of snacks, while 'Café' usually provides guests with a more extensive menu. However, in real life, these two words are often used interchangeable at the discretion of the shop owner as well as the locals.

This study defines a 'Coffee Establishment' as a venue that is either categorized as 'Coffee Shop' or 'Café' by Foursquare. Further analysis will be conducted to determine whether these names would impact the popularity or the quality of the venues.

Admittedly, 'Coffee Shop' or 'Café' are not the only places for purchasing coffee. Bakery, bagel shop, small restaurant or even convenience store can all be reasonable choice for a good cup of coffee. Therefore, by restricting the definition of 'Coffee Establishment' to 'Coffee Shop' and 'Café', I may inevitably limit the scope of this study. Nonetheless, I would also argue that this choice is both reasonable and practical. First, if the venues from categories other than 'Coffee Shop' or 'Café' are included in the study, it can be very difficult or even impossible to determine whether the main features of these venues could actually reflect the popularity and quality of the coffee served in these venues. For instance, it is quite plausible that a bagel store can attract a large number of customers even if its coffee is of mediocre quality. Thus, adding other categories to this study may actually hurt its quality. Second, due to the limitation on the calls of premium endpoints, it is also not practical to expand my definition of 'coffee establishment' as doing so

[1] See: https://developer.foursquare.com/docs/resources/categories

would significantly increase the number of venues I need to examine with premium endpoints.

2) Strategy for retrieving a complete (or nearly complete) list of coffee establishments in Manhattan?

While a call to the 'search' endpoint in Foursquare API can return a list of venues of specified categories ('Coffee Shop' or 'Café') around any location, this call only returns at most 50 venues within a given radius. Therefore, if the goal is to obtain a complete (or more practically, a nearly complete) list of coffee establishments in Manhattan, multiple searches centered at various locations on Manhattan are needed. It is also necessary to specify the radius of the search so that our searches could cover the whole Manhattan area. The the following strategy is then developed for this purpose:

- I borrowed the geographical data of New York from the Week 3 module of this class to obtain the latitudes and longitudes of 40 neighborhoods in Manhattan.
- 40 search requests were sent to Foursquare API. Each search request is conducted around the locations of one of the neighborhoods.
- The radius of each search was proportional to the distance from the location of the corresponding neighborhood to the location of its nearest neighborhood. Therefore, the search radiuses around the more isolated neighborhoods are generally larger.
- The radiuses of searches are further adjusted to ensure that all the 40 searches together cover every inch of the Manhattan Island. The centers and radiuses of all searches are displayed in Figure 2.

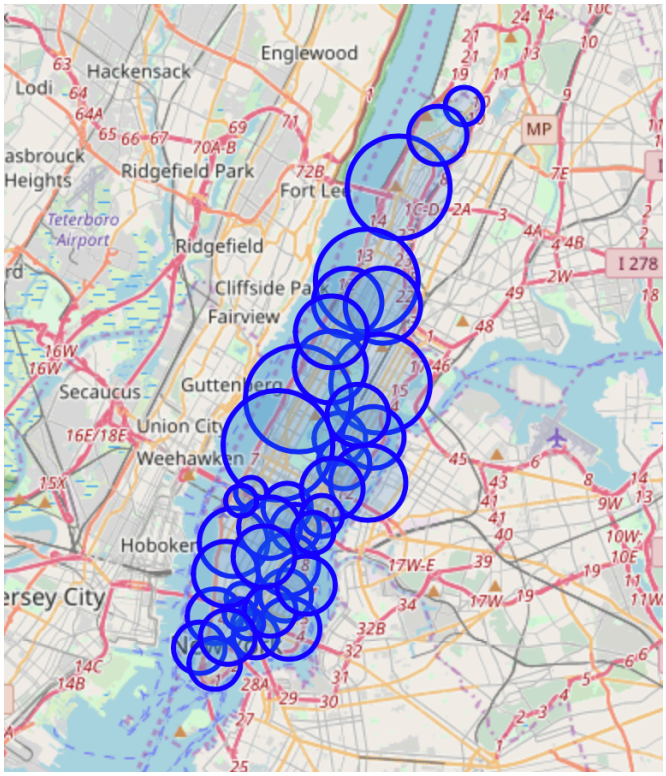

Figure 2: Centers and Radius of All Search Calls in Manhattan

As these searches cover all Manhattan Island, and the number of venues return from each search was smaller than the limit of 50, there is good reason to believe that this strategy produced a relatively complete list of ‘Coffee Shop’ and ‘Café’ in Manhattan.

Initially, each venue was assigned to the neighborhood that serves as the center in the corresponding search call. However, this design also produced significant number of duplicated venues from different search calls. Some searches also returned venues located outside of Manhattan. To future clean up the data, the unique ID of each venue was used to remove duplicates in this list and each unique venue was assigned to the nearest neighborhood. Venues located outside of Manhattan were also removed. Eventually, a list of 633 unique ‘Coffee Shop’ and ‘Café’ in Manhattan was obtained. See Figure 3 for their locations.

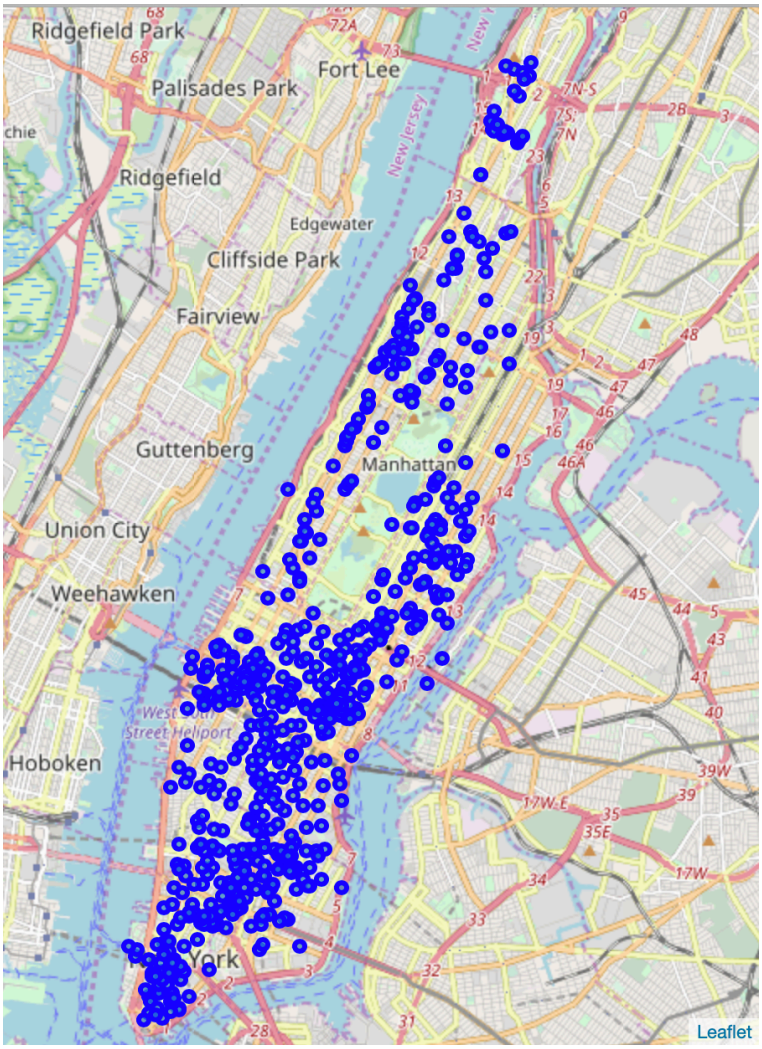

Figure 3: Coffee Shops and Cafés in Manhattan

3) Information collected for each coffee establishment

Given a list of coffee establishments along with their unique ids, it is relatively simple to retrieve the details information using the premium endpoint ‘detail’. To study the popularity and quality of these coffee establishments, the following features were recorded along side with the geographical information:

Feature Name	Explanation of Feture
Rating	Numerical rating of the venue (0 through 10)
Ratings Count	Number of ratings the venue has received
Likes Count	The count of Foursquare users who have liked this venue
Tip Count	Number of tips for this venue
Photos Count	Number of photos for this venue
Price Tier	Price tier of the value, from 1 (least pricey) to 4 (most pricey).

Preliminary Analysis

This section will include a few plots on the data set which can give us some basic insight on the dataset we have.

1) Popularity vs Quality, ‘Coffee Shop’ vs ‘Café’

Figure 4 displays the relationship between the rating, a measure of quality, and the number of likes, a measure of popularity. Different colors are use to distinguish ‘Coffee Shop’ and ‘Café’.

Based on this plot, while a high number of likes does not always translates into

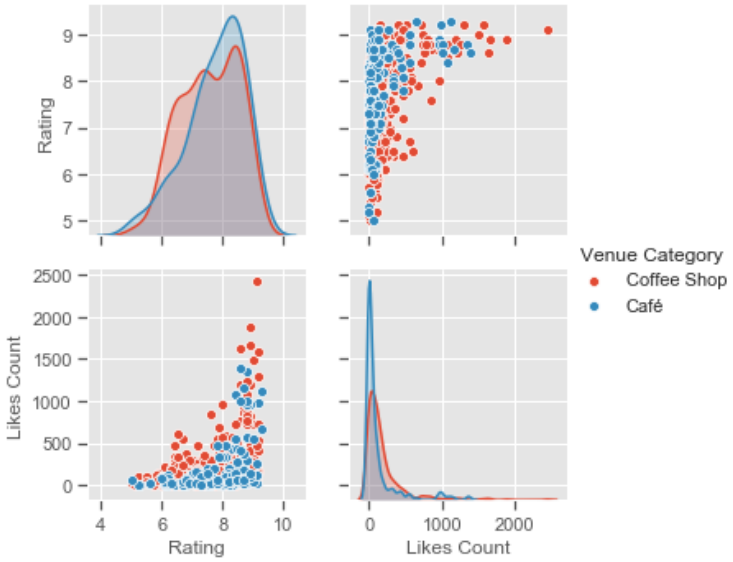

Figure 4: Popularity vs Quality

good rating, coffee establishments with likes count greater than 1000 all have relatively high ratings (greater than 8) . That is to say, only the coffee establishments with really good rating have a chance of gain tremendous popularity. And for most of the coffee establishment, the relationship between popularity and quality might be weak.

As for the difference between ‘Coffee Shop’ and ‘Café’, it seems that, while venues categorized as ‘Café’ tend to have slightly higher rating, venues categorized as ‘Coffee Shop’ are more likely to receive more likes. It is also worthy to note that, all the venues with likes count greater than 1500 are categorized as ‘Coffee Shop’. This is an interesting phenomenon worthy further investigation.

In Figure 5, the relationships between Tip Count, Likes Count, Ratings Count and Photos Count are displayed in the pairwise scatter plots. It is quite clear that all these counts are strongly positively correlated. And it is safe to assume that these features all reflect the popularity of the corresponding venues.

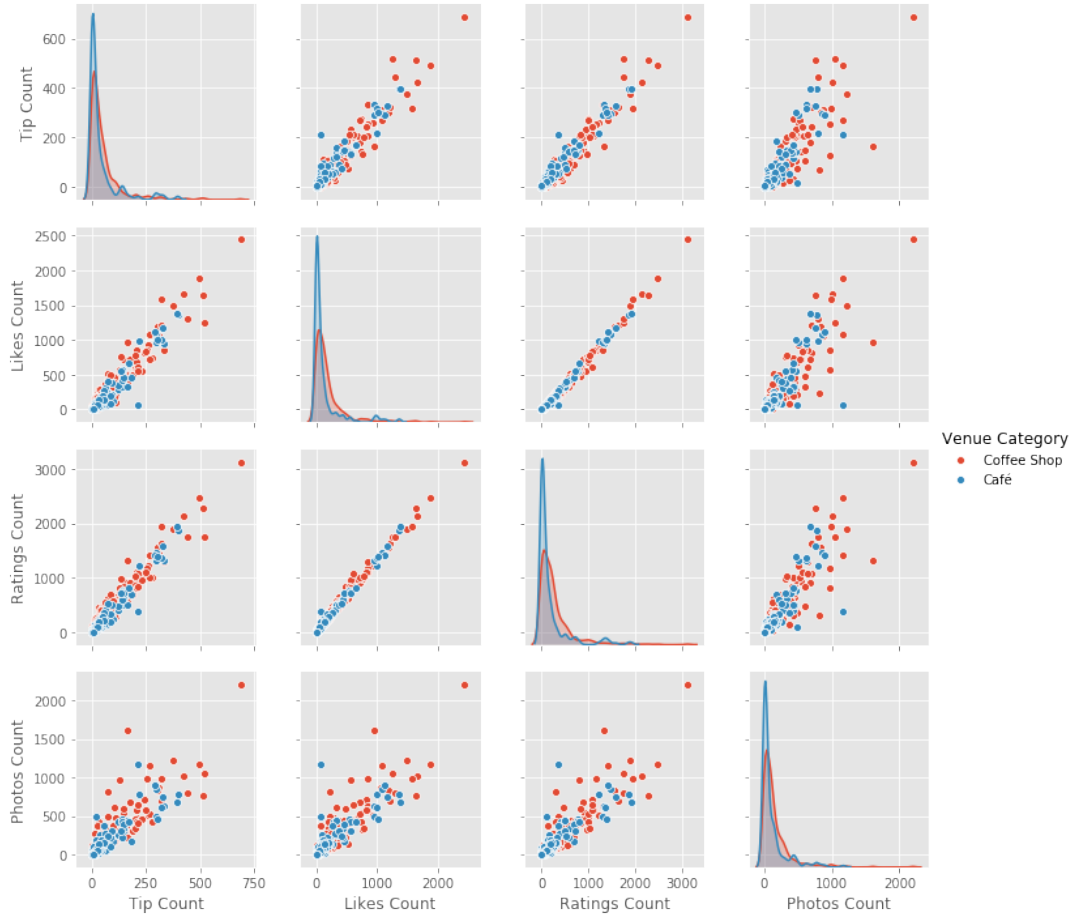

Figure 5: Various Measures of Popularity

2) Effect of Price Tier

Coffee establishments are generally cheap. In Manhattan, 83% coffee establishments belongs to price tier 1, the least expensive tier. Around 16% coffee establishments falls into price tier 2. Only 4 coffee establishment belongs to tier 3. The distribution of rating and the number of likes for different tiers are summarized in Figure 6.

Based on these figures, it can be conclude that, on average, coffee establishments with higher price tiers tend to offer better quality coffee and also attract larger popularity. Nonetheless, there are plenty of good choices among lower-price tier coffee establishments. In fact, the coffee shop with the highest number of likes (near 2500) is a price tier 1 establishment. The price is by no means a barrier for enjoying good coffee!

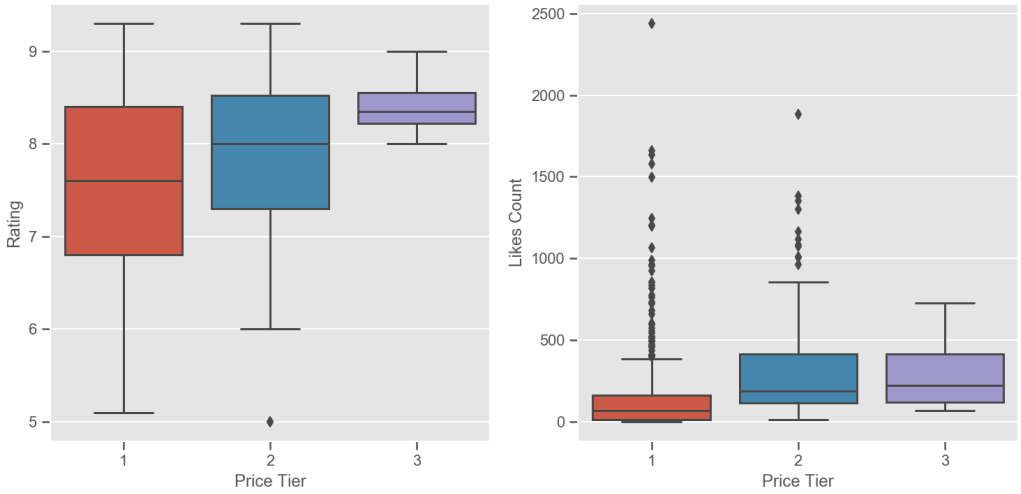

Figure 6: Effect of Price

3) Starbucks vs Others

The Starbucks certainly dominates the coffee market. Based on the list I composed, 157 out of 633 coffee establishments in Manhattan are Starbucks. Is there any major difference between Starbucks and other coffee establishments in term of the quality and popularity? Check Figure 7 to find out.

Unfortunately, Starbucks' average rating and number of likes all fall behind the averages of other coffee establishments. If you are looking for best coffee in Manhattan or looking for the most popular coffee shops, ignore Starbucks. However, it must be admitted that, the variation among Starbucks shops is considerably lower than the variation among

other coffee establishments. So, if you do not like any surprise (good or bad), Starbucks could be the right choice.

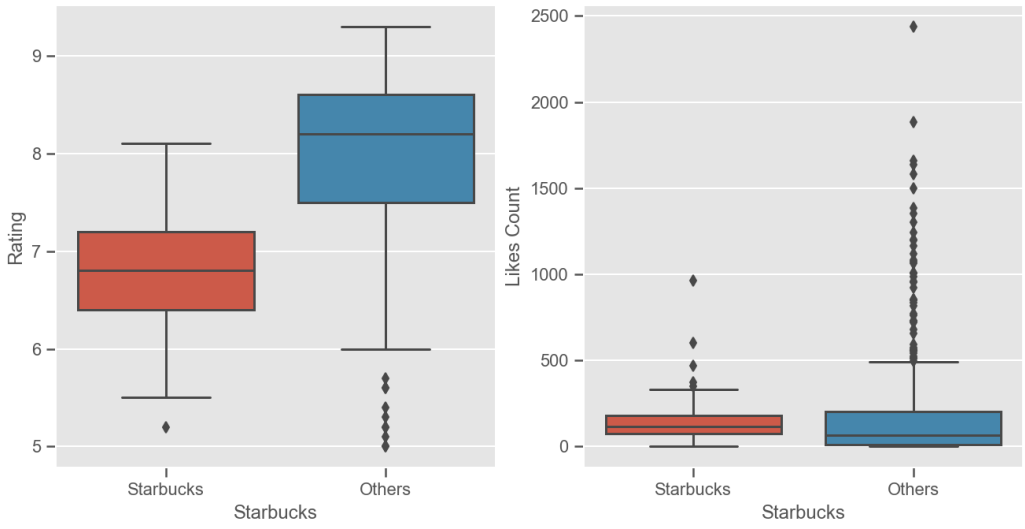

Figure 7: Starbucks vs Others

4) Battle between Neighborhoods

So what about the neighborhoods? Can we expect different experience in coffee establishments at different neighborhoods of Manhattan? Figure 8 shown the average of rating and number of likes in different neighborhoods of Manhattan (the sizes of circle markers are proportional to the average ratings or number of likes).

The maps in Figure 8 suggested that, while the coffee establishments in lower Manhattan (notably in Midtown South, Soho, West Village and Chelsea) tend to be much more popular than the coffee establishments in other neighborhoods, the quality of coffee (as measured by the average ratings) do not differ much between neighborhoods. Thus, if you are going to post a few selfies of you and your friends in coffee shop to Instagram, be sure to visit lower Manhattan for coffee. But if your only concern is to get a good cup of coffee, then it does not matter much on which neighborhoods you are currently visit.

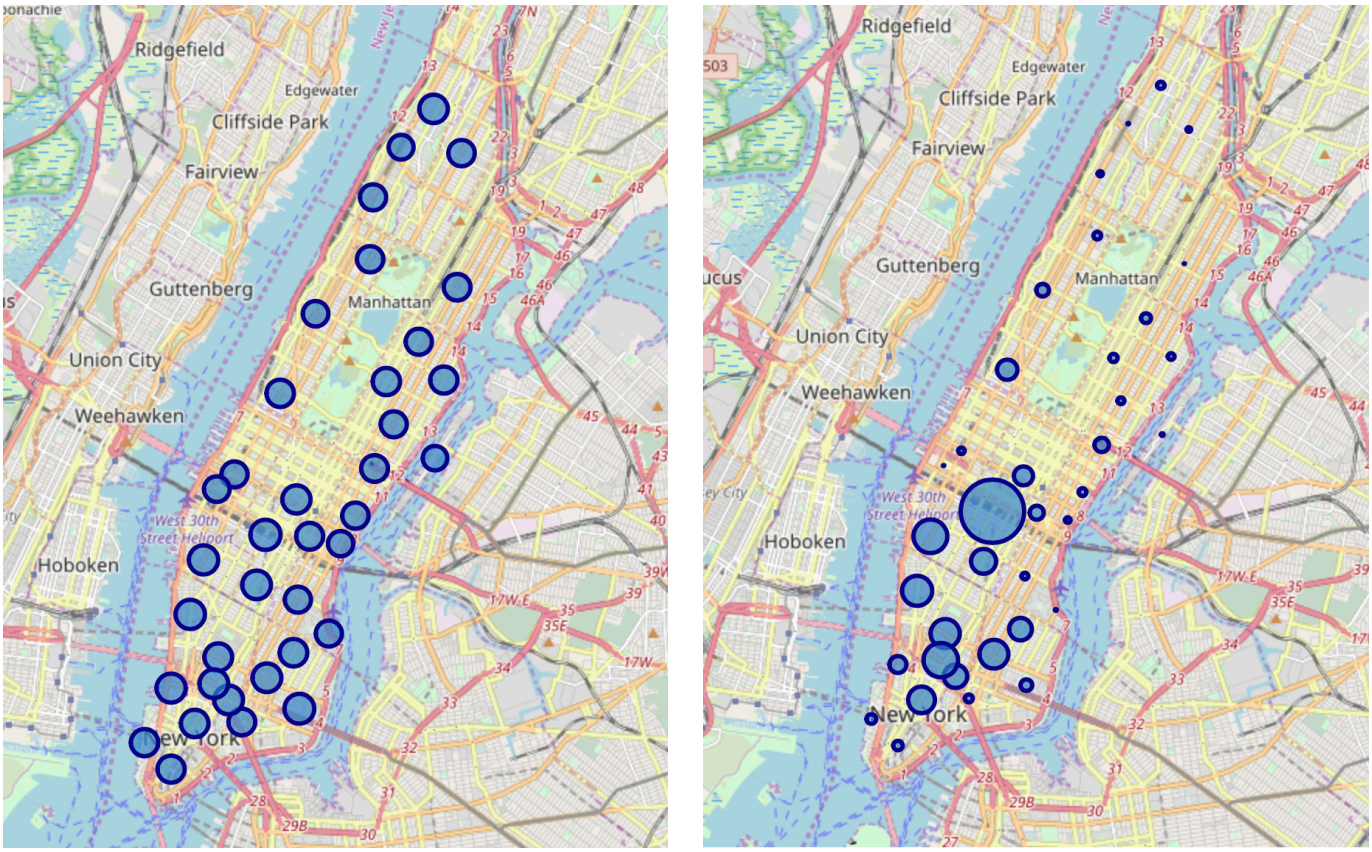

Figure 8: Average Ratings (Left) and Average Number of Likes (Right) of Coffee Establishments in Different Neighborhoods of Manhattan.

Again, many thanks for your time for reviewing my work and I would greatly appreciate any comments you might give!